Towards Practical Alternating Least-Squares for CCA

Zhiqiang Xu Ping Li

Cognitive Computing Lab

Baidu Research

{xuzhiqiang04,liping11}@baidu.com

November 3, 2019

Canonical Correlation Analysis (CCA)

- $\mathbf{X} \in \mathbb{R}^{d_x \times n}$ and $\mathbf{Y} \in \mathbb{R}^{d_y \times n}$: given data matrix pair
- Empirical cross-covariance matrix

$$\mathbf{C}_{xy} = \frac{1}{n}\mathbf{X}\mathbf{Y}^\top$$

- Two empirical auto-covariance matrices

$$\mathbf{C}_{xx} = \frac{1}{n}\mathbf{X}\mathbf{X}^\top + r_x\mathbf{I}, \quad \mathbf{C}_{yy} = \frac{1}{n}\mathbf{Y}\mathbf{Y}^\top + r_y\mathbf{I}$$

CCA

$$\max_{\mathbf{\Phi}^\top\mathbf{C}_{xx}\mathbf{\Phi}=\mathbf{\Psi}^\top\mathbf{C}_{yy}\mathbf{\Psi}=\mathbf{I}} \mathrm{tr}(\mathbf{\Phi}^\top\mathbf{C}_{xy}\mathbf{\Psi})$$

where $\mathbf{\Phi} \in \mathbb{R}^{d_x \times k}$ and $\mathbf{\Psi} \in \mathbb{R}^{d_y \times k}$, $k \geq 1$

Canonical Correlation Analysis (CCA)

Optimal Solution

$$(\mathbf{\Phi}^\star, \mathbf{\Psi}^\star) = (\mathbf{C}_{xx}^{-\frac{1}{2}}\mathbf{P}, \mathbf{C}_{yy}^{-\frac{1}{2}}\mathbf{Q}) \triangleq (\mathbf{U}, \mathbf{V}),$$

- $\mathbf{P} \in \mathbb{R}^{d_x \times k}$ and $\mathbf{Q} \in \mathbb{R}^{d_y \times k}$: top-$k$ left and right singular subspaces of $\mathbf{C} = \mathbf{C}_{xx}^{-\frac{1}{2}}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-\frac{1}{2}}$ in Euclidean metrics, respectively,

$$\mathbf{C} \overset{\text{SVD}}{=\!=} \mathbf{P}\mathbf{\Sigma}\mathbf{Q}^\top + \mathbf{P}_\perp\mathbf{\Sigma}_\perp\mathbf{Q}_\perp^\top,$$

  where $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \cdots, \sigma_k)$ with σ_i being the i-th largest singular value, and $\mathbf{\Sigma}_\perp \in \mathbb{R}^{(r-k)\times(r-k)}$ is diagonal with $r = \mathrm{rank}(\mathbf{C})$.
- $\mathbf{U} \in \mathbb{R}^{d_x \times k}$ and $\mathbf{V} \in \mathbb{R}^{d_y \times k}$: top-$k$ canonical subspaces, which are in metrics $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$, respectively,

$$\mathbf{C}_{xy} = \mathbf{C}_{xx}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{C}_{yy} + \mathbf{C}_{xx}\mathbf{U}_\perp\mathbf{\Sigma}_\perp\mathbf{V}_\perp^\top\mathbf{C}_{yy}$$

Alternating Least-Squares (ALS)

Generalized Eigenvalue Problem Formulation

$$\mathbf{A}\mathbf{w} = \lambda \mathbf{B}\mathbf{w} \quad \text{(cca-stationarity)}$$

where

$$\mathbf{A} = \begin{pmatrix} & \mathbf{C}_{xy} \\ \mathbf{C}_{xy}^{\top} & \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} \mathbf{C}_{xx} & \\ & \mathbf{C}_{yy} \end{pmatrix}.$$

