Penalized and weighted K-means for clustering with noise and prior information incorporation

George C. Tseng
Department of Biostatistics
Department of Human Genetics
University of Pittsburgh
Outline

- Intro of cluster analysis
  - Model-based clustering
  - Heuristic methods
    - Hierarchical clustering
    - K-means & K-memoids
    - ......
- A motivating example (yeast cell cycle microarray data)
- Penalized weighted K-means
  - Penalty term and weights
  - Some properties
  - Estimate parameters ($k$ and $\lambda$)
- Applications
  - Simulation
  - Yeast cell cycle microarray data
  - CID fragmentation patterns in MS/MS
- Discussion
Cluster analysis:
Data $X=\{x_i, i=1, \ldots, n\}$, each object $x_i \in \mathbb{R}^p$.

Given a dissimilarity measure $d(x_i, x_j)$, assign the $n$ objects into $k$ disjoint clusters; i.e. $C=\{C_1, \ldots, C_k\}$

$$X = \bigcup_{j=1}^{k} C_j$$
Intro. of cluster analysis

Cluster analysis:

1. Estimate the number of clusters $k$.

2. Decide which clustering method to use.

3. Evaluation and re-validation of clustering results.

Long history in statistics, computer science and applied math literature.
Intro. of cluster analysis

Model-based clustering:

1. Mixture maximum likelihood (ML):

\[
L(\pi, \mu, \Sigma) = \log \left\{ \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j f(x_i; \mu_j, \Sigma_j) \right\}
\]

2. Classification maximum likelihood (CML):

\[
L(C, \mu, \Sigma) = \log \left\{ \prod_{j=1}^{k} \prod_{x_i \in C_j} f(x_i; \mu_j, \Sigma_j) \right\}
\]

\[C = \{C_1, \ldots, C_k\}, \quad X = \bigcup_{j=1}^{k} C_k\]
Intro. of cluster analysis

Model selection of model-based clustering:
Bayesian Information criterion (BIC) to determine $k$ and $\Sigma_j$

$$2 \log p(x \mid M) + \text{const.} \approx 2 \cdot l_M(x, \hat{\theta}) - m_M \log(n) \equiv BIC$$

$p(x|M)$ is the (integrated) likelihood of the data for the model $M$

$l_M(x, \theta)$ is the maximized mixture loglikelihood for the model

$m_M$ is the number of independent parameters to be estimated in the model
Intro. of cluster analysis

Hierarchical clustering:
Intro. of cluster analysis

Hierarchical clustering:

Iteratively agglomerate nearest nodes to form bottom-up tree.

Single Linkage: shortest distance between points in the two nodes.
Complete Linkage: largest distance between points in the two nodes.

Note: Clusters can be obtained by cutting the hierarchical tree.
Intro. of cluster analysis

**K-means criterion:**
Minimize the within-group sum-squared dispersion to obtain $C$:

$$W_{K\text{-}means}(C; k) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - \overline{x}^{(j)} \right\|^2$$

$\overline{x}^{(j)}$ is the center of cluster $j$.

**K-memoids criterion:**

$$W_{K\text{-}memoids}(C; k) = \sum_{j=1}^{k} \sum_{x_i \in C_j} d(x_i, x^{(j)})$$

$x^{(j)} \in X$ is the median point in cluster $j$. 
Intro. of cluster analysis

Proposition: K-means is a special case of CML under Gaussian model of identical spherical clusters.

K-means:

\[ W_{K-\text{means}}(C; k) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - \bar{x}^{(j)} \right\|^2 \]

\[ \bar{x}^{(j)} \] is the center of cluster \( j \).

CML:

\[ C_1(C, \theta) = f(x|C, \theta) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \log f(x_i|\mu_j, \Sigma_j) \]

\[ f(x_i|\mu_j, \Sigma_j) = \frac{\exp\{-\frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1}(x_i - \mu_j)\}}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \]

\[ \Sigma_j = \sigma^2 I \ (j = 1, \ldots, k) \]
Clusters contain many false positives because the algorithm has to assign all genes into clusters.

Many genes are noise (scattered) genes!!
A motivating example
Yeast cell cycle microarray data

Traditional:
- Assign all genes into clusters.

