DiHypeSC: Data-Informed Hyperparameter Adaptation for Self-Expressive Subspace Clustering

Anonymous authors
Paper under double-blind review

Abstract

Many subspace clustering methods require careful balancing of self-expressive reconstruction and structural regularization. Improper weighting can produce either dense cross-subspace connections or overly sparse and disconnected representations, both of which degrade clustering performance. However, the corresponding trade-off parameter is often selected through costly searches using labeled validation data, and its optimal value depends strongly on the dataset, noise level, representation, and regularizer. Label-free parameter selection remains challenging because deep clustering is stochastic and non-convex, while internal validation measures may not align with the ground-truth partition. We propose DiHypeSC, a data-informed strategy derived from subspace-preserving conditions that relate regularization strength to cross-subspace correlations and reconstruction residuals. During training, DiHypeSC estimates a proxy for the subspace dimension and uses it to adapt the regularization weight to each dataset. When integrated into two state-of-the-art deep subspace clustering methods, DiHypeSC provides a unified, dimension-aware weighting principle for both elastic-net and block-diagonal regularization. Even when using only a single hyperparameter configuration, DiHypeSC often achieves better results than conventional hyperparameter searches.

1 Introduction

Clustering partitions unlabeled data into groups with high intra-cluster similarity and low inter-cluster similarity. For high-dimensional data, subspace clustering assumes that clusters have low intrinsic dimensionality and are represented by a union of linear or affine subspaces. This model is widely used for tasks, such as motion segmentation (Vidal et al., 2003; Liu et al., 2010; Rao et al., 2010; Lauer & Schnörr, 2009; Lochman et al., 2024), image clustering and representation learning (Peng et al., 2016; 2017; Abavisani et al., 2020), and gene-expression analysis (McWilliams & Montana, 2014; Griffith et al., 2009).

While there are many paradigms for subspace clustering (Parsons et al., 2004; Kriegel et al., 2009; Vidal, 2011), the subfield employing spectral-clustering-based techniques is especially popular (Meng et al., 2025). These methods often exploit *self-expression*, whereby each sample is best represented by samples sharing its own subspace. This representation is called *subspace-preserving* if there are no cross-subspace connections. Under suitable conditions, this property is guaranteed theoretically (Li et al., 2018; You et al., 2016; Elhamifar & Vidal, 2013).

Self-expression reconstructs each sample from others sharing its subspace. However, simply minimizing the corresponding reconstruction error alone may yield cross-subspace connections, particularly under noise, thereby violating subspace preservation and degrading clustering. The regularization term imposes structural priors, with a weighting parameter β balancing reconstruction fidelity and regularization. A small β may produce dense cross-subspace connections, whereas a large β may yield overly sparse or degenerate representations that lose important within-subspace connections. Thus, selecting an appropriate value for β is essential for subspace preservation.

Despite variation across datasets and methods, the authors of prior methods provide only default values for their specific datasets (You et al., 2016; Zhang et al., 2021) or rely on extensive sweeps across a wide range of the hyperparameter (Yin et al., 2016; Meng et al., 2025). In the latter case, the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

default values were selected based on the evaluation using labeled validation data together with an external metric, such as clustering accuracy (Yang et al., 2010) or Normalized Mutual Information (NMI) (Strehl & Ghosh, 2002). Such a strategy is, however, infeasible in real-world use cases, as labels are typically unavailable. Instead, label-free internal indices (Ikotun et al., 2025), such as the silhouette score (Rousseeuw, 1987), could be used as substitutes.

Regardless of the metric, each candidate evaluation requires rerunning the full clustering pipeline, which is costly for deep models. Moreover, internal indices need not correspond to label information, and optimal weights are dataset-dependent; thus, the analysis would need to be repeated for any new dataset. Although recent work (Fan et al., 2020) derives weights from intrinsic data properties, it remains unclear how to apply such strategies to deep subspace clustering and adapt them to different regularizers. We address this problem by proposing **D**ata-**I**nformed **Hype**rparameter Learning for **S**ubspace **C**lustering (DiHypeSC), a framework for adaptive regularization to suppress cross-subspace connections, and evaluating its sensitivity, efficiency, and clustering performance. Our main contributions are:

- We analyze the sensitivity of deep subspace clustering to regularization weights across datasets and regularizers.
- We adapt the theoretical insights of Soltanolkotabi et al. (2014b) to data-informed weight selection for elastic-net regularization in the ℓ_1-dominant regime.
- We derive a sufficient condition for subspace preservation under block-diagonal regularization with a locally fixed spectral embedding and use it to select the regularization weight.

2 Related Work

In this section, we review deep subspace clustering methods for learning noise-robust self-expressive representations, as well as hyperparameter tuning approaches for unsupervised learning.

Subspace clustering The curse of dimensionality complicates clustering in high-dimensional spaces (Parsons et al., 2004). Subspace clustering addresses this by identifying clusters within low-dimensional structures (Parsons et al., 2004; Elhamifar & Vidal, 2013). Existing approaches (Kriegel et al., 2009; Vidal, 2011) include grid-based CLIQUE (Agrawal et al., 1998) and self-expression-based Sparse Subspace Clustering (Elhamifar & Vidal, 2009) with the latter underpinning many modern methods.

Self-Expressive Subspace Clustering exploits self-expression by representing each sample as a linear or affine combination of samples from the same subspace (Elhamifar & Vidal, 2009; Vidal, 2011). The resulting subspace-preserving representations reveal the underlying subspaces and their clustering (Elhamifar & Vidal, 2013). Different regularizers impose distinct structures: the ℓ_1-norm promotes sparsity (Elhamifar & Vidal, 2013), the nuclear norm favors low rank (Liu et al., 2013), and elastic-net (You et al., 2016), block-diagonal (Meng et al., 2025), and graph-based regularization (Yin et al., 2016) promote cluster-separable affinities. Recent work also projects the coefficient matrix onto the set of doubly stochastic matrices (Ding et al., 2022). The resulting affinity matrix is typically partitioned using spectral clustering (von Luxburg, 2007).

Deep Self-Expressive Subspace Clustering Deep clustering uses neural network architectures to learn clustering-oriented representations (Ren et al., 2025), a trend also adopted in subspace clustering despite ongoing criticism (Haeffele et al., 2021). DSC-Net (Ji et al., 2017) introduced a self-expressive layer within an autoencoder, whereas SENet (Self-Expressive Network) (Zhang et al., 2021) uses two multilayer perceptrons to generate coefficients for unseen samples and enable mini-batch training. DCSSC (de Souza Oliveira et al., 2025) combines neighborhood graphs with contrastive variational learning, PRO-DSC (Principled fRamewOrk for Deep Subspace Clustering) (Meng et al., 2025) prevents representation collapse while promoting block-diagonal coefficients, and UIDSSC (Huang et al., 2026) utilizes a self-supervised framework. All of these methods must balance representation learning, self-expression, and regularization. However, they generally lack a proper heuristic for determining this balance, which restricts their practical applicability. To address this limitation, we propose DiHypeSC, which balances this trade-off for elastic-net and block-diagonal regularization in a data-informed manner.

Hyperparameter Optimization in Clustering Clustering is sensitive to hyperparameter choices (Fan et al., 2020; Ditton et al., 2022). AutoML automates machine learning tasks (Barbudo et al., 2023), including hyperparameter optimization (Bergstra & Bengio, 2012; Feurer & Hutter, 2019). Although primarily focused on supervised learning, it also supports clustering (Poulakis et al., 2024; da Silva et al., 2024). Unlike DiHypeSC, these methods typically rely on task-related meta-features (Aljoud et al., 2025; 2026) or repeated clustering runs (Fan et al., 2020; Poulakis et al., 2020; Tschechlov et al., 2021; Ditton et al., 2022).

Algorithm-specific strategies estimate, for example, the number of clusters k (Pelleg & Moore, 2000; Kalogeratos & Likas, 2012) for k-means (Lloyd, 1982) or the neighborhood radius ε (Nir et al., 2025; Beer et al., 2025) for DBSCAN (Ester et al., 1996). In self-expressive subspace clustering, CESM (Hashemi & Vikalo, 2018) tunes a smoothing parameter α balancing historical and current data in evolutionary settings. For the regularization parameter β targeted by DiHypeSC, existing tuning guidance (Soltanolkotabi et al., 2014b) is limited to the ℓ_1-norm. DiHypeSC overcomes this limitation, which we will go more in-depth on in the Data-informed Regularization section.

3 Theoretical Background

3.1 Preliminaries

Throughout this paper, $\ln$ denotes the natural logarithm, D the ambient feature dimension, $\mathcal{S}$ a subspace of dimension d, n the number of samples in $\mathcal{S}$, and N the total number of samples in the dataset. Then we represent the dataset as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D\times N}$, where $\mathbf{x}_i \in \mathbb{R}^D$ denotes the ith sample. Furthermore, $\mathbf{A} = \mathbf{X}_{-i} \in \mathbb{R}^{D\times(N-1)}$ denotes the matrix obtained by removing the ith column of $\mathbf{X}$, which is also refered to as a *dictionary* matrix. For a query sample $\mathbf{x}_i$, let $\mathcal{S}_i$ denote its underlying subspace and define $T = \{j \in [N] \setminus \{i\} : \mathbf{x}_j \in \mathcal{S}_i\}$. Its complement $T^c = ([N] \setminus \{i\}) \setminus T$ indexes the *off-support samples*, i.e., samples outside $\mathcal{S}_i$. Further, let d_i denote the dimension of $\mathcal{S}_i$ and N_i the number of samples in $\mathcal{S}_i$.

Self-expressive models aim to learn a coefficient matrix $\mathbf{C}$ by solving the following optimization problem with $\mathcal{R}(\mathbf{C})$ denoting a regularization term:

$$\min_{\mathbf{C}} \|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 + \beta\mathcal{R}(\mathbf{C}), \quad \text{s.t.} \quad \mathrm{Diag}(\mathbf{C}) = 0 \tag{1}$$

Definition 1 (Subspace-preserving property)**.** *A self-expressive representation $\mathbf{x}_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} \mathbf{x}_j c_{i,j}$ is* subspace-preserving *if $c_{i,j} = 0$ whenever $\mathbf{x}_j$ does not belong to the same subspace as $\mathbf{x}_i$.*

In this paper, we focus on the elastic net regularization and the block-diagonal regularizer:

Elastic-net regularization considering $\mathbf{c}$ as a column vector of $\mathbf{C}$, the elastic-net regularization is defined as: $\mathcal{R}(\mathbf{c}) = \lambda\|\mathbf{c}\|_1 + \frac{(1-\lambda)}{2}\|\mathbf{c}\|_2^2$

Block diagonal regularizer given the affinity matrix $\mathbf{Z} = \frac{1}{2}\left(|\mathbf{C}| + |\mathbf{C}|^\top\right)$, the block-diagonal regularizer is defined as $\mathcal{R}(\mathbf{C}) = \|\mathbf{Z}\|_\kappa$ where $\|\mathbf{Z}\|_\kappa$ denotes the sum of the k smallest eigenvalues of the graph Laplacian associated with the affinity matrix $\mathbf{Z}$.

Given the general structure of these methods, we are interested in determining an appropriate weighting parameter β such that the algorithm preferentially selects points belonging to the same cluster and sharing the same underlying subspace, rather than points from different subspaces, to linearly express one another. To facilitate the theoretical analysis, we introduce the following definitions:

Definition 2. *(Independent subspaces) A collection of subspaces $\{\mathcal{S}_\ell\}_{\ell_1=1}^n$ is said to be independent if $dim(\bigcup_{\ell=1}^n \mathcal{S}_\ell) = \sum_{\ell=1}^n dim(\mathcal{S}_\ell)$.*

Definition 3 (Additive isotropic Gaussian noise model)**.** *Let $\mathbf{x}_{observe} \in \mathbb{R}^n$ denote an observed data point and $\mathbf{x}_{true}$ its clean counterpart with $\|\mathbf{x}_{true}\|_2 = 1$. Assume that $\mathbf{x}_{observe} = \mathbf{x}_{true} + \mathbf{z}$, where $\mathbf{z}$ is independent of $\mathbf{x}_{true}$ and satisfies $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \frac{\sigma^2}{n}\mathbf{I}_n)$. Thus, $\mathbb{E}[\|\mathbf{z}\|_2^2] = \sigma^2$, and σ measures the root-mean-square noise magnitude relative to the unit-norm signal.*

Definition 4. *(Affinity condition) (Soltanolkotabi et al., 2014b) A subspace S_i fulfills the affinity condition, when $\max_{j:j\neq i} aff(S_i, S_j) \leq \frac{\kappa_0}{\ln N} - \sigma\sqrt{\frac{d_{max}}{(n_i+n_j)\ln N}}$ with n_i and n_j as number of samples on subspace S_i, S_j, σ denote the noise variance, κ_0 is some fixed numerical constant.*

Intuitively, the condition requires each subspace to be sufficiently separated from all others. The permitted affinity decreases as the noise level or subspace dimension increases.

Definition 5. *(Sampling condition) A subspace S_i fulfills the sampling condition, when $\rho_i = N_i/d_i \geq \rho^\star$ with $\rho^\star$ as another numerical constant. It denotes the minimum sample density for a subspace S_i to be identifiable and depends on the subspace dimension and the noise level.*

4 Data-informed Regularization

In this section, we present our algorithms for estimating the regularization strength at each training epoch, thereby enhancing the noise robustness of deep subspace clustering and mitigating overfitting. First, we introduce an adaptive strategy for selecting the weighting parameter β, which balances the self-expressive reconstruction loss and the elastic-net regularization of the self-expressive coefficients in SENet (Zhang et al., 2021). We then examine how to extend this strategy to the block-diagonal regularizer introduced by Lu et al. (2019) using PRO-DSC (Meng et al., 2025).

4.1 Algorithms

Consider a generic deep-learning-based subspace clustering method. Such a method typically consists of two modules: a *feature-learning module* that extracts clustering-friendly representations by mapping the original nonlinear features onto a union of subspaces (Ding et al., 2023), and a *coefficient-learning module* that uses the learned representations to obtain noise-robust self-expressive coefficients. Subspace clustering algorithms mainly differ in the design of their feature-learning modules, in how they learn self-expressive coefficients from the latent representations, or in the self-expressive coefficient generation mechanisms.

4.1.1 SENet: The Self-Expressive Network

SENet (Zhang et al., 2021) employs two multilayer perceptrons to generate query and key representations, denoted by $\mathbf{Q}$ and $\mathbf{K}$, respectively. Their pairwise inner products $\mathbf{Q}\mathbf{K}^\top$ are passed through a learnable soft-thresholding operator to produce sparse self-expressive coefficients. The network parameters and resulting coefficients are learned by minimizing the elastic-net-regularized objective in Eq. 3.1 using gradient-based optimization.

We propose selecting β according to how accurately $\mathbf{Q}$ can be linearly reconstructed from $\mathbf{K}$ at each epoch. Under the noiseless model, we solve

$$\widehat{\mathbf{C}} = \arg\min_{\mathbf{C}} \|\mathbf{C}\|_F^2 \quad \text{s.t.} \quad \mathbf{Q} = \mathbf{C}\mathbf{K}, \quad \operatorname{diag}(\mathbf{C}) = \mathbf{0}.$$

The zero-diagonal constraint prevents trivial self-representation. This problem is separable across queries. Let $\mathbf{q}_i$ be the ith query and $\mathbf{K}_{-i}$ the key matrix excluding its corresponding key. Then

$$\widehat{\mathbf{c}}_i = (\mathbf{K}_{-i}^\top)^\dagger \mathbf{q}_i^\top = \mathbf{K}_{-i}(\mathbf{K}_{-i}^\top \mathbf{K}_{-i})^{-1}\mathbf{q}_i^\top,$$

where † denotes the Moore–Penrose pseudoinverse, and the second equality holds when $\mathbf{K}_{-i}$ has full column rank.

Under substantial noise, β should balance suppressing noise-induced coefficients and preserving sufficient within-subspace connections. Since a generic point in a d-dimensional subspace typically requires d linearly independent samples for exact representation, we use d as a reference for the effective number of significant coefficients rather than as a strict lower bound.

In the following, we show that the ℓ_1-norm of each estimated coefficient vector $\widehat{\mathbf{c}}_i$ is bounded above and below by quantities proportional to $\sqrt{d}$. When these bounds are sufficiently tight, $\|\widehat{\mathbf{c}}_i\|_1$ can be used to estimate the subspace dimension d.

We first recall the following lemma from the supplementary material of Liu et al. (2013).

Lemma 1. *Let $\tau \geq 0$ be a noise tolerance, and fix a query sample $\mathbf{x}_i$. Define the common feasible set $\mathcal{F}_i(\tau) = \{\mathbf{c} \in \mathbb{R}^n : \|\mathbf{x}_i - \mathbf{X}_{-i}\mathbf{c}\|_2 \leq \tau\}$, and assume that $\mathcal{F}_i(\tau)$ is nonempty. Consider the constrained ℓ_1-minimization problem $\mathbf{c}_{\ell_1}^\star \in \arg\min_{\mathbf{c}\in\mathcal{F}_i(\tau)} \|\mathbf{c}\|_1$ the noiseless ℓ_2-minimization problem with $\tau = 0$*

$$\widehat{\mathbf{c}} \in \arg\min_{\mathbf{c}\in\mathcal{F}_i(\tau=0)} \|\mathbf{c}\|_2^2. \tag{2}$$

and the constrained elastic-net problem $\mathbf{c}^\star_{\text{elastic}} \in \arg\min_{\mathbf{c}\in\mathcal{F}_i(\tau)} \left(\lambda\|\mathbf{c}\|_1 + \frac{1-\lambda}{2}\|\mathbf{c}\|_2^2\right)$, $0 < \lambda < 1$. *Suppose that* $\mathbf{x}_i$ *lies in a subspace* S_i *of dimension* d_i *and that the sampling density* $\rho_i = N_i/d_i$ *of* S_i *satisfies* $\rho_i > \rho^\star = 16$. *It follows that,*

$$\frac{1}{10}\sqrt{\frac{d_i}{\ln \rho_i}} \le \|\mathbf{c}^\star_{\ell_1}\|_1 \le \|\widehat{\mathbf{c}}\|_1 \le 2\sqrt{d_i}. \tag{3}$$

and $\frac{\lambda}{10}\sqrt{\frac{d_i}{\ln \rho_i}} \le \lambda\|\mathbf{c}^\star_{\ell_1}\|_1 \le \lambda\|\mathbf{c}^\star_{\text{elastic}}\|_1 \le 2\lambda\sqrt{d_i} + \kappa_i$, *where* $\kappa_i = \frac{1-\lambda}{2\left(\sqrt{N_i/d_i}-2\right)^2} \le \frac{1-\lambda}{8}$.

The proof of the lemma is in the supplementary material, Section E. We have shown that the ℓ_1-norm of the closed-form minimum-ℓ_2-norm solution is bounded by quantities of order $\sqrt{d_i}$ and, when these bounds are tight, serves as a proxy for the subspace dimension. It remains to select β so that the representation is subspace-preserving. The following lemma provides a sufficient condition.

Lemma 2 (Off-support residual-correlation condition)**.** *Let* $\mathbf{A}$ *denote the dictionary matrix and assume that* $\beta > 0$ *and* $0 < \lambda \le 1$. *Define* $F(\mathbf{c}) = \frac{1}{2}\|\mathbf{x}_i - \mathbf{A}\mathbf{c}\|_2^2 + \beta\lambda\|\mathbf{c}\|_1 + \frac{\beta(1-\lambda)}{2}\|\mathbf{c}\|_2^2$ *Suppose that* $\mathbf{c}^\star \in \arg\min_{\mathbf{c}:\,\mathbf{c}_{T^c}=\mathbf{0}} F(\mathbf{c})$ *is an optimal solution of the restricted problem, and define its residual by* $\mathbf{r}^\star = \mathbf{x}_i - \mathbf{A}\mathbf{c}^\star$ *If* $\left\|\mathbf{A}_{T^c}^\top \mathbf{r}^\star\right\|_\infty < \beta\lambda$ *then* $\mathbf{c}^\star$ *is also an optimal solution of the unrestricted problem* $\min_{\mathbf{c}\in\mathbb{R}^N} F(\mathbf{c})$.

