

Recap

- Univariate distributions
- Delete outliers, StandardScaler
- Retain outliers, RobustScaler
- Or use distribution transformation on features to make outliers into inliers
 - Log Transformer
 - FunctionTransformers
 - PowerTransformers – Box Cox, Yeo Johnson

3rd method - Analyzing outliers for detection

- Focus on outliers and not regular data
- No deletion or scaling
- Exclusively analyze data > 2.5 standard deviation or later by applying specific mathematical approaches
- Interesting & relevant topic in industry
- We will only be briefly looking at this 😞

Cons of our approach so far (from outlier perspective)

- Can we just look at the standard deviations or IQR of each feature individually?
- Each data point is multivariate in reality
- Analysis also needs to be holistic
- Enter multivariate probability density functions

Why look at probability at all in ML course?

- Needed for generative ML (soon in Sem1)
 - Maximum Likelihood Estimation (MLE/MAP)
- Needed for information theory refresher
 - Used in Decision Trees & ensembles
 - Used in Feature Selection
- Generative AI
- Probabilistic Deep Learning (Bayesian Neural Networks)
- Using this opportunity to introduce multivariate distributions

EfficientNet trained on ImageNet images



Prediction: Typewriter
Certainty: 85%

**ImageNet is a 1000
class dataset !!**



Prediction: Stonewall
Certainty: 87%



Multivariate distributions

Expected value

- Weighted average to probability based formulation

$$\mathbb{E}[X] = \sum_x xp(x)$$

$$\mathbb{E}[f(X)] = \sum_x f(x)p(x)$$

$$\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y]$$

- Why Expectation?
 - Hint: Linear operator

Standard deviation method (contd.)

- Standard deviation is the typical deviation of feature value from mean

Put μ for $E[X]$
and then
expand to see
for yourself

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2} = \frac{\|x - \mu \mathbf{1}\|}{\sqrt{n}}$$

$$\sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] + \mathbb{E}[\mathbb{E}[X]]^2 - 2\mathbb{E}[X\mathbb{E}[X]]$$

Note for M.E
students. This
is the Linear
Algebra
equivalent

$$= \mathbb{E}[X^2] + \mathbb{E}[X]^2 - 2\mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

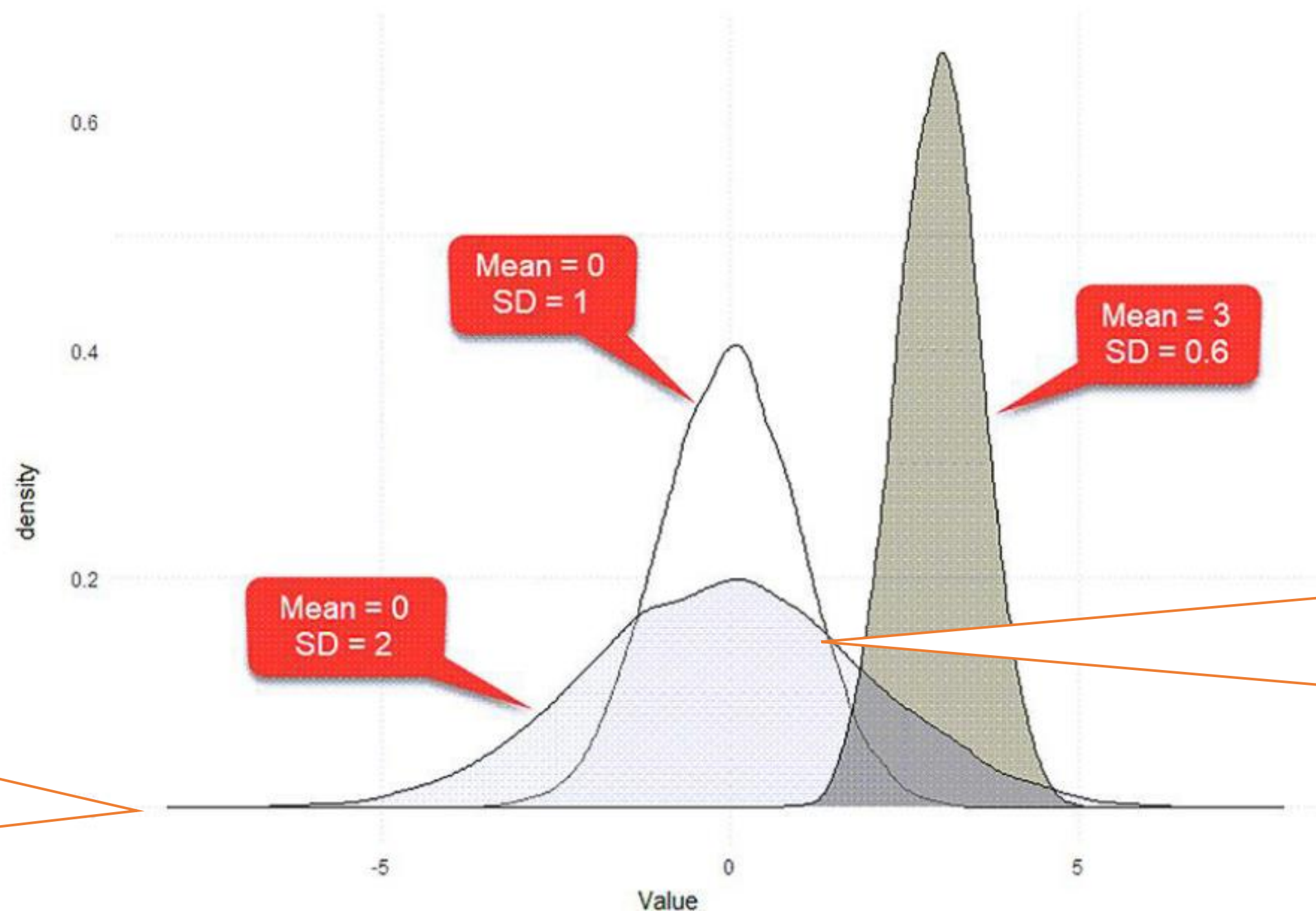
$$std(x)^2 = rms(x)^2 - avg(x)^2$$

Univariate distribution

- A univariate Gaussian

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

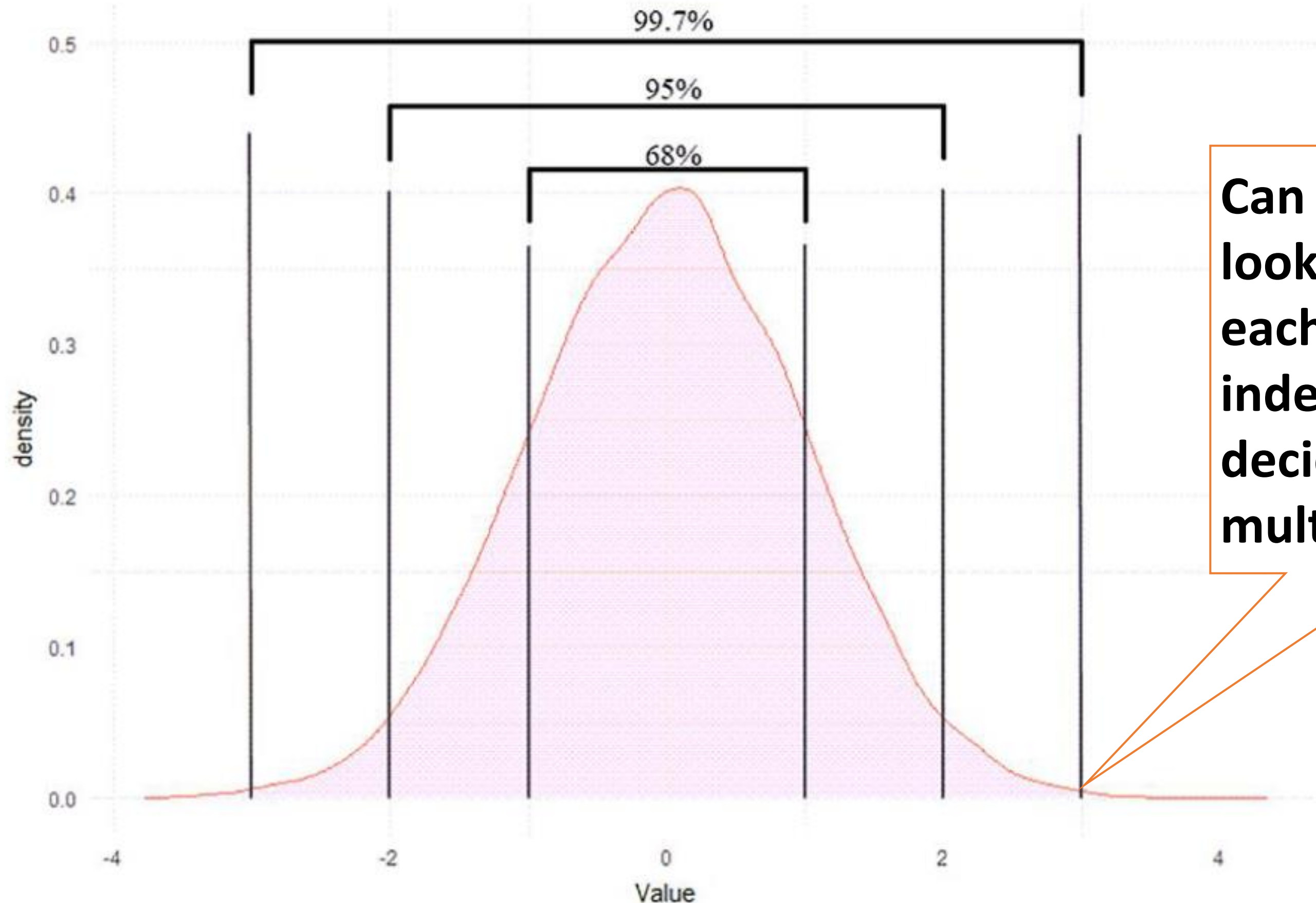
$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}$$



Random
variable X

Why does
height decrease
as the
distribution
becomes wide?

