Robust Multi-view Graph Embedding

Akifumi Okuno $^{1,2}$  Hidetoshi Shimodaira $^{1,2}$

$^1$Graduate School of Informatics, Kyoto University, Japan
$^2$RIKEN Center for Advanced Intelligence Project, Japan

International Conference on Robust Statistics 2017
Table of contents

Existing methods
  Graph Embedding (GE)
  Cross-Domain Matching Correlation Analysis (CDMCA)
  Relation to Canonical Correlation Analysis (CCA)

Iteratively-Reweighted CDMCA (Proposed method)
  Purpose of this study
  Iteratively-Reweighted CDMCA (IR-CDMCA)
  Theoretical guarantee of convergence

Numerical experiments
  Setting
  Experiment 1: Verification of robustness
  Experiment 2: Comparison with existing methods

Conclusion

References
Graph embedding (GE)

Yan et al. (2007) proposed a method for dimensionality reduction based on graph-embedding with known graph-structured links.
Cross-Domain Matching Correlation Analysis (CDMCA)

Shimodaira (2016) extended Yan et al. (2007) as CDMCA.
Cross-Domain Matching Correlation Analysis (CDMCA)

- $x_i^{(d)} \in \mathbb{R}^{p_d}$: data vector,
Cross-Domain Matching Correlation Analysis (CDMCA)

- $x_i^{(d)} \in \mathbb{R}^{p_d}$: data vector,
- $w_{ij}^{(de)} \geq 0$ represents the strength of association between $x_i^{(d)}$ and $x_j^{(e)}$. 
Cross-Domain Matching Correlation Analysis (CDMCA)

- \( x_i^{(d)} \in \mathbb{R}^{p_d} \): data vector,
- \( w_{ij}^{(de)} \geq 0 \) represents the strength of association between \( x_i^{(d)} \) and \( x_j^{(e)} \).

\[ i \in [n_d], j \in [n_e], d \in [D], e \in [D], \]

where \([n]\) represents a set \(\{1, 2, \ldots, n\}\).
Cross-Domain Matching Correlation Analysis (CDMCA)

$\mathbf{A}^{(d)} \in \mathbb{R}^{p_d \times K}$: linear transform matrices to be estimated, so that

$$
\begin{align*}
\mathbf{w}_{ij}^{(de)} > 0 & \implies \mathbf{A}^{(d)\top} \mathbf{x}_i^{(d)} \approx \mathbf{A}^{(e)\top} \mathbf{x}_j^{(e)}.
\end{align*}
$$

- $\mathbf{x}_i^{(d)} \in \mathbb{R}^{p_d}$: data vector,
- $\mathbf{w}_{ij}^{(de)} \geq 0$ represents the strength of association between $\mathbf{x}_i^{(d)}$ and $\mathbf{x}_j^{(e)}$.

$i \in [n_d], j \in [n_e], d \in [D], e \in [D]$, where $[n]$ represents a set $\{1, 2, \ldots, n\}$. 

$\mathbf{w}_{ij}^{(de)}$.
Cross-Domain Matching Correlation Analysis (CDMCA)

CDMCA finds \( \{ \hat{\mathbf{A}}^{(d)} \} \) that minimizes

\[
\phi_0(\mathbf{A}; \mathbf{X}, \mathbf{W}) := \sum_{d=1}^{D} \sum_{e=1}^{D} \sum_{i=1}^{n_d} \sum_{j=1}^{n_e} \hat{w}_{ij}^{(de)} \| \mathbf{A}^{(d)\top} \mathbf{x}_i^{(d)} - \mathbf{A}^{(e)\top} \mathbf{x}_j^{(e)} \|_2^2,
\]

where \( \mathbf{W} \succ 0 \) and \( \hat{w}_{ij}^{(de)} := w_{ij}^{(de)} \sum_{d=1}^{D} \sum_{e=1}^{D} \sum_{i=1}^{n_d} \sum_{j=1}^{n_e} w_{ij}^{(de)} \). It can efficiently be solved by eigendecomposition. For \( D = 2 \), CDMCA is equivalent to Cross-view Graph Embedding (Huang et al., 2012; CvGE).
Cross-Domain Matching Correlation Analysis (CDMCA)

