Dual Principal Component Pursuit for Robust Subspace Learning: Theory and Algorithms for a Holistic Approach

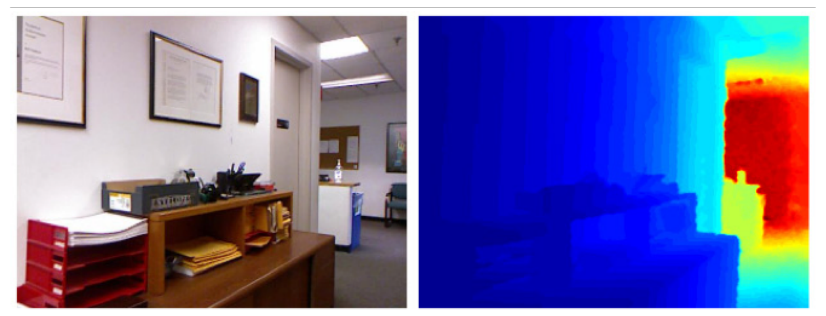
ICML | 2021
Thirty-eighth International Conference on Machine Learning

Tianyu Ding[1], Zhihui Zhu[2], Rene Vidal[1], Daniel Robinson[3]

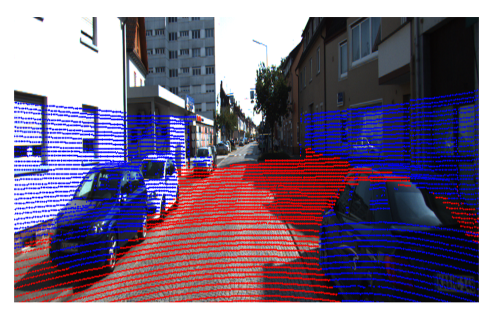
[1]Johns Hopkins University [2]University of Denver [3]Lehigh University

Motivation

- Problem: Fit a subspace of high relative dimension to corrupted data
- Prior work: Dual Principal Component Pursuit (DPCP) finds a normal vector to a single hyperplane that contains the inliers
- Challenges: Learning a subspace via DPCP requires to recursively find a new basis element of the orthogonal complement subspace, which is inefficient



Estimating the geometry of a room from its depth map by fitting multiple hyperplanes

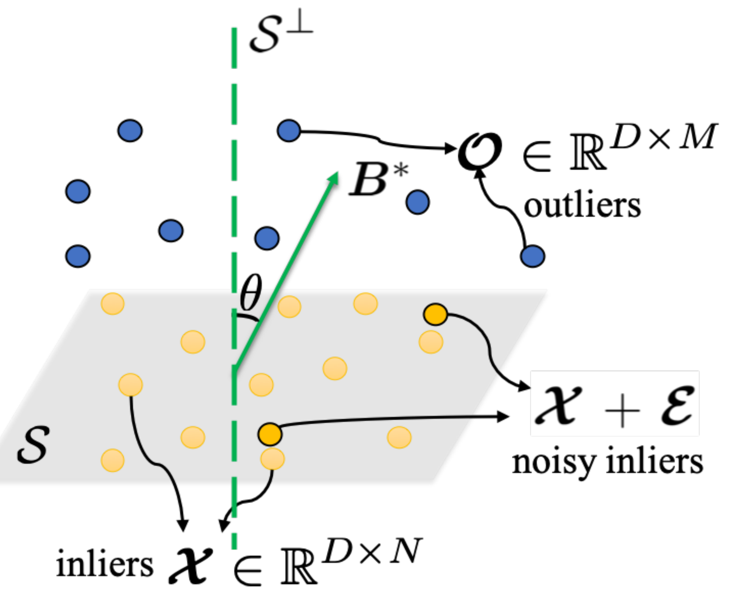


Estimating the road plane from points collected by a laser scanner

Dual Principal Component Pursuit (DPCP): a Holistic Approach

- Inliers $\mathcal{X}$ span a d-dimensional subspace $\mathcal{S}$
- Outliers $\mathcal{O}$ lie in ambient space $\mathbb{R}^D$
- The codimension of $\mathcal{S}$ is $c := D - d$
- Ambient noise $\mathcal{E}$ are additive on inliers
- Dataset $\widetilde{\mathcal{X}} = [\mathcal{X} + \mathcal{E},\ \mathcal{O}]$
- The holistic DPCP approach simultaneously estimates the entire basis of $\mathcal{S}^\perp$ by

$$\min_{B \in \mathbb{R}^{D\times c}} f(B) := \left\|\widetilde{\mathcal{X}}^\top B\right\|_{1,2} = \sum_j \|\widetilde{x}_j^\top B\|_2 \quad \text{s.t.} \quad B^\top B = \mathbf{I} \qquad (1)$$