Question:
- Can we allow a set of scattered (noise) genes?
A motivating example
Yeast cell cycle microarray data
**A motivating example**

**Yeast cell cycle microarray data**

**Prior information:**

**Six groups of validated cell cycle genes:**

| M/G1 Boundary: | AGA1 | ASH1 | CDC46 | CDC47 | CDC6 | CHS1 | CLN3 | CTS1 | EGT2 | FUS1 | MFA2 | PCL2 | PCL9 | RME1 | SIC1 | SST2 | STE2 | SWI4 | TEC1 |
| Late G1, SCB regulated: | CLN1 | CLN2 | CSD2 | CHS3 | FKS1 | CWH53 | GAS1 | HO | KAR4 | KRE6 | MNN1 | PCL1 | PSA1 | SWE1 | TIP1 | VAN2 | GOG5 |
| Late G1, MCB regulated: | ASF1 | ASF2 | CDC21 | CDC45 | CDC8 | CDC9 | CLB5 | CLB6 | DBF4 | DPB2 | DPB3 | GIC2 | MCD1 | MSPH2 | NIK1 | HSL1 | PDS1 | PMS1 | POL1 | POL12 | POL2 | POL3 | CDC2 | POL30 | PRI1 | PRI2 | RAD17 | RAD27 | RAD51 | RAD54 | RFA1 | RFA2 | RFA3 | RNR1 | RNR3 | SPC110 | NUF1 | SPC42 | SPK1 | SRS2 | HPR5 | UNG1 |
| S-phase: | HHT1 | HHT2 | HHF1 | HHF2 | HTA1 | HTA2 | HTB1 | HTB2 |
| S/G2-phase: | CDC14 | CIK1 | CLB3 | CLB4 | CWP1 | CWP2 | KAR3 | NUM1 | TIR1 |
| G2/M-phase: | ACE2 | ASE1 | CDC20 | CDC5 | CLB1 | CLB2 | DBF2 | FAR1 | KIN3 | MOB1 | YRO2 | (MST1) | MRH1 | (MST2) | SED1 | SPO12 | SWI5 |
Goal 1:
- Allow a set of scattered genes without being clustered.

Goal 2:
- Incorporation of prior information in cluster formation.

A motivating example
Yeast cell cycle microarray data
PW-Kmeans

Formulation:
Assign $n$ objects into $k$ clusters and a possible noise set.
i.e. $C=\{C_1, \ldots, C_k, S\}$, \[ X = (\bigcup_{j=1}^{k} C_j) \cup S \]
Extend K-means criterion to:

$$W(C; k, \lambda) = \sum_{j=1}^{k} \sum_{x_i \in C_j} w(x_i; P) \cdot d(x_i, C_j) + \lambda |S|$$

$d(x_i, C_j)$: dispersion of point $x_i$ in $C_j$.
$|S|$: # of objects in noise set $S$.
$w(x_i; P)$: weight function to incorporate prior info $P$.
$\lambda$: a tuning parameter
PW-Kmeans

\[ W(C; k, \lambda) = \sum_{j=1}^{k} \sum_{x_i \in C_j} w(x_i; P) \cdot d(x_i, C_j) + \lambda |S| \]

How does it work?

Penalty term \( \lambda \): assign outlying objects of a cluster to the noise set \( S \).

Weighting term \( w \): utilize prior knowledge of preferred or prohibited patterns \( P \).
Proposition:
Denote $C^* (k, \lambda) = \{C_1^* (k, \lambda), ..., C_k^* (k, \lambda), S^* (k, \lambda)\}$ the minimizer given $k$ and $\lambda$.

1. If $k_1 > k_2$, $W(C^* (k_1, \lambda); k_1, \lambda) < W(C^* (k_2, \lambda); k_2, \lambda)$.

2. If $\lambda_1 > \lambda_2$, $|S^* (k, \lambda_1)| \leq |S^* (k, \lambda_2)|$.

3. If $\lambda_1 > \lambda_2$, $W(C^* (k, \lambda_1); k, \lambda_1) > W(C^* (k, \lambda_2); k, \lambda_2)$.

$$W(C; k, \lambda) = \sum_{j=1}^{k} \sum_{x_i \in C_j} w(x_i; P) \cdot d(x_i, C_j) + \lambda |S|$$
K-means and K-memoids are two special cases of the new generalized form of PW-Kmeans. (i.e. \( w(\cdot,\cdot)=1, \lambda=\infty \))

\[
W_{K\text{-means}}(C; k) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \bar{x}^{(j)}\|^2
\]

\[
W_{K\text{-memoids}}(C; k) = \sum_{j=1}^{k} \sum_{x_i \in C_j} d\left(x_i, x^{(j)}\right)
\]
PW-Kmeans

