Robust Multi-view Graph Embedding

Akifumi Okuno [1,2] Hidetoshi Shimodaira [1,2]

[1]Graduate School of Informatics, Kyoto University, Japan

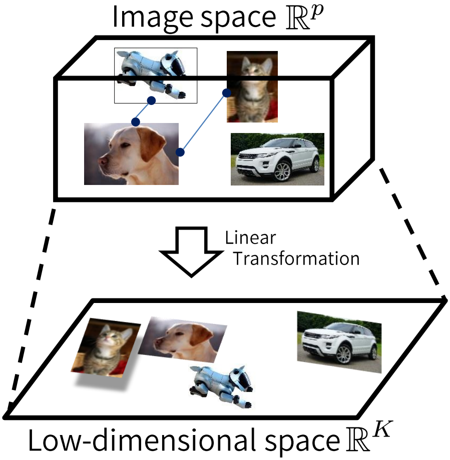
[2]RIKEN Center for Advanced Intelligence Project, Japan

International Conference on Robust Statistics 2017

Table of contents

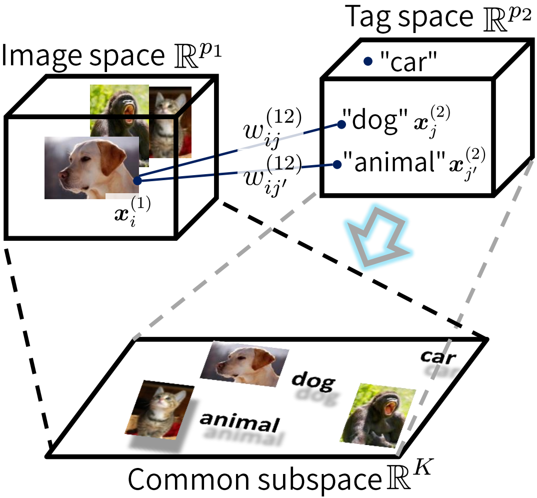
Graph embedding (GE)

Yan et al. (2007) proposed a method for dimensionality reduction based on graph-embedding with known graph-structured links.

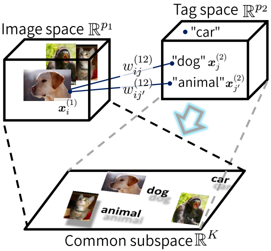


Cross-Domain Matching Correlation Analysis (CDMCA)

Shimodaira (2016) extended Yan et al. (2007) as CDMCA.

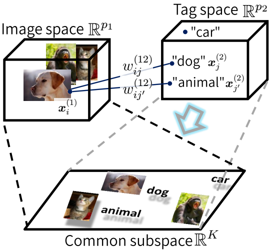


Cross-Domain Matching Correlation Analysis (CDMCA)



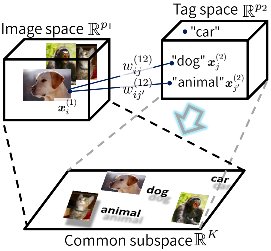
- $\mathbf{x}_i^{(d)} \in \mathbb{R}^{p_d}$:data vector,

Cross-Domain Matching Correlation Analysis (CDMCA)



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

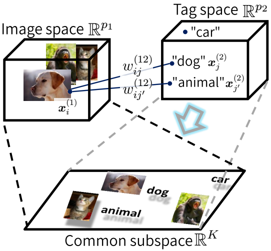
Cross-Domain Matching Correlation Analysis (CDMCA)



- $\mathbf{x}_i^{(d)} \in \mathbb{R}^{p_d}$:data vector,
- $w_{ij}^{(de)} \geq 0$ 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 $\{1, 2, \ldots, n\}$.

Cross-Domain Matching Correlation Analysis (CDMCA)



- $\mathbf{x}_i^{(d)} \in \mathbb{R}^{p_d}$:data vector,
- $w_{ij}^{(de)} \geq 0$ 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 $\{1, 2, \ldots, n\}$.

