A Critique of Self-Expressive Deep Subspace Clustering

Benjamin D. Haeffele[1], Chong You[2], René Vidal[1]

[1]Johns Hopkins University, [2]Univ. of California Berkeley

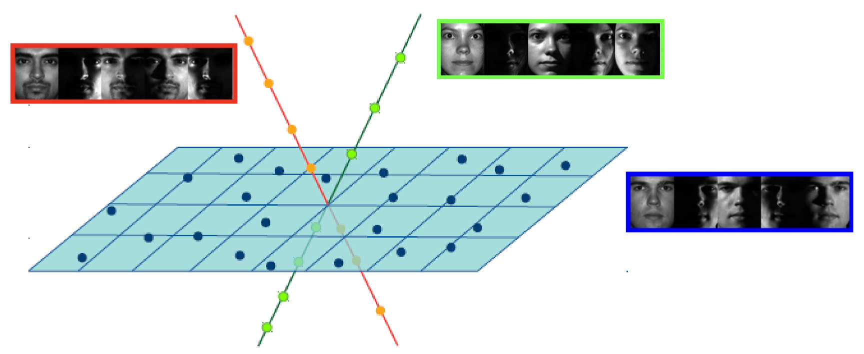
Summary

- Self-expressive deep subspace clustering (SEDSC) [1] is a popular method for clustering data from a union of non-linear manifolds.
- SEDSC is based on using a neural network to try to map data from *non-linear manifolds* into *linear* subspaces.
- Standard techniques can then cluster data from the linear subspaces.
- Here we show theoretically and experimentally that there are numerous problems with many previously proposed SEDSC formulations.
- Globally optimal solution of SEDSC models will map to trivial geometries.
- Experimental performance gain is attributable to ad-hoc post-processing.

Background

Linear Subspace Clustering

- Designed for data supported on multiple linear subspaces.
- Each subspace defines a cluster.
- Many methods exploit the 'self-expressive' property of linear subspaces:
  - A point in a subspace is a linear combinations of other points in the subspace.
- Provably correct clustering under fairly mild conditions [2].



$X = [X_1, X_2, \dots X_N] \in \mathbb{R}^{d \times N}$

$C \in \mathbb{R}^{N \times N}$

$$\min_C \underbrace{\tfrac{1}{2}\|X - XC\|_F^2}_{\text{Self-Expressive Term}} + \lambda\theta(C)$$

Regularization to avoid trial solutions

Example $\theta(\mathbf{C})$:

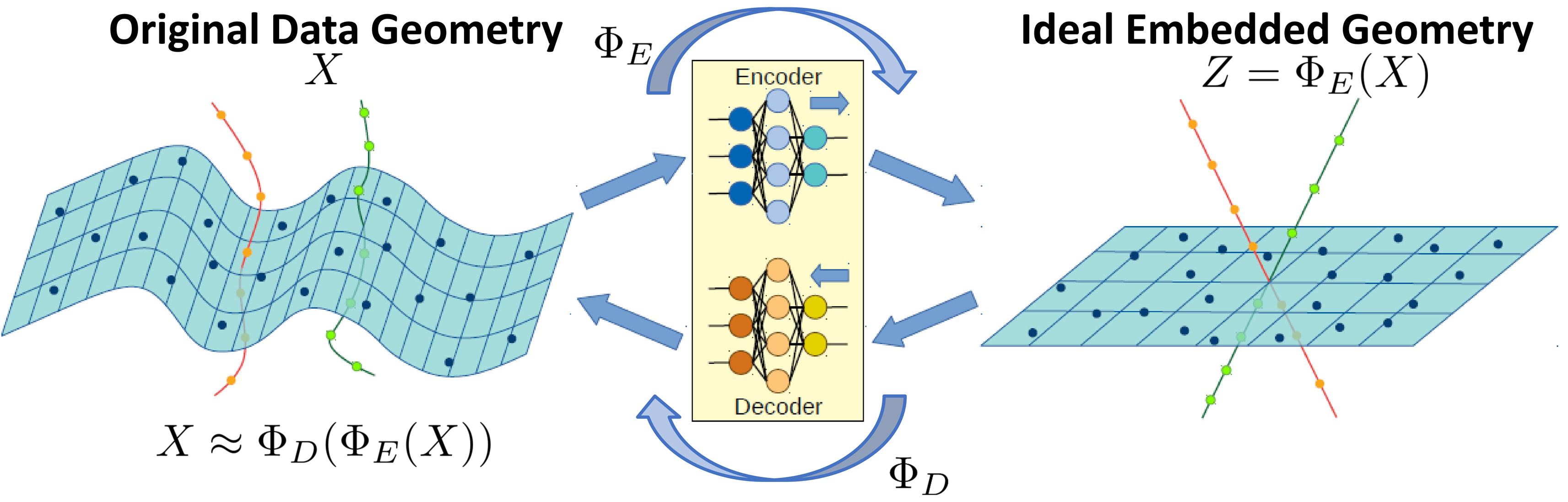
$\theta_{LRR}(C) = \|C\|_*$ [3]