Relation to classification likelihood

**K-means loss function:**

\[ W_{\text{Kmeans}}(C; k) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \| x_i - \bar{x}(j) \|^2. \]

**Classification likelihood:** (Gaussian model)

\[ C_1(C, \theta) = f(x|C, \theta) = \sum_{j=1}^{k} \sum_{i \in C_k} \log f(x_i|\mu_j, \Sigma_j) \]

\[ f(x_i|\mu_j, \Sigma_j) = \frac{\exp\{-\frac{1}{2}(x_i-\mu_j)^T \Sigma_j^{-1}(x_i-\mu_j)\}}{(2\pi)^{p/2} |\Sigma_j|^{1/2}} \quad \Sigma_j = \sigma^2 I. \]
PW-Kmeans

Relation to classification likelihood

Penalized K-means loss function:

\[
W_p(C; k, \lambda_0) = \sum_{j=1}^{k} \sum_{i \in C_j} \|x_i - \bar{x}^{(j)}\|^2 + \left(\frac{H}{\sqrt{k}}\right)^2 \cdot \lambda_0 |S|
\]

Classification likelihood: (Gaussian model)

\[
f(x|C, \theta) = \prod_{j=1}^{k} \prod_{i \in C_j} f(x_i|\mu_j, \Sigma_j) \prod_{i \in S} \frac{1}{|V|}
\]

\[
f(x_i|\mu_j, \Sigma_j) = \frac{\exp\{-\frac{1}{2}(x_i-\mu_j)^T \Sigma_j^{-1}(x_i-\mu_j)\}}{(2\pi)^{p/2} |\Sigma_j|^{1/2}} \quad \Sigma_j = \sigma_0^2 I
\]

\[
\lambda_0 = 2\sigma_0^2 \cdot (\sqrt{k} / H)^2 \cdot \log |V|
\]

\(V\) is the space where noise set is uniformly distributed.
PW-Kmeans

Estimate $k$ and $\lambda$

Tibshirani et al. 2001
PW-Kmeans

Estimate $k$ and $\lambda$

$C(X_{tr},k) =$ clustering operation when cluster $X_{tr}$ into $k$ clusters.

$D[C(X_{tr},k),X_{tr}]$ an $n \times n$ matrix denoting "co-memberships".

$D[C(X_{tr},k),X_{tr}]_{ii'} = 1$ if $i$ and $i'$ are in the same cluster.

$X_{tr} = (A_1^T, \ldots A_k^T)^T$ be the $k$ cluster sets; $n_k = |A_k|$.

$ps(k) = \min_{1 \leq j \leq k} \frac{1}{n_j(n_j - 1)} \sum_{i \neq i' \in A_j} I(D[C(X_{tr},k),X_{te}]_{ii'} = 1)$

For each test cluster, we compute the proportion of observation pairs in that cluster that are also assigned to the same cluster by the training set centroids.
Simulation

Penalized K-means (no weight term)

\( \lambda \) is inversely related to the number of noise genes \(|S|\).
Simulation

