ML Hodgepodge

CSE 312 Winter 25
Lecture 26

Multidimensional Gaussians

Preliminary: Random Vectors

In ML, our data points are often multidimensional.

For example:

To predict housing prices, each data point might have: number of rooms, number of bathrooms, square footage, zip code, year built, …

To make movie recommendations, each data point might have: ratings of existing movies, whether you started a movie and stopped after 10 minutes,…

A single data point is a full vector

Preliminary: Random Vectors

A random vector X is a vector where each entry is a random variable.

$\mathbb{E}[X]$ is a vector, where each entry is the expectation of that entry.

For example, if X is a uniform vector from the sample space

$$\left\{\begin{bmatrix}1\\2\\3\end{bmatrix}, \begin{bmatrix}-1\\2\\3\end{bmatrix}, \begin{bmatrix}0\\2\\6\end{bmatrix}\right\}$$

$$\frac{1}{3}\, 1 + \frac{1}{3} \cdot -1 + \frac{1}{3} \cdot 0 = \begin{bmatrix}0\\2\\4\end{bmatrix}$$

$\mathbb{E}[X] = [0,2,4]^T$

Covariance Matrix

$X_1, X_2, \dots, X_n$

Remember Covariance?

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

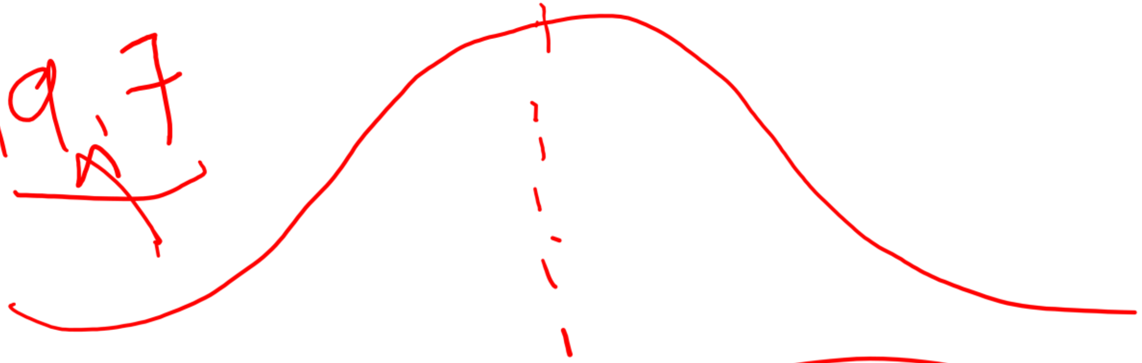
We'll want to talk about covariance between entries:

Define the "covariance matrix"

$$\Sigma = \begin{bmatrix} \mathrm{Cov}(X_1,X_1) & \cdots & \mathrm{Cov}(X_1,X_n) \\ \vdots & \mathrm{Cov}(X_i,X_j) & \vdots \\ \mathrm{Cov}(X_n,X_1) & \cdots & \mathrm{Cov}(X_n,X_n) \end{bmatrix}$$

Covariance

68-95-99.7

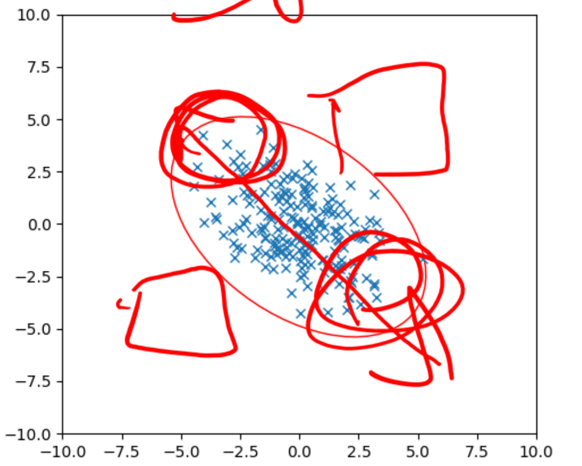


Let's think about 2 dimensions.

