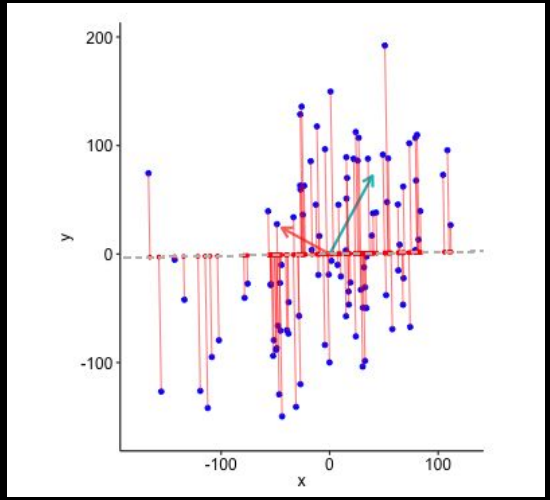
Dimensionality reduction



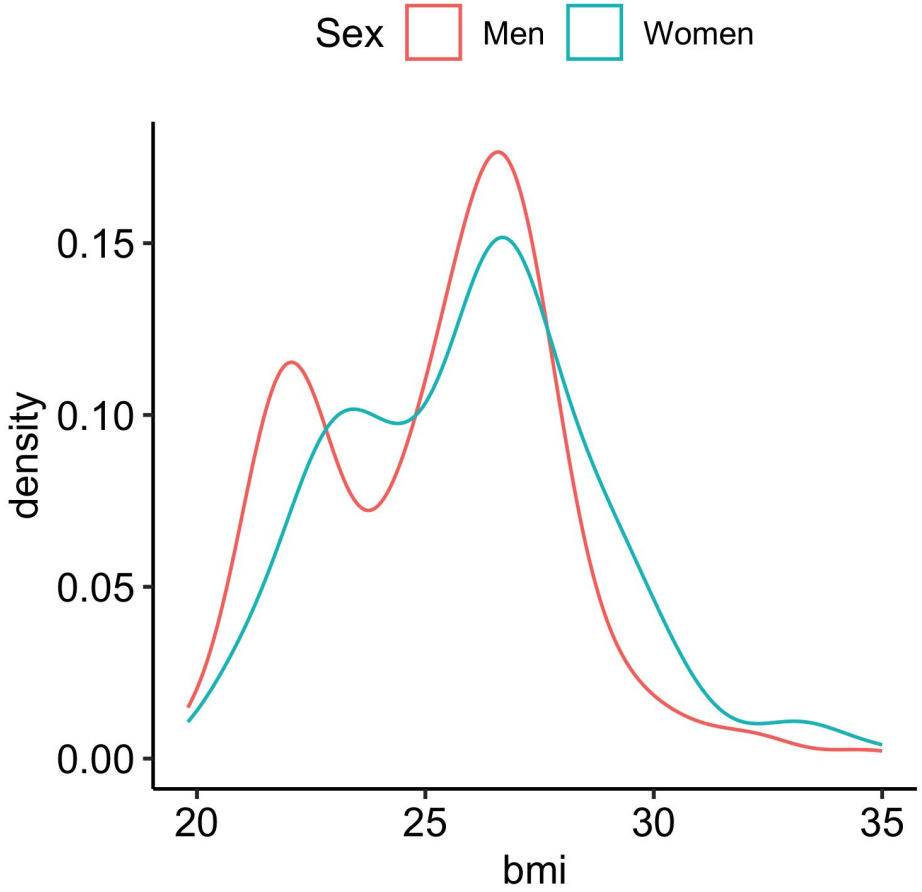
Saket Choudhary
saketc@iitb.ac.in

Introduction to Public Health Informatics
DH 302
Lecture 08 || Friday, 31st January 2025

From last lecture…

- Testing for difference of means
- **Dimensionality reduction primer**

Testing for difference of means



<u>Question</u>: Is there statistically significant difference in mean between men and women BMI?

What is the null hypothesis?

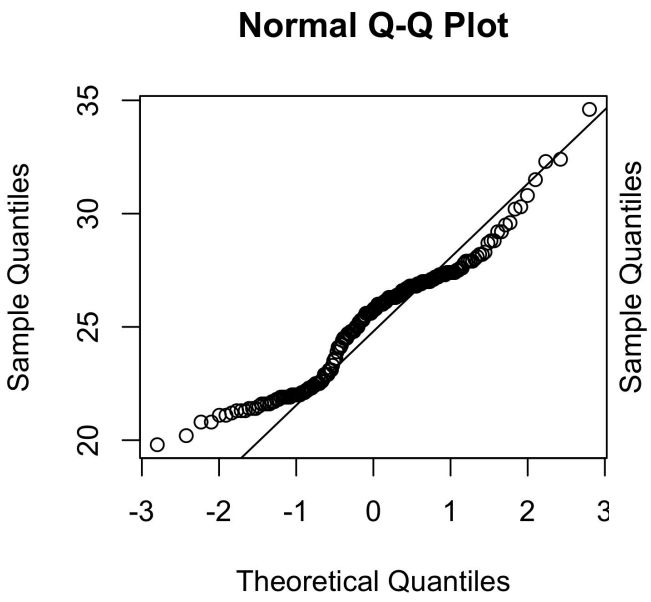
Null Hypothesis: The mean bmi is same for men and womean

**Data shows mean BMI distribution across countries in 2017

Data source

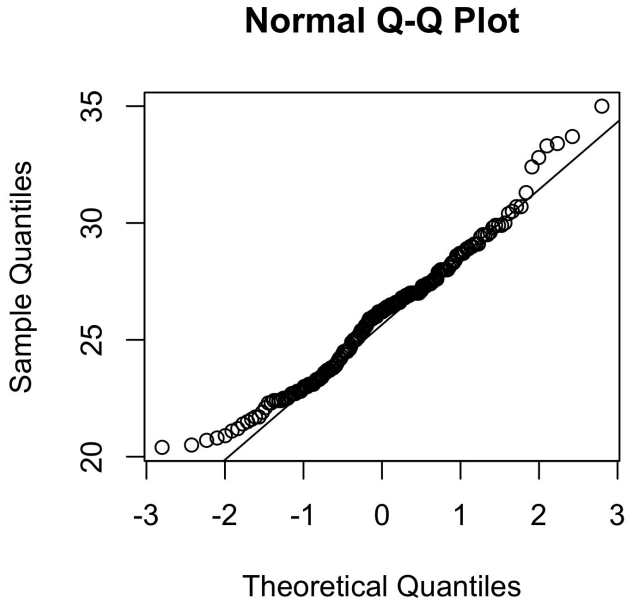
Verifying the ‘normality’ assumption using QQplot

```
qqnorm(men_rural_2017$bmi)
qqline(men_rural_2017$bmi)
```



Men

```
qqnorm(women_rural_2017$bmi)
qqline(women_rural_2017$bmi)
```



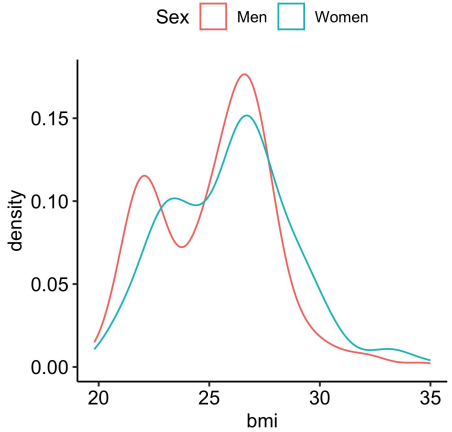
Women

- We can test the normality assumption by plotting the sample quantiles against the expected “theoretical” quantiles of a standard normal in a QQ plot
- QQ plot is the plot of the quantiles of the first dataset against the quantiles on the second dataset
- What is a Quantile? Fraction of points below the given value
- For example, 30% quantile value represents that 30% of the values in dataset lie below this point (and 70% lie above)

