The Information Cut and Graph Spectral Clustering

Robert Jenssen\textsuperscript{1}, Jose C. Principe\textsuperscript{2} and Torbjørn Eltoft\textsuperscript{1}

\textsuperscript{1} Electrical Engineering Group, University of Tromsø, Norway

\textsuperscript{2} Computational NeuroEngineering Laboratory, University of Florida, USA

Many thanks to Deniz Erdogmus and Kenneth E. Hild
Goal of this Presentation

❄ Show a map of the Arctic Region
   ✐ and a couple of pictures...

➲ Introduce the Information Cut
   ✐ A link between Information Theory and Graph Theory

◼ Develop a Graph Spectral Clustering Algorithm
   ✐ in an Information Theoretic framework
The Arctic
The Arctic
Outline

1. Clustering by the Graph CUT
2. The Information Cut
3. Information Theoretic Framework for Clustering by Graph Eigenvectors
4. Concluding Remarks
The Graph Theoretic CUT

- Clustering by removing edges of a graph $C$, resulting in the disjoint sets $C_1$ and $C_2$.

- Minimize the CUT;

\[
CUT(C_1, C_2) = \sum_{i,j=1}^{N_1,N_2} G_{ij},
\]

- $N_1$ ($N_2$) - number of nodes in subgraph $C_1$ ($C_2$)
- $G_{ij}$ - similarity weight between nodes $i$ and $j$
The Graph Theoretic CUT

The **CUT**: Sum of the red similarity weights.

Graph Theoretic CUTs

Forefront CUT: sum of the red similarity weights divided by the sum of the blue.

- Foreground CUT: \( \frac{\text{CUT}}{\sum_1} \)
- Affinity (graph) matrix \( G = \begin{bmatrix} G_{ij} \end{bmatrix}_{i,j=1,...,N} \)
- \( m : m_i = 1(0) \) if \( x_i \in C_1(C_2) \)
CUTs: Clustering by Eigenvectors

- Rayleigh Quotient: \[ \max_{\mathbf{m}} R = \frac{\mathbf{m}^T G \mathbf{m}}{\mathbf{m}^T \mathbf{m}} \]
  - \[ \mathbf{m} = \mathbf{e}_{\text{max}} \text{ and } R = \lambda_{\text{max}} \]

- \[ \lambda_1, \ldots, \lambda_n \]: The spectrum of \( G \)

- The eigenvectors of \( G \) exhibit discriminatory power!

- “Hot” research area in segmentation and clustering during the last 5-10 years, with impressive results.
CUTs: Open Questions 😞

- Discriminatory power only for a suitably chosen edge-weight function
  - Often chosen to be the Gaussian. Why? How to select $\sigma$?

- It is often not clear what criteria are optimized in spectral clustering...

- Oft-posed question: Why does spectral clustering work?
The Information Cut

- The Cauchy-Schwarz pdf Distance (Principe et. al ’00)

\[ D_{CS} = - \log \frac{\int p(x)q(x)dx}{\sqrt{\int p^2(x)dx \int q^2(x)dx}} \geq 0. \quad (2) \]

- Gockay and Principe (’02) used the numerator for clustering, termed the CEF.
The Information Cut

Parzen pdf estimation

\[ \hat{p}(x) = \frac{1}{N_p} \sum_{i=1}^{N_p} G(x - x_i, \sigma^2 I), \]  

(3)

\[ \text{and } \hat{q}(x) = \frac{1}{N_q} \sum_{j=1}^{N_q} G(x - x_j, \sigma^2 I), \text{ for Gaussian } G(x, \sigma^2 I). \]

Convolution theorem for Gaussians

\[ \int G(x - x_i, \sigma^2 I) G(x - x_j, \sigma^2 I) dx = G_{ij,2\sigma^2 I}, \]  

(4)

\[ \text{where } G_{ij,2\sigma^2 I} = G(x_i - x_j, 2\sigma^2 I). \]
The Information Cut

Numerator of (2)

$$\int p(x)q(x)dx = \frac{1}{N_pN_q} \sum_{i,j=1}^{N_p,N_q} G_{ij},2\sigma^2I$$ \hspace{1cm} (5)

Denominator of (2)

$$\int p^2(x)dx = \frac{1}{N_p^2} \sum_{i,i'=1}^{N_p,N_p} G_{ii'},2\sigma^2I$$ \hspace{1cm} (6)

and likewise for $\int q^2(x)dx$. 
The Information Cut (IC)

\[
IC = \frac{\sum_{i,j=1}^{N_p,N_q} G_{ij,2\sigma^2I}}{\sqrt{\sum_{i,i'=1}^{N_p,N_p} G_{ii',2\sigma^2I} \sum_{j,j'=1}^{N_q,N_q} G_{jj',2\sigma^2I}}}. \quad (7)
\]

The numerator is exactly the CUT, with \( G_{ij,2\sigma^2I} \) as the similarity weight between nodes \( i \) and \( j \).
The Information Cut

