压缩数据集上子空间聚类的离群点检测

2020 年电子系科创年会 · 中文版

刘坤瓒

清华大学电子工程系

更新于 2020 年 5 月 17 日

清華大學 Tsinghua University

总览

什么是子空间聚类？[1] [2]

图：运动分割。左：一帧图像中待聚类的点；中：聚类结果；右：子空间的并集模型，属于同一类的点分布于同一个低维子空间附近。

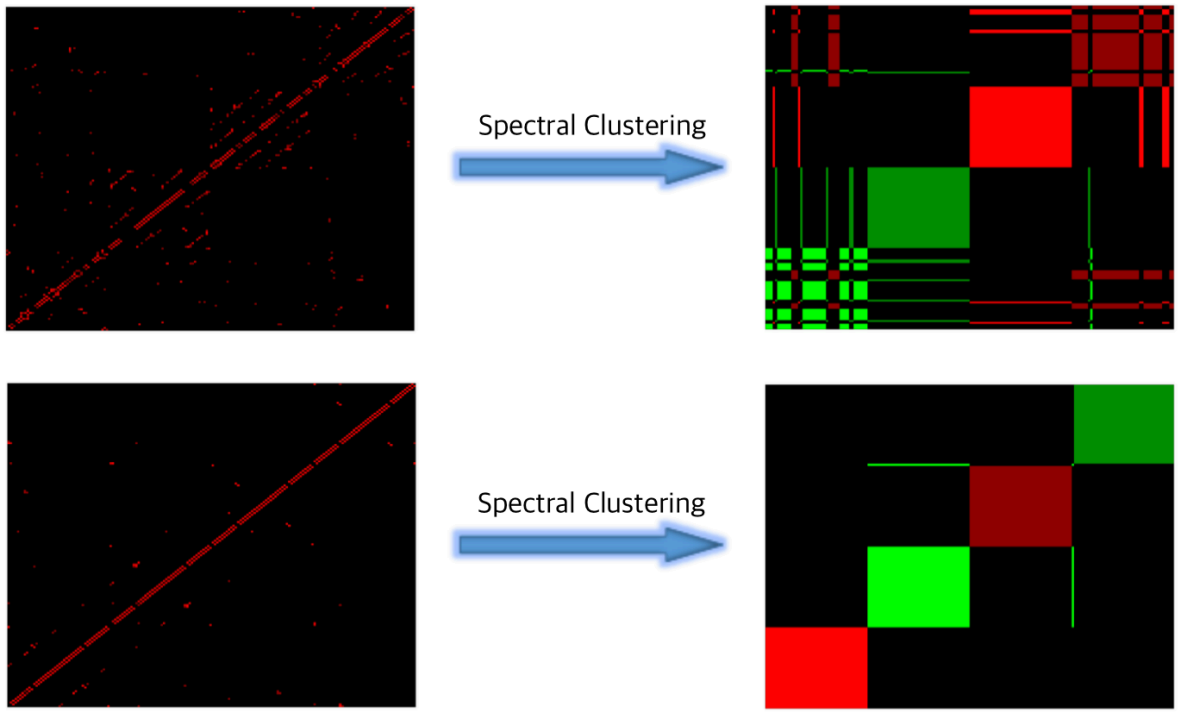
[1] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765–2781, 2013.

[2] M. Soltanolkotabi and E. J. Candès, "A geometric analysis of subspace clustering with outliers," *Annals of Statistics*, vol. 40, no. 4, pp. 2195–2238, 2012.

什么是离群点检测？

图: 然而，在真实数据集中往往存在大量离群点（a.k.a 野点），它们不属于任何一个子空间或表现出任何低维特征，在进一步数据处理前应被筛除。

为什么要做离群点检测？[3]



图：在有噪（上）和无噪（下）数据集上的聚类。左：相似度矩阵，即数据集情况；右：聚类结果。

[3] Y. Wang, Y. Wang and A. Singh, "Graph connectivity in noisy sparse subspace clustering," *In Artificial Intelligence and Statistics*, pp. 538-546, 2016.

