# Information Theoretic Co-clustering

Inderjit S. Dhillon Subramanyam Mallela Dharmendera S. Modha
Department of Computer Sciences
University of Texas, Austin

---

Presenter: Omur Ozel

# Introduction and Motivation

- Co-occurrence of two events, contingency tables
- Clustering correlated data sets jointly.
- E.g. word-document clustering
- Common issue: Sparsity and high-dimensionality.
- An information-theoretic approach and algorithm

# Information-Theoretic Approach

- Consider two-dimensional data.
- Row variables and column variables.
- View the contingency table as two-dimensional PMF
- Objective: minimize information loss after clustering
- Measure of information comes from Information Theory

# Remarks on Information Theory

- $I(X;Y)$: mutual information between $X$ and $Y$
- A measure of how $X$ and $Y$ are inter-related.

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right)$$

- We can also express $I(X;Y)$ as

$$I(X;Y) = D(p(X,Y)||p(X)p(Y))$$

- $D(.||.)$: Kullback-Leibler divergence, relative entropy

$$D(p_1(X)||p_2(X)) = \sum_{x \in \mathcal{X}} p_1(x) \log \left( \frac{p_1(x)}{p_2(x)} \right)$$

- $D(.||.)$ is nonnegative and zero iff $p_1(X) \equiv p_2(X)$.

# Remarks on Information Theory

- $I(X;Y)$ is the channel capacity for $X \to Y$.
- Another way of expressing $I(X;Y)$ is

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \end{aligned}$$

- $H(X)$ is the entropy of $X$:

$$H(X) = -\sum_x p(x)\log(p(x))$$

- Let $X_1, \ldots, X_n$ be an i.i.d. sequence with joint PMF $\Pi_{i=1}^n p(x_i)$
- $H(X)$ is the minimum average number of bits necessary to code $X_1, \ldots, X_n$

## Problem Formulation

- $X \in \{x_1, \ldots, x_m\}, \quad Y \in \{y_1, \ldots, y_n\}$
- $p(X, Y)$,
  - joint PMF
  - two-dimensional contingency table
  - two-way frequency table
- Simultaneous clustering of $X$ and $Y$
- $k$: number of $X$ clusters, $\ell$: number of $Y$ clusters

$$C_X : \{x_1, \ldots, x_m\} \rightarrow \{\hat{x}_1, \ldots, \hat{x}_k\}$$
$$C_Y : \{y_1, \ldots, y_n\} \rightarrow \{\hat{y}_1, \ldots, \hat{y}_\ell\}$$

- $(C_X, C_Y)$: co-clustering
- Each division created in $p(X, Y)$ is called *co-cluster*.

# Example Co-clustering

- Consider the following contingency matrix:

$$p(X, Y) = \begin{pmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{pmatrix}$$

- $\hat{x}_1 = \{x_1, x_2\}$, $\hat{x}_2 = \{x_3, x_4\}$ and $\hat{x}_3 = \{x_5, x_6\}$
- $\hat{y}_1 = \{y_1, y_2, y_3\}$ and $\hat{y}_2 = \{y_4, y_5, y_6\}$.
- Resulting distributions

$$p(\widehat{X}, \widehat{Y}) = \begin{pmatrix} 0.3 & 0 \\ 0 & 0.3 \\ 0.2 & 0.2 \end{pmatrix}$$

# Information Theoretic Co-clustering

- $X, Y$: original random variables
- $\widehat{X}, \widehat{Y}$: clustered random variables
- They are related as $\widehat{X} = C_X(X)$ and $\widehat{Y} = C_Y(Y)$
- Minimize the following objective function:

$$I(X;Y) - I(\widehat{X};\widehat{Y})$$

- That is, we minimize information loss due to co-clustering.

## Example Co-clustering

- Consider the following contingency matrix:

$$p(X, Y) = \begin{pmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{pmatrix}$$

- $\hat{x}_1 = \{x_1, x_2\}$, $\hat{x}_2 = \{x_3, x_4\}$ and $\hat{x}_3 = \{x_5, x_6\}$
- $\hat{y}_1 = \{y_1, y_2, y_3\}$ and $\hat{y}_2 = \{y_4, y_5, y_6\}$.
- $I(X; Y) - I(\widehat{X}; \widehat{Y}) = 0.0957$ bits

# Information Theoretic Co-clustering

- Co-clustering causes a nonnegative information loss:

$$I(X;Y) - I(\hat{X};\hat{Y}) \geq 0$$

- Specifically, for fixed $(C_X, C_Y)$

$$I(X;Y) - I(\hat{X};\hat{Y}) = D(p(X,Y)||q(X,Y))$$

- Here, $q(X,Y)$ is of the form

$$q(x,y) = p(\hat{x},\hat{y})p(x|\hat{x})p(y|\hat{y})$$

- The structure of $q$ is desirable in approximating $p$

$$q(\hat{x},\hat{y}) = p(\hat{x},\hat{y}), \quad q(x,\hat{x}) = p(x,\hat{x}), \quad q(y,\hat{y}) = p(y,\hat{y})$$

- Use this structure to derive an algorithm.