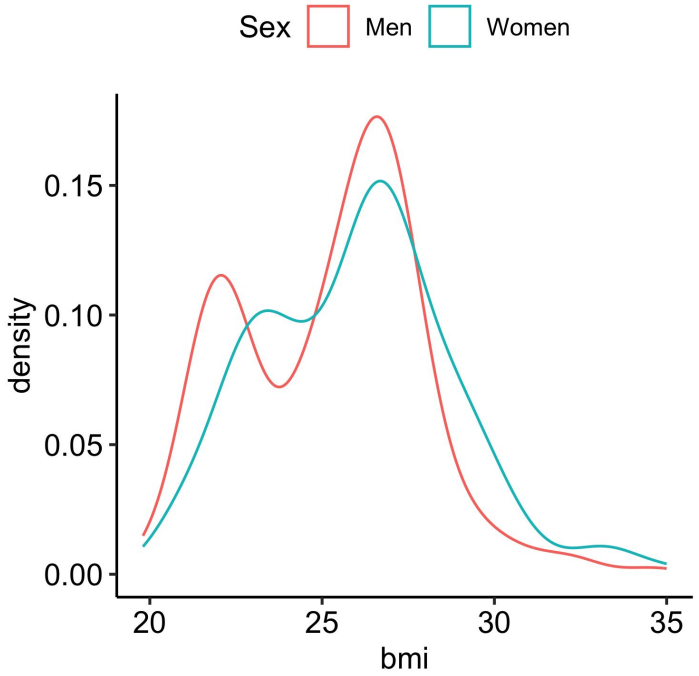
T-test: BMI differences between males and females are statistically significant?

```
> t.test(x = men_rural_2017$bmi, y = women_rural_2017$bmi)

    Welch Two Sample t-test

data:  men_rural_2017$bmi and women_rural_2017$bmi
t = -2.6774, df = 388.7, p-value = 0.007734
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -1.2741967 -0.1951911
sample estimates:
mean of x mean of y
 25.24490  25.97959
```

We reject the null hypothesis that the mean bmi of men and women is equal



What explains the bimodality?



Problem

Twenty-two volunteers at a cold research institute caught a cold after having been exposed to various cold viruses. A random selection of 10 of these volunteers was given tablets containing 1 gram of vitamin C. These tablets were taken four times a day. The control group consisting of the other 12 volunteers was given placebo tablets that looked and tasted exactly the same as the vitamin C tablets. This was continued for each volunteer until a doctor, who did not know if the volunteer was receiving the vitamin C or the placebo tablets, decided that the volunteer was no longer suffering from the cold. The length of time the cold lasted was then recorded.

Treated with Vitamin C	Treated with Placebo
5.5	6.5
6.0	6.0
7.0	8.5
6.0	7.0
7.5	6.5
6.0	8.0
7.5	7.5
5.5	6.5
7	7.5
6.5	6.0
	8.5
	7.0

This can further be summarized as follows:

	Treated with Vitamin C	Treated with Placebo
$\bar{y}$	6.45	7.125
n	10	12
SD	0.761	0.882
SE	0.240	0.254

In the context of this study, state the null and alternative hypotheses.

Test the hypothesis.

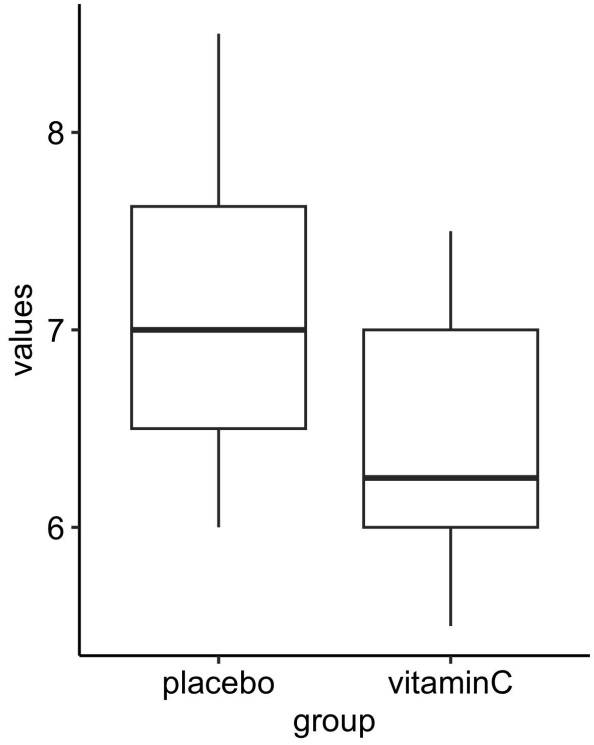
Problem

Twenty-two volunteers at a cold research institute caught a cold after having been exposed to various cold viruses. A random selection of 10 of these volunteers was given tablets containing 1 gram of vitamin C. These tablets were taken four times a day. The control group consisting of the other 12 volunteers was given placebo tablets that looked and tasted exactly the same as the vitamin C tablets. This was continued for each volunteer until a doctor, who did not know if the volunteer was receiving the vitamin C or the placebo tablets, decided that the volunteer was no longer suffering from the cold. The length of time the cold lasted was then recorded.

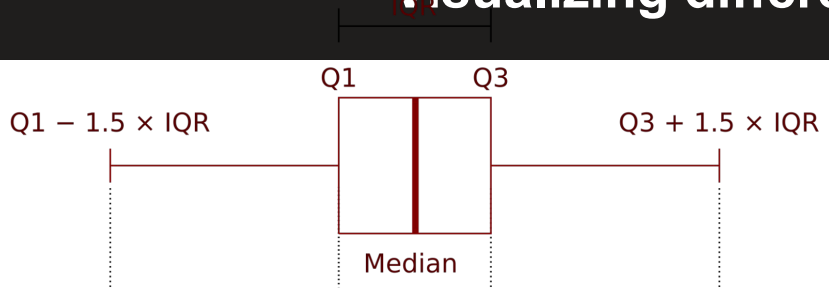
Treated with Vitamin C	Treated with Placebo
5.5	6.5
6.0	6.0
7.0	8.5
6.0	7.0
7.5	6.5
6.0	8.0
7.5	7.5
5.5	6.5
7	7.5
6.5	6.0
	8.5
	7.0

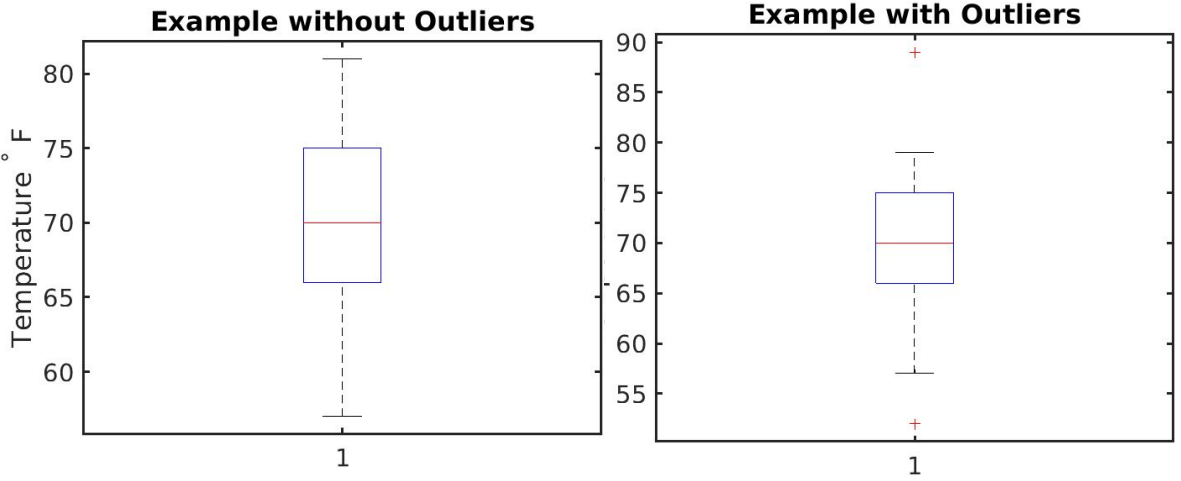
This can further be summarized as follows:

	Treated with Vitamin C	Treated with Placebo
$\bar{y}$	6.45	7.125
n	10	12
SD	0.761	0.882
SE	0.240	0.254



Visualizing difference of means





A box plot of the data set can be generated by calculating five relevant values of this data set: minimum, maximum, median (Q_2), first quartile (Q_1), and third quartile (Q_3).