Empirical Formula for Gaussian Distribution

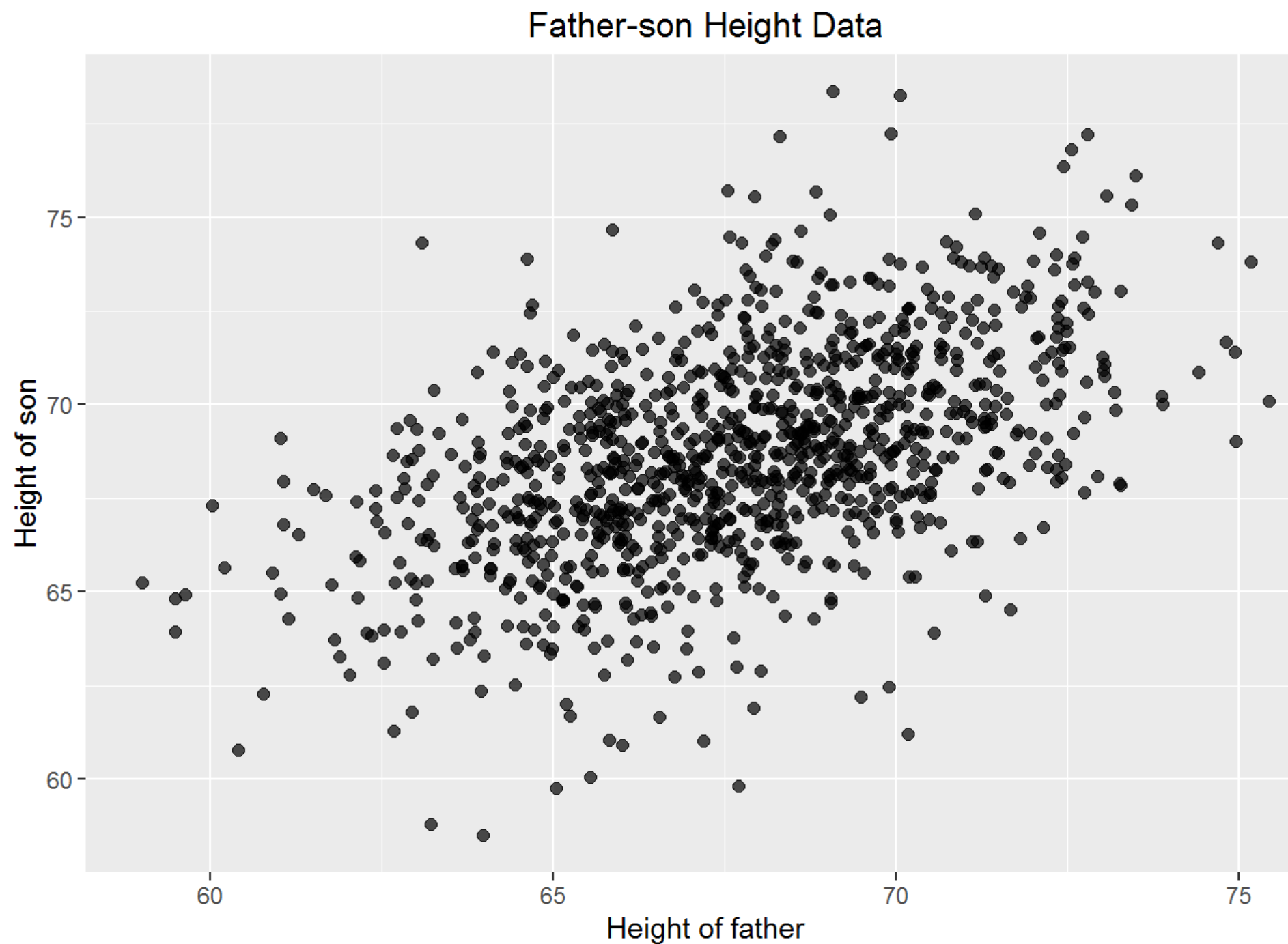


**Can we REALLY
look at 3 SD for
each feature
independently to
decide and delete
multivariate data?**

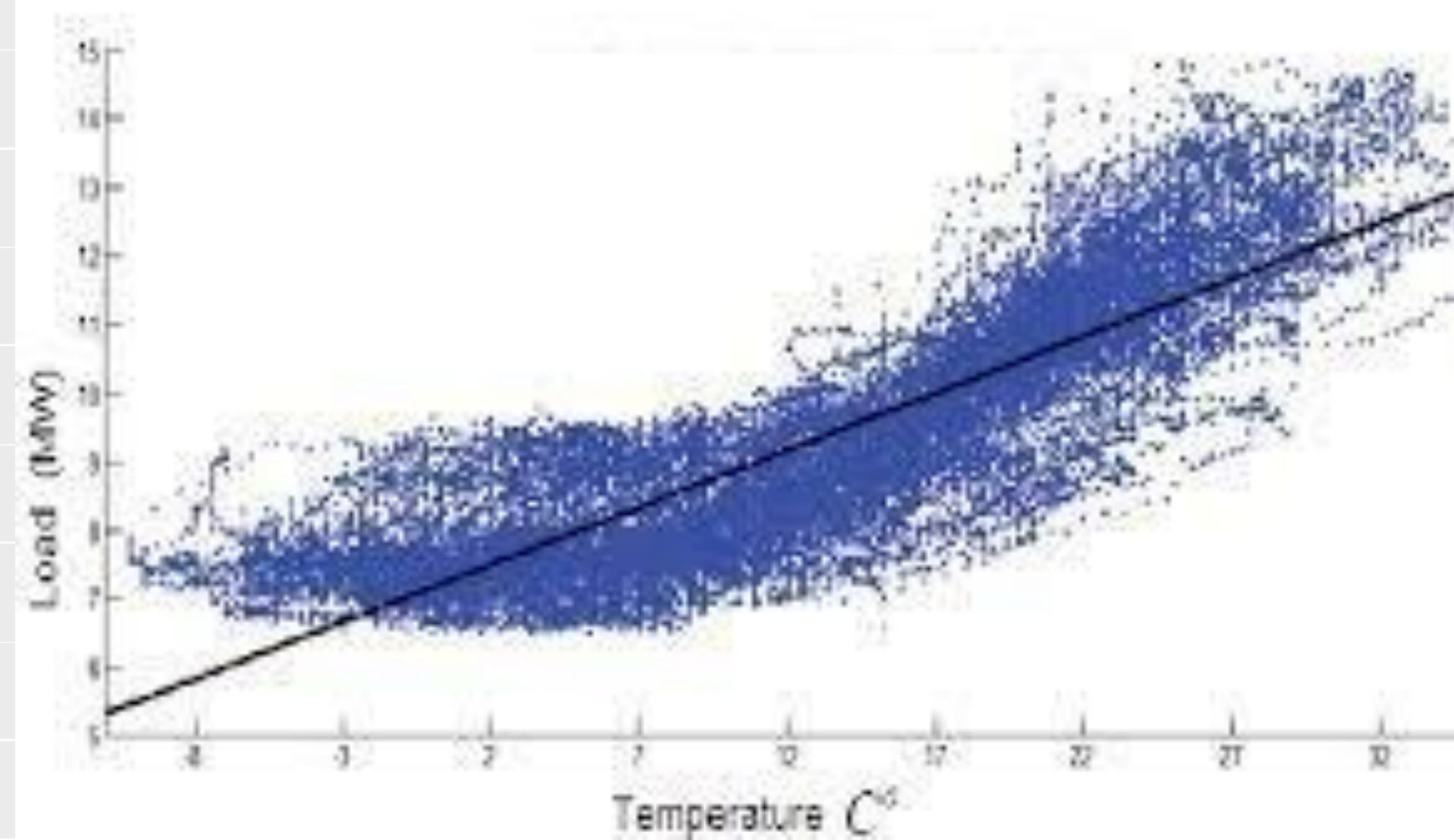


Correlation

- Father-son heights

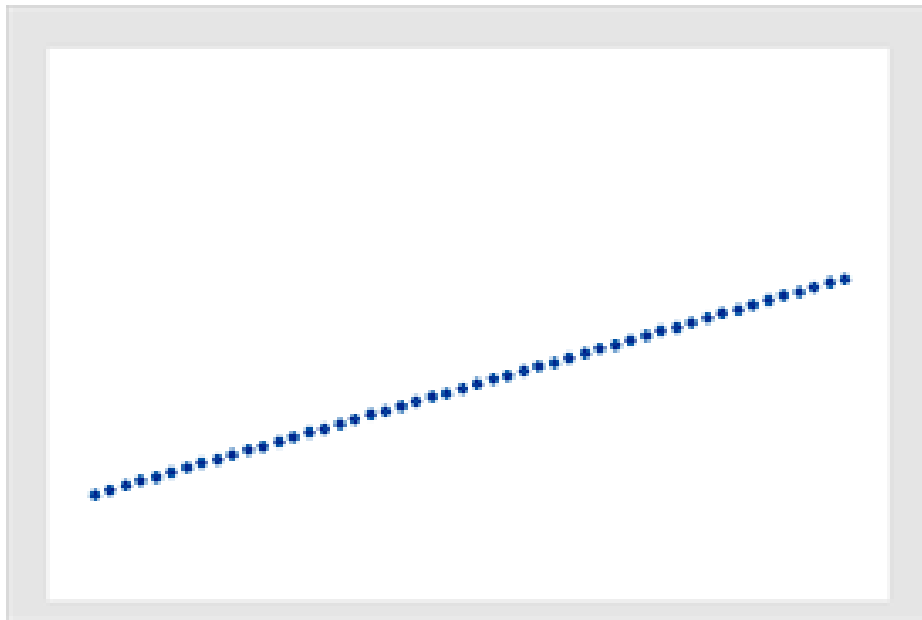


- Temperature-Electric bill

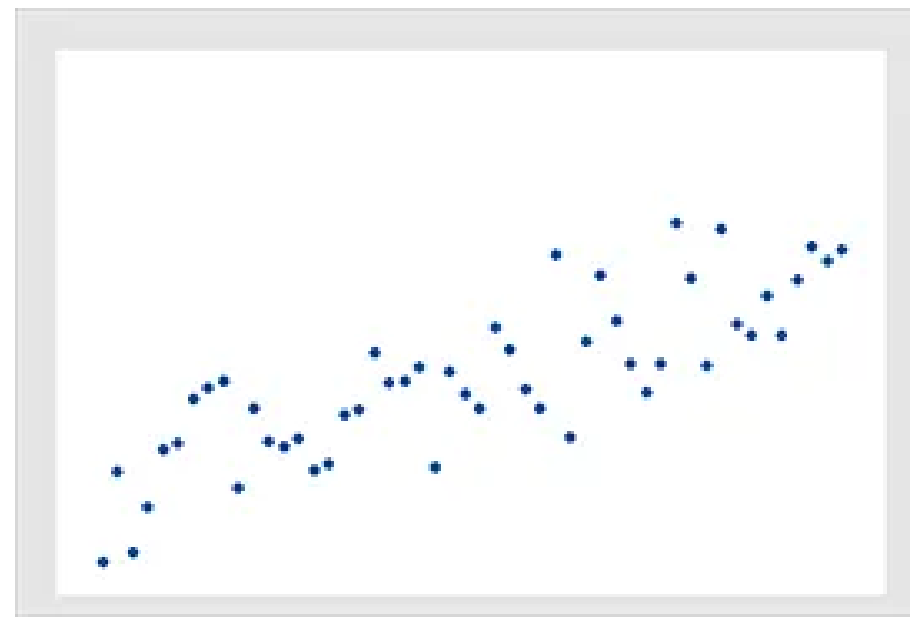


Correlation strength & coefficients

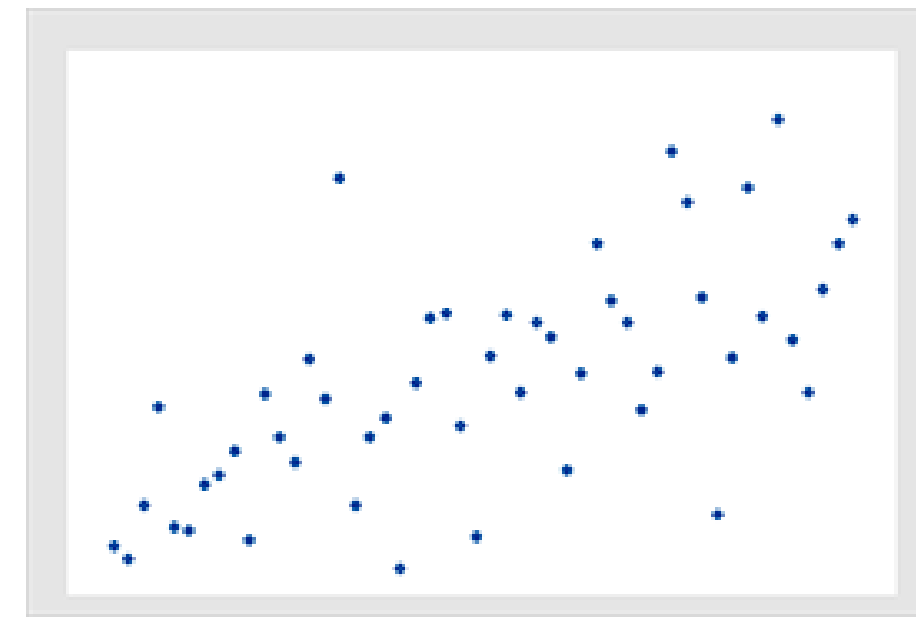
• Very Strong



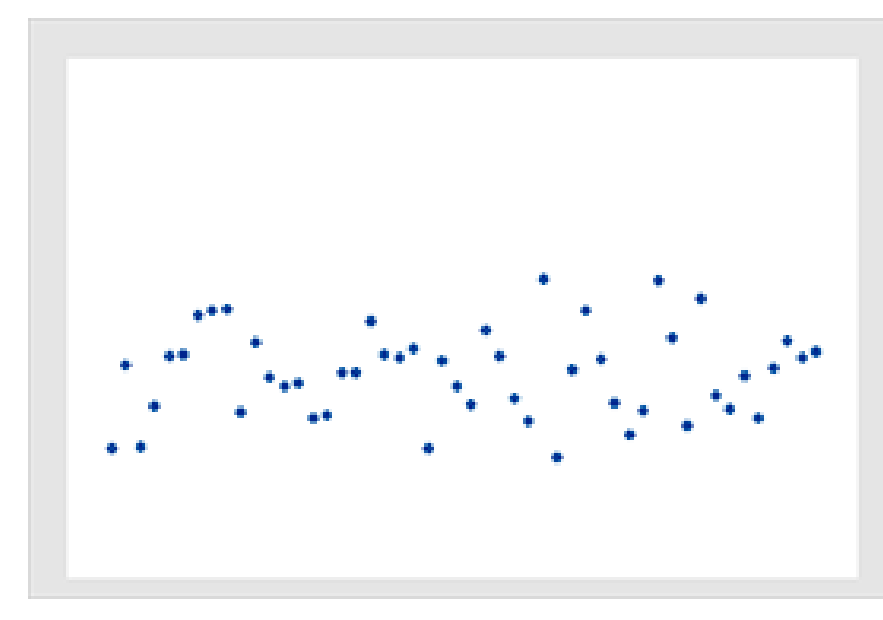
Strong



Moderate



None