Intuitively, $\mathbf{A}_{T^c}^\top \mathbf{r}^\star$ measures how strongly the out-of-subspace samples correlate with the remaining reconstruction error after using only within-subspace samples. If $|\mathbf{x}_j^\top \mathbf{r}^\star| < \beta\lambda$, activating c_j cannot reduce the reconstruction loss enough to offset its regularization cost. A detailed proof of Lemma 2 is provided in Section F of the supplementary material. We next derive an upper bound on the off-support residual correlations, which $\beta\lambda$ must exceed to suppress all off-support coefficients.

Theorem 1 (Upper bound on correlations between off-subspace samples and reconstruction residuals)**.** *Let* $\mathbf{x}_i$ *lie in a subspace* S_i *of dimension* d_i*, and suppose that* S_i *satisfies the affinity and sampling conditions in Definitions 7 and 8, respectively, and that the noise level* σ *from Definition 3 is sufficiently small. Let* $\mathbf{c}^\star$ *solve the following elastic-net problem*

$$\min_{\mathbf{c}} \frac{1}{2}\|\mathbf{x}_i - \mathbf{A}\mathbf{c}\|_2^2 + \beta\left(\lambda\|\mathbf{c}\|_1 + \frac{1-\lambda}{2}\|\mathbf{c}\|_2^2\right),$$

and define $\mathbf{r}^\star = \mathbf{x}_i - \mathbf{A}\mathbf{c}^\star$*, there exists a constant* $m > 0$ *such that if* $\beta > \frac{m}{\lambda\sqrt{d_i}}$ *Then the optimal solution* $c^\star$ *obtained with these regularization strengths satisfies the subspace-preserving property, i.e.,* $\mathbf{c}_{T^c} = \mathbf{0}$ *Further let* $\mathbf{C}^\star$ *solve the restricted weighted* ℓ_1 *norm problem:*

$$F_{\mathbf{U}}(\mathbf{C}) = \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 + \beta\sum_{i,j} w_{ij}|c_{ij}|, \quad \operatorname{diag}(\mathbf{C}) = \mathbf{0}. \tag{4}$$

and $\mathbf{C}^\star \in \arg\min\{F_{\mathbf{U}}(\mathbf{C}) : \operatorname{diag}(\mathbf{C}) = \mathbf{0}\}$*, and define* $\mathbf{E}^\star = \mathbf{X} - \mathbf{X}\mathbf{C}^\star$. *Then, there exists a constant* m' *such that, if* $\beta > \frac{m'}{\min_{i\neq j}\omega_{i,j}\sqrt{d_i}}$ *the resulting optimal solution is subspace-preserving.*

Intuitively, sufficiently sampled and well-separated subspaces allow each point to be reconstructed using samples from its own subspace, while off-subspace samples correlate only weakly with the residual. Cross-subspace coefficients remain zero when these absolute correlations fall below $\beta\lambda$ for elastic net or βw_{ij} for weighted ℓ_1 regularization. Under the stated assumptions, the proposed bounds on off-subspace residual correlations express this threshold through the dimension-dependent scale $\sqrt{d_i}$. Choosing $\beta \propto 1/\lambda\sqrt{d_i}$ for elastic-net regularization or $\beta \propto 1/\min_{i\neq j}\omega_{i,j}\sqrt{d_i}$ compensates for its dependence on the subspace dimension.

Moreover, by Lemma 1, the ℓ_1 norm of the minimum-ℓ_2-norm solution in the noise-free case, i.e., $\|\widehat{\mathbf{c}}\|_1$ in Eq. 3 , serves as a proxy for $\sqrt{d_i}$. Combining these results, we estimate $\widehat{\beta}_i = \frac{a}{\lambda\left(\|\widehat{\mathbf{c}}\|_1+\varepsilon\right)}$ for elastic-net regularization, where $\varepsilon > 0$ ensures numerical stability and $a > 0$ is a constant determined empirically. The detailed algorithm for data-informed hyperparameter learning in SENet with elastic-net regularization is provided in Section C of the supplementary material.

4.1.2 DEEP SUBSPACE CLUSTERING WITH BLOCK-DIAGONAL REGULARIZATION (PRO-DSC)

PRO-DSC combines an encoder with an MCR^2-inspired coding-rate term (Yu et al., 2020) to promote diverse, non-collapsing representations and a Laplacian-based regularizer to strengthen within-subspace connectivity while suppressing cross-subspace connections (Meng et al., 2025). We focus on balancing the self-expressive loss and the block-diagonal regularizer, assuming that the weight of the self-expressive loss satisfies the non-collapse condition in Theorem 2 of Meng et al. (2025). We then establish the following lemma.

Lemma 3 (Subspace preservation for fixed Laplacian eigenvectors)**.** *Let* $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N] \in \mathbb{R}^{D\times N}$ *and* $\mathbf{C} \in \mathbb{R}^{N\times N}$ *with* $\operatorname{diag}(\mathbf{C}) = \mathbf{0}$*. Define* $\mathcal{A}(\mathbf{C}) = (|\mathbf{C}| + |\mathbf{C}|^\top)/2$ *and* $\mathcal{L}(\mathbf{C}) = \operatorname{Diag}(\mathcal{A}(\mathbf{C})\mathbf{1}) - \mathcal{A}(\mathbf{C})$*. Let* $\mathbf{U} \in \mathbb{R}^{N\times k}$ *contain the eigenvectors associated with the* k *smallest eigenvalues of a fixed Laplacian, with* $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$*. Denote its* i*th row by* $\mathbf{u}_i^\top$ *and set* $w_{ij} = \|\mathbf{u}_i - \mathbf{u}_j\|_2^2/2$*. For fixed* $\mathbf{U}$*, consider*

$$F_{\mathbf{U}}(\mathbf{C}) = \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 + \beta \sum_{i,j} w_{ij}|c_{ij}|, \qquad \operatorname{diag}(\mathbf{C}) = \mathbf{0}. \tag{5}$$

Let Ω *and* Ω^c *denote the off-block-diagonal within- and cross-subspace index pairs, respectively. Suppose* $\mathbf{C}^\star \in \arg\min\{F_{\mathbf{U}}(\mathbf{C}) : \operatorname{diag}(\mathbf{C}) = \mathbf{0}, \}$*, and define* $\mathbf{E}^\star = \mathbf{X} - \mathbf{X}\mathbf{C}^\star$*. If* $|[\mathbf{X}^\top\mathbf{E}^\star]_{ij}| < \beta w_{ij} = \frac{\beta}{2}\|\mathbf{u}_i - \mathbf{u}_j\|_2^2 \quad \forall (i,j) \in \Omega^c$ *then* $\mathbf{C}^\star$ *satisfies* $\mathbf{C}^\star_{\Omega^c} = \mathbf{0}$ *and is therefore subspace-preserving.*

Think of $\mathbf{u}_i$ as the spectral coordinates of sample i. The weight w_{ij} makes a connection more expensive when two samples are farther apart in these coordinates. Starting from the best reconstruction using only within-subspace connections, $|\mathbf{x}_i^\top \mathbf{e}_j^\star|$ measures the potential first-order improvement from introducing a cross-subspace coefficient c_{ij}. Its regularization cost per unit magnitude is βw_{ij}. The lemma requires this cost to exceed the potential improvement for every cross-subspace pair. Convexity then guarantees that allowing such connections cannot improve the objective, and the strict inequalities ensure that every optimum excludes them. The proof of the lemma is provided in Section G of the supplementary material.

Combining Lemma 3 and Theorem 1 motivates $\beta > m'/(\min_{i,j\in\Omega^c} \omega_{i,j}\sqrt{d})$, where $\min_{i,j\in\Omega^c} \omega_{i,j}$ is the minimum spectral distance between cross-subspace samples. Using $\|\widehat{\mathbf{c}}\|_1$ as a proxy for $\sqrt{d_i}$, we obtain $\widehat{\beta}_{\mathrm{BDR}} = b/(\min_{i,j\in\Omega^c} \omega_{i,j}\|\widehat{\mathbf{c}}\|_1) > m'/(\min_{i,j\in\Omega^c} \omega_{i,j}\sqrt{d})$ with $b > 0$ as a constant to be selected empirically. Note that a very small cross-subspace spectral weight ω_{ij} weakens the corresponding penalty. In the extreme case $\omega_{ij} = 0$, no finite value of β can suppress c_{ij} through the weighted ℓ_1-regularizer. We therefore replace such degenerate weights by $\omega_{ij} = 1$, recovering an unweighted ℓ_1-penalty, and refer this estimator as $\widehat{\beta}_{\ell_1} = b/\|\widehat{\mathbf{c}}\|_1 \le \widehat{\beta}_{\mathrm{BDR}}$

We apply Lemma 3 by fixing the spectral embedding in Eq. 1. To initialize the coefficients, we use the minimum-Frobenius-norm solution $\widehat{\mathbf{C}}$, which is block-diagonal under exact noiseless self-expression, independent subspaces, and sufficient within-subspace sampling (Lu et al., 2012). We assess the quality of linear self-expression using $\|\mathbf{X} - \mathbf{X}\widehat{\mathbf{C}}\|_F$, with $\operatorname{diag}(\widehat{\mathbf{C}}) = \mathbf{0}$. If this error exceeds a threshold τ, we instead estimate d_i using local PCA, which approximates the tangent space of a smooth manifold (Fan et al., 2010). We denote the resulting estimate of the regularization weight by $\widehat{\beta}_{\mathrm{PCA}}$. Finally, using the empirically selected range $[0.001, 0.1]$, we set

$$\widehat{\beta} = \begin{cases} \widehat{\beta}_{\mathrm{BDR}}, & \text{if } \|\mathbf{X} - \mathbf{X}\widehat{\mathbf{C}}\|_F \le \tau, \\ \widehat{\beta}_{\ell_1}, & \text{if } \widehat{\beta}_{\mathrm{BDR}} > 0.1, \\ \widehat{\beta}_{\mathrm{PCA}}, & \text{if } \widehat{\beta}_{\ell_1} \notin [0.001, 0.1]. \end{cases} \tag{6}$$

The detailed procedure is provided in Section D in the supplementary material. Both the elastic-net and block-diagonal estimators use a Moore–Penrose pseudoinverse solution for initialization. Since directly computing the pseudoinverse is costly for large batches, we use the IMqrginv algorithm (Ataei, 2014) to approximate it. The estimation of $\widehat{\beta}_{\mathrm{BDR}}$ requires an input for b. Using synthetic data, (Soltanolkotabi et al., 2014b) show that choosing $b = 0.25$ to estimate the ℓ_1 regularization weight β provides the best trade-off between maximizing within-subspace connections and minimizing false connections across subspaces.

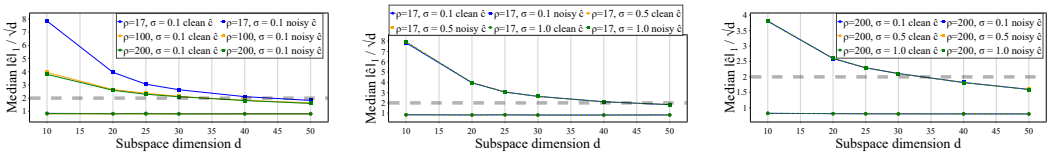
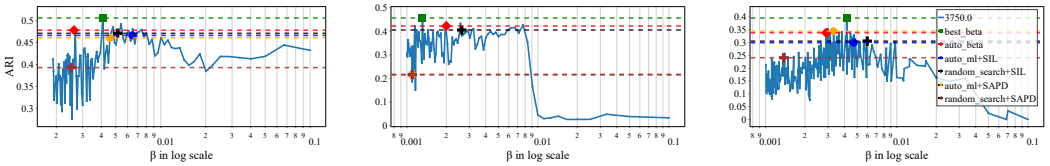
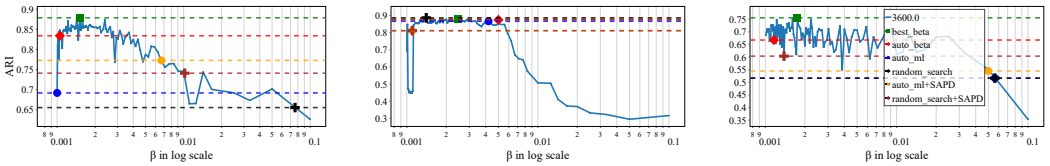


Figure 1: Median normalized coefficient norm $\|\widehat{\mathbf{c}}\|_1/\sqrt{d}$ versus subspace dimension d for different sampling densities ρ and noise levels σ, with ambient dimension $D = 200$. Pseudoinverse coefficients are computed separately from clean and noisy data. The dashed line at $y = 2$ marks the theoretical upper bound in Eq. 3, which requires a minimum sampling density of $\rho = 17$ under Lemma 1.



(a) PRO-DSC CIFAR-10 MCR (b) PRO-DSC-EN CIFAR-10 MCR (c) SeNet CIFAR-10 MCR



(d) PRO-DSC COIL-100 (e) PRO-DSC-EN COIL-100 (f) SeNet COIL-100

Figure 2: Sensitivity of clustering performance to β across datasets for SENet, PRO-DSC with block-diagonal regularization, and its elastic-net variant (PRO-DSC-EN). The horizontal and vertical axes show β and the Adjusted Rand index (ARI), respectively. The green square denotes label-guided selection, whereas the red diamond, blue circle, and black triangle denote unsupervised selection methods. Our estimated β consistently outperforms silhouette-based selection, particularly on CIFAR-10 (MCR) and COIL-100.

5 Experiments

We conduct experiments on synthetic and real-world datasets to validate our theory, evaluate $\|\widehat{\mathbf{c}}\|_1$ from the pseudoinverse solution as a proxy for $\sqrt{d}$, and assess our parameter adaptation framework.

5.1 Experiments on Synthetic Datasets

We generate random linear subspaces in $\mathbb{R}^n$ using QR decomposition of Gaussian matrices to obtain orthonormal bases. For each d-dimensional subspace, we sample ρd unit-norm points by normalizing Gaussian coefficient vectors and mapping them through the basis, where ρ denotes the sampling density. Independent noise $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n/n)$ is then added to each point, yielding $\mathbb{E}\|\mathbf{z}\|_2^2 = \sigma^2$. Let $\mathbf{y} = \mathbf{x} + \mathbf{z}$ denote the noisy data. We compute the minimum-ℓ_2-norm solution $\widehat{\mathbf{c}}$ in Eq. 13 separately from clean and noisy data. Fig. 1 shows the results. For clean data, the median ratio $\|\widehat{\mathbf{c}}\|_1/\sqrt{d}$ remains near 0.82 across sampling densities. For noisy data, the ratio is larger and can exceed the clean-data theoretical upper bound. This is consistent with noise increasing the data matrix's rank, particularly when d is small relative to the ambient dimension. Higher sampling densities and larger subspace dimensions reduce the ratio, bringing it closer to or below the bound.

5.2 Experiments on Real-World Datasets

We evaluate our method on MNIST (LeCun et al., 1998), CIFAR-10 and CIFAR-100 (Krizhevsky, 2009), COIL-100 (Nene et al., 1996), and ORL (Samaria & Harter, 1994). **MNIST** contains 70,000 grayscale images from 10 digit classes. Following Zhang et al. (2021), we extract 3,472-dimensional scattering features (Bruna & Mallat, 2013) and reduce them to 500 dimensions using PCA. **CIFAR-10** and **CIFAR-100** each contain 60,000 RGB images of size 32×32, grouped into 10 and 100 classes, respectively. We use pretrained CLIP image embeddings (Radford et al., 2021) for both datasets and additionally evaluate features extracted by an autoencoder trained with the MCR^2 objective (Yu et al., 2020) on CIFAR-10. **COIL-100** contains 7,200 color images of 100 objects, with 72 views per object captured over a 360° rotation. We use the pretrained features from (Ji et al., 2017). **ORL** contains 400 grayscale face images of size 92×112 from 40 individuals, with 10 images per subject. We use the pretrained CNN features from Ji et al. (2017). **Hopkins 155** (Tron & Vidal, 2007) is a motion-segmentation benchmark containing 155 video sequences with two or three independent motions. Each sample represents the $2F$-dimensional trajectory of a tracked feature point across F frames. We use the provided trajectories and motion labels without additional feature extraction.

5.2.1 Empirical Study of Hyperparameter Sensitivity

First, we investigate whether β requires careful tuning or whether arbitrary values can achieve competitive performance, thereby assessing the need for adaptive estimation. **Implementation details:** PRO-DSC combines an MCR^2-based representation term, a self-expressive reconstruction loss, and a block-diagonal regularizer: $-\frac{1}{2}\ln\det\big(\mathbf{I}+\alpha\mathbf{Z}_{\boldsymbol{\Theta}}^\top\mathbf{Z}_{\boldsymbol{\Theta}}\big)+\frac{\gamma}{2}\|\mathbf{Z}_{\boldsymbol{\Theta}}-\mathbf{Z}_{\boldsymbol{\Theta}}\mathbf{C}_{\boldsymbol{\Psi}}\|_F^2+\beta\, r(\mathbf{C}_{\boldsymbol{\Psi}})$. Here, α controls the coding-rate term, while γ and β weight the self-expressive loss and block-diagonal regularizer, respectively. We focus on their relative weighting, $\beta' = \beta/\gamma$. Following Meng et al. (2025), we fix γ to a default value satisfying the sufficient non-collapse condition in Theorem 2 and adaptively estimate β'. The coefficient in the original PRO-DSC objective is then recovered as $\beta = \gamma\beta'$, which is the value reported in our evaluation. **Results:** Figure 2 presents the results of a grid search over $\beta \in [0.001, 0.1]$, with $1/\beta$ sampled at intervals of 10. Figure 2 shows irregular, multimodal performance curves, with small changes in β often causing substantial ARI fluctuations. Selecting β through a limited search is therefore challenging and depends strongly on the dataset and feature representation. DiHypeSC (red diamonds) consistently approaches the best grid-search ARI (green squares) across datasets and regularizers. In contrast, baselines guided by internal validation metrics perform inconsistently, with substantial gaps on COIL-100 (Figs. 2d and 2f). Despite accounting for subspace structure, SAPD sometimes underperforms the silhouette score on CIFAR-10 MCR (Figs. 2a and 2b), suggesting that its advantage depends on the representation geometry. DiHypeSC nevertheless performs well using a single trial shown in Fig. 3.

5.2.2 Performance comparison

We compare DiHypeSC with two generic AutoML-based hyperparameter-optimization baselines guided by the silhouette score (SIL) (Rousseeuw, 1987) and the subspace clustering-specific internal metric introduced by Lipor & Balzano (2020b) (SAPD), which considers zero distances for points lying on the same subspace.

Gaussian process-based Bayesian optimization (GP-BO). GP-BO (Snoek et al., 2012) uses a Gaussian process surrogate and an acquisition function to select hyperparameters. We evaluate two label-free objectives: the silhouette score (SIL) and subspace-aware pairwise distance (SAPD), a pseudometric for subspace data (Lipor & Balzano, 2020a). Both variants, denoted GP-BO+SIL and GP-BO+SAPD, use Expected Improvement with five initial random evaluations and five subsequent optimization iterations.

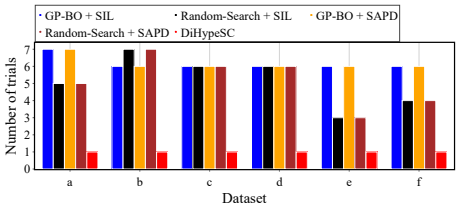


Figure 3: Number of trials used by each method to select its best hyperparameters across datasets (x-axis): (a) MNIST, (b) CIFAR-10 MCR, (c) CIFAR-10 CLIP, (d) CIFAR-100 CLIP, (e) COIL-100, (f) ORL, and (g) Hopkins 155.

Table 1: Experimental results for SENet evaluated using ARI (↑; mean ± standard deviation).