现存方法存在哪些问题？[4]

SOTA 工作
基于自表示的离群点检测

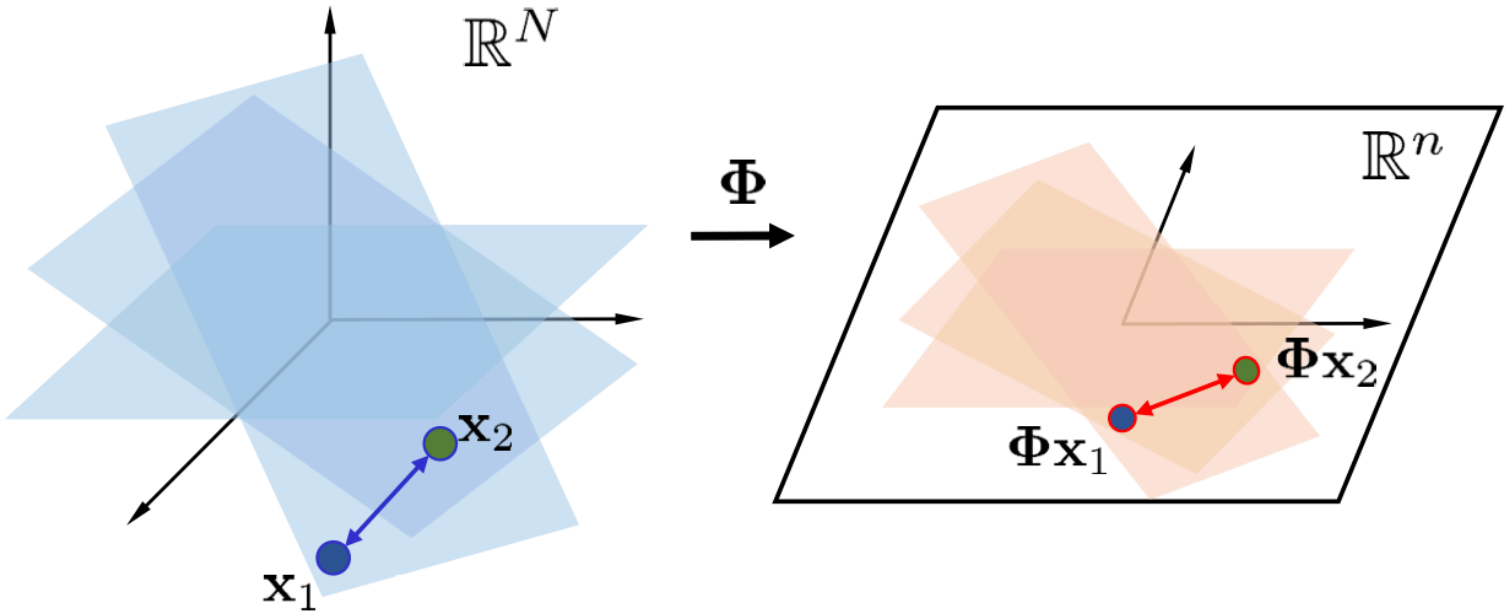
- 已经在真实数据集上获得了较高的准确率
- 效率会随着数据维数的增长而大大降低

本工作
压缩离群点检测

- 保持住真实数据集上的检测准确率
- 大幅降低计算成本提高算法效率

[4] C. You, D. P. Robinson, and R. Vidal, "Provable self-representation based outlier detection in a union of subspaces," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4323–4332, 2017.

