Lecture 2. Sufficient Dimensionality Reduction: Supervised PCA, LDA, and SIR

Yuan Yao

Hong Kong University of Science and Technology

Outline

Sufficient Dimensionality Reduction
PCA as Sufficient Dimensionality Reduction

Supervised PCA
Linear Discriminant Analysis
Sliced Inverse Regression
Localized SIR

Sufficient Dimensionality Reduction

Definition (Cook 2005)

A **sufficient dimension reduction** Γ ($\Gamma \in \mathbb{R}^{p \times d}$, $\Gamma^T \Gamma = I_d$) refers to the setting that the conditional distribution of $Y|X$ is the same as the distribution of $Y|\Gamma^T X$ for all X, i.e.

$$\mathbb{P}(Y|X) = \mathbb{P}(Y|\Gamma^T X).$$

- Example: in regression $Y = f(X, \varepsilon)$, for some unknown function f, sufficient dimensionality reduction implies that $Y = f(\Gamma^T X, \varepsilon)$.

Sufficient Dimensionality Reduction

Definition (Cook 2005)

A **sufficient dimension reduction** Γ ($\Gamma \in \mathbb{R}^{p\times d}$, $\Gamma^T\Gamma = I_d$) refers to the setting that the conditional distribution of $Y|X$ is the same as the distribution of $Y|\Gamma^T X$ for all X, i.e.

$$\mathbb{P}(Y|X) = \mathbb{P}(Y|\Gamma^T X).$$

- Example: in regression $Y = f(X, \varepsilon)$, for some unknown function f, sufficient dimensionality reduction implies that $Y = f(\Gamma^T X, \varepsilon)$.
- Can you find Γ without knowing f?

Sufficient Dimensionality Reduction

Definition (Cook 2005)

A **sufficient dimension reduction** Γ ($\Gamma \in \mathbb{R}^{p \times d}$, $\Gamma^T \Gamma = I_d$) refers to the setting that the conditional distribution of $Y|X$ is the same as the distribution of $Y|\Gamma^T X$ for all X, i.e.

$$\mathbb{P}(Y|X) = \mathbb{P}(Y|\Gamma^T X).$$

- Example: in regression $Y = f(X, \varepsilon)$, for some unknown function f, sufficient dimensionality reduction implies that $Y = f(\Gamma^T X, \varepsilon)$.

- Can you find Γ without knowing f?

- Yes! Consider the inverse problem, with conditional distribution $\mathbb{P}(X|Y)$.

An Inverse Model

Example (Inverse model)

For each value in response variable y,

$$X_y = \mu + \Gamma \nu_y + \varepsilon \tag{1}$$

where

- $X_y \in \mathbb{R}^p$,
- $\nu_y \in \mathbb{R}^d$, $d < p$,
- $\Gamma \in \mathbb{R}^{p \times d}$ such that $\Gamma^T \Gamma = I_d$,
- $\varepsilon \sim N_p(0, \sigma^2 I_p)$,
- assume $\sum_y \nu_y = 0$ for removing the degree of freedom in translation.

Sufficient Dimensionality Reduction

Lemma (Cook 2005)

Under the inverse model, $\mathbb{P}(Y|X) = \mathbb{P}(Y|\Gamma^T X)$*, i.e.* Γ *is a sufficient dimensionality reduction.*

Proof

- First, $X|(Y=y) \sim N_p(\mu + \Gamma\nu_y, \sigma^2 I_p)$.

- By Bayesian formula, we have for any f

$$
\begin{aligned}
f_{Y|X}(y|x) &\propto f_{X|Y}(x|y) f_Y(y) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\|x-\mu-\Gamma\nu_y\|^2\right) f_Y(y) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}(\nu_y^T\nu_y - 2\nu_y^T\Gamma^T(x-\mu)\right) f_Y(y)
\end{aligned}
$$

where the last line is given by the orthogonality $\Gamma^T\Gamma = I$.

Proof (continued)

- Similarly, since $\Gamma^T X | (Y = y) \sim N_d(\Gamma^T \mu + \nu_y, \sigma^2 I_d)$, we have

$$\begin{aligned} f_{Y|\Gamma^T X}(y|\Gamma^T x) &\propto f_{\Gamma^T X|Y}(\Gamma^T x|y) f_Y(y) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\|\Gamma^T x - \Gamma^T \mu - \nu_y\|^2\right) f_Y(y) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(\nu_y^T \nu_y - 2\nu_y^T \Gamma^T (x-\mu))\right) f_Y(y) \end{aligned}$$

  by the orthogonality $\Gamma^T \Gamma = I$.

