Ensemble K-Subspaces

Laura Balzano

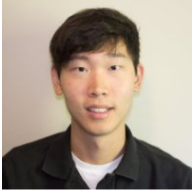
University of Michigan

Simons Institute for the Theory of Computing
Randomized Numerical Linear Algebra and Applications 2018

work with John Lipor, David Hong, and Yan Shuo Tan

Collaborators

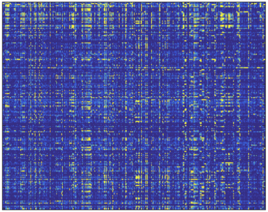
John Lipor

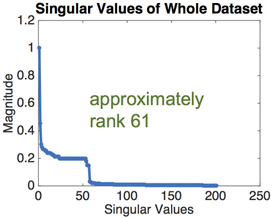


David Hong

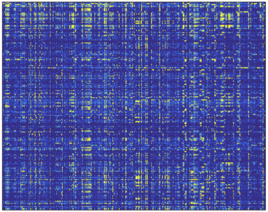
Yan Shuo Tan

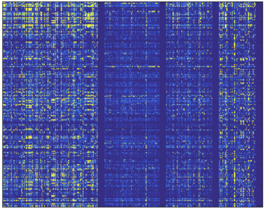
Subspace Clustering

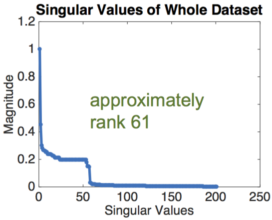


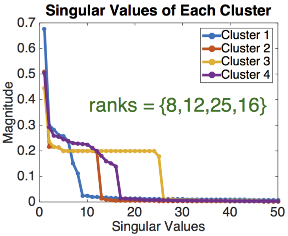


Subspace Clustering

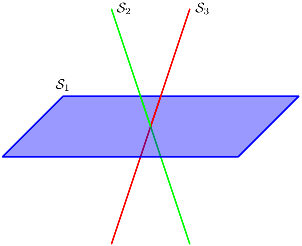




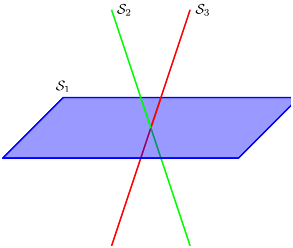


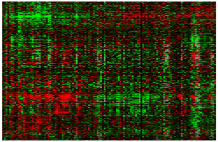


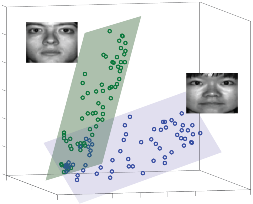
Subspace Clustering



Subspace Clustering







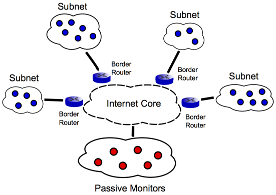


Image courtesy Hopkins 155

Context for subspace clustering

- data can be clustered into meaningful groups (cell type, image content, object features, subnet)
- but we *do not have labels* (at least for this work)
- each cluster has low-rank structure

K-Subspace Clustering Objective

Let $x_i \in \mathbb{R}^d$, $i = 1, \ldots, n$ be data points that we wish to cluster into K low-rank clusters (rank $r \ll \min d, n$).

$$\min_{\mathcal{C},\mathcal{U}} \sum_{k=1}^{K} \sum_{i:x_i \in c_k} \left\| x_i - U_k U_k^T x_i \right\|_2^2, \tag{1}$$

$\mathcal{C} = \{c_1, \ldots, c_K\}$ is a partition on $\{1, \cdots, n\}$, denoting the set of estimated clusters

$\mathcal{U} = \{U_1, \ldots, U_K\}$ with $U_k \in \mathbb{R}^{d \times r}$ denotes the corresponding set of orthonormal subspace bases

This is a generalization of the K-Means objective to clustering with planes as "centers."

