# Statistical consistency of multi-view correlation analysis with many-to-many associations


Akifumi Okuno[†,††] and Hidetoshi Shimodaira[†,††]

[†]Graduate School of Informatics, Kyoto University [††]RIKEN Center for AI Project

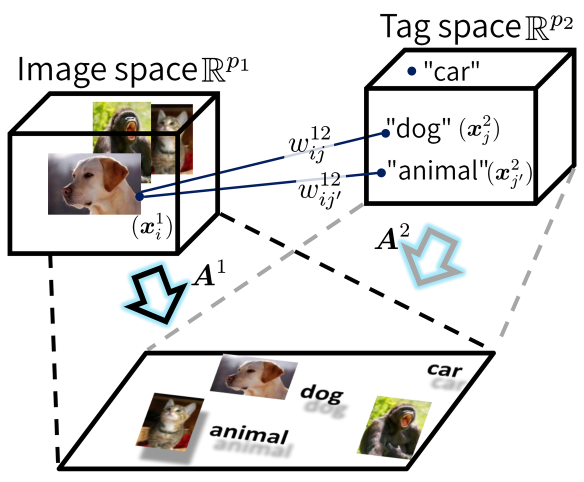


## ▷ Abstract

We have $D$-view data vectors $\{\boldsymbol{x}_i^d\}_{i=1}^{n_d} \subset \mathbb{R}^{p_d}$ and specify a non-negative real value $w_{ij}^{de}$ called matching weight, which represents the strength of association between $\boldsymbol{x}_i^d$ and $\boldsymbol{x}_j^e$, for $i = 1, 2, \ldots, n_d; j = 1, 2, \ldots, n_e; d, e = 1, 2, \ldots, D$. Cross-Domain Matching Correlation Analysis (Shimodaira, 2016, Neural Networks; CDMCA) finds linear transformation matrices $\boldsymbol{A}^d \in \mathbb{R}^{p_d \times K}$ $(d = 1, 2, \ldots, D)$ so that

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

by extending Canonical Correlation Analysis (CCA). The dimensionality $K \le \min_d\{p_d\}$ of $\boldsymbol{y}_i^d := (\boldsymbol{A}^d)^\top \boldsymbol{x}_i^d$ is manually chosen so that the transformed vectors $\{\boldsymbol{y}_i^d\}$ geometrically reflect the association structure well.




**Ex. 1** Let $\{\boldsymbol{x}_i^1\}_{i=1}^{n_1} \subset \mathbb{R}^{p_1}$ be image feature vectors, $\{\boldsymbol{x}_j^2\}_{j=1}^{n_2} \subset \mathbb{R}^{p_2}$ be tag vectors, and $w_{ij}^{12} = 1$ if the image $\boldsymbol{x}_i^1$ is tagged with the word $\boldsymbol{x}_j^2$ and $w_{ij}^{12} = 0$ if not. CDMCA finds linear transformation $\mathbb{R}^{p_d} \ni \boldsymbol{x}_i^d \mapsto (\boldsymbol{A}^d)^\top \boldsymbol{x}_i^d \in \mathbb{R}^K$ so that associated vectors get closer in $\mathbb{R}^K$.

Although some studies have already shown the advantages of using CDMCA by application experiments, its theoretical aspect is still less understood. In this presentation, we give a theoretical guarantee of CDMCA. Especially, we show following three results in this presentation.

(A) Original CDMCA requires strong assumptions to be consistent.
(B) We propose CDMCA based on Maximum Likelihood.
(C) Our proposed algorithm is consistent with weaker assumptions.


## ▷ CDMCA

Let $\tilde{w}_{ij}^{de} := w_{ij}^{de} / \sum_{d=1}^d \sum_{e=1}^D \sum_{i=1}^{n_d} \sum_{j=1}^{n_e} w_{ij}^{de}$. CDMCA minimizes

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

with respect to $\boldsymbol{A} = ((\boldsymbol{A}^1)^\top, \cdots, (\boldsymbol{A}^D)^\top)^\top$ satisfying a quadratic constraint $\boldsymbol{I}_K = \sum_{d=1}^D \sum_{i=1}^{n_d} (\sum_{e=1}^D \sum_{j=1}^{n_e} \tilde{w}_{ij}^{de})[(\boldsymbol{A}^d)^\top \boldsymbol{x}_i^d][(\boldsymbol{A}^d)^\top \boldsymbol{x}_i^d]^\top$, then we obtain $\hat{\boldsymbol{A}}$. The solution can efficiently be computed by eigenvalue decomposition.

