Principal Components Analysis

MATH 271.1: Statistical Methods

Jakov Ivan S. Dumbrique

Ateneo de Manila University

Monday

Principal Components Analysis: An Intuition
http://setosa.io/ev/principal-component-analysis/

Principal Components Analysis

- PCA is a technique for **dimensionality reduction** as it produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have maximal variance, and are mutually uncorrelated.
- Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization.

Preliminaries

Suppose we have an $n \times p$ data set $\hat{\mathbf{A}}$ (also called data matrix) consisting of n observations:

$$\hat{\mathbf{A}} = \begin{bmatrix} \hat{a}_{11} & \hat{a}_{12} & \dots & \hat{a}_{1p} \\ \hat{a}_{21} & \hat{a}_{22} & \dots & \hat{a}_{2p} \\ \vdots & & \ddots & \vdots \\ \hat{a}_{n1} & \hat{a}_{n2} & \dots & \hat{a}_{np} \end{bmatrix}$$
$$= \begin{bmatrix} \hat{\mathbf{a_1}} & \hat{\mathbf{a_2}} & \dots & \hat{\mathbf{a_p}} \end{bmatrix},$$

where $\hat{\mathbf{a_1}}, \hat{\mathbf{a_2}}, \dots, \hat{\mathbf{a_p}}$ are called the feature vectors or simply *features* or *variables* of $\hat{\mathbf{A}}$. In this case, we will assume that $n > 1$–that is, our data matrix contains more than one observation.

Centering the Data Matrix

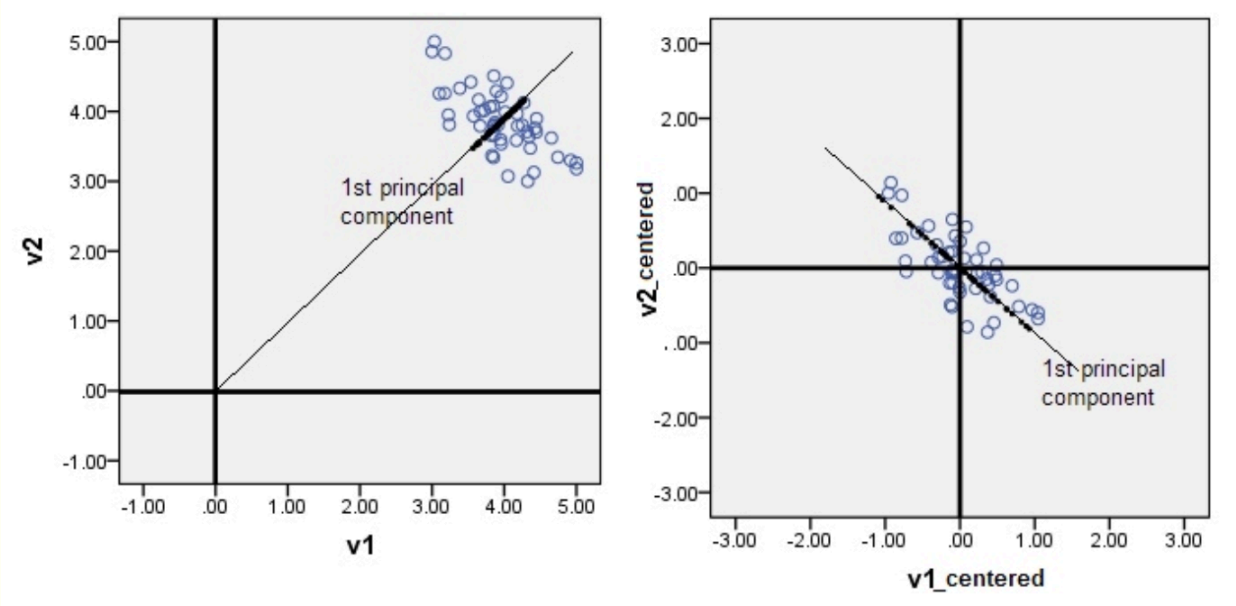
Since we are only interested in the variance of this data matrix, we make it centered by transforming each column of $\hat{\mathbf{A}}$ to have a mean of zero.