CDMCA finds \( \{ \hat{A}^{(d)} \} \) that minimizes

\[
\phi_0(A; X, W) := \sum_{d=1}^{D} \sum_{e=1}^{D} \sum_{i=1}^{n_d} \sum_{j=1}^{n_e} \tilde{w}_{ij}^{(de)} \| A^{(d)\top} x_i^{(d)} - A^{(e)\top} x_j^{(e)} \|_2^2,
\]

with a constraint

\[
\sum_{d=1}^{D} A^{(d)\top} C^{(d)} A^{(d)} = I_K,
\]

where \( C^{(d)} \succ 0 \) and \( \tilde{w}_{ij}^{(de)} := w_{ij}^{de} / \sum_{d=1}^{D} \sum_{e=1}^{D} \sum_{i=1}^{n_d} \sum_{j=1}^{n_e} w_{ij}^{de} \).

It can efficiently be solved by eigendecomposition. For \( D = 2 \), CDMCA is equivalent to Cross-view Graph Embedding (Huang et al., 2012; CvGE).
CDMCA is an extension of Canonical Correlation Analysis (CCA)

Figure: one-to-one relationship (↔ CCA)

Figure: many-to-many relationship (↔ CDMCA)
Purpose of this study

Our purpose is to reduce the adverse effect of improper associations.
What we do:

We **downweight** wrong associations.
Proposed algorithm

Iteratively-Reweighted CDMCA (IR-CDMCA)

- $\gamma > 0$ is a tuning parameter.
  - $\hat{A}_{(0)} \leftarrow \text{CDMCA}(X, W)$.
  - $t \leftarrow 0$.
  - Compute a weight $R_{(t)} := (r_{ij}^{(de)})$ by
    \[ r_{ij}^{(de)} := \exp \left( -\gamma \| \hat{A}_{(t)}^{(d)^\top} x_{i}^{(d)} - \hat{A}_{(t)}^{(e)^\top} x_{j}^{(e)} \|^2_2 \right) \]
  - update transformation matrix
    $\hat{A}_{(t+1)} \leftarrow \text{CDMCA}(X, W \circ R_{(t)})$.
  - $t \leftarrow t + 1$
  - Iterate these steps until convergence

$w_{ij}^{(de)} r_{ij}^{(de)}$ is expected to be small if $w_{ij}^{(de)}$ is false-positive.
IR-CDMCA monotonically reduces a loss function

\[
\phi_{\gamma}(A; X, W) := -\frac{1}{\gamma} \log \sum_{d=1}^{D} \sum_{e=1}^{D} \sum_{i=1}^{n_d} \sum_{j=1}^{n_e} \tilde{w}_{ij}^{(de)} \times \exp \left( -\gamma \| A^{(d)}^\top x_i^{(d)} - A^{(e)}^\top x_j^{(e)} \|_2^2 \right)
\]

as \( \phi_{\gamma}(\hat{A}(t); X, W) \geq \phi_{\gamma}(\hat{A}(t+1); X, W) \).

This function \( \phi_{\gamma}(A; X, W) \) is analogous to \( \gamma \)-divergence (Fujisawa and Eguchi, 2008).

Theorem

\( \phi_{\gamma}(\hat{A}(t); X, W), \ (t = 1, 2, \ldots) \) converges.

These theorems indicate the termination of our algorithm.
Due to the following theorem, IR-CDMCA can be regarded as a generalization of CDMCA.