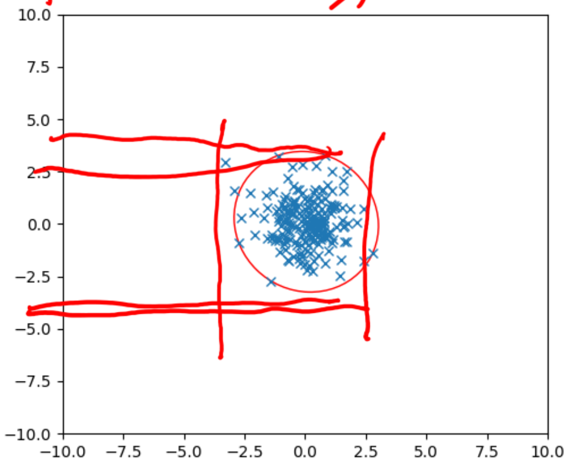
Let $X = [X_1, X_2]^T$ where $X_i \sim \mathcal{N}(0,1)$ and X_1 and X_2 are independent.

What is Σ? Which of these pictures are 200 i.i.d. samples of X?

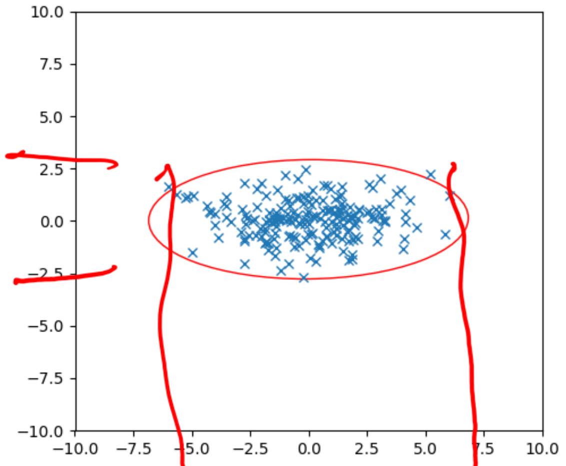
neg. covariance

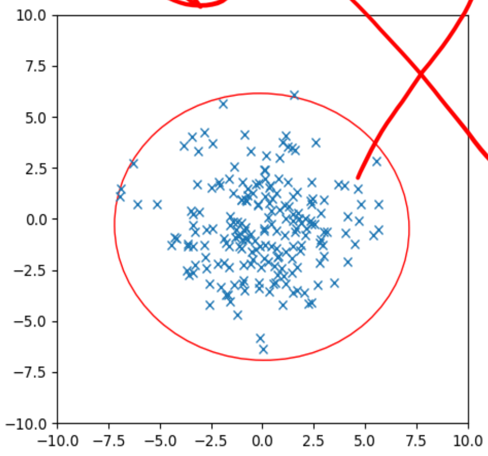


B



C





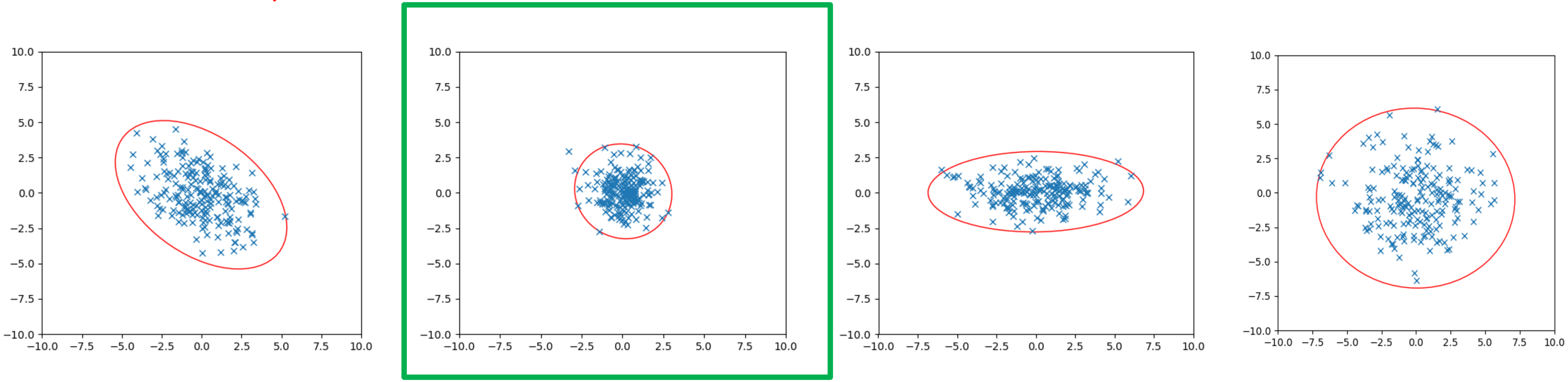
Covariance

Let's think about 2 dimensions.

Let $X = [X_1, X_2]^T$ where $X_i \sim \mathcal{N}(0,1)$ and X_1 and X_2 are independent.