- $(\mathbf{A}, \mathbf{B})$'s generalized eigenvectors

$$\frac{1}{\sqrt{2}} \left\{ \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{v}_1 \end{pmatrix}, \begin{pmatrix} \mathbf{u}_2 \\ \mathbf{v}_2 \end{pmatrix}, \dots, \begin{pmatrix} \mathbf{u}_r \\ \mathbf{v}_r \end{pmatrix}, \begin{pmatrix} -\mathbf{u}_r \\ \mathbf{v}_r \end{pmatrix}, \begin{pmatrix} -\mathbf{u}_2 \\ \mathbf{v}_2 \end{pmatrix}, \dots, \begin{pmatrix} -\mathbf{u}_1 \\ \mathbf{v}_1 \end{pmatrix} \right\},$$

corresponding to generalized eigenvalues $\sigma_1, \sigma_2, \cdots, \sigma_r, -\sigma_r, -\sigma_2, \cdots, -\sigma_1$, respectively, where

$$\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_k), \quad \mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k)$$

Power Method for GEV

$$\tilde{\mathbf{\Omega}}_{t+1} = \mathbf{B}^{-1}\mathbf{A}\mathbf{\Omega}_t + \varsigma_t, \quad \mathbf{\Omega}_{t+1} = \tilde{\mathbf{\Omega}}_{t+1}(\tilde{\mathbf{\Omega}}_{t+1}^\top \mathbf{B}\tilde{\mathbf{\Omega}}_{t+1})^{-\frac{1}{2}}$$

[1]

- Suppose $\mathbf{\Omega}_t \in \mathbb{R}^{(d_x+d_y)\times j}$, where j is even.
- Column space of $\mathbf{\Omega}_t$ converges to a top-j generalized eigenspace spanned by $\frac{1}{2}\left\{\begin{pmatrix}\mathbf{u}_1\\ \mathbf{v}_1\end{pmatrix}, \begin{pmatrix}-\mathbf{u}_1\\ \mathbf{v}_1\end{pmatrix} \cdots, \begin{pmatrix}\mathbf{u}_{\frac{j}{2}}\\ \mathbf{v}_{\frac{j}{2}}\end{pmatrix}, \begin{pmatrix}-\mathbf{u}_{\frac{j}{2}}\\ \mathbf{v}_{\frac{j}{2}}\end{pmatrix}\right\}$ corresponding to j largest generalized eigenvalues $\sigma_1, -\sigma_1, \cdots, \sigma_{\frac{j}{2}}, -\sigma_{\frac{j}{2}}$ in magnitude.

[1] Rong Ge et al. "Efficient Algorithms for Large-scale Generalized Eigenvector Computation and Canonical Correlation Analysis". In: *International Conference on Machine Learning*. 2016, pp. 2741–2750.

Alternating Least-Squares (ALS)

Update Equations

$$\begin{cases} \tilde{\mathbf{\Phi}}_t = \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{\Psi}_{t-1} + \xi_{t-1}, & \mathbf{\Phi}_t = \tilde{\mathbf{\Phi}}_t(\tilde{\mathbf{\Phi}}_t^\top \mathbf{C}_{xx}\tilde{\mathbf{\Phi}}_t + \tilde{\mathbf{\Psi}}_t^\top \mathbf{C}_{yy}\tilde{\mathbf{\Psi}}_t)^{-\frac{1}{2}} \\ \tilde{\mathbf{\Psi}}_t = \mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\mathbf{\Phi}_{t-1} + \eta_{t-1}, & \mathbf{\Psi}_t = \tilde{\mathbf{\Psi}}_t(\tilde{\mathbf{\Phi}}_t^\top \mathbf{C}_{xx}\tilde{\mathbf{\Phi}}_t + \tilde{\mathbf{\Psi}}_t^\top \mathbf{C}_{yy}\tilde{\mathbf{\Psi}}_t)^{-\frac{1}{2}} \end{cases}$$

