Variants of Canonical Correlation Analysis

Speaker : Semin Choi

Department of Statistics, Seoul National University, South Korea

December 3, 2016

Canonical Correlation Analysis (Hotelling, 1936)

- Let $(X, Y) \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ denote random vectors with covariances $(\Sigma_{11}, \Sigma_{22})$ and cross-covariance Σ_{12}.
- CCA finds pairs of linear projections of the two views, $(v'X, u'Y)$ that are maximally correlated:

$$
\begin{aligned}
(v^*, u^*) &= \operatorname*{argmax}_{v,u} \operatorname{corr}(v'X, u'Y) \\
&= \operatorname*{argmax}_{v,u} \frac{v'\Sigma_{12}u}{\sqrt{v'\Sigma_{11}v u'\Sigma_{22}u}} \\
&= \operatorname*{argmax}_{v'\Sigma_{11}v = u'\Sigma_{22}u = 1} v'\Sigma_{12}u
\end{aligned}
$$

Canonical Correlation Analysis

- When finding multiple pairs of vectors (v^i, u^i), subsequent projections are also constrained to be uncorrelated with previous ones:

$$v^i \Sigma_{11} v^j = u^i \Sigma_{22} u^j = 0 \text{ for } i < j.$$

- We obtain the following formulation to identify the top $k \leq \min(p_1, p_2)$ projections:

$$\begin{aligned} \text{maximize:} \quad & \text{tr}(V' \Sigma_{12} U) \\ \text{subject to:} \quad & V' \Sigma_{11} V = U' \Sigma_{22} U = I. \end{aligned}$$

where $V \in \mathbb{R}^{p_1 \times k}$ and $U \in \mathbb{R}^{p_2 \times k}$.

Canonical Correlation Analysis

- Define $T_1 \triangleq \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ and $T_2 \triangleq \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$.
- Then, the optimum objective value is the sum of the top k eigenvalues of T_1 (or T_2).
- Let V_k be the matrix of the first k eigenvectors of T_1 and U_k be the matrix of the first k eigenvectors of T_2.
- Then, the optimum is attained at $(V^*, U^*) = (V_k, U_k)$.

The Two-Block Mode B of Wold's Algorithm (Wold, 1975; Wegelin, 2000)

- Given the centered data $\mathbf{X} \in \mathbb{R}^{n \times p_1}$ and $\mathbf{Y} \in \mathbb{R}^{n \times p_2}$,

$$\hat{T}_1 = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})(\mathbf{Y}'\mathbf{Y})^{-1}(\mathbf{Y}'\mathbf{X}),$$
$$\hat{T}_2 = (\mathbf{Y}'\mathbf{Y})^{-1}(\mathbf{Y}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}).$$

- We can obtain the eigenvectors and eigenvalues of $\hat{T}_1$ and $\hat{T}_2$ by power method.
- It can be viewed as an iterative projection procedure.

The Two-Block Mode B of Wold's Algorithm

- Let $\omega = \mathbf{Y}u$ and $\xi = \mathbf{X}v$.
- Wold's Algorithm :
  1. $r \leftarrow 1$.
  2. Let $\mathbf{X}^{(r)} \leftarrow \mathbf{X}$ and $\mathbf{Y}^{(r)} \leftarrow \mathbf{Y}$.
  3. Standardize $\mathbf{X}^{(r)}$ and $\mathbf{Y}^{(r)}$.
  4. Set $k \leftarrow 0$.
  5. Assign arbitrary normalized values $\hat{v}_r^{(0)}$ and $\hat{u}_r^{(0)}$.
  6. Estimate ξ_r, ω_r, v_r and u_r iteratively, as follows:
     Repeat
     1. $k \leftarrow k + 1$.
     2. $\hat{\xi}_r^k \leftarrow \mathbf{X}^{(r)}\hat{v}_r^{(k-1)}$ and $\hat{\omega}_r^k \leftarrow \mathbf{Y}^{(r)}\hat{u}_r^{(k-1)}$
     3. Compute $\hat{v}_r^{(k)}$ and $\hat{u}_r^{(k)}$ by performing multiple regression:

$$\hat{u}_r^{(k)} = \underset{u_r^{(k)}}{\text{argmin}} |\hat{\xi}_r^k - \mathbf{Y}^{(r)}u_r^{(k)}|^2$$

$$\hat{v}_r^{(k)} = \underset{v_r^{(k)}}{\text{argmin}} |\hat{\omega}_r^k - \mathbf{X}^{(r)}v_r^{(k)}|^2$$

     4. Normalize $\hat{v}_r^{(k)}$ and $\hat{u}_r^{(k)}$.

The Two-Block Mode B of Wold's Algorithm

- Wold's algorithm(cont')
  7 Fit the simple linear regression :

$$\mathbf{X}_j^{(r)} \approx \hat{\gamma}_j \hat{\xi}_r, \quad j = 1, \ldots, p_1$$
$$\mathbf{Y}_j^{(r)} \approx \hat{\theta}_j \hat{\omega}_r, \quad j = 1, \ldots, p_2.$$

