

- The present form of support vector machine (SVM) was largely developed at AT&T Bell Laboratories by Vapnik and co-workers.
- Known as a **maximum margin classifier**.
- Originally proposed for classification and soon applied to regression and time series prediction.
- One of the most efficient **supervised learning** methods.

Problem

- Given a set of training samples

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N), x_i \in R^n, y_i \in \{-1, 1\},$$

find a function $f(x, \alpha)$ to classify the samples, such that

$$f(x_i, \alpha) \begin{cases} > 0, & \forall y_i = +1; \\ < 0, & \forall y_i = -1, \end{cases}$$

where α denotes the parameters.

Problem

- Given a set of training samples

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N), x_i \in R^n, y_i \in \{-1, 1\},$$

find a function $f(x, \alpha)$ to classify the samples, such that

$$f(x_i, \alpha) \begin{cases} > 0, & \forall y_i = +1; \\ < 0, & \forall y_i = -1, \end{cases}$$

where α denotes the parameters.

- For a testing sample x , we can predict its label by $\text{sign}[f(x, \alpha)]$.

Problem

- Given a set of training samples

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N), x_i \in R^n, y_i \in \{-1, 1\},$$

find a function $f(x, \alpha)$ to classify the samples, such that

$$f(x_i, \alpha) \begin{cases} > 0, & \forall y_i = +1; \\ < 0, & \forall y_i = -1, \end{cases}$$

where α denotes the parameters.

- For a testing sample x , we can predict its label by $\text{sign}[f(x, \alpha)]$.
- $f(x, \alpha) = 0$ is called the separation hyperplane.

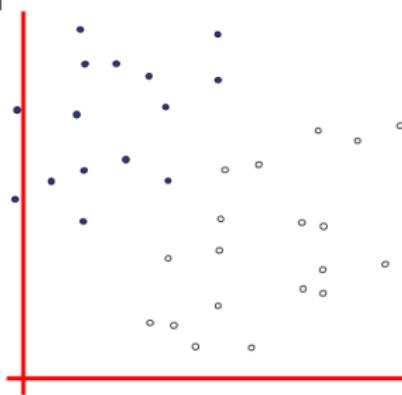
Linear classifiers

Linear hyperplane

$$f(x, w, b) = \langle x, w \rangle + b = 0$$

Consider the linearly separable case, there are infinite number of hyperplanes that can do the job.

- denotes +1
- denotes -1



How would you
classify this data?

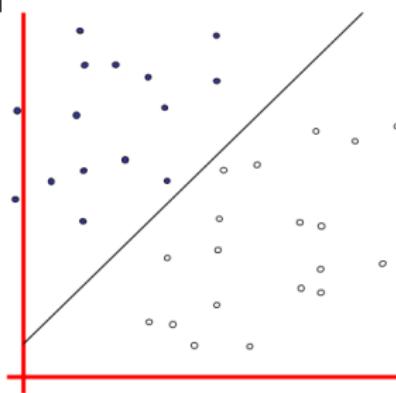
Linear classifiers

Linear hyperplane

$$f(x, w, b) = \langle x, w \rangle + b = 0$$

Consider the linearly separable case, there are infinite number of hyperplanes that can do the job.

- denotes +1
- denotes -1



How would you
classify this data?

Linear classifiers

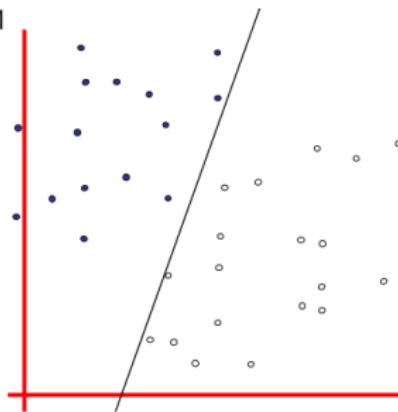
Linear hyperplane

$$f(x, w, b) = \langle x, w \rangle + b = 0$$

Consider the linearly separable case, there are infinite number of hyperplanes that can do the job.

- denotes +1

- denotes -1



How would you
classify this data?

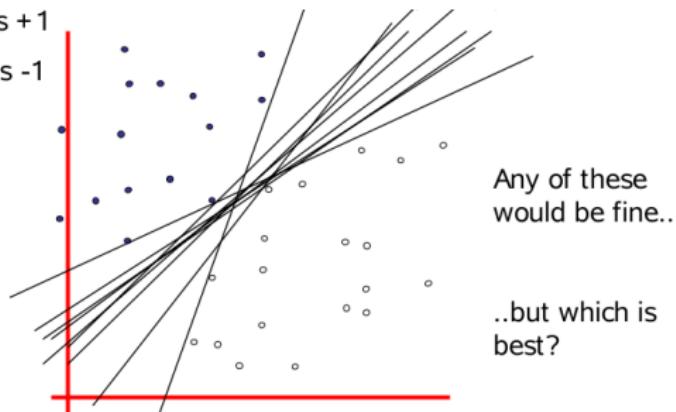
Linear classifiers

Linear hyperplane

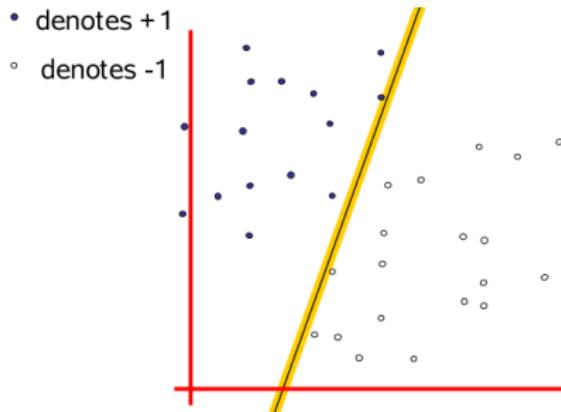
$$f(x, w, b) = \langle x, w \rangle + b = 0$$

Consider the linearly separable case, there are infinite number of hyperplanes that can do the job.

- denotes +1
- denotes -1

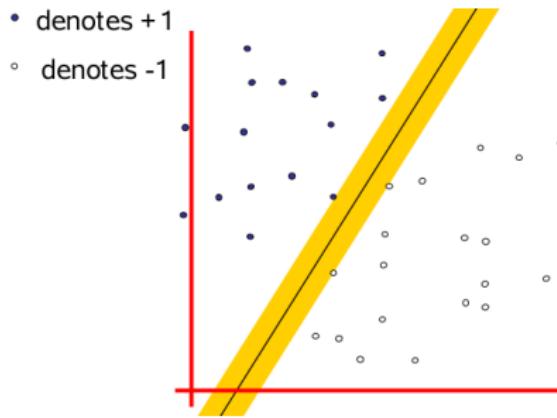


Margin of a linear classifier



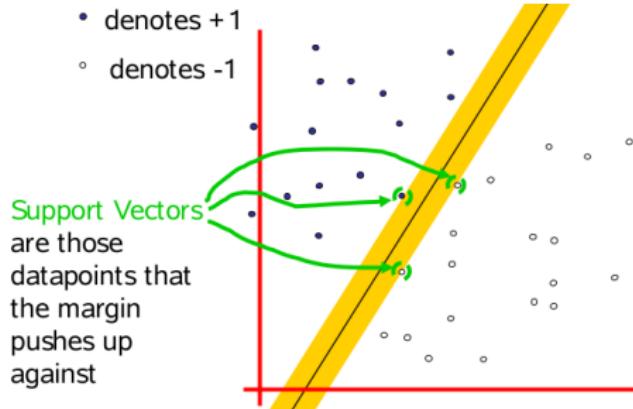
Definition: the width that the boundary could be increased by before hitting a data point.

Maximum margin linear classifier



Definition: the linear classifier with the maximum margin.

Support vectors



Problem formulation

To formulate the margin, we further requires that for all samples

$$f(x_i, \alpha) = \langle x_i, w \rangle + b \begin{cases} \geq +1, & \forall y_i = +1; \\ \leq -1, & \forall y_i = -1. \end{cases}$$

or

$$y_i(\langle x_i, w \rangle + b) \geq 1, \quad i = 1, \dots, N.$$

Problem formulation

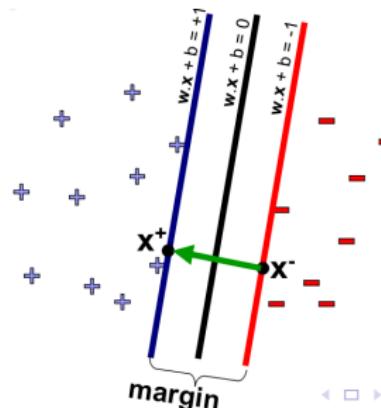
To formulate the margin, we further require that for all samples

$$f(x_i, \alpha) = \langle x_i, w \rangle + b \begin{cases} \geq +1, & \forall y_i = +1; \\ \leq -1, & \forall y_i = -1. \end{cases}$$

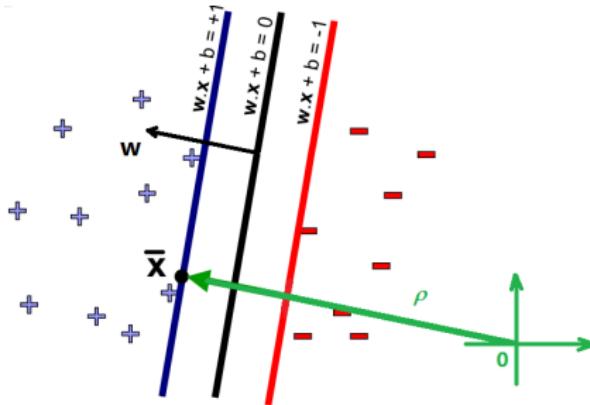
or

$$y_i(\langle x_i, w \rangle + b) \geq 1, \quad i = 1, \dots, N.$$

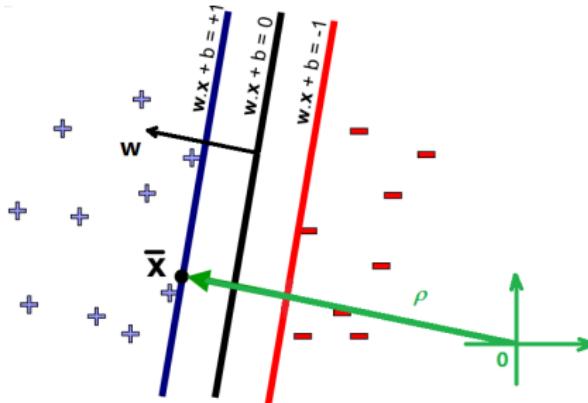
- We have introduced two additional hyperplanes
 $\langle x, w \rangle + b = \pm 1$ parallel to the separation hyperplane
 $\langle x, w \rangle + b = 0$



What is the margin? The distance between the two new hyperplanes.

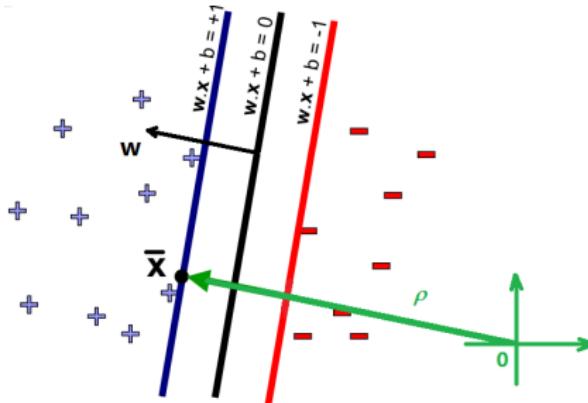


What is the margin? The distance between the two new hyperplanes.



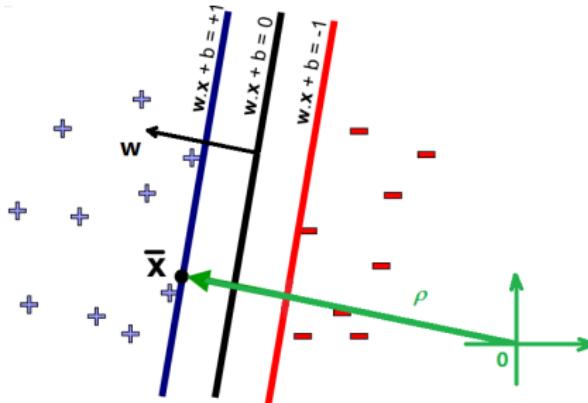
- The minimum distance between the hyperplane $\langle x, w \rangle + b = 1$ and the origin is $\rho_1 = \frac{1-b}{\|w\|}$. (why?)

What is the margin? The distance between the two new hyperplanes.



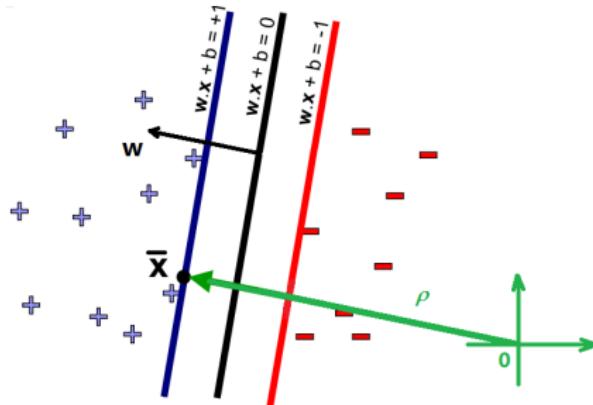
- The minimum distance between the hyperplane $\langle x, w \rangle + b = 1$ and the origin is $\rho_1 = \frac{1-b}{\|w\|}$. (why?)
- The minimum distance between the hyperplane $\langle x, w \rangle + b = -1$ and the origin is $\rho_2 = \frac{-1-b}{\|w\|}$.

What is the margin? **The distance between the two new hyperplanes.**



- The minimum distance between the hyperplane $\langle x, w \rangle + b = 1$ and the origin is $\rho_1 = \frac{1-b}{\|w\|}$. (why?)
- The minimum distance between the hyperplane $\langle x, w \rangle + b = -1$ and the origin is $\rho_2 = \frac{-1-b}{\|w\|}$.
- The margin is $|\rho_1 - \rho_2| = 2/\|w\|$.

How to calculate ρ_1 and ρ_2 ?



Note $\bar{x} = \rho_1 w / \|w\|$, where $w / \|w\|$ is the unit vector along the direction w . Since \bar{x} is on the blue hyperplane, then

$$\langle \rho_1 w / \|w\|, w \rangle + b = 1$$

which follows $\rho_1 = \frac{1-b}{\|w\|}$. Similarly, we obtain $\rho_2 = \frac{-1-b}{\|w\|}$.

Maximizing the margin is the same thing as minimizing the norm of \mathbf{w}

Our goal is to maximize the margin. Among all possible hyperplanes meeting the constraints, we will choose the hyperplane with the smallest $\|\mathbf{w}\|$ because it is the one which will have the biggest margin.

This give us the following [optimization problem](#):

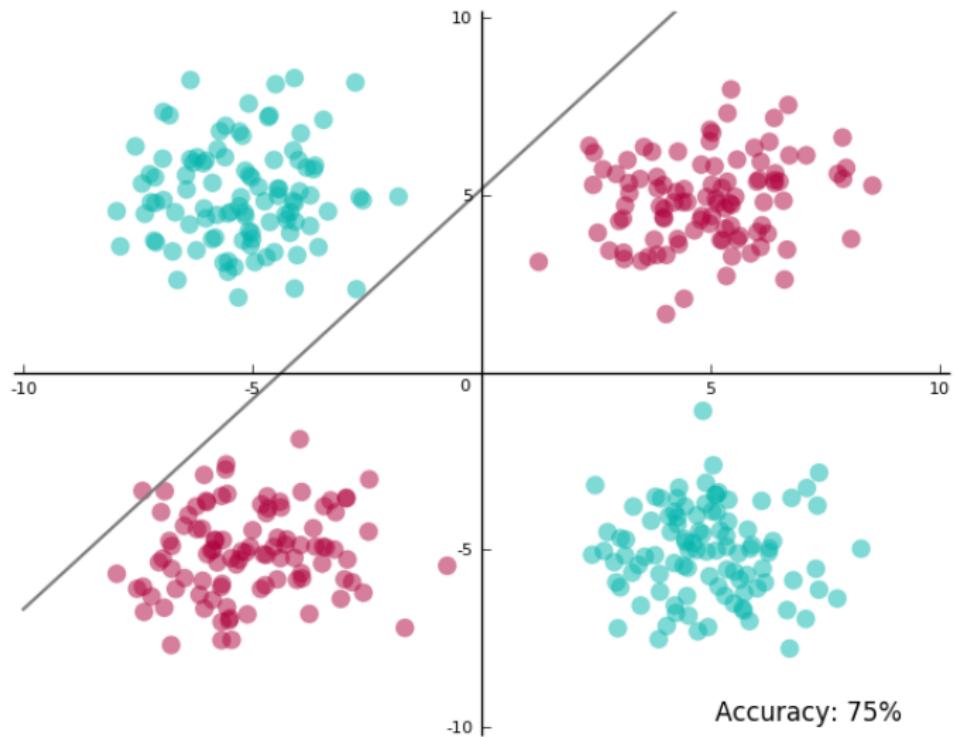
Minimize in (\mathbf{w}, b)

$$\|\mathbf{w}\|$$

$$\text{subject to } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$$

(for any $i = 1, \dots, n$)

Solving this problem is like solving and equation. Once we have solved it, we will have found the couple (\mathbf{w}, b) for which $\|\mathbf{w}\|$ is the smallest possible and the constraints we fixed are met. Which means we will have the equation of the optimal hyperplane !



The kernel-based function is exactly equivalent to preprocessing the data by applying similarity function to all inputs, then learning a linear model in the new transformed space.

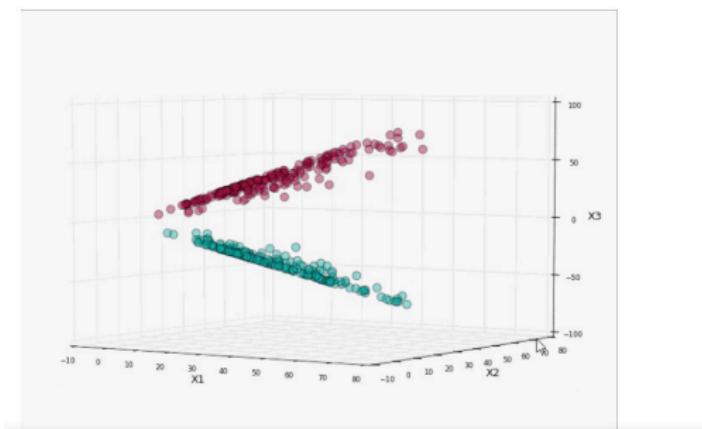
We start with the dataset in the above figure, and project it into a three-dimensional space where the new coordinates are:

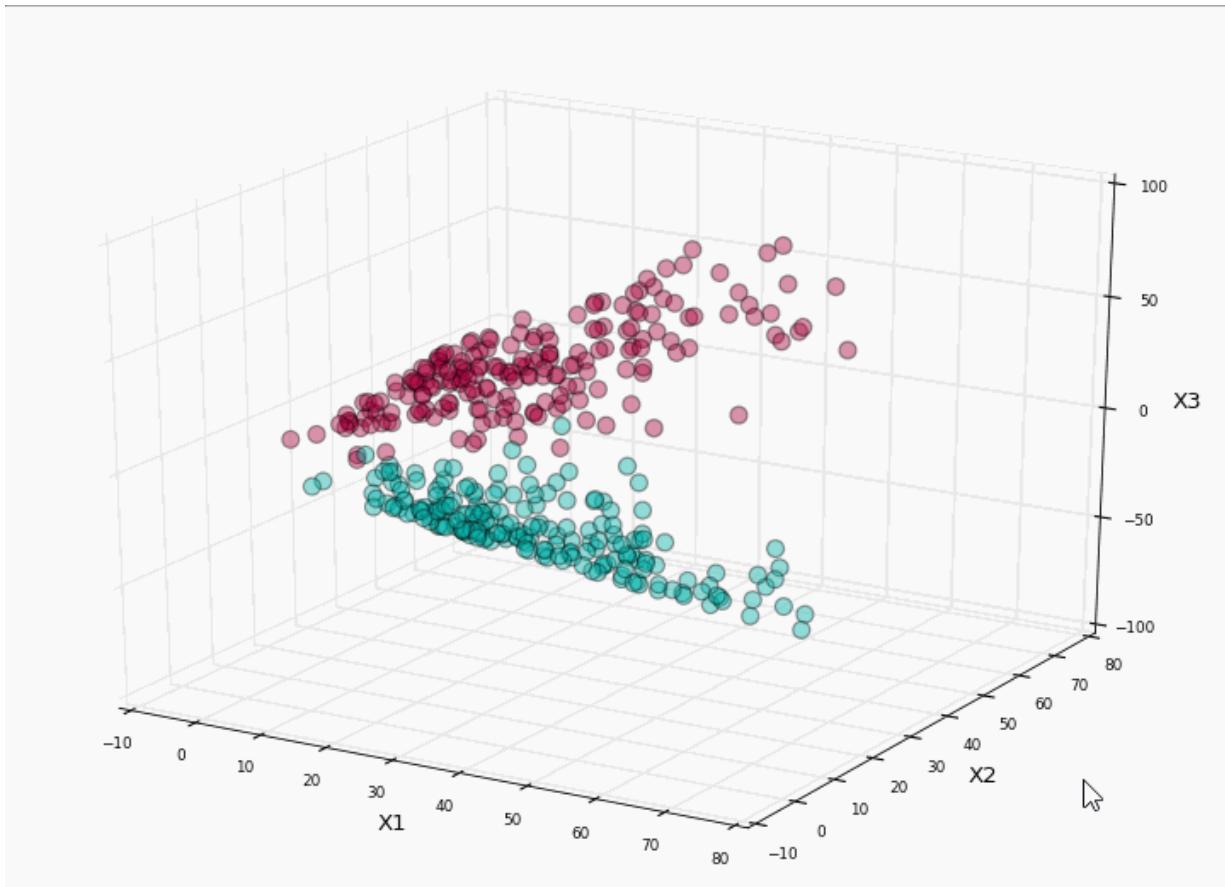
$$X_1 = x_1^2$$

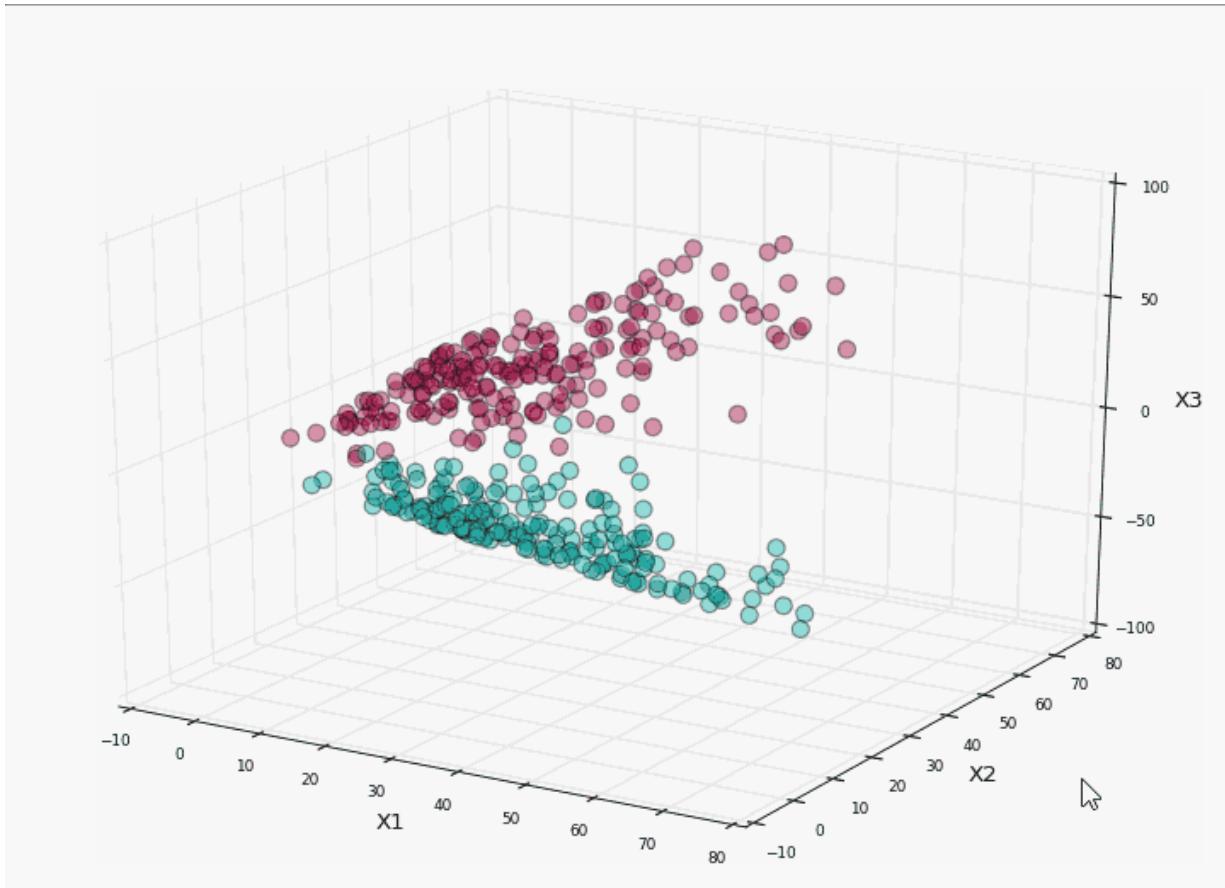
$$X_2 = x_2^2$$

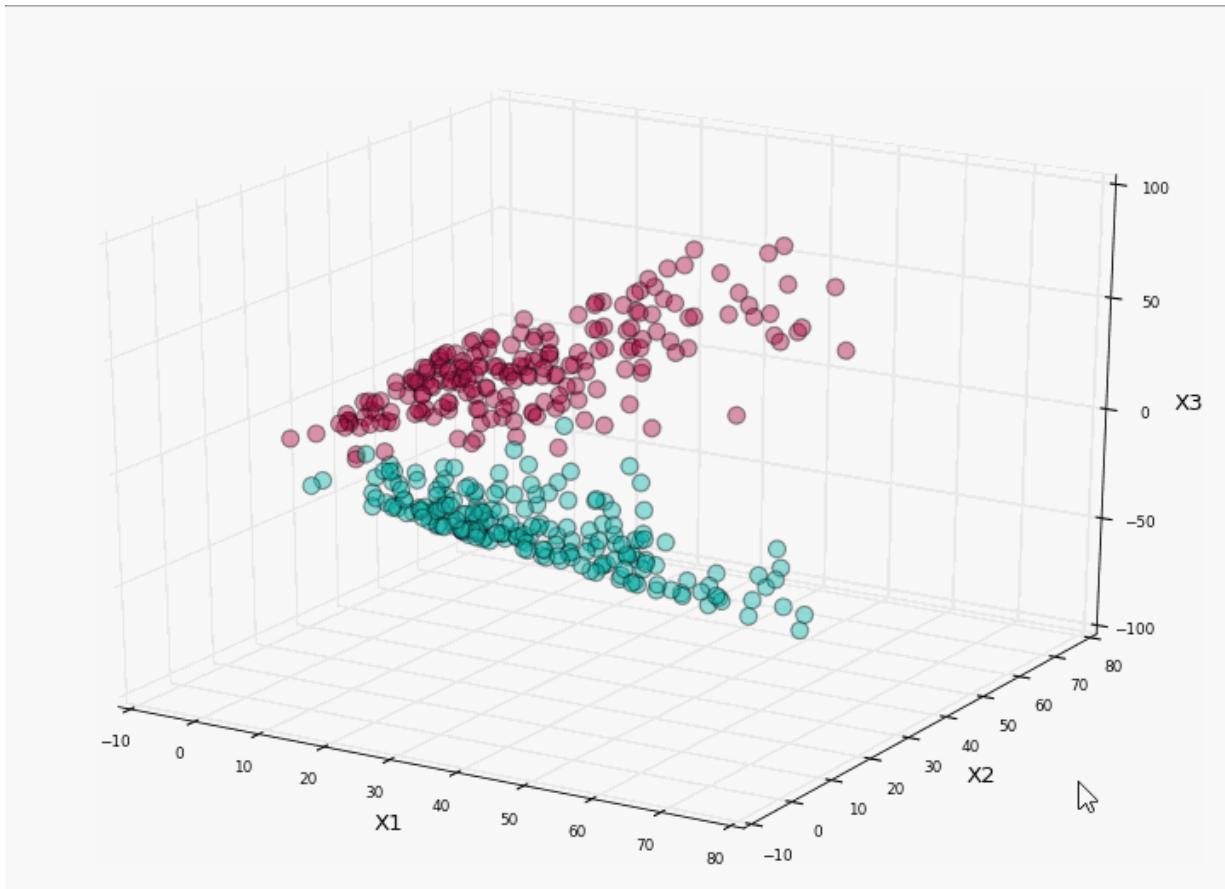
$$X_3 = \sqrt{2}x_1x_2$$

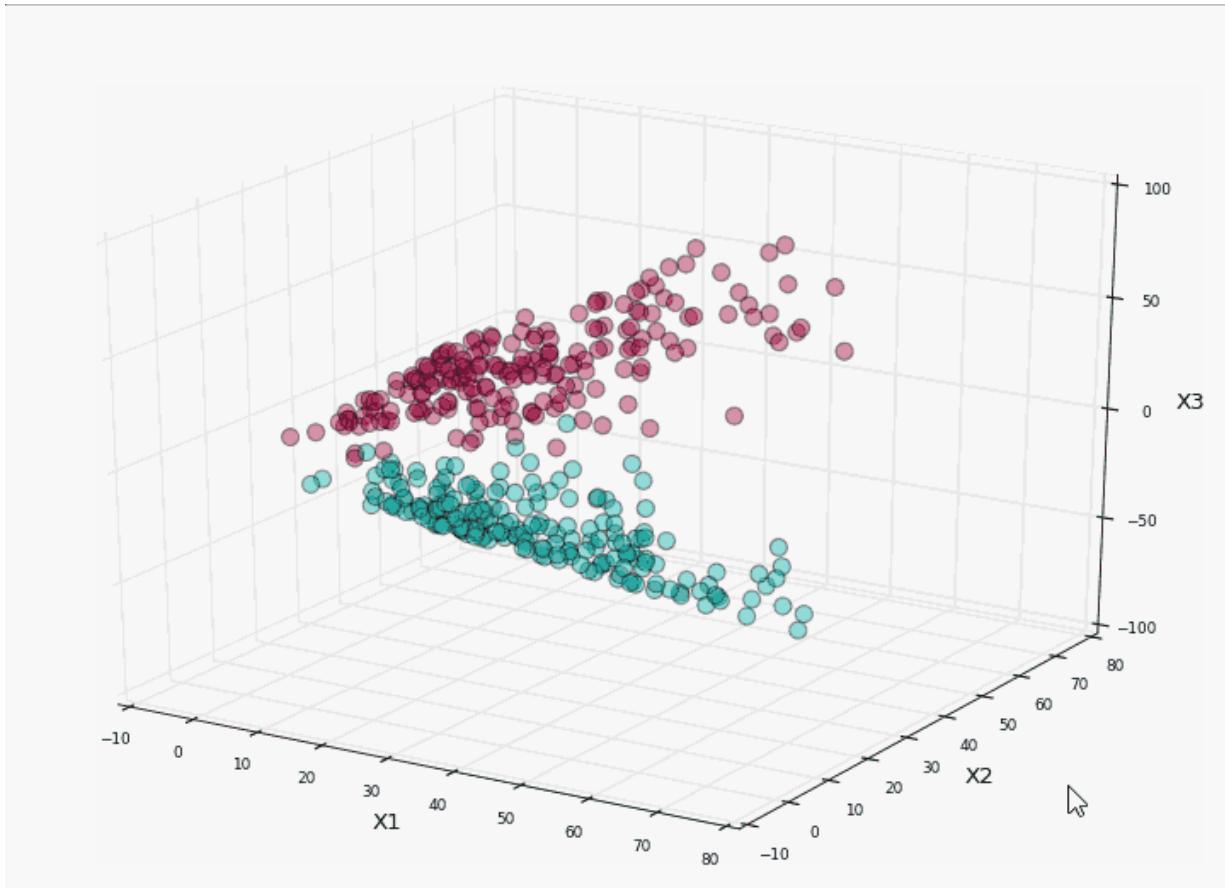
This is what the projected data looks like. Do you see a place where we just might be able to slip in a plane?