如何降低运算成本？[5]



图：随机投影将现有的高维空间 $\mathbb{R}^N$ 投影到低维空间 $\mathbb{R}^n$，它可以保证投影前后任两个点距离几乎不变（a.k.a JL 引理）、任两个子空间距离几乎不变（a.k.a RIP 性质）。

[5] Y. Jiao, G. Li, and Y. Gu, "Principal angles preserving property of Gaussian random projection for subspaces," *IEEE Global Conference on Signal and Information Processing*, pp. 318-322, 2017.

算法提出

Algorithm 1 压缩离群点检测

输入：数据集 $\boldsymbol{X} \in \mathbb{R}^{N \times M}$，压缩后维数 n，迭代次数 T，阈值 ζ，其他参数

Step I: 随机投影

Step II: 自表示【时间瓶颈】

Step III: 随机游走

输出：离群点 $\mathbb{I}(\bar{\pi}_j < \zeta)$

$$\underbrace{\boldsymbol{r}_j}_{\text{自表示向量}} = \arg\min_{\boldsymbol{c}} \underbrace{\lambda\|\boldsymbol{c}\|_1 + \frac{1-\lambda}{2}\|\boldsymbol{c}\|_2^2}_{\text{稀疏性}} + \underbrace{\frac{\gamma}{2}\|\boldsymbol{\Phi}\boldsymbol{x}_j - \boldsymbol{Y}\boldsymbol{c}\|_2^2}_{\text{误差}}$$

