PCA

AW

Lecture Overview

Last Class – Correlation

Z-score of indiviual (raw) value of random variable X is the distance of this value from the mean measured in the number of standard deviations: $z_X = \frac{X - E(X)}{\sigma_X}$

Pearson correlation coefficient for random variables X and Y denoted by ρ_{XY} is the expectation of a product of Z-scores of X and Y, i.e.

$$\rho_{XY} = E\left(\frac{X - E(X)}{\sigma_X} \cdot \frac{Y - E(Y)}{\sigma_Y}\right) = \frac{cov(XY)}{\sigma_X \cdot \sigma_Y}$$

For sample value vectors $\vec{s}_X,\ \vec{s}_y \in \mathbb{R}^n$ drawn from random variables X and Y we have $\hat{\rho}_{XY} = \frac{\vec{s}_X^c \bullet \vec{s}_Y^c}{\|\vec{s}_X^c\| \cdot \|\vec{s}_Y^c\|}$. Observe that $-1 \leq \hat{\rho}_{XY} \leq 1$

- Pearson correlation is +1 in the case of a perfect positive (increasing) linear relationship between X and Y.
- Pearson correlation is -1 in the case of a perfect decreasing (negative) linear relationship (anti-correlation)
- If the variables areindependent, Pearson's correlation coefficient is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables.
- Value between -1 and 1 indicate a the degree oflinear (in)dependencebetween the variables

Last Class – Data Normalizzation

To normalize the data is to make it consistent in some way. Two basic types of normalization:

- feature normalization is adjusting each value of a feature the same way across all examples
  - Typical feature normalizations are centering (i.e. $x_i \rightarrow x_i - \mu_X$), variance scaling (i.e. $x_i \rightarrow \frac{x_i}{\hat{\sigma}_X}$) and asolute scaling (i.e. $x_i \rightarrow \frac{x_i}{\max_i x_i}$)
- example normalization treats each example as a vector in some space (usually $\mathbb{R}^n$). Normalization is then linear automorfism of this space
- The main advantage to example normalization is that it makes comparisons more straightforward across data sets
  - Most common example normalization is vector normalization (i.e. $\vec{x}_i \rightarrow \frac{\vec{x}_i}{\|\vec{x}_i\|}$)

Last Class – Irrelevant and Redundunt Features

- Intuitively, an irrelevant feature is one that is completely uncorrelated with the prediction. We'd like to prune irrelevant features. A reasonable definition of real valued irrelevant feature is a feature with low variance.
- Intuitively, a redundant feature is one that is a function of another feature(s). A reasoable definition of a redundunt feature is a feature that is highly correlated with another feature

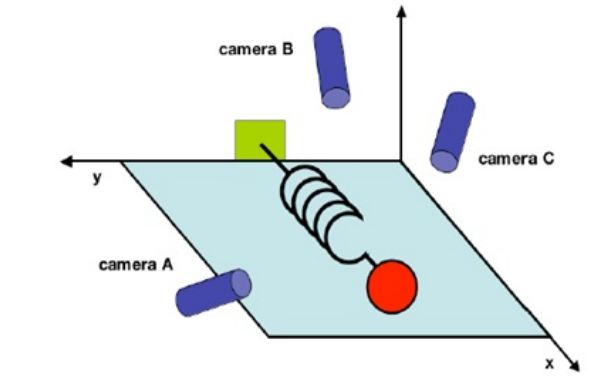
For a given real-valued data D the best representation is using basis that does not include redundant and irrelevant features.

PCA asks: Is there another basis that best re-expresses our data set?

Lecture Overview

Guiding example by Jonathon Shlens

- A ball of mass m attached to a massless, frictionless spring. The ball is released a small distance away from equilibrium (i.e. the spring is stretched). Because the spring is ideal, it oscillates indefinitely along the x-axis about its equilibrium at a set frequency
- We are unaware of this we do not know which axes and dimensions are important to measure.
- We measure the ball's position in a three-dimensional space by placing 3 video cameras around our system of interest. At 120 Hz each camera records an image indicating a two dimensional position of the ball on a screen (a projection).