KSS

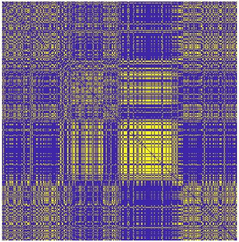
Alternating algorithm generalizing K-Means[1]:

Input: $X \in \mathbb{R}^{d \times n}$: data, K: number of clusters, r: subspace rank, $\{U_1, \dots, U_K\}$: initial subspaces
Output: $\{c_1, \dots, c_K\}$: clusters of X
while Clustering changes and KSS objective decreases **do**
  # Cluster by projection
  $c_k \leftarrow \{x \in X \;:\; \forall j \; \|U_k^T x\|_2 \geq \|U_j^T x\|_2\}$ for $k = 1, \dots, K$
  # Best-fit rank-r subspace from cluster data
  $U_k \leftarrow \text{PCA}\,(c_k, r)$ for $k = 1, \dots, K$
end while

[1] First derived in [Bradley and Mangasarian, 2000]

Initialization

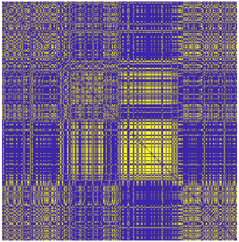
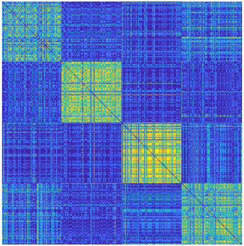
Like K-Means, the K-Subspaces algorithm depends heavily on the initialization. Random init for $d = 100, n = 400, r = 5, K = 4$, additive noise variance 0.1 for each entry of the $d \times n$ matrix.



Indeed, it is known that there is a set of initializations of nonzero measure that provably lead to a local optimal point.

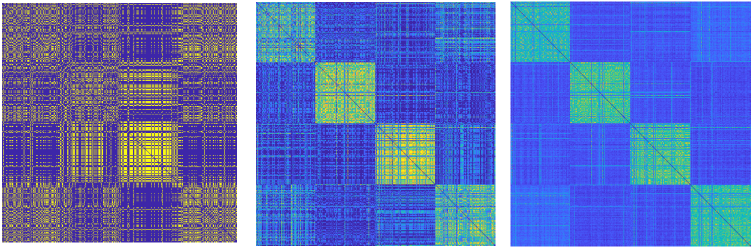
Initialization

Use ideas from consensus clustering and add together the affinity matrices.

Initialization

Average $B = 1, 5, 50$ runs.



Clustering error using spectral clustering $K = 4$: 53%, 12%, 2%.

error definition

EKSS

Input: $X \in \mathbb{R}^{d \times n}$: data, $\mathcal{F}$: distribution on subspaces
$\bar{K}$: number of candidate sets, K: number of output clusters,
q: threshold parameter, B: number of base clusterings
Output: $\widetilde{\mathcal{C}} := \{c_1, \dots, c_K\}$: clusters of X
for $b = 1, \dots, B$ (in parallel) **do**
    $\widetilde{\mathcal{S}} = \{U_1, \dots, U_{\bar{K}}\}$ where $U_k \overset{iid}{\sim} \mathcal{F}, k = 1, \dots, \bar{K}$
    $\widetilde{\mathcal{C}}^{(b)} \leftarrow \text{KSS}(X, \bar{K}, \widetilde{\mathcal{S}})$. Cluster using KSS
end for
$A_{i,j} \leftarrow \frac{1}{B} \left| \{b : x_i, x_j \text{ are co-clustered in } \widetilde{\mathcal{C}}^{(b)}\} \right|$ for $i, j = 1, \dots, n$
$\bar{A} \leftarrow \text{Thresh}(A, q)$ Keep top q entries per row/column
$\mathcal{C} \leftarrow \text{SpectralClustering}(\bar{A}, K)$ Final Clustering

EKSS Performance

Algorithm	Hopkins 	Yale B 	COIL-20 	COIL-100 	USPS 	MNIST-10k 
EKSS	**0.26**	**14.31**	**13.47**	**28.57**	**15.84**	**2.58**
KSS	**0.35**	54.28	33.12	74.53	**18.31**	**2.60**
CoP-KSS	0.69	56.00	29.10	51.38	**10.12**	**8.80**
MKF	**0.24**	46.22	39.24	66.49	28.62	43.49
TSC	2.07	22.20	15.28	**29.82**	31.57	15.98
SSC-ADMM	1.07	**9.83**	**13.19**	44.06	56.61	19.17
SSC-OMP	25.25	**13.28**	27.29	34.79	77.94	19.19
EnSC	9.75	18.87	**8.26**	**28.75**	33.66	17.97

Table: Clustering error comparison. The lowest three clustering errors are given in bold.

data sets error definition

K-Subspaces Theory

Hardness results[2]:

For $r = 1$, with d, n, K input to the problem, $\exists \epsilon > 0$ such that it is NP-hard to approximate the KSS objective within $(1 + \epsilon)$.