What is Σ? Which of these pictures are 200 i.i.d. samples of X?

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

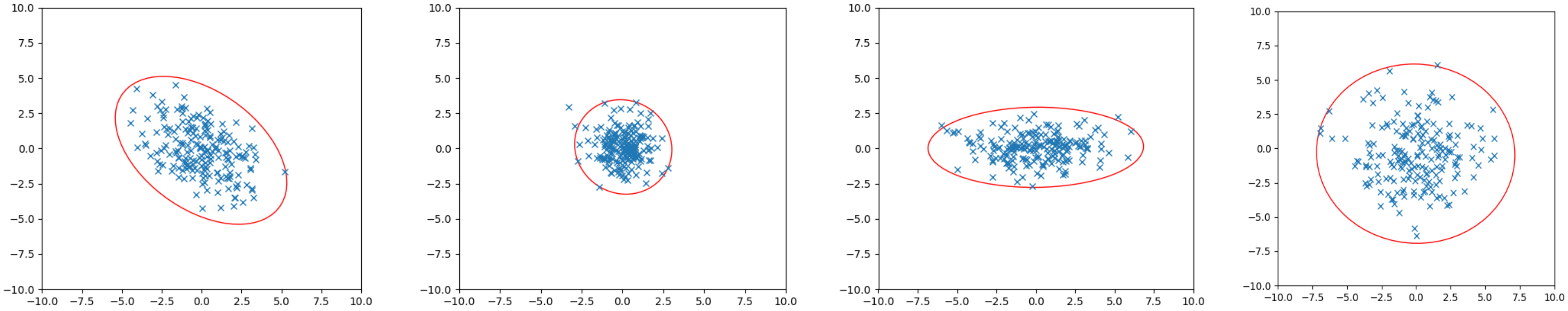


Unequal Variances, Still Independent

Let's think about 2 dimensions.

Let $X = [X_1, X_2]^T$ where $X_1 \sim \mathcal{N}(0,5)$, $X_2 \sim \mathcal{N}(0,1)$ and X_1 and X_2 are independent.

What is Σ? Which of these pictures are i.i.d. samples of X?



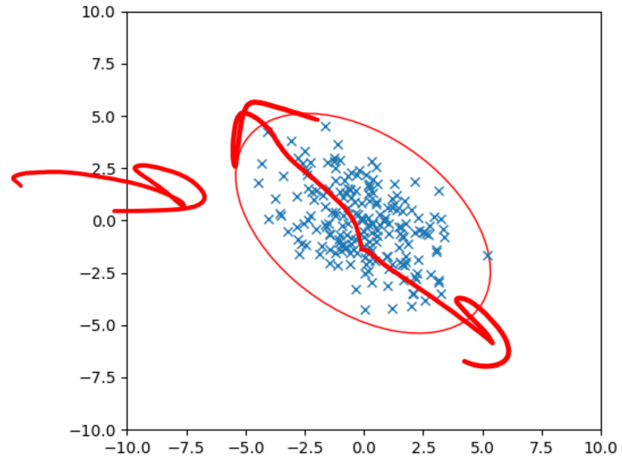
Unequal Variances, Still Independent

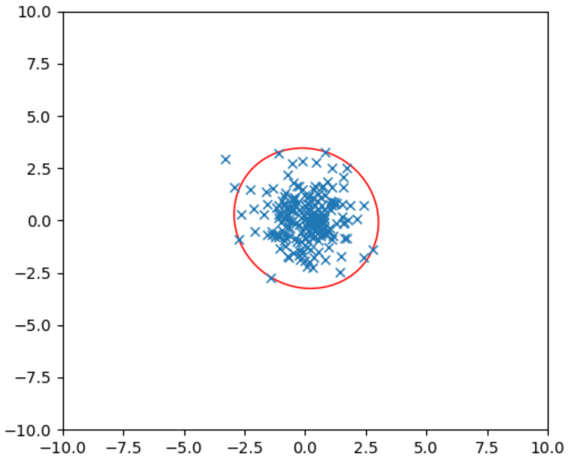
Let's think about 2 dimensions.

