Homework 2

16.10.9

Chuan Lu, 13300180056, chuanlu13@fudan.edu.cn

Problem 1. Explain the relationship between spectral clustering, normalized spectral clustering and graph cut.

Solution. We assume $k = 2$ in the following explanation.

For spectral clustering, we already know that the eigenvector u_2 corresponding to the second smallest eigenvalue λ_2 of the Laplacian matrix of the graph is the solution for the minimization

$$\min_{\mathbf{f}} \ \mathbf{f}^\top \mathbf{L}\mathbf{f},$$
$$s.t. \quad \mathbf{f}^\top \mathbf{f} = 1, \mathbf{f}^\top \mathbf{1} = 0$$

In fact, if we define

$$G(f) = G(f_1, \cdots, f_n) = f^\top \mathbf{L} f - \lambda_1(f^\top f - 1) - \lambda_2 f^\top \mathbf{1}$$
$$= \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2 - \lambda_1(\sum_{i=1}^{n} f_i^2 - 1) - \lambda_2 \sum_{i=1}^{n} f_i$$

Then according to Lagrange Theorm, we let

$$\frac{\partial G}{\partial f_i} = \sum_{j=1}^{n}(f_i - f_j) - 2\lambda_1 f_i - \lambda_2 = 0, \quad i = 1:n$$

If we add the n equations, we get

$$2\lambda_1 \sum_{i,j=1}^{n} f_i + n\lambda_2 = 0$$

Since $f^\top \mathbf{1} = 0$, then $\lambda_2 = 0$. So

$$\sum_{j=1}^{n} w_{ij} f_j = \lambda f_i, \quad \lambda = \sum_{j=1}^{n} w_{ij} - 2\lambda_1, \quad i = 1:n,$$

which means f is an eigenvector of $\mathbf{L}$.

For Min Cut, if we define $\mathbf{f} = (f_1, f_2 \cdots f_n)^\top \in \mathbb{R}^n$,

$$f_i = \begin{cases} 1, & if\ v_i \in \mathbb{A} \\ -1, & if\ v_i \in \bar{\mathbb{A}} \end{cases}$$

Then

$$(f_i - f_j)^2 = \begin{cases} 4, & i \in \mathbb{A}, j \in \bar{\mathbb{A}} \\ 0, & else \end{cases}$$

So

$$MCut(A_1, \cdots, A_k) = \sum_{i=1}^{k} cut(A_i, \bar{A_i}) = \frac{1}{4} f^\top \mathbf{L} f$$

We can minimize this with the f from spectral clustering.

For Ratio Cut, we can define f as

$$f_i = \begin{cases} \sqrt{|\bar{A}|/|\bar{A}|}, & v_i \in \mathbb{A} \\ -\sqrt{|A|/|\bar{A}|}, & v_i \in \bar{\mathbb{A}} \end{cases}$$

Then

$$(f_i - f_j)^2 = \begin{cases} \frac{|A|}{|\bar{A}|} + \frac{|\bar{A}|}{|A|} + 2, & v_i, v_j \; in \; the \; same \; part \\ 0, & else \end{cases}$$

So

$$RCut(A, \bar{A}) = \frac{1}{2} \frac{|A||\bar{A}|}{(|A| + |\bar{A}|)^2} \frac{1}{|A| + |\bar{A}|} \sum_{i,j} (f_i - f_j)^2 w_{ij} = \frac{f^\top \mathbf{L} f}{2n^2}$$

For normalized spectral clustering, we have

$$L_{sym} = I - D^{-1/2} W D^{-1/2} = D^{-1/2} L D^{-1/2}$$

and normalized cut can be done using g^*, the Fiedler vector of $D^{-1/2} L D^{-1/2}$.

When $k \geq 3$, we can find k points x_i in $\mathbb{R}^k$, s.t. $|x_i - x_j| = |x_k - x_l|$, for each i, j, k, l.
Then define $f_i = x_i$, and it is same with $k = 2$.

Problem 2. Cluster Data.csv with spectral clustering, construct three different graphs and compare the difference with result of K-means.

Solution. For Data1.csv, the result is shown as follows.