Estimate $k$ and $\lambda$
### Applications

I: Yeast cell cycle microarray data

**Prior information**

**Six groups of validated cell cycle genes:**

<table>
<thead>
<tr>
<th>M/G1 Boundary:</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGA1 ASH1 CDC46 CDC47 CDC6 CHS1 CLN3 CTS1 EGT2 FUS1 MFA2</td>
</tr>
<tr>
<td>PCL2 PCL9 RME1 SIC1 SST2 STE2 SWI4 TEC1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Late G1, SCB regulated:</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLN1 CLN2 CSD2 CHS3 FKS1 CWH53 GAS1 HO KAR4 KRE6 MNN1</td>
</tr>
<tr>
<td>PCL1 PSA1 SWE1 TIP1 VAN2 GOG5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Late G1, MCB regulated:</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASF1 ASF2 CDC21 CDC45 CDC8 CDC9 CLB5 CLB6 DBF4 DPB2 DPB3</td>
</tr>
<tr>
<td>GIC2 MCD1 MSH2 MSH6 NIK1 HSL1 PDS1 PMS1 POL1 POL12 POL2</td>
</tr>
<tr>
<td>POL3 CDC2 POL30 PRI1 PRI2 RAD17 RAD27 RAD51 RAD54 RFA1</td>
</tr>
<tr>
<td>RFA2 RFA3 RNB1 RNB3 SPC110 NUF1 SPC42 SPK1 SBS2 HPR5 UNG1</td>
</tr>
</tbody>
</table>

**S-phase:**

| HHT1 HHT2 HHF1 HHF2 HTA1 HTA2 HTB1 HTB2 |

**S/G2-phase:**

| CDC14 CIK1 CLB3 CLB4 CWP1 CWP2 KAR3 NUM1 TIR1 |

**G2/M-phase:**

| ACE2 ASE1 CDC20 CDC5 CLB1 CLB2 DBF2 FAR1 KIN3 MOB1 YRO2 |
| (MST1) MRH1 (MST2) SED1 SPO12 SWI5 |
Applications
I: Yeast cell cycle microarray data

Prior information
Six groups of validated cell cycle genes:

- $F_{1:M/G1}$
- $F_{2:\text{late G1}}$
- SCB regulated
- $F_{3:\text{late G1}}$
- MCB regulated
- $F_{4:S}$
- $F_{5:S/G2}$
- $F_{6:G2/M}$

8 histone genes tightly coregulated in S phase
Applications
I: Yeast cell cycle microarray data

Penalized K-means

The 8 histone genes are left in noise set $S$ without being clustered.
Applications

I: Yeast cell cycle microarray data

Penalized weighted K-means

\[ W_{pw}(C; k, \lambda_0) = \sum_{j=1}^{k} \sum_{i \in C_j} w_{pw}(x_i; \mathcal{P}) \|x_i - \bar{x}^{(j)}\|^2 + \left( \frac{H}{\sqrt{k}} \right)^2 \cdot \lambda_0 |S| \]

\[ \mathcal{P} = \left( (\mathcal{P}_1^{(1)}, \ldots, \mathcal{P}_{n_1}^{(1)}), \ldots, (\mathcal{P}_1^{(p)}, \ldots, \mathcal{P}_{n_p}^{(p)}) \right) \quad \text{Prior knowledge of } p \text{ pathways} \]

The weight is designed as a transformation of logistic function:

\[ w_{pw}(x; \mathcal{P}) = \alpha + (1 - \alpha) \cdot \frac{1 - e^{-\tau h(x_i; \mathcal{P})}}{1 + e^{-\tau h(x_i; \mathcal{P})}} \]

\[ h(x_i; \mathcal{P}) = \min_{l} \left(\frac{1}{n_l}\right) \sum_{m} ||x_i - \mathcal{P}_m^{(l)}|| \]
Applications
I: Yeast cell cycle microarray data

Design of weight function

\[ w_{pw}(x; p) = \alpha + (1 - \alpha) \cdot \frac{1 - e^{-\tau h(x; p)}}{1 + e^{-\tau h(x; p)}} \]
Applications
I: Yeast cell cycle microarray data

Take three randomly selected histone genes as prior information, P. Then perform penalized weighted K-means.