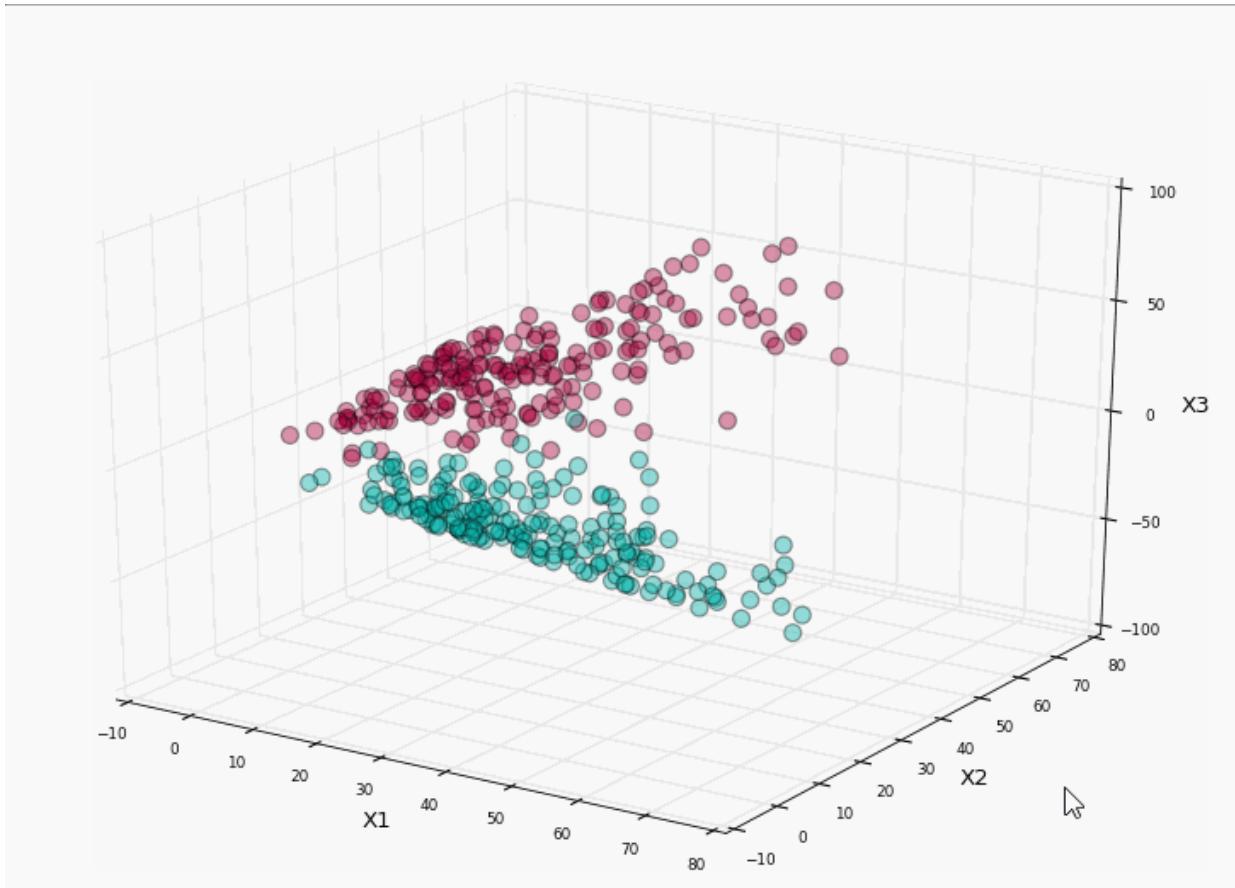


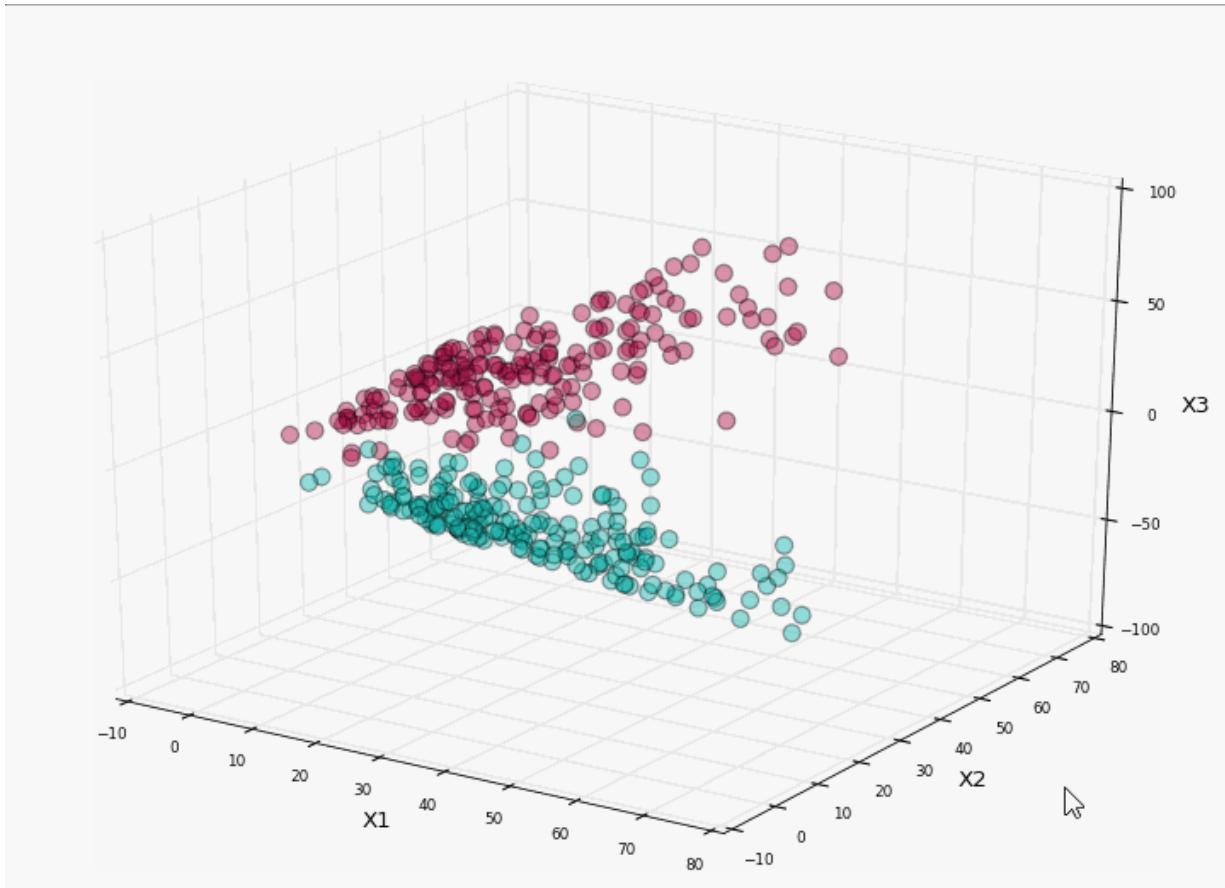


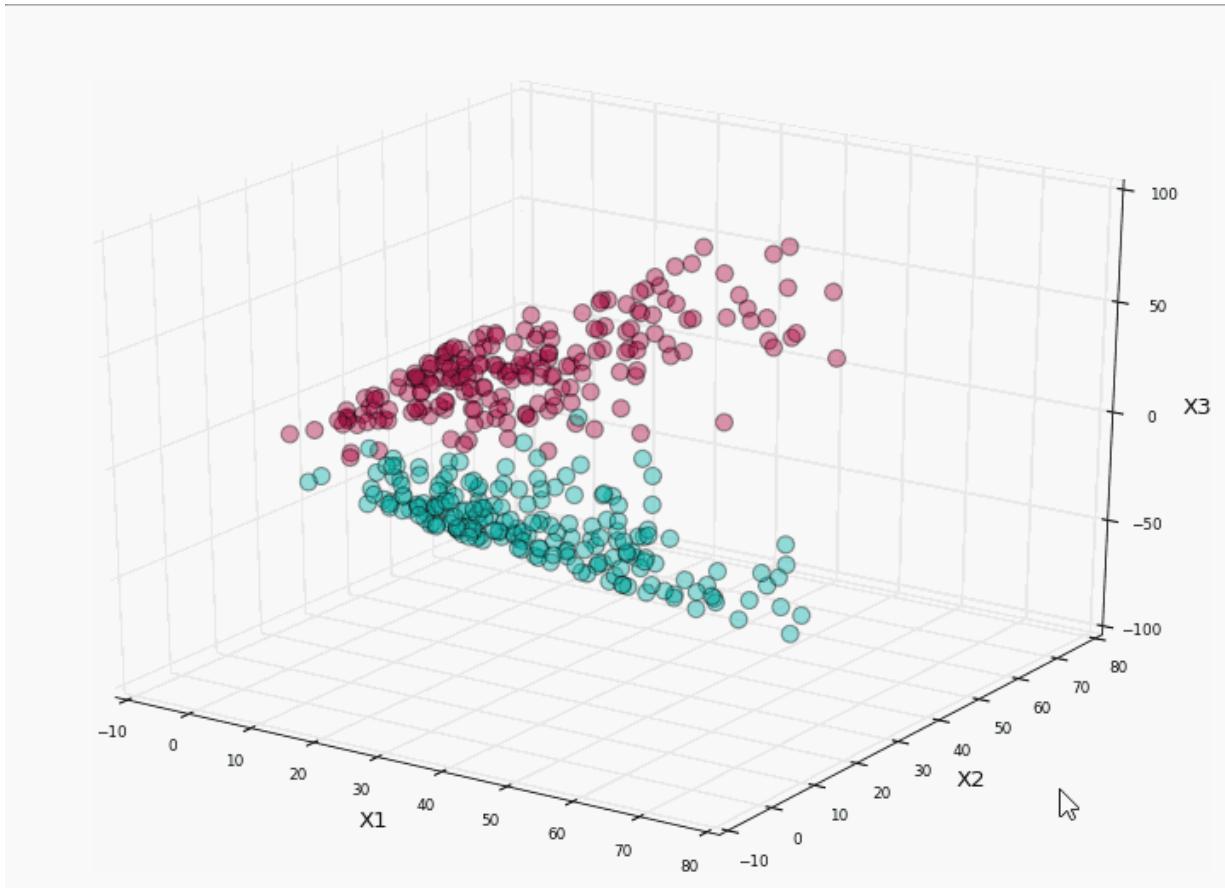


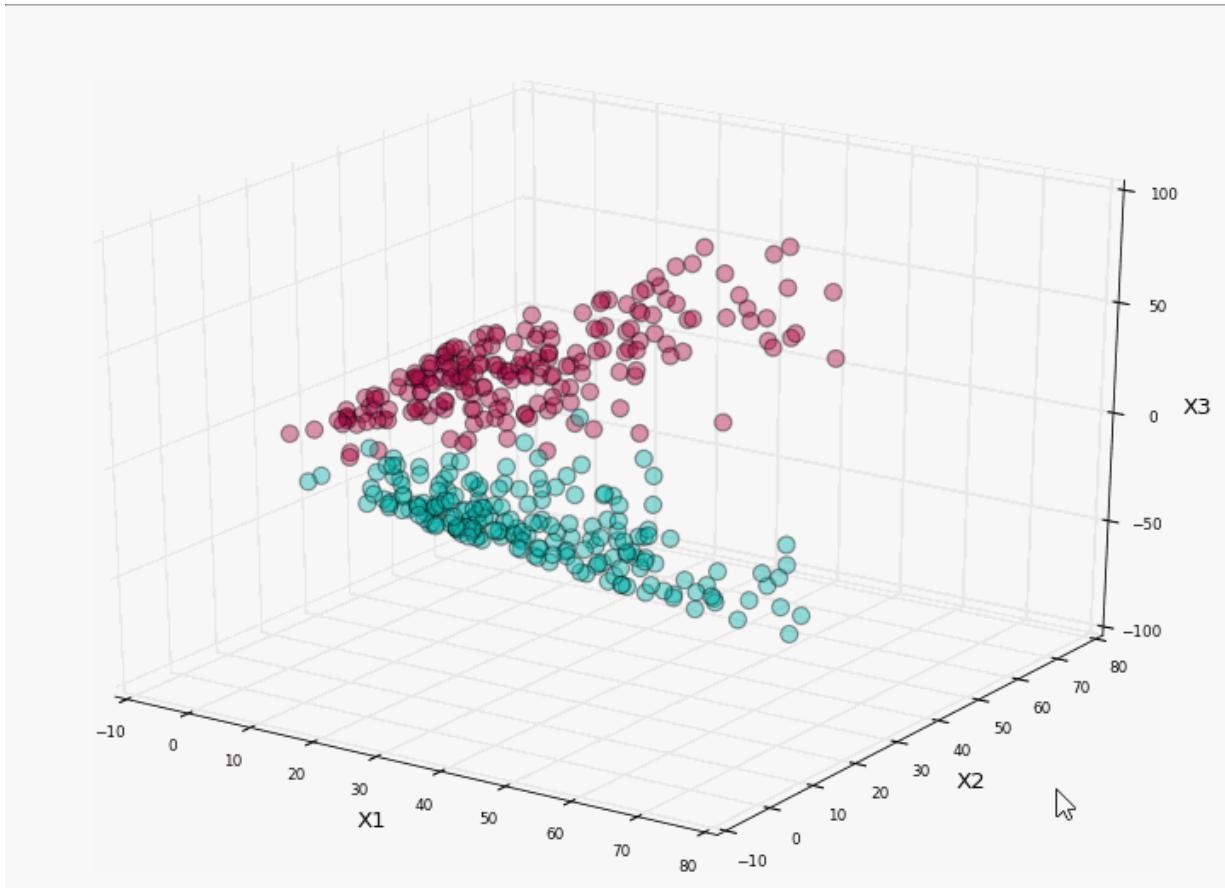


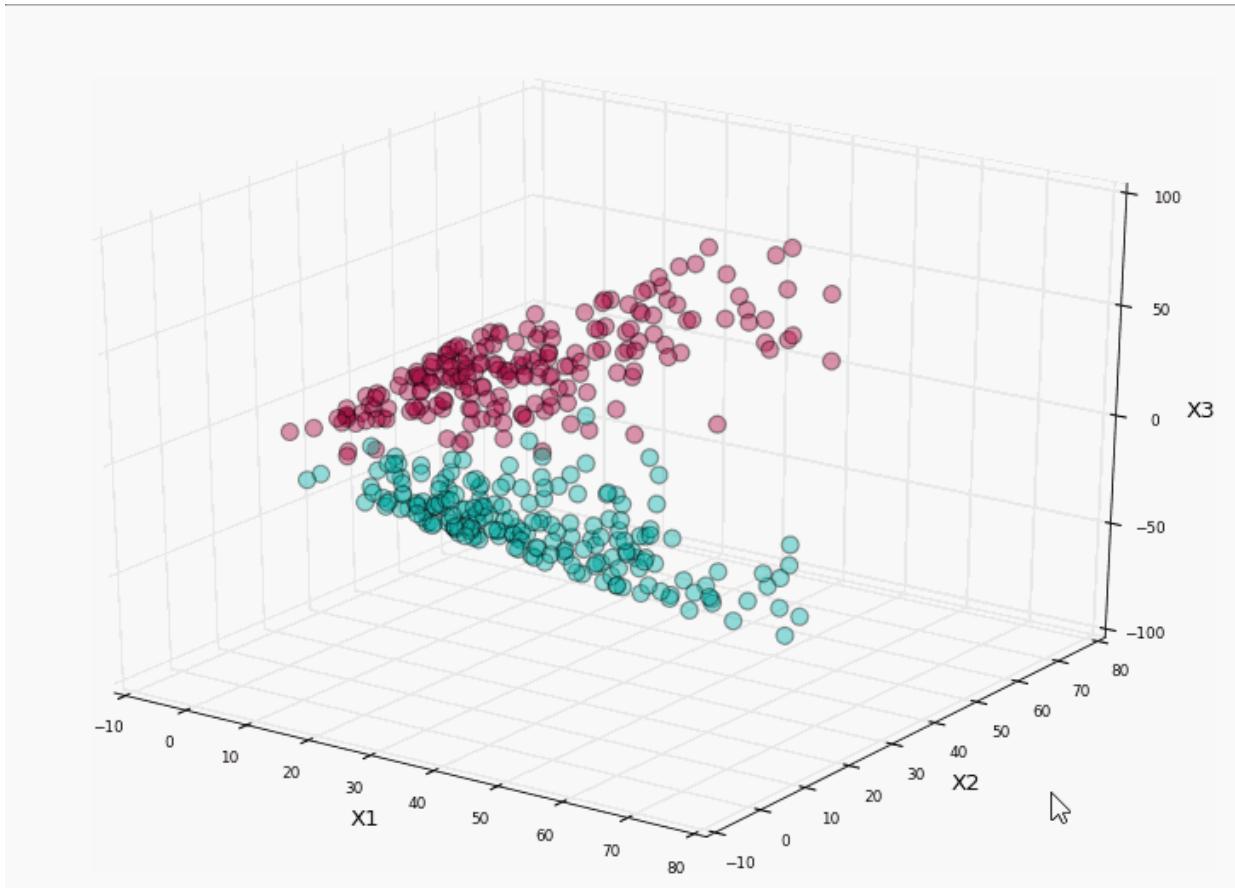


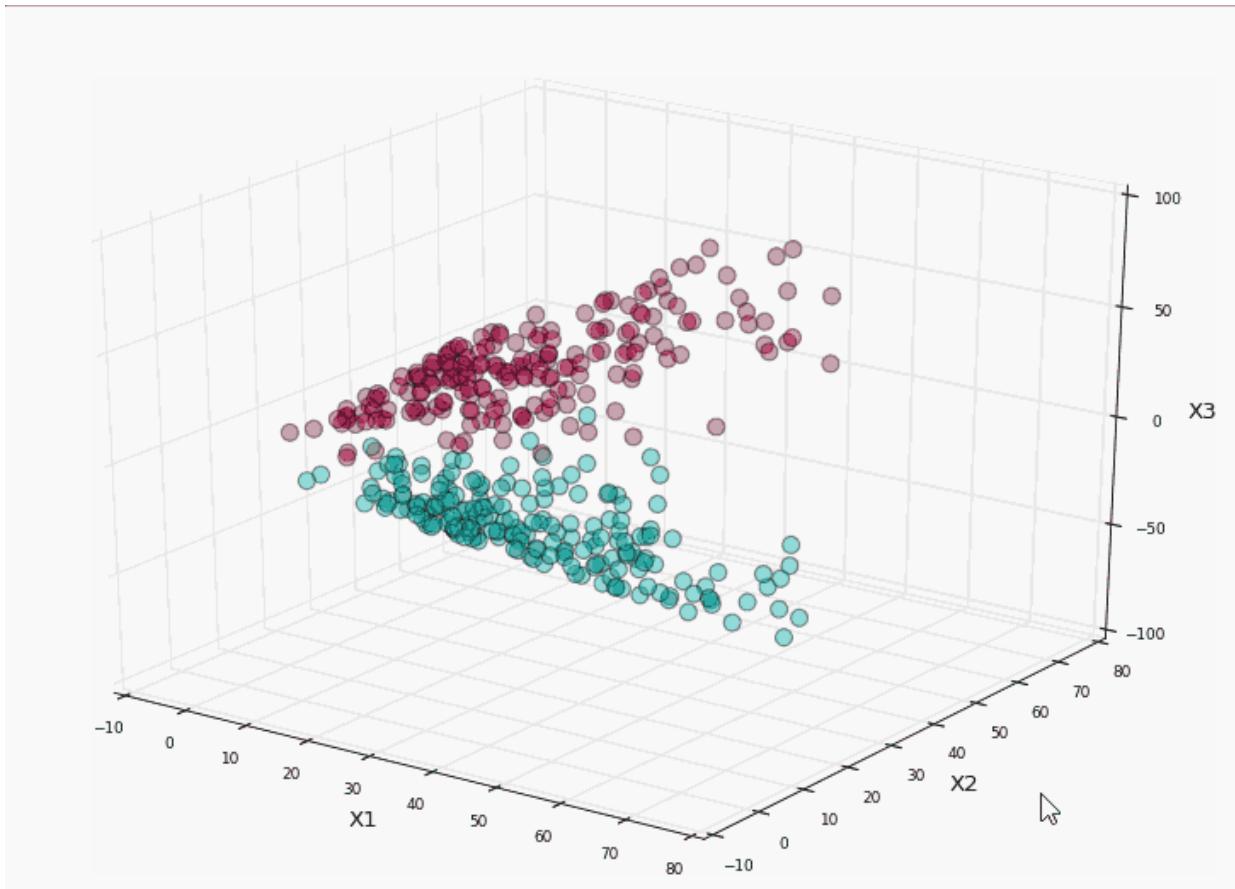


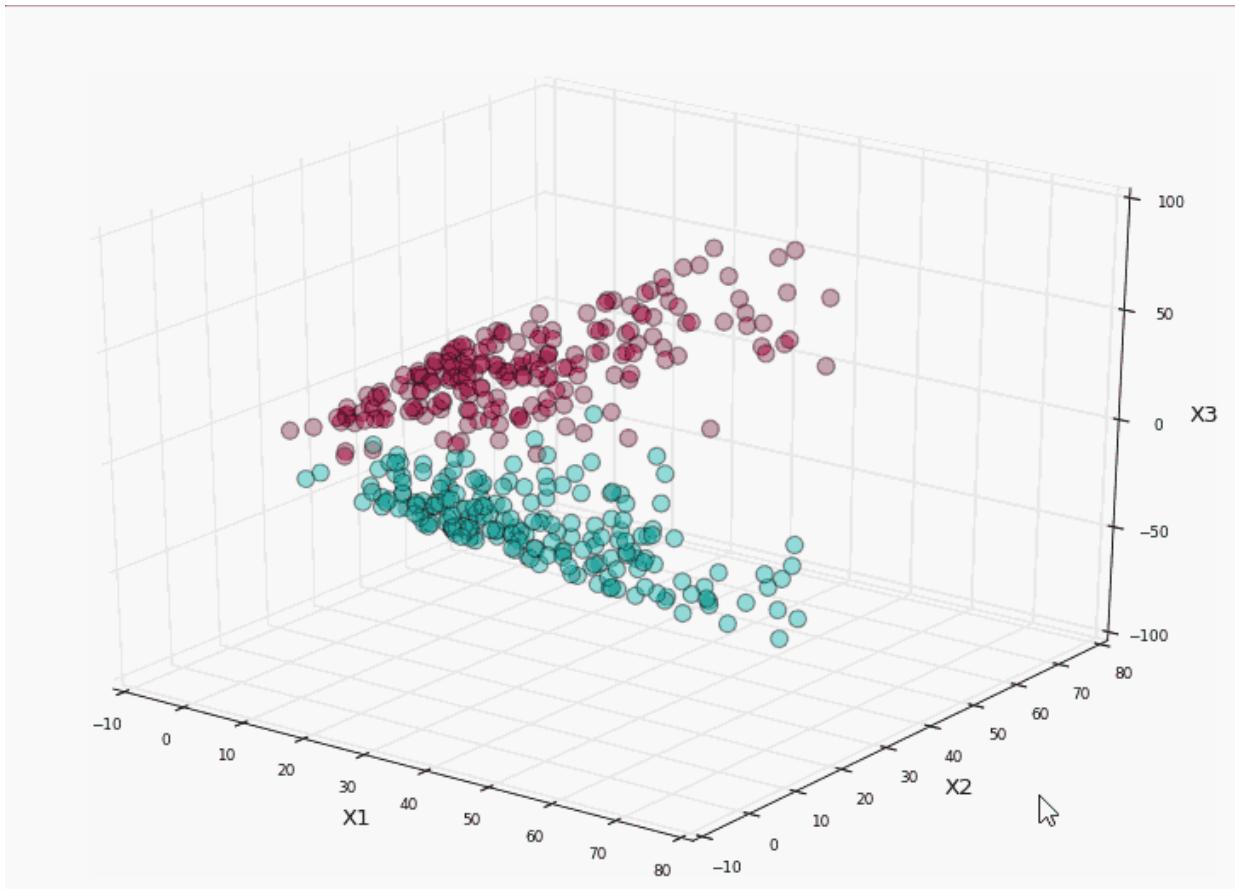


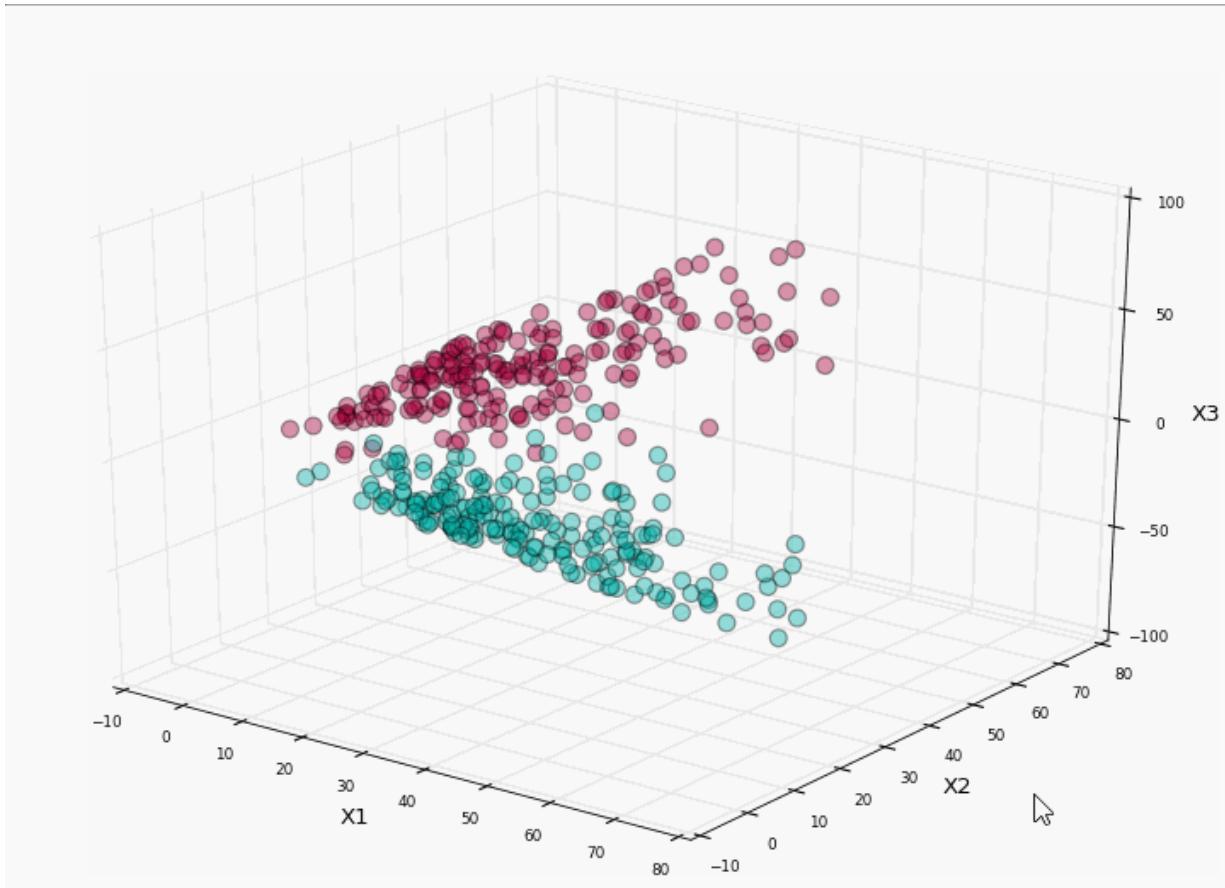


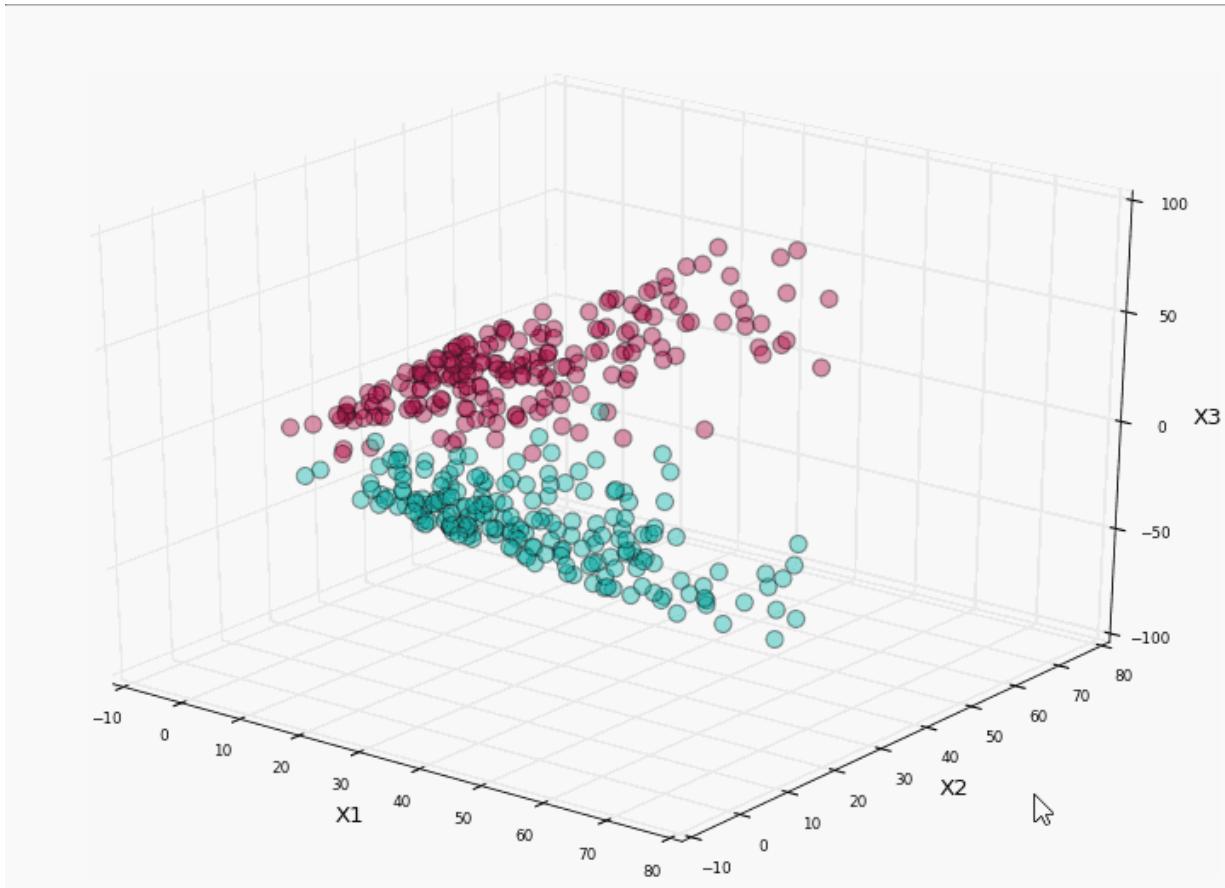


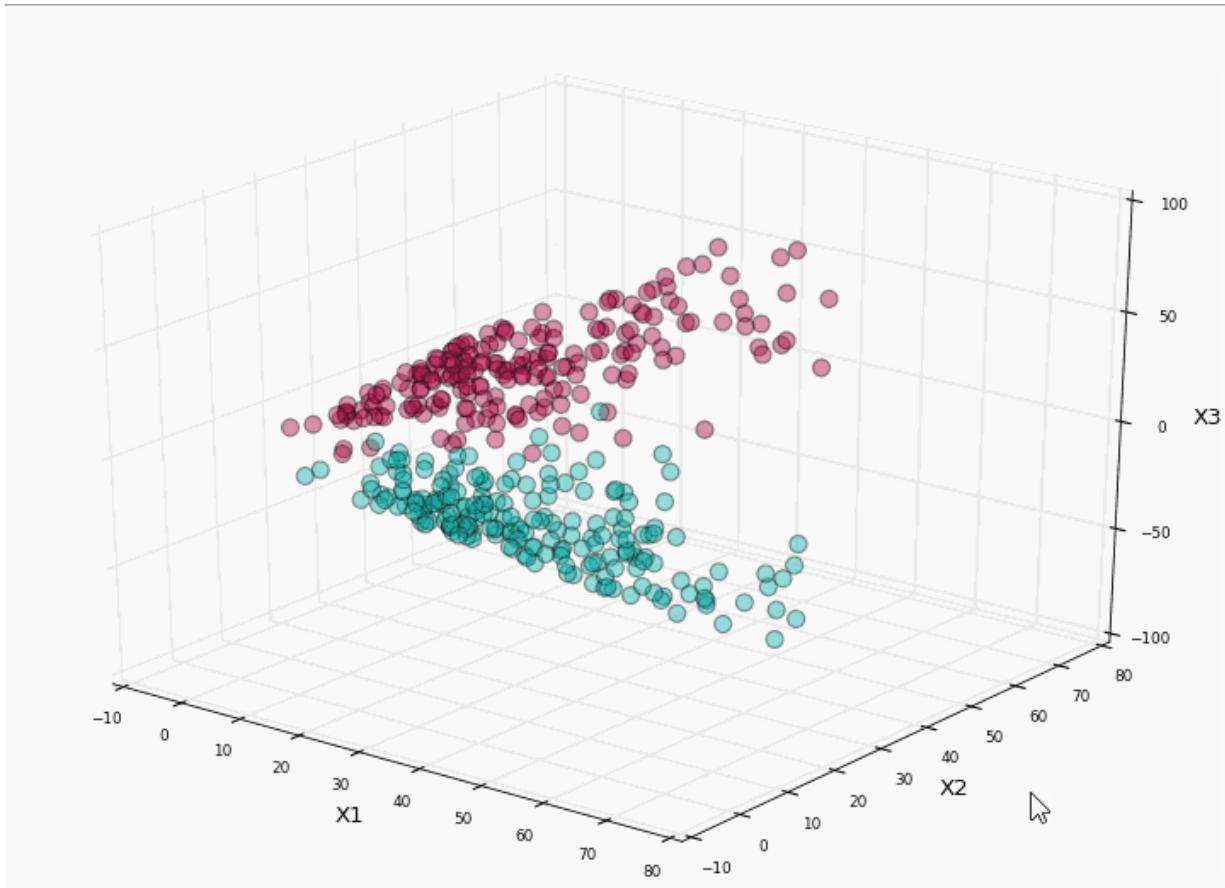


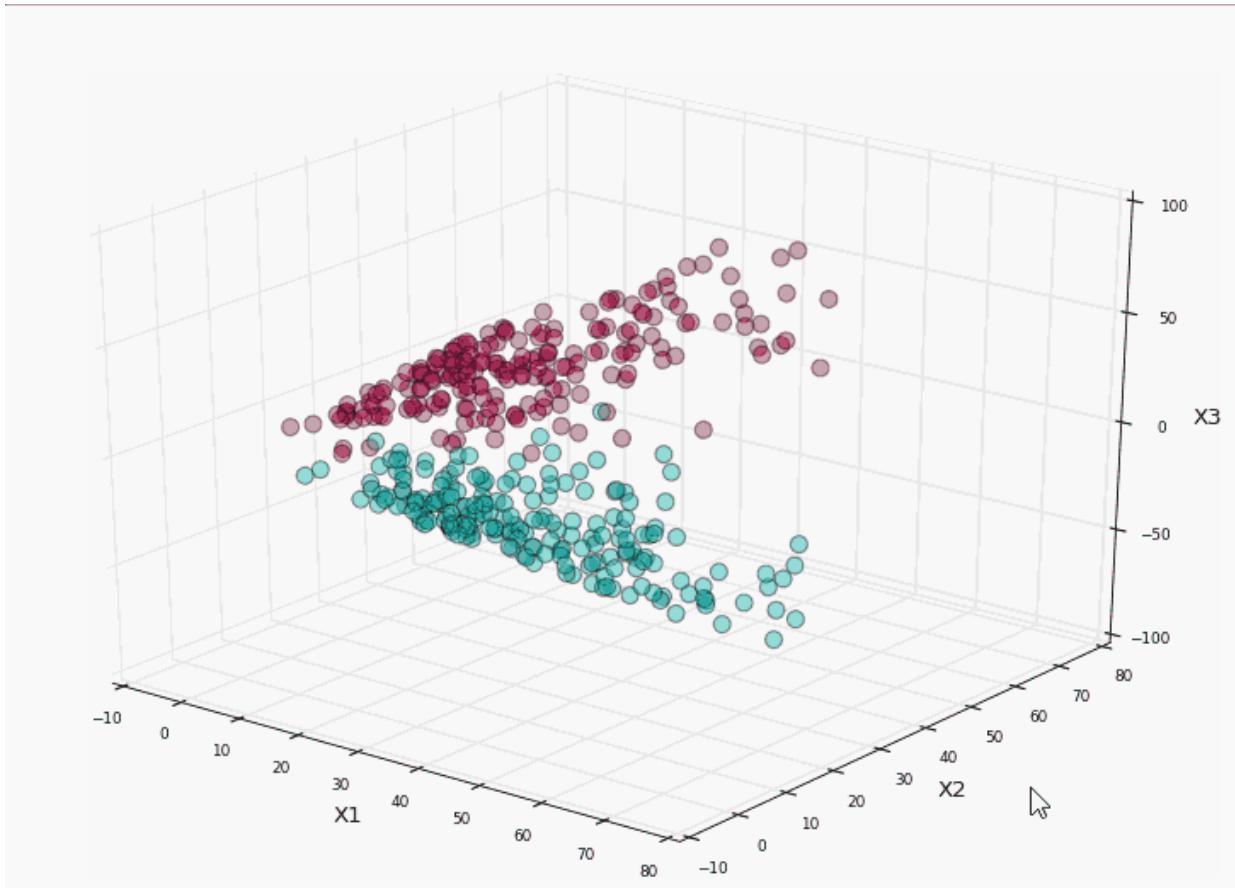


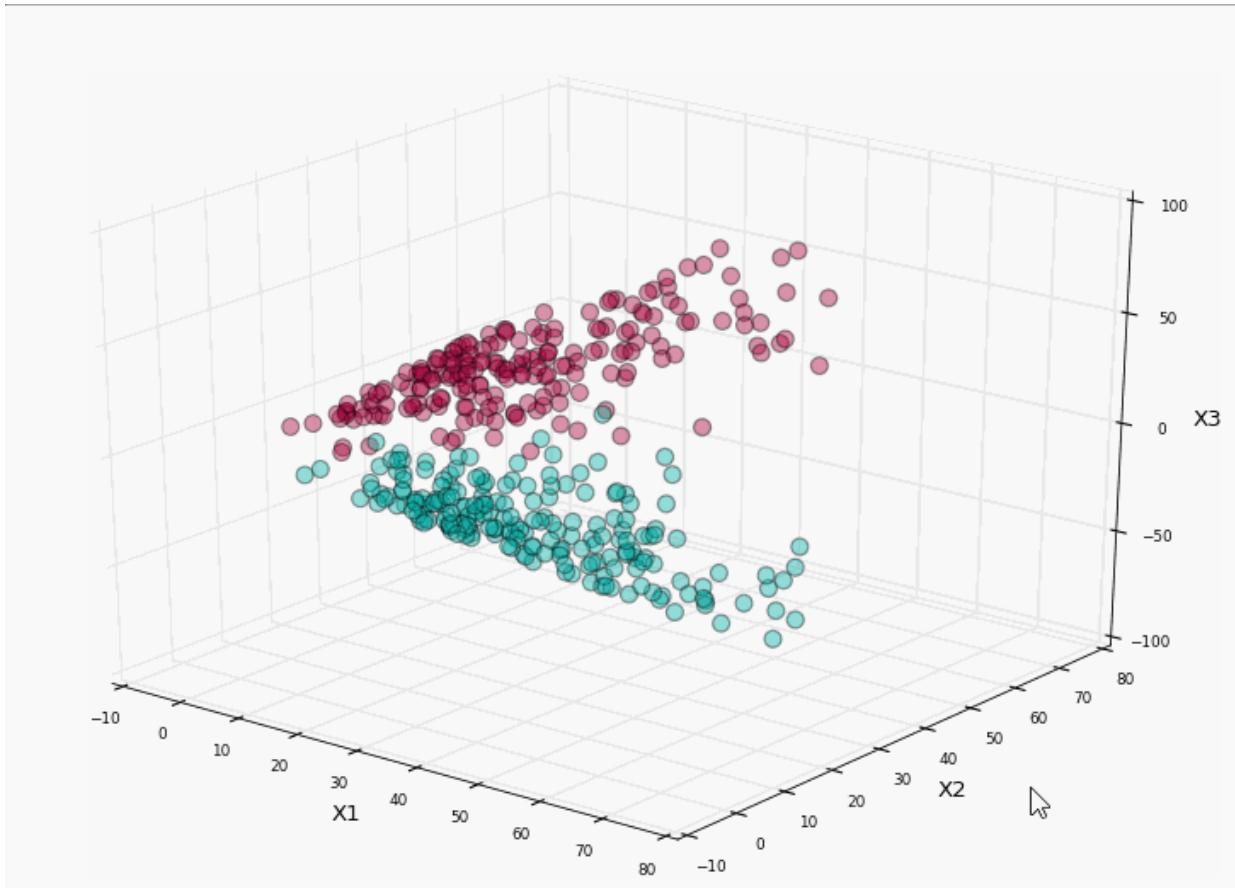


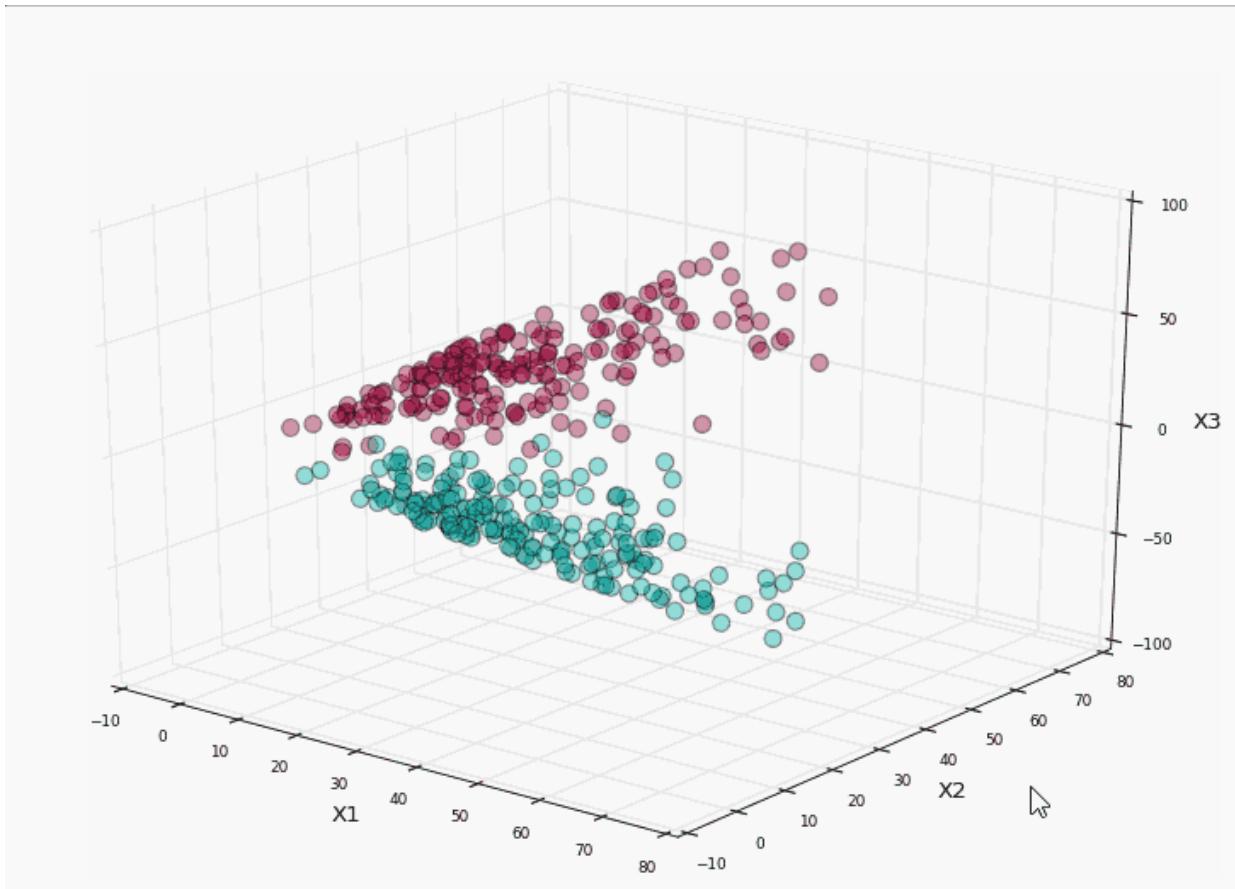


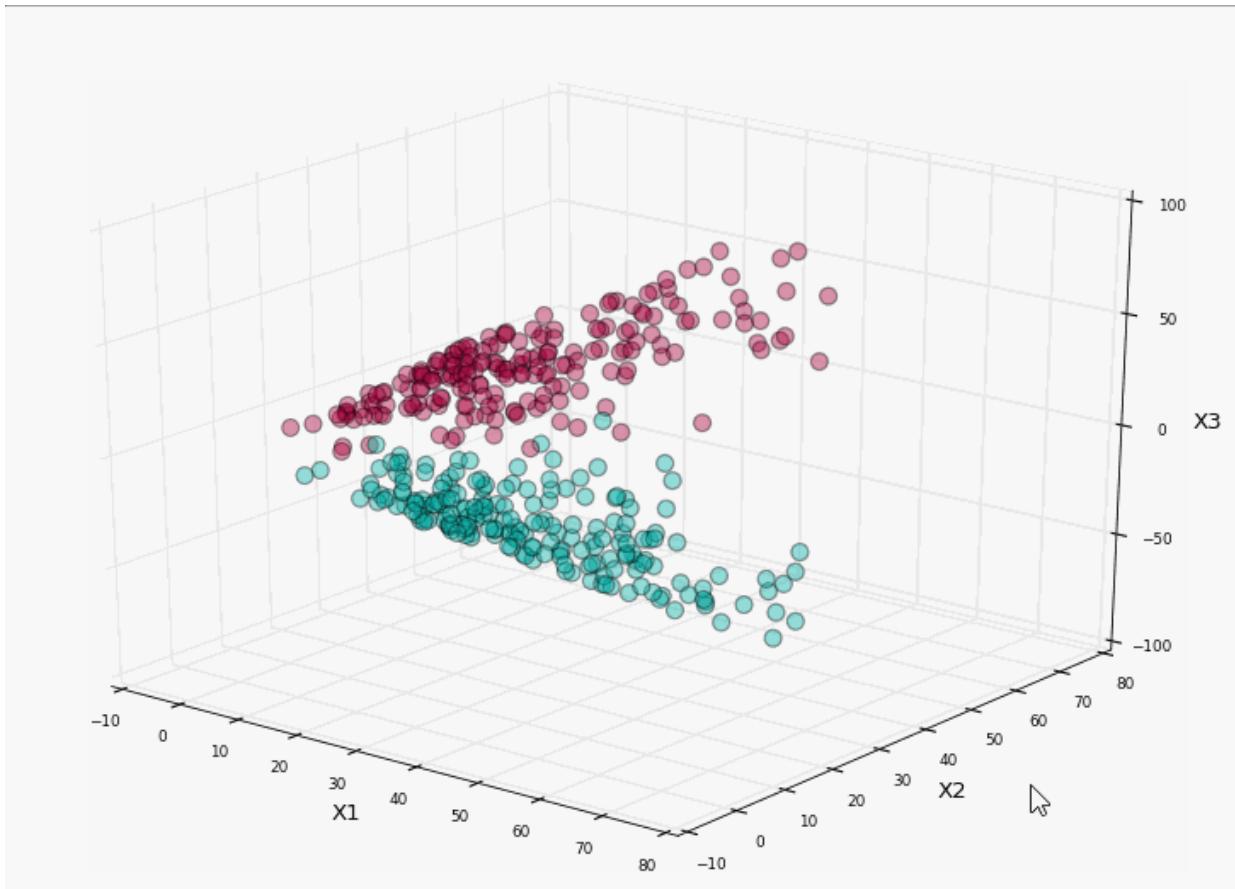


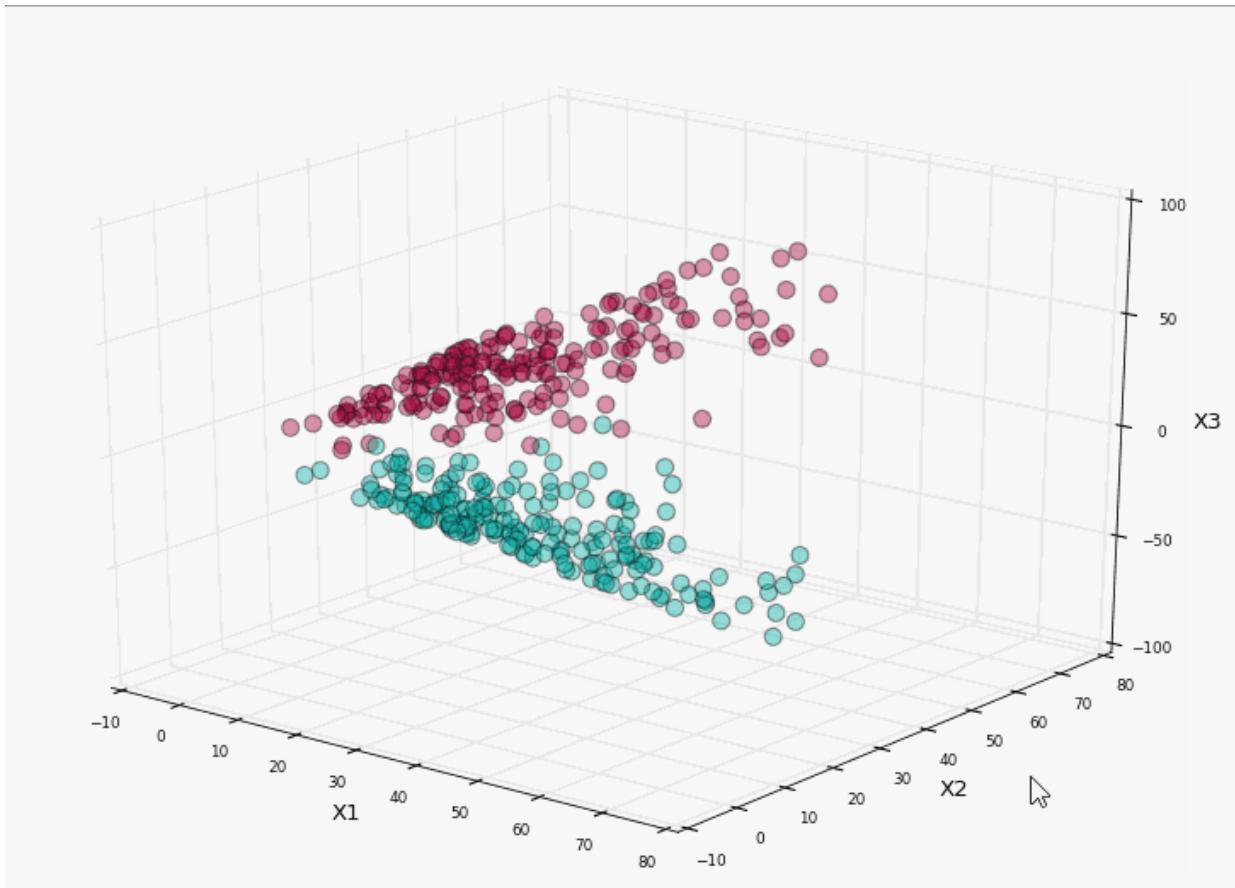


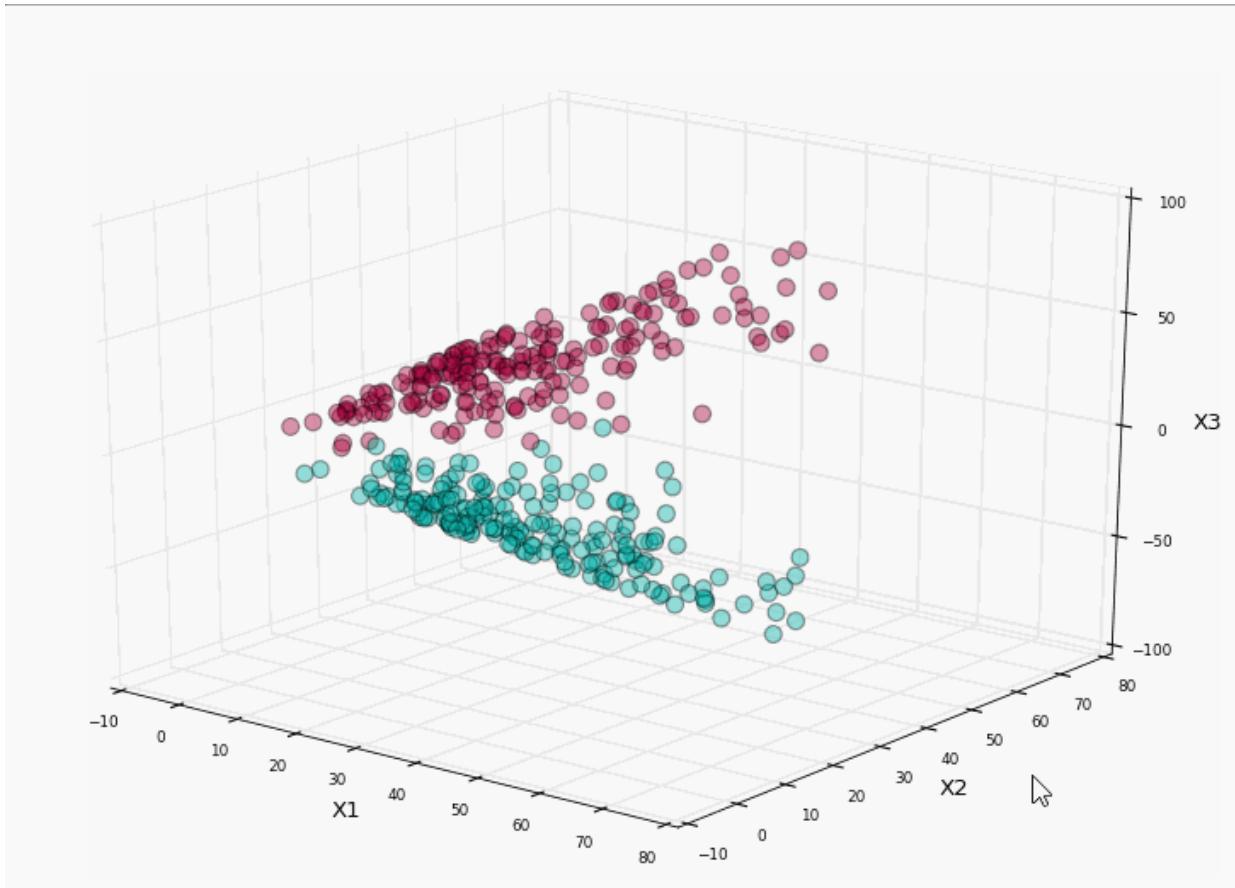


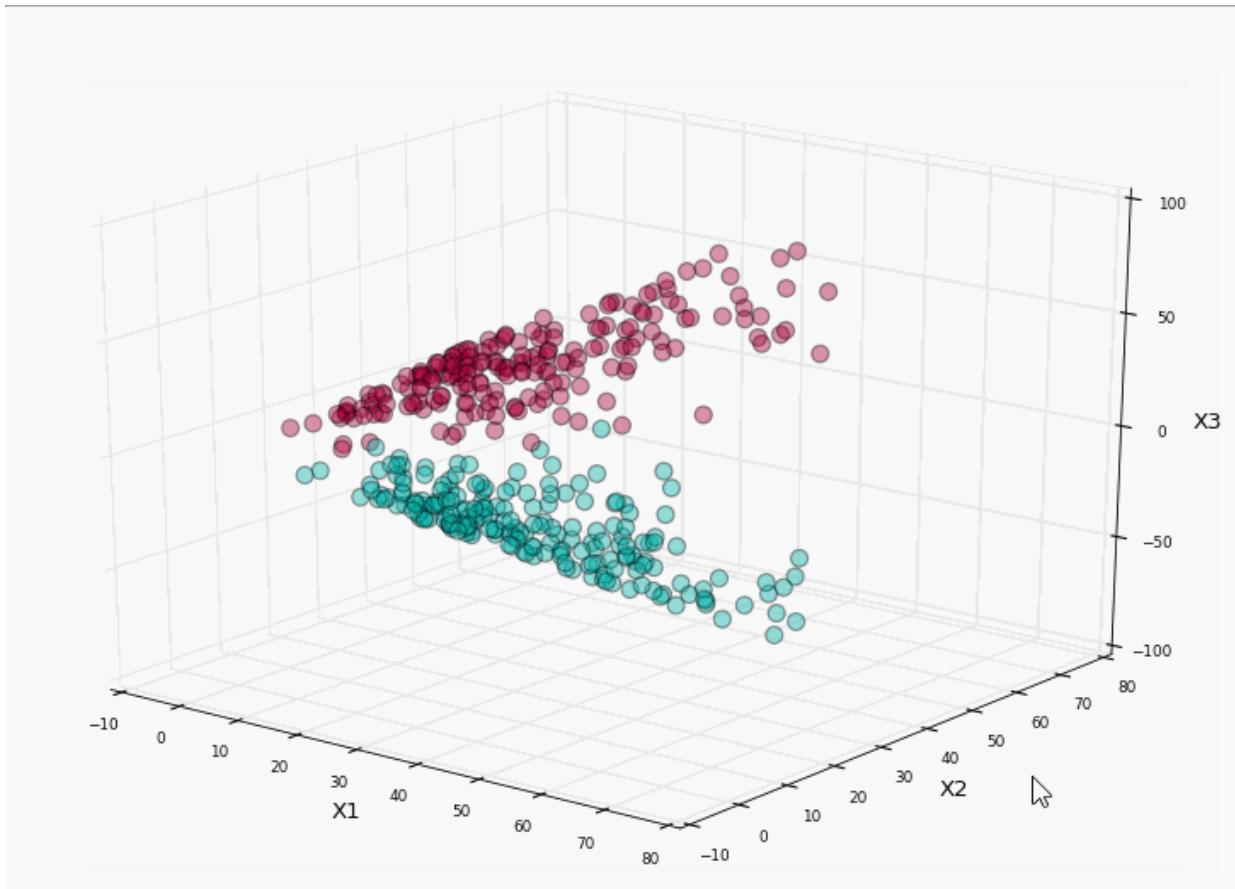


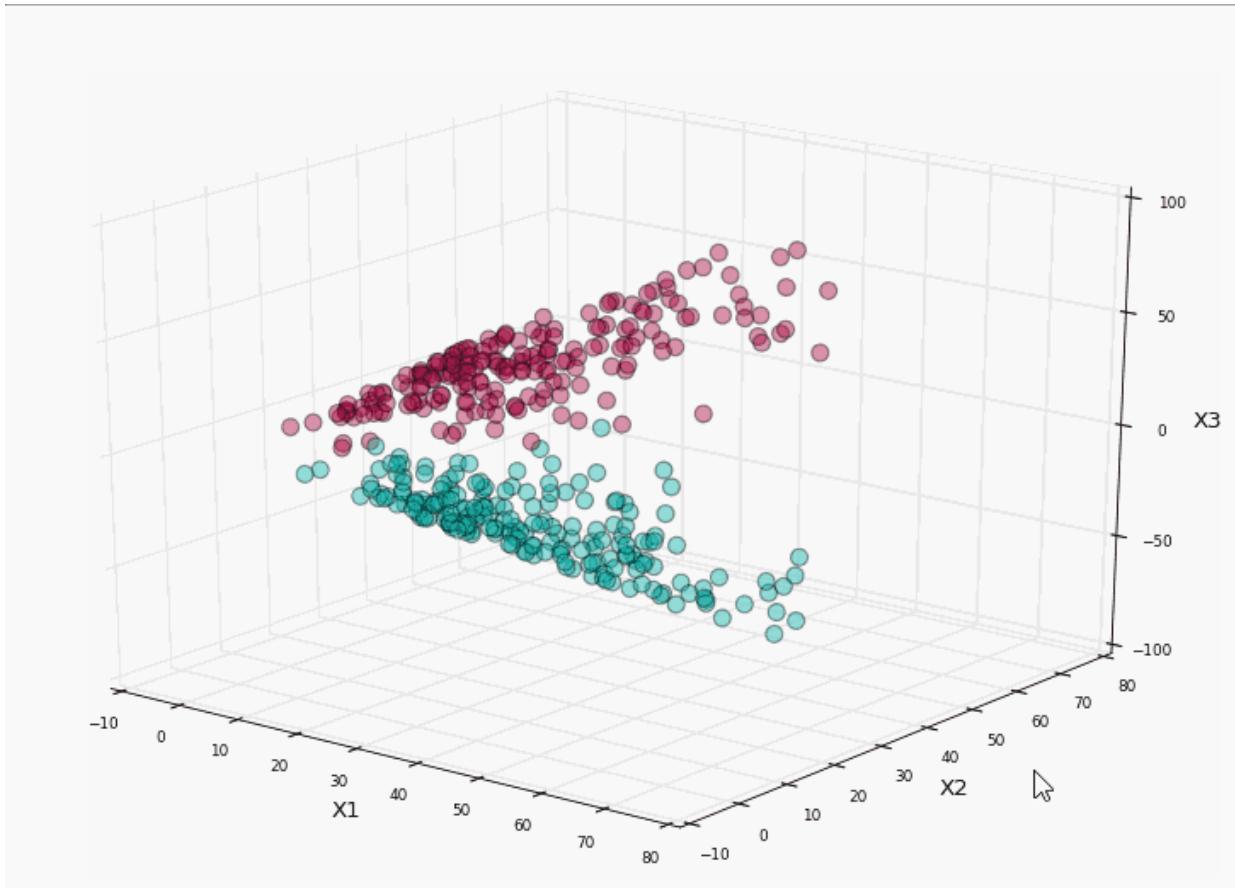


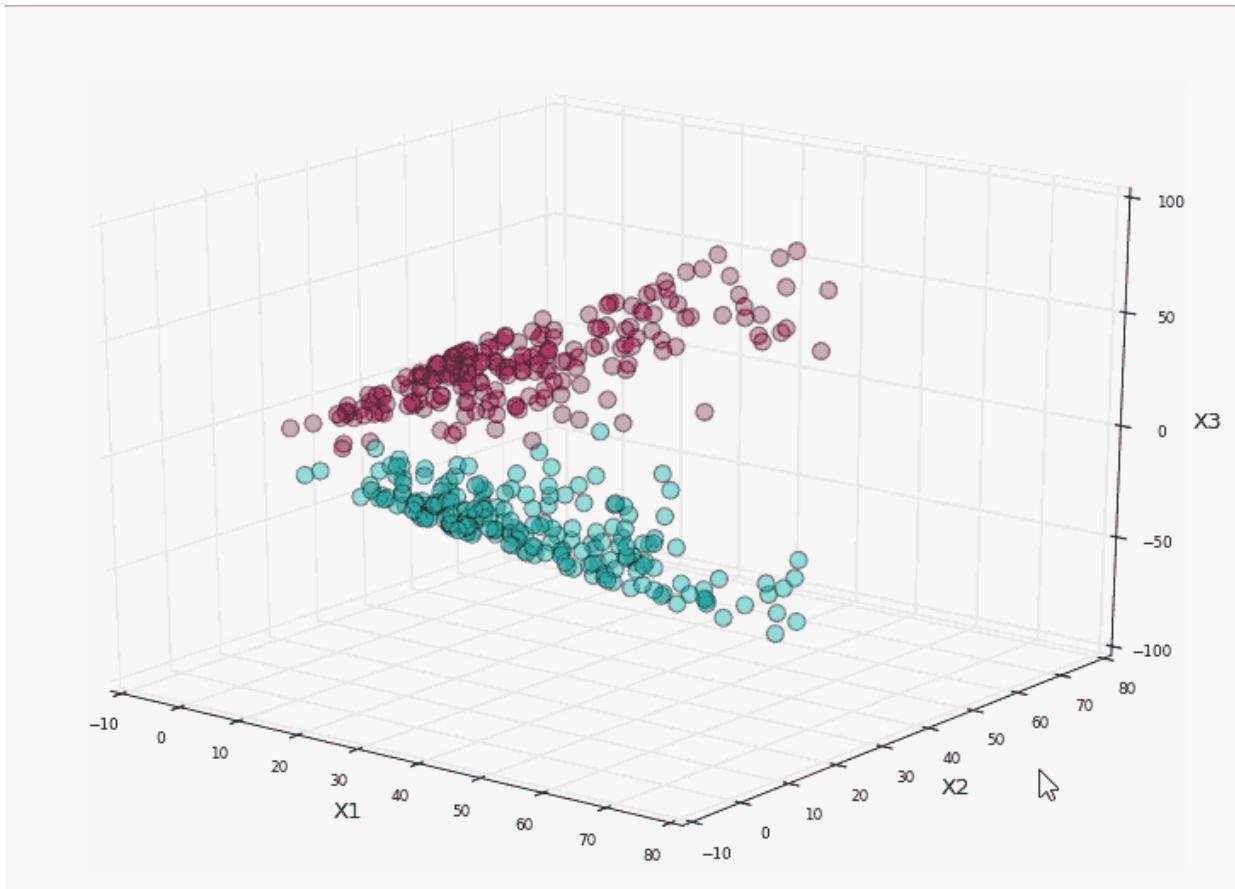


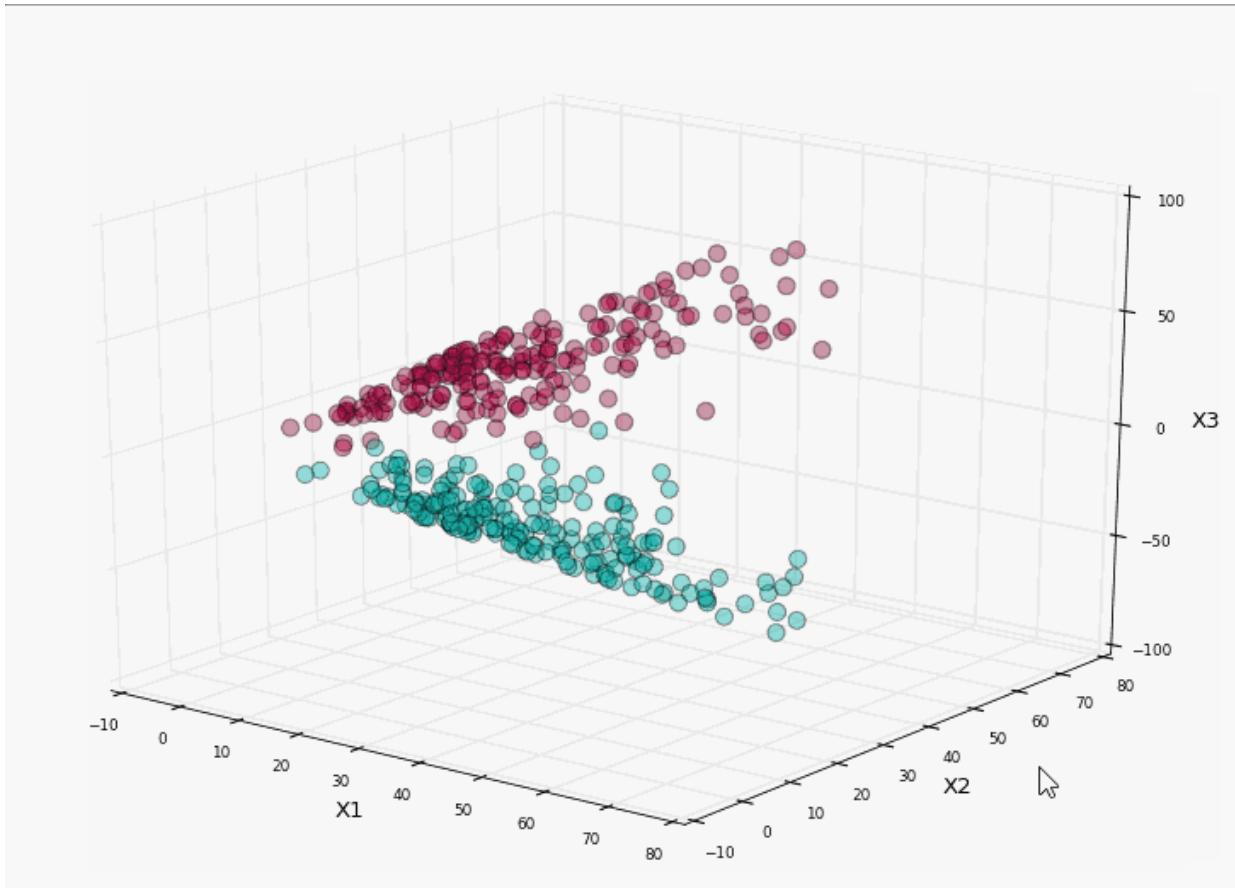


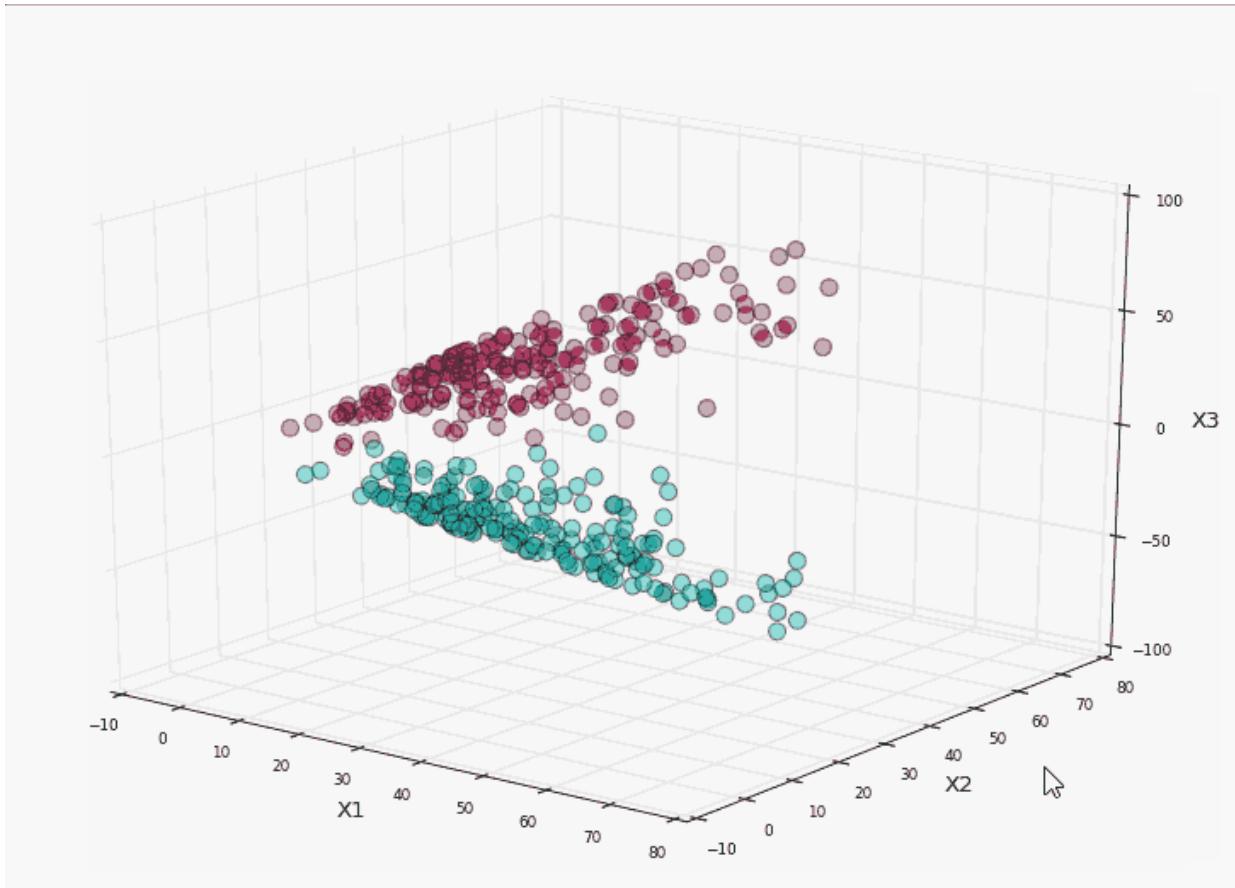


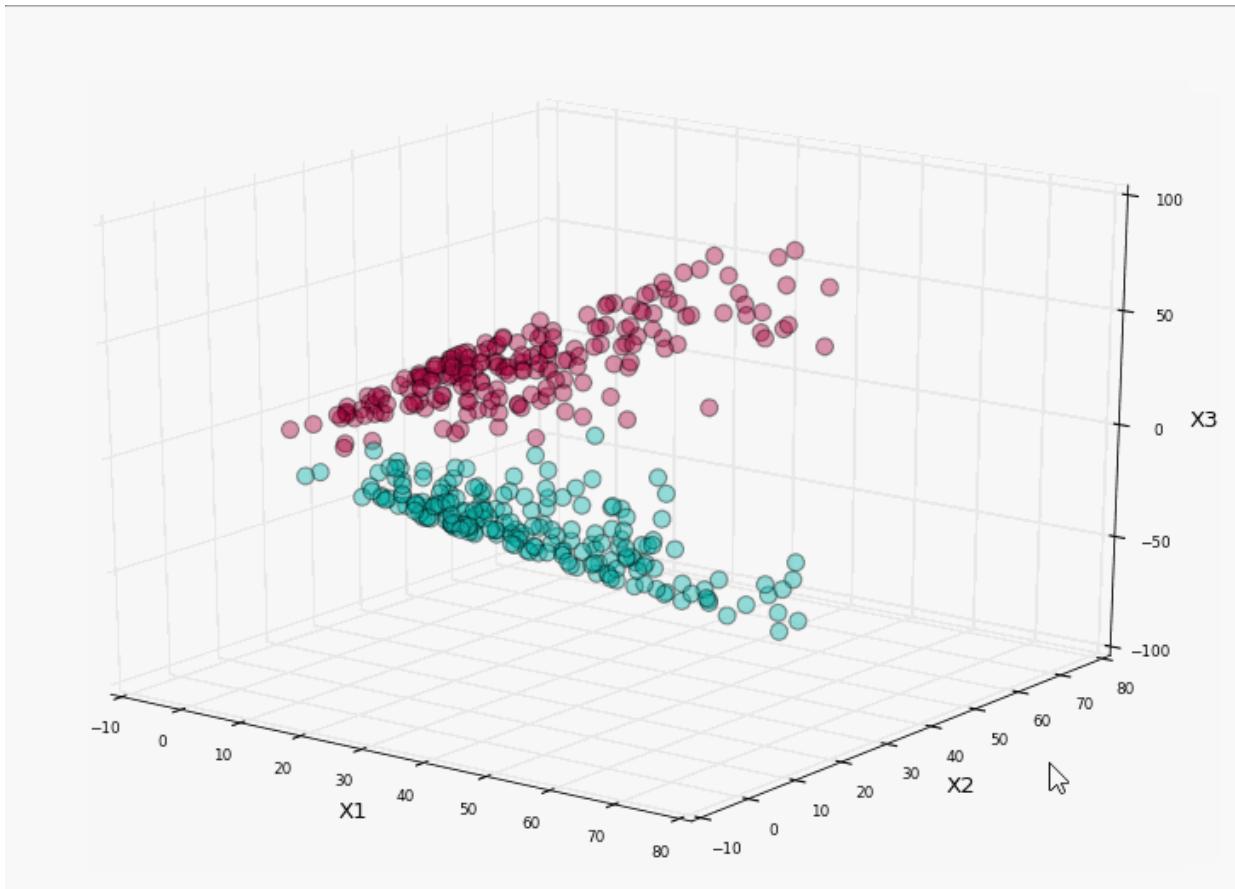


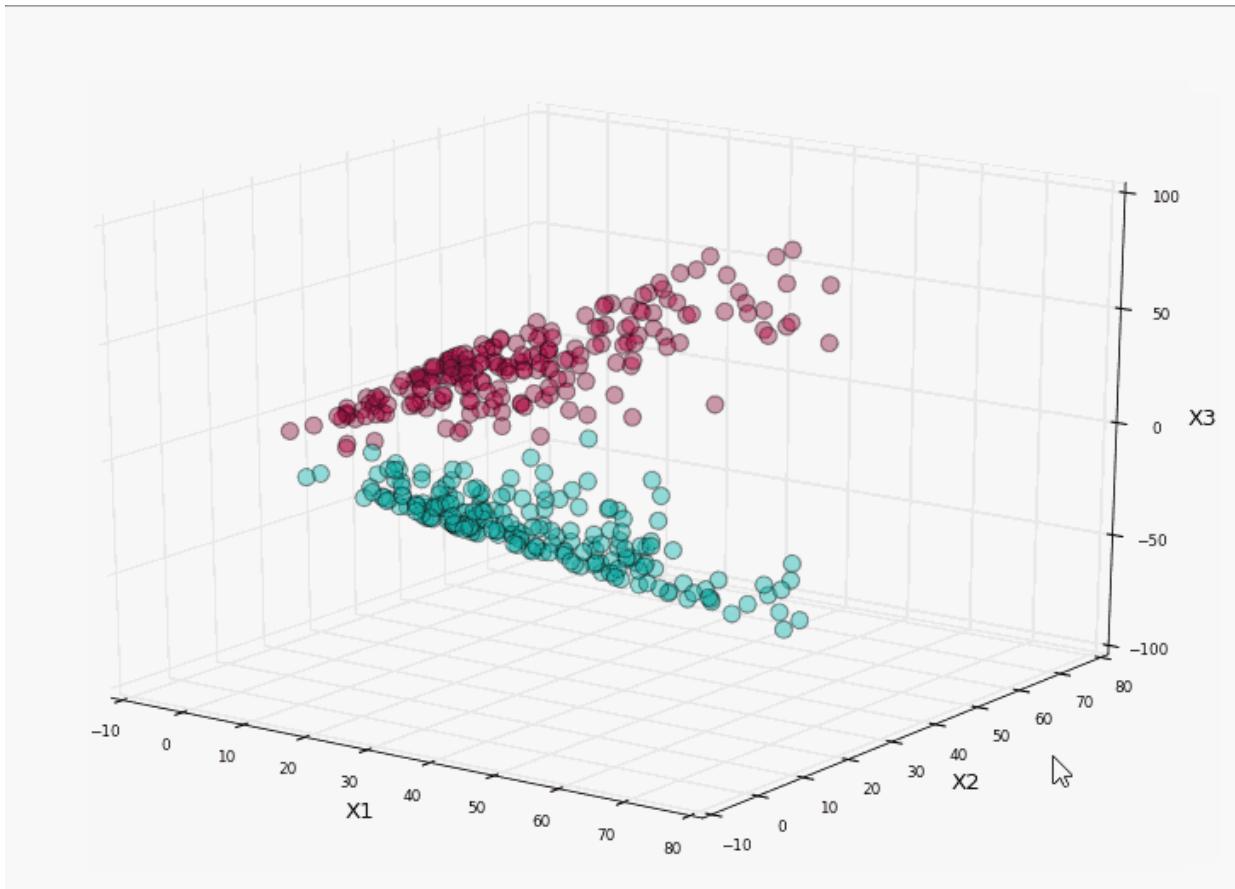


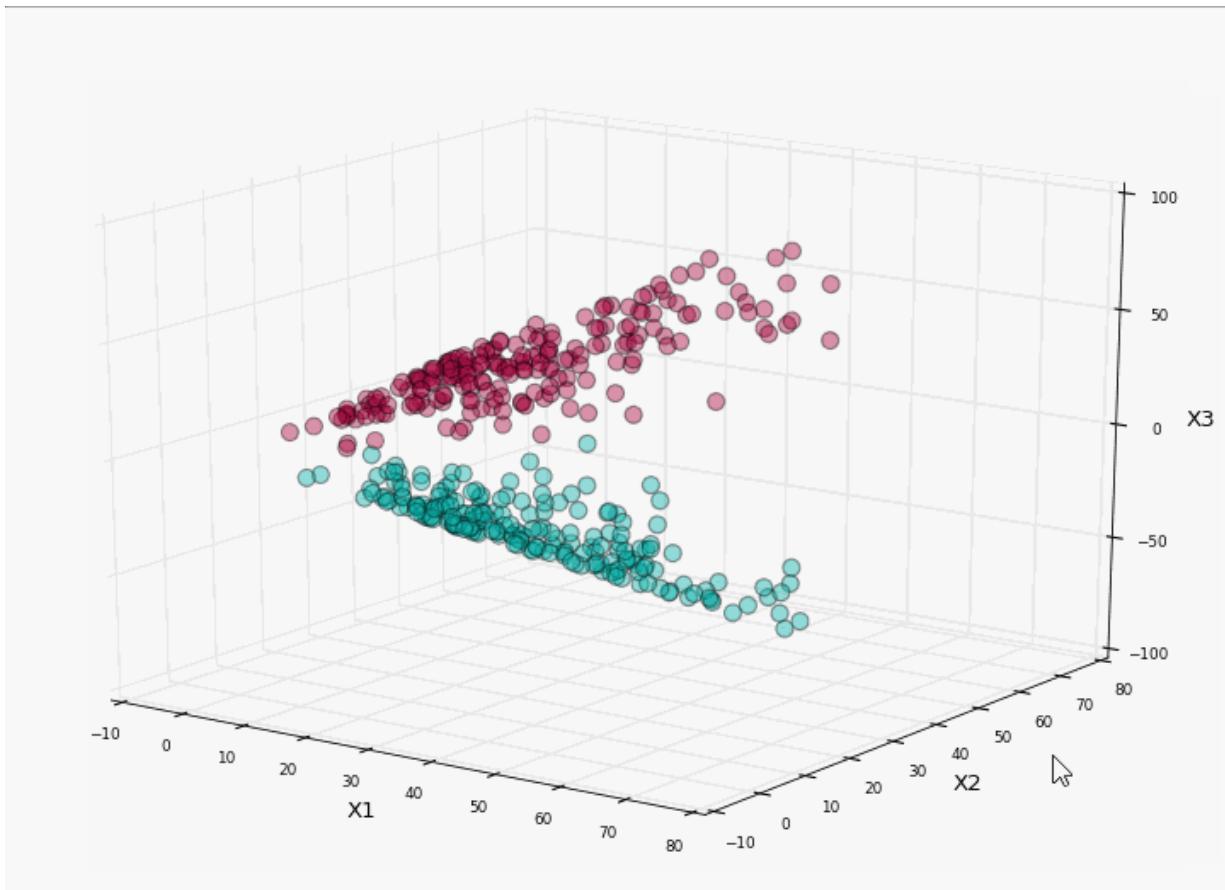


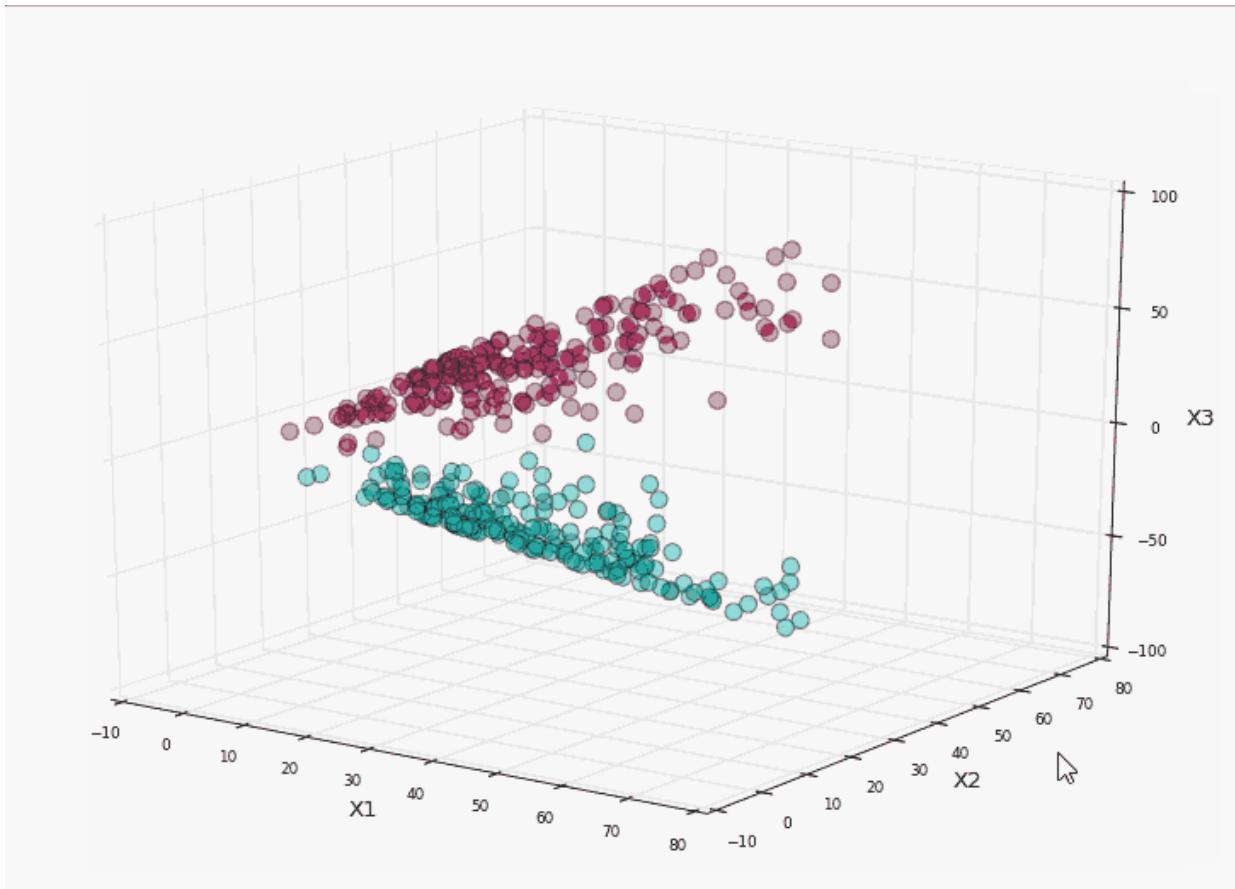


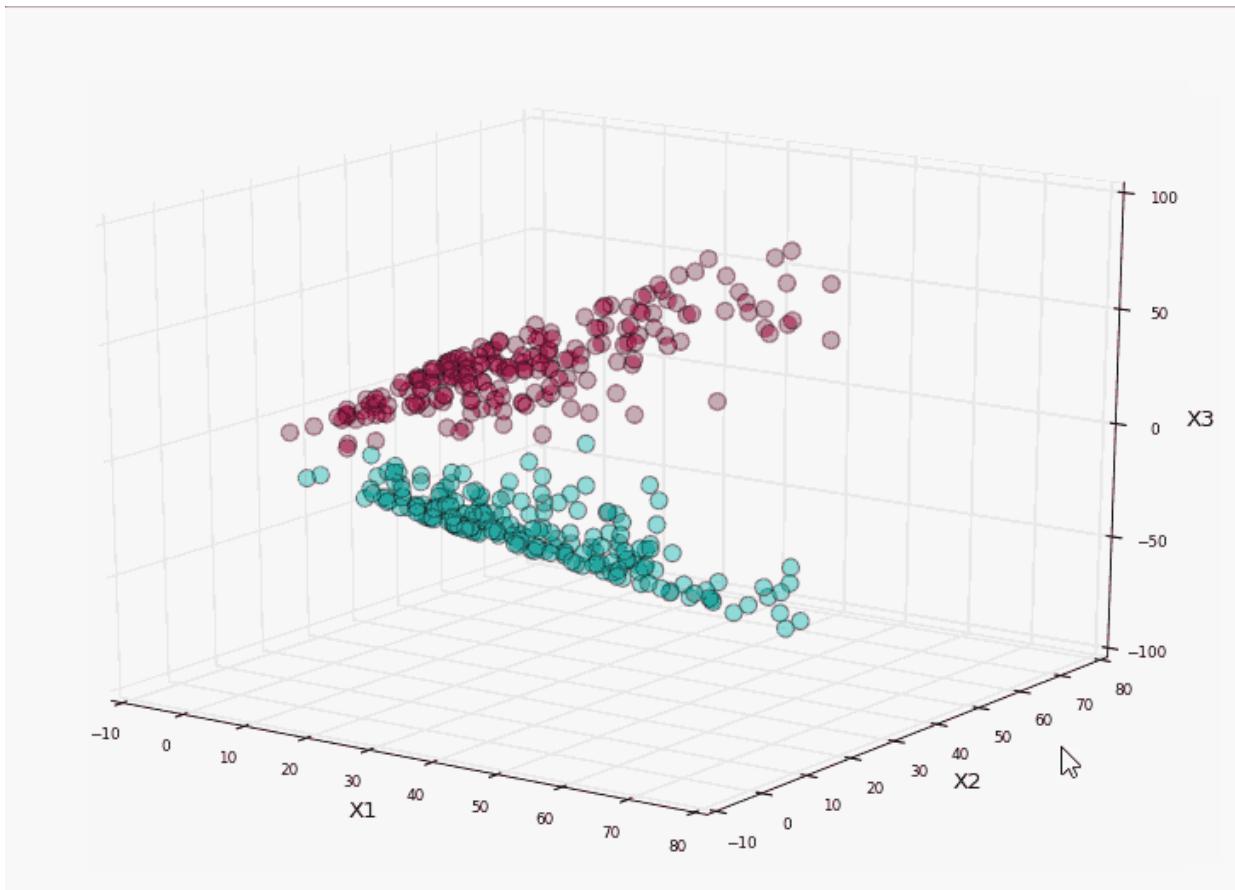


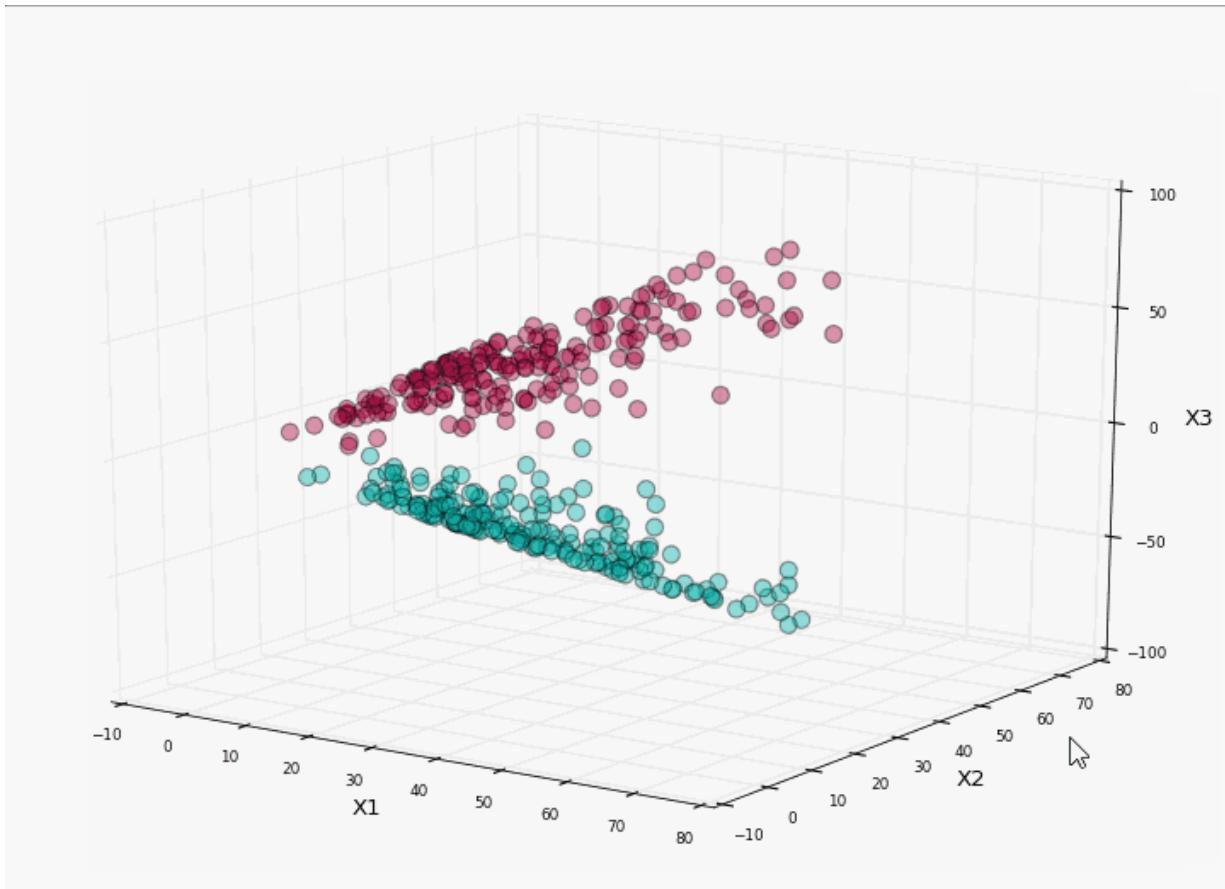


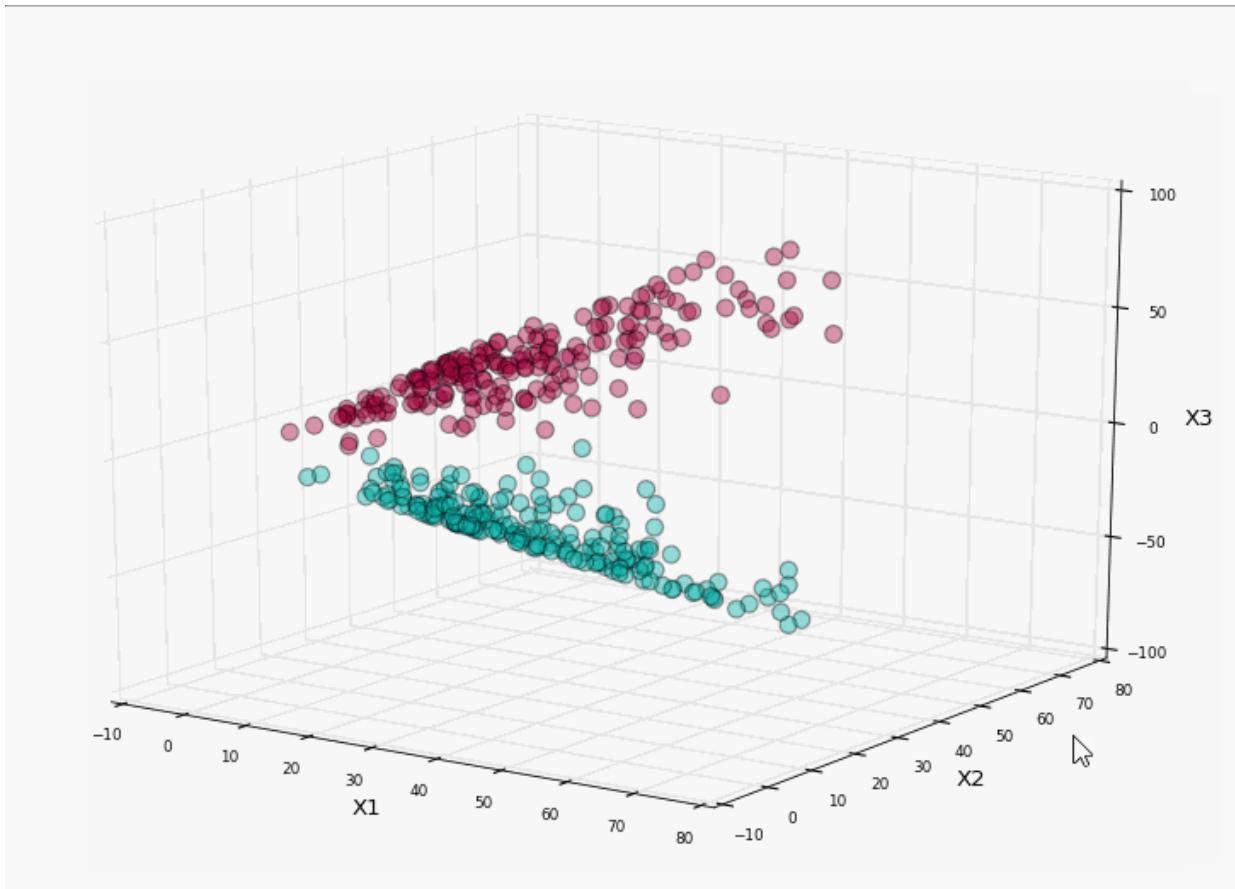


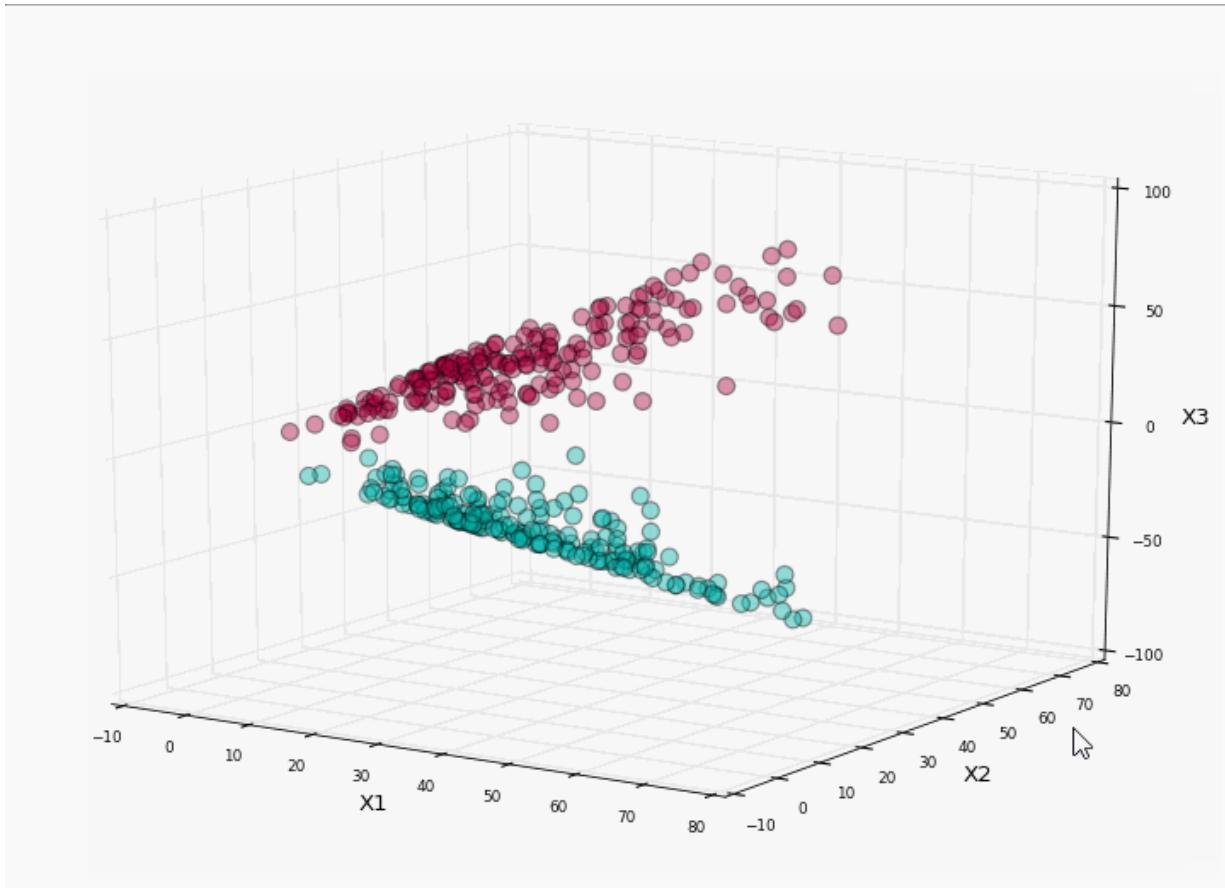


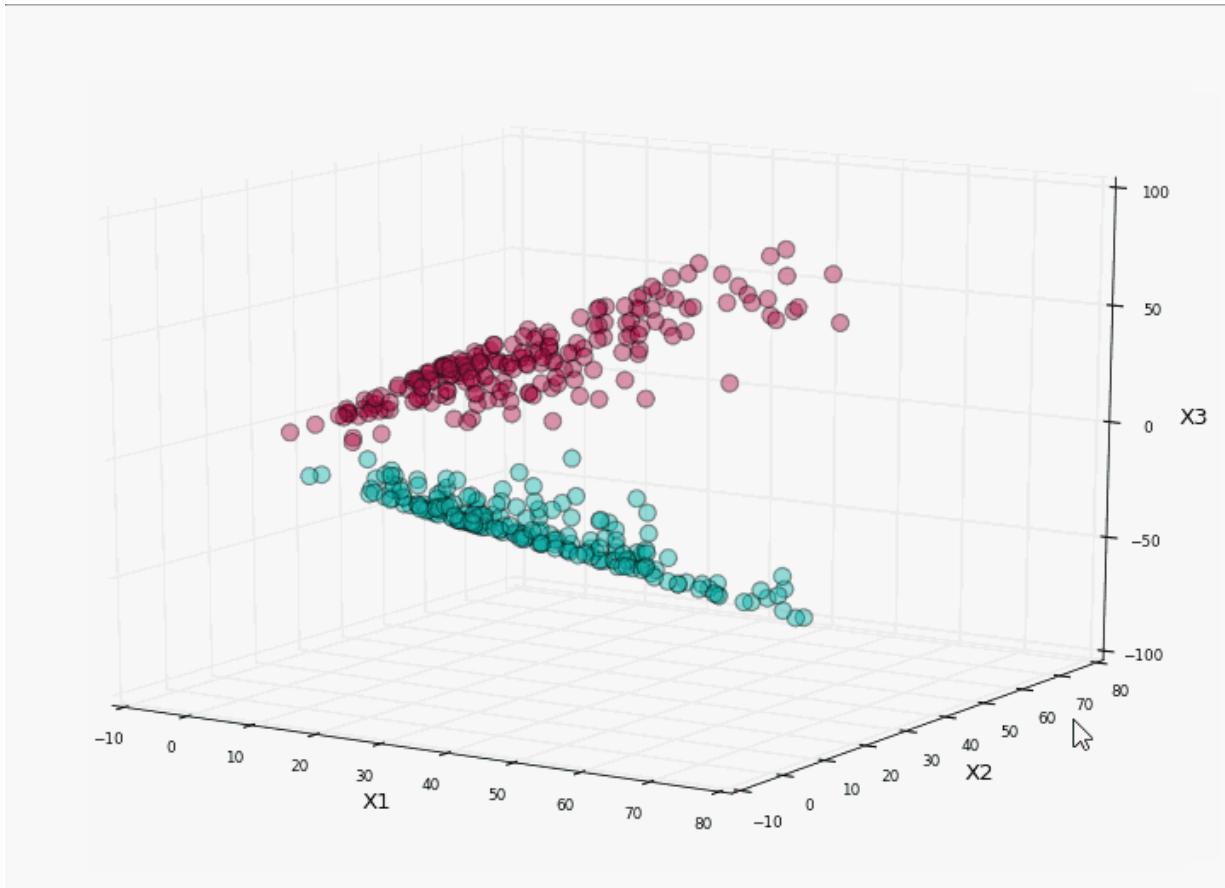


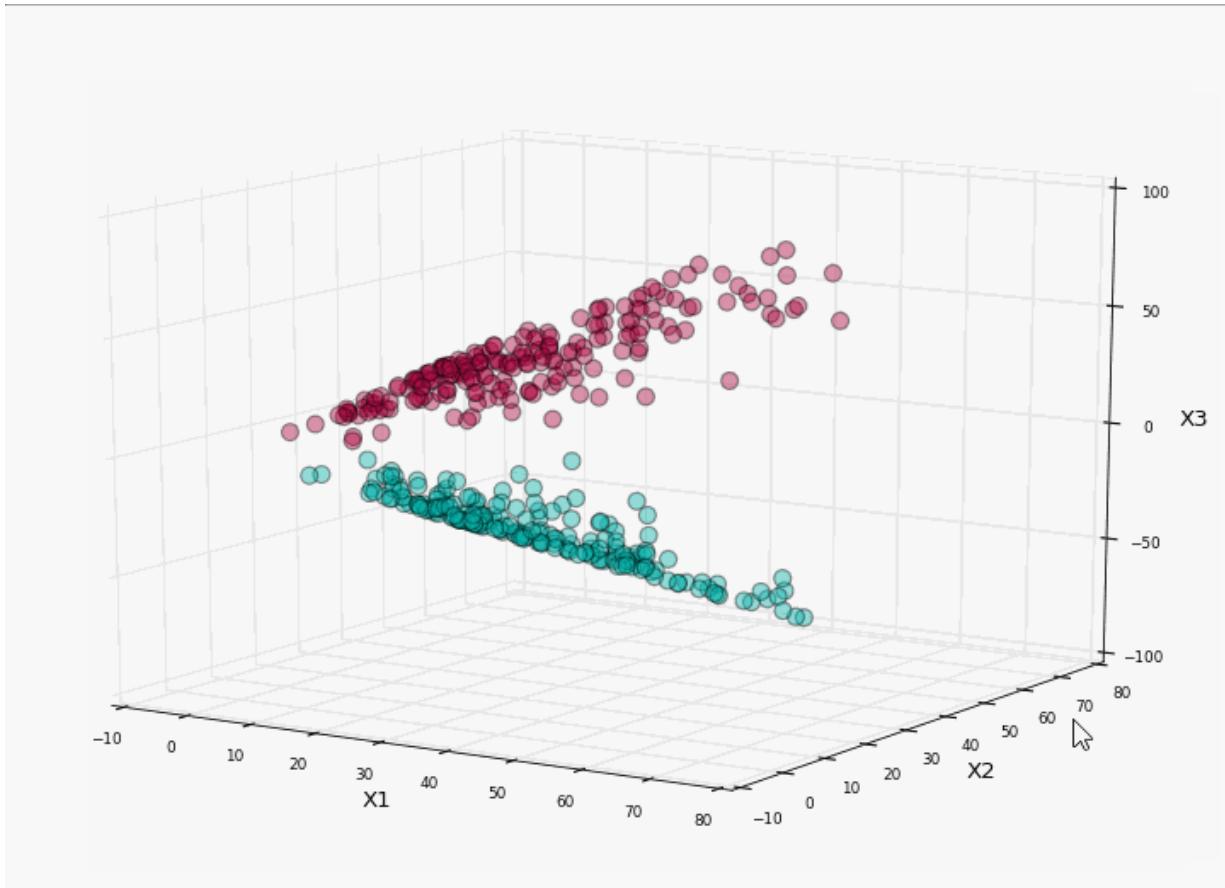


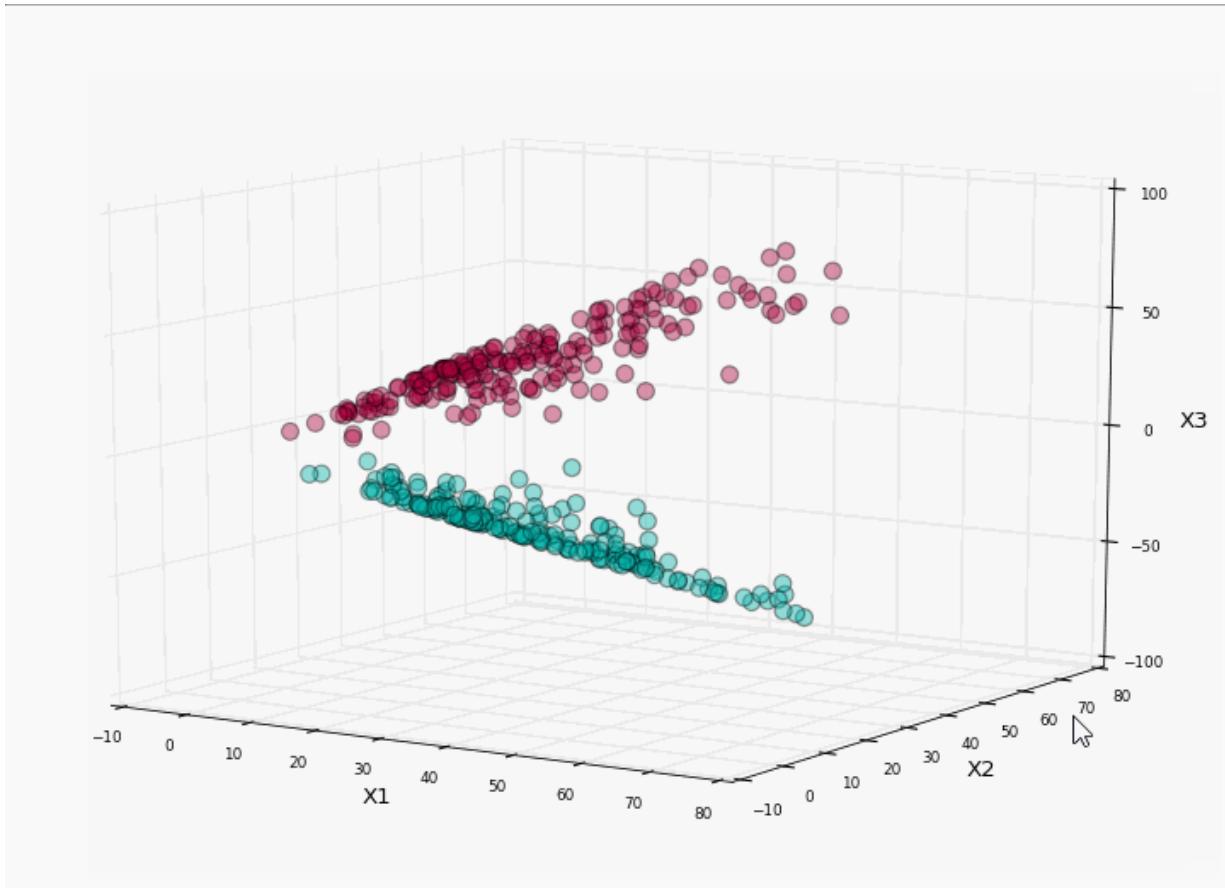


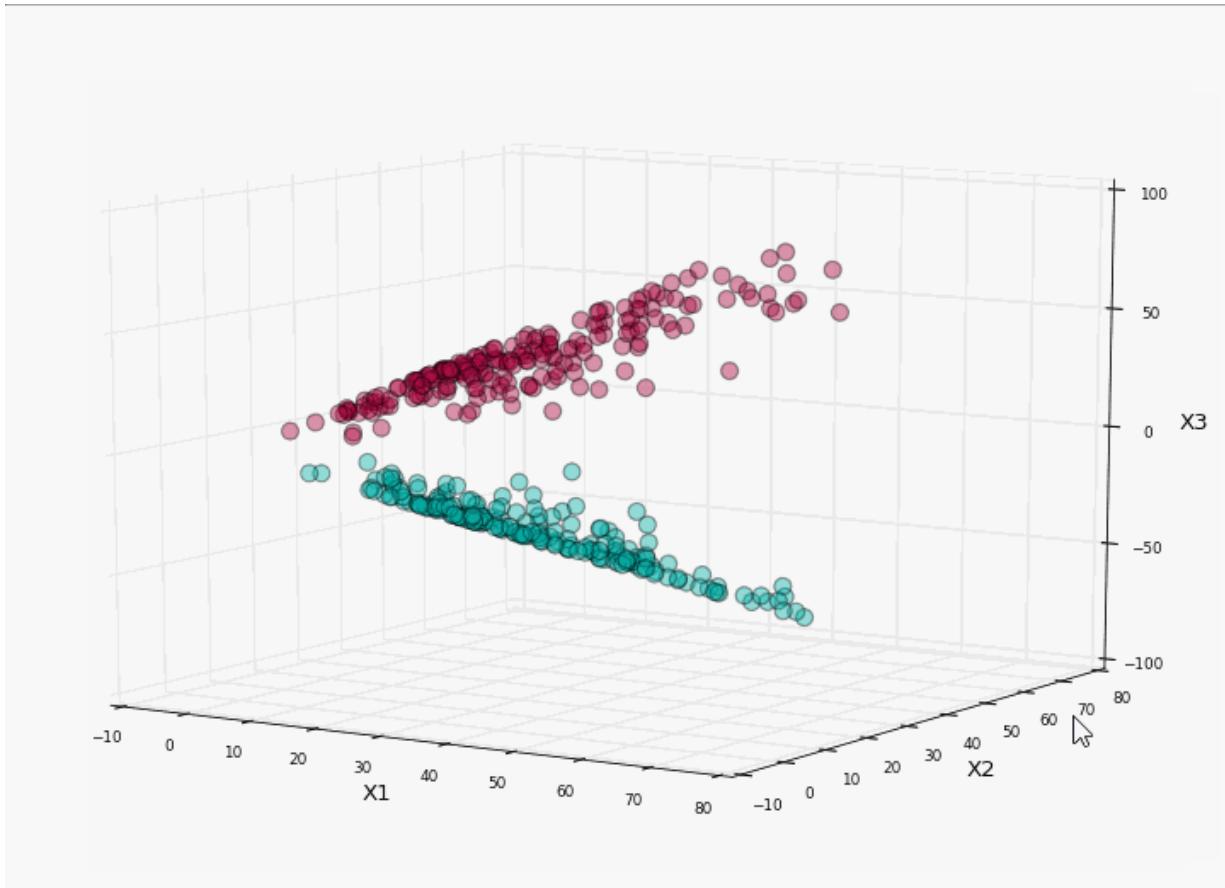


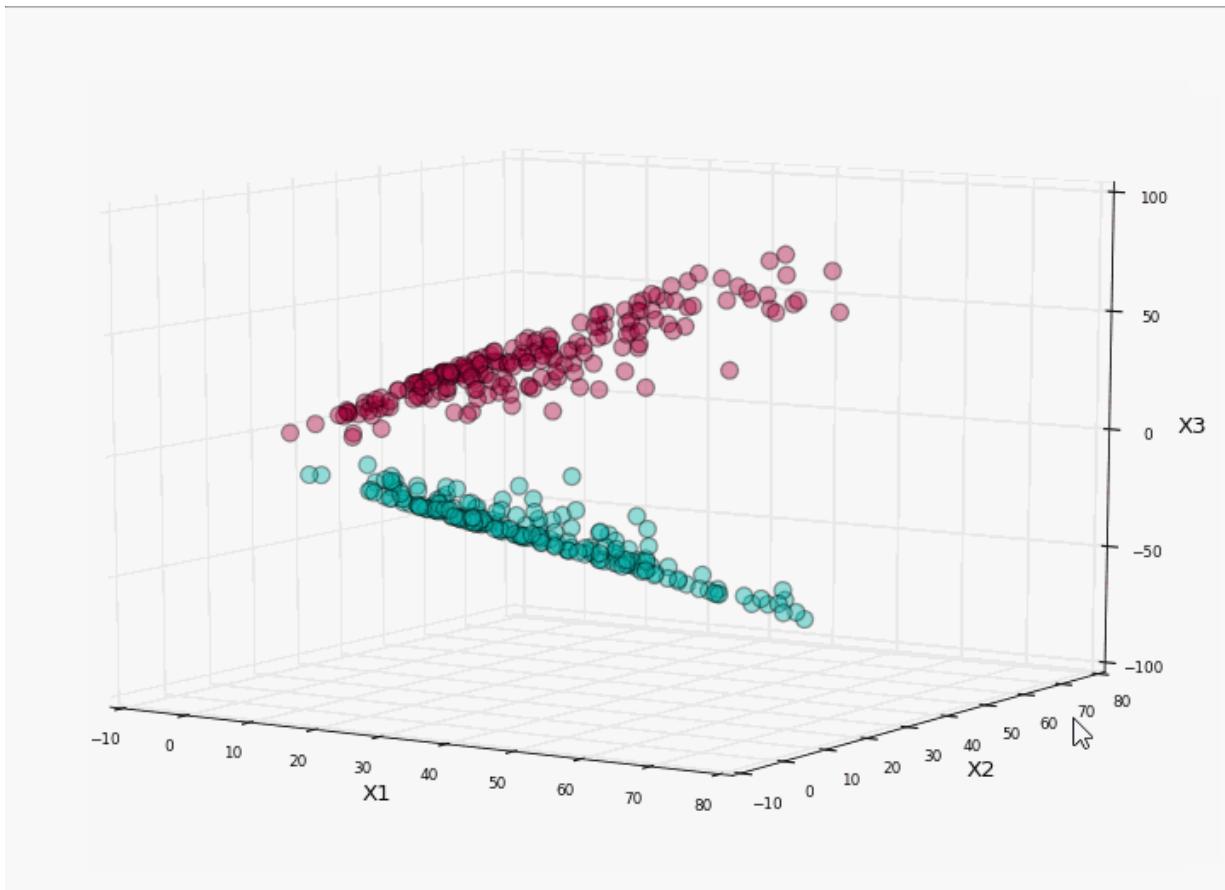


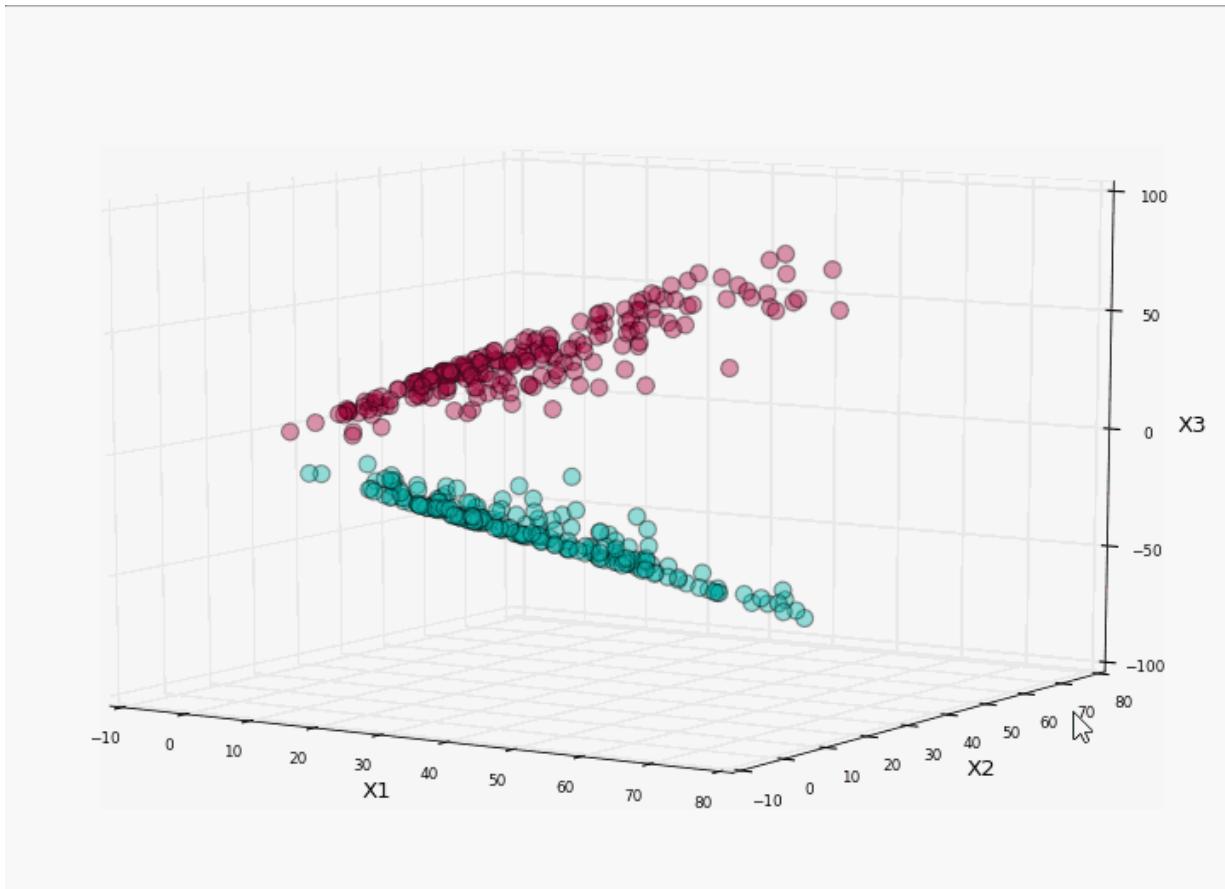


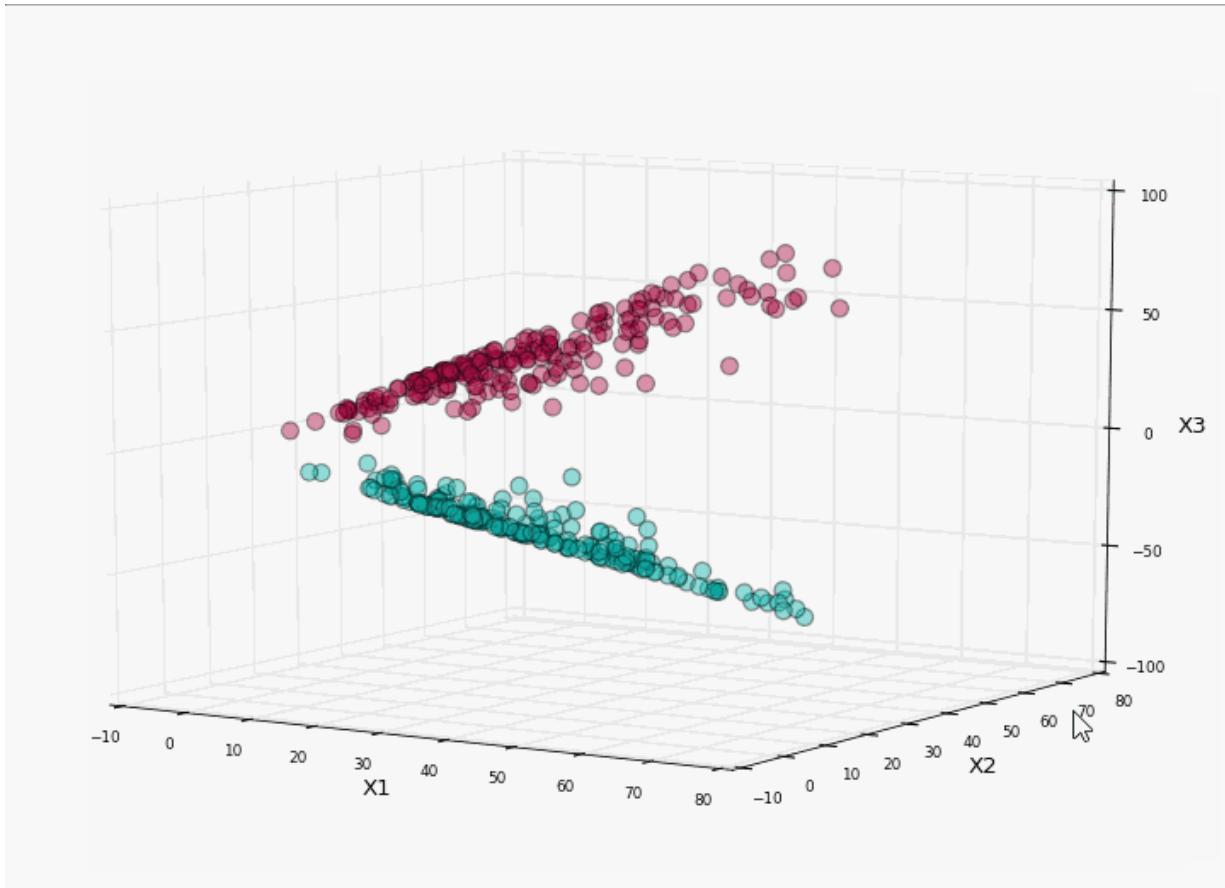


