ON THE BEHAVIOUR OF K-MEANS CLUSTERING OF A DISSIMILARITY MATRIX BY MEANS OF FULL MULTIDIMENSIONAL SCALING

Supplementary material

Table 1.

Averaged ARI values when the K-means solutions on $\boldsymbol{X}_c$, and on $\boldsymbol{X}_f$, are compared with the simulated values (columns $\boldsymbol{A}_c$ and $\boldsymbol{A}_f$ respectively), as well as for the SYNCLUS method on the dissimilarities ($\boldsymbol{A}_{syn}$). The values of $K^* = 4, 6, 8, 10$ and $N = 100, 150, 200$, for different degree of overlap were considered for equal and unequal-sized homogeneous clusters.

Average overlap, .01																		
	$N = 100$						$N = 150$						$N = 200$					
K^*	**Equal**			**Diff**			**Equal**			**Diff**			**Equal**			**Diff**		
	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	.999	1	.999	.999	1	.999	1	1	1	1	1	1	1	1	1	1	1	1
8	.998	1	.998	.999	1	.999	1	1	1	1	1	1	1	1	1	1	1	1
10	.987	.998	.962	.986	1	.978	.999	1	.982	.998	1	.994	.999	1	.991	.999	1	1
Average overlap, .05																		
	$N = 100$						$N = 150$						$N = 200$					
K^*	**Equal**			**Diff**			**Equal**			**Diff**			**Equal**			**Diff**		
	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$
4	.989	.999	.989	.942	.982	.942	1	1	1	.986	1	.986	.999	1	1	.998	1	0.999
6	.986	1	.986	.918	.982	.915	.999	1	.999	.963	.994	.960	.999	1	1	.972	.999	0.972
8	.883	.983	.887	.875	.985	.874	.960	.998	.961	.938	.999	.935	.981	1	.981	.985	1	0.985
10	.766	.891	.772	.755	.886	.747	.825	.952	.815	.801	.949	.803	.845	.985	.843	.828	.973	0.834
Average overlap, .1																		
	$N = 100$						$N = 150$						$N = 200$					
K^*	**Equal**			**Diff**			**Equal**			**Diff**			**Equal**			**Diff**		
	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$
4	.757	.873	.758	.571	.758	.579	.824	.961	.824	.711	.922	.726	.816	.976	.822	.802	.957	.818
6	.783	.804	.785	.794	.841	.796	.805	.849	.805	.816	.880	.814	.809	.878	.810	.819	.877	.817
8	.731	.823	.732	.756	.827	.756	.757	.860	.759	.777	.853	.767	.775	.875	.779	.782	.877	.785
10	.615	.771	.608	.475	.607	.467	.651	.794	.656	.514	.661	.506	.677	.810	.685	.535	.675	.530
Average overlap, .2																		
	$N = 100$						$N = 150$						$N = 200$					
K^*	**Equal**			**Diff**			**Equal**			**Diff**			**Equal**			**Diff**		
	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$
4	.667	.696	.668	.588	.640	.587	.679	.755	.678	.574	.642	.575	.674	.747	.676	.563	.671	.568
6	.625	.660	.630	.640	.679	.640	.702	.730	.705	.719	.735	.717	.735	.756	.732	.762	.778	.757
8	.575	.708	.576	.653	.766	.641	.631	.756	.636	.700	.794	.699	.663	.792	.669	.722	.805	.713
10	.443	.602	.444	.376	.494	.371	.500	.665	.506	.406	.534	.401	.515	.684	.516	.422	.562	.418
Average overlap, .3																		
	$N = 100$						$N = 150$						$N = 200$					
K^*	**Equal**			**Diff**			**Equal**			**Diff**			**Equal**			**Diff**		
	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$	$\mathbf{A}_c$	$\mathbf{A}_f$	$\mathbf{A}_{Syn}$
4	.660	.675	.660	.567	.585	.568	.671	.700	.672	.563	.592	.564	.669	.696	.670	.557	.597	.558
6	.383	.465	.390	.372	.483	.378	.463	.537	.460	.450	.550	.433	.530	.585	.519	.507	.582	.495
8	.419	.556	.412	.518	.673	.516	.487	.634	.498	.572	.729	.575	.534	.688	.536	.620	.747	.619
10	.335	.469	.341	.309	.401	.301	.388	.541	.385	.344	.455	.347	.404	.566	.407	.359	.476	.360

TABLE 2.