  8 Determine the residual matrices of $\mathbf{X}^{(r)}$ and $\mathbf{Y}^{(r)}$.

$$\mathbf{X}^{(r+1)} \leftarrow \mathbf{X}^{(r)} - \hat{\xi}_r \hat{\gamma}'$$
$$\mathbf{Y}^{(r+1)} \leftarrow \mathbf{Y}^{(r)} - \hat{\omega}_r \hat{\theta}'$$

  9 $r \leftarrow r + 1$ and return to Step 3.

Penalized CCA(Waaijenborg et al., 2008)

- Penalized linear regression techniques can be easily adapted to Wold's algorithm, by modifying step 6-3.
- We used the elastic net.
- Selection of the penalty parameters : minimize Δ_{cor}.

$$\Delta_{\text{cor}} = \frac{\sum_{j=1}^{k} ||cor(\mathbf{X}_{-j}\hat{v}^{-j}, \mathbf{Y}_{-j}\hat{u}^{-j})| - |cor(\mathbf{X}_{j}\hat{v}^{-j}, \mathbf{Y}_{j}\hat{u}^{-j})||}{k}$$

Sparse CCA via Precision Adjusted Iterative Thresholding (Chen et al., 2013)

- Waaijenborg(2008), Wiesel et al.(2008) :
  - based on heuristics to avoid the non-convex nature of CCA problem.
  - there is no guarantee whether these algorithms would lead to consistent estimators.
- Witten et al.(2009), Parkhomenko et al.(2009) :
  - using diagonal matrix or even identity matrix to approximate the unknown matrices $(\Sigma_1^{-1}, \Sigma_2^{-1})$.

Sparse CCA via Precision Adjusted Iterative Thresholding

Proposition 1.

When Σ_{12} is of rank 1, the solution (up to sign jointly) of CCA problem is (θ, η) if and only if the covariance structure between X and Y can be written as

$$\Sigma_{12} = \lambda \Sigma_{11} \theta \eta^T \Sigma_{22}$$

where $0 < \lambda \leq 1$, $\theta^T \Sigma_{11} \theta = 1$ and $\eta^T \Sigma_{22} \eta = 1$. In other words, the correlation between $a^T X$ and $b^T Y$ are maximized by $\mathrm{corr}(\theta^T X, \eta^T Y)$, and λ is the canonical correlation between X and Y.

Sparse CCA via Precision Adjusted Iterative Thresholding

Proposition 2.

For general Σ_{12} with rank $r \geq 1$, the solution (up to sign jointly) of CCA problem is (θ_1, η_1) if and only if the covariance structure between X and Y can be written as

$$\Sigma_{12} = \lambda \Sigma_{11} \left(\sum_{i=1}^{r} \lambda_i \theta_i \eta_i^T \right) \Sigma_{22}$$

where $\lambda_1 > \lambda_2 > ... > \lambda_r > 0$, $\theta_i^T \Sigma_{11} \theta_j = \mathbb{I}(i = j) = \eta_i^T \Sigma_{22} \eta_j$.

Sparse CCA via Precision Adjusted Iterative Thresholding

- We propose a probabilistic model of (X, Y), so that the canonical directions (θ, η) are explicitly modeled in the joint distribution of (X, Y).

The Single Canonical Pair Model

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \lambda\Sigma_{11}\theta\eta^T\Sigma_{22} \\ \lambda\Sigma_{22}\eta\theta^T\Sigma_{11} & \Sigma_{22} \end{pmatrix} \right)$$

with $\Sigma_{11} > 0, \Sigma_{22} > 0, \theta^T\Sigma_{11}\theta = \eta^T\Sigma_{22}\eta = 1$ and $0 < \lambda \leq 1$.

Sparse CCA via Precision Adjusted Iterative Thresholding

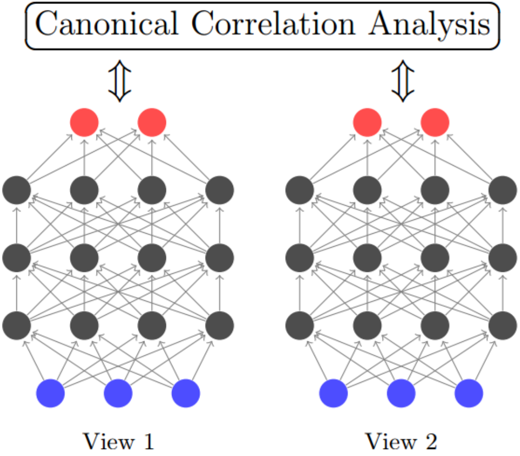
Algorithm : CAPIT

Input : Sample covariance matrices $\hat{\Sigma}_{12}$;
Estimators of precision matrix $\hat{\Omega}_{11}, \hat{\Omega}_{22}$;
Initialization pair $\alpha^{(0)}, \beta^{(0)}$;
Thresholding level γ_1, γ_2.
Output : Canonical direction estimator $\alpha^{(\infty)}, \beta^{(\infty)}$.
Set $\hat{A} = \hat{\Omega}_{11}\hat{\Sigma}_{12}\hat{\Omega}_{22}$;
repeat

