Supplementary Information: Principal Component Analysis (PCA)

July 2021

Principal Component Analysis (PCA) can be used to emphasize variation and bring out strong pattern in a dataset. It is often used for dimension reduction by projecting each data point onto only the first few principal components to obtain a lower-dimensional representation of the data while preserving as much of the variation of the data as possible. Some additional details regarding PCA is showed below:

- Given a set of data $\{x_j\}_{j=1}^n$, where each x_j is of dimension m, i.e. $x_j \in \mathbb{R}^m$
- We have the **sample mean** defined to be $\hat{\mu} = \frac{1}{n} x_j$, and this $\hat{\mu}$ is the best single point approximation to the dataset under the sum of square errors metric
- We also have the **sample variance matrix** [1] defined to be $\hat{\Sigma} = \frac{1}{n} \sum_{j=1}^n (x_j - \hat{\mu})(x_j - \hat{\mu})^T$ and $\hat{\Sigma} \in \mathbb{R}^{m \times m}$
- Note that: the sample covariance matrix $\hat{\Sigma}$ is symmetric positive semi-definite
- Thus, $\hat{\Sigma}$ has real nonnegative eigenvalues, and we can order them as $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m \geq 0$
- Moreover, $\hat{\Sigma}$ has a corresponding set of **orthonormal eigenvectors**, denoted as $u_1, u_2, ..., u_m$ with $\hat{\Sigma} u_i = \lambda_i u_i$ [2]
- Therefore, by **eigenvalue decomposition** of $\hat{\Sigma}$, we can write

$$\hat{\Sigma} = \sum_{i=1}^m \lambda_i u_i u_i^T = PDP^T$$

where

$$P = \begin{bmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_n^T \end{bmatrix}, \; D = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$$

- We call these m orthonormal directions $u_1, u_2, ..., u_m$ the **principle components** of the data.
- Note that the proportion of variance explained by the i-th principal component, i.e. u_i, is: $\frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_m}$
- Every data point can be decomposed (exactly) into a weighted linear combination of these orthonormal principal components. In particular,

$$x_j = \hat{\mu} + \sum_{i=1}^m u_i^T (x_j - \hat{\mu}) u_i = \hat{\mu} + \sum_{i=1}^m \alpha_i u_i$$

where $\alpha_i = u_i^T (x_j - \hat{\mu})$ is a scalar

[1] For simplicity in notation, we use $1/n$ rather than $1/(n-1)$ to normalize sample variances and covariances. The latter normalization is preferred in some contexts since it is statistically unbiased.

[2] Note that u_i here is a column vector while u_i^T is a row vector.

- By truncating the above sum to the leading d terms (i.e. $d < m$), one can obtain the d-th order approximation, denoted as $\hat{x}_j$, to the given data point x_j:

$$\hat{x}_j = \hat{\mu} + \sum_{i=1}^{d} u_i^T (x_j - \hat{\mu}) u_i = \hat{\mu} + \sum_{i=1}^{d} \alpha_i u_i$$

  In this case, we are able to represent the original data $\{x_j\}_{j=1}^n$ with a d-dimensional approximate $\{\hat{x}_j\}_{j=1}^n$. Moreover, such approximate explains the most variation of the dataset than any other d-dimensional approximates.

References

[1] Rasmus Bro and Age K Smilde. Principal component analysis. *Analytical methods*, 6(9):2812–2831, 2014.

[2] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.