Recovery percentage of the simulated number of clusters according to the K-means solutions on $\boldsymbol{X}_f$, for clustered simulated datasets with high degree of overlap ($\overline{\omega} = 0.1$) and values of $N = 100, 150, 200$ and $K^* = 4, 6, 8, 10$ for non-homogeneous clusters of equal and unequal size.

Average overlap, .1						
Data generated for 10 distinct blocks, associated with $K^* = 4$						
Criteria	Equal size	Unequal size	Equal size	Unequal size	Equal size	Unequal size
	$N = 100$		**$N = 150$**		**$N = 200$**	
CH	0	0	0	0	0	0
H	3	0	7	2	18	9
SIL	0	0	0	0	0	0
KL	0	0	0	0	0	0
r	0	0	0	0	0	0
CH^*	0	0	0	0	0	0
H^*	6	0	6	1	15	4
r^*	7	5	7	2	12	6
Data generated for 21 distinct blocks, associated with $K^* = 6$						
Criteria	Equal size	Unequal size	Equal size	Unequal size	Equal size	Unequal size
	$N = 100$		**$N = 150$**		**$N = 200$**	
CH	0	0	0	0	0	0
H	0	0	0	2	0	2
SIL	0	0	0	0	0	0
KL	0	0	0	0	0	0
r	0	0	0	0	0	0
CH^*	0	0	0	0	0	0
H^*	0	0	0	0	0	3
r^*	7	7	6	9	11	11
Data generated for 36 distinct blocks, associated with $K^* = 8$						
Criteria	Equal size	Unequal size	Equal size	Unequal size	Equal size	Unequal size
	$N = 100$		**$N = 150$**		**$N = 200$**	
CH	0	0	0	0	0	0
H	0	0	0	0	1	2
SIL	0	0	0	0	0	0
KL	0	0	0	0	0	0
r	0	0	0	0	0	0
CH^*	0	0	0	0	0	0
H^*	0	0	0	0	1	0
r^*	7	3	4	3	7	5
Data generated for 55 distinct blocks, associated with $K^* = 10$						
Criteria	Equal size	Unequal size	Equal size	Unequal size	Equal size	Unequal size
	$N = 100$		**$N = 150$**		**$N = 200$**	
CH	0	0	0	0	0	0
H	0	0	0	0	0	0
SIL	0	0	0	0	0	0
KL	0	0	0	0	0	0
r	0	0	0	0	0	0
CH^*	0	0	0	0	0	0
H^*	0	0	0	0	0	0
r^*	5	9	4	7	5	10

Table 3.

Recovery percentage of the simulated number of clusters according to the K-means solutions on $\boldsymbol{X}_f$, for clustered simulated datasets with very high degree of overlap ($\overline{\omega} = 0.2$) and values of $N = 100, 150, 200$ and $K^* = 4, 6, 8, 10$ for non-homogeneous clusters of equal and unequal size.

