Dimensionality reduction and matrix factorization

Makoto Yamada
myamada@i.kyoto-u.ac.jp

Kyoto University

June/20/2018

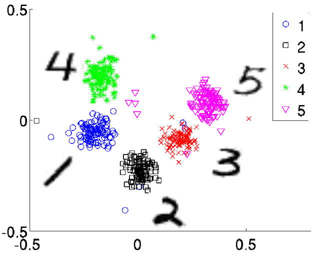
Review of the last lecture (Feature selection and sparsity)

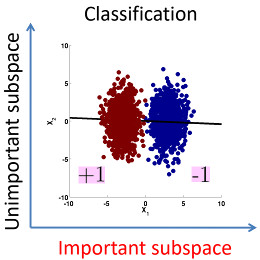
- Feature selection: Wrapper method, Filter method, and Embedded method
- Wrapper method (Selecting features that maximize prediction accuracy. Computationally expensive.)
- Filter method (Use mutual information to select features, e.g., MR, mRMR, QPFS, etc.)
- Embedded method (Selecting features during training. e.g., Lasso)
- Alternating Direction Method of Multipliers (ADMM).
- Advanced method: HSIC Lasso

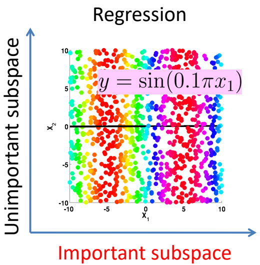
Dimensionality and Feature selection

Dimensionality reduction is a method to reduce the dimensionality of data.

- Feature selection is a dimensionality reduction method. Select a set of r features among d features ($r < d$).
- We tend to use feature selection if we want to interpret features.
- In dimensionality reduction, we may not need to interpret each feature.
- We tend to use dimensionality reduction to compress data, to visualize data, etc.







Multivariate Regression

Multivariate regression is a regression problem to predict multiple output variables $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^m \ \ (m > 1)$. If $m = 1$, it is a uni-variate regression.

Training dataset $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$

- $\boldsymbol{x}_i \in \mathbb{R}^d$: feature vector
- $\boldsymbol{y}_i \in \mathbb{R}^m$: real-valued target vector

Multivariate linear regression model:

$$\boldsymbol{y} = \boldsymbol{W}^\top \boldsymbol{x},$$

where $\boldsymbol{W} \in \mathbb{R}^{d \times m}$ and $^\top$ is matrix transpose.

Solution of the multivariate regression

The optimization problem can be written as

$$\min_{\boldsymbol{W} \in \mathbb{R}^{d \times m}} \quad \sum_{i=1}^{n} \|\boldsymbol{y}_i - \boldsymbol{W}^\top \boldsymbol{x}_i\|_2^2,$$

where $\|\boldsymbol{x}\|_2 = \sqrt{\sum_{k=1}^{d} (x^{(k)})^2}$ is the ℓ_2 norm.

Let us denote $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n] \in \mathbb{R}^{m \times n}$ and $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$. Then, the optimization problem can be written as

$$\min_{\boldsymbol{W}} \|\boldsymbol{Y} - \boldsymbol{W}^\top \boldsymbol{X}\|_F^2,$$

where $\|\boldsymbol{W}\|_F^2 = \sum_{(i,j)} [\boldsymbol{W}]_{ij}^2$: Frobenius norm.

The solution is given as

$$\widehat{\boldsymbol{W}} = (\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}\boldsymbol{Y}.$$

We can use $\frac{\partial \mathrm{tr}(\boldsymbol{A}\boldsymbol{B})}{\partial \boldsymbol{A}} = \boldsymbol{B}^\top$

Reduced rank regression

Multivariate regression is equivalent to m-independent univariate regressions. exploits no shared information.