Method	MNIST	CIFAR-10 MCR	CIFAR-10 CLIP	CIFAR-100 CLIP	COIL-100	ORL	Hopkins 155
R-Search + SIL	0.771 ±0.048	0.305 ±0.023	0.734 ±0.026	0.480 ±0.036	0.243 ±0.015	0.171 ±0.018	0.570 ±0.380
GP-BO + SIL	0.730 ±0.072	0.299 ±0.053	0.718 ±0.024	0.466 ±0.030	0.248 ±0.019	0.154 ±0.029	0.573 ±0.379
R-Search + SAPD	**0.821 ±0.043**	0.241 ±0.028	0.732 ±0.039	0.479 ±0.025	0.545 ±0.023	0.516 ±0.021	0.516 ±0.373
GP-BO + SAPD	0.784 ±0.069	0.267 ±0.066	0.708 ±0.033	**0.494** ±0.015	0.554 ±0.161	0.144 ±0.032	0.539 ±0.385
DiHypeSC (Ours)	0.794 ±0.052	**0.510 ±0.005**	**0.792 ±0.055**	0.439 ±0.021	**0.590 ±0.017**	**0.187 ±0.040**	**0.581 ±0.371**

Table 2: Experimental results for PRO-DSC evaluated using ARI (↑; mean ± standard deviation).

Method	MNIST	CIFAR-10 MCR	CIFAR-10 CLIP	CIFAR-100 CLIP	COIL-100	ORL	Hopkins 155
R-Search + SIL	0.939 ±0.020	0.463 ±0.057	0.921 ±0.032	**0.582 ±0.017**	0.697 ±0.106	0.621 ±0.114	0.460 ±0.36
GP-BO + SIL	0.942 ±0.013	0.430 ±0.053	0.893 ±0.065	0.533 ±0.089	0.689 ±0.088	0.499 ±0.106	0.452 ±0.368
R-Search + SAPD	0.950 ±0.007	0.440 ±0.045	0.921 ±0.038	0.563 ±0.012	0.733 ±0.049	0.500 ±0.082	**0.595 ±0.348**
GP-BO + SAPD	0.942 ±0.013	0.458 ±0.058	0.923 ±0.033	0.384 ±0.023	0.763 ±0.058	0.432 ±0.093	0.584 ±0.346
DiHypeSC (Ours)	**0.950 ±0.012**	**0.476 ±0.004**	**0.935 ±0.002**	0.579 ±0.010	**0.830 ±0.020**	**0.684 ±0.130**	0.563 ±0.351

Random search (R-Search). Random search samples ten values of β from the empirical range $[0.001, 0.1]$ and selects the value maximizing SIL or SAPD (Lipor & Balzano, 2020a), matching GP-BO's evaluation budget. **DiHypeSC.** For SENet, we set $\lambda = 0.9$. Theorem 1 bounds residual correlations by $m'/\sqrt{d}$ for a numerical constant m'. Following Soltanolkotabi et al. (2014b), we set $m' = 0.25$ for both SENet and PRO-DSC in the good-locality regime ($\widehat{\beta} = \widehat{\beta}_{\mathrm{BDR}}$). Otherwise, we use $m' = 0.8$, motivated by their choice $\widehat{\beta} = 1/\sqrt{d}$ and our synthetic finding $\|\widehat{\mathbf{c}}\|_1 \approx 0.8\sqrt{d}$. Appendix Figs. 6 and 7 assess these choices across datasets.

Table 2 shows that our method achieves better results than the AutoML approaches for most of the examined cases, with only slight decreases on CIFAR-100 CLIP and Hopkins 155. The trial yielding the best result is reported in parentheses. DiHypeSC requires only a single run, whereas the AutoML baselines often achieve their best results after multiple trials. Additional results for normalized mutual information (NMI), subspace discovery error (SDE), and the silhouette score (SI) are provided in the appendix. Notably, a high silhouette score does not necessarily correspond to a high ARI.

6 Conclusion

We introduced DiHypeSC, a data-informed method for adapting regularization in deep self-expressive subspace clustering. While adopting the assumptions of (Soltanolkotabi et al., 2014b), our contribution lies in new sufficient conditions for subspace preservation and their translation into practical regularization rules for learned representations. These rules scale regularization inversely with an estimated subspace-dimension proxy and support elastic-net and block-diagonal regularization, as demonstrated with SENet and PRO-DSC.

DiHypeSC replaces dataset-specific searches over β with calibration of a more transferable constant, avoiding the assumptions of standard internal clustering indices. Experiments show performance comparable to AutoML-based optimization without repeated clustering runs. Future work will relax the fixed-eigenvector assumption and improve cluster-specific dimension estimation.

AI USE STATEMENT

We have used Grammarly and ChatGPT for editing the research paper to improve readability. We used ChatGPT in the early stages to find related papers, but manually verified the included related work. We have not used AI tools for other tasks that require or are recommended for disclosure. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide our code as a ZIP file in the supplementary material for the empirical analysis of β sensitivity and experimental comparisons with baseline methods.

REFERENCES

Mahdi Abavisani, Alireza Naghizadeh, Dimitris N. Metaxas, and Vishal M. Patel. Deep subspace clustering with data augmentation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/753a043674f0193523abc1bbce678686-Abstract.html.

Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA*, pp. 94–105. ACM Press, 1998. doi: 10.1145/276304.276314. URL https://doi.org/10.1145/276304.276314.

Mamdouh Aljoud, Gabriel Marques Tavares, Collin Leiber, and Thomas Seidl. Dcmatch - identify matching architectures in deep clustering through meta-learning. In *Advances in Knowledge Discovery and Data Mining - 29th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2025, Sydney, NSW, Australia, June 10-13, 2025, Proceedings, Part I*, volume 15870 of *Lecture Notes in Computer Science*, pp. 213–224. Springer, 2025. doi: 10.1007/978-981-96-8170-9_17. URL https://doi.org/10.1007/978-981-96-8170-9_17.

Mamdouh Aljoud, Gabriel Marques Tavares, Collin Leiber, and Thomas Seidl. Dcmatchboosted - improving deep clustering by architecture recommendation. *Int. J. Data Sci. Anal.*, 22 (1):58, 2026. doi: 10.1007/S41060-025-01009-1. URL https://doi.org/10.1007/s41060-025-01009-1.

Alireza Ataei. Improved qrginv algorithm for computing moore-penrose inverse matrices. *International Scholarly Research Notices*, 2014(1):641706, 2014. doi: https://doi.org/10.1155/2014/641706. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2014/641706.

Rafael Barbudo, Sebastián Ventura, and José Raúl Romero. Eight years of automl: categorisation, review and trends. *Knowl. Inf. Syst.*, 65(12):5097–5149, 2023. doi: 10.1007/S10115-023-01935-1. URL https://doi.org/10.1007/s10115-023-01935-1.

Anna Beer, Ekaterina Kuznetsova, and Claudia Plant. Parameter selection for DBSCAN: insights from persistent homology. In *ECAI 2025 - 28th European Conference on Artificial Intelligence, 25-30 October 2025, Bologna, Italy - Including 14th Conference on Prestigious Applications of Intelligent Systems (PAIS 2025)*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 3250–3257. IOS Press, 2025. doi: 10.3233/FAIA251192. URL https://doi.org/10.3233/FAIA251192.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13:281–305, 2012. doi: 10.5555/2503308.2188395. URL https://dl.acm.org/doi/10.5555/2503308.2188395.

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1872–1886, 2013. doi: 10.1109/TPAMI.2012.230. URL https://doi.org/10.1109/TPAMI.2012.230.

Matheus Camilo da Silva, Biagio Licari, Gabriel Marques Tavares, and Sylvio Barbon Junior. Benchmarking autoML clustering frameworks. In *AutoML Conference 2024 (ABCD Track)*, 2024. URL https://openreview.net/forum?id=RzUKJnph1g.

Marcos de Souza Oliveira, Sérgio Ricardo de Melo Queiroz, Cleber Zanchettin, and Francisco de Assis Tenório de Carvalho. Deep contrastive variational subspace clustering. *Neurocomputing*, 650:130901, 2025. doi: 10.1016/J.NEUCOM.2025.130901. URL https://doi.org/10.1016/j.neucom.2025.130901.

Tianjiao Ding, Derek Lim, René Vidal, and Benjamin D. Haeffele. Understanding doubly stochastic clustering. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5153–5165. PMLR, 2022. URL https://proceedings.mlr.press/v162/ding22a.html.

Tianjiao Ding, Shengbang Tong, Kwan Ho Ryan Chan, Xili Dai, Yi Ma, and Benjamin D. Haeffele. Unsupervised manifold linearizing and clustering. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 5427–5438. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00502. URL https://doi.org/10.1109/ICCV51070.2023.00502.

Elizabeth Ditton, Anne Swinbourne, and Trina S. Myers. Selecting a clustering algorithm: A semi-automated hyperparameter tuning framework for effective persona development. *Array*, 14: 100186, 2022. doi: 10.1016/J.ARRAY.2022.100186. URL https://doi.org/10.1016/j.array.2022.100186.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 2790–2797. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206547. URL https://doi.org/10.1109/CVPR.2009.5206547.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2765–2781, 2013. doi: 10.1109/TPAMI.2013.57. URL https://doi.org/10.1109/TPAMI.2013.57.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, pp. 226–231. AAAI Press, 1996. URL http://www.aaai.org/Library/KDD/1996/kdd96-037.php.

Mingyu Fan, Nannan Gu, Hong Qiao, and Bo Zhang. Intrinsic dimension estimation of data by principal component analysis. *CoRR*, abs/1002.2050, 2010. URL http://arxiv.org/abs/1002.2050.

Xinjie Fan, Yuguang Yue, Purnamrita Sarkar, and Y. X. Rachel Wang. On hyperparameter tuning in general clustering problemsm. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2996–3007. PMLR, 2020. URL http://proceedings.mlr.press/v119/fan20b.html.

Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated Machine Learning - Methods, Systems, Challenges*, The Springer Series on Challenges in Machine Learning, pp. 3–33. Springer, 2019. doi: 10.1007/978-3-030-05318-5_1. URL https://doi.org/10.1007/978-3-030-05318-5_1.

Obi L. Griffith, Byron J. Gao, Mikhail Bilenky, Yuliya Prichyna, Martin Ester, and Steven J. M. Jones. Kiwi: A scalable subspace clustering algorithm for gene expression analysis. *CoRR*, abs/0904.1931, 2009. URL http://arxiv.org/abs/0904.1931.

Benjamin David Haeffele, Chong You, and René Vidal. A critique of self-expressive deep subspace clustering. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=FOyuZ26emy.

Abolfazl Hashemi and Haris Vikalo. Evolutionary self-expressive models for subspace clustering. *IEEE J. Sel. Top. Signal Process.*, 12(6):1534–1546, 2018. doi: 10.1109/JSTSP.2018.2877478. URL https://doi.org/10.1109/JSTSP.2018.2877478.

Tianhao Huang, Tianyang Xu, Xiao-Jun Wu, and Josef Kittler. Unfolded ISTA for deep sparse subspace clustering. *Pattern Recognit.*, 180:113880, 2026. doi: 10.1016/J.PATCOG.2026.113880. URL https://doi.org/10.1016/j.patcog.2026.113880.

Abiodun M. Ikotun, Faustin Habyarimana, and Absalom E. Ezugwu. Cluster validity indices for automatic clustering: A comprehensive review. *Heliyon*, 11(2):e41953, 2025. ISSN 2405-8440. doi: https://doi.org/10.1016/j.heliyon.2025.e41953. URL https://www.sciencedirect.com/science/article/pii/S2405844025003330.

Pan Ji, Tong Zhang, Hongdong Li, Mathieu Salzmann, and Ian D. Reid. Deep subspace clustering networks. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 24–33, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/e369853df766fa44e1ed0ff613f563bd-Abstract.html.

Argyris Kalogeratos and Aristidis Likas. Dip-means: an incremental clustering method for estimating the number of clusters. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pp. 2402–2410, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/a8240cb8235e9c493a0c30607586166c-Abstract.html.

Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discov. Data*, 3(1):1:1–1:58, 2009. doi: 10.1145/1497577.1497578. URL https://doi.org/10.1145/1497577.1497578.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Fabien Lauer and Christoph Schnörr. Spectral clustering of linear subspaces for motion segmentation. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pp. 678–685. IEEE Computer Society, 2009. doi: 10.1109/ICCV.2009.5459173. URL https://doi.org/10.1109/ICCV.2009.5459173.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://doi.org/10.1109/5.726791.

Chun-Guang Li, Chong You, and René Vidal. On geometric analysis of affine sparse subspace clustering. *IEEE J. Sel. Top. Signal Process.*, 12(6):1520–1533, 2018. doi: 10.1109/JSTSP.2018.2867446. URL https://doi.org/10.1109/JSTSP.2018.2867446.

John Lipor and Laura Balzano. Clustering quality metrics for subspace clustering. *Pattern Recognit.*, 104:107328, 2020a. doi: 10.1016/J.PATCOG.2020.107328. URL https://doi.org/10.1016/j.patcog.2020.107328.

John Lipor and Laura Balzano. Clustering quality metrics for subspace clustering. *Pattern Recognit.*, 104:107328, 2020b. doi: 10.1016/J.PATCOG.2020.107328. URL https://doi.org/10.1016/j.patcog.2020.107328.

Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pp. 663–670. Omnipress, 2010. URL https://icml.cc/Conferences/2010/papers/521.pdf.

Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35 (1):171–184, 2013. doi: 10.1109/TPAMI.2012.88. URL https://doi.org/10.1109/TPAMI.2012.88.

Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–136, 1982. doi: 10.1109/TIT.1982.1056489. URL https://doi.org/10.1109/TIT.1982.1056489.

Yaroslava Lochman, Carl Olsson, and Christopher Zach. Learned trajectory embedding for subspace clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 19092–19102. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01806. URL https://doi.org/10.1109/CVPR52733.2024.01806.

Can-Yi Lu, Hai Min, Zhong-Qiu Zhao, Lin Zhu, De-Shuang Huang, and Shuicheng Yan. Robust and efficient subspace segmentation via least squares regression. In *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VII*, volume 7578 of *Lecture Notes in Computer Science*, pp. 347–360. Springer, 2012. doi: 10.1007/978-3-642-33786-4_26. URL https://doi.org/10.1007/978-3-642-33786-4_26.

Canyi Lu, Jiashi Feng, Zhouchen Lin, Tao Mei, and Shuicheng Yan. Subspace clustering by block diagonal representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):487–501, 2019. doi: 10.1109/TPAMI.2018.2794348. URL https://doi.org/10.1109/TPAMI.2018.2794348.

Brian McWilliams and Giovanni Montana. Subspace clustering of high-dimensional data: a predictive approach. *Data Min. Knowl. Discov.*, 28(3):736–772, 2014. doi: 10.1007/S10618-013-0317-Y. URL https://doi.org/10.1007/s10618-013-0317-y.

Xianghan Meng, Zhiyuan Huang, Wei He, Xianbiao Qi, Rong Xiao, and Chun-Guang Li. Exploring a principled framework for deep subspace clustering. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=7psWohxvxp.

Sameer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia object image library (coil-100). Technical Report CUCS-006-96, Dept. of Computer Science, Columbia University, February 1996.

Oron Nir, Jay Tenenbaum, and Ariel Shamir. Unimodal strategies in density-based clustering. In *Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD 2025, Porto, Portugal, September 15-19, 2025, Proceedings, Part I*, volume 16013 of *Lecture Notes in Computer Science*, pp. 498–513. Springer, 2025. doi: 10.1007/978-3-032-05962-8_29. URL https://doi.org/10.1007/978-3-032-05962-8_29.

Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *SIGKDD Explor.*, 6(1):90–105, 2004. doi: 10.1145/1007730.1007731. URL https://doi.org/10.1145/1007730.1007731.

Dan Pelleg and Andrew W. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pp. 727–734. Morgan Kaufmann, 2000.

Xi Peng, Shijie Xiao, Jiashi Feng, Wei-Yun Yau, and Zhang Yi. Deep subspace clustering with sparsity prior. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pp. 1925–1931. IJCAI/AAAI Press, 2016. URL http://www.ijcai.org/Abstract/16/275.

Xi Peng, Jiashi Feng, Shijie Xiao, Jiwen Lu, Zhang Yi, and Shuicheng Yan. Deep sparse subspace clustering. *CoRR*, abs/1709.08374, 2017. URL http://arxiv.org/abs/1709.08374.

Yannis Poulakis, Christos Doulkeridis, and Dimosthenis Kyriazis. Autoclust: A framework for automated clustering based on cluster validity indices. In *20th IEEE International Conference on Data Mining, ICDM 2020, Sorrento, Italy, November 17-20, 2020*, pp. 1220–1225. IEEE, 2020. doi: 10.1109/ICDM50108.2020.00153. URL https://doi.org/10.1109/ICDM50108.2020.00153.

Yannis Poulakis, Christos Doulkeridis, and Dimosthenis Kyriazis. A survey on automl methods and systems for clustering. *ACM Trans. Knowl. Discov. Data*, 18(5):120:1–120:30, 2024. doi: 10.1145/3643564. URL https://doi.org/10.1145/3643564.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Shankar R. Rao, Roberto Tron, René Vidal, and Yi Ma. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(10): 1832–1845, 2010. doi: 10.1109/TPAMI.2009.191. URL https://doi.org/10.1109/TPAMI.2009.191.

Yazhou Ren, Jingyu Pu, Zhimeng Yang, Jie Xu, Guofeng Li, Xiaorong Pu, Philip S. Yu, and Lifang He. Deep clustering: A comprehensive survey. *IEEE Trans. Neural Networks Learn. Syst.*, 36 (4):5858–5878, 2025. doi: 10.1109/TNNLS.2024.3403155. URL https://doi.org/10.1109/TNNLS.2024.3403155.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(87)90125-7. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

Ferdinand Samaria and Andy Harter. Parameterisation of a stochastic model for human face identification. In *Proceedings of Second IEEE Workshop on Applications of Computer Vision, WACV 1994, Sarasota, FL, USA, December 5-7, 1994*, pp. 138–142. IEEE, 1994. doi: 10.1109/ACV.1994.341300. URL https://doi.org/10.1109/ACV.1994.341300.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pp. 2960–2968, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html.

Mahdi Soltanolkotabi and Emmanuel J. Candés. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195 – 2238, 2012. doi: 10.1214/12-AOS1034. URL https://doi.org/10.1214/12-AOS1034.

Mahdi Soltanolkotabi, Ehsan Elhamifar, and Emmanuel J. Candès. Supplement to "robust subspace clustering", 2014a. URL https://doi.org/10.1214/13-AOS1199SUPP. Supplementary material.

Mahdi Soltanolkotabi, Ehsan Elhamifar, and Emmanuel J. Candès. Robust subspace clustering. *The Annals of Statistics*, 42(2):669 – 699, 2014b. doi: 10.1214/13-AOS1199. URL https://doi.org/10.1214/13-AOS1199.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles — A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617, 2002. URL https://jmlr.org/papers/v3/strehl02a.html.

Roberto Tron and René Vidal. A benchmark for the comparison of 3-d motion segmentation algorithms. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*. IEEE Computer Society, 2007. doi: 10.1109/CVPR.2007.382974. URL https://doi.org/10.1109/CVPR.2007.382974.

Dennis Tschechlov, Manuel Fritz, and Holger Schwarz. Automl4clust: Efficient automl for clustering analyses. In *Proceedings of the 24th International Conference on Extending Database Technology, EDBT 2021, Nicosia, Cyprus, March 23 - 26, 2021*, pp. 343–348. OpenProceedings.org, 2021. doi: 10.5441/002/EDBT.2021.32. URL https://doi.org/10.5441/002/edbt.2021.32.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, pp. 210–268. Cambridge University Press, 2012. doi: 10.1017/CBO9780511794308.006. URL https://doi.org/10.1017/cbo9780511794308.006.

René Vidal. Subspace clustering. *IEEE Signal Process. Mag.*, 28(2):52–68, 2011. doi: 10.1109/MSP.2010.939739. URL https://doi.org/10.1109/MSP.2010.939739.

René Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (GPCA). In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), 16-22 June 2003, Madison, WI, USA*, pp. 621–628. IEEE Computer Society, 2003. doi: 10.1109/CVPR.2003.1211411. URL https://doi.org/10.1109/CVPR.2003.1211411.

Ulrike von Luxburg. A tutorial on spectral clustering. *Stat. Comput.*, 17(4):395–416, 2007. doi: 10.1007/S11222-007-9033-Z. URL https://doi.org/10.1007/s11222-007-9033-z.

Yi Yang, Dong Xu, Feiping Nie, Shuicheng Yan, and Yueting Zhuang. Image clustering using local discriminant models and global integration. *IEEE Trans. Image Process.*, 19(10):2761–2773, 2010. doi: 10.1109/TIP.2010.2049235. URL https://doi.org/10.1109/TIP.2010.2049235.