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

$$w_{ij}^{(de)} > 0 \quad \Rightarrow \quad \mathbf{A}^{(d)\top}\mathbf{x}_i^{(d)} \approx \mathbf{A}^{(e)\top}\mathbf{x}_j^{(e)}.$$

Cross-Domain Matching Correlation Analysis (CDMCA)

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

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

Cross-Domain Matching Correlation Analysis (CDMCA)

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

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

with a constraint

$$\sum_{d=1}^{D} \mathbf{A}^{(d)\top}\mathbf{C}^{(d)}\mathbf{A}^{(d)} = \mathbf{I}_K,$$

where $\mathbf{C}^{(d)} \succ 0$ 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)

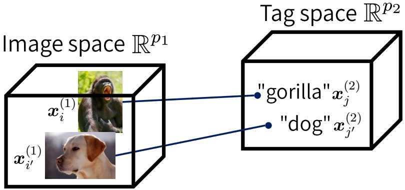


Figure: one-to-one relationship ($\Leftrightarrow$ CCA)

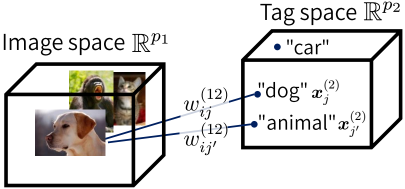


Figure: many-to-many relationship ($\Leftrightarrow$ CDMCA)

Purpose of this study

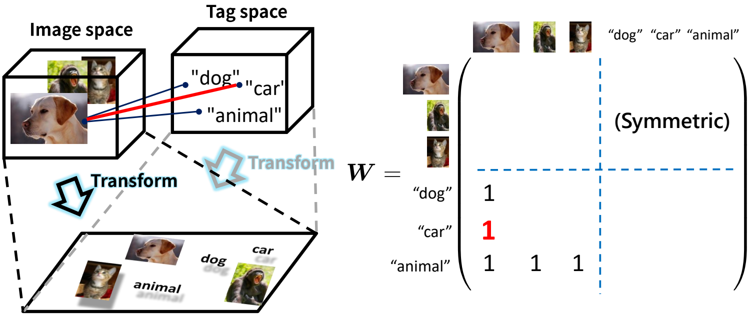
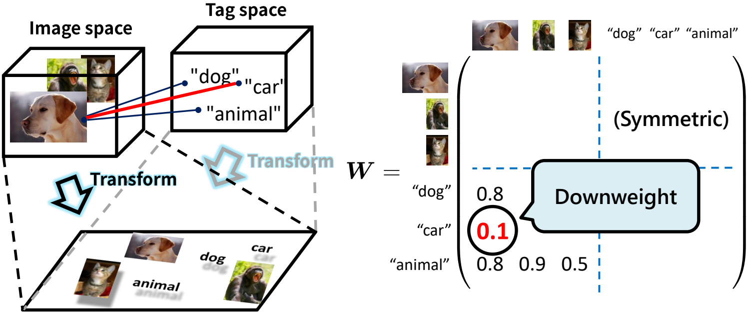


Figure: A image "Dog" is wrongly tagged with a word "car".

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.

- $\hat{\mathbf{A}}_{(0)} \leftarrow \mathrm{CDMCA}(\mathbf{X}, \mathbf{W})$.
- $t \leftarrow 0$.
- Compute a weight $\mathbf{R}_{(t)} := (r_{ij}^{(de)})$ by

$$r_{ij}^{(de)} := \exp\left(-\gamma \|\hat{\mathbf{A}}_{(t)}^{(d)\top} \mathbf{x}_i^{(d)} - \hat{\mathbf{A}}_{(t)}^{(e)\top} \mathbf{x}_j^{(e)}\|_2^2\right)$$

- update transformation matrix $\hat{\mathbf{A}}_{(t+1)} \leftarrow \mathrm{CDMCA}(\mathbf{X}, \mathbf{W} \circ \mathbf{R}_{(t)})$.
- $t \leftarrow t + 1$
- Iterate these steps until convergence

$w_{ij}^{(de)} r_{ij}^{(de)}$ is expected to be small if $w_{ij}^{(de)}$ is false-positive.

Theorem

IR-CDMCA monotonically reduces a loss function

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

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

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

Theorem

$\phi_\gamma(\hat{\mathbf{A}}_{(t)};\mathbf{X},\mathbf{W})$, $(t = 1, 2, \ldots)$ *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

$\phi_\gamma(\mathbf{A}; \mathbf{X}, \mathbf{W}) \to \phi_0(\mathbf{A}; \mathbf{X}, \mathbf{W})$, *as* $\gamma \downarrow 0$.