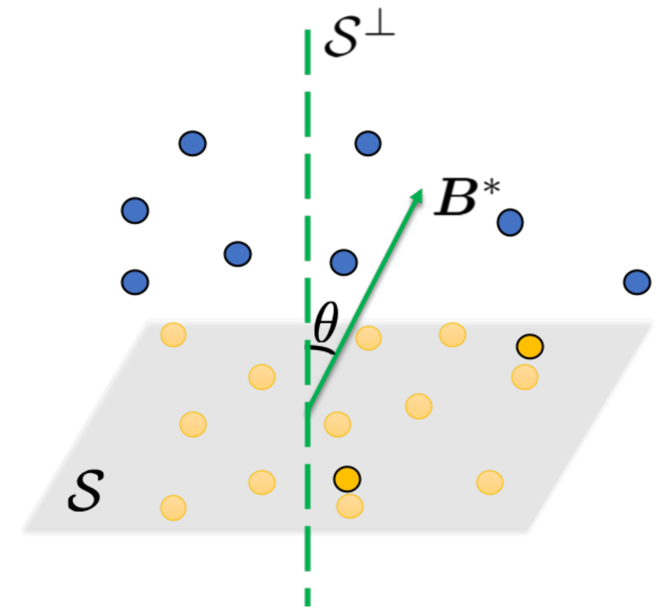
- Intuition: It finds a solution B^* with orthonormal columns that are orthogonal to as many data points as possible

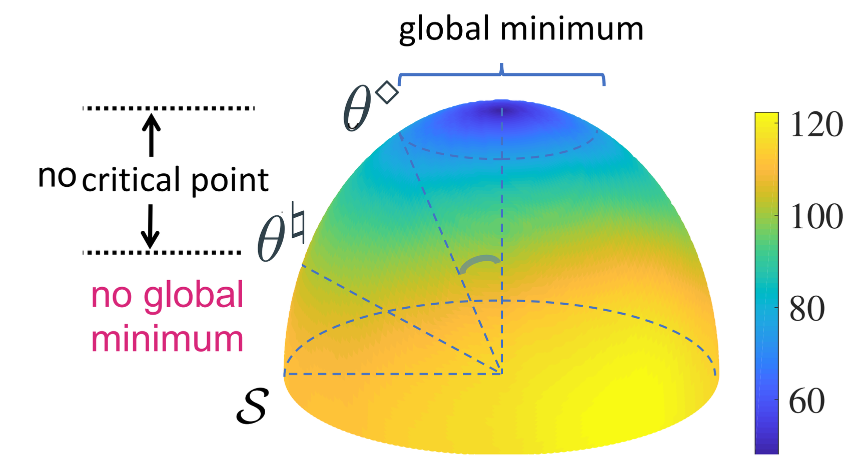
Main Contributions

- We provide the landscape analysis of the holistic DPCP problem for any codimension $c \geq 1$, while prior work only considered the problem with $c = 1$
- We establish the convergence theory of a Projected Riemannain SubGradient Method for solving the problem under the noisy setting, while prior work only showed it converges to $\mathcal{S}^\perp$ with noiseless data

		Main Contributions
Theory	Geometric Deterministic Analysis	$\mathrm{dist}(B^*, \mathcal{S}^\perp) \propto$ noise level
	Geometric Probabilistic Analysis	$\mathrm{dist}(B^*, \mathcal{S}^\perp) \propto$ noise level with high probability if #outliers $= O((\text{\#inliers})^2)$
Algorithm	Projected Riemannian SubGradient Method	locally linear convergence $\mathrm{dist}(B^*, \mathcal{S}^\perp) = O(\beta^k) + \text{const}$, $\beta < 1$, const $\propto$ noise level

Deterministic Global Optimality Analysis





Landscape of DPCP problem (1)

Lemma: Any critical point B of (1) spans a subspace that has an angle θ from $\mathcal{S}^\perp$ s.t.

$$\theta \leq \theta^\diamond \quad \text{or} \quad \theta \geq \theta^\natural$$

where $0 \leq \theta^\diamond \leq \theta^\natural \leq \pi/2$ are determined by outlier-to-inlier and noise-to-inlier ratio.