**Theorem**

\[
\phi_\gamma(A; X, W) \rightarrow \phi_0(A; X, W), \text{ as } \gamma \downarrow 0.
\]

Recall that

- CDMCA minimizes \( \phi_0(A; X, W) \) s.t. \( A \in S(C) \),
- IR-CDMCA minimizes \( \phi_\gamma(A; X, W) \) s.t. \( A \in S(C) \),

where

\[
S(C) := \left\{ A = (A^{(1)^T}, \ldots, A^{(D)^T})^T \left| \sum_{d=1}^{D} A^{(d)^T} C^{(d)} A^{(d)} = I \right. \right\}.
\]
Simulation settings

(1) Underlying common data structure in $\mathbb{R}^{p_0} = \mathbb{R}^2$:

$$x_i^{(0)} := (\cos 2\pi i/10, \sin 2\pi i/10) \in \mathbb{R}^2.$$

(2) Generate vectors sharing the structure by

$$x^{(d)}_{ij} \sim \mathcal{N}[B^{(d)\top} x_i^{(0)}, \sigma^2 I_{p_d}],$$

$$j = 1, 2, \ldots, 10; \ i = 1, 2, \ldots, 10.$$

(3) Associate all vectors in the same class across views ($= \bar{W}_0$).

(4) Resample these links at rate $\alpha \in (0, 1)$ ($= W_0$).

(5) Associate vectors in the different class at rate $\xi \geq 0$ ($= W_\xi$).
Illustrative example \((\alpha = 0.5, \sigma = 0.2)\)

(a) \(W_0\)  

(b) \(W_1\) (Cont.)

Figure: CDMCA (existing method)

(a) \(W_0\)  

(b) \(W_1\) (Cont.)

Figure: IR-CDMCA with \(\gamma = 1\) (proposed method)
Experiment 1: Verification of robustness

Setting: \( D = 3, p_1 = p_2 = p_3 = 10, n_1 = n_2 = n_3 = 100 \)

\[
\hat{A}_\gamma := \arg \min_{A \in S(X^\top X)} \phi_\gamma(A; X, W_\xi)
\]

Error := \( \phi_0(\hat{A}_\gamma; X, W_0) \)

---

Table: Avg. and s.d. of errors over 100 experiments when few associations are observed (\( \alpha = 0.05 \)).

<table>
<thead>
<tr>
<th>St.Dev.</th>
<th>Method</th>
<th>( \xi = 0 )</th>
<th>( \xi = 0.2 )</th>
<th>( \xi = 0.6 )</th>
<th>( \xi = 1.0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma = 0.4 )</td>
<td>CDMCA (( \gamma = 0 ))</td>
<td>0.027 ± 0.008</td>
<td>0.043 ± 0.013</td>
<td>0.070 ± 0.026</td>
<td>0.087 ± 0.030</td>
</tr>
<tr>
<td></td>
<td>IR-CDMCA (( \gamma = 0.5 ))</td>
<td>0.027 ± 0.008</td>
<td>0.031 ± 0.010</td>
<td>0.039 ± 0.015</td>
<td>0.045 ± 0.016</td>
</tr>
<tr>
<td></td>
<td>IR-CDMCA (( \gamma = 1 ))</td>
<td>0.027 ± 0.008</td>
<td><strong>0.028</strong> ± 0.009</td>
<td><strong>0.030</strong> ± 0.010</td>
<td><strong>0.033</strong> ± 0.011</td>
</tr>
<tr>
<td>( \sigma = 1.0 )</td>
<td>CDMCA (( \gamma = 0 ))</td>
<td>0.141 ± 0.042</td>
<td>0.181 ± 0.055</td>
<td>0.227 ± 0.058</td>
<td>0.274 ± 0.063</td>
</tr>
<tr>
<td></td>
<td>IR-CDMCA (( \gamma = 0.5 ))</td>
<td><strong>0.140</strong> ± 0.041</td>
<td>0.160 ± 0.050</td>
<td>0.194 ± 0.059</td>
<td>0.243 ± 0.071</td>
</tr>
<tr>
<td></td>
<td>IR-CDMCA (( \gamma = 1 ))</td>
<td>0.141 ± 0.041</td>
<td><strong>0.157</strong> ± 0.051</td>
<td><strong>0.187</strong> ± 0.063</td>
<td><strong>0.229</strong> ± 0.072</td>
</tr>
</tbody>
</table>

IR-CDMCA is more robust than CDMCA in this experiment.
Experiment 2: Comparison with existing methods ($D = 2$)

By resampling data vectors and links across views so that associations become one-to-one, we can apply existing methods:

- **CCA**: Canonical Correlation Analysis (Hotelling, 1936)
- **KCCA**: Kernel CCA (Lai and Fyfe, 2000)
- **RCCA**: CCA with robust covariance estimators
  - **MCD**: Minimum Covariance Discriminator (Rousseeuw, 1985)
  - **OGK**: Orthogonal Ggenendian Kettering (Maronna and Zammar, 2002)
  - **MVE**: Minimum Volume Ellipsoid (Rousseeuw, 1985)
  - **S-bi**: S-estimator with biweight (Huber, 2011)

We assess these methods by mean Average Precision score (Baeza-Yates and Ribeiro-Neto, 1999; mAP). Higher mAP indicates better retrieval precision.
Experiment 2: Comparison with existing methods ($D = 2$)

Table: Many associations are observed ($\alpha = 0.5$) and $\sigma = 1.0$.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\xi = 0$</th>
<th>$\xi = 0.25$</th>
<th>$\xi = 0.5$</th>
<th>$\xi = 0.75$</th>
<th>$\xi = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCA</td>
<td>0.484 ± 0.055</td>
<td>0.408 ± 0.066</td>
<td>0.346 ± 0.061</td>
<td>0.291 ± 0.056</td>
<td>0.256 ± 0.054</td>
</tr>
<tr>
<td>KCCA ($\beta = 1$)</td>
<td><strong>0.616</strong> ± 0.060</td>
<td><strong>0.530</strong> ± 0.054</td>
<td>0.453 ± 0.062</td>
<td>0.415 ± 0.066</td>
<td>0.372 ± 0.049</td>
</tr>
<tr>
<td>KCCA ($\beta = 1.5$)</td>
<td>0.556 ± 0.076</td>
<td>0.444 ± 0.058</td>
<td>0.371 ± 0.052</td>
<td>0.337 ± 0.055</td>
<td>0.310 ± 0.050</td>
</tr>
<tr>
<td>RCCA (MCD)</td>
<td>0.443 ± 0.059</td>
<td>0.384 ± 0.072</td>
<td>0.313 ± 0.070</td>
<td>0.270 ± 0.056</td>
<td>0.230 ± 0.047</td>
</tr>
<tr>
<td>RCCA (OGK)</td>
<td>0.477 ± 0.054</td>
<td>0.434 ± 0.065</td>
<td>0.379 ± 0.068</td>
<td>0.327 ± 0.052</td>
<td>0.285 ± 0.059</td>
</tr>
<tr>
<td>RCCA (MVE)</td>
<td>0.454 ± 0.057</td>
<td>0.388 ± 0.076</td>
<td>0.323 ± 0.064</td>
<td>0.272 ± 0.059</td>
<td>0.240 ± 0.048</td>
</tr>
<tr>
<td>RCCA (S-bi)</td>
<td>0.488 ± 0.057</td>
<td>0.436 ± 0.059</td>
<td>0.384 ± 0.061</td>
<td>0.336 ± 0.062</td>
<td>0.293 ± 0.053</td>
</tr>
<tr>
<td>CDMCA ($\gamma = 0$)</td>
<td>0.518 ± 0.054</td>
<td>0.509 ± 0.053</td>
<td>0.499 ± 0.053</td>
<td>0.494 ± 0.049</td>
<td>0.487 ± 0.048</td>
</tr>
<tr>
<td>IR-CDMCA ($\gamma = 0.5$)</td>
<td>0.519 ± 0.052</td>
<td>0.518 ± 0.052</td>
<td>0.512 ± 0.053</td>
<td>0.511 ± 0.052</td>
<td>0.507 ± 0.050</td>
</tr>
<tr>
<td>IR-CDMCA ($\gamma = 1$)</td>
<td>0.521 ± 0.051</td>
<td>0.519 ± 0.052</td>
<td>0.516 ± 0.052</td>
<td>0.516 ± 0.052</td>
<td>0.514 ± 0.051</td>
</tr>
<tr>
<td>IR-CDMCA ($\gamma = 1.5$)</td>
<td>0.522 ± 0.051</td>
<td>0.520 ± 0.052</td>
<td><strong>0.517</strong> ± 0.052</td>
<td><strong>0.517</strong> ± 0.052</td>
<td><strong>0.515</strong> ± 0.051</td>
</tr>
</tbody>
</table>