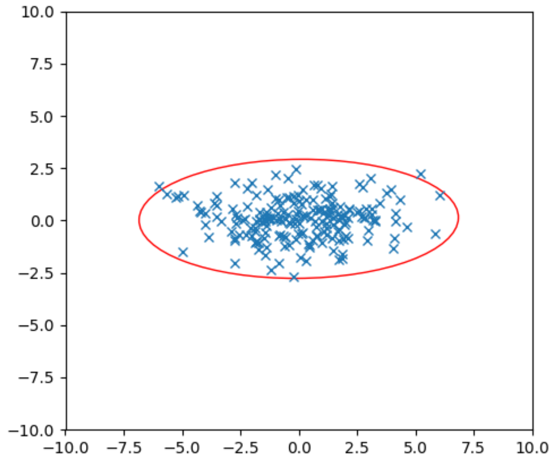
Let $X = [X_1, X_2]^T$ where $X_1 \sim \mathcal{N}(0,5)$, $X_2 \sim \mathcal{N}(0,1)$ and X_1 and X_2 are independent.

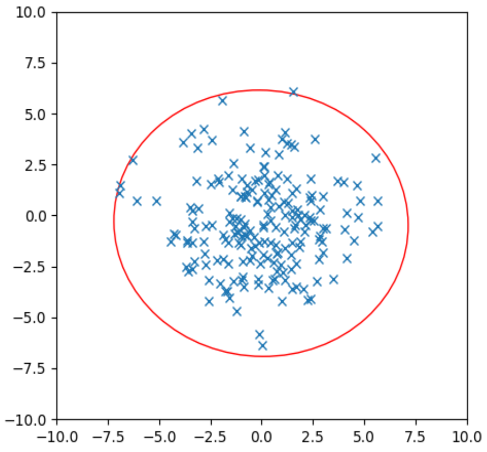
What is Σ? Which of these pictures are i.i.d. samples of X?

$$\Sigma = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}$$









What about dependence.

When we introduce dependence, we need to know the mean vector and the covariance matrix to define the distribution

(instead of just the mean and the variance).

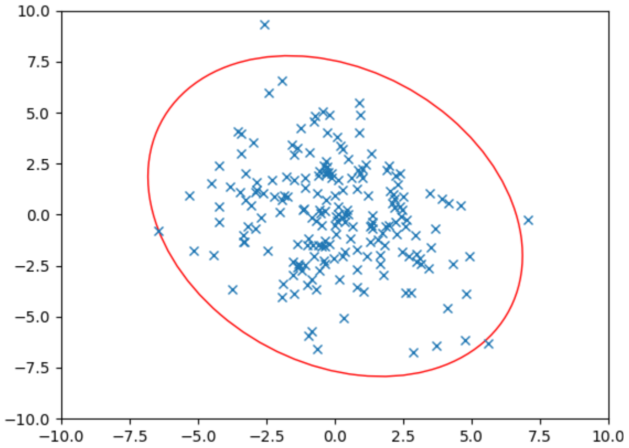
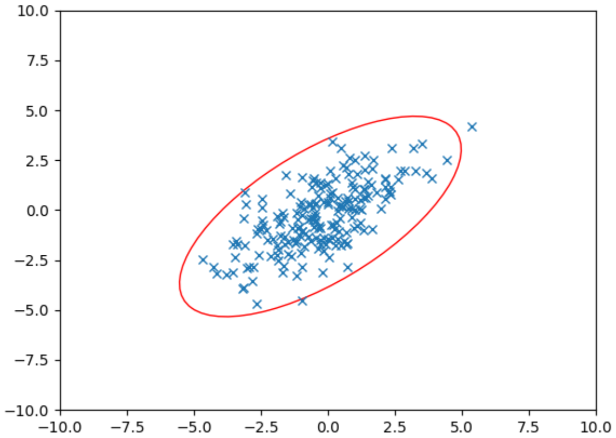
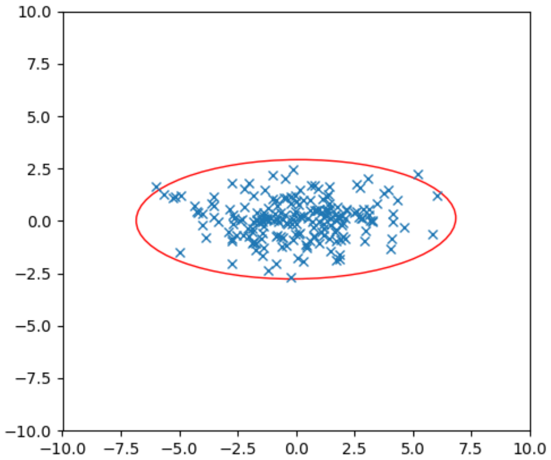
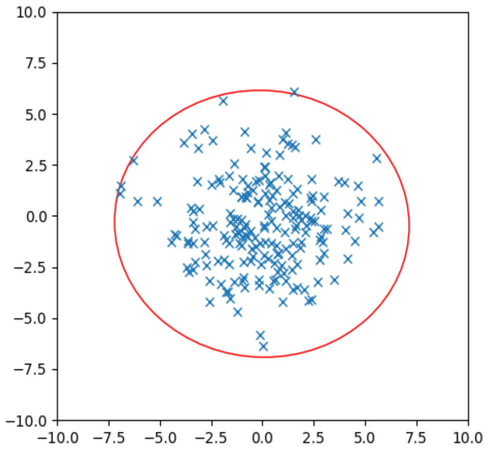
Let’s see a few examples…