Ming Yin, Junbin Gao, and Zhouchen Lin. Laplacian regularized low-rank representation and its applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(3):504–517, 2016. doi: 10.1109/TPAMI.2015.2462360. URL https://doi.org/10.1109/TPAMI.2015.2462360.

Chong You, Chun-Guang Li, Daniel P. Robinson, and René Vidal. Oracle based active set algorithm for scalable elastic net subspace clustering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 3928–3937. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.426. URL https://doi.org/10.1109/CVPR.2016.426.

Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6ad4174eba19ecb5fed17411a34ff5e6-Abstract.html.

Shangzhi Zhang, Chong You, René Vidal, and Chun-Guang Li. Learning a self-expressive network for subspace clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 12393–12403. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01221. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_Learning_a_Self-Expressive_Network_for_Subspace_Clustering_CVPR_2021_paper.html.

A Appendix

Definition 6. *(Independent subspaces) A collection of subspaces $\{\mathcal{S}_\ell\}_{\ell_1=1}^n$ is said to be independent if $dim(\bigcup_{\ell=1}^n \mathcal{S}_\ell) = \sum_{\ell=1}^n dim(\mathcal{S}_\ell)$.*

Let $\mathbf{U}_i \in \mathbb{R}^{D\times d_i}$ and $\mathbf{U}_j \in \mathbb{R}^{D\times d_j}$ be orthonormal bases of subspaces S_i and S_j, respectively, and let $d_{\min} = \min(d_i, d_j)$. Their principal angles $0 \le \theta^{(1)} \le \cdots \le \theta^{(d_{\min})} \le \pi/2$ satisfy $\cos\theta^{(\ell)} = \sigma_\ell(\mathbf{U}_i^\top \mathbf{U}_j)$, where σ_ℓ is the ℓth-largest singular value. The subspace affinity defined as:

$$\mathrm{aff}(S_i, S_j) = \sqrt{\frac{1}{d_{\min}} \sum_{\ell=1}^{d_{\min}} \cos^2\theta^{(\ell)}} = \frac{\|\mathbf{U}_i^\top \mathbf{U}_j\|_F}{\sqrt{d_{\min}}}. \tag{7}$$

With this we now turn to the definition of affinity condition

Definition 7. *(Affinity condition) (Soltanolkotabi et al., 2014b) A subspace S_i is considered as obeying the affinity condition, when:*

$$\max_{j:j\neq i} \mathit{aff}(S_i, S_j) \le \frac{\kappa_0}{\ln N} - \sigma\sqrt{\frac{d_{max}}{(n_i + n_j)\ln N}} \tag{8}$$

with n_i and n_j as number of samples on subspace S_i, S_j, σ denote the noise variance, that is $\mathbf{x}_{observe} = \mathbf{x}_{true} + \mathbf{z}$ with $\mathbf{z}$ as a random variable sample from a unit Gaussian distribution with variance σ. κ_0 is some fixed numerical constant.

Intuitively, the condition requires each subspace to be sufficiently separated from all others. The permitted affinity decreases as the noise level or subspace dimension increases.

Definition 8. *(Sampling condition) A subspace S_i is considered to fulfill the sampling condition, when*

$$\rho_i \ge \rho^\star \tag{9}$$

with $\rho^\star$ as another numerical constant. It says the minimum sample density for a subspace S_i to be identifiable, which is related to the subspace dimension and the noise level.

B Robust Subspace Clustering

We first recap the two-step procedure with data-driven regularization from Soltanolkotabi et al. (2014b) as shown in Algorithm 1. The algorithm involves two steps.

For each query sample $\mathbf{x}_i \in \mathcal{X}$, it minimizes the l_1 norm of its self-expressive coefficients while satisfying the hard constraints. For the self-expressive loss that needs to be bounded, often two times the noise variance, i.e., $\tau = 2\sigma$. They show theoretically as well as practically that the optimal solution from the hard constraint version of LASSO is roughly of the size $\sqrt{d}$. Then later they show theoretically that by defining the function $f(|\mathbf{c}|_1) \ge 0.707\sigma \cdot |\mathbf{c}|_1^{-1}$ with high probability there is no false discovery in $\mathbf{c}_i$ and with $f(|\mathbf{c}|_1) \le \alpha_0 \cdot |\mathbf{c}|_1^{-1}$ For some numerical constant α_0, there are at least $c_1\frac{d_i}{\rho_i}$ many true discoveries with c_1 as another positive numerical constant. Since $|\mathbf{c}_i^\star|_1$ is the optimal solution, and we know $\hat{\mathbf{c}} = \mathbf{X}_{-i}^T(\mathbf{X}_{-i}\mathbf{X}_{-i}^T)^\dagger\mathbf{x}$ is the minimum l_2-norm solution to the noiseless problem $\mathbf{X}_{-i}\mathbf{C} = \mathbf{x}_i$ and they show in the lemma $\frac{1}{10}\sqrt{\frac{d_i}{log\rho_i}} \le |\mathbf{c}^\star|_1 \le |\hat{\mathbf{c}}|_1 \le 2\sqrt{d_i}$, we use the l_1 norm of the pseudo-inverse approximation as an estimation For the optimal β and adapt the algorithm in 1 For deep subspace clustering framework that is illustrated in Algorithm 3.

Lemma 4 (Sub-Gaussian rows (Vershynin, 2012))**.** *Let A be an $N \times d$ matrix, where $N \ge d$, whose rows are independent sub-Gaussian isotropic random vectors in $\mathbb{R}^d$. Then, for every $t \ge 0$,*

$$\sigma_{\min}(A) \ge \sqrt{N} - C\sqrt{d} - t$$

with probability at least

$$1 - e^{-ct^2}.$$

Here, $\sigma_{\min}(A)$ denotes the minimum singular value of A, and $C = C_K$ and $c = c_K > 0$ depend only on the sub-Gaussian norm

$$K = \max_i \|A_i\|_{\Psi_2}$$

of the rows.

Algorithm 1: Two-Step Procedure with Data-Driven Regularization

Data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, noise threshold τ, function $f(\cdot)$

Output: Self-expressive coefficient matrix $\mathbf{C}$

$i = 1, \ldots, n$

Solve the constrained sparse recovery problem:

$$\mathbf{c}^\star = \arg\min_{\mathbf{c} \in \mathbb{R}^n} \|\mathbf{c}\|_1 \tag{10}$$

$$\text{s.t.} \|\mathbf{x}_i - \mathbf{X}\boldsymbol{c}\|_2 \leq \tau, \quad c_i = 0. \tag{11}$$

Estimate the regularization parameter

$$\beta_i = f(\|\boldsymbol{c}^\star\|_1)\,.$$

Solve the LASSO problem

$$\tilde{\boldsymbol{c}} = \arg\min_{\boldsymbol{c} \in \mathbb{R}^n} \frac{1}{2}\|\mathbf{x}_i - \mathbf{X}\boldsymbol{c}\|_2^2 + \beta_i \|\boldsymbol{c}\|_1$$

subject to

$$c_i = 0.$$

Set the i-th column of $\mathbf{C}$ as

$$\mathbf{C}_{:,i} = \tilde{\boldsymbol{c}}.$$

return $\mathbf{C}$

Lemma 5. *Let $\tau \geq 0$ be a noise tolerance, and fix a query sample $\mathbf{x}_i$. Define the common feasible set*

$$\mathcal{F}_i(\tau) = \{\mathbf{c} \in \mathbb{R}^n : \|\mathbf{x}_i - \mathbf{X}_{-i}\mathbf{c}\|_2 \leq \tau, \quad c_i = 0\}\,,$$

and assume that $\mathcal{F}_i(\tau)$ is nonempty. Consider the constrained ℓ_1-minimization problem

$$\mathbf{c}^\star_{\ell_1} \in \arg\min_{\mathbf{c} \in \mathcal{F}_i(\tau)} \|\mathbf{c}\|_1 \tag{12}$$

and the noiseless ℓ_2-minimization problem with $\tau = 0$

$$\widehat{\mathbf{c}} \in \arg\min_{\mathbf{c} \in \mathcal{F}_i(\tau=0)} \|\mathbf{c}\|_2^2. \tag{13}$$

Suppose that $\mathbf{x}_i$ lies in a subspace S_i of dimension d_i and that the sampling density ρ_i of S_i satisfies $\rho_i > \rho^\star$. Under the sampling assumptions specified above, we have

$$\frac{1}{10}\sqrt{\frac{d_i}{\ln \rho_i}} \leq \|\mathbf{c}^\star_{\ell_1}\|_1 \leq \|\widehat{\mathbf{c}}\|_1 \leq 2\sqrt{d_i}. \tag{14}$$

The upper bound holds with probability at least $1 - \exp(-\gamma_1 n_i) - \exp(-\gamma_2 d_i)$, whereas the lower bound holds with probability at least $1 - \exp(-\gamma_3 d_i) - \frac{10}{n_i^2}$. Here, d_i, ρ_i, and n_i denote the dimension, sampling density, and number of samples associated with S_i, respectively. Moreover, γ_1, γ_2, and γ_3 are fixed positive numerical constants.

C SENET WITH DATA-DEPENDENT REGULARIZATION

D SUBSPACE CLUSTERING WITH BLOCK-DIAGONAL REGULARIZATION

E PROOF OF LEMMA 1

Lemma 6 (Elastic-net coefficient-norm bound)**.** *Under the notation and assumptions of Lemma 5, Let N_i denote the number of columns of $\mathbf{X}^{(i)}_{-i}$, that is, the number of available samples from the subspace $\mathcal{S}_i$ containing $\mathbf{x}_i$, excluding $\mathbf{x}_i$ itself. Let $d_i = \dim(\mathcal{S}_i)$, assume that $N_i/d_i \geq 16$, and set $\tau = 2\sigma$.*

Algorithm 2: SENet with data-dependent elastic-net regularization

Training and test sets $\mathbf{X}_{\text{train}}$ and $\mathbf{X}_{\text{test}}$
Batch size n_b, number of training iterations T,
learning rate η, and number of clusters K
Elastic-net parameter $\lambda \in (0, 1]$
Calibration constant $m > 0$, coefficient scale $\alpha > 0$
Numerical constant $\varepsilon > 0$

Initialization: Randomly initialize the parameters Θ and Ψ of the query network $f(\cdot; \Theta)$ and key network $h(\cdot; \Psi)$, respectively. Initialize the learnable threshold $b > 0$ of

$$\mathcal{T}_b(s) = \operatorname{sgn}(s) \max\{0, |s| - b\}.$$

Training: $t = 1, \ldots, T$
Sample a minibatch $\mathbf{X}_b = [\mathbf{x}_1, \ldots, \mathbf{x}_{n_b}] \in \mathbb{R}^{D \times n_b}$ from $\mathbf{X}_{\text{train}}$.
Compute the query and key embeddings

$$\mathbf{Z}_q = [f(\mathbf{x}_1; \Theta) \quad \cdots \quad f(\mathbf{x}_{n_b}; \Theta)] \tag{15}$$

$$\mathbf{Z}_k = [h(\mathbf{x}_1; \Psi) \quad \cdots \quad h(\mathbf{x}_{n_b}; \Psi)]\,. \tag{16}$$

Estimate the minimum-ℓ_2 self-expressive coefficients. $i = 1, \ldots, n_b$ Let $\mathbf{Z}_{k,-i}$ denote $\mathbf{Z}_k$ with its i-th column removed. Compute

$$\widehat{\mathbf{c}}_{-i,i} = \mathbf{Z}_{k,-i}^\top \left(\mathbf{Z}_{k,-i}\mathbf{Z}_{k,-i}^\top\right)^\dagger \mathbf{z}_{q,i}, \qquad \widehat{c}_{ii} = 0.$$

Estimate the batchwise regularization parameter

$$\widehat{\beta}_t = \operatorname{median}_{1 \le i \le n_b} \left\{ \frac{m}{\lambda\big(\|\widehat{\mathbf{c}}_i\|_1 + \varepsilon\big)} \right\}.$$

Treat $\widehat{\beta}_t$ as a constant during backpropagation:

$$\beta_t = \operatorname{stopgrad}\left(\widehat{\beta}_t\right).$$

Compute the SENet coefficient matrix

$$\mathbf{C}_b = \alpha\, \mathcal{T}_b\left(\mathbf{Z}_k^\top \mathbf{Z}_q\right),$$

where $\mathcal{T}_b$ is applied elementwise.
Remove trivial self-representations:

$$\mathbf{C}_b \leftarrow \mathbf{C}_b - \operatorname{Diag}\left(\operatorname{diag}(\mathbf{C}_b)\right).$$

Compute the adaptive SENet loss

$$\mathcal{L}_b = \frac{1}{n_b} \sum_{i=1}^{n_b} \Bigg[\frac{1}{2} \left\|\mathbf{x}_i - \mathbf{X}_b \mathbf{c}_{b,i}\right\|_2^2 + \beta_t \left(\lambda \|\mathbf{c}_{b,i}\|_1 + \frac{1-\lambda}{2} \|\mathbf{c}_{b,i}\|_2^2 \right) \Bigg]. \tag{17}$$

Compute the gradients:
$\nabla_\Theta \mathcal{L}_b, \quad \nabla_\Psi \mathcal{L}_b, \quad \nabla_b \mathcal{L}_b$ Update the trainable parameters:

$$\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathcal{L}_b, \quad \Psi \leftarrow \Psi - \eta \nabla_\Psi \mathcal{L}_b, \quad b \leftarrow b - \eta \nabla_b \mathcal{L}_b.$$

adaptiveSENet

Consider the constrained elastic-net problem

$$\mathbf{c}^\star_{\text{elastic}} \in \arg \min_{\mathbf{c} \in \mathcal{F}_i(\tau)} \left\{ \lambda \|\mathbf{c}\|_1 + \frac{1-\lambda}{2} \|\mathbf{c}\|_2^2 \right\}, \qquad 0 < \lambda < 1. \tag{18}$$

Algorithm 2: SENet with data-dependent regularization (continued)

adaptiveSENet **Testing:** Compute

$$\mathbf{Z}_{q,\text{test}} = f(\mathbf{X}_{\text{test}}; \Theta), \qquad \mathbf{Z}_{k,\text{test}} = h(\mathbf{X}_{\text{test}}; \Psi).$$

Compute the test coefficient matrix

$$\mathbf{C}_{\text{test}} = \alpha\, \mathcal{T}_b \left(\mathbf{Z}_{k,\text{test}}^\top \mathbf{Z}_{q,\text{test}} \right), \qquad \operatorname{diag}(\mathbf{C}_{\text{test}}) = \mathbf{0}.$$

Construct the symmetric affinity matrix

$$\mathbf{A}_{\text{test}} = \frac{1}{2} \left(|\mathbf{C}_{\text{test}}| + |\mathbf{C}_{\text{test}}|^\top \right).$$

Apply spectral clustering with K clusters to $\mathbf{A}_{\text{test}}$.

Output: Predicted cluster assignments For $\mathbf{X}_{\text{test}}$.

Algorithm 3: PRO-DSC with data-dependent regularization

Dataset $\mathbf{X} = \mathbf{X}_{\text{train}} \cup \mathbf{X}_{\text{test}}$,
Batch size n_b,
Hyper-parameter β,
Number of iterations T,
Learning rate η **Initialization:** Randomly initialize the parameters Ψ, Θ in the feature representation networks $f(\cdot; \Theta)$ and coeffient learning networks $h(\cdot; \Psi)$ Initialize internal tracking parameter For the smallest hyper-parameter so far $\beta_{\text{intermediate}} = 0$
Training: $t = 1, \ldots, T$ Sample a batch $\mathbf{X}_b \in \mathbb{R}^{D \times n_b}$ from $\mathbf{X}_{\text{train}}$ *# Forward propagation* Compute representations $\mathbf{Z} \in \mathbb{R}^{d \times n_b}$ by Forward propagation in network $f(\cdot; \Theta)$ Estimate the minimum-ℓ_2 self-expressive coefficients. $i = 1, \ldots, n_b$ Let $\mathbf{Z}_{-i}$ denote $\mathbf{Z}$ with its i-th column removed. Compute

$$\widehat{\mathbf{c}}_{-i,i} = \mathbf{Z}_{-i}^\top \left(\mathbf{Z}_{-i} \mathbf{Z}_{-i}^\top \right)^\dagger \mathbf{z}_i, \qquad \widehat{c}_{ii} = 0.$$

Let $\widehat{\mathbf{C}}_b = [\widehat{\mathbf{c}}_{-1,1}, \cdots, \widehat{\mathbf{c}}_{-i,i}, \cdots \widehat{\mathbf{c}}_{-n,n}]$, construct the symmetric affinity matrix

$$\mathbf{A}_{\text{b}} = \frac{1}{2} \left(|\widehat{\mathbf{C}}_b| + |\widehat{\mathbf{C}}_b|^\top \right).$$

Calculate the Laplacian matrix from the affinity matrix $\mathbf{L} = \operatorname{Diag}(\mathbf{A} \cdot \mathbf{1}^T) - \mathbf{A}$ PerForm eigenvalue decompositions and get the eigenvectors corresponding to k smallest eigenvalues as spectral embedding For samples $\mathbf{U}$ Apply spectral clustering with K clusters to $\mathbf{A}_{\text{b}}$, to get the predicted cluster label $\hat{y}$. Convert the predicted cluster label to an indicator matrix $\mathbf{M}$ as:

$$M_{ij} = \mathbb{I}(y_i \neq y_j)\, \mathbb{I}(i \neq j) = \begin{cases} 1, & y_i \neq y_j \text{ and } i \neq j, \\ 0, & \text{otherwise.} \end{cases}$$

start the estimation of $\hat{\beta}$ by setting initially as zero $\beta = 0$ $i = 1, \ldots, n_b$ calculate error residual $\mathbf{r}_i = \mathbf{z}_i - \widehat{\mathbf{c}}_{-i,i} \mathbf{Z}_{k,-i}$
Calculate the inner-product between the error residual and all the sample embeddings $\mathbf{e} = \mathbf{Z}_{k,-i}$ as $\mathbf{Z}_{k,-i} \cdot \mathbf{r}_i^T$ $j = 1, \ldots, n_b$ Calculate the pairwise Euclidean distance as weight matrix $\mathbf{W}$ with $\omega_{i,j} = \frac{1}{2} \left(\|\mathbf{u}_i - \mathbf{u}_j\|_2^2 \right)$ $\hat{\beta} = \max(\hat{\beta}, \max_{j=1, j \neq i} \frac{e_{i,j}}{\omega_{i,j}}$
$\hat{\beta} > \beta_{\max}$ set the regularization strength as the ℓ_1-norm estimation $\beta = 0.25 \frac{1}{(\|\widehat{\mathbf{C}}_b\|_1)}$
Compute self-expressive matrix $\mathbf{C}_b \in \mathbb{R}^{n_b \times n_b}$ by Forward propogation in network $h(\cdot; \Psi)$ pro_dsc_break_point

Then,

$$\frac{\lambda}{10} \sqrt{\frac{d_i}{\ln \rho_i}} \leq \lambda \|\mathbf{c}_{\ell_1}^\star\|_1 \leq \lambda \|\mathbf{c}_{\text{elastic}}^\star\|_1 \leq 2\lambda \sqrt{d_i} + \kappa_i, \tag{19}$$

where $\kappa_i = \frac{1-\lambda}{2\left(\sqrt{N_i/d_i} - 2\right)^2} \leq \frac{1-\lambda}{8}$. *Consequently,* $\lambda \|\mathbf{c}_{\text{elastic}}^\star\|_1 \leq 2\lambda \sqrt{d_i} + (1-\lambda)/8$.

Algorithm 3: PRO-DSC with data-dependent regularization (continued)

pro_dsc_break_point *# Backward propagation* Update loss function with intermediate hyperparameter $\beta_{\text{intermediate}}$ Compute gradients: $\nabla_\Psi := \frac{\partial \mathcal{L}}{\partial \Psi}$, $\nabla_\Theta := \frac{\partial \mathcal{L}}{\partial \Theta}$ Update Ψ and Θ via: $\Psi \leftarrow \Psi - \eta \nabla_\Psi$, $\Theta \leftarrow \Theta - \eta \nabla_\Theta$ **Testing:** Compute self-expressive matrix $\mathbf{C}_{\text{test}}$ by Forward propagation $\mathbf{X}_{\text{test}}$ through $h(\cdot;\Psi) \circ f(\cdot;\Theta)$ Apply spectral clustering on the affinity $\mathbf{A}_{\text{test}}$

Proof. Because $\mathbf{c}^\star_{\ell_1}$ minimizes the ℓ_1-norm over $\mathcal{F}_i(\tau)$ and $\mathbf{c}^\star_{\text{elastic}} \in \mathcal{F}_i(\tau)$, we have $\|\mathbf{c}^\star_{\ell_1}\|_1 \le \|\mathbf{c}^\star_{\text{elastic}}\|_1$. The lower bound in Lemma 5 therefore gives $\frac{1}{10}\sqrt{d_i/\ln \rho_i} \le \|\mathbf{c}^\star_{\ell_1}\|_1 \le \|\mathbf{c}^\star_{\text{elastic}}\|_1$.