$\theta_{SSC}(C) = \|C\|_1 + \delta(\mathrm{diag}(C) = \mathbf{0})$ [4]

$\theta_{EnSC}(C) = \|C\|_1 + \mu\|C\|_F^2 + \delta(\mathrm{diag}(C) = \mathbf{0})$ [5]

Self-Expressive Deep Subspace Clustering (SEDSC) [1]

- Designed for data supported on multiple *non-linear manifolds*
- Use a neural network to map non-linear manifolds to linear subspaces.
  - Typically an auto-encoder is used.
- The latent geometry is regularized with a self-expressive clustering term.



Neural Network (typically an auto-encoder)

$$(1) \quad \min_{C,D,E} \underbrace{\tfrac{1}{2}\|X - \Phi_D(\Phi_E(X)C)\|_F^2}_{\text{Auto-Encoder Loss}} + \underbrace{\tfrac{\gamma}{2}\|\Phi_E(X) - \Phi_E(X)C\|_F^2 + \lambda\theta(C)}_{\text{Self-Expressive Clustering Loss in Embedded Space}}$$

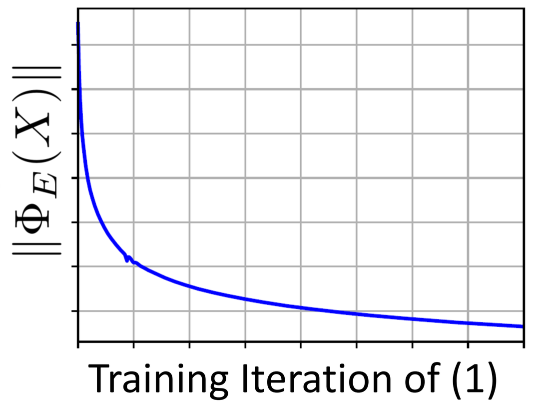
Theoretical Analysis

Basic Issues – Weight Scaling

- The formulation in (1) is often ill-posed from the beginning.
- Typically (1) can be reduced by decreasing the magnitude of $\Phi_E(X)$.
- Can scale down weights in Φ_E and scale up weights in Φ_D to keep auto-encoder loss constant but make clustering loss arbitrarily small.

Scale Down	E	$\|\Phi_E(X)\|$	Goes Down
Scale Up	D	$\Phi_D(\Phi_E(X))$	Stays Constant
		$\frac{1}{2}\|X - \Phi_D(\Phi_E(X)C)\|_F^2$	Stays Constant
		$\frac{\gamma}{2}\|\Phi_E(X) - \Phi_E(X)C\|_F^2 + \lambda\theta(C)$	Goes Down

Seen in Real Experiments



Proposition: (Informal) Most previous formulations based on (1) are ill-posed. For any choice of weights the clustering loss can be made arbitrarily small by scaling.

