
Co-clustering with Bregman Loss Functions

Srujana Merugu

Joint work with

Arindam Banerjee, Inderjit Dhillon, Joydeep Ghosh, Dharmendra Modha (IBM)

Agenda

- Introduction
- Bregman Co-clustering Problem
 - Loss functions based on Bregman Divergences
 - Matrix Approximation Schemes
 - Minimum Bregman Information Principle
- Co-clustering Algorithm
- Experimental Results

Co-clustering

Simultaneous clustering of rows and columns

U, V	1	2	3	4	5	6
1	-66	54	-63	93	51	96
2	35	87	37	-26	84	-22
3	-68	56	-64	92	52	94
4	30	83	32	-24	80	-21
5	-63	55	-60	92	53	95

Original Matrix Z

U, V	1	3	5	2	4	6
4	30	32	80	83	-24	-21
2	35	37	84	87	-26	-22
5	-63	-60	53	55	92	95
1	-66	-63	51	54	93	96
3	-68	-64	52	56	92	94

Reordered Matrix \tilde{Z}

U : row random variable, V : column random variable

(ρ, γ) : Co-clustering

u	1	2	3	4	5
$\rho(u)$	2	1	2	1	2

Row clustering $\rho, k = 2$

v	1	2	3	4	5	6
$\gamma(v)$	1	2	1	3	2	3

Column clustering $\gamma, \ell = 3$

Data Matrices

Natural representation in many domains

- Gene microarray analysis (genes and conditions)
- Recommender systems (users and items)
- Text analysis (words and documents)
- Market basket analysis (customers and products)
- Any finite dimensional vector data (objects and features)

Applications

- Low parameter matrix approximation
 - e.g., Data compression
- Missing value prediction
 - e.g., Recommender systems
- Clustering sparse, high dimensional data
 - e.g., Document clustering

Low Parameter Matrix Approximation

Singular Value Decomposition

- Approximation $\hat{Z} = L\Sigma R^T$ L, R -singular vector matrices
- Accurate and efficient compression
- Not always meaningful, e.g., -ve values in word-document co-occurrence matrices and joint probability matrices

U, V	1	2	3	4	5	6
1	.05	.05	.05	0	0	0
2	.05	.05	.05	0	0	0
3	0	0	0	.05	.05	.05
4	0	0	0	.05	.05	.05
5	.04	.04	0	.04	.04	.04
6	.04	.04	.04	0	.04	.04

Original Matrix Z

U, V	1	2	3	4	5	6
1	.051	.051	.047	-.007	.003	.003
2	.051	.051	.047	-.007	.003	.003
3	.003	.003	-.007	.047	.051	.051
4	.003	.003	-.007	.047	.051	.051
5	.030	.030	.020	.035	.044	.044
6	.044	.044	.035	.020	.030	.030

SVD Approximation

Low Parameter Matrix Approximation

Clustering

- Approximation $\hat{Z} = LR^T$, L -cluster centroids, R -membership matrix
- Meaningful approximations, i.e., in the original data domain
- Compression only on one of the dimensions-not very efficient

U, V	1	2	3	4	5	6
1	.05	.05	.05	0	0	0
2	.05	.05	.05	0	0	0
3	0	0	0	.05	.05	.05
4	0	0	0	.05	.05	.05
5	.04	.04	0	.04	.04	.04
6	.04	.04	.04	0	.04	.04

Original Matrix Z

U, V	1	2	3	4	5	6
1	.05	.05	.05	0	0	0
2	.05	.05	.05	0	0	0
3	0	0	0	.05	.05	.05
4	0	0	0	.05	.05	.05
5	.027	.027	.027	.04	.04	.04
6	.04	.04	.04	.027	.027	.027

Clustering Approximation

Low Parameter Matrix Approximation

Co-clustering provides efficient compression and meaningful approximations

U, V	1	2	3	4	5	6
1	.05	.05	.05	0	0	0
2	.05	.05	.05	0	0	0
3	0	0	0	.05	.05	.05
4	0	0	0	.05	.05	.05
5	.04	.04	0	.04	.04	.04
6	.04	.04	.04	0	.04	.04

Original Matrix Z

U, V	1	3	5	2	4	6
1	.05	.05	.05	0	0	0
2	.05	.05	.05	0	0	0
3	0	0	0	.05	.05	.05
4	0	0	0	.05	.05	.05
5	.033	.033	.033	.033	.033	.033
6	.033	.033	.033	.033	.033	.033

