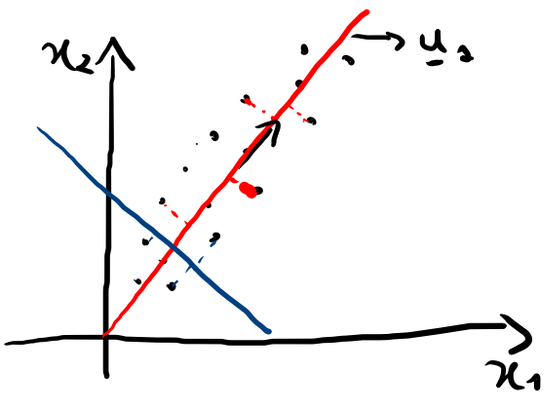
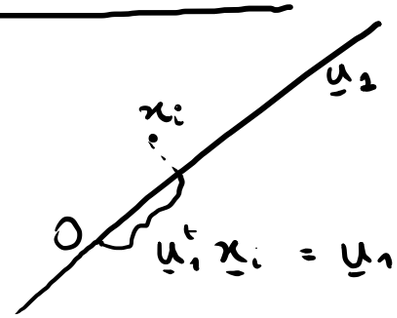
ALTERNATIVE DERIVATION OF PCA

10/05/21





score

$$u_1^t x_i = u_1 \cdot x_i = u_{11}x_{i2} + u_{21}x_{i2} + \cdots + u_{p1}x_{ip} = z_{i2} \quad , i = 1, \ldots, n$$

$$u_1 = \begin{pmatrix} u_{11} \\ \vdots \\ u_{p1} \end{pmatrix} \rightarrow \text{loadings}$$

Assume that the data are centered ($\bar{x}_j = 0 \ \forall j$) $\Rightarrow$ 2 scores are also centered

$$\left(\frac{\sum_{i=1}^{n} z_{i2}}{n} = u_{11}\underbrace{\bar{x}_1}_{0} + \cdots + u_{p1}\underbrace{\bar{x}_p}_{0} = 0 \right)$$

Total variance of projections:

$$\frac{1}{n}\sum_{i=1}^{n}(z_{i2})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} u_{j2}x_{ij}\right)^2$$

OPTIMIZATION PROBLEM

$$\underset{u_2 \in \mathbb{R}^p, \|u_1\|=1}{\operatorname{argmax}} \ \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} u_{j2}x_{ij}\right)^2$$

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} u_{j1}x_{ij}\right)^2}_{\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2} = \frac{1}{n}\sum_{i=1}^{n}\left(\underline{u}_1^t \underline{x}_i\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\underline{u}_1^t \underline{x}_i\right)\cdot\left(\underline{x}_i^t \underline{u}_1\right)$$

$$= \frac{1}{n}\,\underline{u}_1^t \sum_{i=1}^{n} \underline{x}_i \cdot \underline{x}_i^t\, \underline{u}_1$$

($\underline{x}_i$: $p\times 1$, $\underline{x}_i^t$: $1\times p$, so $\underline{x}_i \cdot \underline{x}_i^t$: $p\times p$)

$$= \underline{u}_1^t \left[\frac{\sum_{i=1}^{n} \underline{x}_i \cdot \underline{x}_i^t}{n}\right] \underline{u}_1$$

SAMPLE COVARIANCE (S)

$$= \underline{u}_1^t \frac{X^t X}{n}\, \underline{u}_1 = \lambda_1$$

Setting $\underline{u}_1^t S \underline{u}_1 = \lambda_1 \Rightarrow \underbrace{\underline{u}_1 \underline{u}_1^t}_{I} S \underline{u}_1 = \lambda_1 \underline{u}_1 \Rightarrow S\underline{u}_1 = \lambda_1 \underline{u}_1$

So $\underline{u}_1$ is the eigenvector of S corresponding to the largest eigenvalue.

What about the second component?