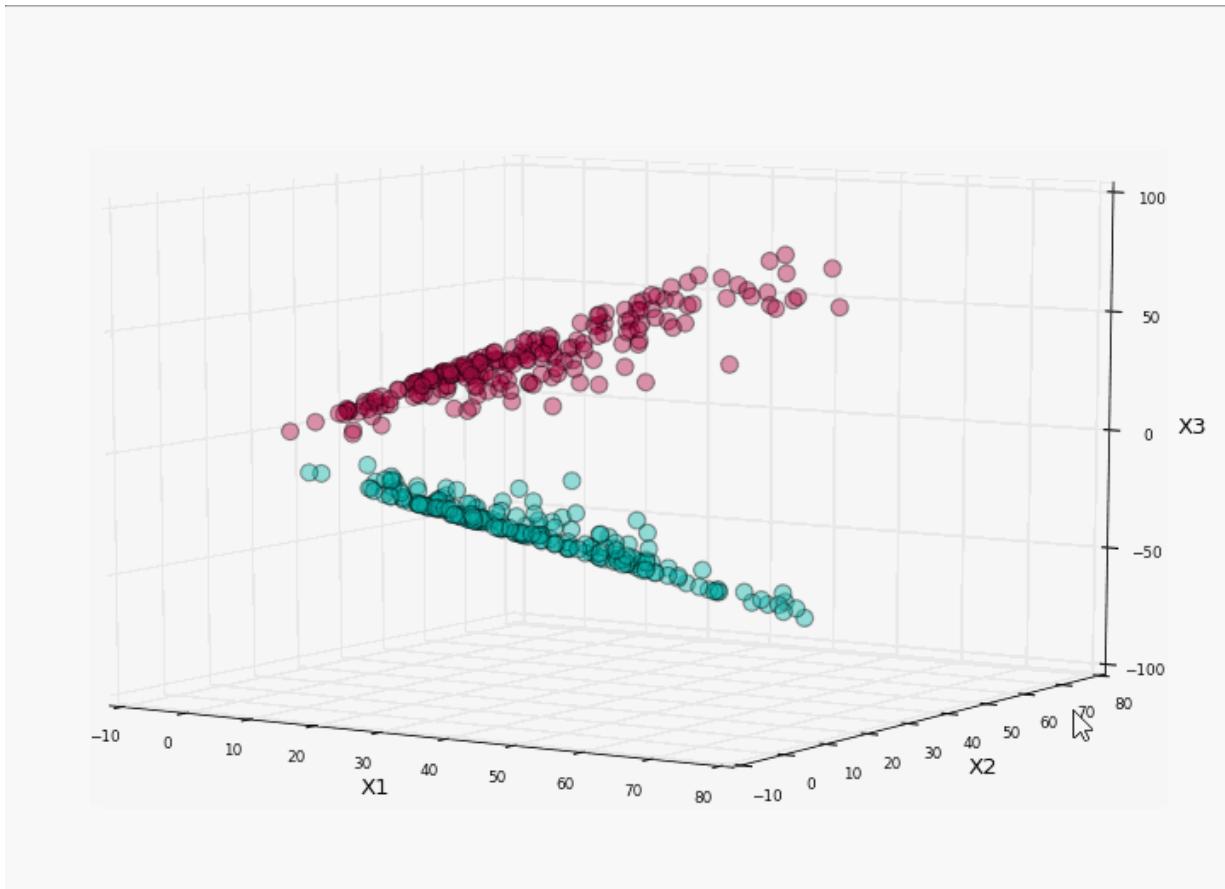


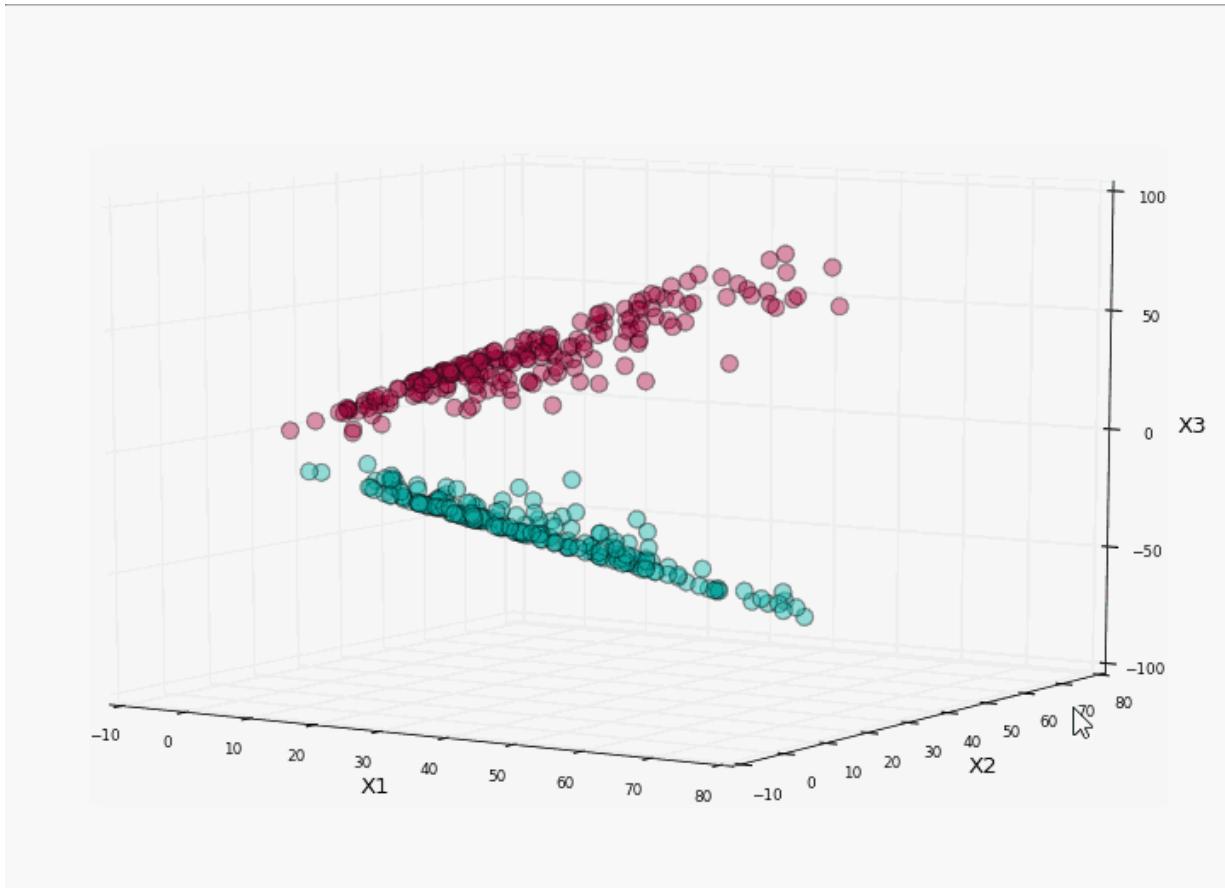


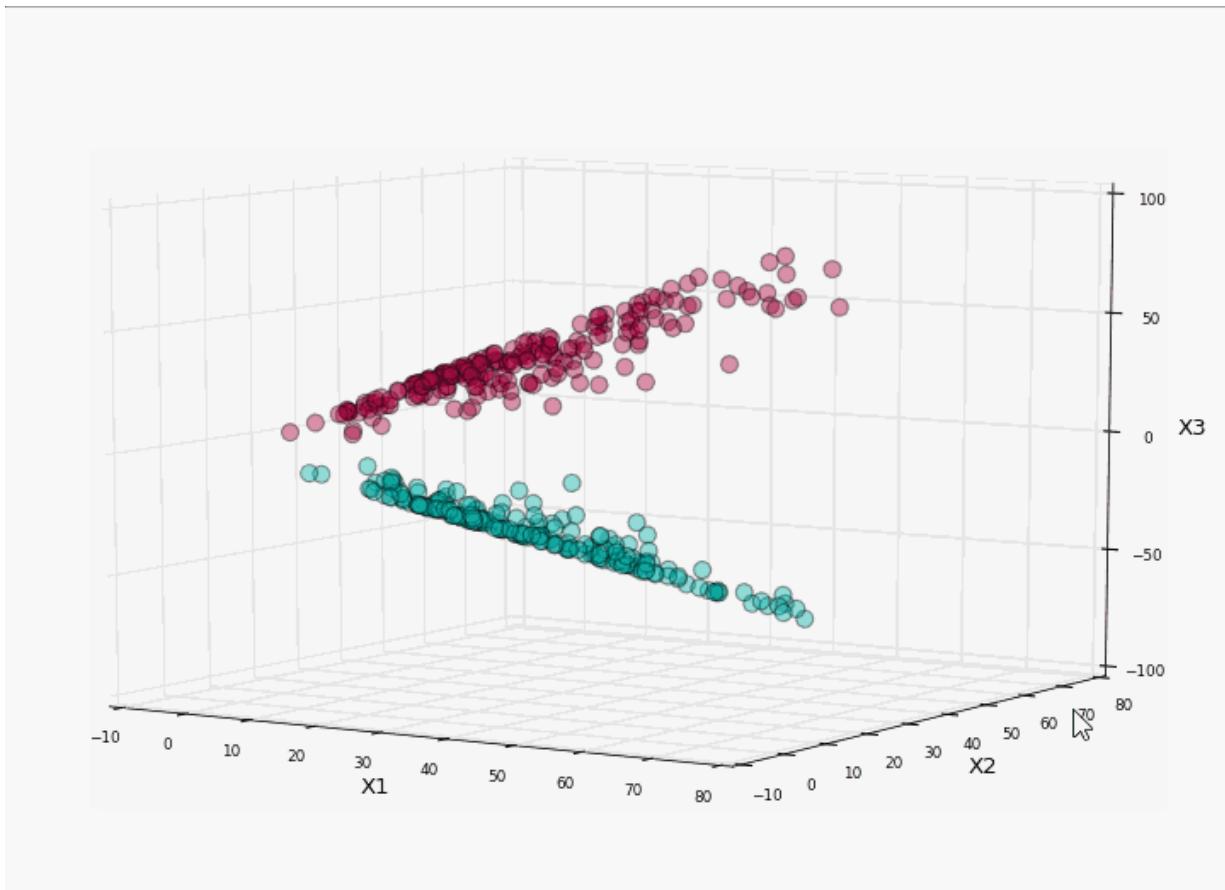


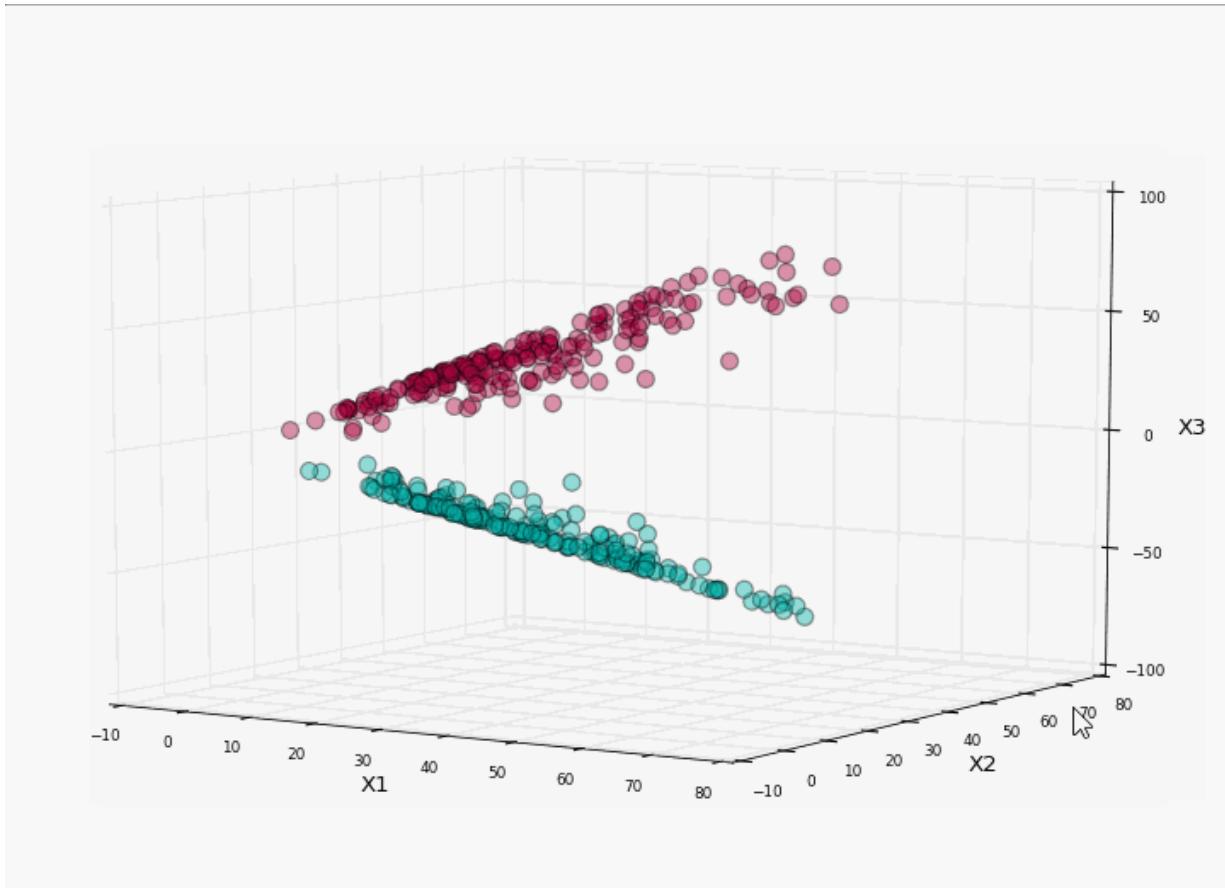


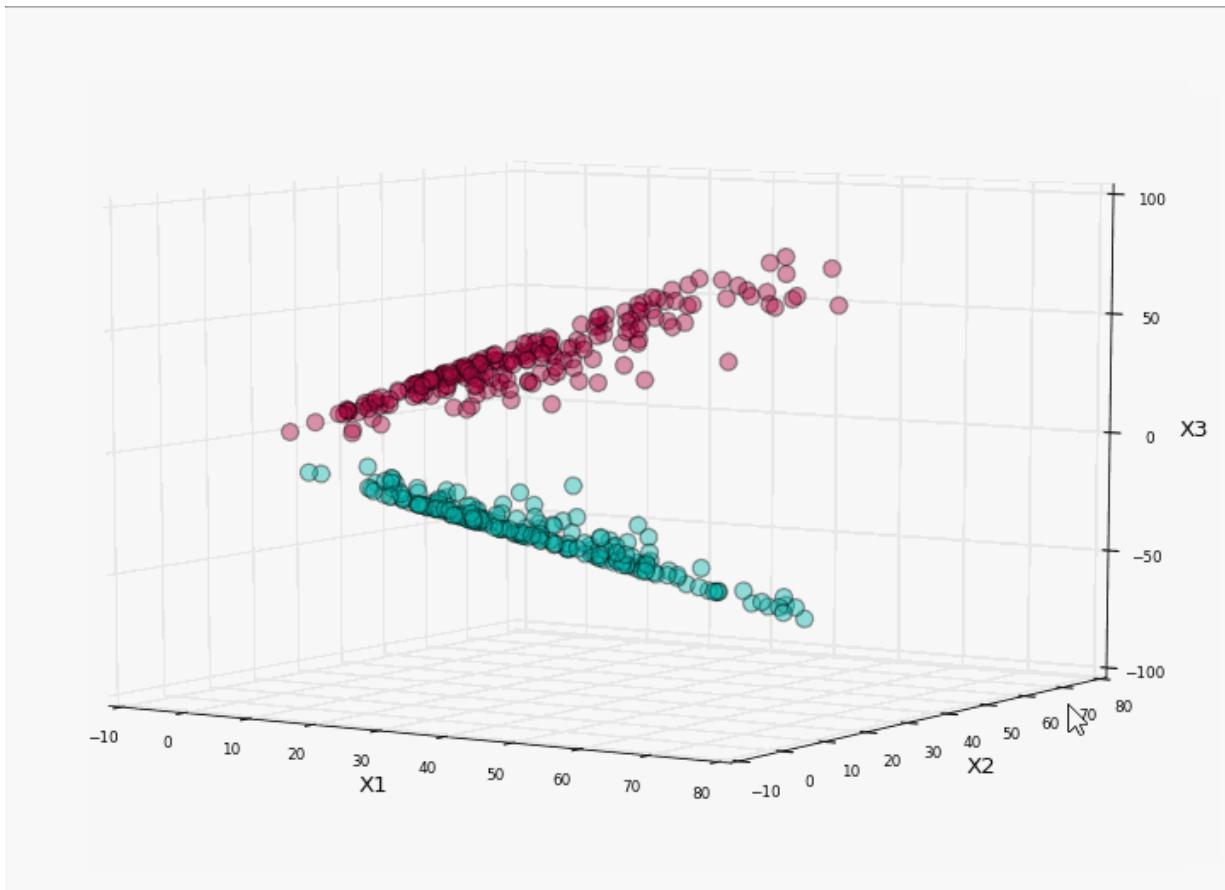


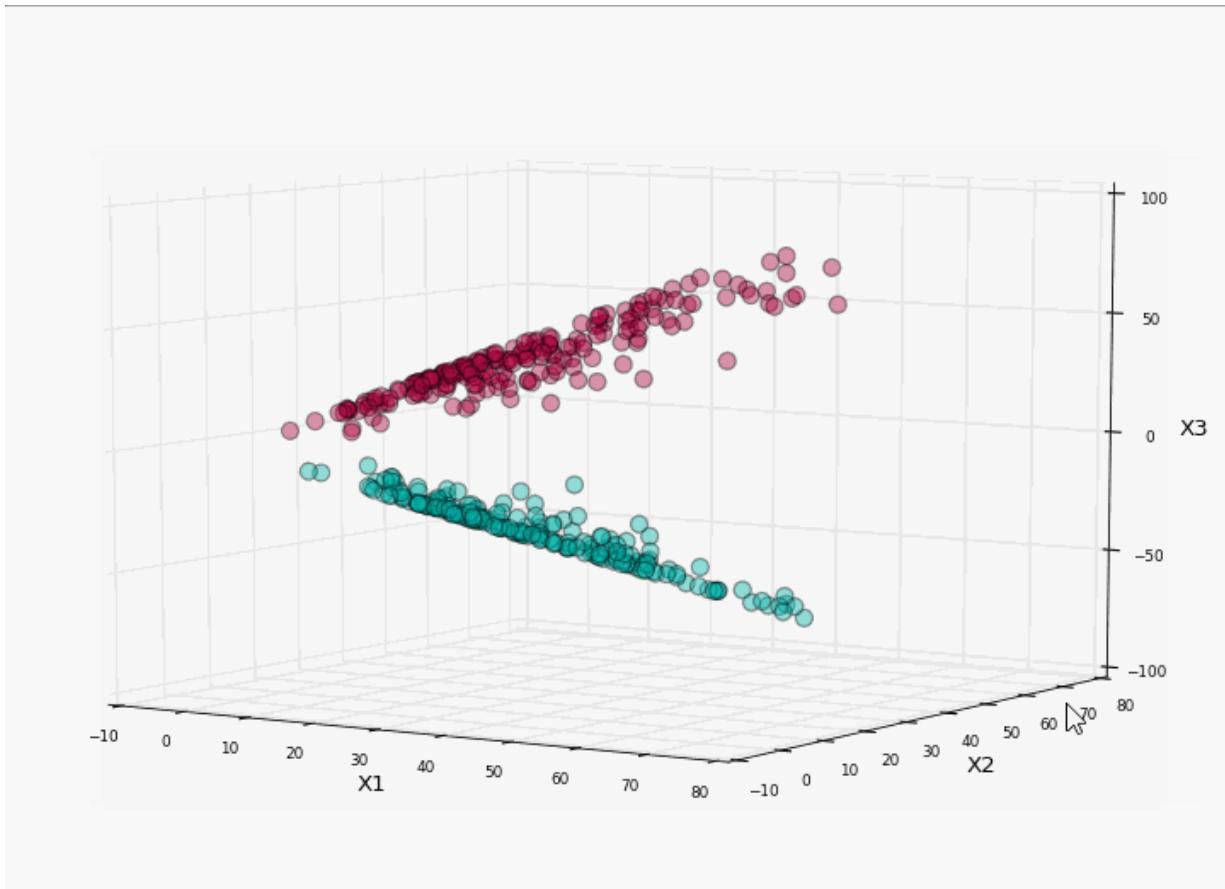


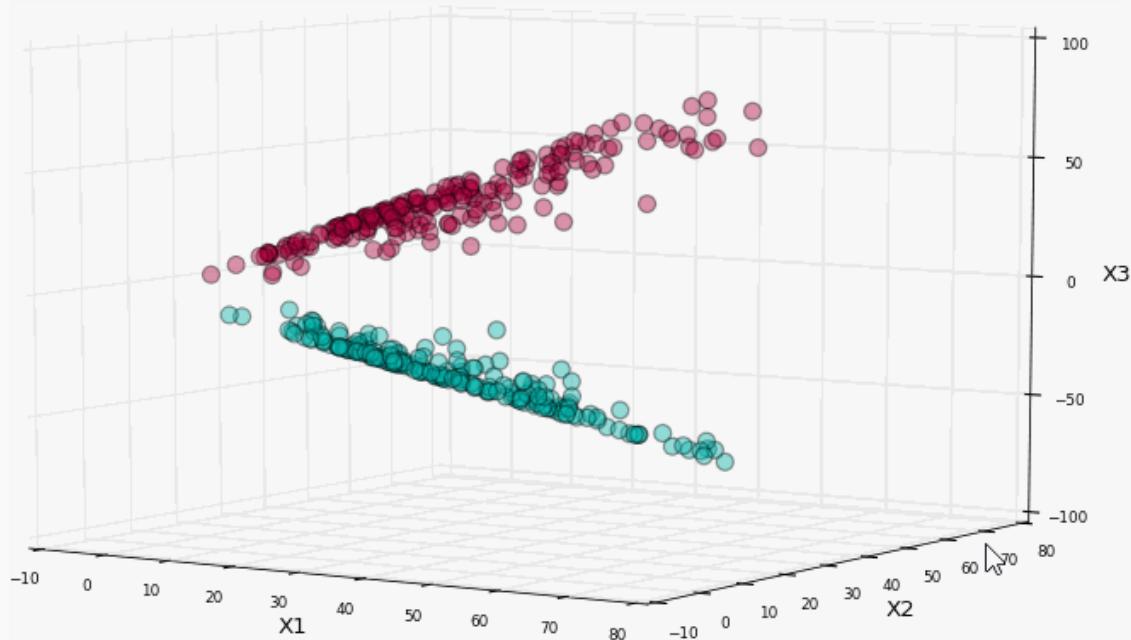


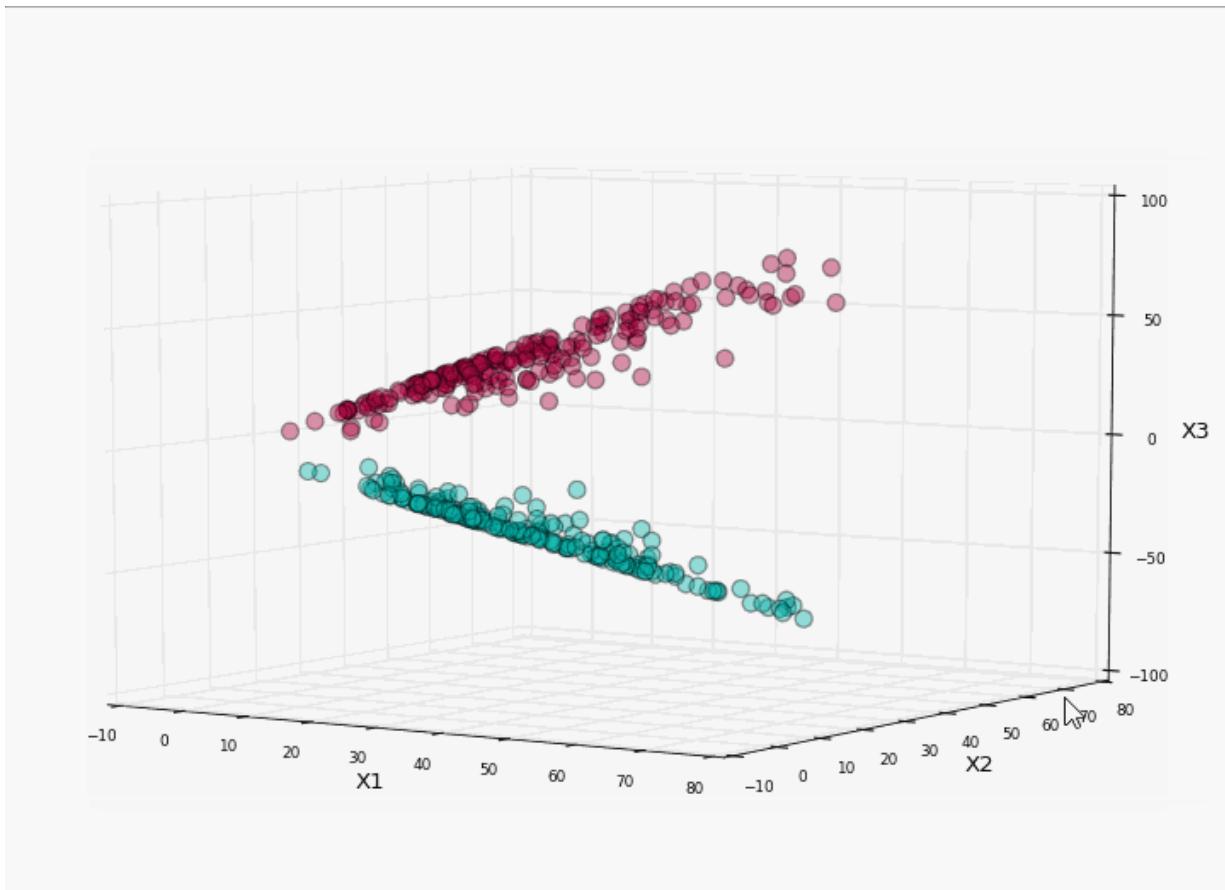


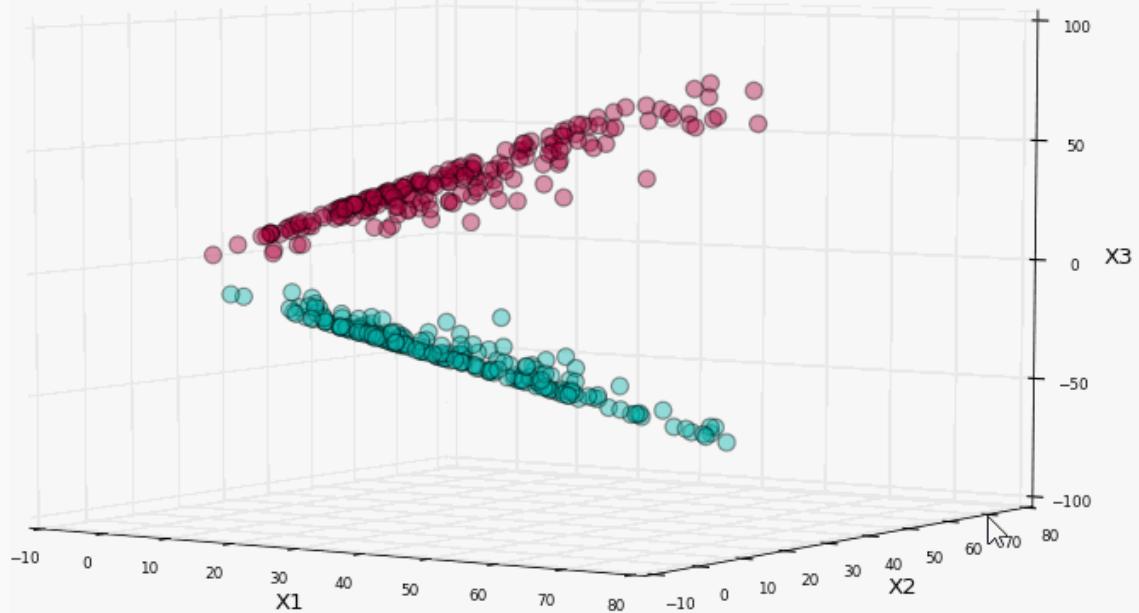


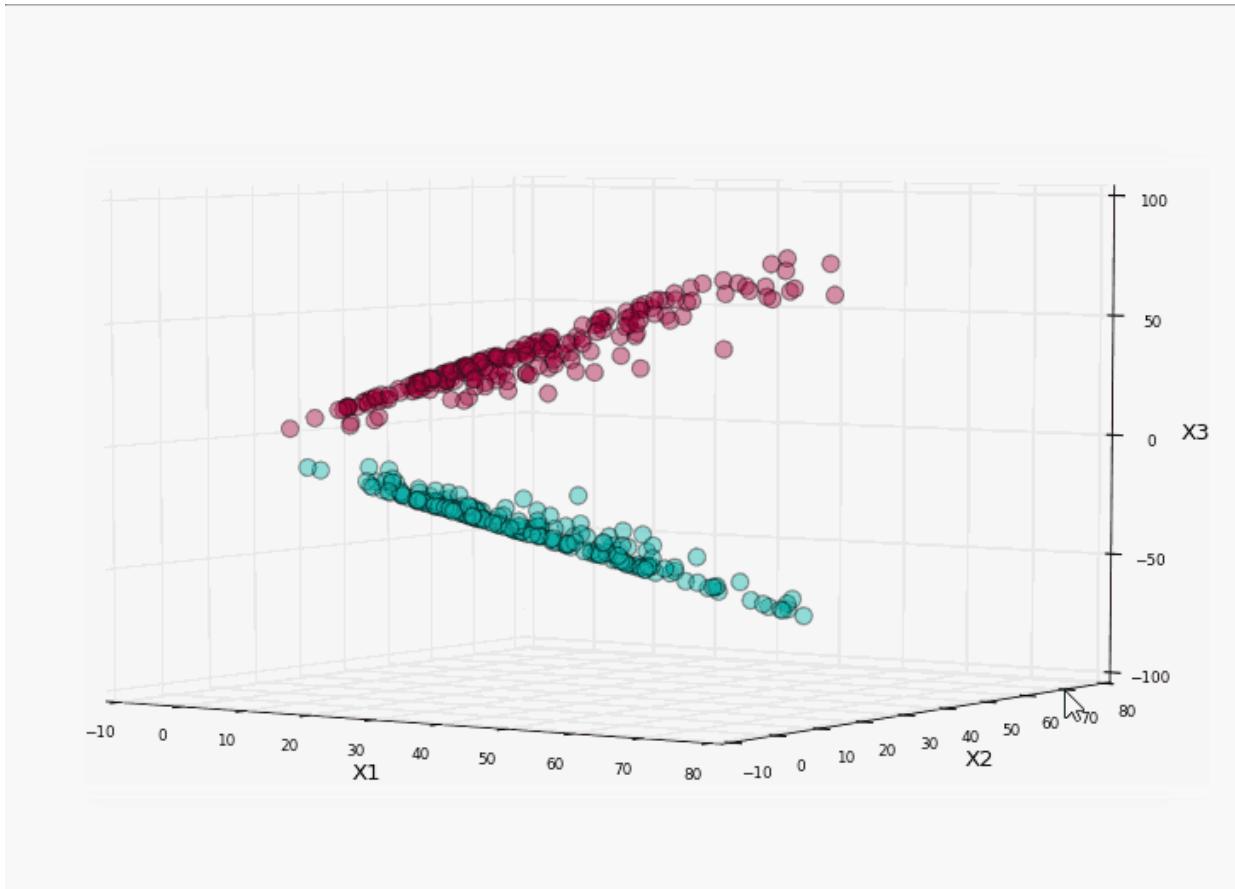


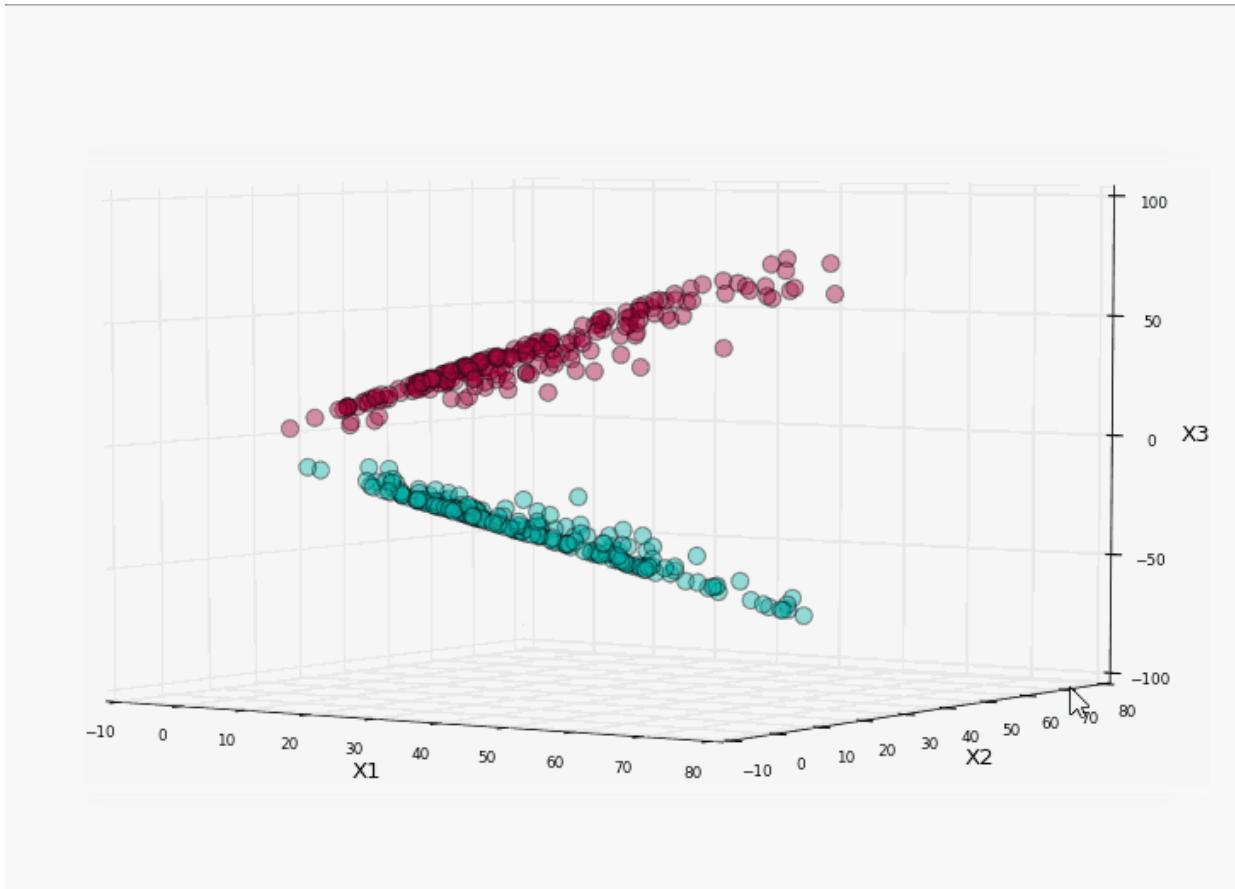


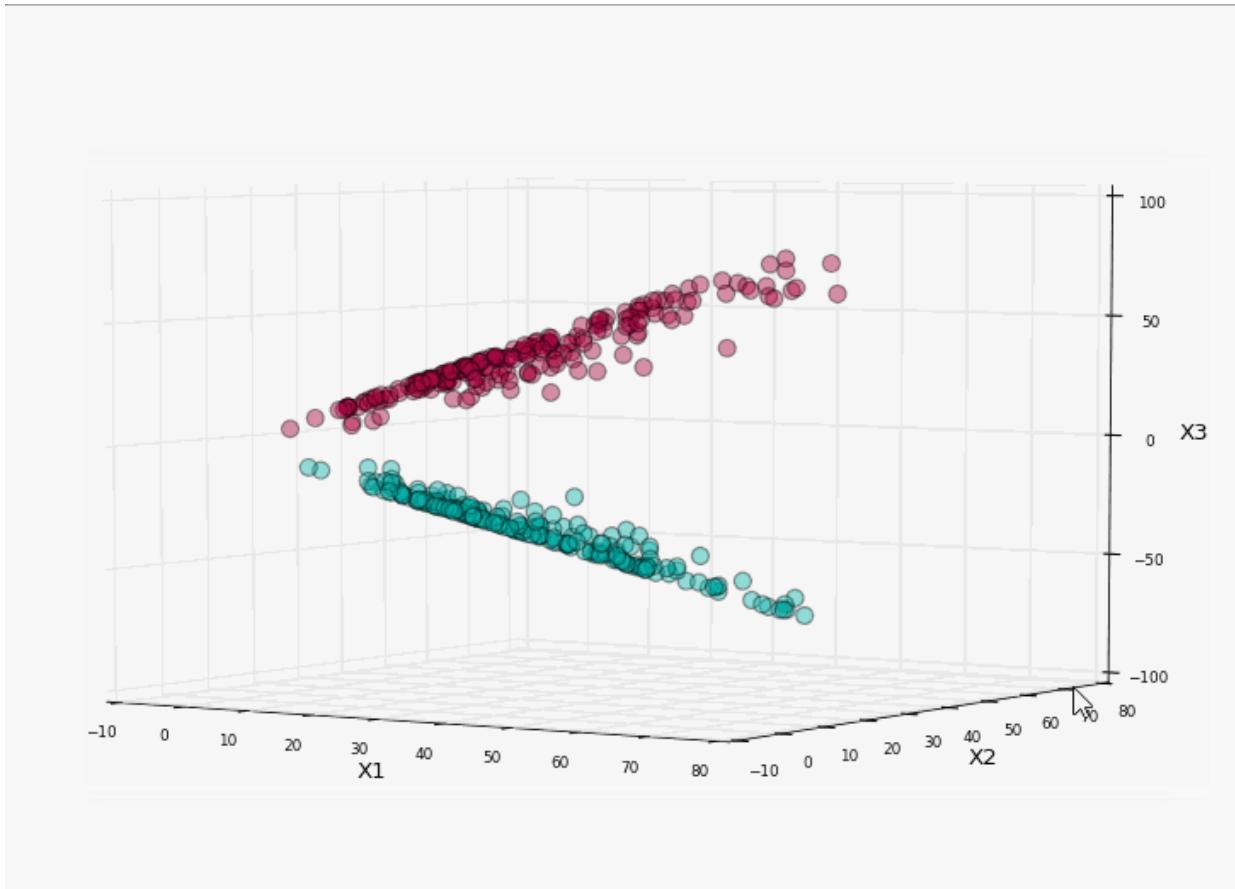


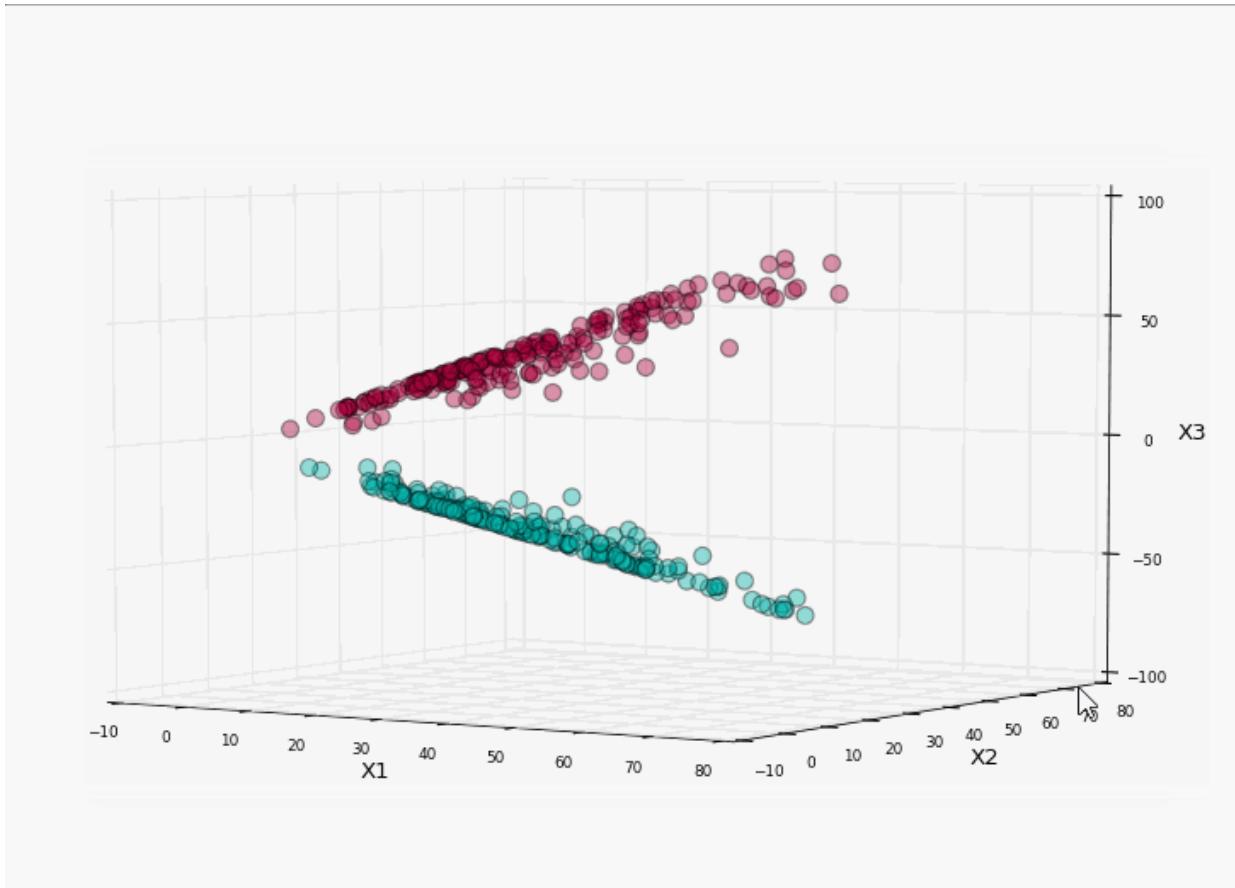


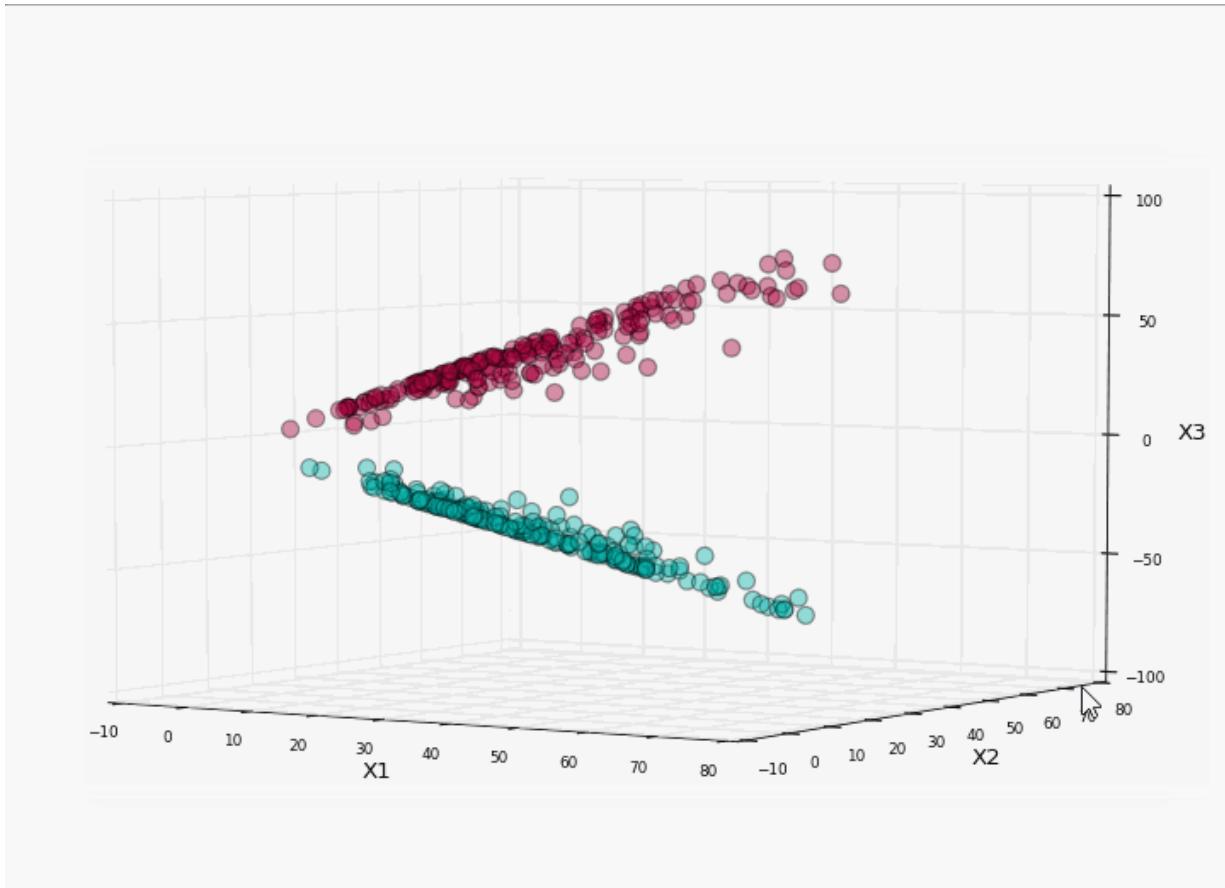


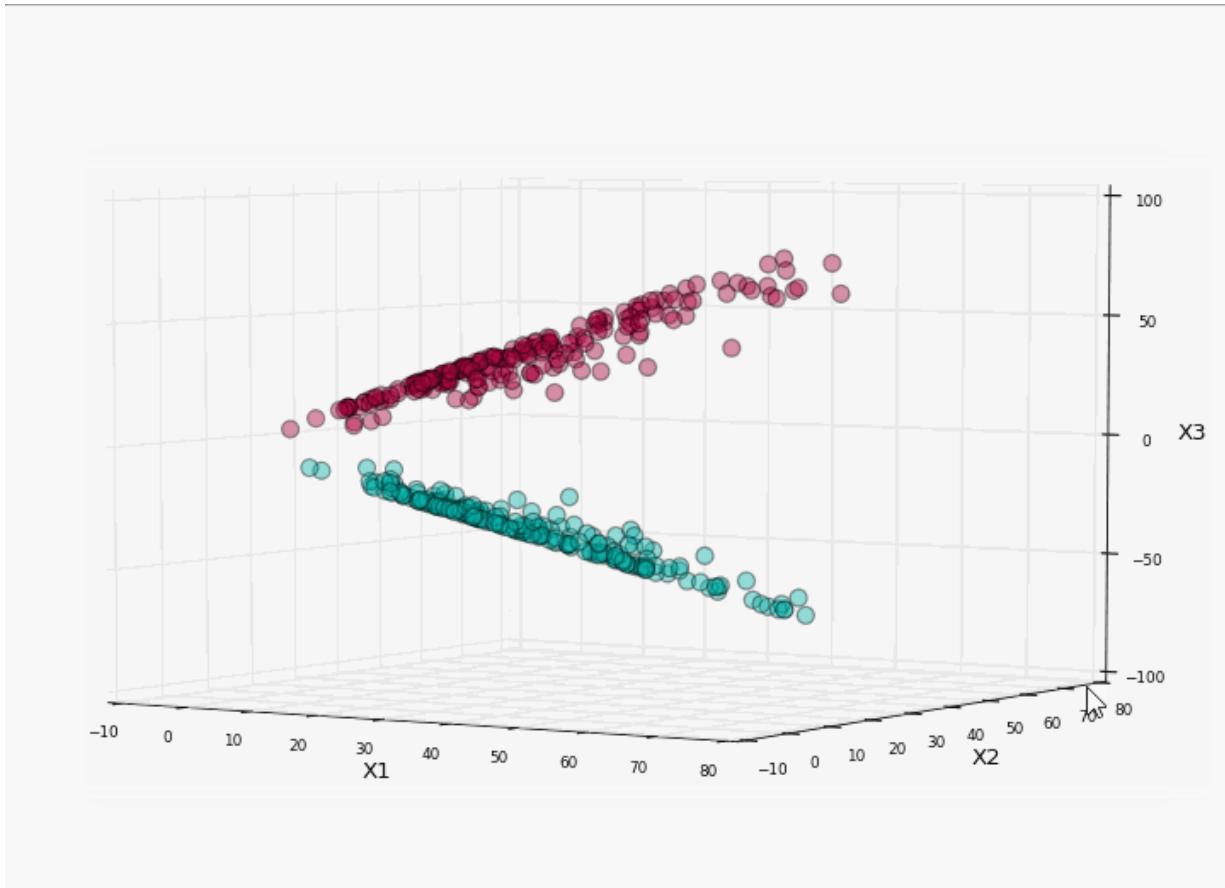


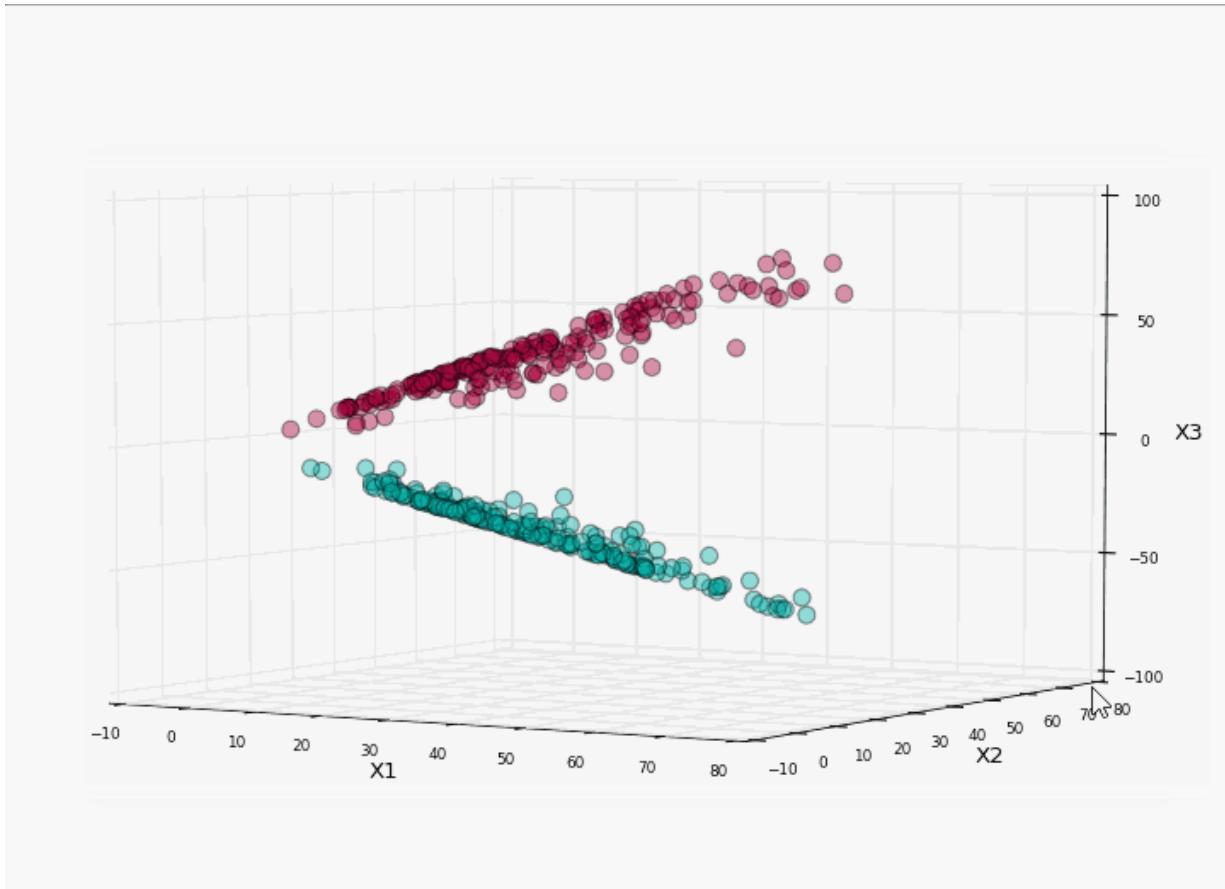


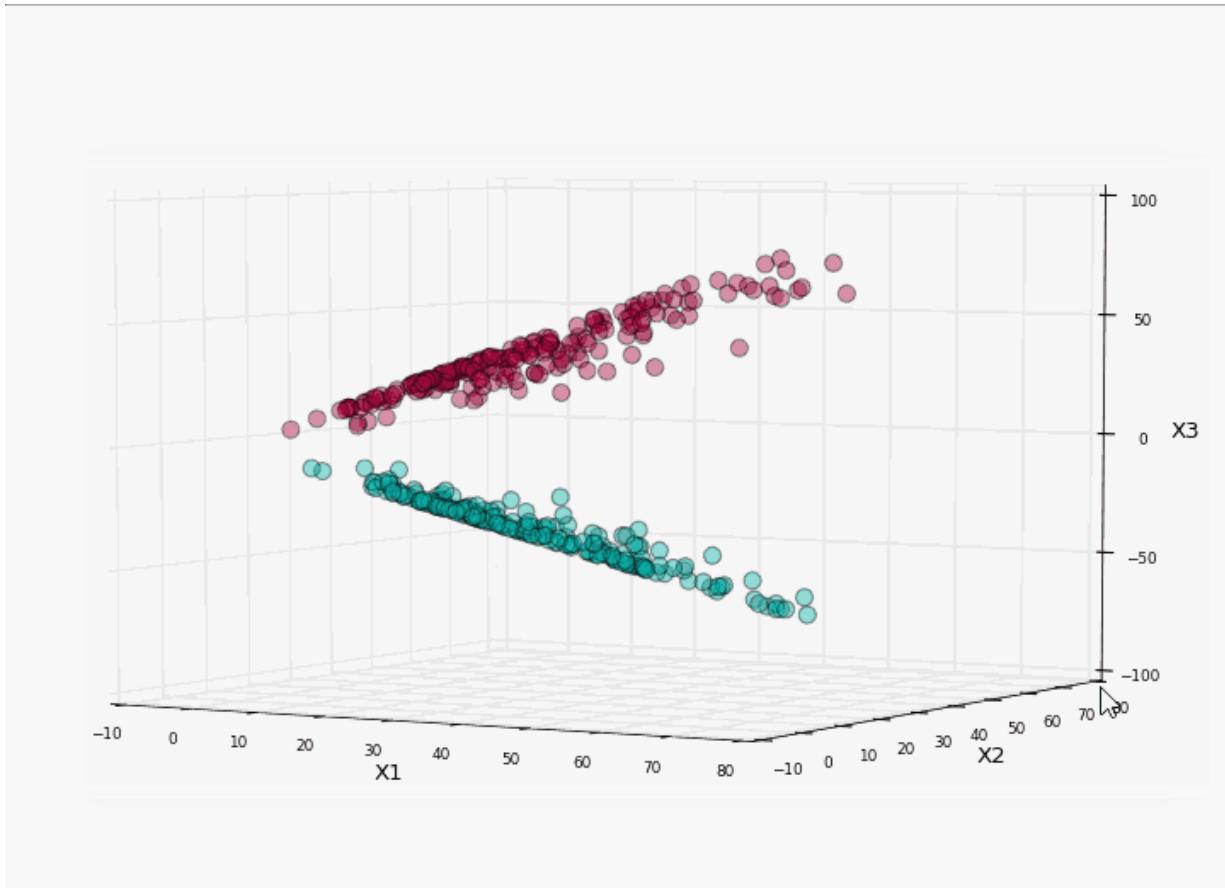


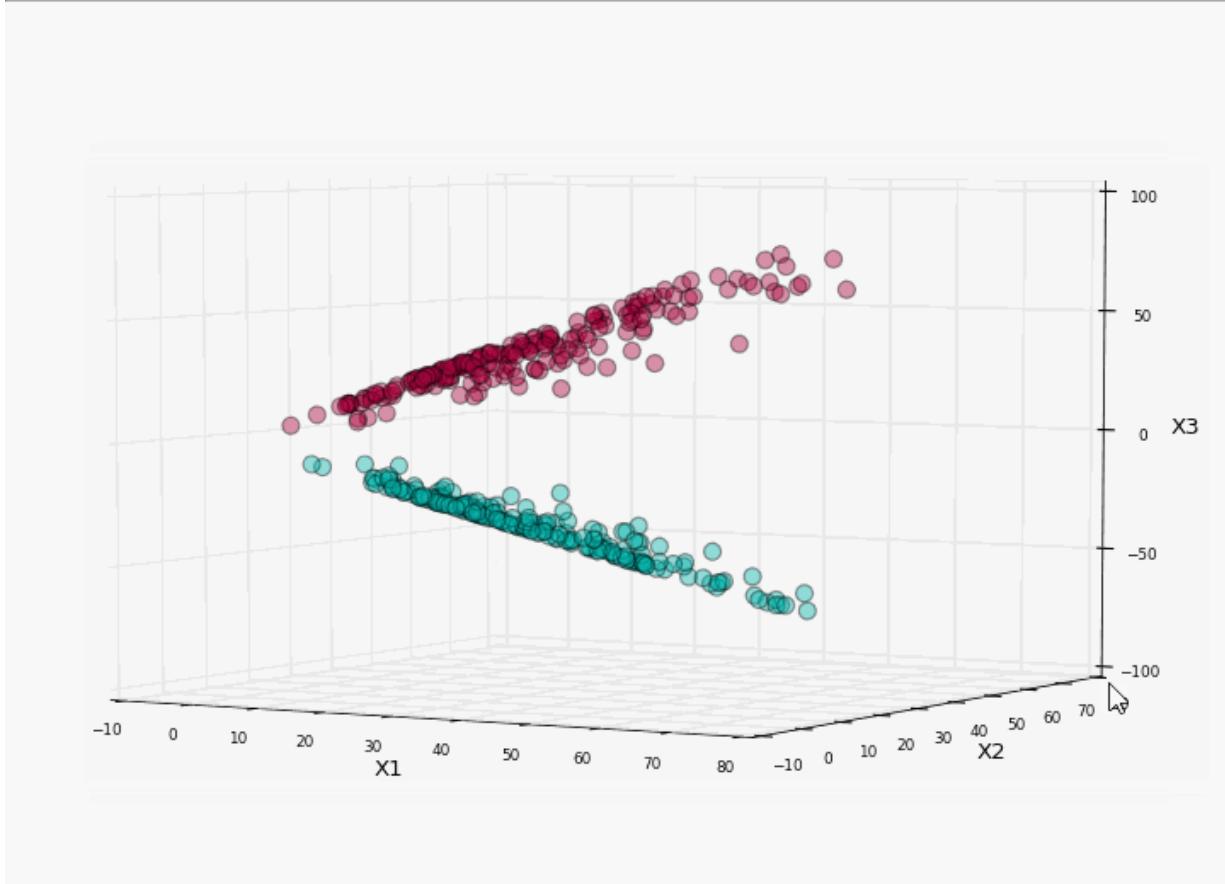


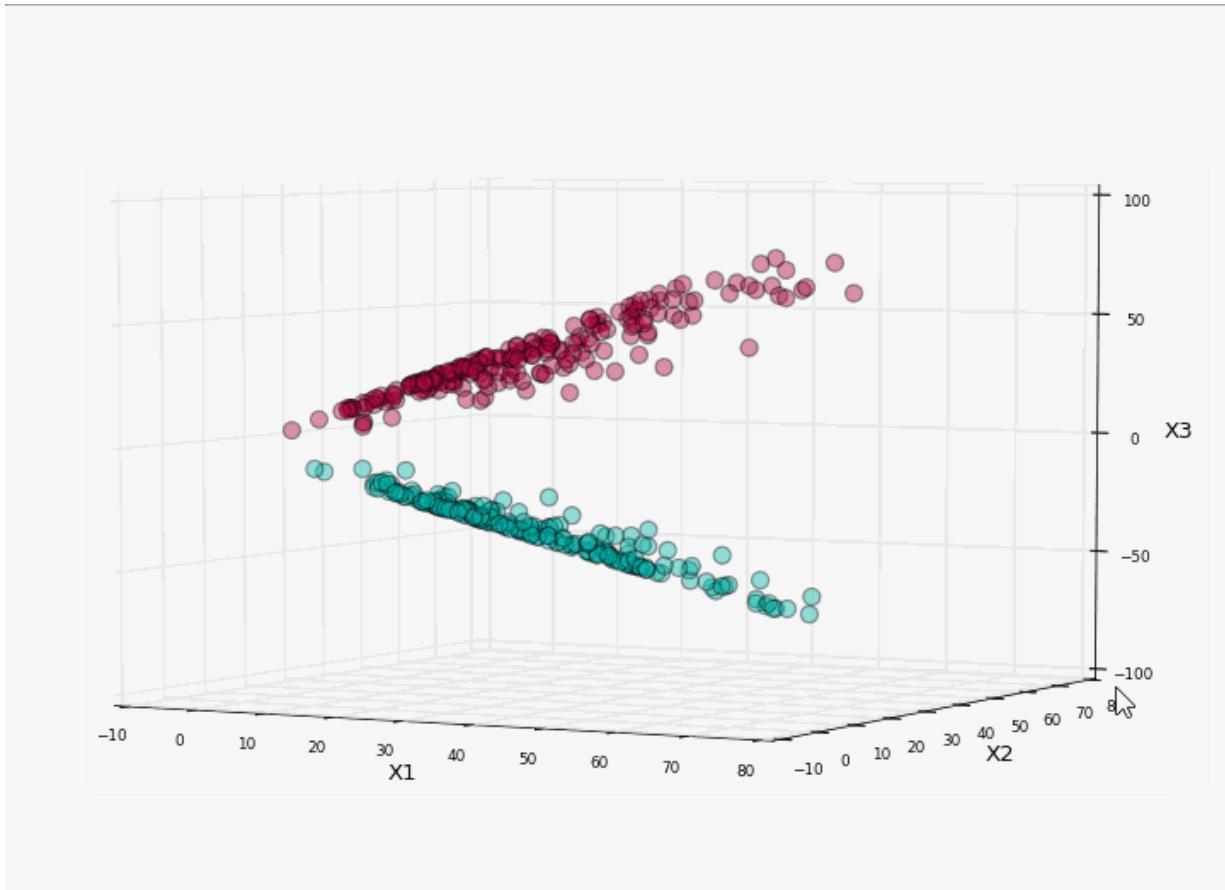


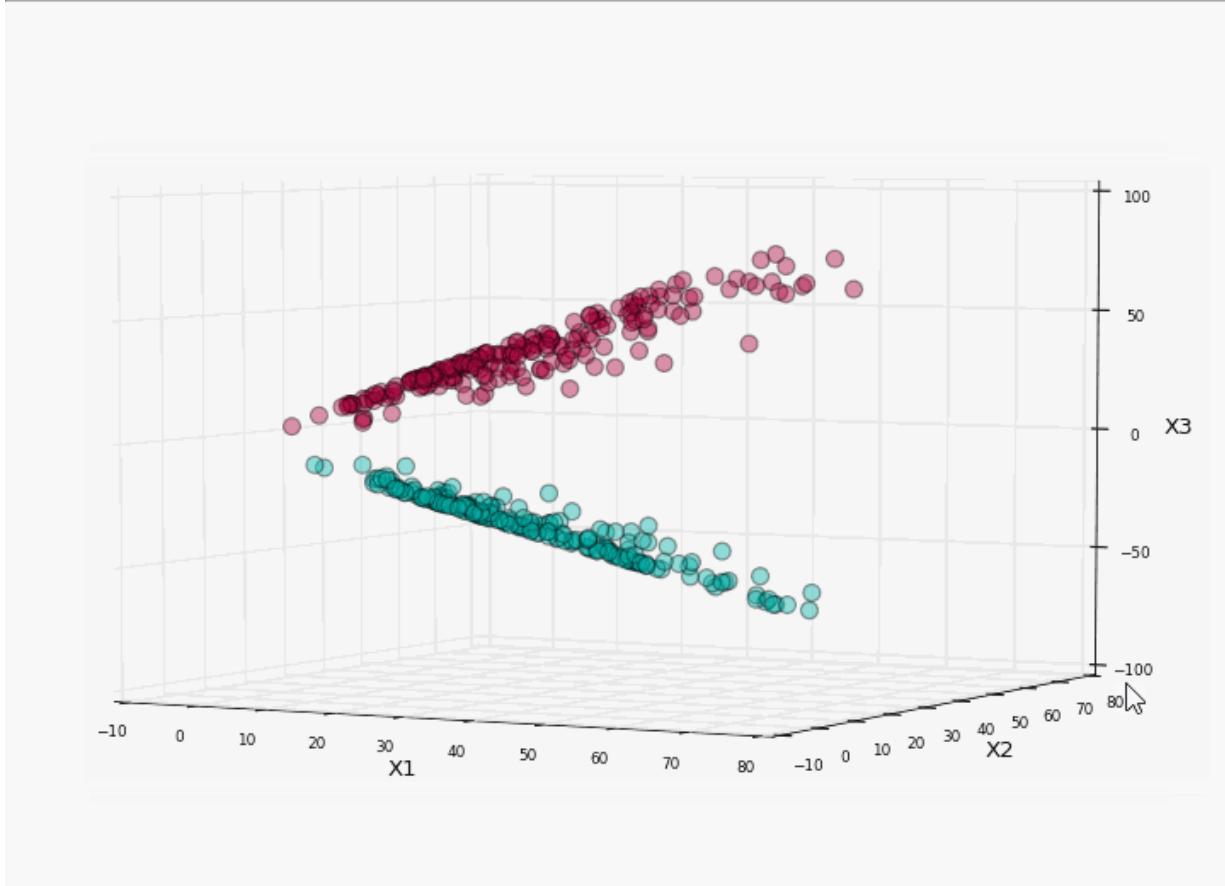


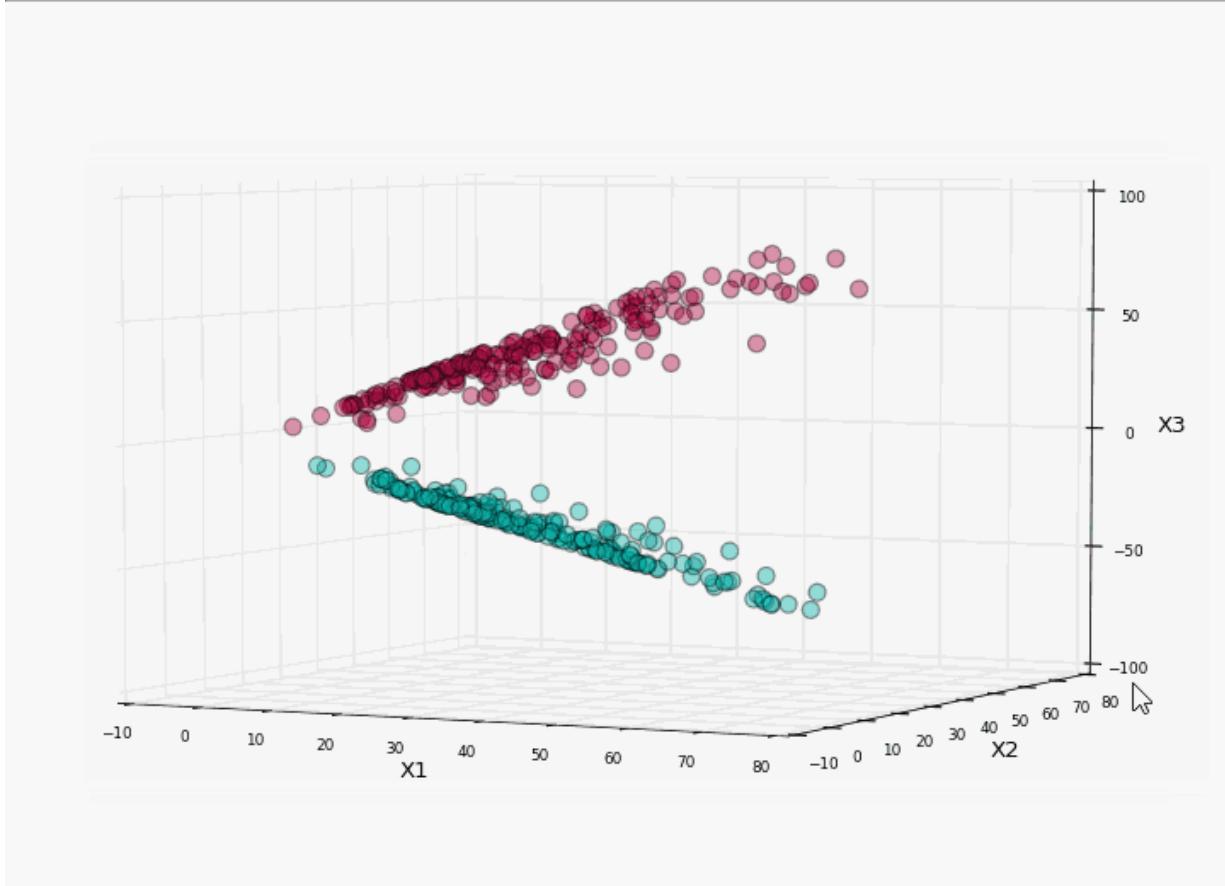


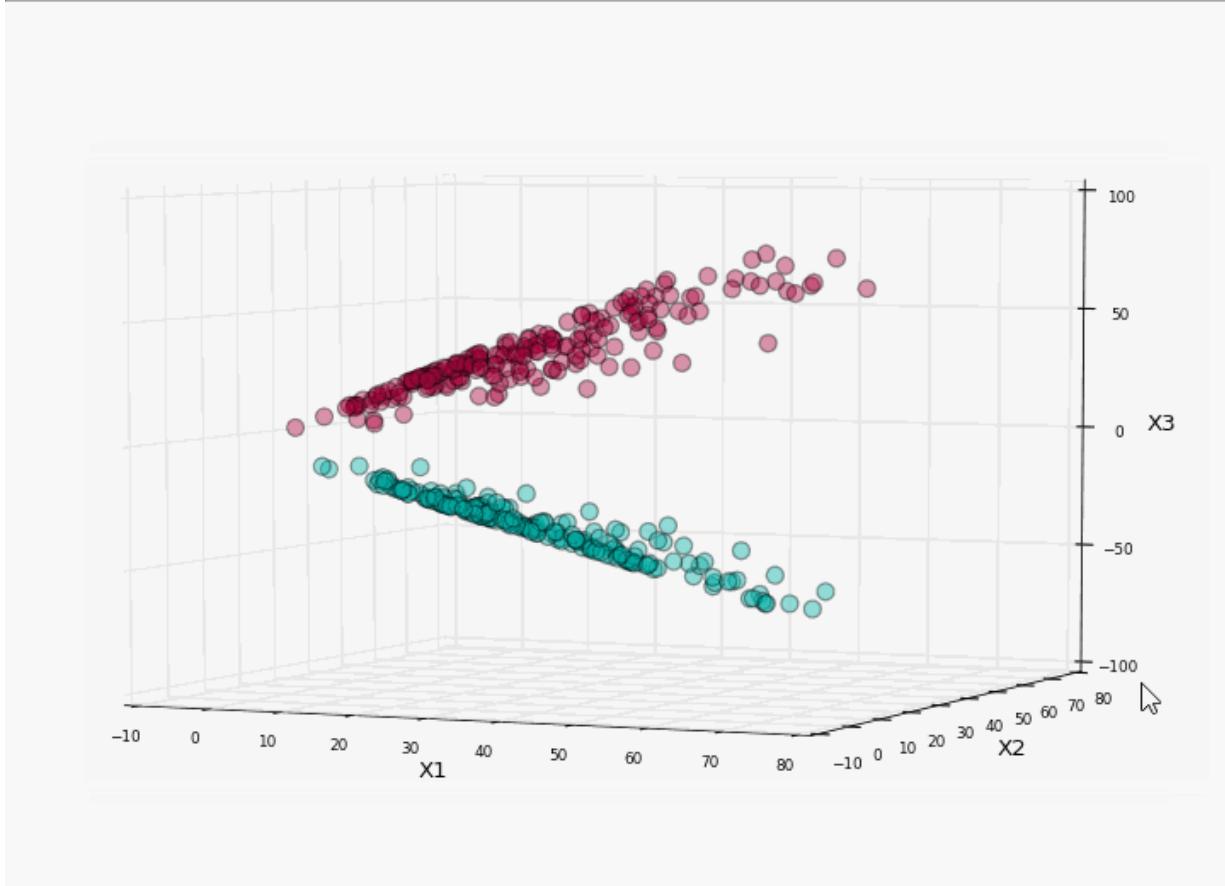


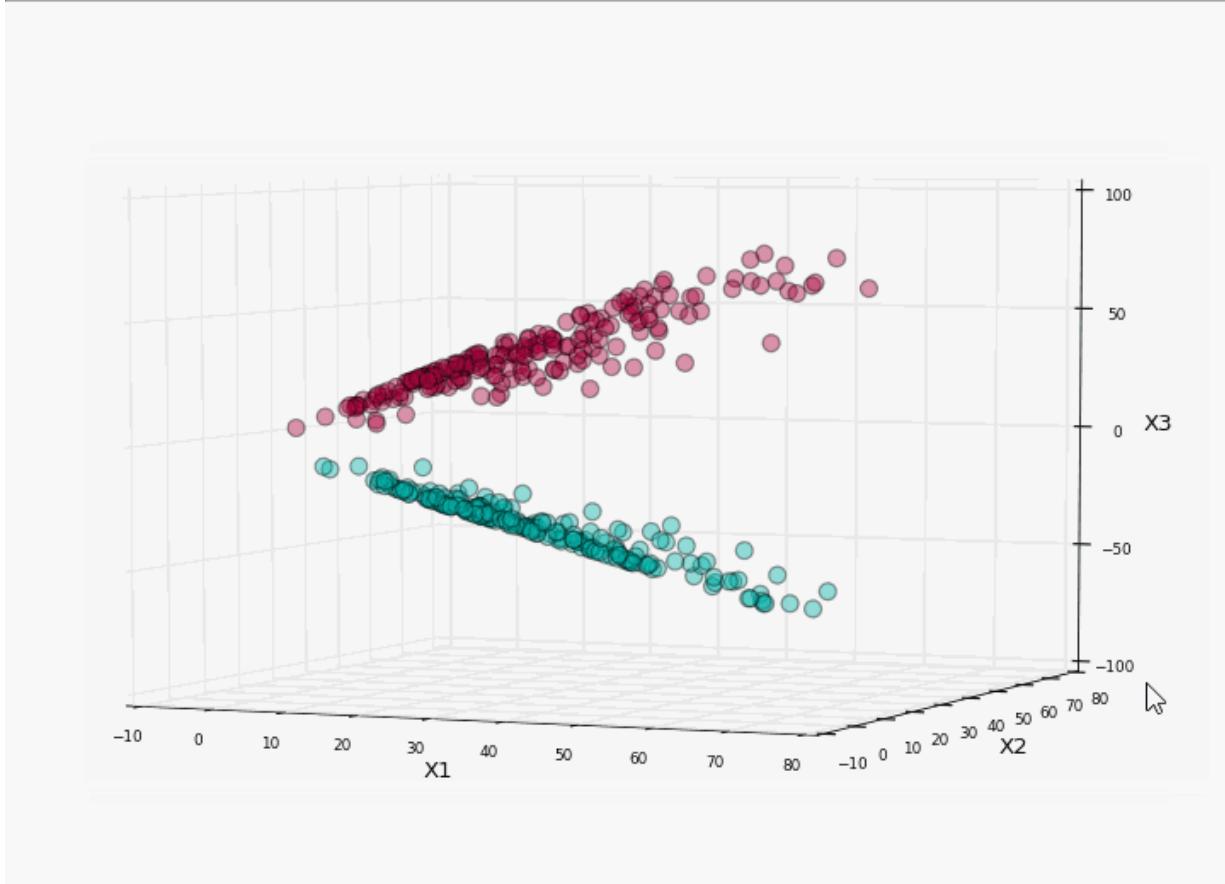


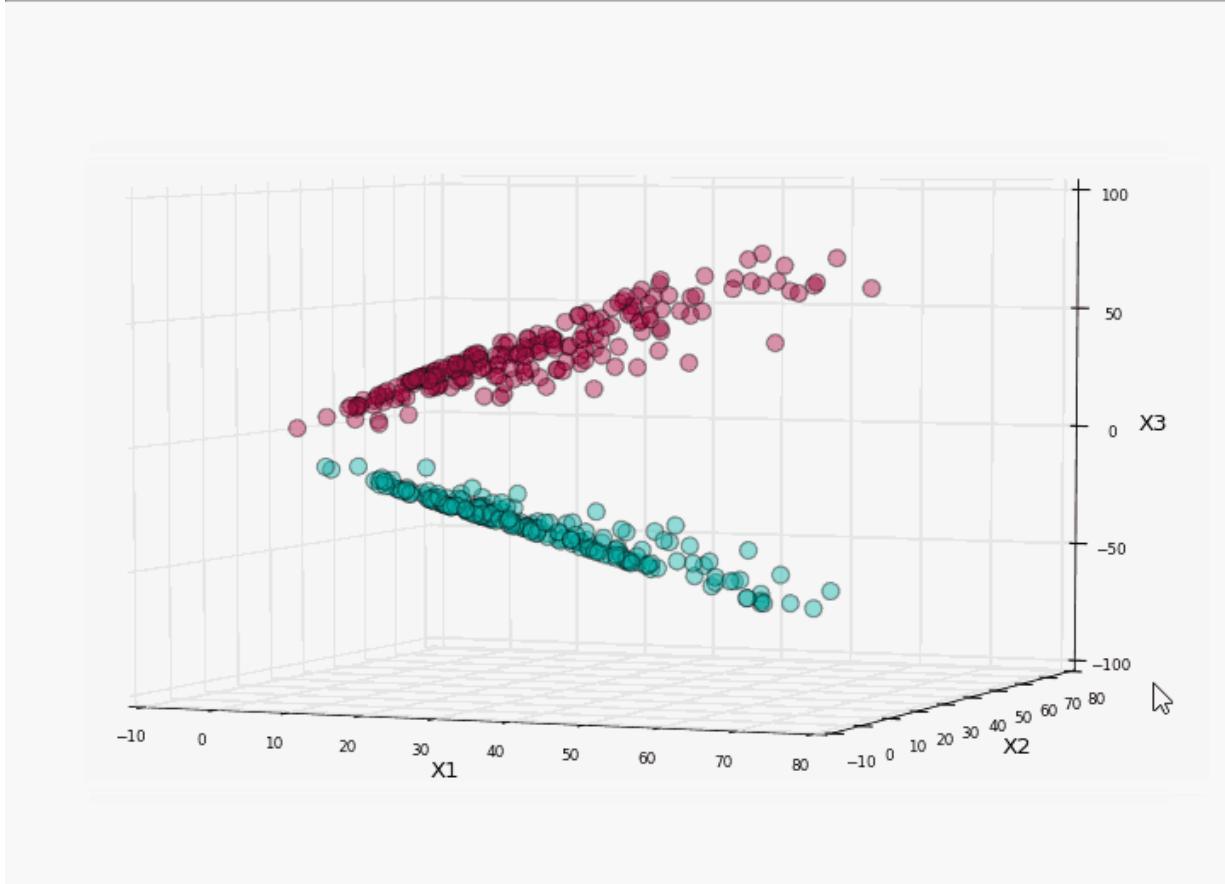


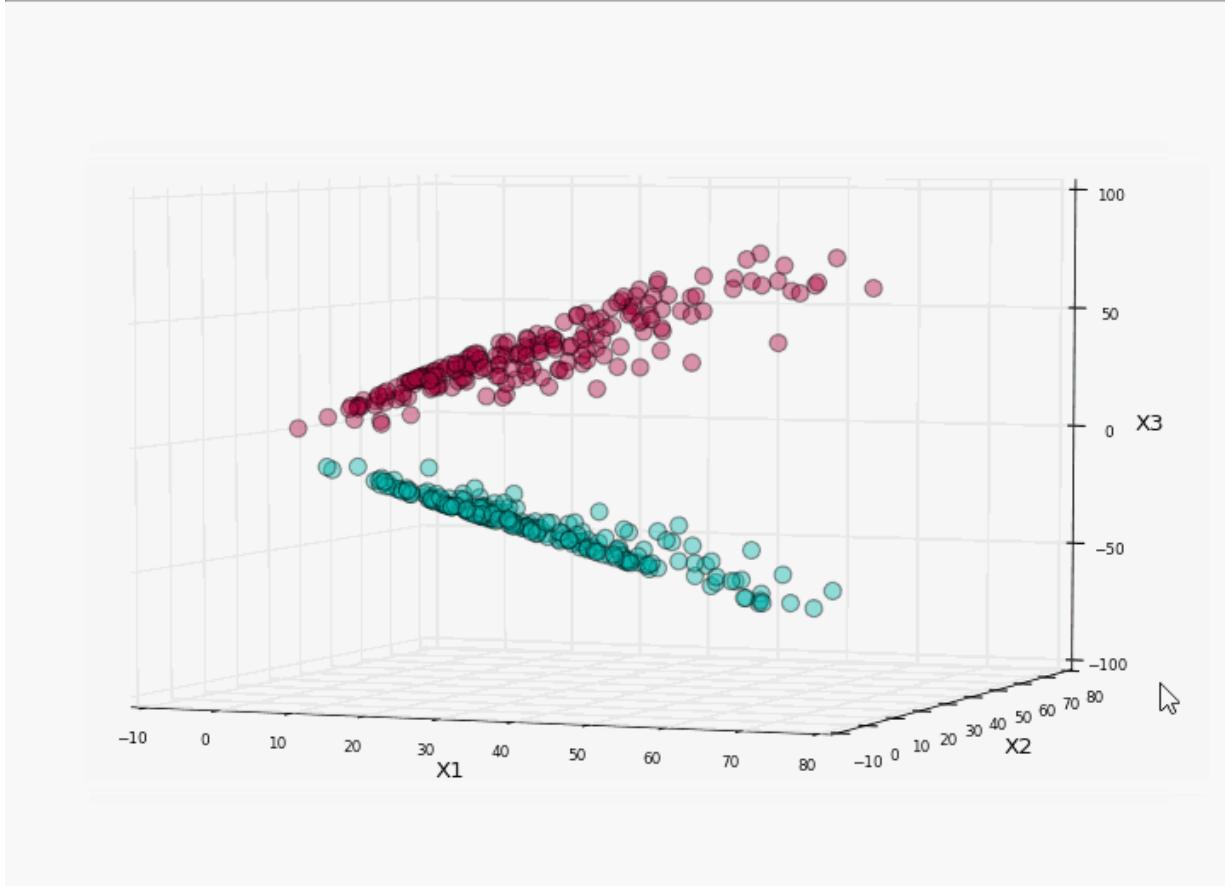


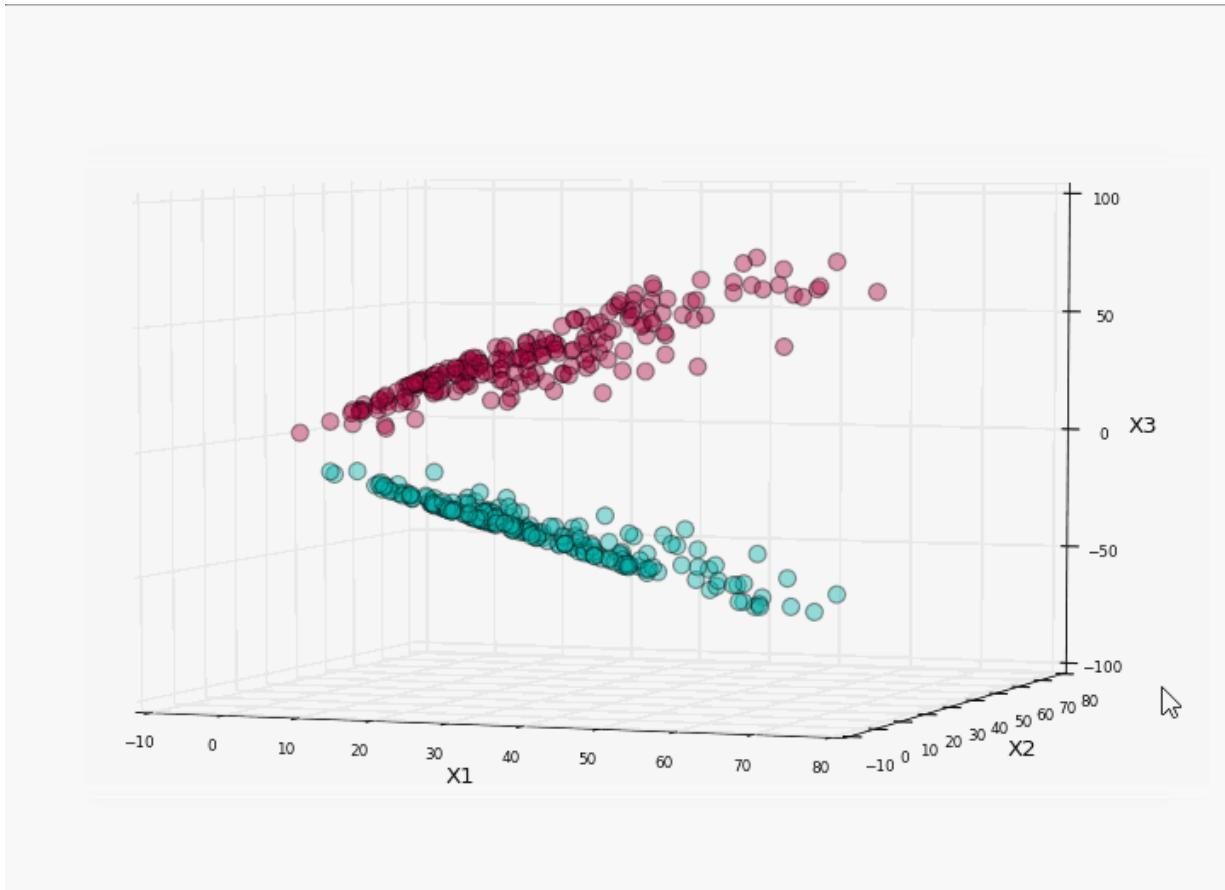


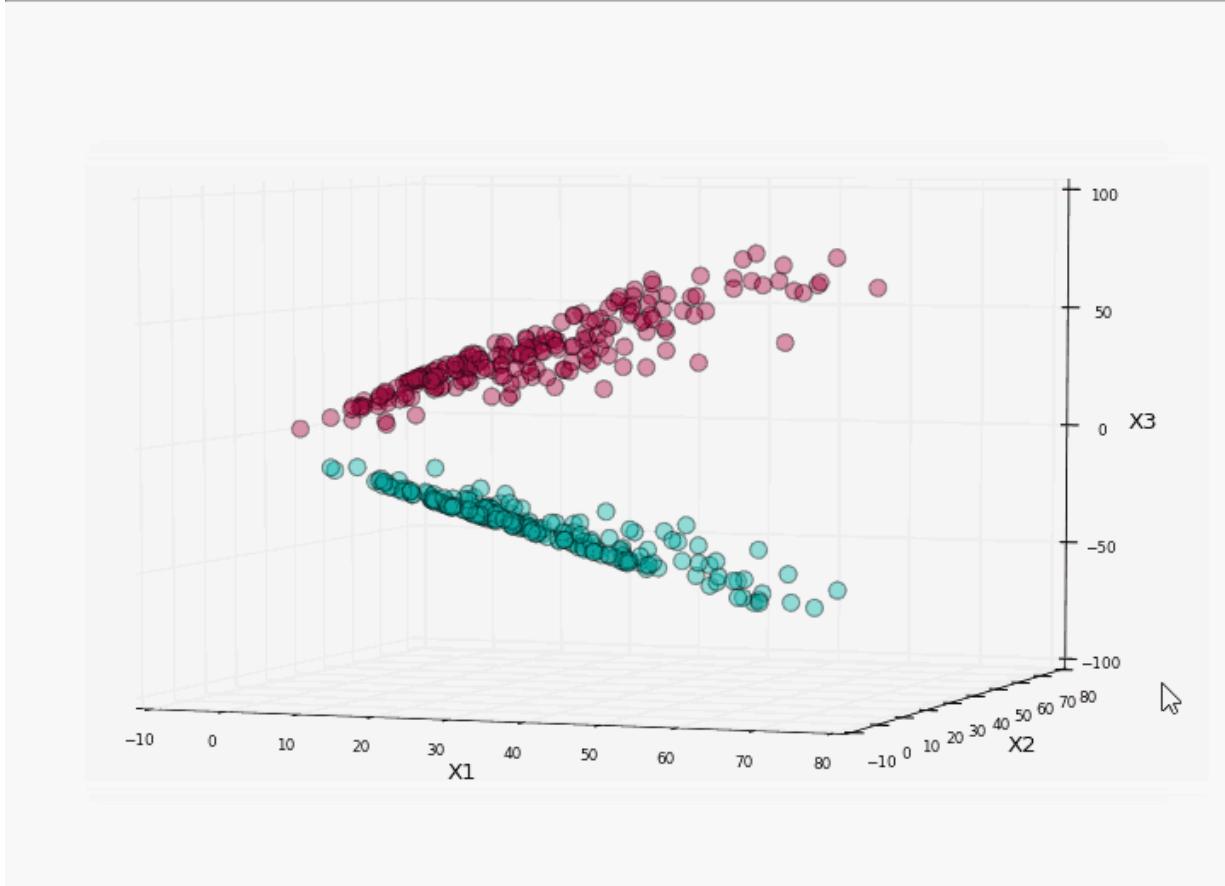


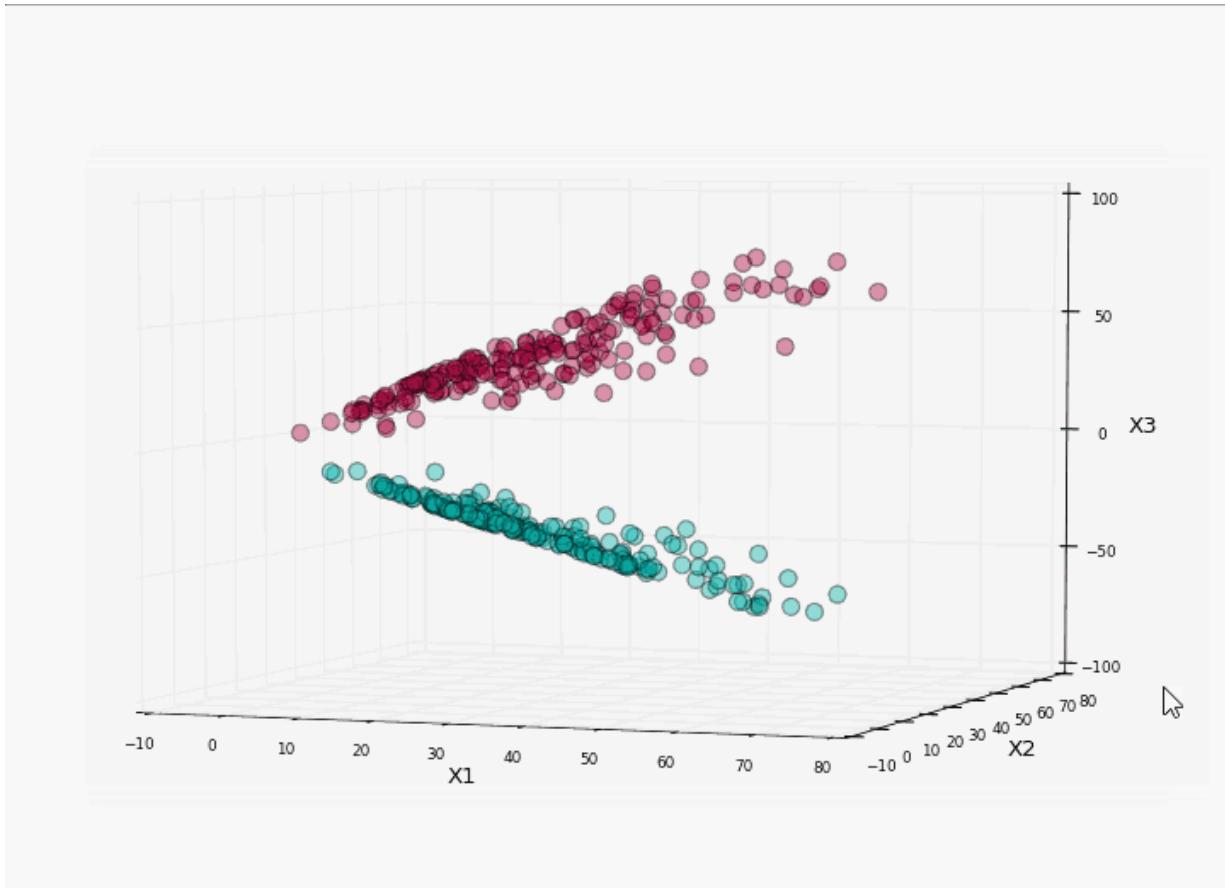


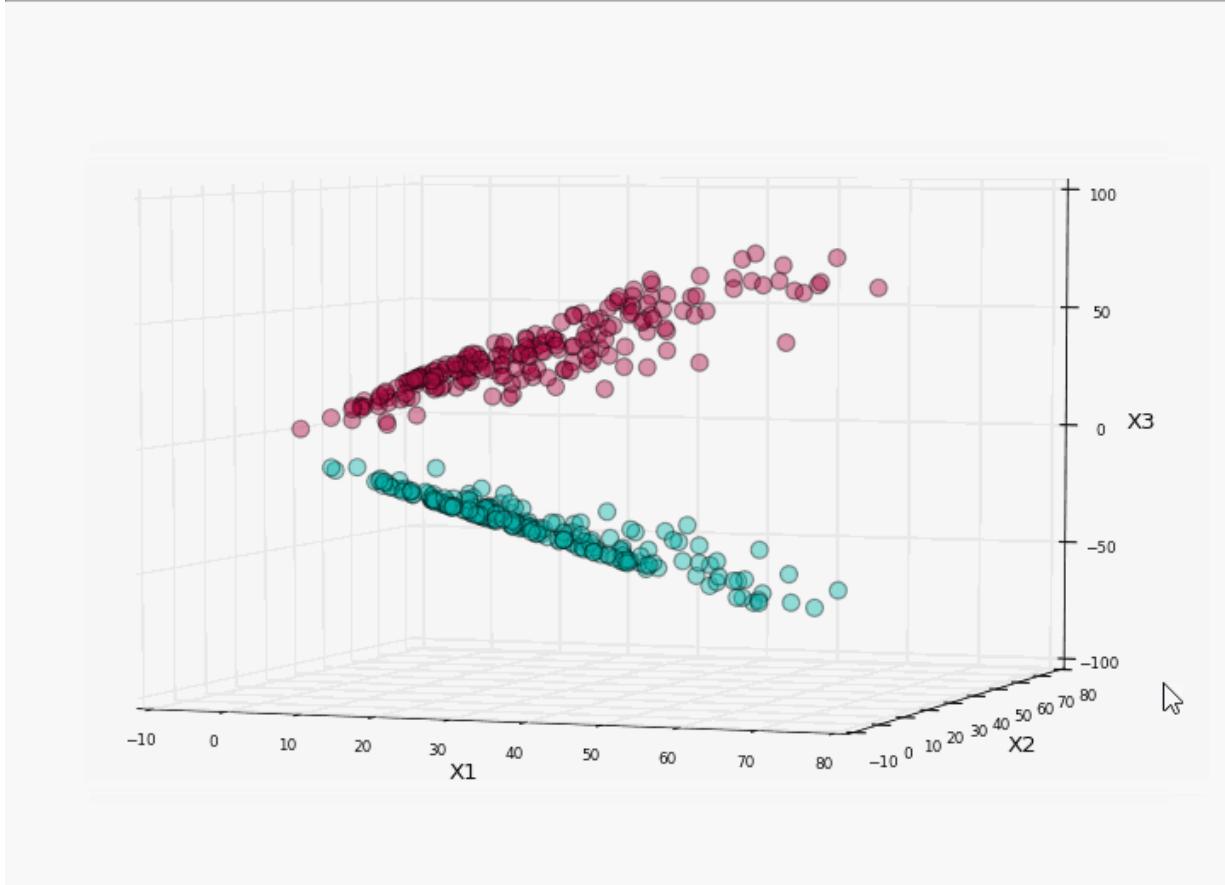


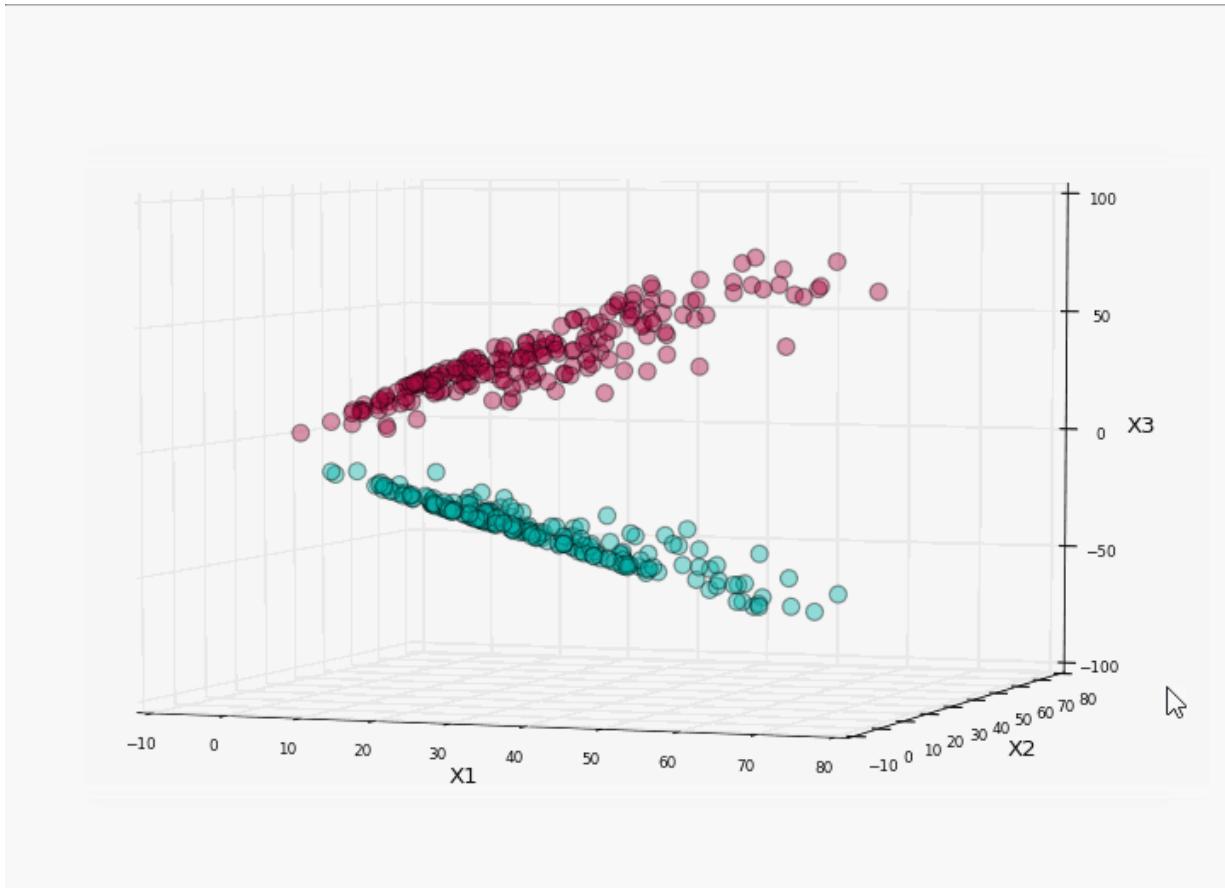


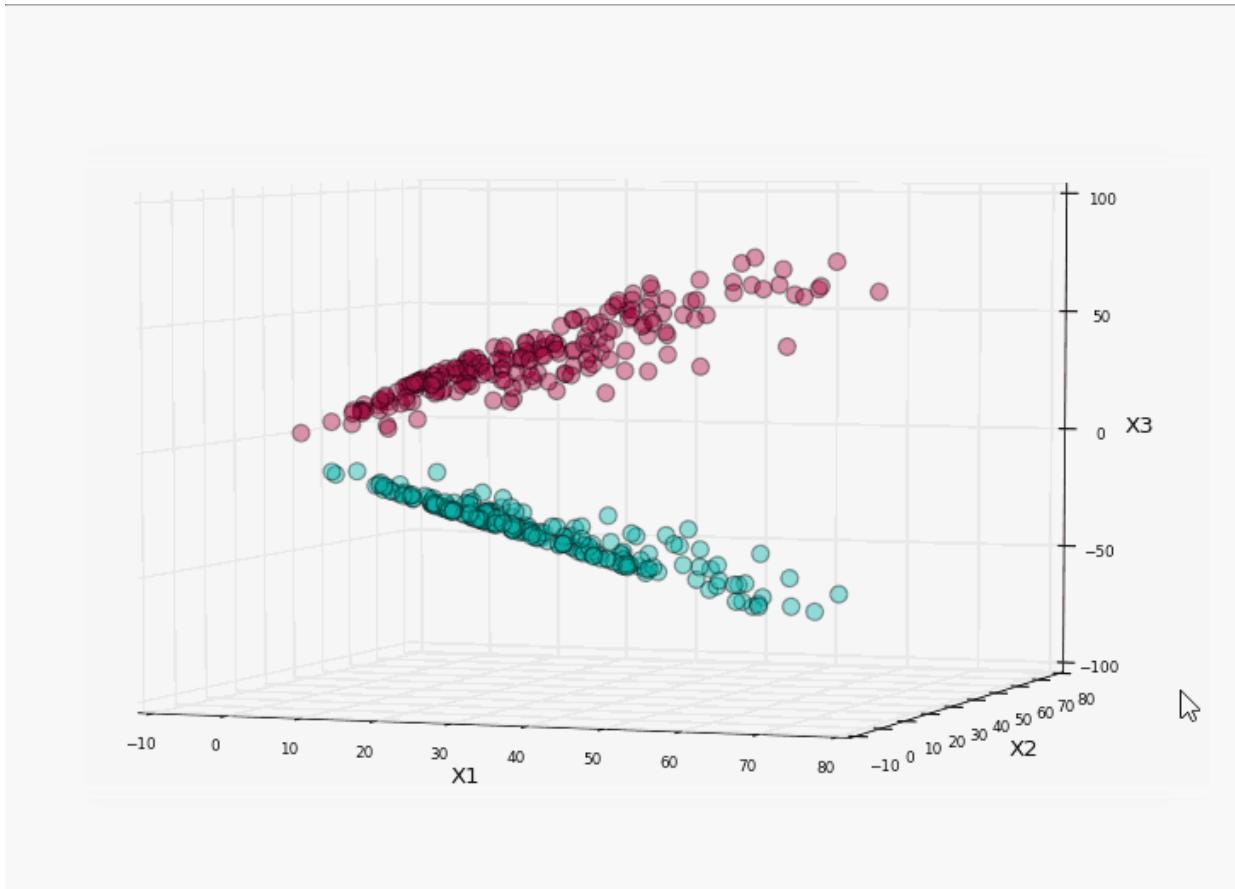


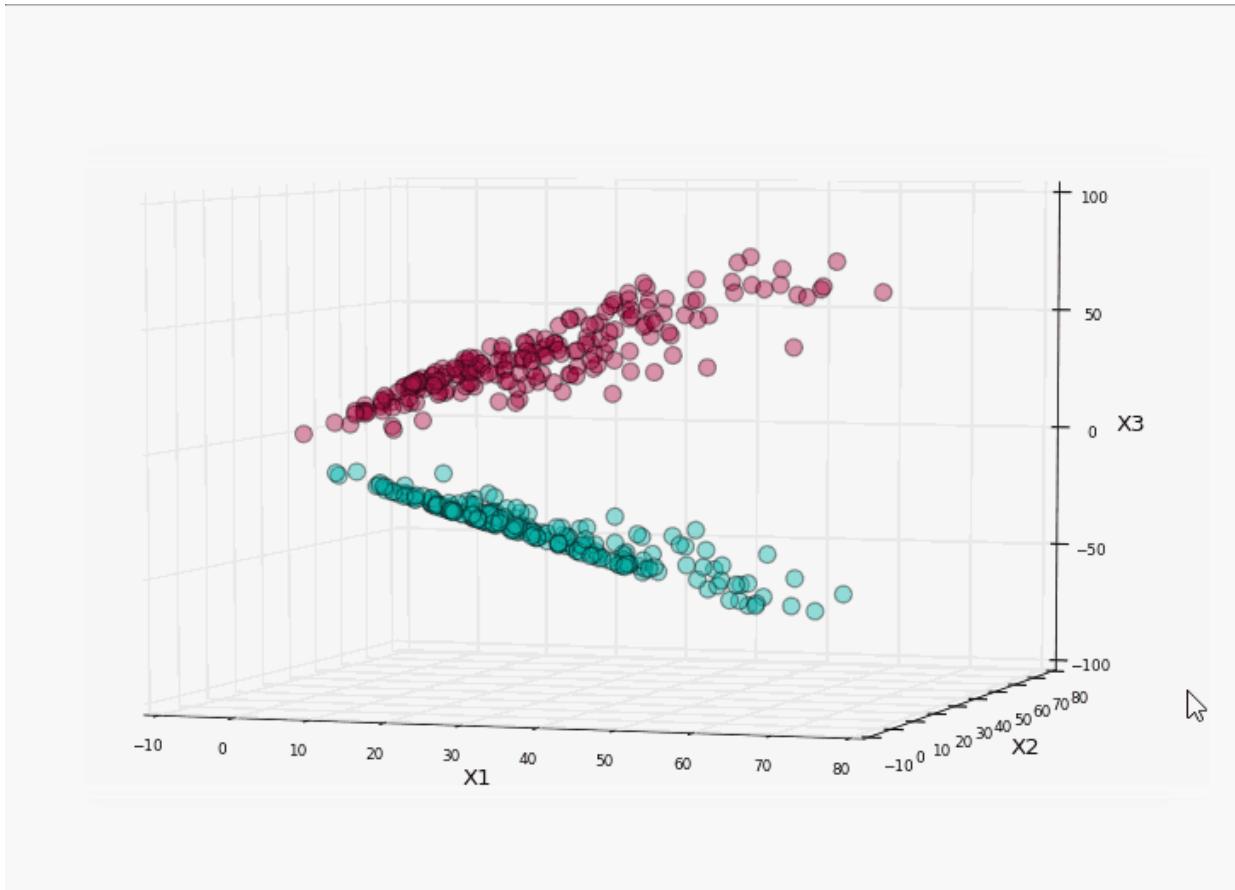


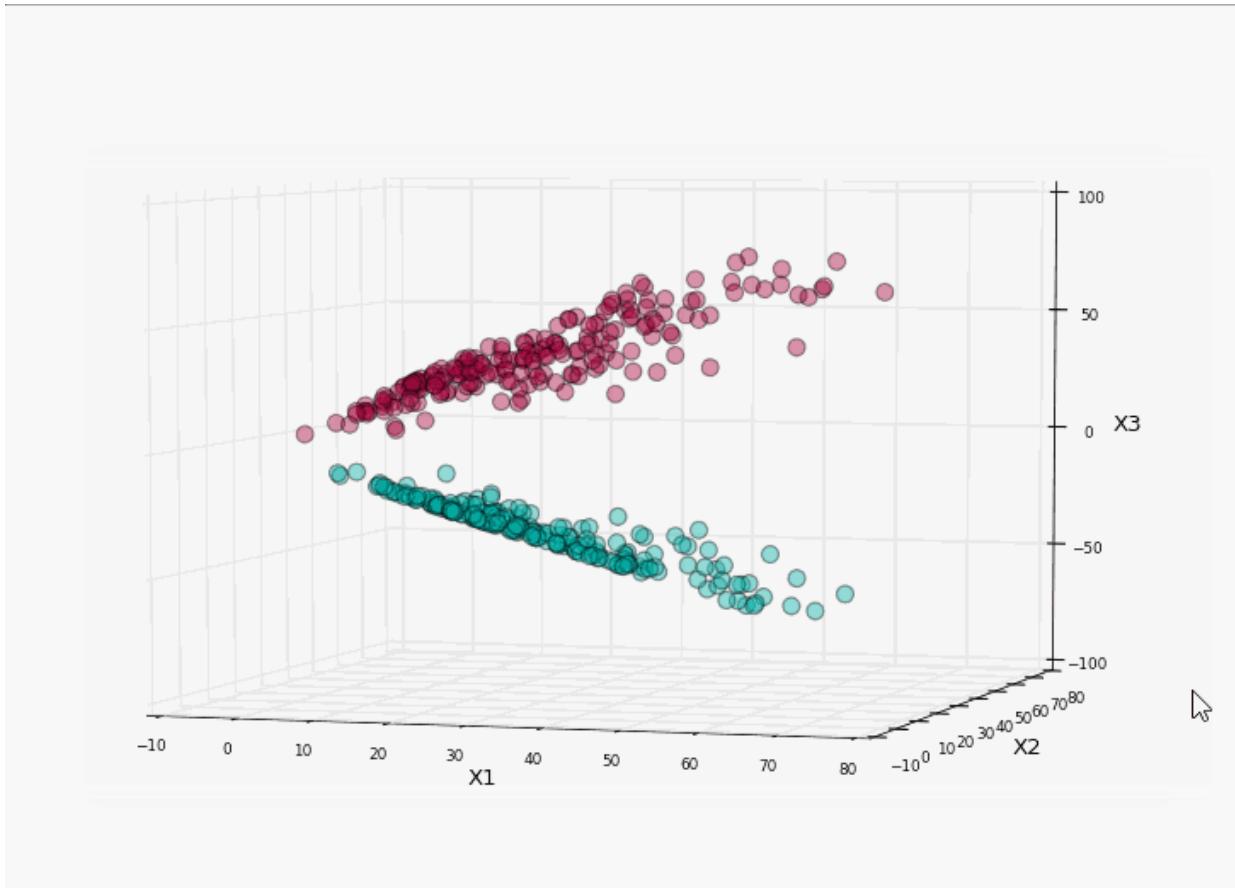


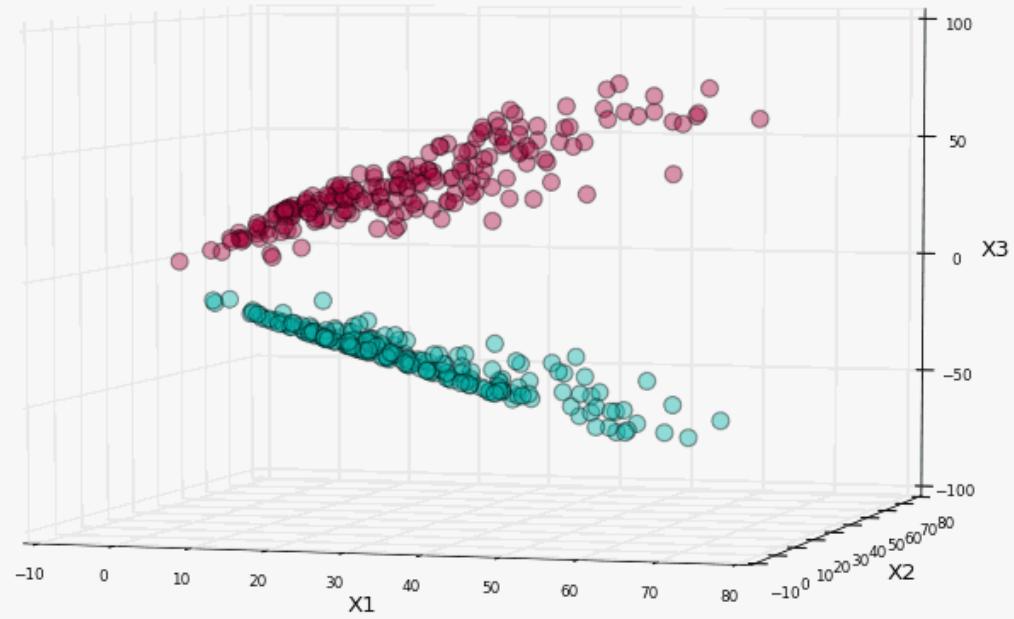


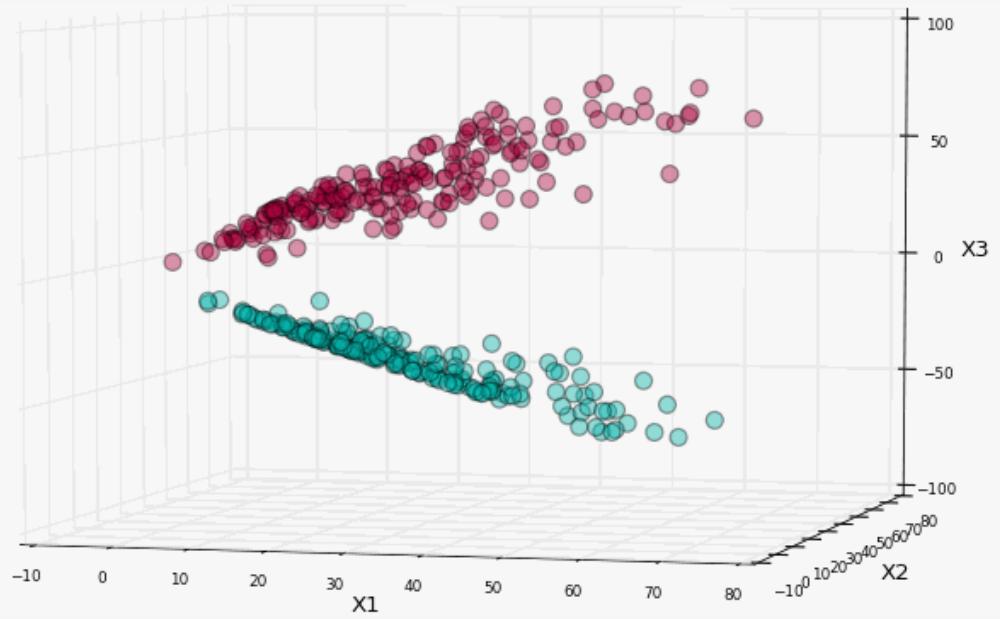


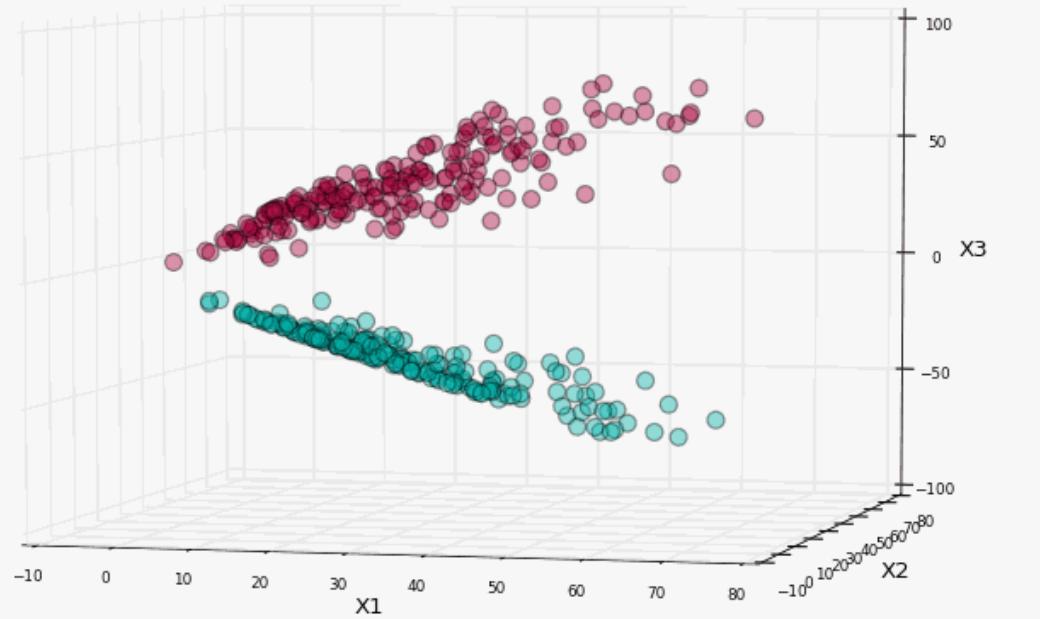