Low-rank assumption

$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{V}^\top$$

$\boldsymbol{U} \in \mathbb{R}^{d \times k}, \boldsymbol{V} \in \mathbb{R}^{m \times k}$ (i.e., rank of $\boldsymbol{W}$ is K and $k < \min(d, m)$) m output variables share K-dimensional latent space

Reduced rank regression: Sparsity in the dim of latent space

$$\min_{\boldsymbol{W}} \quad \|\boldsymbol{Y} - \boldsymbol{W}^\top \boldsymbol{X}\|_F^2$$
$$\text{s.t.} \quad \text{rank}(\boldsymbol{W}) \leq k.$$

Sparsity in reduced rank regression

Parameter $\boldsymbol{W}$ in the reduced rank regression $\boldsymbol{y} = \boldsymbol{W}^\top \boldsymbol{x}$ is dense in terms of matrix elements.

$\boldsymbol{W}$ is sparse in terms of singular values $\rightarrow$ $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{V}^\top$ is low-rank.
$\boldsymbol{U} \in \mathbb{R}^{d \times k}$, $\boldsymbol{V} \in \mathbb{R}^{m \times k}$, $k < \min(d, m)$

Rank is the number of non-zero singular values. That is, Rank is the ℓ_0 norm of singular values:

$$\mathrm{rank}(\boldsymbol{W}) = \|\boldsymbol{\sigma}(\boldsymbol{W})\|_0,$$

where $\boldsymbol{\sigma}(\boldsymbol{W}) = [\sigma_1(\boldsymbol{W}), \sigma_2(\boldsymbol{W}), \ldots, \sigma_{\min(d,m)}(\boldsymbol{W})]^\top \in \mathbb{R}^{\min(d,m)}$ and $\sigma_i(\boldsymbol{W})$ is the i-th singular value of $\boldsymbol{W}$.

$$\|\boldsymbol{\sigma}\|_0 = \sum_{\ell=1}^{d} \delta(\sigma_i), \quad \delta(\sigma) = \begin{cases} 1 & (\sigma \neq 0) \\ 0 & (\sigma = 0) \end{cases}$$

Review of eigenvalue decomposition of symmetric matrix

Symmetric matrix can be diagonalized using an orthogonal matrix.

The eigenvalue decomposition of symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$:

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top$$

- $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d) \in \mathbb{R}^{d \times d}$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. If $\boldsymbol{A}$ is a positive definite matrix $\lambda_d \geq 0$.
- $\boldsymbol{P} \in \mathbb{R}^{d \times d}$ is an orthogonal matrix $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}_d$
- $\mathrm{tr}(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top) = \mathrm{tr}(\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{\Lambda}) = \sum_{i=1}^{d} \lambda_i$.

Review of Singular value decomposition (SVD)

A matrix $\boldsymbol{B} \in \mathbb{R}^{m \times n}$ can be decomposed by

$$\boldsymbol{B} = \boldsymbol{U}\bar{\boldsymbol{\Sigma}}\boldsymbol{V}^\top$$

where $\boldsymbol{U} \in \mathbb{R}^{m \times m}$, $\boldsymbol{V} \in \mathbb{R}^{n \times n}$,

$$\bar{\boldsymbol{\Sigma}} = \begin{cases} \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{O} \end{bmatrix} & (m < n) \\ \boldsymbol{\Sigma} & (m = n) \\ \begin{bmatrix} \boldsymbol{\Sigma} \\ \boldsymbol{O} \end{bmatrix} & (m > n) \end{cases}, \quad \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_q) \in \mathbb{R}^{q \times q},$$

and $q = \min(m, n)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_q \geq 0$ are singular values.

- $\boldsymbol{U}$ is the eigenvectors of $\boldsymbol{B}\boldsymbol{B}^\top$ and $\boldsymbol{V}$ is the eigenvectors of $\boldsymbol{B}^\top\boldsymbol{B}$, respectively.
- $\bar{\boldsymbol{\Sigma}}\bar{\boldsymbol{\Sigma}}^\top = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_q^2, 0, \ldots, 0) \in \mathbb{R}^{m \times m}$

Best rank-K approximation

Best rank-k approximation problem of matrix $\boldsymbol{B}$:

$$\min_{\widehat{\boldsymbol{B}}} \ \|\boldsymbol{B} - \widehat{\boldsymbol{B}}\|_F^2, \ \text{s.t. } \operatorname{rank}(\widehat{\boldsymbol{B}}) \leq k$$