For $K = 2$, with d, n, r input to the problem, it is NP-hard to approximate the KSS objective within $(1 + \epsilon)$ for any $\epsilon > 0$.

[2] [Tao and Balzano, 2018]

K-Subspaces Theory

For the KSS alternating algorithm, we know that KSS objective function decreases at every iteration (by definition) and it reaches a local optimum[3].

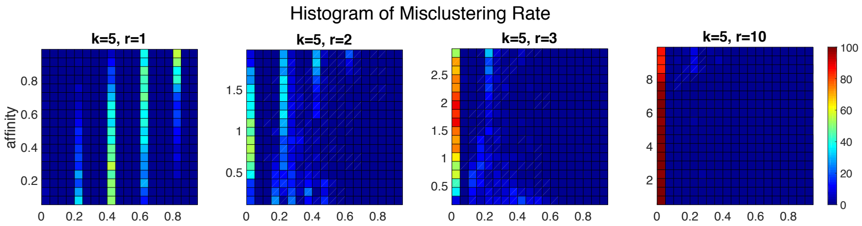
There is a set of initializations of nonzero measure that provably lead to a local optimal point.

[3] [Bradley and Mangasarian, 2000]

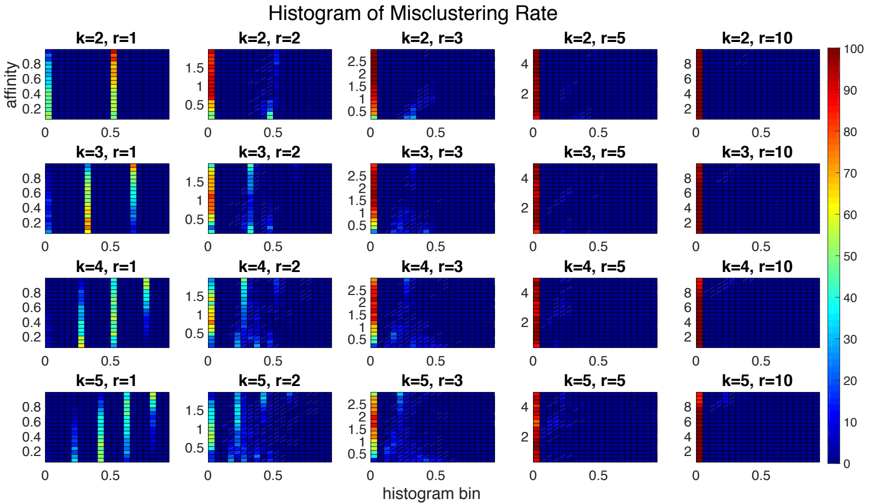
What we see from random initialization

Generate data $X \in \mathbb{R}^{d \times n}$ from a union of subspaces with no noise, $d = 500$, $n = 1000$, and vary rank r, number of subspaces k (in this slide $k = 5$), affinity between subspaces pairwise.

Subspace affinity: $\|U_i^T U_j\|_F^2 \in [0, r]$ for orthonormal U_i, U_j



What we see from random initialization



Overview of our results

- Random initializations cluster a pair of points with probability monotonic in their inner product
- We proved conditions under which one can correctly subspace cluster with any (possibly perturbed) monotonic function of inner products (generalizing TSC)

$$A_{ij} = f\left(\left|\left\langle x_i^{(l)}, x_j^{(k)}\right\rangle\right|\right) + \tau_{i,j}^{(l,k)}$$

- We proved that the EKSS-0* affinity matrix concentrates to a monotonic function of inner products. (*with consensus applied only to the clustering from the projection onto random initialization)

$$\mathbb{E}[A_{ij}] = f\left(\left|\left\langle x_i^{(l)}, x_j^{(k)}\right\rangle\right|\right)$$