- Therefore, $\mathbb{P}(Y|X) = \mathbb{P}(Y|\Gamma^T X)$ of the same density kernels. $\square$

Estimate of Γ

- Can we estimate Γ from finite sample without knowing f?

Estimate of Γ

- Can we estimate Γ from finite sample without knowing f?
- PCA gives the Maximum Likelihood Estimate of Γ

Maximum Likelihood Estimate

- Under the inverse model, the conditional likelihood function

$$f(X_y|\mu, \Gamma, \nu_y) = \frac{1}{\sigma^p\sqrt{(2\pi)^p}} \exp\left[-\frac{1}{2\sigma^2}(X_y - \mu - \Gamma\nu_y)^T(X_y - \mu - \Gamma\nu_y)\right]$$

- MLE

$$\max_{\mu,\Gamma,\nu_y} \prod_y f(X_y|\mu, \Gamma, \nu_y)$$

$$\Leftrightarrow \max_{\mu,\Gamma,\nu_y} -\frac{1}{2\sigma^2}\sum_y \|X_y - \mu - \Gamma\nu_y\|^2 - \sum_y p\log\sigma + C.$$

Maximum Likelihood Estimate (continued)

- MLE solution

$$\widehat{\Gamma} = \arg\min_{\Gamma^T\Gamma = I} \sum_y \|X_y - \hat{\mu} - P_\Gamma(X_y - \hat{\mu})\|^2, \quad P_\Gamma = \Gamma\Gamma^T. \quad (2)$$

where $\widehat{\mu} = \frac{1}{n}\sum_y X_y$, $\nu_y = \widehat{\Gamma}^T(X_y - \hat{\mu})$.

- If y is of distinct values (e.g. the unknown f is injective), PCA (top d eigen-decomposition of $\widehat{\Sigma}$) gives $\widehat{\Gamma}$.

- If y is of discrete values (e.g. classification), discriminant analysis (eigen-decomposition of $\widehat{\Sigma}_B = \frac{1}{K}\sum_{y=1}^{K}(\hat{\mu}_y - \hat{\mu})(\hat{\mu}_y - \hat{\mu})^T$) gives $\widehat{\Gamma}$.

Maximum Likelihood Estimate (continued)

- In general
$$X_y = \mu + \Gamma \nu_y + \epsilon \tag{3}$$
where $\varepsilon \sim N_p(0, \Sigma)$, $\widehat{\mu}_y = \widehat{E}[X_y | y]$.

- Rescale $Z_y = \Sigma^{-1/2} X_y$.

- Eigen-decomposition of $\Sigma^{-1/2} \widehat{\Sigma}_B \Sigma^{-1/2}$ (with $\widehat{\Sigma}$ for the estimate of Σ) meets Fisher's Linear Discriminant Analysis for $\widehat{\Gamma}$.

- Therefore *PCA/LDA can be also derived as a sufficient dimensionality reduction in supervised learning, even the function f is unknown here*.

Outline

Sufficient Dimensionality Reduction
PCA as Sufficient Dimensionality Reduction

Supervised PCA
Linear Discriminant Analysis
Sliced Inverse Regression
Localized SIR

Linear Discriminant Analysis

- Data: $\{X_i, y_i\}_{i=1}^N$ where y_i is discrete in $\{1, 2, \ldots, K\}$ but not ordered

- Compute sample mean and within class means

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} X_i, \quad \hat{\mu}_k = \frac{1}{N_k} \sum_{y_i = k} X_i;$$

- Compute Between class covariance matrix

$$\widehat{\Sigma}_B^{p \times p} = \frac{1}{K} \sum_{k=1}^{K} (\hat{\mu}_k - \hat{\mu})(\hat{\mu}_k - \hat{\mu})^T;$$

- Compute Within class covariance matrix

$$\hat{\Sigma}_W^{p \times p} = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{y_i = k} (X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^T;$$