The optimal solution is given as

$$\widehat{\boldsymbol{B}} = \boldsymbol{U}_k \boldsymbol{\Sigma}_k \boldsymbol{V}_k^\top.$$

If $\widehat{\boldsymbol{B}} = \boldsymbol{O}$, we have

$$\|\boldsymbol{B}\|_F^2 = \sum_{i=1}^{\min(m,n)} \sigma_i^2$$

Since $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_q \geq 0$, the rank-$k$ solution that minimizes the loss function is given as

$$\|\boldsymbol{B} - \widehat{\boldsymbol{B}}\|_F^2 = \|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top - \boldsymbol{U}\boldsymbol{\Sigma}_k\boldsymbol{V}^\top\|_F^2 = \sum_{i=k+1}^{\min(m,q)} \sigma_i^2$$

Solution of reduced rank regression

The objective function to be minimized:

$$\begin{aligned}\|\boldsymbol{Y} - \boldsymbol{W}^\top \boldsymbol{X}\|_F^2 &= \mathrm{tr}[(\boldsymbol{Y} - \boldsymbol{W}^\top \boldsymbol{X})^\top (\boldsymbol{Y} - \boldsymbol{W}^\top \boldsymbol{X})] \\ &= \mathrm{tr}(\boldsymbol{Y}^\top \boldsymbol{Y} - 2\boldsymbol{Y}^\top \boldsymbol{W}^\top \boldsymbol{X} + \boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{X}) \\ &= \mathrm{tr}(\boldsymbol{Y}\boldsymbol{Y}^\top - 2\boldsymbol{W}^\top \boldsymbol{X}\boldsymbol{Y}^\top + \boldsymbol{W}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W}),\end{aligned}$$

where we used $\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA})$.
We further decompose $\boldsymbol{X}\boldsymbol{X}^\top$ as

$$\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top.$$

Let us denote $\widetilde{\boldsymbol{W}} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^\top \boldsymbol{W}$. Then, we have

$$\begin{aligned}\|\boldsymbol{Y} - \boldsymbol{W}^\top \boldsymbol{X}\|_F^2 &= \mathrm{tr}(\boldsymbol{Y}\boldsymbol{Y}^\top - 2\boldsymbol{W}^\top \boldsymbol{X}\boldsymbol{Y}^\top + \boldsymbol{W}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W}) \\ &= \mathrm{tr}(\boldsymbol{Y}\boldsymbol{Y}^\top - 2\widetilde{\boldsymbol{W}}^\top \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top \boldsymbol{X}\boldsymbol{Y}^\top + \widetilde{\boldsymbol{W}}^\top \widetilde{\boldsymbol{W}}) \\ &= \|\widetilde{\boldsymbol{W}} - \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top \boldsymbol{X}\boldsymbol{Y}^\top\|_F^2 + \mathrm{Const.}\end{aligned}$$

Dimensionality Reduction

Dimension reduction is to find a low-dimensional mapping $f : \mathbb{R}^d \to \mathbb{R}^k$ $(d > r)$.

- It is useful for data visualization, computational/space efficiency, etc.
- Compression: keep the original information as much as possible
- The feature selection selects a set of features, while the dimensionality reduction outputs the combination of features.

Typically, dimensionality reduction can be categorized as

- Linear dimension reduction $\boldsymbol{z} = \boldsymbol{U}^\top \boldsymbol{x}$ $(\boldsymbol{U} \in \mathbb{R}^{d \times m})$.

$$m\left\{\boldsymbol{z}\right. \quad = \quad m\left\{\underbrace{\boldsymbol{U}^\top}_{d}\right. \left.\boldsymbol{x}\right\}d$$

- Nonlinear dimension reduction $\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{x})$. For example, deep learning model: $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{\sigma}(\boldsymbol{W}_1(\boldsymbol{\sigma}(\boldsymbol{W}_2)))$

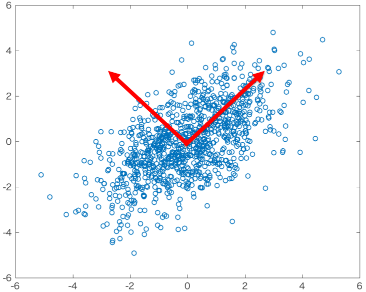
Dimensionality Reduction (Principal Component Analysis)

The principal component analysis (PCA) is a popular method:

$$\max_{\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}.} \quad \mathrm{tr}(\boldsymbol{U}^\top \boldsymbol{R} \boldsymbol{U}),$$

where $\boldsymbol{R} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \in \mathbb{R}^{d \times d}$ (we assume $\mathbb{E}[\boldsymbol{x}] = \mathbf{0}$) is the covariance matrix.