$$
\begin{aligned}
\mathbf{A} &= \hat{\mathbf{A}} - \bar{\boldsymbol{x}}^{\mathbf{T}} \\
&= \begin{bmatrix} \hat{\mathbf{a_1}} & \hat{\mathbf{a_2}} & \dots & \hat{\mathbf{a_p}} \end{bmatrix} - \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_p \end{bmatrix} \\
&= \begin{bmatrix} \hat{\mathbf{a_1}} - \bar{x}_1 & \hat{\mathbf{a_2}} - \bar{x}_2 & \dots & \hat{\mathbf{a_p}} - \bar{x}_p \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{a_1} & \mathbf{a_2} & \dots & \mathbf{a_p} \end{bmatrix},
\end{aligned}
$$

where $\mathbf{a_1}, \mathbf{a_2}, \dots, \mathbf{a_p}$ are the features of $\mathbf{A}$. We are interested in finding the principal components of this centered data matrix $\mathbf{A}$.

Why center?

If the data is not centered, the first principal component may pierce the cloud of data not along the main direction of the cloud, and thus will be statistically misleading.



Source:
https://stats.stackexchange.com/questions/22329/how-does-centering-the-data-get-rid-of-the-intercept-in-regression-and-pca

Variance-Covariance Matrix

Given an $n \times p$ data matrix $\mathbf{A}$, the sample variance-covariance matrix $\mathbf{S}$, often called sample covariance matrix, refers to the following symmetric $p \times p$ matrix:

$$\mathbf{S} = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{bmatrix},$$

the variances

covariances of the features

Variance-Covariance Matrix

- $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} a_{ij}$ is the sample mean of the jth feature $\hat{\mathbf{a_j}}$,
- $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(a_{ij} - \bar{x}_j)^2$ is the sample variance of the jth feature $\hat{\mathbf{a_j}}$,
- $s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(a_{ij} - \bar{x}_j)(a_{ik} - \bar{x}_k)$ is the sample covariance between the jth and kth features $\hat{\mathbf{a_j}}$ and $\hat{\mathbf{a_k}}$.

Variance-Covariance Matrix

AMF/31

Given a centered data matrix $\mathbf{A}$ where rows are observations and columns are features, it can be shown that the corresponding sample covariance matrix $\mathbf{S}$ is given by

$$\mathbf{S} = \frac{1}{n-1}\mathbf{A}^{\mathbf{T}}\mathbf{A}.$$

quadratic form

Finding the First PC

The *first principal component* (first PC) of a set of features $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$ of a centered data matrix $\mathbf{A}$ is the normalized linear combination of the features

first PC

$$\mathbf{z_1} = \phi_{11}\mathbf{a}_1 + \phi_{21}\mathbf{a}_2 + \ldots + \phi_{p1}\mathbf{a}_p$$

$= A\vec{\phi}_1$

$$\vec{\phi}_1 = \begin{bmatrix} \phi_{11} \\ \phi_{21} \\ \vdots \\ \phi_{p1} \end{bmatrix}$$

that has the largest variance. By *normalized*, we mean that

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

unit vector

$\|\vec{\phi}_1\| = 1$

$\vec{\phi}_1^T \vec{\phi}_1 = 1$

first PC loading vector

optimization problem

$$\max_{\vec{\phi}} \; Var(\vec{z}_1)$$

constraint: $\vec{\phi}_1^T \vec{\phi}_1 = 1$

how does this look like?

Theorem 3

Given a centered data matrix $\mathbf{A}$, the sample variance of its first principal component $\mathbf{z_1}$ is given by

$$\mathrm{Var}(\mathbf{z_1}) = s^2_{\mathbf{z_1}} = \boldsymbol{\phi}_1^{\mathbf{T}} \mathbf{S} \boldsymbol{\phi}_1,$$

where $\boldsymbol{\phi}_1$ is the first principal component loading vector and $\mathbf{S}$ is the sample covariance matrix of $\mathbf{A}$.

Finding the First PC

Therefore, in order to find the first principal component $\mathbf{z_1} = \mathbf{A}\boldsymbol{\phi_1}$, we need to find the loadings that will maximize the sample variance $s^2_{\mathbf{z_1}} = \boldsymbol{\phi}_\mathbf{1}^\mathbf{T}\mathbf{S}\boldsymbol{\phi_1}$ subject to the normalization constraint $\boldsymbol{\phi}_\mathbf{1}^\mathbf{T}\boldsymbol{\phi_1} = 1$. We solve this constrained optimization problem using the method of Lagrange multipliers.