Q_1= Quantity such that exactly 25% entries are less than equal to this quantity

Q_2= Median

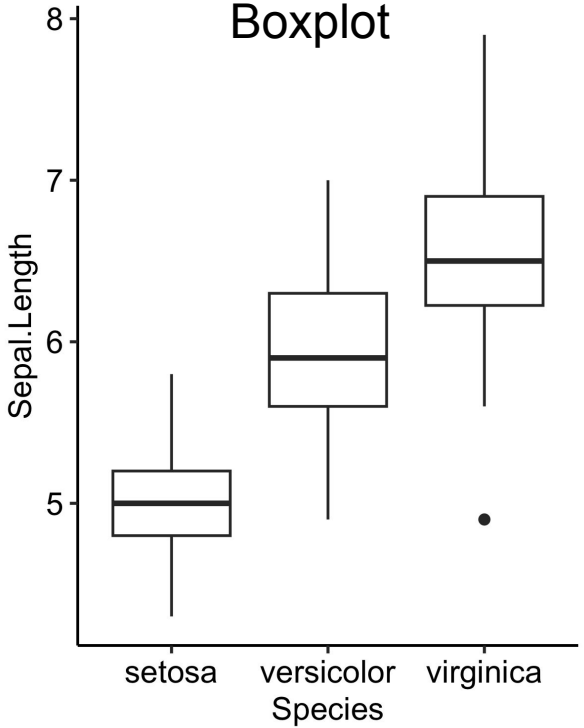
Q_3 = Quantity such that exactly 75% entries are less than equal to this quantity

IQR = Q_3 - Q_1

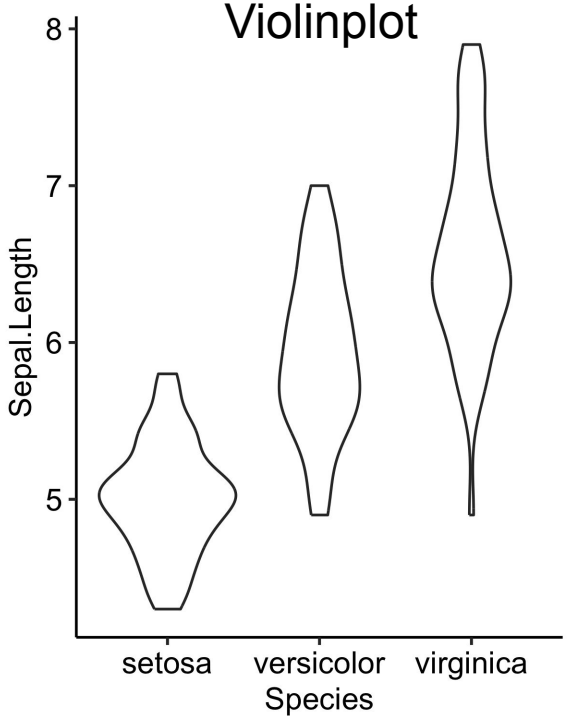
Points beyond **Q_3 + 1.5IQR or Q_1- 1.5IQR** are outliers

What's so special about 1.5?

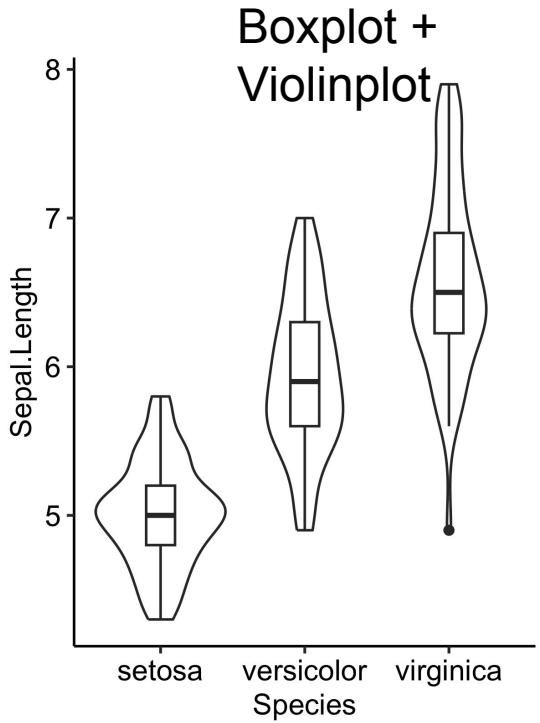
Visualizing difference of means - Boxplots and Violin plots




```
ggplot(iris,
aes(Species, Sepal.Length)) + geom_boxplot()
```

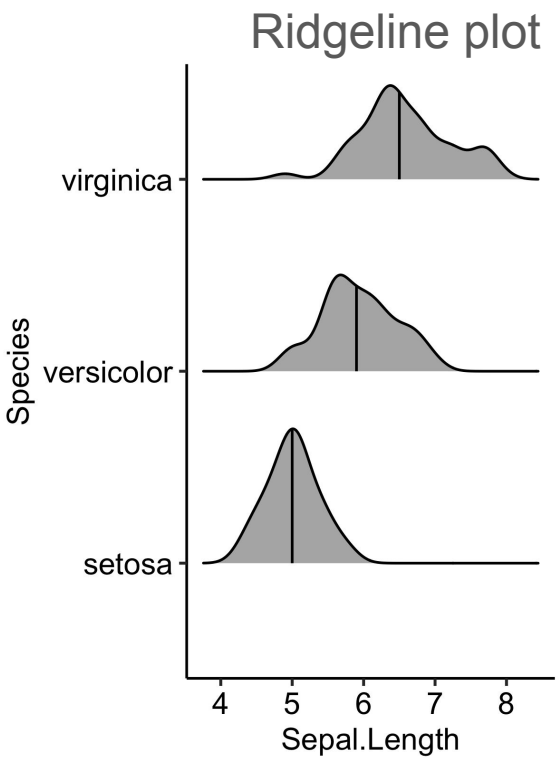



```
ggplot(iris,
aes(Species, Sepal.Length)) +
geom_violin()
```

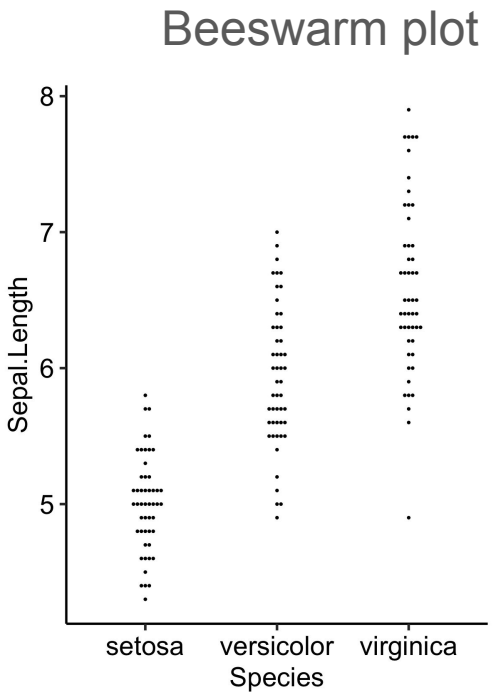



```
ggplot(iris,
aes(Species, Sepal.Length)) +
geom_violin() +
geom_boxplot(width=0.2)
```

