Canonical Correlation Analysis (CCA)

Ke Chen

Department of Computer Science, The University of Manchester

Ke.Chen@manchester.ac.uk

Outline

Background

History, motivation, task and scope of applications

Principle

CCA Derivation: learning maximum correlation between two random vectors

Algorithm

CCA algorithms and Proportion of Variance (PoV)

Illustrative Example

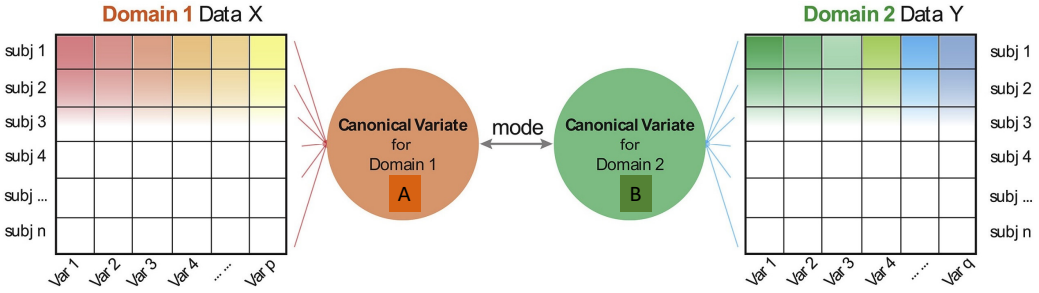
CCA on synthetic and real datasets

Extension

Sparse CCA, kernel CCA, Deep CCA and Probabilistic CCA

- Concept originated by Camille Jordan in 1875, independently developed and name coined by Harold Hotelling in 1930s
- An approach to exploratory data analysis especially for two relevant random vectors, e.g., visual and audio channels in video
- An important linear model for contemporary representation learning
- Applied in economics, medicine, meteorology, neuroscience, pattern recognition, psychology and many other fields
- In context of ML and statistics, widely used for feature extraction and data interpretation

- Find optimal linear projections for two relevant random vectors to maximise their correlation in the new spaces
- New low-dimensional representations of two random vectors achieved by projecting data points to new CCA spaces



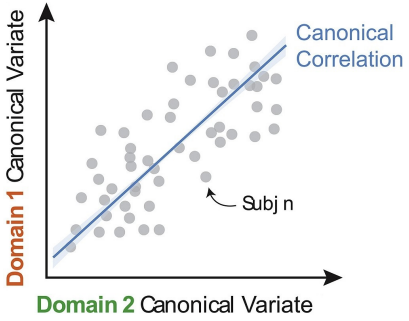
- Find optimal linear projections for two relevant random vectors to maximise their correlation in the new spaces
- New low-dimensional representations of two random vectors achieved by projecting data points to new CCA spaces

For $X_{p\times N} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$ and $Y_{q\times N} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N\}$, find $\boldsymbol{a}_{p\times 1}$ and $\boldsymbol{b}_{q\times 1}$:

$$\text{Domain 1:} \quad z_{X,n} = \boldsymbol{x}_n^T \boldsymbol{a} \quad n = 1, 2, \cdots N,$$
$$\text{Domain 2:} \quad z_{Y,n} = \boldsymbol{y}_n^T \boldsymbol{b} \quad n = 1, 2, \cdots N.$$

i.e., $(\boldsymbol{z}_X, \boldsymbol{z}_Y) = (X^T \boldsymbol{a}, Y^T \boldsymbol{b})$ such that

maximise correlation$(\boldsymbol{z}_X, \boldsymbol{z}_Y)$



Find out the 1st pair of canonical vectors, a and b

Given a dataset of N data points, (X, Y), $X_{p\times N} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$ and $Y_{q\times N} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N\}$ are centralised, respectively.