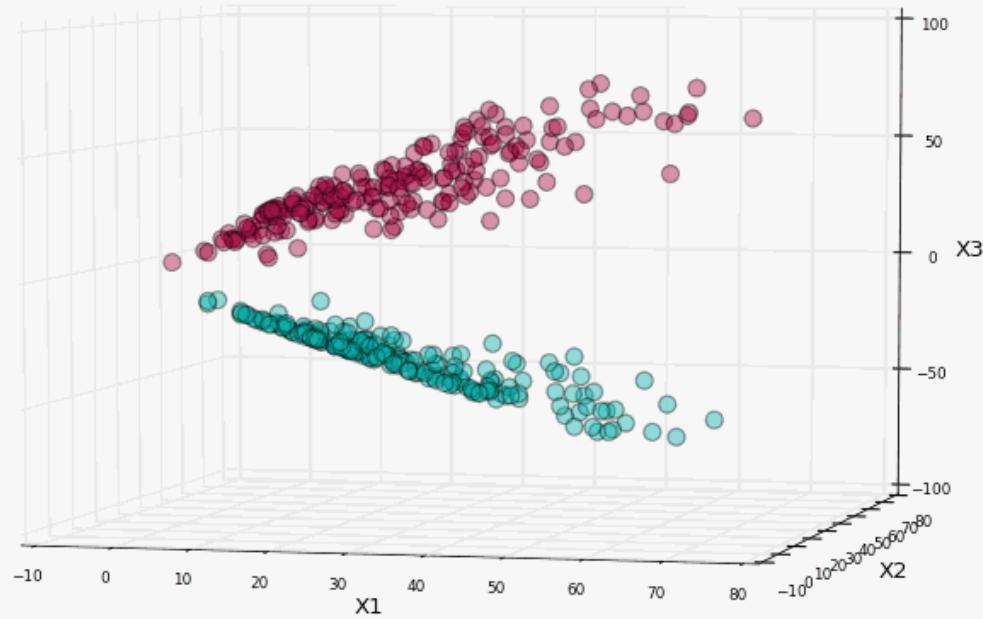


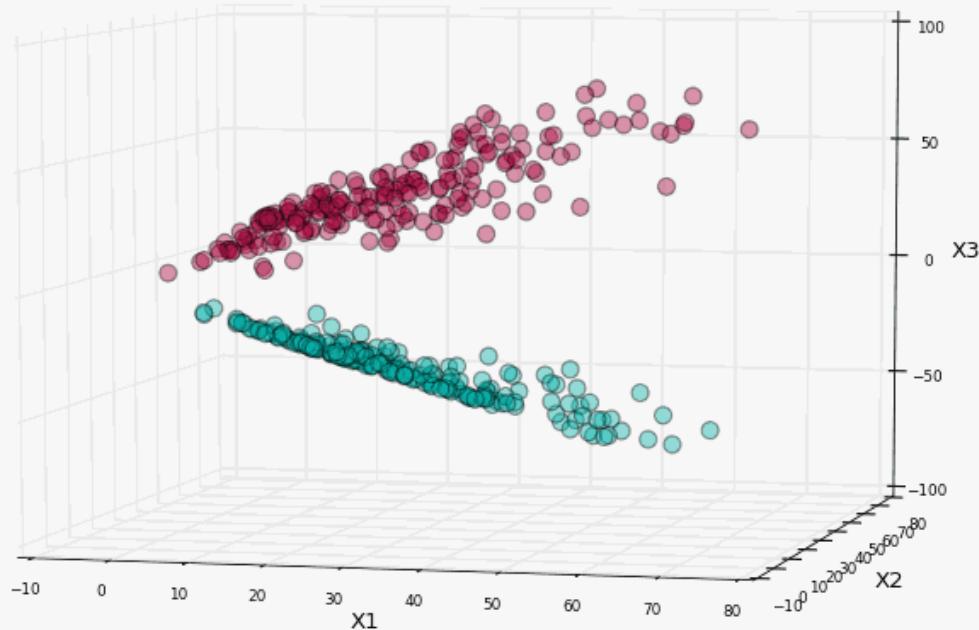


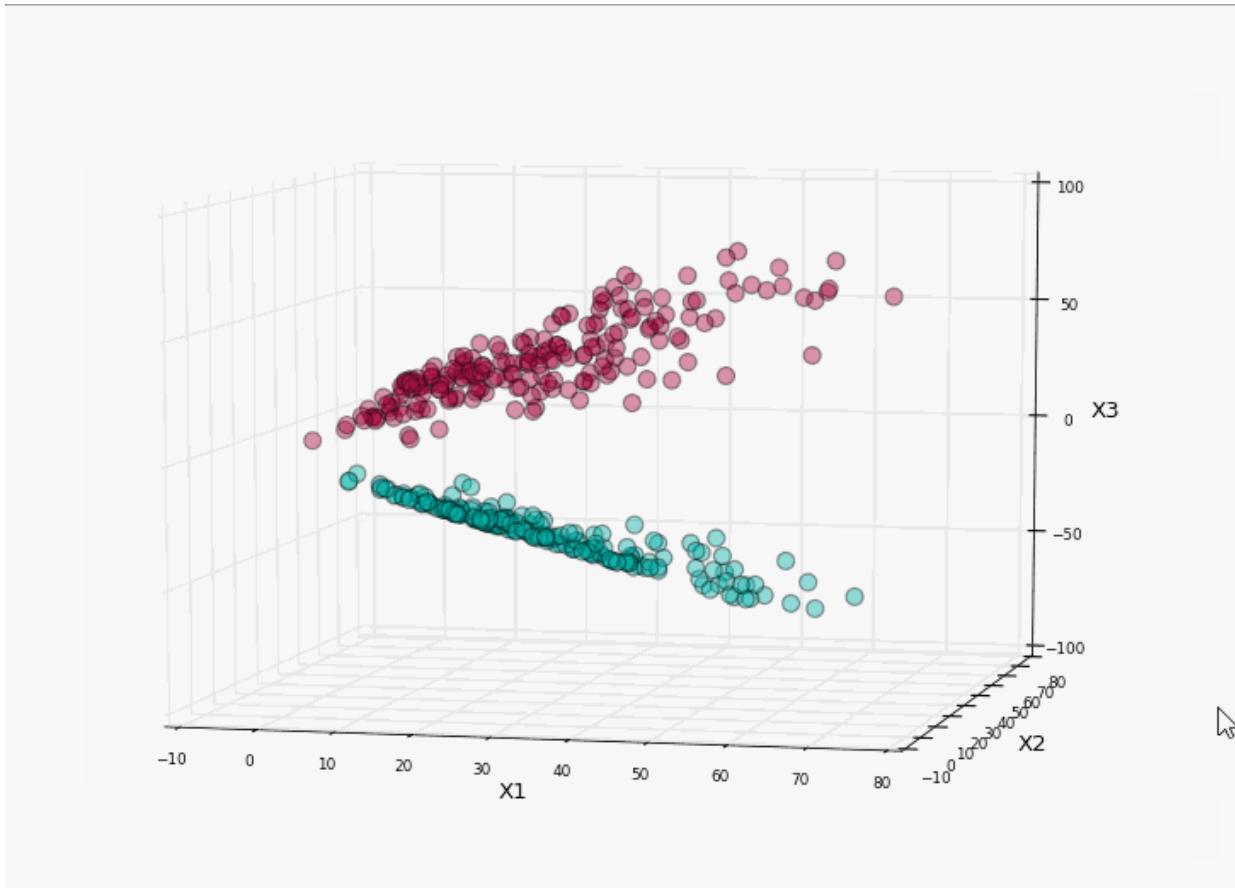


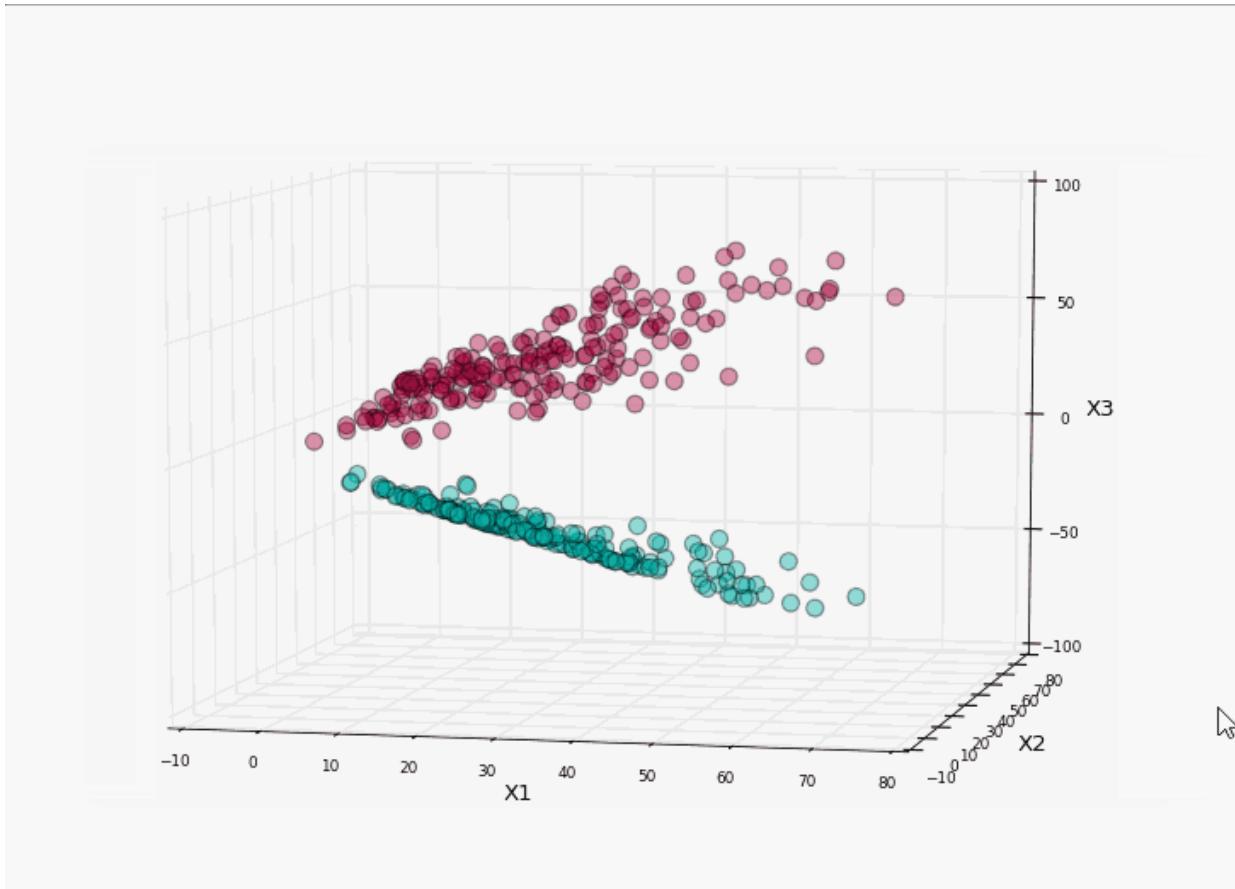


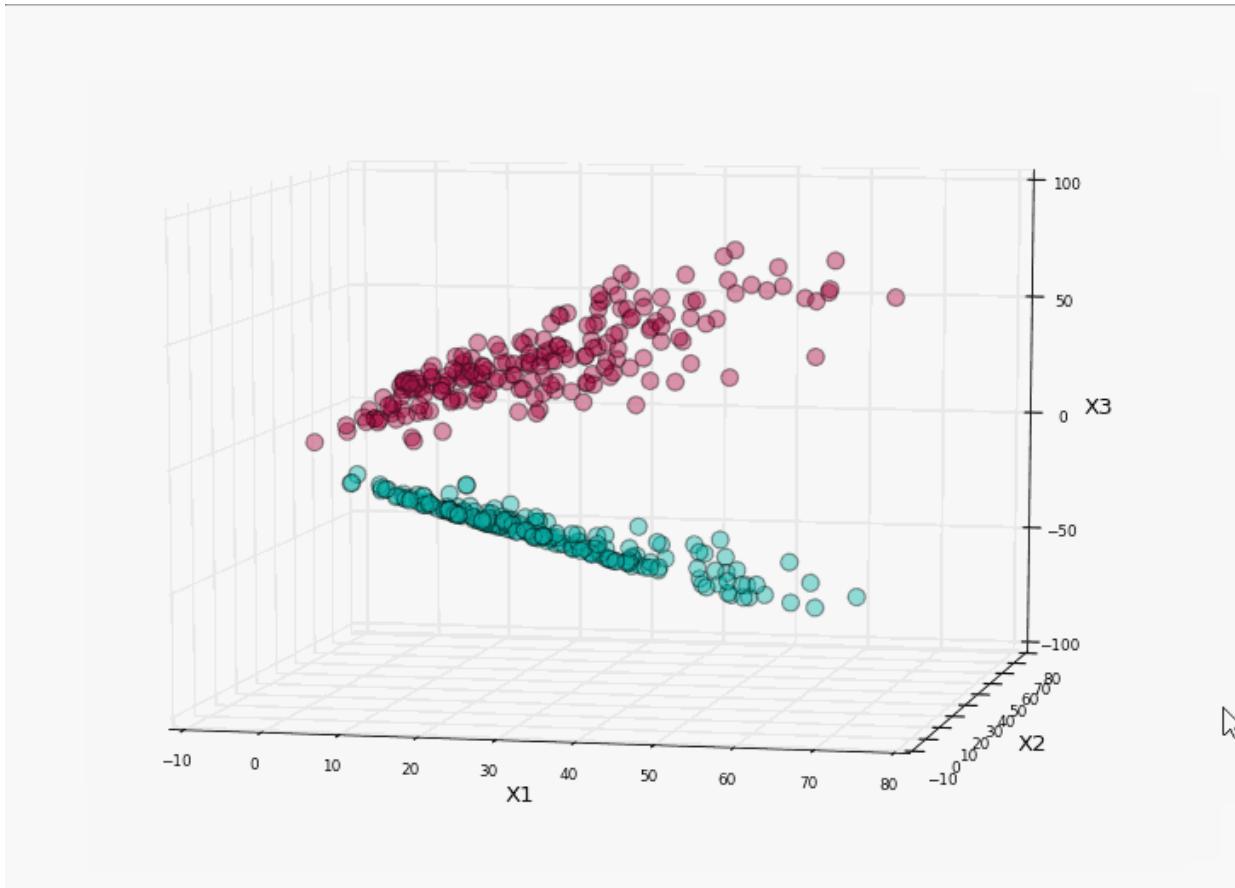


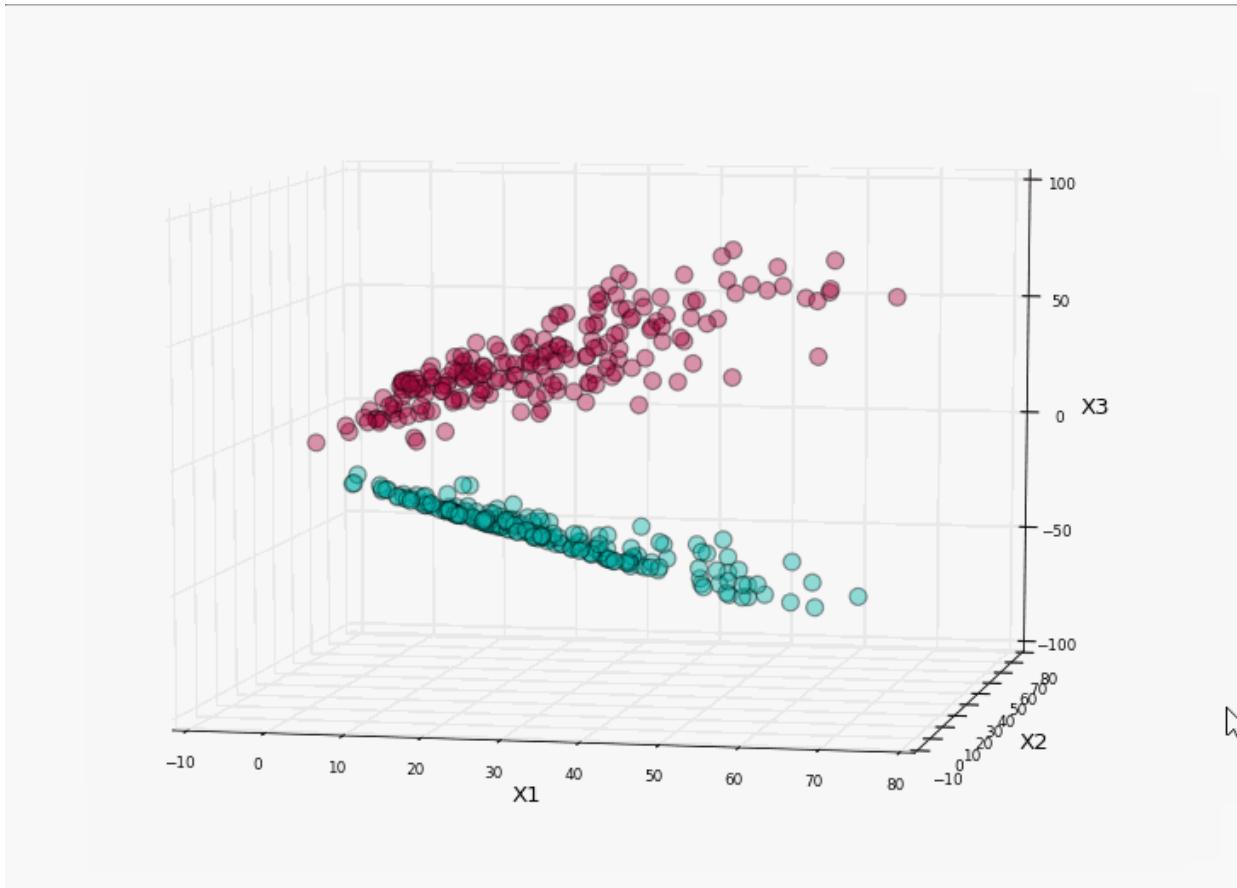


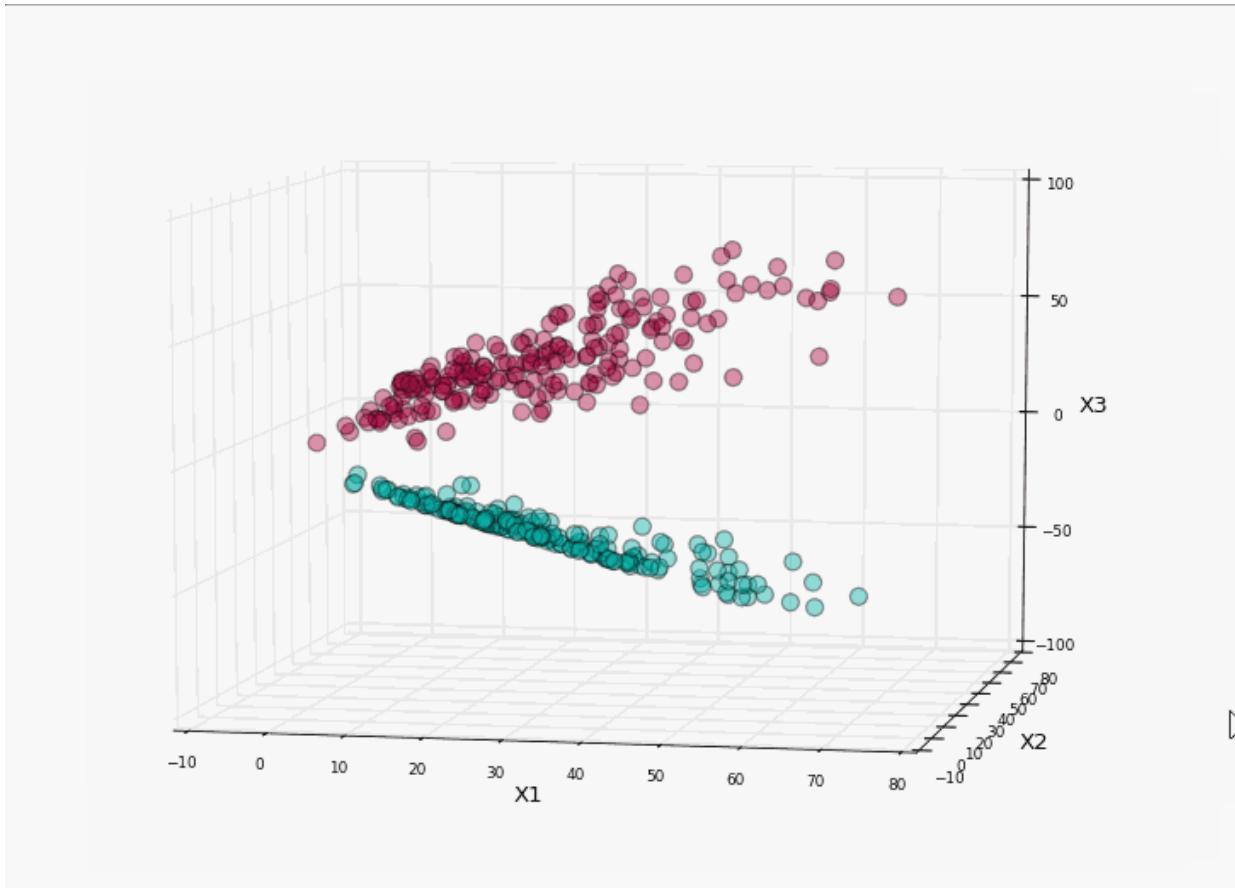


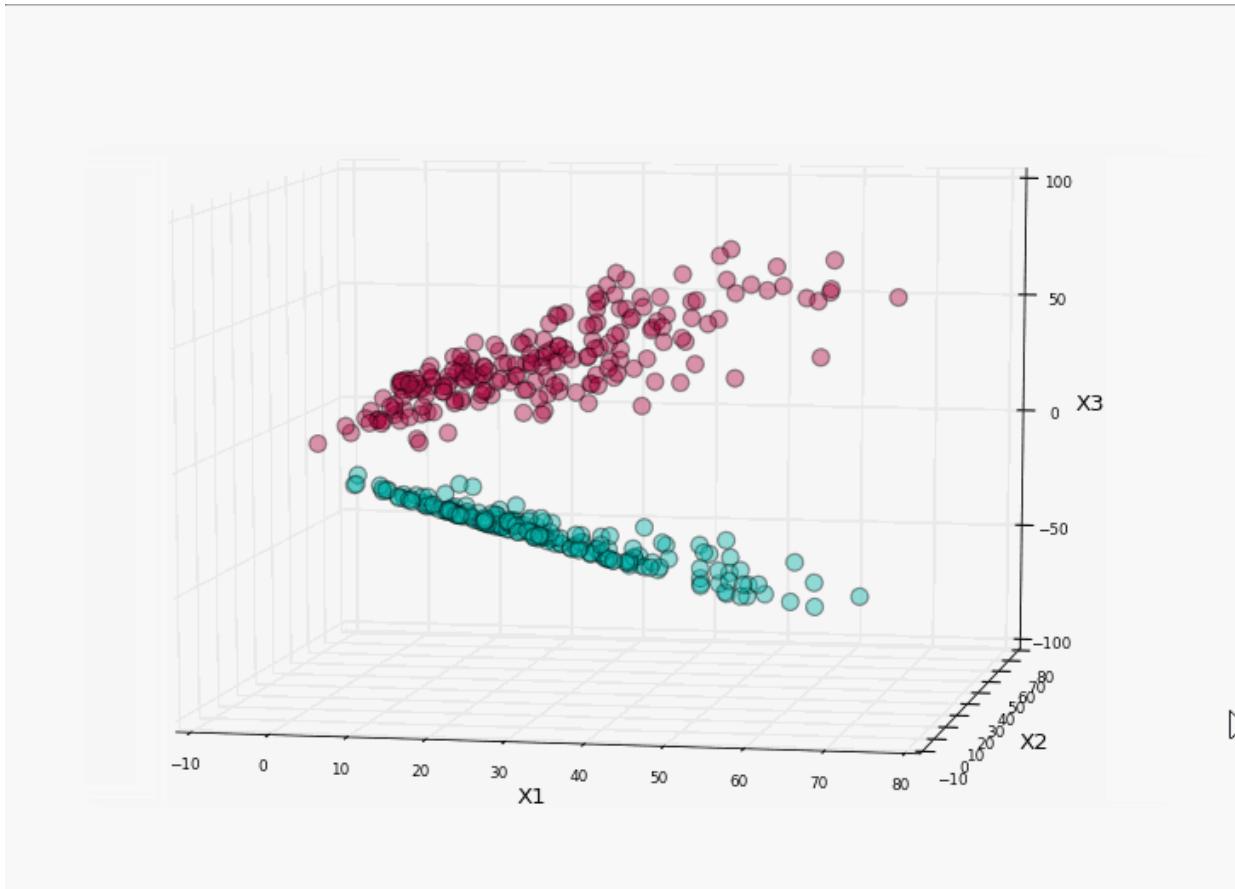


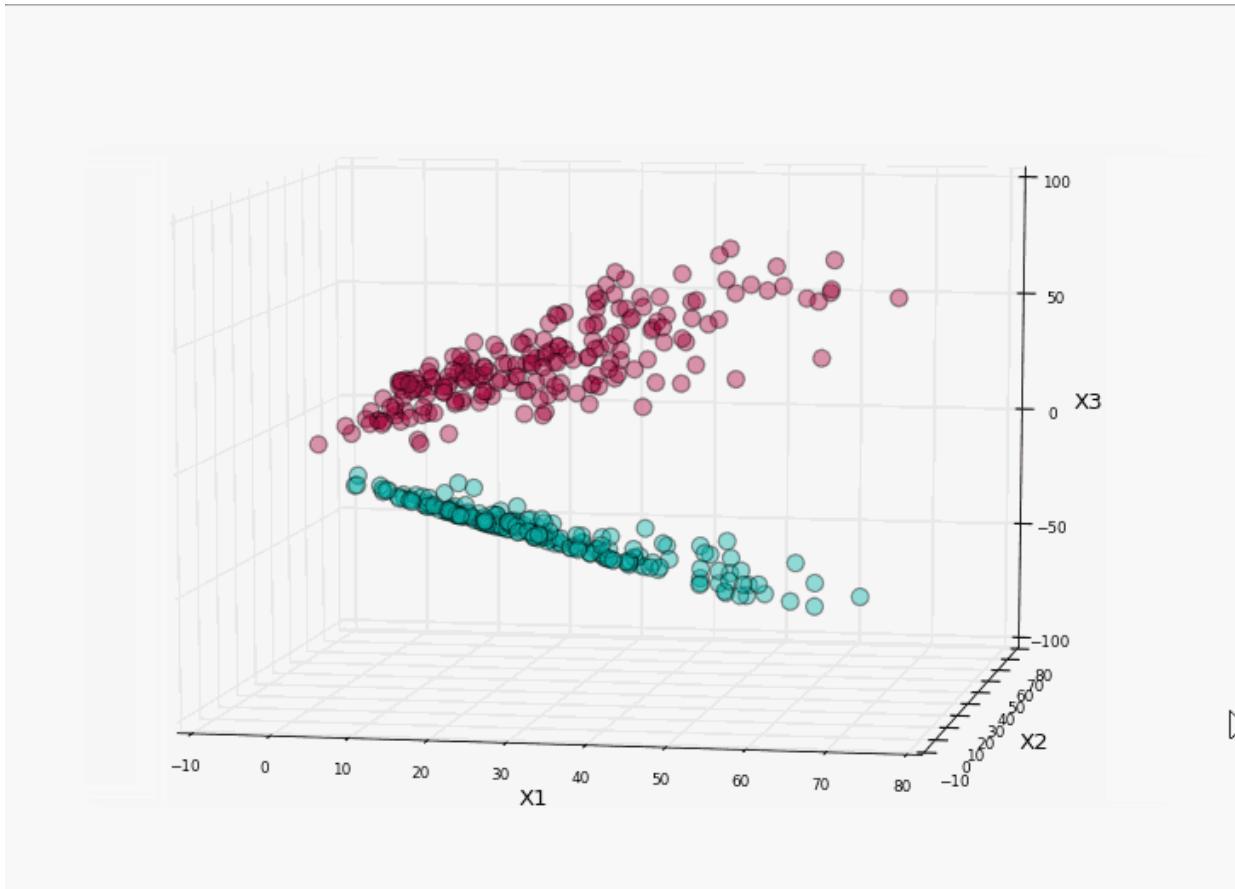


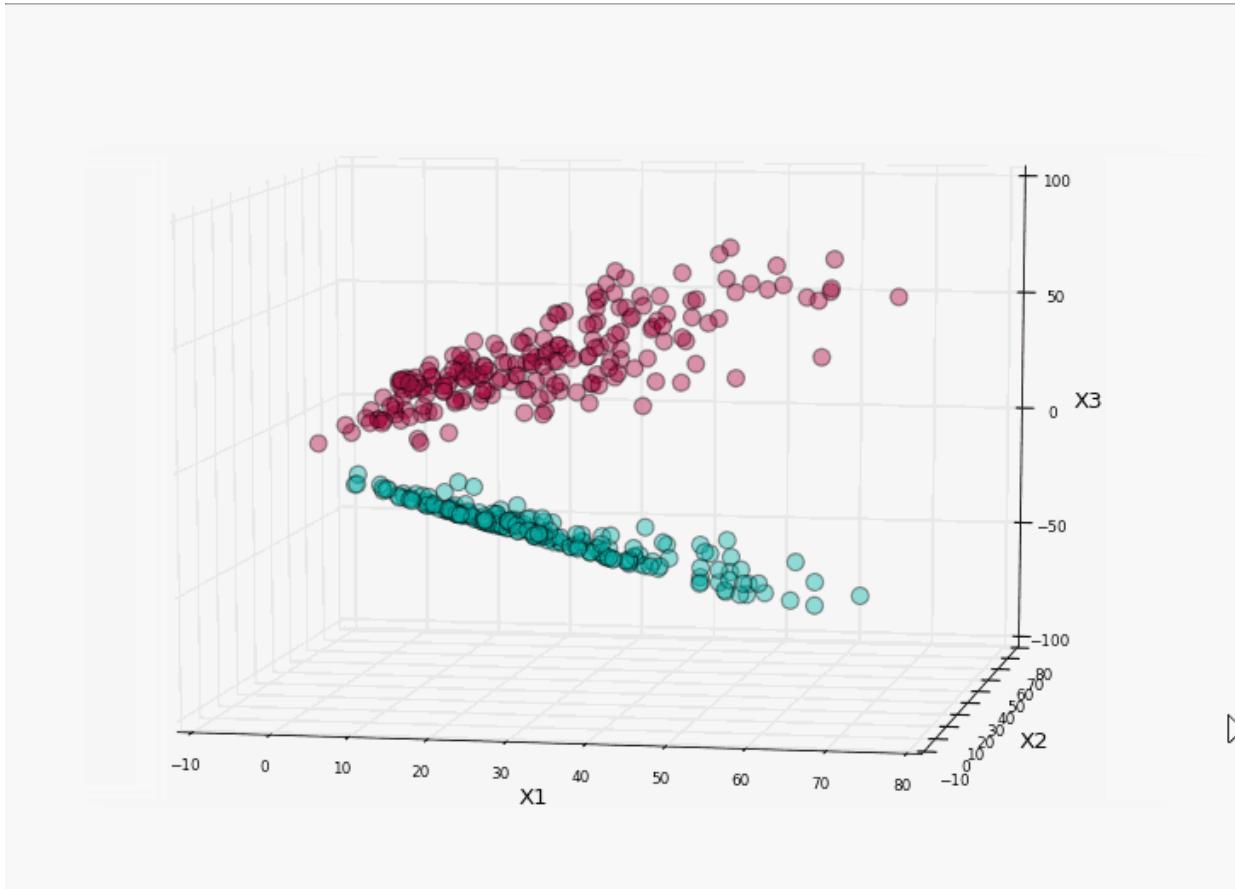


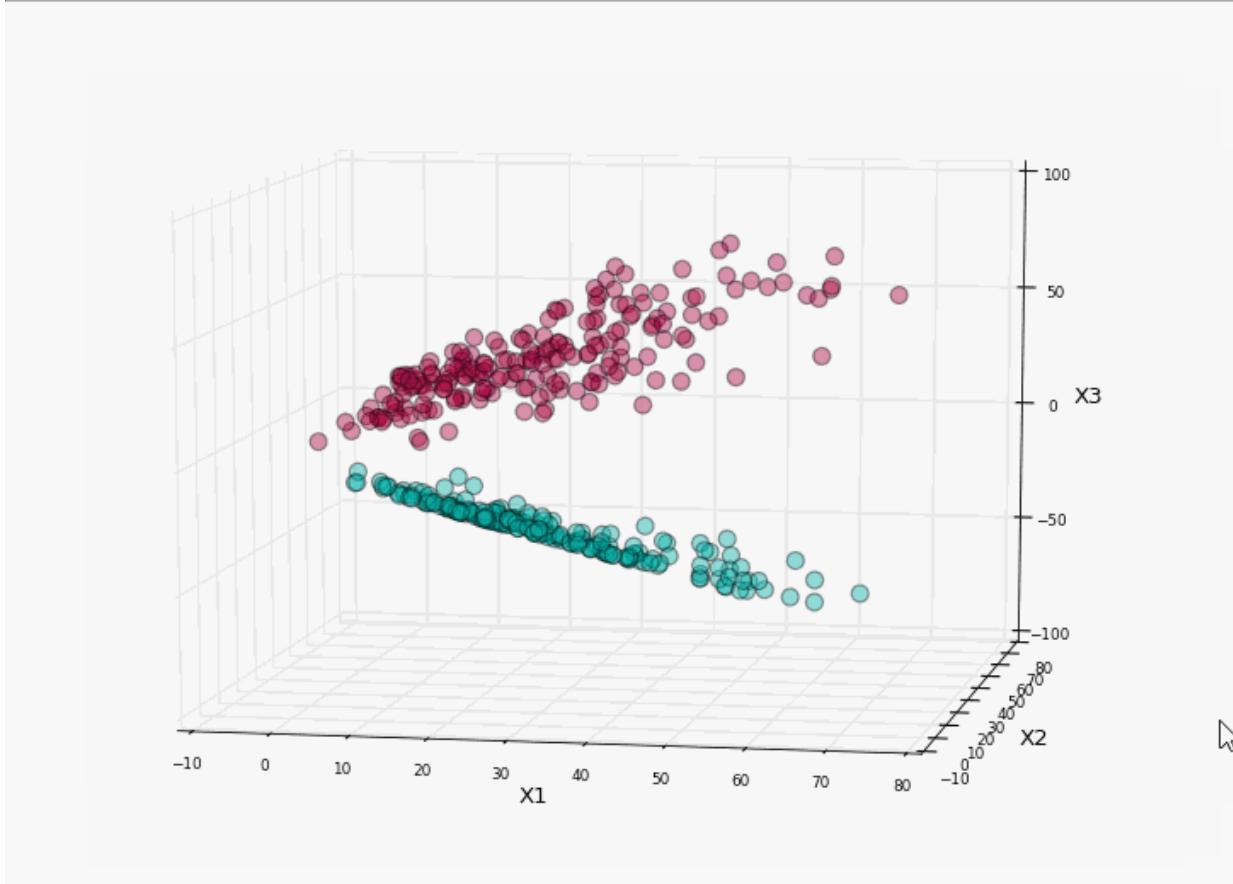


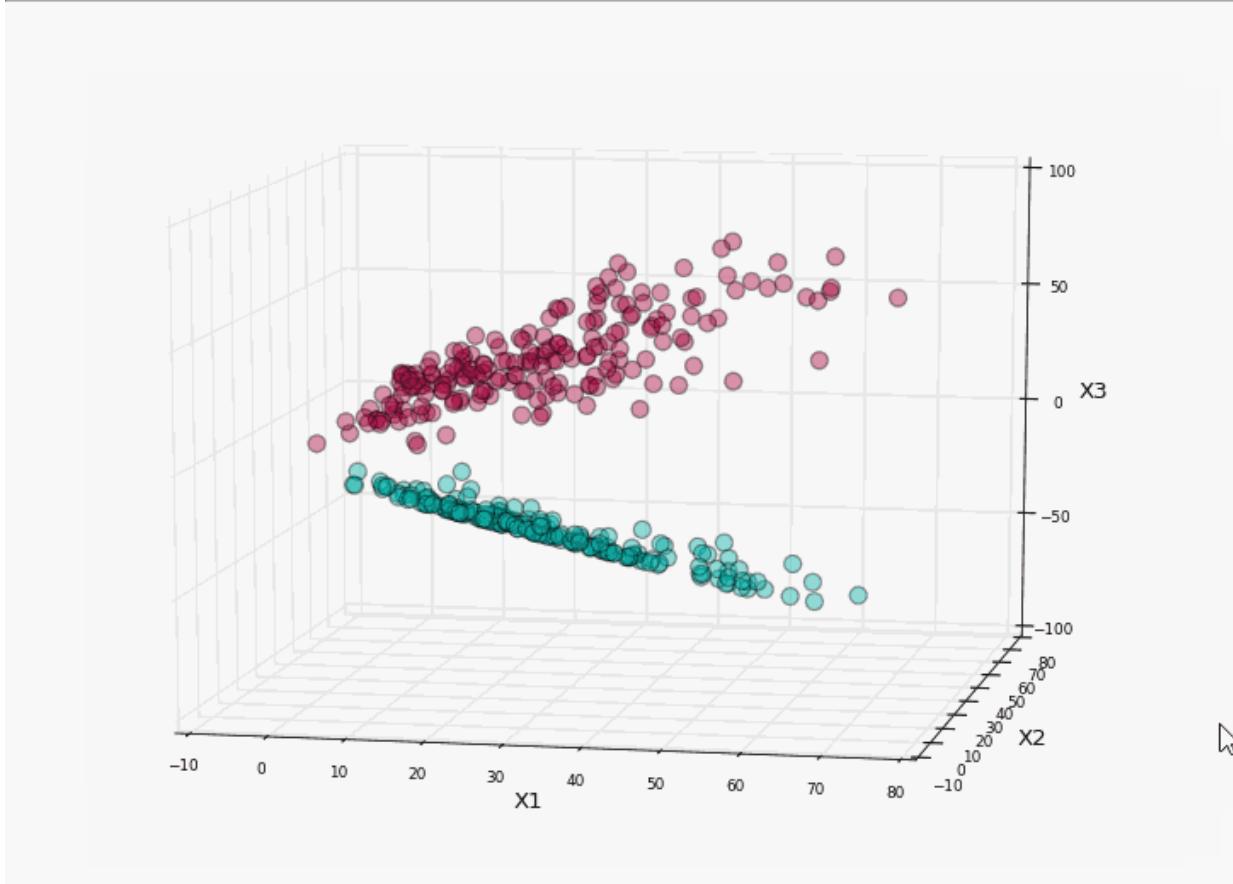


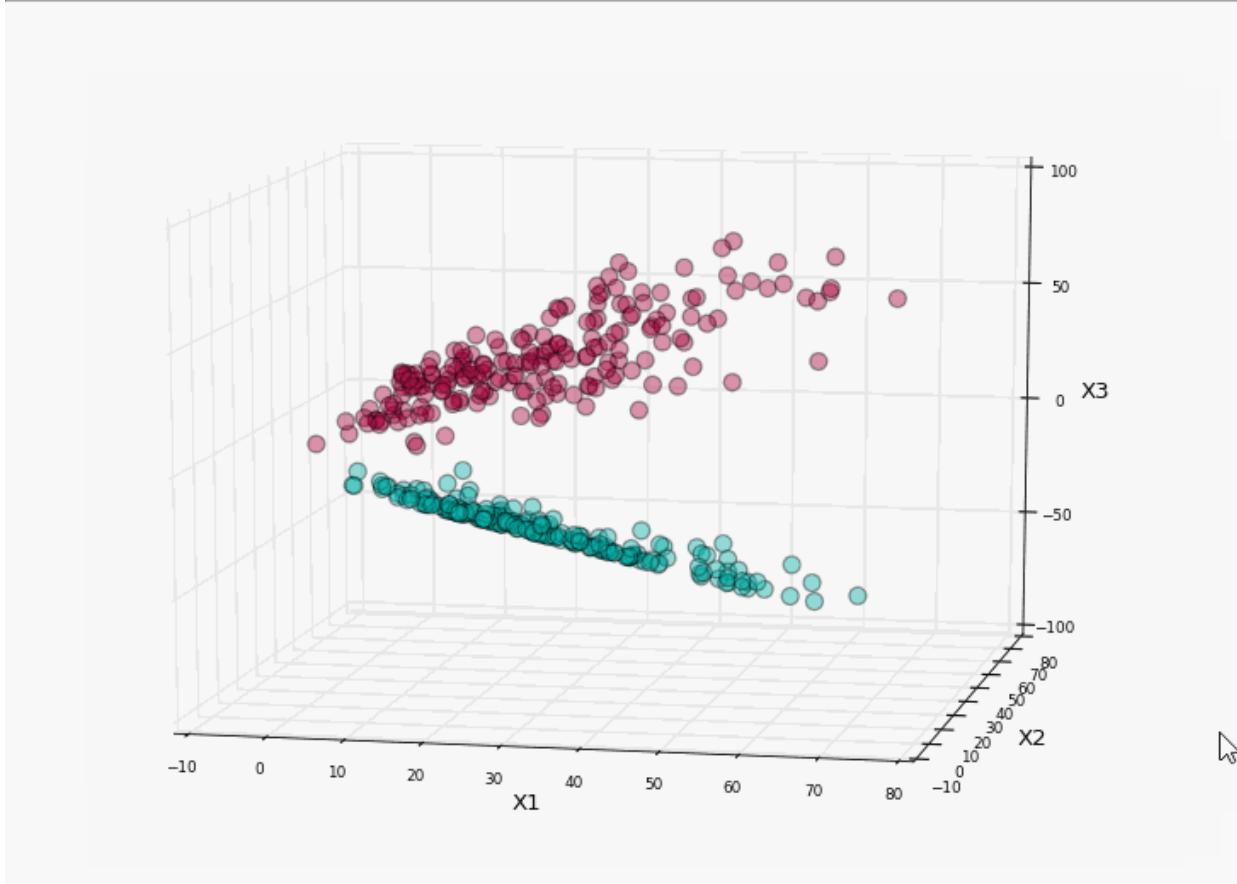


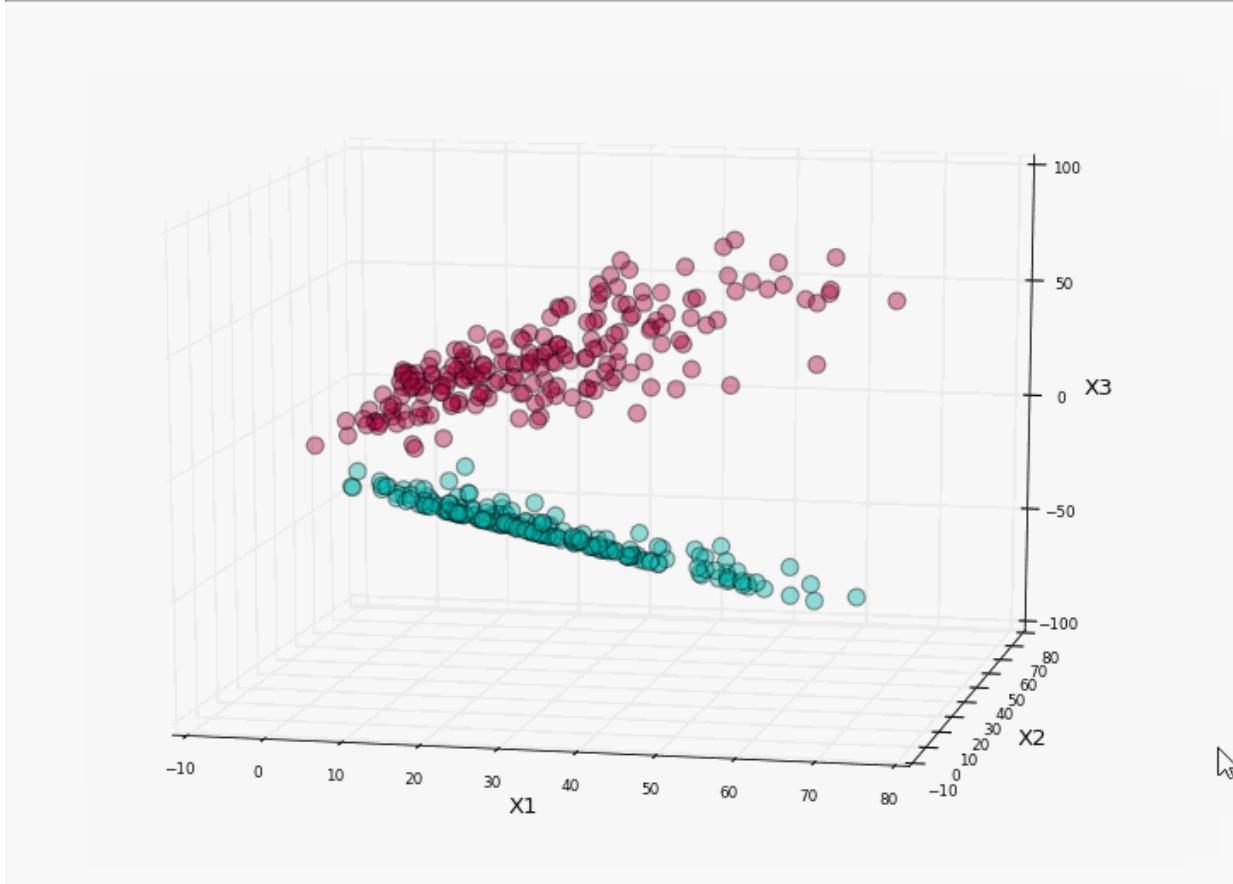


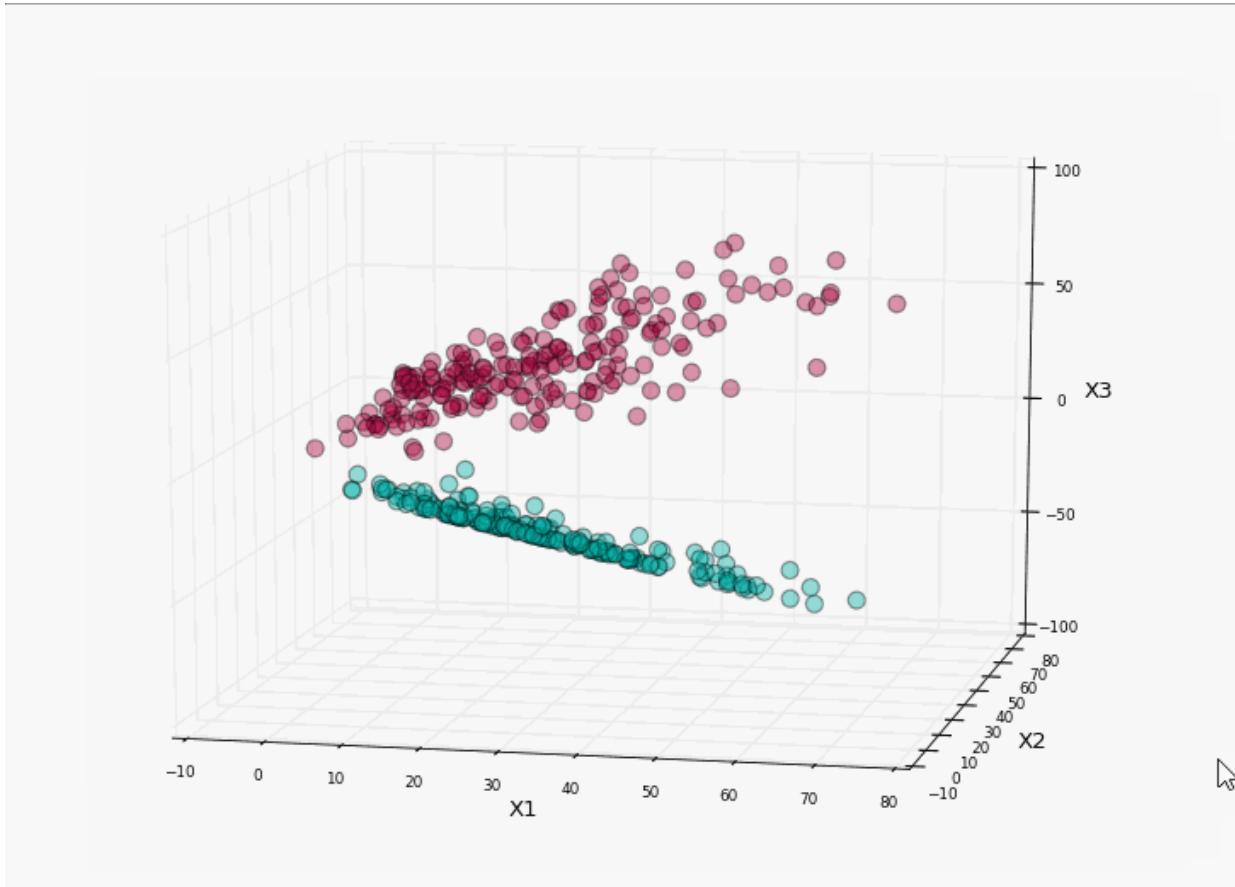


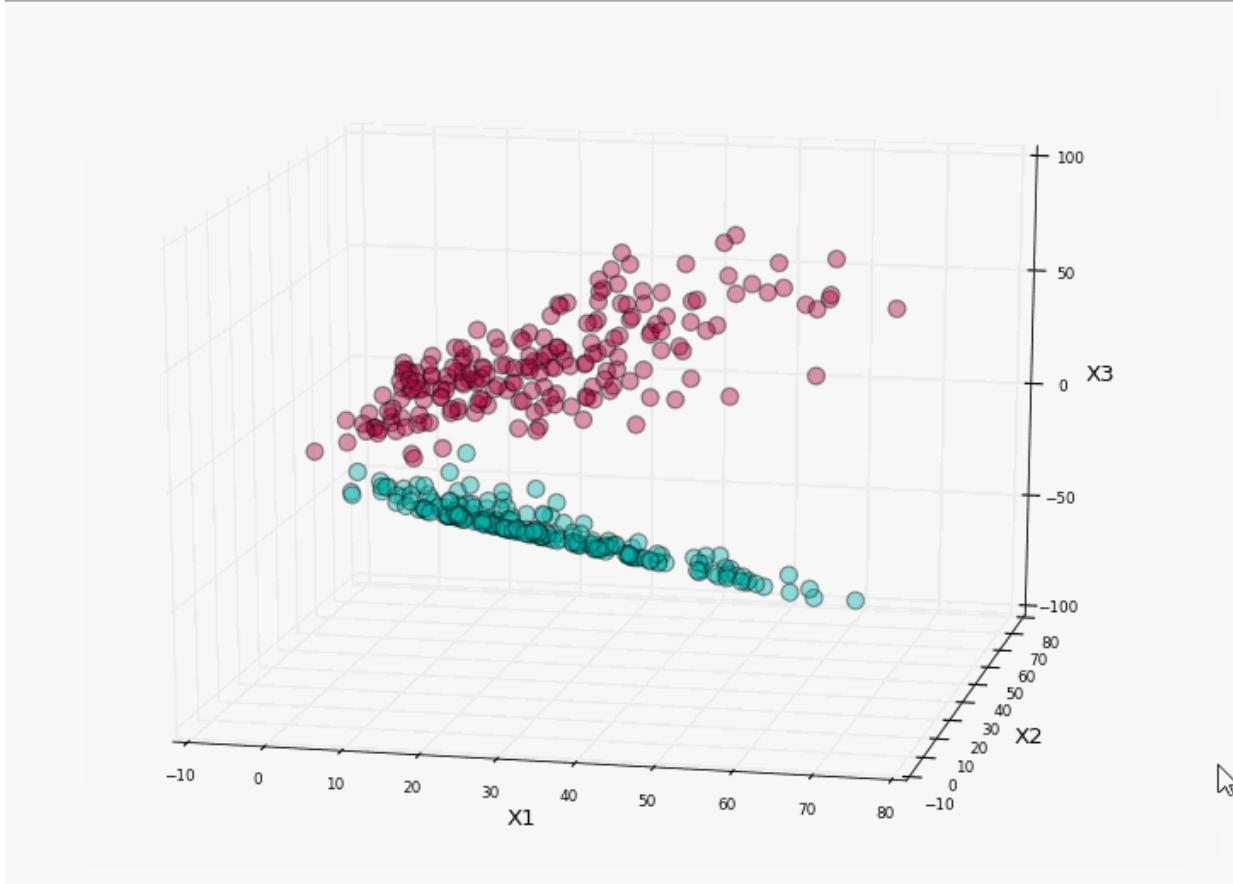


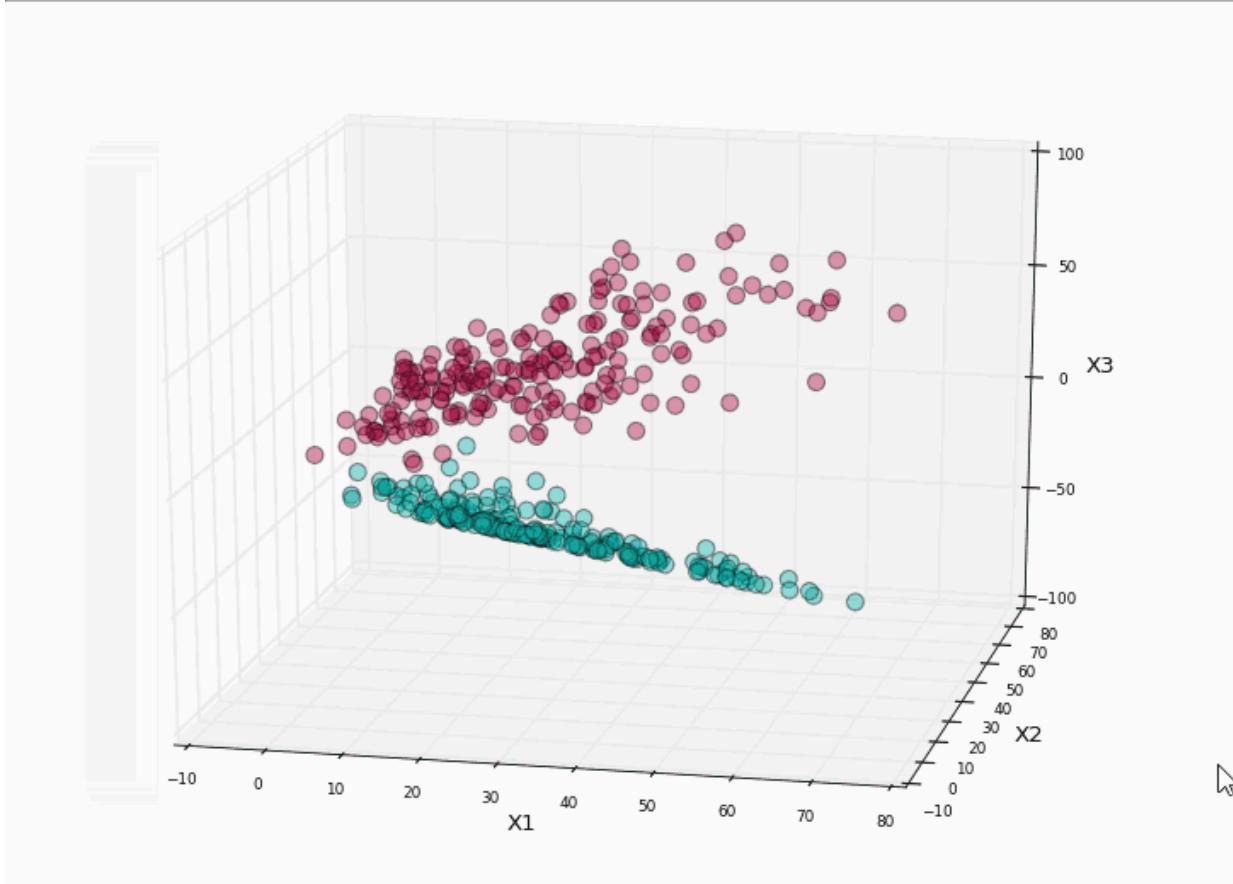


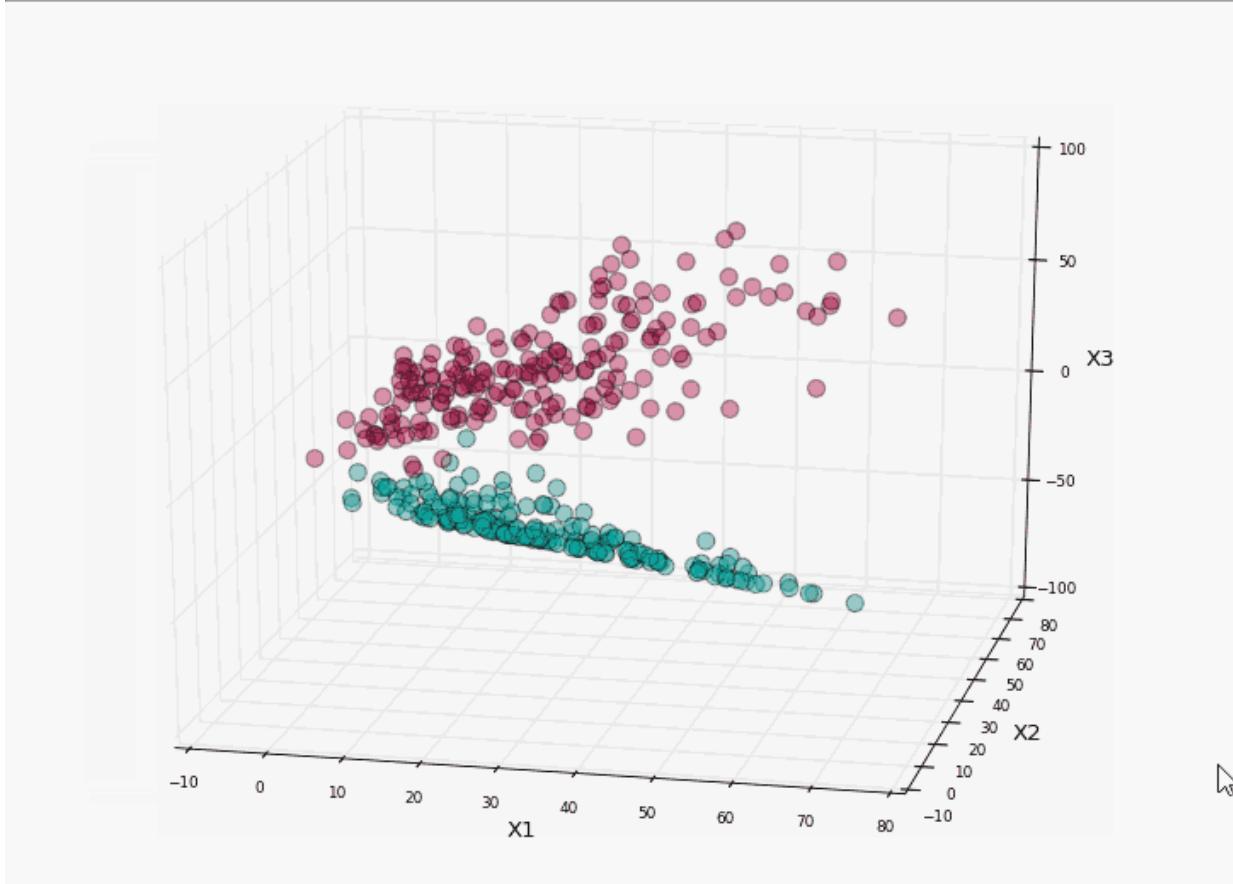


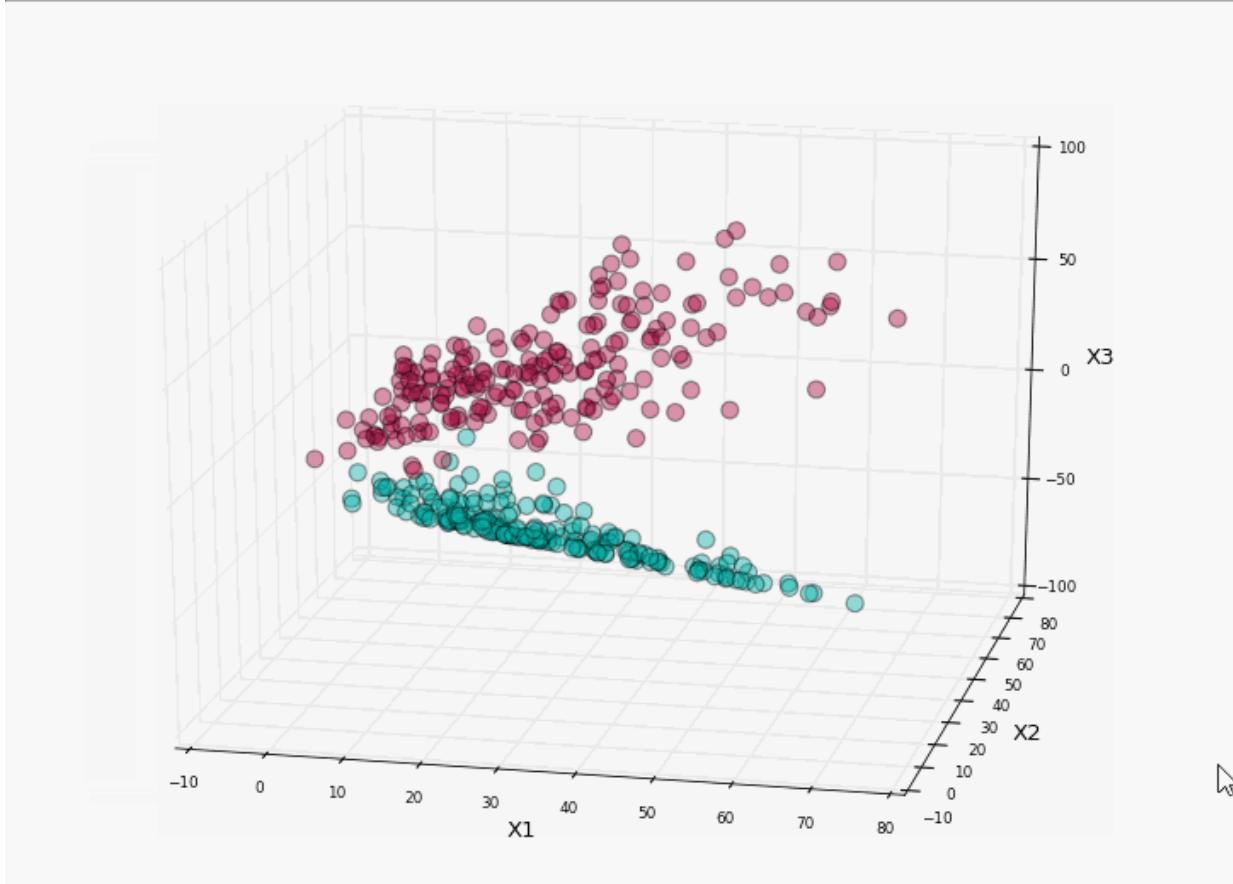


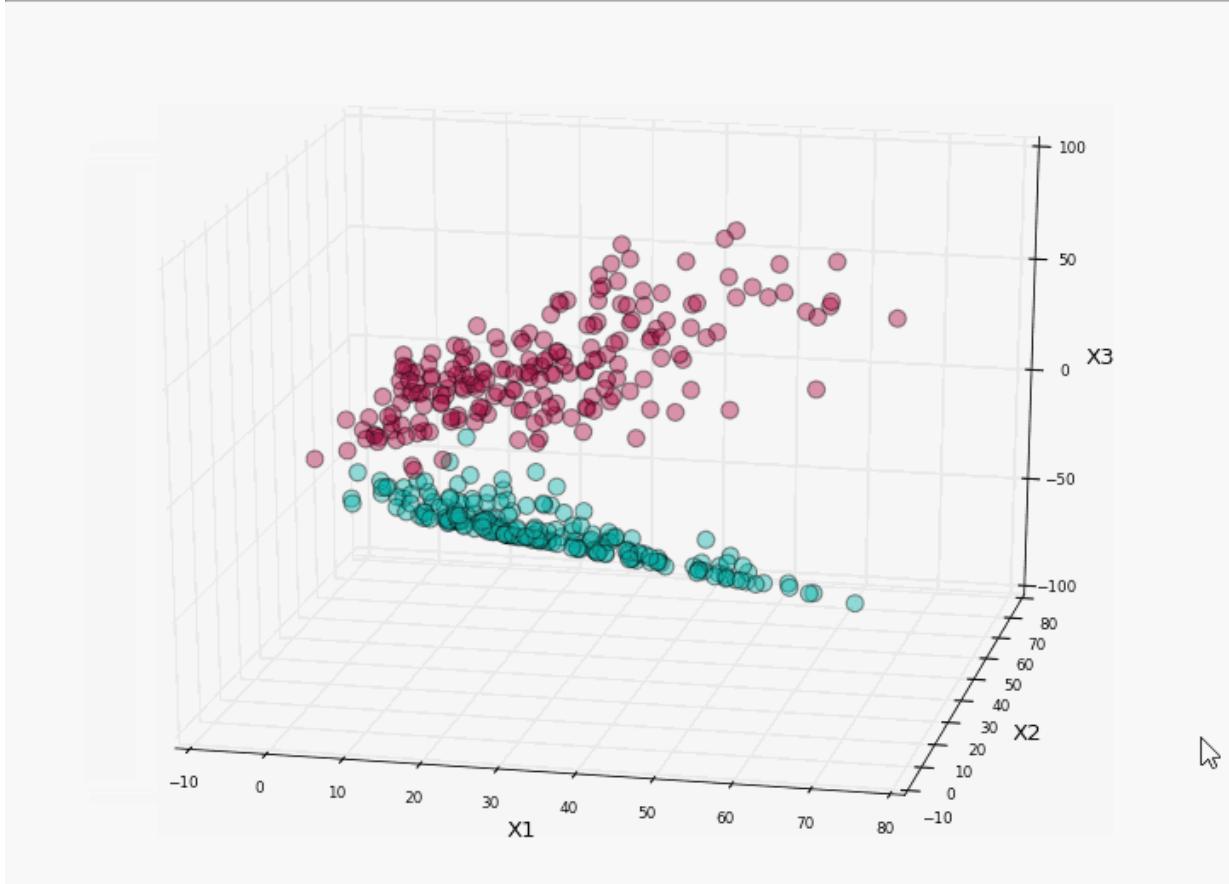


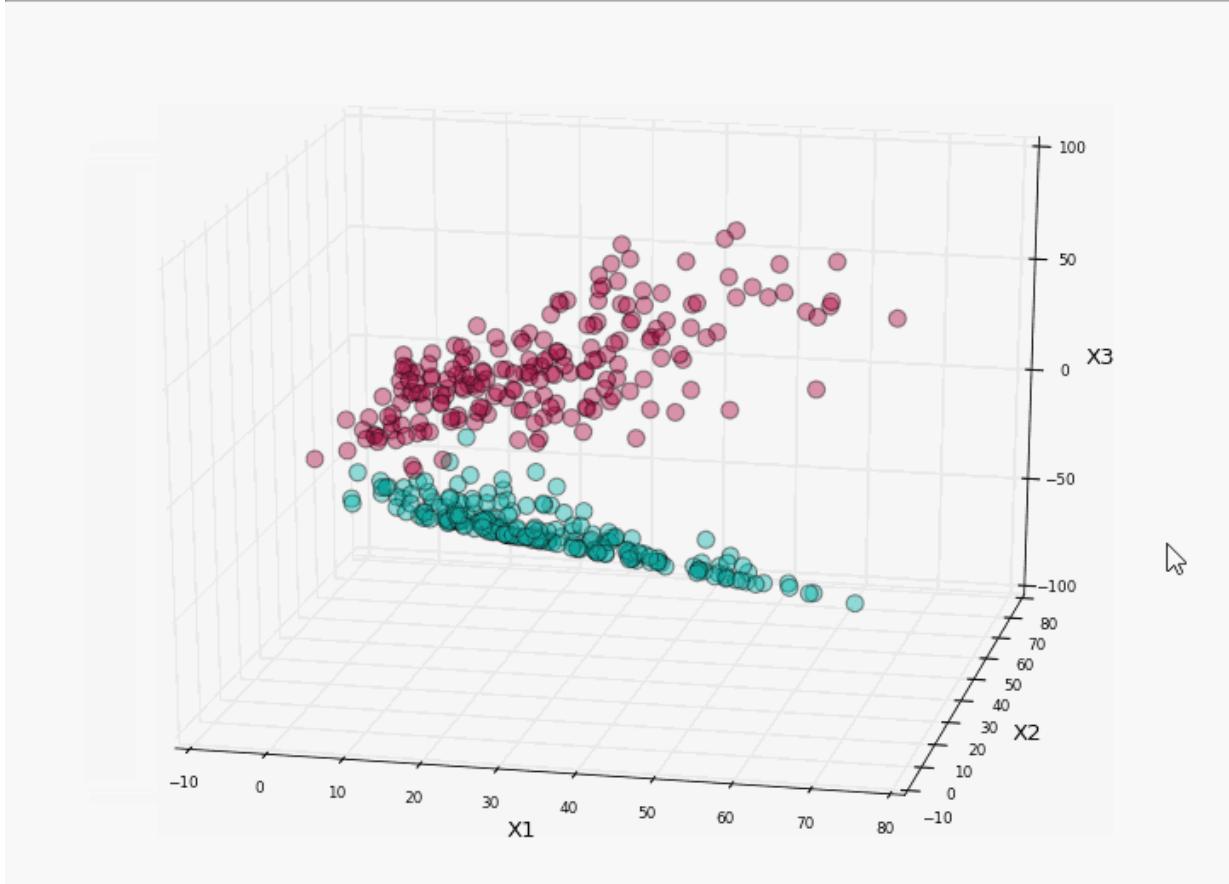


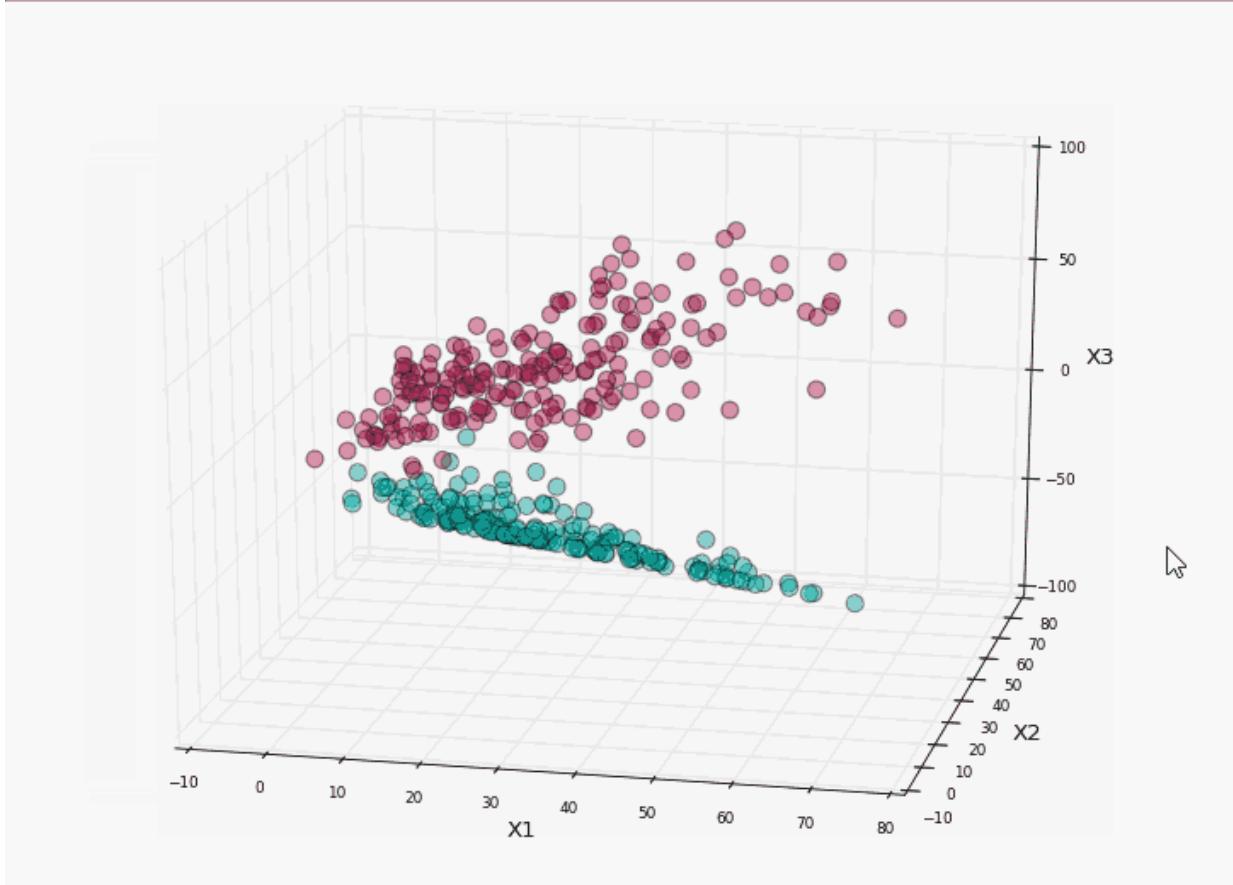


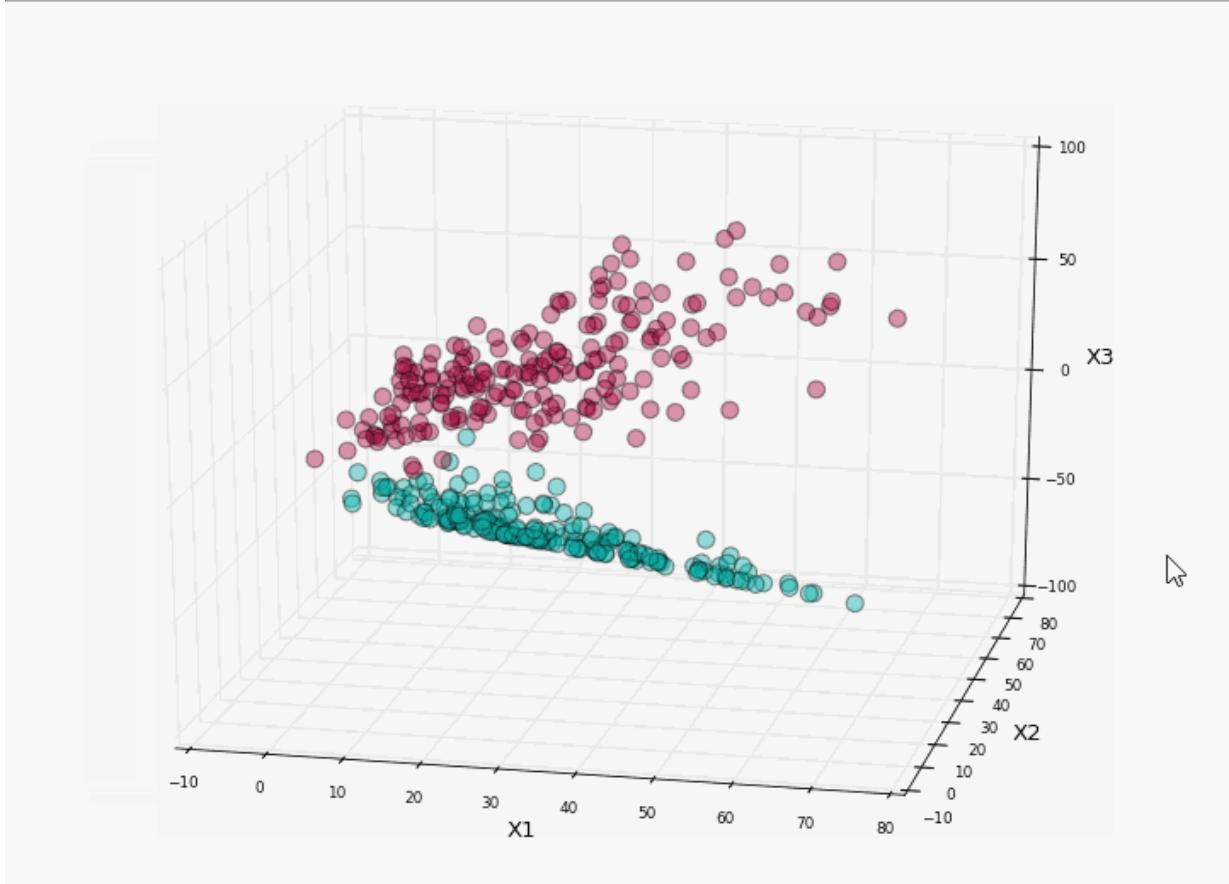


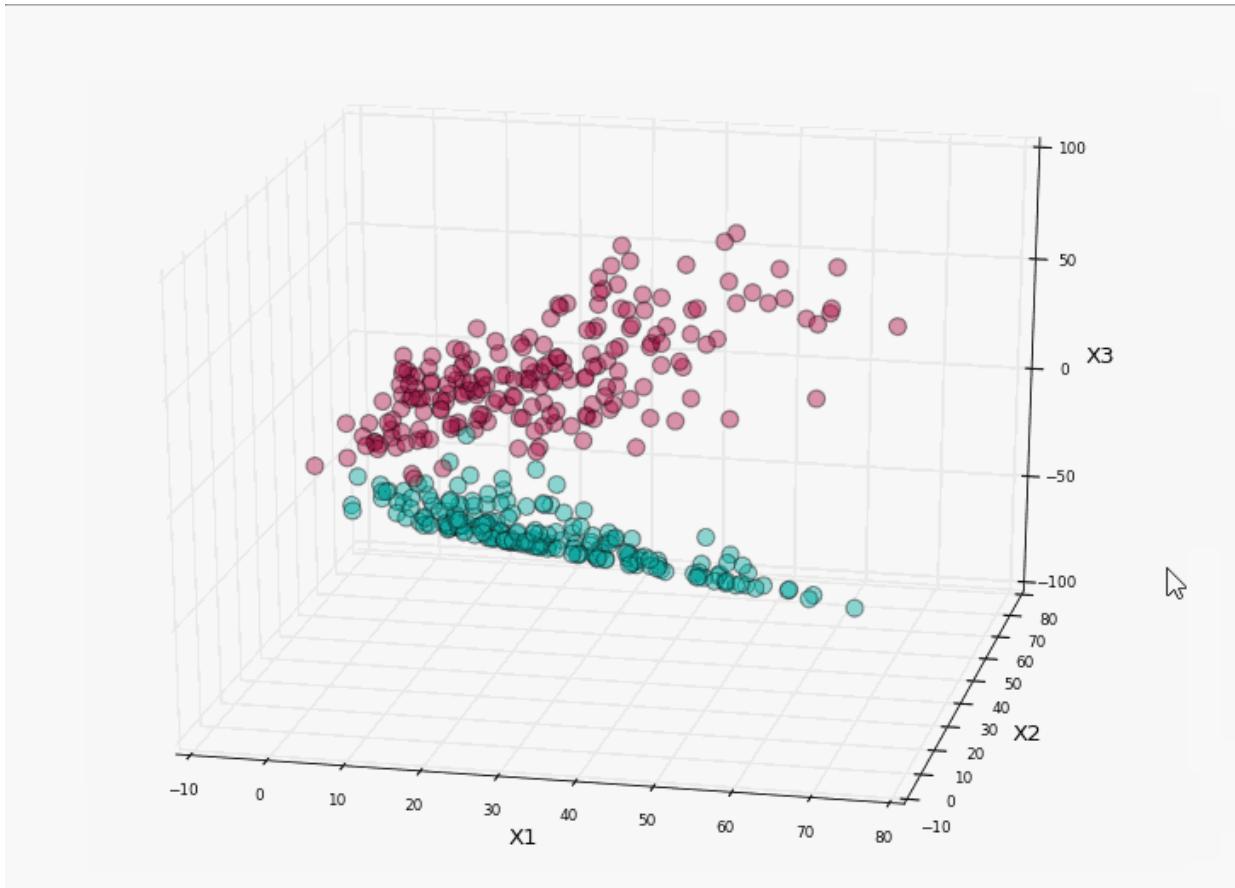


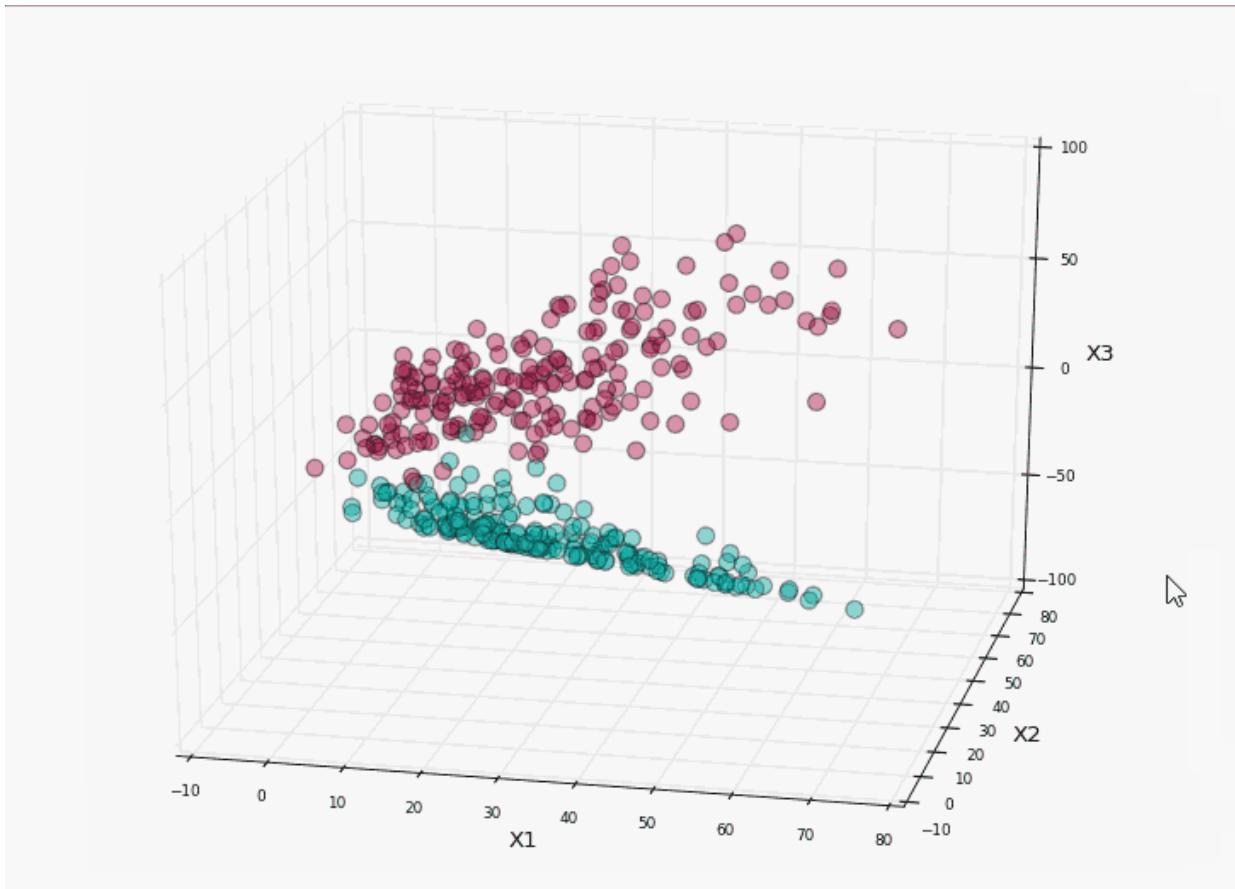


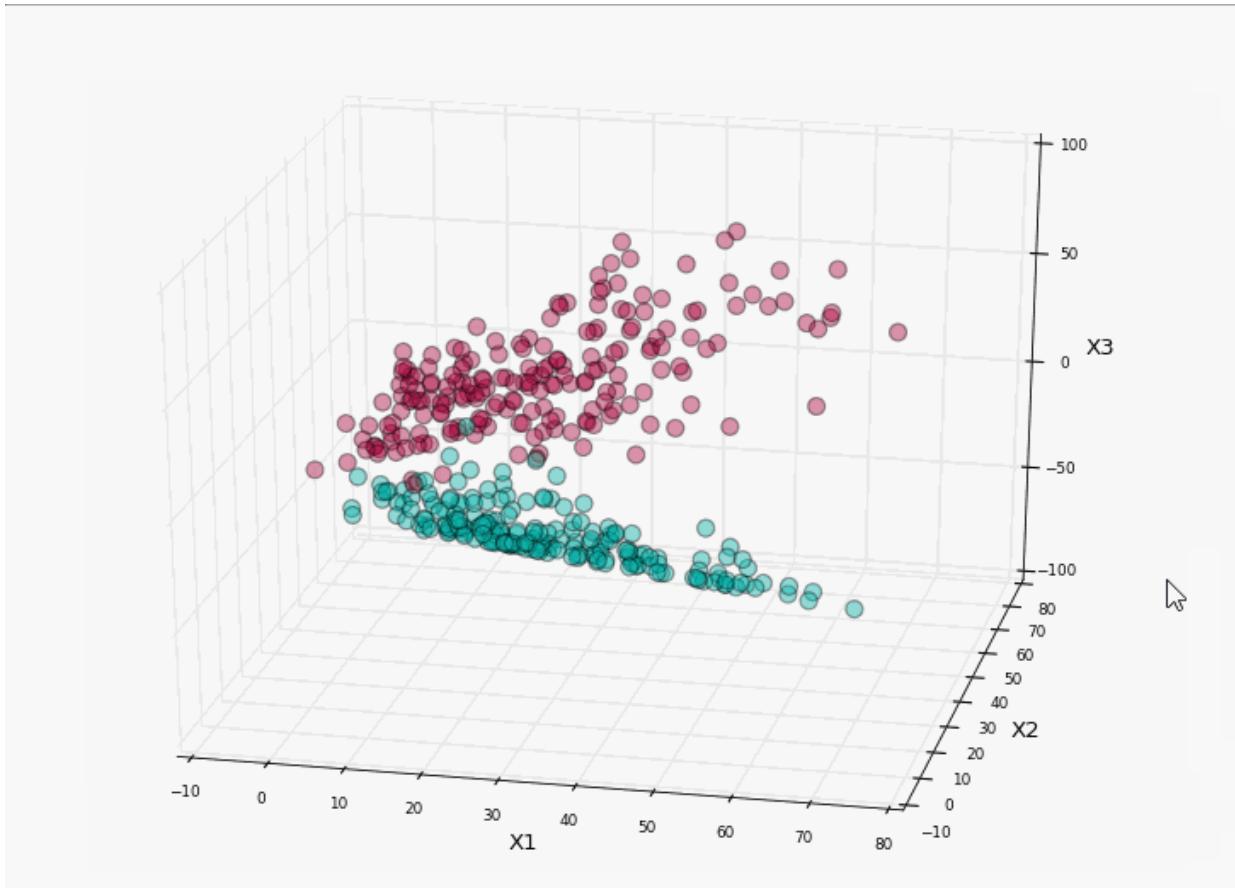


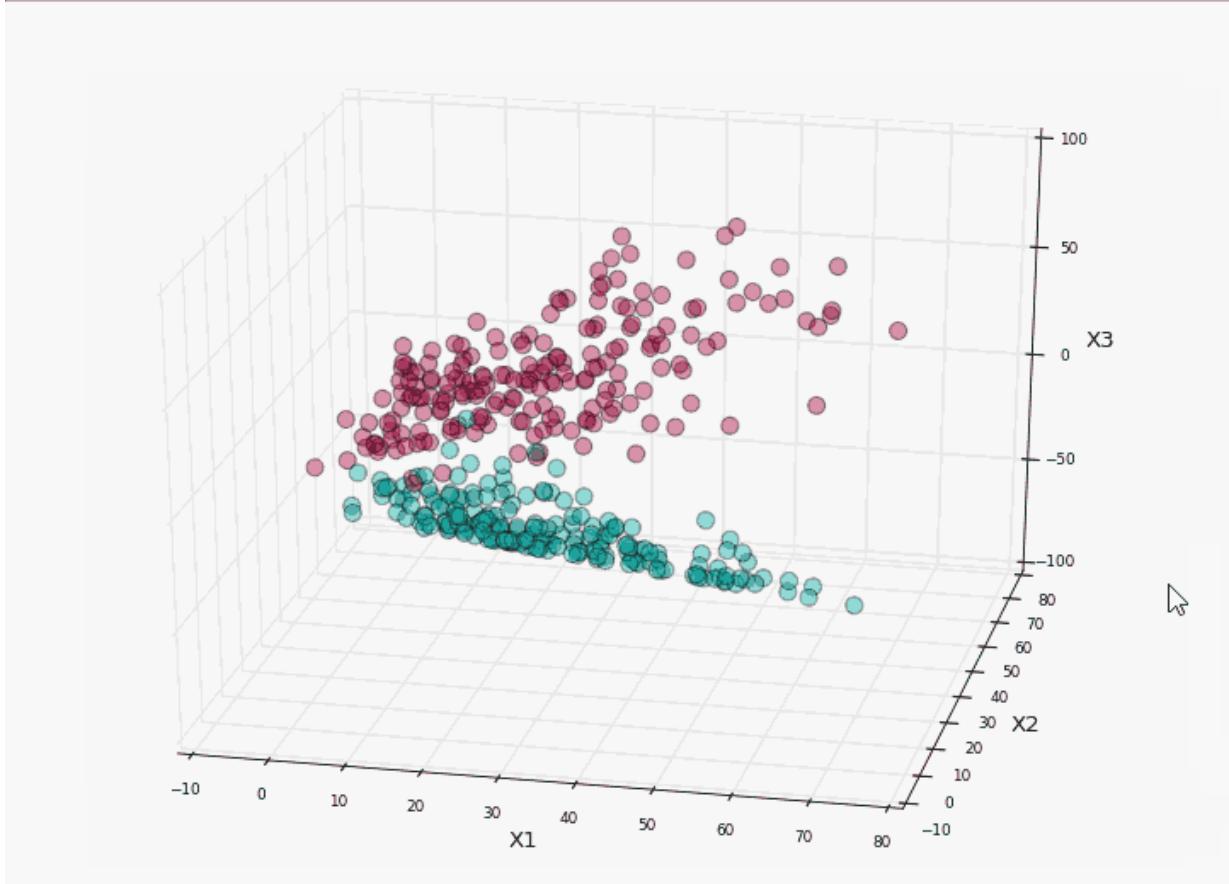


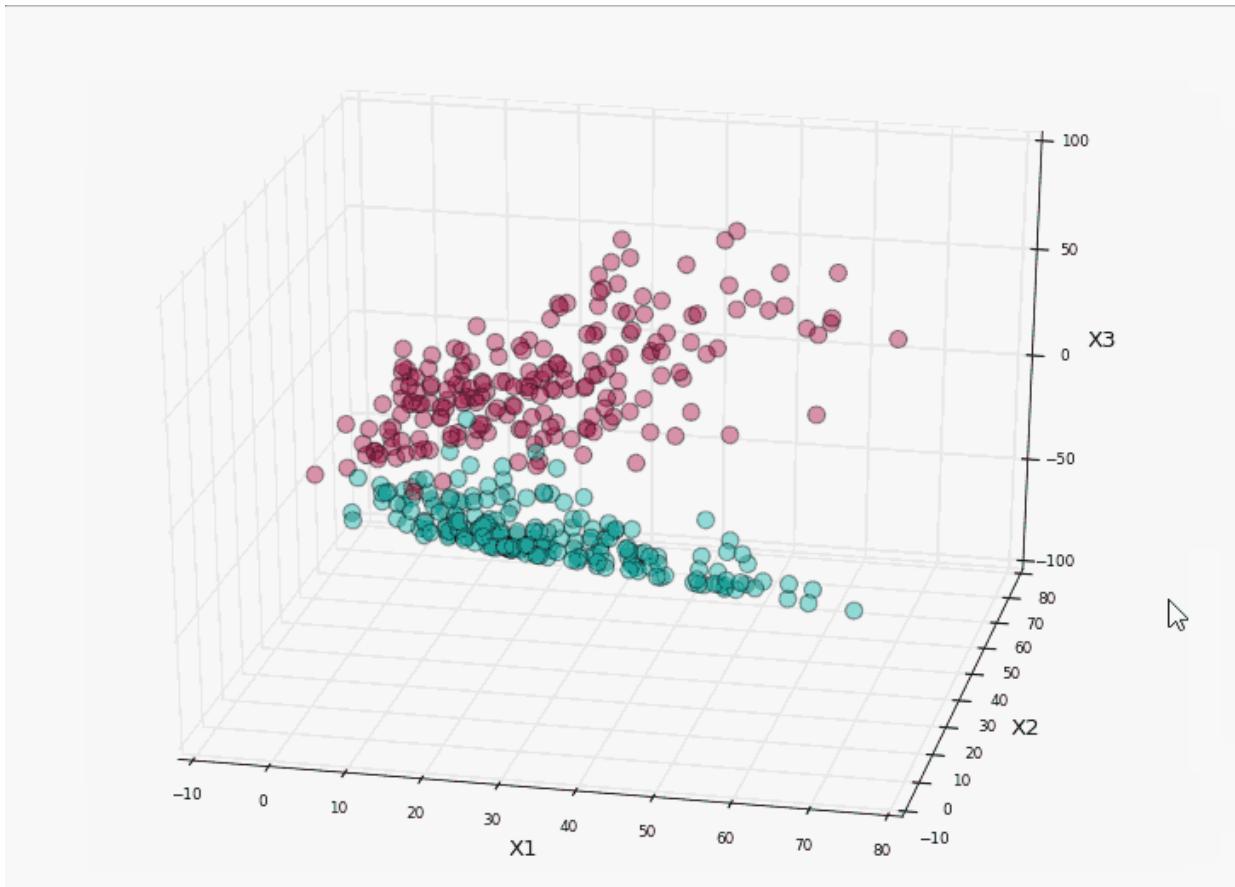


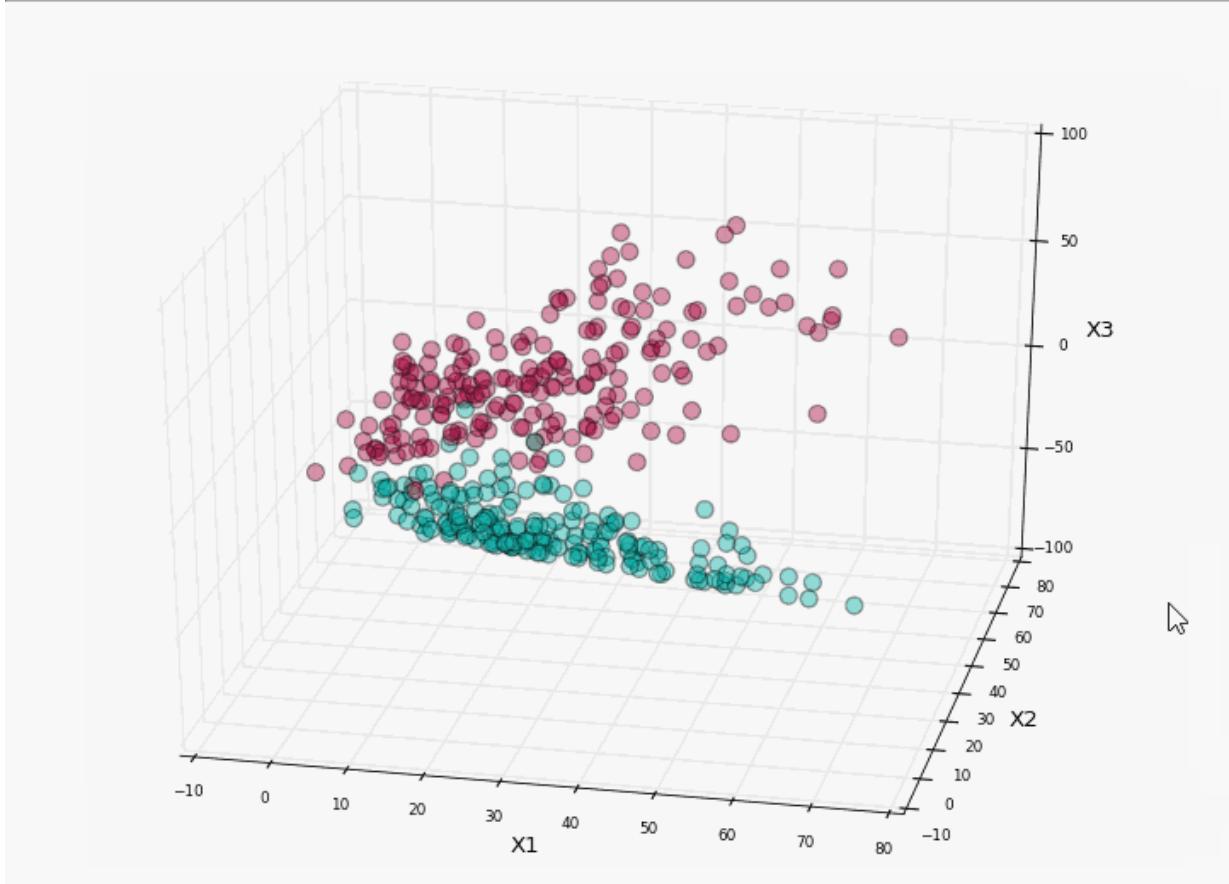


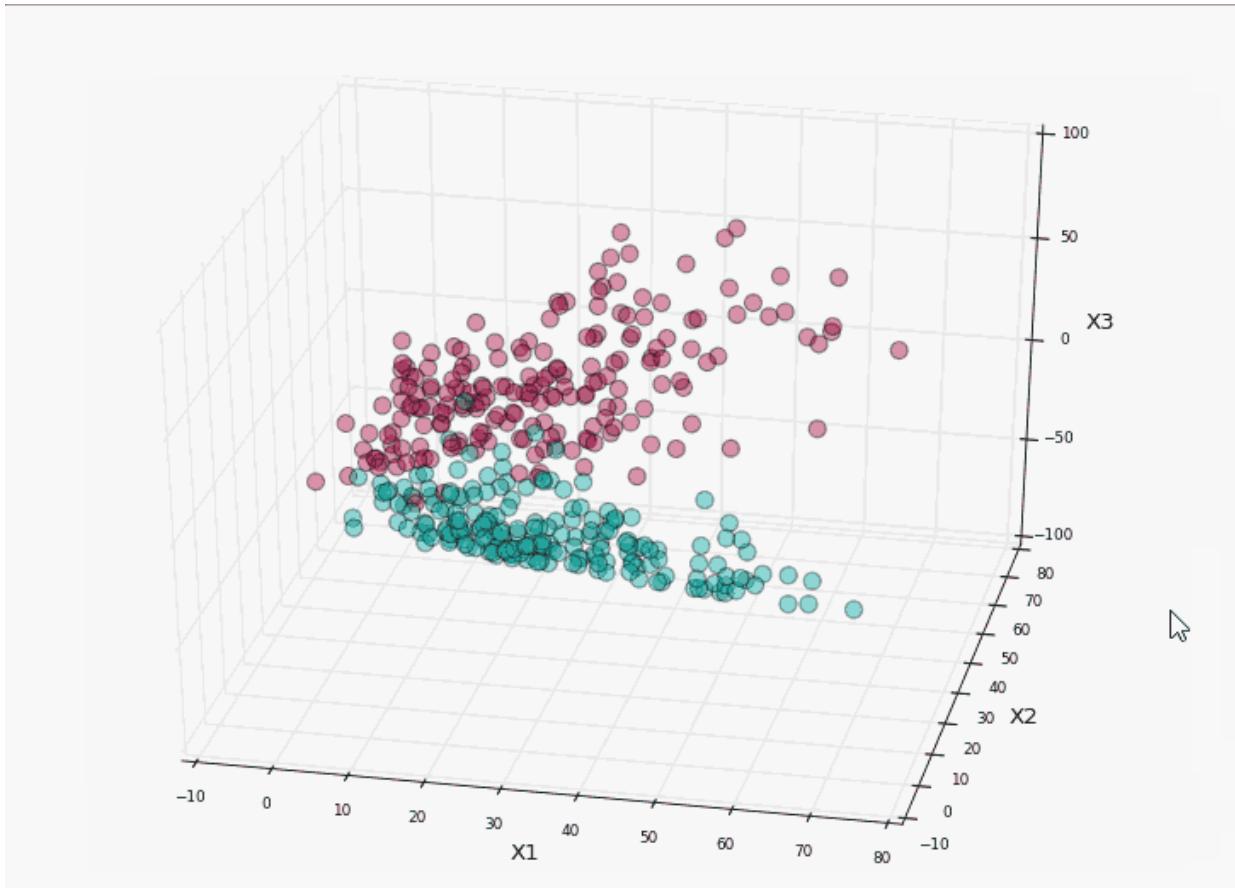


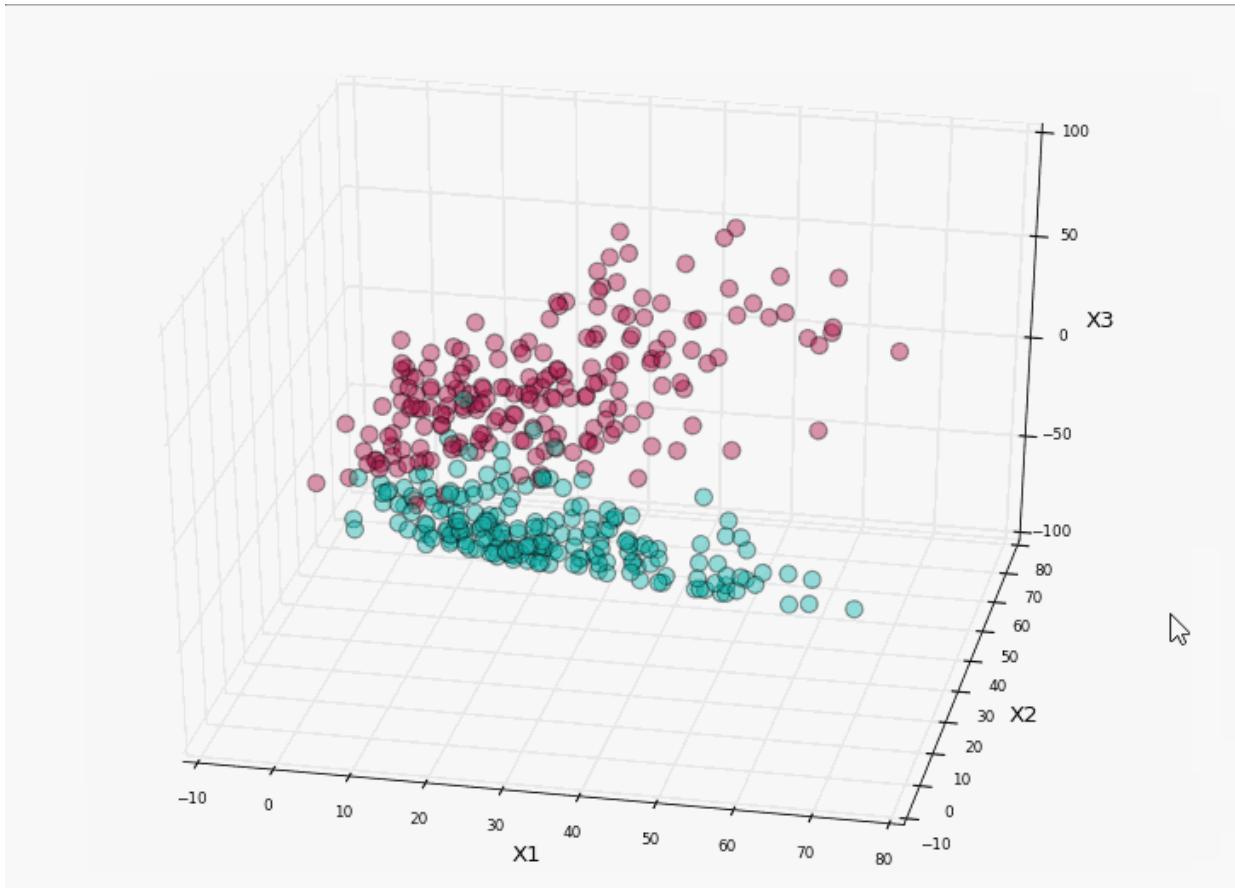


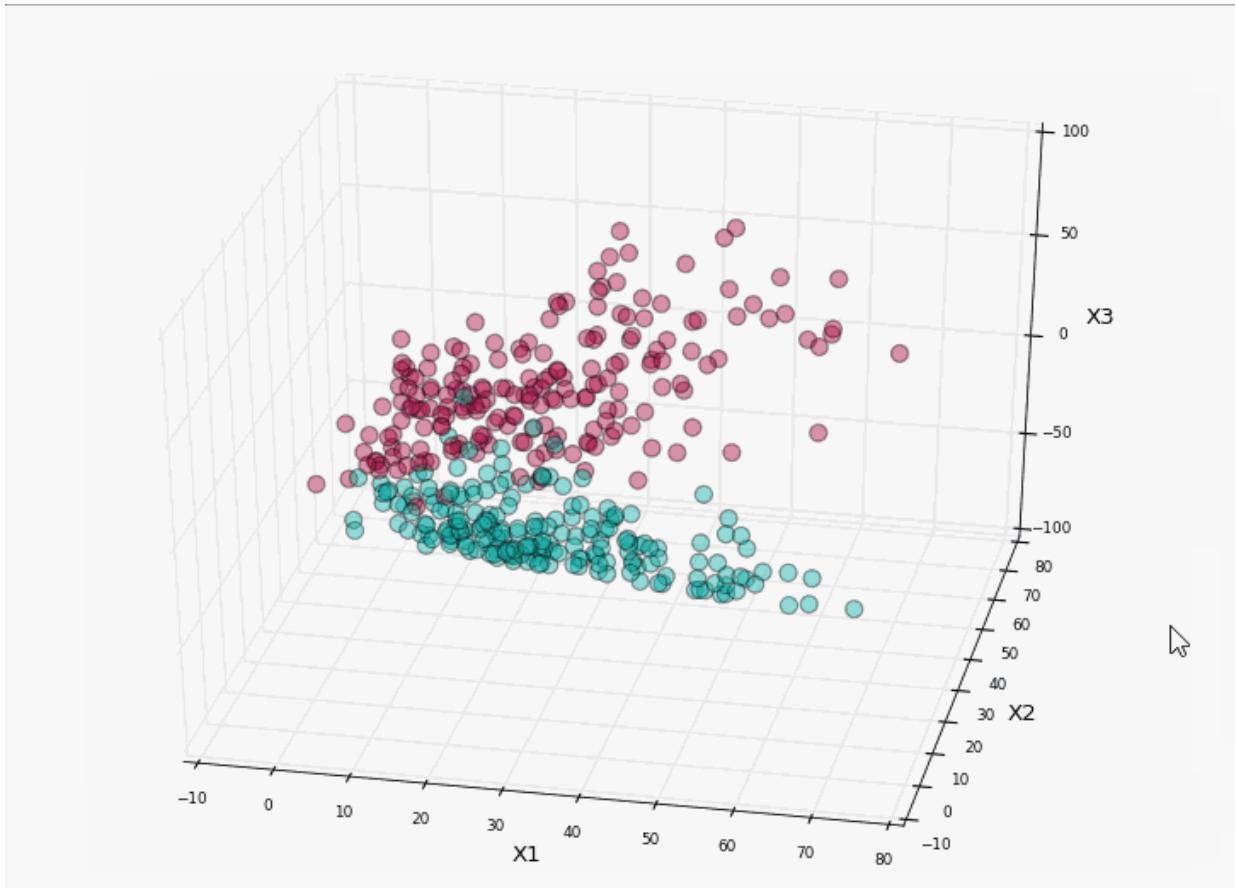


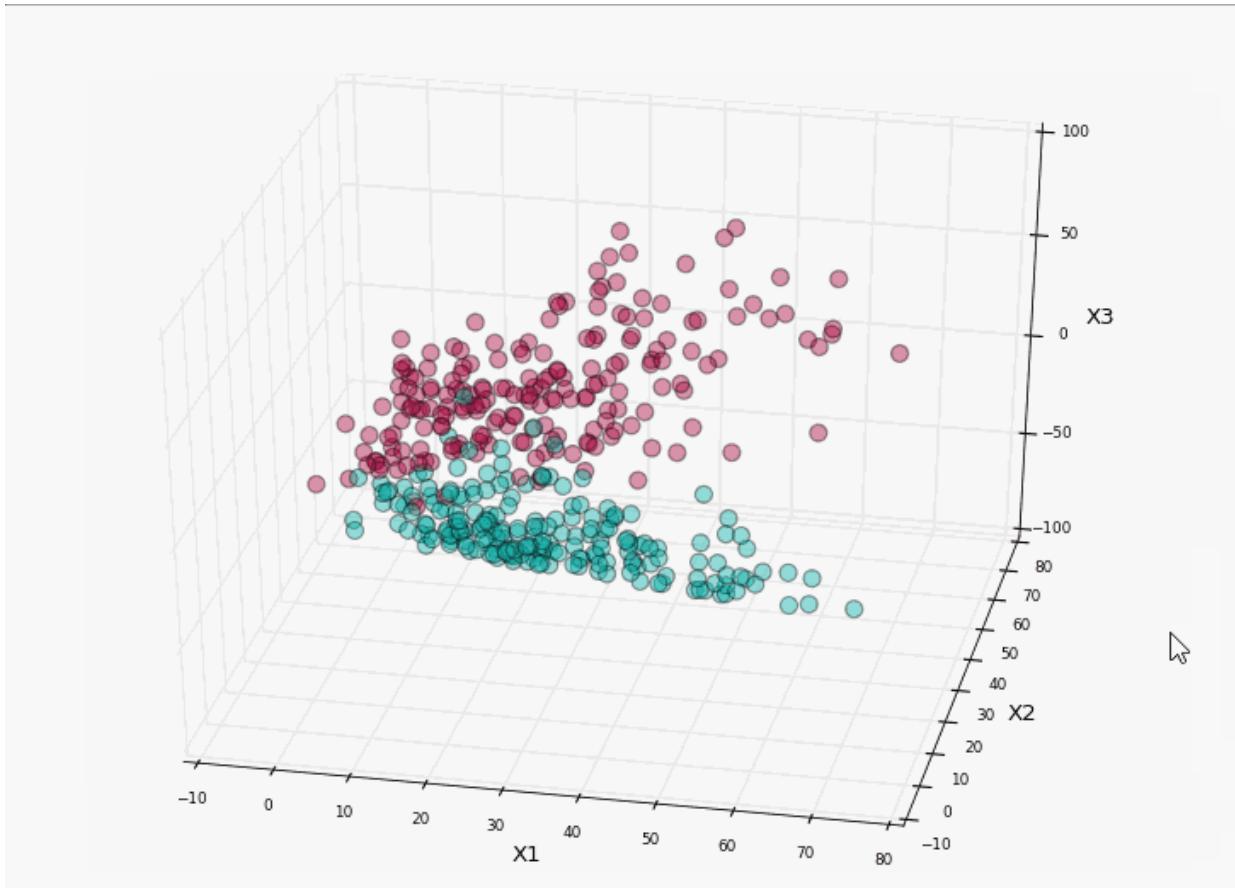


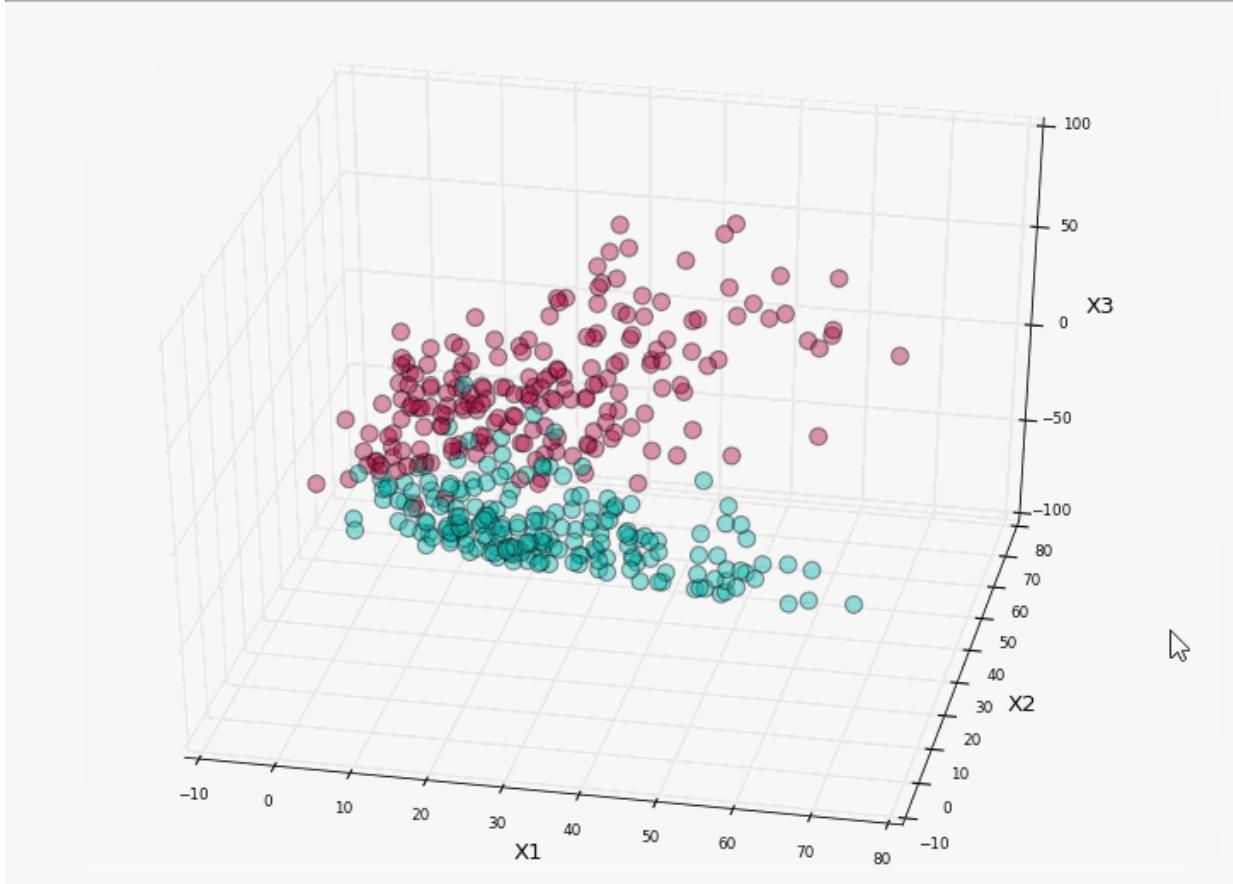


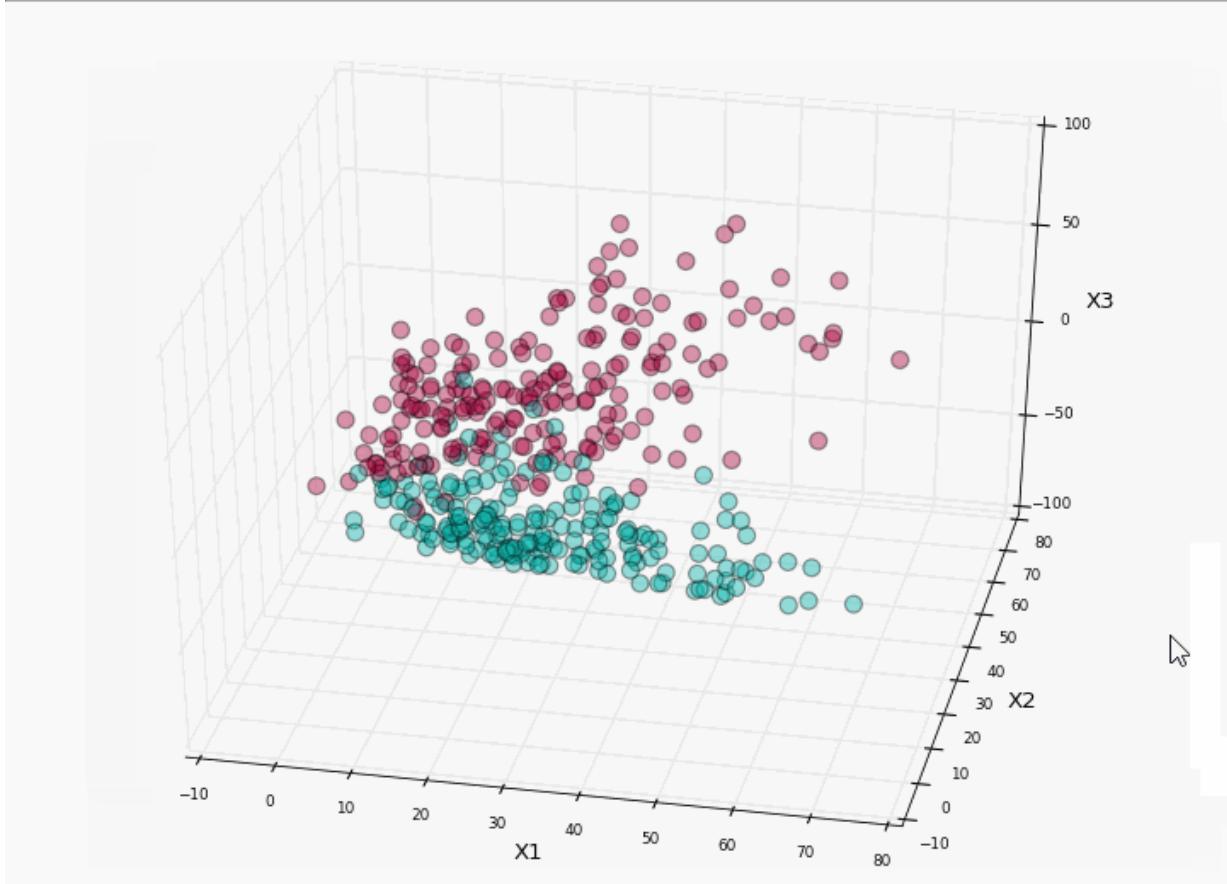


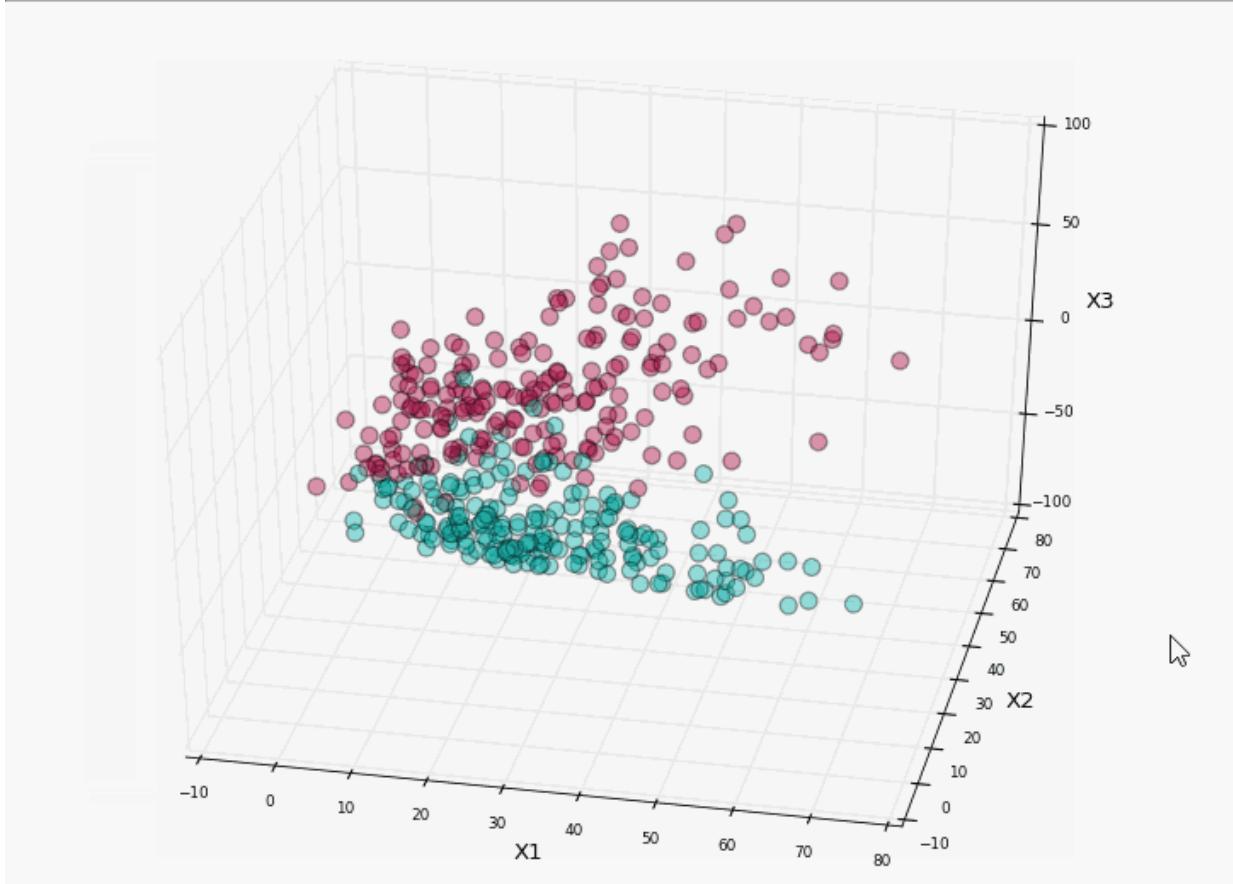