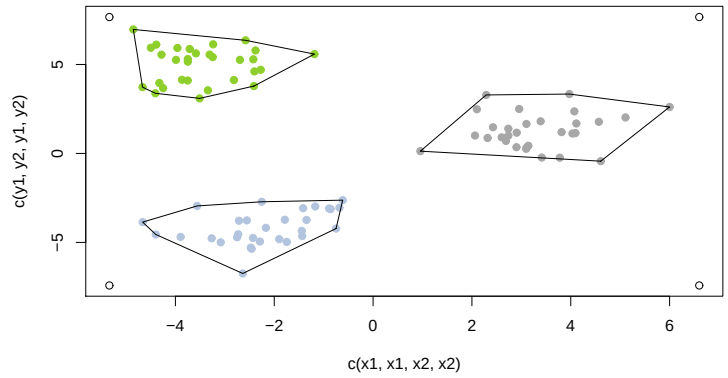
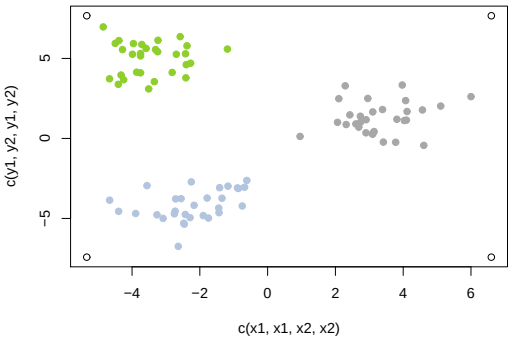
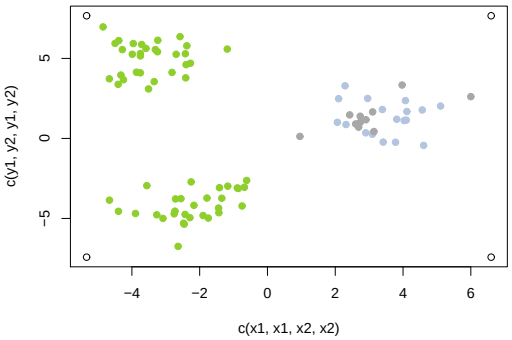


Figure 1: Kmeans

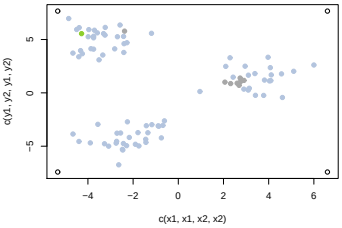
For Data3.csv, the result is shown as follows.