More Significant Issues – Data Geometry

- The above issue can be corrected by normalization or regularization.
- The question remains: What is the optimal embedded data geometry?
- If the encoder and decoder are highly expressive then the embedded geometry can be essentially arbitrary while having 0 auto-encoder loss.
- The optimal embedded geometry will minimize the clustering loss term.
  - Results in trivial geometries in most cases. One example case below (others in paper).

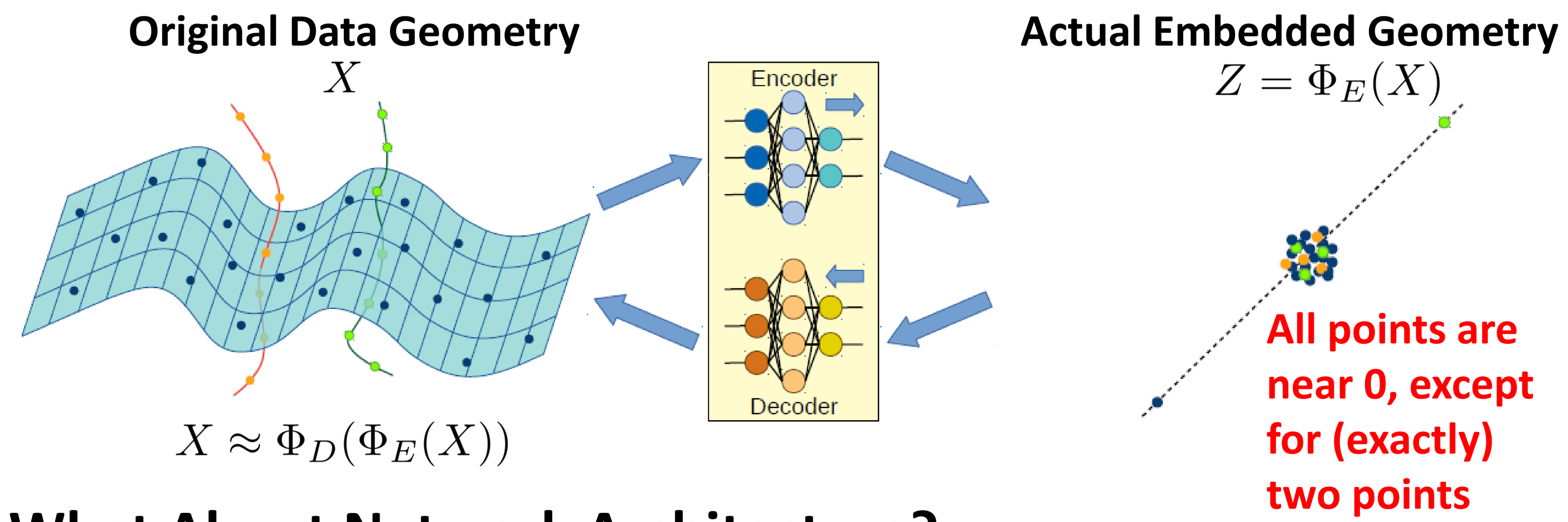
Theorem: Optimal solutions to the problem

$$\min_{Z,C} \theta_{SSC}(C) = \|C\|_1 + \delta(\mathrm{diag}(C) = \mathbf{0}) \quad \text{s.t.} \quad Z = ZC, \ \|Z\|_F^2 \geq \tau$$

are characterized by the set

$$(Z^*, C^*) \in \left\{ \begin{bmatrix} \mathbf{z} & \mathbf{z} & \mathbf{0}_{d \times N-2} \end{bmatrix} P \right\} \times \left\{ P^\top \begin{bmatrix} 0 & 1 & \cdot & 0 \\ 1 & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} P \right\}$$

where P is any signed permutation matrix and $\mathbf{z}$ is any vector such that $\|\mathbf{z}\|_F^2 \geq \tau/2$



What About Network Architecture?

