Differentially Private Subspace Clustering

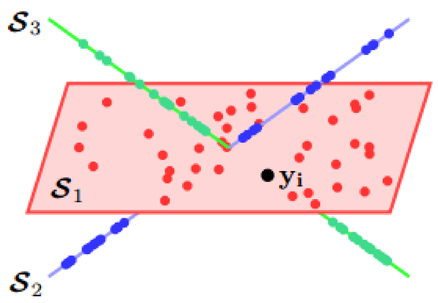
Yining Wang, Yu-Xiang Wang and Aarti Singh

Carnegie Mellon University

Machine Learning Department

Subspace Clustering

Find k low-dimensional linear subspaces to approximate a set of unlabeled data points.



- k-means objective: $\min_{\mathcal{C}} \mathrm{cost}(\mathcal{C}; \mathcal{X})$, where $\mathrm{cost}(\mathcal{C}; \mathcal{X}) = \sum_{i=1}^{n} \min_{j=1}^{k} d^2(\boldsymbol{x}_i, \mathcal{S}_j)/n$.

Differential Privacy

Definition. A randomized algorithm $\mathcal{A}$ is (ε, δ)-differentially private if for all $\mathcal{X}, \mathcal{Y}$ satisfying $d(\mathcal{X}, \mathcal{Y}) = 1$ and all measurable sets S, we have

$$\Pr[\mathcal{A}(\mathcal{X}) \in S] \leq e^{\varepsilon} \Pr[\mathcal{A}(\mathcal{Y}) \in S] + \delta.$$

Objective. Develop differentially private subspace clustering algorithms and characterize the tradeoff between statistical efficiency and guaranteed privacy.

- Important as subspace clustering is an increasingly popular tool for anlyzing sensitive medical and social network data.

DP tools. Sample-and-aggregate [1], exponential mechanism [2], SuLQ [3], etc.

Methods

1. **Sample-and-aggregate based private subspace clustering**

   1: **Input**: $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^n \subseteq \mathbb{R}^d$, number of subsets m, privacy parameters ε, δ; f, $d_{\mathcal{M}}$.
   2: **Initialize**: $s = \sqrt{m}$, $\alpha = \varepsilon/(5\sqrt{2\ln(2/\delta)})$, $\beta = \varepsilon/(4(D + \ln(2/\delta)))$.
   3: **Subsampling**: Select m random subsets of size n/m of $\mathcal{X}$ independently and uniformly at random without replacement. Repeat this step until no single data point appears in more than $\sqrt{m}$ of the sets. Mark the subsampled subsets $\mathcal{X}_{S_1}, \cdots, \mathcal{X}_{S_m}$.
   4: **Separate queries**: Compute $\mathcal{B} = \{\boldsymbol{s}_i\}_{i=1}^m \subseteq \mathbb{R}^D$, where $\boldsymbol{s}_i = f(\mathcal{X}_{S_i})$.
   5: **Aggregation**: Compute $g(\mathcal{B}) = \boldsymbol{s}_{i^*}$ where $i^* = \mathrm{argmin}_{i=1}^m r_i(t_0)$ with $t_0 = (\frac{m+s}{2} + 1)$. Here $r_i(t_0)$ denotes the distance $d_{\mathcal{M}}(\cdot,\cdot)$ between $\boldsymbol{s}_i$ and the t_0-th nearest neighbor to $\boldsymbol{s}_i$ in $\mathcal{B}$.
   6: **Noise calibration**: Compute $S(\mathcal{B}) = 2\max_k(\rho(t_0 + (k+1)s) \cdot e^{-\beta k})$, where $\rho(t)$ is the mean of the top $\lfloor s/\beta \rfloor$ values in $\{r_1(t), \cdots, r_m(t)\}$.
   7: **Output**: $\mathcal{A}(\mathcal{X}) = g(\mathcal{B}) + \frac{S(\mathcal{B})}{\alpha}\boldsymbol{u}$, where $\boldsymbol{u}$ is a standard Gaussian random vector.

   - $f(\cdot)$ can be any approximation algorithm for subspace clustering.
   - Requires stability of f (i.e., $f(\mathcal{X}_S) \approx f(\mathcal{X})$) to work.