Dependence

Let's think about 2 dimensions.

Let $X = [X_1, X_2]^T$ where $\mathrm{Var}(X_1) = 3$, $\mathrm{Var}(X_2) = 3$ BUT X_1 and X_2 are dependent. $\mathrm{Cov}(X_1, X_2) = 2$

What is Σ? Which of these pictures are i.i.d. samples of X?

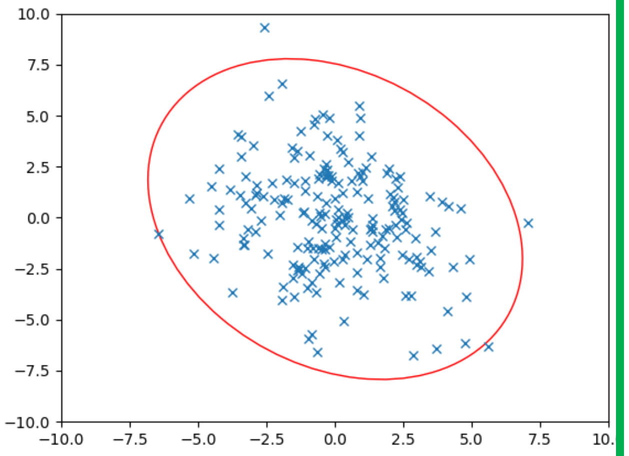
Dependence

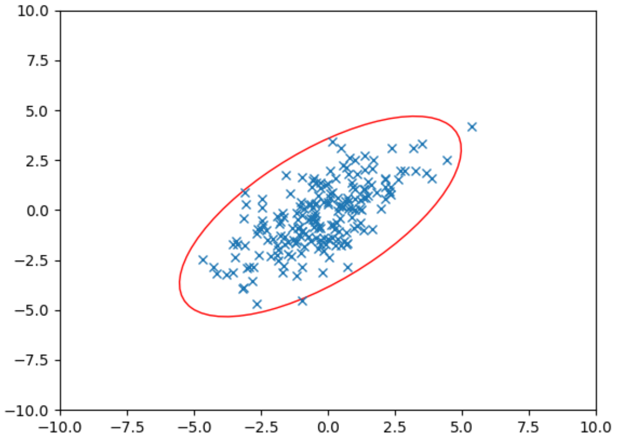
Let’s think about 2 dimensions.

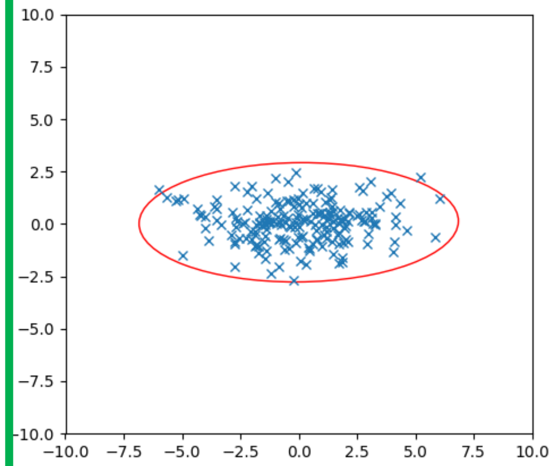
Let $X = [X_1, X_2]^T$ where $\mathrm{Var}(X_1) = 3$, $\mathrm{Var}(X_2) = 3$ BUT X_1 and X_2 are dependent. $\mathrm{Cov}(X_1, X_2) = 2$