MCD    Minimum Covariance Discriminator (Rousseeuw, 1985)
OGK    Orthogonal Gnenendian Kettenring (Maronna and Zamar, 2002)
MVE    Minimum Volume Ellipsoid (Rousseeuw, 1985)
S-bi   biweight-type S-estimator (Huber, 2011)
Experiment 2: Comparison with existing methods ($D = 2$)

Table: **Few** associations are observed ($\alpha = 0.05$) and $\sigma = 1.0$.

<table>
<thead>
<tr>
<th></th>
<th>$\xi = 0$</th>
<th>$\xi = 0.25$</th>
<th>$\xi = 0.5$</th>
<th>$\xi = 0.75$</th>
<th>$\xi = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCA</td>
<td>0.162 ± 0.022</td>
<td>0.159 ± 0.026</td>
<td>0.162 ± 0.019</td>
<td>0.163 ± 0.019</td>
<td>0.158 ± 0.022</td>
</tr>
<tr>
<td>KCCA ($\beta = 1$)</td>
<td>0.171 ± 0.018</td>
<td>0.173 ± 0.017</td>
<td>0.171 ± 0.018</td>
<td>0.165 ± 0.012</td>
<td>0.173 ± 0.018</td>
</tr>
<tr>
<td>KCCA ($\beta = 1.5$)</td>
<td>0.165 ± 0.014</td>
<td>0.169 ± 0.013</td>
<td>0.166 ± 0.014</td>
<td>0.161 ± 0.012</td>
<td>0.164 ± 0.009</td>
</tr>
<tr>
<td>RCCA (MCD)</td>
<td>0.157 ± 0.022</td>
<td>0.166 ± 0.029</td>
<td>0.165 ± 0.032</td>
<td>0.163 ± 0.023</td>
<td>0.163 ± 0.024</td>
</tr>
<tr>
<td>RCCA (OGK)</td>
<td>0.173 ± 0.030</td>
<td>0.176 ± 0.028</td>
<td>0.167 ± 0.027</td>
<td>0.170 ± 0.027</td>
<td>0.173 ± 0.023</td>
</tr>
<tr>
<td>RCCA (MVE)</td>
<td>0.168 ± 0.027</td>
<td>0.168 ± 0.027</td>
<td>0.161 ± 0.022</td>
<td>0.164 ± 0.018</td>
<td>0.164 ± 0.023</td>
</tr>
<tr>
<td>RCCA (S-bi)</td>
<td>0.162 ± 0.022</td>
<td>0.166 ± 0.023</td>
<td>0.170 ± 0.026</td>
<td>0.174 ± 0.029</td>
<td>0.173 ± 0.027</td>
</tr>
<tr>
<td><strong>CDMCA ($\gamma = 0$)</strong></td>
<td>0.412 ± 0.073</td>
<td>0.331 ± 0.066</td>
<td>0.300 ± 0.060</td>
<td>0.282 ± 0.060</td>
<td>0.262 ± 0.052</td>
</tr>
<tr>
<td><strong>IR-CDMCA ($\gamma = 0.5$)</strong></td>
<td>0.418 ± 0.073</td>
<td>0.377 ± 0.070</td>
<td>0.358 ± 0.071</td>
<td>0.339 ± 0.076</td>
<td>0.321 ± 0.061</td>
</tr>
<tr>
<td><strong>IR-CDMCA ($\gamma = 1$)</strong></td>
<td>0.419 ± 0.071</td>
<td>0.402 ± 0.072</td>
<td>0.383 ± 0.073</td>
<td>0.379 ± 0.076</td>
<td>0.366 ± 0.072</td>
</tr>
<tr>
<td><strong>IR-CDMCA ($\gamma = 1.5$)</strong></td>
<td><strong>0.420 ± 0.070</strong></td>
<td><strong>0.408 ± 0.071</strong></td>
<td><strong>0.395 ± 0.072</strong></td>
<td><strong>0.394 ± 0.073</strong></td>
<td><strong>0.387 ± 0.075</strong></td>
</tr>
</tbody>
</table>