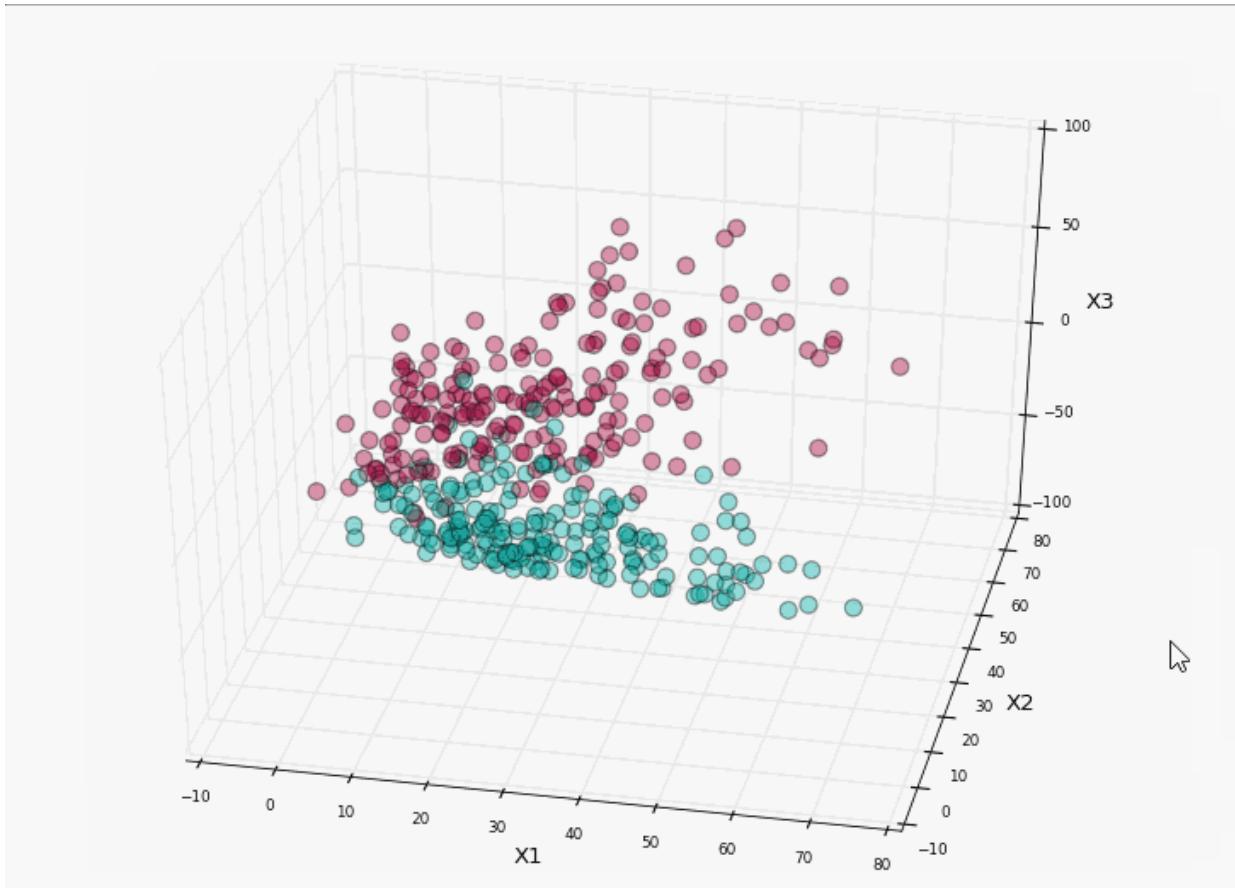


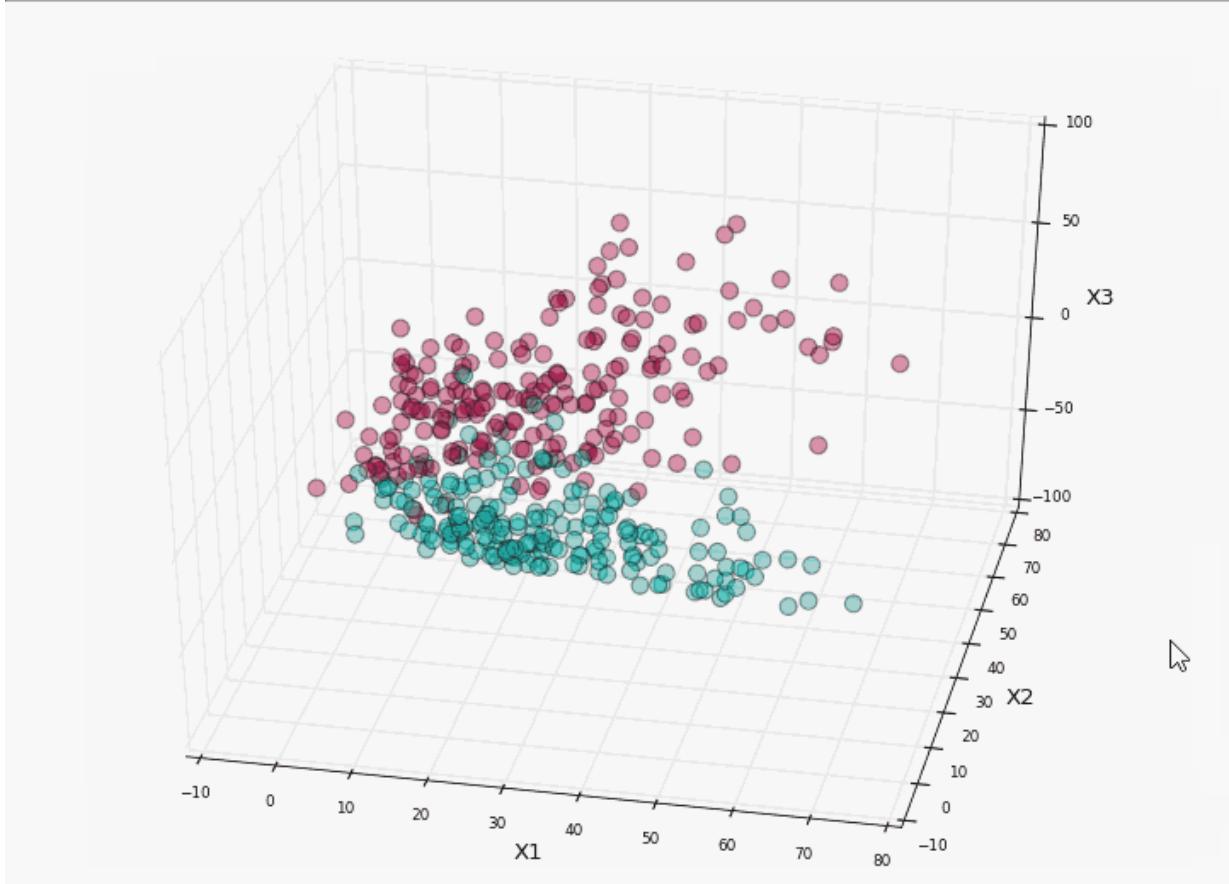


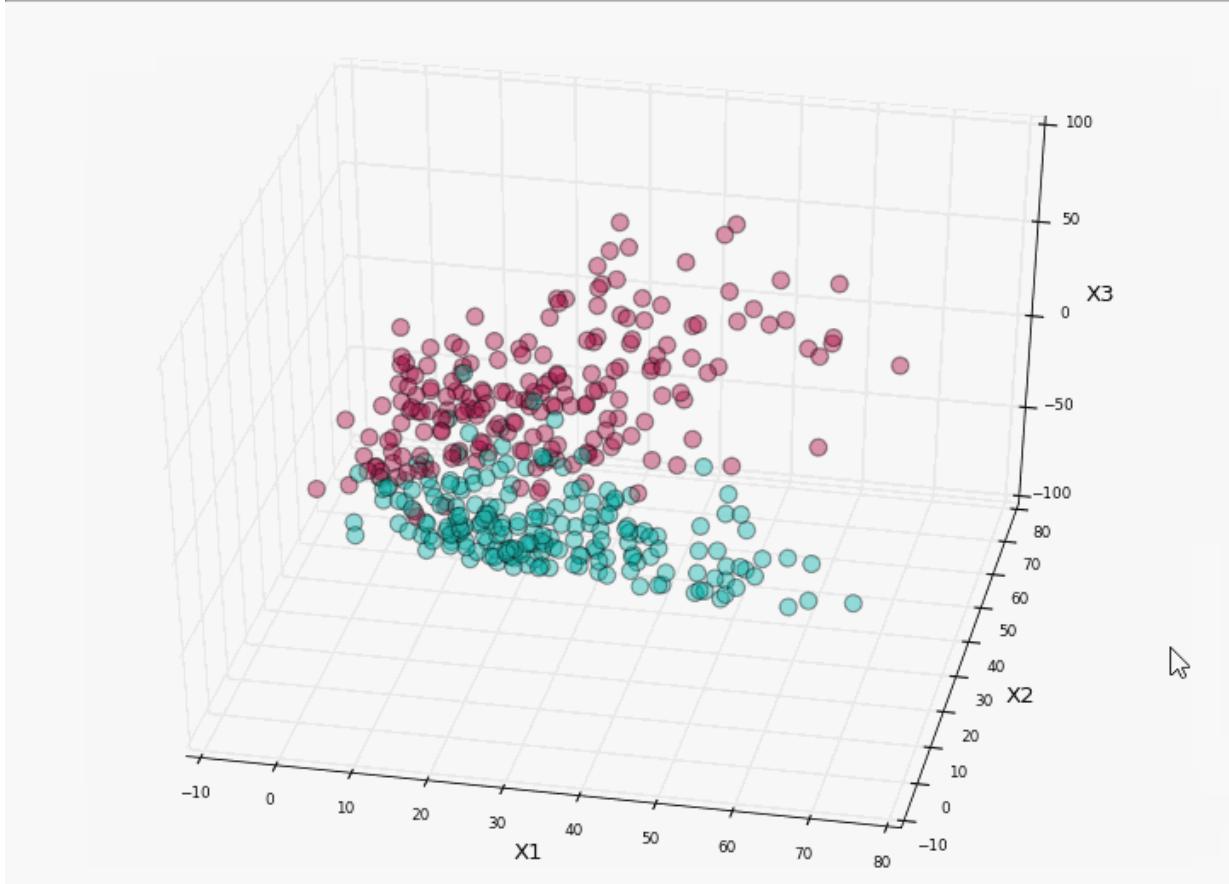


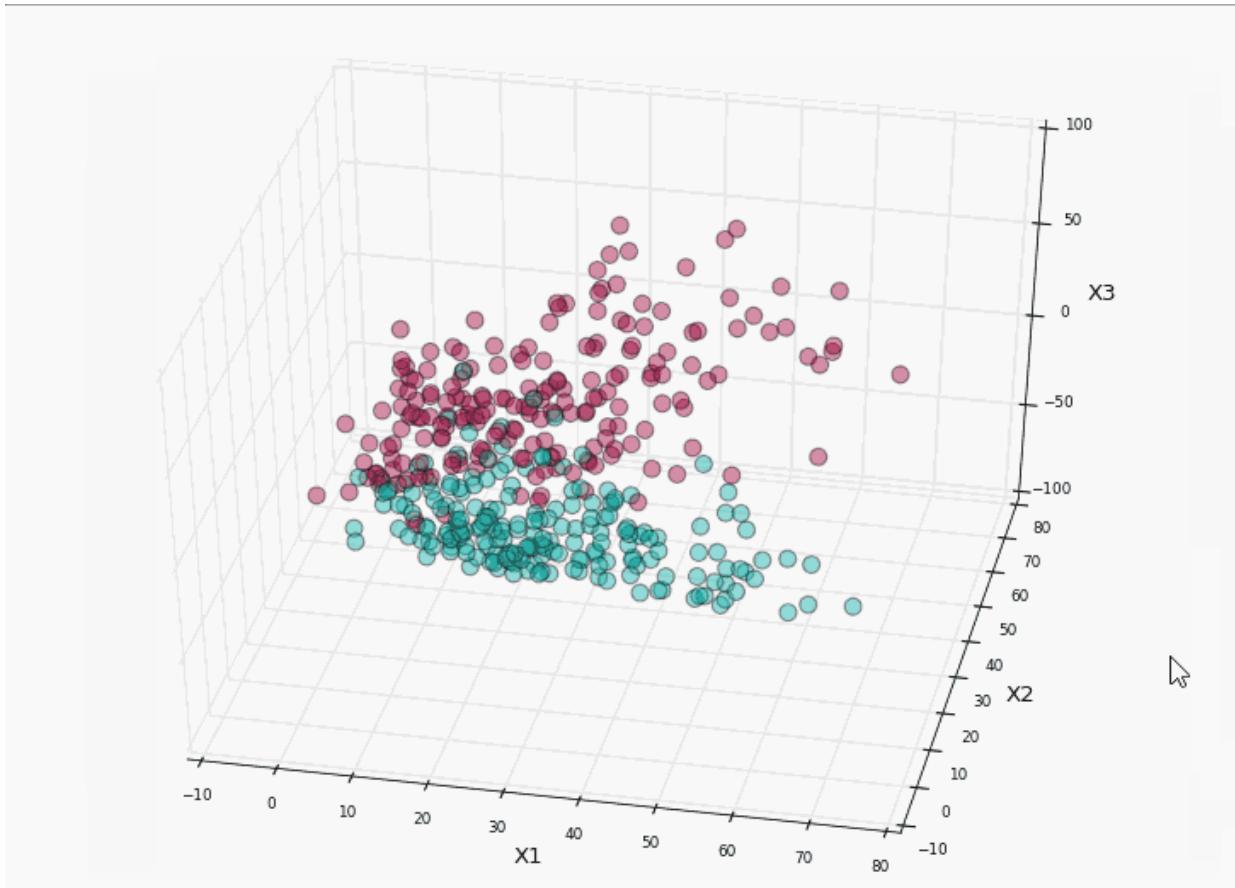


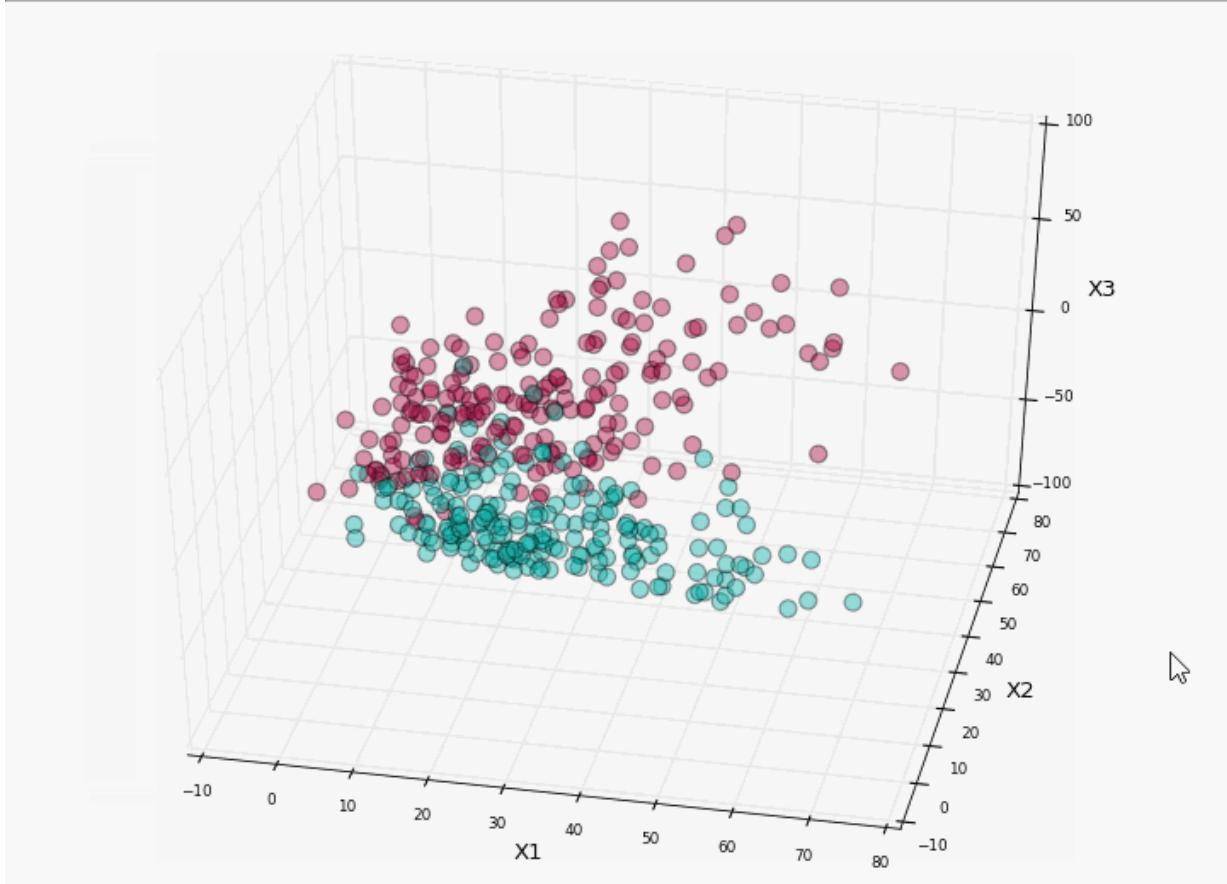


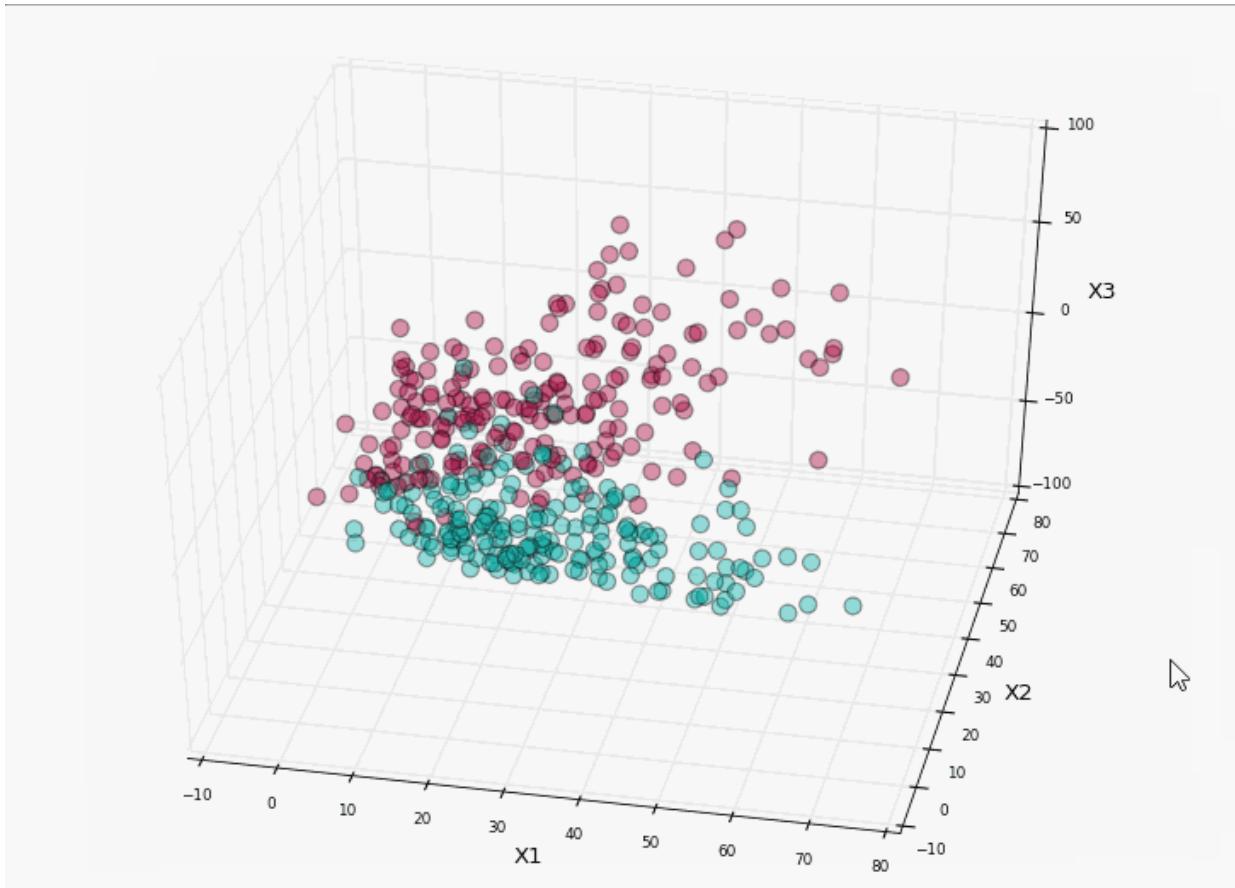


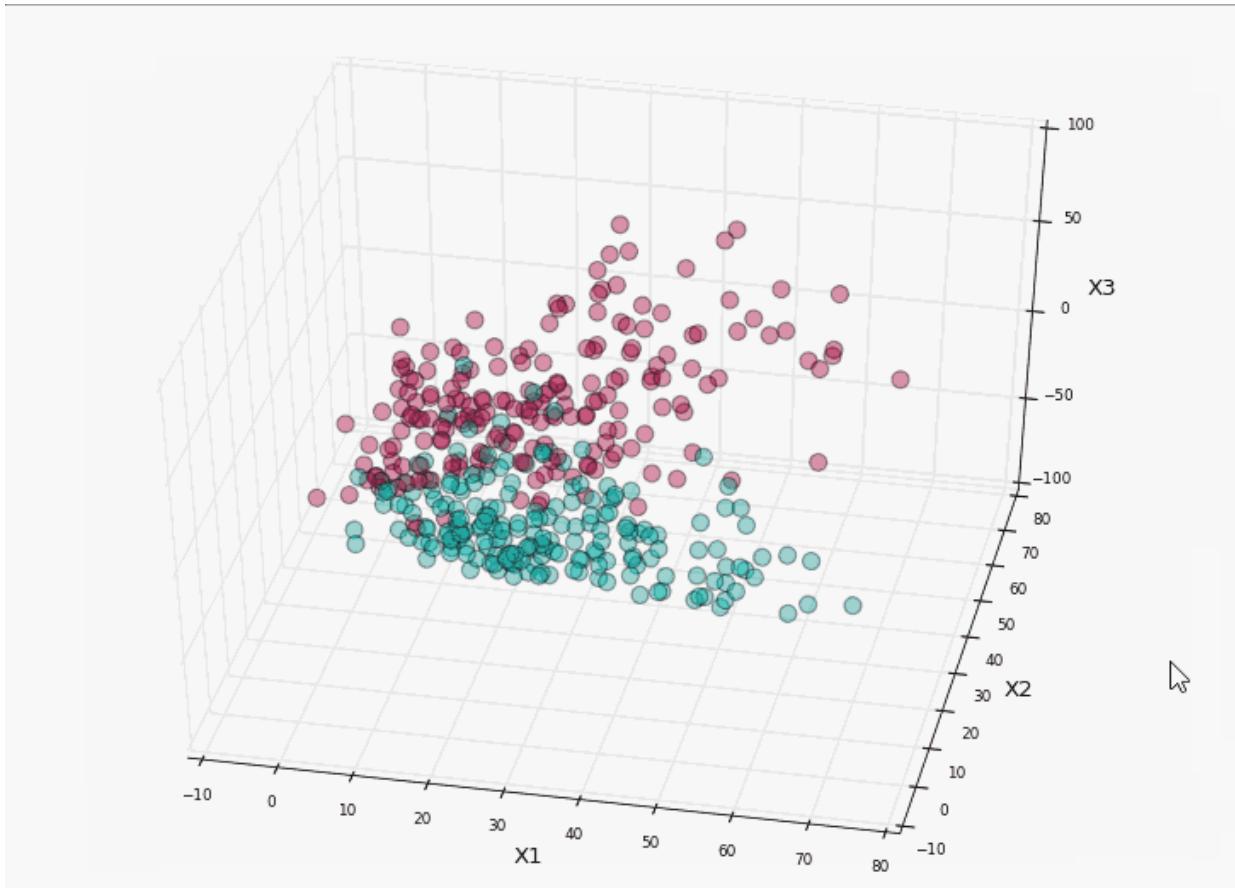


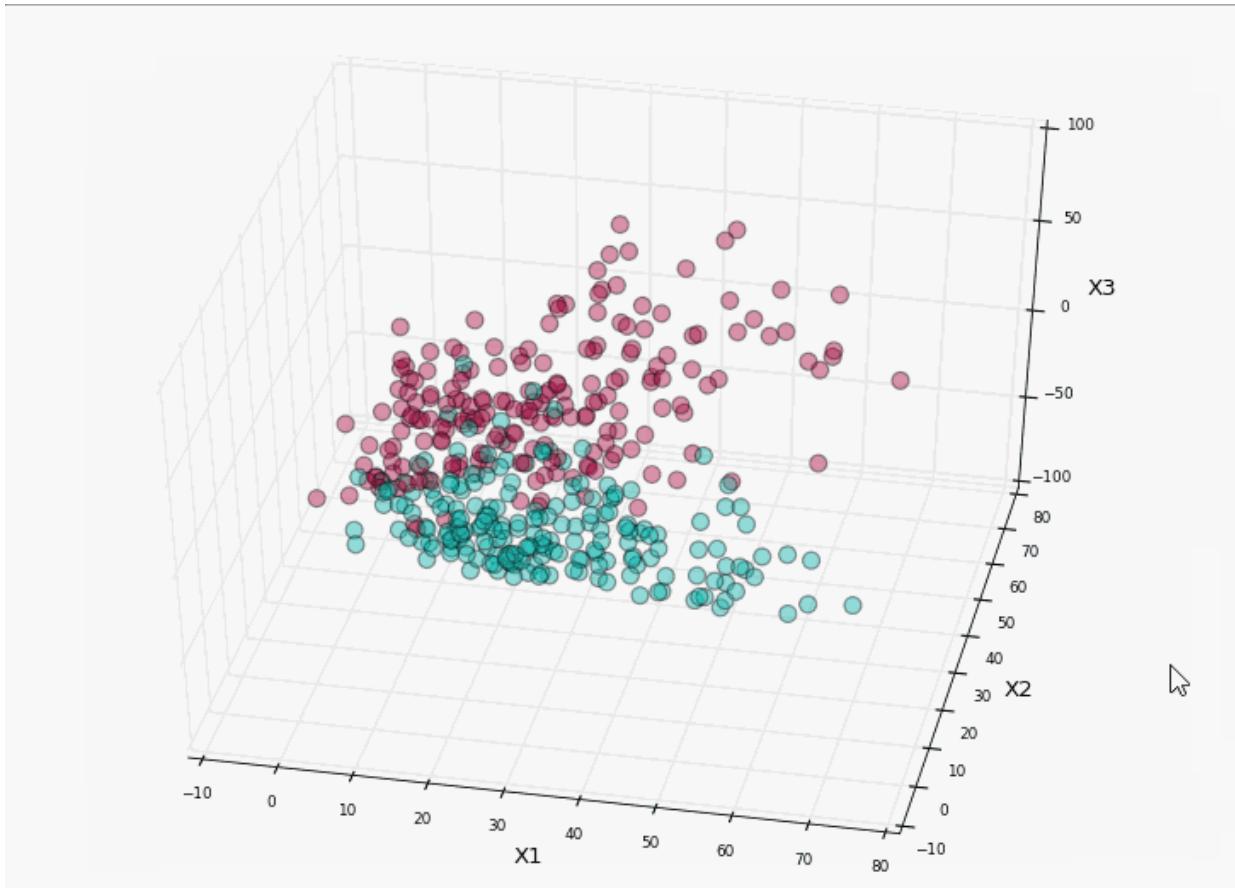


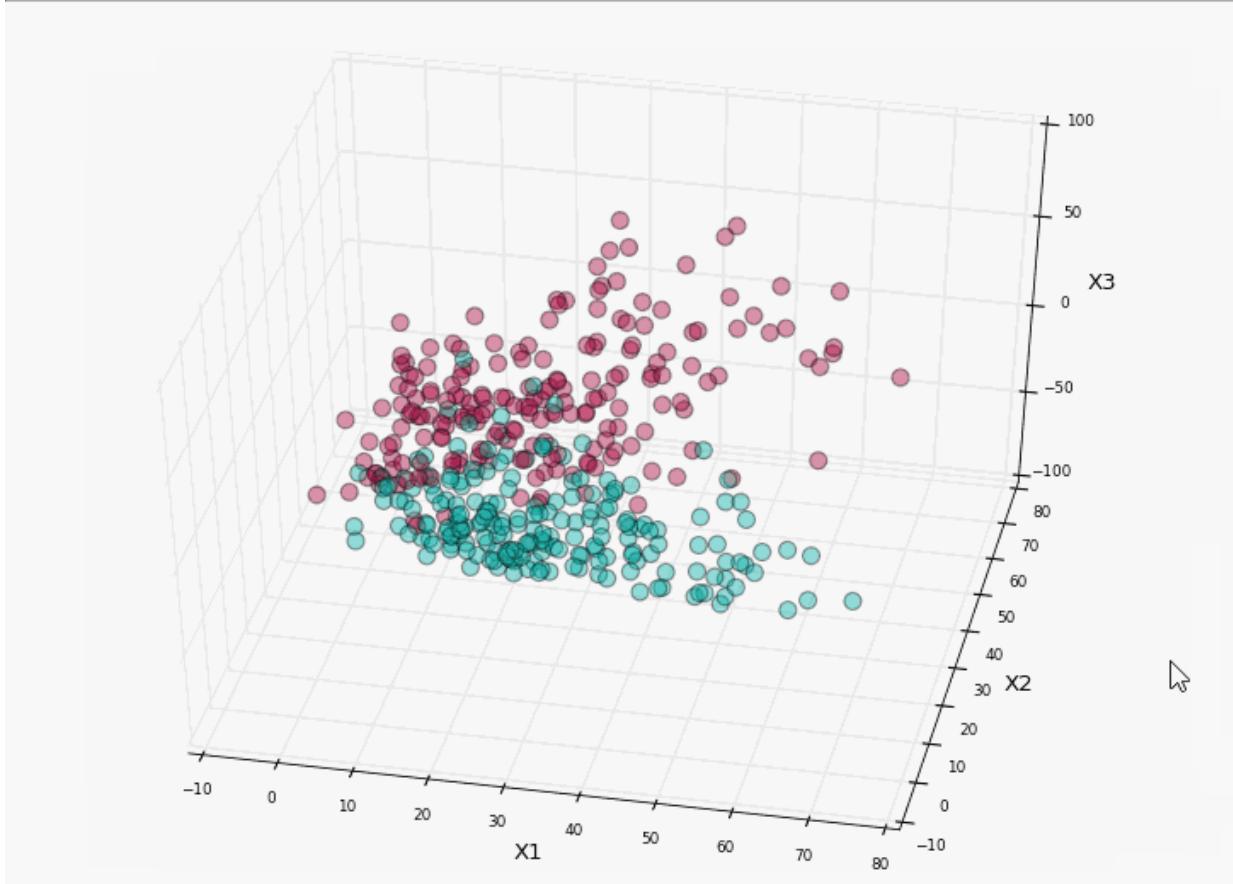


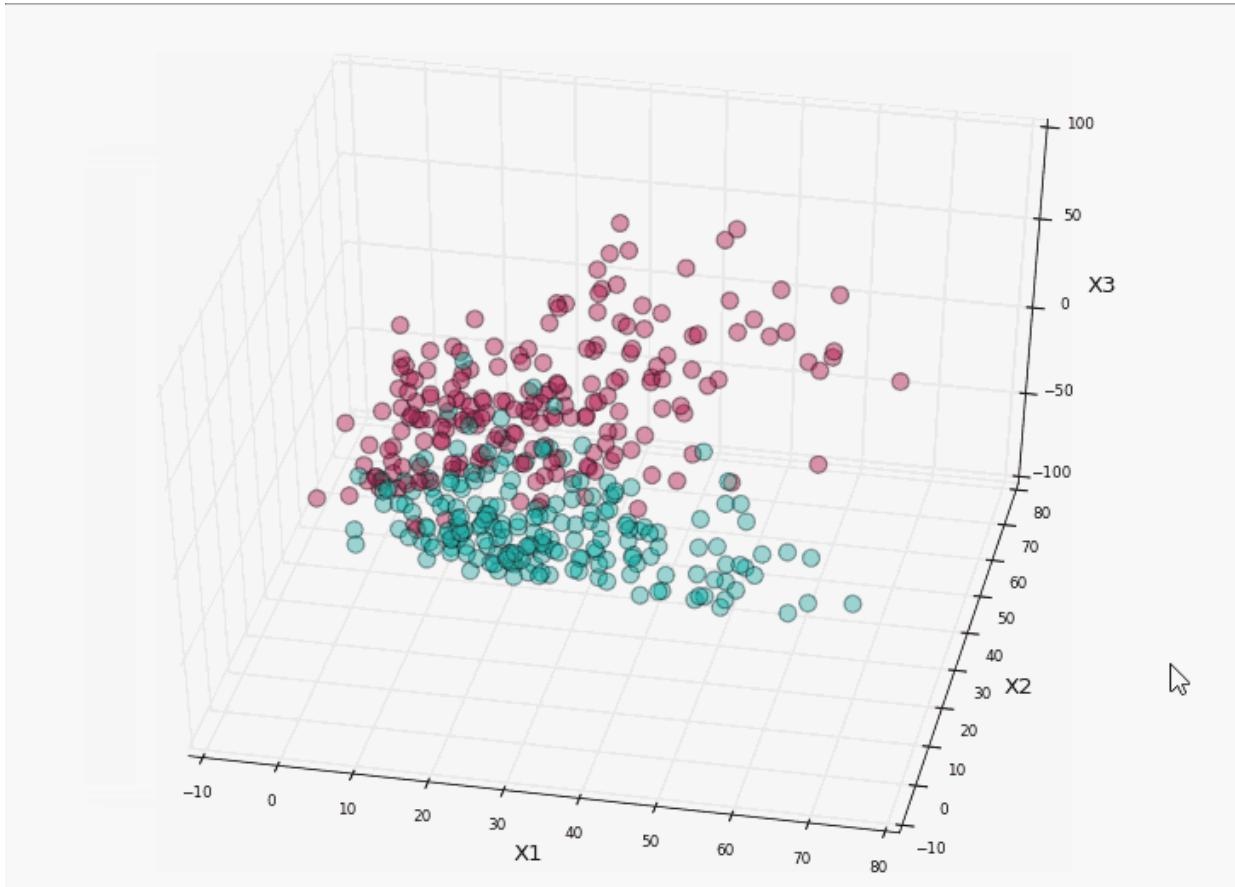


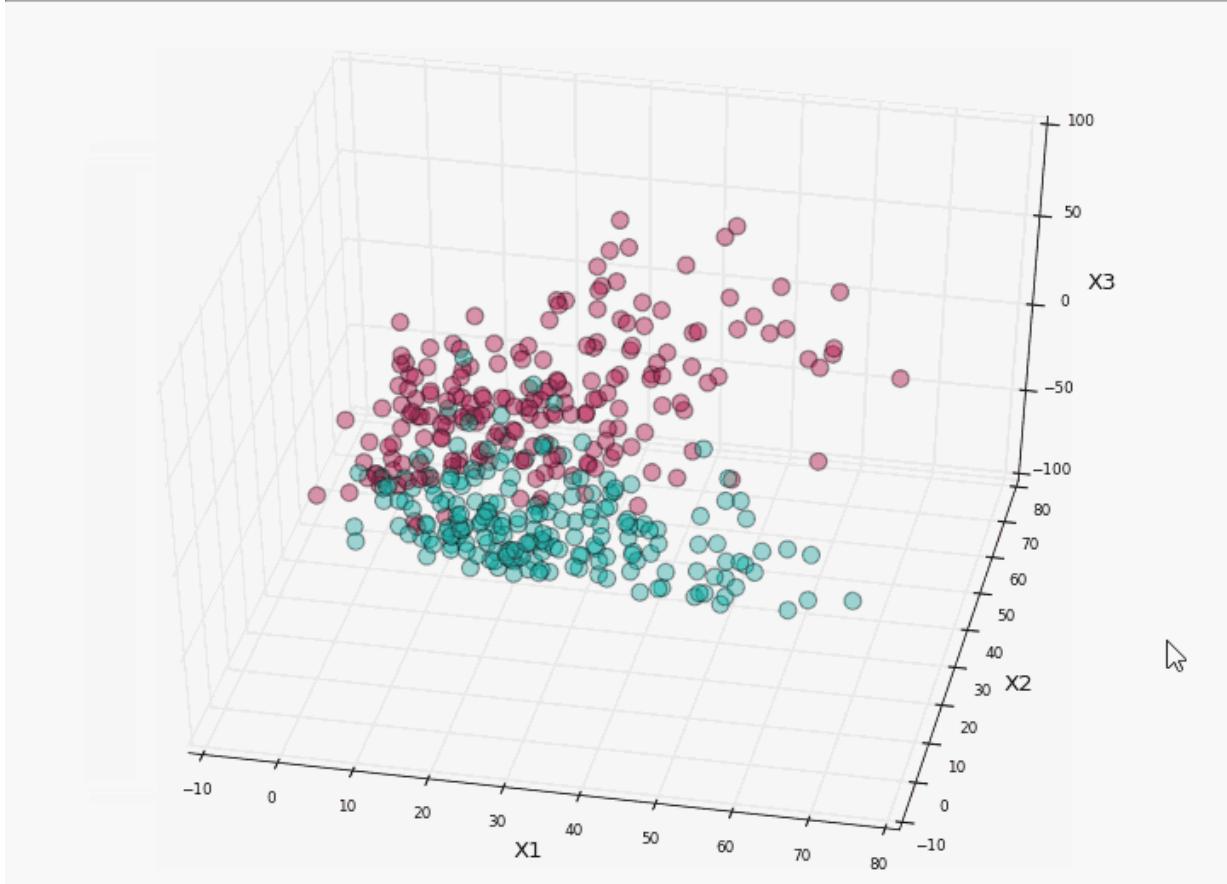


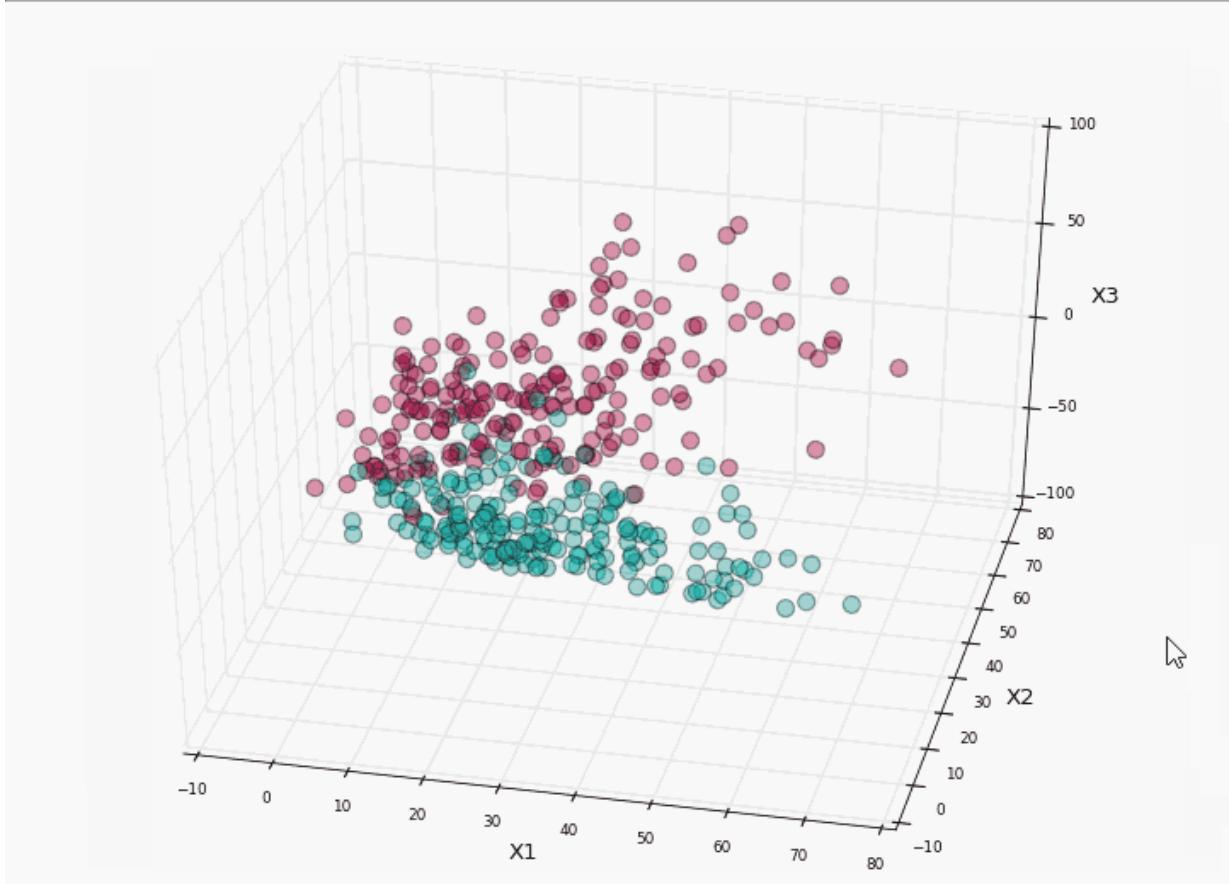


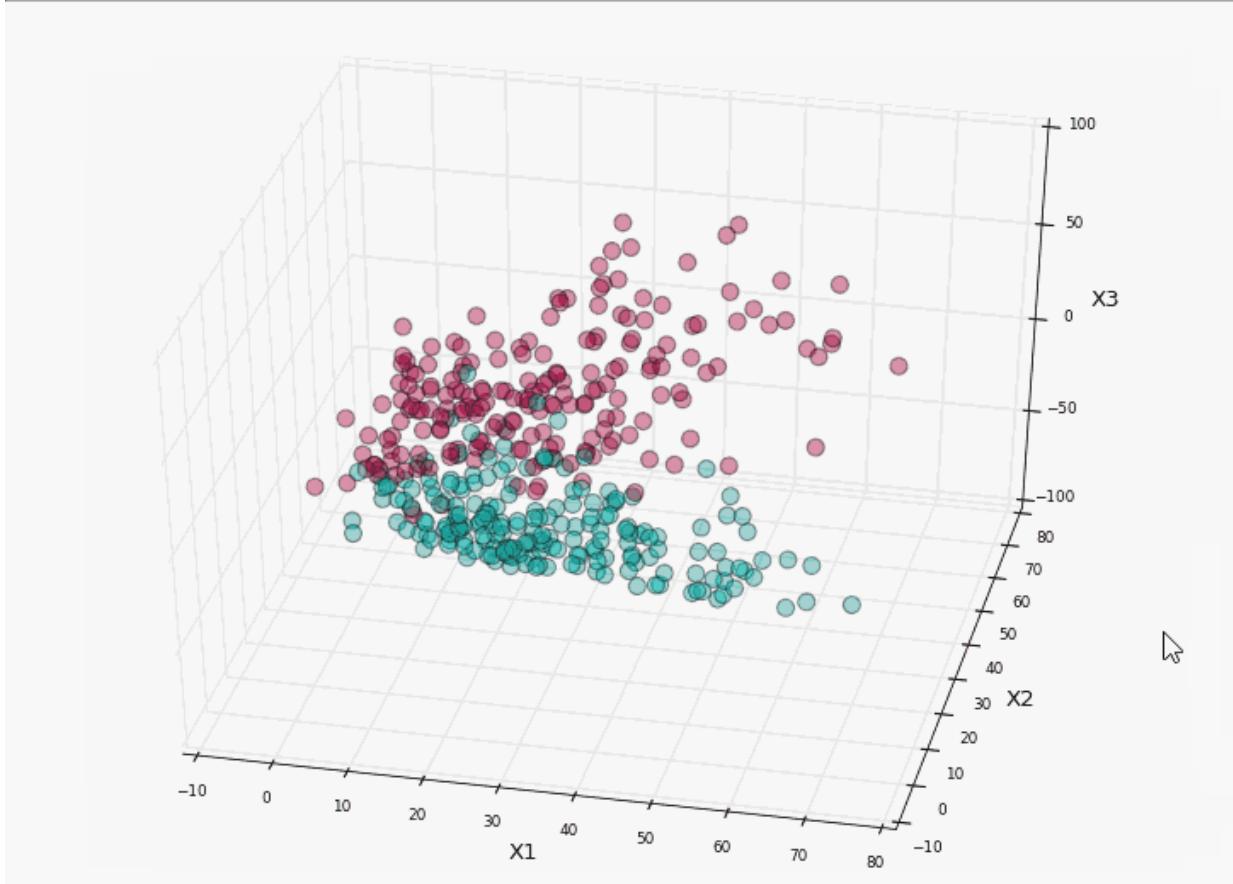


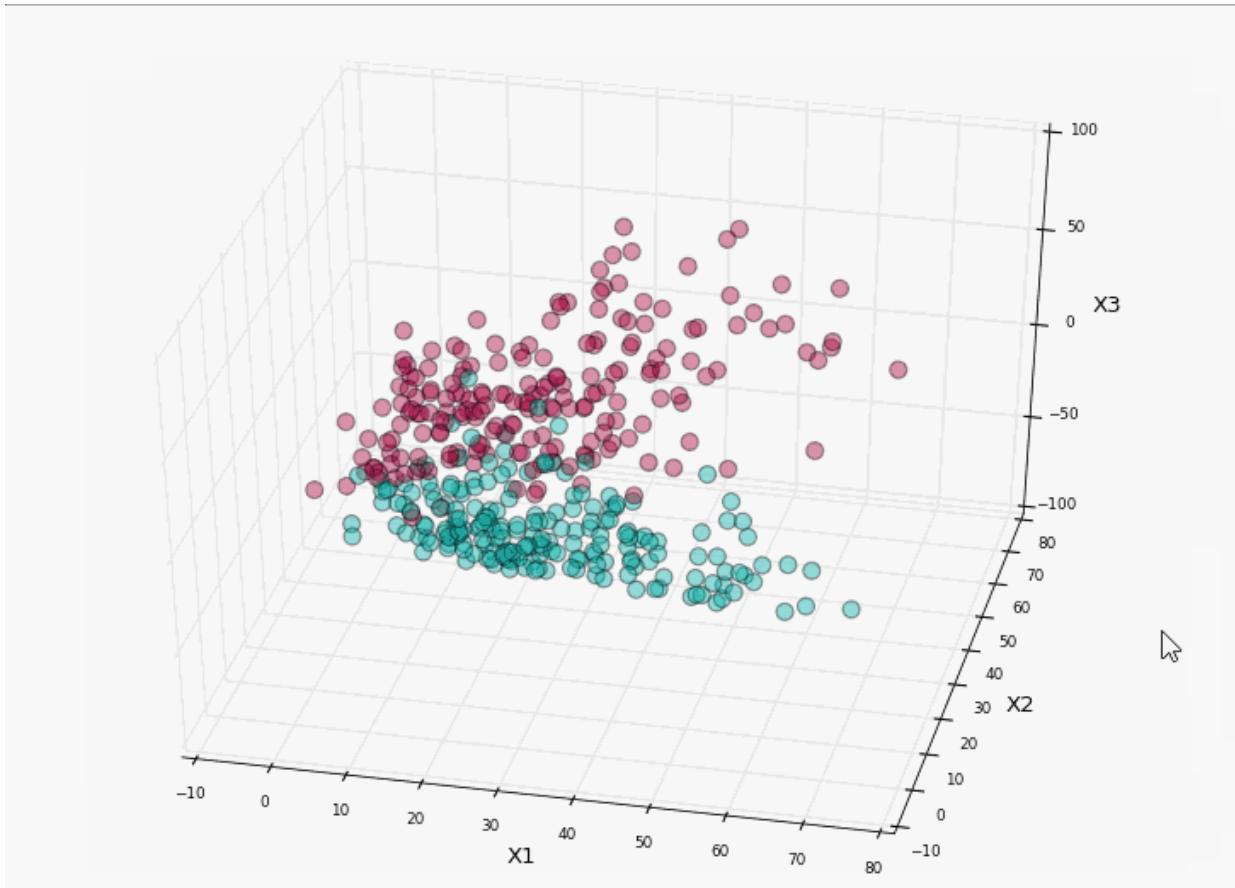


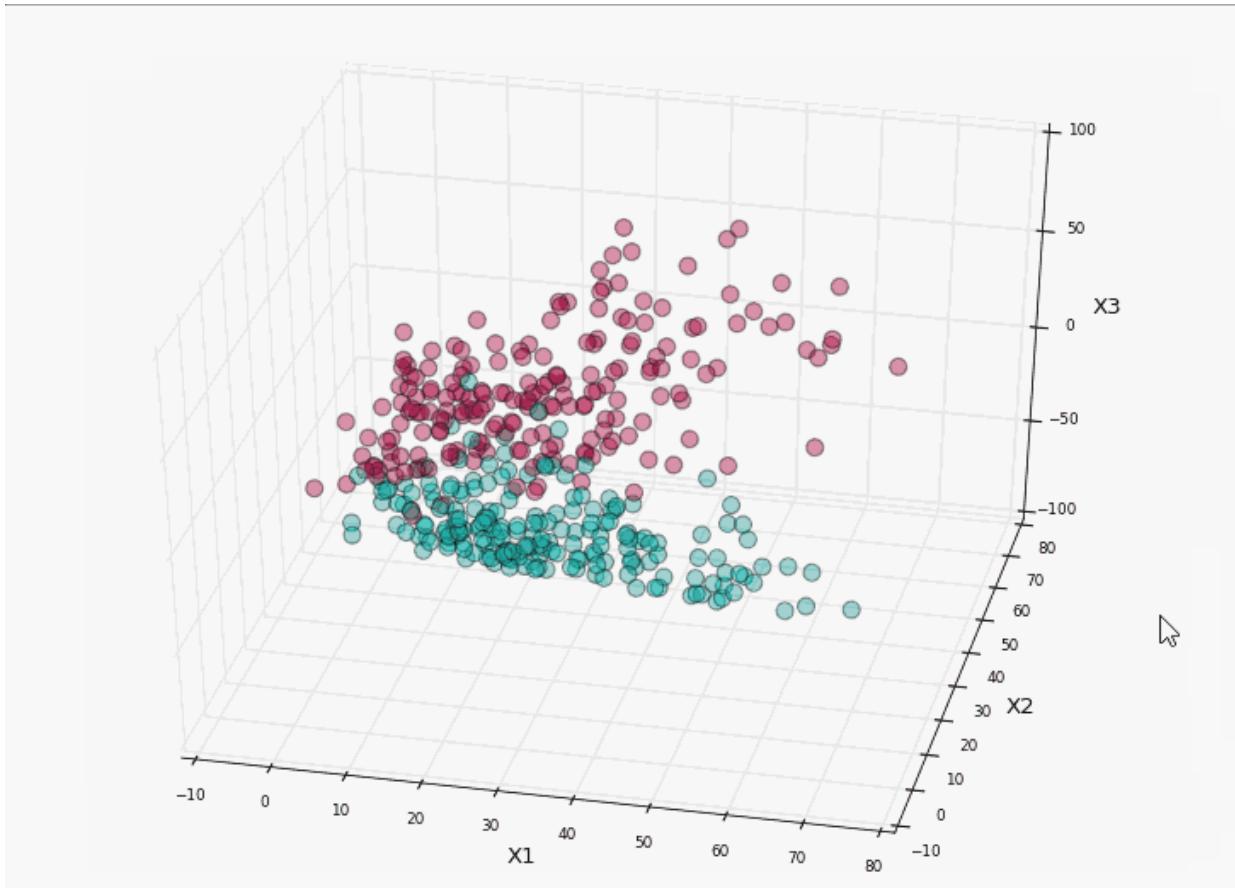


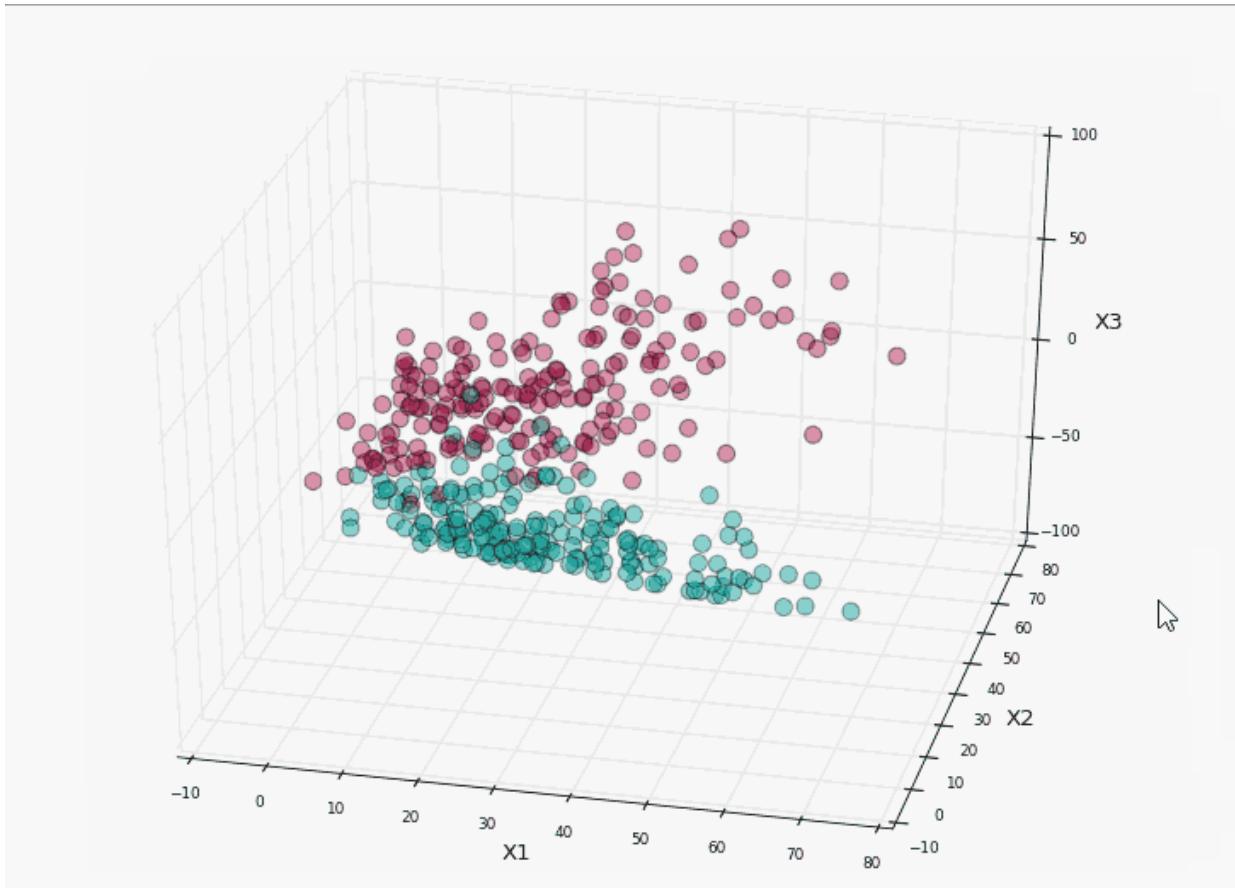


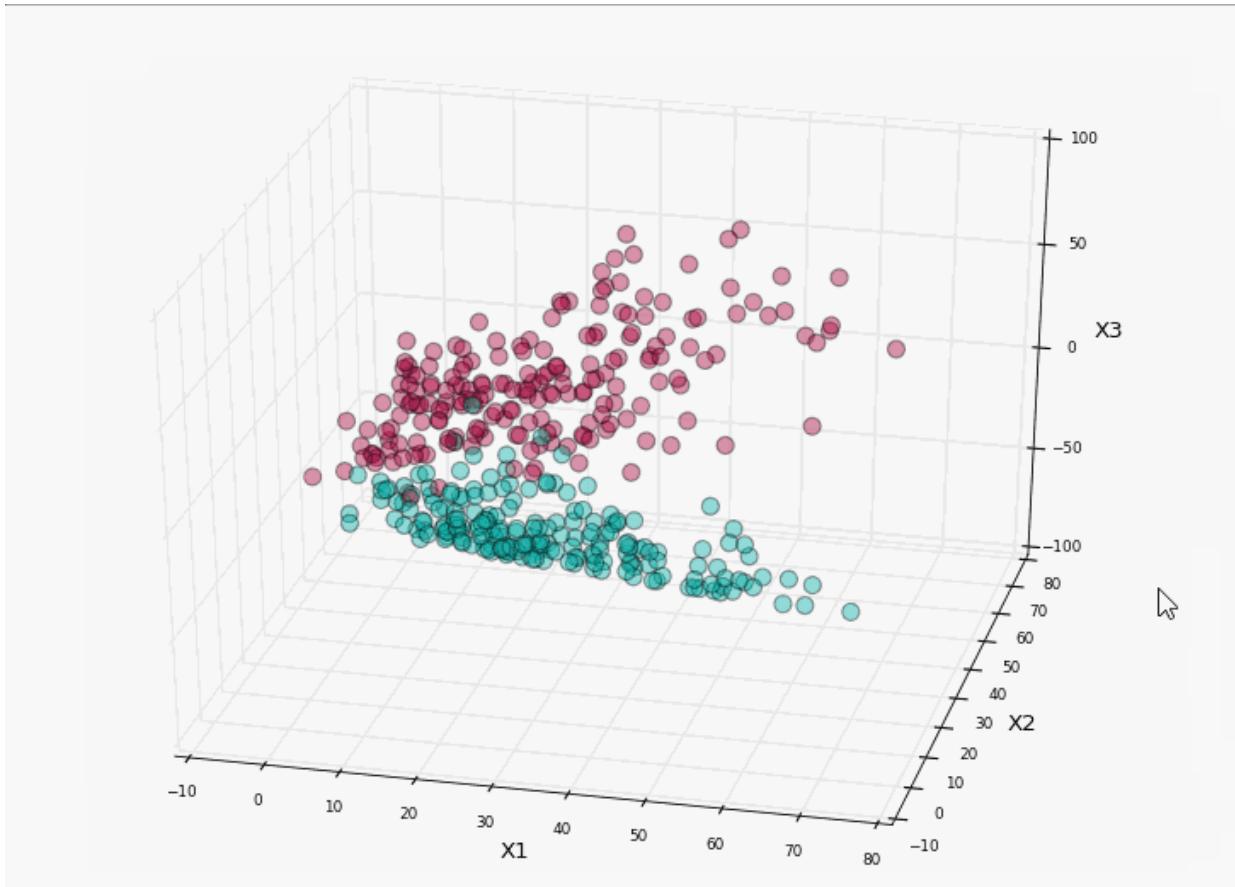


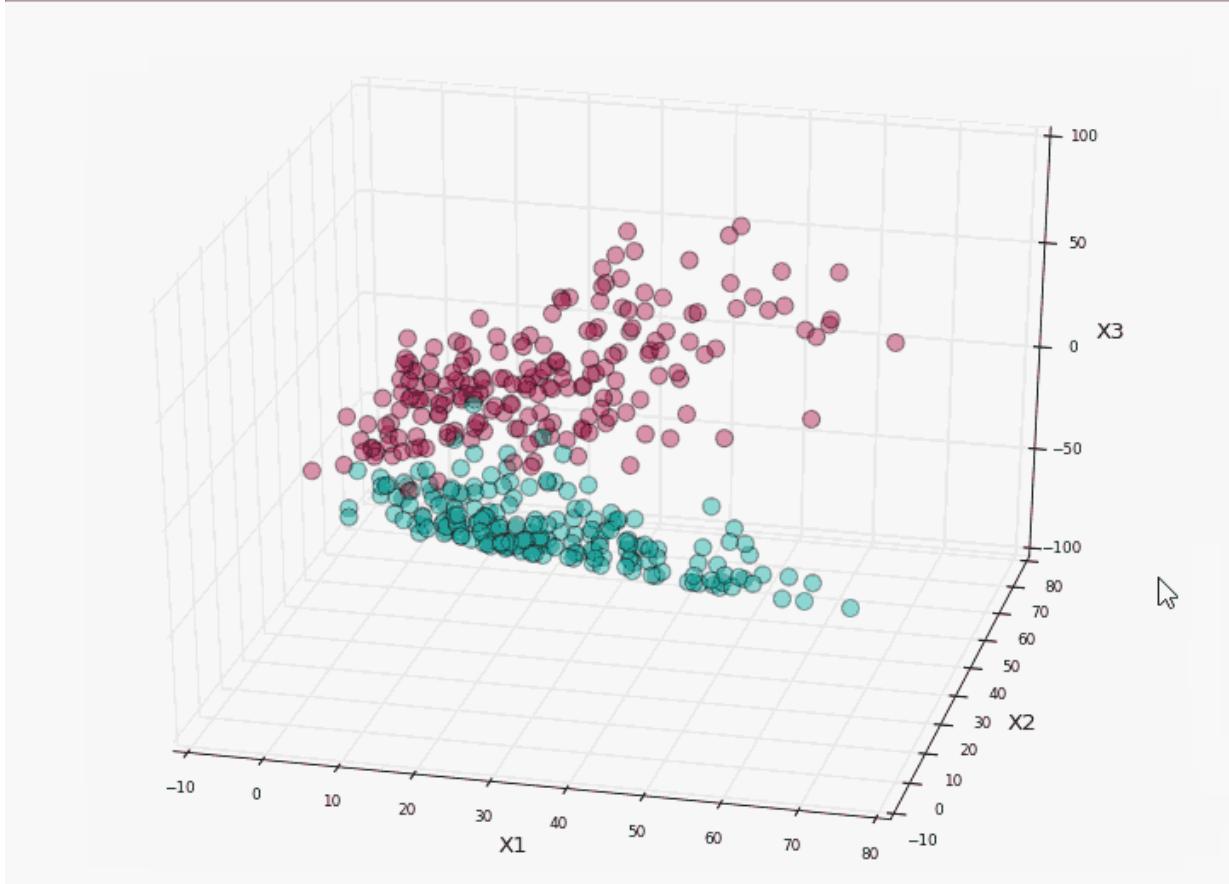


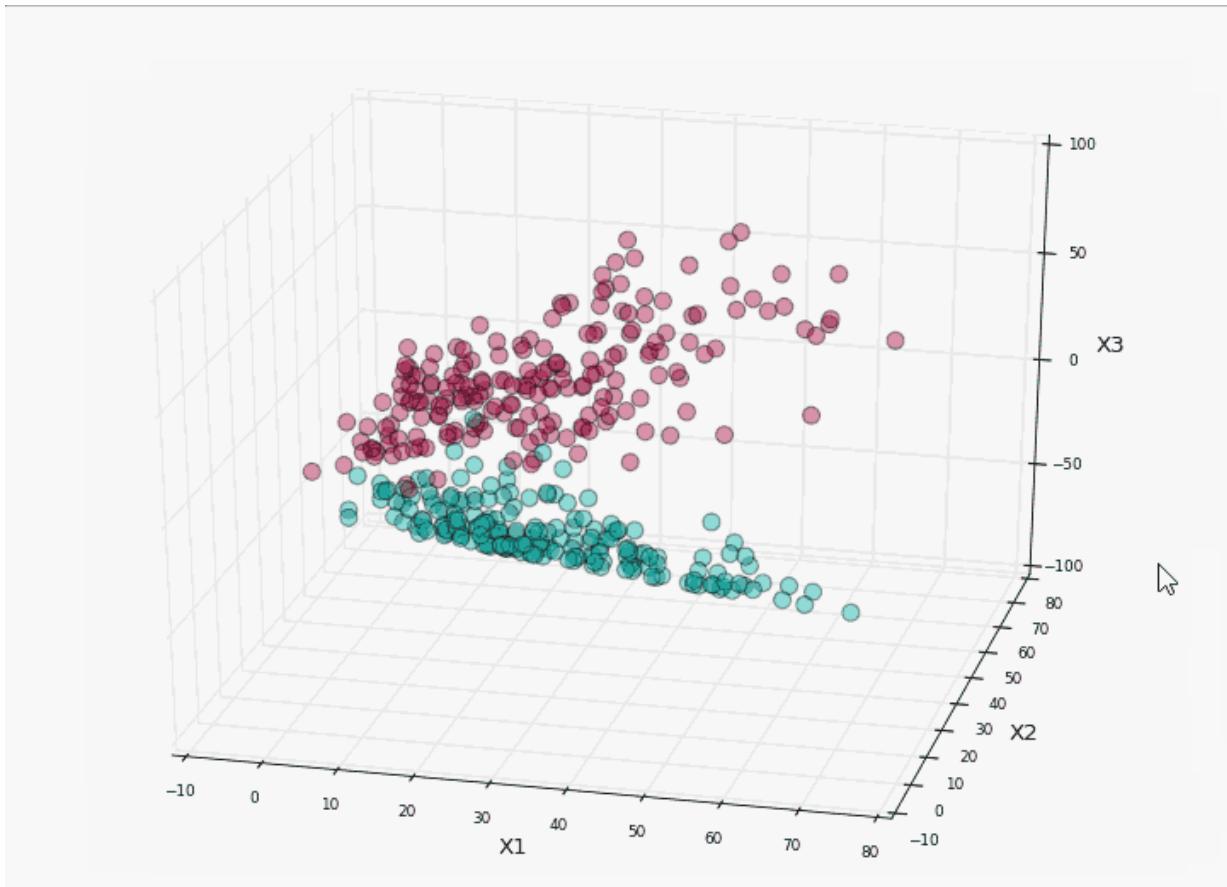


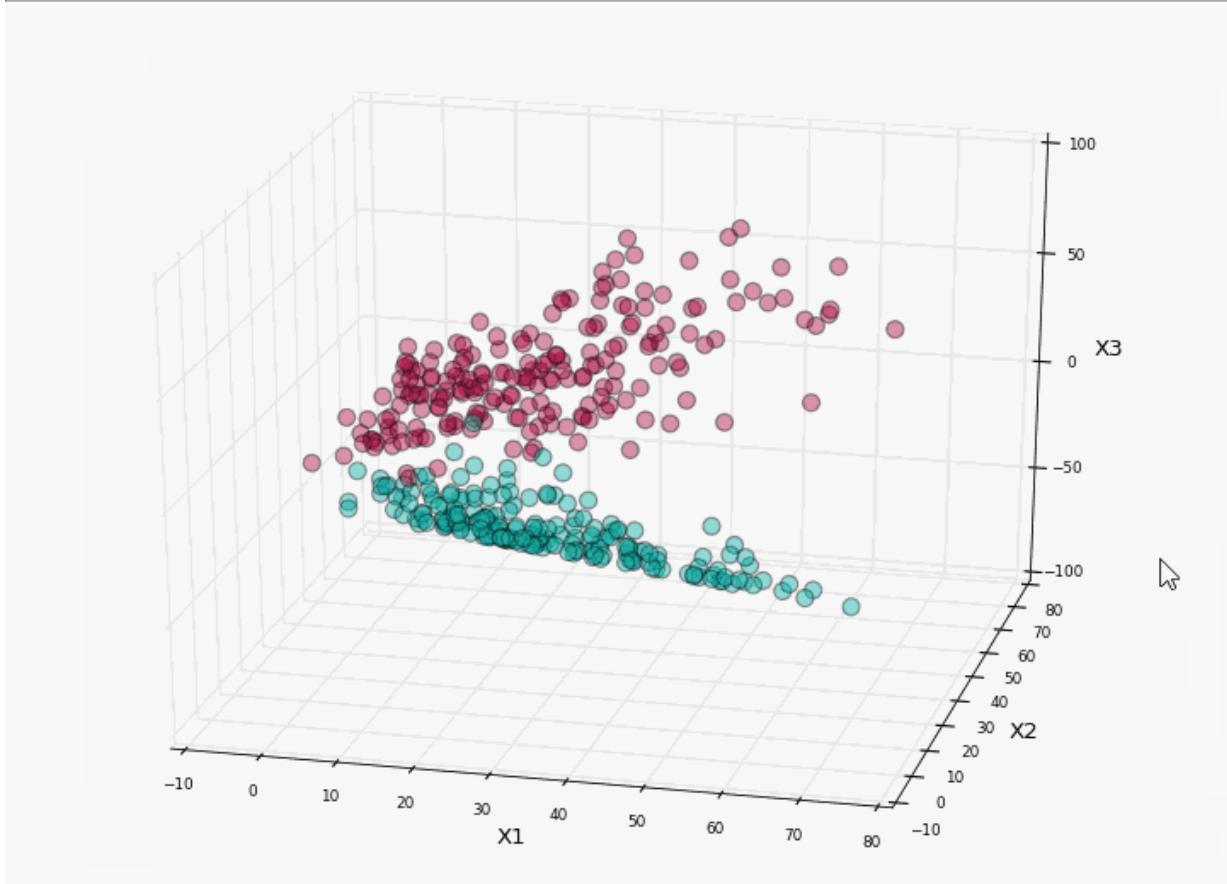


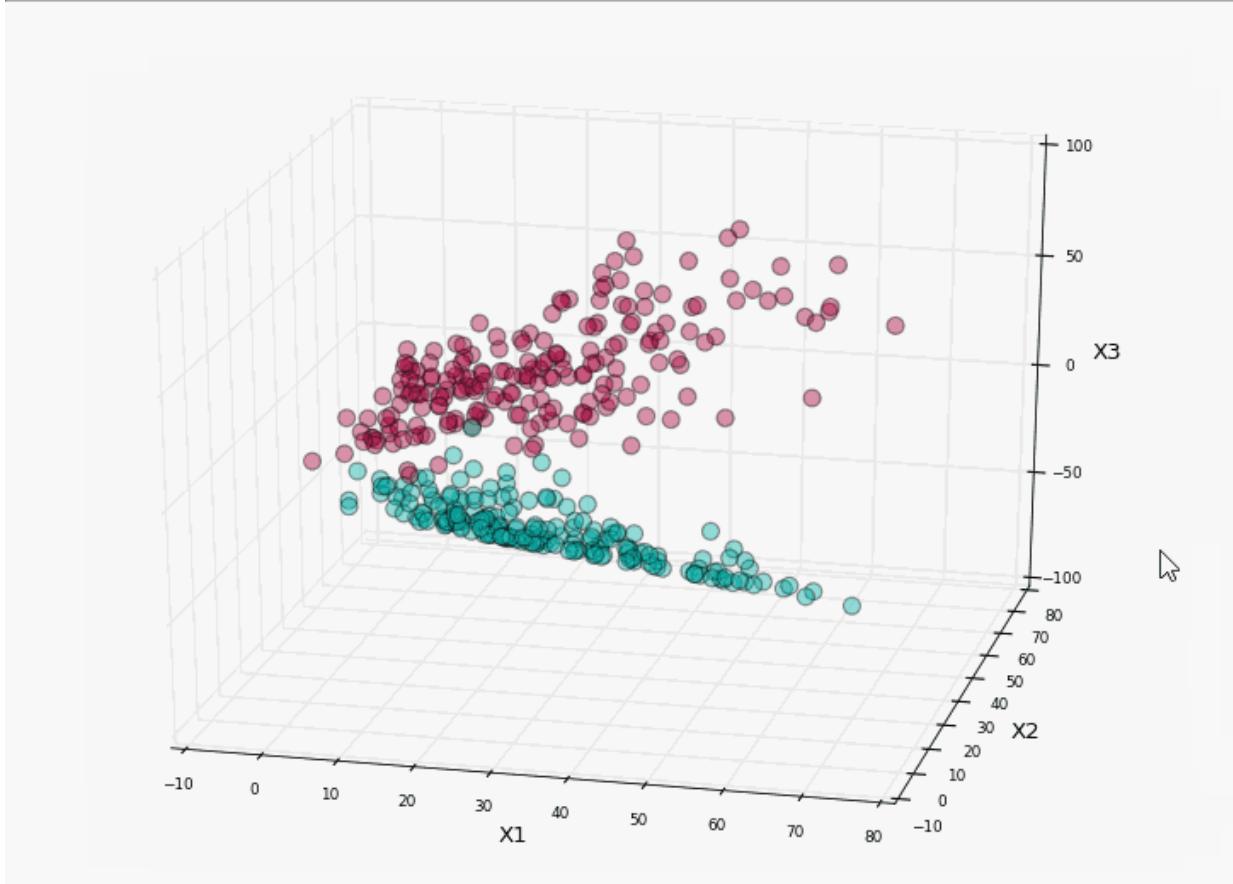


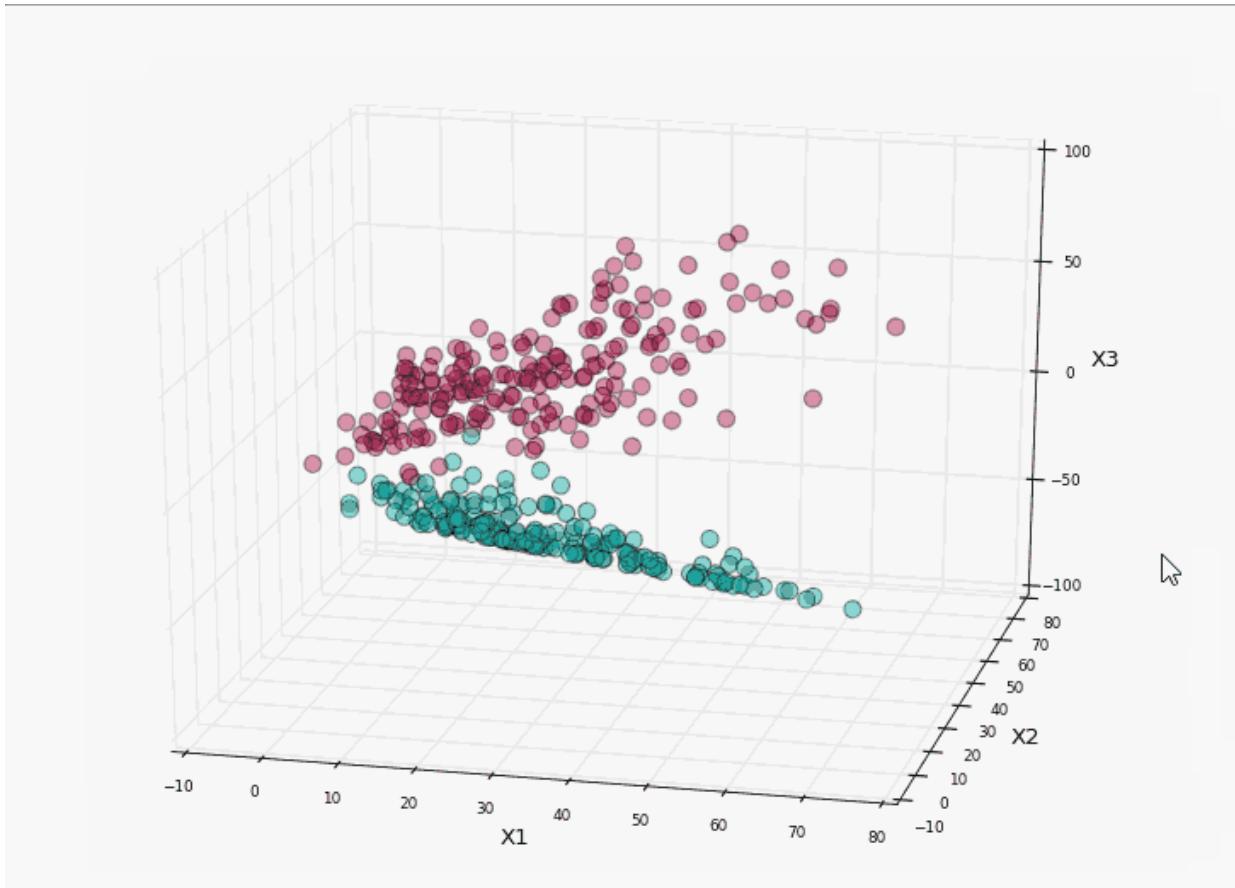


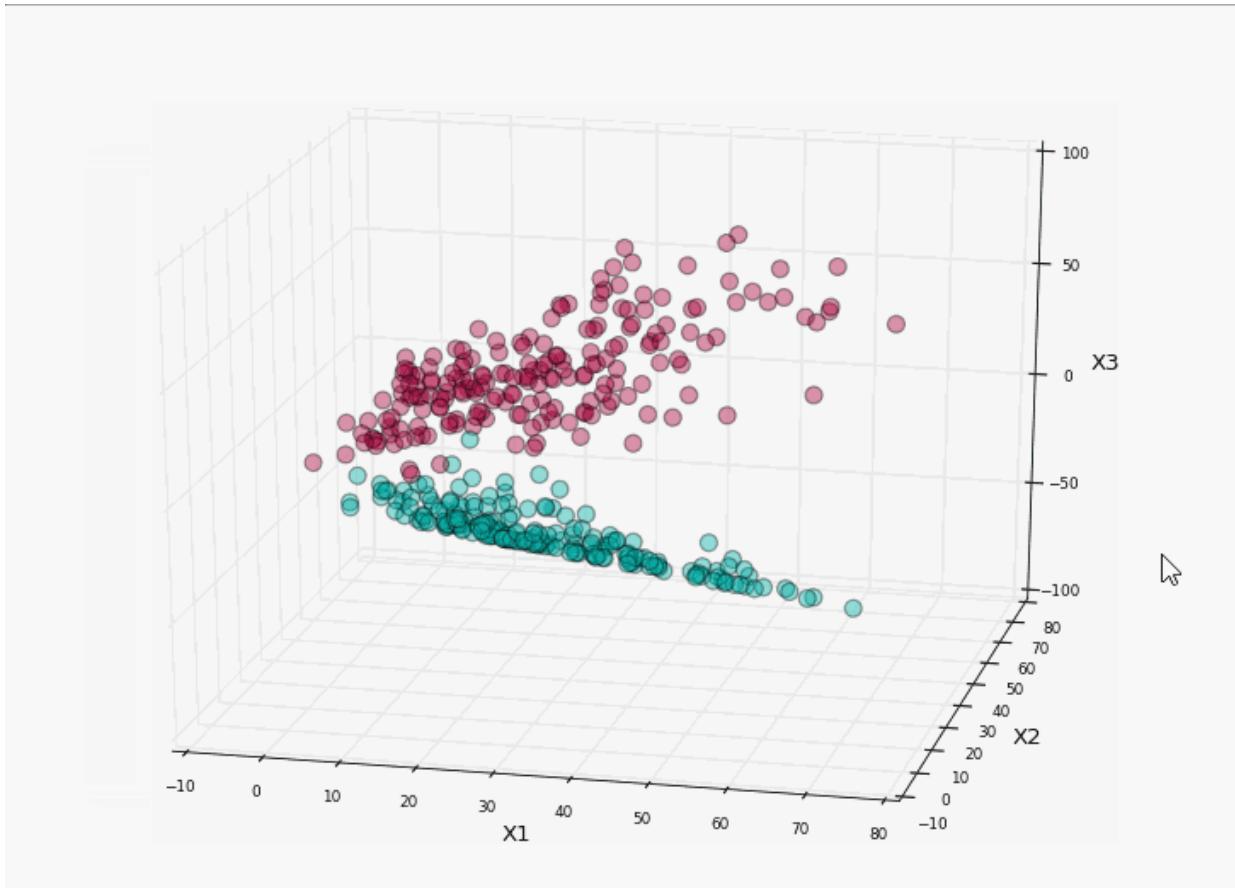


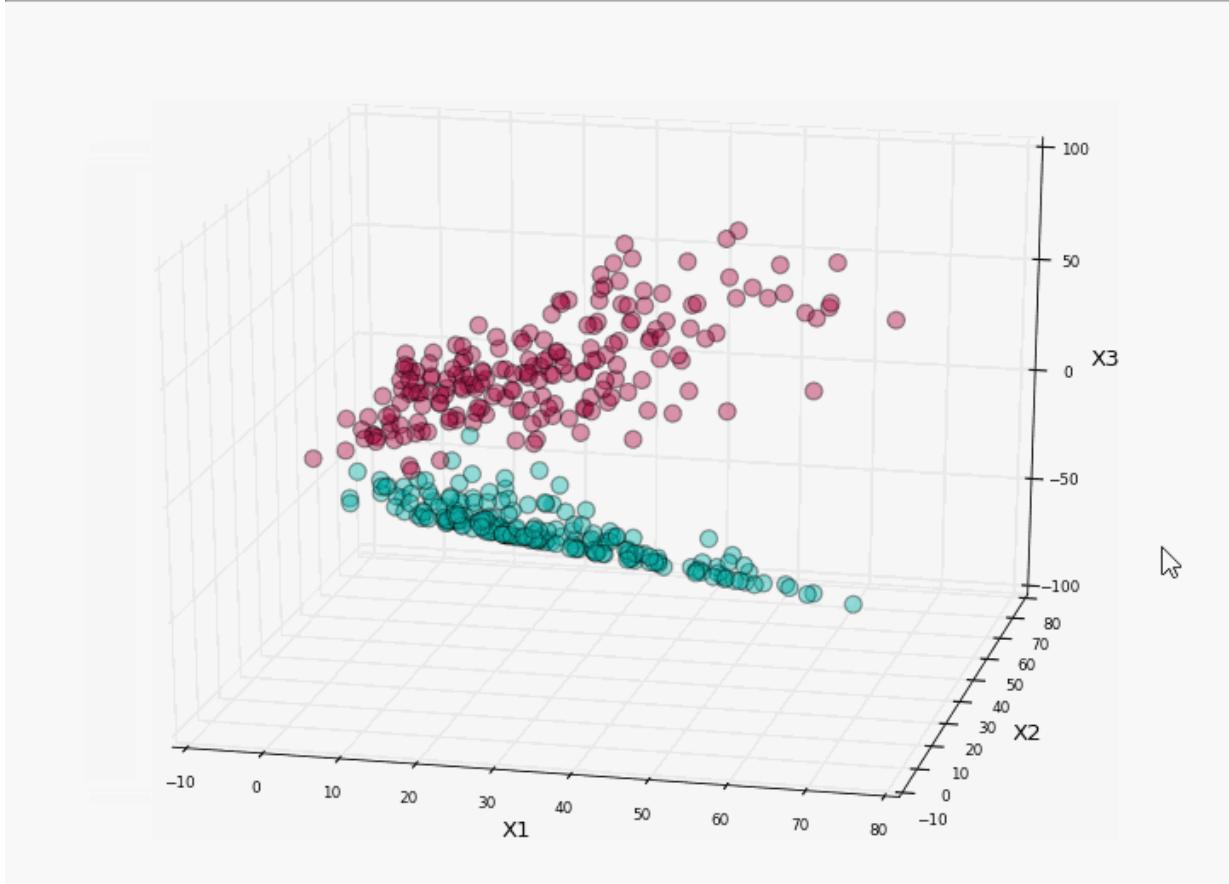


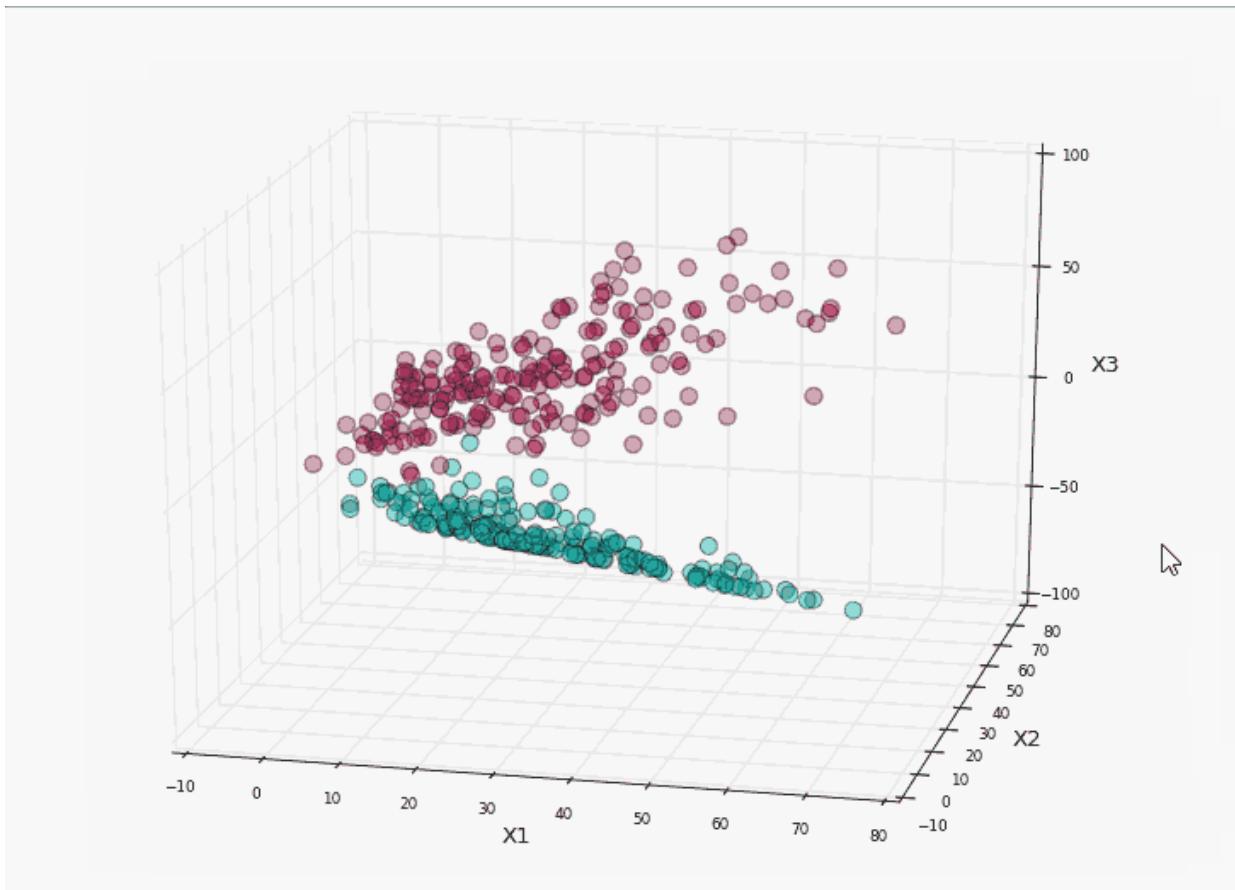


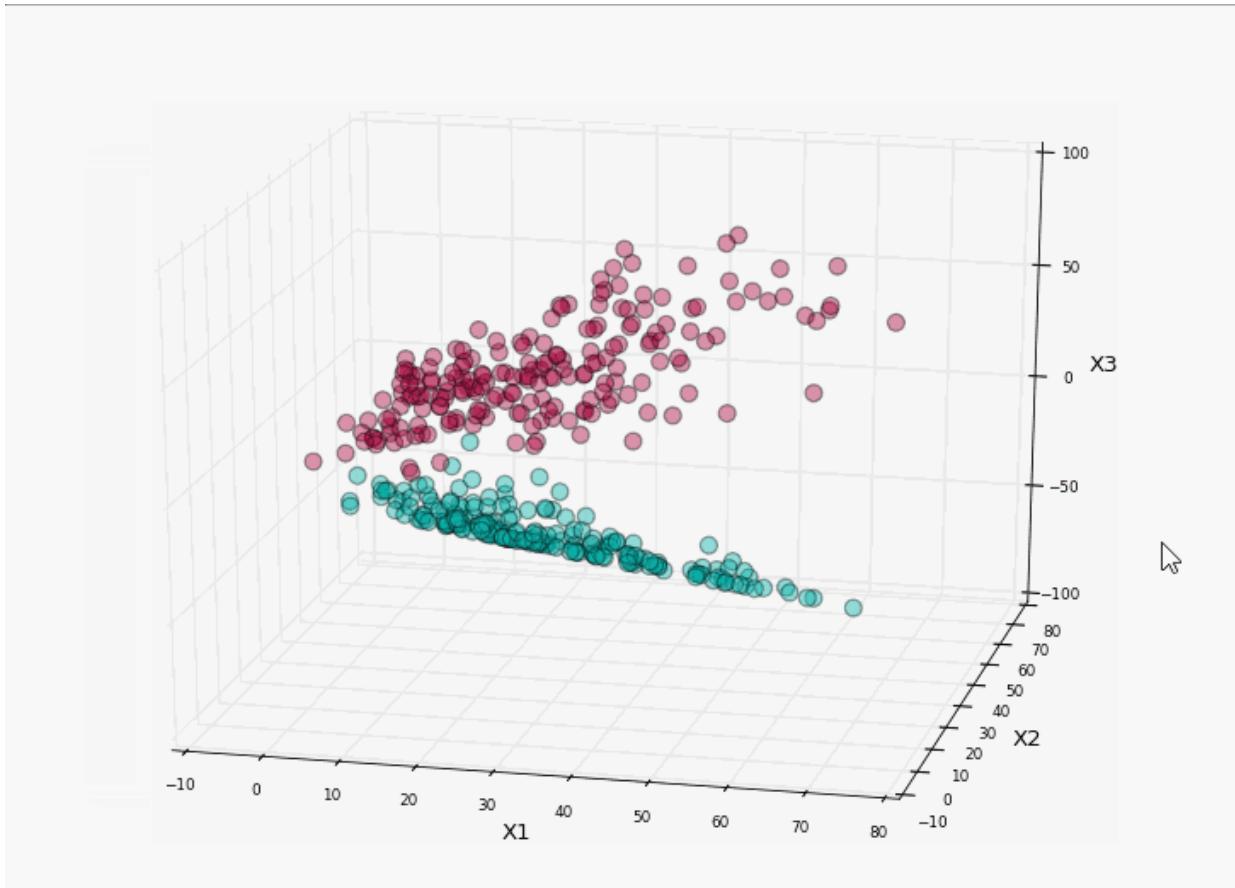


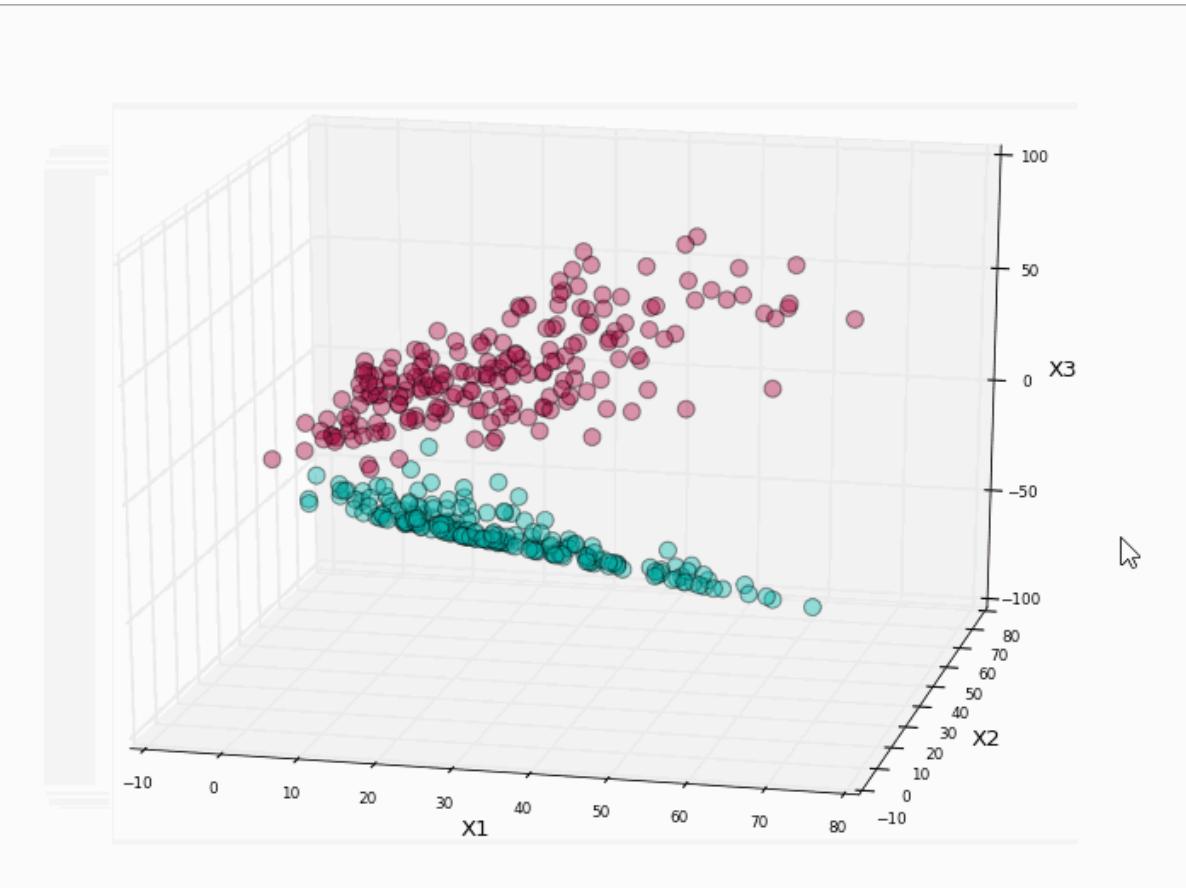


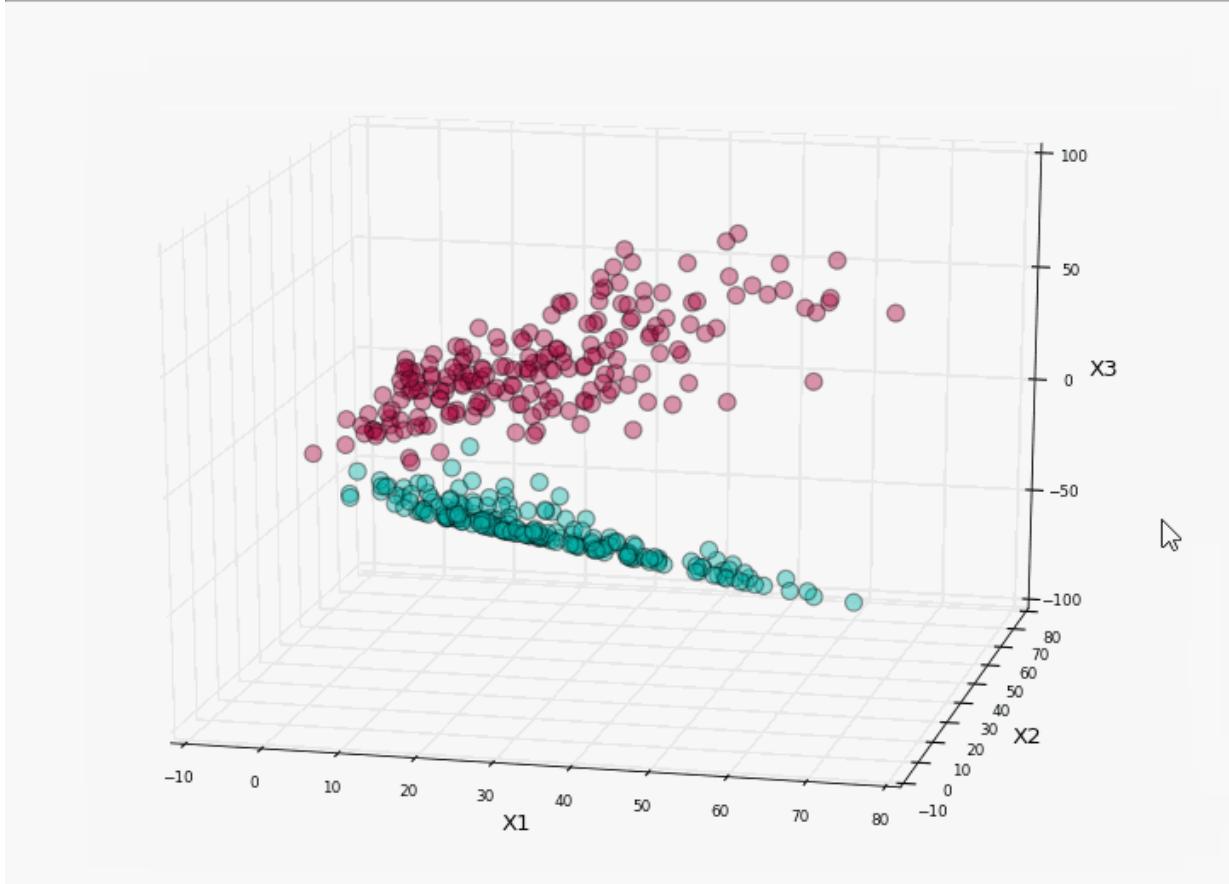


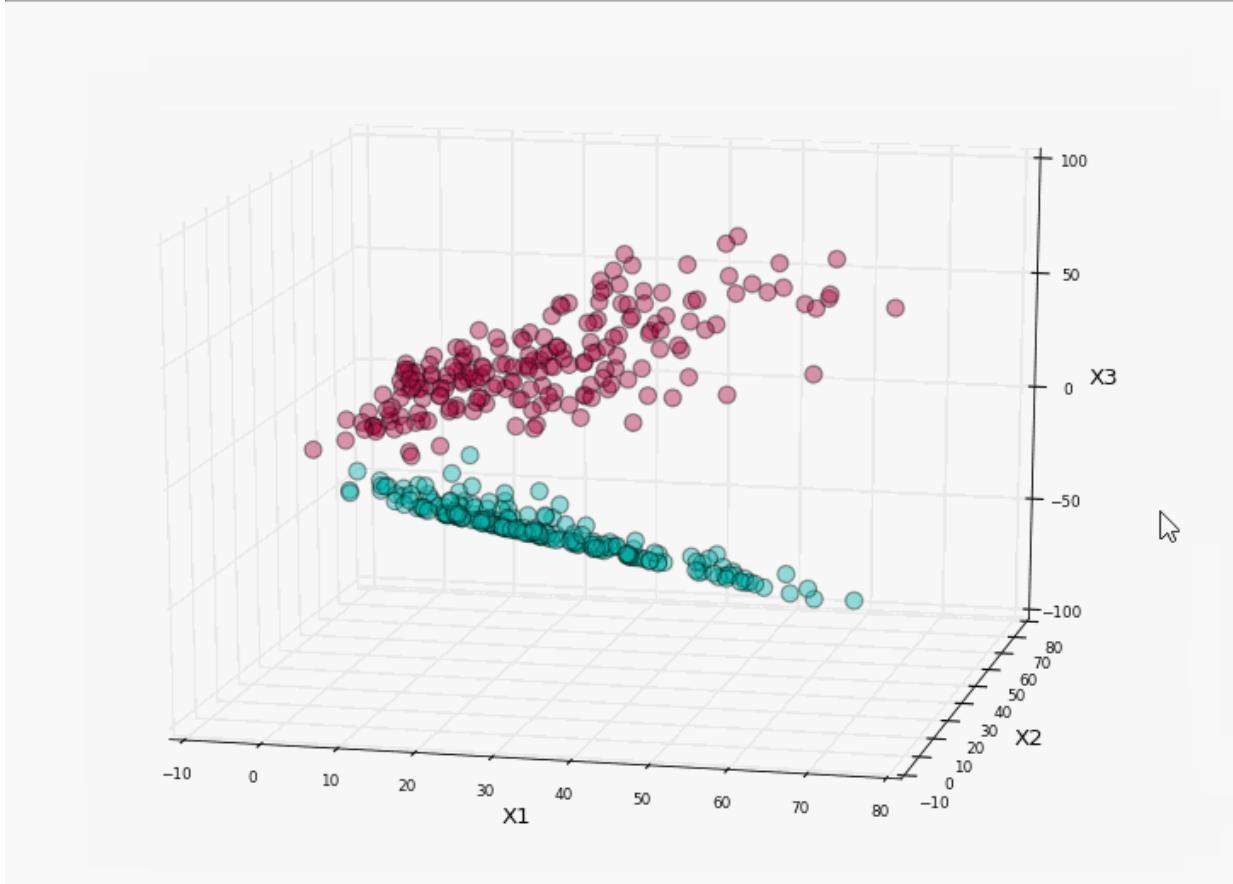


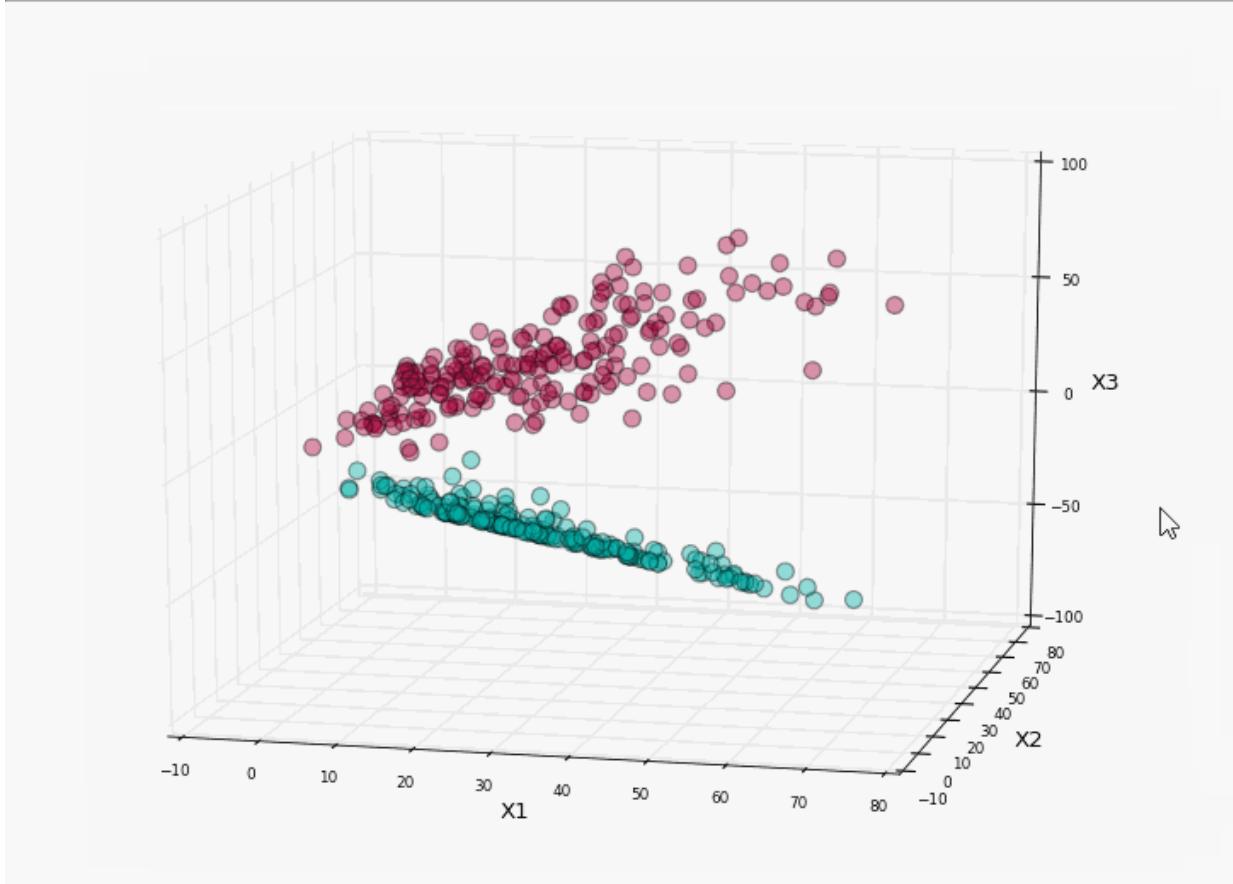


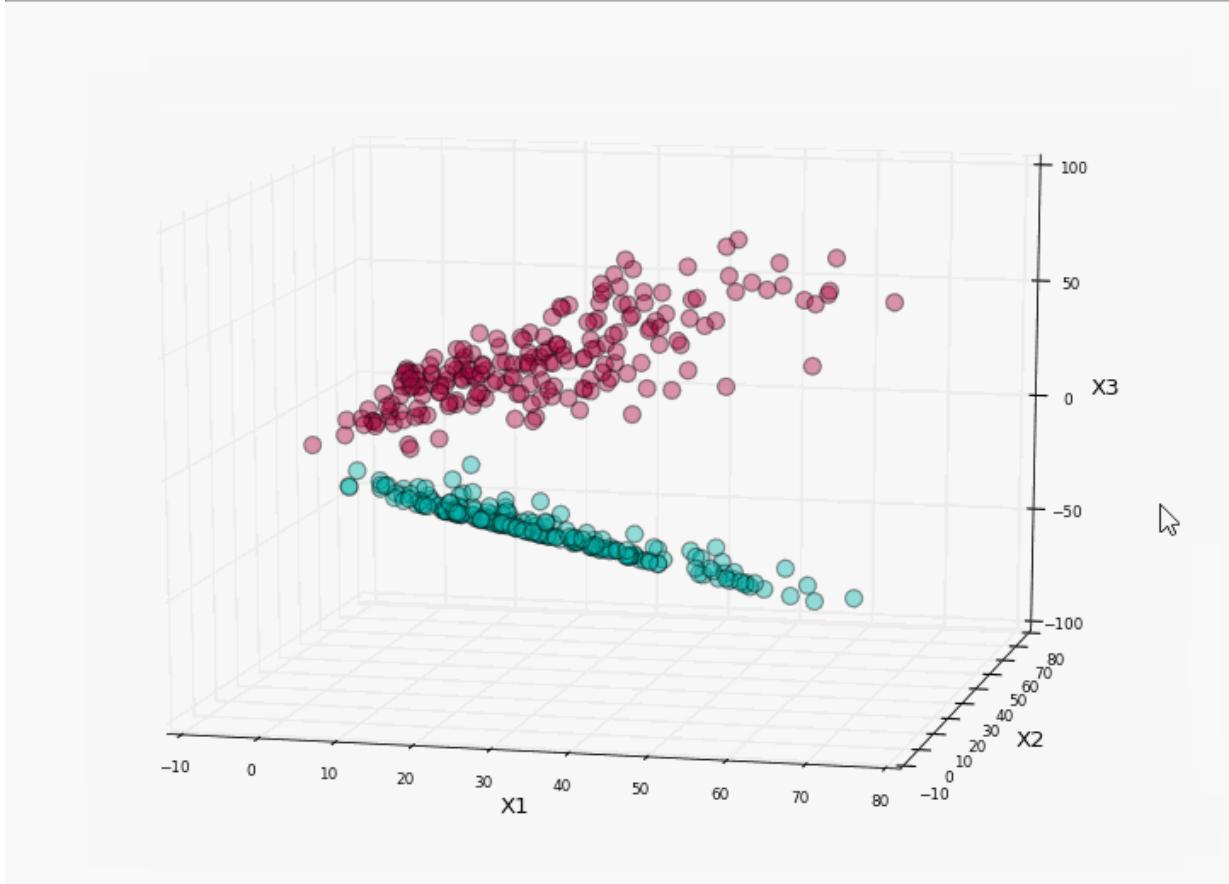


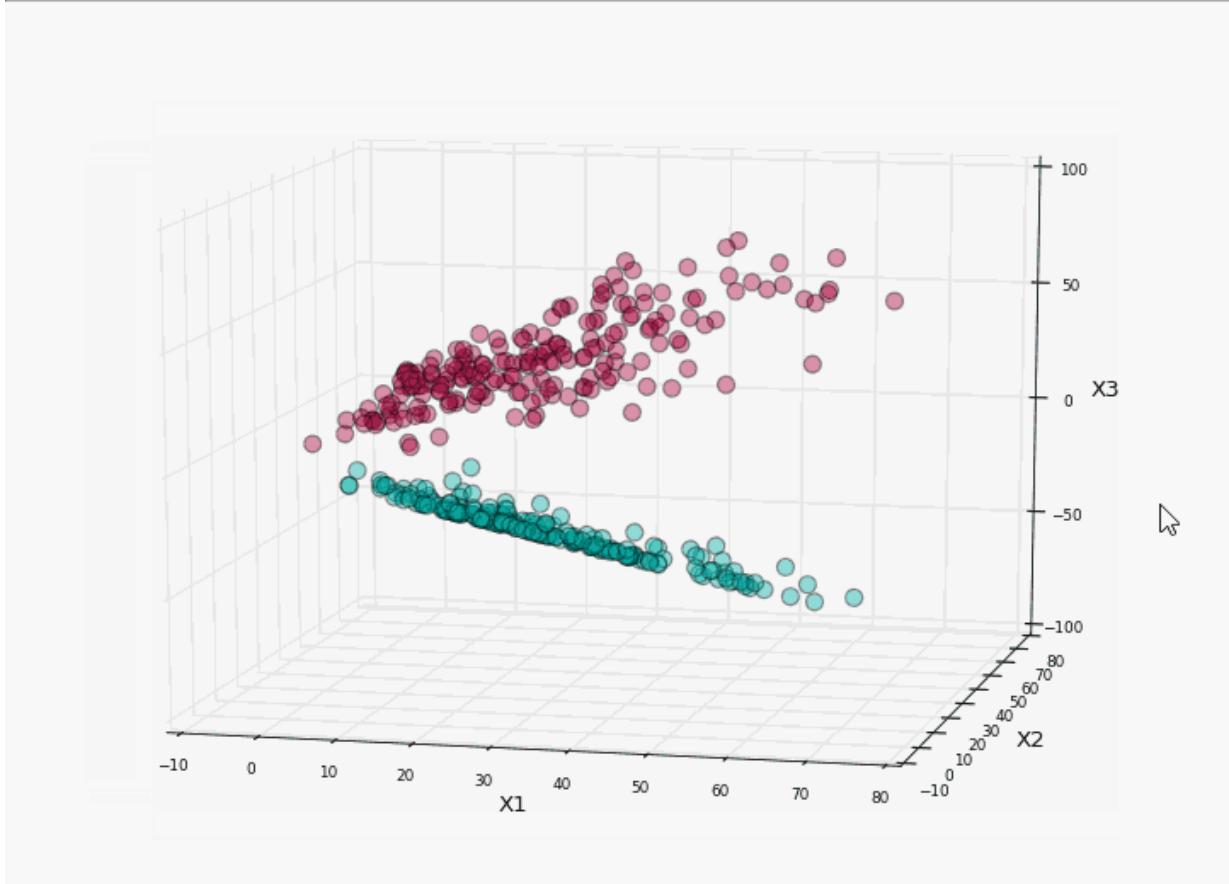


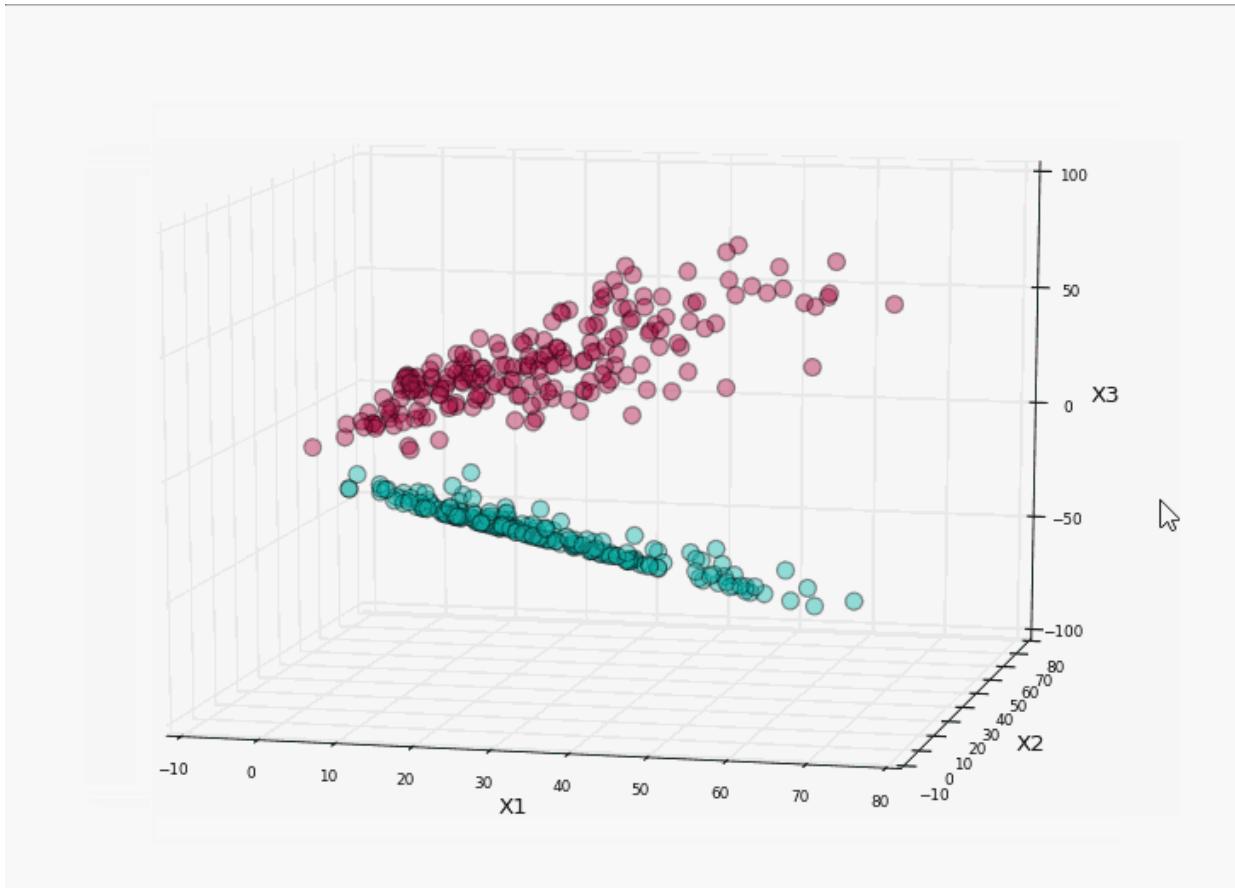


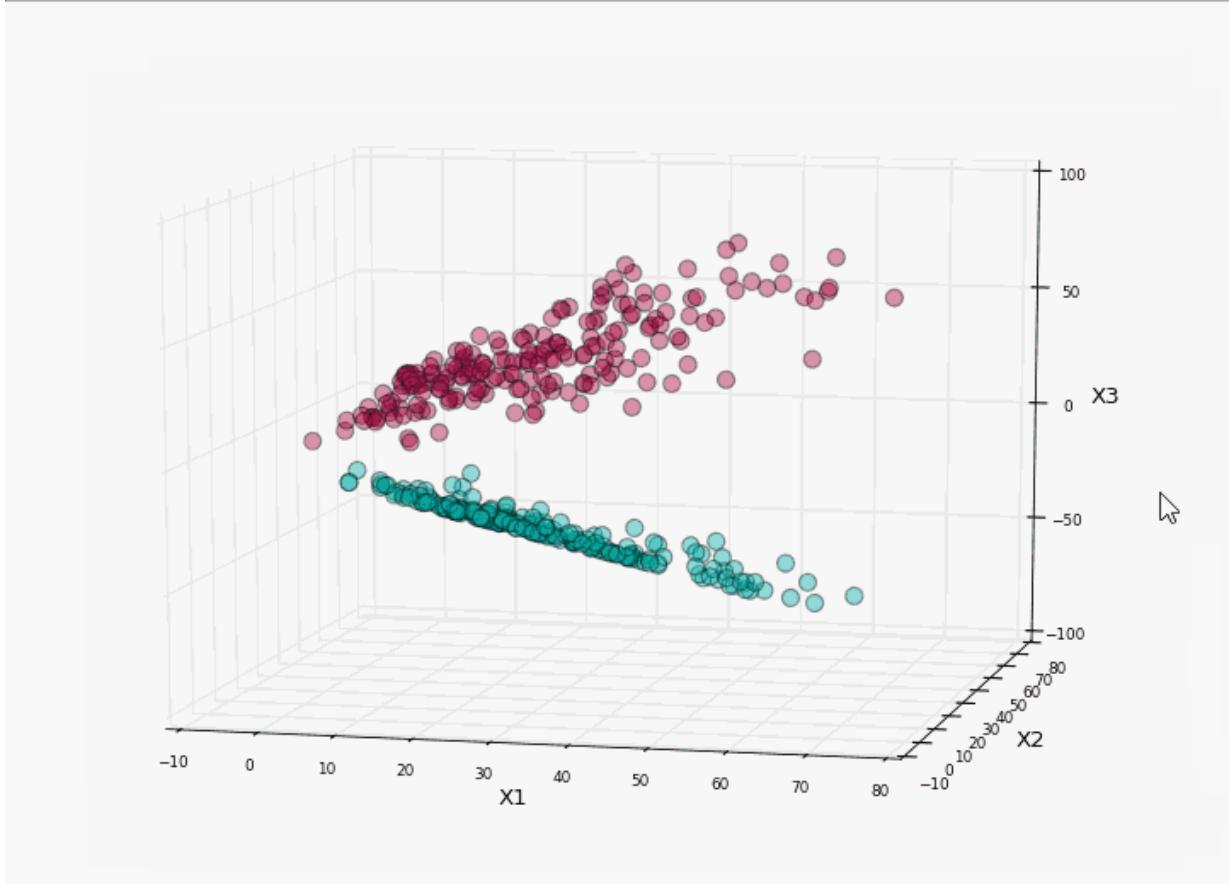


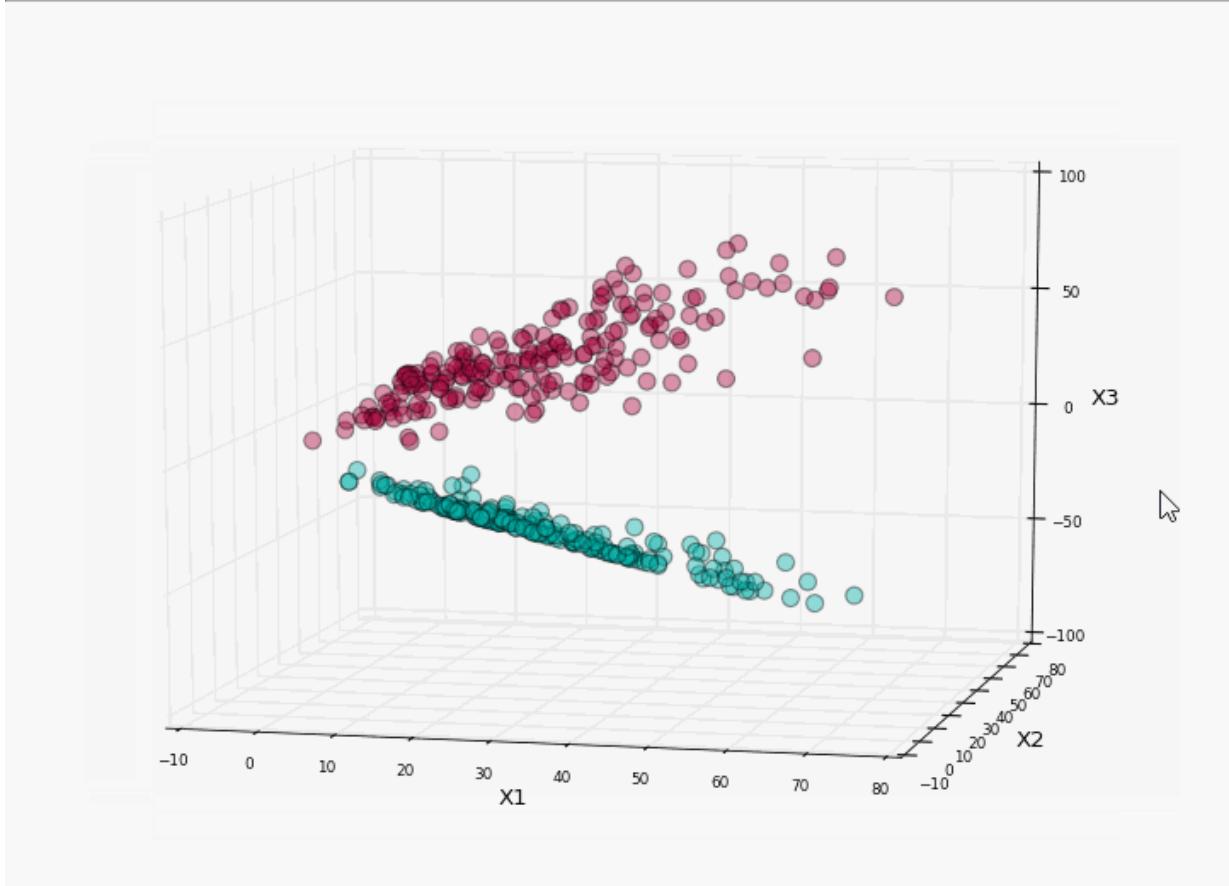