如何证明算法优势？

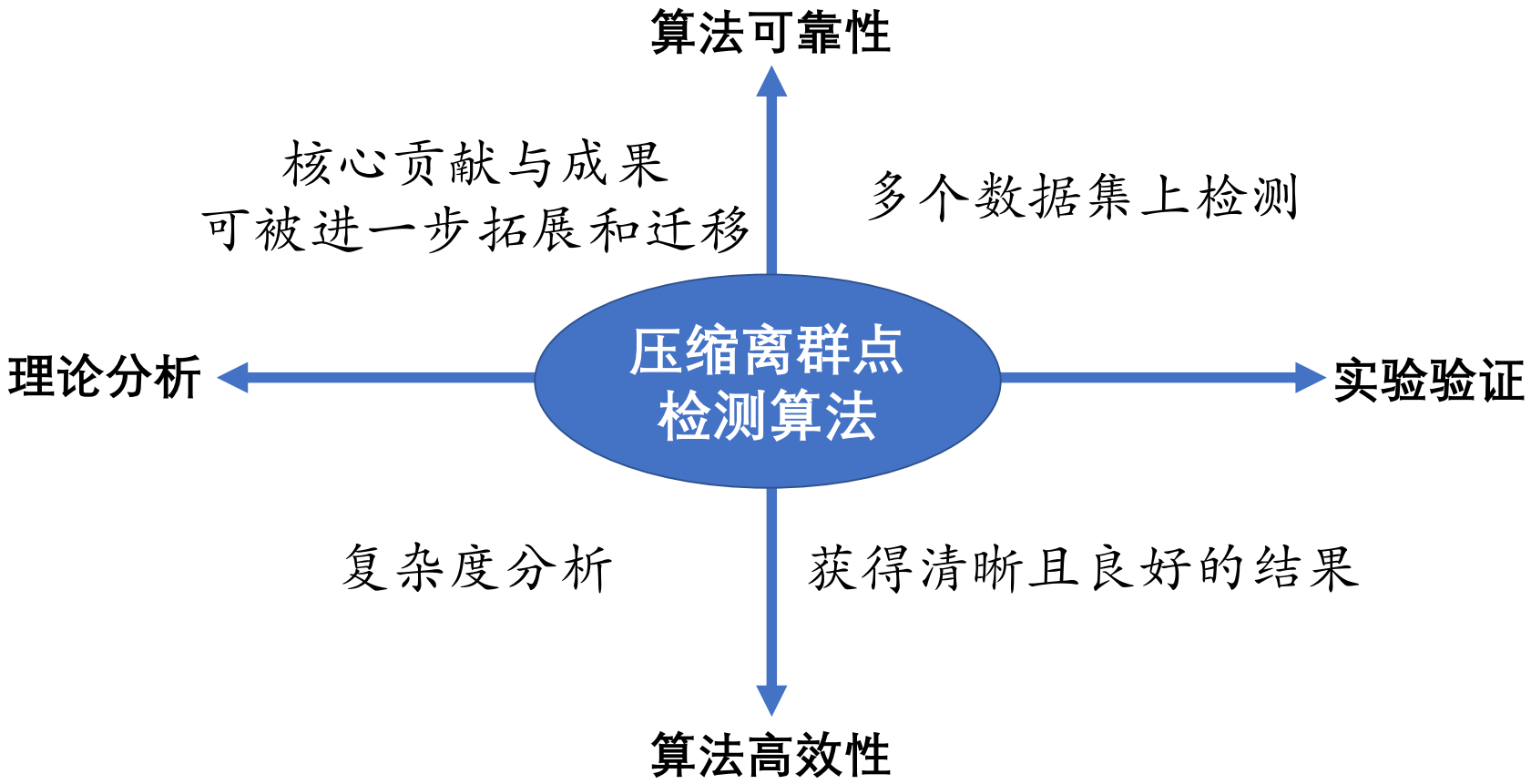
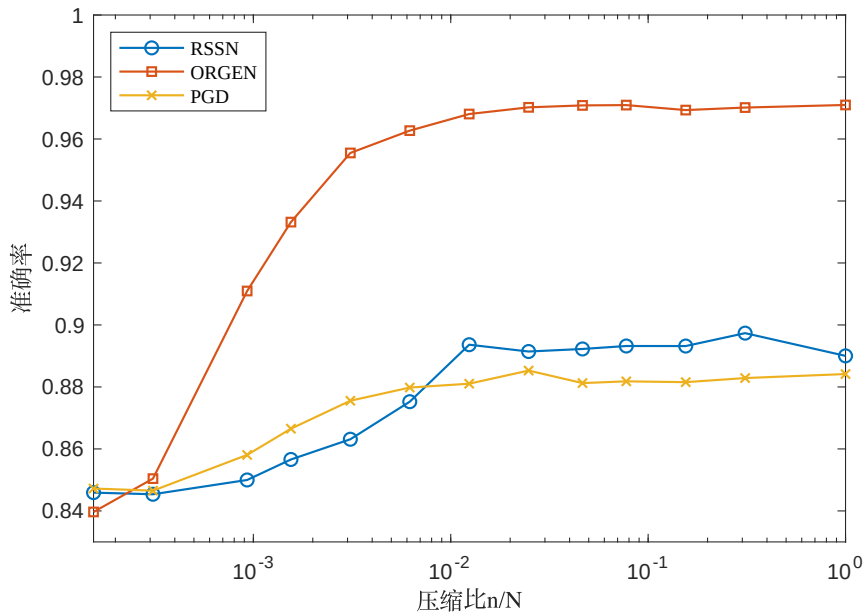