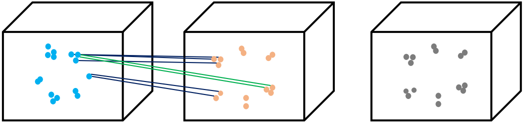
Recall that

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

where

$$\mathcal{S}(\mathbf{C}) := \left\{ \mathbf{A} = (\mathbf{A}^{(1)\top}, \cdots, \mathbf{A}^{(D)\top})^\top \,\middle|\, \sum_{d=1}^{D} \mathbf{A}^{(d)\top} \mathbf{C}^{(d)} \mathbf{A}^{(d)} = \mathbf{I} \right\}.$$

Simulation settings



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

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

(2) Generate vectors sharing the structure by

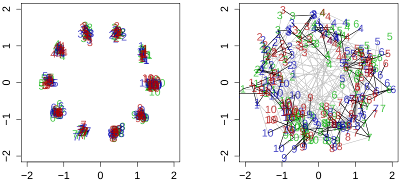
$$\mathbf{x}_{ij}^{(d)} \sim N[\mathbf{B}^{(d)\top}\mathbf{x}_i^{(0)}, \sigma^2 \mathbf{I}_{p_d}], \quad (j = 1, 2, \ldots, 10; i = 1, 2, \ldots, 10).$$

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

(4) Resample these links at rate $\alpha \in (0, 1)$ ($=\mathbf{W}_0$).

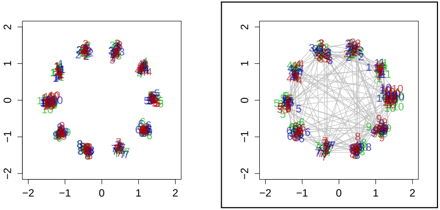
(5) Associate vectors in the different class at rate $\xi \geq 0$ ($=\mathbf{W}_\xi$).

Illustrative example ($\alpha = 0.5, \sigma = 0.2$)



(a) $\mathbf{W}_0$ (b) $\mathbf{W}_1$ (Cont.)

Figure: CDMCA (existing method)



(a) $\mathbf{W}_0$ (b) $\mathbf{W}_1$ (Cont.)

Figure: IR-CDMCA with $\gamma = 1$ (proposed method)

Experiment 1: Verification of robustness

Setting: $D = 3, p_1 = p_2 = p_3 = 10, n_1 = n_2 = n_3 = 100$

$$\hat{\mathbf{A}}_\gamma := \operatorname*{arg\,min}_{\mathbf{A} \in \mathcal{S}(\mathbf{X}^\top \mathbf{X})} \phi_\gamma(\mathbf{A}; \mathbf{X}, \mathbf{W}_\xi)$$

$$\text{Error} := \phi_0(\hat{\mathbf{A}}_\gamma; \mathbf{X}, \mathbf{W}_0)$$

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

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

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 Gnenendian 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 $\sigma = 1.0$.

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

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 $\sigma = 1.0$.

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

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.

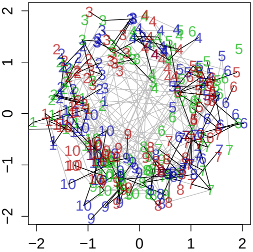


Figure: CDMCA with cont.

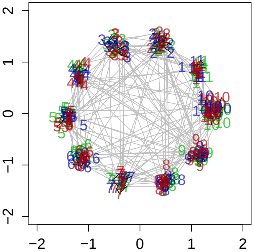


Figure: IR-CDMCA with cont.

References I

[1] Yan, S., Xu, D., Zhang, B., Zhang, H. J., Yang, Q. and Lin, S. (2007). Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, **29**(1), 40-51.

[2] Huang, Z., Shan, S., Zhang, H., Lao, S. and Chen, X. (2012). Cross-view graph embedding. In *Asian Conference on Computer Vision* (pp. 770-781). Springer Berlin Heidelberg.

[3] Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, **99**(9), 2053-2081.

[4] Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, **8**, 283-297.

References II

[5] Maronna, R. A. and Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, **44**(4), 307-317.

[6] Huber, P. J. (2011). Robust statistics. Springer Berlin Heidelberg.