Fisher's Linear Discriminant Analysis

We choose the k-th class such that the following *linear* score function is the largest:

$$\hat{\delta}_k(x) = \hat{\mu}_k^T \hat{\Sigma}^{-1} x - \frac{1}{2}\hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu}_k + \log \hat{\pi}_k, \tag{4}$$

where given data $(x_i, y_i), i = 1, ..., n,$

- $\hat{\pi}_k = n_k/n$ is the sample proportion of class k where n_k is the number of subjects in class k
- $\hat{\mu}_k$ is the sample mean of class k

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i;$$

- $\hat{\Sigma}$ is the pooled (overall) sample covariance

$$\hat{\Sigma} = \widehat{\Sigma}_B + \widehat{\Sigma}_W = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T,$$

Fisher's LDA

- Fisher's LDA (1920s) aims to capture dominant variations between different classes of data:
  - Compute **generalized Eigen-decomposition** $\widehat{\Sigma}_B = \widehat{\Sigma} U \Lambda U^T$ with $\Lambda = \mathbf{diag}(\lambda_1, \lambda_2, ...\lambda_n)$ where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$;
  - Choose top-d generalized eigenvectors corresponding to top $d \leq K$ nonzero eigenvalues,

$$U_d = [u_1, \ldots, u_d], \quad u_j \in \mathbb{R}^p.$$

Sliced Inverse Rgression

- Data: $\{X_i, y_i\}_{i=1}^N$, where $X_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ is continuous (or ordered discrete)
- Divide the range of y_i into S non-overlapping slices $H_s(s = 1, ..., S)$. N_s is the number of observations within each slice.
- Compute the sample mean and total covariance matrix

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i, \qquad \hat{\Sigma}^{p\times p} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{\mu})(X_i - \hat{\mu})^T;$$

- Compute the mean of X_i over all slices and Between slices covariance matrix

$$\hat{\mu}_k = \frac{1}{N_s}\sum_{y_i \in H_s} X_i, \qquad \hat{\Sigma}_B^{p\times p} = \frac{1}{K}\sum_{h}^{K}(\hat{\mu_k} - \hat{\mu})(\hat{\mu_k} - \hat{\mu})^T;$$

Li's SIR

- K.-C. Li's Slice Inverse Regression (1991) aims to capture dominant variations between different slices of data:
  - Compute **Generalized Eigen-decomposition** $\hat{\Sigma}_B = \hat{\Sigma} U \Lambda U^T$ with $\Lambda = \mathbf{diag}(\lambda_1, \lambda_2, ...\lambda_n)$ where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$;
  - Choose top-d generalized eigenvectors corresponding to top $d \leq K$ nonzero eigenvalues,

$$\Gamma_d = [u_1, \ldots, u_d], \quad u_k \in \mathbb{R}^p.$$

Localized Sliced Inverse Rgression

- Data: $\{X_i, y_i\}_{i=1}^{N}$, where $X_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ is continuous (or ordered discrete)
- Divide the range of y_i into S non-overlapping slices $H_s(s = 1, ..., S)$. N_s is the number of observations within each slice.
- Compute the sample mean $\hat(\mu)$ and total covariance $\hat{\Sigma}$ as in SIR
- Compute the **localized** mean of X_i over all slices and **localized** Between-slice covariance matrix

$$\hat{\mu}_{i,loc} = \frac{1}{|s_i|} \sum_{j \in s_i} X_j, \qquad \hat{\Sigma}_{locB} = \frac{1}{N} \sum_i (\hat{\mu}_{i,loc} - \hat{\mu})(\hat{\mu}_{i,loc} - \hat{\mu})^T \ ;$$

where $s_i = \{j : x_j$ belongs to the k nearest neighbours of x_i in $H_s\}$ and s indexes the slice H_s to which i belongs.

LSIR

- Wu-Liang-Mukherjee Localized Slice Inverse Regression (2009) aims to capture nonlinear variations between different slices of data:
  - Compute **Generalized Eigen-decomposition** $\hat{\Sigma}_{locB} = \hat{\Sigma} U \Lambda U^T$ with $\Lambda = \mathbf{diag}(\lambda_1, \lambda_2, ...\lambda_n)$ where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$;
  - Choose top-d generalized eigenvectors corresponding to top $d \leq K$ nonzero eigenvalues,

$$\Gamma_d = [u_1, \ldots, u_d], \quad u_k \in \mathbb{R}^p.$$