- To recover $\mathbf{U}$ and $\mathbf{V}$, one needs to first set $j = 2k$ for having $\mathrm{col}(\mathbf{\Omega}_\infty)$ spanned by $\frac{1}{2}\begin{pmatrix} \mathbf{U} & -\mathbf{U} \\ \mathbf{V} & \mathbf{V} \end{pmatrix}$
- that is, $\mathbf{\Phi}_t \in \mathbb{R}^{d_x \times 2k}$ and $\mathbf{\Psi}_t \in \mathbb{R}^{d_y \times 2k}$
- then do random projection with random Gaussian matrix $\mathbf{G} \in \mathbb{R}^{2k \times k}$,

$$\begin{cases} \tilde{\tilde{\mathbf{\Phi}}}_T = \mathbf{\Phi}_T\mathbf{G} \\ \tilde{\tilde{\mathbf{\Psi}}}_T = \mathbf{\Psi}_T\mathbf{G} \end{cases}, \qquad \begin{cases} \widehat{\mathbf{U}} = \tilde{\tilde{\mathbf{\Phi}}}_T(\tilde{\tilde{\mathbf{\Phi}}}_T^\top \mathbf{C}_{xx}\tilde{\tilde{\mathbf{\Phi}}}_T)^{-\frac{1}{2}} \\ \widehat{\mathbf{V}} = \tilde{\tilde{\mathbf{\Psi}}}_T(\tilde{\tilde{\mathbf{\Psi}}}_T^\top \mathbf{C}_{yy}\tilde{\tilde{\mathbf{\Psi}}}_T)^{-\frac{1}{2}} \end{cases}$$

Truly Alternating Least-Squares

Coupled Equations of Half the Size

$$\begin{cases} \tilde{\mathbf{\Phi}}_t = \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{\Psi}_{t-1} + \xi_{t-1}, & \mathbf{\Phi}_t = \tilde{\mathbf{\Phi}}_t(\tilde{\mathbf{\Phi}}_t^\top \mathbf{C}_{xx}\tilde{\mathbf{\Phi}}_t)^{-\frac{1}{2}} \\ \tilde{\mathbf{\Psi}}_t = \mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\mathbf{\Phi}_t + \eta_t, & \mathbf{\Psi}_t = \tilde{\mathbf{\Psi}}_t(\tilde{\mathbf{\Psi}}_t^\top \mathbf{C}_{yy}\tilde{\mathbf{\Psi}}_t)^{-\frac{1}{2}} \end{cases}$$

- $\mathbf{\Phi}_t \in \mathbb{R}^{d_x \times k}$ and $\mathbf{\Psi}_t \in \mathbb{R}^{d_y \times k}$, more memory-efficient
- No need of the random projection any more

$$\widehat{\mathbf{U}} = \mathbf{\Phi}_T \text{ and } \widehat{\mathbf{V}} = \mathbf{\Psi}_T$$

- Faster roughly by a factor of $\frac{\sigma_k}{\sigma_k + \sigma_{k+1}}$, especially for cases of a small singular value gap

Truly Alternating Least-Squares (TALS)

TALS-CCA

Algorithm 1: TALS-CCA

Input: T, k, data matrices $\mathbf{X}$, $\mathbf{Y}$

Output: approximate top-k canonical subspaces $(\mathbf{\Phi}_T, \mathbf{\Psi}_T)$

$\mathbf{\Phi}_0 = \mathrm{GS}_{\mathbf{C}_{xx}}(\mathbf{\Phi}_{\mathrm{init}})$, $\mathbf{\Phi}_{\mathrm{init}} \in \mathbb{R}^{d_x \times k}$ is random Gaussian

$\mathbf{\Psi}_0 = \mathrm{GS}_{\mathbf{C}_{yy}}(\mathbf{\Psi}_{\mathrm{init}})$, $\mathbf{\Psi}_{\mathrm{init}} \in \mathbb{R}^{d_y \times k}$ is random Gaussian

for $t = 1, 2, \cdots, T$ **do**

$\quad \tilde{\mathbf{\Phi}}_t \approx \arg\min l_t(\tilde{\mathbf{\Phi}}) = \frac{1}{2n}\|\mathbf{X}^\top \tilde{\mathbf{\Phi}} - \mathbf{Y}^\top \mathbf{\Psi}_{t-1}\|_F^2 + \frac{r_x}{2}\|\tilde{\mathbf{\Phi}}\|_F^2$