Choose $\underline{u}_2$ so that $\underbrace{\underline{u}_2^t S \underline{u}_2}_{\text{variance of projections on } \underline{u}_2}$ is maximised and $\underline{u}_2 \perp \underline{u}_1$

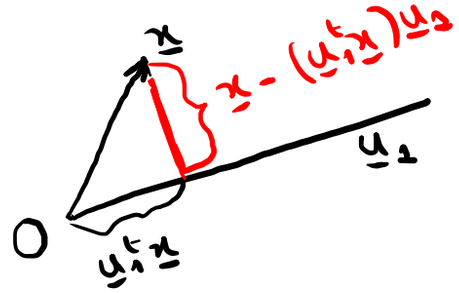
$\Rightarrow$ $\underline{u}_2$ is the eigenvector corresponding to the second largest eigenvalue.

And so forth:

$$\underline{u}_1, \underline{u}_2, \underline{u}_3, \dots, \underline{u}_p$$

We've reconstructed the same basis as last time!

There is a final justification for this particular choice of vectors.



We can state the problem as finding $\underline{u}_2$ that minimizes the average squared distances of the points to the chosen direction.

OPTIMIZATION PROBLEM

$$\underset{\substack{\underline{u}_1 \in \mathbb{R}^p \\ \|\underline{u}_1\|=1}}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \| \underline{x}_i - (\underline{u}_1^t \underline{x}_i)\, \underline{u}_2 \|^2$$

$$\frac{1}{n} \sum_{i=1}^{n} \| \underline{x}_i - (\underline{u}_1^t \underline{x}_i)\underline{u}_1 \|^2 = \frac{1}{n} \sum_{i=1}^{n} (\underline{x}_i - (\underline{u}_1^t \underline{x}_i)\underline{u}_1)^t (\underline{x}_i - (\underline{u}_1^t \underline{x}_i)\underline{u}_2)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[\underline{x}_i^t \underline{x}_i - \underline{x}_i^t (\underline{u}_1^t \underline{x}_i)\underline{u}_1 - (\underline{u}_1^t \underline{x}_i)\underline{u}_1^t \underline{x}_i + \underline{u}_1^t (\underline{u}_1^t \underline{x}_i)(\underline{u}_1^t \underline{x}_i)\underline{u}_2 \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[\underline{x}_i^t \underline{x}_i - (\underline{u}_1^t \underline{x}_i)(\underline{x}_i^t \underline{u}_2) - (\underline{u}_1^t \underline{x}_i)^2 + (\underline{u}_1^t \underline{x}_i)^2 \underline{u}_1^t \underline{u}_1 \right]$$

($\underline{x}_i^t \underline{u}_2 = \underline{u}_1^t \underline{x}_i$; $\underline{u}_1^t \underline{u}_1 = \|\underline{u}_1\|^2 = 1$)

$$= \frac{1}{n} \sum_{i=1}^{n} \left[\underline{x}_i^t \underline{x}_i - (\underline{u}_1^t \underline{x}_i)^2 \right]$$

$$= \underbrace{\frac{1}{n} \sum_{i=1}^{n} \underline{x}_i^t \underline{x}_i}_{\text{does not depend on } \underline{u}_2} - \underbrace{\frac{1}{n} \sum_{i=1}^{n} \underline{u}_1^t \underline{x}_i \underline{x}_i^t \underline{u}_1}_{\underline{u}_1^t S \underline{u}_1}$$

$\Rightarrow$ So $\underline{u}_2$ should maximize $\underline{u}_1^t S \underline{u}_2$ (as before!)

With two components, we would look at the plane that is closest to the data points:

$$\frac{1}{n}\sum_{i=1}^{n} \| \underline{x}_i - \underbrace{z_{i1}}_{\underline{u}_1^t \underline{x}_i}\underline{u}_1 - \underbrace{z_{i2}}_{\underline{u}_2^t \underline{x}_i}\underline{u}_2 \|^2$$