Co-clustering Approximation

U, \hat{U}	1	2	3
1	1	0	0
2	1	0	0
3	0	1	0
4	0	1	0
5	0	0	1
6	0	0	1

Row Clustering

×

\hat{U}, \hat{V}	1	2
1	.05	0
2	0	.05
3	.033	.033

Low Parameter Matrix

×

\hat{V}, V	1	2	3	4	5	6
1	1	1	1	0	0	0
2	0	0	0	1	1	1

Column Clustering

Missing Value Prediction

- Traditional Approach
 - Extrapolate from the column values of k nearest neighbors(rows)
- Co-clustering
 - Uses the matrix approximation for predicting missing values
 - Implicitly exploits correlations between data values

U, V	1	2	3	4	5	6
1	-66	54	-63	93	51	96
2	35	87	37	-26	84	-22
3	-68	56	-64	92	52	94
4	30	83	?	-24	80	-21
5	?	55	-60	92	53	95

Original Matrix Z

U, V	1	3	5	2	4	6
4	30	?	80	83	-24	-21
2	35	37	84	87	-26	-22
5	?	-60	53	55	92	95
1	-66	-63	51	54	93	96
3	-68	-64	52	56	92	94

Reordered Matrix \tilde{Z}

Missing Value Prediction

- Traditional Approach
 - Extrapolate from the column values of k nearest neighbors(rows)
- Co-clustering
 - Uses the matrix approximation for predicting missing values
 - Implicitly exploits correlations between data values

U, V	1	2	3	4	5	6
1	-66	54	-63	93	51	96
2	35	87	37	-26	84	-22
3	-68	56	-64	92	52	94
4	30	83	?	-24	80	-21
5	?	55	-60	92	53	95

Original Matrix Z

U, V	1	3	5	2	4	6
4	30	34	80	83	-24	-21
2	35	37	84	87	-26	-22
5	-64.2	-60	53	55	92	95
1	-66	-63	51	54	93	96
3	-68	-64	52	56	92	94

Reordered Matrix \tilde{Z}

Missing Value Prediction

- Traditional Approach
 - Extrapolate from the column values of k nearest neighbors(rows)
- Co-clustering
 - Use the matrix approximation for predicting missing values
 - Implicitly exploits correlations between data values

U, V	1	2	3	4	5	6
1	-66	54	-63	93	51	96
2	35	87	37	-26	84	-22
3	-68	56	-64	92	52	94
4	30	83	34	-24	80	-21
5	-64.2	55	-60	92	53	95

Original Matrix Z

U, V	1	3	5	2	4	6
4	30	34	80	83	-24	-21
2	35	37	84	87	-26	-22
5	-64.2	-60	53	55	92	95
1	-66	-63	51	54	93	96
3	-68	-64	52	56	92	94

Reordered Matrix \tilde{Z}

Clustering Sparse High Dimensional Data

- Traditional Approach
 - Dimensionality reduction followed by clustering
- Co-clustering
 - Simultaneously clusters dimensions as well as objects
 - Loss specific dimensionality reduction

Bregman Co-clustering

Goodness of Co-clustering

- Quality (e.g., homogeneity) of groups obtained through co-clustering

U, V	1	2	3	4	5	6
1	-66	54	-63	93	51	96
2	35	87	37	-26	84	-22
3	-68	56	-64	92	52	94
4	30	83	32	-24	80	-21
5	-63	55	-60	92	53	95

Original Matrix Z

U, V	1	3	5	2	4	6
4	30	32	80	83	-24	-21
2	35	37	84	87	-26	-22
5	-63	-60	53	55	92	95
1	-66	-63	51	54	93	96
3	-68	-64	52	56	92	94

Reordered Matrix \tilde{Z}

or, equivalently,

- Accuracy of a low parameter, co-clustering based approximation \hat{Z}

U, \hat{U}	1	2
1	0	1
2	1	0
3	0	1
4	1	0
5	0	1

Row Clustering

\hat{U}, \hat{V}	1	2	3
1	33.5	83.5	-23.3
2	-64.0	53.5	93.7

Low Parameter Matrix

\hat{V}, V	1	2	3	4	5	6
1	1	0	1	0	0	0
2	0	1	0	0	1	0
3	0	0	0	1	0	1

Column Clustering

Goodness of Co-clustering

- Quality (e.g., homogeneity) of groups obtained through co-clustering

U, V	1	2	3	4	5	6
1	-66	54	-63	93	51	96
2	35	87	37	-26	84	-22
3	-68	56	-64	92	52	94
4	30	83	32	-24	80	-21
5	-63	55	-60	92	53	95