The Information Cut is a normalized CUT.

\[
IC = \frac{CUT}{\sqrt{\Sigma_1 \Sigma_2}} = \frac{m^T G (1 - m)}{\sqrt{m^T G m (1 - m)^T G (1 - m)}}.
\]
The Information Cut

- A new graph cut based on information theory
- Parzen pdf estimation with Gaussian kernel $\Rightarrow$ Gaussian graph edge-weight function
- $\sigma$: pdf estimates relatively accurate, e.g. $MISE$ (not unproblematic)
- Can the IC be utilized as a theoretically well defined graph spectral cost function for clustering by eigenvectors?
Spectral Solution

- Schur decomposition:

\[ G = E \Lambda E^T = E \Lambda^{\frac{1}{2}} \Lambda^{\frac{1}{2}} E^T, \]  

\( \Rightarrow \) \( E: \) eigenvectors, \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_N): \) eigenvalues

- Define \( u = \Lambda^{\frac{1}{2}} E^T m, \quad t = \Lambda^{\frac{1}{2}} E^T 1 \)

- Hence

\[ IC = \frac{u^T (t - u)}{\sqrt{||u||^2 ||t - u||^2}} = \cos \angle(u, t - u). \]  

(10)
Spectral Solution

Find \( \min_u IC(u) \) subject to binary \( m \):

\[
m = \Lambda^{-\frac{1}{2}}Eu = \sum_{i=1}^{N} \frac{u_i}{\sqrt{\lambda_i}} e_i. \tag{11}
\]

Note that any \( \hat{u} : \hat{u}_i \in \{t_i, 0\} \)

\[
\Rightarrow \hat{u} \perp t - \hat{u} \iff \cos \angle(\hat{u}, t - \hat{u}) = 0. \tag{12}
\]

Idea: Our \( u \) should be “close” to one particular such \( \hat{u} \) (but cannot be equal to it)
Use \( \hat{u} \) to approximate the discrete solution:

\[
\hat{m} = \sum_{i=1}^{N} \frac{\hat{u}_i}{\sqrt{\lambda_i}} e_i = \sum_{i=1}^{N} w_i (e_i 1) e_i, \quad (13)
\]

\( w_i = 1 \) (0) iff \( u_i = t_i \) (0).

Our solution given by: \( m = \text{threshold}(\hat{m}) \big|_{1/2} \)

How do we find \( \hat{u} \)?
Spectral Solution: Reducing Complexity

- Only $M \ll N$ of the $t_i = \sqrt{\lambda_i} (e_i 1)$, $i = 1, \ldots, N$, deviate significantly from zero!

- Only a few $\lambda_i > 0$

- Frequently $e_i 1 \approx 0$

- We are faced with a low-dimensional optimization problem of finding $\hat{u}_k \in \{t_k, 0\}$, $k = 1, \ldots, M$. 
The approximate solution is either

$$\hat{m} = (e_1 \mathbf{1}) e_1 \quad \text{or} \quad \hat{m} = (e_2 \mathbf{1}) e_2.$$  

(14)
Spectral Solution: A Novel Algorithm

- Eigendecompose $G$
- Calculate $t$
- $M = \# \text{ significant } t_i$'s
- Initialize: $\hat{\mathbf{m}} = (e_1^T \mathbf{1}) e_1$
- Determine $IC(\mathbf{m})$
- while loop
  * for $j = 1 : M-1$
    - $\hat{\mathbf{m}}_j = \hat{\mathbf{m}} + (e_j'^T \mathbf{1}) e_j$
    - Store $IC(\mathbf{m}_j)$
  end for
- Find $\hat{\mathbf{m}}_{\text{min}} : \min_{\mathbf{m}_j} IC(\mathbf{m}_j), \ j = 1, \ldots, M - 1$
- if $IC(\mathbf{m}_{\text{min}}) < IC(\mathbf{m})$
  - $\hat{\mathbf{m}} = \hat{\mathbf{m}}_{\text{min}}$
  - $M = M - 1$
else stop
end if
until stop
Some Results

1. Clustering based on the “largest” eigenvector
Clustering based on the 1st and the 5th “largest” eigenvectors
Some Results

Clusterings based on the 1st, 5th and 15th “largest” eigenvectors
Some Results

4 Clustering based on the 1st, 5th, 15th and 3rd "largest" eigenvectors
Some Results

Clustering based on the 1st, 5th, 15th, 3rd and 10th “largest” eigenvectors
Challenging “rings” dataset clustered by the information contained in the eigenvectors of $G$. 
Concluding Remarks

- A new graph-theoretic cost function, called the Information Cut, was introduced.
- A novel spectral clustering algorithm was developed.
- There is a close link between spectral clustering and non-parametric density estimation.
- A natural criterium for creating the data-affinity matrix can be provided.
Other recent methods

- **Grow and re-cluster procedure**
  

- **Lagrange optimization of memberships, kernel annealing**
  
Some Results

“Largest” eigenvector for the two half-circles.