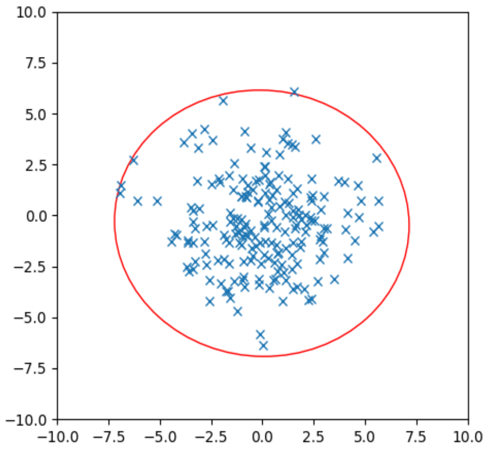
What is Σ? Which of these pictures are i.i.d. samples of X?

$$\Sigma = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}$$







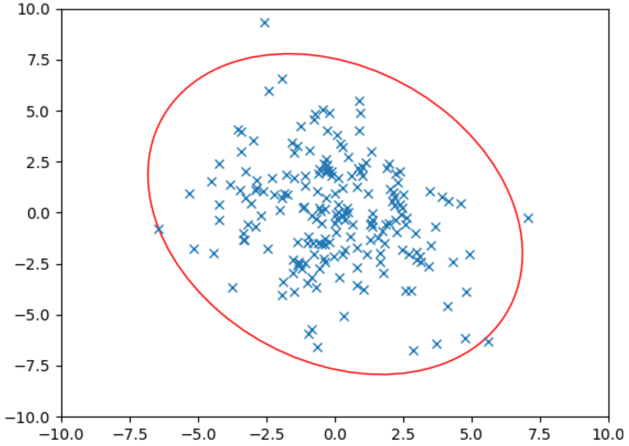


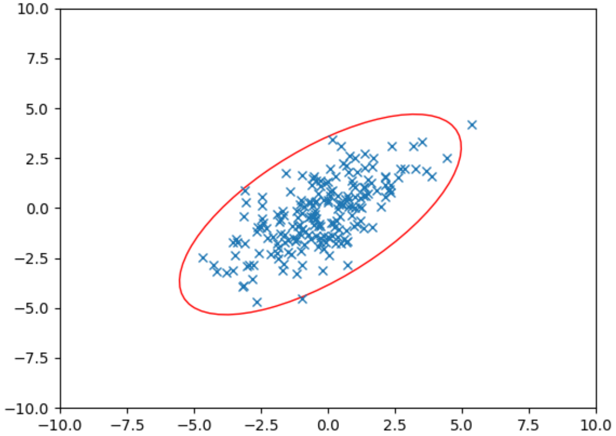
Dependence

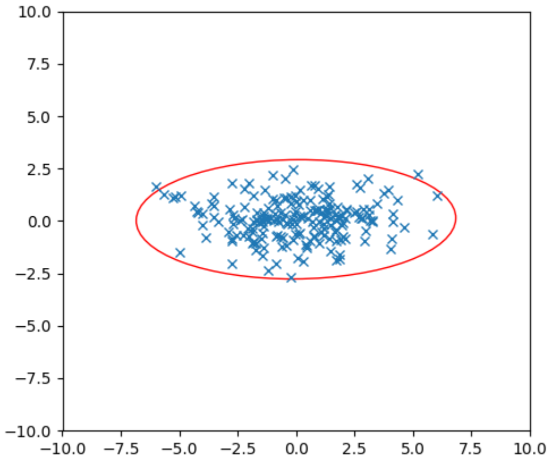
Let's think about 2 dimensions.

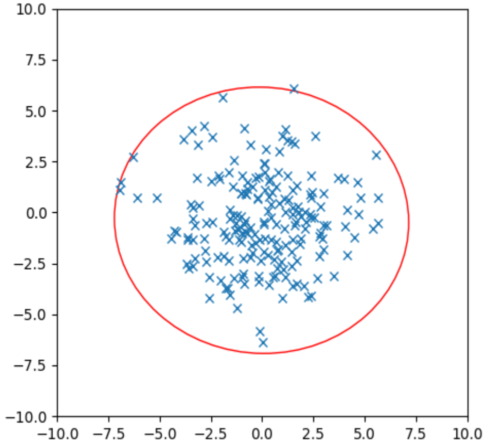
Let $X = [X_1, X_2]^T$ where $\mathrm{Var}(X_1) = 5$, $\mathrm{Var}(X_2) = 7$ BUT X_1 and X_2 are dependent. $\mathrm{Cov}(X_1, X_2) = -2$

What is Σ? Which of these pictures are i.i.d. samples of X?