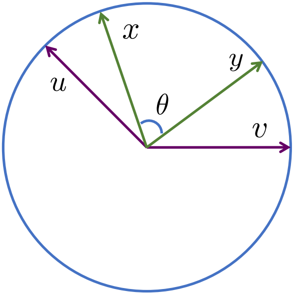
A Simple Problem

Suppose we have two unit norm data points $x, y \in \mathbb{R}^d$, and two random candidate subspaces, $U, V \in \mathbb{R}^{d \times r}$.
What is the probability that both points are closer to the same subspace?

$$\|U^T x\| > \|V^T x\| \text{ and } \|U^T y\| > \|V^T y\| \quad \text{(or flip } U, V\text{)}$$

A Simpler Problem

Suppose we have two unit norm data points $x, y \in \mathbb{R}^d$, and two random candidate subspaces, $u, v \in \mathbb{R}^{d \times 1}$.

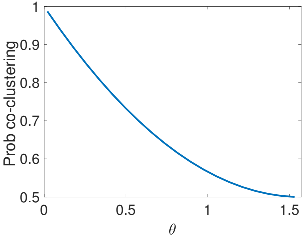


A Simpler Problem

Theorem 1

Let $x, y \in \mathbb{R}^d$ be unit norm and $|x^T y| = \cos\theta$ for $\theta \in [0, pi/2]$. The probability that both x and y have larger projection on either u or v is

$$\mathbb{P}(\theta) = 1 - 2\frac{\theta}{\pi}\left(1 - \frac{\theta}{\pi}\right) .$$



Generalize the model

What if one is only able to observe some noisy version of a monotonic function of the inner products? (as in noisy data, missing data, compressed data etc).

- x_j^k is the j^{th} point in the k^{th} subspace,
- $f(\cdot)$ is a monotonic function,
- τ is a bounded deviation term.

$$f\left(\left|\left\langle x_i^{(l)}, x_j^{(k)}\right\rangle\right|\right) + \tau_{i,j}^{(l,k)}, \quad k \in 1, \ldots, K \tag{2}$$

Results

Definition 2 (Angular separation)

Let $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_K$ be a set of points with the ith point of $\mathcal{X}_l$ denoted as $x_i^{(l)}$. Then we define the *q-angular separation* as

$$\phi_q = \min_{l \in [K], i} \frac{f\left(\left|\left\langle x_i^{(l)}, x_{\neq i}^{(l)}\right\rangle\right|_{[q]}\right) - f\left(\max_{k \neq l, j}\left|\left\langle x_i^{(l)}, x_j^{(k)}\right\rangle\right|\right)}{2} \qquad (3)$$

where $\left|\left\langle x_i^{(l)}, x_{\neq i}^{(l)}\right\rangle\right|_{[q]}$ denotes the q^{th} largest absolute inner product between $x_i^{(l)}$ and others in subspace l.

Results

Lemma 3 (Expected affinity matrix)

The (i,j)th entry of the affinity matrix A formed by EKSS-0 has expected value

$$\mathbb{E}\left[A_{i,j}\right] = f(|\langle x_i, x_j \rangle|) \tag{4}$$

where $f : \mathbb{R}_+ \to \mathbb{R}_+$ is a strictly increasing function, and the expectation is taken with respect to the random subspaces drawn in EKSS-0.

We can prove concentration/deviation $\tau < \phi_q$ for different assumptions on the subspaces and random data models, *e.g.*, with additive noise or missing data.

Results

Theorem 4 (EKSS-0 provides correct clustering for subspaces with bounded affinity)

Let $\mathcal{S}_k$, $k = 1, \dots, K$ be subspaces of dimension r in $\mathbb{R}^d$. Let the points in $\mathcal{X}_k$ be a set of points drawn uniformly from the unit sphere in subspace k. Let $q \in [c_4 \log n_{max}, n_{min}/6)$, where $c_4 = 12(24\pi)^{r-1}$. If

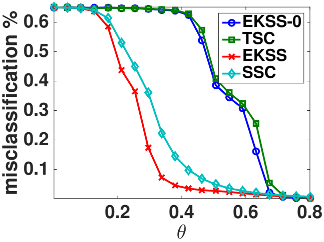
$$\max_{k,l:k\neq l} \text{aff}(\mathcal{S}_k, \mathcal{S}_l) \leq \frac{1}{15 \log n},$$

then $\bar{A}$ obtained by EKSS-0 results in correct clustering of the data with probability at least $1 - \frac{10}{n} - ne^{-c_2 n_{min}} - n^2 e^{-c_3 \gamma B}$, where $c_2, c_3 > 0$ are numerical constants, and (roughly) $0 < \gamma < \phi_q$.