We next prove the upper bound. Let $\mathbf{X}^{(i)}_{-i}$ be the clean data matrix whose columns are the samples from $\mathcal{S}_i$, excluding $\mathbf{x}_i$, and define the minimum-ℓ_2-norm exact representation

$$\widehat{\mathbf{c}}_0 = {\mathbf{X}^{(i)}_{-i}}^\top \left(\mathbf{X}^{(i)}_{-i}{\mathbf{X}^{(i)}_{-i}}^\top\right)^\dagger \mathbf{x}_i. \tag{20}$$

Here, † denotes the Moore–Penrose pseudoinverse. We embed $\widehat{\mathbf{c}}_0$ into the full coefficient space by setting its entries corresponding to all other subspaces and to $\mathbf{x}_i$ equal to zero. By construction, $\mathbf{X}^{(i)}_{-i}\widehat{\mathbf{c}}_0 = \mathbf{x}_i$.

The proof of Lemma 5 in (Soltanolkotabi et al., 2014a) shows that, with the stated probability, $\|\mathbf{y}_i - \mathbf{Y}_{-i}\widehat{\mathbf{c}}_0\|_2 \le 2\sigma = \tau$. Hence, $\widehat{\mathbf{c}}_0 \in \mathcal{F}_i(\tau)$. The same proof also establishes

$$\begin{aligned} \|\widehat{\mathbf{c}}_0\|_2 &\le \frac{1}{\sqrt{N_i/d_i} - 2}, \\ \|\widehat{\mathbf{c}}_0\|_1 &\le \sqrt{N_i}\,\|\widehat{\mathbf{c}}_0\|_2 \le \frac{\sqrt{d_i}}{1 - 2\sqrt{d_i/N_i}} \le 2\sqrt{d_i}, \end{aligned} \tag{21}$$

where the last inequality follows from $N_i/d_i \ge 16$.

Because $\widehat{\mathbf{c}}_0$ is feasible, the optimality of $\mathbf{c}^\star_{\text{elastic}}$ implies

$$\begin{aligned} \lambda\|\mathbf{c}^\star_{\text{elastic}}\|_1 &\le \lambda\|\mathbf{c}^\star_{\text{elastic}}\|_1 + \frac{1-\lambda}{2}\|\mathbf{c}^\star_{\text{elastic}}\|_2^2 \\ &\le \lambda\|\widehat{\mathbf{c}}_0\|_1 + \frac{1-\lambda}{2}\|\widehat{\mathbf{c}}_0\|_2^2 \\ &\le 2\lambda\sqrt{d_i} + \frac{1-\lambda}{2\left(\sqrt{N_i/d_i} - 2\right)^2} \\ &= 2\lambda\sqrt{d_i} + \kappa_i. \end{aligned} \tag{22}$$

Finally, $N_i/d_i \ge 16$ implies $\left(\sqrt{N_i/d_i} - 2\right)^2 \ge 4$, and therefore $\kappa_i \le (1-\lambda)/8$. Combining the lower and upper bounds proves the result. □

F Proof of Lemma 2

Lemma 7 (Off-support residual-correlation condition)**.** *Let*

$$\mathbf{A} = \mathbf{X}_{-i} = [\mathbf{a}_1, \dots, \mathbf{a}_N] \in \mathbb{R}^{D\times N}$$

denote the matrix of candidate key samples after excluding the query sample $\mathbf{x}_i$. Let

$$T \subseteq [N] := \{1, \dots, N\}$$

be the index set of key samples lying in the same subspace as $\mathbf{x}_i$, and let $T^c = [N] \setminus T$. Assume that $\beta > 0$ and $0 < \lambda \le 1$.

Define

$$F(\mathbf{c}) = \frac{1}{2}\|\mathbf{x}_i - \mathbf{A}\mathbf{c}\|_2^2 + \beta\lambda\|\mathbf{c}\|_1 + \frac{\beta(1-\lambda)}{2}\|\mathbf{c}\|_2^2. \tag{23}$$

Suppose that

$$\mathbf{c}^\star \in \arg\min_{\mathbf{c}:\, \mathbf{c}_{T^c}=\mathbf{0}} F(\mathbf{c}) \tag{24}$$

is an optimal solution of the restricted problem, and define its residual by

$$\mathbf{r}^\star = \mathbf{x}_i - \mathbf{A}\mathbf{c}^\star.$$

If

$$\left\|\mathbf{A}_{T^c}^\top \mathbf{r}^\star\right\|_\infty < \beta\lambda, \tag{25}$$

then $\mathbf{c}^\star$ is also an optimal solution of the unrestricted problem

$$\min_{\mathbf{c}\in\mathbb{R}^N} F(\mathbf{c}). \tag{26}$$

Moreover, every optimal solution $\widehat{\mathbf{c}}$ of Eq. equation 26 satisfies

$$\widehat{\mathbf{c}}_{T^c} = \mathbf{0}.$$

Thus, every optimal solution is subspace-preserving.

If $0 < \lambda < 1$, then F is strongly convex, and hence its optimal solution is unique:

$$\widehat{\mathbf{c}} = \mathbf{c}^\star.$$

Proof. Because $\mathbf{c}^\star$ minimizes the restricted problem, there exists

$$\mathbf{s}_T \in \partial\|\mathbf{c}_T^\star\|_1$$

such that

$$\mathbf{A}_T^\top \mathbf{r}^\star = \beta\lambda\mathbf{s}_T + \beta(1-\lambda)\mathbf{c}_T^\star. \tag{27}$$

For each $j \in T$, the components of $\mathbf{s}_T$ satisfy

$$s_j = \begin{cases} \operatorname{sgn}(c_j^\star), & c_j^\star \neq 0, \\ s_j \in [-1,1], & c_j^\star = 0. \end{cases}$$

Consider an arbitrary perturbation

$$\mathbf{c} = \mathbf{c}^\star + \mathbf{h}.$$

The change in the reconstruction term is

$$\begin{aligned} &\frac{1}{2}\|\mathbf{x}_i - \mathbf{A}(\mathbf{c}^\star + \mathbf{h})\|_2^2 - \frac{1}{2}\|\mathbf{x}_i - \mathbf{A}\mathbf{c}^\star\|_2^2 \\ &\quad = -\left\langle \mathbf{A}^\top \mathbf{r}^\star, \mathbf{h}\right\rangle + \frac{1}{2}\|\mathbf{A}\mathbf{h}\|_2^2. \end{aligned} \tag{28}$$

Since $\mathbf{c}_{T^c}^\star = \mathbf{0}$, the change in the ℓ_1-norm satisfies

$$\begin{aligned} \|\mathbf{c}^\star + \mathbf{h}\|_1 - \|\mathbf{c}^\star\|_1 &= \|\mathbf{c}_T^\star + \mathbf{h}_T\|_1 - \|\mathbf{c}_T^\star\|_1 + \|\mathbf{h}_{T^c}\|_1 \\ &\geq \langle \mathbf{s}_T, \mathbf{h}_T\rangle + \|\mathbf{h}_{T^c}\|_1. \end{aligned} \tag{29}$$

The squared ℓ_2-term satisfies

$$\|\mathbf{c}^\star + \mathbf{h}\|_2^2 - \|\mathbf{c}^\star\|_2^2 = 2\langle \mathbf{c}^\star, \mathbf{h}\rangle + \|\mathbf{h}\|_2^2. \tag{30}$$

Because $\mathbf{c}_{T^c}^\star = \mathbf{0}$,

$$\langle \mathbf{c}^\star, \mathbf{h}\rangle = \langle \mathbf{c}_T^\star, \mathbf{h}_T\rangle.$$

Combining Eqs. equation 28–equation 30 gives

$$\begin{aligned} F(\mathbf{c}^\star + \mathbf{h}) - F(\mathbf{c}^\star) \geq\ & -\langle \mathbf{A}_T^\top \mathbf{r}^\star, \mathbf{h}_T\rangle - \langle \mathbf{A}_{T^c}^\top \mathbf{r}^\star, \mathbf{h}_{T^c}\rangle \\ & + \beta\lambda\langle \mathbf{s}_T, \mathbf{h}_T\rangle + \beta\lambda\|\mathbf{h}_{T^c}\|_1 \\ & + \beta(1-\lambda)\langle \mathbf{c}_T^\star, \mathbf{h}_T\rangle + \frac{1}{2}\|\mathbf{A}\mathbf{h}\|_2^2 \\ & + \frac{\beta(1-\lambda)}{2}\|\mathbf{h}\|_2^2. \end{aligned} \tag{31}$$

By Eq. equation 27, all terms that are linear in $\mathbf{h}_T$ cancel. Therefore,

$$F(\mathbf{c}^\star + \mathbf{h}) - F(\mathbf{c}^\star) \geq \beta\lambda\|\mathbf{h}_{T^c}\|_1 - \left\langle \mathbf{A}_{T^c}^\top \mathbf{r}^\star, \mathbf{h}_{T^c} \right\rangle + \frac{1}{2}\|\mathbf{A}\mathbf{h}\|_2^2 + \frac{\beta(1-\lambda)}{2}\|\mathbf{h}\|_2^2. \tag{32}$$

By Hölder's inequality,

$$\left\langle \mathbf{A}_{T^c}^\top \mathbf{r}^\star, \mathbf{h}_{T^c} \right\rangle \leq \left\|\mathbf{A}_{T^c}^\top \mathbf{r}^\star\right\|_\infty \|\mathbf{h}_{T^c}\|_1.$$

Consequently,

$$F(\mathbf{c}^\star + \mathbf{h}) - F(\mathbf{c}^\star) \geq \left(\beta\lambda - \left\|\mathbf{A}_{T^c}^\top \mathbf{r}^\star\right\|_\infty\right)\|\mathbf{h}_{T^c}\|_1 + \frac{1}{2}\|\mathbf{A}\mathbf{h}\|_2^2 + \frac{\beta(1-\lambda)}{2}\|\mathbf{h}\|_2^2. \tag{33}$$

By assumption,

$$\beta\lambda - \left\|\mathbf{A}_{T^c}^\top \mathbf{r}^\star\right\|_\infty > 0.$$

All terms on the right-hand side of Eq. equation 33 are therefore nonnegative. Hence,

$$F(\mathbf{c}^\star + \mathbf{h}) \geq F(\mathbf{c}^\star)$$

for every $\mathbf{h}$, proving that $\mathbf{c}^\star$ is an optimal solution of the unrestricted problem.

Furthermore, if $\mathbf{h}_{T^c} \neq \mathbf{0}$, then

$$F(\mathbf{c}^\star + \mathbf{h}) > F(\mathbf{c}^\star).$$

Thus, no other optimal solution can have nonzero coefficients on T^c. Therefore, every optimal solution $\widehat{\mathbf{c}}$ satisfies

$$\widehat{\mathbf{c}}_{T^c} = \mathbf{0}.$$

Finally, if $0 < \lambda < 1$, then

$$\frac{\beta(1-\lambda)}{2}\|\mathbf{c}\|_2^2$$

is strongly convex. Therefore, the complete objective F is strongly convex and has the unique minimizer

$$\widehat{\mathbf{c}} = \mathbf{c}^\star.$$

□

G Proof of Lemma 3

Lemma 8 (Subspace preservation for fixed Laplacian eigenvectors)**.** *Let*

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D\times N}$$

and let $\mathbf{C} \in \mathbb{R}^{N\times N}$ *be a self-expressive coefficient matrix satisfying*

$$\operatorname{diag}(\mathbf{C}) = \mathbf{0}.$$

Define

$$\mathbf{A}(\mathbf{C}) = \frac{1}{2}\left(|\mathbf{C}| + |\mathbf{C}|^\top\right)$$

and

$$\mathbf{L}(\mathbf{C}) = \operatorname{Diag}\left(\mathbf{A}(\mathbf{C})\mathbf{1}^T\right) - \mathbf{A}(\mathbf{C}).$$

Fix a matrix

$$\mathbf{U} = [\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)}] \in \mathbb{R}^{N\times k}, \qquad \mathbf{U}^\top\mathbf{U} = \mathbf{I}_k,$$

whose columns are the eigenvectors associated with the k *smallest eigenvalues of a fixed Laplacian matrix. Let* $\mathbf{u}_i^\top$ *denote the* i*th row of* $\mathbf{U}$*, and define*

$$w_{ij} = \frac{1}{2}\|\mathbf{u}_i - \mathbf{u}_j\|_2^2. \tag{34}$$

For fixed $\mathbf{U}$*, consider*

$$F_{\mathbf{U}}(\mathbf{C}) = \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 + \beta \sum_{i,j=1}^{N} w_{ij}|c_{ij}|, \tag{35}$$

subject to $\operatorname{diag}(\mathbf{C}) = \mathbf{0}$.

Let

$$\Omega = \{(i,j) : i \neq j,\ \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ lie in the same subspace}\},$$

and let Ω^c *denote the set of cross-subspace index pairs. Suppose that*

$$\mathbf{C}^\star \in \arg \min_{\substack{\operatorname{diag}(\mathbf{C})=\mathbf{0} \\ \mathbf{C}_{\Omega^c}=\mathbf{0}}} F_{\mathbf{U}}(\mathbf{C}) \tag{36}$$

is an optimal solution of the restricted problem. Define

$$\mathbf{E}^\star = \mathbf{X} - \mathbf{X}\mathbf{C}^\star.$$

If

$$\left| \left[\mathbf{X}^\top \mathbf{E}^\star\right]_{ij} \right| < \beta w_{ij} = \frac{\beta}{2}\|\mathbf{u}_i - \mathbf{u}_j\|_2^2 \tag{37}$$

for every $(i,j) \in \Omega^c$*, then* $\mathbf{C}^\star$ *is also an optimal solution of the unrestricted fixed-*$\mathbf{U}$ *problem. Moreover, every optimal solution* $\widehat{\mathbf{C}}$ *satisfies*

$$\widehat{\mathbf{C}}_{\Omega^c} = \mathbf{0}.$$

Thus, every optimal solution is subspace-preserving.

Proof. We first show that the fixed-eigenvector block-diagonal regularizer is a weighted ℓ_1-regularizer. For any symmetric affinity matrix $\mathbf{A}$, the Laplacian identity gives

$$\begin{aligned}
\operatorname{Tr}\left(\mathbf{U}^\top \mathbf{L}\mathbf{U}\right) &= \operatorname{Tr}(\mathbf{U}^\top \mathbf{D}\mathbf{U}) - \operatorname{Tr}(\mathbf{U}^\top \mathbf{A}\mathbf{U}) && (38)\\
&= \sum_{i=1}^{N} d_i \|\mathbf{u}_i\|_2^2 - \sum_{i,j=1}^{N} a_{ij}\mathbf{u}_i^\top \mathbf{u}_j && (39)\\
&= \frac{1}{2}\sum_{i,j=1}^{N} a_{ij}(\|\mathbf{u}_i\|_2^2 + \|\mathbf{u}_j\|_2^2) - \sum_{i,j=1}^{N} a_{ij}\mathbf{u}_i^\top \mathbf{u}_j && (40)\\
&= \frac{1}{2}\sum_{i,j=1}^{N} a_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2^2. && (41)
\end{aligned}$$

Substituting

$$a_{ij} = \frac{1}{2}\left(|c_{ij}| + |c_{ji}|\right)$$

into Eq. equation 41 yields

$$\begin{aligned}
\operatorname{Tr}\left(\mathbf{U}^\top \mathbf{L}(\mathbf{C})\mathbf{U}\right) &= \frac{1}{4}\sum_{i,j=1}^{N} \left(|c_{ij}| + |c_{ji}|\right)\|\mathbf{u}_i - \mathbf{u}_j\|_2^2 \\
&= \frac{1}{2}\sum_{i,j=1}^{N} |c_{ij}|\|\mathbf{u}_i - \mathbf{u}_j\|_2^2 \\
&= \sum_{i,j=1}^{N} w_{ij}|c_{ij}|. && (42)
\end{aligned}$$

The second equality follows by exchanging the indices i and j in the terms involving $|c_{ji}|$.

Now consider an arbitrary feasible perturbation

$$\mathbf{C} = \mathbf{C}^\star + \mathbf{H}, \qquad \operatorname{diag}(\mathbf{H}) = \mathbf{0}.$$

The change in the reconstruction term is

$$\begin{aligned}&\frac{1}{2}\|\mathbf{X}-\mathbf{X}(\mathbf{C}^\star+\mathbf{H})\|_F^2-\frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{C}^\star\|_F^2\\&\quad=-\left\langle\mathbf{X}^\top\mathbf{E}^\star,\mathbf{H}\right\rangle+\frac{1}{2}\|\mathbf{X}\mathbf{H}\|_F^2.\end{aligned}\tag{43}$$

Because $\mathbf{C}^\star$ minimizes the restricted convex problem, there exists a matrix $\mathbf{S}_\Omega$ such that

$$s_{ij}\in\partial|c_{ij}^\star|,\qquad(i,j)\in\Omega,$$

and

$$\left[\mathbf{X}^\top\mathbf{E}^\star\right]_{ij}=\beta w_{ij}s_{ij},\qquad(i,j)\in\Omega.\tag{44}$$

In particular,

$$s_{ij}=\begin{cases}\operatorname{sgn}(c_{ij}^\star), & c_{ij}^\star\neq 0,\\ s_{ij}\in[-1,1], & c_{ij}^\star=0.\end{cases}$$

Since $\mathbf{C}_{\Omega^c}^\star=\mathbf{0}$, the subgradient inequality gives

$$\begin{aligned}&\sum_{i,j}w_{ij}\left(|c_{ij}^\star+h_{ij}|-|c_{ij}^\star|\right)\\&\quad\geq\sum_{(i,j)\in\Omega}w_{ij}s_{ij}h_{ij}+\sum_{(i,j)\in\Omega^c}w_{ij}|h_{ij}|.\end{aligned}\tag{45}$$

Combining Eqs. equation 43, equation 44, and equation 45, the terms indexed by Ω cancel, and we obtain

$$\begin{aligned}&F_{\mathbf{U}}(\mathbf{C}^\star+\mathbf{H})-F_{\mathbf{U}}(\mathbf{C}^\star)\\&\quad\geq\sum_{(i,j)\in\Omega^c}\left(\beta w_{ij}|h_{ij}|-\left[\mathbf{X}^\top\mathbf{E}^\star\right]_{ij}h_{ij}\right)+\frac{1}{2}\|\mathbf{X}\mathbf{H}\|_F^2.\end{aligned}\tag{46}$$

For every $(i,j)\in\Omega^c$,

$$-\left[\mathbf{X}^\top\mathbf{E}^\star\right]_{ij}h_{ij}\geq-\left|\left[\mathbf{X}^\top\mathbf{E}^\star\right]_{ij}\right||h_{ij}|.$$

Therefore,

$$\begin{aligned}&F_{\mathbf{U}}(\mathbf{C}^\star+\mathbf{H})-F_{\mathbf{U}}(\mathbf{C}^\star)\\&\quad\geq\sum_{(i,j)\in\Omega^c}\left(\beta w_{ij}-\left|\left[\mathbf{X}^\top\mathbf{E}^\star\right]_{ij}\right|\right)|h_{ij}|+\frac{1}{2}\|\mathbf{X}\mathbf{H}\|_F^2.\end{aligned}\tag{47}$$

By assumption, every coefficient in parentheses is strictly positive. Thus, the right-hand side is nonnegative for every feasible perturbation $\mathbf{H}$, showing that $\mathbf{C}^\star$ is also an optimal solution of the unrestricted fixed-$\mathbf{U}$ problem.