← We want to find $\underline{u}_1$ and $\underline{u}_2$ that minimix this quantity with $\underline{u}_1 \perp \underline{u}_2$

$$= \frac{1}{n}\sum_{i=1}^{n} (\underline{x}_i - z_{i1}\underline{u}_1 - z_{i2}\underline{u}_2)^t (\underline{x}_i - z_{i1}\underline{u}_1 - z_{i2}\underline{u}_2)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \Big(\underline{x}_i^t \underline{x}_i - z_{i1}\underline{x}_i^t \underline{u}_1 - z_{i2}\underline{x}_i^t \underline{u}_2 - z_{i1}\underline{u}_1^t \underline{x}_i + z_{i1}^2 \underbrace{\underline{u}_1^t \underline{u}_1}_{\|\underline{u}_1\|^2 = 1} + z_{i1}z_{i2}\underbrace{\underline{u}_1^t \underline{u}_2}_{=0 \text{ since } \underline{u}_1 \perp \underline{u}_2}$$

$$- z_{i2}\underline{u}_2^t \underline{x}_i + z_{i1}z_{i2}\underbrace{\underline{u}_2^t \underline{u}_1}_{0} + z_{i2}^2 \underbrace{\underline{u}_2^t \underline{u}_2}_{=\|\underline{u}_2\|^2 = 1} \Big)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \Big[\underline{x}_i^t \underline{x}_i - z_{i1}^2 - z_{12}^2 - \cancel{z_{i1}^2} + \cancel{z_{i1}^2} - \cancel{z_{i2}^2} + \cancel{z_{i2}^2} \Big]$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n} \underline{x}_i^t \underline{x}_i}_{\text{does not depend on } \underline{u}_1 \text{ nor } \underline{u}_2} - \underbrace{\frac{\sum_{i=1}^{n} z_{i1}^2}{n}}_{\substack{\text{maximized by } \underline{u}_2 \\ \text{eigenvector corresponding to} \\ \text{largest eigenvalue of } S}} - \underbrace{\frac{\sum_{i=1}^{n} z_{i2}^2}{n}}_{\rightarrow \text{ second largest eigenvalue}}$$

CHOOSING THE NUMBER OF COMPONENTS

TOTAL VARIANCE

$$\text{Trace}\left(\frac{X^tX}{n}\right)$$

Sum of diagonal terms of sample covariance

$$= \text{Trace}\left(\frac{UD^t\overbrace{V^tV}^{I}DU^t}{n}\right)$$

$X = VDU^t$

$$= \frac{\text{Trace}(UD^tDU^t)}{n} = \frac{\text{Trace}(D^tD\overbrace{U^tU}^{I})}{n}$$

$$= \frac{\text{Trace}(D^tD)}{n} =$$

$$D_{n\times p} = \begin{bmatrix} d_{11} & & 0 \\ & \ddots & \\ 0 & & d_{pp} \end{bmatrix}, \quad d_{ii} = \sqrt{\lambda_i}$$

$$= \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_p}{n}$$

$$D^tD_{p\times p} = \begin{bmatrix} d_{11}^2 & & 0 \\ & \ddots & \\ 0 & & d_{pp}^2 \end{bmatrix}$$

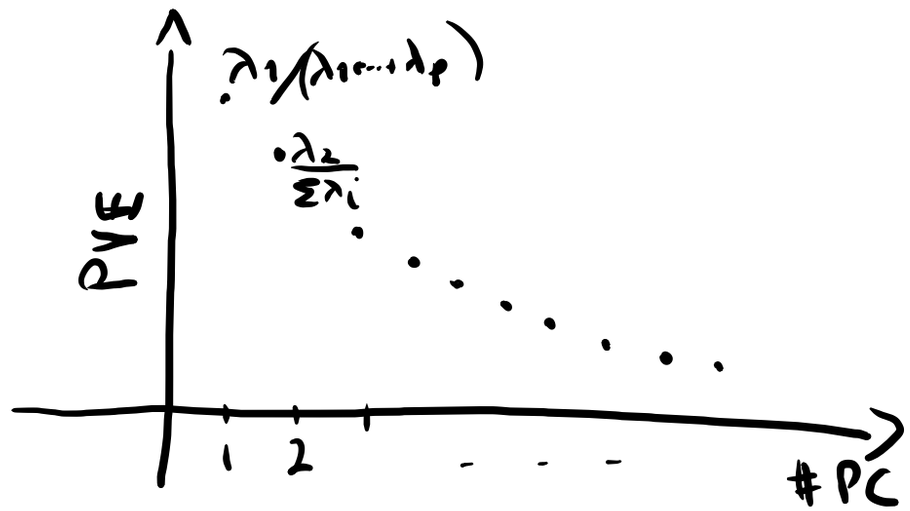
VARIANCE EXPLAINED BY k^{th} COMPONENT

$$\frac{1}{n}\sum_{i=1}^{n} z_{ik}^2 = \frac{\lambda_k}{n}$$ where λ_k is the k^{th} eigenvalue of X^tX