- To find out the 1st pair of canonical vectors, $\boldsymbol{a}$ and $\boldsymbol{b}$, of maximum correlation, we need to establish a utility function by encoding this learning goal.
- Projecting X and Y onto $\boldsymbol{a}$ and $\boldsymbol{b}$ leads to $\boldsymbol{z}_X = X^T\boldsymbol{a}$ and $\boldsymbol{z}_Y = Y^T\boldsymbol{b}$.
- The correlation between two vectors, $\boldsymbol{z}_X$ and $\boldsymbol{z}_Y$, is

$$\text{correlation}(\boldsymbol{z}_X, \boldsymbol{z}_Y) = \frac{\boldsymbol{z}_X^T\boldsymbol{z}_Y}{||\boldsymbol{z}_X||||\boldsymbol{z}_Y||} = \frac{\boldsymbol{a}^T XY^T\boldsymbol{b}}{\sqrt{\boldsymbol{a}^T XX^T\boldsymbol{a}}\sqrt{\boldsymbol{b}^T YY^T\boldsymbol{b}}} = \frac{\boldsymbol{a}^T S_{XY}\boldsymbol{b}}{\sqrt{\boldsymbol{a}^T S_{XX}\boldsymbol{a}}\sqrt{\boldsymbol{b}^T S_{YY}\boldsymbol{b}}},$$

where S_{XX}, S_{YY} and S_{XY} are covariance matrices on X, Y and XY (defined on both X and Y with size $p \times q$), respectively; i.e.

$$S_{XX} \equiv \frac{1}{N-1}XX^T,\ S_{YY} \equiv \frac{1}{N-1}YY^T,\ S_{XY} \equiv \frac{1}{N-1}XY^T.$$

Find out the 1st pair of canonical vectors, $\boldsymbol{a}$ and $\boldsymbol{b}$

- The denominators in $\text{correlation}(\boldsymbol{z}_X, \boldsymbol{z}_Y)$ simply play a role of normalisation to make the correlation invariant with respect to rescaling $\boldsymbol{a}$ and $\boldsymbol{b}$.

$$\text{correlation}(\boldsymbol{z}_X, \boldsymbol{z}_Y) = \frac{\boldsymbol{z}_X^T \boldsymbol{z}_Y}{||\boldsymbol{z}_X|| ||\boldsymbol{z}_Y||} = \frac{\boldsymbol{a}^T S_{XY} \boldsymbol{b}}{\sqrt{\boldsymbol{a}^T S_{XX} \boldsymbol{a}} \sqrt{\boldsymbol{b}^T S_{YY} \boldsymbol{b}}}$$

- Thus, the learning objective to find out the 1st pair of $\boldsymbol{a}$ and $\boldsymbol{b}$ for maximum correlation is formulated along with two normalisation constraints:

$$\boldsymbol{z}_X^T \boldsymbol{z}_Y \ \text{ s.t. } ||\boldsymbol{z}_X||^2 = 1, \ ||\boldsymbol{z}_Y||^2 = 1 \implies \boldsymbol{a}^T S_{XY} \boldsymbol{b} \ \text{ s.t. } \boldsymbol{a}^T S_{XX} \boldsymbol{a} = 1, \ \boldsymbol{b}^T S_{YY} \boldsymbol{b} = 1.$$

- Apply the Lagrangian multipliers to form the unconstrained utility function:

$$\mathcal{L}(\boldsymbol{a}, \boldsymbol{b}, \lambda_a, \lambda_b; X, Y) = \boldsymbol{a}^T S_{XY} \boldsymbol{b} + \frac{\lambda_a}{2}(1 - \boldsymbol{a}^T S_{XX} \boldsymbol{a}) + \frac{\lambda_b}{2}(1 - \boldsymbol{b}^T S_{YY} \boldsymbol{b}).$$

Find out the 1st pair of canonical vectors, a and b

- To find out the optimal $\boldsymbol{a}$ and $\boldsymbol{b}$, apply the optimality conditions:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{a}} = 0 \implies S_{XY}\boldsymbol{b} - \lambda_a S_{XX}\boldsymbol{a} = 0 \implies S_{XY}\boldsymbol{b} = \lambda_a S_{XX}\boldsymbol{a}, \quad (1)$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}} = 0 \implies S_{YX}\boldsymbol{a} - \lambda_b S_{YY}\boldsymbol{b} = 0 \implies S_{YX}\boldsymbol{a} = \lambda_b S_{YY}\boldsymbol{b}, \quad (2)$$