Furthermore, if

$$\mathbf{H}_{\Omega^c}\neq\mathbf{0},$$

then at least one term in the sum is strictly positive. Hence,

$$F_{\mathbf{U}}(\mathbf{C}^\star+\mathbf{H})>F_{\mathbf{U}}(\mathbf{C}^\star).$$

Therefore, no optimal solution can contain a nonzero cross-subspace coefficient, and every optimal solution satisfies

$$\widehat{\mathbf{C}}_{\Omega^c}=\mathbf{0}.$$

□

H PROOF OF LEMMA 4

Lemma 9 (Block-diagonal property of minimum-Frobenius-norm self-expression)**.** *Let*

$$\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2 \quad \cdots \quad \mathbf{X}_K] \in \mathbb{R}^{D\times N},$$

where the columns of $\mathbf{X}_k$ *lie in the subspace* $\mathcal{S}_k \subseteq \mathbb{R}^D$*. Assume that the subspaces are independent, i.e.,*

$$\dim\left(\sum_{k=1}^{K}\mathcal{S}_k\right) = \sum_{k=1}^{K}\dim(\mathcal{S}_k).$$

Furthermore, assume that every sample can be represented as a linear combination of other samples from the same subspace.

Consider the exact self-expression problem

$$\mathbf{C}^\star = \arg\min_{\mathbf{C}\in\mathbb{R}^{N\times N}} \|\mathbf{C}\|_F^2 \quad \textit{subject to}\ \mathbf{X} = \mathbf{X}\mathbf{C},\ \mathrm{diag}(\mathbf{C}) = \mathbf{0}. \tag{48}$$

Then the unique optimal solution $\mathbf{C}^\star$ *satisfies the block-diagonal property:*

$$\mathbf{C}^\star = \mathrm{blkdiag}\left(\mathbf{C}_1^\star, \ldots, \mathbf{C}_K^\star\right).$$

Equivalently, $c_{ij}^\star = 0$ *whenever* $\mathbf{x}_i$ *and* $\mathbf{x}_j$ *belong to different subspaces.*

Proof. Consider a query sample $\mathbf{x}_j \in \mathcal{S}_\ell$ and partition its coefficient vector according to the subspaces:

$$\mathbf{c}_j = \begin{bmatrix}\mathbf{c}_{1j}^\top & \cdots & \mathbf{c}_{Kj}^\top\end{bmatrix}^\top.$$

The exact self-expression constraint gives

$$\mathbf{x}_j = \sum_{k=1}^{K}\mathbf{X}_k\mathbf{c}_{kj}.$$

Since $\mathbf{x}_j \in \mathcal{S}_\ell$ and the subspaces are independent, this equality implies

$$\mathbf{x}_j = \mathbf{X}_\ell\mathbf{c}_{\ell j}, \qquad \mathbf{X}_k\mathbf{c}_{kj} = \mathbf{0} \quad \text{for every } k \neq \ell.$$

Define $\widetilde{\mathbf{c}}_j$ by retaining $\mathbf{c}_{\ell j}$ and setting all out-of-subspace coefficient vectors to zero. Then

$$\mathbf{X}\widetilde{\mathbf{c}}_j = \mathbf{X}\mathbf{c}_j = \mathbf{x}_j,$$

and the zero-diagonal constraint remains satisfied. Moreover,

$$\|\widetilde{\mathbf{c}}_j\|_2^2 = \|\mathbf{c}_{\ell j}\|_2^2 \leq \sum_{k=1}^{K}\|\mathbf{c}_{kj}\|_2^2 = \|\mathbf{c}_j\|_2^2,$$

with strict inequality whenever an out-of-subspace coefficient vector is nonzero. Therefore, no minimum-norm solution can contain nonzero out-of-subspace coefficients.

Applying this argument to every column of $\mathbf{C}^\star$ shows that $\mathbf{C}^\star$ is block diagonal. Finally, because the squared Frobenius norm is strictly convex and the feasible set is affine, the optimal solution is unique. □

[Oracle reconstruction and residual-correlation bounds] Let

$$\mathbf{U} \in \mathbb{R}^{N\times k}, \qquad \mathbf{U}^\top\mathbf{U} = \mathbf{I}_k,$$

and define

$$w_{ij} = \frac{1}{2}\|\mathbf{u}_i - \mathbf{u}_j\|_2^2, \qquad P_{\mathbf{W}}(\mathbf{C}) = \sum_{i\neq j} w_{ij}|c_{ij}|.$$

Consider the fixed-eigenvector problem

$$\widehat{\mathbf{C}} \in \arg\min_{\mathrm{diag}(\mathbf{C})=\mathbf{0}}\left\{\frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 + \beta P_{\mathbf{W}}(\mathbf{C})\right\}, \tag{49}$$

and define

$$\widehat{\mathbf{E}} = \mathbf{X} - \mathbf{X}\widehat{\mathbf{C}}.$$

For any feasible coefficient matrix $\mathbf{C}^0$ satisfying $\operatorname{diag}(\mathbf{C}^0) = \mathbf{0}$, define

$$\mathbf{E}^0 = \mathbf{X} - \mathbf{X}\mathbf{C}^0.$$

Then

$$\|\widehat{\mathbf{E}}\|_F^2 + 2\beta P_{\mathbf{W}}(\widehat{\mathbf{C}}) \le \|\mathbf{E}^0\|_F^2 + 2\beta P_{\mathbf{W}}(\mathbf{C}^0). \tag{50}$$

Consequently,

$$\|\widehat{\mathbf{E}}\|_F^2 \le \|\mathbf{E}^0\|_F^2 + 2\beta P_{\mathbf{W}}(\mathbf{C}^0). \tag{51}$$

Moreover, for every $i \neq j$,

$$\left| \left[\mathbf{X}^\top \widehat{\mathbf{E}} \right]_{ij} \right| \le \beta w_{ij}. \tag{52}$$

Since $0 \le w_{ij} \le 1$, it follows that

$$\left\| \mathcal{P}_{\text{off}} \left(\mathbf{X}^\top \widehat{\mathbf{E}} \right) \right\|_\infty \le \beta, \tag{53}$$

where

$$\mathcal{P}_{\text{off}}(\mathbf{M}) = \mathbf{M} - \operatorname{Diag}(\operatorname{diag}(\mathbf{M}))$$

and, throughout this proposition,

$$\|\mathbf{M}\|_\infty := \max_{i,j} |m_{ij}|$$

denotes the entrywise maximum norm.

I Proof of Theorem 1

Before introducing the proof of Theorem 1, the following lemma (Soltanolkotabi et al., 2014b) is recalled:

Lemma 10. *Let $\mathbf{A} \in \mathbb{R}^{d_1 \times N_1}$ be a matrix with columns sampled uniformly at random from the unit sphere of $\mathbb{R}^{d_1}$, $\omega \in \mathbb{R}_2^d$ be a vector sampled at random from the unit sphere of $\mathbb{R}_2^d$ and independent of $\mathbf{A}$ and $\Sigma \in \mathbb{R}^{d_1 \times d_2}$ be a deterministic matrix. For some constants a, $b > 1$, we have:*

$$\left\| \mathbf{A}^\top \mathbf{\Sigma} \mathbf{w} \right\|_\infty \le \sqrt{\ln(a)\ln(b)} \frac{\|\mathbf{\Sigma}\|_F}{\sqrt{d_1}\sqrt{d_2}} \tag{54}$$

with probability at least

$$1 - \frac{2}{\sqrt{a}} - \frac{2N_1}{\sqrt{b}}.$$

The proof of the lemma is provided in Soltanolkotabi & Candés (2012).

Lemma 11 (Gaussian inner-product bound)**.** *Suppose that $\mathbf{A} \in \mathbb{R}^{d \times N}$ has i.i.d. $\mathcal{N}(0,1)$ entries, and let $\mathbf{z} \in \mathbb{R}^d$ be a unit-norm vector. Then*

$$\left\| \mathbf{A}^\top \mathbf{z} \right\|_\infty \le 2\sqrt{2 \ln N}$$

with probability at least

$$1 - \frac{2}{N^2}.$$

The result also holds when $\mathbf{z}$ is a random vector independent of $\mathbf{A}$.

Theorem 2 (Upper bound on correlations between off-subspace samples and reconstruction residuals)**.** *Let $\mathbf{x}_i$ lie in a subspace S_i of dimension d_i, $\mathbf{A}$ as dataset samples excluding $\mathbf{x}_i$ the query sample, i.e., $\mathbf{X}_{-i}$ Suppose that S_i satisfies the affinity and sampling conditions in Definitions 7 and 8, respectively, and that the noise level σ is sufficiently small. Let $\mathbf{c}^\star$ solve the restricted elastic-net problem*

$$\min_{\mathbf{c}:\, \mathbf{c}_{T^c} = \mathbf{0}} \frac{1}{2}\|\mathbf{x}_i - \mathbf{A}\mathbf{c}\|_2^2 + \beta \left(\lambda \|\mathbf{c}\|_1 + \frac{1-\lambda}{2} \|\mathbf{c}\|_2^2 \right),$$

and define $\mathbf{r}^\star = \mathbf{x}_i - \mathbf{A}\mathbf{c}^\star$. *Suppose that, under the sampling and noise model in the preliminary, there exists a constant* $m > 0$ *such that if*

$$\beta > \frac{m}{\lambda\sqrt{d_i}} \tag{55}$$

then the optimal solution $c^\star$ *given these regularization strength have zero cross-subspace coefficients. Further let* $\mathbf{C}^\star$ *solve the restricted weighted* ℓ_1 *norm problem:*

$$F_{\mathbf{U}}(\mathbf{C}) = \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 + \beta \sum_{i,j} w_{ij}|c_{ij}|, \quad \operatorname{diag}(\mathbf{C}) = \mathbf{0}. \tag{56}$$

and $\mathbf{C}^\star \in \arg\min\{F_{\mathbf{U}}(\mathbf{C}) : \operatorname{diag}(\mathbf{C}) = \mathbf{0},\ \mathbf{C}_{T^c} = \mathbf{0}\}$, *and define* $\mathbf{E}^\star = \mathbf{X} - \mathbf{X}\mathbf{C}^\star$. *Then, there exists a constant* m' *such that, if*

$$\beta > \frac{m'}{\min_{i\neq j} \omega_{i,j}\sqrt{d_i}}, \tag{57}$$

the resulting optimal solution is subspace-preserving.

Proof. The original elastic-net problem is

$$\begin{aligned} \widehat{\boldsymbol{c}} \in \arg\min_{\boldsymbol{\beta}} \quad & \frac{1}{2}\left\|\mathbf{x} - \sum_{\ell=1}^{L} \mathbf{X}^{(\ell)}\mathbf{c}^{(\ell)}\right\|_2^2 \\ & + \lambda\beta \sum_{\ell=1}^{L} \left\|\mathbf{c}^{(\ell)}\right\|_1 \\ & + \beta\frac{(1-\lambda)}{2} \sum_{\ell=1}^{L} \left\|\boldsymbol{c}^{(\ell)}\right\|_2^2 \end{aligned} \tag{58}$$

Suppose that $\mathbf{x}$ belongs to the subspace $\mathcal{S}_1$. To prove that equation 58 has no false discoveries, one needs to show that

$$\widehat{\boldsymbol{c}}^{(\ell)} = \mathbf{0}, \qquad \ell \neq 1.$$

For this purpose, let us first consider the reduced oracle, problem

$$\widehat{\boldsymbol{c}}^{(1)} \in \arg\min_{\boldsymbol{c}^{(1)}} \frac{1}{2}\left\|\mathbf{x} - \mathbf{X}^{(1)}\boldsymbol{c}^{(1)}\right\|_2^2 + \lambda\left\|\boldsymbol{c}^{(1)}\right\|_1 + \beta\frac{1-\lambda}{2}\|\boldsymbol{c}^{(1)}\|_2^2, \tag{59}$$

which is obtained by constraining all off-subspace coefficients to zero. Its solution can be embedded into the full coefficient space as

$$\boldsymbol{c}^{\mathrm{red}} = \left(\widehat{\boldsymbol{c}}^{(1)}, \mathbf{0}, \ldots, \mathbf{0}\right),$$

with residual

$$\widehat{\mathbf{r}} = \mathbf{x} - \mathbf{X}^{(1)}\widehat{\boldsymbol{c}}^{(1)}.$$

Optimality of the reduced problem gives the within-subspace KKT condition

$$\mathbf{X}^{(1)\top}\widehat{\mathbf{r}} \in \lambda\beta\,\partial\left\|\widehat{\boldsymbol{c}}^{(1)}\right\|_1.$$

For a zero coefficient associated with an off-subspace sample, the KKT condition for the original problem requires

$$\left|\mathbf{x}_j^\top\widehat{\mathbf{r}}\right| \leq \lambda\beta.$$

Consequently, if

$$\left\|\mathbf{X}^{(\ell)\top}\left(\mathbf{x} - \mathbf{X}^{(1)}\widehat{\boldsymbol{c}}^{(1)}\right)\right\|_\infty < \lambda\beta, \qquad \ell \neq 1, \tag{60}$$

then the padded reduced solution $\boldsymbol{c}^{\mathrm{red}}$ satisfies the KKT conditions of the original problem. Lemma A.2 further shows that the strict inequality in equation 60 implies that every optimal solution of the original problem satisfies

$$\widehat{\boldsymbol{c}}^{(\ell)} = \mathbf{0}, \qquad \ell \neq 1.$$

Therefore, we now switch to analyzing the reduced problem in equation 59. Let $\mathbf{x} = \mathbf{x}^\star + \mathbf{z}$ with $\mathbf{z}$ as noise variable $\mathbf{z} \sim \mathcal{N}(0, \sigma^2)$ and $\mathbf{x}^\star$ as the uncorrupted data

$$\begin{aligned}\mathbf{X}^{(\ell)\top}\left(\mathbf{x} - \mathbf{X}^{(1)}\widehat{\boldsymbol{c}}^{(1)}\right) &= \mathbf{X}^{\star(\ell)\top}\left(\mathbf{x}_\parallel - \mathbf{X}_\parallel^{(1)}\widehat{\boldsymbol{c}}^{(1)}\right) \\ &+ \mathbf{X}^{\star(\ell)\top}\left(\mathbf{x}_\perp - \mathbf{X}_\perp^{(1)}\widehat{\boldsymbol{c}}^{(1)}\right) \\ &+ \mathbf{z}^T\left(\mathbf{x}_\parallel - \mathbf{X}_\parallel^{(1)}\widehat{\boldsymbol{c}}^{(1)}\right) \\ &+ \mathbf{z}^T\left(\mathbf{x}_\perp - \mathbf{X}_\perp^{(1)}\widehat{\boldsymbol{c}}^{(1)}\right)\end{aligned} \tag{61}$$

We now calculate the ℓ_∞ norm of each term by using Lemma 10 and estimate the size of each residuals:

The term $\|\mathbf{X}^{\star(\ell)\top}\left(\mathbf{x}_\parallel - \mathbf{X}_\parallel^{(1)}\widehat{\boldsymbol{c}}^{(1)}\right)\|_\infty$. Using Lemma 10 with $a = 2\sqrt{\ln N}$, $b = 2\sqrt{2\ln N}$ and assume $d_1 > d_\ell$ we have:

$$\begin{aligned}&\|\mathbf{X}^{\star(\ell)\top}\left(\mathbf{x}_\parallel - \mathbf{X}_\parallel^{(1)}\widehat{\boldsymbol{c}}^{(1)}\right)\|_{\ell_\infty} \\ &\leq \sqrt{32}\ln N\frac{\mathrm{aff}(S_1, S_\ell)}{\sqrt{d_1}}\|\mathbf{x}_\parallel - \mathbf{X}_\parallel^{(1)}\widehat{c}^{(1)}\|_{\ell_2}\end{aligned} \tag{62}$$

We now turn to bound the projected residual $\|\mathbf{x}_\parallel - \mathbf{X}_\parallel^{(1)}\widehat{c}^{(1)}\|_{\ell_2}$ To this aim, we consider the projected problem:

$$\widetilde{\mathbf{c}}^{(1)} \in \arg\min_{\boldsymbol{c}^{(1)}} \frac{1}{2}\left\|\mathbf{x}_\parallel - \mathbf{X}_\parallel^{(1)}\boldsymbol{c}^{(1)}\right\|_2^2 + \beta\lambda\left\|\boldsymbol{c}^{(1)}\right\|_1 + \beta\frac{1-\lambda}{2}\|\boldsymbol{c}^{(1)}\|_2^2 \tag{63}$$

and assume $\mathbf{x}_{i,\parallel} = \mathbf{X}_\parallel^{(1)}\widetilde{\mathbf{c}}^{(1)} + \epsilon_{\mathbf{i}}^\star$ with ϵ as the noise for the projected resiudal, further let $\eta_1 = \beta\lambda$, $\eta_2 = \beta(1-\lambda)$ and $\widehat{\mathbf{r}}_i^{(1)} = \mathbf{x}_{i,\parallel} - \mathbf{X}_\parallel^{(1)}\widehat{\mathbf{c}}^{(1)} = \mathbf{X}_\parallel^{(1)}\widetilde{\mathbf{c}}^{(1)} + \epsilon_{\mathbf{i}}{}^\star - \mathbf{X}_\parallel^{(1)}\widehat{\mathbf{c}}^{(1)}$

Basic oracle reconstruction bound. By the optimality of $\widehat{\mathbf{c}}_i^{(1)}$,

$$\begin{aligned}&\frac{1}{2}\|\widehat{\mathbf{r}}_i^{(1)}\|_2^2 + \eta_1\|\widehat{\mathbf{c}}_i^{(1)}\|_1 + \frac{\eta_2}{2}\|\widehat{\mathbf{c}}_i^{(1)}\|_2^2 \\ &\quad\leq \frac{1}{2}\|\boldsymbol{\epsilon}_i^\star\|_2^2 + \eta_1\|\widetilde{\mathbf{c}}^{(1)}\|_1 + \frac{\eta_2}{2}\|\widetilde{\mathbf{c}}^{(1)}\|_2^2.\end{aligned} \tag{64}$$

Dropping the nonnegative regularization terms on the left gives

$$\boxed{\|\widehat{\mathbf{r}}_i^{(1)}\|_2^2 \leq \|\boldsymbol{\epsilon}_i^\star\|_2^2 + 2\eta_1\|\widetilde{\mathbf{c}}^{(1)}\|_1 + \eta_2\|\widetilde{\mathbf{c}}^{(1)}\|_2^2.}$$

This bound does not require a restricted eigenvalue condition or a particular choice of η_1.