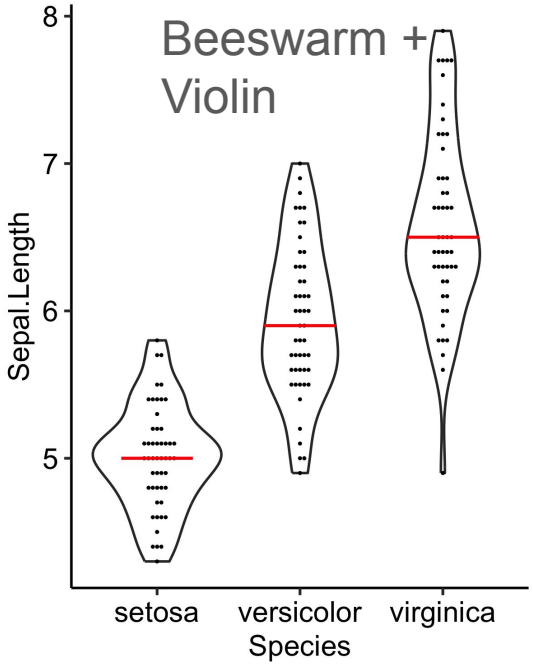
Visualizing difference of means - Bringing in more granularity




```
ggplot(iris,aes(x = Sepal.Length, y=Species)) +
    ggridges::stat_density_ridges(quantile_lines = T,
                                  quantiles = 2,
                                  scale=0.7)
```



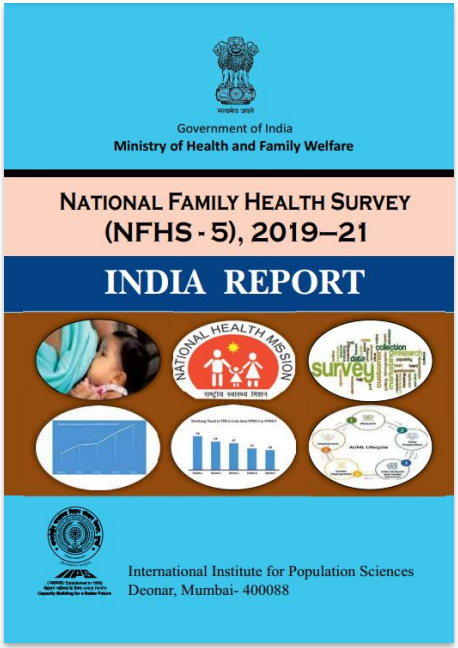

```
ggplot(iris,aes(Species,
Sepal.Length))  +

ggbeeswarm::geom_beeswarm(size=
0.1)
```




```
ggplot(iris,aes(Species, Sepal.Length))  +
geom_violin() +
    geom_beeswarm(size=0.1) +
    ggbeeswarm::geom_beeswarm( size = 0.1)
+
    stat_summary(
        geom = "crossbar", color = "red",
        fun = median, fun.min = median,
fun.max = median,
        fatten = 1.2, width = 0.5
```

Dimensionality reduction

Why reduce the dimensionality at all?



131 variables measured across states (districts)

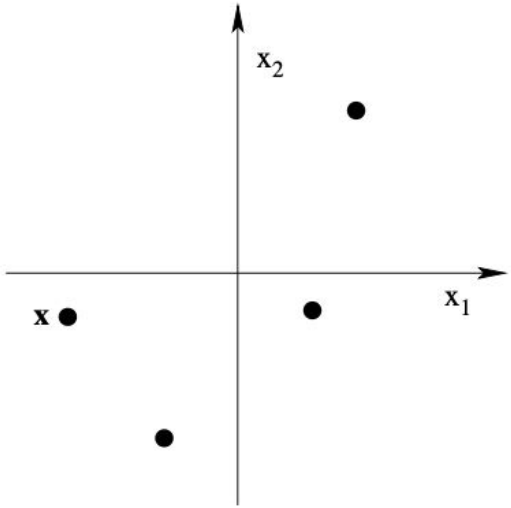
state_code	indicator	nfhs5_urban	nfhs5_rural	nfhs5_total	nfhs4_total
	1. Female population age 6 years and above who ever attended school (%)	82.5	66.8	71.8	68.8
	2. Population below age 15 years (%)	23.1	28.1	26.5	28.6
	3. Sex ratio of the total population (females per 1,000 males)	985.0	1037.0	1020.0	991.0
	4. Sex ratio at birth for children born in the last five years (females per 1,000 males)	924.0	931.0	929.0	919.0
	5. Children under age 5 years whose birth was registered with the civil authority (%)	93.3	87.5	89.1	79.7
	6. Deaths in the last 3 years registered with the civil authority (%)	83.2	65.8	70.8	
	7. Population living in households with electricity (%)	99.1	95.7	96.8	88.0
	8. Population living in households with an improved drinking-water source1 (%)	98.7	94.6	95.9	94.4
	9. Population living in households that use an improved sanitation facility2 (%)	81.5	64.9	70.2	48.5
	10. Households using clean fuel for cooking3 (%)	89.7	43.2	58.6	43.8
	11. Households using iodized salt (%)	96.9	93.0	94.3	93.1
	12. Households with any usual member covered under a health insurance/financing scheme (%)	38.1	42.4	41.0	28.7
	13. Children age 5 years who attended pre-primary school during the school year 2019-20 (%)	18.1	12.0	13.6	
	14. Women who are literate4 (%)	83.0	65.9	71.5	

Data

Is there a global structure in the National survey data?
Are different states performing similarly on multiple health metrics?

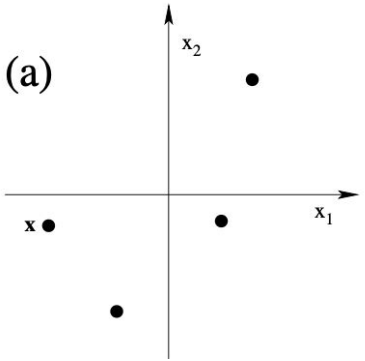
The Goal - 1D PCA

- Reduce the “dimensionality” of data: represent the data using a small set of numbers per instance
- Perform a linear transformation of the data such that we end up retaining the maximum information
- Ideally we should be able to use the reduced dimensional space to reconstruct the original high dimensional space suffering minimum error in the process



The Goal - 1D PCA

Goal: Reduce the dimensionality of data by projecting the data into lower dimensional space such that the reconstruction error is minimized

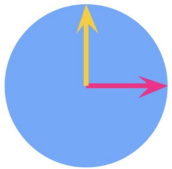


data points in 2D

Source

LinAlg101: Bases of a vector space

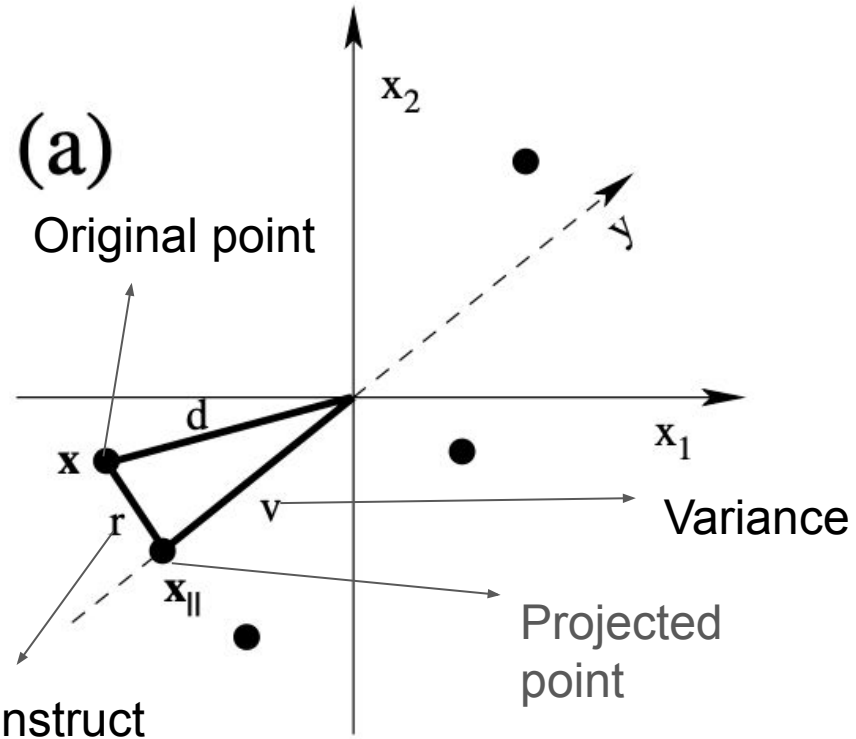
- Basis of a vector space: Consists of vectors that span the vector space
- A vector space can have multiple bases vectors but all of them have same number of elements → the dimension of the vector space
- Any vector in R^n can be represented as a sum of the vectors (1,0,0,...,0), (0,1,0,..,0), (0,0,0,1,...,0) … (0,0,0,0,...,1) → standard basis vector of R^n