图: 从理论与实验论证可靠性与高效性两个维度

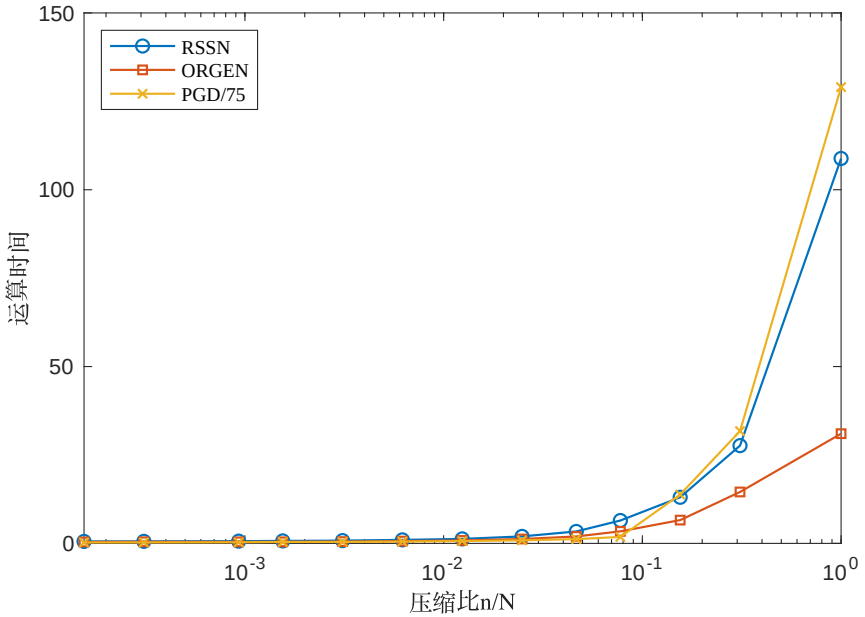
实验验证

262 个数据中包含 70 个离群点，数据维数是 32256

$$\arg\min_{\boldsymbol{c}} \lambda\|\boldsymbol{c}\|_1 + \frac{1-\lambda}{2}\|\boldsymbol{c}\|_2^2 + \frac{\gamma}{2}\|\boldsymbol{\Phi x}_j - \boldsymbol{Y c}\|_2^2$$



(a) 可靠性：准确率



(b) 高效性：运算时间

理论分析

记号 N：原始数据维数，n：压缩后数据维数，M：数据量
定义 $\varepsilon \to 0$ 表示小量，c_1 和 c_2 是大于 0 的常量

	SOTA 工作	本工作
算法	基于自表示的离群点检测	压缩离群点检测
假设	连通性假设	连通性假设
条件	$\max_j \max_{k\neq j, \boldsymbol{x}_k \in \mathcal{S}_{\ell_j}} \lvert\langle \boldsymbol{x}_k, \boldsymbol{\delta}_j\rangle\rvert < \lambda$	$\max_j \max_{k\neq j, \boldsymbol{x}_k \in \mathcal{S}_{\ell_j}} \lvert\langle \boldsymbol{x}_k, \boldsymbol{\delta}_j\rangle\rvert < \lambda - \tau$
附加条件		$n > \max\left(c_1\varepsilon^{-2}, \ln M\right)$
成功率	1	$1 - \exp\left(-c_2\varepsilon^2 n\right) \to 1$
复杂度	$\mathcal{O}\left(N^2M^3\right)$	$\mathcal{O}\left(n^2M^3\right)$

$$\tau := \max_j \left\{ \varepsilon\left(1 + \|\boldsymbol{\delta}_j\|_2^2\right) + \sqrt{6\varepsilon}\,(1+\varepsilon)\,\gamma\,\|\boldsymbol{\delta}_j\|_2 \right\} \to 0$$

