DOWNSAMPLING FOR SPARSE SUBSPACE CLUSTERING

Xianghui Mao, Xiaohan Wang, and Yuantao Gu
Department of Electronic Engineering, Tsinghua University, Beijing, China
Email: {maoxh14, wangxiaohan11}@mails.tsinghua.edu.cn, gyt@tsinghua.edu.cn

Background & Motivation

- **Sparse Subspace clustering (SSC)**
  - **Cluster samples lying on a union of subspaces**
- **Problem:**
  - **Recovering subspaces from huge amount of samples without labels is time-consuming**
  - **Sample amounts are often unbalanced in different subspaces**
- **Our solution: SSC with downsampling (SSCD)**

Basic Idea

- **Notations:** $\mathbf{x}_i$, **the** i^{th} **sample;** l_i, **the label of the subspace which** $\mathbf{x}_i$ **belongs to;** N_l, **amount of samples in the** l^{th} **subspace;** p_i, **downsampling probability of** $\mathbf{x}_i$.
- p_i **should be in reverse ratio to** N_{l_i}.
- **Mutual linear representation**

$$\hat{\mathbf{c}}_i = \operatorname{argmin}\|\mathbf{c}\|_1, \quad s.t. \quad \mathbf{X}_i\mathbf{c} = \mathbf{x}_i, \quad \forall i \in \{1, \cdots, N\} \qquad (1)$$

- $\|\hat{\mathbf{c}}_i\|_1$ **is closely related to** N_{l_i}:
  - **Lemma [3]. Assuming that** $B^{(d)}$ **is the Euclidean ball in** $\mathcal{R}^d$, **and** A_N **is a random inscribed polytope with all of its** N **vertices independently uniformly chosen on the boundary of** $B^{(d)}$, **i.e.,** S^{d-1}, **one has**

$$\mathbb{E}\{Vol_d(A_N)\} = Vol_d(B^{(d)}) - (c_{B^{(d)}} + o(1))N^{-\frac{2}{d-1}},$$

where $Vol_d(\cdot)$ **denotes the** d**-dimensional volume and** $c_{B^{(d)}}$ **is a constant depending on** d.

Theoretical Analysis: Balancing Guarantee

- **Atomic norm and volume of polytope**
  - **Proposition1. For a deterministic polytope** $conv(\mathcal{A})$ **with all vertices fixed on** S^{d-1} **and a random variable** $\mathbf{x} \in \mathcal{R}^d$ **satisfies a uniform distribution on** S^{d-1}, **one has**

$$\mathbb{E}\left\{\frac{1}{\|\mathbf{x}\|_{\mathcal{A}}^d}\right\} = \frac{Vol_d(conv(\mathcal{A}))}{Vol_d(B^{(d)})},$$

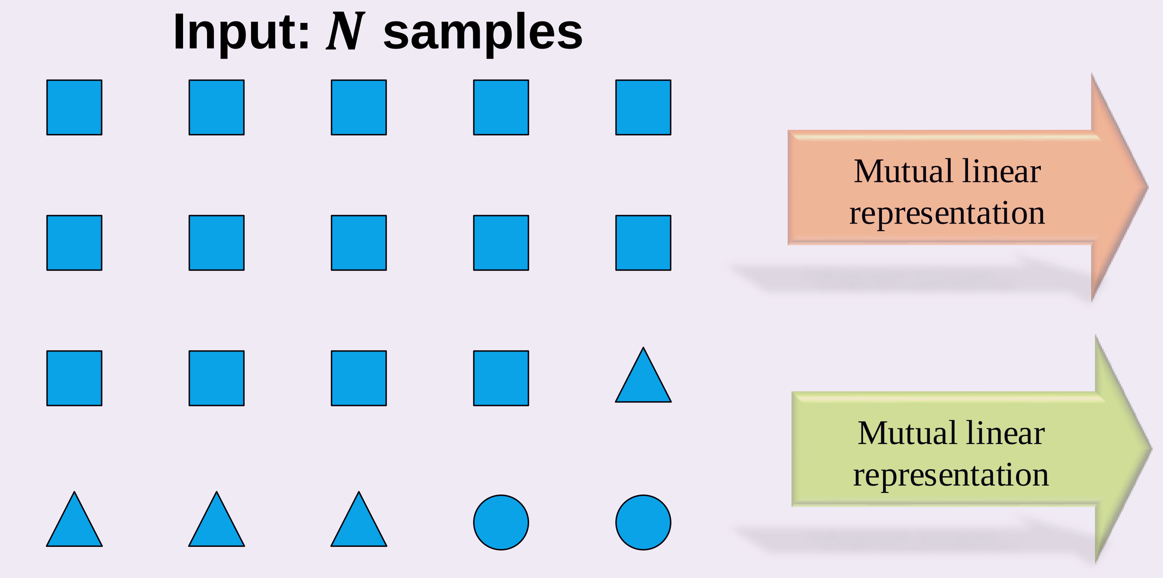
where $B^{(d)} = \{\mathbf{x} \in \mathcal{R}^d \,|\, \|\mathbf{x}\|_2 \le 1\}$.