Other algorithms

- (CoP-KSS) Coherence Pursuit K-Subspaces [Gitlin et al., 2018]
- (MKF) Median K-Flats [Zhang et al., 2009]
- (TSC) Thresholded Subspace Clustering [Heckel and Bölcskei, 2015]
- (SSC-ADMM) Sparse Subspace Clustering with its ADMM implementation [Elhamifar and Vidal, 2013]
- (SSC-OMP) SSC with Orthogonal Matching Pursuit [You et al., 2016b]
- (EnSC) Elastic Net Subspace Clustering [You et al., 2016a]

Synthetic data



Problem params: $d = 100, r = 10, K = 3, N_k = 500, \sigma^2 = 0.05$.

Although our experiments indicate that EKSS-0 appears to have no benefits over TSC, we do find that by running a small number of KSS iterations, significant performance improvements are achieved.

EKSS Performance

Algorithm	Hopkins	Yale B	COIL-20	COIL-100	USPS	MNIST-10k
EKSS	**0.26**	**14.31**	**13.47**	**28.57**	**15.84**	**2.58**
KSS	**0.35**	54.28	33.12	74.53	**18.31**	**2.60**
CoP-KSS	0.69	56.00	29.10	51.38	**10.12**	**8.80**
MKF	**0.24**	46.22	39.24	66.49	28.62	43.49
TSC	2.07	22.20	15.28	**29.82**	31.57	15.98
SSC-ADMM	1.07	**9.83**	**13.19**	44.06	56.61	19.17
SSC-OMP	25.25	**13.28**	27.29	34.79	77.94	19.19
EnSC	9.75	18.87	**8.26**	**28.75**	33.66	17.97

Table: Clustering error of subspace clustering algorithms for a variety of benchmark datasets. The lowest three clustering errors are given in bold. No other algorithm is in the top five for all datasets.

Conclusion

Subspace Clustering using Ensembles of K-Subspaces
John Lipor, David Hong, Yan Shuo Tan, Laura Balzano
https://arxiv.org/abs/1709.04744

- We have presented a new subspace clustering algorithm based on ensembles of K-Subspaces with random initialization.
- It has theoretical guarantees as strong as state-of-the-art.
- Its performance exceeds those guarantees.

- We have not analyzed the alternating steps of KSS. Showing the impact of this improvement is a matter of ongoing work.

References I

Bradley, P. S. and Mangasarian, O. L. (2000).
k-Plane clustering.
Journal of Global Optimization, 16:23–32.

Elhamifar, E. and Vidal, R. (2013).
Sparse subspace clustering: Algorithm, theory, and applications.
Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35(11):2765–2781.

Gitlin, A., Tao, B., Balzano, L., and Lipor, J. (2018).
Improving k-subspaces via coherence pursuit.
Accepted to the Journal of Selected Topics in Signal Processing.

References II

Heckel, R. and Bölcskei, H. (2015).
Robust subspace clustering via thresholding.
IEEE Trans. Inf. Theory, 24(11):6320–6342.

Tao, B. and Balzano, L. (2018).
On the hardness of k-subspaces.
http://www-personal.umich.edu/~bstao/Biaoshuai%20Tao_files/hardnessProofReport.pdf.

You, C., Li, C.-G., Robinson, D. P., and Vidal, R. (2016a).
Oracle based active set algorithm for scalable elastic net subspace clustering.
In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*.

References III

You, C., Robinson, D. P., and Vidal, R. (2016b).
Scalable sparse subspace clustering by orthogonal matching pursuit.
In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*.

Zhang, T., Szlam, A., and Lerman, G. (2009).
Median k-flats for hybrid linear modeling with many outliers.
In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 234–241. IEEE.

Clustering error

Let Q^{out} and Q^{true} be the output and ground-truth labelings of the data, with $Q_{i,j} = 1$ if point j belongs to cluster i and zero otherwise. Then we measure error by

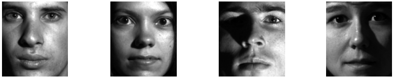
$$\text{err} = \frac{100}{n} \left(1 - \max_{\pi} \sum_{i,j} Q^{\text{out}}_{\pi(i)j} Q^{\text{true}}_{ij} \right),$$

where π is a permutation of the cluster labels.