$\quad$ starting from $\tilde{\mathbf{\Phi}}^{(0)} = \mathbf{\Phi}_{t-1}(\mathbf{\Phi}_{t-1}^\top \mathbf{C}_{xx} \mathbf{\Phi}_{t-1})^{-1}(\mathbf{\Phi}_{t-1}^\top \mathbf{C}_{xy} \mathbf{\Psi}_{t-1})$

$\quad \mathbf{\Phi}_t = \mathrm{GS}_{\mathbf{C}_{xx}}(\tilde{\mathbf{\Phi}}_t)$

$\quad \tilde{\mathbf{\Psi}}_t \approx \arg\min s_t(\tilde{\mathbf{\Psi}}) = \frac{1}{2n}\|\mathbf{Y}^\top \tilde{\mathbf{\Psi}} - \mathbf{X}^\top \mathbf{\Phi}_t\|_F^2 + \frac{r_y}{2}\|\tilde{\mathbf{\Psi}}\|_F^2$

$\quad$ starting from $\tilde{\mathbf{\Psi}}^{(0)} = \mathbf{\Psi}_{t-1}(\mathbf{\Psi}_{t-1}^\top \mathbf{C}_{yy} \mathbf{\Psi}_{t-1})^{-1}(\mathbf{\Psi}_{t-1}^\top \mathbf{C}_{xy}^\top \mathbf{\Phi}_t)$

$\quad \mathbf{\Psi}_t = \mathrm{GS}_{\mathbf{C}_{yy}}(\tilde{\mathbf{\Psi}}_t)$

end for

Faster Alternating Least-Squares (FALS)

Momentum Acceleration

$$\begin{cases} \tilde{\mathbf{\Phi}}_t = \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{\Psi}_{t-1} - \beta\mathbf{\Phi}_{t-2} + \xi_{t-1}, & \mathbf{\Phi}_t = \tilde{\mathbf{\Phi}}_t(\tilde{\mathbf{\Phi}}_t^\top \mathbf{C}_{xx}\tilde{\mathbf{\Phi}}_t)^{-\frac{1}{2}} \\ \tilde{\mathbf{\Psi}}_t = \mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\mathbf{\Phi}_t - \beta\mathbf{\Psi}_{t-1} + \eta_t, & \mathbf{\Psi}_t = \tilde{\mathbf{\Psi}}_t(\tilde{\mathbf{\Psi}}_t^\top \mathbf{C}_{yy}\tilde{\mathbf{\Psi}}_t)^{-\frac{1}{2}} \end{cases}$$

- Memory-efficient: $\mathbf{\Phi}_t \in \mathbb{R}^{d_x \times k}$ and $\mathbf{\Psi}_t \in \mathbb{R}^{d_y \times k}$
- No need of the random projection: $\widehat{\mathbf{U}} = \mathbf{\Phi}_T$ and $\widehat{\mathbf{V}} = \mathbf{\Psi}_T$
- Least-squares becomes

$$\begin{cases} \min l_t(\tilde{\mathbf{\Phi}}) = \frac{1}{2n}\|\mathbf{X}^\top(\tilde{\mathbf{\Phi}} + \beta\mathbf{\Phi}_{t-2}) - \mathbf{Y}^\top\mathbf{\Psi}_{t-1}\|_F^2 + \frac{r_x}{2}\|\tilde{\mathbf{\Phi}} + \beta\mathbf{\Phi}_{t-2}\|_F^2, \\ \min s_t(\tilde{\mathbf{\Psi}}) = \frac{1}{2n}\|\mathbf{Y}^\top(\tilde{\mathbf{\Psi}} + \beta\mathbf{\Psi}_{t-1}) - \mathbf{X}^\top\mathbf{\Phi}_t\|_F^2 + \frac{r_y}{2}\|\tilde{\mathbf{\Psi}} + \beta\mathbf{\Psi}_{t-1}\|_F^2 \end{cases}$$