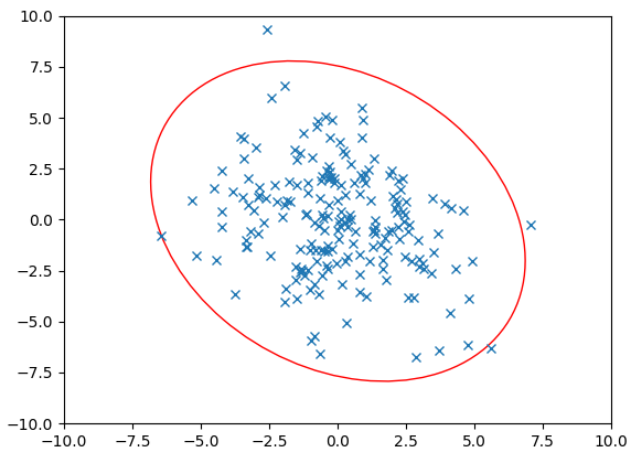
Dependence

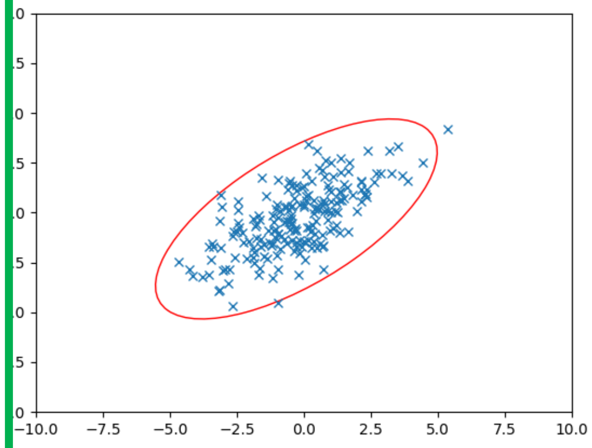
Let's think about 2 dimensions.

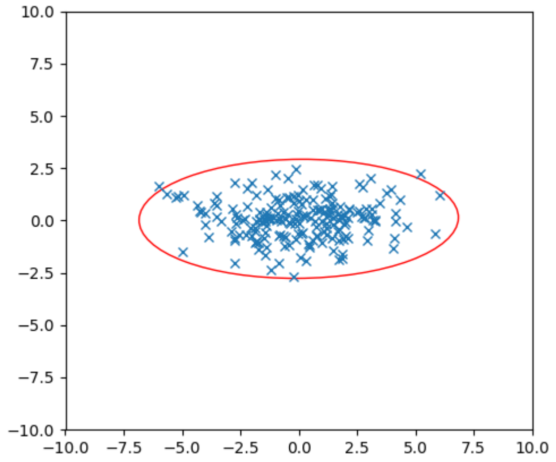
Let $X = [X_1, X_2]^T$ where $\text{Var}(X_1) = 5$, $\text{Var}(X_2) = 7$ BUT X_1 and X_2 are dependent. $\text{Cov}(X_1, X_2) = -2$

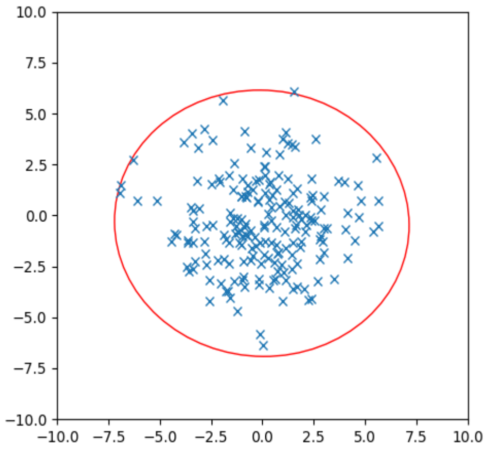
What is Σ? Which of these pictures are i.i.d. samples of X?

$$\Sigma = \begin{bmatrix} 5 & -2 \\ -2 & 7 \end{bmatrix}$$









Using the Covariance Matrix

What were those ellipses in those datasets?

How do we know how many standard deviations from the mean a 2D point is, for the independent, variance 1 ones

Well $(x_1 - \mathbb{E}[X_1])$ is the distance from x to the center in the x-direction.

And $(x_2 - \mathbb{E}[x_2])$ is the distance from x to the center in the y-direction.

So the number of standard deviations is $\sqrt{(x_1 - \mathbb{E}[X_1])^2 + (x_2 - \mathbb{E}[x_2])^2}$

That's just the distance!

In general, the major/minor axes of those ellipses were the eigenvectors of the covariance matrix. And the associated eigenvalues tell you how the directions should be weighted.

