1 simulation

Table 1: Simulation results

Number of clusters	1	2	3	4	5	6	7	8	9	10	PE
Setting 1											
Oracle	**100**	0	0	0	0	0	0	0	0	0	1
Gap	**92**	0	1	1	1	2	1	0	1	1	8.0 ± 28.3
Gaussian-Mix	**100**	0	0	0	0	0	0	0	0	0	1.0 ± 0
CH	–	–	–	–	–	–	–	–	–	–	–
Hartigan	–	–	–	–	–	–	–	–	–	–	–
Jump	**0**	0	0	0	0	0	1	8	34	57	102.3 ± 58.7
Prediction strength	**100**	0	0	0	0	0	0	0	0	0	1.0 ± 0
Stability	–	–	–	–	–	–	–	–	–	–	–
Gabriel CV	**100**	0	0	0	0	0	0	0	0	0	1.0 ± 0
Wold CV	**100**	0	0	0	0	0	0	0	0	0	1.0 ± 0
Setting 2											
Oracle	0	**100**	0	0	0	0	0	0	0	0	1
Gap	0	**98**	1	1	0	0	0	0	0	0	1.2 ± 1.3
Gaussian-Mix	0	**100**	0	0	0	0	0	0	0	0	1.0 ± 0
CH	0	**100**	0	0	0	0	0	0	0	0	1.0 ± 0
Hartigan	0	**0**	6	16	21	14	12	10	8	13	28.1 ± 50.8
Jump	0	**70**	0	0	0	0	0	2	7	21	15.5 ± 55.4
Prediction strength	0	**100**	0	0	0	0	0	0	0	0	1.0 ± 0
Stability	0	**100**	0	0	0	0	0	0	0	0	1.0 ± 0
Gabriel CV	0	**99**	1	0	0	0	0	0	0	0	1.1 ± 1.0
Wold CV	0	**66**	3	14	11	4	1	0	1	0	6.0 ± 10.6
Setting 3											
Oracle	0	0	0	**100**	0	0	0	0	0	0	1
Gap	0	0	0	**69**	28	3	0	0	0	0	1.3 ± 0.5
Gaussian-Mix	85	2	3	**3**	1	0	4	2	0	0	35.9 ± 13.7
CH	0	36	28	**36**	0	0	0	0	0	0	12.1 ± 9.9
Hartigan	0	0	13	**75**	11	0	1	0	0	0	2.5 ± 3.5
Jump	0	0	0	**0**	0	0	0	0	14	86	5.6 ± 0.4
Prediction strength	56	0	0	**44**	0	0	0	0	0	0	23.5 ± 20.6
Stability	0	0	0	**33**	59	8	0	0	0	0	1.7 ± 0.5
Gabriel CV	0	0	0	**100**	0	0	0	0	0	0	1.0 ± 0
Wold CV	0	0	0	**100**	0	0	0	0	0	0	1.0 ± 0
Number of clusters	≤ 6	7	8	9	10	11	12	13	14	15	
Setting 4											
Oracle	0	0	0	0	**100**	0	0	0	0	0	1
Gap	0	0	0	0	**12**	39	23	16	7	3	1.5 ± 0.4
Gaussian-Mix	71	6	4	2	**4**	4	0	2	3	4	22.2 ± 13.5
CH	84	6	6	4	**0**	0	0	0	0	0	20.8 ± 8.9
Hartigan	13	4	13	21	**17**	12	5	9	1	5	4.4 ± 4.8
Jump	0	0	0	20	**80**	0	0	0	0	0	1.4 ± 0.9
Prediction strength	100	0	0	0	**0**	0	0	0	0	0	35.2 ± 4.4
Stability	0	0	0	0	**0**	3	24	29	28	16	2.0 ± 0.3
Gabriel CV	0	0	0	0	**100**	0	0	0	0	0	1.0 ± 0
Wold CV	0	0	0	0	**100**	0	0	0	0	0	1.0 ± 0

In setting 1, "-" means the method can not be used with parameter $k = 1$.