Finding the First PC

Constrained optimization problem

$$\max_{\vec{\phi}_1} \; Var(\vec{z}_1) = \vec{\phi}_1^T S \vec{\phi}_1 := f(\vec{\phi}_1) = \lambda$$

Lagrange Multiplier

subject to $\vec{\phi}_1^T \vec{\phi}_1 = 1$.

$g(\vec{\phi}_1) = c$

Method of Lagrange Multipliers:

new objective function

$$u(\vec{\phi}_1, \lambda) = f(\vec{\phi}_1) - \lambda[g(\vec{\phi}_1) - c]$$

where λ is a Lagrange multiplier.

$$u(\vec{\phi}_1, \lambda) = \vec{\phi}_1^T S \vec{\phi}_1 - \lambda[\vec{\phi}_1^T \vec{\phi}_1 - 1]$$

$$\begin{cases} \dfrac{\partial u}{\partial \vec{\phi}_1} = 0 \\ \dfrac{\partial u}{\partial \lambda} = 0 \end{cases}$$

$$\frac{\partial u}{\partial \vec{\phi}_1} = \underbrace{\frac{\partial(\vec{\phi}_1^T S \vec{\phi}_1)}{\partial \vec{\phi}_1}}_{①} - \lambda \underbrace{\left[\frac{\partial(\vec{\phi}_1^T \vec{\phi}_1)}{\partial \vec{\phi}_1} - 0\right]}_{②} = 0$$

$$= 2S\vec{\phi}_1 - \lambda(2\vec{\phi}_1) = 0$$

$$\frac{\partial u}{\partial \lambda} = 0 - [\vec{\phi}_1^T \vec{\phi}_1 - 1] = 0$$

$$\vec{\phi}_1^T \vec{\phi}_1 = 1$$

(normalization constraint)

$$2S\vec{\phi}_1 = 2\lambda\vec{\phi}_1$$

$$S\vec{\phi}_1 = \lambda\vec{\phi}_1$$

λ eigenvalue of S

① $\dfrac{\partial(\vec{\phi}_1^T S \vec{\phi}_1)}{\partial \vec{\phi}_1}$

$$\boxed{\frac{\partial(\vec{u}^T A \vec{v})}{\partial \vec{x}} = \frac{\partial \vec{u}}{\partial \vec{x}} A\vec{v} + \frac{\partial \vec{v}}{\partial \vec{x}} A^T \vec{u}}$$

$$= \frac{\partial \vec{\phi}_1}{\partial \vec{\phi}_1} S\vec{\phi}_1 + \frac{\partial \vec{\phi}_1}{\partial \vec{\phi}_1} S^T \vec{\phi}_1$$

$$= S\vec{\phi}_1 + S^T\vec{\phi}_1 = 2S\vec{\phi}_1$$

② $\dfrac{\partial(\vec{\phi}_1^T \vec{\phi}_1)}{\partial \vec{\phi}_1}$

$$\boxed{\frac{\partial(\vec{u}^T \vec{v})}{\partial \vec{x}} = \frac{\partial \vec{u}}{\partial \vec{x}} \vec{v} + \frac{\partial \vec{v}}{\partial \vec{x}} \vec{u}}$$

$$= \frac{\partial \vec{\phi}_1}{\partial \vec{\phi}_1} \vec{\phi}_1 + \frac{\partial \vec{\phi}_1}{\partial \vec{\phi}_1} \vec{\phi}_1$$

$$= 2\vec{\phi}_1$$

Thus, the first PC loading vector $\vec{\phi}_1$ is an eigenvector of S. But which eigenvector?

$$Var(\vec{z}_1) = \vec{\phi}_1^T S \vec{\phi}_1$$

$$= \vec{\phi}_1^T \lambda \vec{\phi}_1$$

$$= \lambda \vec{\phi}_1^T \vec{\phi}_1$$

$$= \lambda$$

Since we want to maximize λ, λ = largest eigenvalue of S.