• 1

0.8

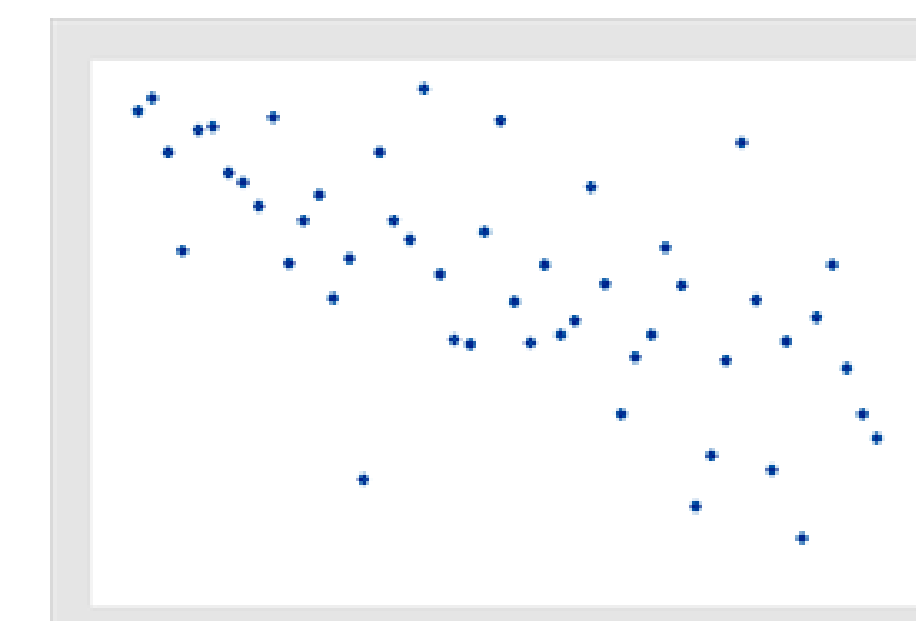
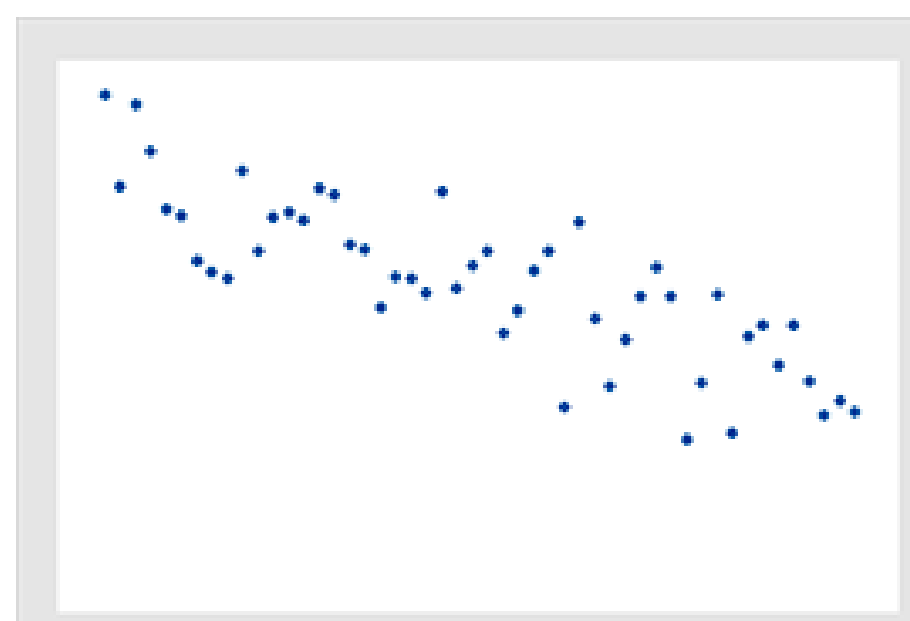
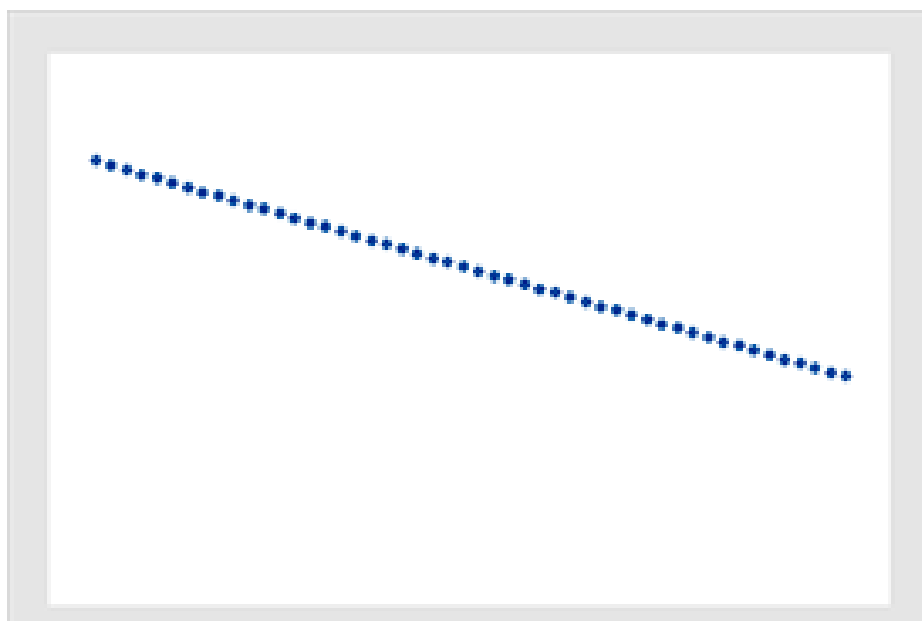
0.6

0

• -1

-0.8

-0.6




Correlation coefficient

- Covariance and Correlation are bivariate

$$\begin{aligned} Cov(x, y) &= \frac{\sum_{i=1}^n (x - \bar{x})(y - \bar{y})}{n} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned}$$

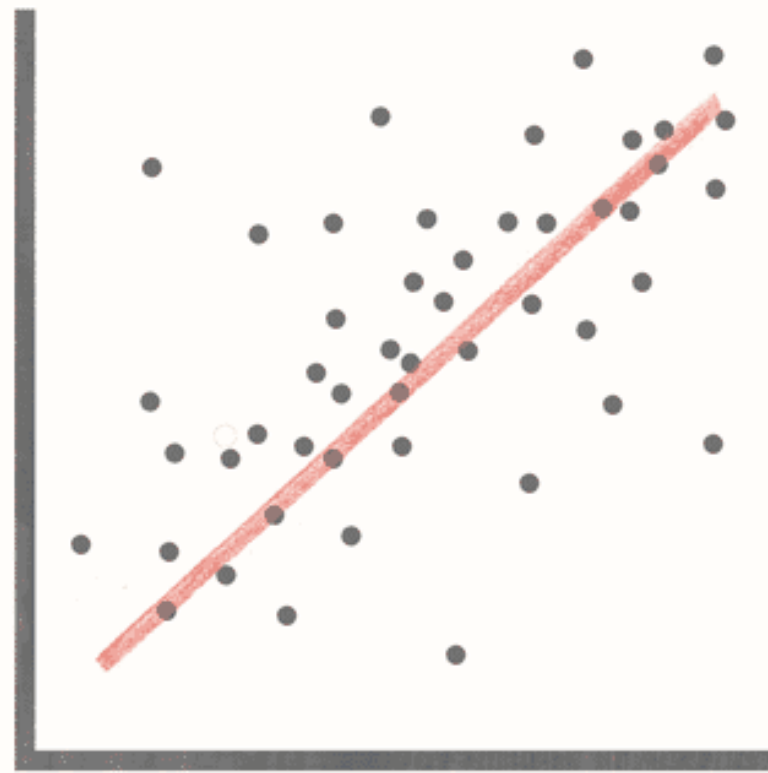
$Cov(x, x) = Var(x)$



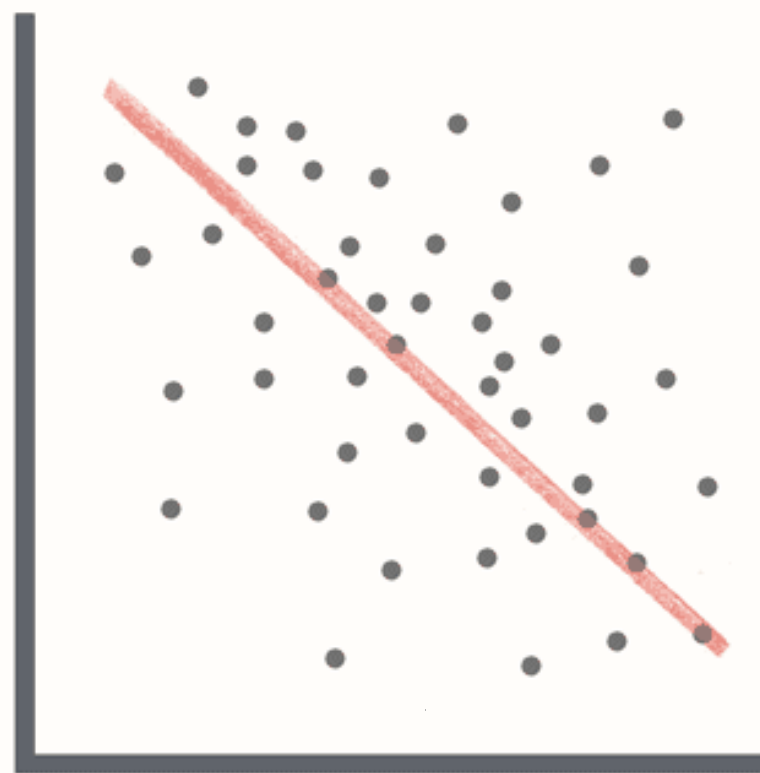
$$\rho = Correl(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$