FALS-CCA

Algorithm 2: FALS-CCA

Input: T, k, momentum parameter β, data matrices $\mathbf{X}$, $\mathbf{Y}$

Output: approximate top-k canonical subspaces $(\mathbf{\Phi}_T, \mathbf{\Psi}_T)$

$\mathbf{\Phi}_{-1} = \mathbf{0} \in \mathbb{R}^{d_x \times k}$

$\mathbf{\Phi}_0 = \mathrm{GS}_{\mathbf{C}_{xx}}(\mathbf{\Phi}_{\mathrm{init}})$, $\mathbf{\Phi}_{\mathrm{init}} \in \mathbb{R}^{d_x \times k}$ is random Gaussian

$\mathbf{\Psi}_0 = \mathrm{GS}_{\mathbf{C}_{yy}}(\mathbf{\Psi}_{\mathrm{init}})$, $\mathbf{\Psi}_{\mathrm{init}} \in \mathbb{R}^{d_y \times k}$ is random Gaussian

for $t = 1, 2, \cdots, T$ **do**

  $\tilde{\mathbf{\Phi}}_t \approx \arg\min l_t(\tilde{\mathbf{\Phi}}) = \frac{1}{2n}\|\mathbf{X}^\top(\tilde{\mathbf{\Phi}} + \beta\mathbf{\Phi}_{t-2}) - \mathbf{Y}^\top\mathbf{\Psi}_{t-1}\|_F^2 + \frac{r_x}{2}\|\tilde{\mathbf{\Phi}} + \beta\mathbf{\Phi}_{t-2}\|_F^2$

  starting from $\tilde{\mathbf{\Phi}}^{(0)} = \mathbf{\Phi}_{t-1}(\mathbf{\Phi}_{t-1}^\top \mathbf{C}_{xx}\mathbf{\Phi}_{t-1})^{-1}(\mathbf{\Phi}_{t-1}^\top \mathbf{C}_{xy}\mathbf{\Psi}_{t-1})$

  $\mathbf{\Phi}_t = \mathrm{GS}_{\mathbf{C}_{xx}}(\tilde{\mathbf{\Phi}}_t)$

  $\tilde{\mathbf{\Psi}}_t \approx \arg\min s_t(\tilde{\mathbf{\Psi}}) = \frac{1}{2n}\|\mathbf{Y}^\top(\tilde{\mathbf{\Psi}} + \beta\mathbf{\Psi}_{t-1}) - \mathbf{X}^\top\mathbf{\Phi}_t\|_F^2 + \frac{r_y}{2}\|\tilde{\mathbf{\Psi}} + \beta\mathbf{\Psi}_{t-1}\|_F^2$

  starting from $\tilde{\mathbf{\Psi}}^{(0)} = \mathbf{\Psi}_{t-1}(\mathbf{\Psi}_{t-1}^\top \mathbf{C}_{yy}\mathbf{\Psi}_{t-1})^{-1}(\mathbf{\Psi}_{t-1}^\top \mathbf{C}_{xy}^\top\mathbf{\Phi}_t)$

  $\mathbf{\Psi}_t = \mathrm{GS}_{\mathbf{C}_{yy}}(\tilde{\mathbf{\Psi}}_t)$

end for

Faster Alternating Least-Squares (FALS-T_{tals})