Thus, $\vec{\phi}_1$ is the eigenvector of S associated with the largest eigenvalue of S.

Finding the First PC

- The first principal component is given by $\mathbf{z_1} = A\boldsymbol{\phi_1}$, where the first principal component loading vector $\boldsymbol{\phi_1}$ is the eigenvector of the sample covariance matrix $\mathbf{S}$ that is associated with the largest eigenvalue $\lambda := \lambda_1$.
- If we know the SVD of the centered data matrix $\mathbf{A} = \mathbf{U\Sigma V^T}$, then the first principal component loading vector $\boldsymbol{\phi_1}$ is the right singular vector corresponding to the largest singular value of A. This is simply the first right singular vector $(\mathbf{v_1})$ of A since the singular values in $\boldsymbol{\Sigma}$ are arranged in a non-increasing order.

$= \vec{V}_1$

arranged in non increasing order

first/largest singular value of A

$S = \frac{1}{n-1} A^T A$

eigenvalues : $\frac{1}{n-1}\sigma_i^2$

eigenvectors : right singular vectors of A (columns of V in $A = U\Sigma V^T$)

largest eigenvalue of S : $\frac{1}{n-1}\sigma_1^2$

eigenvector $(\vec{\phi}_1)$: $\vec{V}_1$

Finding the Second PC

The *second principal component* (second PC) of a set of features $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$ of a centered data matrix $\mathbf{A}$ is the *normalized* linear combination of the features

2nd PC

$$\mathbf{z_2} = \phi_{12}\mathbf{a}_1 + \phi_{22}\mathbf{a}_2 + \ldots + \phi_{p2}\mathbf{a}_p = A\vec{\phi}_2.$$

that has the largest variance among all linear combinations that are *uncorrelated* with the first principal component $\mathbf{z_1}$. The loadings $\phi_{12}, \phi_{22}, \ldots, \phi_{p2}$ are the elements of the second principal component loading vector $\boldsymbol{\phi_2}$.

optimization problem

$$\max_{\vec{\phi}_2} \ Var(\vec{z}_2)$$

subject to ① $\vec{\phi}_2^T \vec{\phi}_2 = 1$

② $\vec{\phi}_2^T \vec{\phi}_1 = 0$

② $corr(\vec{z}_1, \vec{z}_2) = 0 \Rightarrow cov(\vec{z}_1, \vec{z}_2) = 0$

$$cov(\vec{z}_1, \vec{z}_2) = cov(A\vec{\phi}_1, A\vec{\phi}_2)$$
$$= \vec{\phi}_1^T S \vec{\phi}_2$$
$$= \vec{\phi}_2^T S \vec{\phi}_1$$
$$= \vec{\phi}_2^T \lambda_1 \vec{\phi}_1$$
$$0 = \lambda_1 \vec{\phi}_2^T \vec{\phi}_1$$

S = var-covar. matrix of A

Finding the Second PC

$\vec{\phi}_2^T \vec{\phi}_1 = 0$

"Principal Components Analysis" by Jolliffe

Exercise: Set up the constrained optimization problem of finding the second principal component $\mathbf{z_2}$ of a centered data matrix $\mathbf{A}$. Using Lagrange multipliers, show that the second principal component loading vector $\boldsymbol{\phi_2}$ is the eigenvector of the sample covariance matrix S that is associated with the second largest eigenvalue.

Var

$\lambda_1 = \frac{1}{n-1}\sigma_1^2 \Leftarrow \vec{z}_1 = A\vec{\phi}_1 = A\vec{v}_1 \rightarrow$ first right singular vector of A

$\lambda_2 = \frac{1}{n-1}\sigma_2^2 \Leftarrow \vec{z}_2 = A\vec{\phi}_2 = A\vec{v}_2 \rightarrow$ second " " " " "

$\vdots$

$\lambda_k = \frac{1}{n-1}\sigma_k^2 \Leftarrow \vec{z}_k = A\vec{\phi}_k = A\vec{v}_k$

Further principal components

In general, the *kth principal component* (kth PC) of a set of features $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$ of a centered data matrix $\mathbf{A}$ is the *normalized* linear combination of the features

$$\mathbf{z_k} = \phi_{1k}\mathbf{a}_1 + \phi_{2k}\mathbf{a}_2 + \ldots + \phi_{pk}\mathbf{a}_p = A\vec{\phi}_k = A\vec{v}_k$$

that has the largest variance among all linear combinations that are *uncorrelated* with the first $k-1$ principal components $\mathbf{z_1}, \mathbf{z_2}, \ldots, \mathbf{z_{k-1}}$. The loadings $\phi_{1k}, \phi_{2k}, \ldots, \phi_{pk}$ are the elements of the kth principal component loading vector $\boldsymbol{\phi_k}$.