Original Matrix Z

U, V	1	3	5	2	4	6
4	30	32	80	83	-24	-21
2	35	37	84	87	-26	-22
5	-63	-60	53	55	92	95
1	-66	-63	51	54	93	96
3	-68	-64	52	56	92	94

Reordered Matrix \tilde{Z}

or, equivalently,

- Accuracy of a low parameter, co-clustering based approximation \hat{Z}

U, V	1	2	3	4	5	6
1	-64.0	53.5	-64.0	93.7	53.5	93.7
2	33.5	83.5	33.5	-23.3	83.5	-23.3
3	-64.0	53.5	-64.0	93.7	53.5	93.7
4	33.5	83.5	33.5	-23.3	83.5	-23.3
5	-64.0	53.5	-64.0	93.7	53.5	93.7

Problem Definition

- **Given:** Matrix $Z_{m \times n}$, #row clusters k , #column clusters ℓ
- **Required:** Optimal co-clustering (ρ^*, γ^*) such that

$$(\rho^*, \gamma^*) = \underset{(\rho, \gamma)}{\operatorname{argmin}} E_\nu [d(Z, \hat{Z})] = \underset{(\rho, \gamma)}{\operatorname{argmin}} \sum_{u=1}^m \sum_{v=1}^n \nu_{uv} d(z_{uv}, \hat{z}_{uv})$$

where

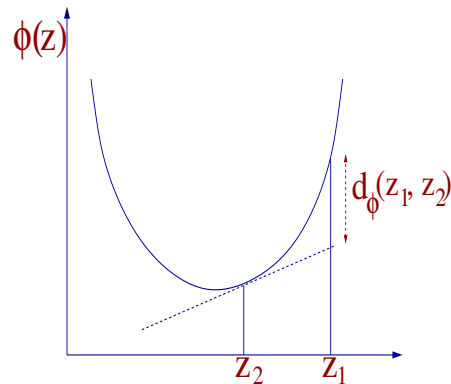
- d – any given Bregman divergence
- \hat{Z} – matrix reconstructed from (ρ, γ) and any (of certain type) given low parameter representation of Z
- ν – any given finite measure (weights) on the matrix elements

Why Bregman divergences?

- Include a number of popular distance measures, *e.g., squared Euclidean distance, KL-divergence, I-divergence, Itakura Saito distance, Mahalanobis distance, etc.*
- Nice convexity properties
- Bijection with exponential families [Banerjee et al. '04]

What are Bregman divergences ?

- Let ϕ be a strictly convex, real-valued, differentiable function



$$\underbrace{d_\phi(z_1, z_2)}_{\text{Bregman divergence}} = \phi(z_1) - \phi(z_2) - \langle z_1 - z_2, \nabla \phi(z_2) \rangle$$

- Example 1:**

- $\phi(z) = z^2$ for $z \in \mathbb{R} \Rightarrow d_\phi(z_1, z_2) = (z_1 - z_2)^2$ (squared Euclidean distance)
- For uniform ν , $E_\nu[d_\phi(Z, \hat{Z})] = \|Z - \hat{Z}\|_F^2$ (squared Frobenius distortion)

- Example 2:**

- $\phi(z) = z \log z$ for $z \in \mathbb{R}_+ \Rightarrow d_\phi(z_1, z_2) = z_1 \log(\frac{z_1}{z_2}) - z_1 + z_2$ (I-divergence)
- For uniform ν and $Z = p(X, Y)$, $\hat{Z} = q(X, Y)$, $E_\nu[d_\phi(Z, \hat{Z})] = KL(p||q)$ (KL-divergence)

Low Parameter Representations

Original matrix size: $m \times n, 5 \times 6$

U, V	1	2	3	4	5	6
1	-66	54	-63	93	51	96
2	35	87	37	-26	84	-22
3	-68	-56	-64	92	52	94
4	-30	-83	32	-24	80	-21
5	-63	-55	-60	92	53	95

Original Matrix Z

U, V	1	3	5	2	4	6
4	30	32	80	83	-24	-21
2	35	37	84	87	-26	-22
5	-63	-60	53	55	92	95
1	-66	-63	51	54	93	96
3	-68	-64	52	56	92	94

Reordered Matrix \tilde{Z}

Scheme 1: $E[Z|\hat{U}, \hat{V}]$ (co-cluster means)