- Optimal[2] momentum parameter β should be around $\frac{\sigma_{k+1}^2}{4}$
- $\boldsymbol{\Sigma}$ estimates by $(\mathbf{U}^\top \mathbf{C}_{xx}\mathbf{U})^{-1}\mathbf{U}^\top \mathbf{C}_{xy}\mathbf{V} = \boldsymbol{\Sigma} = (\mathbf{V}^\top \mathbf{C}_{yy}\mathbf{V})^{-1}\mathbf{V}^\top \mathbf{C}_{xy}^\top \mathbf{U}$:

$$\boldsymbol{\Sigma}^{(t,1)} = (\boldsymbol{\Phi}_t^\top \mathbf{C}_{xx}\boldsymbol{\Phi}_t)^{-1}\boldsymbol{\Phi}_t^\top \mathbf{C}_{xy}\boldsymbol{\Psi}_t,$$
$$\boldsymbol{\Sigma}^{(t,2)} = (\boldsymbol{\Psi}_{t-1}^\top \mathbf{C}_{yy}\boldsymbol{\Psi}_{t-1})^{-1}\boldsymbol{\Psi}_{t-1}^\top \mathbf{C}_{xy}^\top \boldsymbol{\Phi}_t$$

- β estimates by TALS:

$$\hat{\beta} = \frac{1}{4}\min_i(\boldsymbol{\Sigma}_{ii}^{(T_{tals},1)})^2 \text{ or } \frac{1}{4}\min_i(\boldsymbol{\Sigma}_{ii}^{(T_{tals},2)})^2,$$

where T_{tals} is small.

[2] Peng Xu et al. "Accelerated Stochastic Power Iteration". In: *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*. 2018, pp. 58–67. URL: http://proceedings.mlr.press/v84/xu18a.html.

Adaptive Momentum Acceleration

$$\begin{cases} \tilde{\mathbf{\Phi}}_t = \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{\Psi}_{t-1} - \beta_{t-1}^{\phi}\mathbf{\Phi}_{t-2} + \xi_{t-1}, & \mathbf{\Phi}_t = \tilde{\mathbf{\Phi}}_t(\tilde{\mathbf{\Phi}}_t^\top \mathbf{C}_{xx}\tilde{\mathbf{\Phi}}_t)^{-\frac{1}{2}} \\ \tilde{\mathbf{\Psi}}_t = \mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\mathbf{\Phi}_t - \beta_t^{\psi}\mathbf{\Psi}_{t-1} + \eta_t, & \mathbf{\Psi}_t = \tilde{\mathbf{\Psi}}_t(\tilde{\mathbf{\Psi}}_t^\top \mathbf{C}_{yy}\tilde{\mathbf{\Psi}}_t)^{-\frac{1}{2}} \end{cases}$$

- Adaptive momentum parameters:

$$\beta_t^{\phi} = \frac{1}{4}\min_i(\mathbf{\Sigma}_{ii}^{(t,1)})^2 \text{ and } \beta_t^{\psi} = \frac{1}{4}\min_i(\mathbf{\Sigma}_{ii}^{(t,2)})^2$$

- Least-squares becomes

$$\begin{cases} \min l_t(\tilde{\mathbf{\Phi}}) = \frac{1}{2n}\|\mathbf{X}^\top(\tilde{\mathbf{\Phi}} + \beta_{t-1}^{\phi}\mathbf{\Phi}_{t-2}) - \mathbf{Y}^\top\mathbf{\Psi}_{t-1}\|_F^2 + \frac{r_x}{2}\|\tilde{\mathbf{\Phi}} + \beta_{t-1}^{\phi}\mathbf{\Phi}_{t-2}\|_F^2, \\ \min s_t(\tilde{\mathbf{\Psi}}) = \frac{1}{2n}\|\mathbf{Y}^\top(\tilde{\mathbf{\Psi}} + \beta_t^{\psi}\mathbf{\Psi}_{t-1}) - \mathbf{X}^\top\mathbf{\Phi}_t\|_F^2 + \frac{r_y}{2}\|\tilde{\mathbf{\Psi}} + \beta_t^{\psi}\mathbf{\Psi}_{t-1}\|_F^2, \end{cases}$$

with initials $\tilde{\mathbf{\Phi}}^{(0)} = \mathbf{\Phi}_{t-1}\mathbf{\Sigma}^{(t-1,1)}$, $\tilde{\mathbf{\Psi}}^{(0)} = \mathbf{\Psi}_{t-1}\mathbf{\Sigma}^{(t-1,2)}$, respectively