- The above argument was made for highly expressive encoders/decoders.
- **Q:** Will this behavior change with small network architectures? **A:** No.

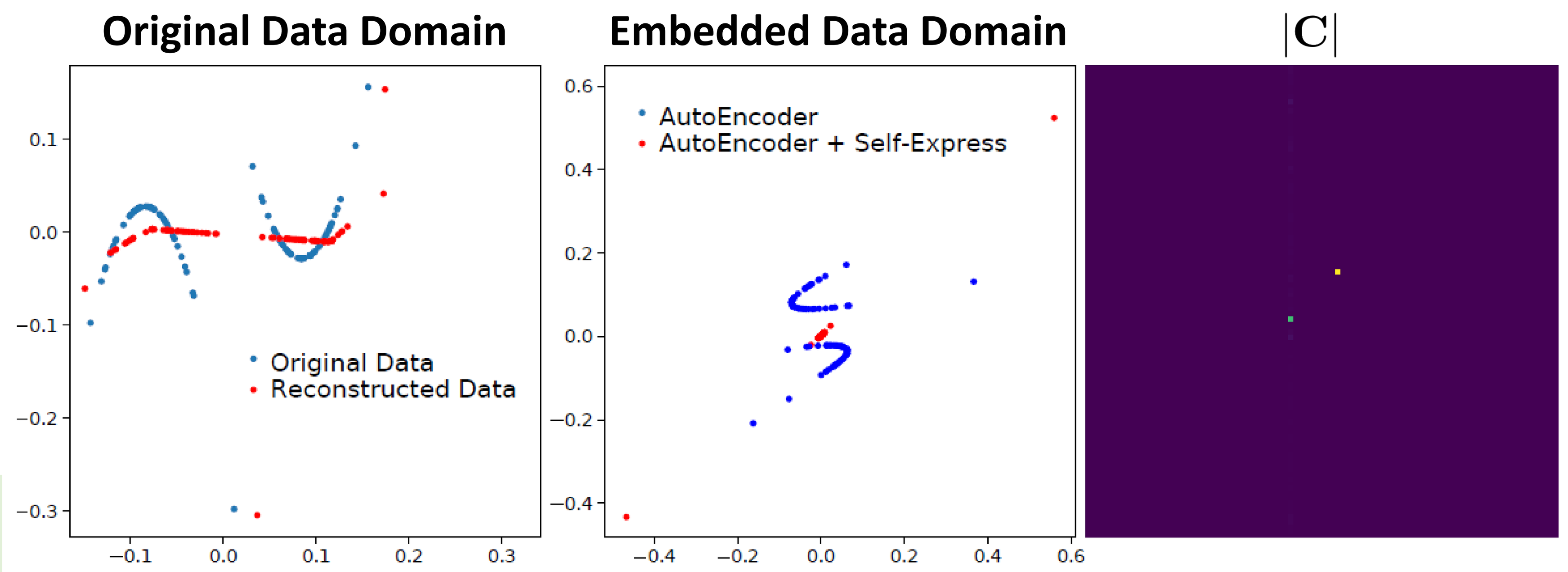
Proposition: (Informal) Suppose the encoder and decoder are both single-hidden layer CNNs with only one convolutional channel and ReLU non-linearities. Then optimal solutions to (1) will have trivial geometries in the embedded space.

Trivial geometries also occur with very simple network architectures.