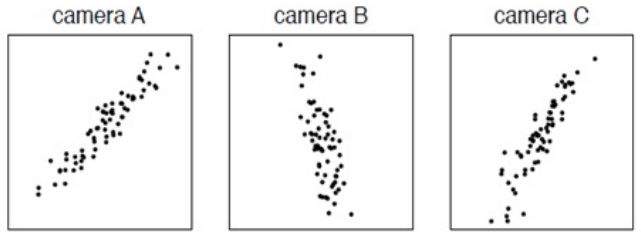


Figure : Our experiment

- We treat every time sample as an individual sample in our data set. At each time sample we record a set of data consisting of multiple measurement: at one point in time, camera A records a ball position (x_A, y_A), camera B (x_B, y_B), camera C (x_C, y_C). So one instance is a time measurement $(x_A, y_A, x_B, y_B, x_C, y_C)$

Geometric View of the Example

- Say we recorded the experiment for 1 min. Then we have $60 * 120 = 7200$ instances.
- Each instance is a 6-dimensional vector $\vec{x}$ in a standard basis, i.e. $\vec{x} = (x_1, \ldots, x_6)^T \in \mathbb{R}^6$
- Why did we choose standard basis $\vec{e}_1, \ldots, \vec{e}_6$ to express our instances? why not any other orthonormal basis? why not for example columns of this orthogonal matrix?

$$\begin{pmatrix} 1/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6} \\ -5/\sqrt{318} & 7/\sqrt{318} & -11/\sqrt{318} & 7/\sqrt{318} & -5/\sqrt{318} & 7/\sqrt{318} \\ 6\sqrt{2/1219} & 15/\sqrt{2438} & -8\sqrt{2/1219} & -19\sqrt{2/1219} & 6\sqrt{2/1219} & 15/\sqrt{2438} \\ 29\sqrt{2/3013} & -31/\sqrt{6026} & -8\sqrt{2/3013} & 4\sqrt{2/3013} & -17\sqrt{2/3013} & 15/\sqrt{6026} \\ 10\sqrt{6/10087} & -\sqrt{14/4323} & -67\sqrt{2/30261} & 67/\sqrt{60522} & 47\sqrt{3/20174} & -20\sqrt{6/10087} \\ -\sqrt{6/77} & -\sqrt{14/33} & -\sqrt{2/231} & 1/\sqrt{462} & 3\sqrt{3/154} & 2\sqrt{6/77} \end{pmatrix}$$

- The reason is that the naive standard basis reflects the method we gathered the data. Pretend we recorded the position $\begin{pmatrix} 2 \\ 2 \end{pmatrix}$ on camera A. We did not record the position

$$2 \begin{pmatrix} 1/\sqrt{6} \\ -5/\sqrt{318} \\ 6\sqrt{2/1219} \\ 29\sqrt{2/3013} \\ 10\sqrt{6/10087} \\ -\sqrt{6/77} \end{pmatrix} + 2 \begin{pmatrix} 1/\sqrt{6} \\ 7/\sqrt{318} \\ 15/\sqrt{2438} \\ -31/\sqrt{6026} \\ -\sqrt{14/4323} \\ -\sqrt{14/33} \end{pmatrix}$$

What are we trying to achieve?