(standard) basis vectors of R2

- **Orthogonality**: Two vector x,y are orthogonal if their dot product x . y = 0
- **Orthogonal basis**: A basis with all basis vectors are mutually orthogonal
- **Orthonormal basis**: An orthogonal basis where each vector has norm 1

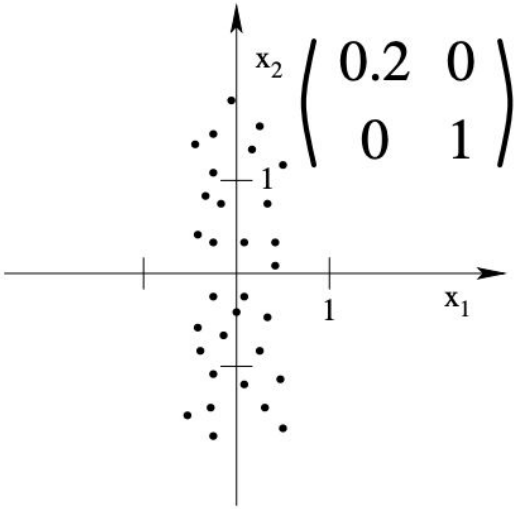
Minimizing reconstruction error = Maximizing variance



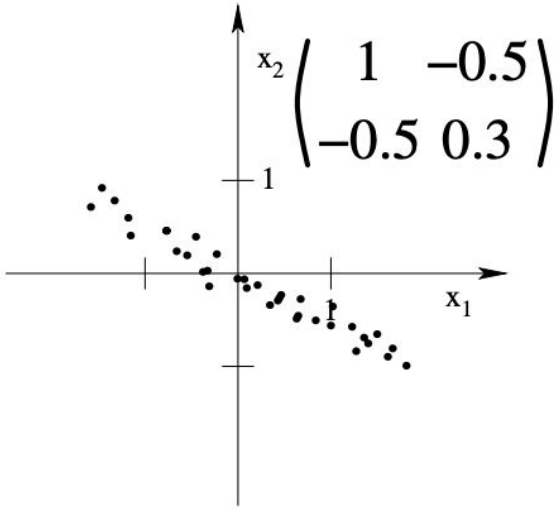
How can we determine the direction of maximal variance?

What does covariance matrix tell you about the data?

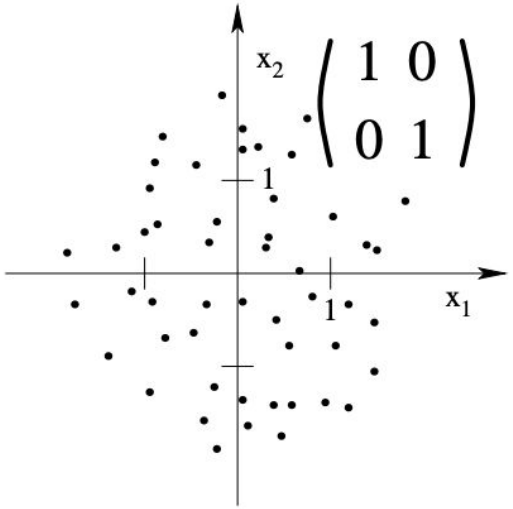
$$\mathrm{K}_{X_i X_j} = \mathrm{cov}[X_i, X_j] = \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])]$$