Prediction-error bound. Since

$$\widehat{\mathbf{r}}_i^{(1)} = \boldsymbol{\epsilon}_i^\star - \mathbf{X}_\parallel^{(1)}\boldsymbol{\Delta}_i,$$

with $\Delta_i = \widehat{\mathbf{c}}^{(1)} - \widetilde{\mathbf{c}}^{(1)}$ the basic inequality can be written as

$$\begin{aligned}\frac{1}{2}\left\|\mathbf{X}_\parallel^{(1)}\boldsymbol{\Delta}_i\right\|_2^2 + \frac{\eta_2}{2}\|\boldsymbol{\Delta}_i\|_2^2 &\leq \left\langle {\mathbf{X}_\parallel^{(\mathbf{1})}}^\top\boldsymbol{\epsilon}_i^\star - \eta_2\widetilde{\mathbf{c}}^{(1)}, \boldsymbol{\Delta}_i\right\rangle \\ &+ \eta_1\left(\|\widetilde{\mathbf{c}}^{(1)}\|_1 - \|\widetilde{\mathbf{c}}^{(1)} + \boldsymbol{\Delta}_i\|_1\right).\end{aligned} \tag{65}$$

Assume that

$$\boxed{\eta_1 \geq 2\left\|{X_\parallel^{(1)}}^\top\boldsymbol{\epsilon}_i^\star - \eta_2\widetilde{\mathbf{c}}^{(1)}\right\|_\infty.}$$

By Hölder's inequality,

$$\left\langle \mathbf{X}_{-i}^\top\boldsymbol{\epsilon}_i^\star - \eta_2\widetilde{\mathbf{c}}^{(1)}, \boldsymbol{\Delta}_i\right\rangle \leq \frac{\eta_1}{2}\|\boldsymbol{\Delta}_i\|_1. \tag{66}$$

Moreover, since

$$\begin{aligned}\|\widehat{\mathbf{c}}^{(1)}\|_1 &= \|\widetilde{\mathbf{c}}^{(1)} + \Delta\|_1 \\ &= \|\widetilde{\mathbf{c}}^{(1)}_{S_i} + \Delta_{S_i}\|_1 + \|\widetilde{\mathbf{c}}^{(1)}_{S_i^c} + \Delta_{S_i^c}\|_1 \\ &= \|\widetilde{\mathbf{c}}^{(1)}_{S_i} + \Delta_{S_i}\|_1 + \|\Delta_{S_i^c}\|\end{aligned} \tag{67}$$

with $\widetilde{\mathbf{c}}^{(1)}_{S_i^c} = 0$

$$\|\widetilde{\mathbf{c}}^{(1)}\|_1 - \|\widetilde{\mathbf{c}}^{(1)} + \boldsymbol{\Delta}_i\|_1 \leq \|(\boldsymbol{\Delta}_i)_{S_i}\|_1 - \|(\boldsymbol{\Delta}_i)_{S_i^c}\|_1. \tag{68}$$

with $(\boldsymbol{\Delta}_i)_{S_i}$ with the coefficient error restricted to sample indices in subspace S_i and $(\boldsymbol{\Delta}_i)_{S_i}$ with the coefficient error restricted to sample indices outside the subspace S_i Consequently,

$$\begin{aligned}&\left\|X_{\|}^{(1)}\boldsymbol{\Delta}_i\right\|_2^2 + \eta_2\|\boldsymbol{\Delta}_i\|_2^2 + \eta_1\|(\boldsymbol{\Delta}_i)_{S_i^c}\|_1 \\ &\leq 3\eta_1\|(\boldsymbol{\Delta}_i)_{S_i}\|_1.\end{aligned} \tag{69}$$

In particular,

$$\|(\boldsymbol{\Delta}_i)_{S_i^c}\|_1 \leq 3\|(\boldsymbol{\Delta}_i)_{S_i}\|_1.$$

due to $\left(\left\|X_{\|}^{(1)}\boldsymbol{\Delta}_i\right\|_2^2 + \eta_2\|\boldsymbol{\Delta}_i\|_2^2\right) \geq 0$

Define the elastic-net restricted curvature

$$\alpha_i = \inf_{\substack{\boldsymbol{\Delta}\neq\mathbf{0} \\ \|\boldsymbol{\Delta}_{S_i^c}\|_1 \leq 3\|\boldsymbol{\Delta}_{S_i}\|_1}} \frac{\|\mathbf{X}_{\|}^{(1)}\boldsymbol{\Delta}\|_2^2 + \eta_2\|\boldsymbol{\Delta}\|_2^2}{\|\boldsymbol{\Delta}\|_2^2}.$$

Because of the ridge component,

$$\alpha_i \geq \eta_2.$$

with s_i as the number of non-zero elements in $(\boldsymbol{\Delta}_i)_{S_i}$, in the ℓ_1 dominated regime, we assume $s_i \asymp d_i$, i.e., s_i and d_i have the same order of magnitude. Using

$$\|(\boldsymbol{\Delta}_i)_{S_i}\|_1 \leq \sqrt{s_i}\|\boldsymbol{\Delta}_i\|_2,$$

we obtain

$$\alpha_i\|\boldsymbol{\Delta}_i\|_2^2 \leq 3\eta_1\sqrt{s_i}\|\boldsymbol{\Delta}_i\|_2.$$

Therefore,

$$\boxed{\|\widehat{\mathbf{c}}_i^{(1)} - \widetilde{\mathbf{c}}^{(1)}\|_2 \leq \frac{3\eta_1\sqrt{s_i}}{\alpha_i}.}$$

Furthermore,

$$\boxed{\left\|\mathbf{X}_{\|}^{(1)}(\widehat{\mathbf{c}}_i^{(1)} - \widetilde{\mathbf{c}}^{(1)})\right\|_2^2 + \eta_2\|\widehat{\mathbf{c}}_i^{(1)} - \widetilde{\mathbf{c}}^{(1)}\|_2^2 \leq \frac{9\eta_1^2 s_i}{\alpha_i}.}$$

In particular, the self-expression prediction error satisfies

$$\boxed{\left\|\mathbf{X}_{\|}^{(1)}(\widehat{\mathbf{c}}_i^{(1)} - \widetilde{\mathbf{c}}^{(1)})\right\|_2^2 \leq \frac{9\eta_1^2 s_i}{\alpha_i} - \eta_2\|\widehat{\mathbf{c}}_i^{(1)} - \widetilde{\mathbf{c}}^{(1)}\|_2^2}$$

Bound on the actual reconstruction loss. Since

$$\widehat{\mathbf{r}}_i^{(1)} = \boldsymbol{\epsilon}_i^\star - \mathbf{X}_{-i}(\widehat{\mathbf{c}}_i^{(1)} - \widetilde{\mathbf{c}}^{(1)}),$$

the triangle inequality gives

$$\|\widehat{\mathbf{r}}_i^{(1)}\|_2 \leq \|\boldsymbol{\epsilon}_i^\star\|_2 + \left\|\mathbf{X}_{\|}^{(1)}(\widehat{\mathbf{c}}_i^{(1)} - \widetilde{\mathbf{c}}^{(1)})\right\|_2.$$

Hence,

$$\boxed{\|\widehat{\mathbf{r}}^{(1)}\|_2^2 \leq \left(\|\boldsymbol{\epsilon}_i^\star\|_2 + 3\eta_1\sqrt{\frac{s_i}{\alpha_i}}\right)^2.}$$

A simpler, but looser, bound is

$$\boxed{\|\widehat{\mathbf{r}}_i^{(1)}\|_2^2 \le 2\|\boldsymbol{\epsilon}_i^\star\|_2^2 + \frac{18\eta_1^2 s_i}{\alpha_i}.}$$

Now, coming back to the bound $\|\mathbf{x}_\| - \mathbf{X}_\|^{(1)}\widehat{c}^{(1)}\|_{\ell_2} \le \sqrt{2\|\boldsymbol{\epsilon}_i^\star\|_2^2 + \frac{18\eta_1^2 s_i}{\alpha_i}}$., when $\eta_1 \gtrsim \epsilon^\star/\sqrt{d_i}$ and $s_i \asymp d_i$, we then have for the term $\|\mathbf{X}^{\star(\ell)\top}\left(\mathbf{x}_\| - \mathbf{X}_\|^{(1)}\widehat{\boldsymbol{c}}^{(1)}\right)\|_\infty \le B$

Similarly, $\mathbf{z}^T\left(\mathbf{x}_\| - \mathbf{X}_\|^{(1)}\widehat{\boldsymbol{c}}^{(1)}\right) \le D\eta_1\sqrt{d_1}$ by applying Lemma 11 and for the two terms involving the projection to orthogonal of subspace S_1, i.e., $\mathbf{X}^{\star(\ell)\top}\left(\mathbf{x}_\perp - \mathbf{X}_\perp^{(1)}\widehat{\boldsymbol{c}}^{(1)}\right)$ and $\mathbf{z}^T\left(\mathbf{x}_\perp - \mathbf{X}_\perp^{(1)}\widehat{\boldsymbol{c}}^{(1)}\right)$ we know there are upper-bounded by some constants involving noise-term, i.e. $\mathbf{X}^{\star(\ell)\top}\left(\mathbf{x}_\perp - \mathbf{X}_\perp^{(1)}\widehat{\boldsymbol{c}}^{(1)}\right) \le E\sigma$ and $\mathbf{z}^T\left(\mathbf{x}_\perp - \mathbf{X}_\perp^{(1)}\widehat{\boldsymbol{c}}^{(1)}\right) \le F\sigma$.

B, D, E, F are some numerical constants. Finally, choosing $\eta_1 \gtrsim 1/\sqrt{d_i}$ bounds all four terms by numerical constants times the noise level, eliminating their dependence on the subspace dimension. When the noise is sufficiently small, the correlations between off-subspace samples and the reconstruction residual are therefore small enough to be suppressed by the threshold $\beta\lambda$, yielding the subspace-preserving condition.

We now turn to the second part of our proof: consider $\mathbf{C}^\star \in \arg\min\{F_\mathbf{U}(\mathbf{C}) : \operatorname{diag}(\mathbf{C}) = \mathbf{0},\ \mathbf{C}_{T^c} = \mathbf{0}\}$, and define $\mathbf{E}^\star = \mathbf{X} - \mathbf{X}\mathbf{C}^\star$ with $F_\mathbf{U}(\mathbf{C})$ as defined in 56, $\beta \gtrsim \frac{1}{\min_{i\neq j}\omega_{i,j}\sqrt{d_i}}$, the resulting optimal solution is subspace-preserving.

Consider the fixed-eigenvector problem

$$\widehat{\mathbf{C}} \in \arg\min_{\operatorname{diag}(\mathbf{C})=\mathbf{0}} \left\{\frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 + \beta P_\mathbf{W}(\mathbf{C})\right\}, \tag{70}$$

with $P_\mathbf{W}$ defined as $\sum_{i,j} w_{ij}|c_{ij}|$ and define

$$\widehat{\mathbf{E}} = \mathbf{X} - \mathbf{X}\widehat{\mathbf{C}}.$$

For any feasible coefficient matrix $\mathbf{C}^0$ satisfying $\operatorname{diag}(\mathbf{C}^0) = \mathbf{0}$, define

$$\mathbf{E}^0 = \mathbf{X} - \mathbf{X}\mathbf{C}^0.$$

Because $\widehat{\mathbf{C}}$ is an optimal solution of equation 56, its objective value cannot exceed the objective value at any feasible comparator $\mathbf{C}^0$. Therefore,

$$\frac{1}{2}\|\mathbf{X} - \mathbf{X}\widehat{\mathbf{C}}\|_F^2 + \beta P_\mathbf{W}(\widehat{\mathbf{C}}) \le \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{C}^0\|_F^2 + \beta P_\mathbf{W}(\mathbf{C}^0). \tag{71}$$

Using

$$\widehat{\mathbf{E}} = \mathbf{X} - \mathbf{X}\widehat{\mathbf{C}}, \qquad \mathbf{E}^0 = \mathbf{X} - \mathbf{X}\mathbf{C}^0,$$

and multiplying equation 71 by 2, we obtain

$$\|\widehat{\mathbf{E}}\|_F^2 + 2\beta P_\mathbf{W}(\widehat{\mathbf{C}}) \le \|\mathbf{E}^0\|_F^2 + 2\beta P_\mathbf{W}(\mathbf{C}^0),$$

Since $P_\mathbf{W}(\widehat{\mathbf{C}}) \ge 0$, dropping this nonnegative term yields

$$\|\widehat{\mathbf{E}}\|_F^2 \le \|\mathbf{E}^0\|_F^2 + 2\beta P_\mathbf{W}(\mathbf{C}^0),$$

Equivalently, because the inequality holds for every feasible $\mathbf{C}^0$,

$$\|\widehat{\mathbf{E}}\|_F^2 \le \inf_{\operatorname{diag}(\mathbf{C})=\mathbf{0}} \left\{\|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 + 2\beta P_\mathbf{W}(\mathbf{C})\right\}. \tag{72}$$

We next derive the residual-correlation bound. Define

$$\ell(\mathbf{C}) = \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2.$$

Its gradient is

$$\begin{aligned}\nabla_{\mathbf{C}}\ell(\mathbf{C}) &= \mathbf{X}^\top(\mathbf{X}\mathbf{C}-\mathbf{X})\\ &= -\mathbf{X}^\top(\mathbf{X}-\mathbf{X}\mathbf{C}).\end{aligned} \tag{73}$$

Evaluating this gradient at $\widehat{\mathbf{C}}$ gives

$$\nabla_{\mathbf{C}}\ell(\widehat{\mathbf{C}}) = -\mathbf{X}^\top\widehat{\mathbf{E}}. \tag{74}$$

For an off-diagonal coefficient $\widehat{c}_{ij}$, the constraint $\mathrm{diag}(\mathbf{C}) = \mathbf{0}$ imposes no restriction. The KKT condition therefore gives

$$0 \in -\left[\mathbf{X}^\top\widehat{\mathbf{E}}\right]_{ij} + \beta w_{ij}\partial|\widehat{c}_{ij}|, \qquad i \neq j. \tag{75}$$

Equivalently, there exists

$$z_{ij} \in \partial|\widehat{c}_{ij}|$$

such that

$$\left[\mathbf{X}^\top\widehat{\mathbf{E}}\right]_{ij} = \beta w_{ij} z_{ij}. \tag{76}$$

The subdifferential of the absolute-value function is

$$\partial|c| = \begin{cases}\{\mathrm{sign}(c)\}, & c \neq 0,\\ [-1,1], & c = 0.\end{cases}$$

Hence,

$$|z_{ij}| \leq 1,$$

and equation 76 implies

$$\left|\left[\mathbf{X}^\top\widehat{\mathbf{E}}\right]_{ij}\right| \leq \beta w_{ij}, \qquad i \neq j.$$

More precisely,

$$\left[\mathbf{X}^\top\widehat{\mathbf{E}}\right]_{ij} = \begin{cases}\beta w_{ij}\,\mathrm{sign}(\widehat{c}_{ij}), & \widehat{c}_{ij} \neq 0,\\ z_{ij}, \quad |z_{ij}| \leq \beta w_{ij}, & \widehat{c}_{ij} = 0.\end{cases} \tag{77}$$

It remains to bound w_{ij}. If $\mathbf{e}_i$ denotes the i-th canonical basis vector in $\mathbb{R}^N$, then

$$\mathbf{u}_i = \mathbf{U}^\top\mathbf{e}_i.$$

Therefore,

$$\begin{aligned}2w_{ij} &= \|\mathbf{u}_i - \mathbf{u}_j\|_2^2\\ &= \left\|\mathbf{U}^\top(\mathbf{e}_i - \mathbf{e}_j)\right\|_2^2\\ &\leq \|\mathbf{U}^\top\|_2^2\|\mathbf{e}_i - \mathbf{e}_j\|_2^2.\end{aligned} \tag{78}$$

Because $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$, we have $\|\mathbf{U}^\top\|_2 = 1$. Moreover,

$$\|\mathbf{e}_i - \mathbf{e}_j\|_2^2 = 2 \qquad \text{for } i \neq j.$$

Consequently,

$$2w_{ij} \leq 2, \qquad \text{and hence} \qquad 0 \leq w_{ij} \leq 1.$$

Combining this with equation 37 gives

$$\max_{i\neq j}\left|\left[\mathbf{X}^\top\widehat{\mathbf{E}}\right]_{ij}\right| \leq \beta \max_{i\neq j} w_{ij} \leq \beta, \tag{79}$$

Let

$$S_j = \mathrm{supp}(\mathbf{c}_j^0), \qquad s_j = |S_j|,$$

and define

$$w_{\max,j} = \max_{i\in S_j} w_{ij}.$$

By the definition of $P_{\mathbf{W}}$,

$$P_{\mathbf{W}}(\mathbf{C}^0) = \sum_{j=1}^{N} \sum_{i \in S_j} w_{ij} |c_{ij}^0|. \tag{80}$$

Applying the Cauchy–Schwarz inequality separately to each column gives

$$\begin{aligned} \sum_{i \in S_j} w_{ij} |c_{ij}^0| &\leq \left(\sum_{i \in S_j} w_{ij}^2 \right)^{1/2} \left(\sum_{i \in S_j} |c_{ij}^0|^2 \right)^{1/2} \\ &\leq w_{\max,j} \sqrt{s_j} \, \|\mathbf{c}_j^0\|_2. \end{aligned} \tag{81}$$

Therefore,

$$P_{\mathbf{W}}(\mathbf{C}^0) \leq \sum_{j=1}^{N} w_{\max,j} \sqrt{s_j} \, \|\mathbf{c}_j^0\|_2.$$

Substituting this inequality into equation 51 proves that:

$$\|\widehat{\mathbf{E}}\|_F^2 \leq \|\mathbf{E}^0\|_F^2 + 2\beta \sum_{j=1}^{N} w_{\max,j} \sqrt{s_j} \, \|\mathbf{c}_j^0\|_2. \tag{82}$$

The remaining bounds follow from $s_j \leq d_j$ and $w_{\max,j} \leq 1$ and hence:

$$\|\widehat{\mathbf{E}}\|_F^2 \leq \|\mathbf{E}^0\|_F^2 + 2\beta \sum_{j=1}^{N} w_{\max,j} \sqrt{d_j} \, \|\mathbf{c}_j^0\|_2. \tag{83}$$

Since $w_{\max,j} \leq 1$, this further implies

$$\|\widehat{\mathbf{E}}\|_F^2 \leq \|\mathbf{E}^0\|_F^2 + 2\beta \sum_{j=1}^{N} \sqrt{d_j} \, \|\mathbf{c}_j^0\|_2. \tag{84}$$

Now we turn to the full residual-correlation bound $\|\mathbf{X}^\top \widehat{\mathbf{E}}\|_\infty$ Let $\widehat{\mathbf{e}}_j$ denote the j-th column of $\widehat{\mathbf{E}}$. By the Cauchy–Schwarz inequality,

$$\begin{aligned} \left| \left[\mathbf{X}^\top \widehat{\mathbf{E}} \right]_{ij} \right| &= |\mathbf{x}_i^\top \widehat{\mathbf{e}}_j| \\ &\leq \|\mathbf{x}_i\|_2 \|\widehat{\mathbf{e}}_j\|_2 \\ &\leq \|\mathbf{X}\|_{2,\infty} \|\widehat{\mathbf{E}}\|_F. \end{aligned} \tag{85}$$

Taking the maximum over i, j gives

$$\|\mathbf{X}^\top \widehat{\mathbf{E}}\|_\infty \leq \|\mathbf{X}\|_{2,\infty} \|\widehat{\mathbf{E}}\|_F.$$

Using the oracle reconstruction bound

$$\|\widehat{\mathbf{E}}\|_F^2 \leq \|\mathbf{E}^0\|_F^2 + 2\beta P_{\mathbf{W}}(\mathbf{C}^0)$$

and taking square roots and assuming data points are normalized such that $\|\mathbf{x}_i\|_2 = 1$, gives:

$$\|\mathbf{X}^\top \widehat{\mathbf{E}}\|_\infty \leq \sqrt{\|\mathbf{E}^0\|_F^2 + 2\beta P_{\mathbf{W}}(\mathbf{C}^0)}. \tag{86}$$

We let $\mathbf{C}^0$ as the minimizing the lasso solution to the problem:

$$\mathbf{C}^0 \in \arg \min_{\operatorname{diag}(\mathbf{C}) = \mathbf{0}} \left\{ \frac{1}{2} \|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 + \beta \omega_{\min}(\mathbf{C}) \right\}, \tag{87}$$

If $\|\mathbf{x}_i\|_2 = 1$ for every i, then $\|\mathbf{X}\|_{2,\infty} = 1$, which gives equation 86.