\hat{U}, \hat{V}	1	2	3
1	33.5	83.5	-23.3
2	-64.0	53.5	93.7

Sizes	Num. Params
$k \times \ell$	$k\ell$
2×3	6

Low Parameter Representations

U, V	1	2	3	4	5	6
1	-66	54	-63	93	51	96
2	35	87	37	-26	84	-22
3	-68	-56	-64	92	52	94
4	-30	-83	32	-24	80	-21
5	-63	-55	-60	92	53	95

Original Matrix Z

U, V	1	3	5	2	4	6
4	30	32	80	83	-24	-21
2	35	37	84	87	-26	-22
5	-63	-60	53	55	92	95
1	-66	-63	51	54	93	96
3	-68	-64	52	56	92	94

Reordered Matrix \tilde{Z}

Scheme 2: $E[Z|\hat{U}]$ (row cluster means), $E[Z|\hat{V}]$ (column cluster means)

\hat{U}	
1	31.3
2	27.7

\hat{V}	1	2	3
	-15.3	68.5	-35.2

Sizes	Num. Params
$k \times 1, 1 \times \ell$	$k + \ell - 1$
$2 \times 1, 1 \times 3$	4

Low Parameter Representations (cont.)

U, V	1	2	3	4	5	6
1	-66	54	-63	93	51	96
2	35	87	37	-26	84	-22
3	-68	-56	-64	92	52	94
4	-30	-83	32	-24	80	-21
5	-63	-55	-60	92	53	95

Original Matrix Z

U, V	1	3	5	2	4	6
4	30	32	80	83	-24	-21
2	35	37	84	87	-26	-22
5	-63	-60	53	55	92	95
1	-66	-63	51	54	93	96
3	-68	-64	52	56	92	94

Reordered Matrix \tilde{Z}

Scheme 3: $E[Z|U]$ (row means), $E[Z|V]$ (column means),
 $E[Z|\hat{U}, \hat{V}]$ (co-cluster means)

U	
1	27.5
2	32.5
3	27.0
4	30.0
5	28.7

V	1	2	3	4	5	6
	-26.4	67.0	-23.6	45.4	64.0	48.4

\hat{U}, \hat{V}	1	2	3
1	33.5	83.5	-23.3
2	-64.0	53.5	93.7

Sizes	Num. Params
$m \times 1, 1 \times n$	$m + n + kl$
$k \times l$	$-k - l$
$5 \times 1, 1 \times 6$	12
2×3	

Low Parameter Representations (cont.)

Scheme 4: $E[Z|U, \hat{V}]$ (row-wise column cluster means),
 $E[Z|\hat{U}, V]$ (col-wise row cluster means)

U, \hat{V}	1	2	3
1	-64.5	52.5	94.5
2	36.0	85.5	24.0
3	-66.0	54.0	93.0
4	31.0	81.5	-22.5
5	-61.5	54.0	93.5

\hat{U}, V	1	2	3	4	5	6
1	32.5	85.0	34.5	-25.0	82.0	-21.5
2	-65.7	55.0	-62.3	92.3	52.0	95

Sizes	Num. Params
$m \times l, k \times n$	$ml + nk - kl$
$5 \times 3, 2 \times 6$	21

Schemes 1-4 are the only non-trivial, symmetric, co-clustering based, summary statistic representations

Matrix Reconstruction

- Ideally, we want reconstructed matrix \hat{Z} to be the “best” among all matrices Z' satisfying the following Markov property (conditional independence)

$$Z \rightarrow \{\text{summary statistics of } Z\} \rightarrow Z'$$

where “best” means closest to original Z , i.e.,

$$\hat{Z} = \underset{Z'}{\operatorname{argmin}} d_{\phi}(Z, Z')$$

- Difficult to optimize over the set of all Z' in general

Matrix Reconstruction Example

U, V	1	2	3	4	5	6
1	-66	54	-63	93	51	96
2	35	87	37	-26	84	-22
3	-68	56	-64	92	52	94
4	30	83	32	-24	80	-21
5	-63	55	-60	92	53	95

Original Matrix Z

U, V	1	3	5	2	4	6
4	30	32	80	83	-24	-21
2	35	37	84	87	-26	-22
5	-63	-60	53	55	92	95
1	-66	-63	51	54	93	96
3	-68	-64	52	56	92	94