Average overlap, .2						
Data generated for 10 distinct blocks, associated with $K^* = 4$						
Criteria	Equal size	Unequal size	Equal size	Unequal size	Equal size	Unequal size
	$N = 100$		**$N = 150$**		**$N = 200$**	
CH	0	0	0	0	0	0
H	1	0	0	0	0	1
SIL	1	0	0	0	0	0
KL	1	0	0	0	1	3
r	1	0	0	0	0	3
CH^*	0	0	0	0	0	0
H^*	0	0	0	0	0	0
r^*	12	1	6	2	8	5
Data generated for 21 distinct blocks, associated with $K^* = 6$						
Criteria	Equal size	Unequal size	Equal size	Unequal size	Equal size	Unequal size
	$N = 100$		**$N = 150$**		**$N = 200$**	
CH	0	0	0	0	0	0
H	0	0	0	0	0	0
SIL	0	0	0	0	0	0
KL	4	8	3	8	4	2
r	1	7	0	5	5	3
CH^*	0	0	0	0	0	0
H^*	0	0	0	0	0	0
r^*	2	0	0	3	3	0
Data generated for 36 distinct blocks, associated with $K^* = 8$						
Criteria	Equal size	Unequal size	Equal size	Unequal size	Equal size	Unequal size
	$N = 100$		**$N = 150$**		**$N = 200$**	
CH	0	0	0	0	0	0
H	0	0	0	0	0	0
SIL	0	0	0	0	0	0
KL	11	6	13	11	12	8
r	6	4	14	7	8	7
CH^*	0	0	0	0	0	0
H^*	0	0	0	0	0	0
r^*	4	8	6	5	2	6
Data generated for 55 distinct blocks, associated with $K^* = 10$						
Criteria	Equal size	Unequal size	Equal size	Unequal size	Equal size	Unequal size
	$N = 100$		**$N = 150$**		**$N = 200$**	
CH	0	0	0	0	0	0
H	0	0	0	0	0	0
SIL	0	0	0	0	0	0
KL	16	9	10	9	16	6
r	8	3	6	3	12	3
CH^*	0	0	0	0	0	0
H^*	0	0	0	0	0	0
r^*	6	1	4	3	3	0

Table 4.

Recovery percentage of the simulated number of clusters according to the K-means solutions on $\boldsymbol{X}_f$, for fully overlapping grouped simulated data sets ($\overline{\omega} = 0.3$) and values of $N = 100, 150, 200$ and $K^* = 4, 6, 8, 10$ for non-homogeneous clusters of equal and unequal size.

Average overlap, .3						
Data generated for 10 distinct blocks, associated with $K^* = 4$						
Criteria	Equal size	Unequal size	Equal size	Unequal size	Equal size	Unequal size
	$N = 100$		**$N = 150$**		**$N = 200$**	
CH	0	0	0	0	0	0
H	0	0	0	0	0	0
SIL	11	0	3	0	1	0
KL	1	0	1	0	0	0
r	1	0	1	0	0	0
CH^*	1	0	0	0	0	0
H^*	0	0	0	0	0	0
r^*	7	4	8	8	5	3
Data generated for 21 distinct blocks, associated with $K^* = 6$						
Criteria	Equal size	Unequal size	Equal size	Unequal size	Equal size	Unequal size
	$N = 100$		**$N = 150$**		**$N = 200$**	
CH	0	0	0	0	0	0
H	0	0	0	0	0	0
SIL	0	0	0	0	0	0
KL	0	0	0	0	0	0
r	0	0	0	0	0	0
CH^*	0	0	0	0	0	0
H^*	0	0	0	0	0	0
r^*	3	6	1	6	2	3
Data generated for 36 distinct blocks, associated with $K^* = 8$						
Criteria	Equal size	Unequal size	Equal size	Unequal size	Equal size	Unequal size
	$N = 100$		**$N = 150$**		**$N = 200$**	
CH	0	0	0	0	0	0
H	0	0	0	0	0	0
SIL	1	1	0	0	0	0
KL	0	0	0	0	0	0
r	0	0	0	0	0	0
CH^*	0	0	0	0	0	0
H^*	0	0	0	0	0	0
r^*	3	6	3	7	3	5
Data generated for 55 distinct blocks, associated with $K^* = 10$						
Criteria	Equal size	Unequal size	Equal size	Unequal size	Equal size	Unequal size
	$N = 100$		**$N = 150$**		**$N = 200$**	
CH	0	0	0	0	0	0
H	0	0	0	0	0	0
SIL	0	0	1	0	0	0
KL	0	0	0	0	0	0
r	0	0	0	0	0	0
CH^*	0	0	0	0	0	0
H^*	0	0	0	0	0	0
r^*	7	6	7	12	7	10

Table 5.