The 8 histone genes are now in cluster 3.
# Applications

## I: Yeast cell cycle microarray data

### Annotation prediction from clusters

<table>
<thead>
<tr>
<th></th>
<th>F₁</th>
<th>F₂</th>
<th>F₃</th>
<th>F₄</th>
<th>F₅</th>
<th>F₆</th>
<th>F₇</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>58</td>
</tr>
<tr>
<td>C₂</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>31</td>
</tr>
<tr>
<td>C₃</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>101</td>
</tr>
<tr>
<td>C₄</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>39</td>
</tr>
<tr>
<td>C₅</td>
<td>0</td>
<td>4</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>71</td>
</tr>
<tr>
<td>S</td>
<td>16</td>
<td>8</td>
<td>5</td>
<td>8</td>
<td>7</td>
<td>5</td>
<td>1276</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(F₁): M/G1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(F₂): late G1 SCB regulated</td>
</tr>
<tr>
<td></td>
<td>(F₃): late G1 MCB regulated</td>
</tr>
<tr>
<td></td>
<td>(F₄): S-phase</td>
</tr>
<tr>
<td></td>
<td>(F₅): S/G2-phase</td>
</tr>
<tr>
<td></td>
<td>(F₆): G2/M-phase</td>
</tr>
<tr>
<td></td>
<td>(F₇): unannotated genes</td>
</tr>
</tbody>
</table>

### p-value calculation (null is hypergeometric distribution):

\[
P \left( G, D(F), n(C), d(F) \right) = 1 - \sum_{i=1}^{d(F)-1} \frac{\binom{D(F)}{i} \binom{G-D(F)}{n(C)-i}}{\binom{G}{n(C)}}
\]

\(G\): total of 1663 genes
\(D(F)\): # of genes in the functional category (23+5=28)
\(n(C)\): # of genes in the cluster (4+23+71=98)
\(d(F)\): # of genes in the cluster and the functional category (23)
Given a p-value threshold $\delta (\delta=0.01)$, we can compute:

**Predictions made**: $42+98+98=238$

**Accuracy**: $(10+4+23)/(42+98+98) = 15.55\%$

Varying $\delta$ gives varying “Predictions made” and “Accuracy”
Applications

I: Yeast cell cycle microarray data

Evaluation of annotation prediction

$\delta = 10^{-4}, \ldots 10^{-20}$

Accuracy of random guess
Applications

I: Yeast cell cycle microarray data

Conclusion: Evaluation of annotation prediction

- P-kmeans generally better than Kmeans.
- P-kmeans makes fewer predictions than Kmeans but produce much higher accuracy.
- Smaller $\lambda$ result in smaller clusters and # of prediction made but with better accuracy.
Applications
II: CID fragmentation pattern in MS/MS

Enzyme → HPLC → MS → MS/MS

Protein #1: SIYDGK, FWSEFR
Protein #2: TLLHPYK

Peptide Sequencing Algorithm

Applications II: CID fragmentation pattern in MS/MS
Collision-Induced Dissociation (CID)

The abundance of such cleavages are recorded as intensities.

Applications
II: CID fragmentation pattern in MS/MS

The abundance of such cleavages are recorded as intensities.
### Applications

**II: CID fragmentation pattern in MS/MS**

One single peptide: AAAMDAQAEAK

<table>
<thead>
<tr>
<th>A</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>P</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1st</th>
<th>2nd</th>
<th>count</th>
<th>intensities</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>2</td>
<td>0 0.381</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>D</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>E</td>
<td>1</td>
<td>0.031</td>
</tr>
<tr>
<td>A</td>
<td>F</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>H</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>I</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>K</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>L</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>M</td>
<td>1</td>
<td>0.514</td>
</tr>
<tr>
<td>A</td>
<td>N</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>P</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>Q</td>
<td>1</td>
<td>0.096</td>
</tr>
<tr>
<td>A</td>
<td>R</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>S</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>T</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>V</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>W</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>Y</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Assume intensities measure the probability of dissociation.

All intensities normalized to [0,1]
### Applications

**II: CID fragmentation pattern in MS/MS**

For a specific set of peptides (720 peptides):