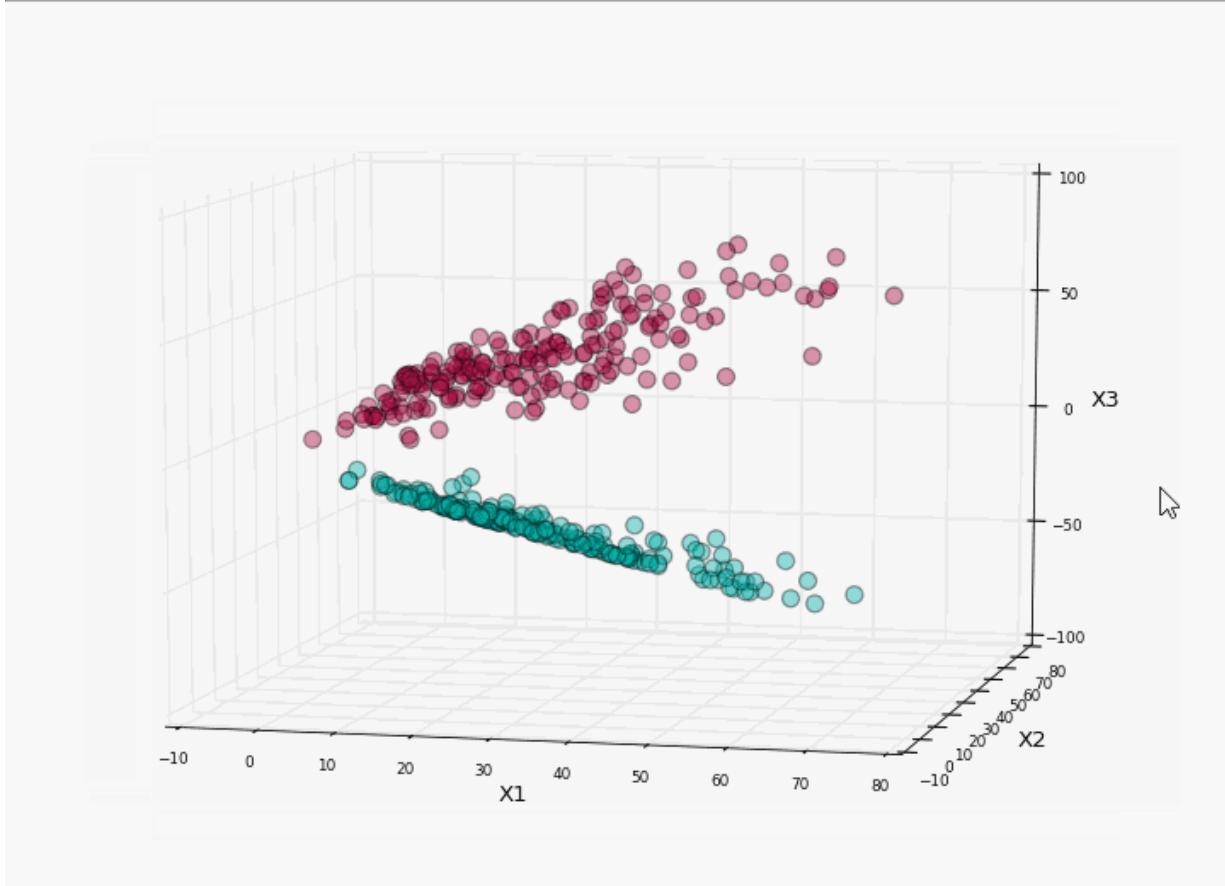


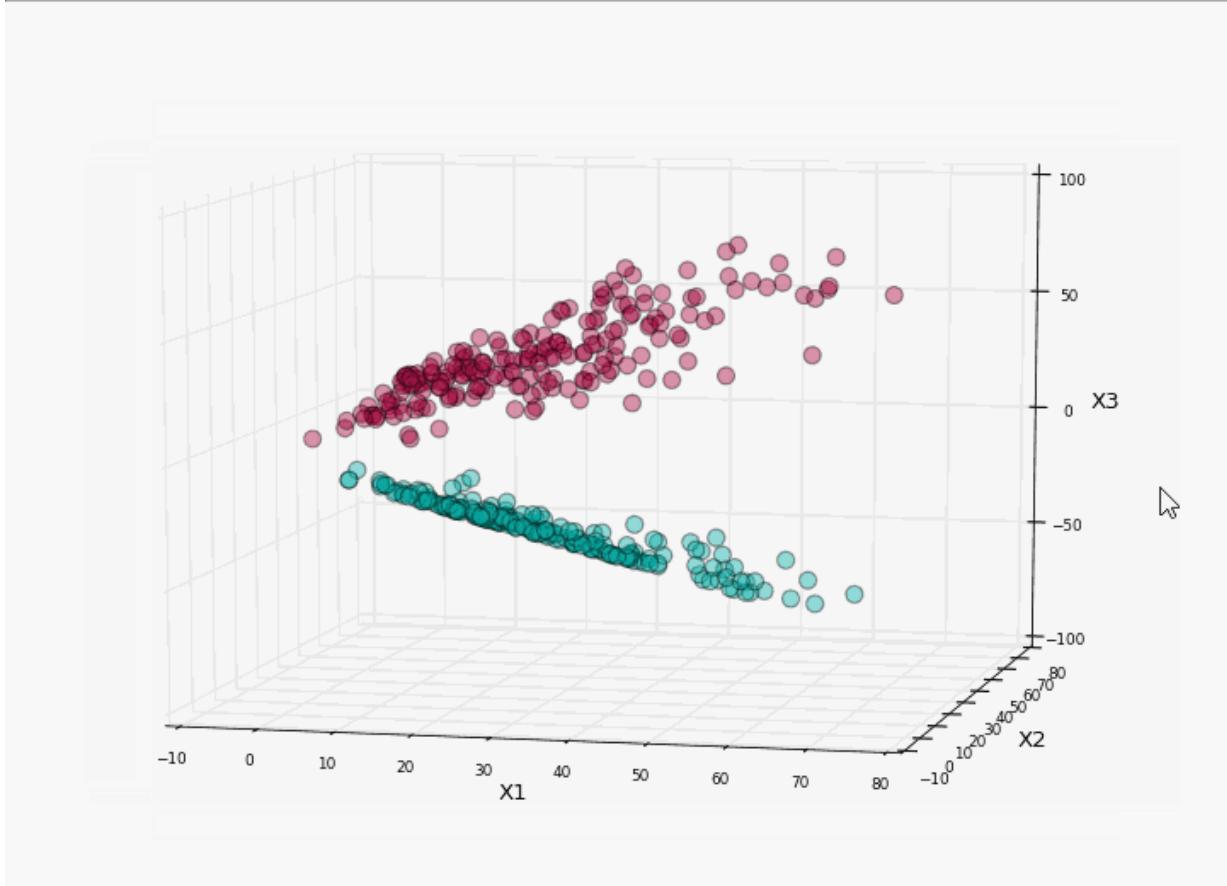


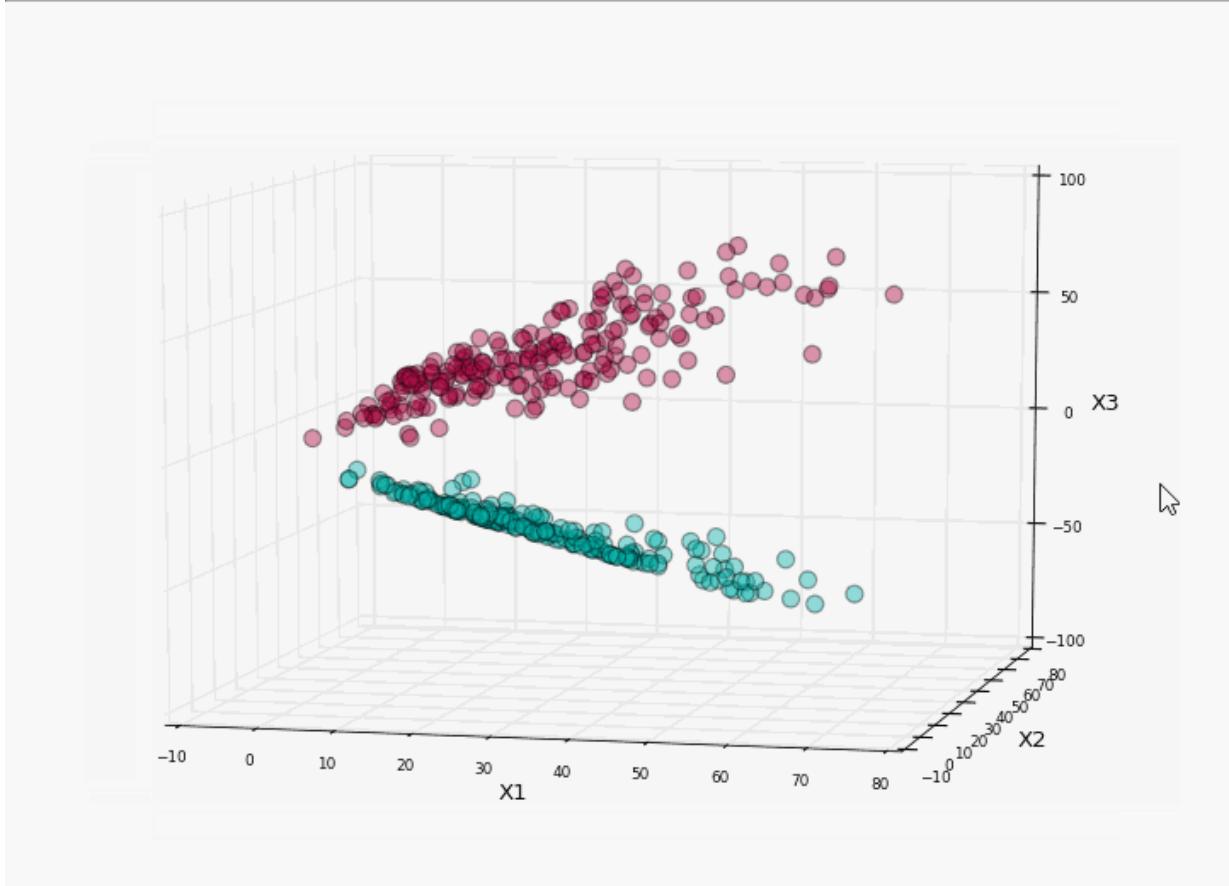


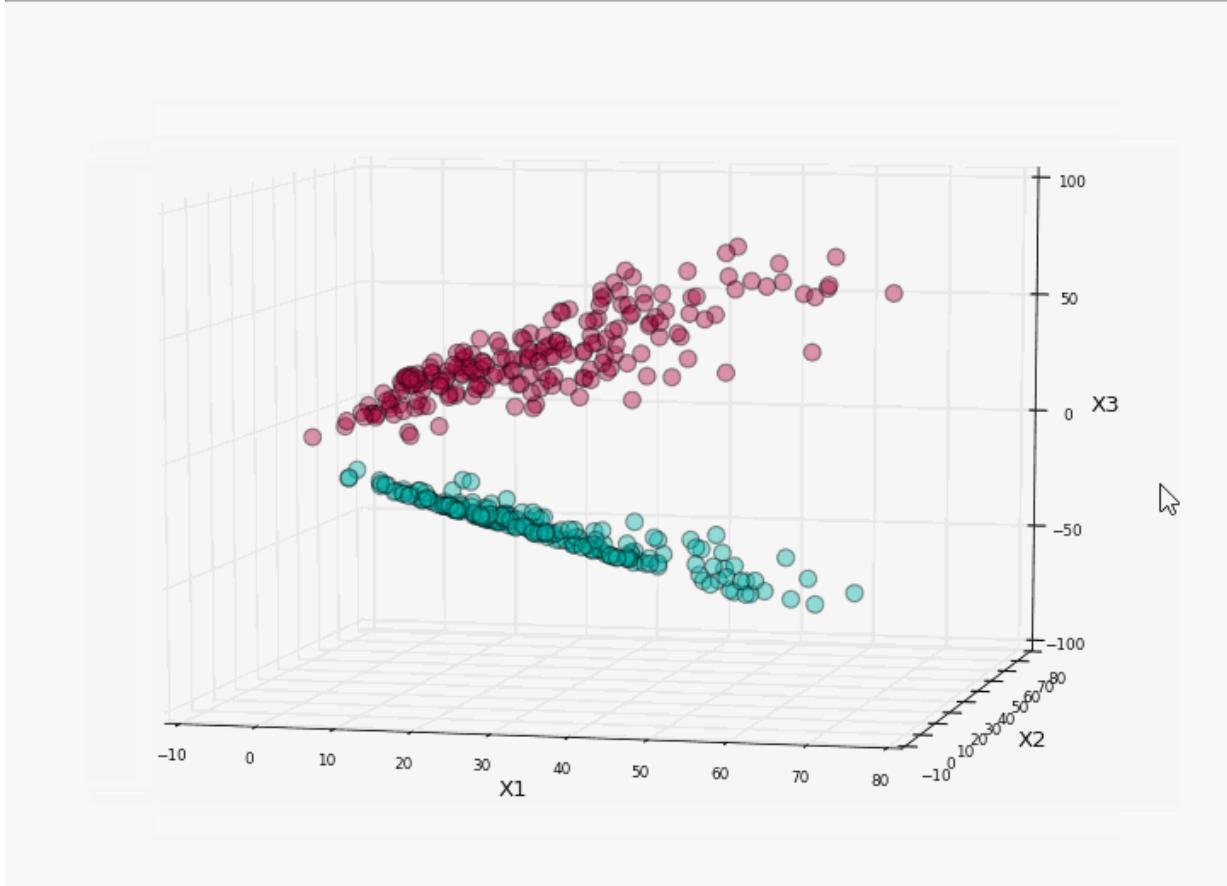


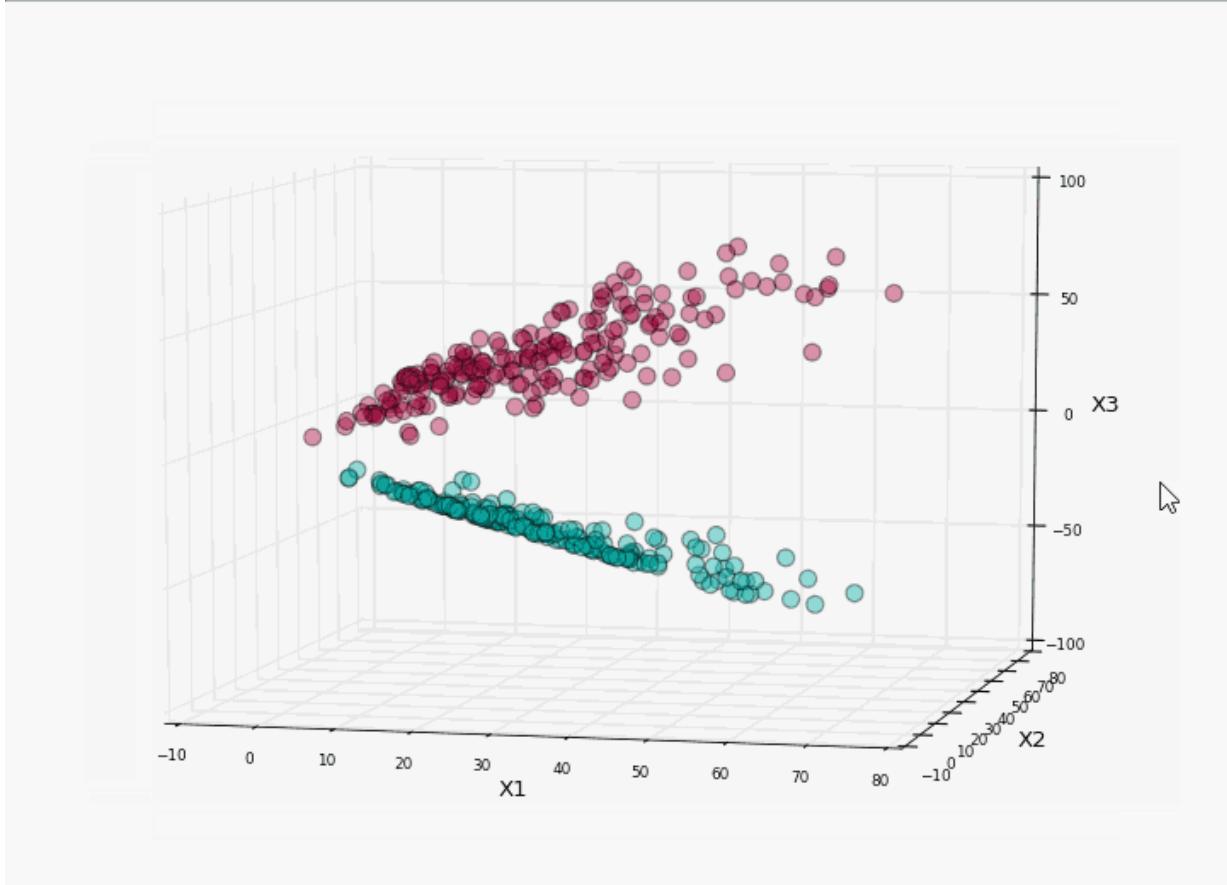


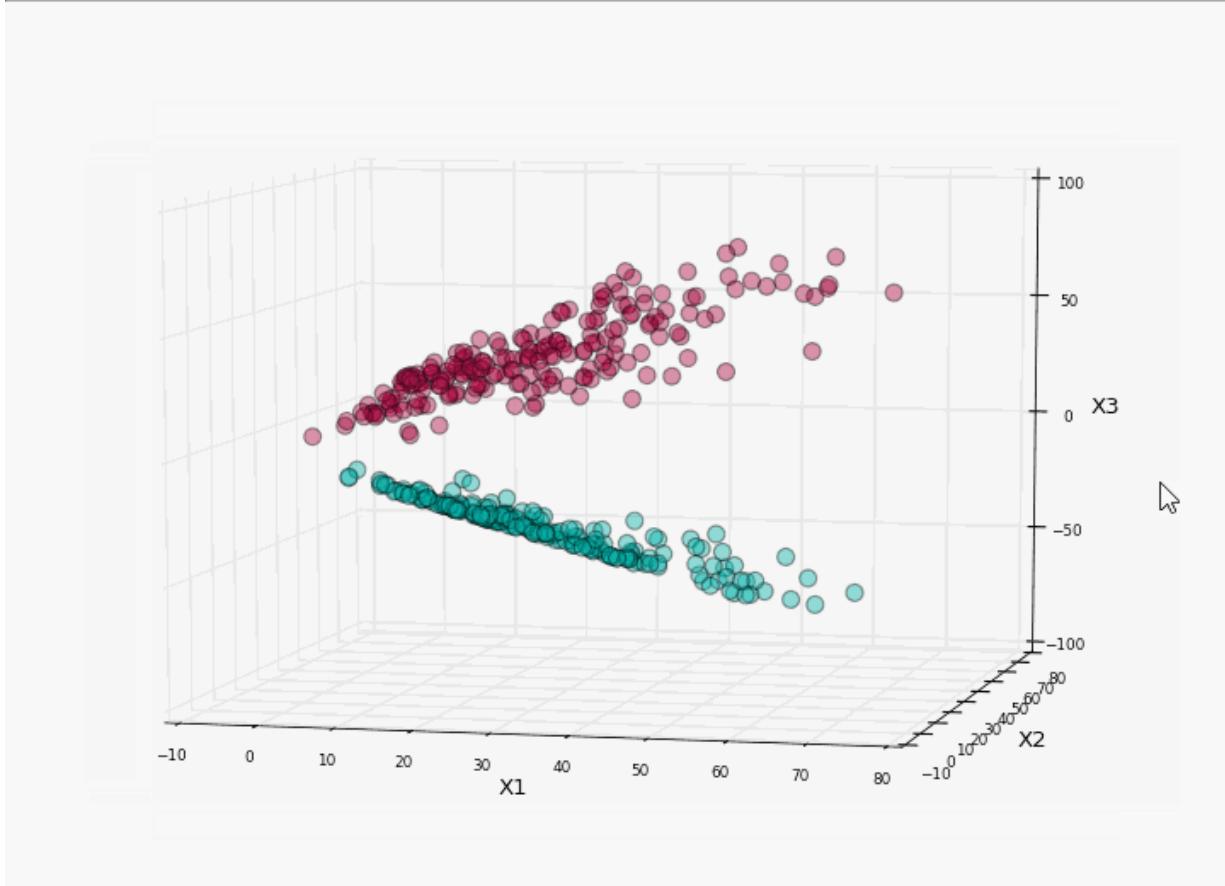


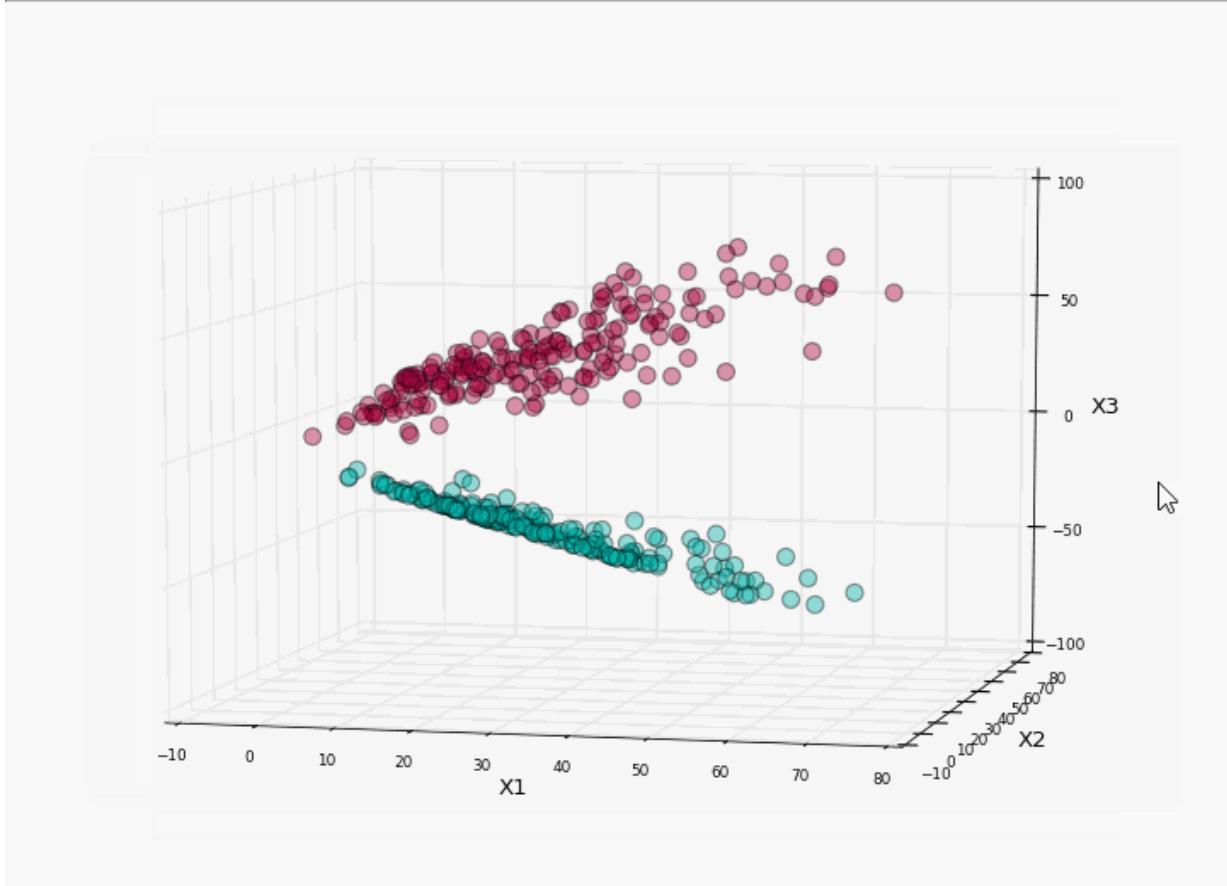


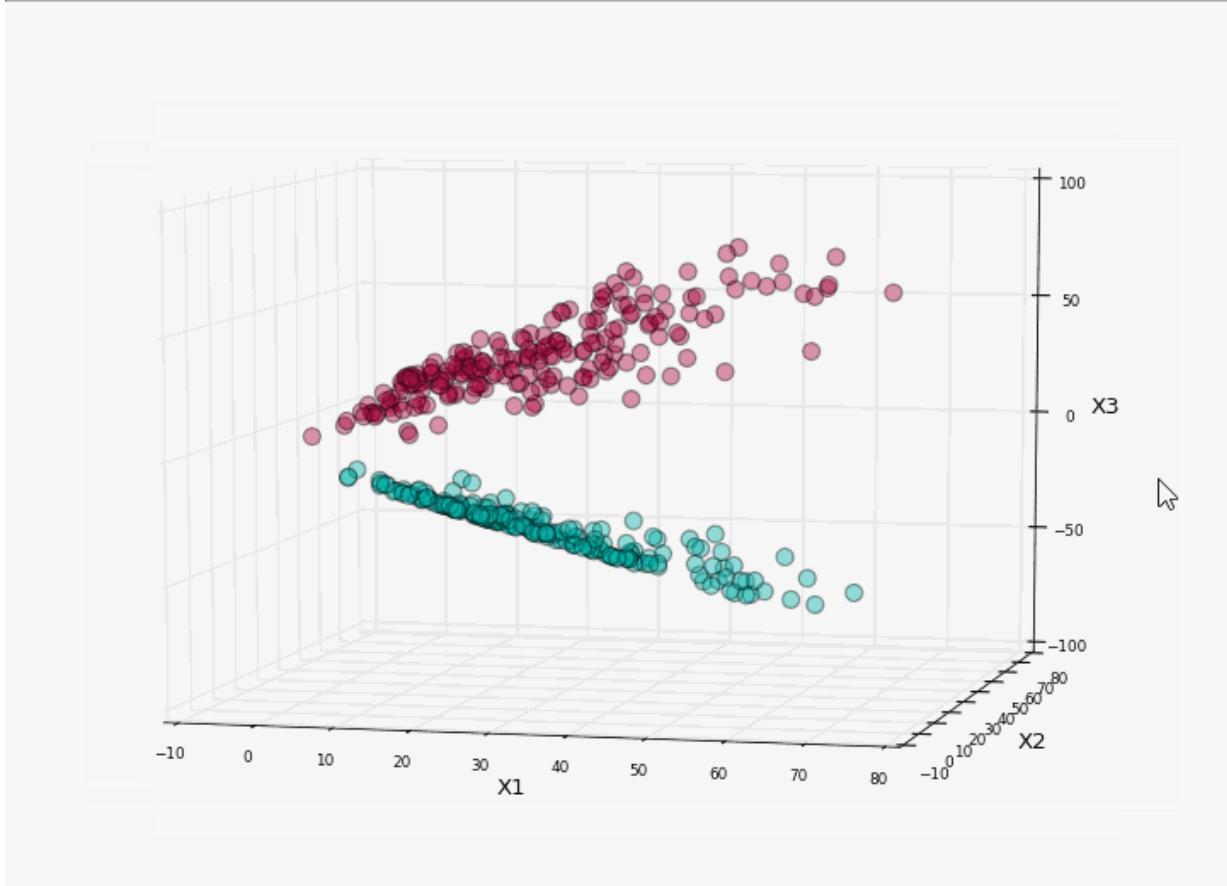


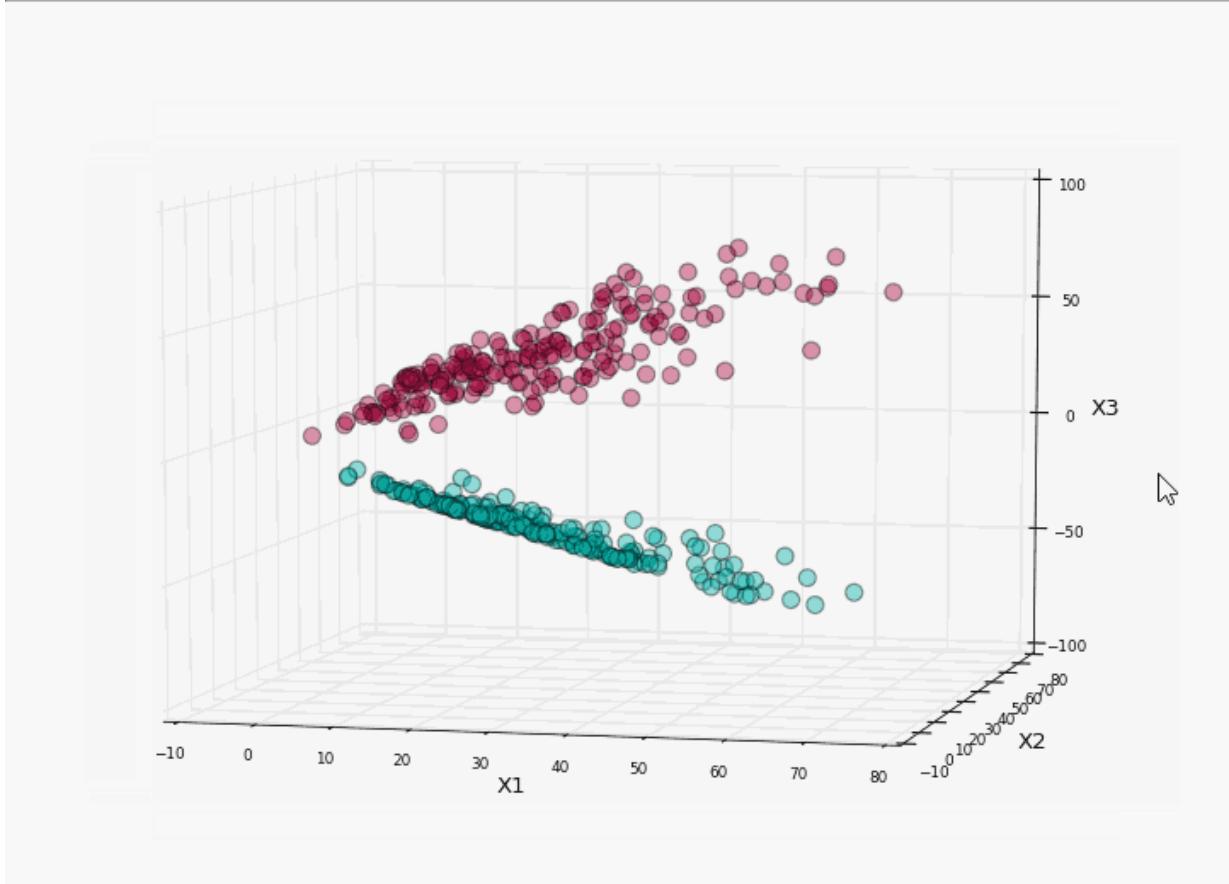


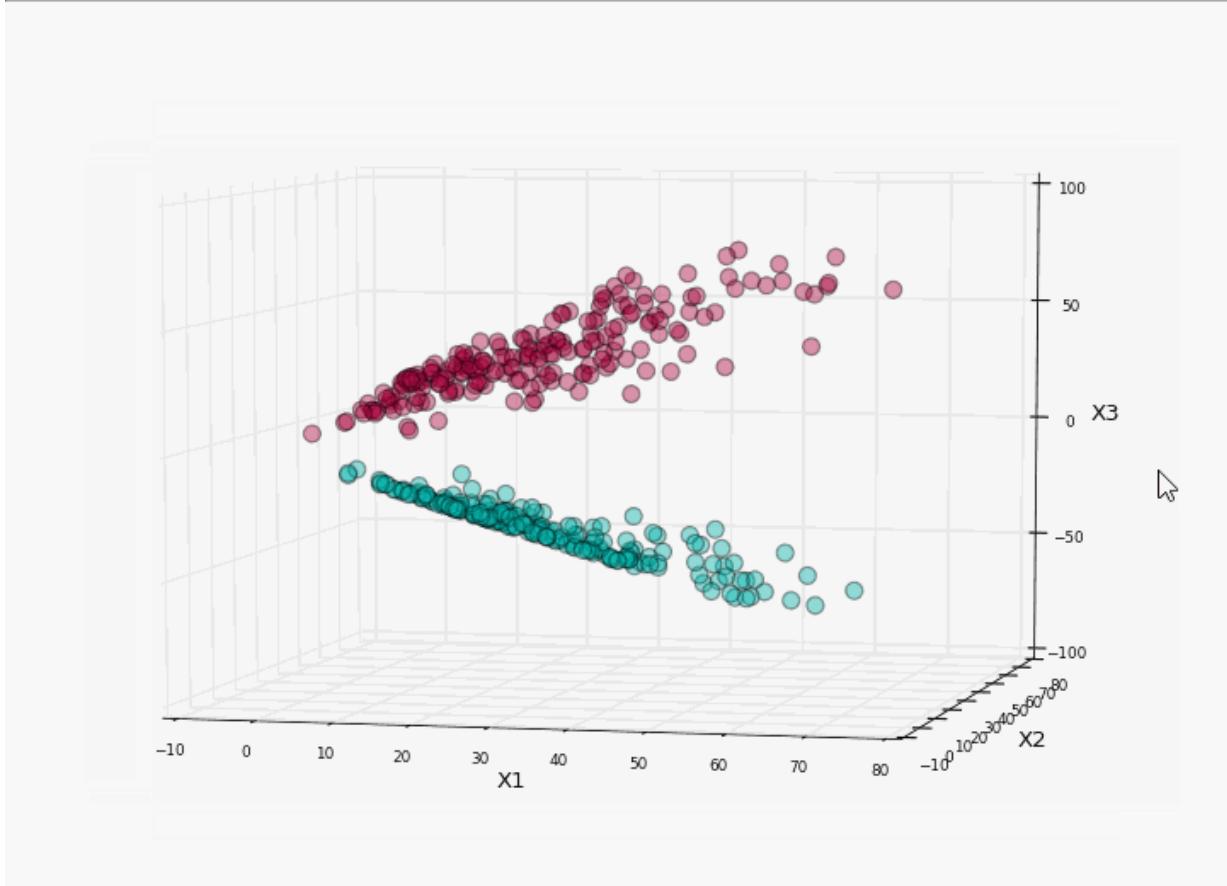


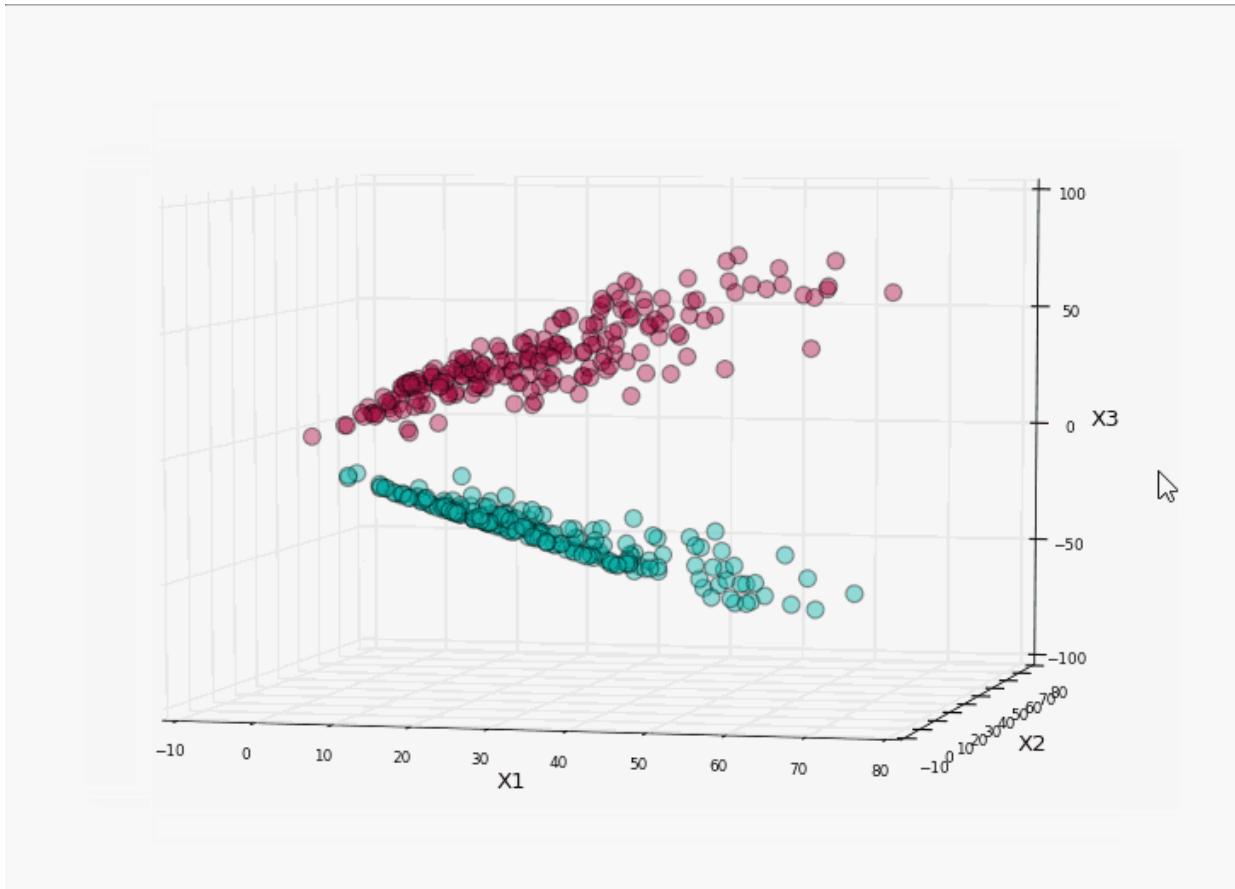


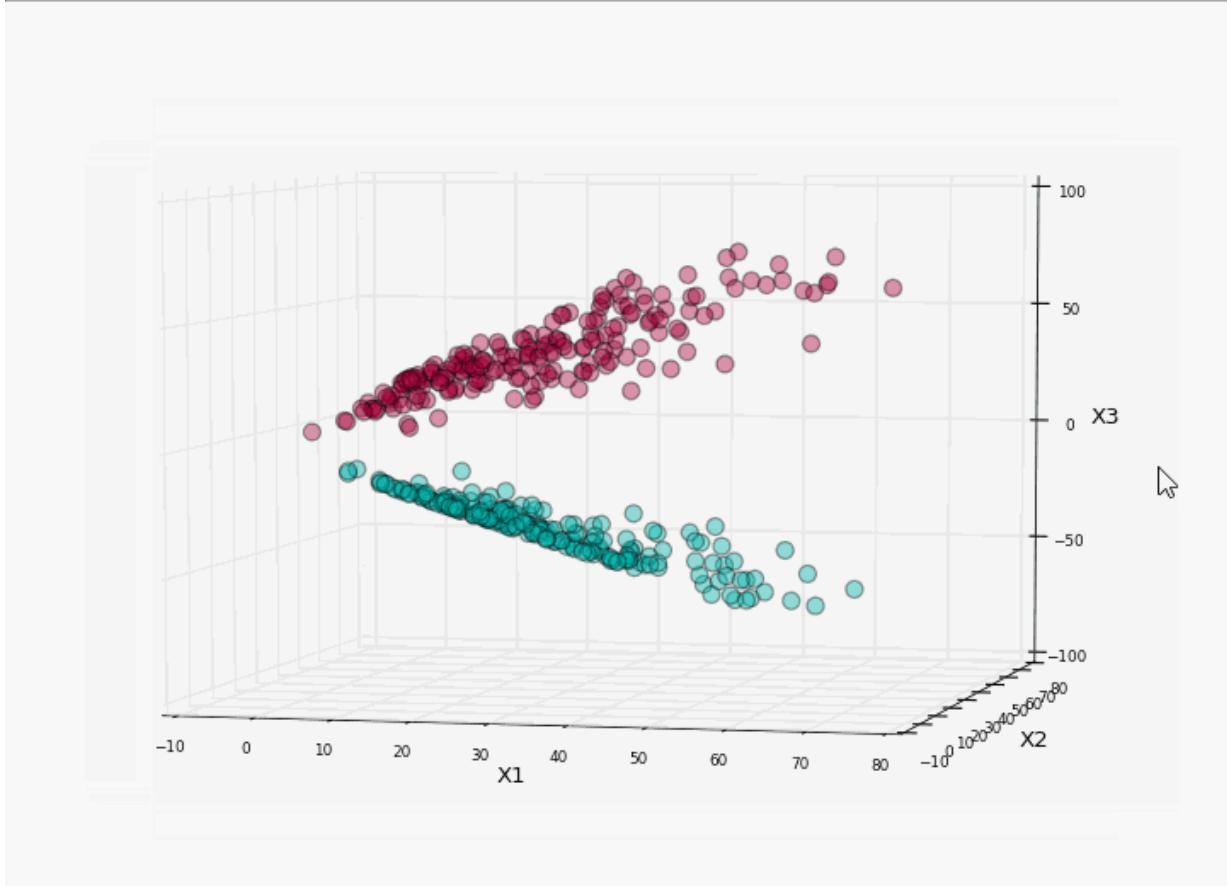


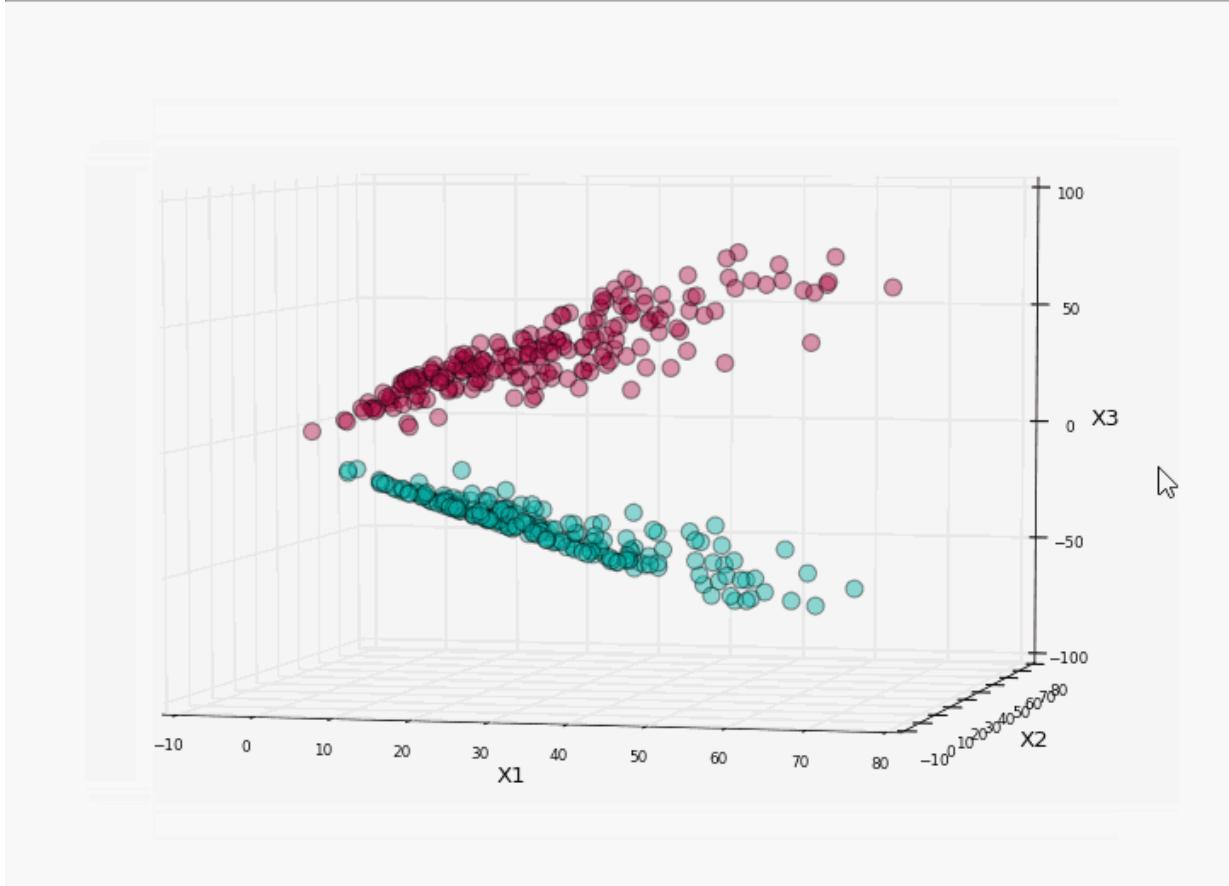


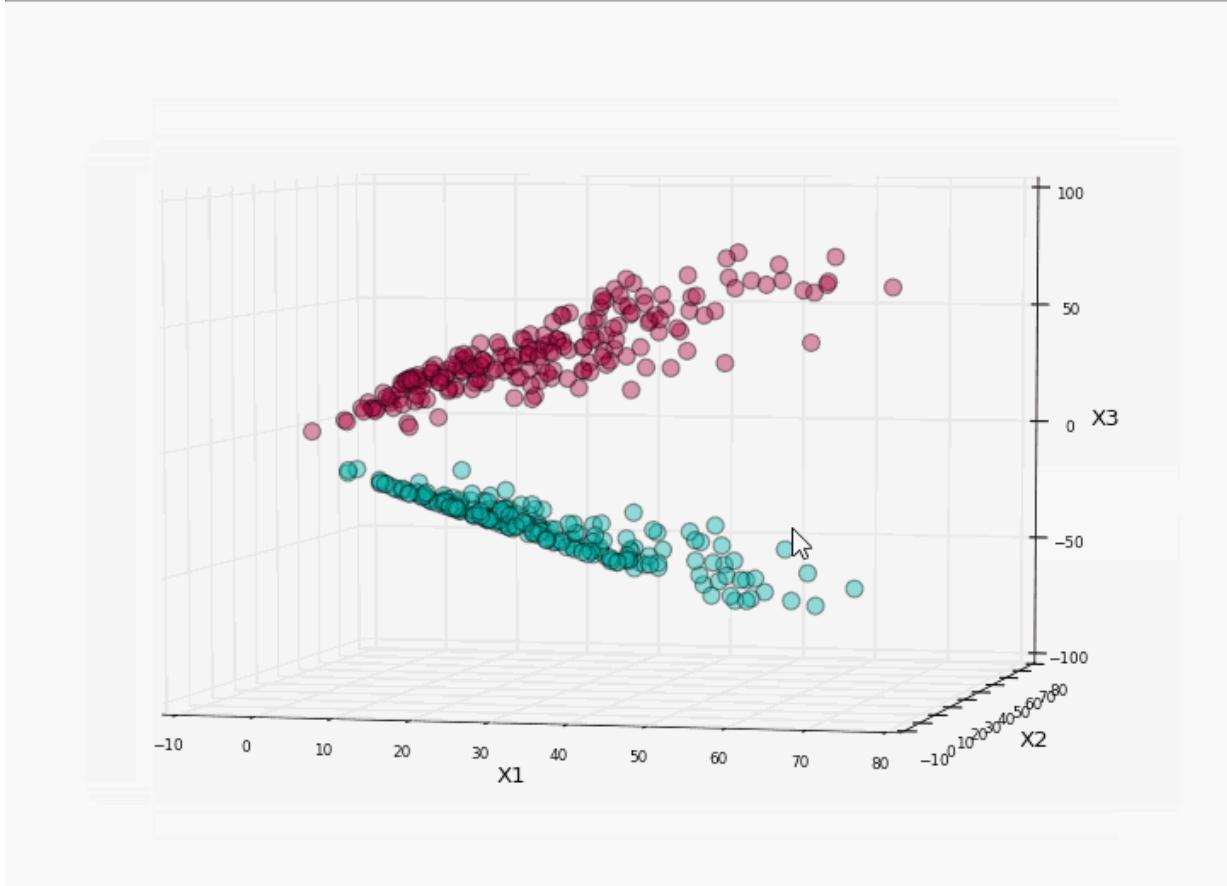


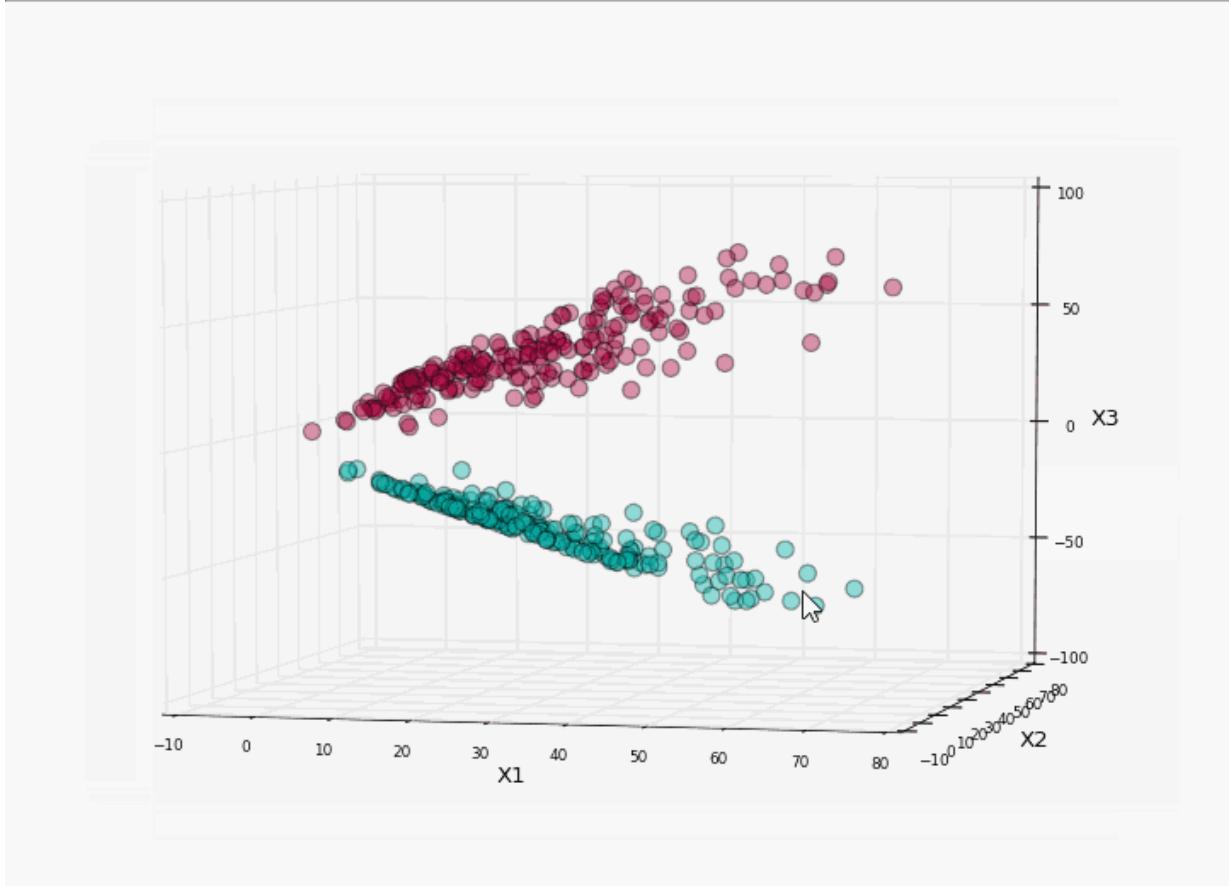


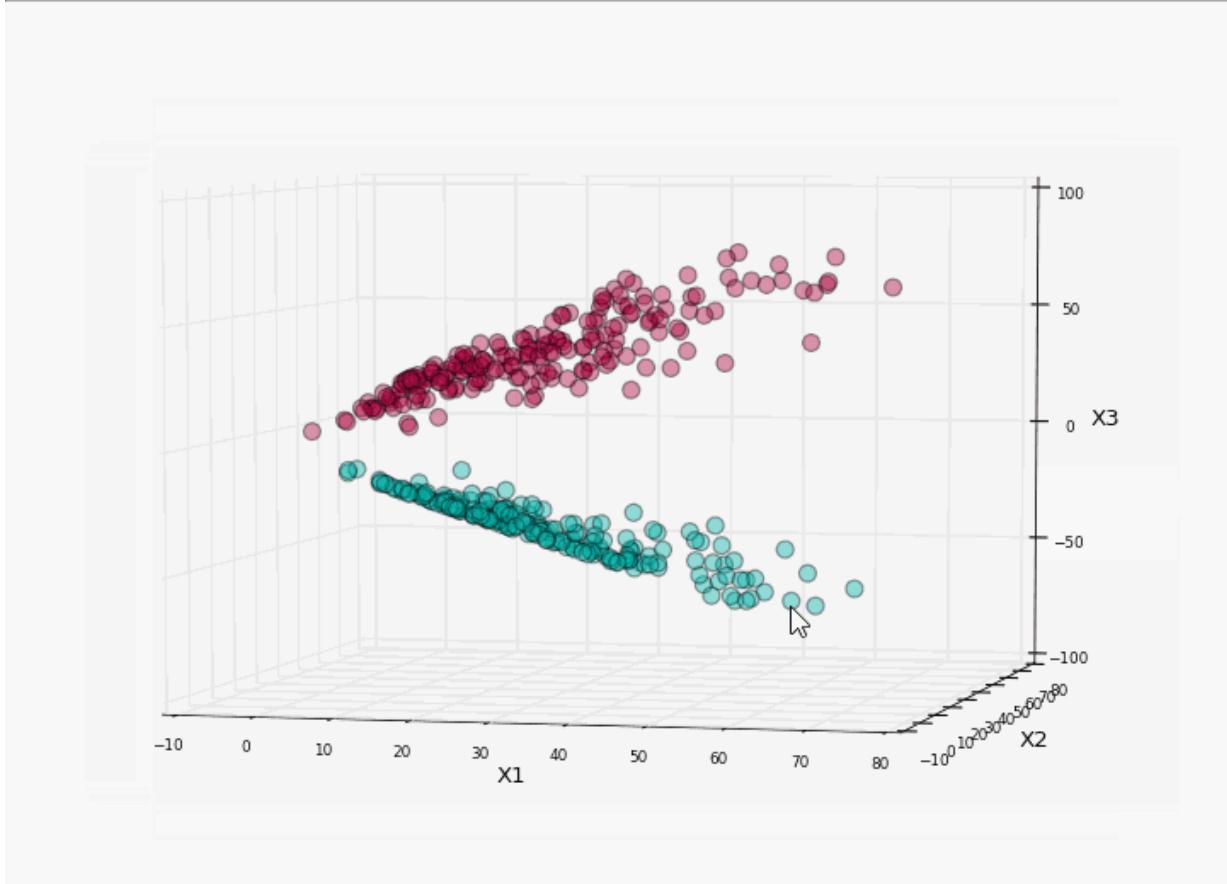


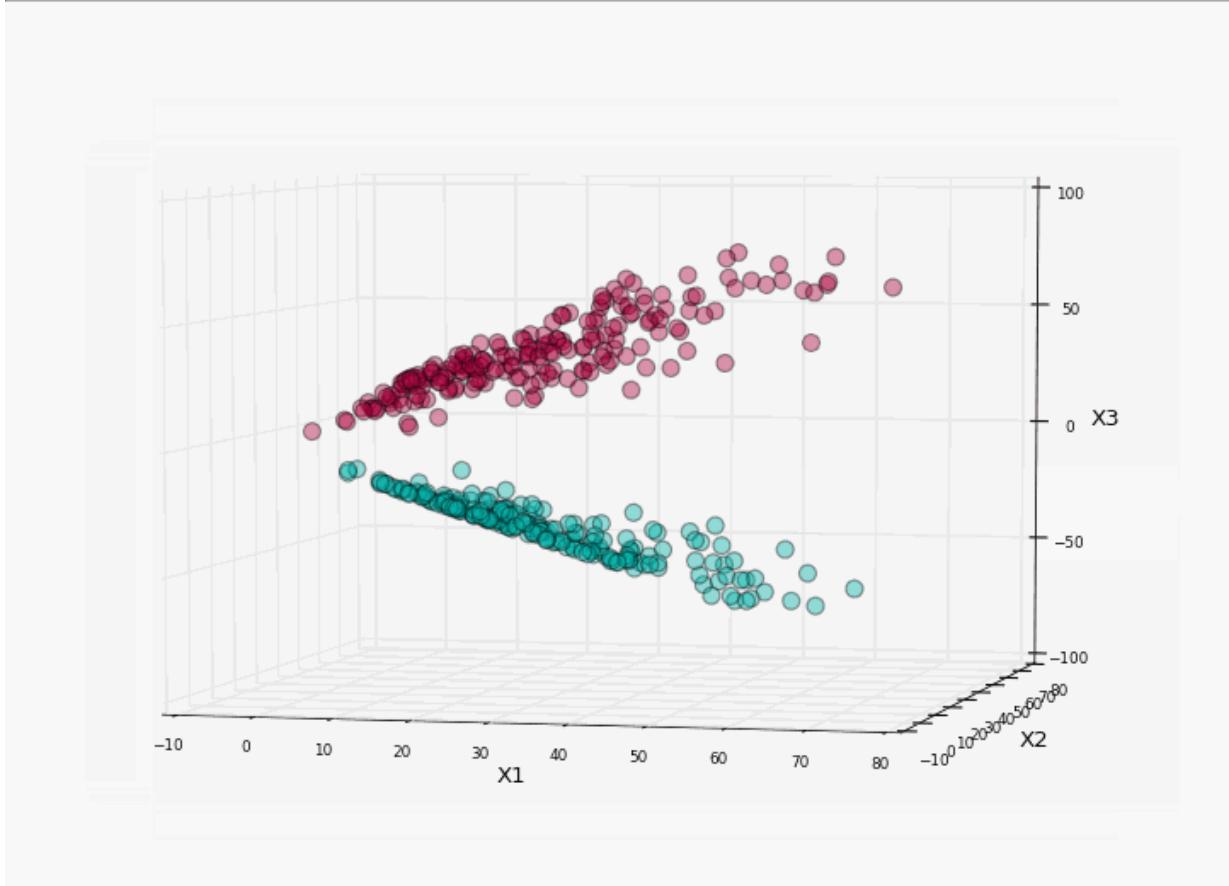


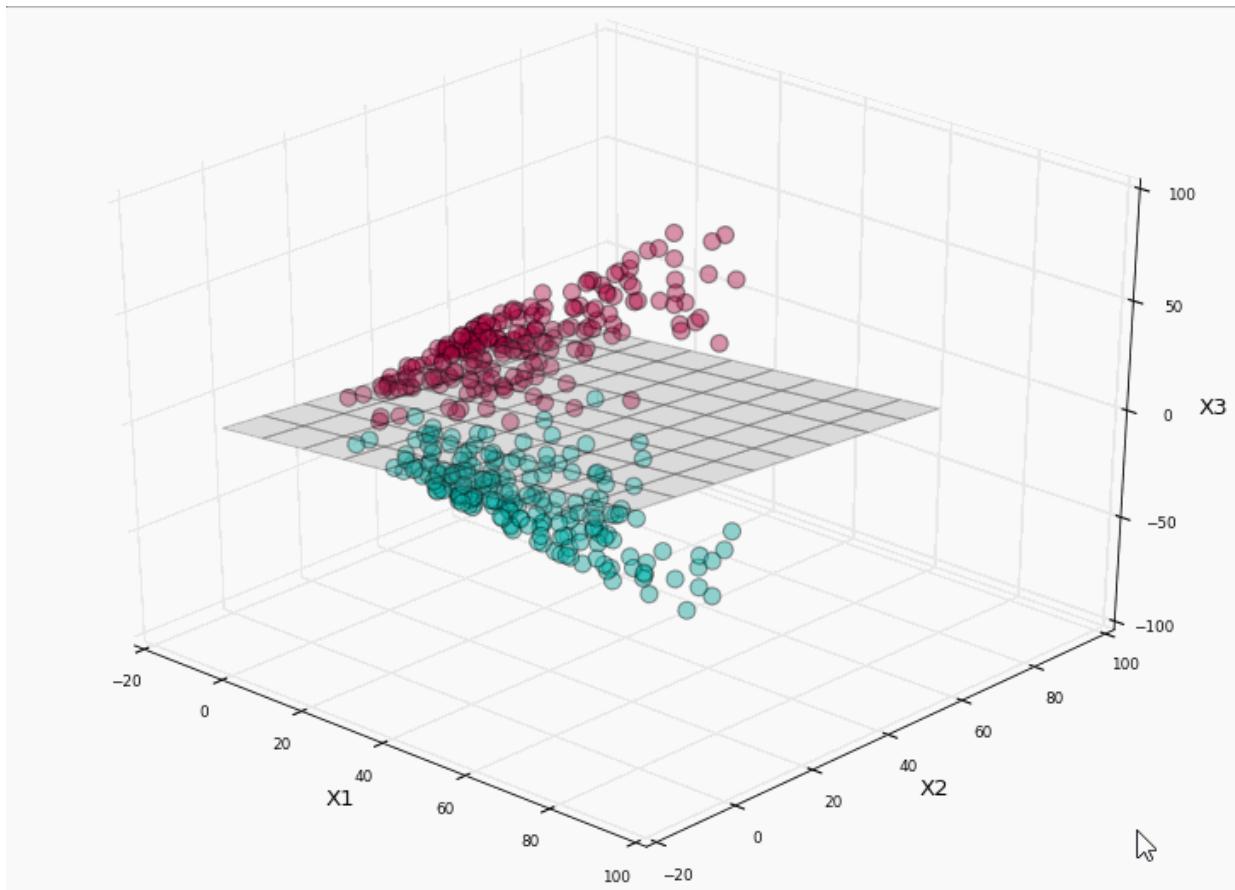


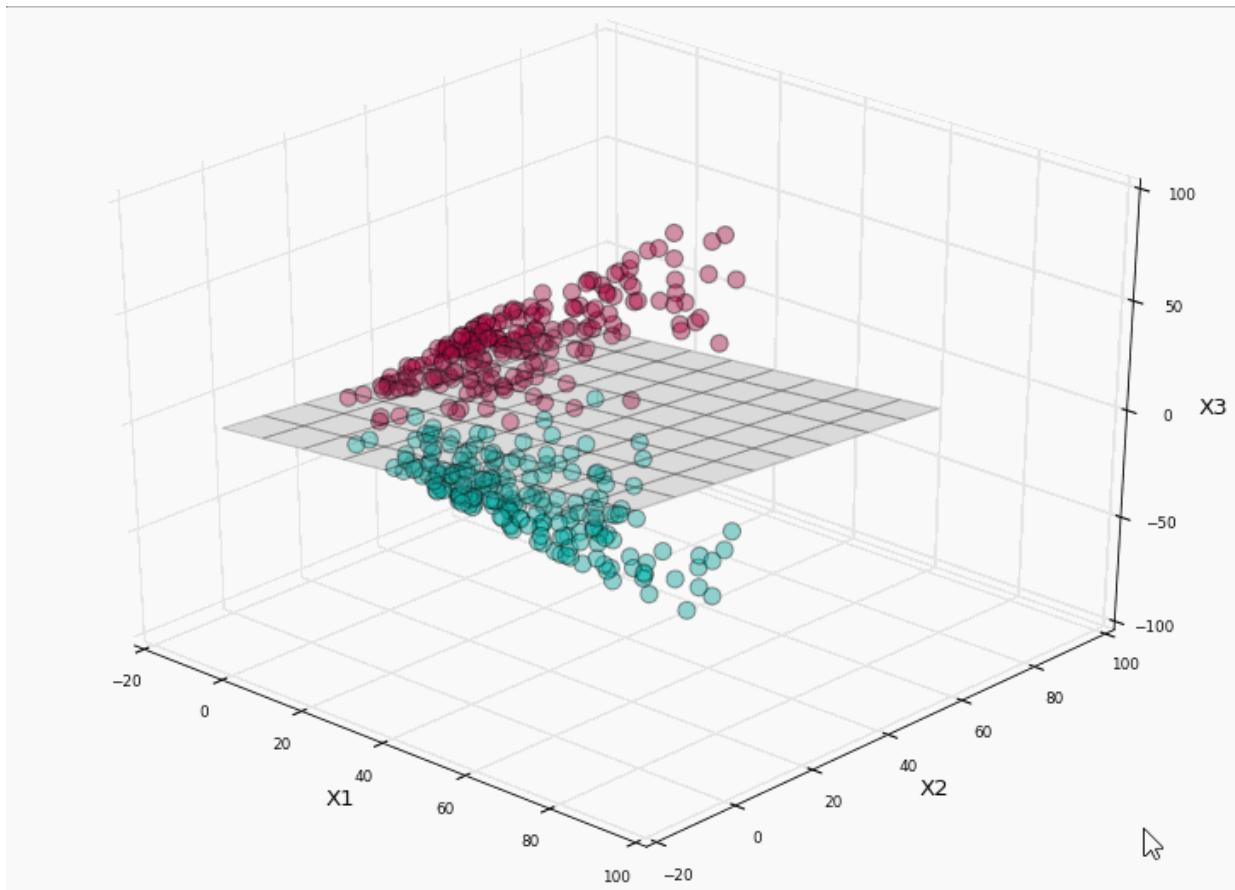


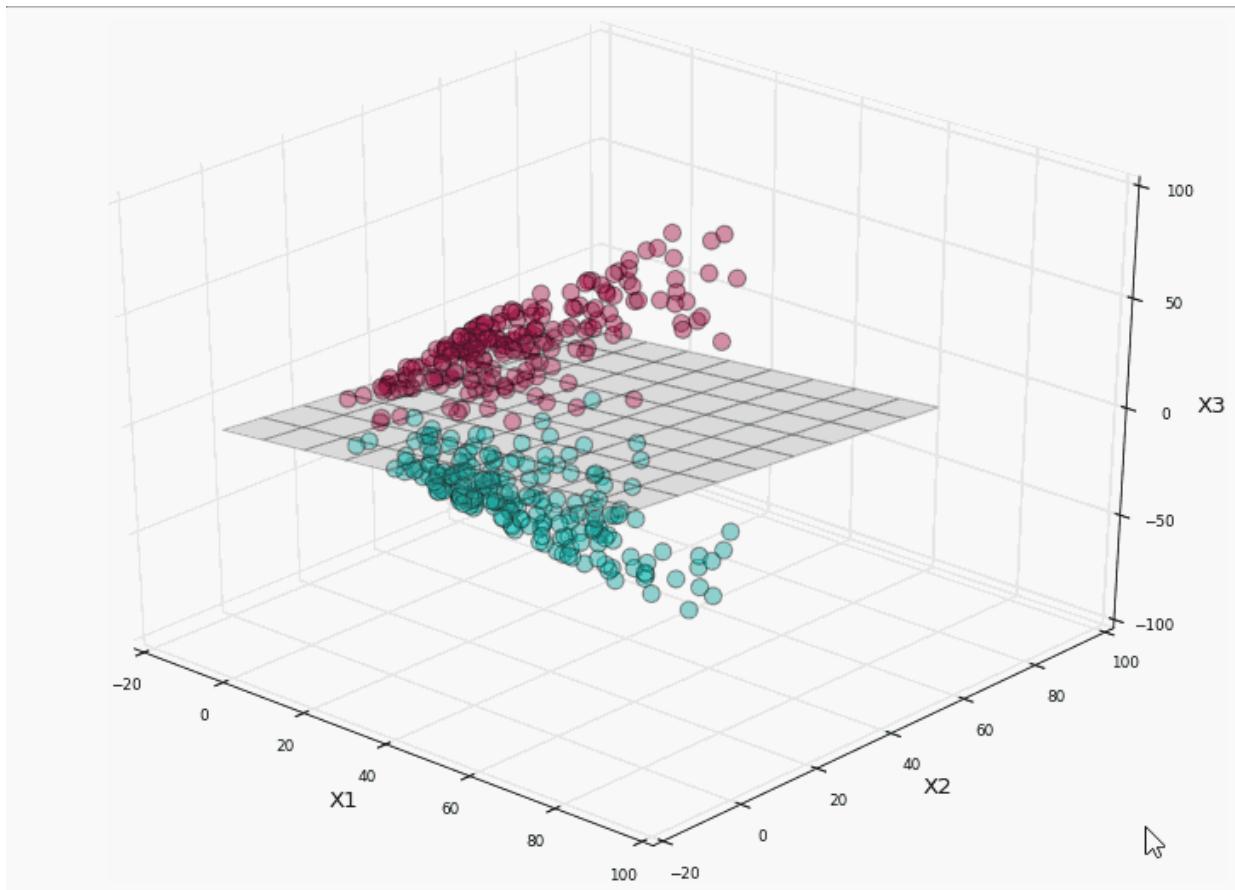


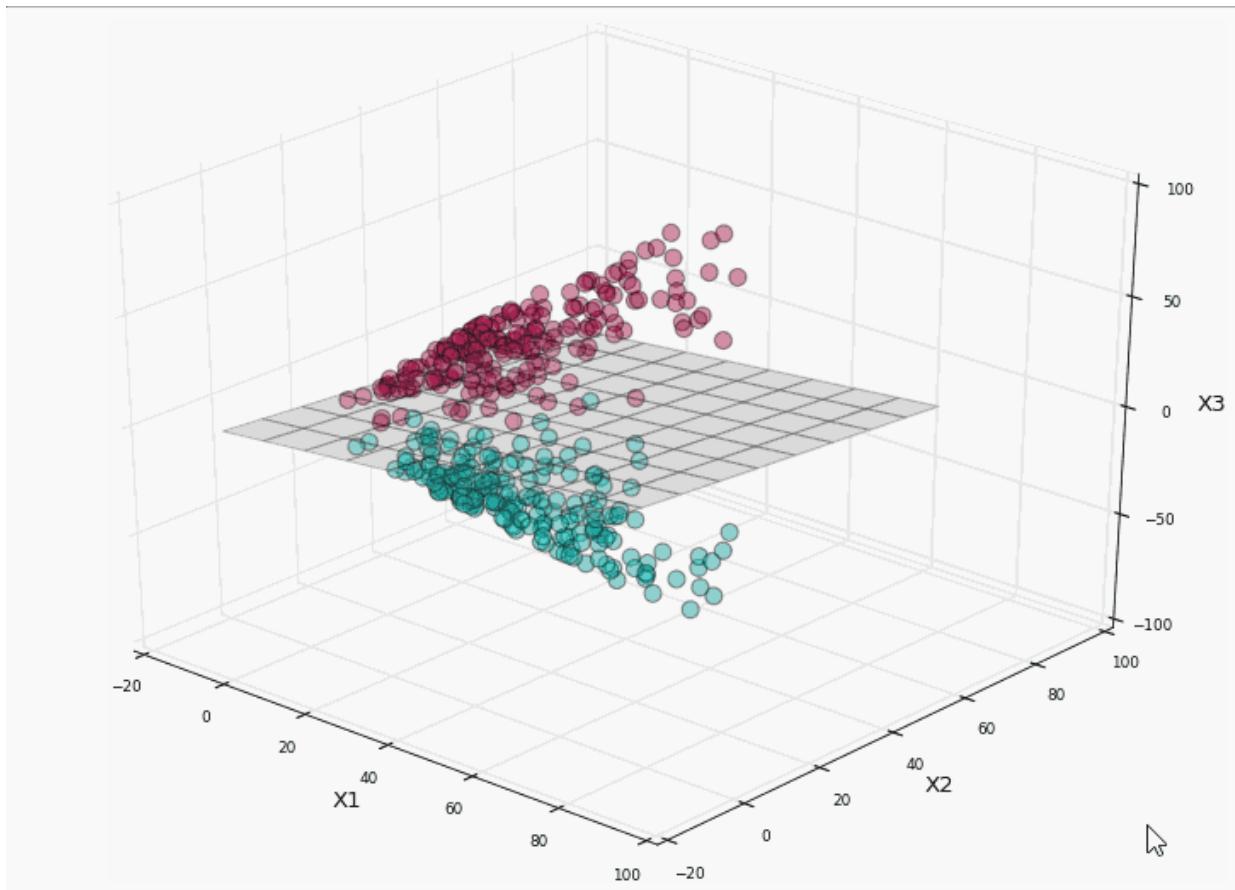


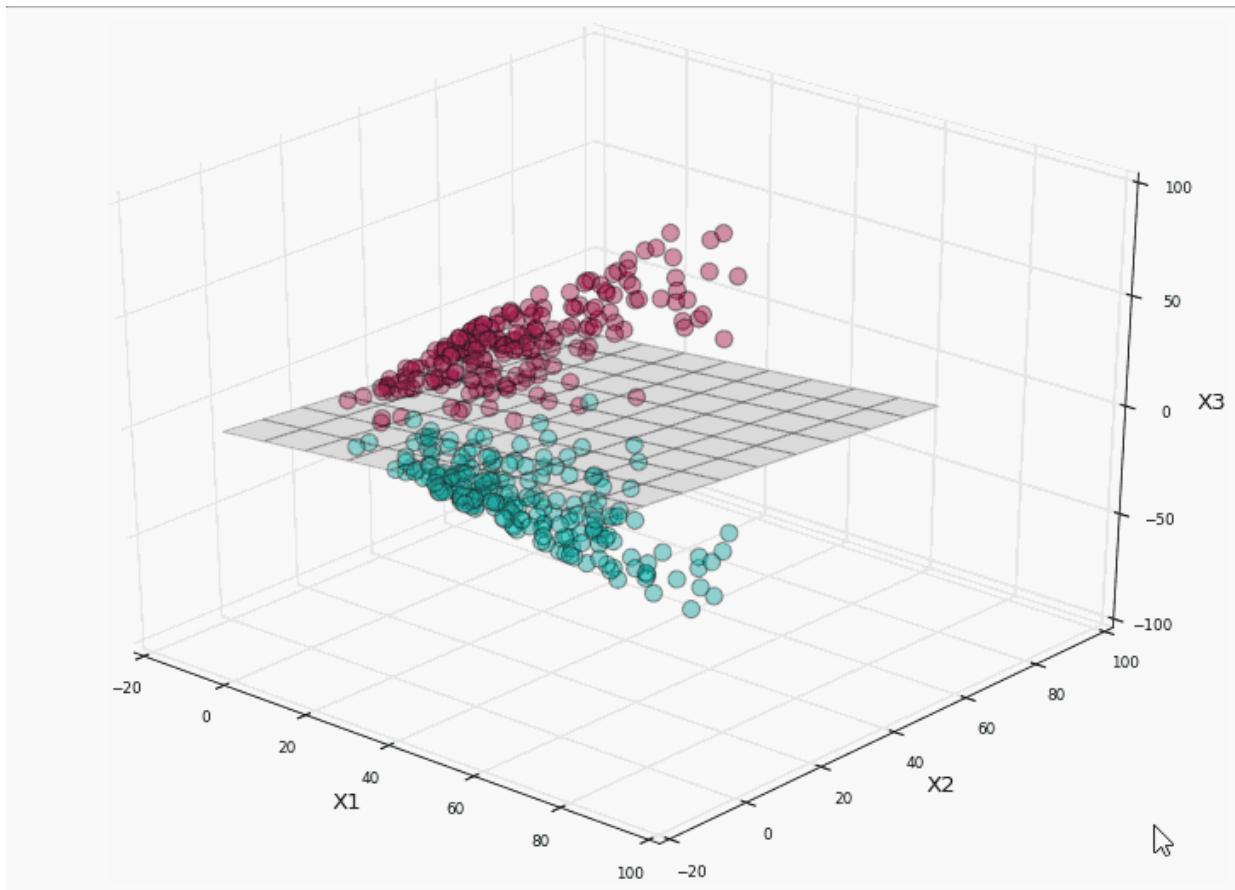


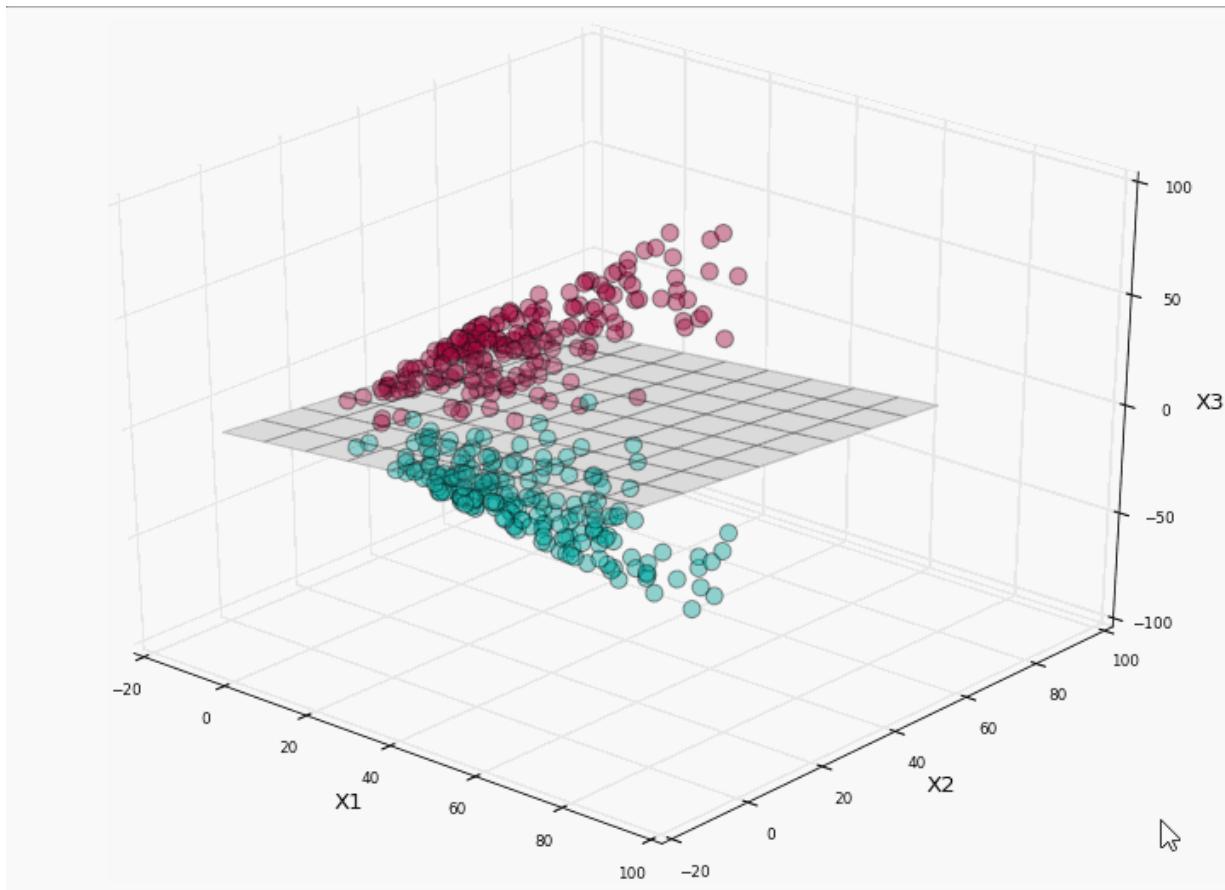


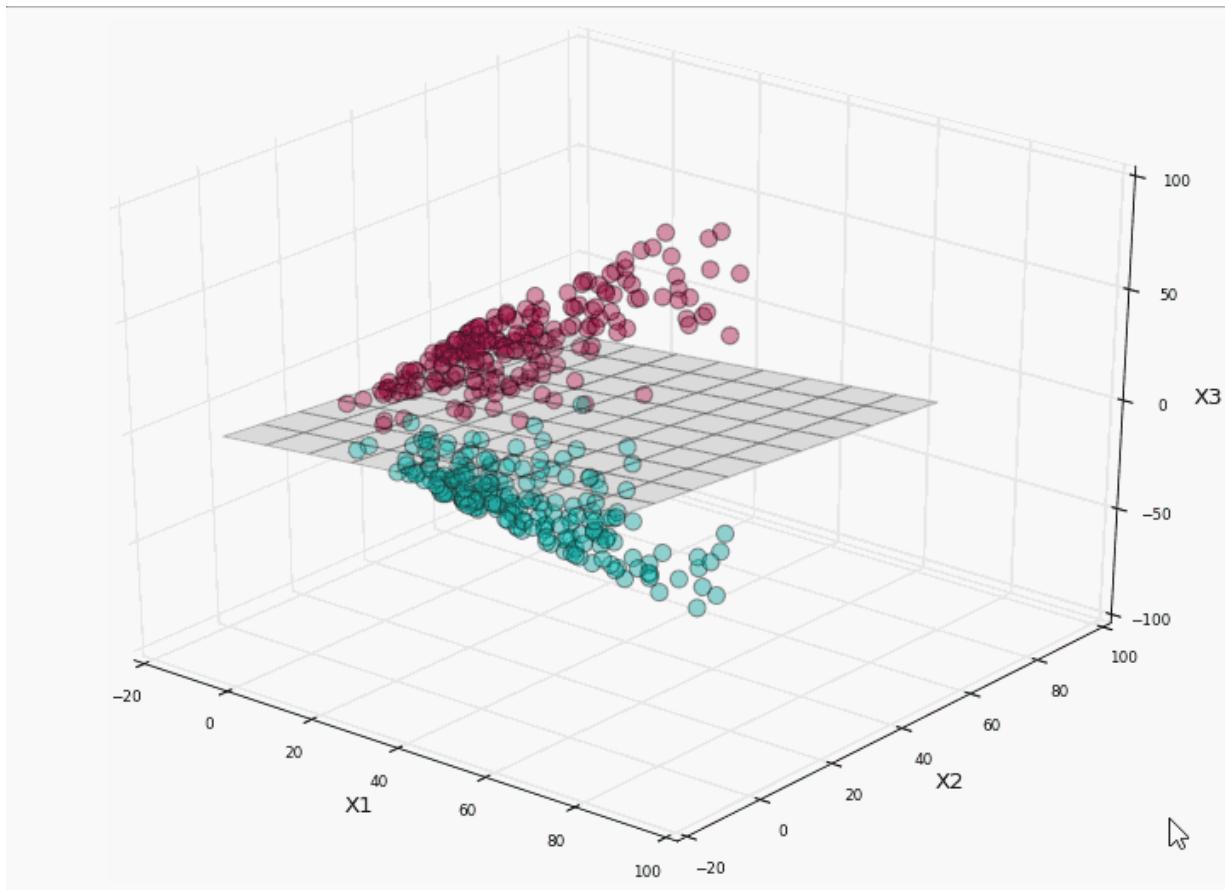


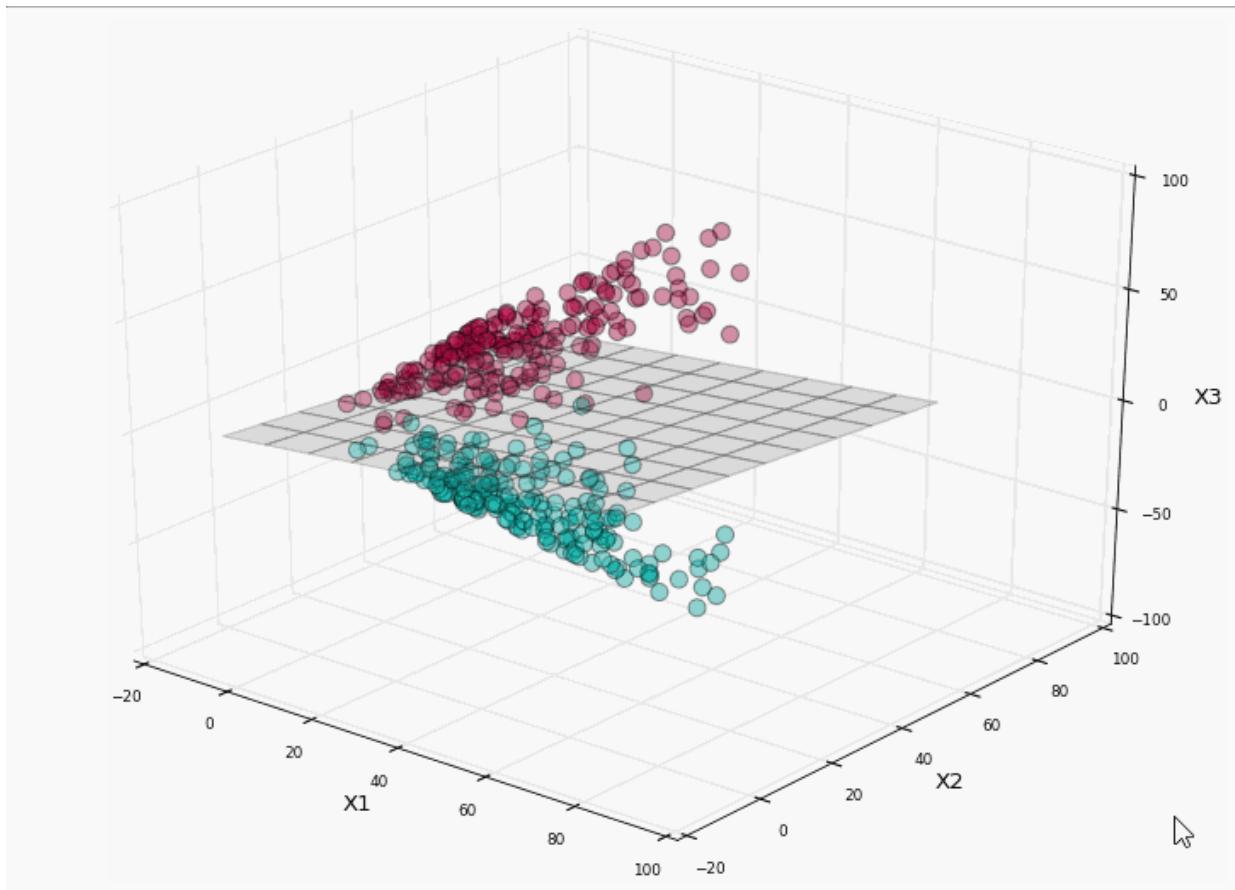


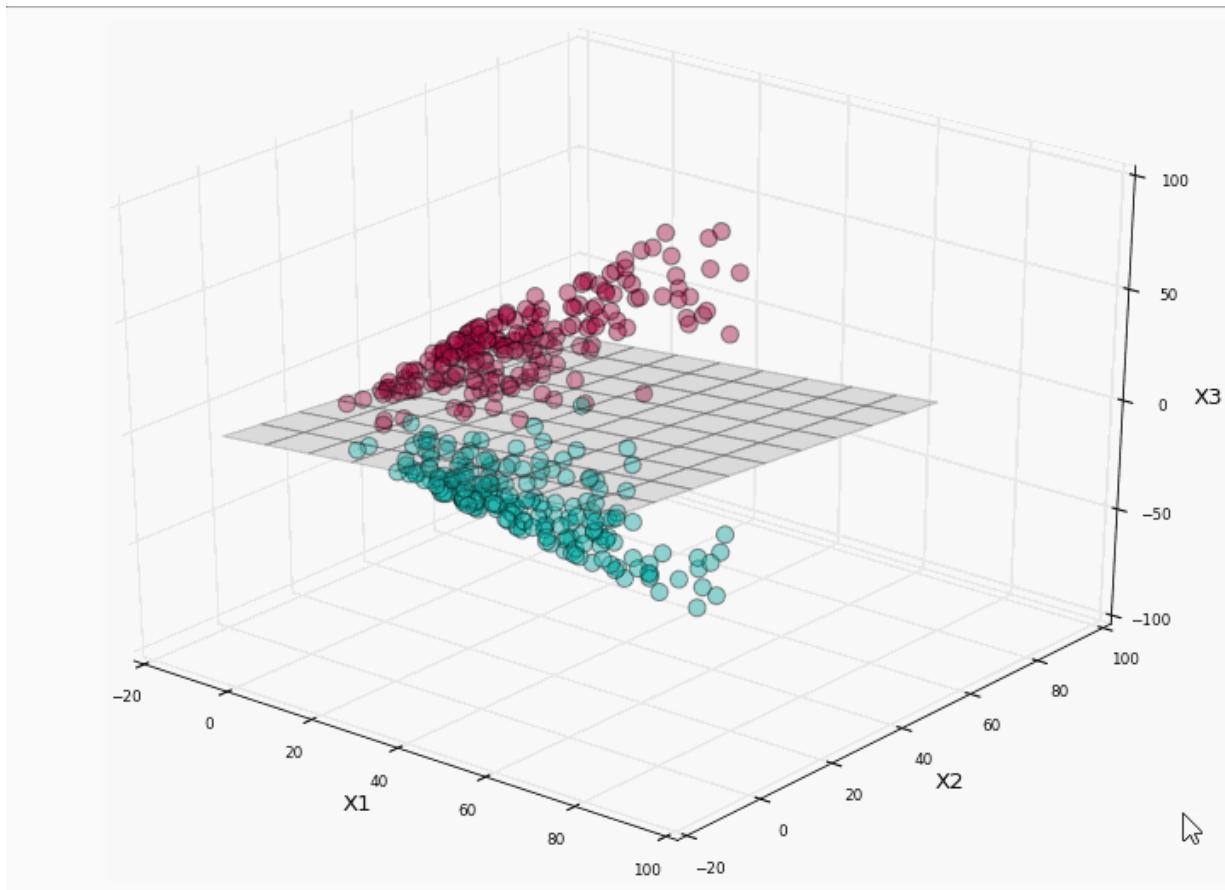


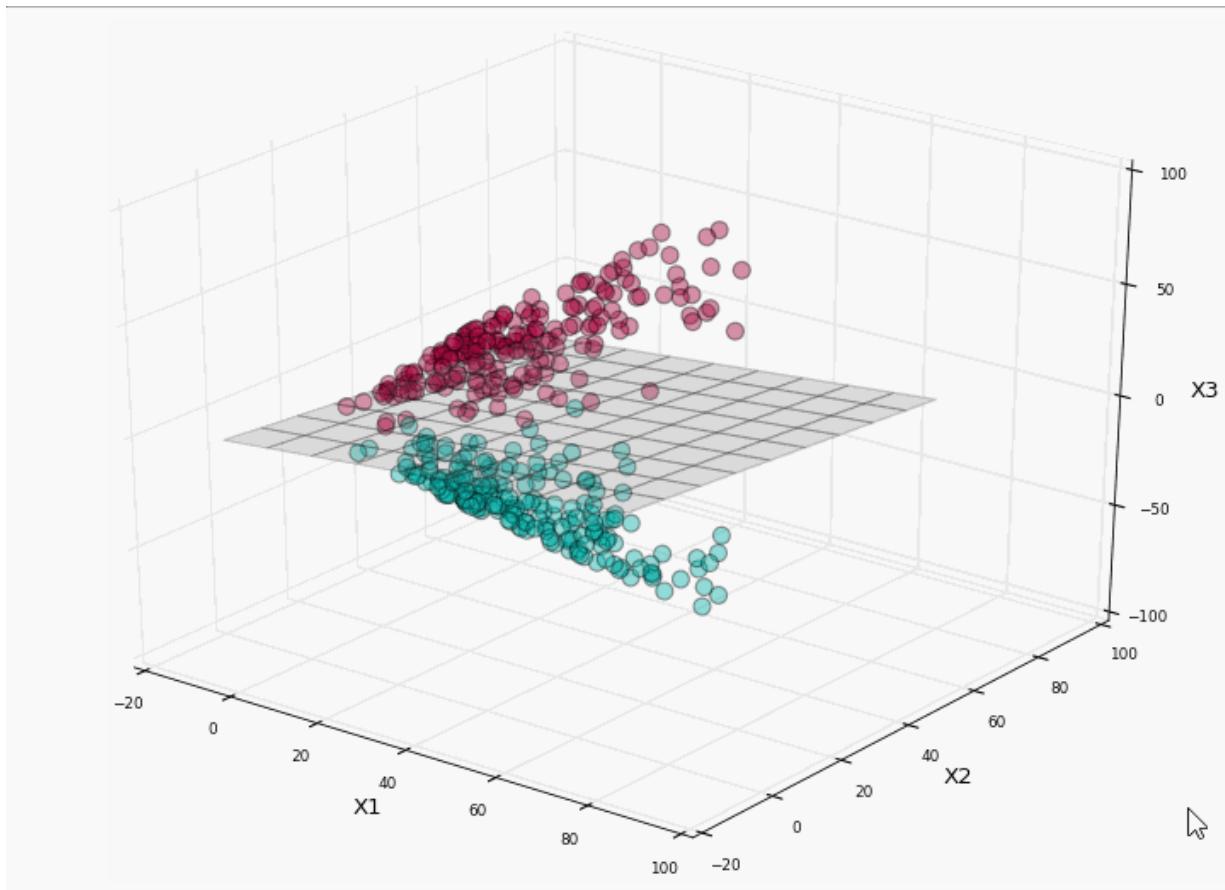


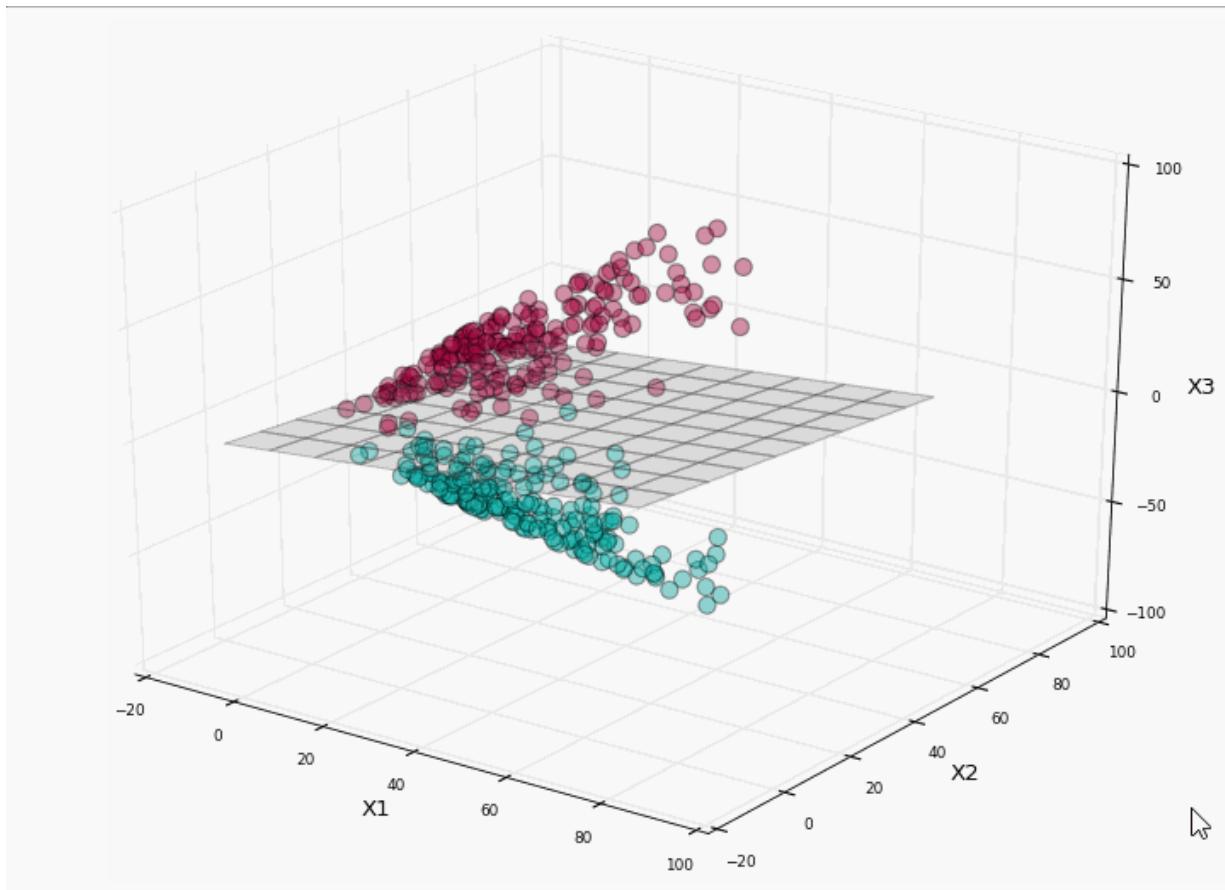


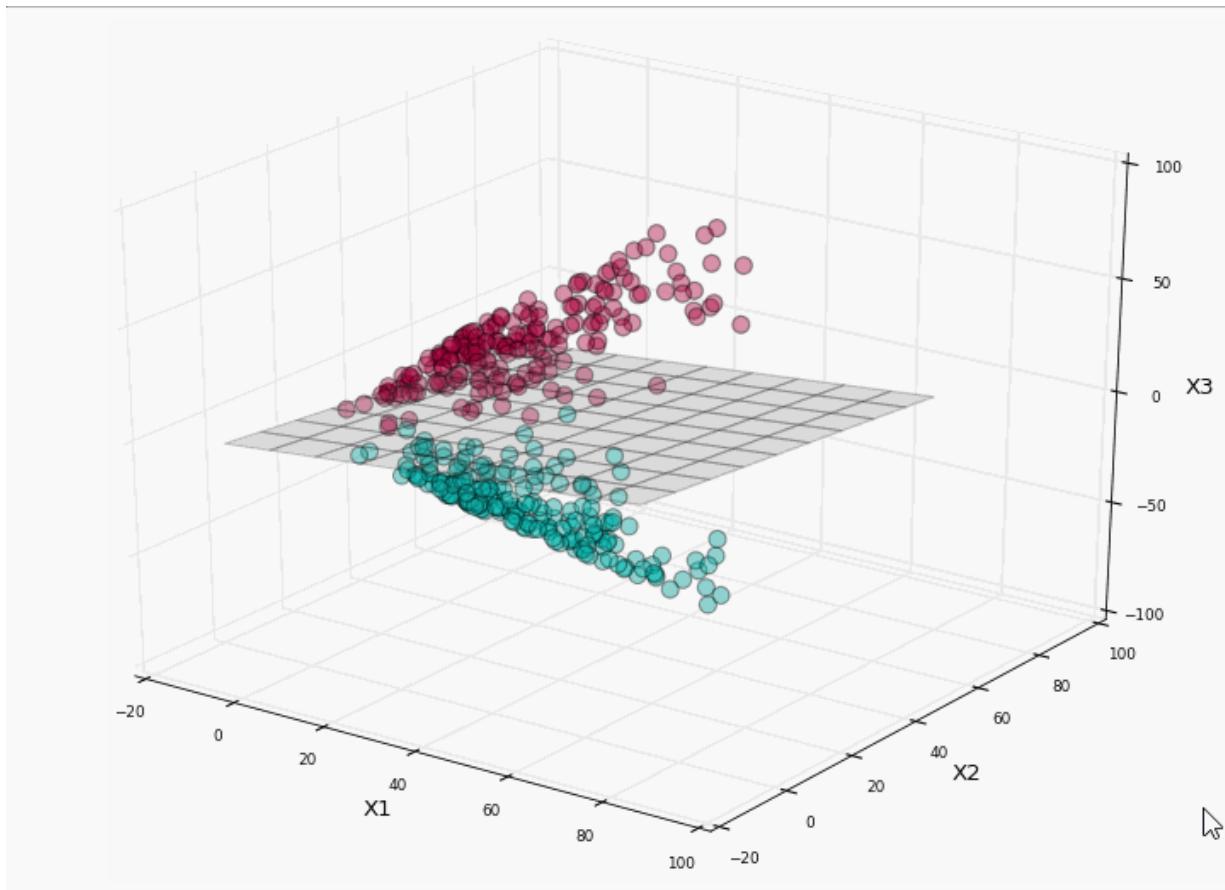


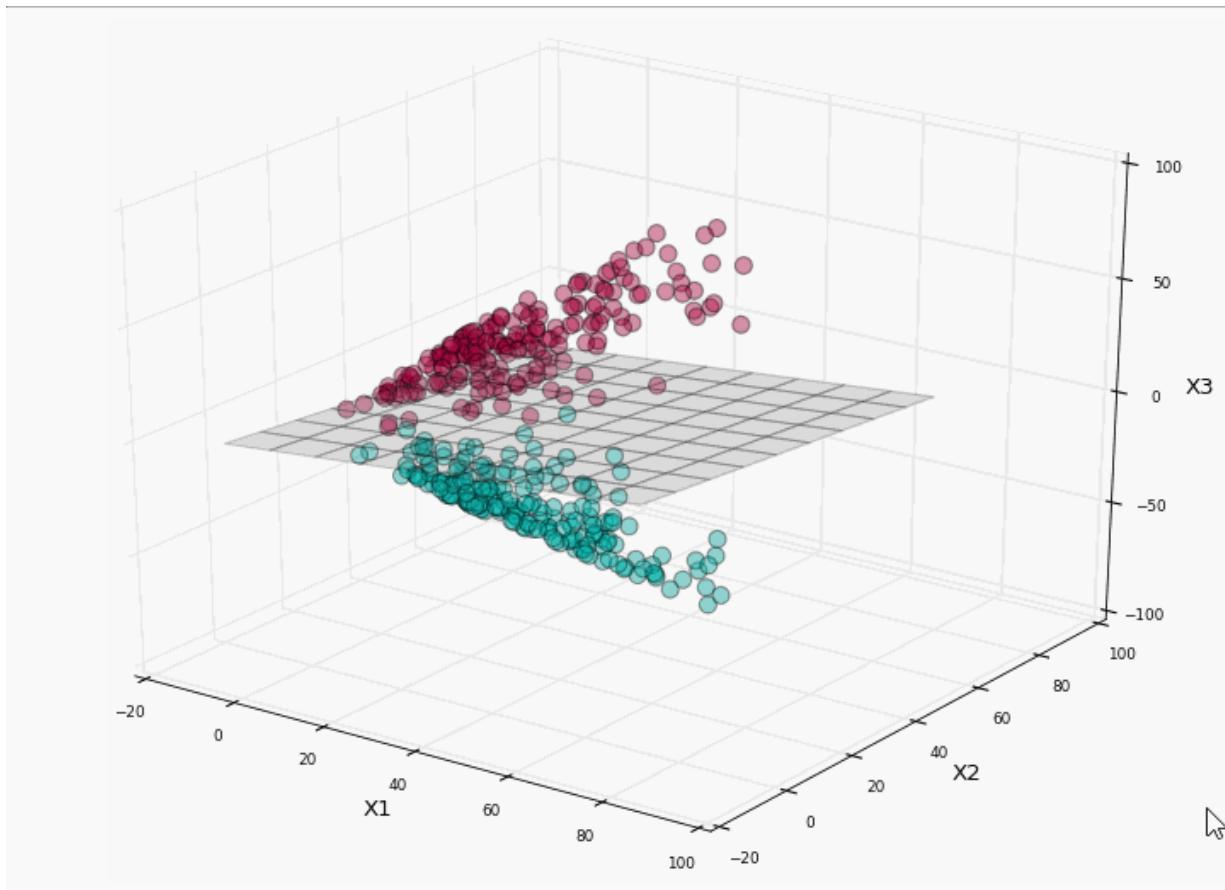


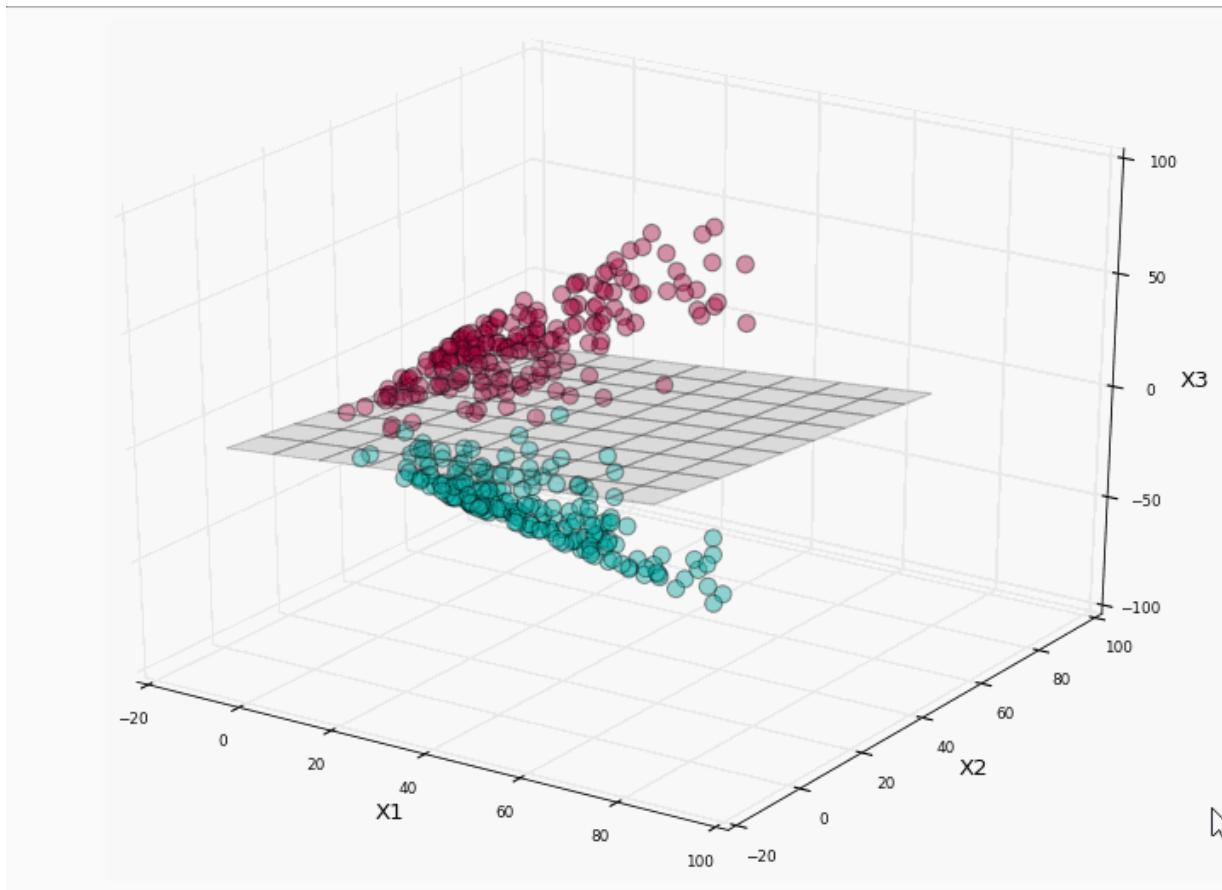


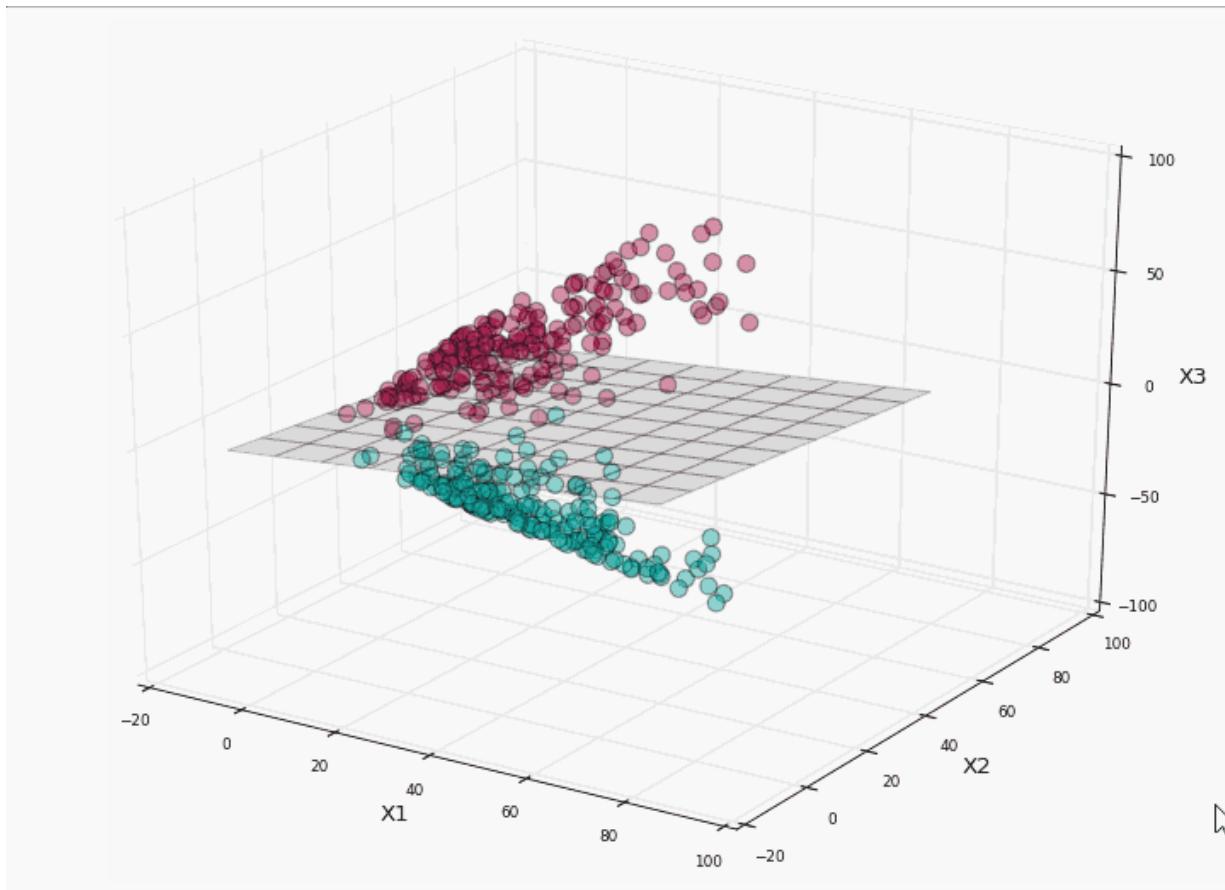


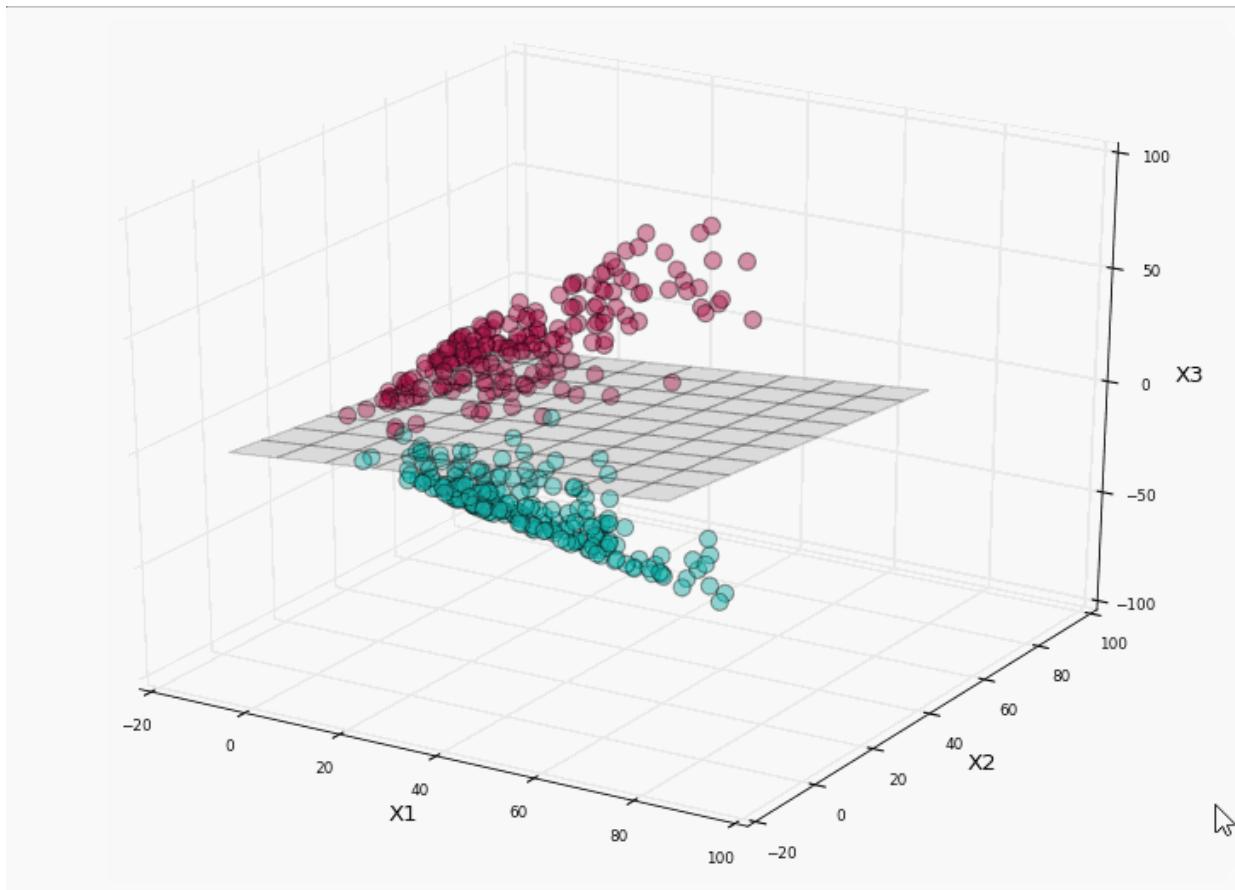


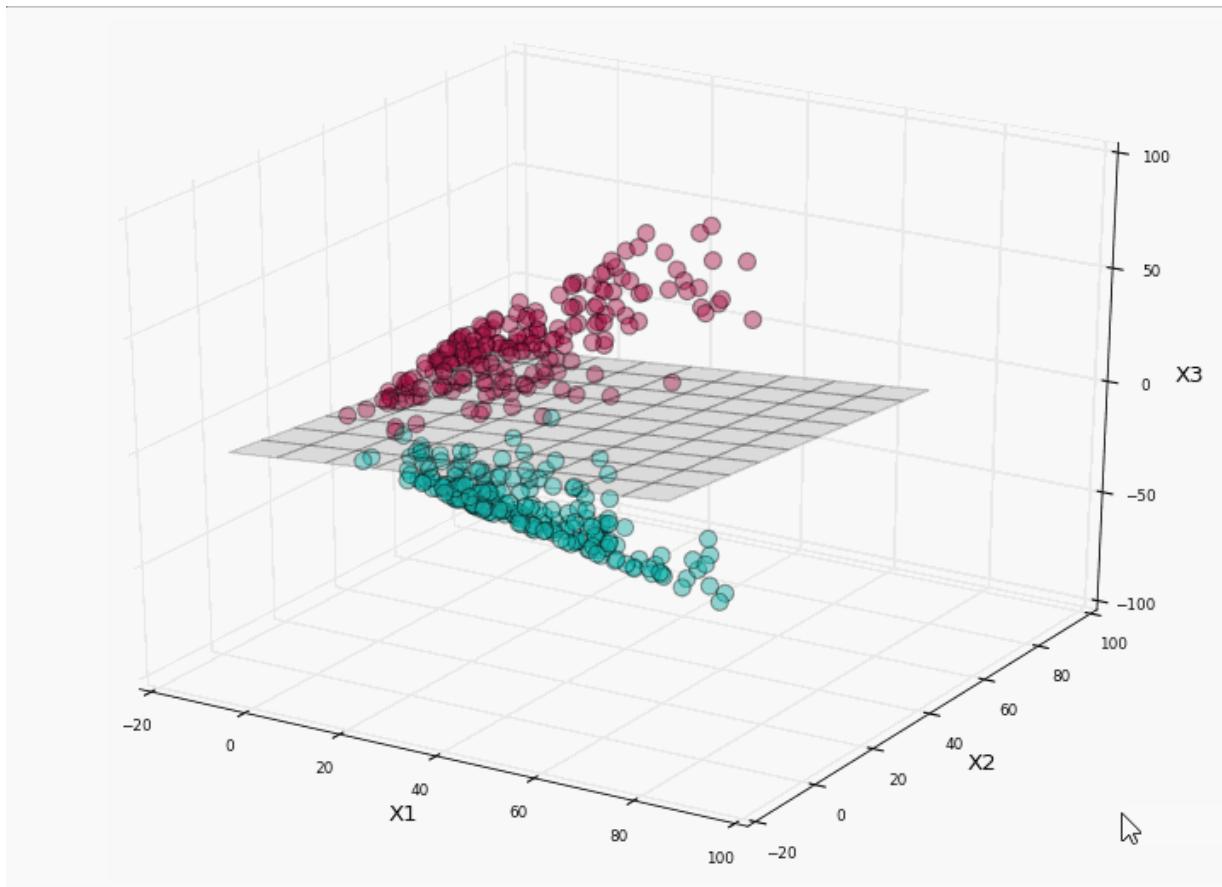


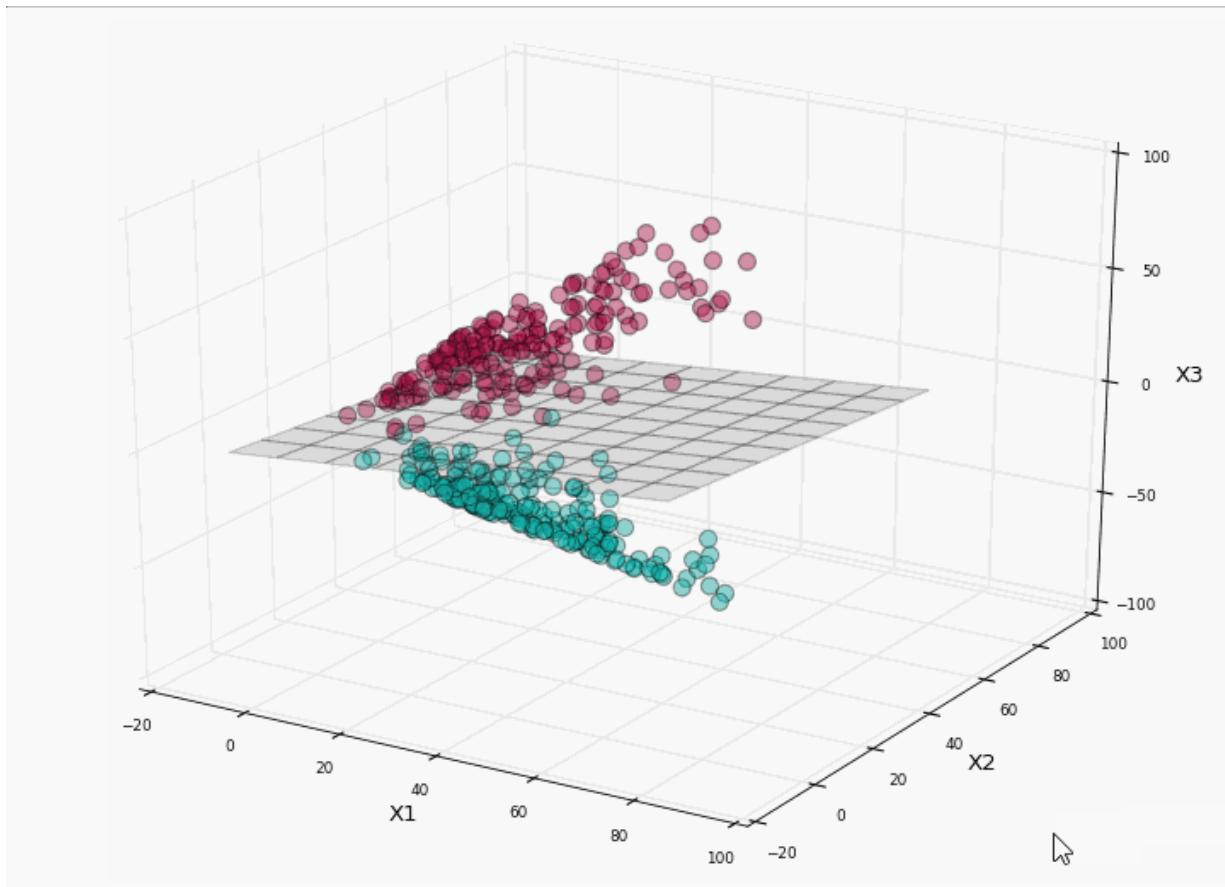


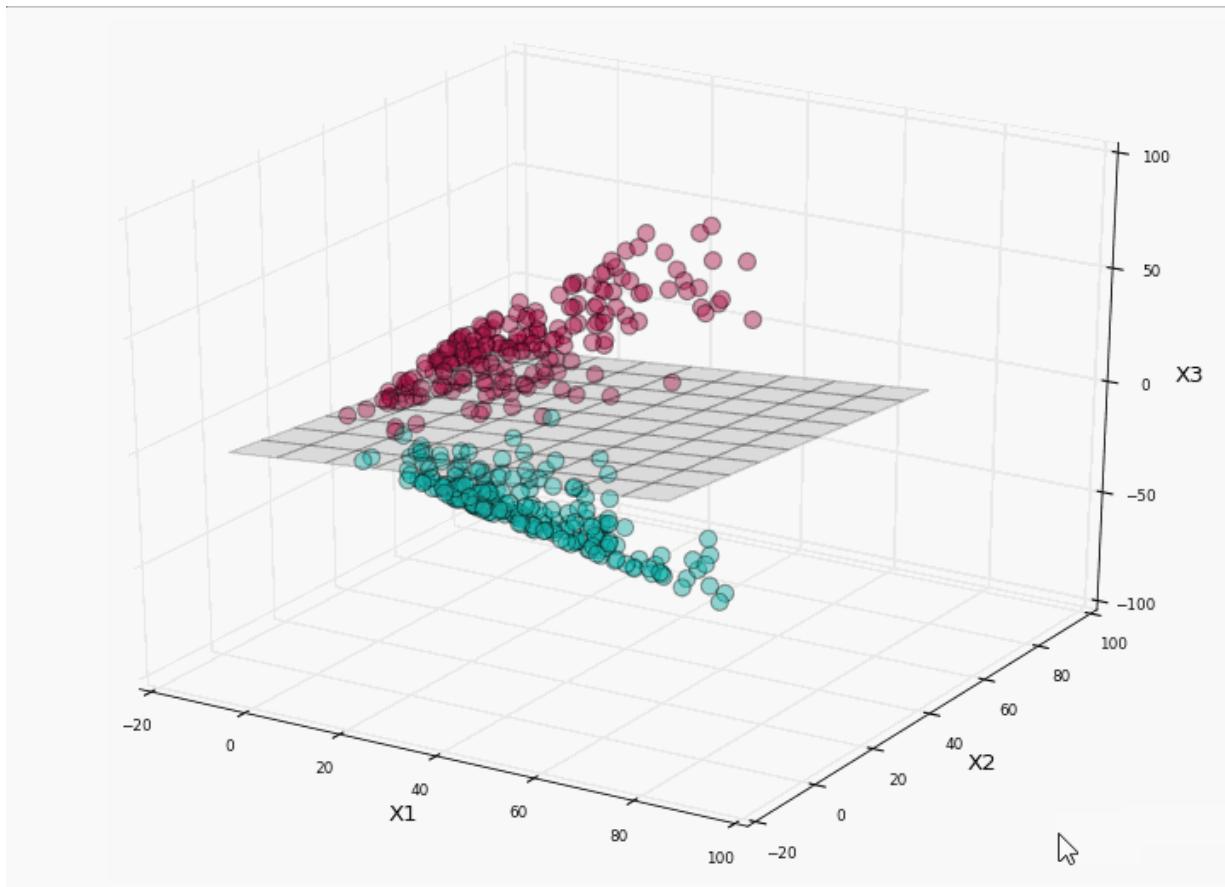