**MCD** Minimum Covariance Discriminator (Rousseeuw, 1985)

**OGK** Orthogonal Gnenendian Kettenring (Maronna and Zamar, 2002)

**MVE** Minimum Volume Ellipsoid (Rousseeuw, 1985)

**S-bi** biweight-type S-estimator (Huber, 2011)
Conclusion

- We propose Iteratively-Reweighted CDMCA (IR-CDMCA), which is a robust extension of CDMCA.
- We prove the convergence of IR-CDMCA.
- IR-CDMCA outperforms CDMCA in numerical experiments.

Figure: CDMCA with cont.  Figure: IR-CDMCA with cont.


References II


Solution of CDMCA

\[ X = \text{Diag}[X^{(1)}, X^{(2)}, \ldots, X^{(D)}] \in \mathbb{R}^{n \times p}, \]

\[ W = [W^{(de)}] \in \mathbb{R}^{n \times n} \quad (W^{(de)} = (w^{(de)}_{ij}) \in \mathbb{R}^{n_d \times n_e}), \]

\[ \hat{G} = X^\top \text{diag}(W1)X \in \mathbb{R}^{p \times p}, \]

\[ \hat{H} = X^\top WX \in \mathbb{R}^{p \times p}, \]

\[ A = (A^{(1)}, A^{(2)}, \ldots, A^{(D)})^\top \in \mathbb{R}^{p \times K}, \]

where \( p = p_1 + p_2 + \cdots + p_D, \quad n = n_1 + n_2 + \cdots + n_D. \)

Solution of CDMCA is

\[ \hat{A} = \hat{G}^{-1/2}(\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_K), \]

where \( \hat{G}^{-1/2} \hat{H} \hat{G}^{-1/2} = \sum_{k=1}^{p} \hat{\lambda}_k \hat{u}_k \hat{u}_k^\top \) is eigendecomposition satisfying \( \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p. \)
Simulation settings

- Number of views: $D = 3$
- Dimension: $p_1 = p_2 = p_3 = 10$
- Sample size: $n_1 = n_2 = n_3 = 100$
- Scatter within cluster: $\sigma > 0$
- Resampling rate: $\alpha \in (0, 1]$
- Contamination rate: $\xi \geq 0$

\[
x_{ij}^{(d)} \sim N[B^{(d)\top} x_i^{(0)}, \sigma^2 I_{p_d}]
\]

Resampling at rate $\alpha$ \quad Contaminate at rate $\xi$

$W_0 \rightarrow W_0 \rightarrow W_\xi$
mean Average Precision (mAP)

For a query \( x_i^1 \in \mathbb{R}^{p1} \), we rank view-2 data vectors \( \{ x_j^2 \}_{j=1}^{n2} \subset \mathbb{R}^{p2} \) by considering euclidean distances from the query \( \{ \| (\hat{A}^1)^\top x_i^1 - (\hat{A}^2)^\top x_j^2 \|_2 \}_{j=1}^{n2} \). We define an index set of associated vectors \( S_i := \{ 1 \leq j \leq n2 \mid w_{ij}^{12} = 1 \} \), and we sort the ranking of \( \{ x_j^2 \mid j \in S_i \} \) so as to be \( q_1^{(i)} \leq q_2^{(i)} \leq \cdots \leq q_{|S_i|}^{(i)} \). Then Average Precision (AP) for a query \( x_i^1 \) is defined by

\[
\text{AP}_i := |S_i|^{-1} \sum_{j=1}^{|S_i|} (j / q_j^{(i)})
\]

and a sample mean of AP scores over all queries,

\[
\text{mAP} := \frac{1}{n1} \sum_{i=1}^{n1} \frac{1}{|S_i|} \sum_{j=1}^{|S_i|} \frac{j}{q_j^{(i)}},
\]

\[
= \text{AP}_i
\]

is called mean Average Precision (mAP). Higher mAP indicates better retrieval precision.