Probability and ML

Many problems in ML: Given a bunch of data points, you'll find a function f that you hope will predict future points well.

We usually assume there is some true distribution $\mathcal{D}$ of data points (e.g. all theoretical possible houses and their prices).

You get a dataset S that you assume was sampled from $\mathcal{D}$ to find f_S

f_S depends on the data (just like our MLEs depended on the data), so before you knew what S was, f was a random variable. You then want to figure out what the true error is if you knew $\mathcal{D}$.

Probability and ML

But $\mathcal{D}$ is a theoretical construct. I can't calculate probabilities.
What can we do instead? Get a second dataset T drawn from $\mathcal{D}$ (drawn independently of S)

(or actually save part of your database before you start).

Then $\mathbb{E}_{\mathcal{D}}[\text{error of } f] = \mathbb{E}_T[\text{error of } f_S | S]$

But how confident can you be? You can make confidence intervals

(statements like the true error is within 5% of our estimate with probability at least $.9$) using concentration inequalities.

One more ML preview

Experiments with the correct expectation

Gradient Descent

How did I “train” the model in the first place?

Lots of options…one is “gradient descent”

Think of the “error on the data-set” as a function we’re minimizing.

Take the gradient (derivative of the error with respect to every coefficient in your function), and move in the direction of lower error.

But finding the gradient is expensive…what if we could just estimate the gradient…like with a subset of the data.

Experiments with the correct expectation

We could find an unbiased estimator of the true gradient!

An experiment where the expectation of the vector we get is the true gradient (but might not be the true gradient itself).

Looking at a random subset of the full data-set would let us do that!

For many optimizations it's faster to approximate the gradient. You'll be less accurate in each step (since your gradient is less accurate), but each step is much faster, and that tradeoff can be worth it.

Conditional Expectation Practice

Practice with conditional expectations

Consider of the following process:

Flip a fair coin, if it's heads, pick up a 4-sided die; if it's tails, pick up a 6-sided die (both fair)

Roll that die independently 3 times. Let X_1, X_2, X_3 be the results of the three rolls.

What is $\mathbb{E}[X_2]$? $\mathbb{E}[X_2|X_1 = 5]$? $\mathbb{E}[X_2|X_3 = 1]$?

Using conditional expectations

Let F be the event "the four sided die was chosen"

$\mathbb{E}[X_2] = \mathbb{P}(F)\mathbb{E}[X_2|F] + \mathbb{P}(\bar{F})\mathbb{E}[X_2|\bar{F}]$

$= \frac{1}{2} \cdot 2.5 + \frac{1}{2} \cdot 3.5 = 3$

$\mathbb{E}[X_2|X_1 = 5]$ event $X_1 = 5$ tells us we're using the 6-sided die.

$\mathbb{E}[X_2|X_1 = 5] = 3.5$

$\mathbb{E}[X_2|X_3 = 1]$ We aren't sure which die we got, but…is it still 50/50?

Setup

Let E *be the event "*$X_3 = 1$"

$\mathbb{P}(E) = \frac{1}{2} \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{1}{4} = \frac{5}{24}$

$\mathbb{P}(F|E) = \frac{\mathbb{P}(E|F) \cdot \mathbb{P}(F)}{\mathbb{P}(E)}$

$= \frac{\frac{1}{4} \cdot \frac{1}{2}}{5/24} = \frac{3}{5}$

$\mathbb{P}(\bar{F}|E) = \frac{\mathbb{P}(E|\bar{F}) \cdot \mathbb{P}(\bar{F})}{\mathbb{P}(E)} = \frac{\frac{1}{6} \cdot \frac{1}{2}}{5/24} = \frac{2}{5}$ (we could also get this with LTP, but it's good confirmation)

Analysis

$\mathbb{E}[X_2|X_3 = 1] = \mathbb{P}(F|X_3 = 1)\mathbb{E}[X_2|X_3 = 1 \cap F] + \mathbb{P}(\bar{F}|X_3 = 1)\mathbb{E}[X_2|X_3 = 1 \cap \bar{F}]$

Wait what?

This is the LTE, applied in the space where we've conditioned on $X_3 = 1$.

Everything is conditioned on $X_3 = 1$. Beyond that conditioning, it's LTE.

$= \frac{3}{5} \cdot 2.5 + \frac{2}{5} \cdot 3.5 = 2.9$.

A little lower than the unconditioned expectation. Because seeing a 1 has made it ever so slightly more probable that we're using the 4-sided die.