Idea: Find a direction that maximizes the variance



Relationship to the Linear Auto-encoder (1/2)

Assume that $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{0}$. Then, consider the following linear Auto-encoder problem:

$$\min_{\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}} \quad \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}_i\|_2^2,$$

The loss function term can be written as

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}_i\|_2^2 &= \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^\top \boldsymbol{x}_i - 2\boldsymbol{x}_i^\top \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}_i + \boldsymbol{x}_i^\top \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}_i \right) \\
&\propto -\frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^\top \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}_i \right) \quad (\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}) \\
&= -\frac{1}{n} \sum_{i=1}^{n} \left(\mathrm{tr}(\boldsymbol{U}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{U}) \right) \quad (\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{B}\boldsymbol{A})) \\
&= -\mathrm{tr}(\boldsymbol{U}^\top \boldsymbol{R} \boldsymbol{U}), \quad (\boldsymbol{R} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top)
\end{aligned}$$

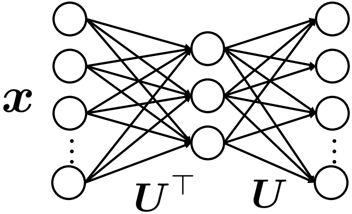
Relationship to the Linear Auto-encoder (2/2)

The minimization problem can be written as the maximization problem:

$$\min_{\boldsymbol{U}^\top \boldsymbol{U}=\boldsymbol{I}} \quad \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}_i\|_2^2, \leftrightarrow \quad \max_{\boldsymbol{U}^\top \boldsymbol{U}=\boldsymbol{I}} \quad \mathrm{tr}(\boldsymbol{U}^\top \boldsymbol{R}\boldsymbol{U})$$

Thus, PCA is related to the linear Auto-encoder.

Idea: Find a direction that maximizes the variance

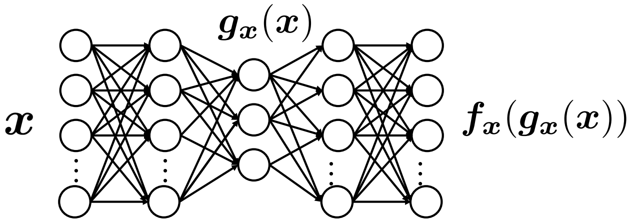


Nonlinear Auto-encoder (Deep auto-encoder)

We consider the following Auto-encoder problem:

$$\min_{\Theta} \quad \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{f}_{\boldsymbol{x}}(\boldsymbol{g}_{\boldsymbol{x}}(\boldsymbol{x}_i))\|_2^2,$$

Idea: Find a direction that maximizes the variance



Multi-modal Dimensionality Reduction Methods

PCA and auto-encoders are for uni-modal input (i.e., only image or only text).
How to do dimensionality reduction for multi-modal data (i.e., image and text)?
We have $(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{x} \in \mathbb{R}^{d_x}$ and $\boldsymbol{y} \in \mathbb{R}^{d_y}$.

- Linear dimension reduction $\boldsymbol{z}_x = \boldsymbol{U}^\top \boldsymbol{x}$ and $\boldsymbol{z}_y = \boldsymbol{V}^\top \boldsymbol{y}$. $\boldsymbol{U} \in \mathbb{R}^{d_x \times m}$ and $\boldsymbol{U} \in \mathbb{R}^{d_y \times m}$.
- Nonlinear dimension reduction $\boldsymbol{z}_x = \boldsymbol{g}_{\boldsymbol{x}}(\boldsymbol{x})$ and $\boldsymbol{z}_y = \boldsymbol{g}_{\boldsymbol{y}}(\boldsymbol{y})$.