Number of clusters	1	2	3	4	5	6	7	8	9	10	PE
Setting 5											
Oracle	0	0	0	**100**	0	0	0	0	0	0	1
Gap	0	0	0	**66**	25	7	1	1	0	0	1.8 ± 1.3
Gaussian-Mix	0	0	0	**56**	36	6	2	0	0	0	2.0 ± 1.3
CH	0	5	25	**53**	14	2	1	0	0	0	10.3 ± 17.0
Hartigan	0	0	24	**61**	5	6	3	1	0	0	7.3 ± 11.5
Jump	0	0	0	**73**	0	0	1	0	5	21	3.1 ± 3.5
Prediction strength	74	8	5	**13**	0	0	0	0	0	0	114.5 ± 67.8
Stability	0	4	5	**23**	36	23	7	2	0	0	5.7 ± 10.2
Gabriel CV	0	0	0	**100**	0	0	0	0	0	0	1.0 ± 0.0
Wold CV	0	0	0	**99**	1	0	0	0	0	0	1.0 ± 0.1
Setting 6											
Oracle	0	0	**100**	0	0	0	0	0	0	0	1
Gap	0	0	**10**	8	13	11	12	11	17	18	9.8 ± 5.9
Gaussian-Mix	0	0	**88**	12	0	0	0	0	0	0	1.4 ± 1.3
CH	0	17	**74**	8	1	0	0	0	0	0	4.0 ± 6.3
Hartigan	0	0	**87**	10	3	0	0	0	0	0	1.6 ± 1.6
Jump	0	0	**0**	0	0	1	2	13	32	52	13.7 ± 5.0
Prediction strength	19	2	**78**	1	0	0	0	0	0	0	10.3 ± 20.5
Stability	0	0	**24**	39	29	7	1	0	0	0	4.5 ± 3.0
Gabriel CV	0	2	**97**	1	0	0	0	0	0	0	1.3 ± 2.1
Wold CV	0	0	**89**	9	1	0	1	0	0	0	1.6 ± 1.9

Last column gives the mean and standard deviation of PE for each algorithm.

2 Real data application

Table 2: Number of clusters selected by each algorithm

	Congress Voting	Breast Cancer	Sonar
CH-index	2	2	2
Hartigan	3	3	3
Jump	9	9	10
Prediction strength	2	2	1
Bootstrap stability	2	2	10
Gap	10	9	10
Gaussian-Mix	7	5	1
Gabriel	2	2	2
Wold	2	3	10

All the algorithms executed with their default parameter settings with k ranges from 1 to 10

3 CV error with two multivariate normal distributed clusters

3.1 Setup

There are two clusters G_1 and G_2, where observations from G_1 are distributed as

$$N\left(\begin{pmatrix}\mu_1^X\\ \mu_1^Y\end{pmatrix}, \mathbf{I}\right)$$

and observations from G_2 are distributed as

$$N\left(\begin{pmatrix}-\mu_1^X\\ -\mu_1^Y\end{pmatrix}, \mathbf{I}\right)$$

where $\mu_1^X > 0$ and $\mu_1^Y > 0$. If the true cluster is single cluster G with

$$P(G = G_1) = P(G = G_2) = 1/2$$

After applying K-means on Y axis with $k = 2$ to the whole data, and assume the observation number $n \to \infty$, we have the estimated center of G_1 be

$$\bar{\mu}_1^Y = 2\varphi(\mu_1^Y) + 2\mu_1^Y\Phi(\mu_1^Y) - \mu_1^Y \tag{1}$$

and the estimated center of G_2 be

$$\bar{\mu}_2^Y = -2\varphi(\mu_1^Y) - 2\mu_1^Y\Phi(\mu_1^Y) + \mu_1^Y \tag{2}$$

where $\varphi()$ and $\Phi()$ are the standard normal probability and cumulative distribution function respectively.

3.2 CV error with $k = 1$ and $k = 2$

By symmetry, the CV error for points from G_1 is same as the points from G_2. Because $P(G = G_1) = P(G = G_2) = 1/2$, the CV error for $k = 2$ can be calculated solely from G_2, that is

$$
\begin{aligned}
CV(2) &= E[(Y - \hat{Y})^2], \quad Y \sim N(-\mu_1^Y, 1) \\
&= P(\hat{G} = 2|G = 2) \cdot E[(Y - \hat{Y})^2|\hat{G} = 2] + P(\hat{G} = 1|G = 2) \cdot E[(Y - \hat{Y})^2|\hat{G} = 1] \\
&= P(\hat{G} = 2|G = 2) \cdot E[(Y - \bar{\mu}_2^Y)^2] + P(\hat{G} = 1|G = 2) \cdot E[(Y - \bar{\mu}_1^Y)^2] \\
&= \Phi(\mu_1^X)[var(Y) + (-\mu_1^Y - \bar{\mu}_2^Y)^2] + [1 - \Phi(\mu_1^X)][var(Y) + (-\mu_1^Y - \bar{\mu}_1^Y)^2] \\
&= \Phi(\mu_1^X)[1 + (\mu_1^Y + \bar{\mu}_2^Y)^2] + [1 - \Phi(\mu_1^X)][1 + (\mu_1^Y + \bar{\mu}_1^Y)^2] \\
\bar{\mu}_2^Y = -\bar{\mu}_1^Y &= 1 + (\mu_1^Y + \bar{\mu}_1^Y)^2 - 4\Phi(\mu_1^X)\mu_1^Y\bar{\mu}_1^Y
\end{aligned}
$$

where $\bar{\mu}_1^Y$ is given in equation (1).

When $k = 1$, the result is straight forward since the estimated center will be $(0, 0)$, so

$$CV(1) = E[Y^2] = 1 + (\mu_1^Y)^2$$

where $Y \sim N(-\mu_1^Y, 1)$ or $Y \sim N(\mu_1^Y, 1)$