Averaged ARI values when the K-means solutions on $\boldsymbol{X}_a$ are compared with the simulated values, where $\boldsymbol{X}_a$ is obtained using the additive constant procedure of Cailliez (left panel). The averaged values of the additive constants provided by the Calliez procedure (central panel) and by the Lingoes procedure (right panel) are shown. The values considered were $K^* = 4, 6, 8, 10$ and $N = 100, 150, 200$, for non-overlapping clusters.

Non-overlapping clusters									
$N = 100$									
K^*	A_a			Cailliez			Lingoes		
	Equal	**%10**	**%60**	**Equal**	**%10**	**%60**	**Equal**	**%10**	**%60**
4	1	1	.990	5.62	5.75	6.10	5.79	6.05	6.96
6	1	.999	.968	4.01	4.05	4.46	3.86	3.98	5.00
8	.997	.997	.718	3.19	3.15	3.65	2.92	2.87	4.02
10	.955	.960	.503	2.66	2.66	3.03	2.33	2.33	3.20
$N = 150$									
K^*	A_a			Cailliez			Lingoes		
	Equal	**%10**	**%60**	**Equal**	**%10**	**%60**	**Equal**	**%10**	**%60**
4	1	1	1	6.93	6.96	7.53	7.33	7.40	8.95
6	.999	1	.971	4.98	5.04	5.43	4.93	5.01	6.34
8	.998	.997	.717	3.88	3.94	4.42	3.66	3.70	4.97
10	.962	.969	.518	3.19	3.28	3.62	2.87	3.03	3.92
$N = 200$									
K^*	A_a			Cailliez			Lingoes		
	Equal	**%10**	**%60**	**Equal**	**%10**	**%60**	**Equal**	**%10**	**%60**
4	1	1	1	8.00	8.34	8.93	8.50	9.16	10.77
6	1	1	.976	5.80	5.73	6.35	5.90	5.80	7.43
8	.998	1	.706	4.52	4.50	5.12	4.37	4.37	5.91
10	.706	.969	.507	5.12	3.75	4.24	3.48	3.50	4.77

Table 6.

Averaged ARI values when the K-means solutions on $\boldsymbol{X}_a$ are compared with the simulated values, where $\boldsymbol{X}_a$ is obtained using the additive constant procedure of Cailliez (left panel). The averaged values of the additive constants provided by the Calliez procedure (central panel) and by the Lingoes procedure (right panel) are shown. The values considered were $K^* = 4, 6, 8, 10$ and $N = 100, 150, 200$, for low ($\overline{\omega} = 0.01$) and moderate ($\overline{\omega} = 0.05$) degrees of overlap. Equal and unequal-sized, non-homogeneous clusters were considered.

			Average overlap, .01			
			$N = 100$			
K^*	A_a		Cailliez		Lingoes	
	Equal	**Unequal**	**Equal**	**Unequal**	**Equal**	**Unequal**
4	.998	1	3.35	4.36	3.01	5.03
6	.872	.911	3.51	3.54	3.08	3.60
8	.911	.916	2.84	2.61	2.21	2.05
10	.803	.656	3.40	3.68	2.93	3.44
			$N = 150$			
K^*	A_a		Cailliez		Lingoes	
	Equal	**Unequal**	**Equal**	**Unequal**	**Equal**	**Unequal**
4	1	1	4.12	5.42	3.83	6.48
6	.875	.927	4.37	4.38	3.92	4.62
8	.908	.925	3.41	2.98	2.72	2.36
10	.804	.671	4.25	4.22	3.76	3.78
			$N = 200$			
K^*	A_a		Cailliez		Lingoes	
	Equal	**Unequal**	**Equal**	**Unequal**	**Equal**	**Unequal**
4	1	1	4.70	6.24	4.43	7.62
6	.880	.936	5.05	5.10	4.57	5.46
8	.918	.930	3.89	3.48	3.15	2.81
10	.811	.674	4.86	4.87	4.35	4.40
			Average overlap, .05			
			$N = 100$			
K^*	A_a		Cailliez		Lingoes	
	Equal	**Unequal**	**Equal**	**Unequal**	**Equal**	**Unequal**
4	.769	.456	6.10	7.68	6.11	9.44
6	.645	.739	6.89	6.78	6.83	7.80
8	.639	.728	7.54	6.64	7.16	6.36
10	.523	.446	7.74	7.74	7.77	8.04
			$N = 150$			
K^*	A_a		Cailliez		Lingoes	
	Equal	**Unequal**	**Equal**	**Unequal**	**Equal**	**Unequal**
4	.775	.487	7.62	9.74	7.71	12.21
6	.669	.781	8.55	8.50	8.52	10.07
8	.673	.747	9.19	8.00	8.79	7.63
10	.525	.434	9.74	9.68	9.92	9.86
			$N = 200$			
K^*	A_a		Cailliez		Lingoes	
	Equal	**Unequal**	**Equal**	**Unequal**	**Equal**	**Unequal**
4	.791	.524	8.79	11.37	8.99	14.42
6	.695	.804	9.92	9.97	9.98	11.82
8	.694	.756	10.75	9.48	10.26	9.13
10	.533	.449	11.36	11.34	11.46	11.53