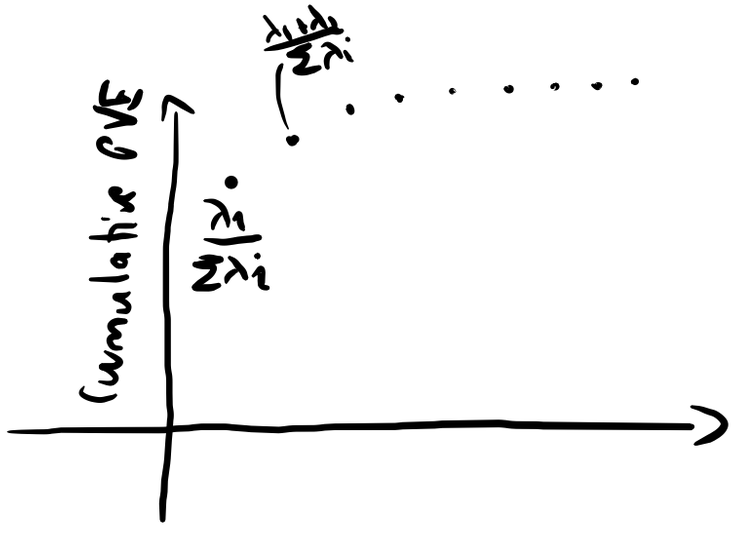
PROPORTION OF VARIANCE EXPLAINED (PVE)

$$PVE_k = \frac{\frac{\lambda_k}{n}}{\frac{\lambda_1 + \cdots + \lambda_p}{n}} = \frac{\lambda_k}{\lambda_1 + \cdots + \lambda_p} \quad \Rightarrow \sum_{k=1}^{p} PVE_k = 1$$

TYPICAL PLOTS:



← Usually looking for the "elbow".

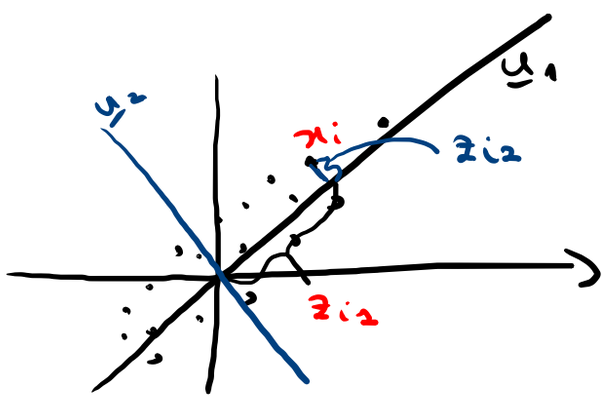


PCA IN PRACTICE

① Get familiar with the terminology!

$$Z_K = \phi_{1K} X_1 + \phi_{2K} X_2 + \dots + \phi_{pK} X_p$$

$$z_{iK} = \sum_{j=1}^{p} \phi_{jK} x_{ij}, \quad i = 1, \dots, n$$



K PRINCIPAL COMPONENT

SCORES OF K^{th} PRINCIPAL COMPONENT

(projections)

$\phi_1, \dots, \phi_p$ PRINCIPAL COMPONENT LOADING VECTORS → directions

$\phi_{1K}, \dots, \phi_{pK}$ LOADINGS OF K^{th} COMPONENT

→ how much each variable $X_1, \dots, X_p$ contributed to component K.

→ scores and loadings are both "visible" in a biplot

② SCALING OF DATA

→ Always advised

③ IMPLEMENTATION

Two functions:

- stats : prcomp (SVD)
- stats : princomp (ED of sample covariance)

Results unique up to sign flip !