整体思路

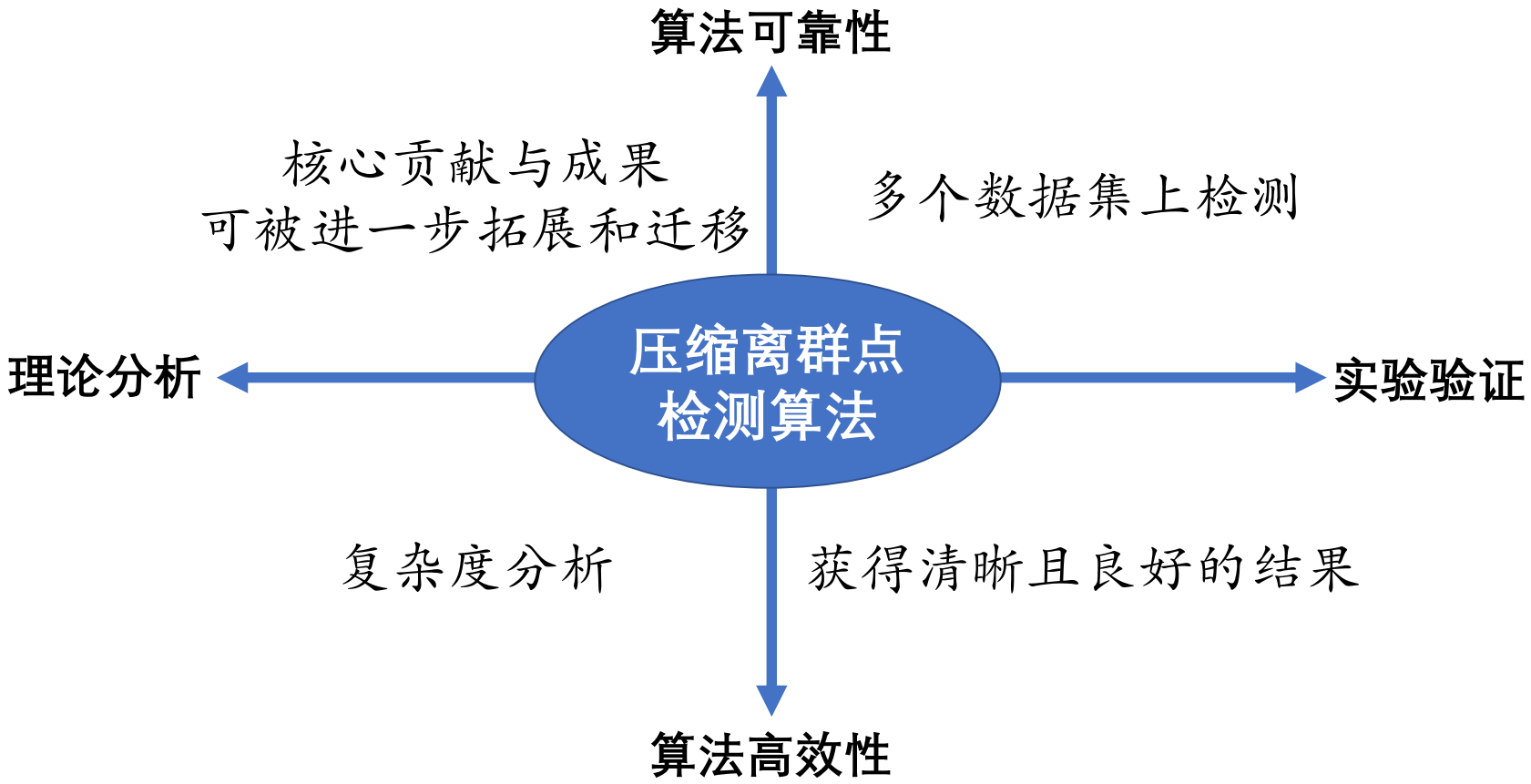


图: 从理论与实验考查可靠性与高效性两个维度

压缩数据集上子空间聚类的离群点检测

2020 年电子系科创年会 · 中文版

刘坤瓒

清华大学电子工程系

更新于 2020 年 5 月 17 日

感谢倾听，恳请指正！

联系方式：lkz18@mails.tsinghua.edu.cn

Step I: Compression[6]