<table>
<thead>
<tr>
<th>A</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>P</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1st</th>
<th>2nd</th>
<th>count</th>
<th>intensities</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>188</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>1</td>
<td>0.105</td>
</tr>
<tr>
<td>A</td>
<td>D</td>
<td>91</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>A</td>
<td>E</td>
<td>94</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>A</td>
<td>F</td>
<td>41</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0.008 0.008 0.01</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>129</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>A</td>
<td>H</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>I</td>
<td>69</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0.019</td>
</tr>
<tr>
<td>A</td>
<td>K</td>
<td>5</td>
<td>0 0.01 0.013 0.122 0.139</td>
</tr>
<tr>
<td>A</td>
<td>L</td>
<td>137</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>A</td>
<td>M</td>
<td>28</td>
<td>0.004 0.02 0.023 0.034 0.034 0.06 0.062 0.068 0.093 0.101 0.122</td>
</tr>
<tr>
<td>A</td>
<td>N</td>
<td>52</td>
<td>0 0 0 0 0 0.006 0.022 0.025 0.025 0.033 0.039 0.046</td>
</tr>
<tr>
<td>A</td>
<td>P</td>
<td>152</td>
<td>0 0 0 0 0.012 0.029 0.041 0.042 0.049 0.059 0.068</td>
</tr>
<tr>
<td>A</td>
<td>Q</td>
<td>58</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>A</td>
<td>R</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>S</td>
<td>59</td>
<td>0 0 0 0 0 0.01 0.021 0.031 0.038 0.04 0.049</td>
</tr>
<tr>
<td>A</td>
<td>T</td>
<td>78</td>
<td>0 0 0 0 0 0 0 0 0 0 0.013 0.018</td>
</tr>
<tr>
<td>A</td>
<td>V</td>
<td>88</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>A</td>
<td>W</td>
<td>11</td>
<td>0 0 0 0 0.029 0.149 0.205 0.321 0.428 0.454 1</td>
</tr>
<tr>
<td>A</td>
<td>Y</td>
<td>26</td>
<td>0 0 0 0 0.008 0.011 0.047 0.063 0.065 0.078 0.079</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

$20 \times 20 = 400$ independent distributions

Each with 0 or multiple (up to hundreds) observations
Applications

II: CID fragmentation pattern in MS/MS

Protein #1:
- SIYDGK
- FWSEFR
- TLLHPYK

Protein #2:

Peptide Sequence

Peptide Sequencing Algorithm

Enzyme → HPLC → MS → Abundance

m/z

(CID)
Current protein identification algorithms assume completely random dissociation probability pattern.

This is, however, found not true and the dissociation pattern depends on the charge state and the peptide sequence motif.
### Applications

#### II: CID fragmentation pattern in MS/MS

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>188</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>1</td>
<td>0.105</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>D</td>
<td>91</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>E</td>
<td>94</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>F</td>
<td>41</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.008</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>129</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>H</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>I</td>
<td>69</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>K</td>
<td>5</td>
<td>0</td>
<td>0.01</td>
<td>0.013</td>
<td>0.122</td>
<td>0.139</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>L</td>
<td>137</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>M</td>
<td>28</td>
<td>0.004</td>
<td>0.02</td>
<td>0.023</td>
<td>0.034</td>
<td>0.034</td>
<td>0.06</td>
<td>0.062</td>
<td>0.068</td>
<td>0.093</td>
</tr>
<tr>
<td>A</td>
<td>N</td>
<td>52</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
<td>0.022</td>
<td>0.025</td>
<td>0.025</td>
</tr>
<tr>
<td>A</td>
<td>P</td>
<td>152</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.012</td>
<td>0.029</td>
<td>0.041</td>
<td>0.042</td>
</tr>
<tr>
<td>A</td>
<td>Q</td>
<td>58</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>R</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>A</td>
<td>73</td>
<td>0.018</td>
<td>0.026</td>
<td>0.032</td>
<td>0.046</td>
<td>0.046</td>
<td>0.093</td>
<td>0.108</td>
<td>0.146</td>
<td>0.155</td>
</tr>
<tr>
<td>D</td>
<td>C</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>D</td>
<td>38</td>
<td>0</td>
<td>0.012</td>
<td>0.056</td>
<td>0.097</td>
<td>0.097</td>
<td>0.118</td>
<td>0.135</td>
<td>0.142</td>
<td>0.156</td>
</tr>
<tr>
<td>D</td>
<td>E</td>
<td>53</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.021</td>
<td>0.026</td>
<td>0.035</td>
<td>0.05</td>
<td>0.073</td>
<td>0.08</td>
</tr>
<tr>
<td>D</td>
<td>F</td>
<td>38</td>
<td>0</td>
<td>0.003</td>
<td>0.044</td>
<td>0.1</td>
<td>0.119</td>
<td>0.214</td>
<td>0.232</td>
<td>0.283</td>
<td>0.41</td>
</tr>
<tr>
<td>D</td>
<td>G</td>
<td>44</td>
<td>0.024</td>
<td>0.128</td>
<td>0.128</td>
<td>0.226</td>
<td>0.239</td>
<td>0.247</td>
<td>0.383</td>
<td>0.395</td>
<td>0.491</td>
</tr>
<tr>
<td>D</td>
<td>H</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>I</td>
<td>38</td>
<td>0.029</td>
<td>0.054</td>
<td>0.063</td>
<td>0.128</td>
<td>0.15</td>
<td>0.173</td>
<td>0.229</td>
<td>0.233</td>
<td>0.257</td>
</tr>
<tr>
<td>D</td>
<td>K</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>L</td>
<td>76</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.057</td>
<td>0.064</td>
<td>0.113</td>
</tr>
<tr>
<td>D</td>
<td>M</td>
<td>15</td>
<td>0</td>
<td>0.147</td>
<td>0.212</td>
<td>0.376</td>
<td>0.419</td>
<td>0.709</td>
<td>0.806</td>
<td>0.841</td>
<td>0.885</td>
</tr>
<tr>
<td>D</td>
<td>N</td>
<td>18</td>
<td>0.047</td>
<td>0.108</td>
<td>0.232</td>
<td>0.442</td>
<td>0.458</td>
<td>0.506</td>
<td>0.508</td>
<td>0.575</td>
<td>0.585</td>
</tr>
<tr>
<td>D</td>
<td>P</td>
<td>63</td>
<td>0.122</td>
<td>0.444</td>
<td>0.481</td>
<td>0.61</td>
<td>0.631</td>
<td>0.675</td>
<td>0.753</td>
<td>0.882</td>
<td>0.883</td>
</tr>
<tr>
<td>D</td>
<td>Q</td>
<td>14</td>
<td>0</td>
<td>0.098</td>
<td>0.163</td>
<td>0.228</td>
<td>0.262</td>
<td>0.297</td>
<td>0.421</td>
<td>0.459</td>
<td>0.463</td>
</tr>
<tr>
<td>D</td>
<td>R</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**AX: low**