# Example Co-clustering

- The resulting matrix is:

$$q(X,Y) = \begin{pmatrix} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ .036 & .036 & .028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{pmatrix}$$

- $\hat{x}_1 = \{x_1, x_2\}$, $\hat{x}_2 = \{x_3, x_4\}$ and $\hat{x}_3 = \{x_5, x_6\}$
- $\hat{y}_1 = \{y_1, y_2, y_3\}$ and $\hat{y}_2 = \{y_4, y_5, y_6\}$.

## An Interpretation from Data Compression and Transmission

- Transmit $X$ and $Y$ from a source to a destination.
- First compute $\widehat{X} = C_X(X)$ and $\widehat{Y} = C_Y(Y)$ jointly.
- Separately transmit
  - $X$ given destination already knows $\widehat{X}$
  - $Y$ given destination already knows $\widehat{Y}$
- First operation requires $H(\widehat{X}, \widehat{Y})$ bits.
- Second operation requires $H(X|\widehat{X}) + H(\widehat{X}|X)$ bits.
- On the other hand, we have

$$H(\widehat{X}, \widehat{Y}) + H(X|\widehat{X}) + H(\widehat{X}|X) \\ = D(p(X, Y)||q(X, Y))$$

## Algorithm *Co_Clustering*

- Define four-dimensional joint PMFs

$$p(x, y, \hat{x}, \hat{y}) = p(\hat{x}, \hat{y}) p(x, y|\hat{x}, \hat{y})$$
$$q(x, y, \hat{x}, \hat{y}) = p(\hat{x}, \hat{y}) p(x|\hat{x}) p(y|\hat{y})$$

- For every fixed hard co-clustering,

$$D(p(X, Y)||q(X, Y)) = D(p(X, Y, \widehat{X}, \widehat{Y})||q(X, Y, \widehat{X}, \widehat{Y}))$$

# Algorithm *Co_Clustering*

- Express the objective function in terms of solely row-clustering or solely column clustering:

$$
\begin{aligned}
&D(p(X, Y, \widehat{X}, \widehat{Y}) || q(X, Y, \widehat{X}, \widehat{Y})) \\
&= \sum_{\hat{x}} \sum_{x : C_X(x) = \hat{x}} p(x) D(p(Y|x) || q(Y|\hat{x})) \\
&= \sum_{\hat{y}} \sum_{y : C_Y(y) = \hat{y}} p(y) D(p(X|y) || q(X|\hat{y}))
\end{aligned}
$$

# Algorithm *Co_Clustering*

- Input:
  - joint PMF $p(X, Y)$,
  - $k$ number of desired row clusters,
  - $\ell$ number of column clusters
- Output: the final partition functions $C_X^f$ and $C_Y^f$
- An iterative algorithm.

## Algorithm *Co_Clustering*: Initialization

- Start with initial partition functions $C_X^{(0)}$ and $C_Y^{(0)}$
- Compute $q^{(0)}(\widehat{X}, \widehat{Y})$, $q^{(0)}(X|\widehat{X})$, $q^{(0)}(Y|\widehat{Y})$
- Compute $q^{(0)}(Y|\hat{x})$, $1 \leq \hat{x} \leq k$ by means of

$$q^{(0)}(y|\hat{x}) = q^{(0)}(y|\hat{y})q^{(0)}(\hat{y}|\hat{x})$$

# Algorithm *Co_Clustering*: Step 2

- Compute row clusters $C_X^{(t+1)}(x)$ for all $1 \leq x \leq m$ as

$$C_X^{(t+1)}(x) = \arg\min_{\hat{x}} D(p(Y|x)||q^{(t)}(Y|\hat{x}))$$

- Let $C_Y^{(t+1)} = C_Y^{(t)}$

# Algorithm *Co_Clustering*: Step 3

- Compute the distributions $q^{(t+1)}(\widehat{X}, \widehat{Y})$, $q^{(t+1)}(X|\widehat{X})$, $q^{(t+1)}(Y|\widehat{Y})$
- Compute $q^{(t+1)}(X|\hat{y})$, $1 \leq \hat{y} \leq \ell$ by means of