PCA IN SUPERVISED LEARNING (Section 6.3)

Consider these two cases:

① The number of variables p is very large, e.g. in genomics

→ Dimensionality reduction to reduce p

② Predictors are highly correlated

→ Summarising information via principal components

→ One component would represent a number of correlated variables

Two methods:

- PRINCIPAL COMPONENT REGRESSION (PCR)
- PARTIAL LEAST SQUARES (PLS)

PCR

$X_1, \dots, X_p \rightarrow$ extract M $(< p)$ components

$Z_1, \dots, Z_M$

$\rightarrow$ Use $Z_1, \dots, Z_n$ as my new predictors

$$y = Z\theta + \varepsilon \qquad \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_M \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad Z = \begin{pmatrix} z_{11} & \cdots & z_{1M} \\ \vdots & & \vdots \\ z_{n1} & & z_{nM} \end{pmatrix}$$

$= z_0 + z_1\theta_1 + \dots + z_M\theta_n + \varepsilon$

(scores)

The vector θ is related to the original regression coefficients β $(y = X\beta + \varepsilon)$.

In order to see this:

$$z_{ik} = u_{1k}x_{i1} + \dots + u_{pk}x_{ip}, \quad i = 1, \dots, n, \quad k = 1, \dots, M$$

In matrix form:

$$Z = XU \qquad \text{with} \quad U = \begin{pmatrix} u_{11} & & u_{1M} \\ \vdots & \cdots & \vdots \\ u_{p1} & & u_{pM} \end{pmatrix}$$

$n \times M \quad n \times p \; p \times M$

So

$$y = Z\theta + \varepsilon$$

$$= XU\theta + \varepsilon$$

$$= X\underbrace{(U\theta)}_{\beta} + \varepsilon$$

So $\underset{p\times 1}{\beta} = \underset{p\times M}{U}\,\underset{M\times 1}{\theta}$ $\qquad \beta_j = \sum_{k=1}^{M} u_{jk}\theta_k$

- $\hat{\theta} \rightarrow \hat{\beta}$ interpretation
- Or check variables most associated to one component

Choice of M $\longrightarrow$ · Either like before
· Or, better, via CV or test error (supervised learning!)

PLS

This method uses y (the response) also in the selection of the directions.
In particular, we look for directions that:

1. Explain/summarise variation in X
2. Are most related with y

FIRST PLS DIRECTION

Find transformation $z_1 = X w_1$ with $\|w_1\| = 1$ and such that $z_1^t y$ is maximised (with maximal correlation with y).

This is found by projecting X on y, ie.

$$w_1 = \frac{X^t y}{\|X^t y\|}, \text{ ie } w_{j1} \propto x_j^t y$$

proportional to regression coefficient of X_j on y

So $$z_{i1} = \sum_{j=1}^{p} w_{j1} x_{ij} = w_{11} x_{i1} + \cdots + w_{p1} x_{ip}$$

maximal contribution by predictors that are most correlated with y

SECOND PLS DIRECTION

$$X_1 = \gamma_1 z_1 + \varepsilon$$

$$\vdots$$

$$X_p = \gamma_p z_1 + \varepsilon$$

Calculate residuals:

$$\hat{\varepsilon}_j = \underline{x}_j - \hat{\gamma}_j \underline{z}_1 \qquad \text{(orthogonal to } \underline{z}_1\text{)}$$

Create a new data matrix

$$E = \begin{pmatrix} \hat{\varepsilon}_{11} & \cdots & \hat{\varepsilon}_{1p} \\ \vdots & & \vdots \\ \hat{\varepsilon}_{n1} & & \hat{\varepsilon}_{np} \end{pmatrix} \quad \begin{matrix} \\ \\ \\ \hat{\varepsilon}_1 \quad \hat{\varepsilon}_p \end{matrix}$$

Remaing information not accounted for by 1st PLS

Find a transformation $\underline{z}_2 = E\underline{w}_2$ with $\|\underline{w}_2\| = 1$, such that $\underline{z}_2^t \underline{y}$ is maximised.

So $\underline{w}_2 \propto E^t \underline{y}$ .... And continue this way ....

The number of PLS components can be decided by CV.