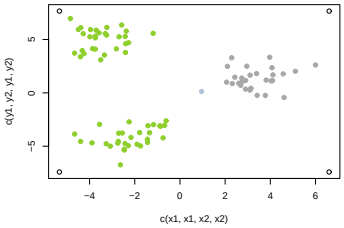
(a) Full-connect1



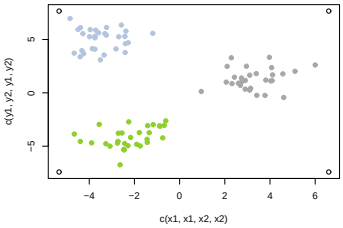
(b) Full-connect2



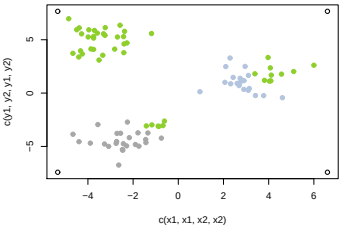
(a) Epsilon = 0.1



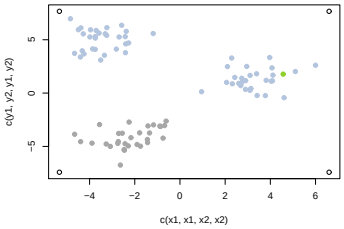
(b) Epsilon = 1



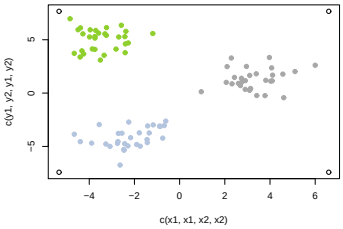
(c) Epsilon = 3



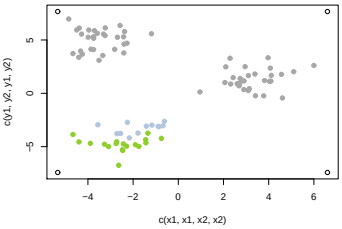
(a) Directed, n = 3



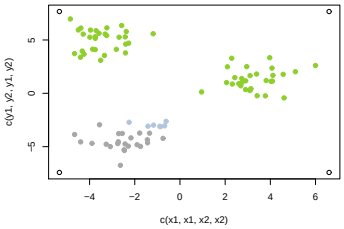
(b) Directed, n = 5



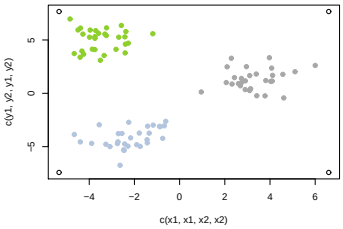
(c) Directed, n = 10



(a) Undirected, n = 3



(b) Undirected, n = 5



(c) Undirected, n = 20

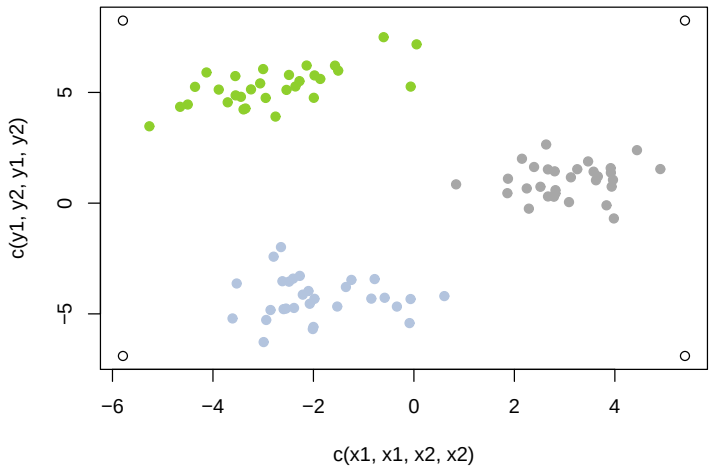
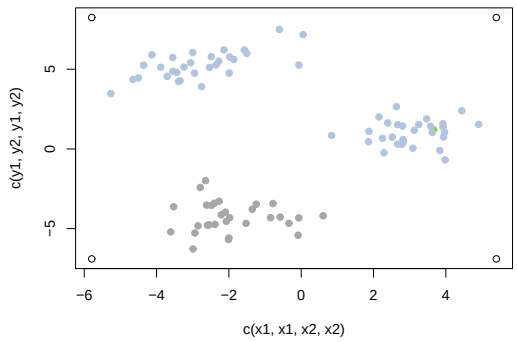
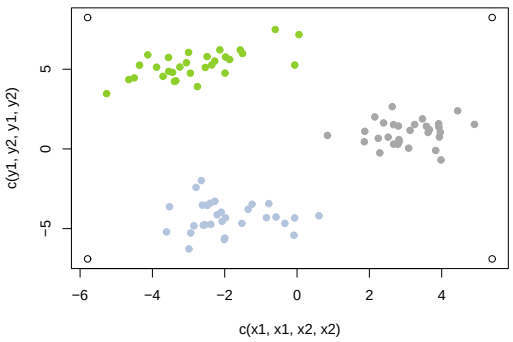