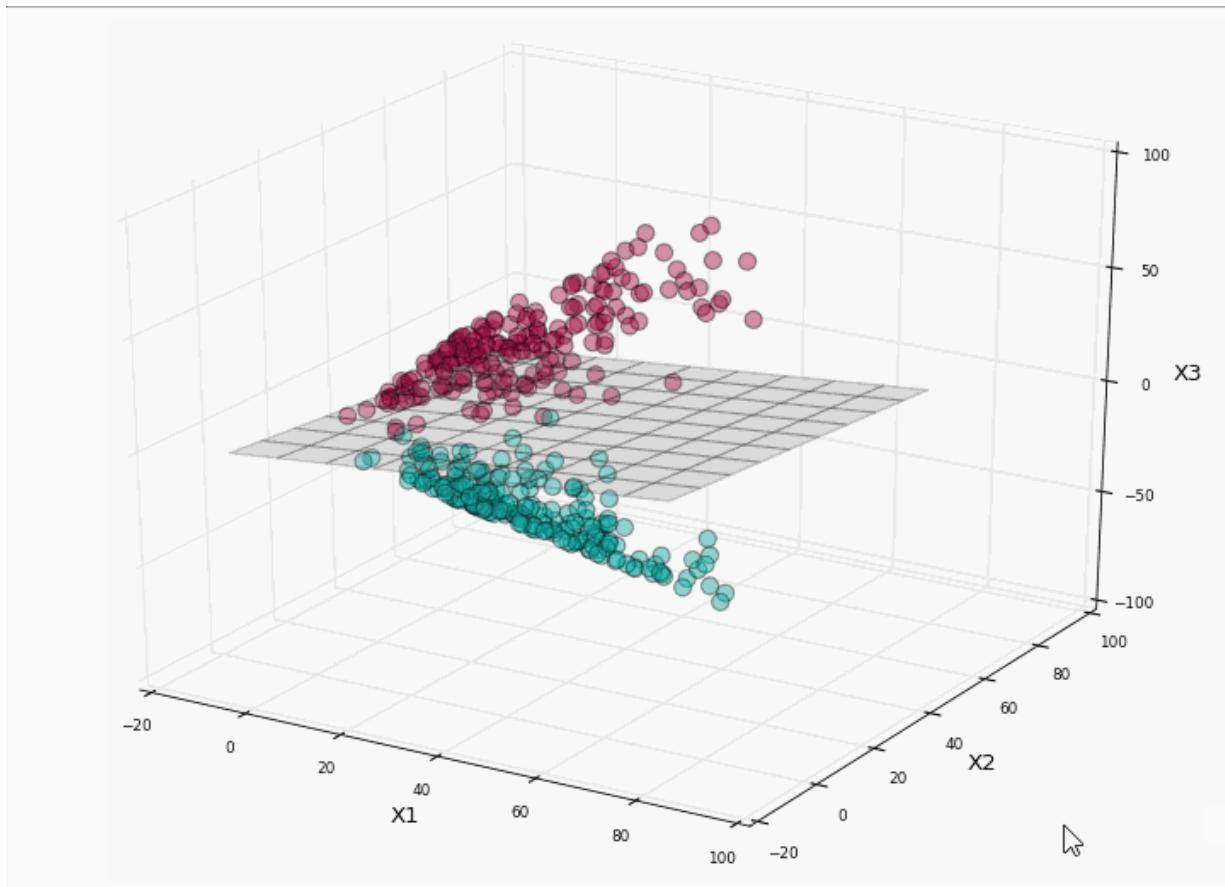


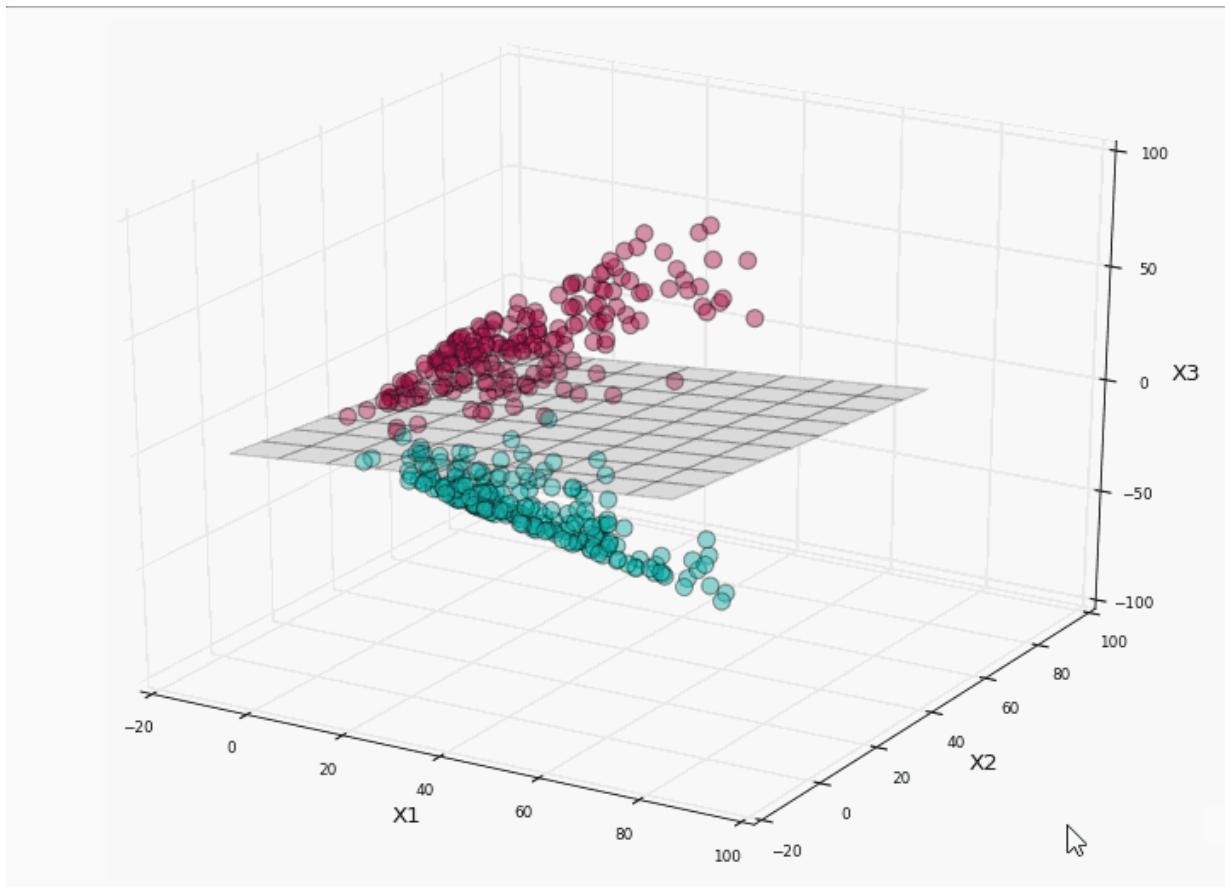


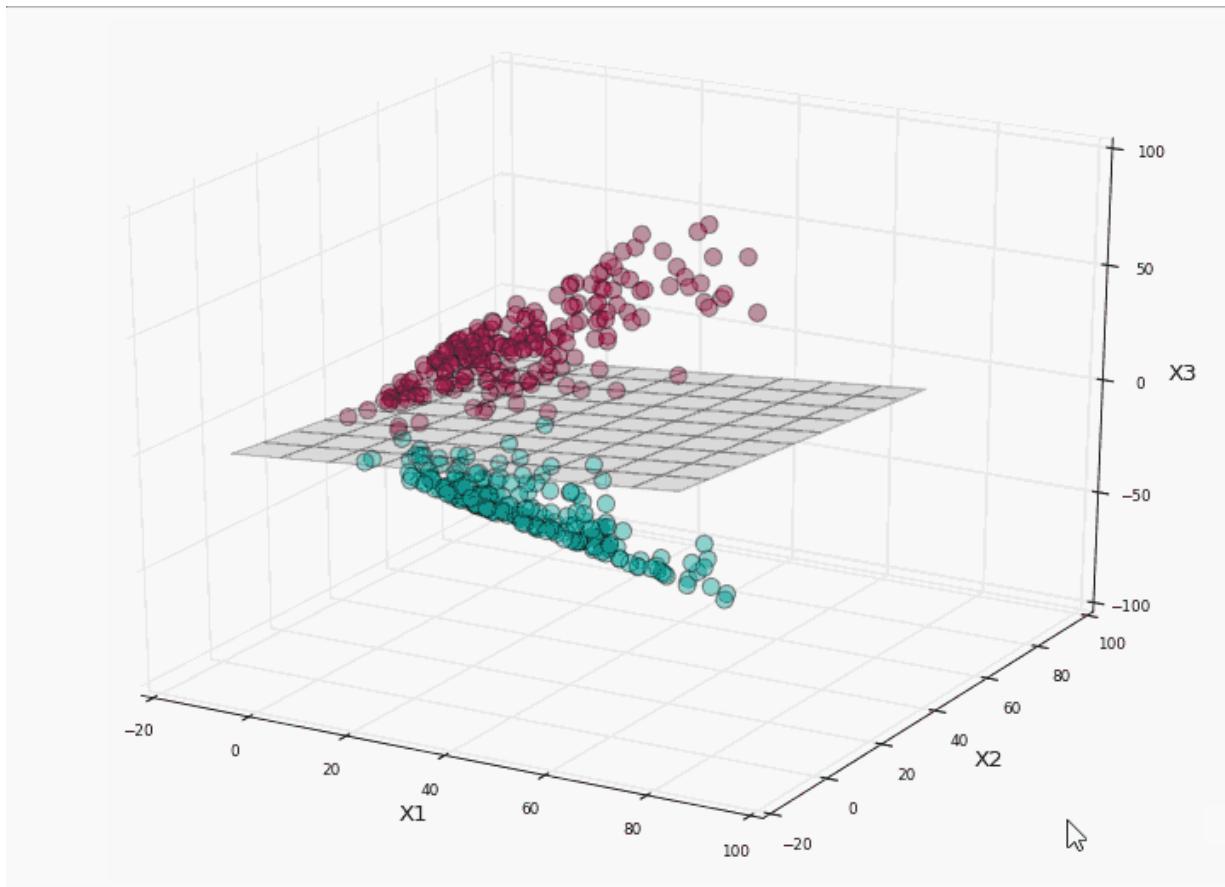


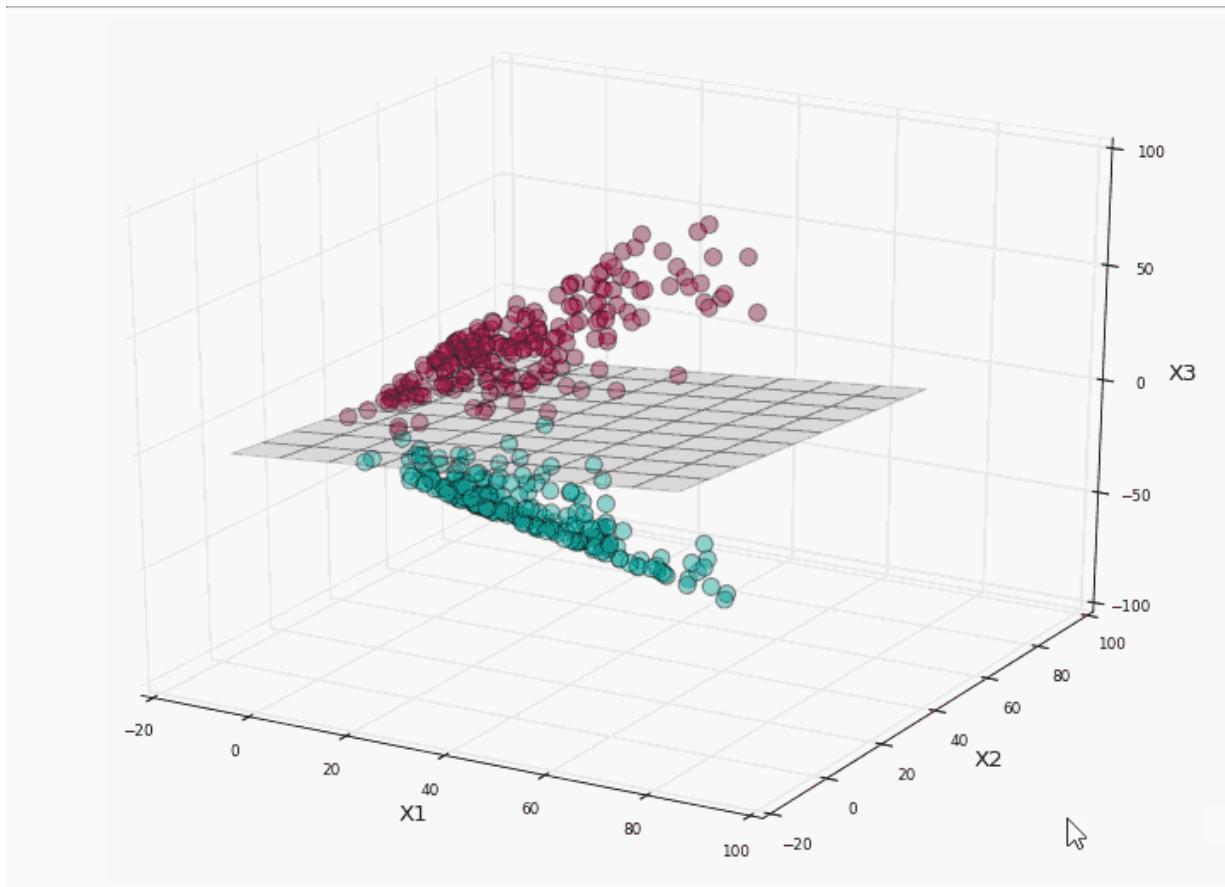


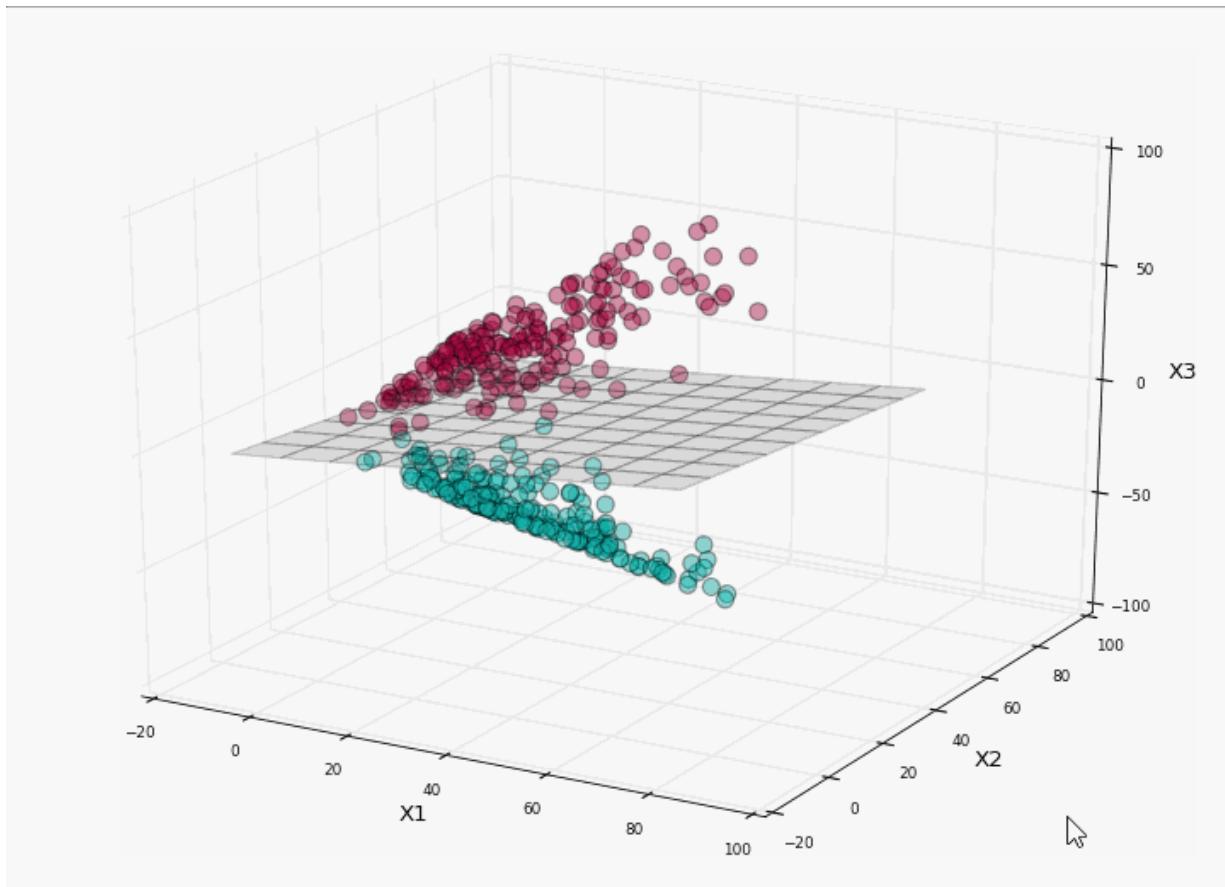


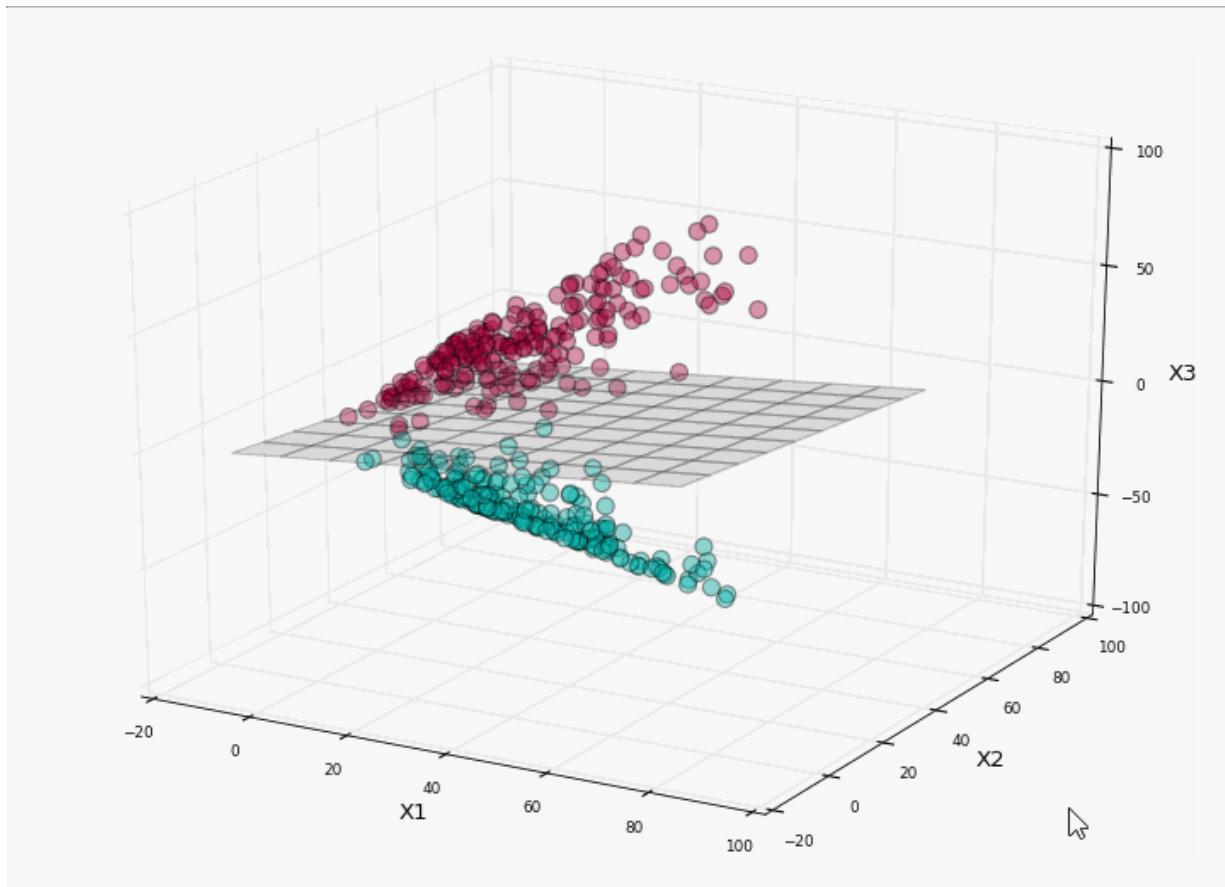


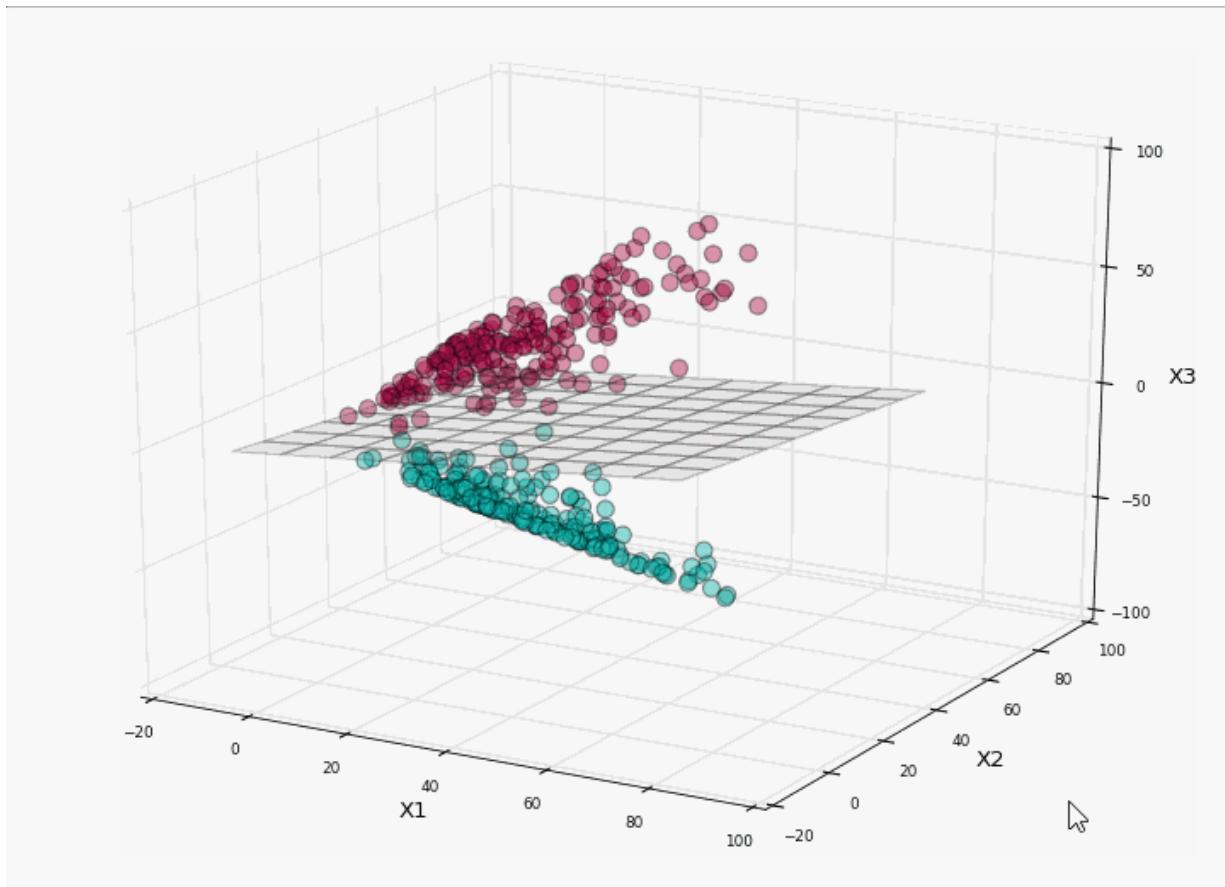


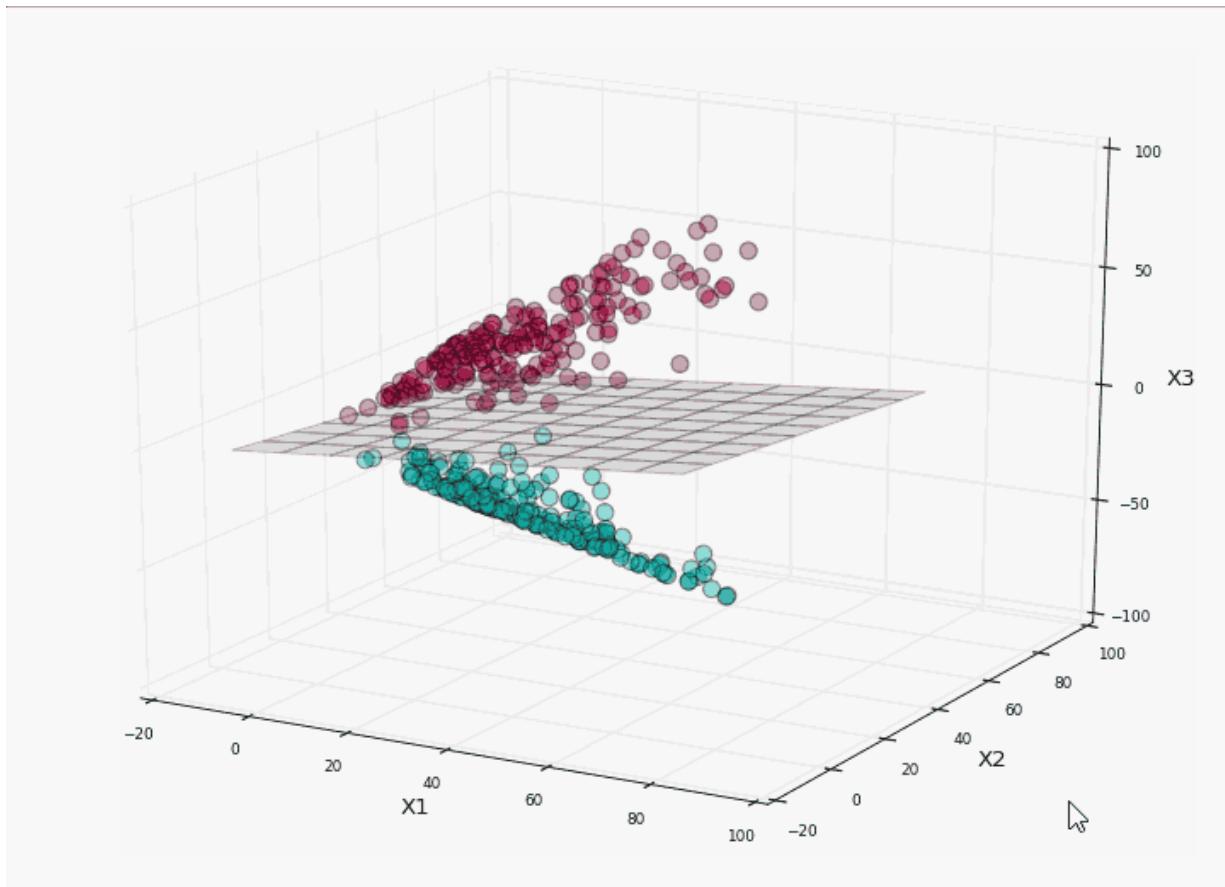


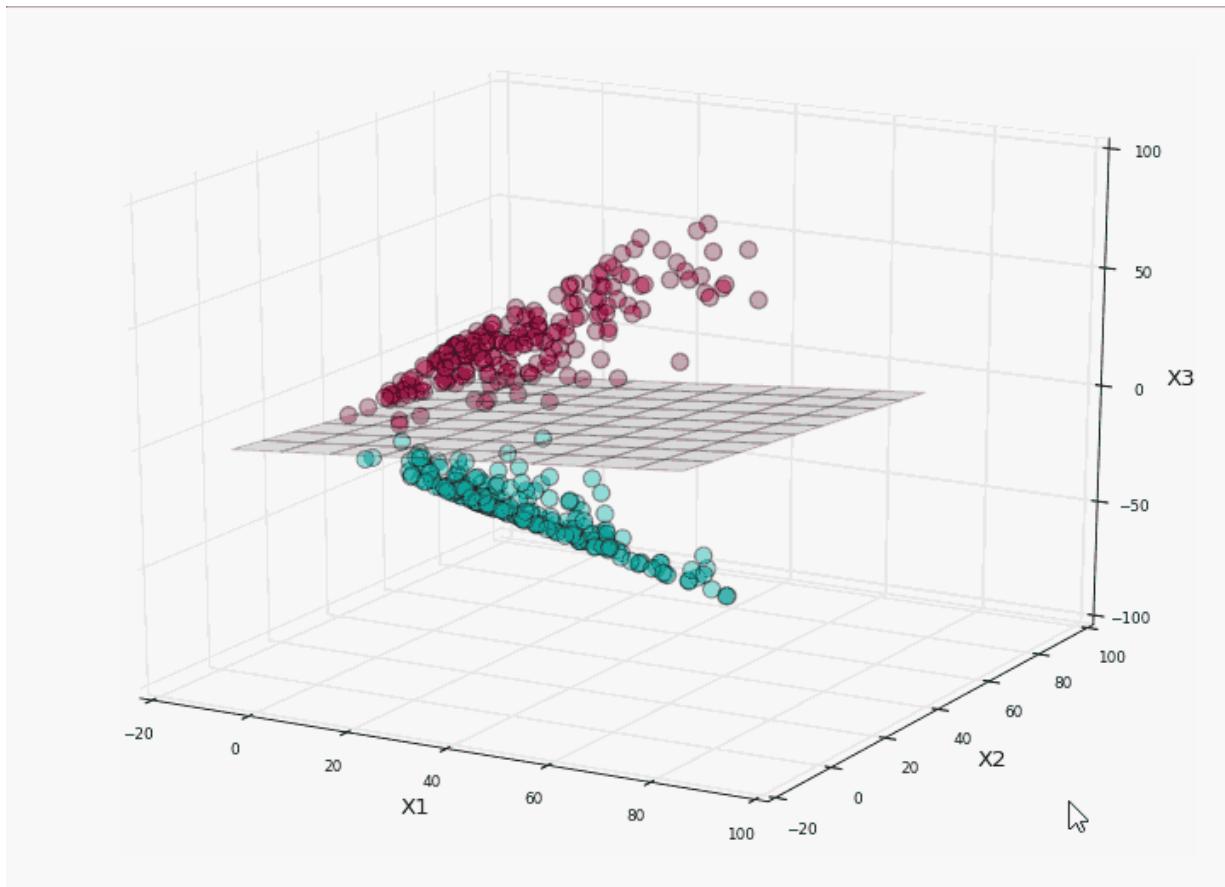


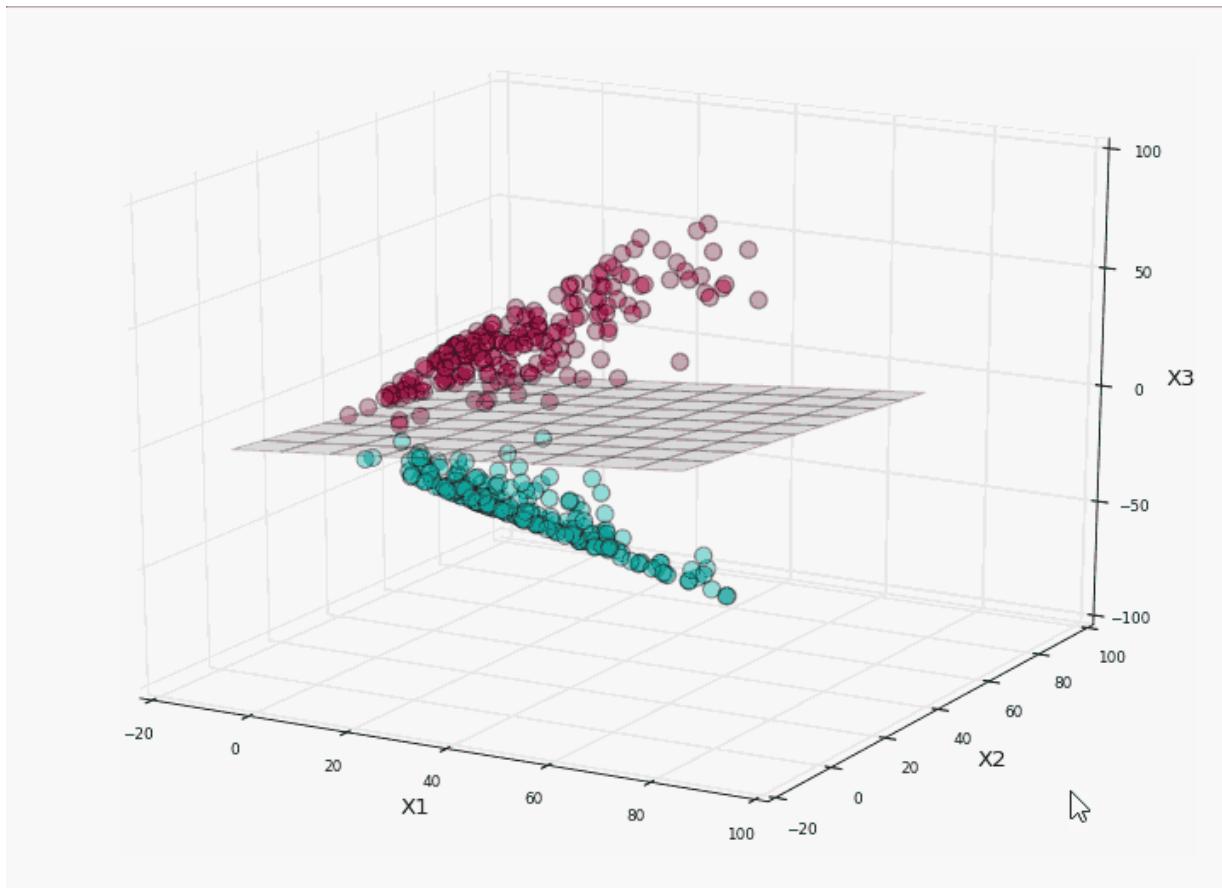


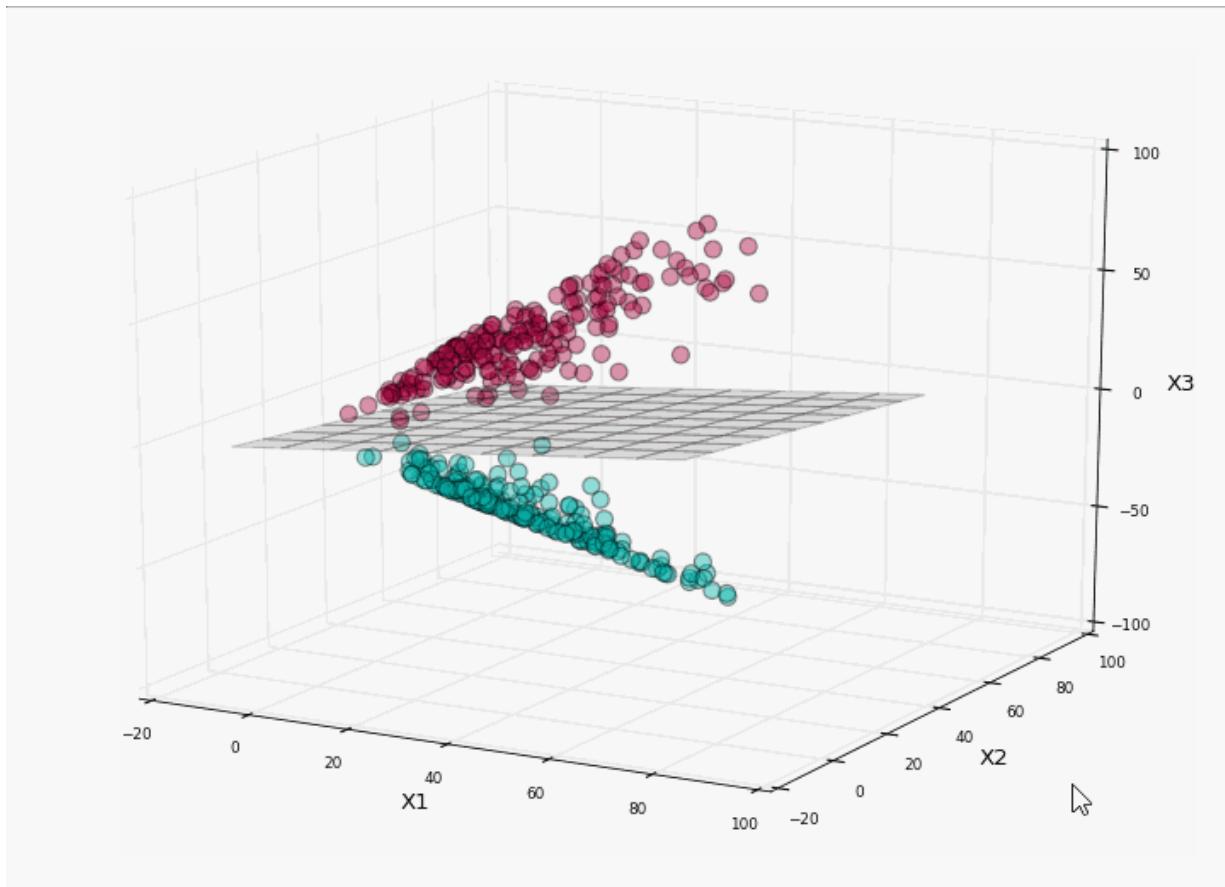


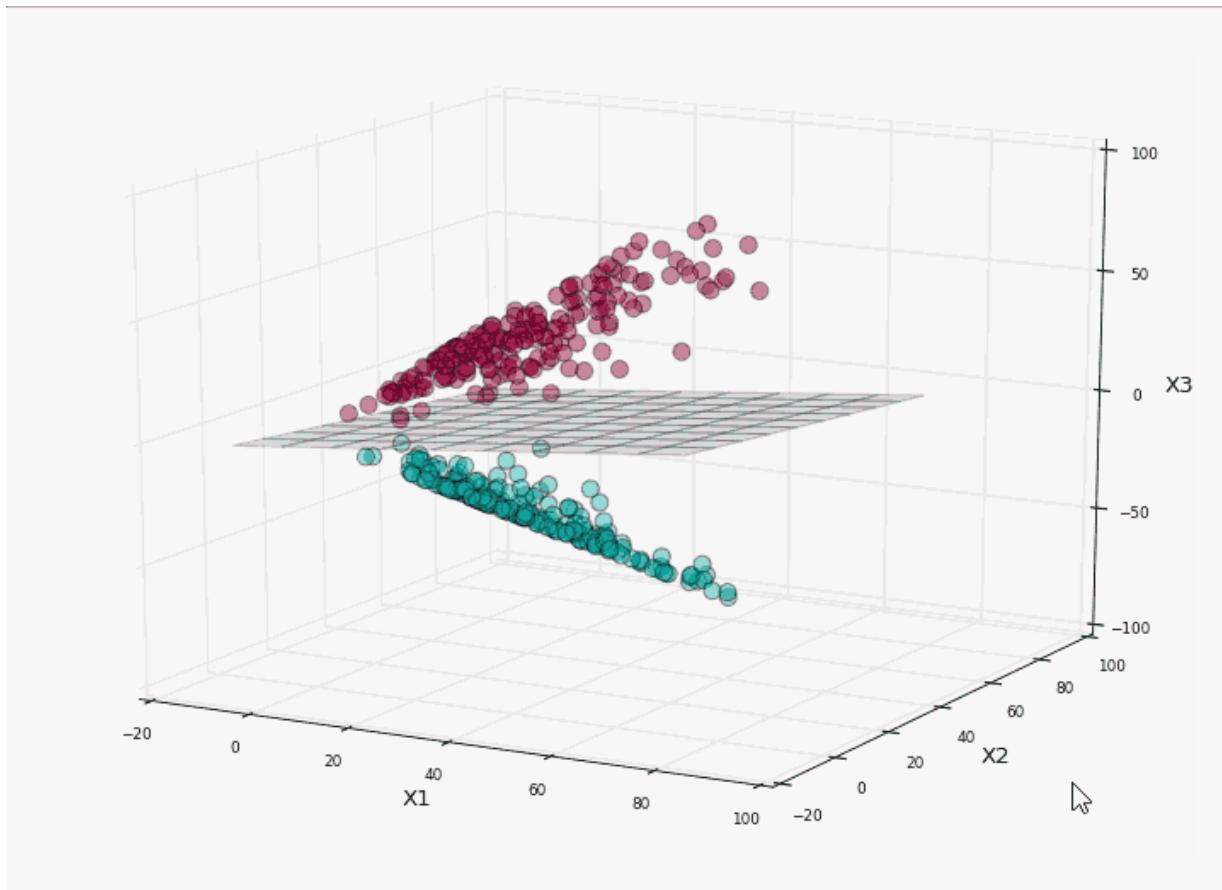


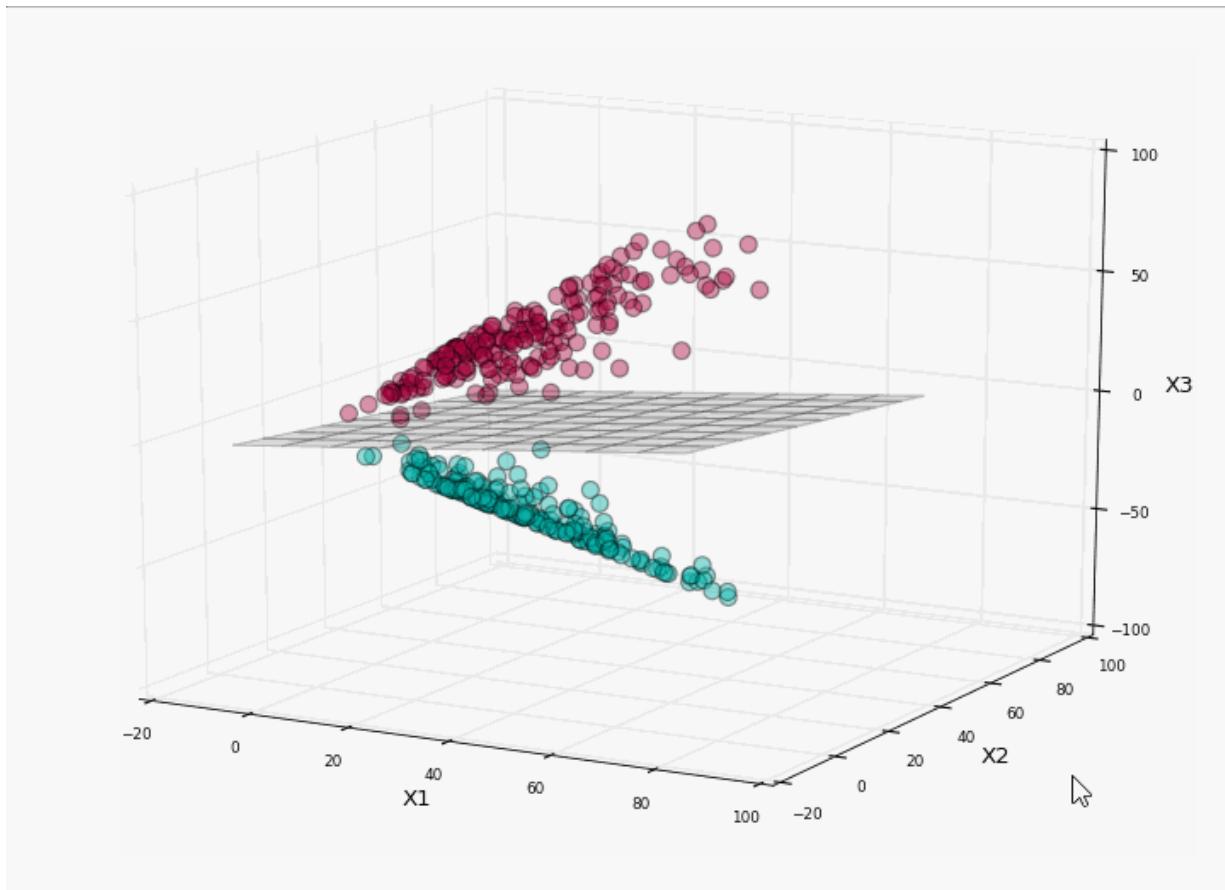


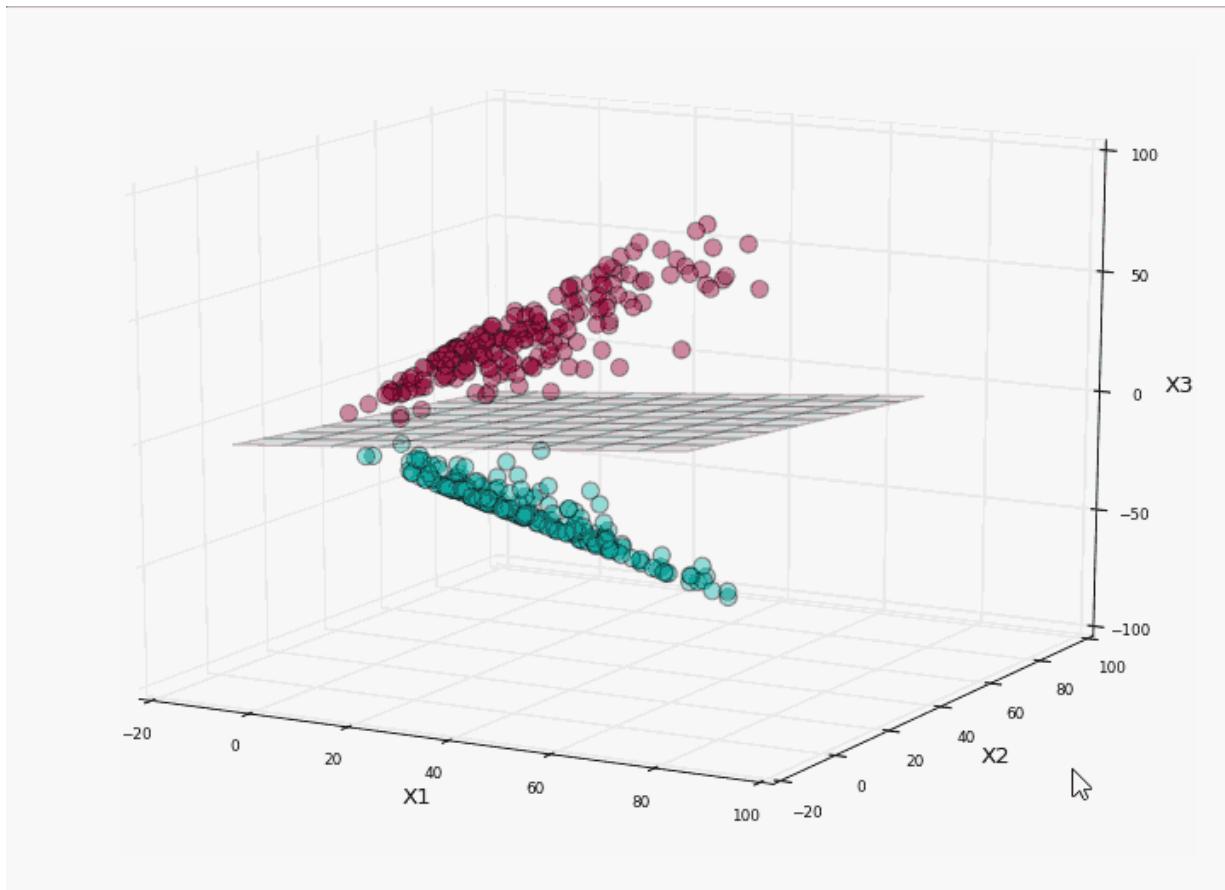


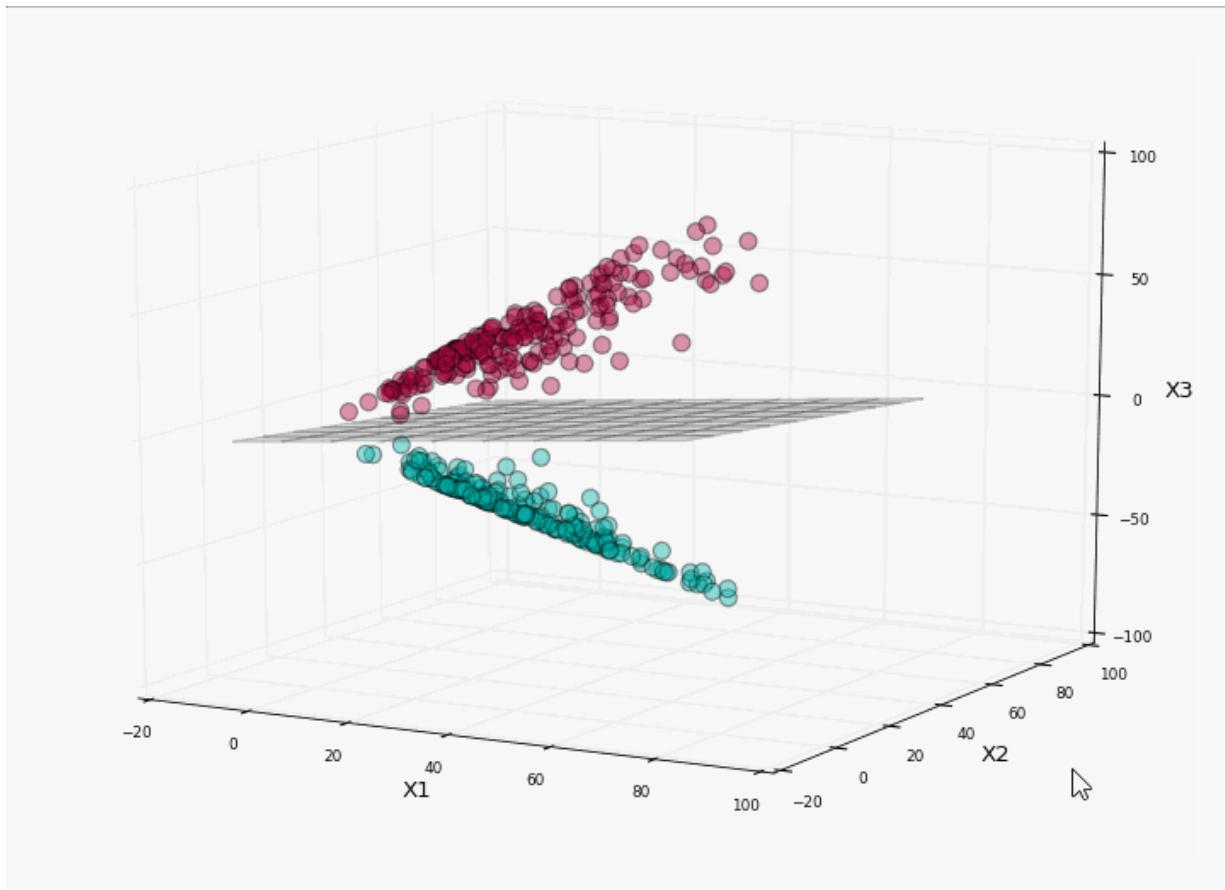


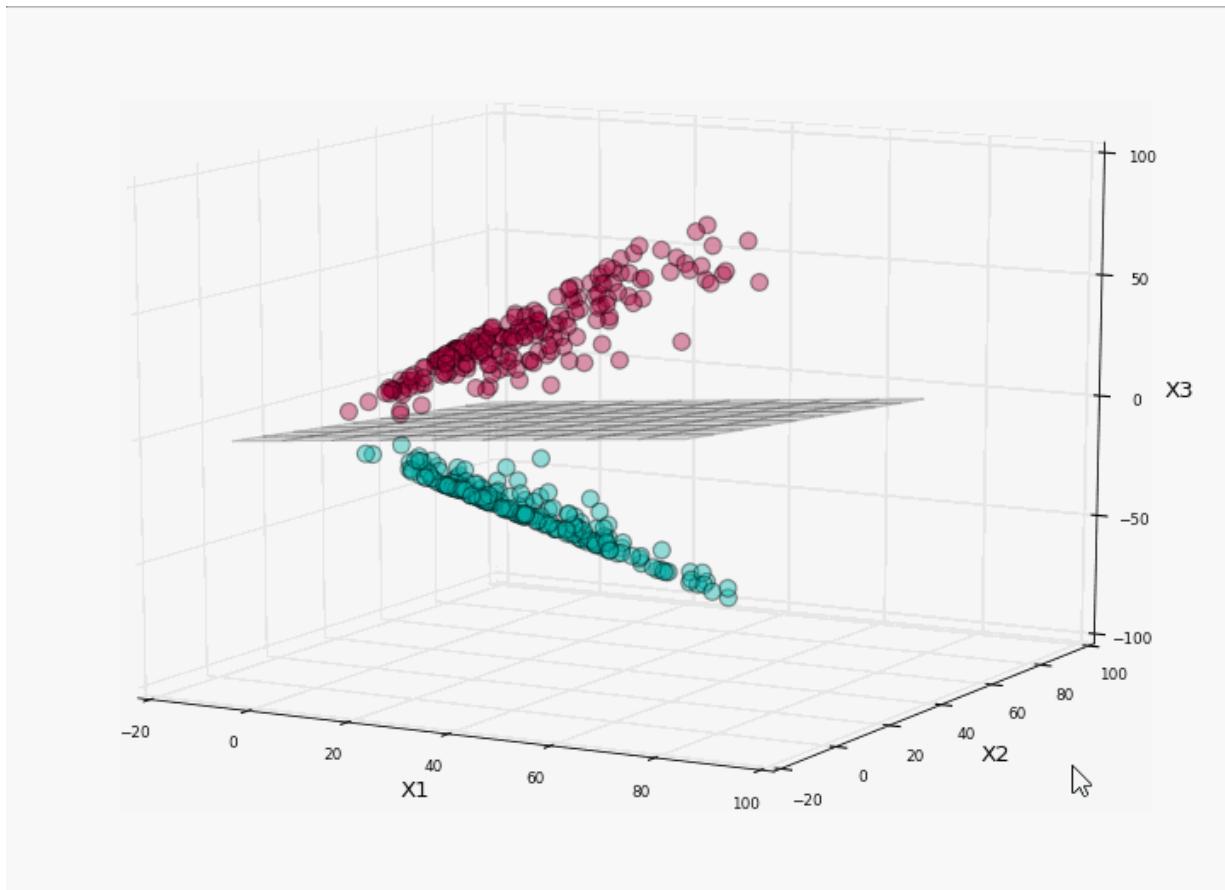


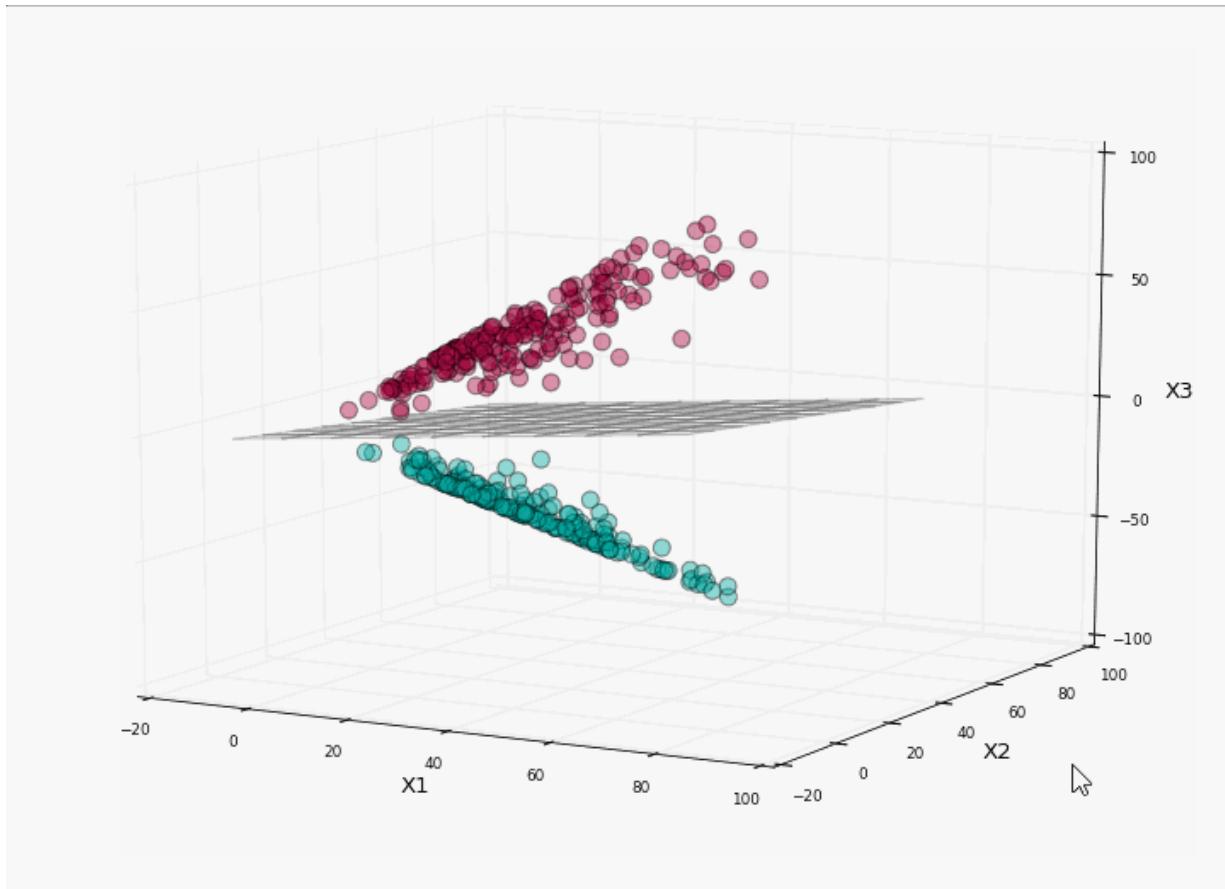


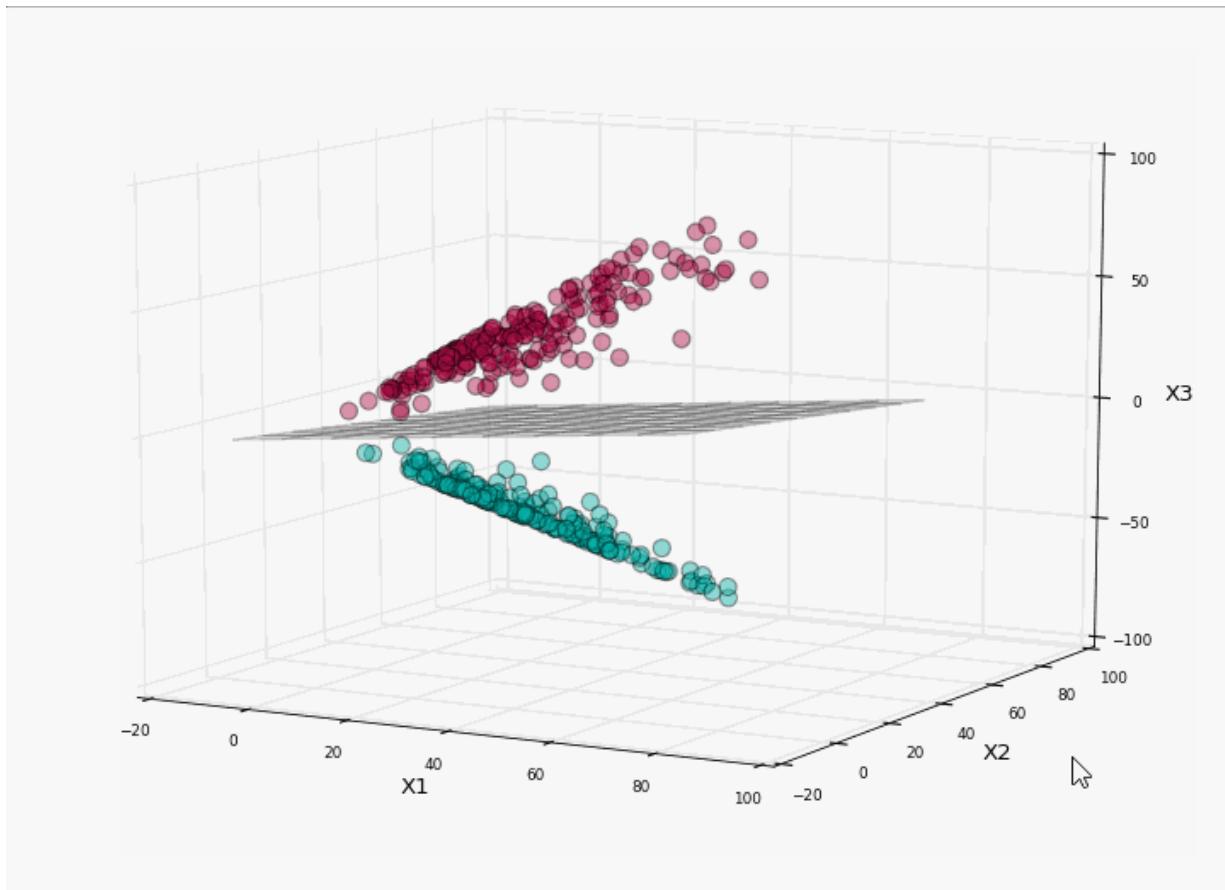


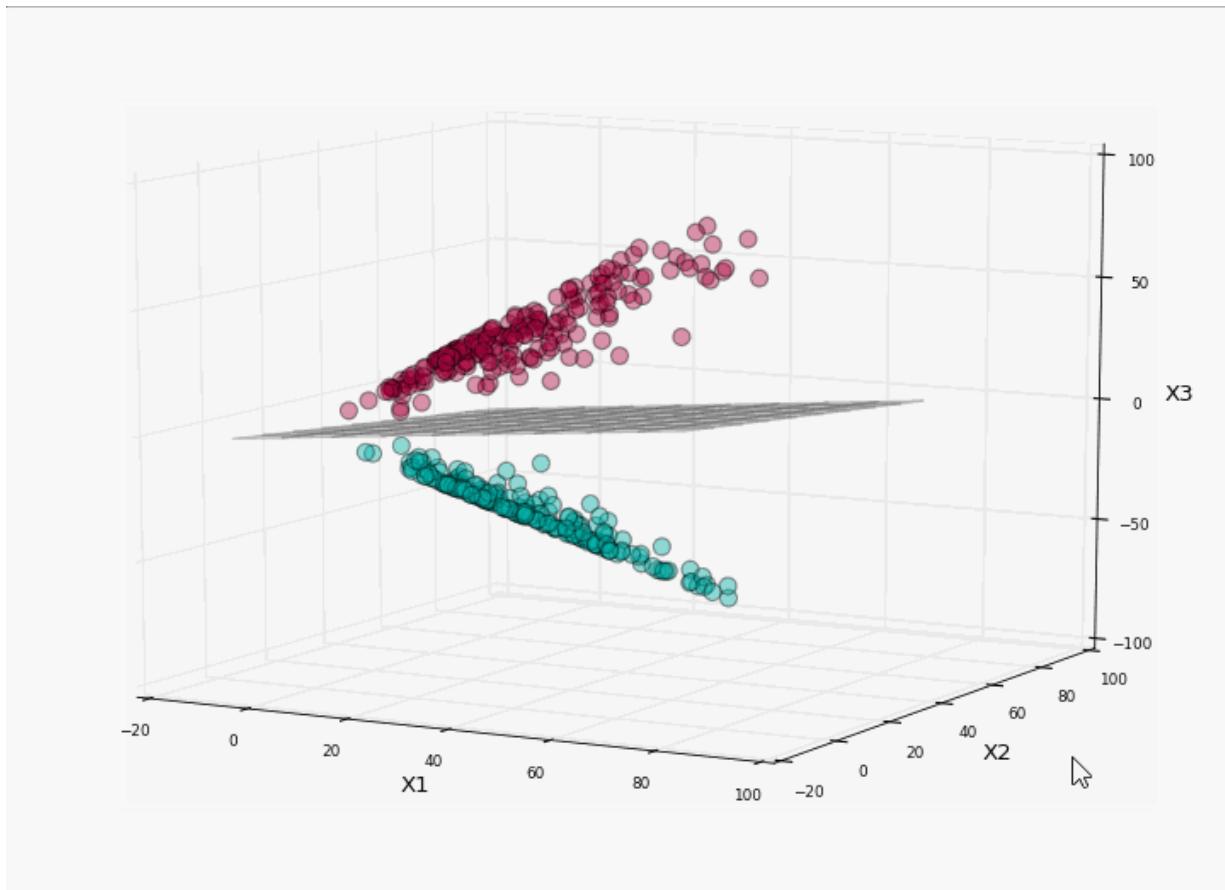


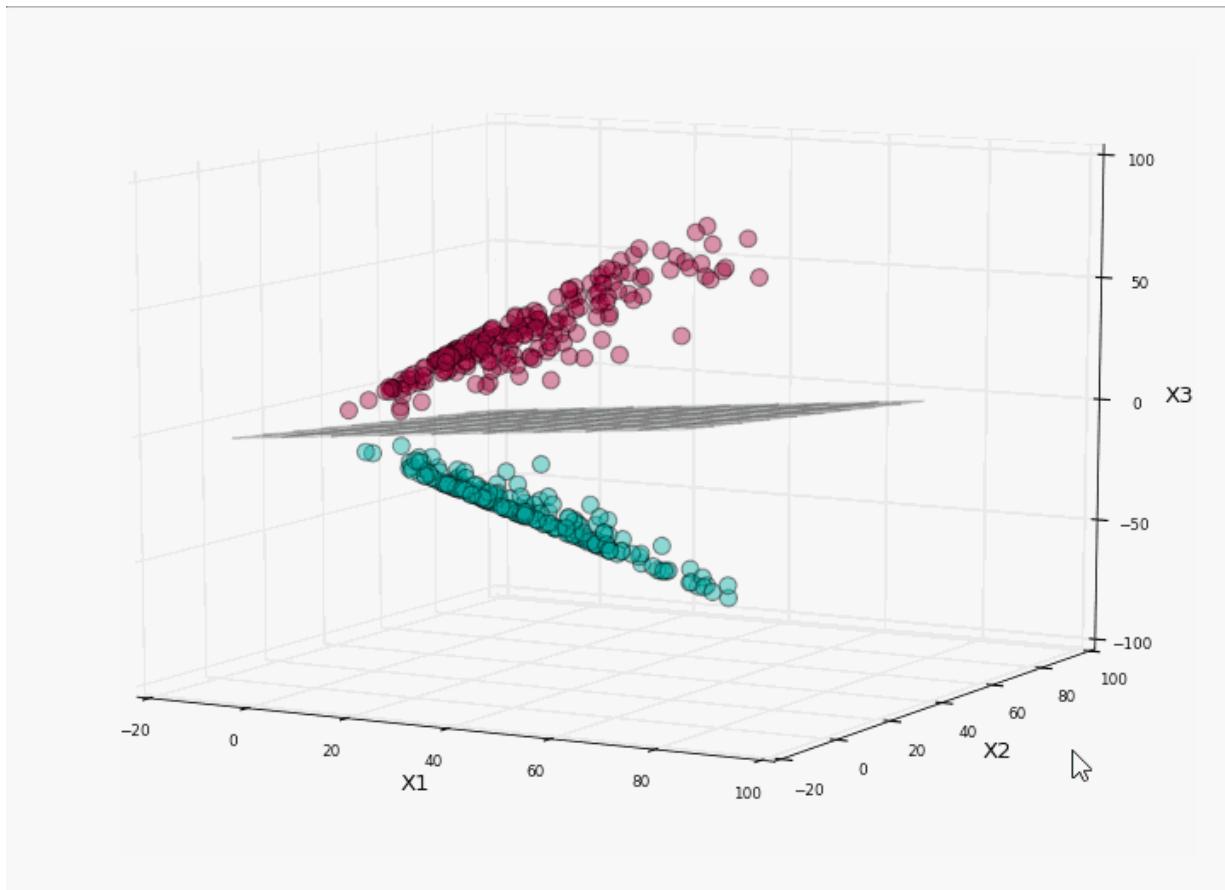


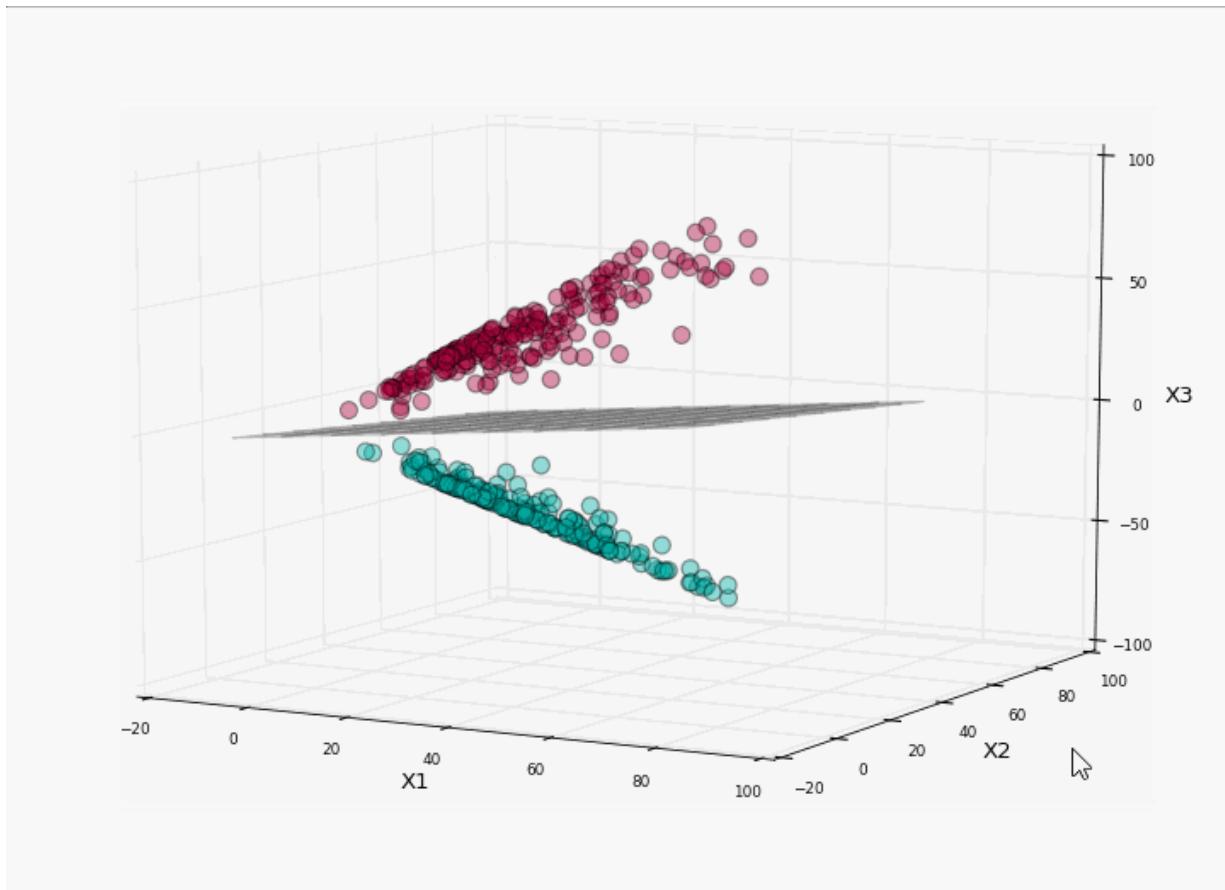


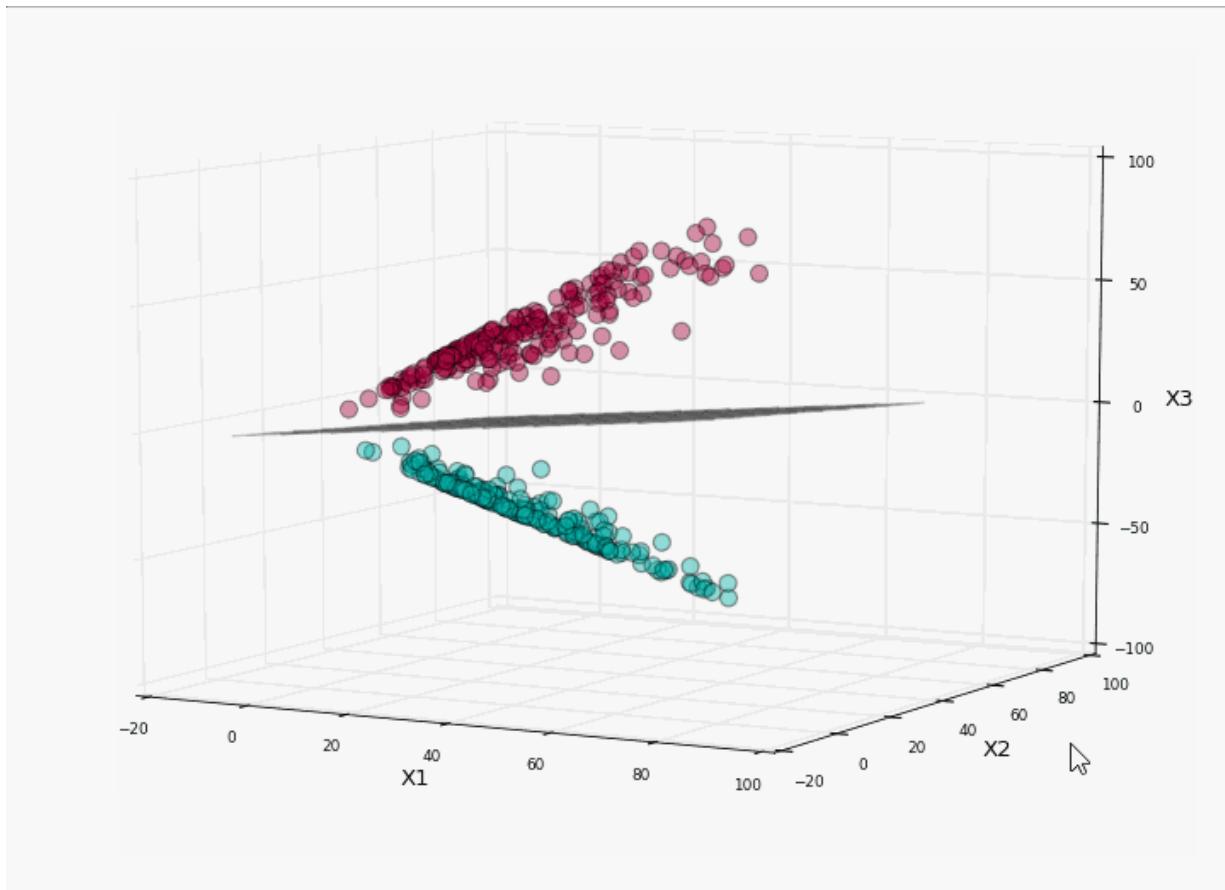


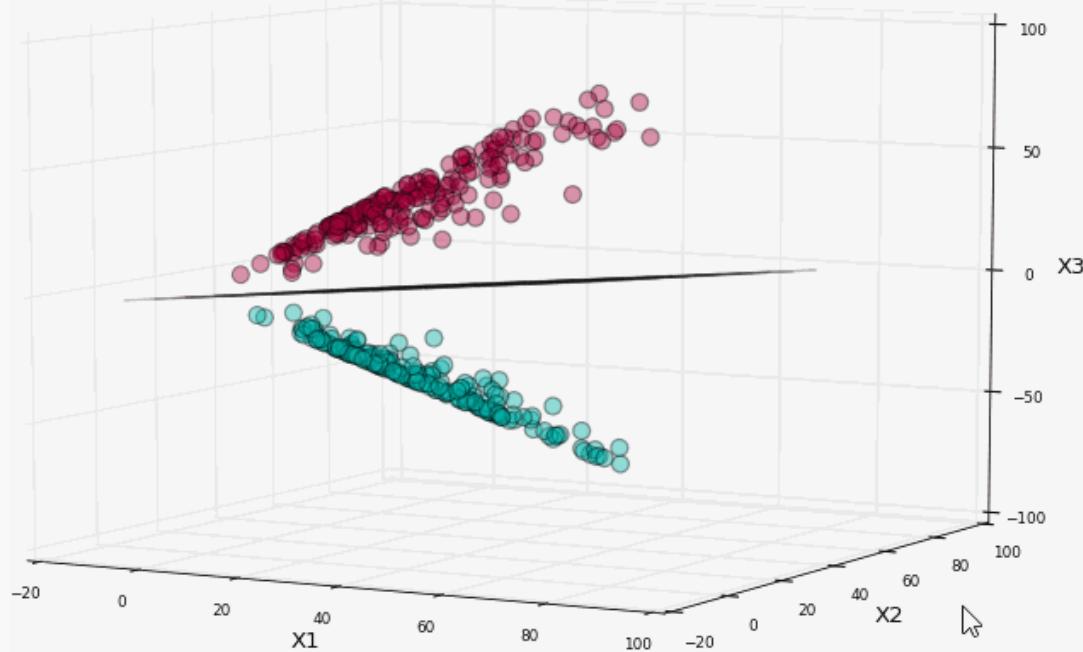


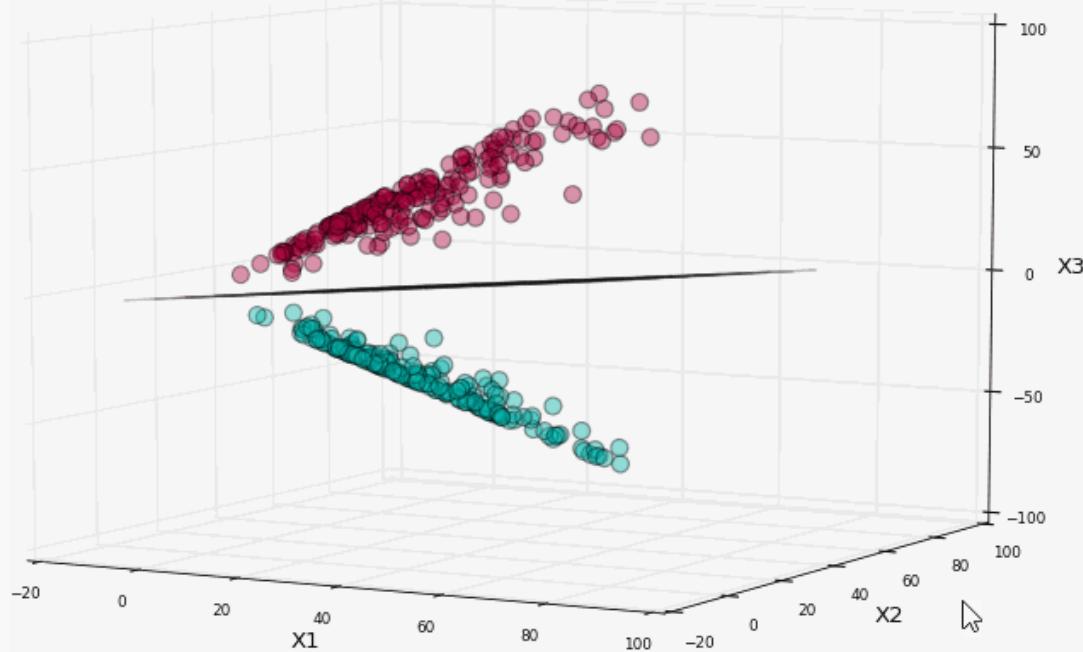


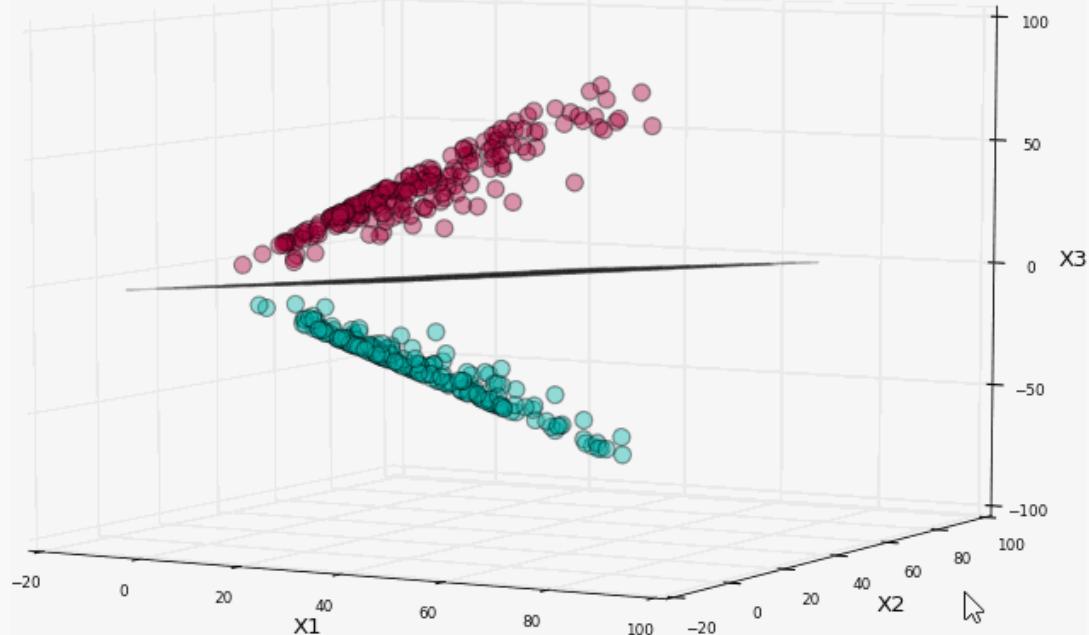


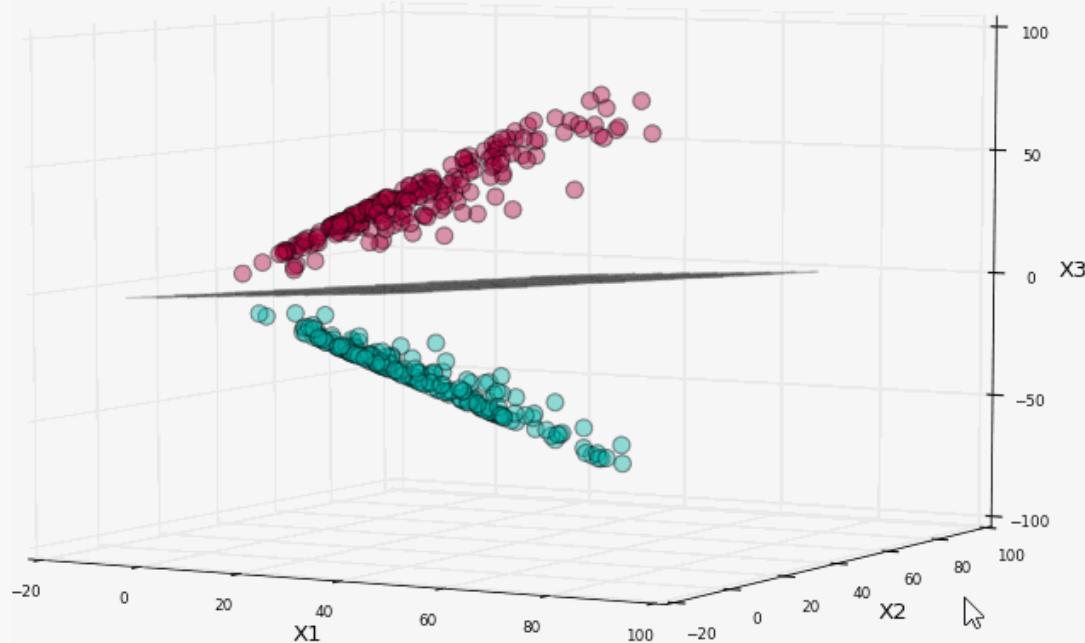


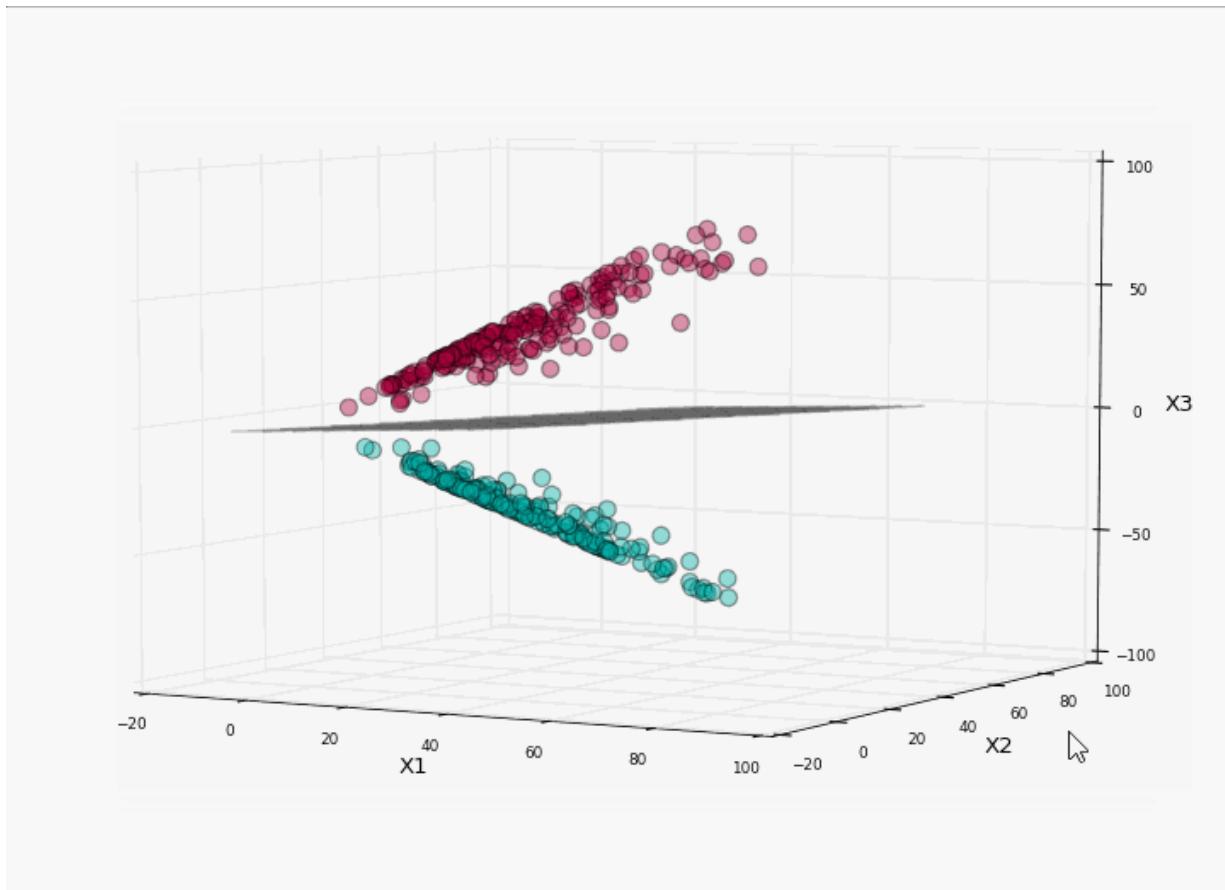


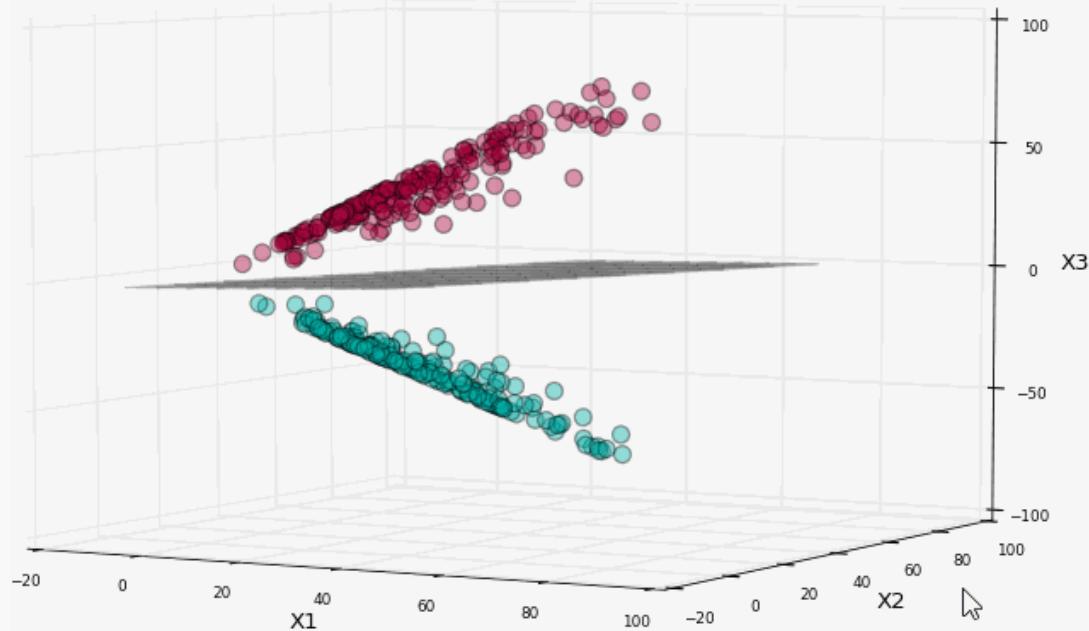


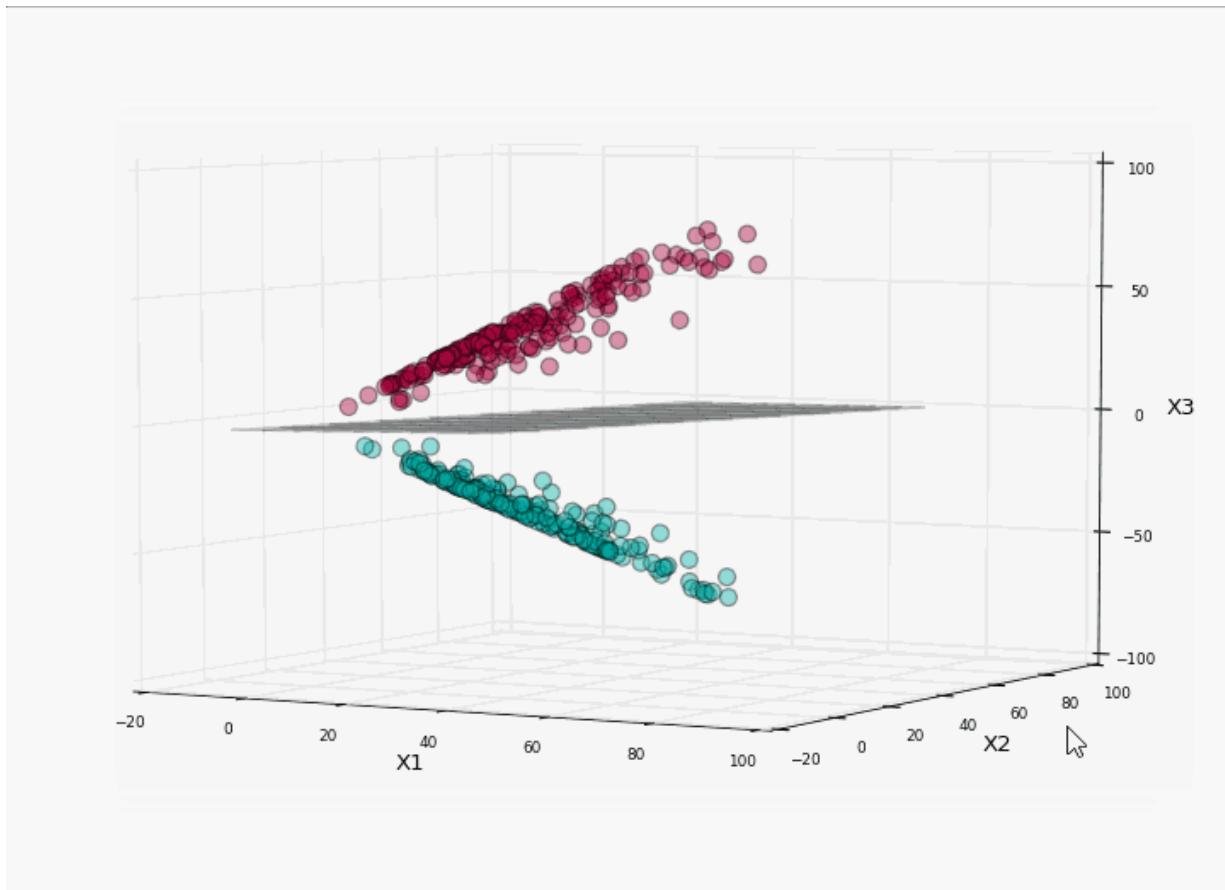


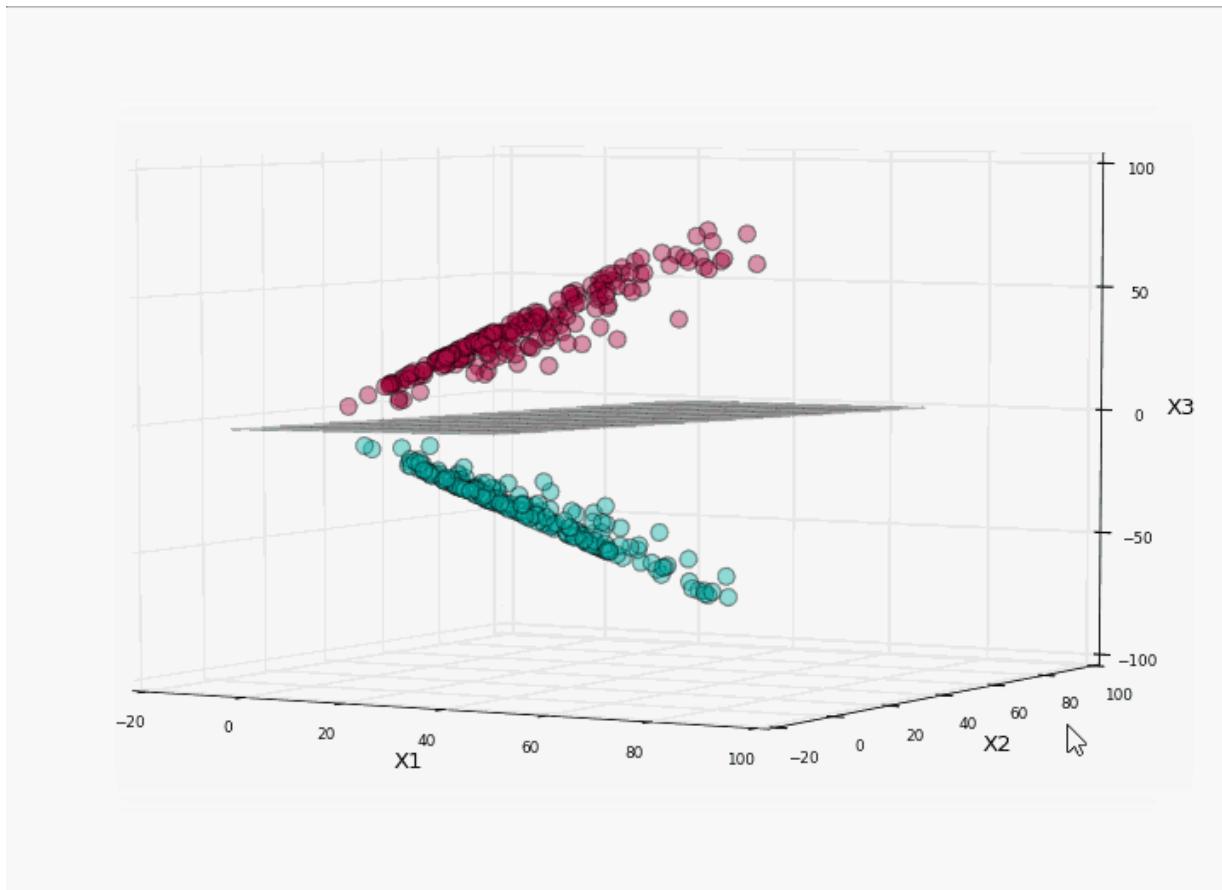


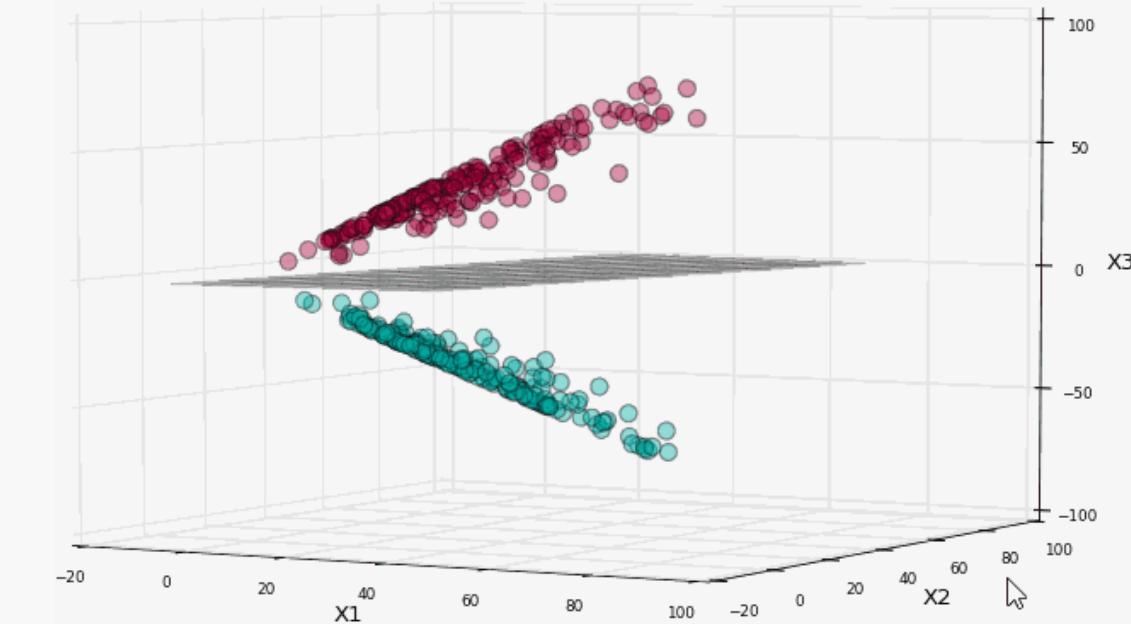


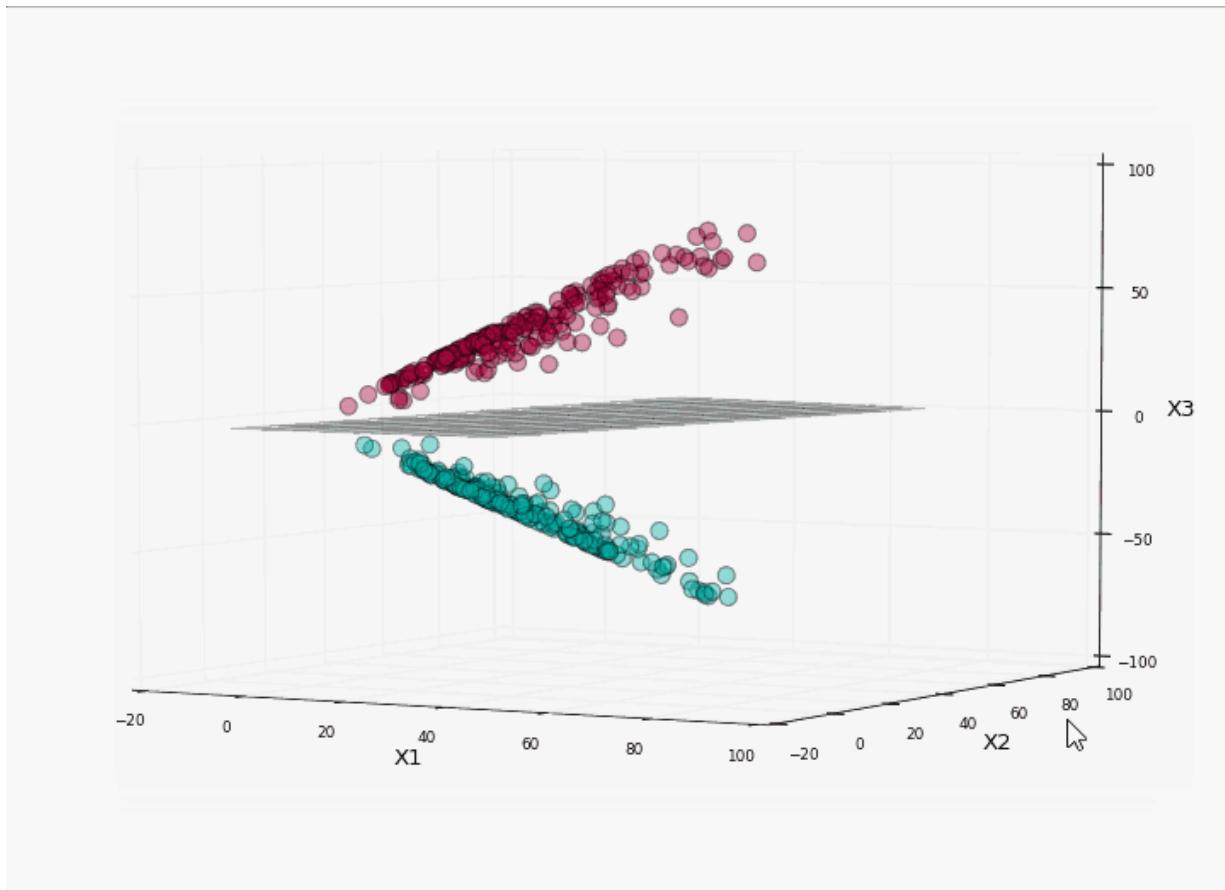


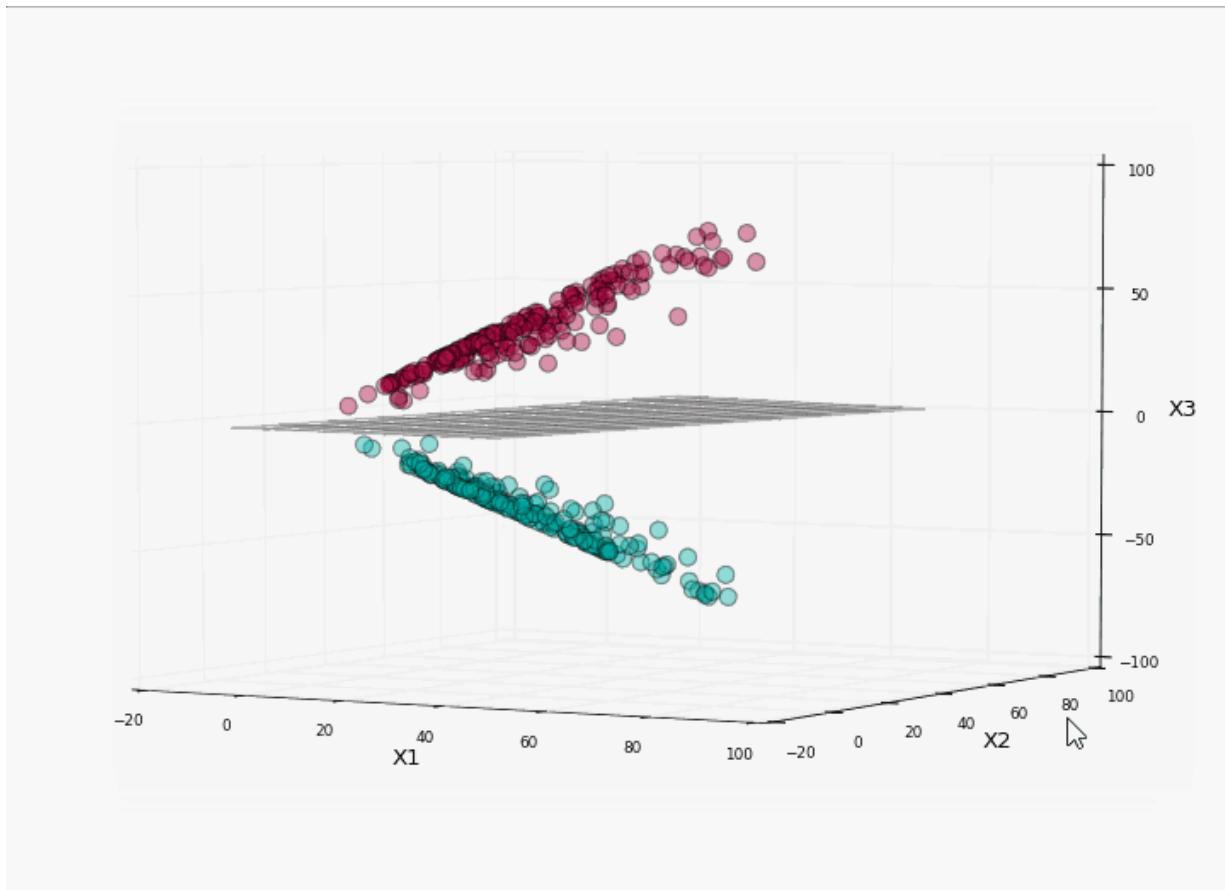


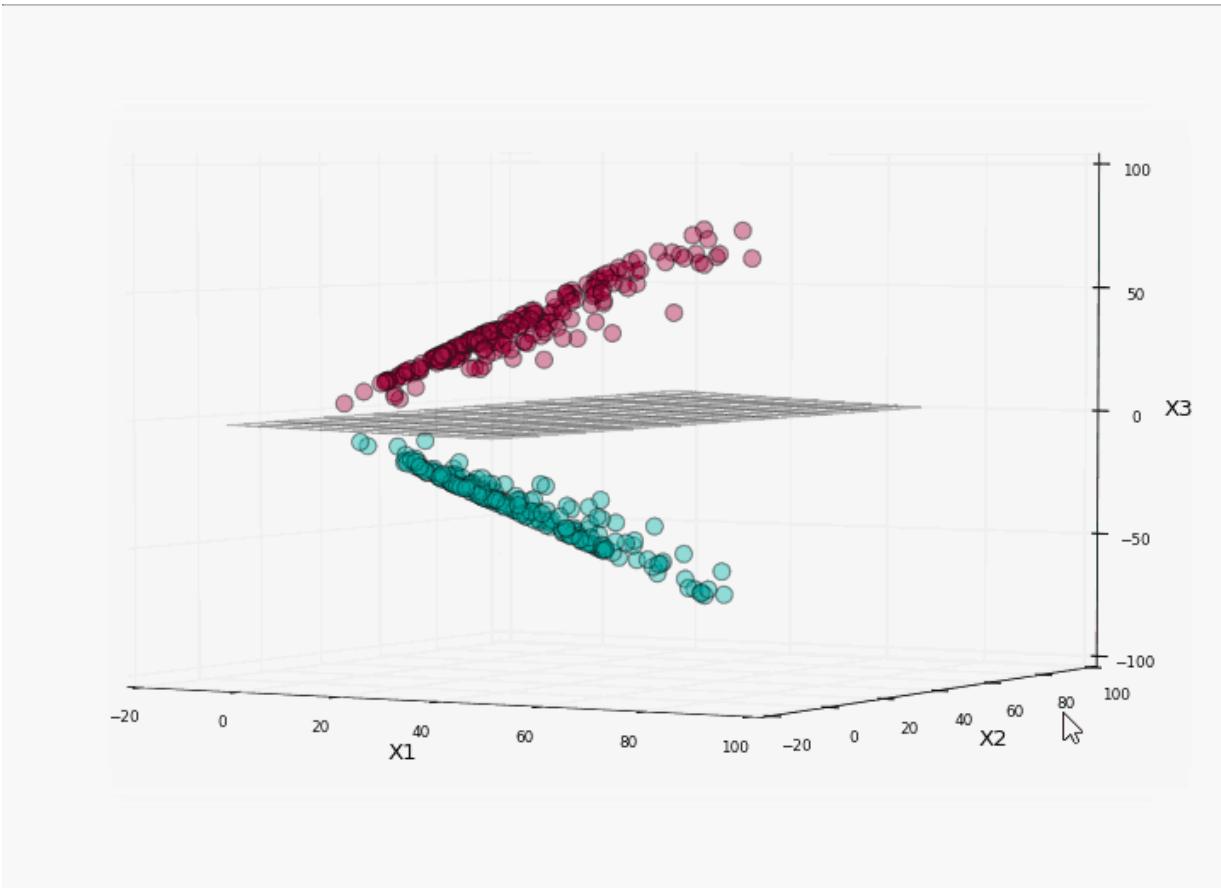


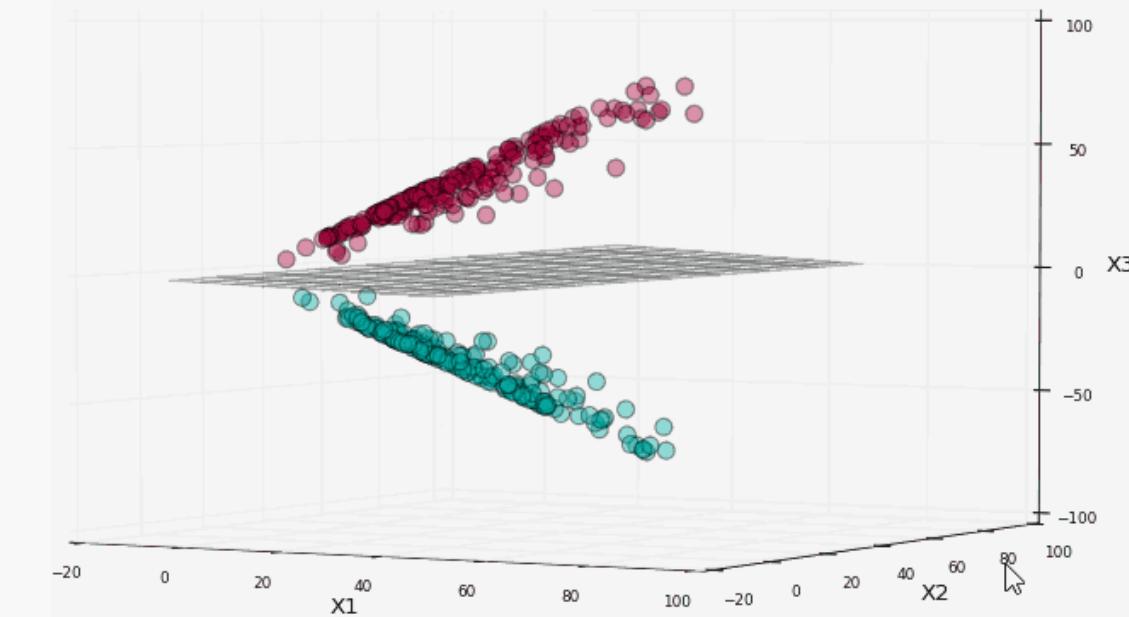


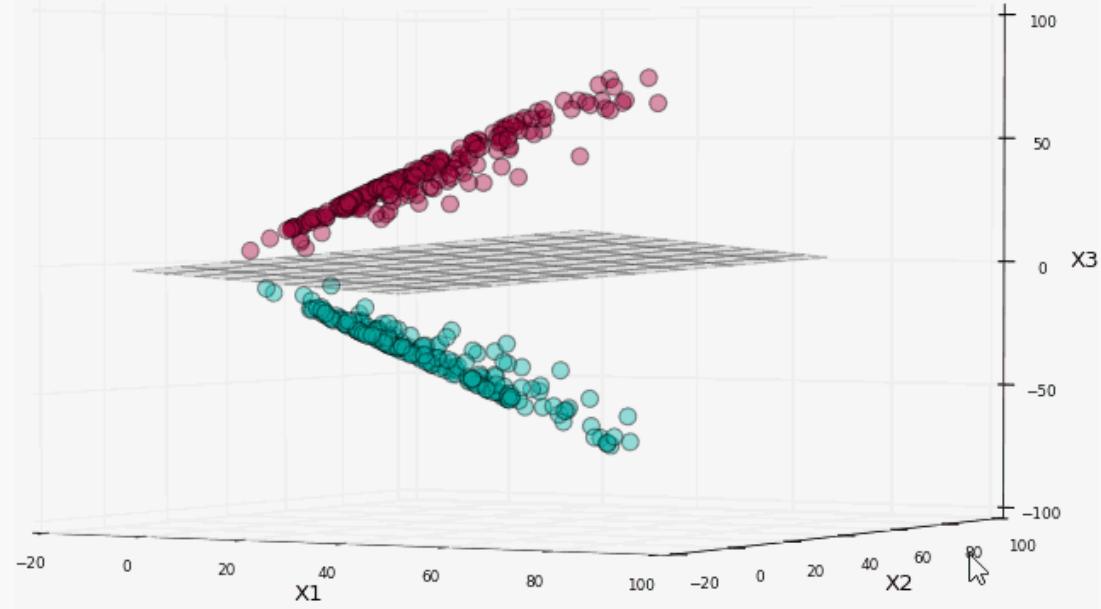