- Obviously in our example it is enough to have one coordinate to fully analyze the data $\mathtt{D}$. So we are looking to reduce the number of dimensions in $\mathbb{R}^n$.
- Is dimesionality reduction always the case? Not,really. We may just be looking for representation of $\mathtt{D}$ in some $\mathbb{R}^d$ in which we can better solve our actual task (classification, regression, ranking, etc.)
- Thus we are looking for a map $W : \mathbb{R}^n \to \mathbb{R}^d$ where d can be any number such that our next task is solved easily. But the resut of the next task should be applied to $\mathtt{D}$, not to $W(\mathtt{D})$. So we need an 'inverse' map $U : \mathbb{R}^d \to \mathbb{R}^n$ that assigns 'results' obtained in $\mathbb{R}^d$ back to original data $\mathtt{D}$.
- What should be properties of a pair of functions U, W? ideally $U \cdot W$ should be identity function. But it may not be possible as it is not possible in our example (if these are linear maps then ranks of matrices would be 1). So what is the next best alternative? $U \cdot W|_{\mathtt{D}} = 1$. But even this may not be possible as in our example! Therefore for each data point x we want to have $\tilde{x} = UWx$ be as close to x as possible (so that $\tilde{x}$ has the same properties as x).
- Data contains many points so we need to minimize sum of absolute values of distances between pairs (data point, its image under UW), or equivalently sum of squares, i.e. we need to obtain

$$(S, T) = \arg\min_{U \in \{\mathbb{R}^d \to \mathbb{R}^n\}, W \in \{\mathbb{R}^n \to \mathbb{R}^d\}} \sum_{i=1}^{m} \|x - \tilde{x}\|^2$$

What is Principal Component Analysis (PCA)?

So re-expressing the data means to find maps $W : \mathbb{R}^n \to \mathbb{R}^d$ where $d \leq n$, and $U : \mathbb{R}^d \to \mathbb{R}^n$ such that

$$(S, T) = \underset{U \in \{\mathbb{R}^d \to \mathbb{R}^n\}, W \in \{\mathbb{R}^n \to \mathbb{R}^d\}}{\arg\min} \sum_{i=1}^{m} \|x - \tilde{x}\|^2$$

Note, we do not seek dimensionality reduction: if variances are non-neglectable everywhere $n = d$, and so be it.

Consider the example again. Let the set of all measurements be $D = \begin{pmatrix} \vec{i}_1^{\,T} \\ \vdots \\ \vec{i}_m^{\,T} \end{pmatrix}$ where each measurement $\vec{i}_j \in \mathbb{R}^n$ is given in standard basis $e_1, \ldots, e_n$ where $e_k = (\underbrace{0, \ldots, 0}_{k-1}, 1, \underbrace{0, \ldots, 0}_{n-k})^T$. Let's assume W is linear transformation that maps $\mathbb{R}^n \to \mathbb{R}^n$.

Then in its simplest form PCA asks: Is there another basis that best re-expresses our data set?

- What means "best basis to express data"?
- Our answer is that it is the basis that does not include redundant and irrelevant features!

What is Principal Component Analysis (PCA)?

So re-expressing the data means to find maps $W : \mathbb{R}^n \to \mathbb{R}^d$ where $d \leq n$, and $U : \mathbb{R}^d \to \mathbb{R}^n$ such that

$$(S, T) = \underset{U \in \{\mathbb{R}^d \to \mathbb{R}^n\}, W \in \{\mathbb{R}^n \to \mathbb{R}^d\}}{\arg\min} \sum_{i=1}^{m} \|x - \tilde{x}\|^2$$

Note, we do not seek dimensionality reduction: if variances are non-neglectable everywhere $n = d$, and so be it.

Consider the example again. Let the set of all measurements be $D = \begin{pmatrix} \vec{i}_1^{\,T} \\ \vdots \\ \vec{i}_m^{\,T} \end{pmatrix}$ where each measurement $\vec{i}_j \in \mathbb{R}^n$ is given in standard basis $e_1, \ldots, e_n$ where $e_k = (\underbrace{0, \ldots, 0}_{k-1}, 1, \underbrace{0, \ldots, 0}_{n-k})^T$. Let's assume W is linear transformation that maps $\mathbb{R}^n \to \mathbb{R}^n$.

Then in its simplest form PCA asks: Is there another basis that best re-expresses our data set?