## ▷ Proposal of a novel probabilistic model

Assuming that $K_* \le \min_{d \in [D]} p_d$, $\alpha_* \in (0, 1]$, $\boldsymbol{A}_*^d \in \mathbb{R}^{p_d \times K_*}$ are given. We propose a novel probabilistic model for $\{\boldsymbol{x}_i^d\}_{i=1}^{n_d} \subset \mathbb{R}^{p_d}$, $\{w_{ij}^{de}\}_{i,j=1}^{n_d, n_e}$, as

$$\boldsymbol{x}_i^d \overset{\text{i.i.d.}}{\sim} Q^d,$$
$$w_{ij}^{de} \mid \boldsymbol{x}_i^d, \boldsymbol{x}_j^e \overset{\text{i.i.d.}}{\sim} \text{Po}(\alpha_* \exp(-\|(\boldsymbol{A}_*^d)^\top \boldsymbol{x}_i^d - (\boldsymbol{A}_*^e)^\top \boldsymbol{x}_j^e\|_2^2)).$$

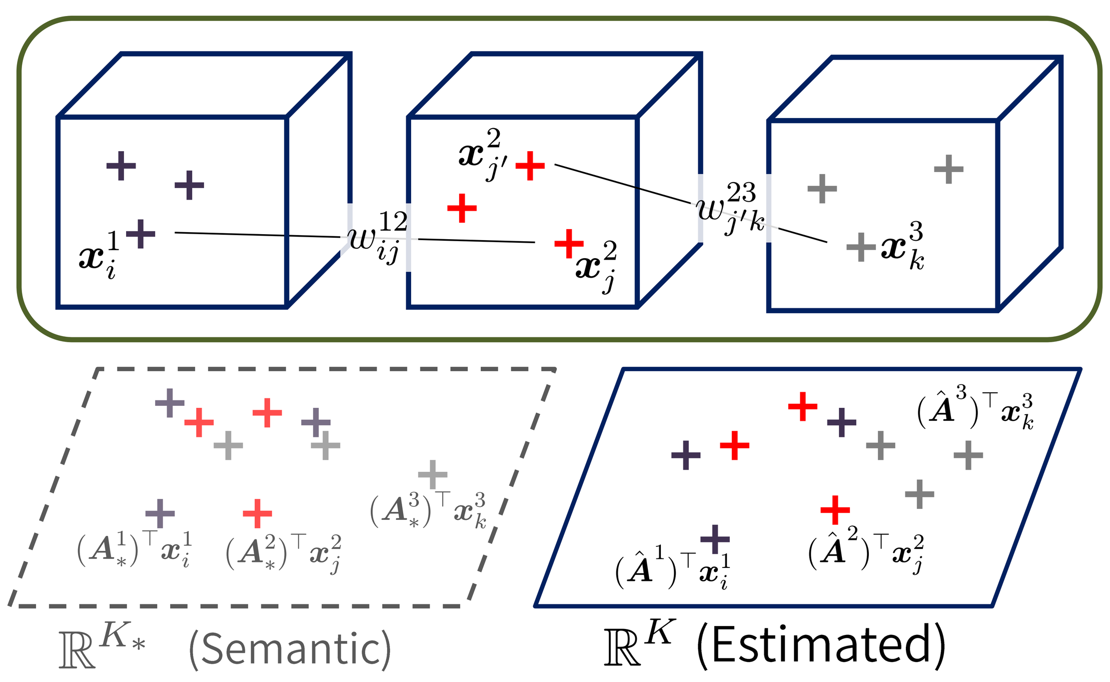



**Fig. 2** This probabilistic model generates $w_{ij}^{de} \in \{0, 1, 2, \ldots\}$, which shows the strength of association between $\boldsymbol{x}_i^d, \boldsymbol{x}_j^e$, by considering distance of their attributes in the semantic space.

We studied if the result of CDMCA $\hat{\boldsymbol{A}}_N$, which is computed with $N = n_1 + n_2 + \cdots + n_D$ data vectors $\{\boldsymbol{x}_i^d\}_{i=1}^{n_d}$ and their associations $\{w_{ij}^{de}\}_{i,j=1}^{n_d, n_e}$, converges in probability to the semantic projection matrix $\boldsymbol{A}^*$.