Reordered Matrix \tilde{Z}

U, V	1	2	3	4	5	6
1	6	6	6	6	6	6
2	6	6	6	6	6	6
3	6	6	6	6	6	6
4	6	6	6	6	6	6
5	6	6	6	6	6	6

Optimal Reconstruction from global mean for all Bregman divergences

Matrix Reconstruction Example

U, V	1	2	3	4	5	6
1	-66	54	-63	93	51	96
2	35	87	37	-26	84	-22
3	-68	56	-64	92	52	94
4	30	83	32	-24	80	-21
5	-63	55	-60	92	53	95

Original Matrix Z

U, V	1	3	5	2	4	6
4	30	32	80	83	-24	-21
2	35	37	84	87	-26	-22
5	-63	-60	53	55	92	95
1	-66	-63	51	54	93	96
3	-68	-64	52	56	92	94

Reordered Matrix \tilde{Z}

U, V	1	2	3	4	5	6
1	-64.0	53.5	-64.0	93.7	53.5	93.7
2	33.5	83.5	33.5	-23.3	83.5	-23.3
3	-64.0	53.5	-64.0	93.7	53.5	93.7
4	33.5	83.5	33.5	-23.3	83.5	-23.3
5	-64.0	53.5	-64.0	93.7	53.5	93.7

Optimal Reconstruction from co-cluster means for all Bregman divergences

Matrix Reconstruction Example

U, V	1	2	3	4	5	6
1	-66	54	-63	93	51	96
2	35	87	37	-26	84	-22
3	-68	56	-64	92	52	94
4	30	83	32	-24	80	-21
5	-63	55	-60	92	53	95

Original Matrix Z

U, V	1	3	5	2	4	6
4	30	32	80	83	-24	-21
2	35	37	84	87	-26	-22
5	-63	-60	53	55	92	95
1	-66	-63	51	54	93	96
3	-68	-64	52	56	92	94

Reordered Matrix \tilde{Z}

Optimal reconstruction from row, column and co-cluster means?

Matrix Reconstruction Example

U, V	1	2	3	4	5	6
1	-66	54	-63	93	51	96
2	35	87	37	-26	84	-22
3	-68	56	-64	92	52	94
4	30	83	32	-24	80	-21
5	-63	55	-60	92	53	95

Original Matrix Z

U, V	1	3	5	2	4	6
4	30	32	80	83	-24	-21
2	35	37	84	87	-26	-22
5	-63	-60	53	55	92	95
1	-66	-63	51	54	93	96
3	-68	-64	52	56	92	94

Reordered Matrix \tilde{Z}

Optimal reconstruction from row, column and co-cluster means ?

- Depends on the Bregman divergence; non-trivial even for squared Euclidean distance
- Depends on the matrix Z as well and is difficult to quantify without restricting the search set

Bregman Information

Bregman information is the expected Bregman divergence to the mean

$$I_\phi(Z) = E_\nu[d_\phi(Z, E[Z])]$$

- Example 1: For $\phi(z) = z^2$, $\forall z \in \mathbb{R}$ and uniform ν ,
 $I_\phi(Z) = \text{squared Frobenius norm} + \text{constant}$
- Example 2: For $\phi(z) = z \log z$, $\forall z \in \mathbb{R}_+$, uniform ν , and $Z = p(X, Y)$,
 $I_\phi(Z) = \text{negative entropy} + \text{constant}$

Min. Bregman Information (MBI) Principle

- The minimum Bregman information matrix is the one with least Bregman information among all matrices Z'' whose relevant summary statistics are identical to that of Z , i.e.,

$$\operatorname{argmin}_{Z''} I_{\phi}(Z'')$$

- **Theorem:** The minimum Bregman information matrix is the “best” reconstruction of the original matrix from the summary statistics

$$\hat{Z} = \operatorname{argmin}_{Z'} E[d_{\phi}(Z, Z')] \quad = \operatorname{argmin}_{Z''} I_{\phi}(Z'')$$

independent given summary constrained on summary

& additive in natural parameter space

Min. Bregman Information (MBI) Problem

- Convex optimization problem with unique solution
- Numerically solvable using iterative scaling techniques
- Closed form solutions
 - Squared Euclidean distance \Rightarrow Least Squares solution \Rightarrow Additive models
 - I-divergence/KL-divergence \Rightarrow Max Entropy solution \Rightarrow Multiplicative models
- Example 1: For $\phi(z) = z^2$, uniform ν , Scheme 4,
 $\hat{Z} = E[Z|U, \hat{V}] + E[Z|\hat{U}, V] - E[Z|\hat{U}, \hat{V}]$ (Cho et al., Cheng et al.)
- Example 2: For $\phi(z) = z \log z$, uniform ν , $Z = p(X, Y)$, Scheme 3,
 $\hat{Z} = q(X, Y) = p(X|\hat{X})p(\hat{X}, \hat{Y})p(Y|\hat{Y})$ (Dhillon et al.)