- What means "best basis to express data"?
- Our answer is that it is the basis that does not include redundant and irrelevant features!

What is Principal Component Analysis (PCA)?

So re-expressing the data means to find maps $W : \mathbb{R}^n \to \mathbb{R}^d$ where $d \leq n$, and $U : \mathbb{R}^d \to \mathbb{R}^n$ such that

$$(S, T) = \underset{U \in \{\mathbb{R}^d \to \mathbb{R}^n\}, W \in \{\mathbb{R}^n \to \mathbb{R}^d\}}{\arg\min} \sum_{i=1}^{m} \|x - \tilde{x}\|^2$$

Note, we do not seek dimensionality reduction: if variances are non-neglectable everywhere $n = d$, and so be it.

Consider the example again. Let the set of all measurements be $D = \begin{pmatrix} \vec{i}_1^{\,T} \\ \vdots \\ \vec{i}_m^{\,T} \end{pmatrix}$ where each measurement $\vec{i}_j \in \mathbb{R}^n$ is given in standard basis $e_1, \ldots, e_n$ where $e_k = (\underbrace{0, \ldots, 0}_{k-1}, 1, \underbrace{0, \ldots, 0}_{n-k})^T$. Let's assume W is linear transformation that maps $\mathbb{R}^n \to \mathbb{R}^n$.

Then in its simplest form PCA asks: Is there another basis that best re-expresses our data set?

- What means "best basis to express data"?
- Our answer is that it is the basis that does not include redundant and irrelevant features!

Why Linear Transformations?

Are we looking for reduced basis? it is only good strategy if data D lies in a subspace of a $\mathbb{R}^n$. But maybe there is no subspace of $\mathbb{R}^n$ that contains D while there is a non-linear surface S that fits D, so why are we doing PCA?

Why Linear Transformations?

Are we looking for reduced basis? it is only good strategy if data D lies in a subspace of a $\mathbb{R}^n$. But maybe there is no subspace of $\mathbb{R}^n$ that contains D while there is a non-linear surface S that fits D, so why are we doing PCA?
For two reasons:

- Possibly our assumption about small subspace isn't true. But assuming that all data is in a subspace in $\mathbb{R}^n$ is a simplest possible hypothesis, so we need to try it, + it is easy to eliminate dimensions this way (unlike in the case of a non-linear subspace)
- What are we going to loose? in the worst case we decouple the basis and use basic vectors that pairwise linearly independent (i.e. non-redundunt features). We'll also learn that all new features are relevant.

Linear Algebra of PCA

- Let $\mathtt{D}$ be $m \times n$ data matrix with n features. Since we are only interested in variances of features we should center data to remove differences introduced by constant factor. Let now D stand for $\mathtt{D}^c$ (i.e. centered data $\mathtt{D}$).
- We need to find a linear transformation $P : \mathbb{R}^n \to \mathbb{R}^n$. Then under it new data represntation is another $m \times n$ matrix Y such that $D^T = PY^T$, where D is our data set.
- P must be one-to-one and since data D theoretically can be any point in $\mathbb{R}^n$, it has to be onto. Hence it has to be invertible. Then $D^T = PY^T$ is the represntation of data D relatively columns of P. Geometrically, P^{-1} is a rotation and a stretch which transforms D into Y.
- Columns of $P = (\vec{p}_1, \ldots, \vec{p}_n)$, are new basis for expressing the columns of D with weights (coordinates) from Y^T.
- For now let us assume that we are only looking for othronomal basis P (i.e. it is orthogonal basis such that vectors are unit vectors, similar to standard basis). For one it is a convenient type of basis. But we'll justify this requirement thoroughly later.
- We want new features to be non-redundunt, which by our definition means that the newly re-expressed data must be uncorreclated, i.e. for $Y = (\vec{y}_1, \ldots, \vec{y}_n)$ should hold that for all $i \neq j$ their covariance $cov_{ij} = 0$

Summary of The PCA problem

Desiderata: we need orthogonal $n \times n$ matrix P (i.e.P is square matrix with orthonormal columns) and $m \times n$ matrix Y such that a.) $D^T = PY^T$ and b.) covariance matrix C_Y has 0's everywhere off diagonal (i.e. have $cov_{ij} = 0$).