## ▷ (A) Consistency of CDMCA

**Theorem 1.1 (Consistency)** Let $Q^d = N_{p_d}[\boldsymbol{0}, \boldsymbol{I}]$, $\alpha_*^d = 1$, $(\boldsymbol{A}_*^d)^\top \boldsymbol{A}_*^d = \boldsymbol{I}_{K_*}$, $(\forall d)$. With some assumptions (i.e. eigengap condition) and $n_d/N \to \exists c_d \in (0, 1)$, there exists $\gamma_* > 0$ such that

$$\|\hat{\boldsymbol{A}}_N \hat{\boldsymbol{A}}_N^\top - \gamma_* \boldsymbol{A}_* \boldsymbol{A}_*^\top\|_{\text{F}} = O_p\left(\frac{1}{N - \max_d n_d}\right).$$

**Theorem 1.2 (Inconsistency)** We use the same settings as in Theorem 1.1 except for the condition of $\boldsymbol{A}_*^1$, which is replaced with $(\boldsymbol{A}_*^1)^\top \boldsymbol{A}_*^1 \ne \boldsymbol{I}_{K_*}$. Then $\forall \gamma_* > 0$,

$$\|\hat{\boldsymbol{A}}_N \hat{\boldsymbol{A}}_N^\top - \gamma_* \boldsymbol{A}_* \boldsymbol{A}_*^\top\|_{\text{F}}^2 \overset{p}{\not\to} 0.$$

CDMCA needs strong conditions to be consistent. ⇒ How to improve it?

## ▷ (B) CDMCA based on Maximum Likelihood

We maximize the likelihood w.r.t. associations $\{w_{ij}^{de}\}$, which is

$$L_N(\alpha, \boldsymbol{A}) := \prod_{d=1}^D \prod_{e=1}^D \prod_{i=1}^{n_d} \prod_{j=1}^{n_e} p(w_{ij}^{de} \mid \boldsymbol{x}_i^d, \boldsymbol{x}_j^e).$$

By taking logarithm, we maximize the equivalent objective function

$$\begin{aligned}\ell_N(\alpha, \boldsymbol{A}) &:= \log L_N(\alpha, \boldsymbol{A}) \\ &= -\underbrace{\sum_{d=1}^D \sum_{e=1}^D \sum_{i=1}^{n_d} \sum_{j=1}^{n_e} \Big[ w_{ij}^{de}(\|(\boldsymbol{A}^d)^\top \boldsymbol{x}_i^d - (\boldsymbol{A}^e)^\top \boldsymbol{x}_j^e\|_2^2 - \log \alpha)}_{\text{CDMCA minimizes this part}} \\ &\quad + \alpha \exp\left(-\|(\boldsymbol{A}^d)^\top \boldsymbol{x}_i^d - (\boldsymbol{A}^e)^\top \boldsymbol{x}_j^e\|_2^2\right) \Big] + \text{Constant}.\end{aligned}$$

We propose a method that finds optimal $\hat{\boldsymbol{A}}_N$ by maximizing $\ell_N(\alpha, \boldsymbol{A})$.

## ▷ (C) Consistency of Maximum Likelihood-CDMCA

**Theorem 1.3** Let $\delta > 0$ be sufficiently small, and let parameter space be $\boldsymbol{\Theta}_{K,\delta} := [\delta, 1] \times \{\boldsymbol{A} \in \mathbb{R}^{P \times K} \mid \delta \le \|\boldsymbol{A}^d\|_{\text{F}} \le 1/\delta\}$. Regular conditions

| | | |
|---|---|---|
| (C-1) | $n_d/N \to \exists c_d \in (0, 1)$, | (Same as Theorem 1.1) |
| (C-2) | $(\alpha_*, \boldsymbol{A}_*) \in \boldsymbol{\Theta}_{K_*,\delta}$, | ($\boldsymbol{A}_*^d$ does not necessarily be orthogonal) |
| (C-3) | $E_{Q^d}[\|\boldsymbol{x}_1^d\|_2^2] < \infty$, | ($Q^d$ can be non-normal) |