Data aligned with axes and covariance is diagonal

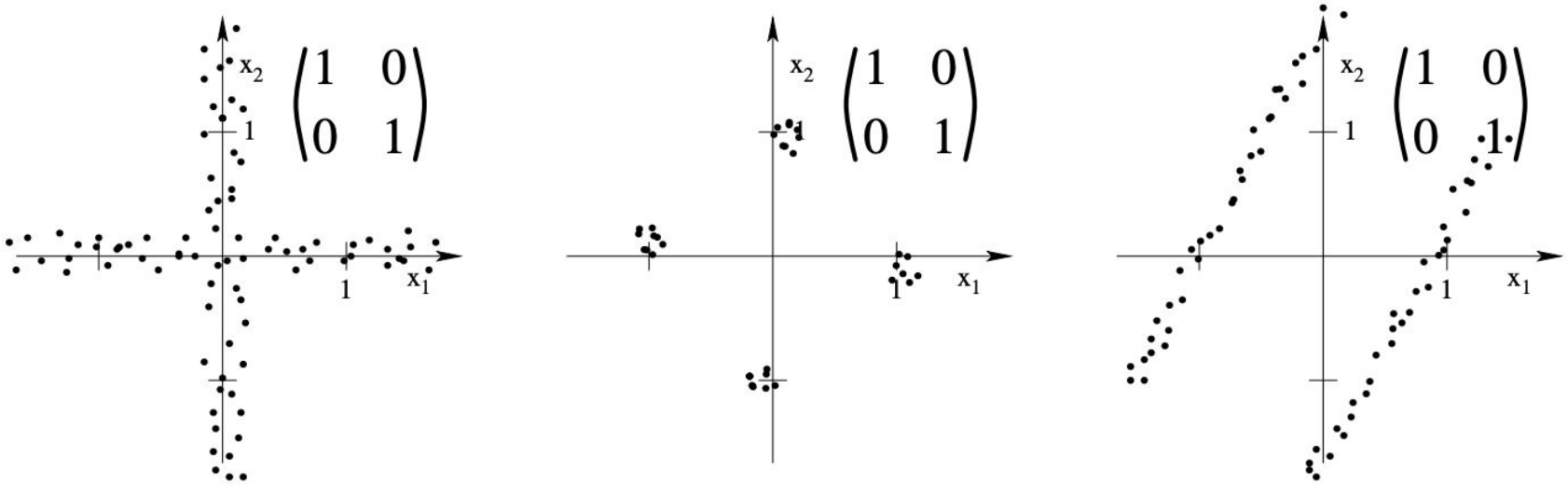


Data oblique wrt axies and covariance is diagonal



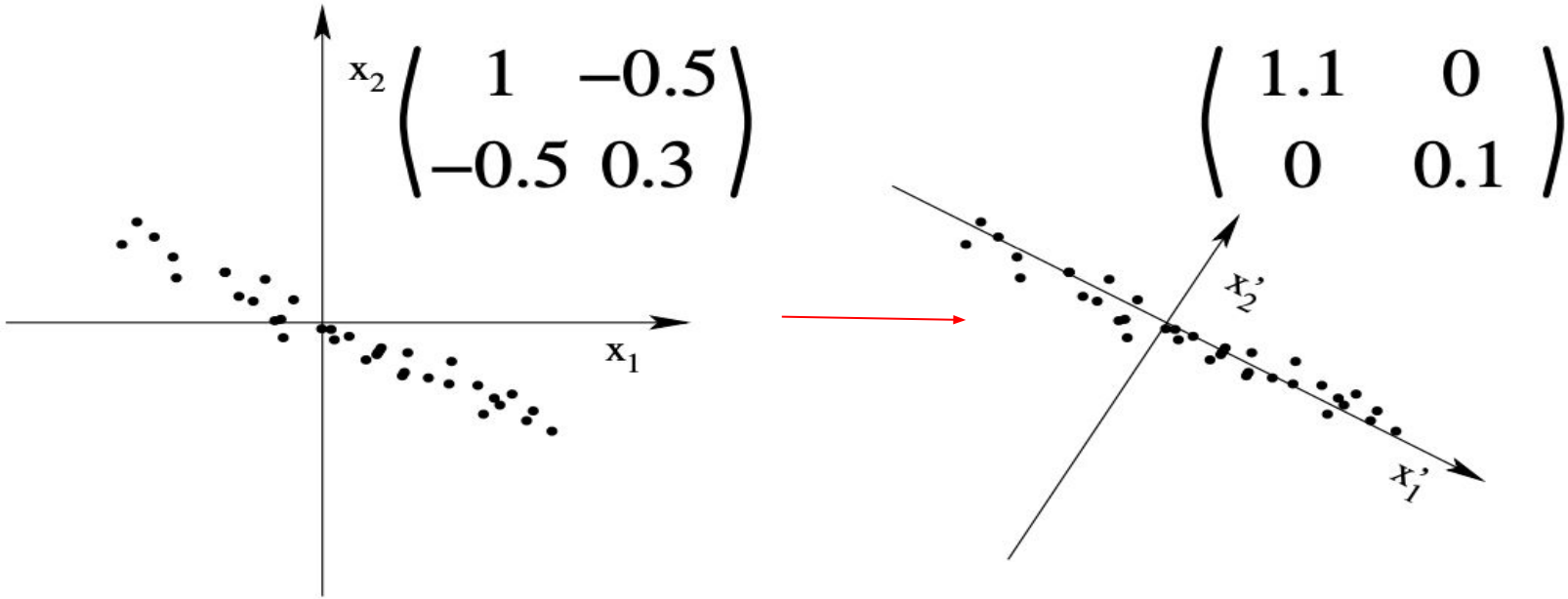
Gaussian cloud

Covariance matrix captures the general extent of data



Different distributions with same covariance matrix

PCA rotates the axes to diagonalize the covariance matrix



SVD Theorem

An arbitrary matrix $A \in \mathbb{R}^{m\times n}$ admits a decomposition of the form

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T = U\tilde{S}V^T, \quad \tilde{S} := \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix},$$

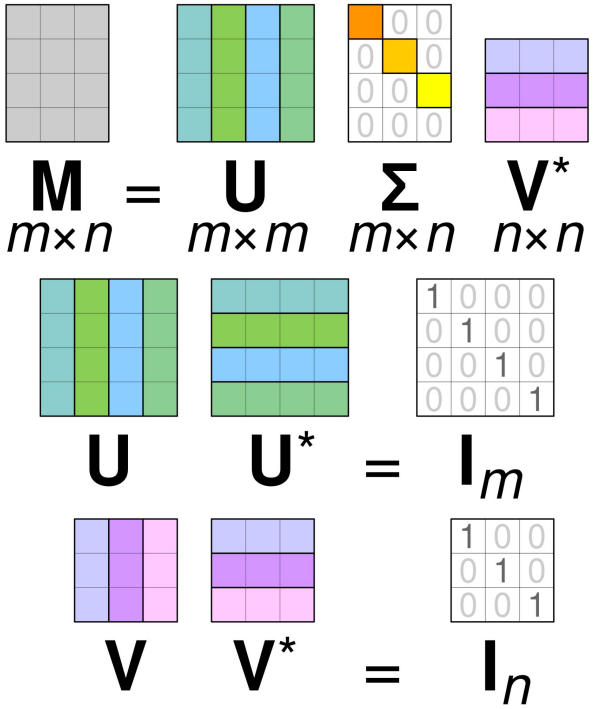
where $U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{n\times n}$ are both orthogonal matrices, and the matrix S is diagonal:

$$S = \mathbf{diag}(\sigma_1, \ldots, \sigma_r),$$

where the positive numbers $\sigma_1 \geq \cdots \geq \sigma_r > 0$ are unique and are called the *singular values* of A. The number $r \leq \min(m, n)$ is equal to the rank of A, and the triplet $(U, \tilde{S}, V)$ is called a *singular value decomposition* (SVD) of A. The first r columns of U: $u_i, i = 1, \ldots, r$ (resp. V: $v_i, i = 1, \ldots, r$) are called left (resp. right) singular vectors of A, and satisfy

$$Av_i = \sigma_i u_i, \quad u_i^T A = \sigma_i v_i, \quad i = 1, \ldots, r.$$

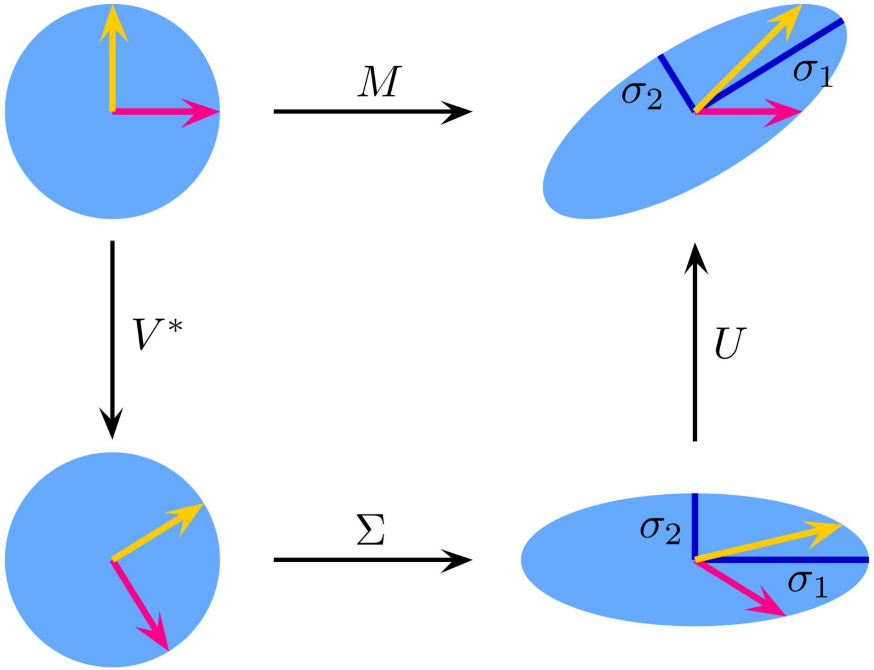
Singular value decomposition



- Input: A matrix
- Output: a set of numbers called *singular values* and two collections of vectors: a set of *right singular vectors* and another set of *left singular vectors*.

Source

Geometry of Singular value decomposition



$$M = U \cdot \Sigma \cdot V^*$$

- A given matrix M transforms the unit vectors into an ellipse
- This can be imagined as
  - 1. Performing rotation of the unit vectors by V
  - 2. Scaling these vectors by scaling factors (singular values of the matrix M)
  - 3. Performing another rotation by U

Source

Singular value decomposition

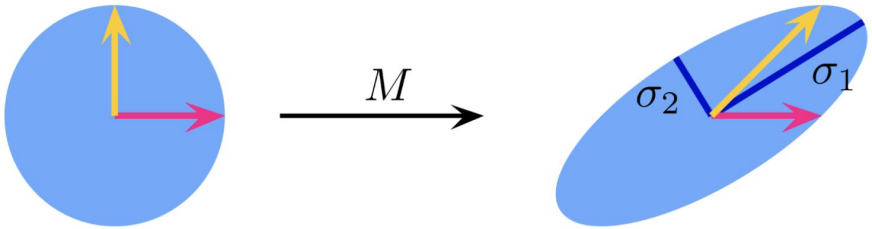
$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & 0 & 2 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{U} = \begin{bmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{bmatrix} \quad \mathbf{\Sigma} = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 \\ 0 & \sqrt{5} & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{0} & 0 \end{bmatrix} \quad \mathbf{V}^* = \begin{bmatrix} 0 & 0 & -1 & 0 & 0 \\ -\sqrt{0.2} & 0 & 0 & 0 & -\sqrt{0.8} \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -\sqrt{0.8} & 0 & 0 & 0 & \sqrt{0.2} \end{bmatrix}$$

- The matrix *M* is rectangular.
- There are three non-zero singular values.
- The rank of *M* is 3
- $UU^T = I$ and $VV^T = I$ where I is the unit vector

Source

What are singular values?