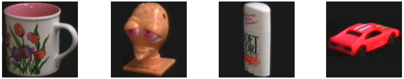
Back to progression
Back to performance

Data sets

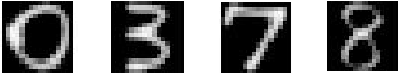
Yale:



COIL:

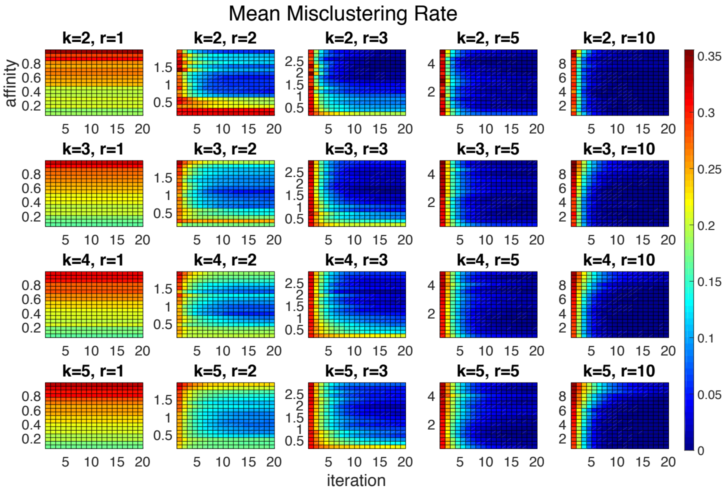


USPS:

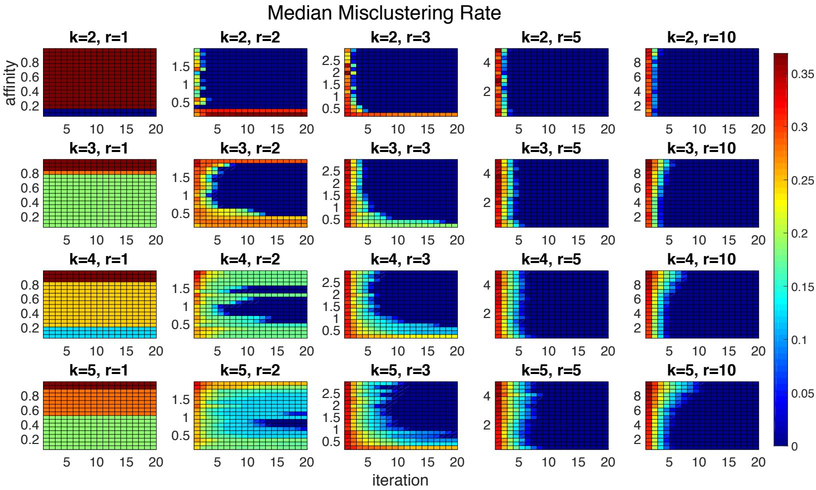


Back to performance

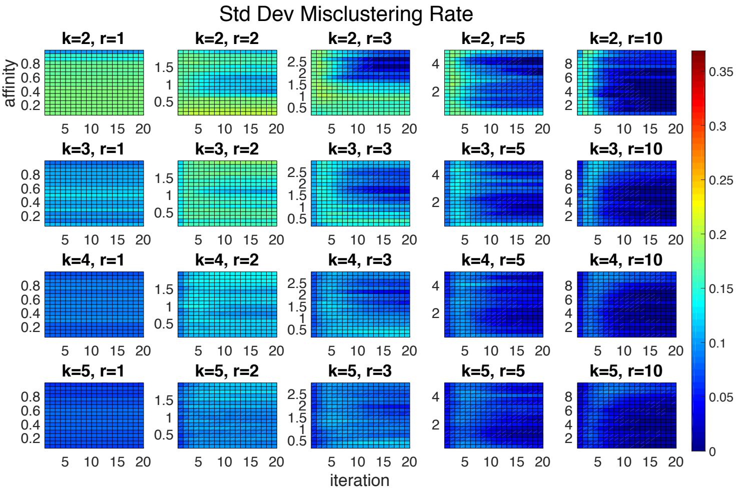
What we see from random initialization



What we see from random initialization



What we see from random initialization



A couple runs