derive the consistency

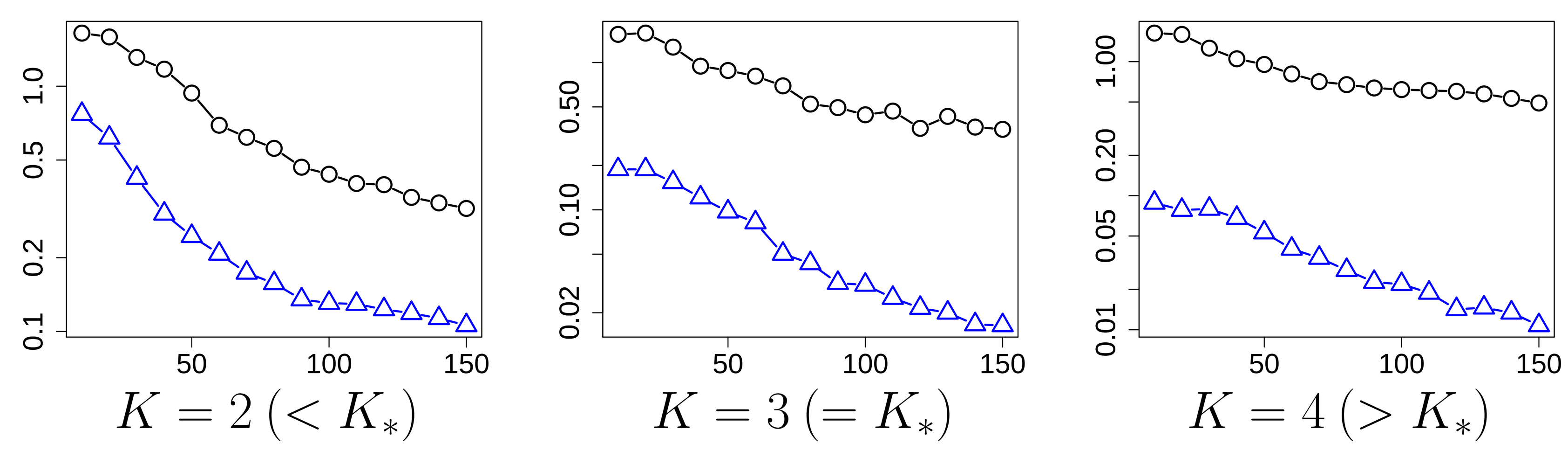
$$d\left(\underset{(\alpha, \boldsymbol{A}) \in \boldsymbol{\Theta}_{K_*,\delta}}{\arg\max}\ N^{-2} \ell_N(\alpha, \boldsymbol{A}), \boldsymbol{\Theta}_0\right) \overset{p}{\to} 0 \quad \text{as} \quad N \to \infty,$$

where $\boldsymbol{\Theta}_0 := \{(\alpha_*, \boldsymbol{A}) \in \boldsymbol{\Theta}_{\delta,K_*} \mid \|(\boldsymbol{A}^d)^\top \boldsymbol{x}_1^d - (\boldsymbol{A}^e)^\top \boldsymbol{x}_1^e\|_2^2 = \|(\boldsymbol{A}_*^d)^\top \boldsymbol{x}_1^d - (\boldsymbol{A}_*^e)^\top \boldsymbol{x}_1^e\|_2^2 \text{ a.s. } \forall d, e\}$.

## ▷ Numerical Experiment

We define the regularized error as $\varepsilon(\boldsymbol{A}) := \left\| \frac{\boldsymbol{A}\boldsymbol{A}^\top}{\|\boldsymbol{A}\boldsymbol{A}^\top\|_{\text{F}}} - \frac{\boldsymbol{A}_*\boldsymbol{A}_*^\top}{\|\boldsymbol{A}_*\boldsymbol{A}_*^\top\|_{\text{F}}} \right\|_{\text{F}}$.

Sample average of the regularized error over 100 times experiments with $K_* = 3$, $K = 2, 3, 4$ are plotted, along with increasing sample size. (Black:CDMCA, Blue:Proposal, $x$-axis:$n$, $y$-axis:$\varepsilon(\hat{\boldsymbol{A}}_n)$.)




[Note 1] $D = 3$, $(n_1, n_2, n_3) = (n, n, n)$, $(p_1, p_2, p_3) = (5, 10, 10)$, $K_* = 3$, $Q^d = N_{p_d}[\boldsymbol{0}, \boldsymbol{I}]$, $\alpha_* = 0.5$, $\text{vec}\boldsymbol{A}_*^d \sim N_{p_d K_*}[\boldsymbol{0}, p_d^{-2}\boldsymbol{I}]$.

[Note 2] Proposed method estimates $\alpha$ and $\boldsymbol{A}$ simultaneously, while original CDMCA estimates only $\boldsymbol{A}$.