- A m x n matrix M can be thought of as mapping a vector *x from* R^n to R^m.
- A unit sphere in R^n is mapped to an ellipsoid in R^m:



- The non-zero *singular values* of *M* are the lengths of the *semi-axes* of the ellipsoid.

Eckart Young Theorem

- Let $M = U\Sigma V^T$ and set $M_r = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \ldots + \sigma_r u_r v_r^T$
- Then M_r is the best low rank (r) approximation of M

- What does the theorem really say?
  - If you want a lower rank matrix, you can perform a SVD and retain the top singular values (and vectors)
  - Top singular values which measure the sizes of the largest of the semi-axes of the resulting ellipsoid from the transformation capture “most of the original information”

Eckart and Young, 1936

What is eigendecomposition of a matrix?

- A n x n square matrix A is diagonalizable if it can be written in the form $A = PDP^{-1}$ where D is a diagonal matrix, alternatively AP = PD

$$D = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{pmatrix} \qquad P = \begin{pmatrix} | & | & & | \\ \mathbf{p}_1 & \mathbf{p}_2 & \cdots & \mathbf{p}_n \\ | & | & & | \end{pmatrix}$$

- Decomposition of a square matrix A into the form $A\mathbf{p}_i = \lambda_i\mathbf{p}_i$ is called eigendecomposition as columns of P, i.e. $\mathbf{p}_i$ are eigen vectors of A and λ_i are eigenvalues

Singular values to eigenvalues (SVD to eigendecomposition)

- SVD of $M = U\Sigma V^T$. Consider $M^TM \rightarrow M^TM = (V\Sigma U^T)(U\Sigma V^T) = V\Sigma^2V^T = V\Sigma^2V^{-1}$
- $V\Sigma^2V^T$ is the eigendecomposition of M^TM
- Thus, singular values of M = square root of the eigenvalues of M^TM
- Recipe (not best) to calculate SVD of a matrix M:
  - Compute M^TM
  - Compute eigenvalues of M^TM, σ = singular values of (M) = square root of eigenvalues of M^TM
  - Find eigenvectors of $M^TM \rightarrow V$
  - Find eigenvectors of $MM^T \rightarrow U$

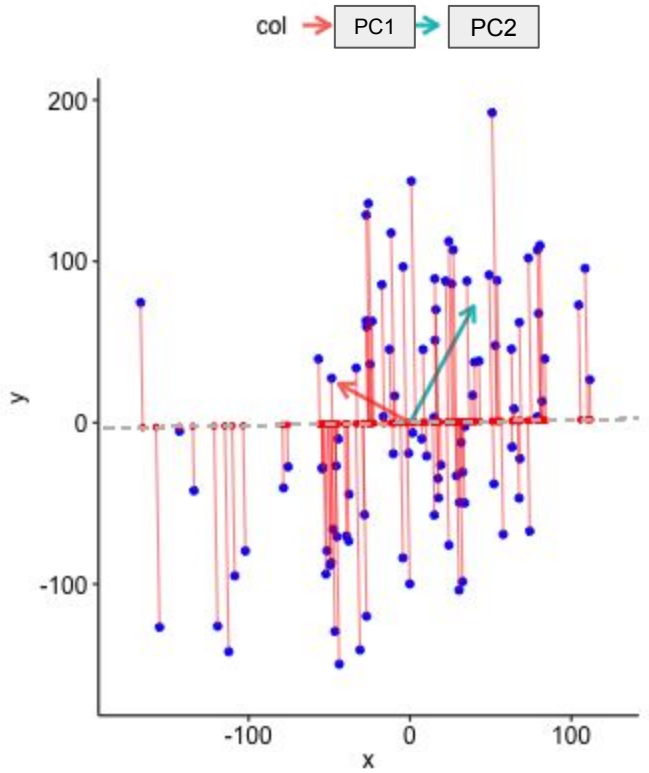
Eigendecomposition to SVD (eigenvalues to singular values)

- SVD is a generalization of eigendecomposition (which is only defined for square matrices)
- **Singular values** of a matrix M are the **square roots of the eigen values of MM^T**
- If M is a square matrix, with non-negative eigenvalues, then the above singular values and eigen values will be same

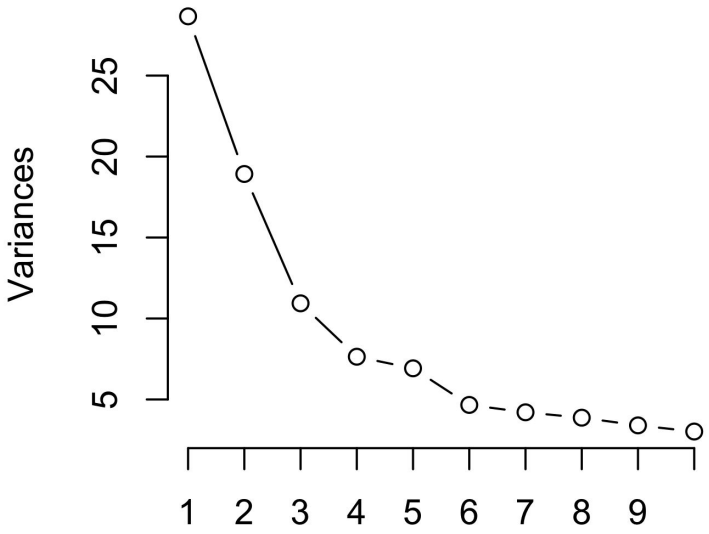
Principal Component Analysis - The recipe

- Start with a data matrix *M.*
- *Center M* by subtracting the column means (each column is a feature)
- Perform SVD of M $\rightarrow$ $M = U\Sigma V^T$
  - *U and V are orthonormal*
  - Σ is a diagonal matrix of singular values
  - *V* is made of eigenvectors that diagonalize the covariance matrix M^TM.
- Truncate V_k to retain the first k columns
  - $M_k = U_k\Sigma V_k^T$ is a good low rank (k) approximation of *M.*
- "Project" the original matrix M onto V_k: MV_k
  - This projection has two properties:
    - It maximises the variance of projected points
    - It results in minimum reconstruction error if the original matrix is to be reconstructed