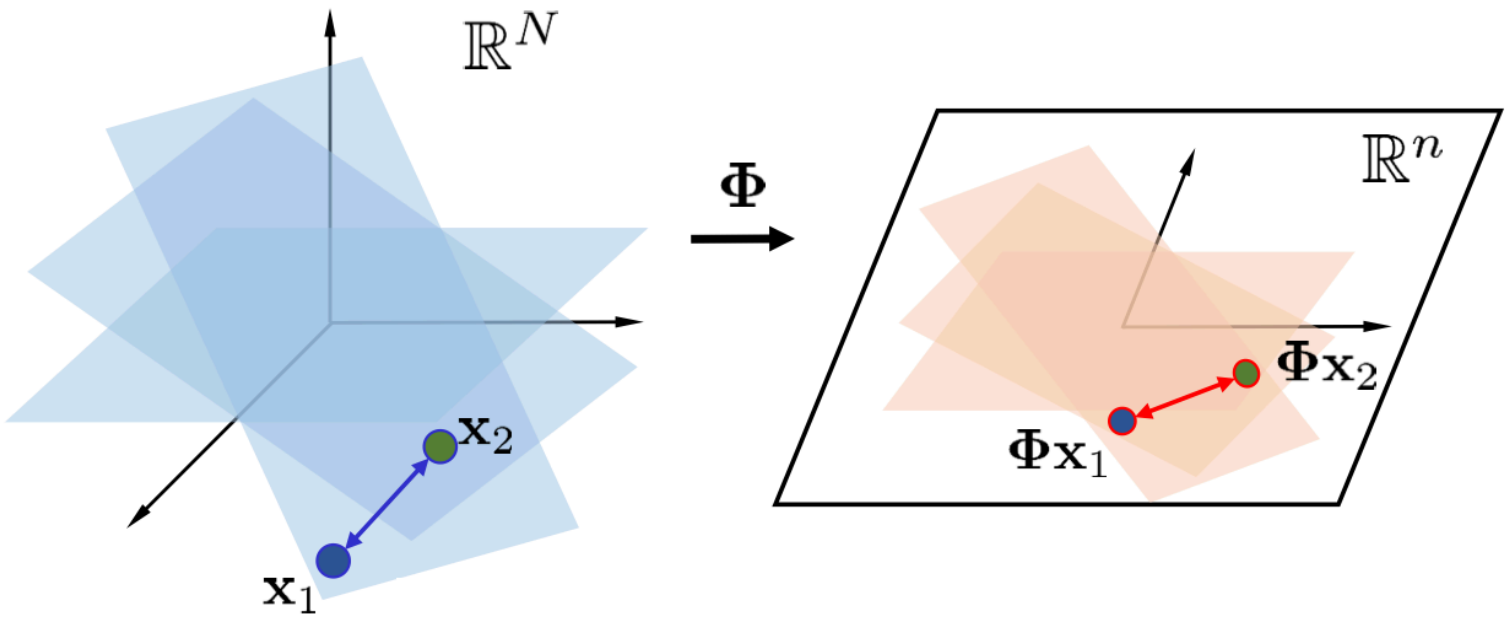


图: A random matrix $\mathbf{\Phi}$ is applied to achieve dimensionality reduction w.r.t. data $\boldsymbol{X}$, which yields a compressed dataset $\boldsymbol{Y} := \mathbf{\Phi}\boldsymbol{X}$.

[6] Y. Jiao, G. Li, and Y. Gu, "Principal angles preserving property of Gaussian random projection for subspaces," *IEEE Global Conference on Signal and Information Processing*, pp. 318-322, 2017.

Step II: Self-Representation

$$\boldsymbol{r}_j = \arg\min_{\boldsymbol{c}} \lambda\|\boldsymbol{c}\|_1 + \frac{1-\lambda}{2}\|\boldsymbol{c}\|_2^2 + \frac{\gamma}{2}\|\boldsymbol{\Phi}\boldsymbol{x}_j - \boldsymbol{Y}\boldsymbol{c}\|_2^2$$