$$q^{(t+1)}(x|\hat{y}) = q^{(t+1)}(x|\hat{x})q^{(t+1)}(\hat{x}|\hat{y})$$

# Algorithm *Co_Clustering*: Step 4

- Compute column clusters $C_Y^{(t+2)}(y)$ for all $1 \leq y \leq n$ as

$$C_Y^{(t+2)}(y) = \arg\min_{\hat{y}} D(p(X|y)||q^{(t)}(X|\hat{y}))$$

- Let $C_X^{(t+2)} = C_X^{(t+1)}$

## Algorithm *Co_Clustering*: Step 5

- Compute the distributions $q^{(t+2)}(\widehat{X}, \widehat{Y})$, $q^{(t+2)}(X|\widehat{X})$, $q^{(t+2)}(Y|\widehat{Y})$
- Compute $q^{(t+2)}(Y|\hat{x})$, $1 \leq \hat{x} \leq k$ by means of

$$q^{(t+2)}(y|\hat{x}) = q^{(t+2)}(y|\hat{y})q^{(t+2)}(\hat{y}|\hat{x})$$

## Algorithm *Co_Clustering*: Step 6

- Compute the change in the objective

$$D(p(X,Y)||q^{(t)}(X,Y)) - D(p(X,Y)||q^{(t+2)}(X,Y))$$

- If it is small enough, say less than $\epsilon = 10^{-4}$, stop
- Otherwise, go to step 2 with $t = t + 2$.

# Algorithm *Co_Clustering*

**Theorem**

*The mutual information loss is monotonically decreasing in each step of the Co_Clustering algorithm.*

$$D(p^{(t)}(X,Y,\widehat{X},\widehat{Y})||q^{(t)}(X,Y,\widehat{X},\widehat{Y})) \geq D(p^{(t+1)}(X,Y,\widehat{X},\widehat{Y})||q^{(t+1)}(X,Y,\widehat{X},\widehat{Y}))$$

**Corollary**

*The Co_Clustering algorithm terminates in a finite number of steps at a cluster assignment that is locally optimal.*

## Numerical Study

| | $q^{(t)}(X,Y)$ | | | | | | $p^{(t)}(\hat{X},\hat{Y})$ | |
|---|---|---|---|---|---|---|---|---|
| | $\hat{y}_1$ | $\hat{y}_1$ | $\hat{y}_2$ | $\hat{y}_1$ | $\hat{y}_2$ | $\hat{y}_2$ | | |
| $\hat{x}_3$ | .029 | .029 | .019 | .022 | .024 | .024 | 0.10 | 0.05 |
| $\hat{x}_1$ | .036 | .036 | .014 | .028 | .018 | .018 | 0.10 | 0.20 |
| $\hat{x}_2$ | .018 | .018 | .028 | .014 | .036 | .036 | 0.30 | 0.25 |
| $\hat{x}_2$ | .018 | .018 | .028 | .014 | .036 | .036 | | |
| $\hat{x}_3$ | .039 | .039 | .025 | .030 | .032 | .032 | | |
| $\hat{x}_3$ | .039 | .039 | .025 | .030 | .032 | .032 | | |

$\downarrow$ steps 2 & 3 of Figure 1

| | $\hat{y}_1$ | $\hat{y}_1$ | $\hat{y}_2$ | $\hat{y}_1$ | $\hat{y}_2$ | $\hat{y}_2$ | | |
|---|---|---|---|---|---|---|---|---|
| $\hat{x}_1$ | .036 | .036 | .014 | .028 | .018 | .018 | 0.20 | 0.10 |
| $\hat{x}_1$ | .036 | .036 | .014 | .028 | .018 | .018 | 0.18 | 0.32 |
| $\hat{x}_2$ | .019 | .019 | .026 | .015 | .034 | .034 | 0.12 | 0.08 |
| $\hat{x}_2$ | .019 | .019 | .026 | .015 | .034 | .034 | | |
| $\hat{x}_3$ | .043 | .043 | .022 | .033 | .028 | .028 | | |
| $\hat{x}_2$ | .025 | .025 | .035 | .020 | .046 | .046 | | |