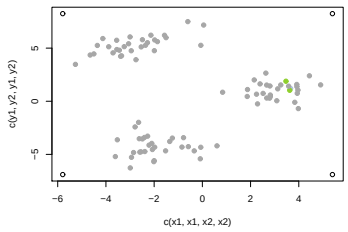
Figure 2: Kmeans



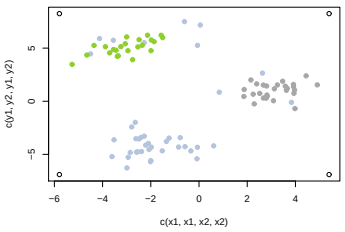
(a) Full-connect1



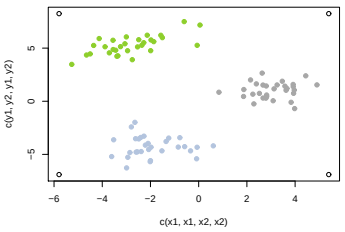
(b) Full-connect2



(a) Epsilon = 0.1



(b) Epsilon = 2



(c) Epsilon = 5

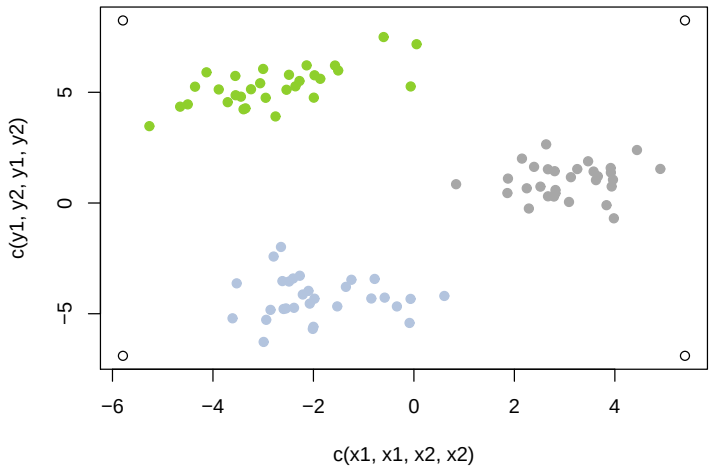


Figure 3: directed, n = 3

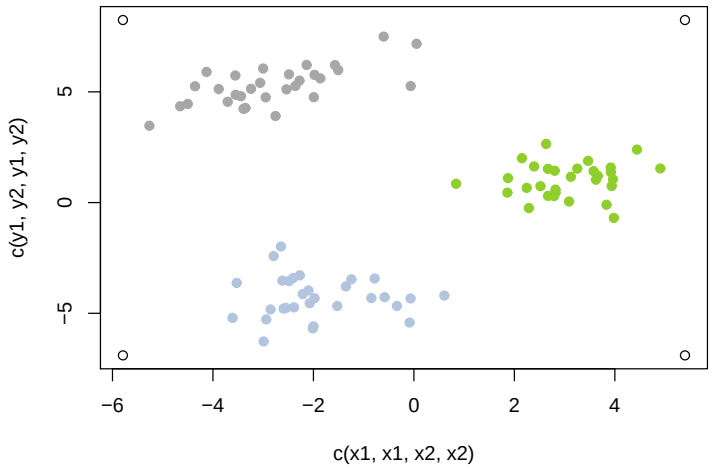


Figure 4: Undirected, n = 3

From the results we can know that for these groups, spectral clustering is not always better than kmeans. The parameters in spectral clustering are also important.

Problem 3. Cluster circles, curves with spectral clustering

Solution. The results of circles.csv is as follows.

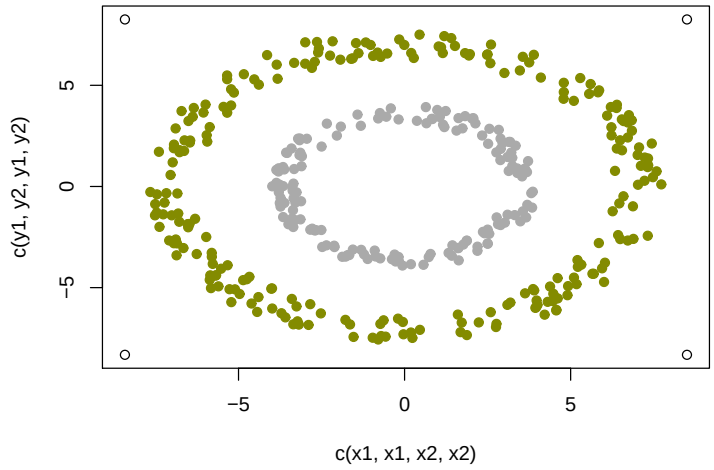
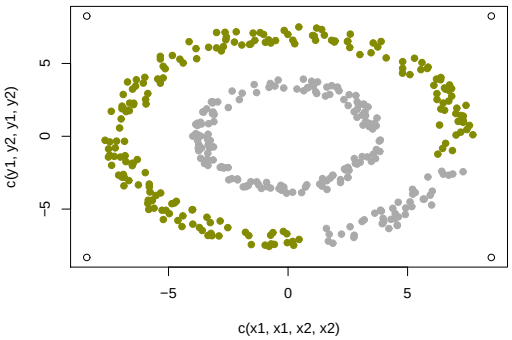
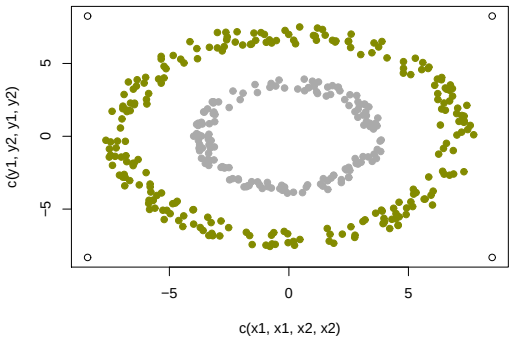