**DX: high**
Applications
II: CID fragmentation pattern in MS/MS

Visualization of a distribution:
Ten concentric donuts to represent 5%, 15%,…, 95% percentiles. Value represented by gradient color.
Applications
II: CID fragmentation pattern in MS/MS

The dissociation pattern depends on the charge state and the peptide sequence motif.
Distances cannot be defined for most pairs of peptides. (more than 95% missing values)

Distance between a peptide and a set of peptides can be defined.

K-means and PW-Kmeans are applicable while most other clustering methods fail.

\[
W(C; k, \lambda) = \sum_{j=1}^{k} \sum_{x_i \in C_j} w(x_i; P) \cdot d(x_i, C_j) + \lambda |S|
\]
Applications
II: CID fragmentation pattern in MS/MS

[...P...R]^+

Original Data
1+, .....P...R

Kmeans
cluster 1

P-Kmeans
cluster 1

674 peptides

1184 peptides

720 peptides

[...P...R]^2+

2+, .....P...R

cluster 2

cluster 2

2182 peptides

1671 peptides

1775 peptides
• Intensity data of each peptide contain >95% missing values. Most clustering methods would not work.
• Dissimilarity between two peptides cannot be defined.
• Fortunately dissimilarity between one peptide and a set of peptides can be calculated and penalized K-means can be used.
Discussion

heuristic

Hierarchical clustering
CLICK
SOM

Model-based

K-memoids
K-means
PW-Kmeans

Gaussian mixture model
Bayesian clustering

Tight clustering
(re-evaluation machine by re-sampling techniques)
Conclusion

- Penalized and weighted K-means (PW-Kmeans) provides a general and flexible way for clustering complex data.

- The penalty term allows a noise set not being clustered, avoiding information dilution by noises.

- The weights can be designed to incorporate biological prior pathway information. Similar to Bayesian approach but avoids specific modelling.

- In the situation of many missing values (MS/MS example), most methods are hard to pursue but P-Kmeans worked well.
Acknowledgement

- MS/MS data collaboration with Yingying Huang from Vicki Wysocki’s lab in University of Arizona

- Discussion and comments from Haiyan Huang, Eleanor Feingold and Wing Wong.