Known facts:

- For orthogonal matrices inverse is equal to transpose. Therefore $P^{-1} = P^T$.
- $C_Y = \frac{1}{m-1} Y^T Y = \frac{1}{m-1}(P^{-1}D^T)(P^{-1}D^T)^T = \frac{1}{m-1} P^T(D^T D)P = P^T C_D P$ where $C_D = \frac{1}{m-1} D^T D$ is covariance matrix of D.
- Observe that C_D is symmetric.
- Theorem (see e.g. D.Lay. Linear Algebra) An $n \times n$ symmetric matrix A has the following properties:
  1. A has n real eigenvalues, counting multiplicities.
  2. The dimension of the eigenspace for each eigenvalue λ equals the algebraic multiplicity of λ as a root of the characteristic equation.
  3. The eigenspaces are mutually orthogonal in the sense that eigenvectors corresponding to different eigenvalues are orthogonal.
  4. A is orthogonally diagonalizable, i.e. $A = F^{-1}SF = F^T SF$ where S is a diagonal matrix of eigenvalues and F is the orthogonal matrix of corresponding eigenvectors.

New Basis

C_D is symmetric so it is orthogoonally diagonalizable, i.e. $C_D = F^T S F$ where S is a diagonal matrix of eigenvalues of C_D and F is the orthogonal matrix of corresponding eigenvectors. Hence by taking $P = F^T$ we have

$$\begin{aligned} C_Y &= \frac{1}{m-1} Y^T Y \\ &= \frac{1}{m-1} \left((F^T)^{-1} D^T \right) \left((F^T)^{-1} D^T \right)^T \\ &= F \left(\frac{1}{m-1} D^T D \right) F^T \\ &= F C_D F^T = F(F^T S F) F^T \\ &= (F F^T) S (F F^T) = S \end{aligned}$$

Therefore we can take the orthogonal matrix of eigenvectors to satisfy our goals as a new basis!
Usually it makes no difference how we arrange eigenvalues in S, but in this case lets arrange them in order of magnitude i.e. $s_{ii} = \lambda_i \geq \lambda_{i+1} = s_{i+1,i+1}$. Respectively i^{th} column in F is an eigenvector of λ_i

Orthonormal Basis

Why does the basis have to be orthonormal even when $n \neq d$? Reminder - Desiderata:

- we'd like to eliminate irrelevant features, so we need a pair of linear maps $W : \mathbb{R}^n \to \mathbb{R}^d$ where $d \leq n$, and $U : \mathbb{R}^d \to \mathbb{R}^n$ such that $\tilde{x} = UWx$ have the same properties as x.
- By our assumption features determine properties. This means that $\tilde{x}$ must be as close as possible to x. In $\mathbb{R}^n$ closedness of points is determined by their Eucledean distance.
- Data contains many points so we need to minimize total distance between all pairs of data points and theire images under UW, or equivalently squares. In other words we need to obtain

$$(S, T) = \underset{U \in \left\{\mathbb{R}^d \to \mathbb{R}^n\right\}, W \in \left\{\mathbb{R}^n \to \mathbb{R}^d\right\}}{\arg\min} \sum_{i=1}^{m} \|x - \tilde{x}\|^2$$

To get the desired result we have
Theorem *Orthonormality* (see Shalev-Shwartz, Ben-David, p. 324).

Let (S, T) be a solution to the above equation. Then the columns of S are orthonormal (namely, $S^T S = I$ is the identity matrix of $\mathbb{R}^n$) and $T = S^T$.