Putting it all together

- **Given:** Matrix $Z_{m \times n}$, #row clusters k , #column clusters ℓ
- **Required:** Optimal co-clustering (ρ^*, γ^*) such that

$$(\rho^*, \gamma^*) = \underset{(\rho, \gamma)}{\operatorname{argmin}} E_{\nu} [d_{\phi}(Z, \hat{Z})] = \underset{(\rho, \gamma)}{\operatorname{argmin}} \sum_{u=1}^m \sum_{v=1}^n \nu_{uv} d_{\phi}(z_{uv}, \hat{z}_{uv})$$

where

- d_{ϕ} – **any given Bregman divergence**
- \hat{Z} – minimum Bregman information matrix corresponding to (ρ, γ) and constraints arising from **any given representation scheme**
- ν – **any given finite measure** on the matrix elements

Co-clustering Algorithm

Input: Matrix Z , measure ν , Bregman divergence d_ϕ , #row clusters l , #column clusters k , representation scheme \mathcal{C}

Output: Locally optimal co-clustering (ρ^*, γ^*)

Algorithm:

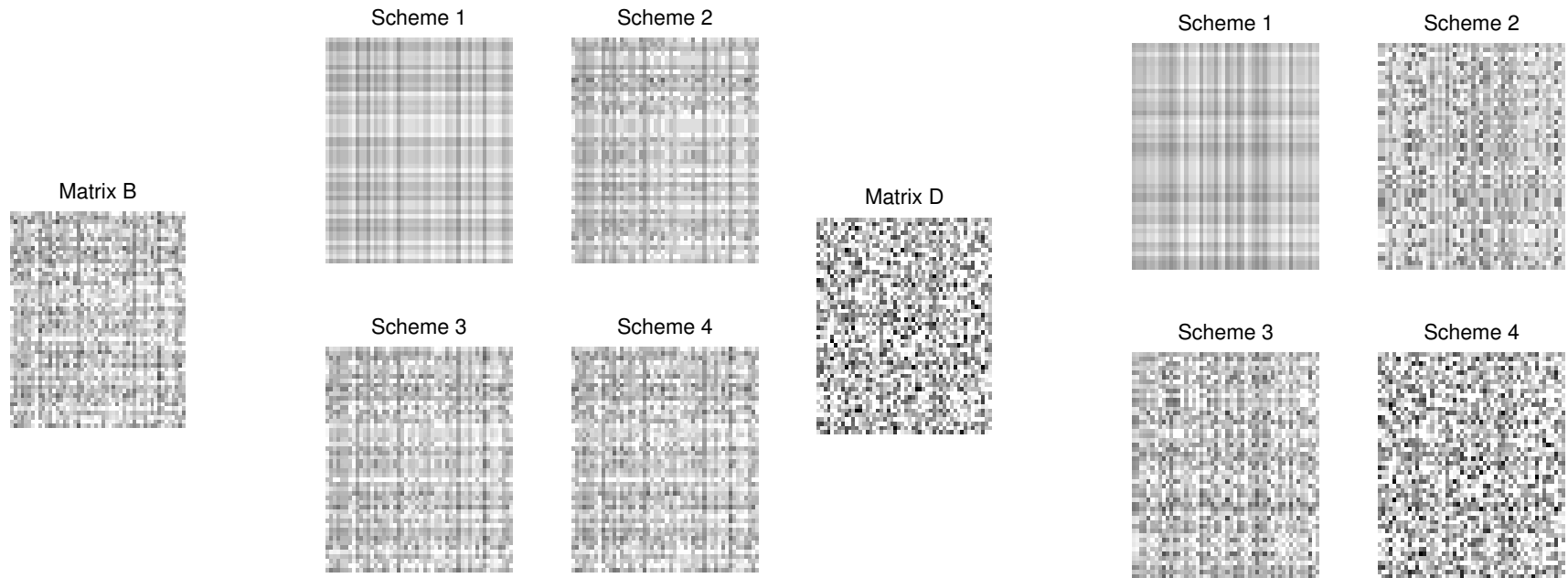
- Randomly initialize ρ^0 and γ^0
- Repeat till convergence
 - $t \leftarrow t + 1$
 - Obtain minimum Bregman information solution $\hat{Z}(\rho^t, \gamma^t, \mathcal{C})$
 - Update row clustering: $\rho^{t+1} \leftarrow \underset{\rho}{\operatorname{argmin}} E[d_\phi(Z, \hat{Z}(\rho, \gamma^t, \mathcal{C}))]$
 - Update column clustering: $\gamma^{t+1} \leftarrow \underset{\gamma}{\operatorname{argmin}} E[d_\phi(Z, \hat{Z}(\rho^{t+1}, \gamma, \mathcal{C}))]$