Remark: Any critical point is either close to $\mathcal{S}^\perp$ or close to $\mathcal{S}$

Theorem: Any global solution B^* of (1) must be close to $\mathcal{S}^\perp$ such that $\theta \leq \theta^\diamond$ as long as the outlier-to-inlier and noise-to-inlier ratios are small.

	$\theta^\diamond$	$\theta^\natural$
smaller $\|\mathcal{E}\|$ smaller $\|\mathcal{O}\|$	$\downarrow 0$	$\uparrow \frac{\pi}{2}$

- $\theta^\diamond \propto$ noise level
- If $\mathcal{E} = 0$, then $\theta^\diamond = 0$, which means the global solution B^* is exactly an orthonormal basis of $\mathcal{S}^\perp$

Probabilistic Analysis

- Columns of outliers $\mathcal{O}$ are drawn uniformly from the unit sphere
- Columns of noisy inliers $\mathcal{X} + \mathcal{E}$ are drawn by first independently generating inliers from $\mathcal{N}(\mathbf{0}, (1/d)\mathcal{P}_\mathcal{S})$ and noise from $\mathcal{N}(\mathbf{0}, (\sigma^2/D)\mathbf{I}_D)$, and then projecting their sum to the unit sphere
- Under this random model, the SNR is $\mathbb{E}[\|\mathcal{X}\|_F]/\mathbb{E}[\|\mathcal{E}\|_F] = 1/\sigma$

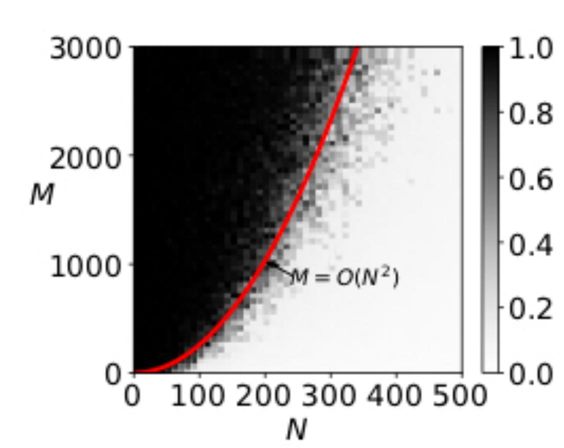
Theorem: Any global solution of (1) must lie in a neighbor hood of $\mathcal{S}^\perp$ such that

$$\sin(\theta) \lesssim \sqrt{\sigma}/(1-\sigma)$$

with high probability if

$$M \lesssim \frac{1-\sigma}{cdD\log^2 D} N^2$$

- The holistic DPCP approach can handle more outliers if $\sigma\downarrow,\ c\downarrow,\ d\downarrow,\ D\downarrow$
- Comparison with state-of-the-art: other existing methods can only handle at most $M = O(N)$ outliers in theory [Lerman and Maunu, 2018]



Plot of $\sin(\theta)$, $\sigma = 0.1$

Projected Riemannian SubGradient Method

Input: $\widetilde{\mathcal{X}}, \{\mu_0, \beta\} \subset (0,1)$, and $k \leftarrow 0$

Spectral initialization: set $B_0 \leftarrow \arg\min_{B \in \mathbb{R}^{D\times c}, B^\top B = \mathbf{I}} \left\|\widetilde{\mathcal{X}}^\top B\right\|_F^2$

Geometrically diminishing step size: $\mu_k \leftarrow \mu_0 \beta^k$

Compute a Riemannian subgradient: $\mathcal{G}(B_k) \leftarrow (\mathbf{I} - B_k B_k^\top) \sum_j \widetilde{x}_j \mathrm{sign}(\widetilde{x}_j^\top B_k)$

Update the iterate as: $\widehat{B}_{k+1} \leftarrow B_k - \mu_k \mathcal{G}(B_k)$