Further principal components

- In general, the kth principal component is given by $\mathbf{z_k} = \mathbf{A}\boldsymbol{\phi_k}$, and $\text{Var}(\mathbf{z_k}) = s^2_{\mathbf{z_k}} = \lambda_k$, where λ_k is the kth largest eigenvalue of $\mathbf{S}$ and $\boldsymbol{\phi_k}$ is the corresponding eigenvector.
- We have shown that finding the principal components of a centered data matrix $\mathbf{A}$ reduces to finding the eigenvalues and eigenvectors of the sample covariance matrix $\mathbf{S} = \dfrac{1}{n-1}\mathbf{A^T A}$. However, the eigenvectors of $\mathbf{S}$ are the same as the ~~eigenvalues~~ eigenvectors of $\mathbf{A^T A}$ and the eigenvalues of $\mathbf{S}$ are just the eigenvalues of $\mathbf{A^T A}$ scaled by a factor of $\dfrac{1}{n-1}$.

Recall Theorem 1

Further principal components

- *How do we find the eigenvalues and eigenvectors of* $\mathbf{A}^{\mathbf{T}}\mathbf{A}$*?*
  - If we know the SVD of A, then by Theorem 1, we already know the eigenvalues and eigenvectors of $\mathbf{A}^{\mathbf{T}}\mathbf{A}$, and hence that of $\mathbf{S}$ as well.
  - Given the SVD of the centered data matrix $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathbf{T}}$, the first principal component loading vector $\boldsymbol{\phi}_{\mathbf{1}}$ is the right singular vector corresponding to the largest singular value of A. This is simply the first right singular vector $(\mathbf{v}_{\mathbf{1}})$ of A since the singular values in $\boldsymbol{\Sigma}$ are arranged in a non-increasing order.
  - The variance of the first principal component $\mathbf{z}_{\mathbf{1}}$ is given by $\text{Var}(\mathbf{z}_{\mathbf{1}}) = s^2_{\mathbf{z}_{\mathbf{1}}} = \lambda_1 = \dfrac{1}{n-1}\sigma_1^2$.
  - The other principal components and their loading vectors can be retrieved in a similar manner.

Further principal components

- *Why do we prefer calculating the SVD of $\mathbf{A}$ over the eigenvalue decomposition of $\mathbf{A}^{\mathbf{T}}\mathbf{A}$?*
  - According to Trefethen and Bau (1997), calculating the SVD of an $n \times p$ matrix $\mathbf{A}$ ($4np^2 - \frac{4}{3}p^3$ flops) is less computationally expensive than forming $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ and computing its eigenvalue decomposition, which takes a number of flops of order $O(p^3)$. Statistical softwares use divide-and-conquer approach for solving the SVD of a matrix, which is more numerically stable than the QR algorithm used in solving for the eigenvalue decomposition of its covariance matrix.

Geometry of the First PC

- The loading vector $\phi_{\mathbf{1}}$ with elements $\phi_{11}, \phi_{21}, \ldots, \phi_{p1}$ defines a direction in feature space along which the data vary the most.
- If we project the n data points $\mathbf{a_1}, \ldots, \mathbf{a_n}$ onto this direction, the projected values are the principal component scores $z_{11}, \ldots, z_{n1}$ themselves.

Further principal components

- Once we have computed the principal components, we can plot them against each other in order to produce low-dimensional views of the data. For instance, we can plot the $\mathbf{z_1}$ against $\mathbf{z_2}$, $\mathbf{z_1}$ and $\mathbf{z_3}$, $\mathbf{z_2}$ against $\mathbf{z_3}$, and so forth. Geometrically, this amounts to projecting the original data down onto the subspace spanned by the orthogonal basis vectors $\boldsymbol{\phi_1}, \boldsymbol{\phi_2}$, and $\boldsymbol{\phi_3}$, and plotting the projected points.