- Multiplying $\boldsymbol{a}^T$ and $\boldsymbol{b}^T$ on both sides in Eqs. (1) and (2) and further applying two constraints, $\boldsymbol{a}^T S_{XX}\boldsymbol{a} = 1,\ \boldsymbol{b}^T S_{YY}\boldsymbol{b} = 1$, we have

$$\boldsymbol{a}^T(S_{XY}\boldsymbol{b}) = \lambda_a \boldsymbol{a}^T S_{XX}\boldsymbol{a} = \lambda_a, \quad \boldsymbol{b}^T(S_{YX}\boldsymbol{a}) = \lambda_b \boldsymbol{b}^T S_{YY}\boldsymbol{b} = \lambda_b.$$

- $\boldsymbol{a}^T S_{XY}\boldsymbol{b}$ and $\boldsymbol{b}^T S_{YX}\boldsymbol{a}$ are scalar, and $(\boldsymbol{a}^T S_{XY}\boldsymbol{b})^T = \boldsymbol{b}^T S_{YX}\boldsymbol{a}$. Therefore, we must have $\boxed{\lambda_a = \lambda_b}$ at the optimum. $\lambda_a = \lambda_b = \lambda$ is the canonical correlation coefficient.

Find out the 1st pair of canonical vectors, $\boldsymbol{a}$ and $\boldsymbol{b}$

- By using the common λ, we rewrite Eqs. (1) and (2) as follows:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{a}} = 0 \implies S_{XY}\boldsymbol{b} = \lambda S_{XX}\boldsymbol{a}, \tag{3}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}} = 0 \implies S_{YX}\boldsymbol{a} = \lambda S_{YY}\boldsymbol{b}, \tag{4}$$

- Assume S_{XX} and S_{YY} are invertible, we rewrite Eq.(4) as $\boldsymbol{b} = \frac{1}{\lambda} S_{YY}^{-1} S_{YX}\boldsymbol{a}$ and insert it into Eq.(3):

$$S_{XY}S_{YY}^{-1}S_{YX}\boldsymbol{a} = \lambda^2 S_{XX}\boldsymbol{a} \implies \boxed{S_{XX}^{-1}S_{XY}S_{YY}^{-1}S_{YX}\boldsymbol{a} = \lambda^2\boldsymbol{a}}. \tag{5}$$

- With eigen-analysis on Eq.(5), choose $\boldsymbol{a}^*$ with the largest λ^* and $\boldsymbol{b}^* = \frac{1}{\lambda^*} S_{YY}^{-1} S_{YX}\boldsymbol{a}^*$.
- The choice of $\boldsymbol{a}^*$ and $\boldsymbol{b}^*$ leads to the **the 1st pair of canonical vectors, $\boldsymbol{a}$ and $\boldsymbol{b}$**.

CCA Formulation

- Similarly, we can find out more pairs of canonical vectors with the same idea, e.g., the learning objectives of the 2nd pair of $\boldsymbol{a}$ and $\boldsymbol{b}$ as follows:

$$\boldsymbol{a}^T S_{XY} \boldsymbol{b} \ \text{ s.t. } \boldsymbol{a}^T S_{XX} \boldsymbol{a} = 1, \ \boldsymbol{b}^T S_{YY} \boldsymbol{b} = 1, \ \boldsymbol{a}^T \boldsymbol{a}^* = 0, \ \boldsymbol{b}^T \boldsymbol{b}^* = 0.$$

- In general, M ($M \leq \min(p, q)$) pairs of canonical vectors, $A = \{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_M\}$ and $B = \{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_M\}$, are achieved by doing the eigen analysis on the composite matrix, $S_{XX}^{-1} S_{XY} S_{YY}^{-1} S_{YX}$, to form $A_{p \times M}$ with the eigen vectors, $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_M$, of top M eigenvalues and $B_{q \times M}$ with $\boldsymbol{b}_i = \frac{1}{\lambda_i} S_{YY}^{-1} S_{YX} \boldsymbol{a}_i, \ i = 1, \cdots, M$
- **Property**: all M canonical vectors in A and B are uncorrelational, i.e., $\boldsymbol{a}_i^T \boldsymbol{a}_j \neq 0$ if $i = j$ and 0 otherwise, and $\boldsymbol{b}_i^T \boldsymbol{b}_j \neq 0$ if $i = j$ and 0 otherwise.
- For dataset $(X_{p \times N}, Y_{q \times N})$, we can use CCA projection matrices, $A_{p \times M} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_M\}$ and $B_{q \times M} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_M\}$, to achieve the M-dimensional representations:

$$(\boldsymbol{z}_{X,i}, \boldsymbol{z}_{Y,i}) = (X^T \boldsymbol{a}_i, Y^T \boldsymbol{b}_i) \ \ i = 1, 2, \cdots, M.$$