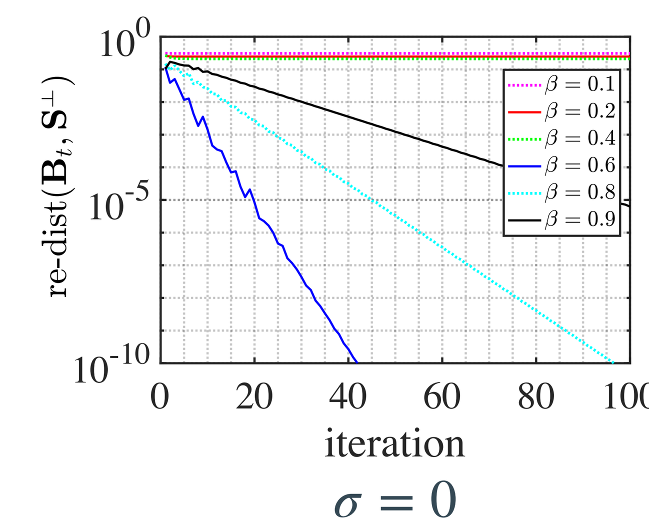
$B_{k+1} \leftarrow \mathrm{orth}(\widehat{B}_{k+1})$

Theorem: B_k converges linearly to $\mathcal{S}^\perp$: $\mathrm{dist}(B_k, \mathcal{S}^\perp) \lesssim \mathrm{dist}(B_0, \mathcal{S}^\perp)\beta^k + \sqrt{\sigma}$

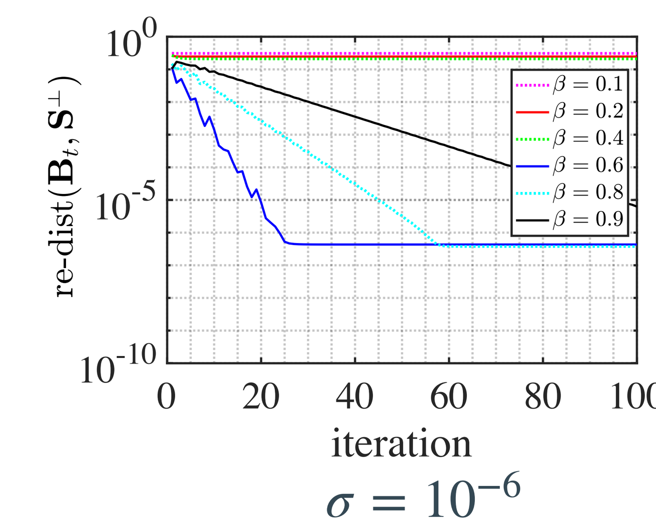
- a large β, slow convergence rate
- a small β, may cause divergence