Experimental Verification

Synthetic Data Experiments Confirm Theory

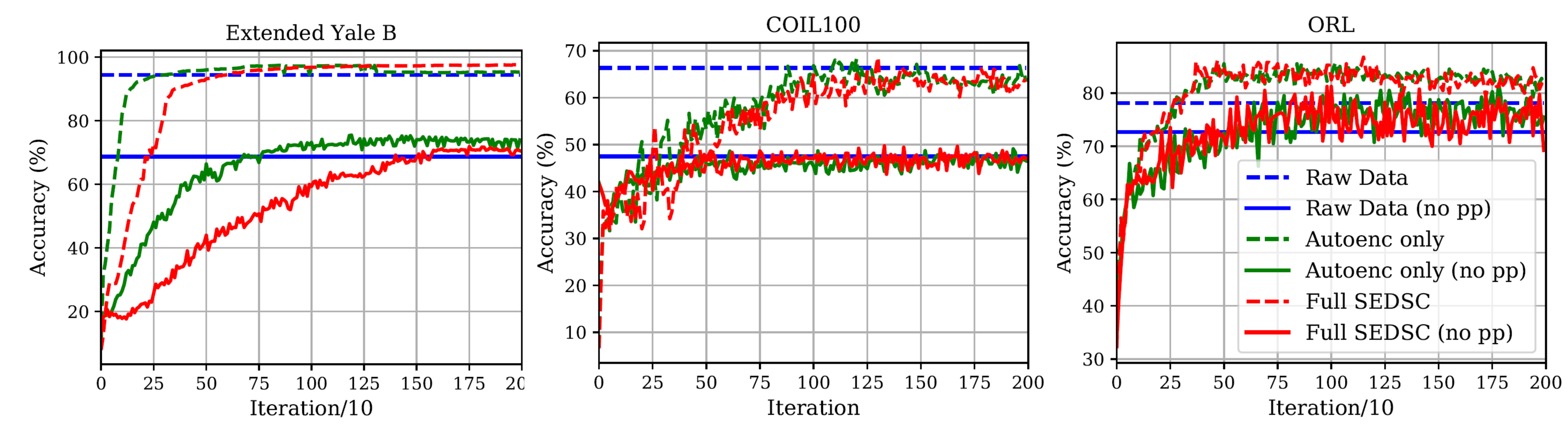
- The predicted trivial geometries can be observed in experiments.
- Even with very simple synthetic data the model in (1) learns poor solutions due to the deficiencies we analyze.



Where Did Reported Performance Come From?

- Previous work has reported reasonable experimental results.
- If the methods are ill-posed where did performance come from?
  1) Not converging to a true solution for the model in (1).
  2) Ad-hoc post processing of the C matrix.

No difference in performance between the model in (1) and simple baselines if ad-hoc post-processing is used (dashed lines) or not (solid lines) in all methods.



Clustering Accuracy

	With post-processing			Without post-processing		
	Raw Data	Autoenc only	Full SEDSC	Raw Data	Autoenc only	Full SEDSC
YaleB	94.40%	**97.12%**	96.79%	68.71%	**71.96%**	59.09%
COIL100	66.47%	**68.26%**	64.96%	**47.51%**	44.84%	45.67%
ORL	78.12%	83.43%	**84.10%**	72.68%	**73.73%**	73.50%

References and Acknowledgements

[1] P Ji, T Zhang, H Li, M Salzmann, I Reid. "Deep subspace clustering networks." Neural Information Processing Systems (NeurIPS), pp. 24–33, 2017.

[2] R Vidal, Y Ma, S Sastry. "Generalized Principal Component Analysis." Springer, 2016.

[3] G Liu, Z Lin, S Yan, J Sun, Y Ma. "Robust recovery of subspace structures by low-rank representation." IEEE T PAMI, 35(1):171–184, 2013.

[4] E Elhamifar, R Vidal. "Sparse subspace clustering: Algorithm, theory, and applications." IEEE T PAMI, 35(11):2765–2781, 2013.

[5] C You, C-G Li, D Robinson, R Vidal. "Oracle based active set algorithm for scalable elastic net subspace clustering." IEEE CVPR, pp. 3928–3937, 2016.

Acknowledgments. We thank Zhihui Zhu and Benjamin Béjar Haro for helpful discussions in the early stages of this work. This work was partially supported by the Northrop Grumman Mission Systems Research in Applications for Learning Machines (REALM) initiative, NSF Grants 1704458, 2031985 and 1934979, and the Tsinghua-Berkeley Shenzhen Institute Research Fund.