Algorithm Properties

- Guaranteed convergence to a local minima
- Computational complexity linear in matrix size assuming constant time for solving the minimum Bregman information problem
- Efficient implementation for sparse matrices and matrices with missing values
- Special cases include:
 - information-theoretic co-clustering (Dhillon et al.)
 - minimum sum squared residue co-clustering (Cho et al.)
 - one-sided Bregman clustering (Banerjee et al.)

Experimental Results

Information Preserving Compression

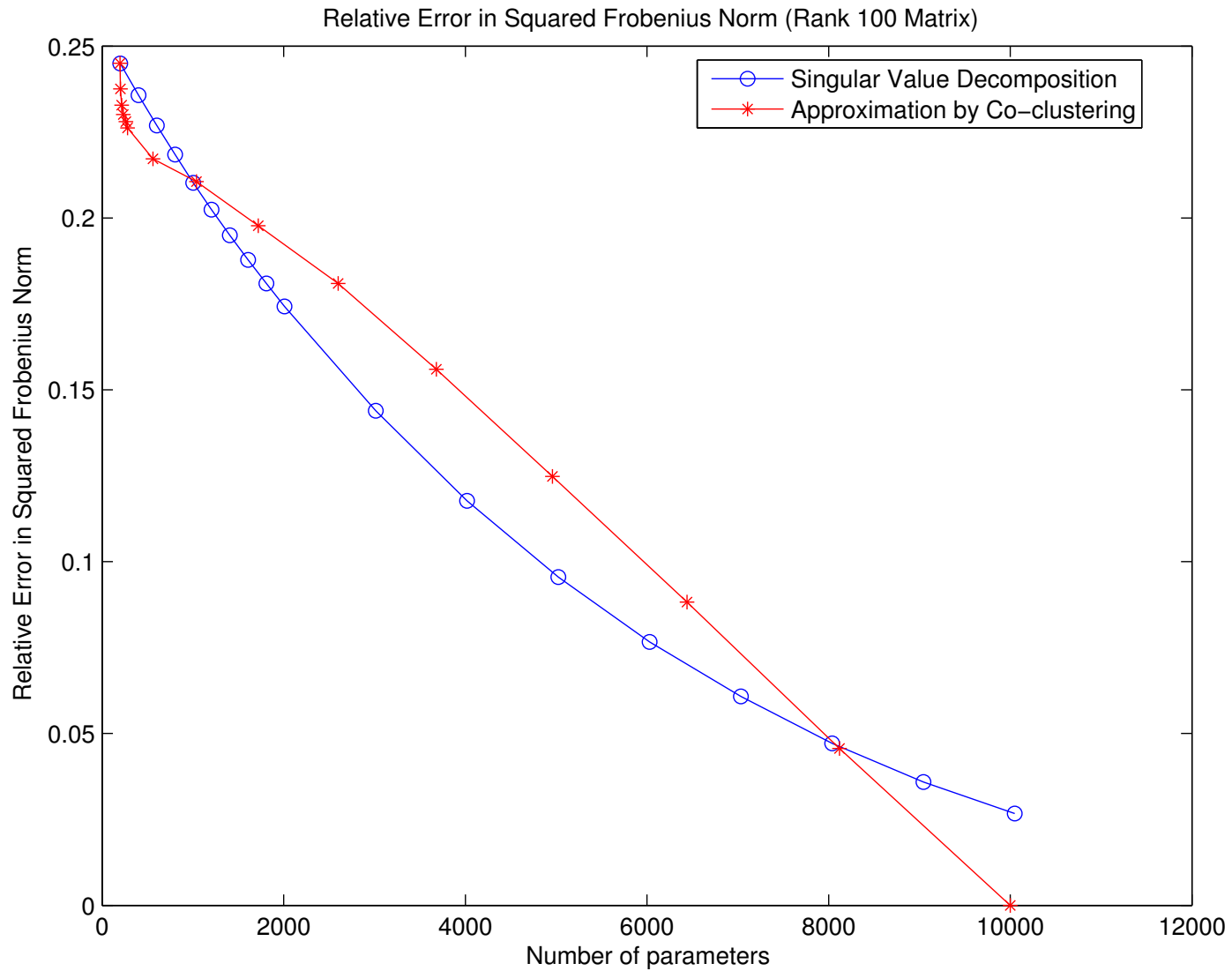


Matrix approximations with squared Euclidean distance; $m = n = 50, k = l = 10$;