Training Phase: Find out CCA projection matrices

- **Data centralisation**
  For a dataset, $X_{p\times N} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$ ($p < N$) and $Y_{q\times N} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N\}$ ($q < N$), calculate $\boldsymbol{m}_X = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n$ and $\boldsymbol{m}_Y = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{y}_n$ to produce $\bar{X}_{p\times N} = \{\boldsymbol{m}_X, \boldsymbol{m}_X, \cdots, \boldsymbol{m}_X\}$ and $\bar{Y}_{q\times N} = \{\boldsymbol{m}_Y, \boldsymbol{m}_Y, \cdots, \boldsymbol{m}_Y\}$. Data centralisation: $\hat{X} = X - \bar{X}$ and $\hat{Y} = Y - \bar{Y}$.
- **Eigen analysis**
  Calculate covariance matrices: $S_{XX} = \frac{1}{N-1}\hat{X}\hat{X}^T$, $S_{YY} = \frac{1}{N-1}\hat{Y}\hat{Y}^T$, $S_{XY} = \frac{1}{N-1}\hat{X}\hat{Y}^T$.
  For $S_{XX}^{-1}S_{XY}S_{YY}^{-1}S_{YX}$, find out and rank all p eigenvalues so that $\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_p^2$ with their corresponding eigenvectors, $\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_p$. Then, generate paired eigenvectors via $\boldsymbol{b}_i = \frac{1}{\lambda_i}S_{YY}^{-1}S_{YX}\boldsymbol{a}_i,\ i = 1, 2, \cdots, \min(p, q)$.
- **Construct projection matrices**
  Select top M ($M \leq \min(p, q)$) eigenvectors to form the paired projection matrices: $A_{p\times M} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_M\}$ and $B_{q\times M} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_M\}$

Deployment Phase: Generate low-dimensional representations in CCA space

$$(\boldsymbol{x}, \boldsymbol{y}) \implies (\boldsymbol{z}_x, \boldsymbol{z}_y), \quad \boldsymbol{z}_x = A^T(\boldsymbol{x} - \boldsymbol{m}_X) \text{ and } \boldsymbol{z}_y = B^T(\boldsymbol{y} - \boldsymbol{m}_Y).$$

Algorithm: SVD-based CCA

Training Phase: Find out CCA projection matrices

- **Data centralisation**
  For a given dataset, $X_{p\times N} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$ and $Y_{q\times N} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N\}$, calculate $\boldsymbol{m}_X = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n$ and $\boldsymbol{m}_Y = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{y}_n$ to produce $\bar{X}_{p\times N} = \{\boldsymbol{m}_X, \boldsymbol{m}_X, \cdots, \boldsymbol{m}_X\}$ and $\bar{Y}_{q\times N} = \{\boldsymbol{m}_Y, \boldsymbol{m}_Y, \cdots, \boldsymbol{m}_Y\}$. Data centralisation: $\hat{X} = X - \bar{X}$ and $\hat{Y} = Y - \bar{Y}$.
- **SVD solution to eigen analysis**
  Calculate covariance matrices: $S_{XX} = \frac{1}{N-1}\hat{X}\hat{X}^T$, $S_{YY} = \frac{1}{N-1}\hat{Y}\hat{Y}^T$, $S_{XY} = \frac{1}{N-1}\hat{X}\hat{Y}^T$.
  Apply SVD so as to $S_{XX}^{-\frac{1}{2}} S_{XY} S_{YY}^{-\frac{1}{2}} = U_{p\times p}\Sigma_{p\times q}V_{q\times q}^T$, where $U_{p\times p} = \{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_p\}$ and $V_{q\times q} = \{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_q\}$.
- **Construct projection matrices**
  Select top M ($M \leq \min(p, q)$) eigenvectors to form the paired projection matrices:
  $A_{p\times M} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_M\}$ and $B_{q\times M} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_M\}$, where $\boldsymbol{a}_i = S_{XX}^{-\frac{1}{2}}\boldsymbol{u}_i$, $\boldsymbol{b}_i = S_{XX}^{-\frac{1}{2}}\boldsymbol{v}_i$