$$\rho = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right)$$

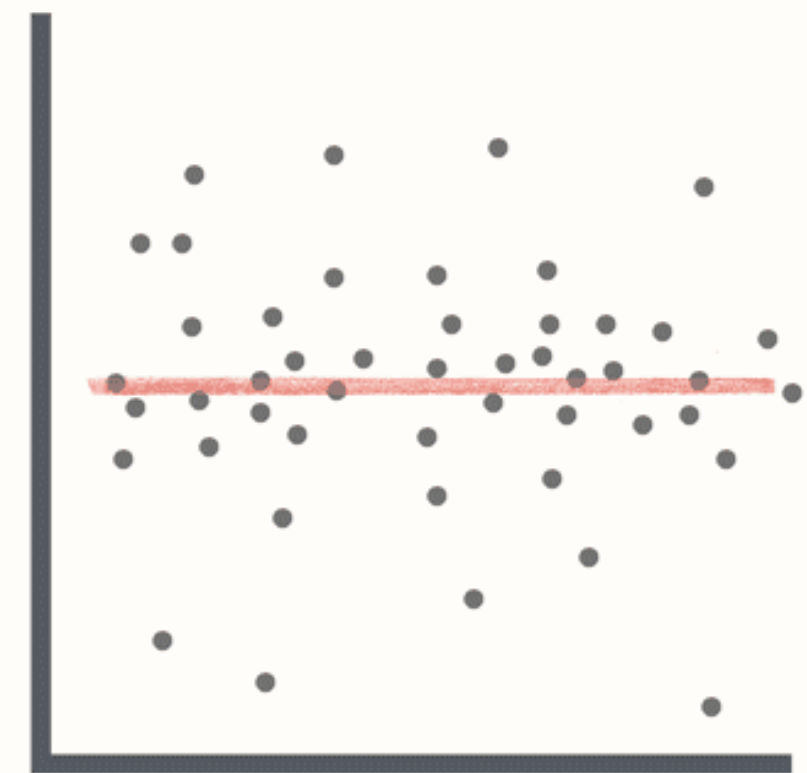
Intuition behind mean centering



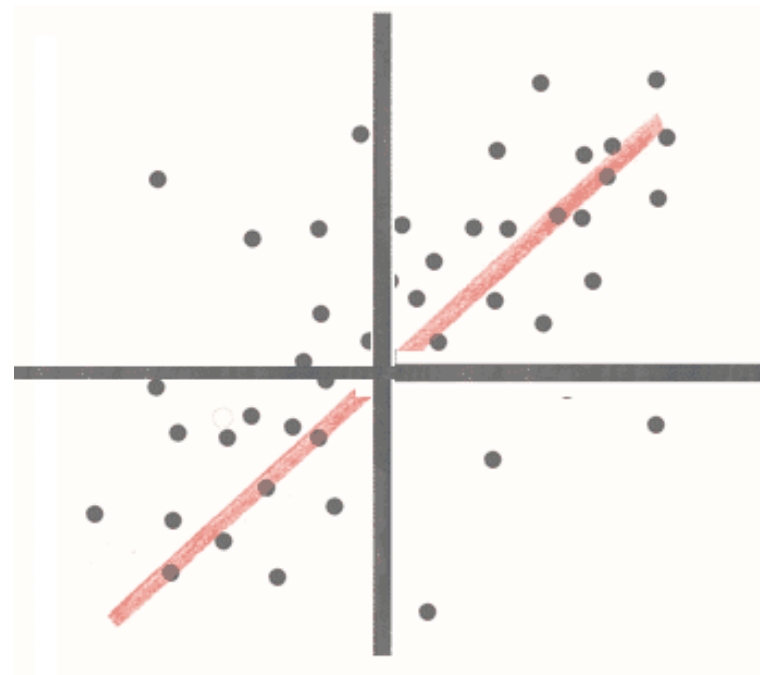
Positive Correlation



Negative Correlation



No Correlation



Positive Correlation

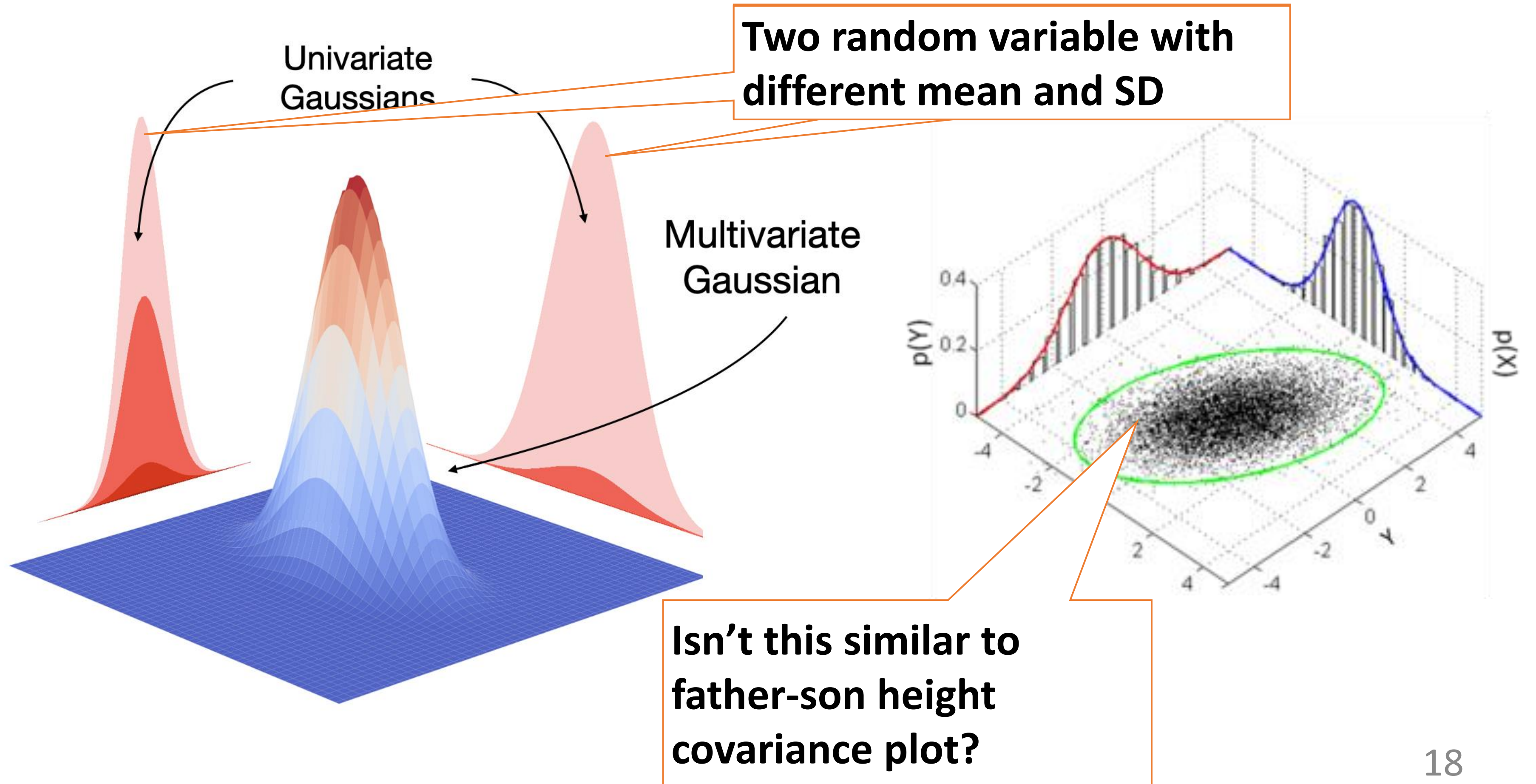


Correlation is not causation



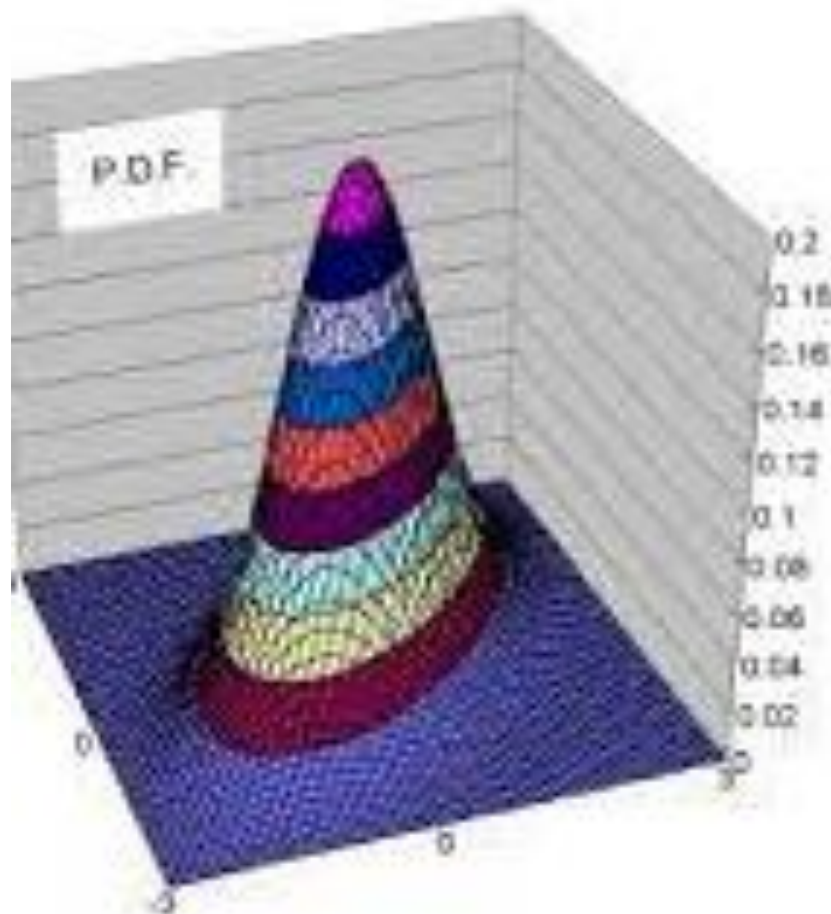


Multivariate Gaussian



Interpreting contour plots

- Multivariate Gaussian
 - <https://www.geogebra.org/m/pO4JcWPz>
- Contour plots (Isocontours)
 - Slicing through the function surface for a fixed z



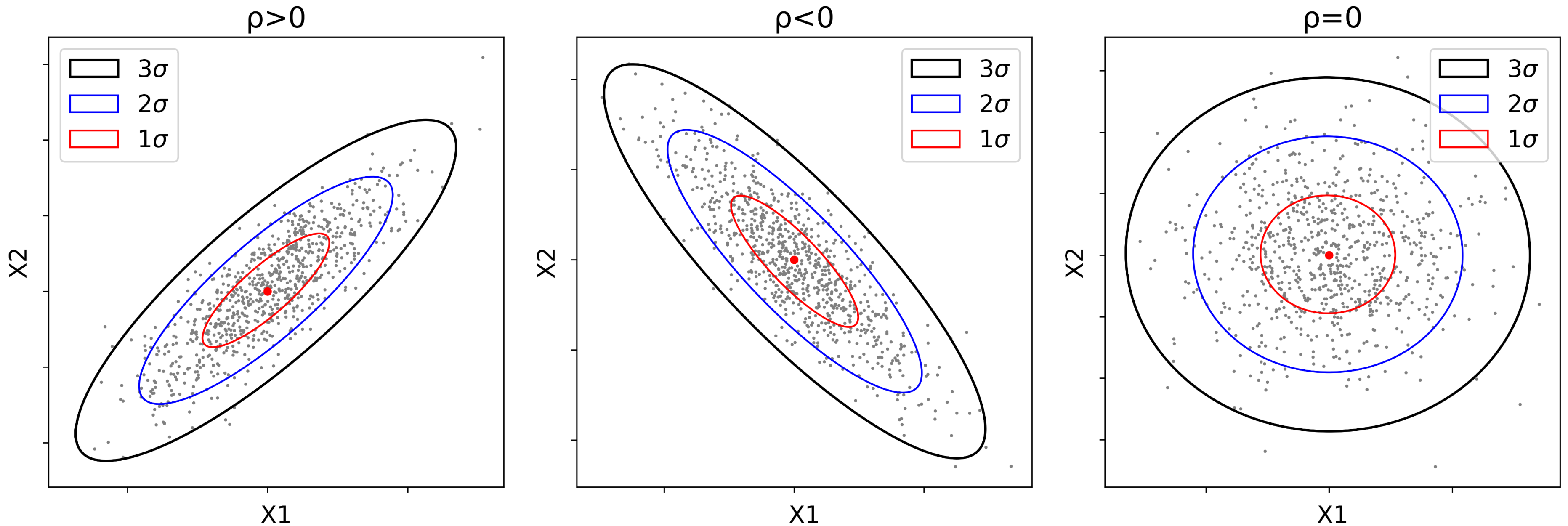
Multivariate Gaussian formula intuition

- We saw bivariate distribution as having two random variables with different mean and SD