Scaling of the variables matters

7 x 5

- We have already mentioned that before PCA is performed, the variables should be centered to have mean zero. Furthermore, the results obtained which we perform PCA will also depend on whether the variables have been individually scaled (each multiplied by a different constant).
- If the variables are in different units, scaling each to have standard deviation equal to one is recommended.

height (m) distance from home to Atenea (m)

person 1

person 2

1 – 2 n

1000 – 100,000

How many PCs should we retain?

If we use principal components as a summary of our data, how many components are sufficient?

- No simple answer to this question, as cross-validation is not available for this purpose.
  - *Why not?*
- To understand the strength of each component, we are interested in knowing the proportion of variance explained (PVE) by each one. We can use the PVEs to determine how many PCs should we retain.

Proportion of Variance Explained

- The *total variance* present in a data set (assuming that the variables have been centered to have mean zero) is defined as the sum of sample variances of each variable:

$$\sum_{j=1}^{p} Var(\mathbf{a_j}) = \sum_{j=1}^{p} \frac{1}{n-1} \sum_{i=1}^{n} a_{ij}^2,$$

- It can be shown that the total variance is also equal to the sum of the sample variances of each PC:

$$\sum_{j=1}^{p} Var(\mathbf{a_j}) = \sum_{m=1}^{M} \underline{Var(\mathbf{z_m})},$$

with $M = \min(n-1, p)$. More often than not, $M = p$ since there are usually more observations than variables.

Proportion of Variance Explained

- Since the sample variance of the mth PC is the mth largest eigenvalue of $\mathbf{S}$,

$$\sum_{j=1}^{p} Var(\mathbf{a_j}) = \sum_{m=1}^{M} Var(\mathbf{z_m}) = \sum_{m=1}^{M} \lambda_m$$

- The proportion of variance explained (PVE) of the kth principal component is the ratio of the sample variance of the kth PC to the total variance:

$$PVE_k = \frac{Var(\mathbf{z_k})}{\sum_{m=1}^{M} Var(\mathbf{z_m})} = \frac{\lambda_k}{\sum_{m=1}^{M} \lambda_m}$$