## Numerical Study

| | $\hat{y}_1$ | $\hat{y}_1$ | $\hat{y}_1$ | $\hat{y}_2$ | $\hat{y}_2$ | $\hat{y}_2$ | | |
|---|---|---|---|---|---|---|---|---|
| $\hat{x}_1$ | .054 | .054 | .042 | 0 | 0 | 0 | 0.30 | 0 |
| $\hat{x}_1$ | .054 | .054 | .042 | 0 | 0 | 0 | 0.12 | 0.38 |
| $\hat{x}_2$ | .013 | .013 | .010 | .031 | .041 | .041 | 0.08 | 0.12 |
| $\hat{x}_2$ | .013 | .013 | .010 | .031 | .041 | .041 | | |
| $\hat{x}_3$ | .028 | .028 | .022 | .033 | .043 | .043 | | |
| $\hat{x}_2$ | .017 | .017 | .013 | .042 | .054 | .054 | | |

$\downarrow$ steps 2 & 3 of Figure 1

| | $\hat{y}_1$ | $\hat{y}_1$ | $\hat{y}_1$ | $\hat{y}_2$ | $\hat{y}_2$ | $\hat{y}_2$ | | |
|---|---|---|---|---|---|---|---|---|
| $\hat{x}_1$ | .054 | .054 | .042 | 0 | 0 | 0 | 0.30 | 0 |
| $\hat{x}_1$ | .054 | .054 | .042 | 0 | 0 | 0 | 0 | 0.30 |
| $\hat{x}_2$ | 0 | 0 | 0 | .042 | .054 | .054 | 0.20 | 0.20 |
| $\hat{x}_2$ | 0 | 0 | 0 | .042 | .054 | .054 | | |
| $\hat{x}_3$ | .036 | .036 | .028 | .028 | .036 | .036 | | |
| $\hat{x}_3$ | .036 | .036 | .028 | .028 | .036 | .036 | | |

# Numerical Study

| Co-clustering | | | 1D-clustering | | |
|---|---|---|---|---|---|
| **992** | 4 | 8 | **944** | 9 | 98 |
| 40 | **1452** | 7 | 71 | **1431** | 5 |
| 1 | 4 | **1387** | 18 | 20 | **1297** |

Figure: Confusion matrix for co-clustering and 1D-clustering. Co-clustering accurately recovers original clusters in the *CLASSIC3* data set.

# Numerical Study

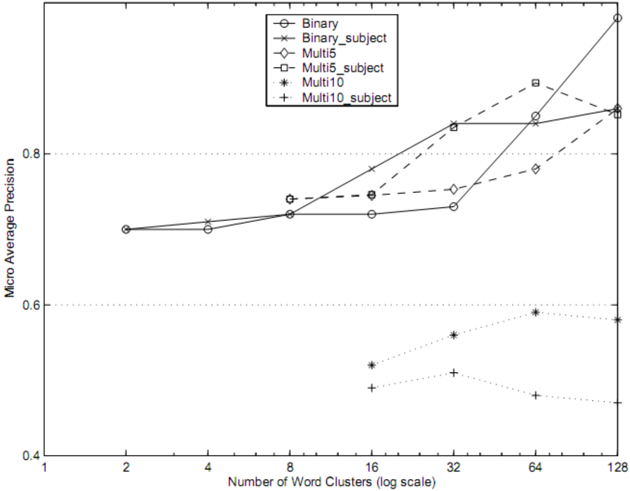


Figure: Micro-averaged precision values with varied number of word clusters with co-clustering on different NG20 data sets.

# Numerical Study

▶

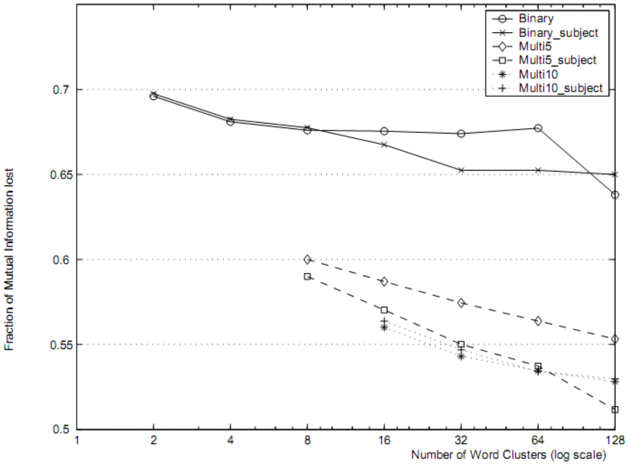


Figure: Fraction of mutual-information lost with varied number of word clusters using co-clustering on different NG20 data sets.