2. **Exponential mechanism based private subspace clustering**

$$p(\{\mathcal{S}_\ell\}_{\ell=1}^k, \{z_i\}_{i=1}^n; \mathcal{X}) \propto \exp\left\{-\frac{\varepsilon}{2} \cdot \sum_{i=1}^{n} d^2(\boldsymbol{x}_i, \mathcal{S}_\ell)\right\}.$$

   - $(\varepsilon, 0)$-differentially private, if sampled exactly from the proposed distribution.
   - A Gibbs sampling implementation (more details in the paper):

$$p(z_i = \ell | \boldsymbol{x}_i, \mathcal{S}) \propto \exp\left\{-\varepsilon/2 \cdot d^2(\boldsymbol{x}_i, \mathcal{S}_\ell)\right\}; \qquad p(\mathbf{U}_\ell | \mathcal{X}, z) \propto \exp\left\{\varepsilon/2 \cdot \mathrm{tr}(\mathbf{U}_\ell^\top \mathbf{X}_\ell \mathbf{U})\ell)\right\}.$$

   - Can also be viewed as the Gibbs sampler for an MPPCA model with prior $\mathcal{N}_d(0, \mathbf{I}_d/\varepsilon)$.

Theoretical results

- **Well-separation conditions for k-means subspace clustering**

  Definition. A dataset $\mathcal{X}$ is (ϕ, η, ψ)-well separated if

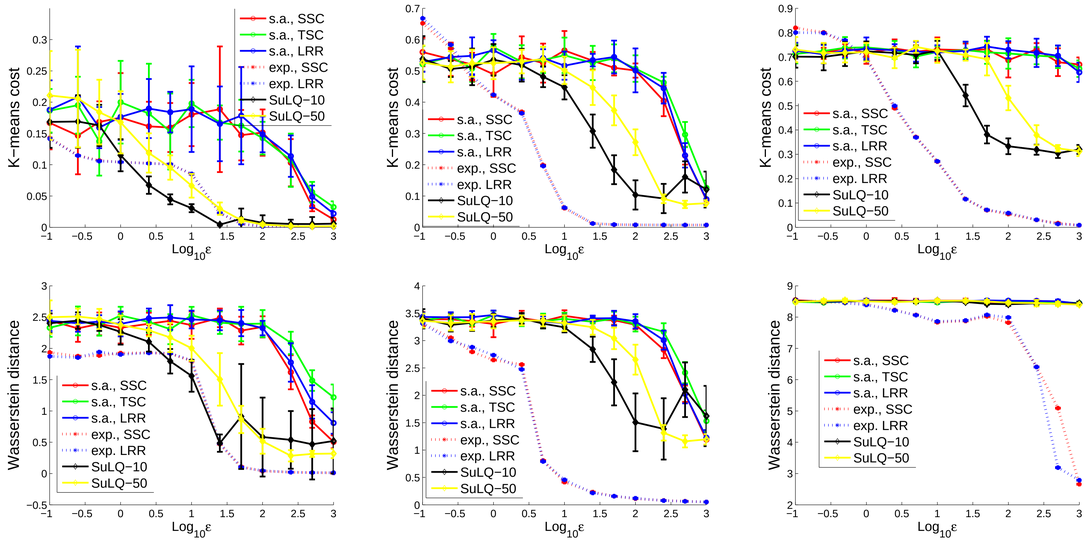
$$\Delta_k(\mathcal{X}) \leq \min\{\phi^2 \Delta_{k-1}^2(\mathcal{X}), \Delta_{k,-}^2(\mathcal{X}) - \psi, \Delta_{k,+}^2(\mathcal{X}) + \eta\}.$$

  Here Δ_k uses k q-dimensional subspaces, Δ_{k-1} uses $(k-1)$ q-dimensional subspaces, $\Delta_{k,-}$ uses $(k-1)$ q-dimensional subspaces and one $(q-1)$-dimensional subspace, $\Delta_{k,+}$ uses $(k-1)$ q-dimensional subspaces and one $(q+1)$-dimensional subspace.

- **Main result (Theorem 3.4)**: If $f(\cdot)$ is a constant approximation algorithm of k-means subspace clustering, then sample-and-aggregate is (ε, δ)-differentially private and the magnitude of per-coordinate Gaussian noise is $O(\frac{\phi^2\sqrt{k}}{\varepsilon(\psi-\eta)})$.

- Additional theoretical results for the stochastic setting available in the paper.

Experiments



From left to right: synthetic dataset, $n = 5000, d = 5, k = 3, q = 3, \sigma = 0.01$; $n = 1000, d = 10, k = 3, q = 3, \sigma = 0.1$; extended Yale Face Dataset B (a subset). $n = 320, d = 50, k = 5, q = 9, \sigma = 0.01$.

References

1. K. Nissim, S. Raskhodnikova and A. Smith. **Smooth sensitivity and sampling in private data analysis.** In *STOC*, 2007.

2. F. McSherry and K. Talwar. **Mechanism design via differential privacy**. In *FOCS*, 2007.

3. A. Blum, C. Dwork, F. McSherry and K. Nissim. **Practical privacy: the SuLQ framework**. In *PODS*, 2005.