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

Distribution of random vector X by also taking into account the interaction between RV

- What do we mean by interaction?
 - Recall father son heights
 - Student absent days versus grade
 - Google stock prices versus ice cream num sold



- Multivariate Gaussian formula should take into account the correlation/covariance based interaction

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

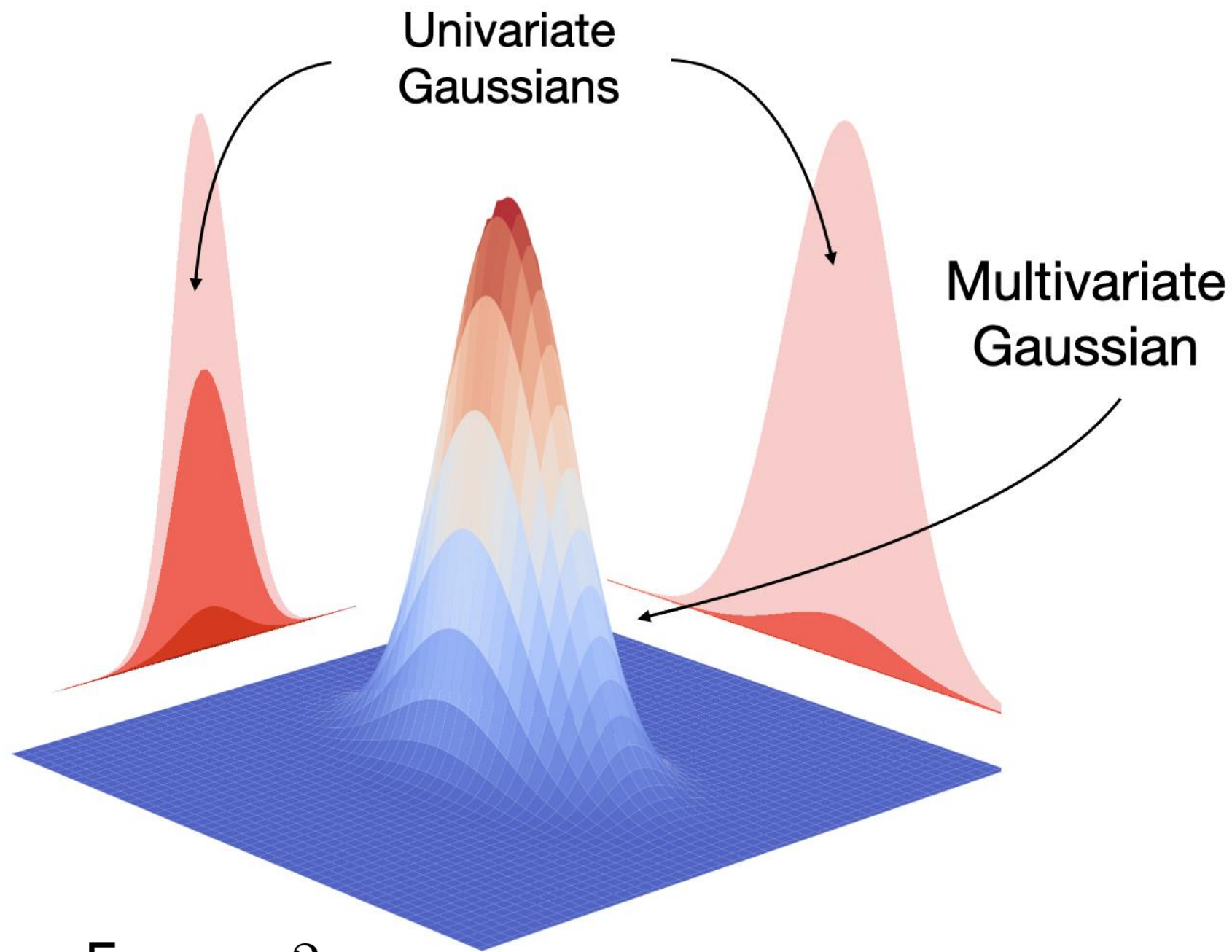
Multivariate Gaussian formula should have mean & SD for both random variables in vector

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \sigma = \begin{bmatrix} \sigma_1 \\ \sigma_2 \end{bmatrix}$$

$$Cov(x, y) = \frac{\sum_{i=1}^n (x - \bar{x})(y - \bar{y})}{n}$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & Cov_{12} & \dots & Cov_{1n} \\ Cov_{21} & \sigma_2^2 & \dots & Cov_{2n} \\ \dots & \dots & \dots & \dots \\ Cov_{(n-1)1} & \dots & \sigma_{n-1}^2 & Cov_{(n-1)n} \\ Cov_{n1} & Cov_{n2} & \dots & \sigma_n^2 \end{bmatrix}$$

Why have a matrix when scalar cov are duplicated?



Univariate

**Normalization
constant**

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}$$

Bell shape bcoz of this

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

$$\sigma = \begin{bmatrix} \sigma_1 \\ \sigma_2 \end{bmatrix}$$

$$\Sigma$$

**Cov matrix
already
holds all SD**

Multivariate

$$\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$$

Multivariate

$$\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$$

$$\frac{1}{\sqrt{\sigma^2 2\pi}} e^{\frac{-1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}$$

**X is a vector. Mu is a vector.
Can you square a vector?**

Quadratic form giving scalar

$$x^T x = \|x\|^2$$

$$e^{\frac{-1}{2} \left(\frac{\mathbf{X}-\mu}{\sigma} \right)^2}$$

$$e^{\frac{-1}{2} (\mathbf{X}-\mu)^T \dots (\mathbf{X}-\mu)}$$

**How to account for spread of
two random variables & their
interaction (joint spread)?**

Account for 3 scalars capturing the spread

Spread goes to denominator

$$e^{\frac{-1}{2} (\mathbf{X}-\mu)^T \Sigma (\mathbf{X}-\mu)}$$

$$e^{\frac{-1}{2} (\mathbf{X}-\mu)^T \Sigma^{-1} (\mathbf{X}-\mu)}$$

**Normalization
constant**

$$\frac{1}{\sqrt{\det(\Sigma)(2\pi)^D}} e^{\frac{-1}{2} (\mathbf{X}-\mu)^T \Sigma^{-1} (\mathbf{X}-\mu)}$$

Uni v/s multivariate similarities

Univariate

$$\frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}$$

$$\sigma > 0$$

Multivariate

$$\frac{1}{\sqrt{\det(\Sigma) (2\pi)^D}} e^{-\frac{1}{2} (\mathbf{X}-\mu)^T \Sigma^{-1} (\mathbf{X}-\mu)}$$

$$\Sigma > 0$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

- Covariance matrix is symmetric positive definite
- Symmetric is easy to see
- Positive definite means Eigen values > 0

Univariate vs multivariate similarities(contd.)

Univariate

$$\frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}$$

$$z = \frac{x - \mu}{\sigma}$$

Multivariate

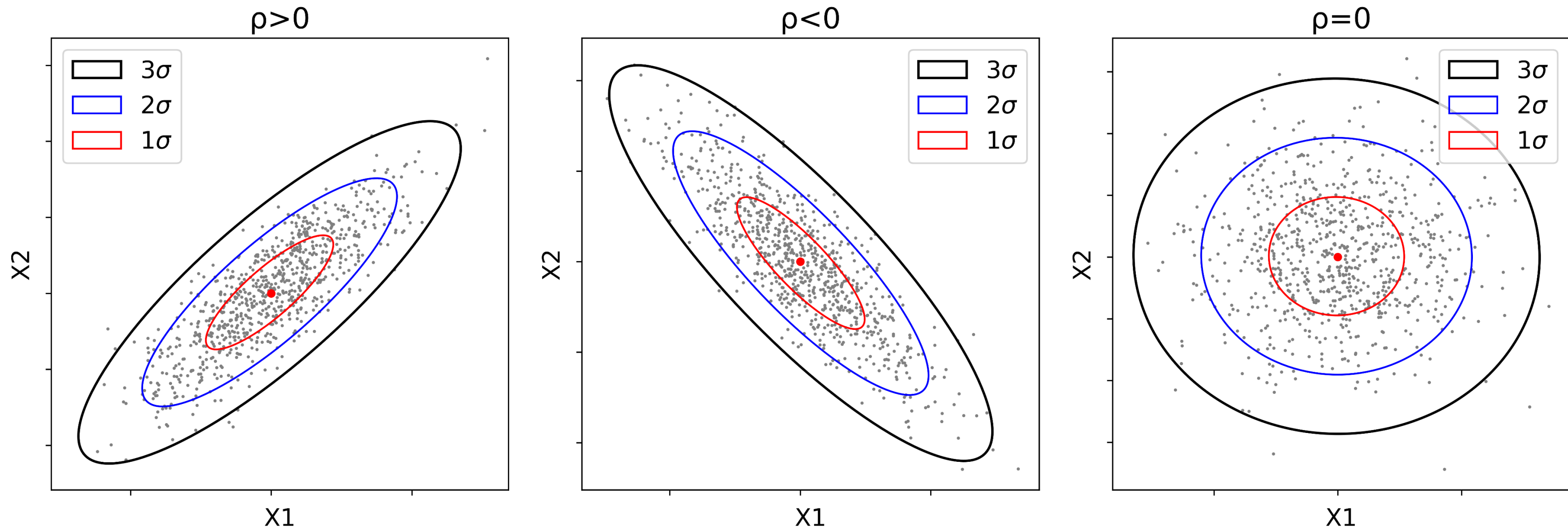
$$\frac{1}{\sqrt{\det(\Sigma)(2\pi)^D}} e^{-\frac{1}{2} (\mathbf{X}-\mu)^T \Sigma^{-1} (\mathbf{X}-\mu)}$$

