Towards Practical Alternating Least-Squares for CCA

Zhiqiang Xu and Ping Li

Cognitive Computing Lab, Baidu Research

NEURAL INFORMATION PROCESSING SYSTEMS

Baidu 百度

Canonical Correlation Analysis (CCA)

Problem: $$\max_{\Phi^\top C_{xx}\Phi=\Psi^\top C_{yy}\Psi=I} \operatorname{tr}(\Phi^\top C_{xy}\Psi)$$

- ☐ Given data matrix pair $(\mathbf{X},\mathbf{Y})\in\mathbb{R}^{d_x\times n}\times\mathbb{R}^{d_y\times n}$
- ☐ Canonical variable pair $(\mathbf{\Phi},\mathbf{\Psi})\in\mathbb{R}^{d_x\times k}\times\mathbb{R}^{d_y\times k}$, $k\geq 1$
- ☐ Cross/auto-covariance matrices

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

- ☐ Ground truth $(\mathbf{\Phi}^\star,\mathbf{\Psi}^\star)=(\mathbf{U},\mathbf{V})$ a.k.a. canonical subspaces

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

Alternating Least-Squares (ALS)

Update equations

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

- ☐ $\begin{pmatrix}\mathbf{\Phi}_t\\ \mathbf{\Psi}_t\end{pmatrix}\to\frac{1}{\sqrt{2}}\begin{pmatrix}\mathbf{u}_1 & -\mathbf{u}_1 & \cdots & \mathbf{u}_k & -\mathbf{u}_k\\ \mathbf{v}_1 & \mathbf{v}_1 & \cdots & \mathbf{v}_k & \mathbf{v}_k\end{pmatrix}$
- ☐ Generalized eigenvalues $\sigma_1,-\sigma_1,\cdots,\sigma_k,-\sigma_k$
- ☐ $\mathbf{A}=\begin{pmatrix} & \mathbf{C}_{xy}\\ \mathbf{C}_{xy}^\top & \end{pmatrix}$ and $\mathbf{B}=\begin{pmatrix}\mathbf{C}_{xx} & \\ & \mathbf{C}_{yy}\end{pmatrix}$
- ☐ Variable size $(\mathbf{\Phi}_t,\mathbf{\Psi}_t)\in\mathbb{R}^{d_x\times\boxed{2k}}\times\mathbb{R}^{d_y\times\boxed{2k}}$

Post-processing

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

- ☐ Projection with random Gaussian matrix $\mathbf{G}\in\mathbb{R}^{2k\times k}$

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

Truly Alternating Least-Squares (TALS)

Coupled equations of half the size

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

- ☐ Minimize $l_t(\tilde{\mathbf{\Phi}})=\frac{1}{2n}\|\mathbf{X}^\top\tilde{\mathbf{\Phi}}-\mathbf{Y}^\top\mathbf{\Psi}_{t-1}\|_F^2+\frac{r_x}{2}\|\tilde{\mathbf{\Phi}}\|_F^2$ for $\tilde{\mathbf{\Phi}}_t$ with $\tilde{\mathbf{\Phi}}^{(0)}=\mathbf{\Phi}_{t-1}(\mathbf{\Phi}_{t-1}^\top\mathbf{C}_{xx}\mathbf{\Phi}_{t-1})^{-1}(\mathbf{\Phi}_{t-1}^\top\mathbf{C}_{xy}\mathbf{\Psi}_{t-1})$
- ☐ Minimize $s_t(\tilde{\mathbf{\Psi}})=\frac{1}{2n}\|\mathbf{Y}^\top\tilde{\mathbf{\Psi}}-\mathbf{X}^\top\mathbf{\Phi}_t\|_F^2+\frac{r_y}{2}\|\tilde{\mathbf{\Psi}}\|_F^2$ for $\tilde{\mathbf{\Psi}}_t$ with $\tilde{\mathbf{\Psi}}^{(0)}=\mathbf{\Psi}_{t-1}(\mathbf{\Psi}_{t-1}^\top\mathbf{C}_{yy}\mathbf{\Psi}_{t-1})^{-1}(\mathbf{\Psi}_{t-1}^\top\mathbf{C}_{xy}^\top\mathbf{\Phi}_t)$
- ☐ $(\mathbf{\Phi}_t,\mathbf{\Psi}_t)\in\mathbb{R}^{d_x\times\boxed{k}}\times\mathbb{R}^{d_y\times\boxed{k}}$ more memory efficient
- ☐ $(\hat{\mathbf{U}},\hat{\mathbf{V}})=(\mathbf{\Phi}_T,\mathbf{\Psi}_T)$ no need of the post-processing
- ☐ Roughly $\frac{\sigma_k}{\sigma_k+\sigma_{k+1}}$ faster esp. when a small gap exists

Faster Alternating Least-Squares (FALS)

Momentum acceleration

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

- ☐ Minimize

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

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

- ☐ Minimize

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

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

- ☐ $\hat{\beta}=\frac{1}{4}\min_i(\mathbf{\Sigma}_{ii}^{(T_{tals},1)})^2$ or $\frac{1}{4}\min_i(\mathbf{\Sigma}_{ii}^{(T_{tals},2)})^2$ with a small T_{tals}

Quality measures

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

Adaptive Alternating Least-Squares (AALS)

Adaptive momentum

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

- ☐ Optimal momentum parameter is around $\sigma_{k+1}^2/4$

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

- ☐ $\beta_t^\phi=\frac{1}{4}\min_i(\mathbf{\Sigma}_{ii}^{(t,1)})^2$ and $\beta_t^\psi=\frac{1}{4}\min_i(\mathbf{\Sigma}_{ii}^{(t,2)})^2$
- ☐ $\tilde{\mathbf{\Phi}}^{(0)}=\mathbf{\Phi}_{t-1}\mathbf{\Sigma}^{(t-1,1)}$ and $\tilde{\mathbf{\Psi}}^{(0)}=\mathbf{\Psi}_{t-1}\mathbf{\Sigma}^{(t,2)}$

Experiments

- ☐ ALS-k is ALS in Wang et al. NIPS 2016 with block size k
- ☐ CCALin-k is CCALin in Ge et al. ICML 2016 with block size k