Lecture Overview

PCA in a Nutshell

- The principal components (new features) are the eigenvectors of covariance matrix $C_D = \frac{1}{m-1} D^T D$
- Transformed data $Y = ((F^T D)^T = DF^T$ where transformation matrix $F = \left(\vec{f}_1, \ldots, \vec{f}_n\right)$ is orthogonal matrix of eigenvectors that are ordered according to order of eigenvalues.
- Covariance matrix C_Y is diagonal and the ith diagonal element is λ_i - ith biggest eigenvalue of $C_D = \frac{1}{m-1} D^T D$.
- ith diagonal element of C_Y is the variance of Y along $\vec{f}_i$.
- Diagonal elements in C_Y are ordered with respect to their value. This ordering form the energy spectrum of D.

PCA algorithm

1. to obtain D center original data matrix $\mathtt{D}$
2. compute covariance matrix $C_D = cov(D)$
3. Diagonalize C_D. If doing it by software then it returns matrices C_Y and F

   If doing it by hand+calculator then do the folowing steps.

   i. Find characteristic polynomial and its roots - they are eignevalues of C_D.
   ii. Construct diagonal matrix C_Y ordering diagonal elements in descending order
   iii. Find eigenvectors for eigenvalues to form transformation matrix F
4. Compute new representation of data (after transformation) $Y = ((F^T D)^T = DF^T$

Example of PCA by hand+ Wolfram Alpha

Suppose we have data frame of 3 attributes 4 records

$$\mathrm{D} = \begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \end{array} \begin{array}{c} \begin{array}{ccc} A_1 & A_2 & A_3 \end{array} \\ \begin{pmatrix} 1 & 2 & 1 \\ 4 & 2 & 13 \\ 7 & 8 & 1 \\ 8 & 4 & 5 \end{pmatrix} \end{array}$$

so multidimensional mean is $(5, 4, 5)$ and mean centered matrix is

$$\mathrm{D} = \begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \end{array} \begin{array}{c} \begin{array}{ccc} A_1 & A_2 & A_3 \end{array} \\ \begin{pmatrix} -4 & -2 & -4 \\ -1 & -2 & 8 \\ 2 & 4 & -4 \\ 3 & 0 & 0 \end{pmatrix} \end{array}$$

Note that we skip scaling here since variability of all features is measured in the same untis and is about the same

$$C_D = \frac{1}{4-1} D^T D = \begin{pmatrix} 10 & -6 & 0 \\ 6 & 8 & -8 \\ 0 & -8 & 32 \end{pmatrix}$$

Example of PCA by hand+ Wolfram Alpha

In Wolfram Alpha command
`charcteristic polynmial` $\{\{10,6,0\},\{6,8,-8\},\{0,-8,32\}\}$ returns
$768 - 556\lambda + 50\lambda^2 - \lambda^3$

Example of PCA by hand+ Wolfram Alpha

In Wolfram Alpha command
`charcteristic polynmial` $\{\{10,6,0\},\{6,8,-8\},\{0,-8,32\}\}$ returns
$768 - 556\lambda + 50\lambda^2 - \lambda^3$

Command `solve` $768 - 556\lambda + 50\lambda^2 - \lambda^3$ returns $\lambda_1 = 34.551$, $\lambda_2 = 13.843$ and $\lambda_3 = 1.60571$, so $S_Y = \begin{pmatrix} 34.551 & 0 & 0 \\ 0 & 13.4833 & 0 \\ 0 & 0 & 1.6057 \end{pmatrix}$

Example of PCA by hand+ Wolfram Alpha

In Wolfram Alpha command
`charcteristic polynmial` $\{\{10,6,0\},\{6,8,-8\},\{0,-8,32\}\}$ returns
$768 - 556\lambda + 50\lambda^2 - \lambda^3$

Command `solve` $768 - 556\lambda + 50\lambda^2 - \lambda^3$ returns $\lambda_1 = 34.551$, $\lambda_2 = 13.843$ and $\lambda_3 = 1.60571$, so $S_Y = \begin{pmatrix} 34.551 & 0 & 0 \\ 0 & 13.4833 & 0 \\ 0 & 0 & 1.6057 \end{pmatrix}$