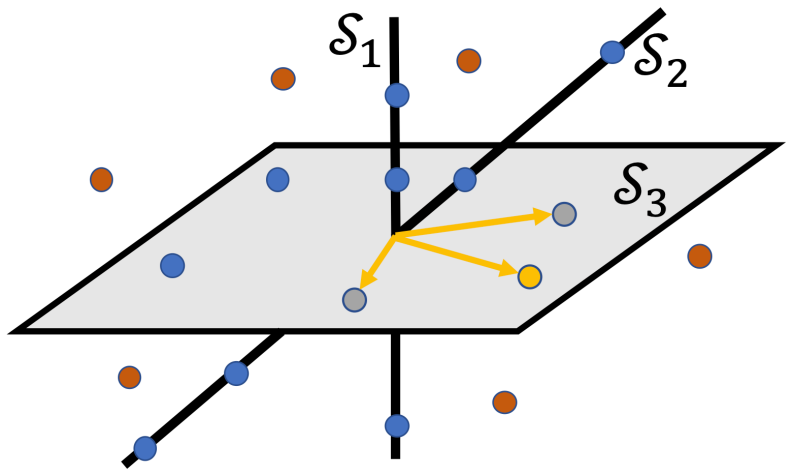


图: The representative vector $\boldsymbol{r}_j$ is solved from the loss function to obtain the representation matrix $\boldsymbol{R}$.

Step III: Random Walks[7]

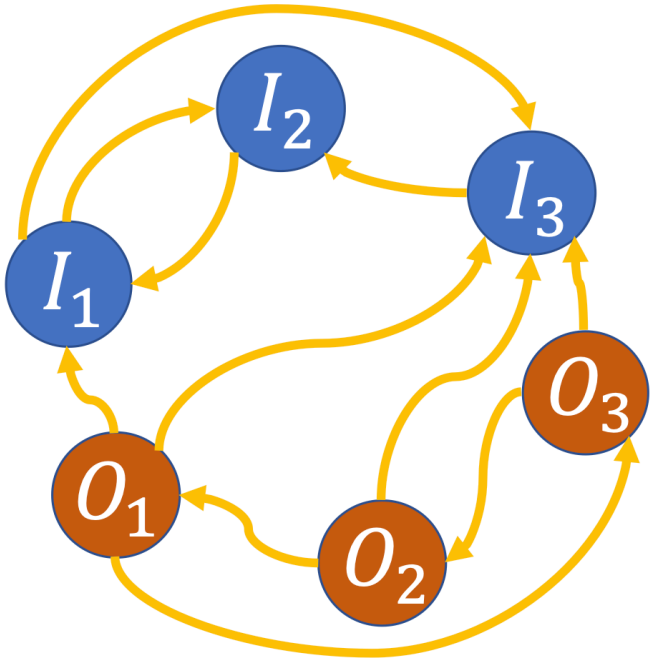


图: We initialize random walks on the representation graph from each state, and every random walk starting from any state will end up at only inlier states.

[7] C. You, D. P. Robinson, and R. Vidal, "Provable self-representation based outlier detection in a union of subspaces," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4323–4332, 2017.

Complexity Analysis

For an $N \times M$-dimensional data matrix and a compression ratio $\frac{n}{N}$,

- Solving the elastic net problem (RSSN, one iteration):

$$\mathcal{O}\left(N^2M^3\right) \to \mathcal{O}\left(n^2M^3\right).$$

- Additional cost (partial Fourier matrix):

$$0 \to \mathcal{O}\left(N\log NM^2\right).$$

Proximal GD[8]

$$\arg\min_{\boldsymbol{c}} \underbrace{\lambda\|\boldsymbol{c}\|_1}_{:=f(x),\text{nonsmooth}} + \underbrace{\frac{1-\lambda}{2}\|\boldsymbol{c}\|_2^2 + \frac{\gamma}{2}\|\boldsymbol{\Phi}\boldsymbol{x}_j - \boldsymbol{Y}\boldsymbol{c}\|_2^2}_{:=g(x),\text{smooth}}$$

Algorithm 2 Proximal GD

Input: Objective function.

1. Randomly generate $\boldsymbol{x}^0$.
2. **for** $t = 1, 2, ..., T$ **do**
3. $\quad \boldsymbol{x}^{t+1} = \text{prox}_{\eta_t h}(\boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t))$

Output: Optimal point.

[8] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, 2013.