[7] Salibian-Barrera, M. and Yohai, V. J. (2006). A fast algorithm for S-regression estimates. Journal of Computational and Graphical Statistics, 15(2), 414-427.

[8] Todorov V and Filzmoser P (2009) An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software* **32**(1):1-47,

[9] Lai, P. L., and Fyfe, C. (2000). Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, **10**(05), 365-377.

[10] Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, **28**(3/4), 321-377.

Solution of CDMCA

$$
\begin{aligned}
\mathbf{X} &= \mathrm{Diag}[\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(D)}] \in \mathbb{R}^{n \times p}, \\
\mathbf{W} &= [\mathbf{W}^{(de)}] \in \mathbb{R}^{n \times n} \ (\mathbf{W}^{(de)} = (w_{ij}^{(de)}) \in \mathbb{R}^{n_d \times n_e}), \\
\hat{\mathbf{G}} &= \mathbf{X}^\top \mathrm{diag}(\mathbf{W}\mathbf{1})\mathbf{X} \in \mathbb{R}^{p \times p}, \\
\hat{\mathbf{H}} &= \mathbf{X}^\top \mathbf{W}\mathbf{X} \in \mathbb{R}^{p \times p}, \\
\mathbf{A} &= (\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \cdots, \mathbf{A}^{(D)})^\top \in \mathbb{R}^{p \times K},
\end{aligned}
$$

where $p = p_1 + p_2 + \cdots + p_D,\ n = n_1 + n_2 + \cdots + n_D$.

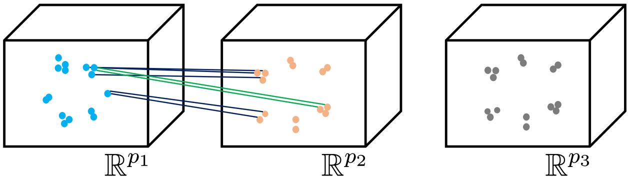
Solution of CDMCA is

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

where $\hat{\mathbf{G}}^{-1/2}\hat{\mathbf{H}}\hat{\mathbf{G}}^{-1/2} = \sum_{k=1}^{p} \hat{\lambda}_k \hat{\mathbf{u}}_k \hat{\mathbf{u}}_k^\top$ 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: $\xi \geq 0$



$$\mathbf{x}_{ij}^{(d)} \sim N[B^{(d)\top}\mathbf{x}_i^{(0)}, \sigma^2 \mathbf{I}_{p_d}]$$

$$\overline{\mathbf{W}}_0 \xrightarrow{\text{Resampling at rate } \alpha} \mathbf{W}_0 \xrightarrow{\text{Contaminate at rate } \xi} \mathbf{W}_\xi$$

mean Average Precision (mAP)

For a query $\mathbf{x}_i^1 \in \mathbb{R}^{p_1}$, we rank view-2 data vectors $\{\mathbf{x}_j^2\}_{j=1}^{n_2} \subset \mathbb{R}^{p_2}$ by considering euclidean distances from the query $\{\|(\hat{\mathbf{A}}^1)^\top \mathbf{x}_i^1 - (\hat{\mathbf{A}}^2)^\top \mathbf{x}_j^2\|_2\}_{j=1}^{n_2}$. We define an index set of associated vectors $\mathcal{S}_i := \{1 \le j \le n_2 \mid w_{ij}^{12} = 1\}$, and we sort the ranking of $\{\mathbf{x}_j^2 \mid j \in \mathcal{S}_i\}$ so as to be $q_1^{(i)} \le q_2^{(i)} \le \cdots \le q_{|\mathcal{S}_i|}^{(i)}$. Then Average Precision (AP) for a query $\mathbf{x}_i^1$ is defined by $\mathrm{AP}_i := |\mathcal{S}_i|^{-1} \sum_{j=1}^{|\mathcal{S}_i|}(j/q_j^{(i)})$, and a sample mean of AP scores over all queries,

$$\mathrm{mAP} := \frac{1}{n_1} \sum_{i=1}^{n_1} \underbrace{\frac{1}{|\mathcal{S}_i|} \sum_{j=1}^{|\mathcal{S}_i|} \frac{j}{q_j^{(i)}}}_{=:\mathrm{AP}_i},$$

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