Deployment Phase: Generate low-dimensional representations in CCA space

$$(\boldsymbol{x}, \boldsymbol{y}) \implies (\boldsymbol{z}_x, \boldsymbol{z}_y), \quad \boldsymbol{z}_x = A^T(\boldsymbol{x} - \boldsymbol{m}_X) \text{ and } \boldsymbol{z}_y = B^T(\boldsymbol{y} - \boldsymbol{m}_Y).$$

Fact: `sklearn.cross_decomposition.CCA` in the `scikit-learn` library.

How to find out a proper dimension, M^*, to form CCA space?

The shared eigenvalues between A and B can still be used in PoV.

- **Proportion of Variance (PoV)**

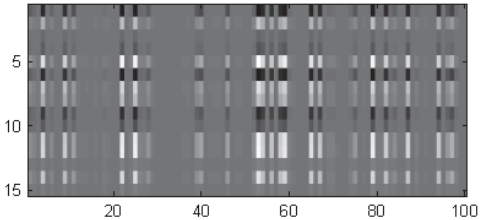
$$\text{PoV}(k) = \frac{\sum_{i=1}^{k} \lambda_i^2}{\sum_{i=1}^{d} \lambda_i^2} = \frac{\lambda_1^2 + \cdots + \lambda_k^2}{\lambda_1^2 + \lambda_2^2 + \cdots + \lambda_{\min(p,q)}^2}, \quad \text{for } k < \min(p, q).$$

- In practice, find out k^* so that $\text{PoV}(k^*) \geq 90\%$. Then, set $M^* = k^*$.
- In CCA, $\text{PoV}(k)$ works up to $\min(p, q)$ non-zero eigenvalues achieved in the squared form with the basic or the SVD-based CCA algorithms.

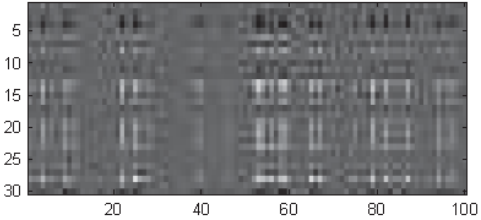
Example 1: Synthetic Data

- Dataset: generated by $X_{15\times100} = \boldsymbol{a}_{15\times1}\boldsymbol{h}^T_{100\times1}$ and $Y_{30\times100} = \boldsymbol{b}_{30\times1}\boldsymbol{h}^T_{100\times1}$. $\boldsymbol{a}, \boldsymbol{b}$ and $\boldsymbol{h}$ are randomly chosen vectors.