Table 7.

Averaged frequency at which the local minimum is reached in K-means clustering. The values considered were 5000 replicates, $K^* = 4, 6, 8, 10$ and $N = 100, 150, 200$. Non-overlapping clusters and non-homogeneous clusters with different degrees of overlap were considered.

Non-overlapping clusters

	$N = 100$			$N = 150$			$N = 200$		
K^*	**Equal**	**%10**	**%60**	**Equal**	**%10**	**%60**	**Equal**	**%10**	**%60**
4	69.6	63.7	35.2	72.4	83.8	62.0	91.0	90.0	81.0
6	54.2	28.6	51.8	50.6	76.4	29.2	50.0	55.0	70.0
8	51.0	58.4	16.0	60.8	51.8	26.4	46.0	54.0	28.0
10	27.0	20.2	18.4	29.4	21.4	2.4	40.0	59.0	33.0

Overlapping clusters

Average overlap, .01

	$N = 100$		$N = 150$		$N = 200$	
K^*	**Equal**	**Unequal**	**Equal**	**Unequal**	**Equal**	**Unequal**
4	87.0	87.0	79.6	86.4	77.4	89.0
6	21.0	15.4	28.8	16.0	22.0	10.6
8	37.0	34.2	42.2	37.0	39.0	30.0
10	23.0	10.4	16.0	13.6	16.4	14.4

Average overlap, .05

	$N = 100$		$N = 150$		$N = 200$	
K^*	**Equal**	**Unequal**	**Equal**	**Unequal**	**Equal**	**Unequal**
4	58.0	40.0	64.0	35.0	54.0	40.0
6	16.6	26.0	28.0	35.0	27.0	31.0
8	18.0	15.0	10.0	14.0	11.6	13.8
10	4.0	3.0	3.0	4.0	3.8	1.2

Average overlap, .1

	$N = 100$		$N = 150$		$N = 200$	
K^*	**Equal**	**Unequal**	**Equal**	**Unequal**	**Equal**	**Unequal**
4	11.2	8.2	17.0	1.0	1.6	2.0
6	4.4	2.0	9.6	3.4	15.2	6.8
8	4.6	0.4	3.6	1.4	0.6	1.2
10	1.8	3.2	1.2	0.2	0.2	1.2

Average overlap, .2

	$N = 100$		$N = 150$		$N = 200$	
K^*	**Equal**	**Unequal**	**Equal**	**Unequal**	**Equal**	**Unequal**
4	27.8	8.8	8.4	0.4	14.8	0.4
6	1.6	5.4	0.2	1.2	0.6	0.2
8	7.4	0.6	1.2	2.8	0.4	0.2
10	0.2	0.2	0.2	0.4	0.2	0.2

Average overlap, .3

	$N = 100$		$N = 150$		$N = 200$	
K^*	**Equal**	**Unequal**	**Equal**	**Unequal**	**Equal**	**Unequal**
4	3.0	2.6	5.4	1.2	2.6	1.8
6	1.6	0.2	0.2	0.8	0.2	0.2
8	0.8	0.8	0.6	0.2	0.2	0.2
10	0.2	0.2	0.2	0.2	0.2	0.2