} trade-off of β

- Example: $D = 30$, $c = 5$, $N = 500$, $M/(M+N) = 0.7$



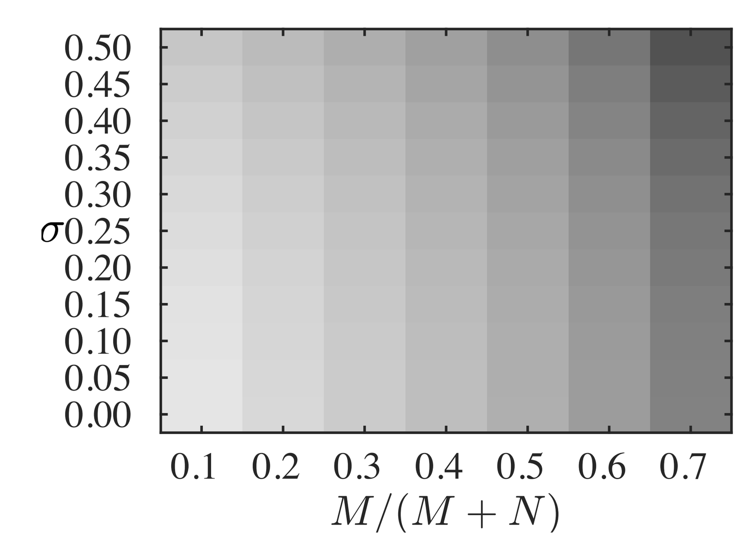
$\sigma = 0$



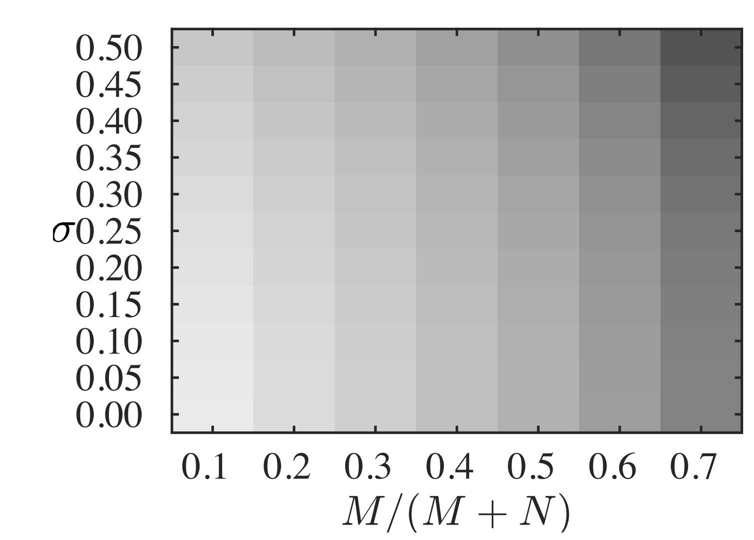
$\sigma = 10^{-6}$

Experiments

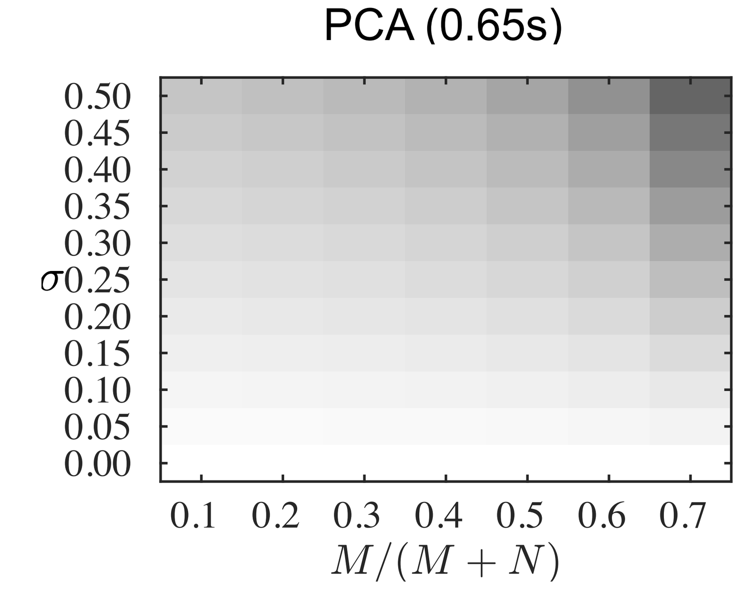
- We plot the phase transition of distance between computed basis and ground-truth basis when varying outlier ratio $M/(M+N)$ and noise level σ
- Example: $D = 1000$, $c = 50$, $N = 10000$; the lighter the better



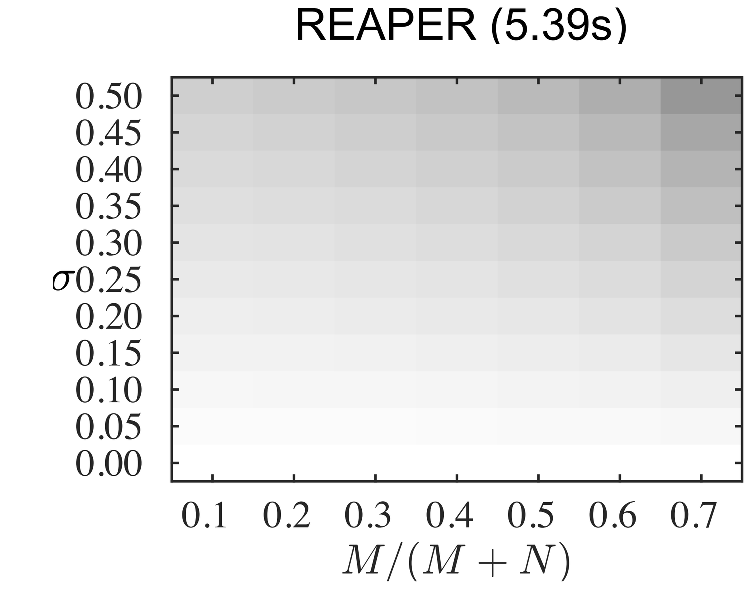
PCA (0.65s)



REAPER (5.39s)



DPCP-recursive (79.74s)



DPCP-holistic (1.78s)