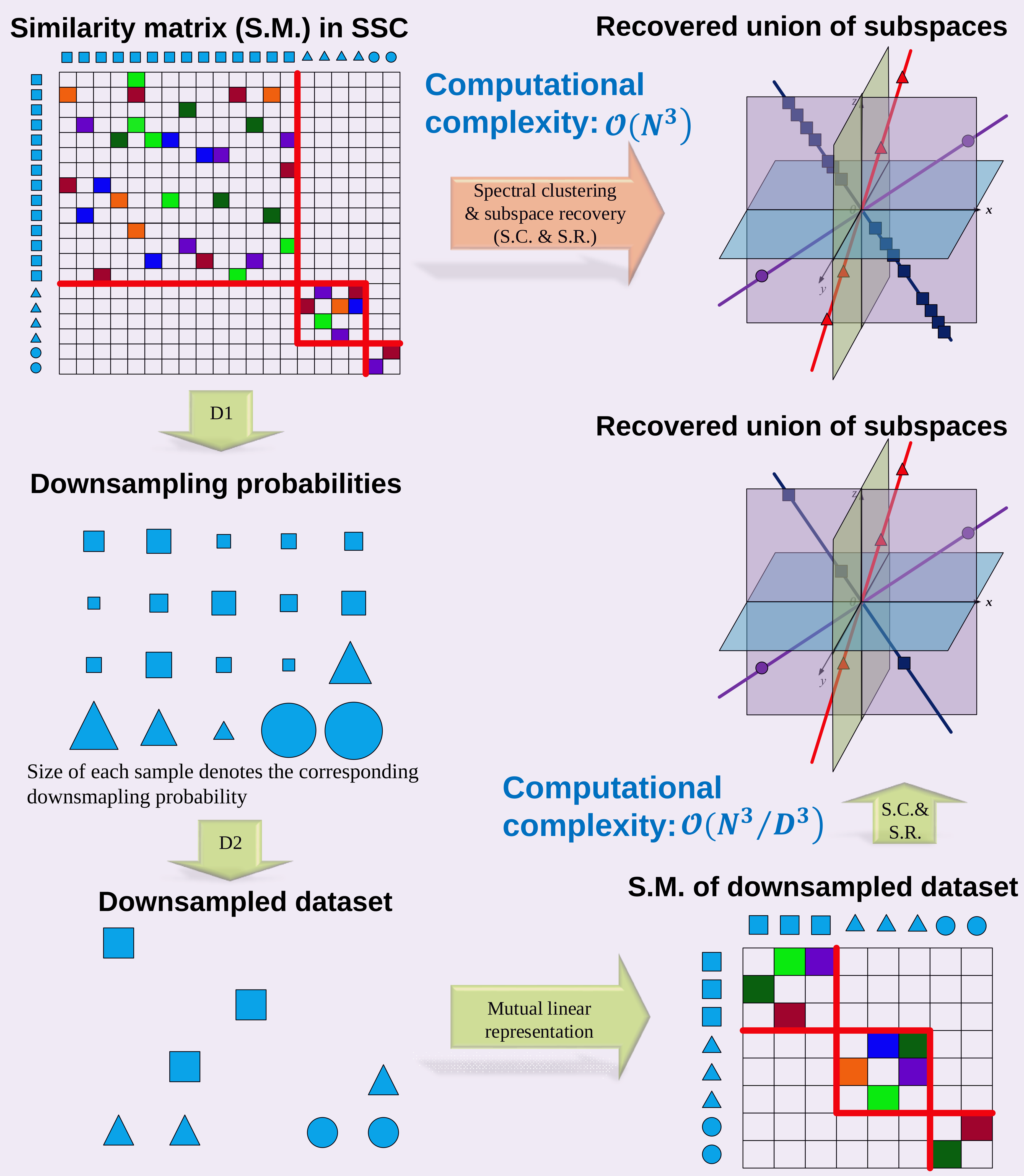
- **Estimation of the amount of samples in the** l^{th} **subspace**
  - **Proposition2.** N_l **can be approximated by**

$$\widehat{N}_l = C_d\left(1 - \mathbb{E}\left\{\frac{1}{\|\hat{\mathbf{c}}_i\|_1^d} \,|\, \mathbf{x}_i \in U_l\right\}\right)^{-\frac{d-1}{2}},$$

where C_d **is a constant only depending on** d.