$X_{15\times100}$

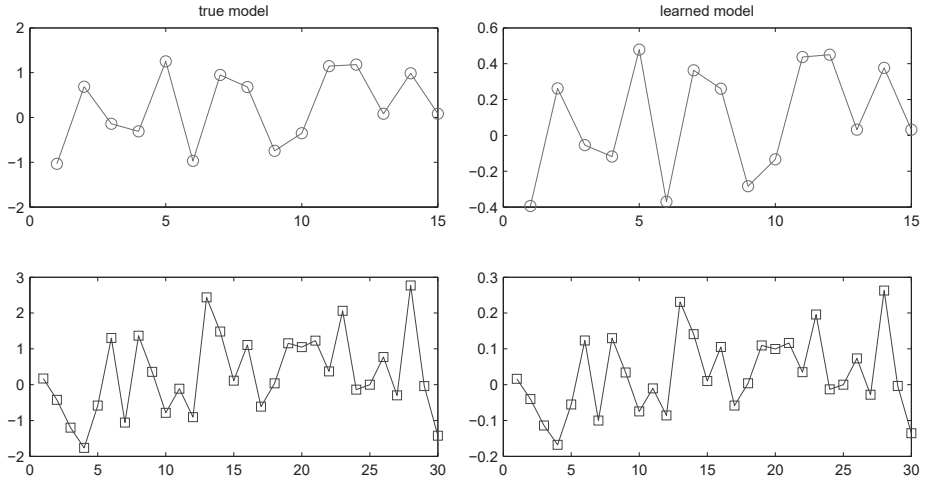


$Y_{30\times100}$



Example 1: Synthetic Data

- Dataset: generated by $X_{15\times100} = \boldsymbol{a}_{15\times1}\boldsymbol{h}^T_{100\times1}$ and $Y_{30\times100} = \boldsymbol{b}_{30\times1}\boldsymbol{h}^T_{100\times1}$.
- Apply CCA to the dataset to learn $\boldsymbol{a}$ and $\boldsymbol{b}$ in contrast to ground-truth



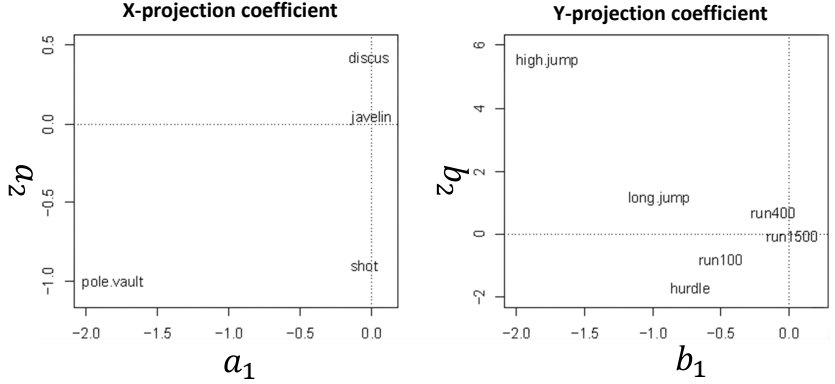
Example 2: Olympic Decathlon Data

- Dataset: men's 1988 Olympic decathlon (34 athletes, 10 events)
- Split the events into two subsets: arm (X) versus leg (Y) events:
  X: {shot, discus, javelin, pole.vault},
  Y: {run100, run400, run1500, hurdle, long.jump, high.jump}
- Apply CCA to achieve low-dimensional representations for data interpretation
- To make interpretation consistent for all events, change sign of all the running time ($+ \to -$) to indicate the larger the the better for all 10 events consistently

	run100	long.jump	shot	high.jump	run400	hurdle	discus	pole.vault	javelin	run1500	score
Schenk	-11.25	7.43	15.48	2.27	-48.90	-15.13	49.28	4.7	61.32	-268.95	8488
Voss	-10.87	7.45	14.97	1.97	-47.71	-14.46	44.36	5.1	61.76	-273.02	8399
Steen	-11.18	7.44	14.20	1.97	-48.29	-14.81	43.66	5.2	64.16	-263.20	8328
Thompson	-10.62	7.38	15.02	2.03	-49.06	-14.72	44.80	4.9	64.04	-285.11	8306
Blondel	-11.02	7.43	12.92	1.97	-47.44	-14.40	41.20	5.2	57.46	-256.64	8286
Plaziat	-10.83	7.72	13.58	2.12	-48.34	-14.18	43.06	4.9	52.18	-274.07	8272
Bright	-11.18	7.05	14.12	2.06	-49.34	-14.39	41.68	5.7	61.60	-291.20	8216
De.Wit	-11.05	6.95	15.34	2.00	-48.21	-14.36	41.32	4.8	63.00	-265.86	8189
Johnson	-11.15	7.12	14.52	2.03	-49.15	-14.66	42.36	4.9	66.46	-269.62	8180