$$d_M = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}$$

- Z score and Mahalanobis distance are equivalent

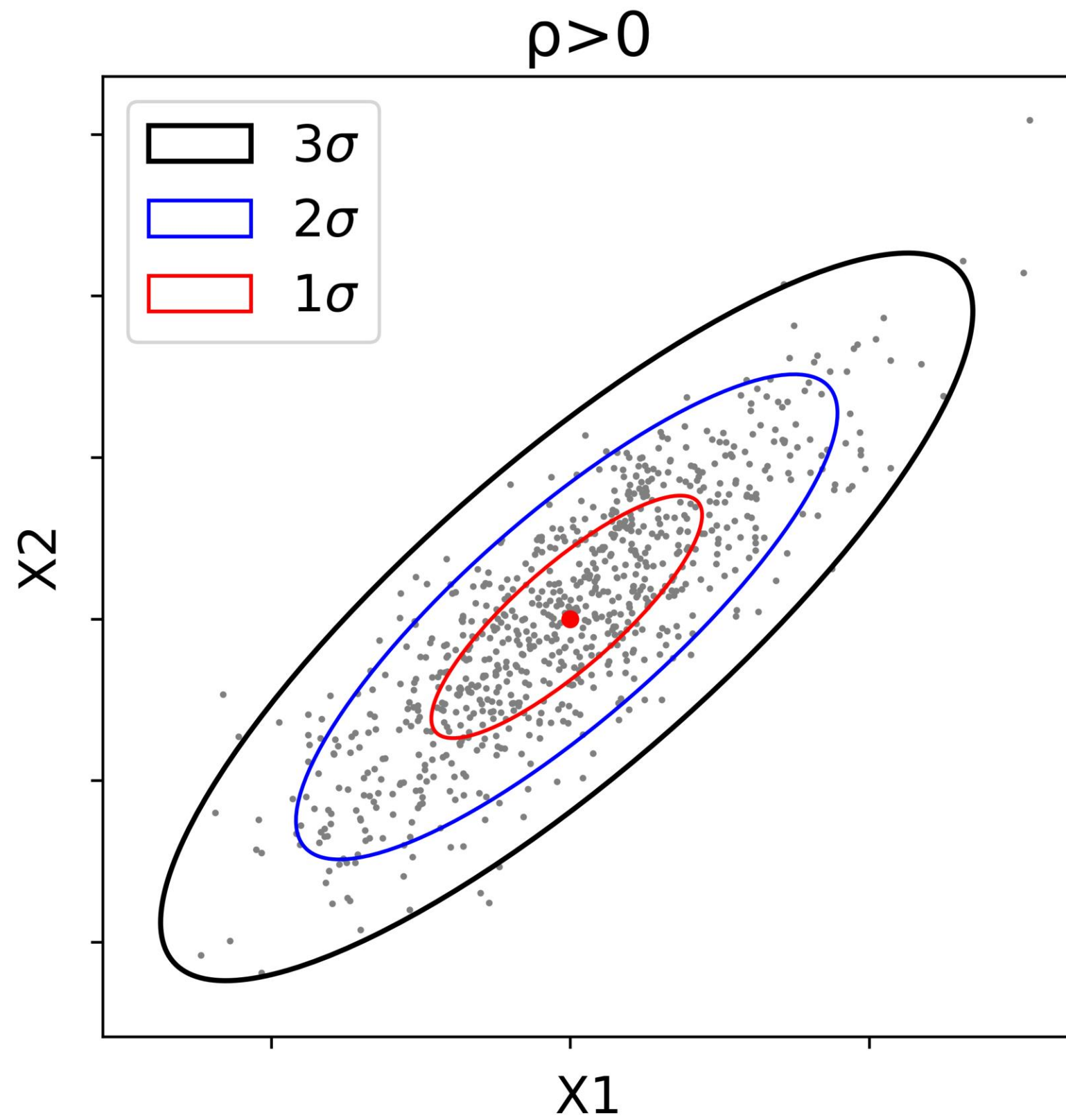
Geometric meaning of contour plots

- Look at Standard deviations in addition to correlation & guess the covariance matrix



- Draw a few more contour plots to familiarize

Geometric meaning of StandardScaler

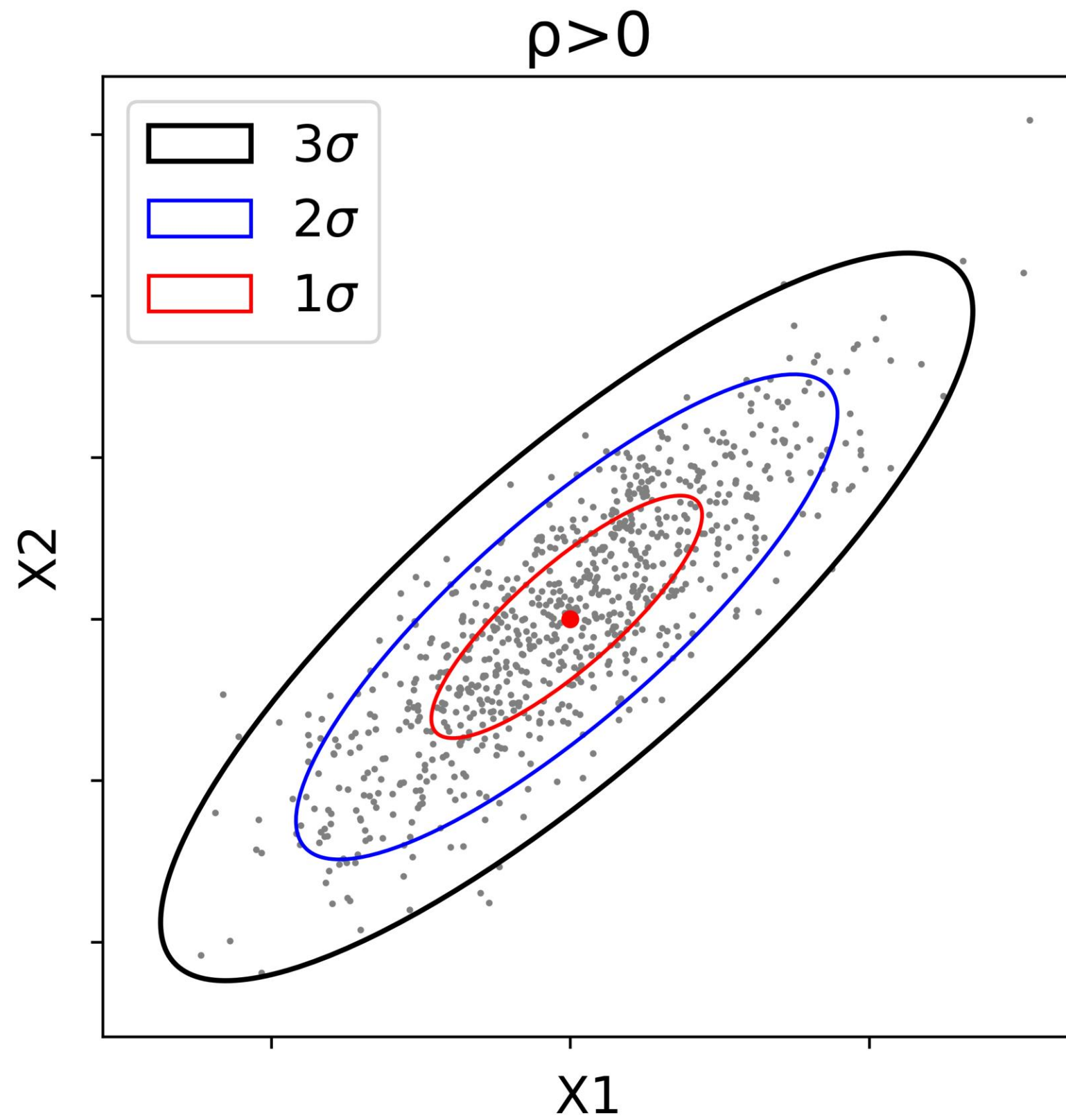


$$\phi(x) = z = \frac{x - \mu}{\sigma}$$

- Demo at <https://projector.tensorflow.org/>



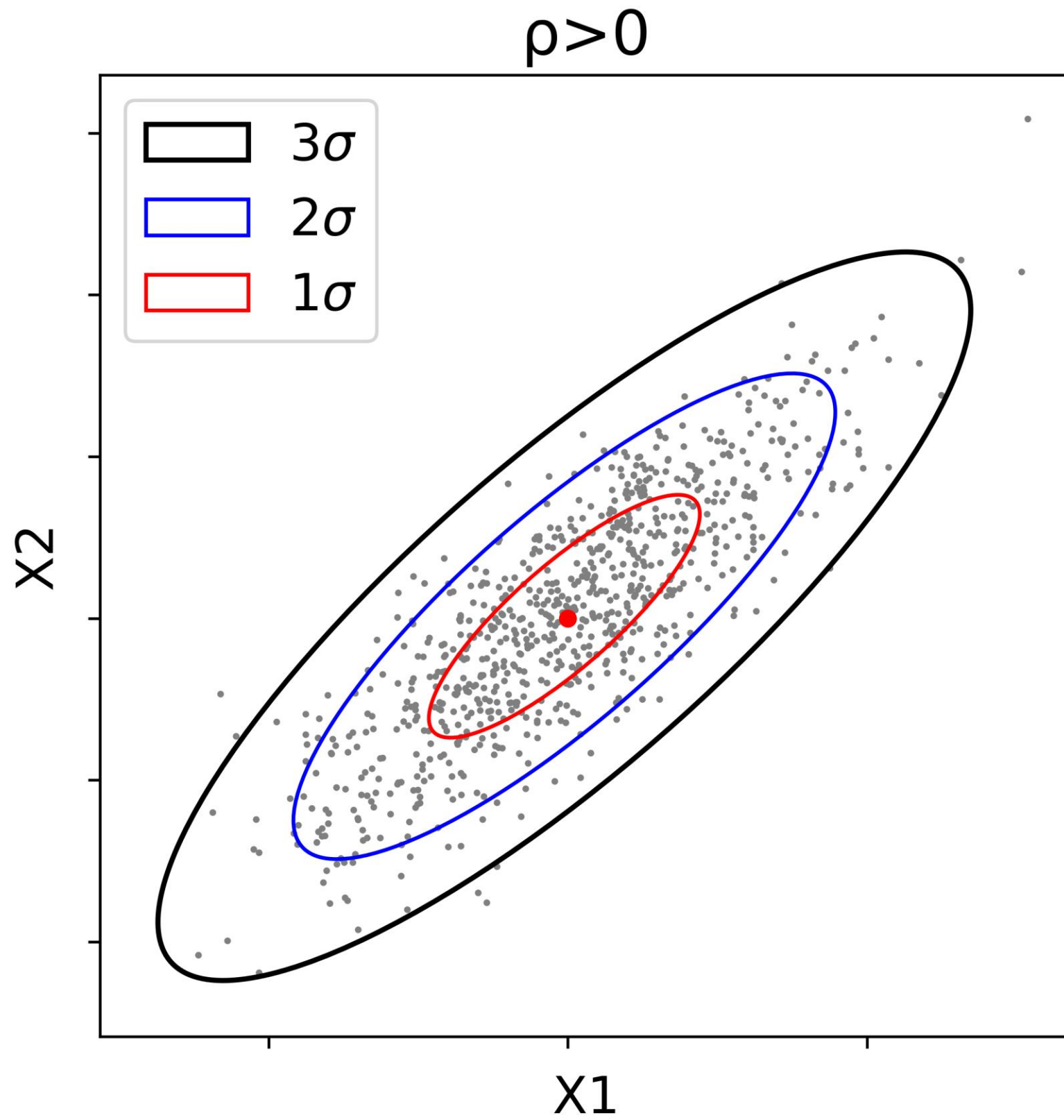
A relook at Mahalanobis distance



- Mark some points and logically see if they are outliers?

$$\sqrt{(\mathbf{X} - \mu)^T \Sigma^{-1} (\mathbf{X} - \mu)}$$

Problems with Mahalanobis distance



- Not robust enough (what does that mean?)
 - Distribution fitted over all points
 - Add an outlier & distribution “bends” towards it