- Right Multiplication: $\omega^{l,(i)} = \hat{A}\beta^{(i-1)}$;
- Left Thresholding : $\omega_{th}^{l,(i)} = T(\omega^{l,(i)}, \gamma_1)$;
- Left Normalization : $\alpha^{(i)} = \omega_{th}^{l,(i)} / \|\omega_{th}^{l,(i)}\|$;
- Left Multiplication : $\omega^{r,(i)} = \alpha^{(i)}\hat{A}$;
- Right Thresholding : $\omega_{th}^{r,(i)} = T(\omega_{th}^{r,(i)}, \gamma_2)$;
- Right Normalization : $\beta^{(i)} = \omega_{th}^{r,(i)} / \|\omega_{th}^{r,(i)}\|$;

until Convergence of $\alpha^{(i)}$ and $\beta^{(i)}$.

Deep Canonical Correlation Analysis (Andrew et al., 2013)



Deep Canonical Correlation Analysis

- If θ_1 is the vector of all parameters of the first view, and similarly for θ_2, then

$$(\theta_1^*, \theta_2^*) = \underset{(\theta_1,\theta_2)}{\operatorname{argmax}} \operatorname{corr}(f(X;\theta_1), g(Y;\theta_2))$$

- $H \in \mathbb{R}^{n\times o}, K \in \mathbb{R}^{n\times o}$: data matrices with top-level representation.
- $\bar{H}$, $\bar{K}$: centered data matrices.
- Define $\hat{\Sigma}_{12} = \frac{1}{n-1}\bar{H}'\bar{K}$ and $\hat{\Sigma}_{11} = \frac{1}{n-1}\bar{H}'\bar{H} + r_1 I$ (resp. $\hat{\Sigma}_{22}$).
- If we take $k = o$, then

$$\operatorname{corr}(H, K) = \|T\|_{tr} = \operatorname{tr}(T'T)^{1/2}$$

where $T = \hat{\Sigma}_{11}^{-1/2}\hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1/2}$.

Deep Canonical Correlation Analysis

- Optimizing this quantity using gradient-based optimization.
- If the singular decomposition of T is $T = UDV'$ then,

$$\frac{\partial \mathrm{corr}(H,K)}{\partial H} = \frac{1}{n-1}(2\Delta_{11}\bar{H} + \Delta_{12}\bar{K}).$$

where

$$\Delta_{12} = \hat{\Sigma}_{11}^{-1/2} U V' \hat{\Sigma}_{22}^{-1/2}$$

and

$$\Delta_{11} = -\frac{1}{2}\hat{\Sigma}_{11}^{-1/2} U D U' \hat{\Sigma}_{11}^{-1/2}$$

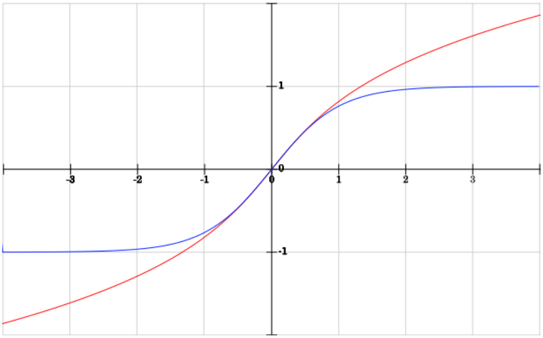
and $\partial \mathrm{corr}(H,K)/\partial K$ has a symmetric expression.

Deep Canonical Correlation Analysis

- The correlation objective is a function of the entire training set that does not decompose into a sum over data points.
- Full-match optimization using the L-BFGS second-order optimization method.
- Pre-training : denoising autoencoder
- Non-linear function : a novel non-saturating sigmoid function based on the cube root.

Deep Canonical Correlation Analysis

- If $g : \mathbb{R} \to \mathbb{R}$ is the function $g(y) = y^3/3 + y$, then our function is $s(x) = g^{-1}(x)$.



Deep Canonical Correlation Analysis

- s is not bounded, and its derivative falls off much more gradually with x.
- We hypothesize that these properties make s better-suited for batch optimization with second-order methods.
- The derivative of s is a simple function of its value.
  - $s'(x) = (s^2(x) + 1)^{-1}$.
- To compute $s(x)$, we use Newton's method.