Command `eigenvectors` $\{\{10,6,0\},\{6,8,-8\},\{0,-8,32\}\}$ returns

$v_1 = \begin{pmatrix} -0.0779385 \\ -0.318916 \\ 1 \end{pmatrix}$, $v_2 = \begin{pmatrix} 3.54356 \\ 2.26963 \\ 1 \end{pmatrix}$, and $v_3 = \begin{pmatrix} -2.771562 \\ 3.79929 \\ 1 \end{pmatrix}$.

Example of PCA by hand+ Wolfram Alpha

In Wolfram Alpha command
`charcteristic polynmial` $\{\{10,6,0\},\{6,8,-8\},\{0,-8,32\}\}$ returns
$768 - 556\lambda + 50\lambda^2 - \lambda^3$

Command `solve` $768 - 556\lambda + 50\lambda^2 - \lambda^3$ returns $\lambda_1 = 34.551$, $\lambda_2 = 13.843$ and $\lambda_3 = 1.60571$, so $S_Y = \begin{pmatrix} 34.551 & 0 & 0 \\ 0 & 13.4833 & 0 \\ 0 & 0 & 1.6057 \end{pmatrix}$

Command `eigenvectors` $\{\{10,6,0\},\{6,8,-8\},\{0,-8,32\}\}$ returns

$v_1 = \begin{pmatrix} -0.0779385 \\ -0.318916 \\ 1 \end{pmatrix}$, $v_2 = \begin{pmatrix} 3.54356 \\ 2.26963 \\ 1 \end{pmatrix}$, and $v_3 = \begin{pmatrix} -2.771562 \\ 3.79929 \\ 1 \end{pmatrix}$.

These vectors are orthogonal but not normal. So they need to be normalized.

Commands `normalize` $(-0.0779385, -0.318916, 1)$, `normalize` $(3.54356, 2.26963, 1)$, and `normalize` $(-2.771562, 3.79929, 1)$ return vectors in matriix

$$F = \begin{pmatrix} -0.07405 & 0.819267 & -0.576457 \\ -0.303005 & 0.524736 & 0.790214 \\ 0.950108 & 0.231199 & 0.20799 \end{pmatrix}$$

When to use different functions in **R**

Functions `prcomp` and `princomp`

- The difference `prcomp` use sample covariance formula $\frac{1}{m-1}D^TD$ while `princomp` uses covariance formula $\frac{1}{n}D^TD$ applicable when theoretical mean is known
- Use the first on empirical data frame where distribution of attributes is unknown and the second when sample is drawn from a known theoretical distribution (for example using `rnorm` in **R**).
- In the latter case known expectation should be supplied to the function
- For `prcomp` parameters are Data frame, 'retx' = true/false - indicates if rotated varible should be returned, 'scale' and 'center' - both true/false, indicate reuest for respective normalization
- The `prcomp` returns `$rot` the matrixx of eigenvectors (aka loadings), `$sdev` suare roots of eigenvalues of covariance = st deviations of principal components + whatever the results of normalizations where requested

Example of Computing PCA Rotation in R

```
attach(iris)
d <- -dim(iris)[2] - 1 # number of features/components
prc <- -prcomp(iris[1 : d], retx = TRUE, center = TRUE, scale. = TRUE)
prc$x # rotated data
prc$sdev # vector of standard deviations of principal components
prc$center # means of iris data features
prc$scale # total z-score of each feature
prc$rot # eigenvectors - new basis
biplot(prc, scale = 0) # plots 2 components with largest variance
                       # note that Petal Length is roughly PC1 and
                       # Sepal Width is roughly correponding to -PC2;
spectrum <- -prc$sdev^2 # vector of variances (eigenvalues) along new basis
totvar <- -sum(spectrum); totvar # total variance in new basis
```