Figure 5: full connection

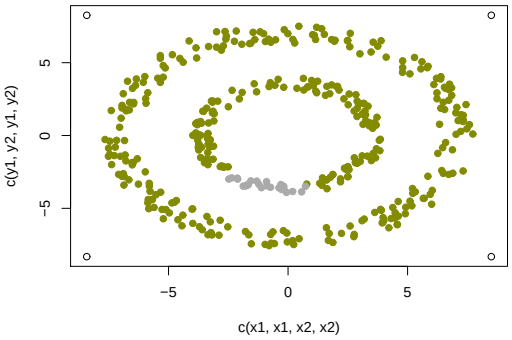


(a) epsilon = 1

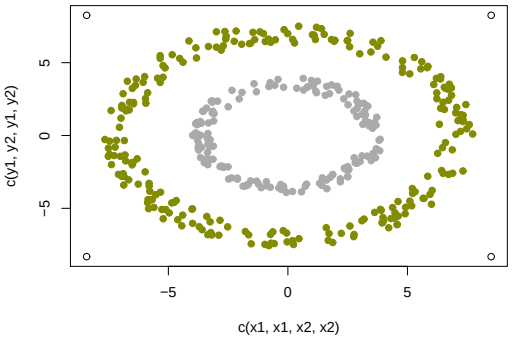


(b) epsilon = 2

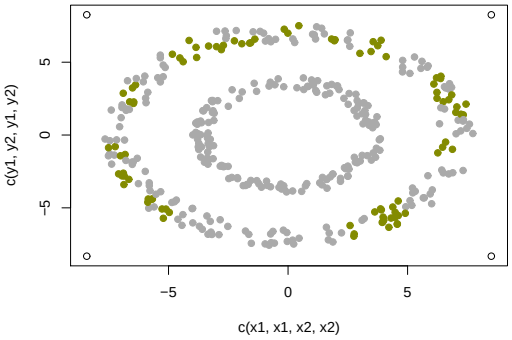
The result of curves.csv is shown as follows.