Original: 2500, Scheme 1: 19, Scheme 2: 100, Scheme 3: 180, Scheme 4: 900

- Co-clustering provides low parameter approximations that preserve critical summary statistics
- Summary statistics match natural structure \Rightarrow Good compression vs. information loss trade-off

Efficiency of Compression



Matrix approximation using squared Euclidean distance and Scheme 3 co-clustering;

Clustering Sparse High Dimensional Data

(k=3,l=20)			(k=3,l=500)			(k=3,l=2500)		
1389	1	2	1364	3	18	920	49	292
9	1455	33	5	1446	21	31	1239	404
0	4	998	29	11	994	447	172	337

Confusion matrices with respect to true labels for the Classic 3 dataset with 3 document clusters and different number of word clusters

- Clustering interleaved with implicit dimensionality reduction
- Superior performance as compared to one-sided clustering

Missing Value Prediction

Algo	SqE1	SqE3	IDiv1	IDiv3	Pearson
Error	0.8398	0.7639	0.8397	0.7723	0.9413

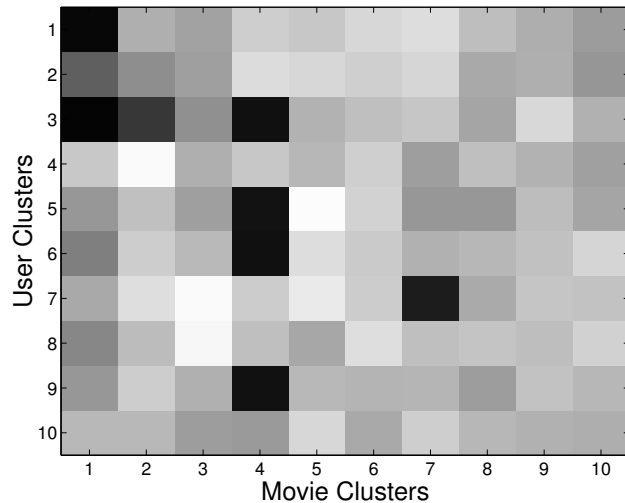
Mean absolute error for ratings (0-5) in EachMovie dataset;

SqE1/SqE3 - squared Euclidean distance with schemes 1 and 3;

IDiv1/IDiv3 - I-Divergence with schemes 1 and 3

- Assign zero measure for missing elements, co-cluster and use reconstructed matrix for prediction
- Implicit discovery of correlated sub-matrices

Learning Correlations



Cluster 1	It is a Wonderful Life, Casablanca, Life is Beautiful, An Affair to Remember
Cluster 4	Usual Suspects, Manhattan Murder Mystery, Pulp Fiction, North by NorthWest
Cluster 7	Star Trek V, Blade Runner, The Terminator, A Clockwork Orange

User cluster - Movie cluster correlations for subset of EachMovie dataset

- Use low parameter representations to discover correlations between row and column entities
- Helpful in decision support systems

Extensions

- Applicable to matrices with general elements
e.g., vectors, sequences
- Applicable to multi-dimensional data cubes
- Generative models and soft co-clustering algorithms

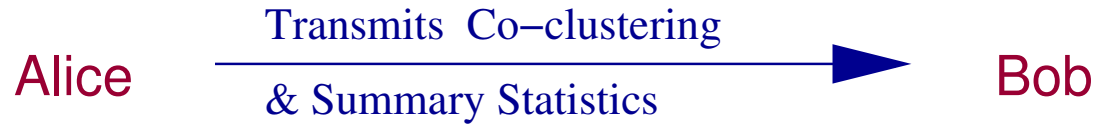
Summary

- Co-clustering framework
 - Bregman loss functions and various representation schemes
 - min. Bregman information principle and meta algorithm
- Potential applications in a number of domains
 - clustering high dimensional data, missing value prediction, etc.
 - microarray analysis, text analysis, recommender systems, etc.
- Extensions to other scenarios
 - general matrices, multi-dimensional data cubes, etc.

Thank You!



Matrix Approximation via Co-clustering



Knows input matrix Z

Does not know Z

Determines a co-clustering

Reconstructs
an approximation \hat{X} given
co-clustering & summary statistics