λ_1: largest eigenvalue of S

$$\sum_{k=1}^{M} PVE_k = 1$$

$$PVE_1 \geq PVE_2 \geq \cdots \geq PVE_k \cdots \geq PVE_M$$

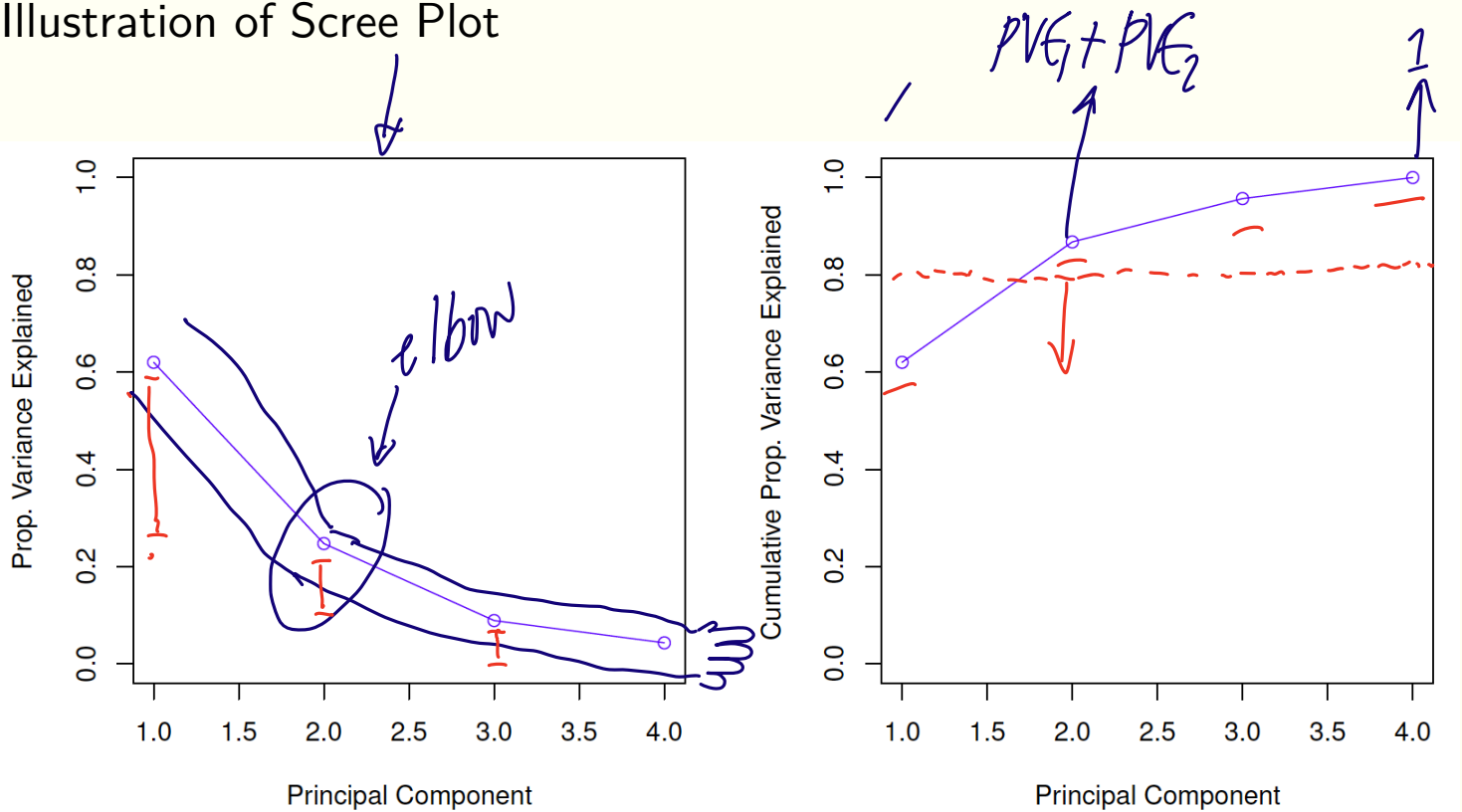
Proportion of Variance Explained

- For any k, $PVE_k \in [0, 1]$.
- The PVEs sum to one. That is,

$$\sum_{k=1}^{M} PVE_k = \sum_{k=1}^{M} \left[\frac{\lambda_k}{\sum_{m=1}^{M} \lambda_m} \right] = 1$$

- The PVEs can be visualized using "scree plots".

Illustration of Scree Plot



(Figure from James et. al., 2013)

2 PCs

2 PCs

How many PCs should we retain?

There are two widely used methods on determining how many principal components should be used:

- we look for an "elbow" in the scree plot: a good threshold is where the PVE drops significantly [Cattell, 1966]. In most of the datasets, a significant drop of PVE occurs.

method 1

- we set a threshold to the cumulative PVE of the chosen PCs (e.g. we retain the first k PCs that explain at least 80% of the variance) [Abdi & Williams, 2010]

method 2

$\frac{1}{4}$

70–80%

References

1. Abdi, H. & L. Williams. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics,* 2(4):433-459.
2. Cattell, R. (1966). The scree test for the number of factors. *Multivariate behavioral research,* 1(2):245-276.
3. Goodfellow, I., Y. Bengio, & A. Courville. (2016). Deep Learning. MIT Press.
4. Hastie, T., R. Tibshirani, & J. Friedman. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.
5. Izenman, A. (2008). Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Springer.

References

6. James, G., D. Witten, T. Hastie, & R. Tibshirani. (2013). An Introduction to Statistical Learning (with Applications in R). Springer.
7. Johnson R. & D. Wichern. (2007). Applied Multivariate Statistical Analysis (6th ed.). Prentice Hall.
8. Jolliffe, I. (2002). Principal Component Analysis (2nd ed.). Springer.
9. Theodoridis, S. & K. Koutroumbas. (2009). Pattern Recognition (4th ed.). Academic Press.
10. Trefethen L. & D. Bau. (1997). Numerical Linear Algebra (1st ed.). SIAM: Society for Industrial and Applied Mathematics.

References

11. Wickham, H. & G. Grolemund. (2017). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media, Inc.