AALS-CCA

Algorithm 3: AALS-CCA

Input: T, k, data matrices $\mathbf{X}$, $\mathbf{Y}$

Output: approximate top-k canonical subspaces $(\mathbf{\Phi}_T, \mathbf{\Psi}_T)$

$\mathbf{\Phi}_{-1} = \mathbf{0} \in \mathbb{R}^{d_x \times k}$

$\mathbf{\Phi}_0 = \mathrm{GS}_{\mathbf{C}_{xx}}(\mathbf{\Phi}_{\mathrm{init}})$, $\mathbf{\Phi}_{\mathrm{init}} \in \mathbb{R}^{d_x \times k}$ is random Gaussian

$\mathbf{\Psi}_0 = \mathrm{GS}_{\mathbf{C}_{yy}}(\mathbf{\Psi}_{\mathrm{init}})$, $\mathbf{\Psi}_{\mathrm{init}} \in \mathbb{R}^{d_y \times k}$ is random Gaussian

for $t = 1, 2, \cdots, T$ **do**

$\quad \beta_{t-1}^{\phi} = \frac{1}{4}\min_i(\mathbf{\Sigma}_{ii}^{(t-1,1)})^2$, where $\mathbf{\Sigma}^{(t-1,1)} = (\mathbf{\Phi}_{t-1}^\top \mathbf{C}_{xx} \mathbf{\Phi}_{t-1})^{-1} \mathbf{\Phi}_{t-1}^\top \mathbf{C}_{xy} \mathbf{\Psi}_{t-1}$

$\quad \tilde{\mathbf{\Phi}}_t \approx \arg\min l_t(\tilde{\mathbf{\Phi}}) = \frac{1}{2n}\|\mathbf{X}^\top(\tilde{\mathbf{\Phi}} + \beta_{t-1}^{\phi}\mathbf{\Phi}_{t-2}) - \mathbf{Y}^\top \mathbf{\Psi}_{t-1}\|_F^2 + \frac{r_x}{2}\|\tilde{\mathbf{\Phi}} + \beta_{t-1}^{\phi}\mathbf{\Phi}_{t-2}\|_F^2$

$\quad$ starting from $\tilde{\mathbf{\Phi}}^{(0)} = \mathbf{\Phi}_{t-1}\mathbf{\Sigma}^{(t-1,1)}$

$\quad \mathbf{\Phi}_t = \mathrm{GS}_{\mathbf{C}_{xx}}(\tilde{\mathbf{\Phi}}_t)$

$\quad \beta_t^{\psi} = \frac{1}{4}\min_i(\mathbf{\Sigma}_{ii}^{(t,2)})^2$, where $\mathbf{\Sigma}^{(t,2)} = (\mathbf{\Psi}_{t-1}^\top \mathbf{C}_{yy} \mathbf{\Psi}_{t-1})^{-1} \mathbf{\Psi}_{t-1}^\top \mathbf{C}_{xy}^\top \mathbf{\Phi}_t$

$\quad \tilde{\mathbf{\Psi}}_t \approx \arg\min s_t(\tilde{\mathbf{\Psi}}) = \frac{1}{2n}\|\mathbf{Y}^\top(\tilde{\mathbf{\Psi}} + \beta_t^{\psi}\mathbf{\Psi}_{t-1}) - \mathbf{X}^\top \mathbf{\Phi}_t\|_F^2 + \frac{r_y}{2}\|\tilde{\mathbf{\Psi}} + \beta_t^{\psi}\mathbf{\Psi}_{t-1}\|_F^2$