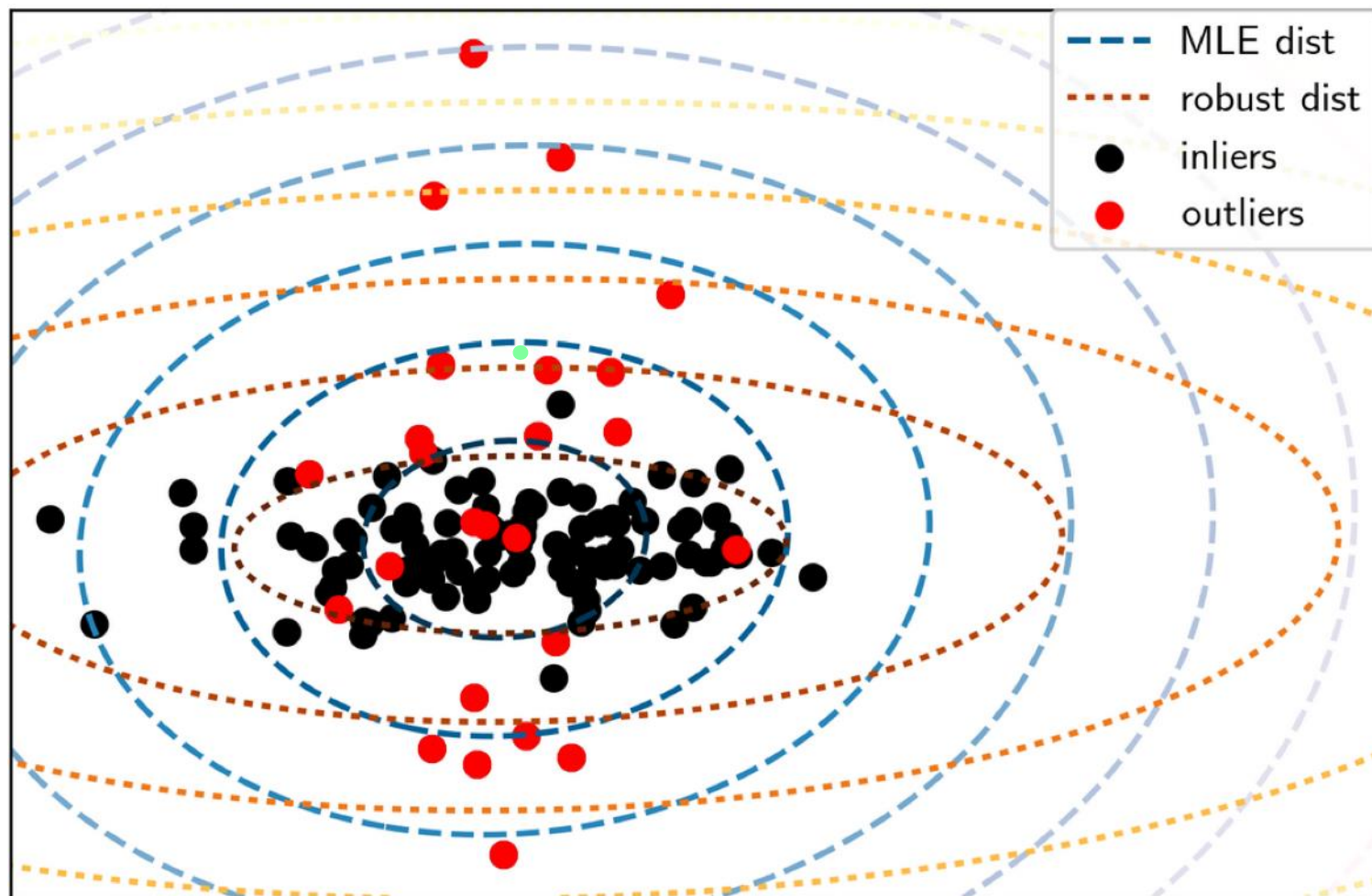
$$\sqrt{(\mathbf{X} - \mu_{MCD})^T \Sigma_{MCD}^{-1} (\mathbf{X} - \mu_{MCD})}$$

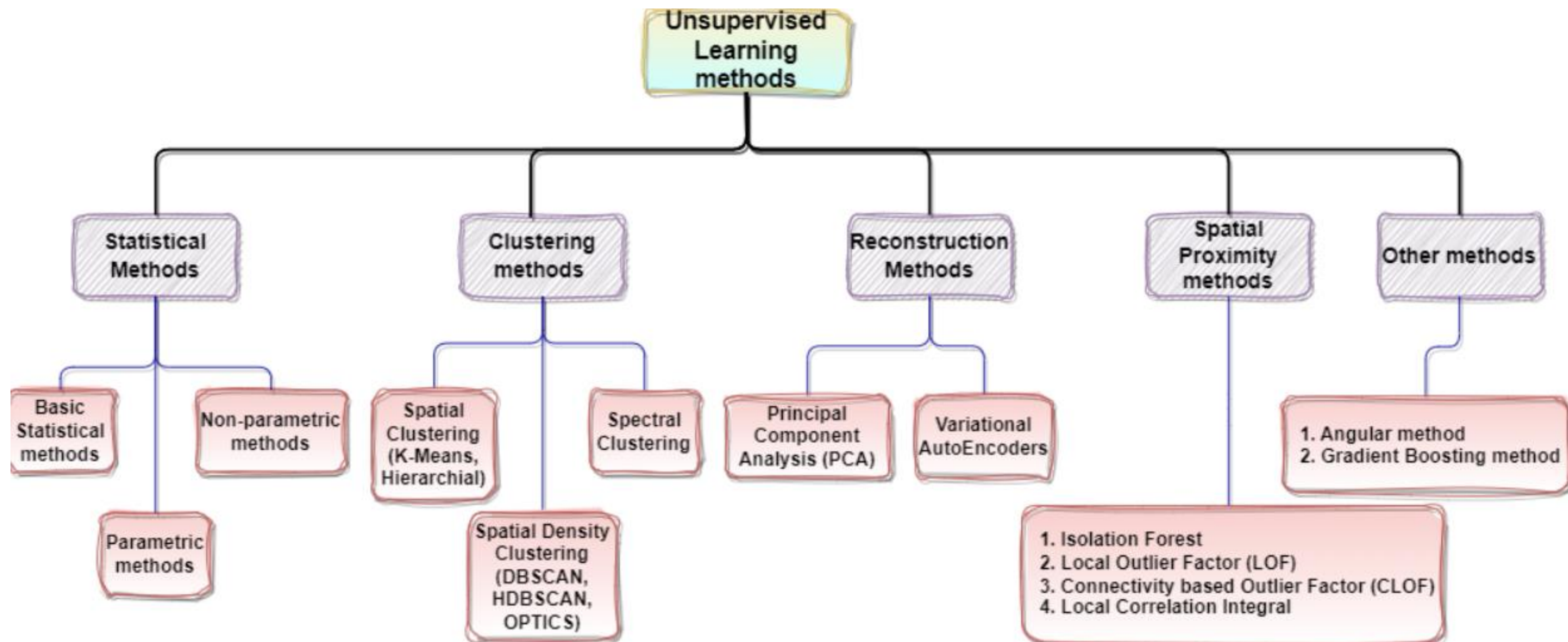
MCD = Minimum Covariance determinant

MCD procedure

- Take k typically $0.75 * n$ data points
- Sample different data points and find their cov matrix
- Find the cov matrix that has least determinant
- This represents the tightest cloud of points

