Statistical Inference and Multivariate Analysis (MA324)

Lecture Slides
Lecture 34

Principal Component Analysis

Indian Institute of Technology Guwahati

Jan-May 2023

Principal Components Obtained from Standardized Variables

$$Z_i = \frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}, i = 1, 2, ..., p$$

$$\Leftrightarrow \quad \underset{\sim}{Z} = \begin{pmatrix} Z_1 \\ . \\ . \\ . \\ Z_p \end{pmatrix} = V^{-1}(\underset{\sim}{X} - \underset{\sim}{\mu}),\ where\ V = diag(\sqrt{\sigma_{11}}, \sqrt{\sigma_{22}}, ..., \sqrt{\sigma_{pp}})$$

- $Cov(\underset{\sim}{Z}) = \rho_{p\times p} = \begin{pmatrix} \rho_{11}(=1) & . & . & . & \rho_{1p} \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ \rho_{p1} & . & . & . & \rho_{pp}(=1) \end{pmatrix}$
- $\sum_{i=1}^{p} Var(Y_i) = p$
- Proportion of standardized population variance due to k^{th} principal component $= \frac{\lambda_k}{p},\ k = 1, 2, ..., p$, where λ_k's are the eigen values pf $\rho_{p\times p}$.

Summarizing Sample Variation using Principal Componenets

- Let $\underset{\sim}{x}_1, \underset{\sim}{x}_2, ..., \underset{\sim}{x}_n$ denote n indepenedent draws from a p-dimensional population with mean vector $\underset{\sim}{\mu}$ and variance-covariance matrix Σ.

- The sample mean is $\overline{\underset{\sim}{x}}$, the sample var-cov matrix S and the sample correlation matrix R.

- 1^{st} sample principal component $=$ Linear combination $\underset{\sim}{a}_1' \underset{\sim}{x}_j$ that maximizes the sample variance of $\underset{\sim}{a}_1' \underset{\sim}{x}_j$ subject to $\underset{\sim}{a}_1' \underset{\sim}{a}_1 = 1$.

- i^{th} sample principal component $=$ Linear combination $\underset{\sim}{a}_i' \underset{\sim}{x}_j$ that maximizes the sample variance of $\underset{\sim}{a}_i' \underset{\sim}{x}_j$ subject to $\underset{\sim}{a}_i' \underset{\sim}{a}_i = 1$ and zero sample correlation between $\underset{\sim}{a}_i' \underset{\sim}{x}_j$ and $\underset{\sim}{a}_k{}' \underset{\sim}{x}_j$ for all $k < i$.

- Note that sample variance of $\underset{\sim}{a}_1' \underset{\sim}{x}_j$ is $\underset{\sim}{a}_1' S \underset{\sim}{a}_1$. The sample covariance between $\underset{\sim}{a}_i' \underset{\sim}{x}_j$ and $\underset{\sim}{a}_k{}' \underset{\sim}{x}_j$ is $\underset{\sim}{a}_i' S \underset{\sim}{a}_k$.

Theorem

Let S be $p \times p$ sample var-cov matrix with eigen-value-eigen-vector pair $(\hat{\lambda_1}, \hat{e_1}), ..., (\hat{\lambda_p}, \hat{e_p})$ with $\hat{\lambda_1} \geq \hat{\lambda_2} \geq ... \geq \hat{\lambda_p} \geq 0$. Then i^{th} sample principal component is given by,

$$\hat{y_i} = \hat{e_i}' \underset{\sim}{x} = \hat{e_{i1}} x_1 + ... + \hat{e_{ip}} x_p, \ i = 1, 2, ..., p$$

Also, sample variance of $\hat{y_i} = \hat{\lambda_i}, \ i = 1, 2, ..., p$ and sample covariance $\hat{y_i}$ and $\hat{y_k}$ is 0, for $i \neq k$.

How many components to retain?

- No definite answer
- Scree plot

Sample principal components based on standardized data

- S matrix needs to be replaced by R.

Application of PCA...

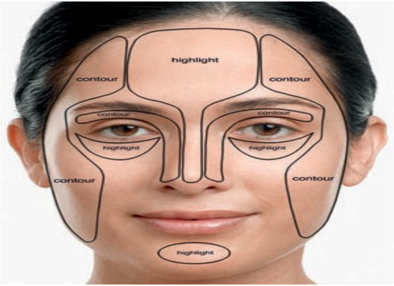

Figure 1. The regions that are either highlighted or contoured when used in PCA.[1]

- Reference: Alkandari, A., & Aljaber, S. J. (2015, April). Principle Component Analysis algorithm (PCA) for image recognition. In 2015 Second International Conference on Computing Technology and Information Management (ICCTIM) (pp. 76-80). IEEE.

Application of PCA...

- Original and compressed image with two, five, and eight principal components. (a) Original image. (b) Compression with two components. (c) Compression with five components. (d) Compression with eight components.
- Reference: Hernandez, W., Mendez, A., & Göksel, T. (2018). Application of principal component analysis to image compression. Statistics-Growing Data Sets and Growing Demand for Statistics.

Application of PCA...

Did you spot the bias in AI?