SSC versus SSCD





Purpose of Downsampling:

- **To Balance the amounts of samples in different subspaces**
- **To Reduce the computational complexity**

Downsampling strategy:

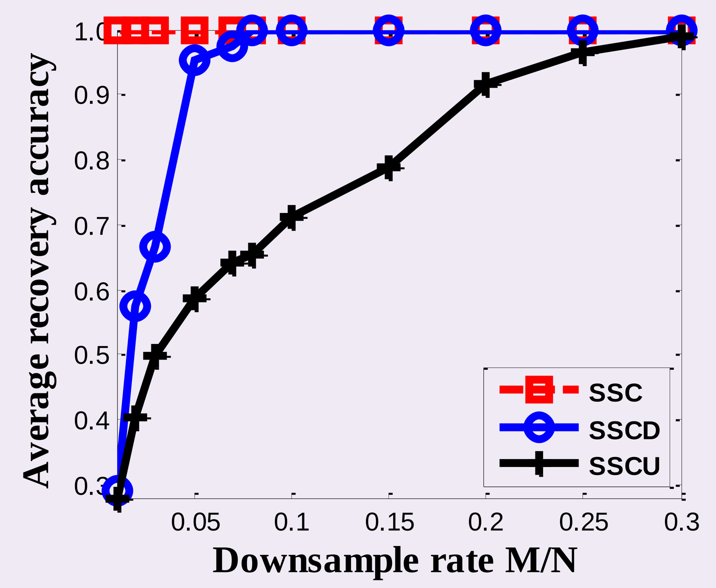
D1: Assingn downsampling probabilities according to the optimal value of (1). The downsampling probability of the i^{th} **sample**

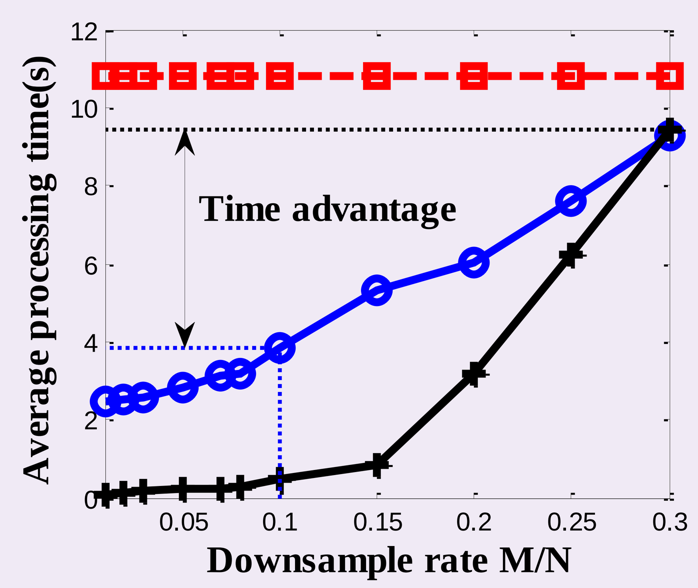
$$p_i = w_i / \sum_{j=1}^{N} w_j,$$

where $w_i = \left(1 - 1/\|\hat{\mathbf{c}}_i\|_1^d\right)^{\frac{d-1}{2}}$.

D2: With downsampling ratio D, **randomly choose** N/D **different samples from the original sample set according to the** p_i**s as assigned in step D1.**

Simulation Results





References

[1] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in IEEE Conference on Computer Vision and Pattern Recognition, pp. 2790–2797, 2009.
[2] B. Cheng, J.C. Yang, S.C. Yan, Y. Fu, and T.S. Huang, "Learning with ℓ_1-Graph for Image Analysis," IEEE Trans. Image Processing vol. 19, no. 4, pp. 858-866, Apr. 2010.
[3] E. W. Carsten Schütt, Polytopes with Vertices Chosen Randomly from the Boundary of a Convex Body. Geometric Aspects of Functional Analysis, 2003.
[4] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," Foundations of Computational Mathematics, vol. 12, no. 6, pp. 805–849, 2012.