Elliptic
Envelope
in sklearn



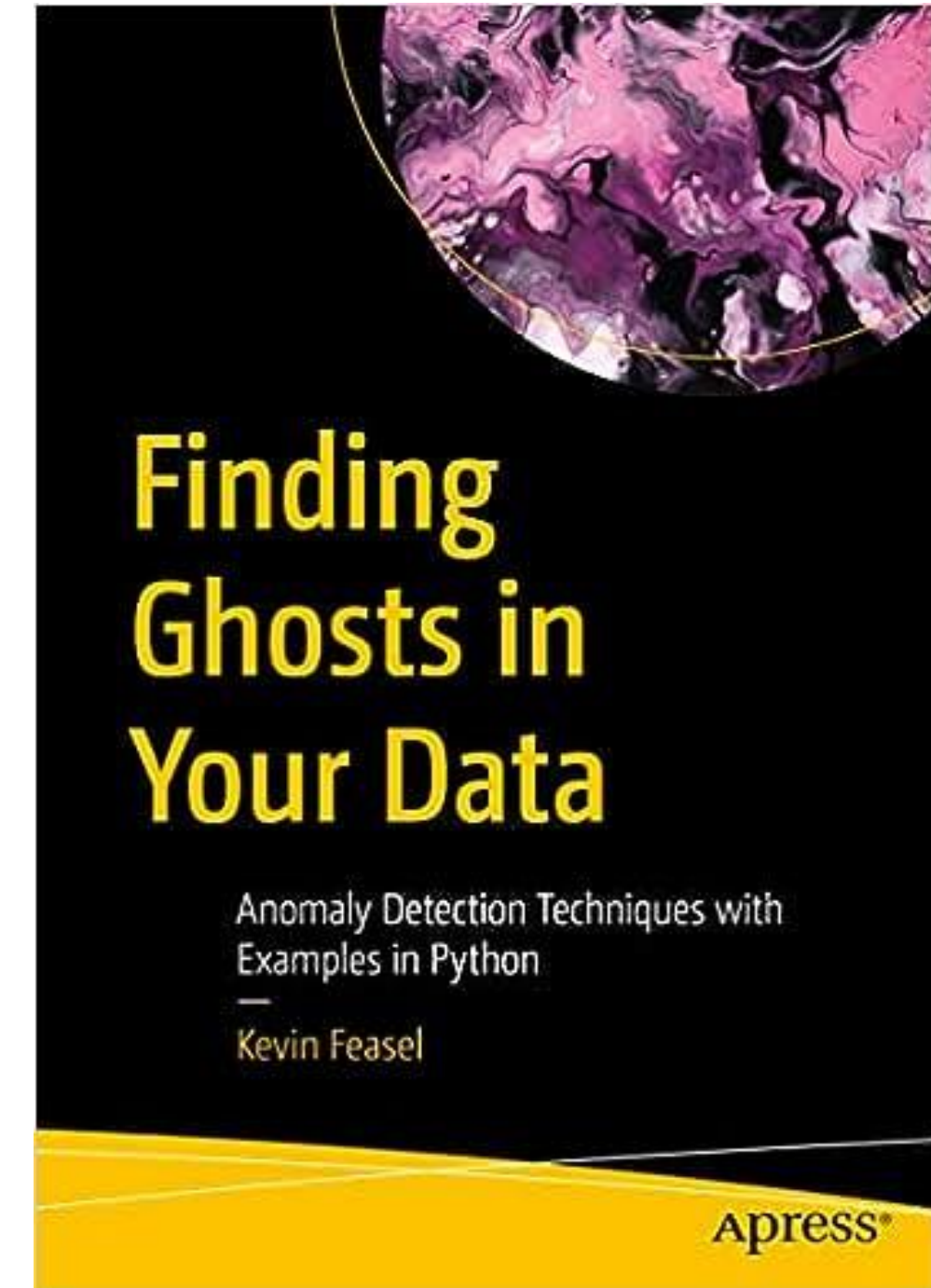
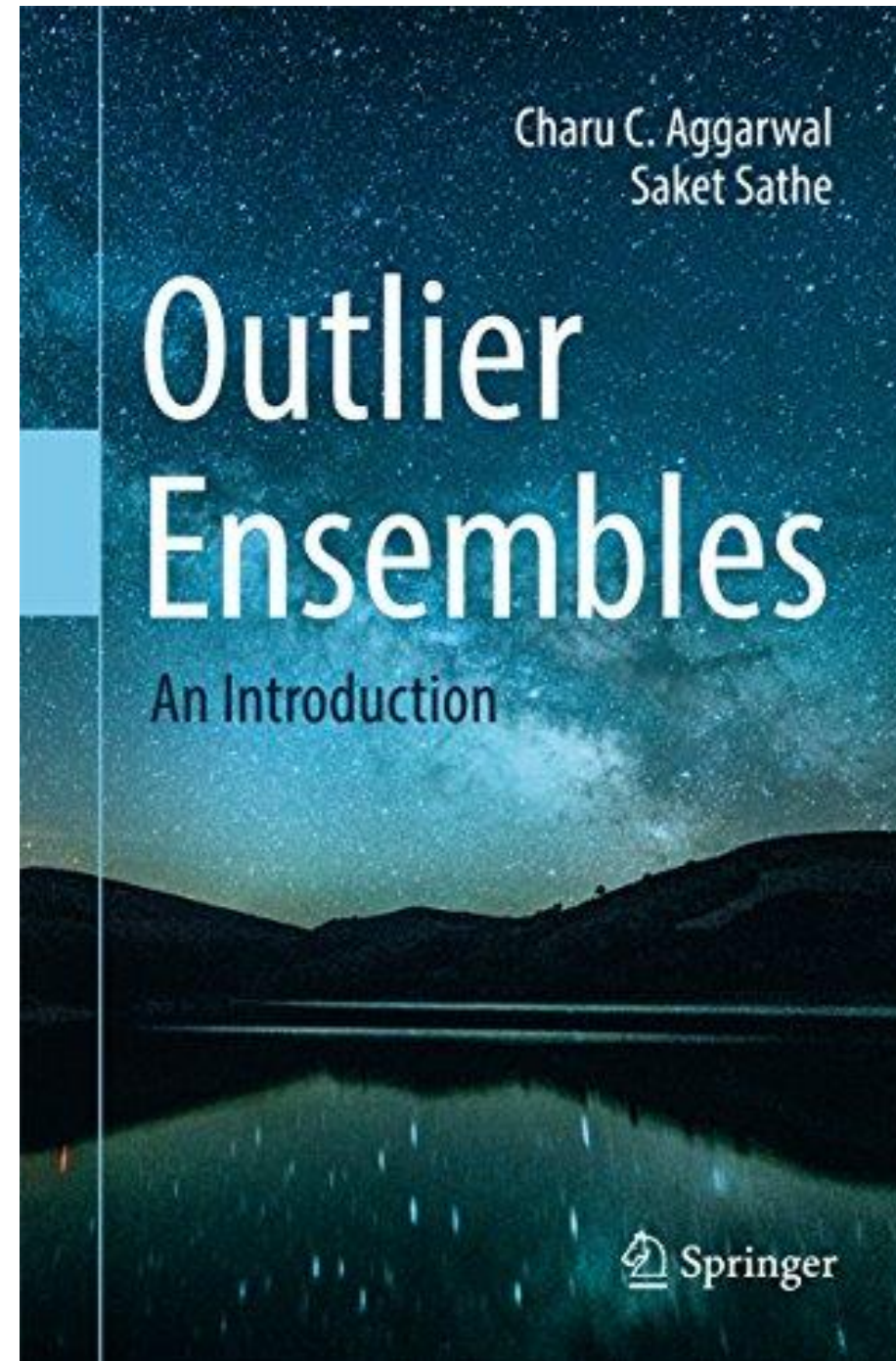
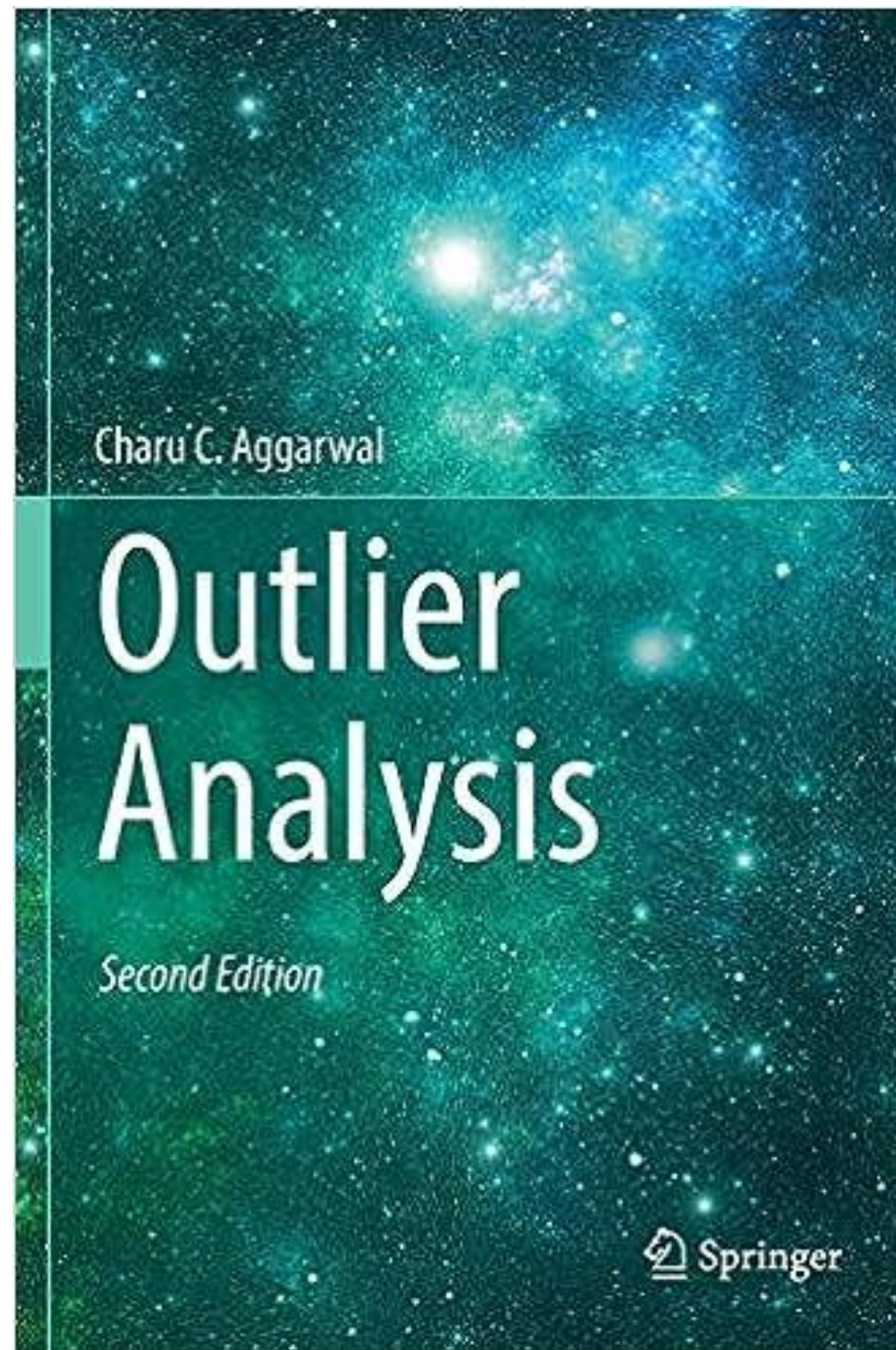


Other outlier detection algorithms

- Proximity based:
 - kNN
 - Isolation Forest, Local Outlier Factor(LOF)
- Clustering based
 - K-Means, Gaussian Mixture Model (GMM) Clustering
- Distance metric based
 - Cook's distance, Gower's distance (mixed data type)
 - MCD on GMM
- Reconstruction based: PCA, Autoencoder
- Take a look at PyOD library

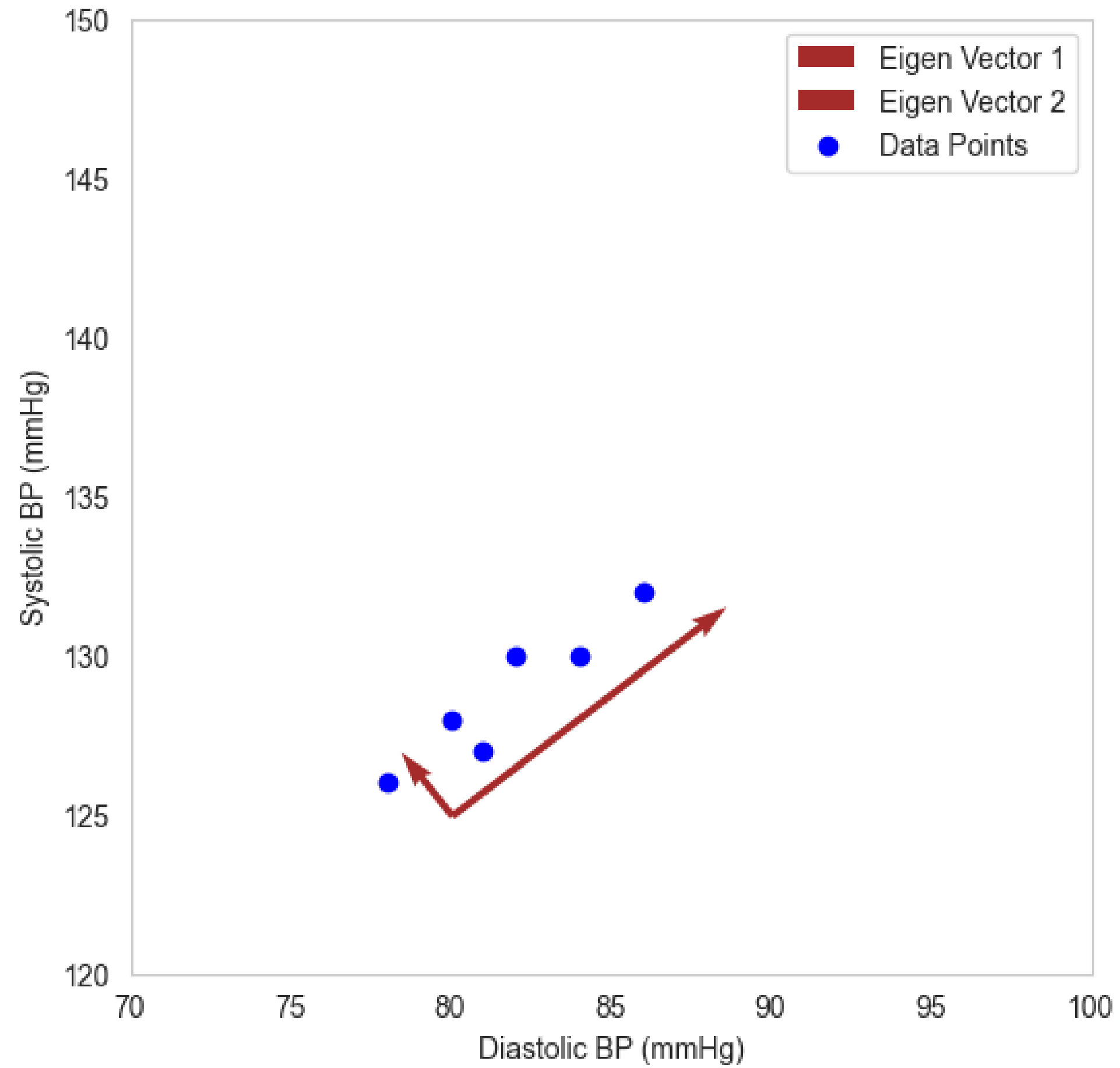
Outlier analysis: Recommended books

- Not part of syllabus. On your own interest



Matrix-Vector product & determinant

- Geogebra demo
 - Matrix-Vector product transforms the vector
 - Extent of transformation given by area of parallelogram of original & transformed vectors
 - aka determinant of matrix
- But we are not multiplying data with Cov matrix
- Enter Eigen values of Cov matrix



Eigen Values & Vectors

- Eigen values of any matrix represents stretch in direction of vector
- Eigen vector for cov matrix represents direction of max variance
- Product of Eigen values = Determinant of square matrix
- Combine these ideas
 - Determinant is a single measure of spread of data
- As an aside: This determinant-spread relation also answers the question why determinant of cov matrix is in denominator of multivariate gaussian PDF formula



QUESTIONS



Thank You!