$\quad$ starting from $\tilde{\mathbf{\Psi}}^{(0)} = \mathbf{\Psi}_{t-1}\mathbf{\Sigma}^{(t,2)}$

$\quad \mathbf{\Psi}_t = \mathrm{GS}_{\mathbf{C}_{yy}}(\tilde{\mathbf{\Psi}}_t)$

end for

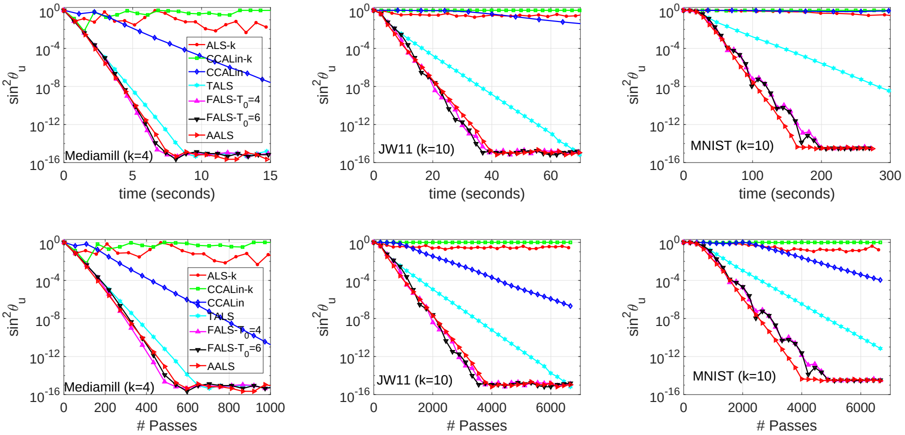
Experiments

- Data

DATA	Description	d_x	d_y	n
Memdiamill	images and its labels	100	120	30000
JW11	acoustic and articulation	273	112	30000
MNIST	left and right halves of images	392	392	60000

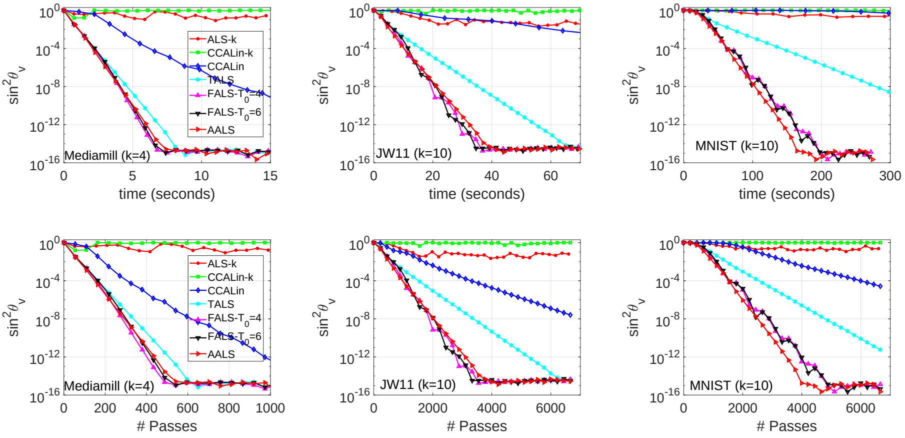
- Quality measures:

$$\sin^2\theta_u \triangleq \sin^2\theta_{\max}(\mathbf{\Phi}_t, \mathbf{U}), \quad \sin^2\theta_v \triangleq \sin^2\theta_{\max}(\mathbf{\Psi}_t, \mathbf{V})$$



10^{0}
10^{-4}
10^{-8}
10^{-12}
10^{-16}
$\sin^2\theta_u$
ALS-k
CCALin-k
CCALin
TALS
FALS-T_0=4
FALS-T_0=6
AALS
Mediamill (k=4)
JW11 (k=10)
MNIST (k=10)
0
5
10
15
20
40
60
100
200
300
time (seconds)
400
600
800
1000
2000
4000
6000
Passes

Experiments



Thank you!