Canonical Correlation Analysis (CCA) is to find dimensionality reduction that maximize the similarity between $\boldsymbol{z}_x = \boldsymbol{U}^\top \boldsymbol{x}$ and $\boldsymbol{z}_y = \boldsymbol{V}^\top \boldsymbol{y}$.

Let us assume that $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{0}$. We want to maximize the correlation:

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_{x,i}^\top \boldsymbol{z}_{y,i} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{U}\boldsymbol{V}^\top \boldsymbol{y}_i \\
&= \mathrm{tr}(\boldsymbol{U}^\top \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{y}_i^\top \boldsymbol{V}) \\
&= \mathrm{tr}(\boldsymbol{U}^\top \boldsymbol{R}_{xy} \boldsymbol{V})
\end{aligned}
$$

where $\boldsymbol{R}_{xy} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{y}_i^\top \in \mathbb{R}^{d_x \times d_y}$.

Canonical Correlation Analysis (CCA) (2/3)

The optimization problem of CCA is given as

$$\max_{\boldsymbol{U},\boldsymbol{V}} \quad \mathrm{tr}(\boldsymbol{U}^\top \boldsymbol{R}_{xy}\boldsymbol{V}),$$
$$\text{s.t.} \quad \boldsymbol{U}^\top \boldsymbol{R}_{xx}\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{V}^\top \boldsymbol{R}_{yy}\boldsymbol{V} = \boldsymbol{I},$$

where $\boldsymbol{R}_{xx} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top$ and $\boldsymbol{R}_{yy} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{y}_i\boldsymbol{y}_i^\top$.
Then, CCA can be written as

$$\max_{\boldsymbol{U},\boldsymbol{V}} \quad \mathrm{tr}\left(\begin{bmatrix} \boldsymbol{U}^\top & \boldsymbol{V}^\top \end{bmatrix}\begin{bmatrix} \boldsymbol{O} & \boldsymbol{R}_{xy} \\ \boldsymbol{R}_{xy}^\top & \boldsymbol{O} \end{bmatrix}\begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}\right),$$
$$\text{s.t.} \quad \begin{bmatrix} \boldsymbol{U}^\top & \boldsymbol{V}^\top \end{bmatrix}\begin{bmatrix} \boldsymbol{R}_{xx} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{R}_{yy} \end{bmatrix}\begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} = \boldsymbol{I},$$

This is a generalized eigenvalue decomposition (GEV) problem.

Canonical Correlation Analysis (CCA) (3/3)

Let us transform the variables as

$$\begin{bmatrix} \bar{\boldsymbol{U}} \\ \bar{\boldsymbol{V}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}_{xx}^{1/2} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{R}_{yy}^{1/2} \end{bmatrix} \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}$$

we can rewrite the CCA optimization problem as

$$\max_{\bar{\boldsymbol{U}}, \bar{\boldsymbol{V}}} \ \frac{1}{2}\mathrm{tr}\left(\begin{bmatrix} \bar{\boldsymbol{U}}^\top & \bar{\boldsymbol{V}}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{O} & \boldsymbol{R}_{xx}^{-1/2}\boldsymbol{R}_{xy}\boldsymbol{R}_{yy}^{-1/2} \\ (\boldsymbol{R}_{xx}^{-1/2}\boldsymbol{R}_{xy}\boldsymbol{R}_{yy}^{-1/2})^\top & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{U}} \\ \bar{\boldsymbol{V}} \end{bmatrix} \right)$$

$$\text{s.t.} \quad \begin{bmatrix} \bar{\boldsymbol{U}}^\top & \bar{\boldsymbol{V}}^\top \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{U}} \\ \bar{\boldsymbol{V}} \end{bmatrix} = \boldsymbol{I},$$

Thus, we can solve the CCA problem by using eigenvalue decomposition!

Summary of today's lecture

- Dimensionality reduction (Reduce the dimensionality of features).
- Feature selection is to interpret features, while dimensionality reduction is to reduce the dimensionality (for compression and visualization).
- Multi-variate Regression and reduced-rank regression
- Principal Component Analysis (PCA)
- Canonical Correlation Analysis (Multi-modal data)