PCA: the optimization



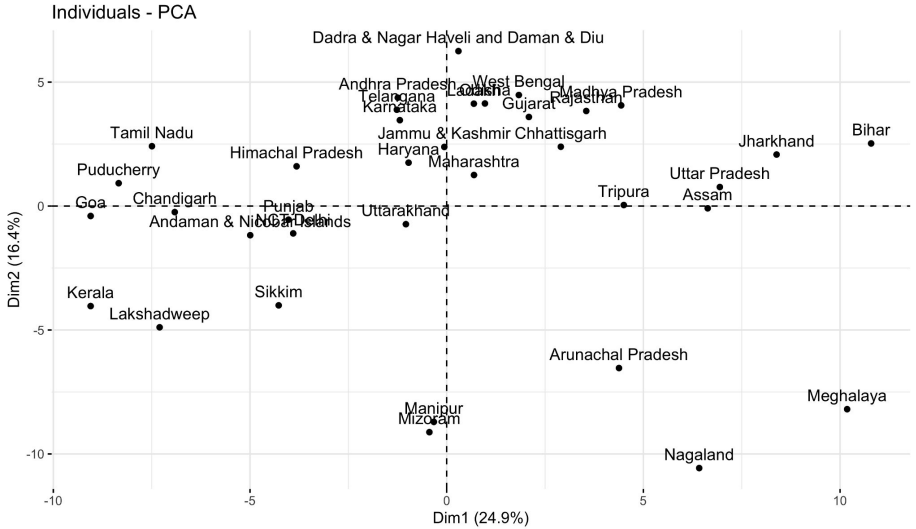
NFHS5 PCA reveals spatial clusters



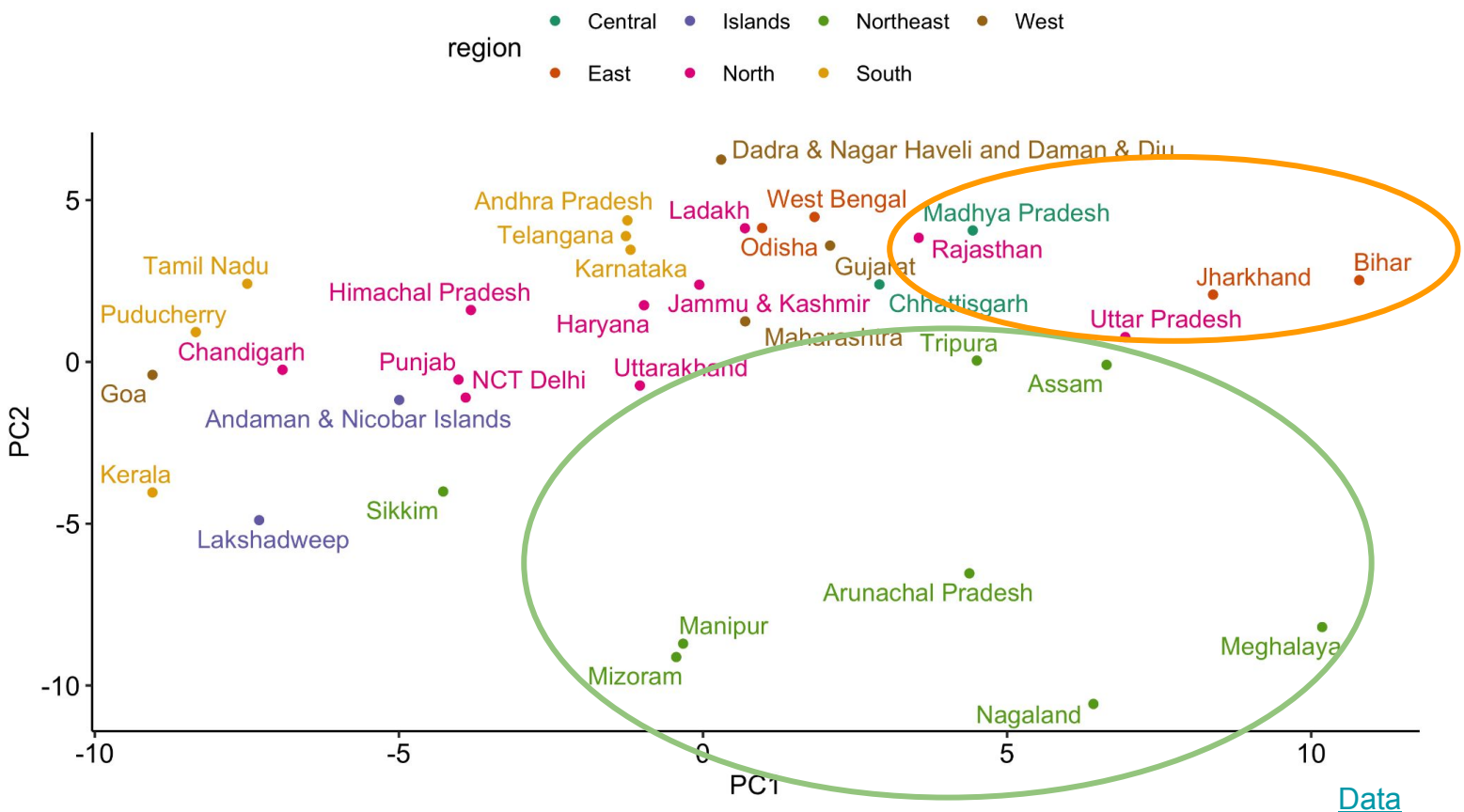
Variance explained by PC
= Eigen value □

Percentage variance explained = 100*□$_i$/Σ□$_i$

```
pca <- prcomp(df, scale = TRUE)
df.pca <- as.data.frame(pca$x)
df.pca$state <- rownames(df.total.long.noindia.nona)
df.pca <- df.pca %>% select(state, PC1, PC2)
ggplot(df.pca, mapping = aes(x = PC1, y = PC2)) +
  geom_point() +
  geom_text_repel(aes(label = state), hjust = 0, vjust = 0)
# For plotting, alternatively
factoextra::fviz_pca_ind(pca)
```



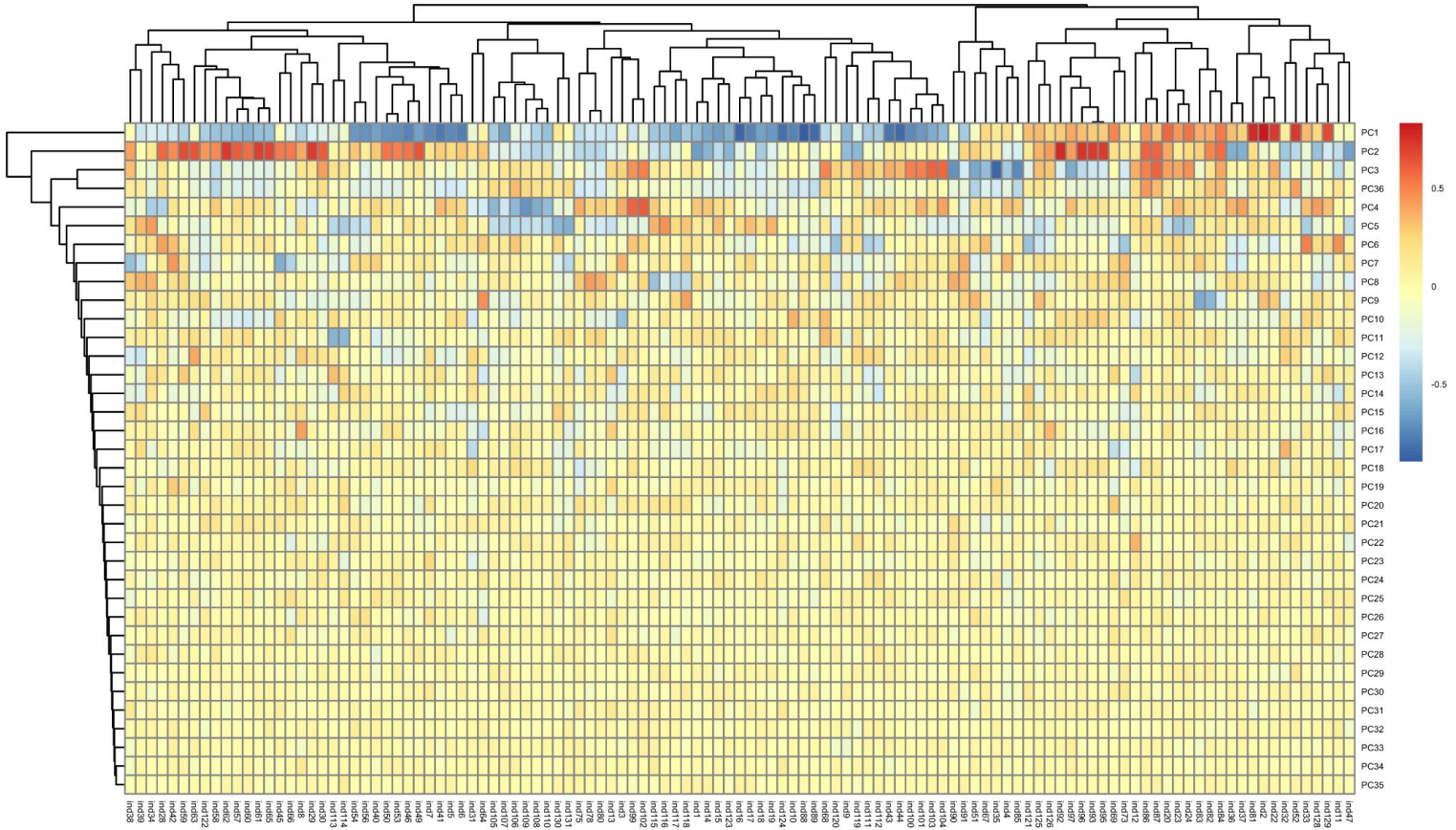
NFHS5 PCA reveals spatial clusters



Data

NFHS5 PCA reveals spatial clusters

Correlation between PCs and original variables



Questions?