Suppose that $\mathbf{C}^\star$ is the restricted solution in Lemma 3, and let

$$\mathbf{E}^\star = \mathbf{X} - \mathbf{X}\mathbf{C}^\star.$$

Once the strict cross-subspace condition

$$\left|\left[\mathbf{X}^\top \mathbf{E}^\star\right]_{ij}\right| < \beta w_{ij}, \qquad (i,j) \in \Omega^c,$$

is satisfied, $\mathbf{C}^\star$ is also a global minimizer. Consequently, all the preceding oracle and KKT bounds apply with

$$\widehat{\mathbf{C}} = \mathbf{C}^\star, \qquad \widehat{\mathbf{E}} = \mathbf{E}^\star.$$

In particular,

$$\left\|\left(\mathbf{X}^\top \mathbf{E}^\star\right)_{\Omega^c}\right\|_\infty < \beta \max_{(i,j)\in\Omega^c} w_{ij} \le \beta.$$

Alternatively, suppose that an independent geometric or probabilistic argument gives

$$\left\|\left(\mathbf{X}^\top \mathbf{E}^\star\right)_{\Omega^c}\right\|_\infty \le \rho.$$

Define

$$w_{\min}^{\text{cross}} = \min_{(i,j)\in\Omega^c} w_{ij}.$$

If $w_{\min}^{\text{cross}} > 0$, the sufficient condition

$$\beta > \frac{\rho}{w_{\min}^{\text{cross}}}$$

implies

$$\left|\left[\mathbf{X}^\top \mathbf{E}^\star\right]_{ij}\right| < \beta w_{ij} \qquad \text{for every } (i,j) \in \Omega^c.$$

If, in addition,

$$\rho \le \frac{m}{\sqrt{d}},$$

then it is sufficient to choose

$$\beta > \frac{m}{w_{\min}^{\text{cross}}\sqrt{d}}.$$

□

J D.1 Experiment Results

J.1 Extended Experiment Setup

Expanded discussion of ARI, NMI, SDE, Silhouette Score, and SAPD

Definition 9 (Subspace discovery error (SDE))**.** *Without loss of generality, we assume samples from the same clusters share the same subspaces and samples on different subspaces do not originate from the same clusters. Further, let Y denote the class labels, then the Subspace discovery error is defined as;*

$$\mathcal{L}_{\text{sde}}(\mathbf{C}, \mathbf{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} |C_{ij}| \cdot \mathbb{1}_{\{\mathbf{Y}_i \neq \mathbf{Y}_j\}}. \tag{88}$$

Here, we describe our hardware configuration and report the runtime on a representative dataset. We run PRO-DSC on a worker equipped with eight NVIDIA Quadro RTX 8000 GPUs and 512 GB of system memory. For the CIFAR-100 CLIP features as an example, we apply early stopping when the relative training loss does not improve by more than 0.1% for 20 consecutive epochs, with a maximum of 2,000 epochs. Averaged over 10 random seeds, PRO-DSC requires $2{,}445.876$ seconds per run. In comparison, SENet requires $5{,}683.243$ seconds for a single run on the same dataset using a subsample of 3,750 instances, corresponding to approximately $1/16$ of the full dataset. Moreover, its training time grows quadratically with the subsample size.

Figures 4 and 5 show that DiHypeSC performs label-free selection using a theoretically motivated dependence on subspace dimension and a transferable constant calibrated once on a reference dataset. We reuse the constant from Soltanolkotabi et al. (2014b), originally derived for ℓ_1

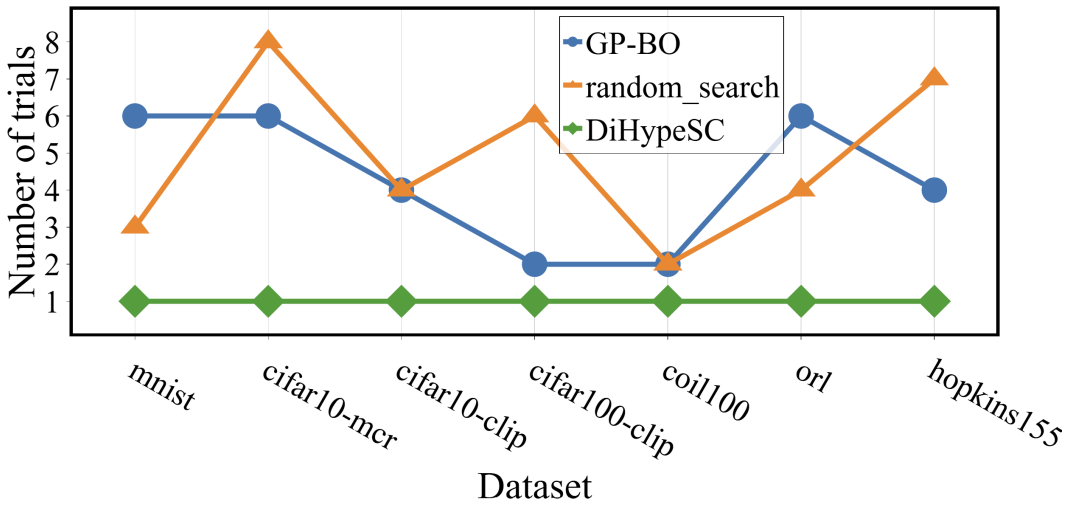


Figure 4: Comparisons of Iterations achieving its best result when running SeNet

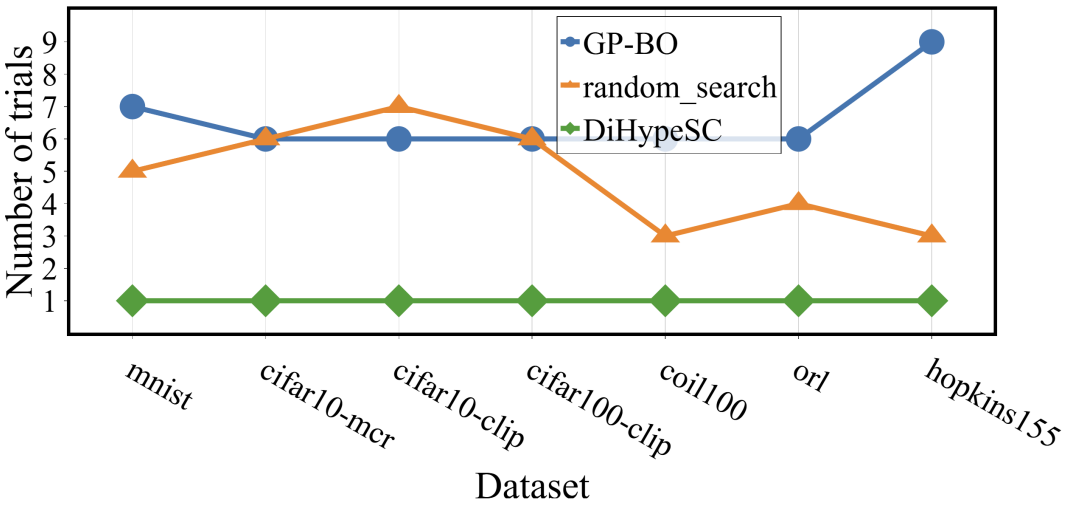


Figure 5: Comparisons of Iterations achieving its best result when running PRO-DSC

Table 3: Experimental results evaluated using nmi ↑

Method	MNIST	CIFAR-10 MCR	CIFAR-10 CLIP	CIFAR-100 CLIP	COIL-100	ORL	Hopkins155
SeNet	0.819	0.551	0.796	0.741	0.654	0.586	0.598
+ R-Search	±0.026	±0.012	±0.022	±0.009	±0.008	±0.017	±0.332
SeNet	0.795	0.533	0.787	0.739	0.657	0.570	**0.602**
+ GP-BO	±0.036	±0.019	±0.016	±0.012	±0.010	±0.028	**±0.337**
SeNet	**0.838**	**0.628**	**0.841**	**0.747**	**0.931**	**0.600**	0.593
+Ours	**±0.022**	**±0.003**	**±0.028**	**±0.005**	**±0.001**	**±0.033**	±0.322
PRO-DSC	0.931	**0.563**	0.919	0.794	0.903	0.852	0.518
+R-Search	±0.016	**±0.027**	±0.012	±0.006	±0.059	±0.060	±0.324
PRO-DSC	0.934	0.543	0.906	0.778	0.899	0.792	0.507
+GP-BO	±0.013	±0.013	±0.032	±0.027	±0.053	±0.055	±0.330
PRO-DSC	**0.942**	0.560	**0.926**	**0.796**	**0.955**	**0.886**	**0.518**
+Ours	**±0.011**	±0.013	**±0.002**	**±0.005**	**±0.011**	**±0.063**	**±0.327**

Table 4: Experimental results evaluated using silhouette score ↑

Method	MNIST	CIFAR-10 MCR	CIFAR-10 CLIP	CIFAR-100 CLIP	COIL-100	ORL	Hopkins 155
SeNet	0.0904	0.1381	**0.0920**	0.0382	-0.1120	**−0.0269**	0.2294
+ R-Search	±0.0004	±0.0223	**±0.0033**	±0.0003	±0.0197	**±0.0009**	±0.0903
SeNet	**0.0940**	**0.1387**	0.0906	**0.0388**	-0.1063	-0.0276	**0.2297**
+GP-BO	**±0.0028**	**±0.0209**	±0.0036	**±0.0017**	±0.0241	±0.0144	**±0.0905**
SeNet	0.0752	0.0636	0.0768	0.0366	**0.0685**	-0.0360	0.2243
+Ours	±0.0000	±0.0182	±0.0053	±0.0020	**±0.0001**	±0.0139	±0.0089
PRO-DSC	0.1012	**0.1014**	0.0946	0.0416	0.1225	**0.0878**	**0.5112**
+R-Search	±0.0032	**±0.0044**	±0.0014	±0.0024	±0.0035	**±0.0092**	±0.1038
PRO-DSC	0.1020	0.0970	**0.0956**	0.0389	**0.1268**	0.0774	0.5081
+GP-BO	±0.0017	±0.0054	**±0.0023**	±0.0064	**±0.0027**	±0.0092	±0.1056
PRO-DSC	**0.1027**	0.0943	0.0939	**0.0434**	0.1148	0.0837	0.4773
+Ours	**±0.0009**	±0.0058	±0.0001	**±0.0017**	±0.0073	±0.0195	±0.1279

regularization, and extend the principle to the ℓ_1-dominated elastic-net regime and block-diagonal regularization. In contrast, the two generic AutoML baselines typically require multiple iterations to attain their best internal validation scores.

Using the generic numerical constant $m = 0.25$ for the fixed-eigenvector block-diagonal regularizer and $m = 1$ in the other regimes, our method requires only a **single run** to achieve competitive performance. In contrast, generic unsupervised hyperparameter-selection methods require multiple trials to identify the configuration with the highest silhouette score, yet the resulting adjusted Rand index (ARI) may still be poor.

In the table below, we further compare their performance in terms of normalized mutual information (NMI), subspace discovery error, and silhouette score. As discussed earlier, the silhouette score does not always align with the ground-truth clustering performance.

We further evaluate the baselines in a one-shot setting. As shown in Table 6, our method consistently outperforms them.

J.2 Sensitivity analysis for constant factors

J.3 Runtime analysis

Table 5: Experimental results evaluated using subspace discovery error ↓

Method	MNIST	CIFAR-10 MCR	CIFAR-10 CLIP	CIFAR-100 CLIP	COIL-100	ORL	Hopkins 155
SeNet + Random	**0.146 ±0.014**	0.859 ±0.024	0.837 ±0.030	0.855 ±0.001	0.923 ±0.001	0.976 ±0.000	0.179 ±0.118
SeNet +GP-BO	0.150 ±0.011	0.877 ±0.009	0.830 ±0.045	0.854 ±0.008	0.923 ±0.000	0.976 ±0.000	**0.177 ±0.115**
SeNet +Ours	0.168 ±0.013	**0.307 ±0.008**	**0.159 ±0.003**	**0.846 ±0.005**	**0.075 ±0.008**	**0.975 ±0.001**	0.220 ±0.090
PRO-DSC +Random	0.560 ±0.020	0.888 ±0.001	0.487 ±0.034	0.823 ±0.012	0.174 ±0.306	0.748 ±0.159	0.337 ±0.172
PRO-DSC +GP-BO	0.570 ±0.015	0.888 ±0.002	0.502 ±0.058	0.817 ±0.026	0.189 ±0.301	0.886 ±0.144	0.346 ±0.188
PRO-DSC +Ours	**0.559 ±0.012**	**0.887 ±0.001**	**0.431 ±0.003**	**0.689 ±0.026**	**0.066 ±0.015**	**0.688 ±0.152**	**0.279 ±0.106**

Table 6: Experimental results evaluated using ARI (↑; mean ± standard deviation), all under one-shot

Method	MNIST	CIFAR-10 MCR	CIFAR-10 CLIP	CIFAR-100 CLIP	COIL-100	ORL	Hopkins 155
SeNet + R-Search	0.764 ±0.065	0.252 ±0.010	0.640 ±0.014	**0.472 ±0.027**	0.437 ±0.024	0.146 ±0.035	0.460 ±0.366
SeNet + GP-BO	0.743 ±0.063	0.265 ±0.036	0.576 ±0.095	0.453 ±0.045	0.437 ±0.024	0.146 ±0.035	0.453 ±0.371
SeNet +Ours	**0.794 ±0.052**	**0.510 ±0.005**	**0.792 ±0.055**	0.439 ±0.021	**0.554 ±0.161**	**0.187 ±0.040**	**0.490 ±0.370**
PRO-DSC + R-Search	0.675 ±0.162	0.385 ±0.059	0.793 ±0.181	0.329 ±0.144	0.747 ±0.118	0.487 ±0.113	0.328 ±0.332
PRO-DSC +GP-BO	0.764 ±0.096	0.352 ±0.092	0.787 ±0.190	0.381 ±0.161	0.741 ±0.117	0.472 ±0.119	0.386 ±0.320
PRO-DSC +Ours	**0.950 ±0.012**	**0.476 ±0.004**	**0.935 ±0.002**	**0.579 ±0.010**	**0.830 ±0.020**	**0.684 ±0.130**	**0.563 ±0.351**

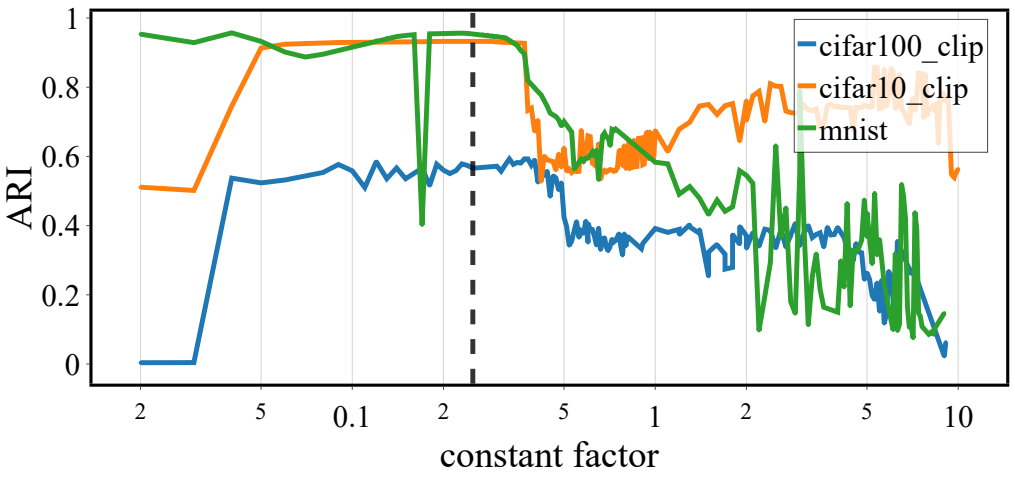


Figure 6: Sensitivity analysis of the constant factor m' under the mode $\widehat{\beta} = \widehat{\beta}_{\mathrm{BDR}}$ in Eq. 6. The vertical dashed line indicates $m' = 0.25$ for $\widehat{\beta}_{\mathrm{BDR}}$, which achieves near-optimal ARI performance.

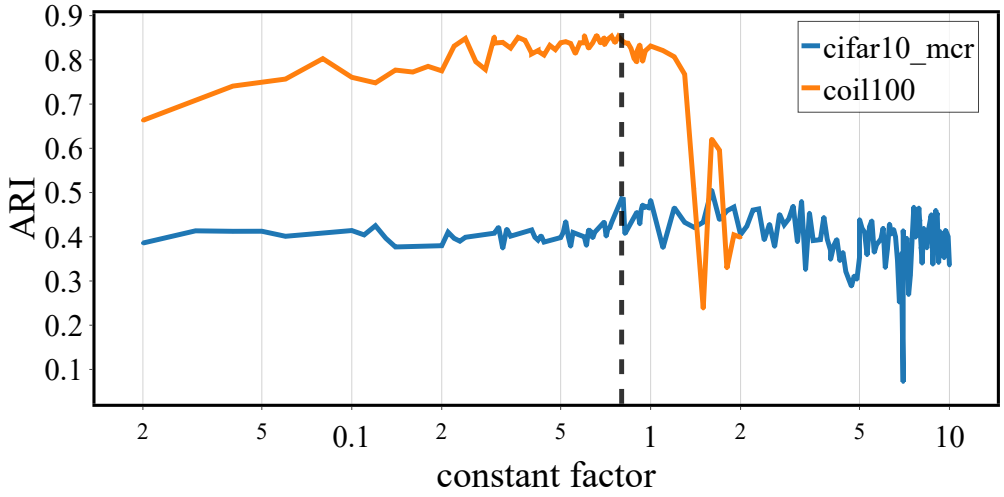


Figure 7: Sensitivity analysis of the constant factor m' under the mode $\widehat{\beta} = \widehat{\beta}_{\ell_1}$ or $\widehat{\beta} = \widehat{\beta}_{\mathrm{PCA}}$ in Eq. 6. The vertical dashed line indicates $m' = 0.8$ for $\widehat{\beta}$, which achieves near-optimal ARI performance.

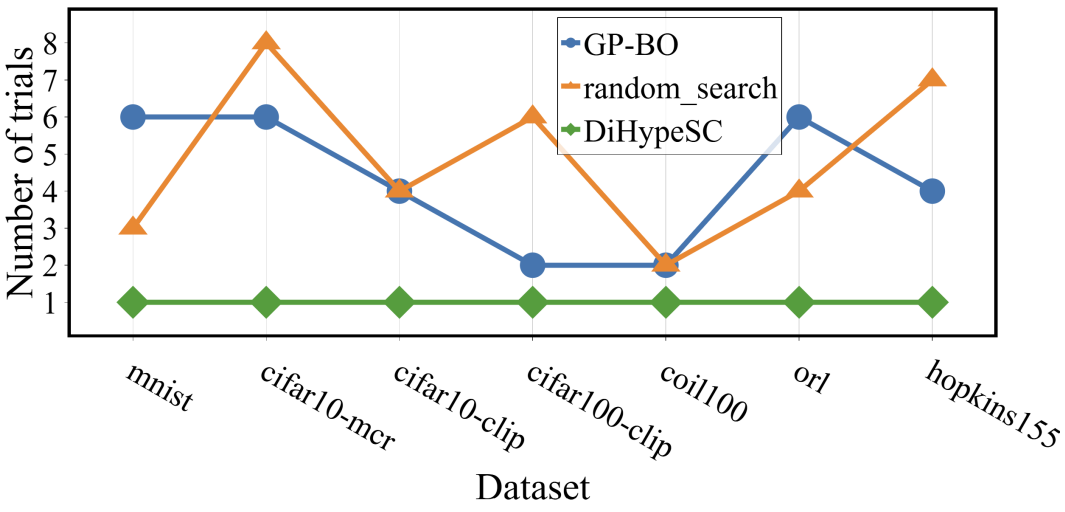


Figure 8: Comparisons of Iterations achieving its best result when running SeNet

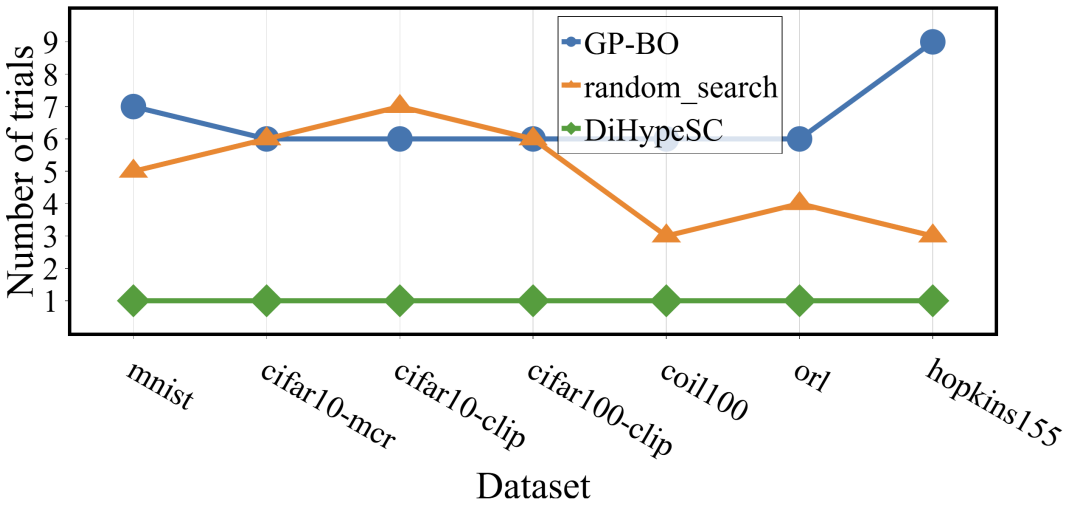


Figure 9: Comparisons of Iterations achieving its best result when running PRO-DSC