# Numerical Study

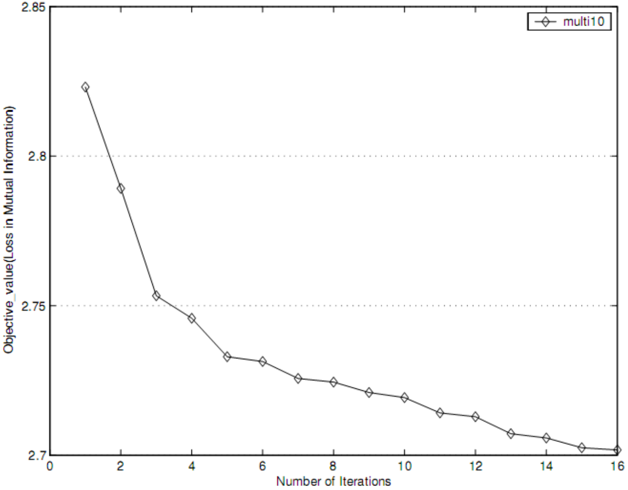


Figure: Loss in mutual information decreases monotonically with the number of iterations on a typical co-clustering run on the Multi10 data set.

# Numerical Results

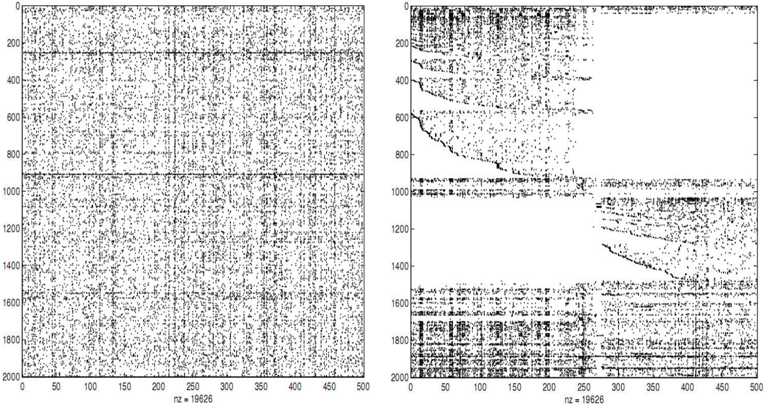


Figure: Sparsity structure of the *Binary_subject* word-document co-occurrence matrix

| $\hat{x}_{13}$ | $\hat{x}_{14}$ | $\hat{x}_{16}$ | $\hat{x}_{23}$ | $\hat{x}_{24}$ | $\hat{x}_{47}$ |
|---|---|---|---|---|---|
| dod | pitching | graphics | space | israel | army |
| ride | season | image | nasa | arab | working |
| rear | players | mac | shuttle | jewish | running |
| riders | scored | ftp | flight | occupied | museum |
| harleys | cubs | color | algorithm | rights | drive |
| camping | fans | cd | orbital | palestinian | visit |
| carbs | teams | package | satellite | holocaust | post |
| bikers | yankees | display | budget | syria | cpu |
| tharp | braves | data | srb | civil | plain |
| davet | starters | format | prototype | racist | mass |

**Table 6: Word Clusters obtained using co-clustering on the Multi5_subject data set. The clusters $\hat{x}_{13}$, $\hat{x}_{14}$, $\hat{x}_{16}$, $\hat{x}_{23}$ and $\hat{x}_{24}$ represent *rec.motorcycles*, *rec.sport.baseball*, *comp.graphics*, *sci.space* and *talk.politics.mideast* newsgroups respectively. For each cluster only top 10 words sorted by mutual information are shown.**

## Conclusions

- An information-theoretic formulation for co-clustering.
- A principled approach and algorithm.
- Converges in finite number of steps.
- Application: word-document co-clustering.
- An extensive numerical study

# References


(1) I. S. Dhillon, S. Mallela and D. S. Modha, **Information-Theoretic Co-clustering**, ACM SIGKDD Conference, August 2003

(2) N. Slonim, N. Friedman and N. Tishby, Agglomerative multivariate information bottleneck, NIPS Conference, 2001

(3) N. Tishby, F.C. Pareira and W. Bialek, The information bottleneck method, ACM SIGIR Conference, 2000

(4) N. Friedman, O. Mosenzon, N. Slonim and N. Tishby, Multivariate information bottleneck method, UAI Conference, 2001

(5) R. El-Yaniv and O. Souroujon, Iterative double clustering for unsupervised and semi-supervised learning, EMCL Conference, January 2001

(6) A. K. Jain and R. C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ, 1988.