NEURAL INFORMATION PROCESSING SYSTEMS

Towards Practical Alternating Least-Squares for CCA

Zhiqiang Xu and Ping Li
Cognitive Computing Lab, Baidu Research

Canonical Correlation Analysis (CCA)

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

- ☐ 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 $(\Phi,\Psi)\in\mathbb{R}^{d_x\times k}\times\mathbb{R}^{d_y\times k}$, $k\ge 1$
- ☐ 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 $(\Phi^\star,\Psi^\star)=(\mathbf{U},\mathbf{V})$ a.k.a. canonical subspaces

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

Alternating Least-Squares (ALS)

Update equations

$$\begin{cases}\tilde\Phi_t=\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\Psi_{t-1}+\xi_{t-1}, & \Phi_t=\tilde\Phi_t(\tilde\Phi_t^\top\mathbf{C}_{xx}\tilde\Phi_t+\tilde\Psi_t^\top\mathbf{C}_{yy}\tilde\Psi_t)^{-\frac12}\\ \tilde\Psi_t=\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\Phi_{t-1}+\eta_{t-1}, & \Psi_t=\tilde\Psi_t(\tilde\Phi_t^\top\mathbf{C}_{xx}\tilde\Phi_t+\tilde\Psi_t^\top\mathbf{C}_{yy}\tilde\Psi_t)^{-\frac12}\end{cases}$$

- ☐ $\begin{pmatrix}\Phi_t\\ \Psi_t\end{pmatrix}\to\frac{1}{\sqrt2}\begin{pmatrix}\mathbf{u}_1 & -\mathbf{u}_1 & \cdots & \mathbf{u}_k & -\mathbf{u}_k\\ \mathbf{v}_1 & \mathbf{v}_1 & \cdots & \mathbf{v}_k & \mathbf{v}_k\end{pmatrix}$
- ☐ 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 $(\Phi_t,\Psi_t)\in\mathbb{R}^{d_x\times\boxed{2k}}\times\mathbb{R}^{d_y\times\boxed{2k}}$

Post-processing

$$\begin{cases}\hat\Phi_T=\Phi_T\mathbf{G}\\ \hat\Psi_T=\Psi_T\mathbf{G}\end{cases},\quad\begin{cases}\hat{\mathbf{U}}=\hat\Phi_T(\hat\Phi_T^\top\mathbf{C}_{xx}\hat\Phi_T)^{-\frac12}\\ \hat{\mathbf{V}}=\hat\Psi_T(\hat\Psi_T^\top\mathbf{C}_{yy}\hat\Psi_T)^{-\frac12}\end{cases}$$

- ☐ 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

$$\begin{cases}\tilde\Phi_t=\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\Psi_{t-1}+\xi_{t-1}, & \Phi_t=\tilde\Phi_t(\tilde\Phi_t^\top\mathbf{C}_{xx}\tilde\Phi_t)^{-\frac12}\\ \tilde\Psi_t=\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\boxed{\Phi_t}+\eta_t, & \Psi_t=\tilde\Psi_t(\tilde\Psi_t^\top\mathbf{C}_{yy}\tilde\Psi_t)^{-\frac12}\end{cases}$$

- ☐ Minimize $l_t(\tilde\Phi)=\frac{1}{2n}\|\mathbf{X}^\top\tilde\Phi-\mathbf{Y}^\top\Psi_{t-1}\|_F^2+\frac{r_x}{2}\|\tilde\Phi\|_F^2$ for $\tilde\Phi_t$ with $\tilde\Phi^{(0)}=\Phi_{t-1}(\Phi_{t-1}^\top\mathbf{C}_{xx}\Phi_{t-1})^{-1}(\Phi_{t-1}^\top\mathbf{C}_{xy}\Psi_{t-1})$
- ☐ Minimize $s_t(\tilde\Psi)=\frac{1}{2n}\|\mathbf{Y}^\top\tilde\Psi-\mathbf{X}^\top\Phi_t\|_F^2+\frac{r_y}{2}\|\tilde\Psi\|_F^2$ for $\tilde\Psi_t$ with $\tilde\Psi^{(0)}=\Psi_{t-1}(\Psi_{t-1}^\top\mathbf{C}_{yy}\Psi_{t-1})^{-1}(\Psi_{t-1}^\top\mathbf{C}_{xy}^\top\Phi_t)$
- ☐ $(\Phi_t,\Psi_t)\in\mathbb{R}^{d_x\times\boxed{k}}\times\mathbb{R}^{d_y\times\boxed{k}}$ more memory efficient
- ☐ $(\hat{\mathbf{U}},\hat{\mathbf{V}})=(\Phi_T,\Psi_T)$ 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

$$\begin{cases}\tilde\Phi_t=\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\Psi_{t-1}-\beta\Phi_{t-2}+\xi_{t-1}, & \Phi_t=\tilde\Phi_t(\tilde\Phi_t^\top\mathbf{C}_{xx}\tilde\Phi_t)^{-\frac12}\\ \tilde\Psi_t=\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\Phi_t-\beta\Psi_{t-1}+\eta_t, & \Psi_t=\tilde\Psi_t(\tilde\Psi_t^\top\mathbf{C}_{yy}\tilde\Psi_t)^{-\frac12}\end{cases}$$

- ☐ Minimize

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

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

- ☐ Minimize

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

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

- ☐ $\hat\beta=\frac14\min_i(\Sigma_{ii}^{(T_{tals},1)})^2$ or $\frac14\min_i(\Sigma_{ii}^{(T_{tals},2)})^2$ with a small T_{tals}

Quality measures

$$\sin^2\theta_u\triangleq\sin^2\theta_{\max}(\Phi_t,\mathbf{U}),\quad\sin^2\theta_v\triangleq\sin^2\theta_{\max}(\Psi_t,\mathbf{V})$$

Adaptive Alternating Least-Squares (AALS)

Adaptive momentum

$$\begin{cases}\tilde\Phi_t=\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\Psi_{t-1}-\beta_{t-1}^\phi\Phi_{t-2}+\xi_{t-1}, & \Phi_t=\tilde\Phi_t(\tilde\Phi_t^\top\mathbf{C}_{xx}\tilde\Phi_t)^{-\frac12}\\ \tilde\Psi_t=\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\Phi_t-\beta_t^\psi\Psi_{t-1}+\eta_t, & \Psi_t=\tilde\Psi_t(\tilde\Psi_t^\top\mathbf{C}_{yy}\tilde\Psi_t)^{-\frac12}\end{cases}$$

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

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

- ☐ $\beta_t^\phi=\frac14\min_i(\Sigma_{ii}^{(t,1)})^2$ and $\beta_t^\psi=\frac14\min_i(\Sigma_{ii}^{(t,2)})^2$
- ☐ $\tilde\Phi^{(0)}=\Phi_{t-1}\Sigma^{(t-1,1)}$ and $\tilde\Psi^{(0)}=\Psi_{t-1}\Sigma^{(t,2)}$

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