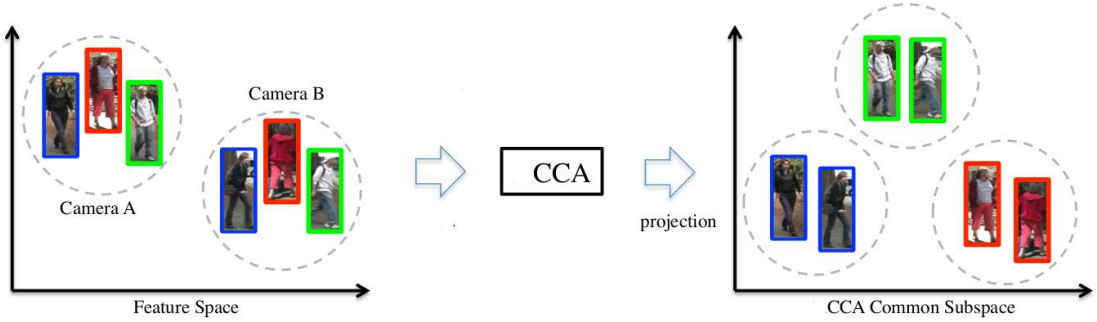
Example 2: Olympic Decathlon Data

- Dataset: men's 1988 Olympic decathlon (34 athletes, 10 events)
- Split the events into two subsets: arm (X) versus leg (Y) events:
  X ={shot, discus, javelin, pole.vault},
  Y ={run100, run400, run1500, hurdle, long.jump, high.jump}
- Projection matrices for $M = 2$, $A_{4\times2}$ and $B_{6\times2}$

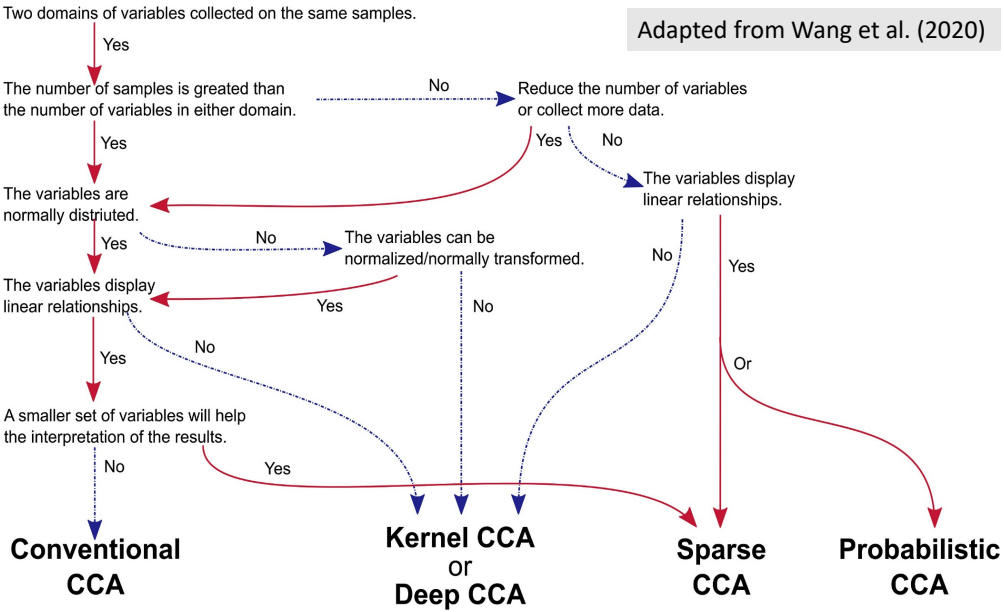


Example 3: Multi-view Alignment in Computer Vision

- Dataset: images captured by two cameras from different angles
- CCA leads to the common latent space that "groups" the correlated "objects".



Extension



Adapted from Wang et al. (2020)

Reference

If you want to deepen your understanding and learn something beyond this lecture, you can self-study the optional references below.

[Alpaydin, 2014] Alpaydin E. (2014): *Introduction to Machine Learning* (3rd Ed.), MIT Press. (Section 6.9)

[Barber, 2012] Barber D. (2012): *Bayesian Reasoning and Machine Learning*, Cambridge University Press. (Sections 15.8)

[Uurtio et al., 2017] Uurtio V. et al. (2017): A tutorial on canonical correlation methods. *ACM Computing Survey*, Vol. 50, No. 6, Article 95.