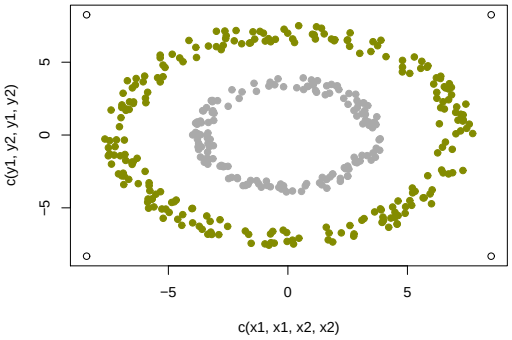
(a) directed, n = 5



(b) directed, n = 10



(a) undirected, n = 2



(b) undirected, n = 5

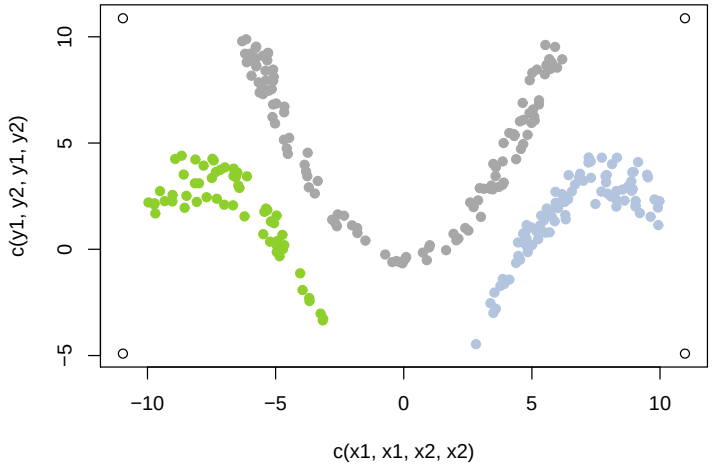
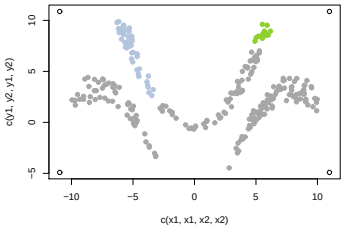
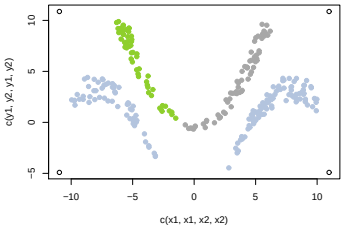


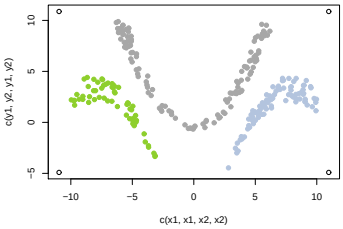
Figure 6: full connection



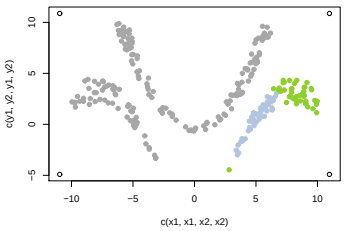
(a) epsilon = 1



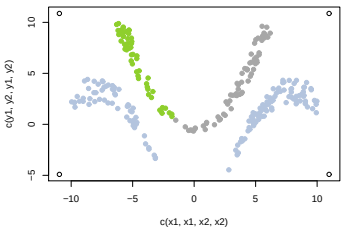
(b) epsilon = 3



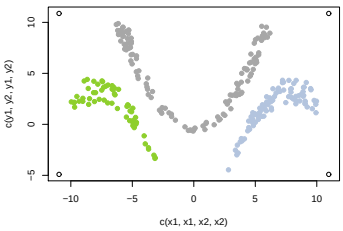
(c) epsilon = 10



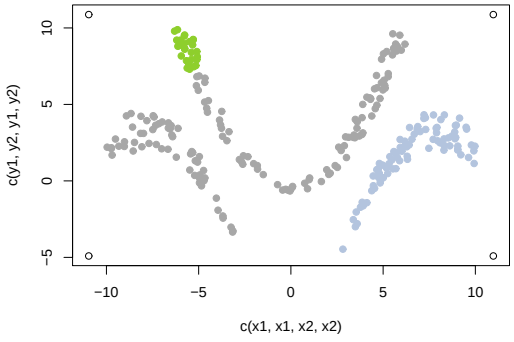
(a) directed, n = 10



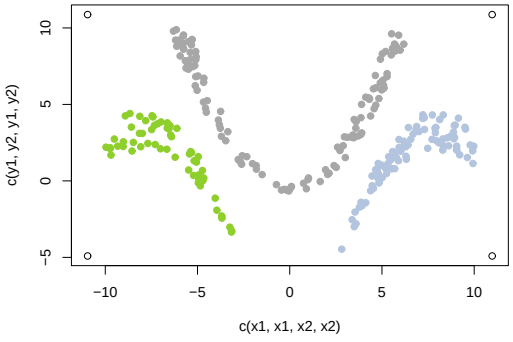
(b) directed, n = 20



(c) directed, n = 30



(a) undirected, n = 5



(b) undirected, n = 10