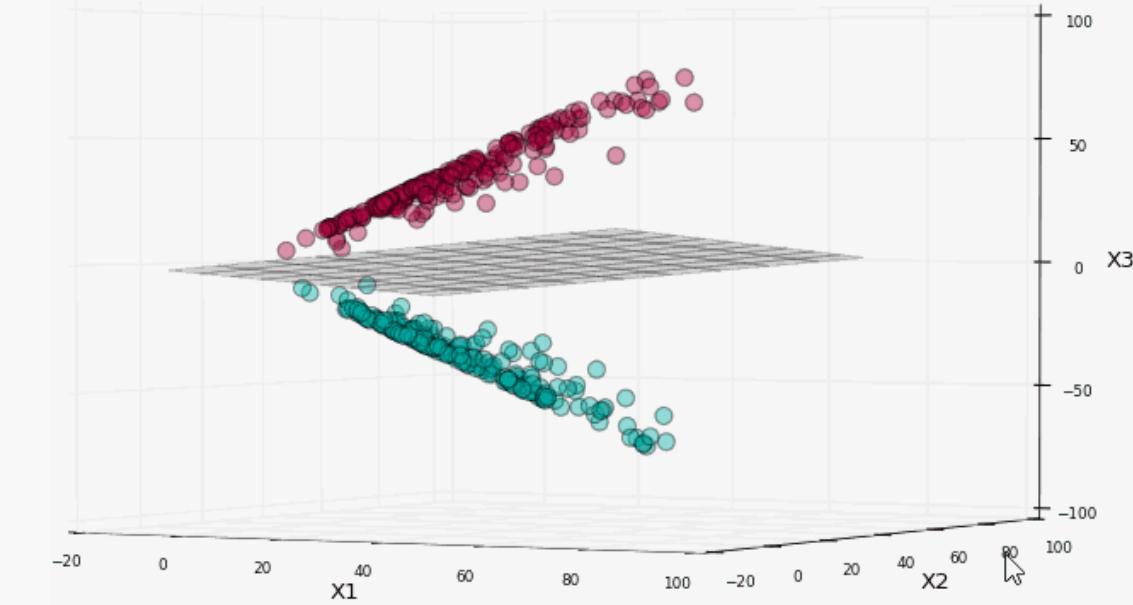


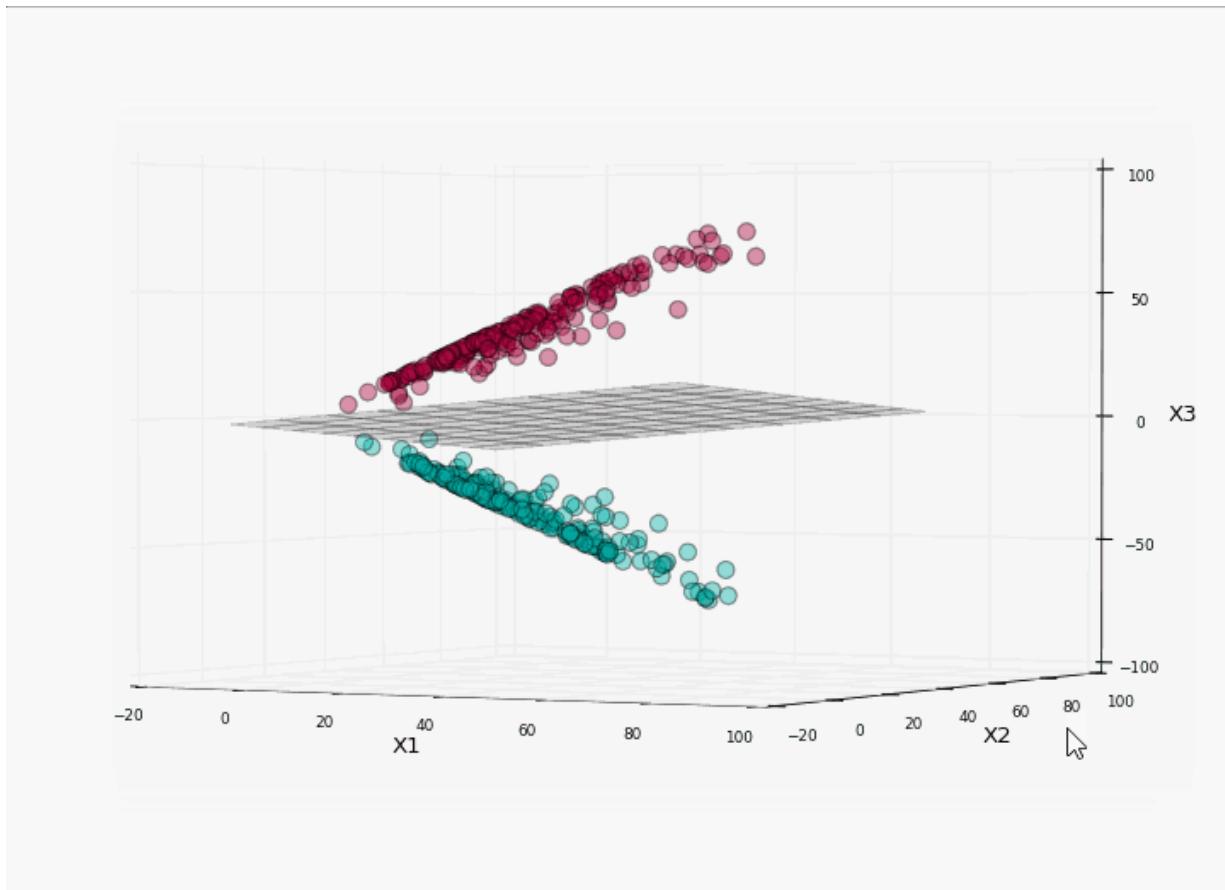


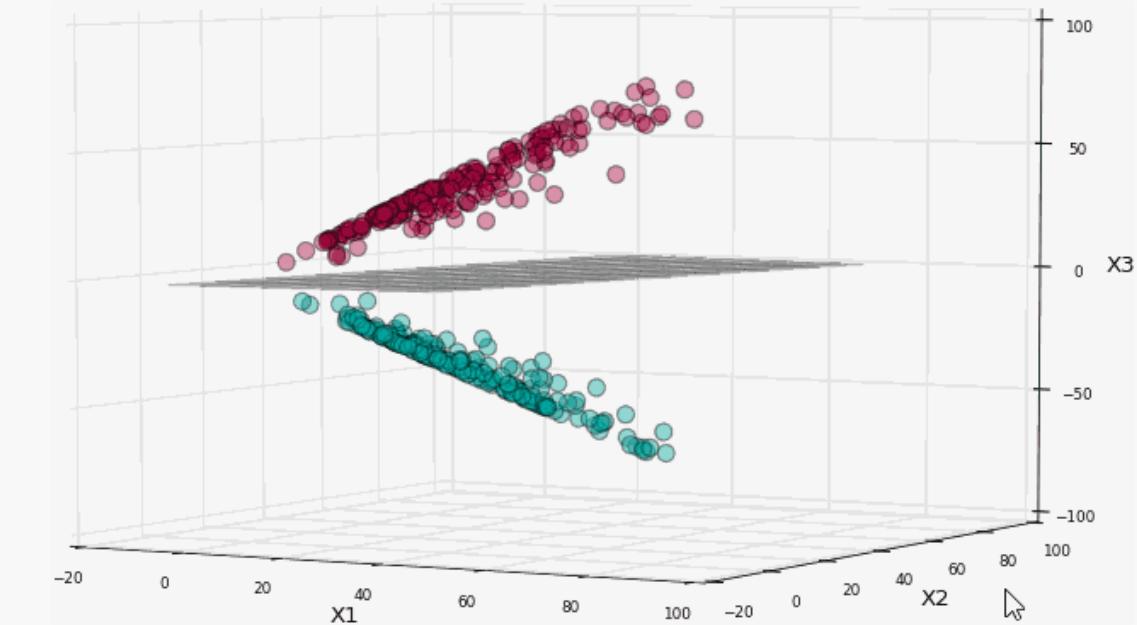


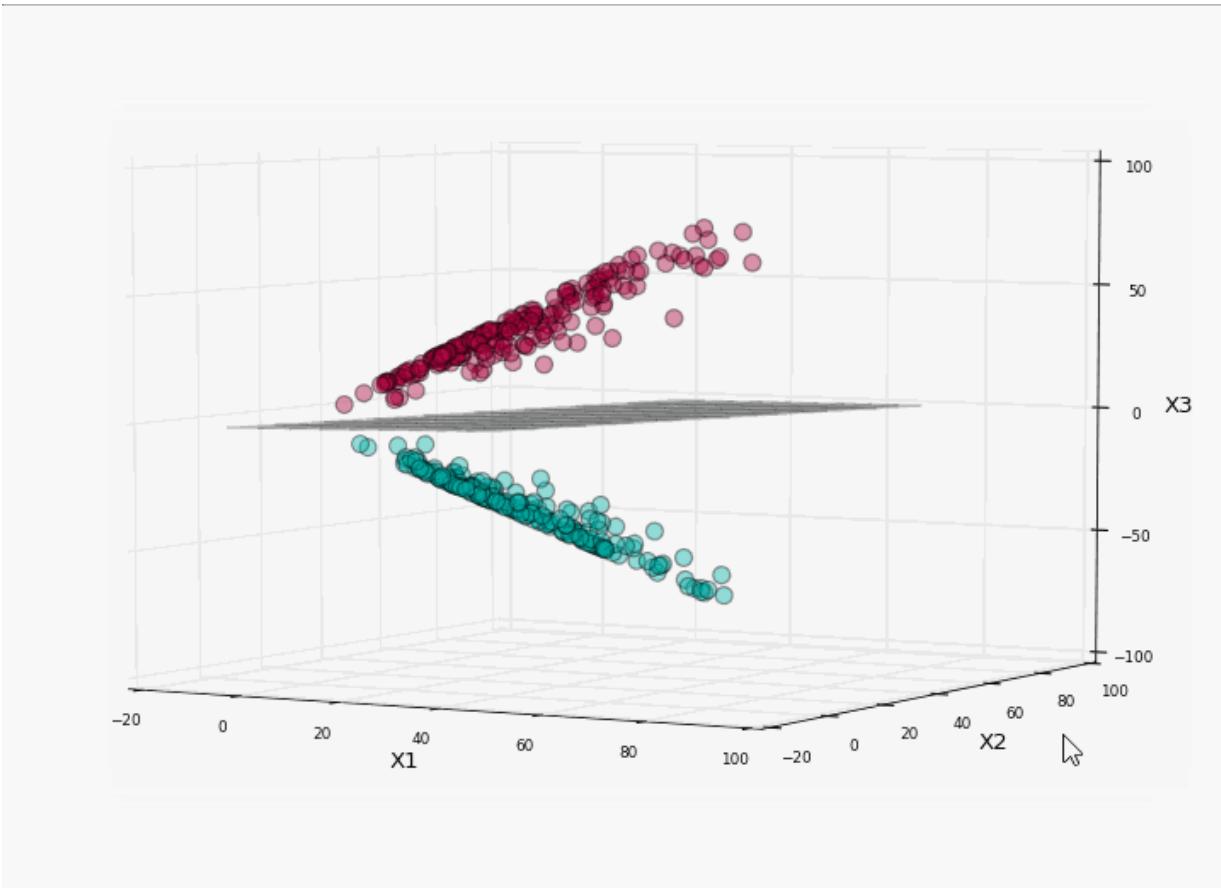


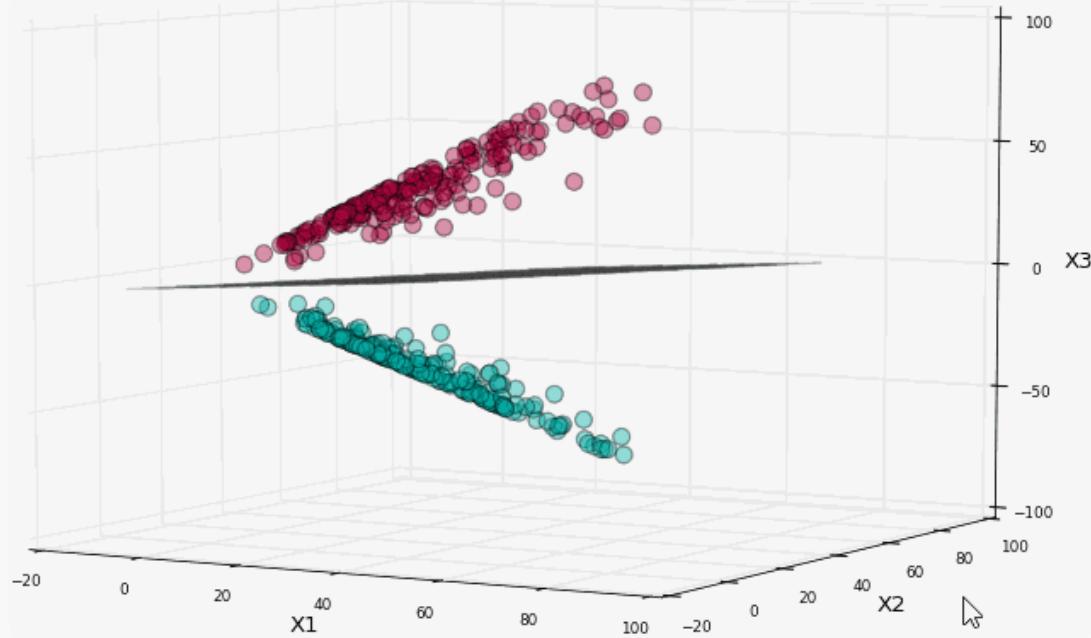


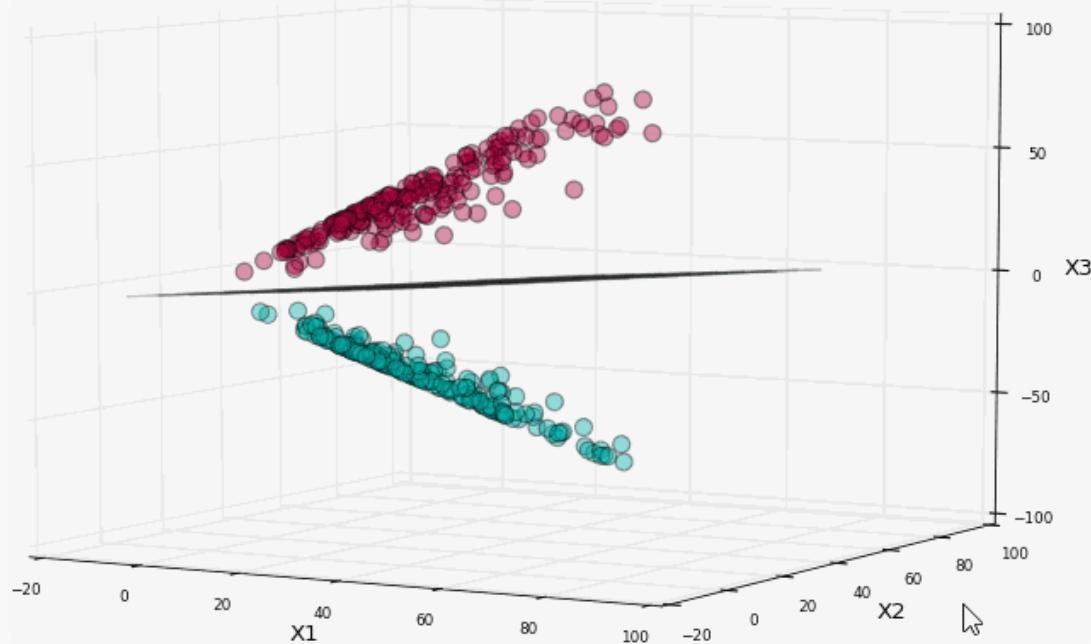


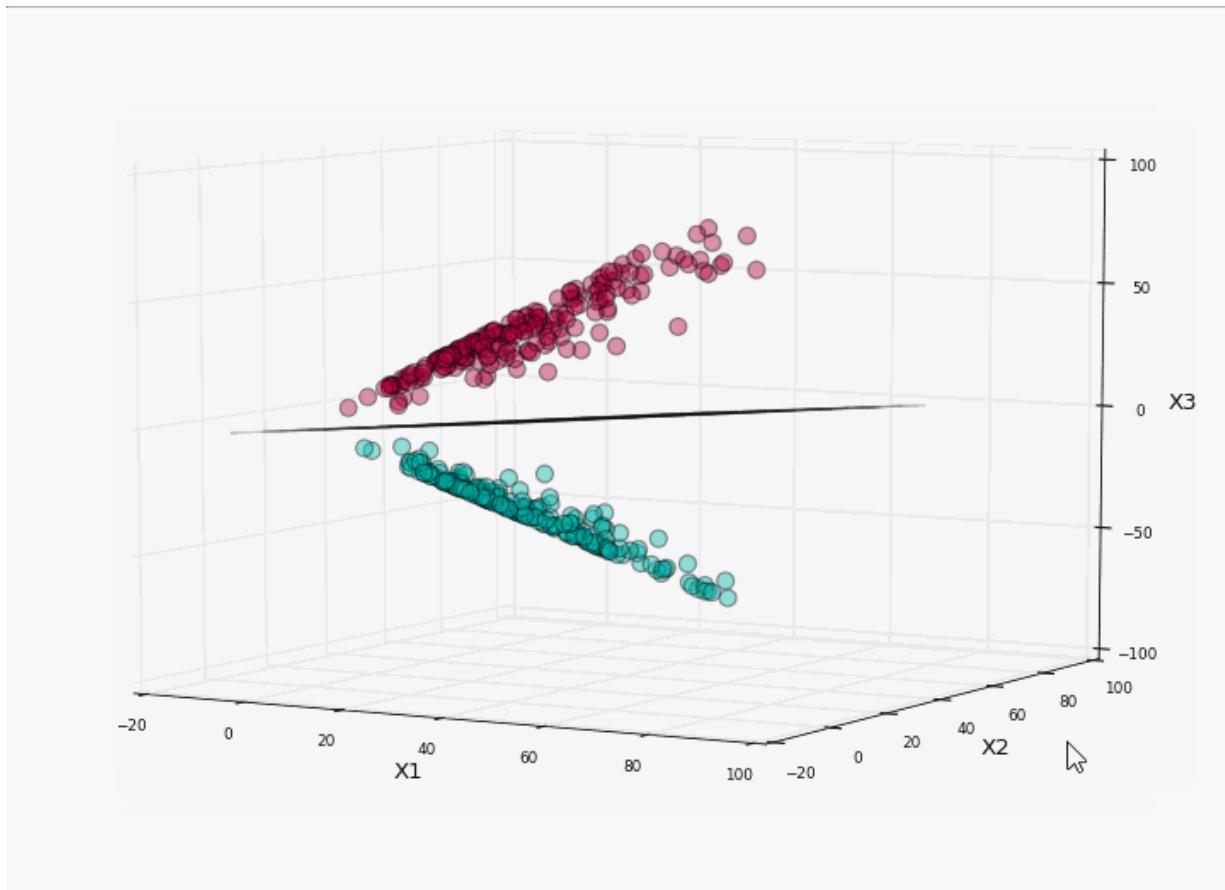


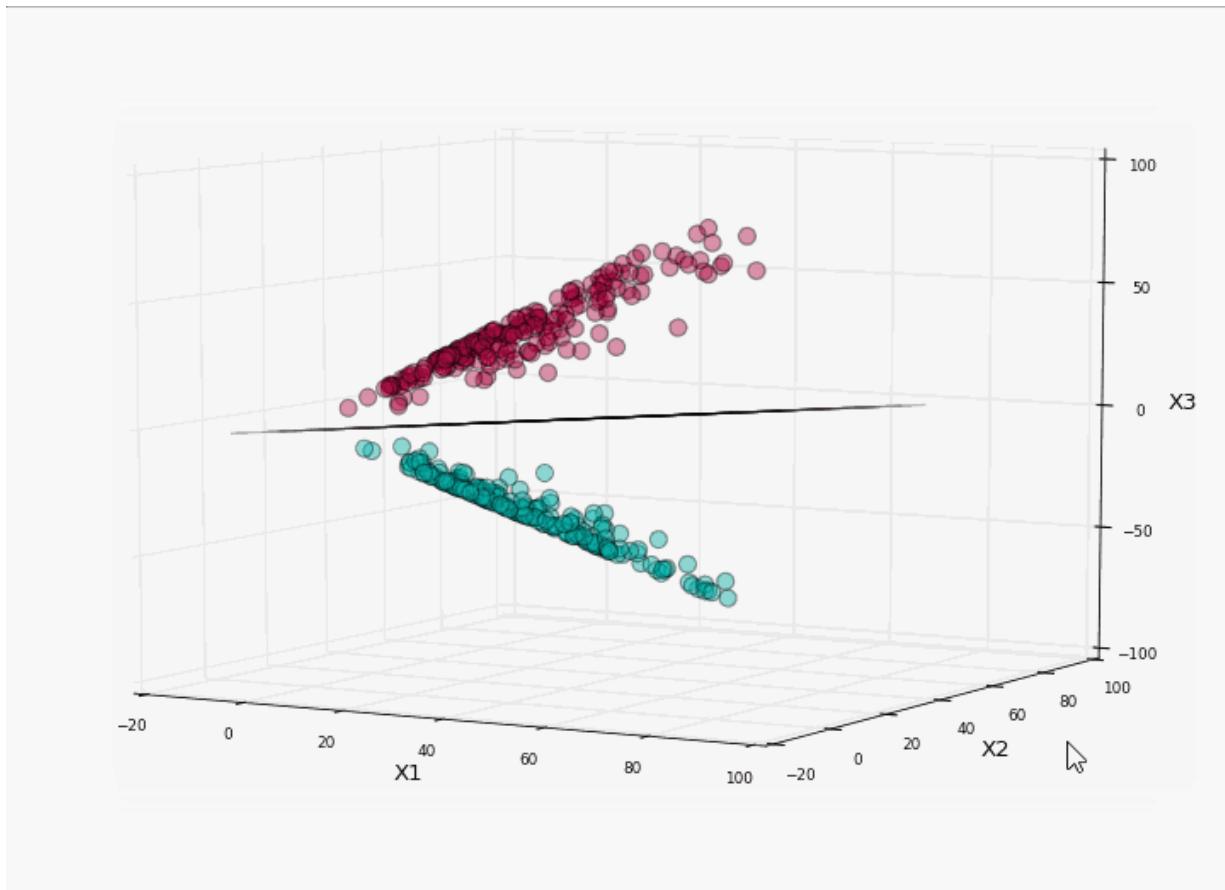


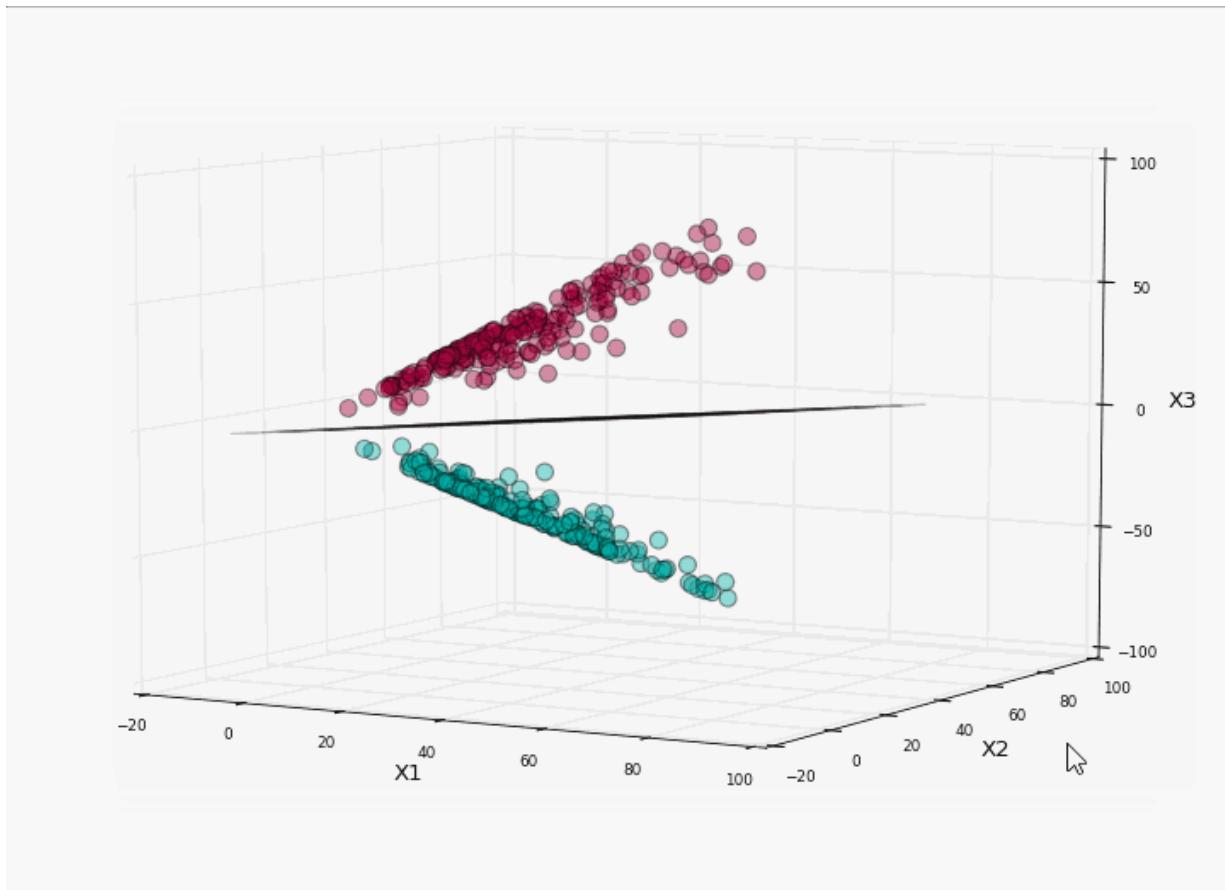


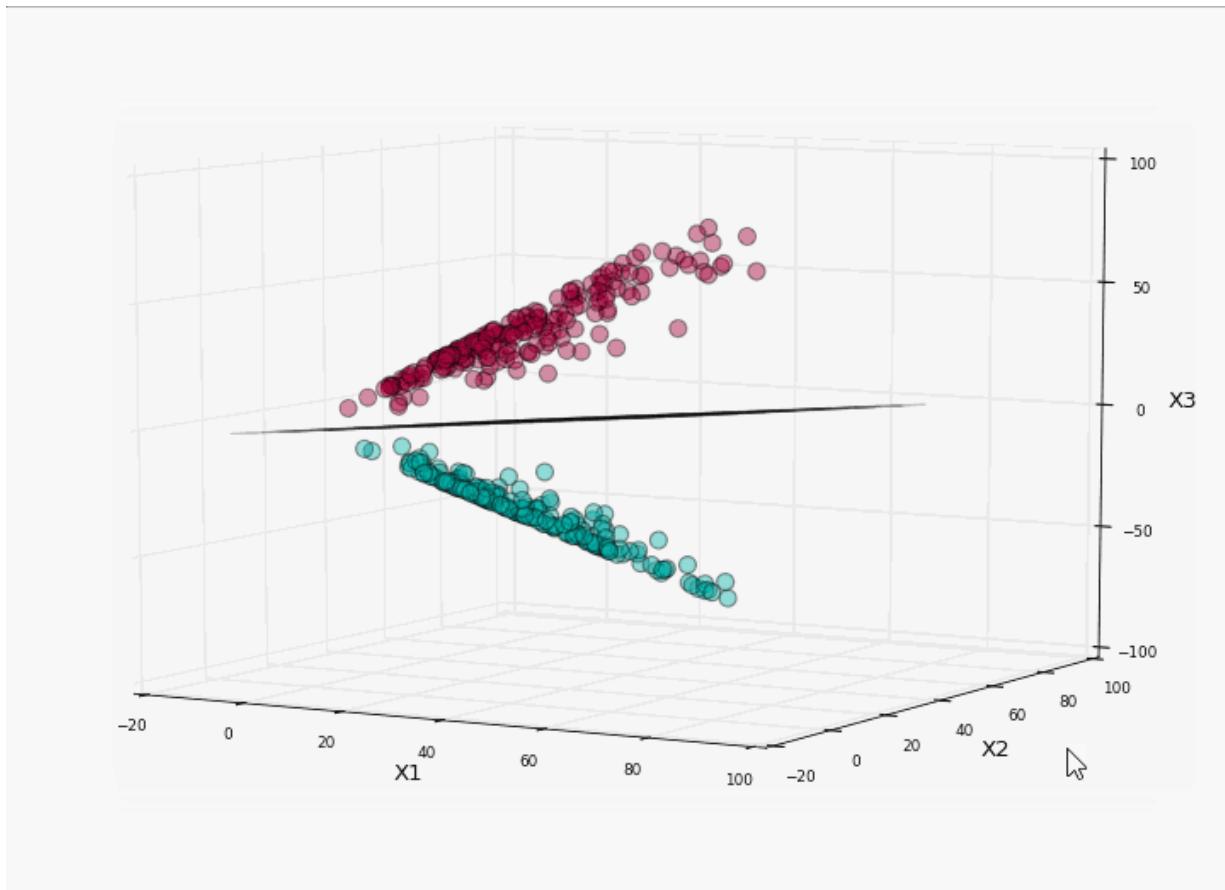


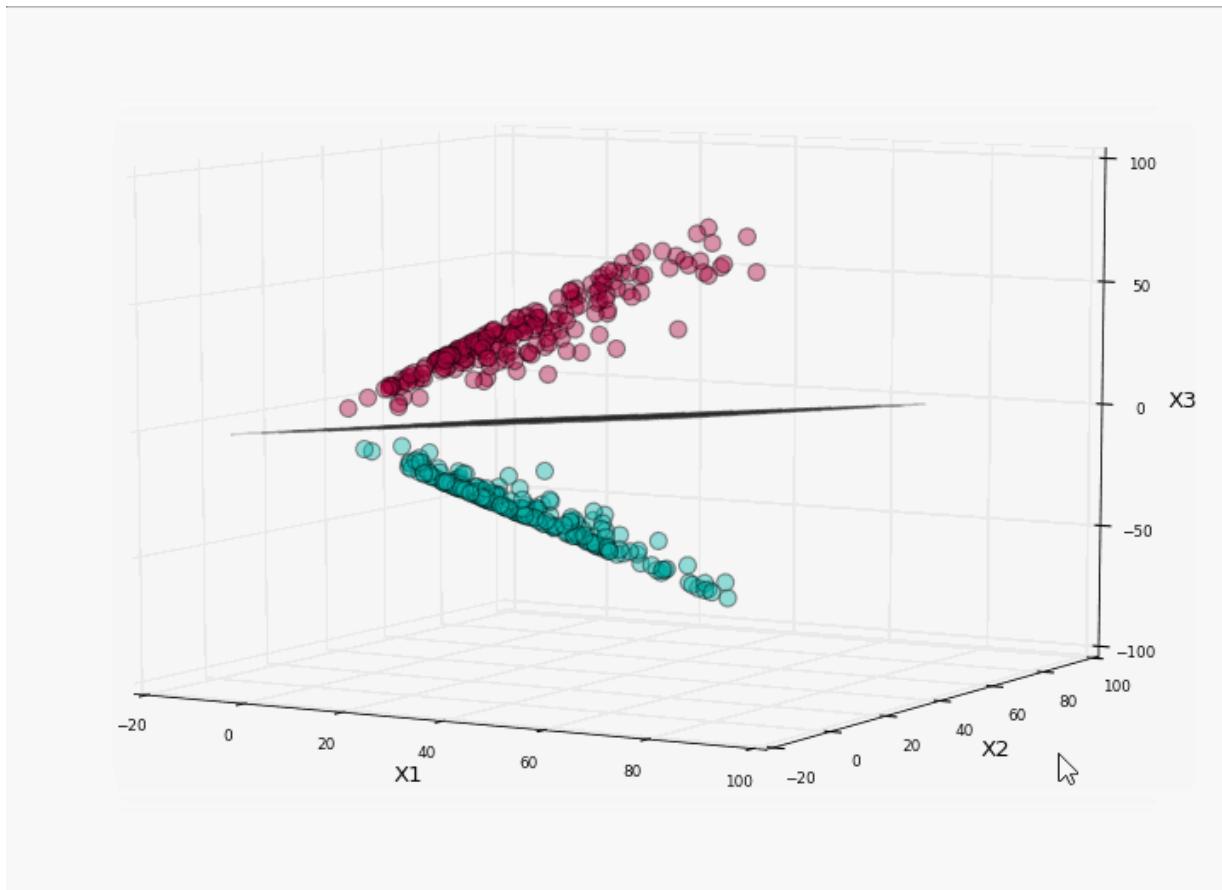


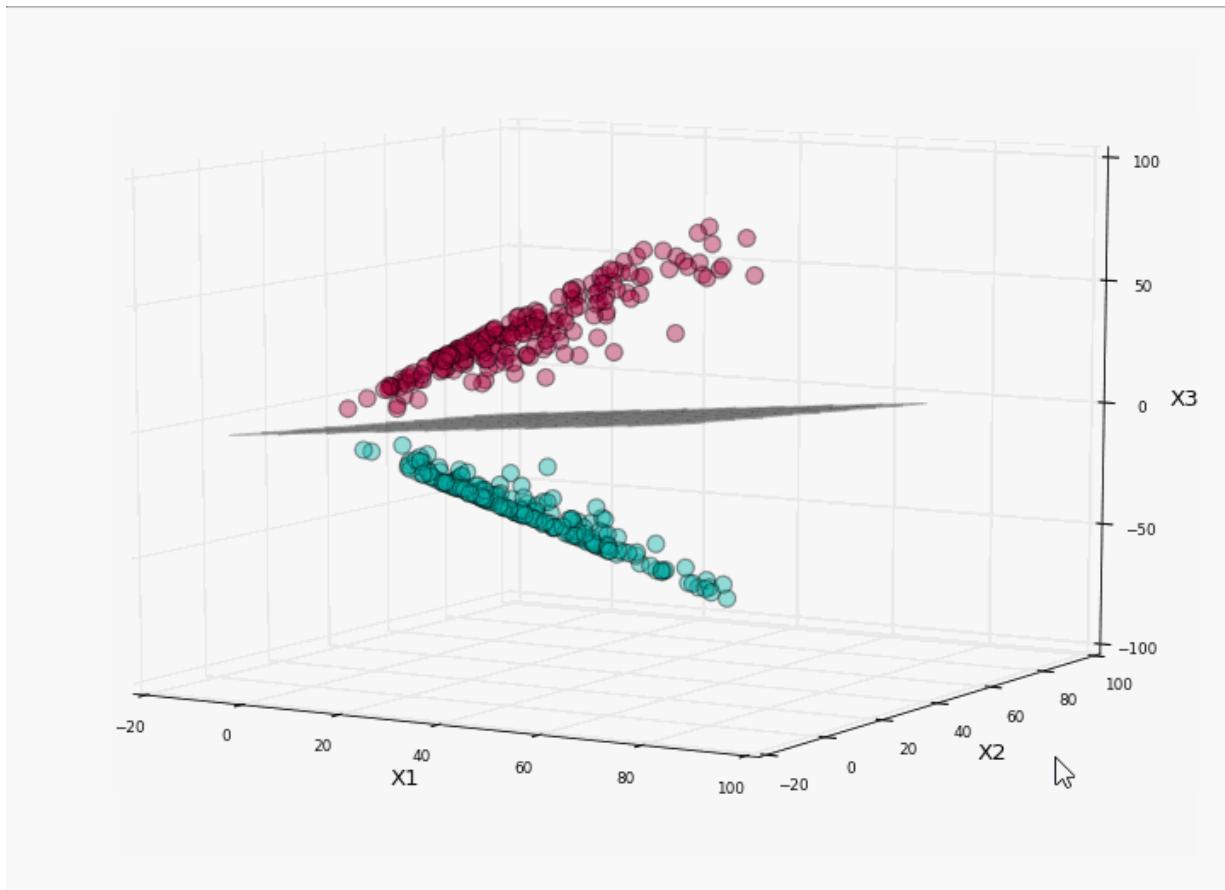


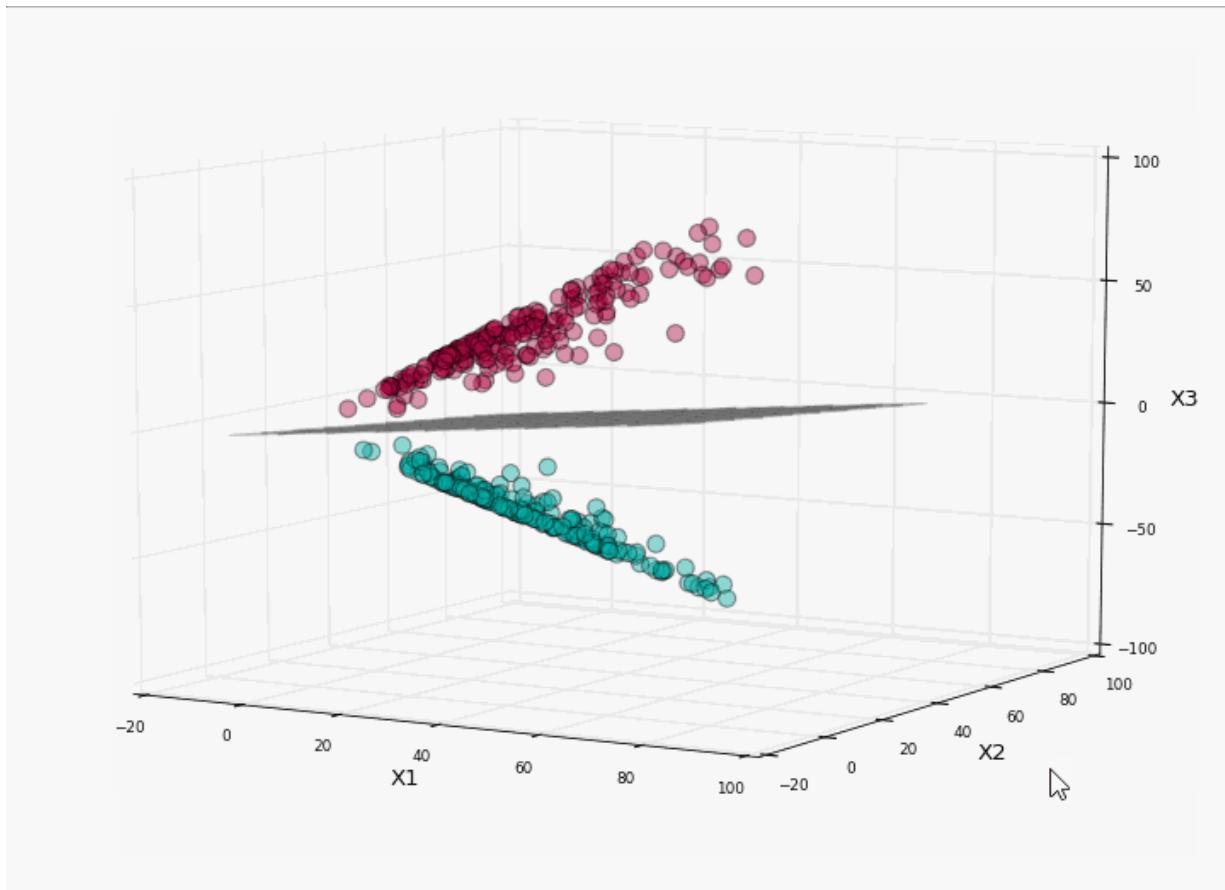


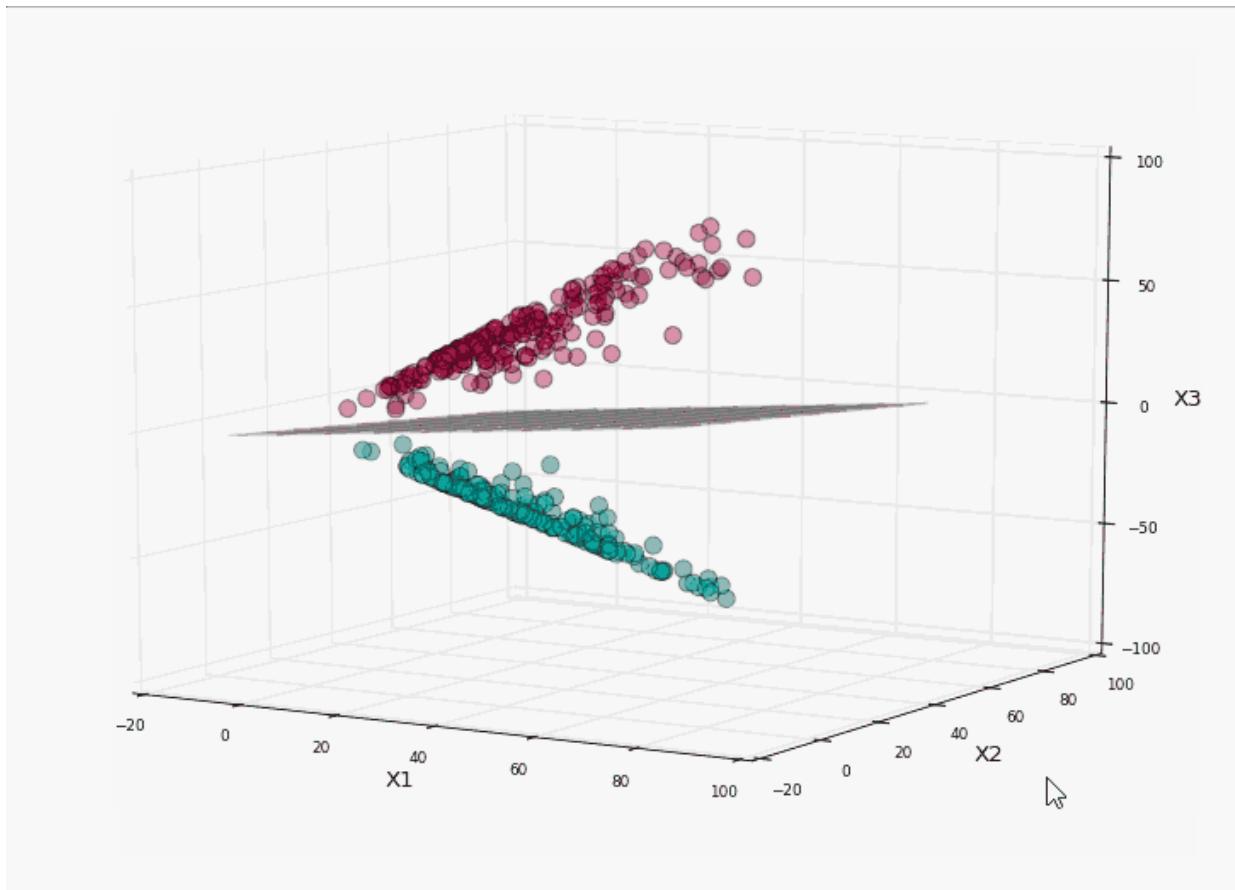


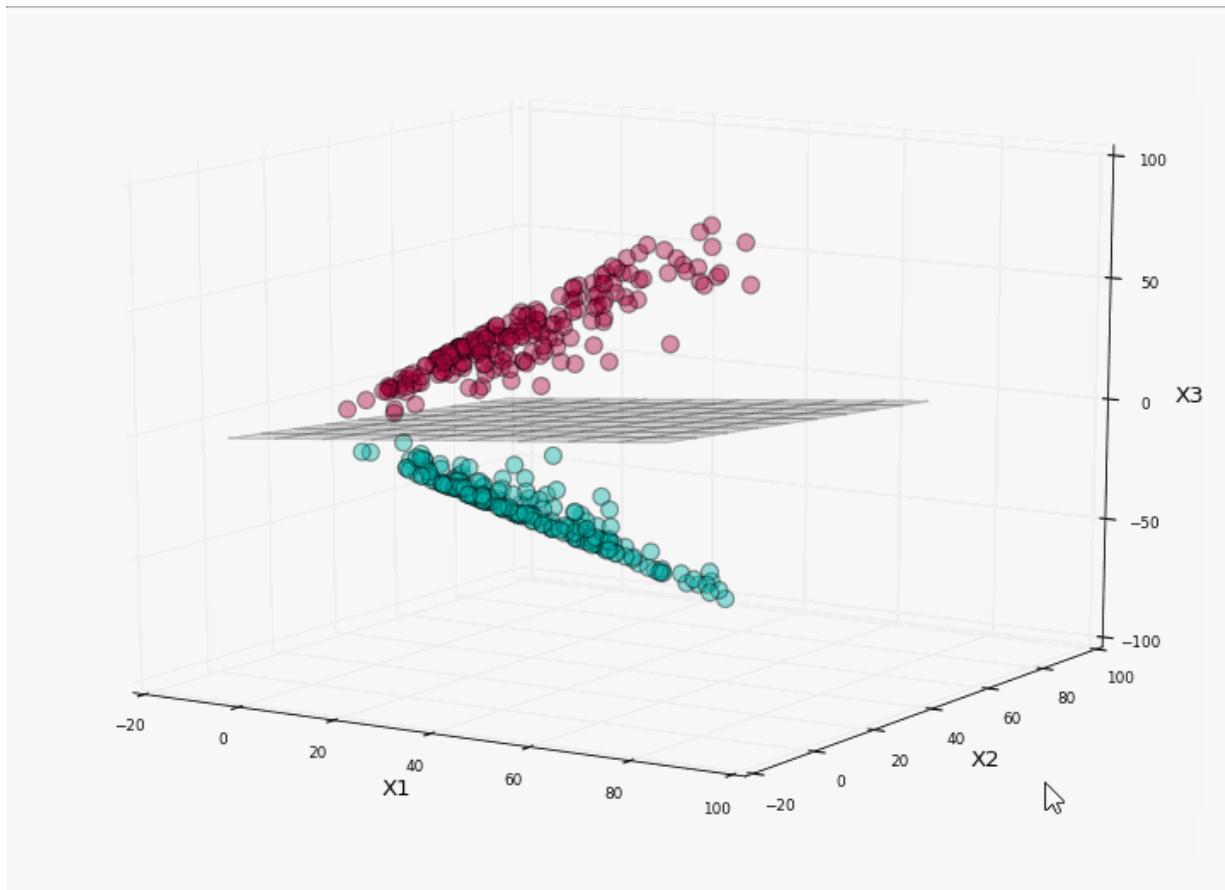


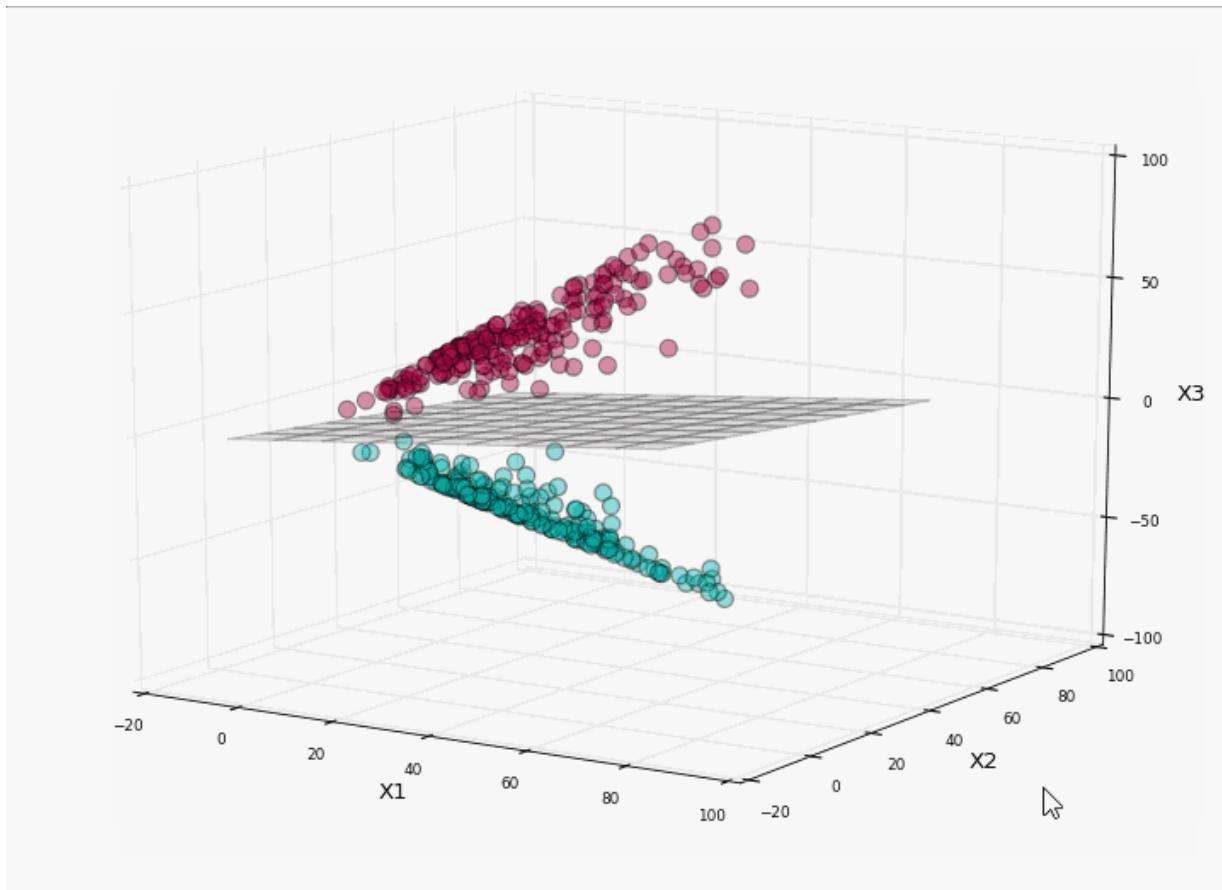


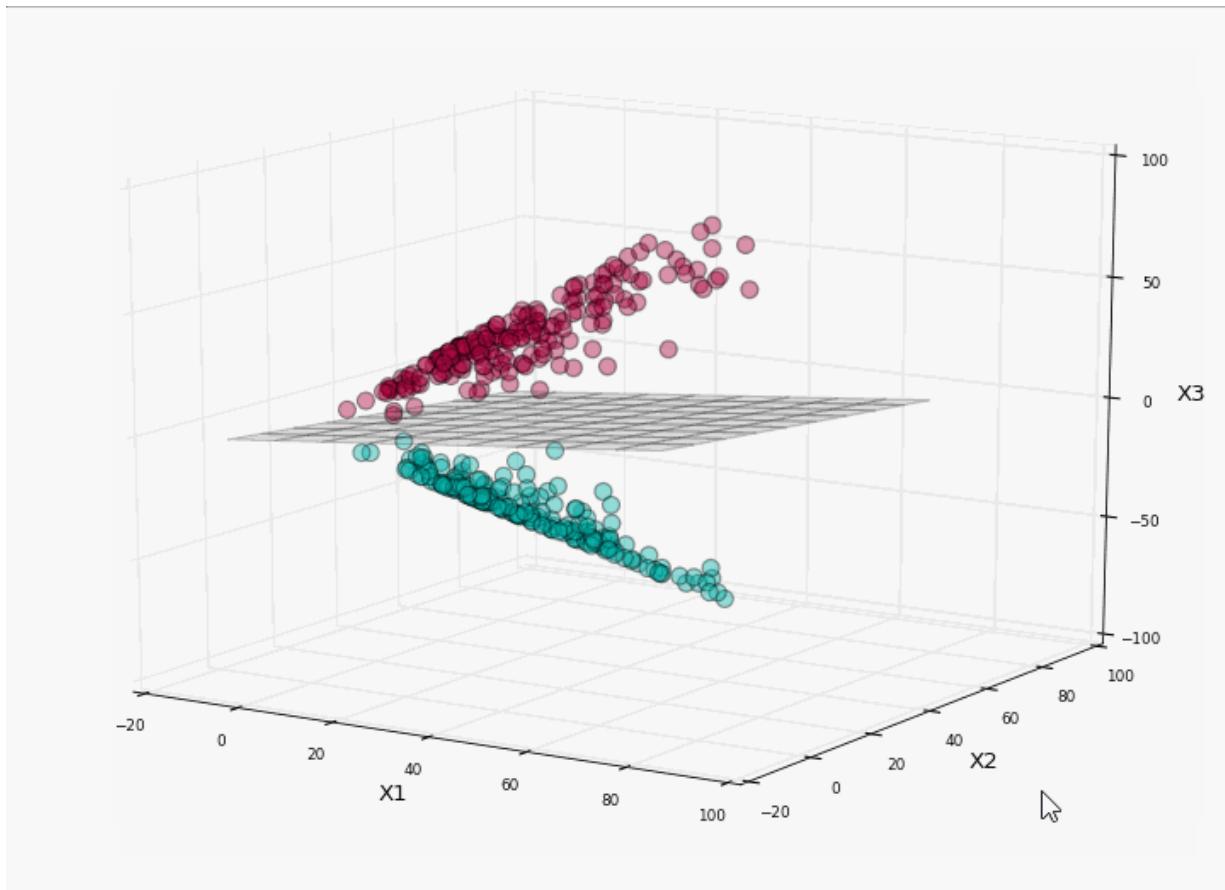


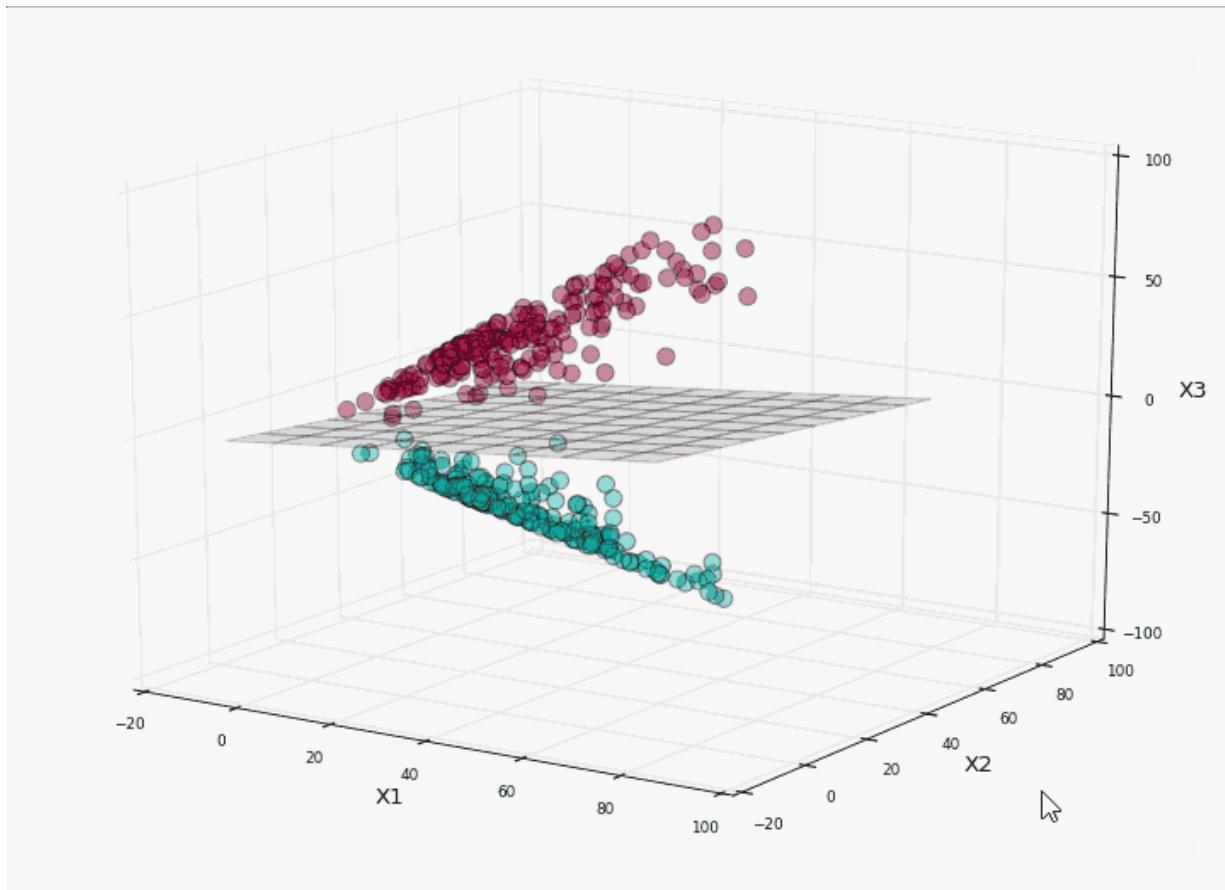


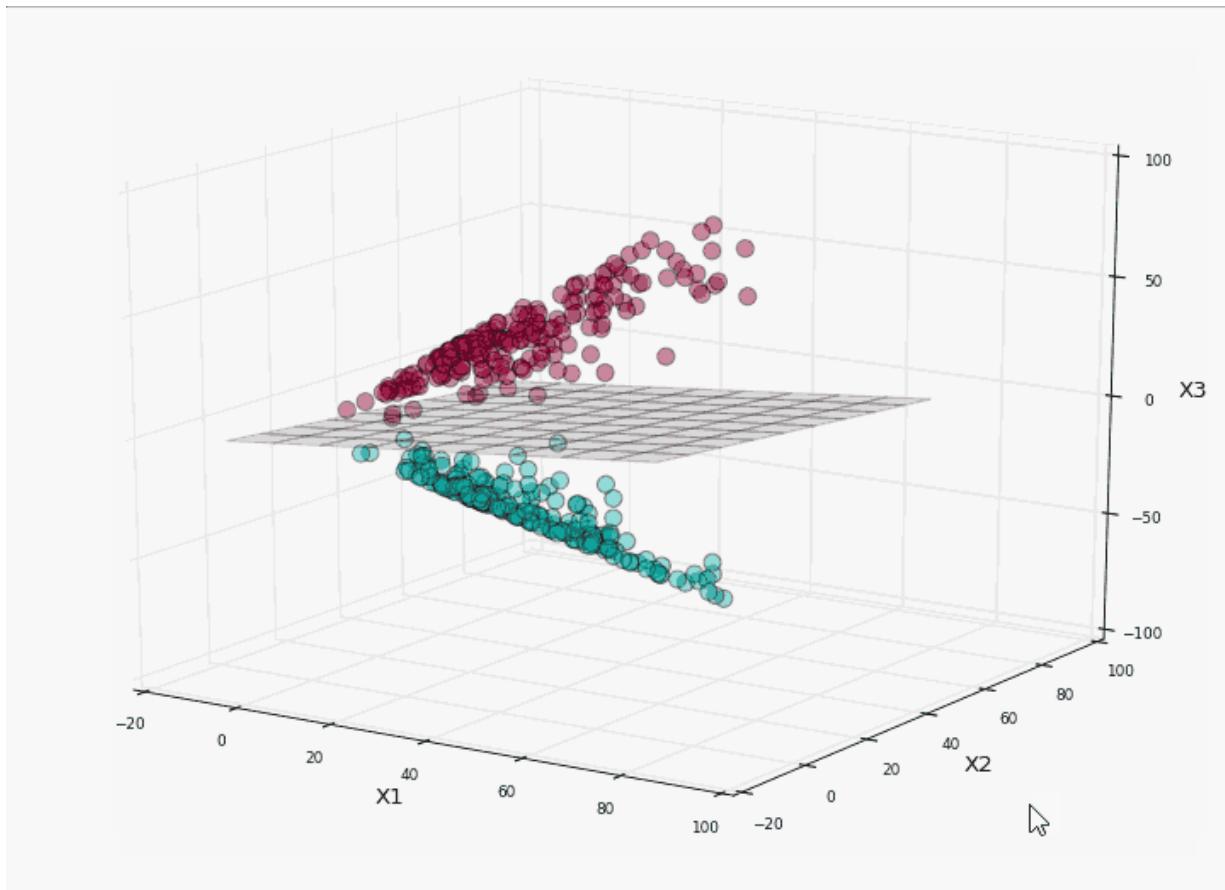


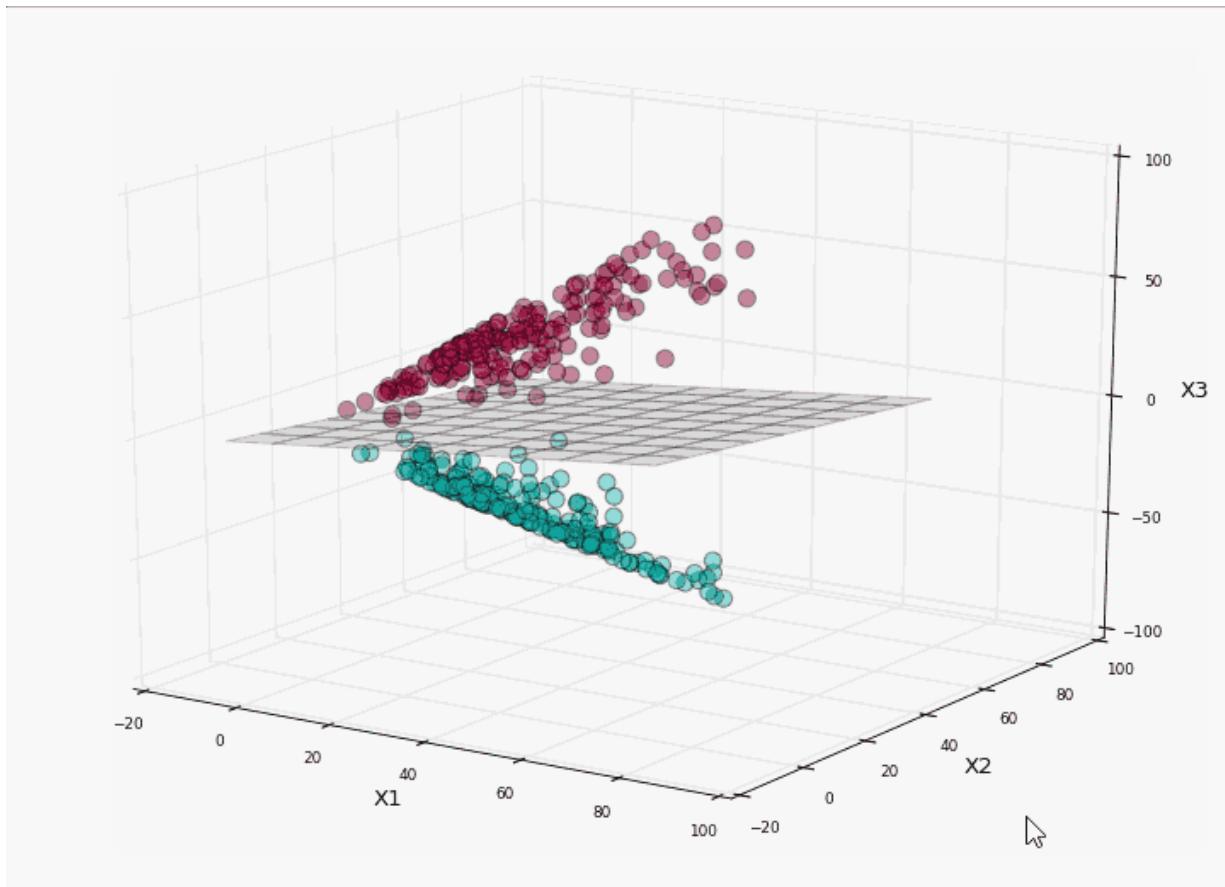


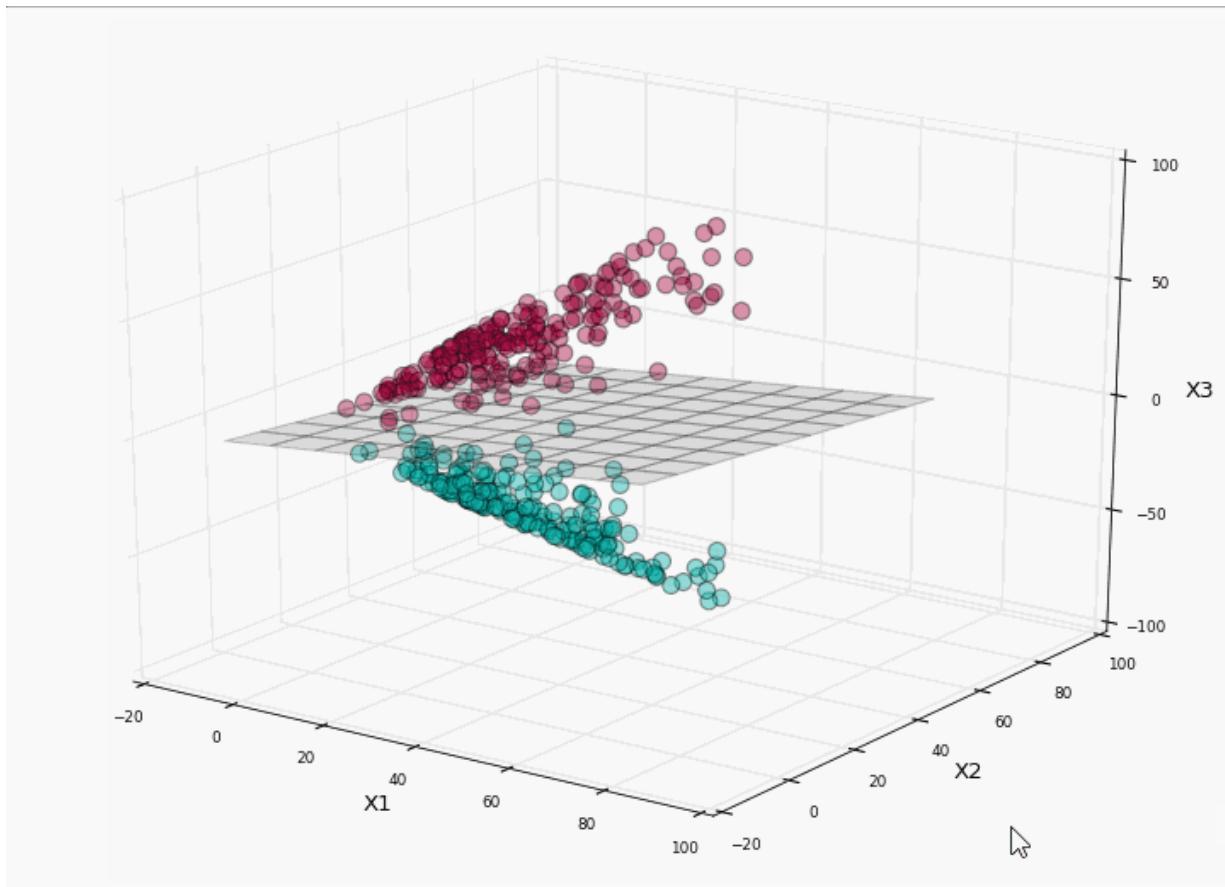


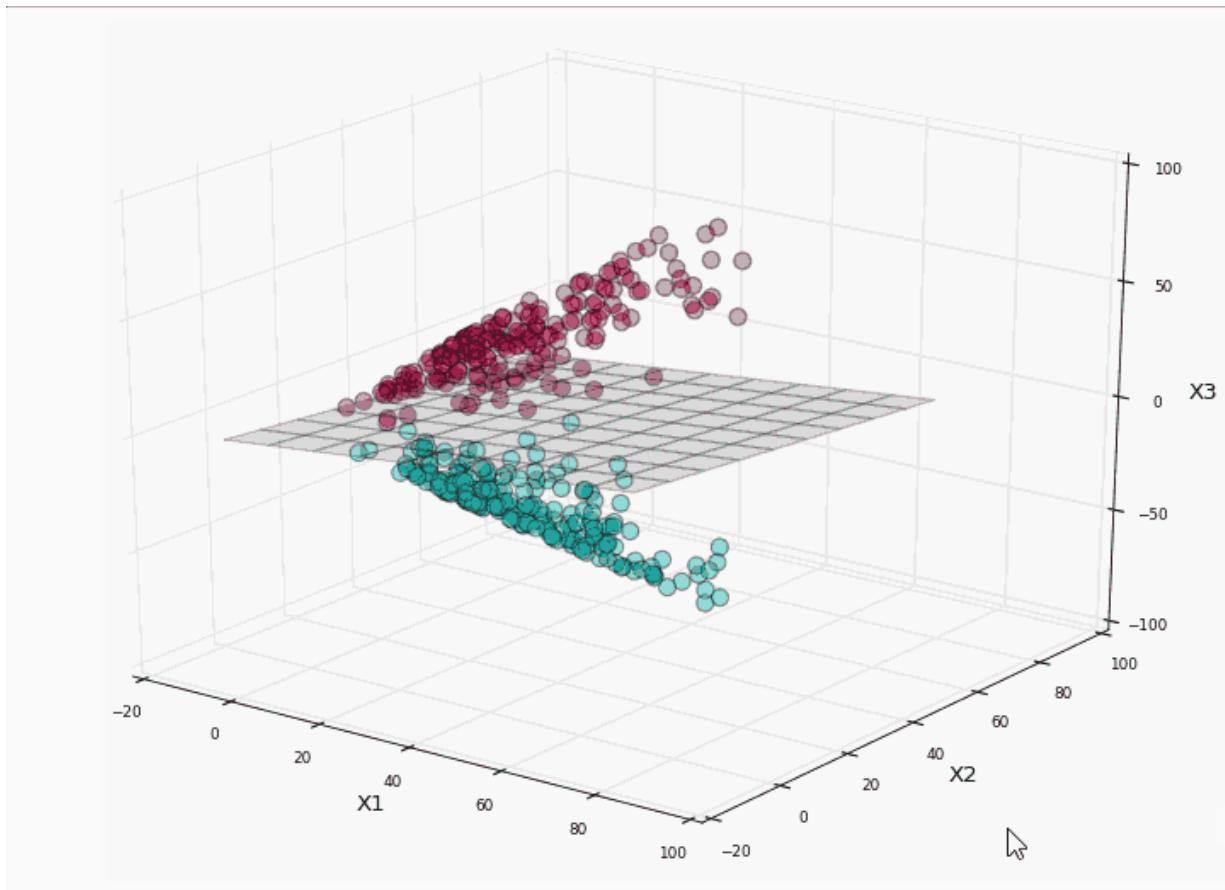


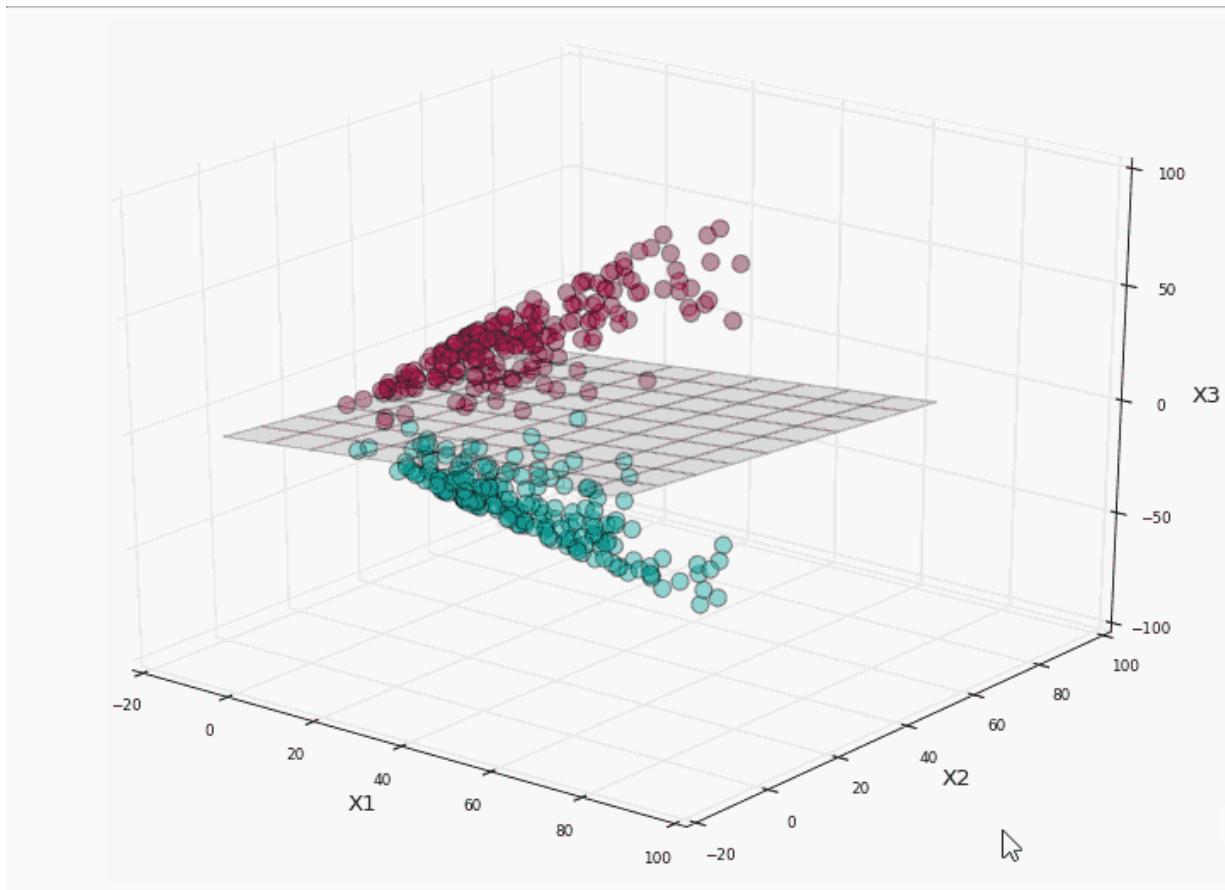


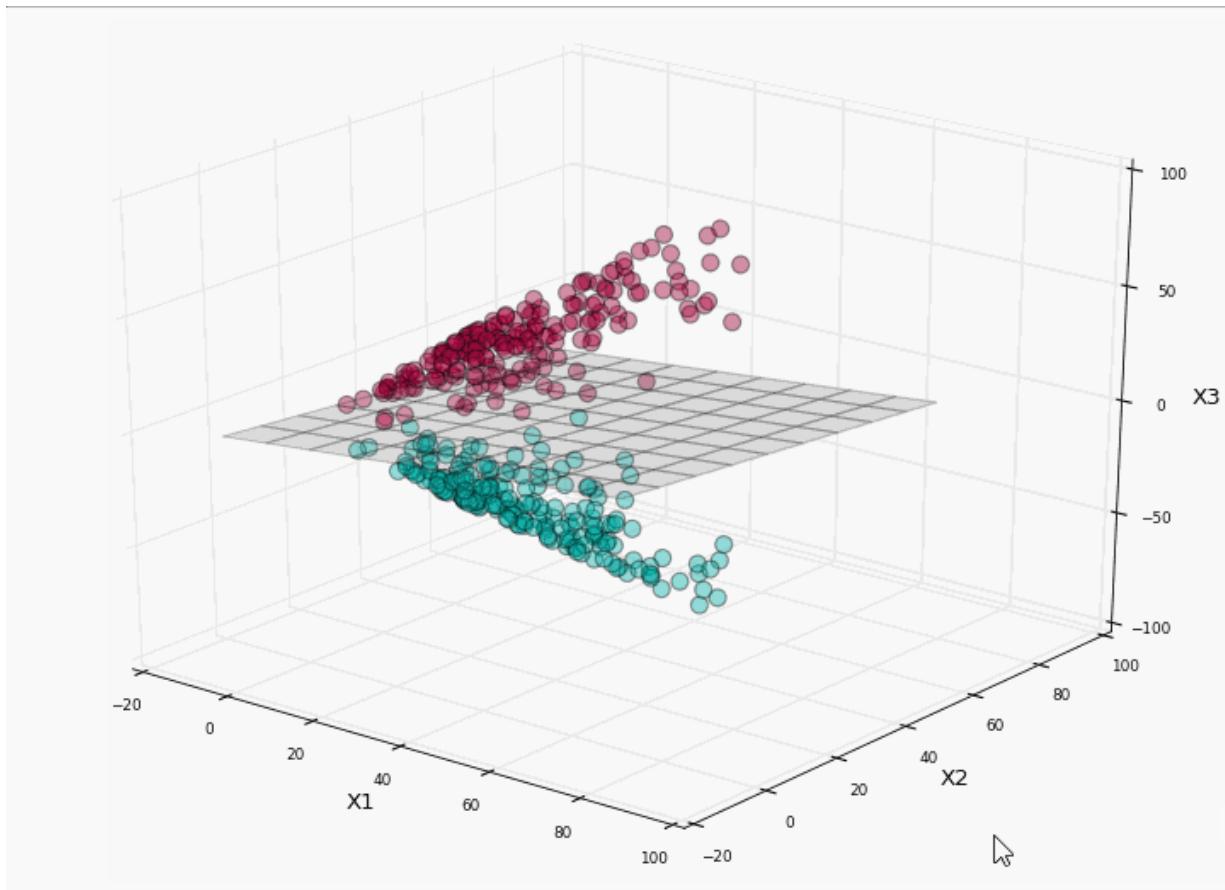


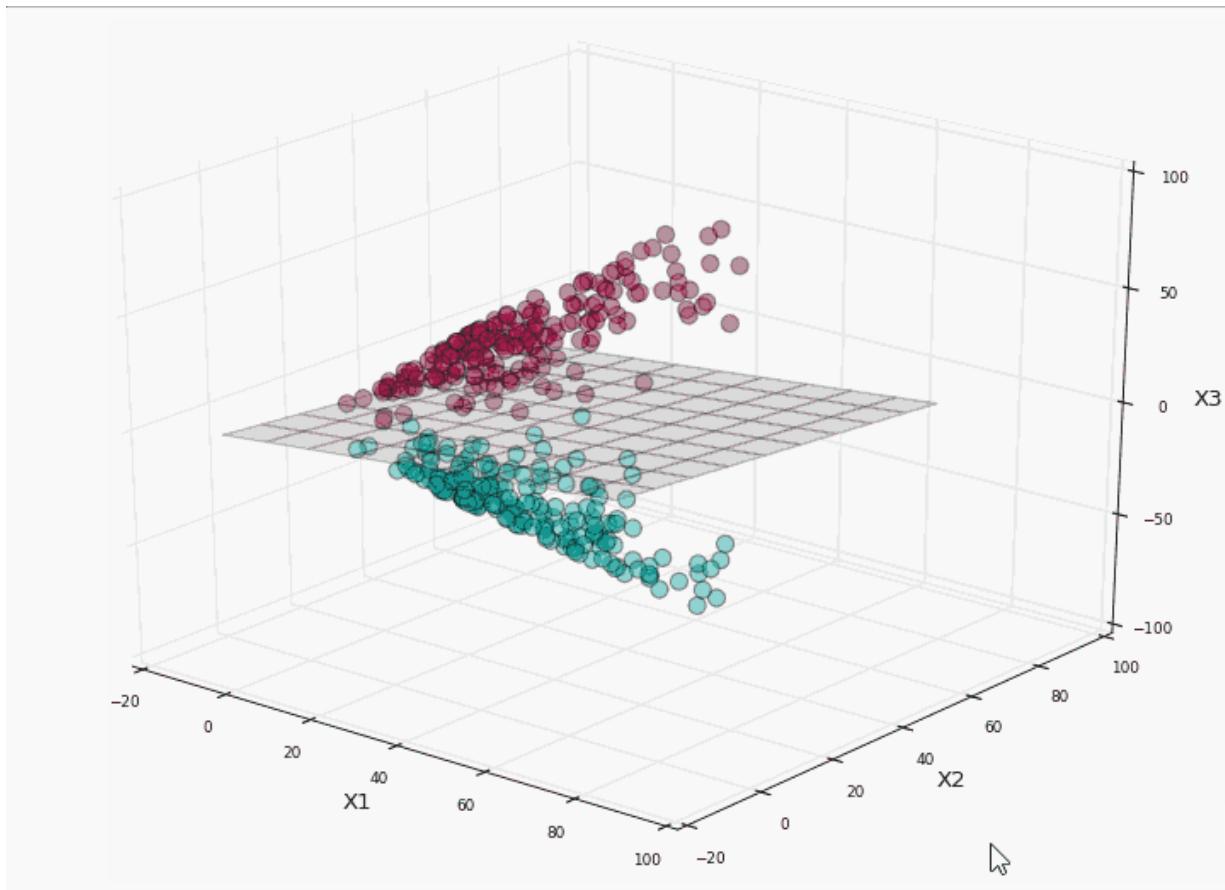


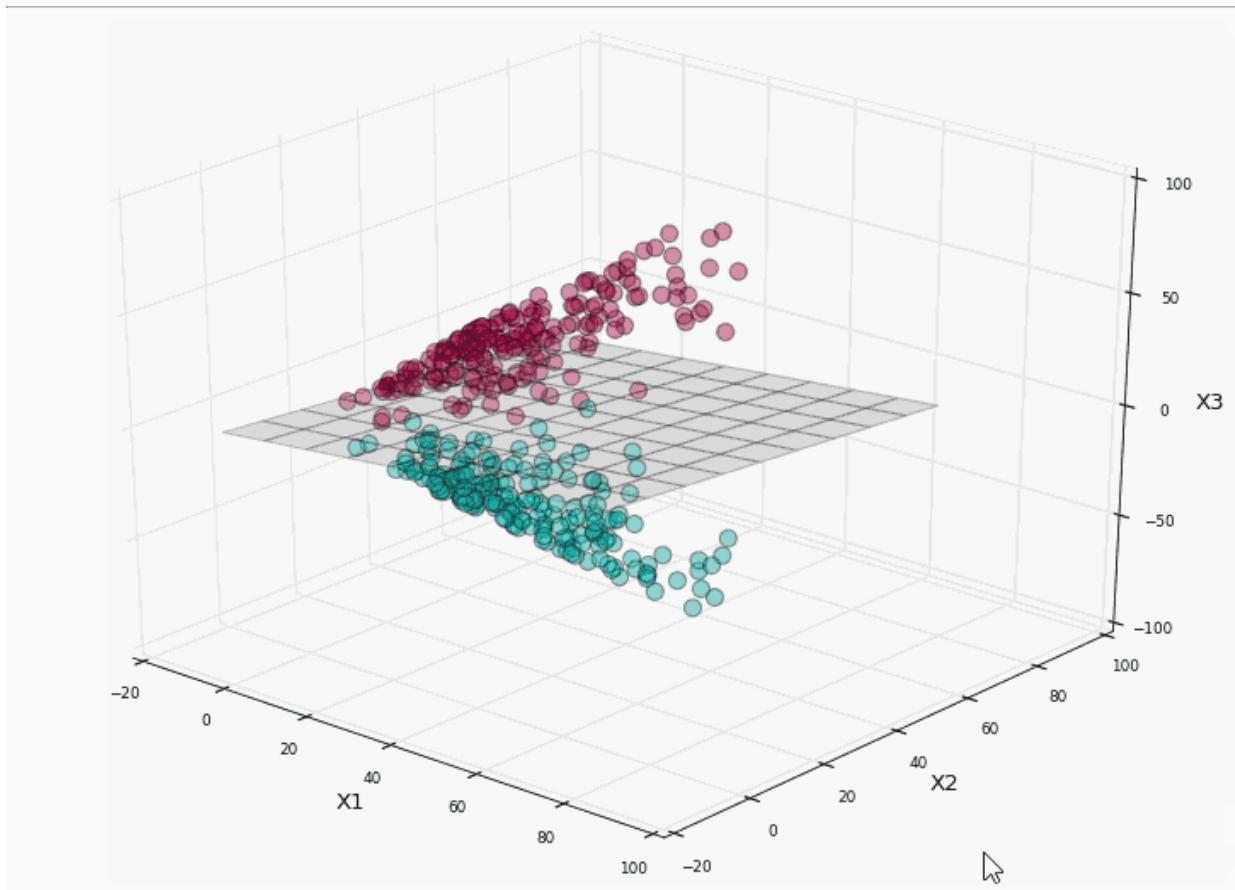


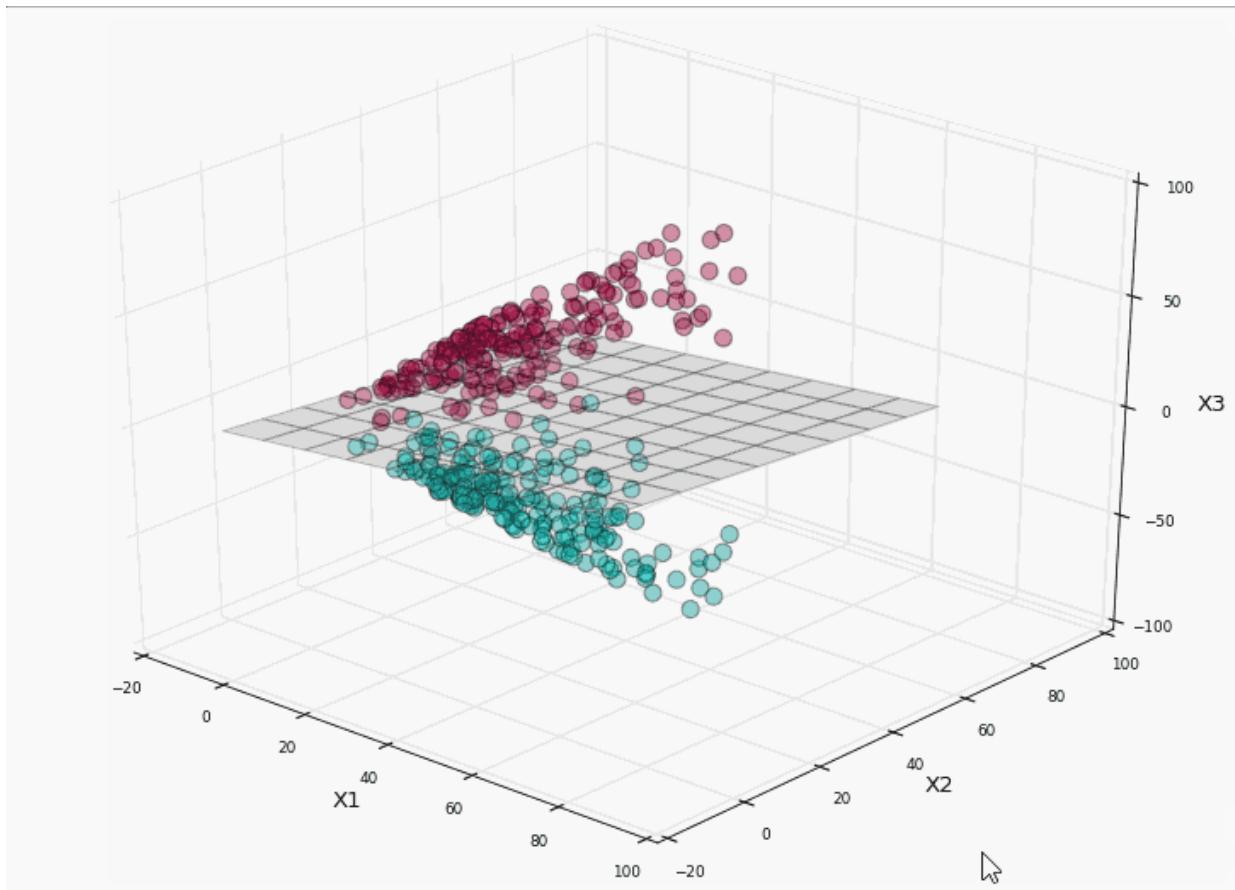


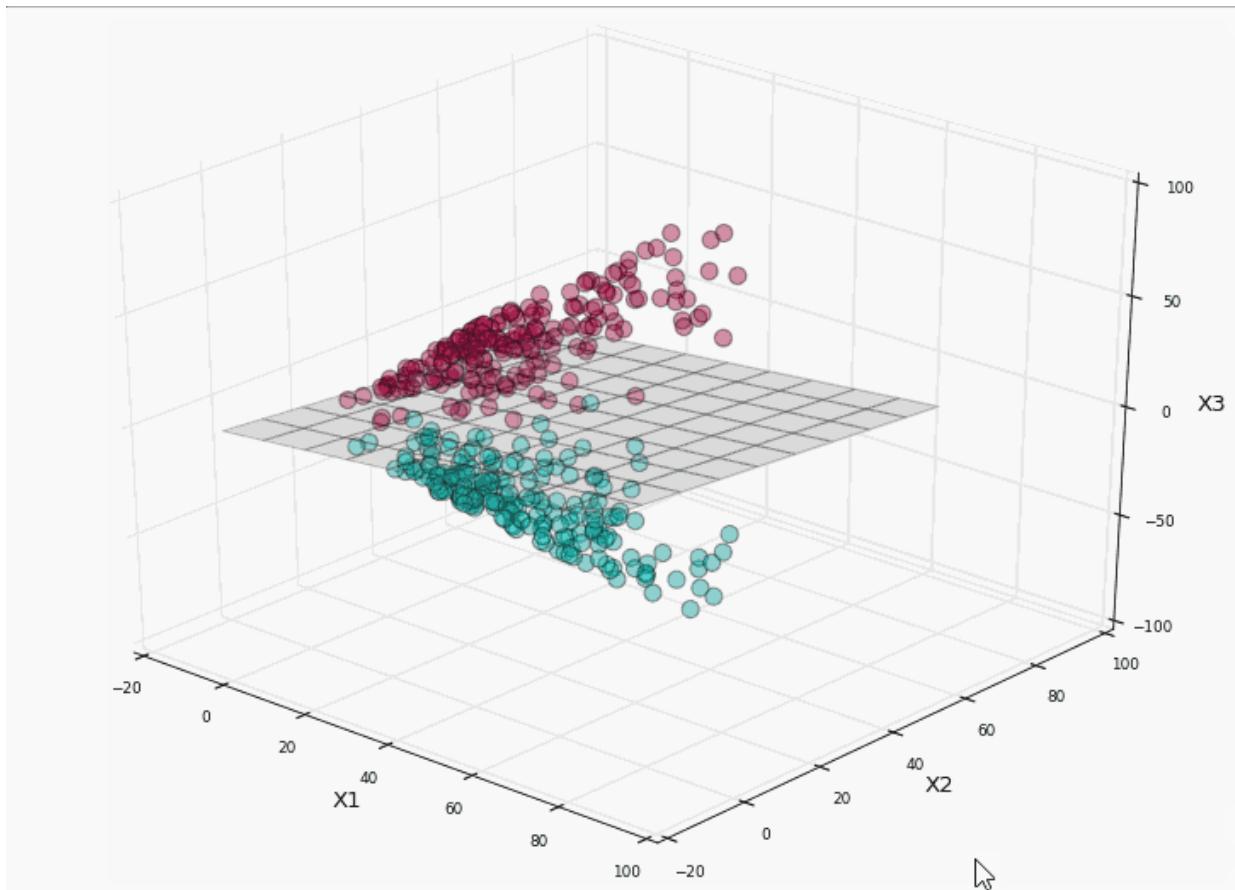


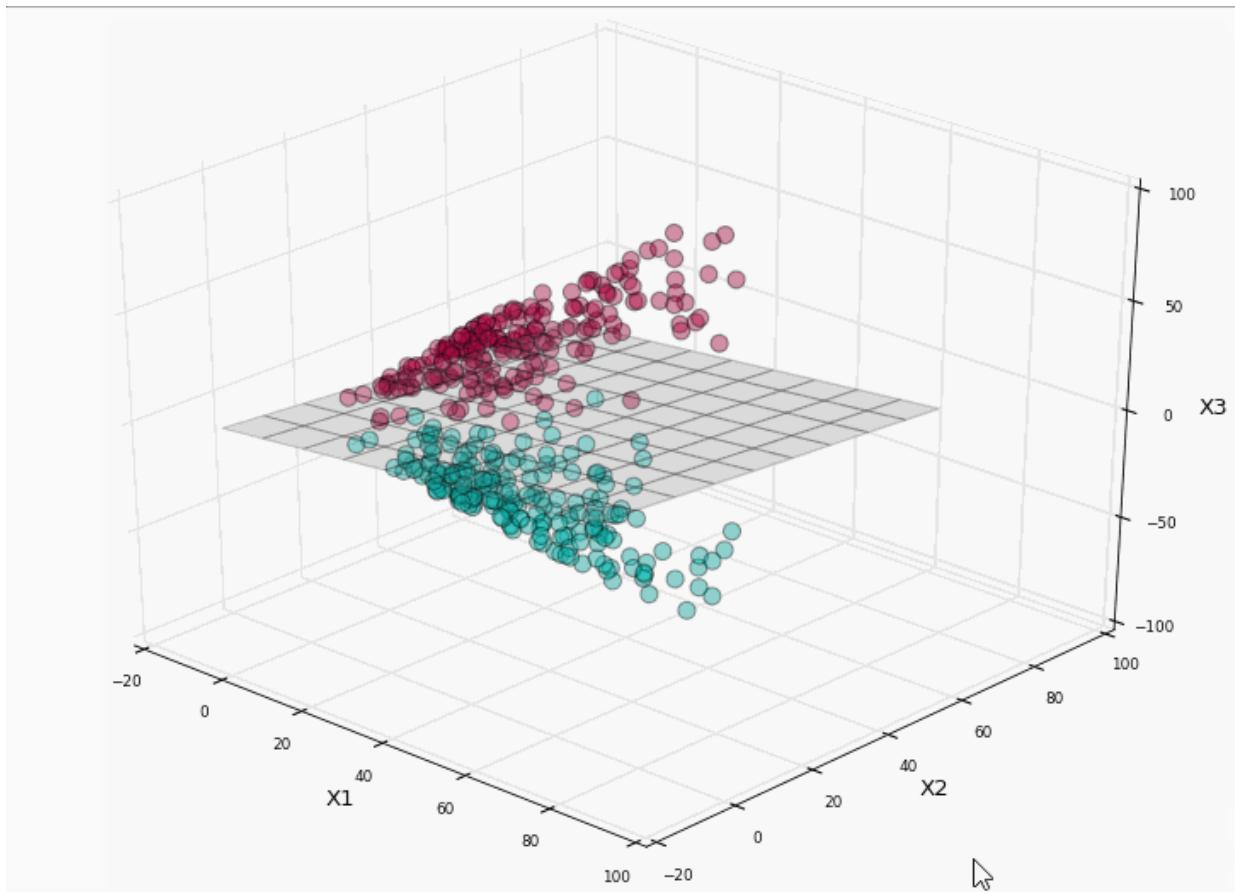


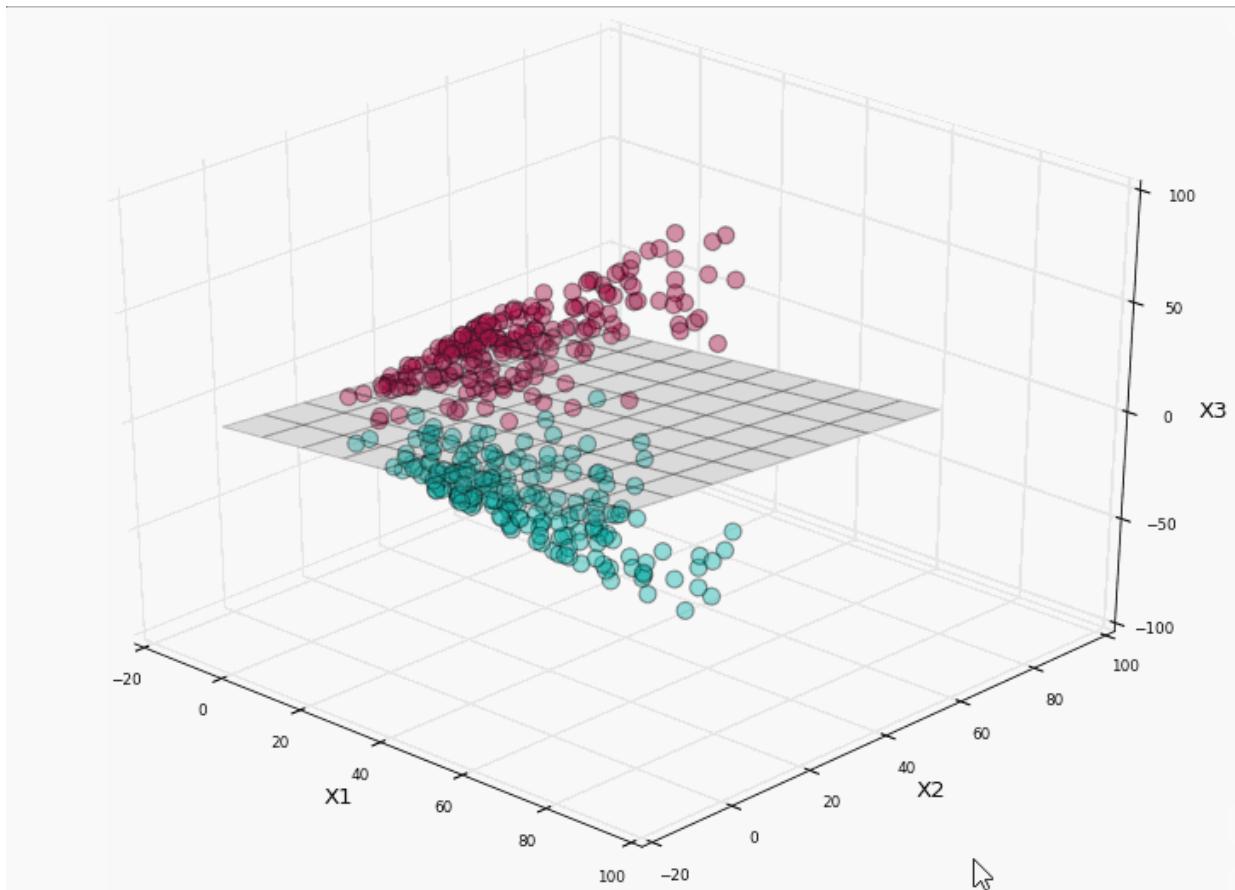


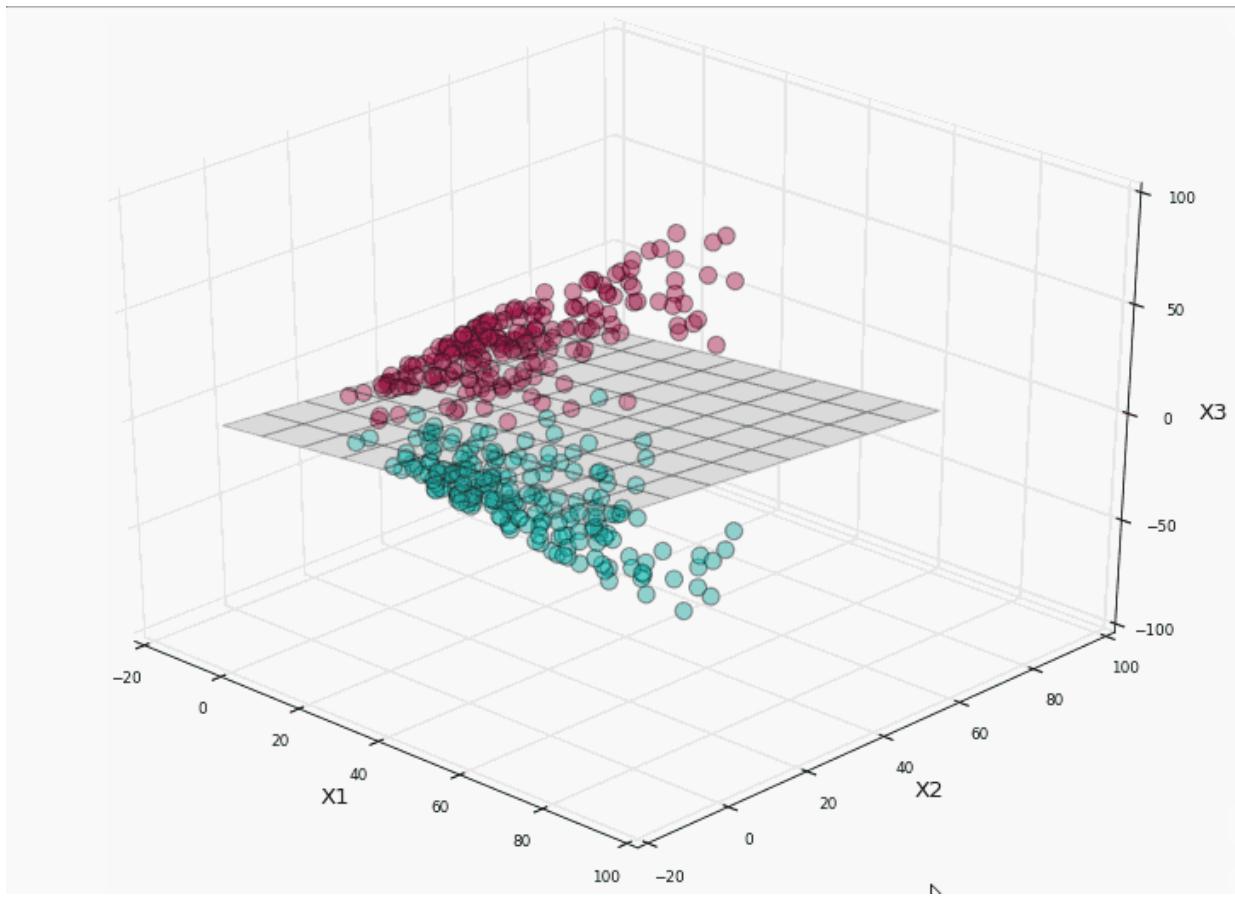


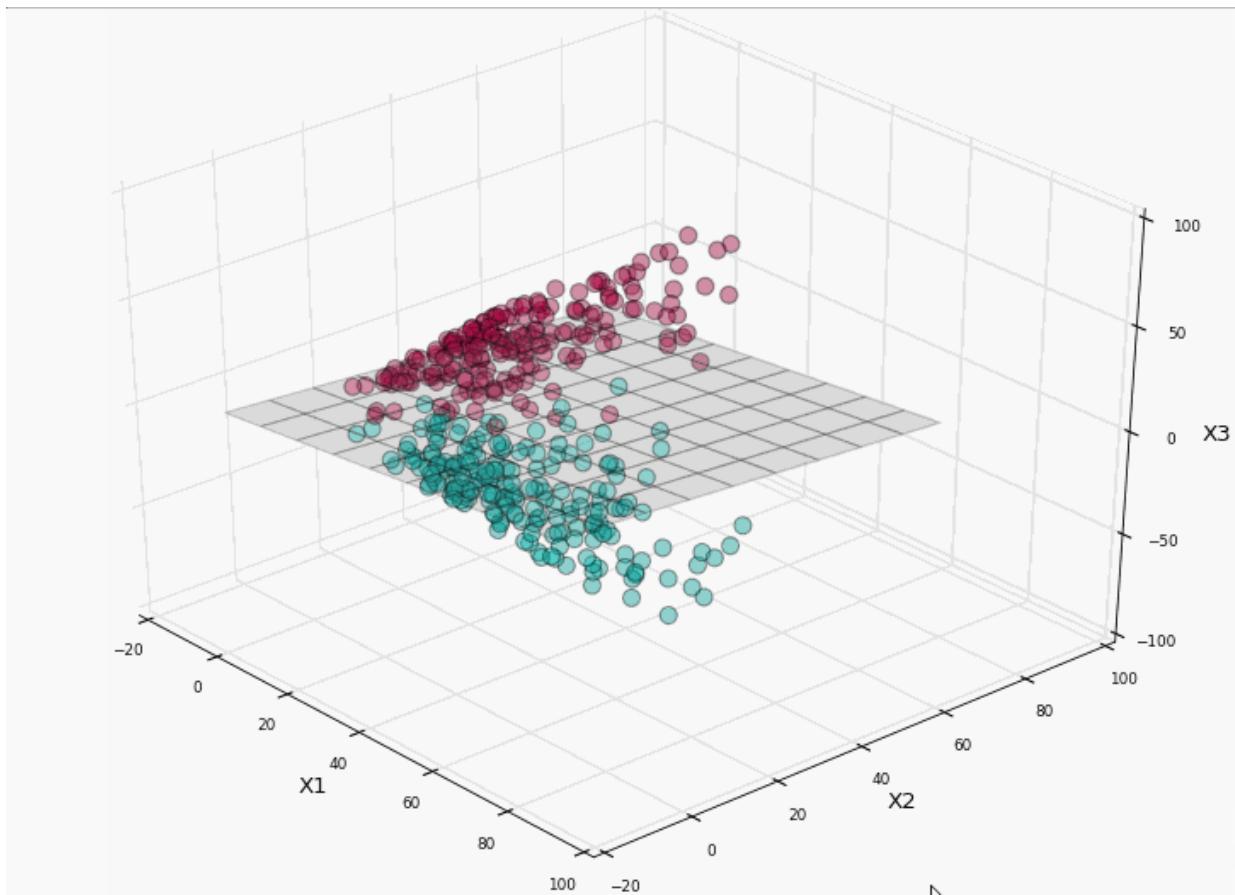


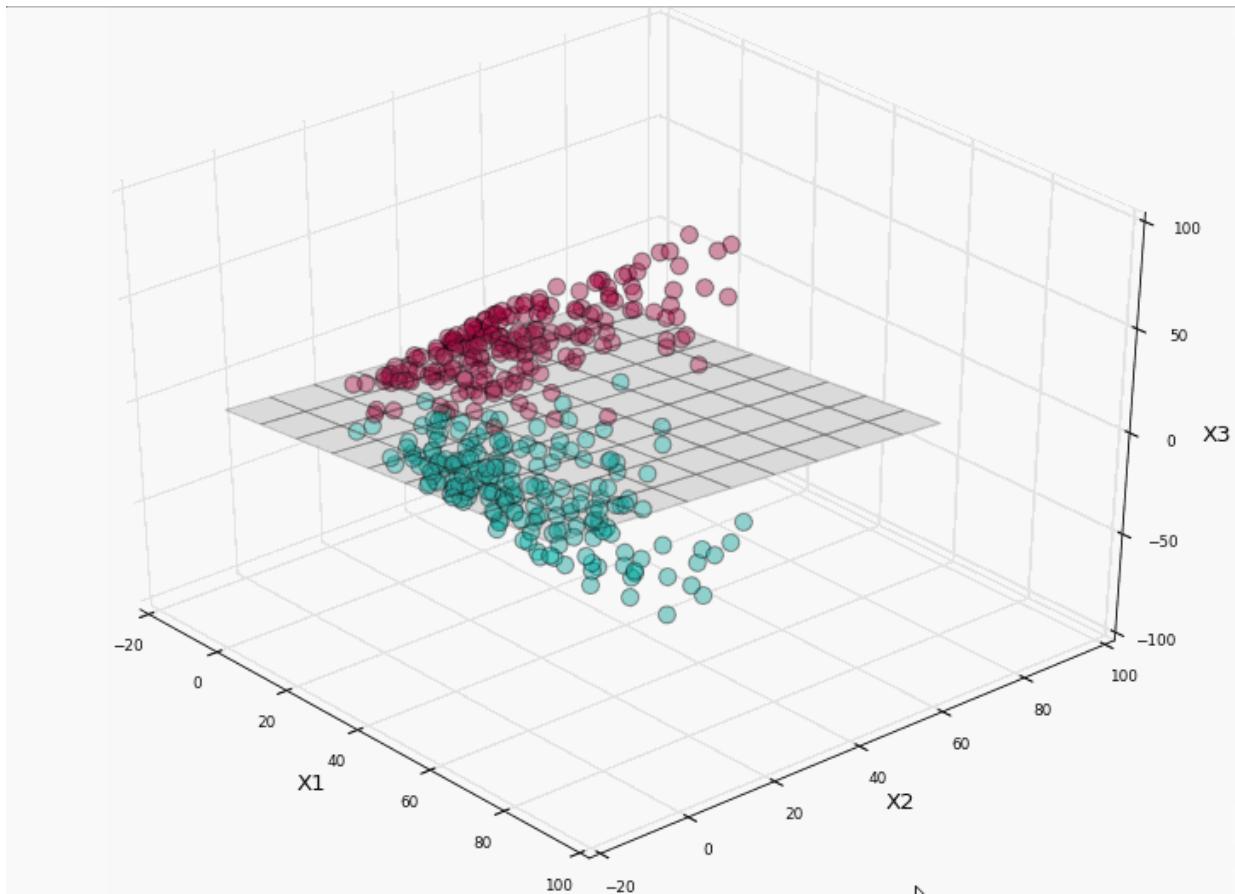


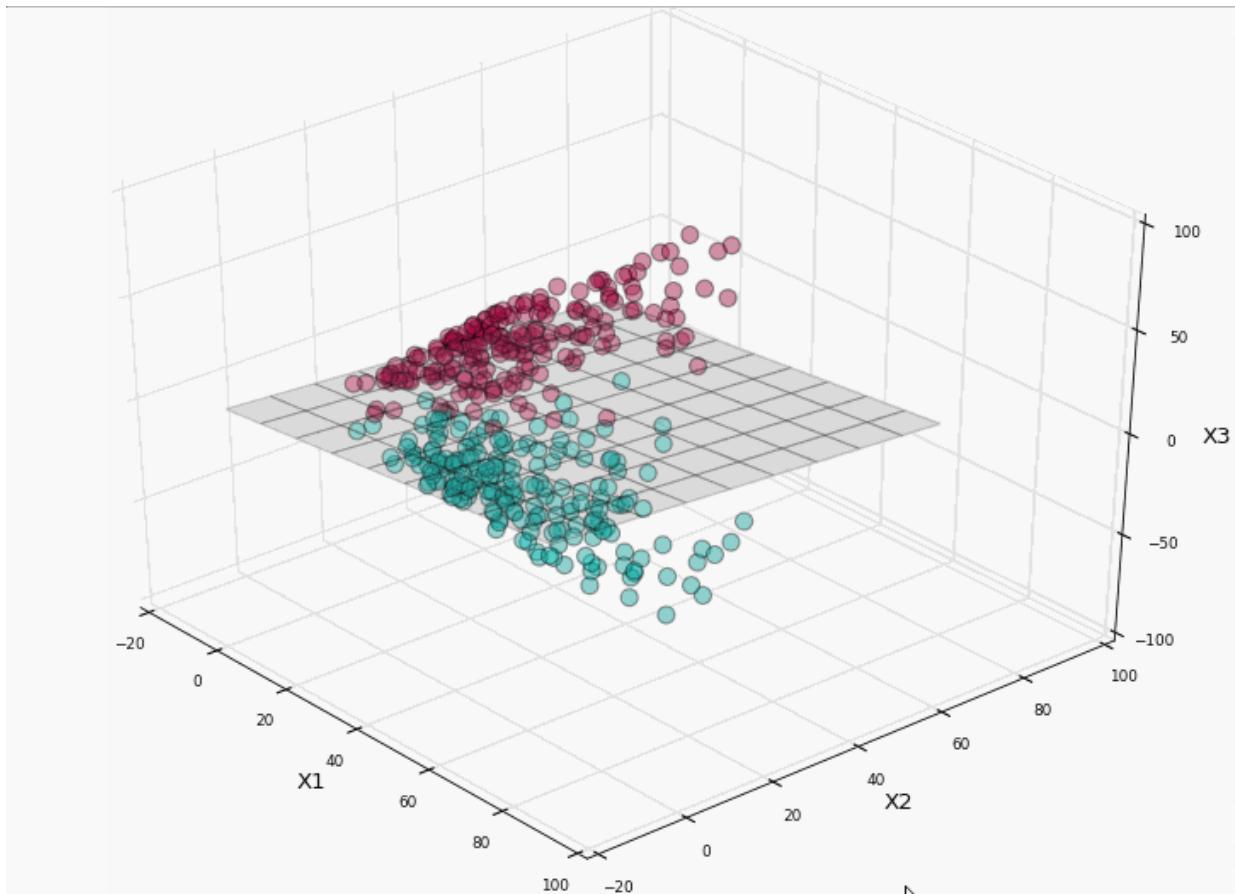


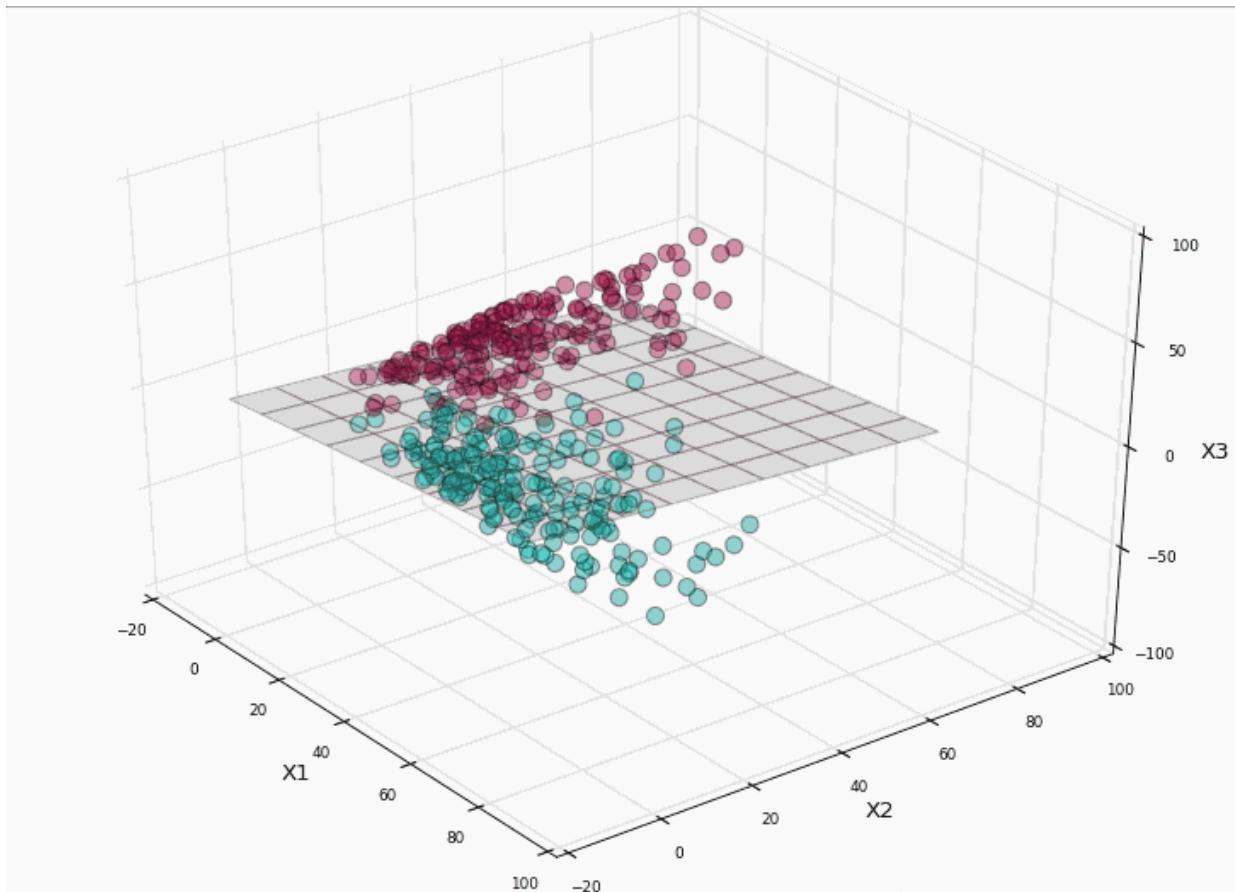


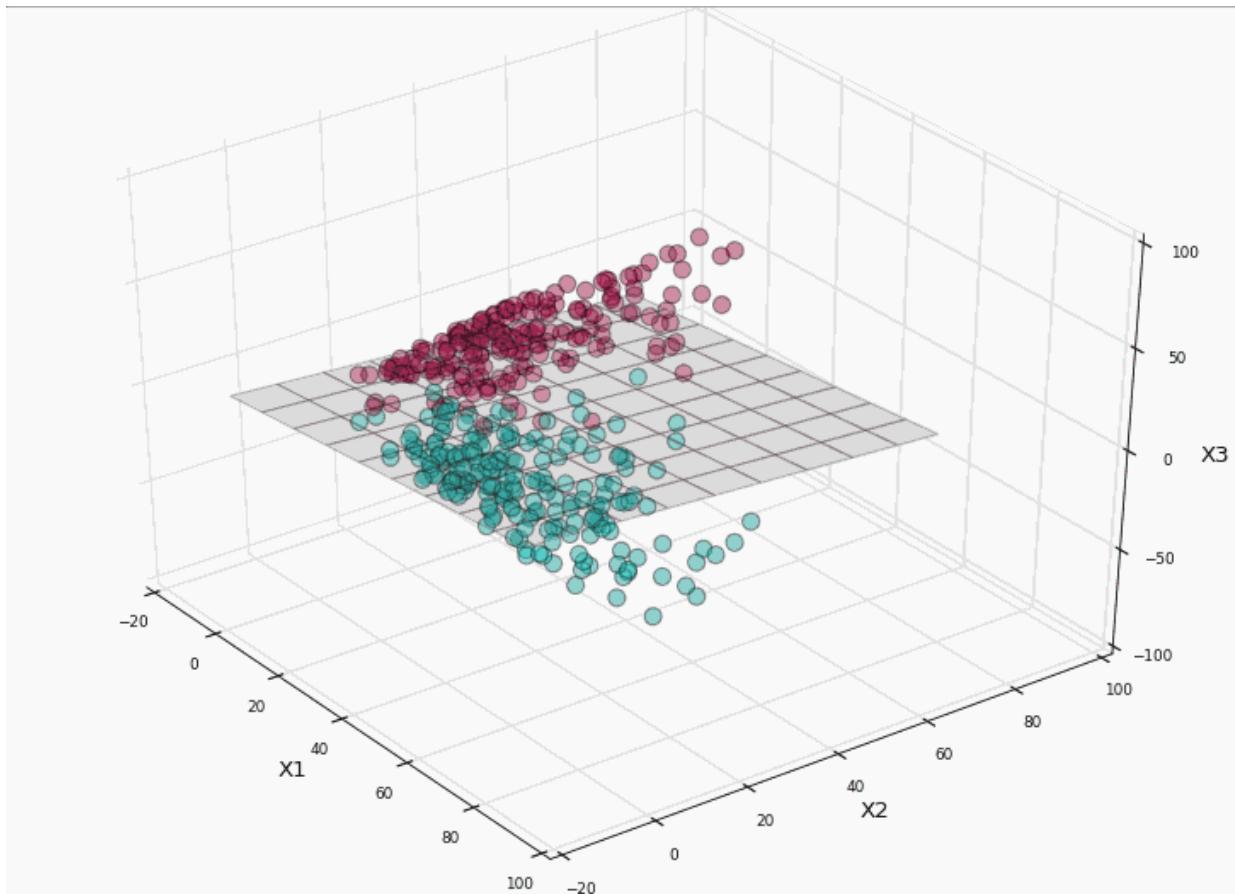


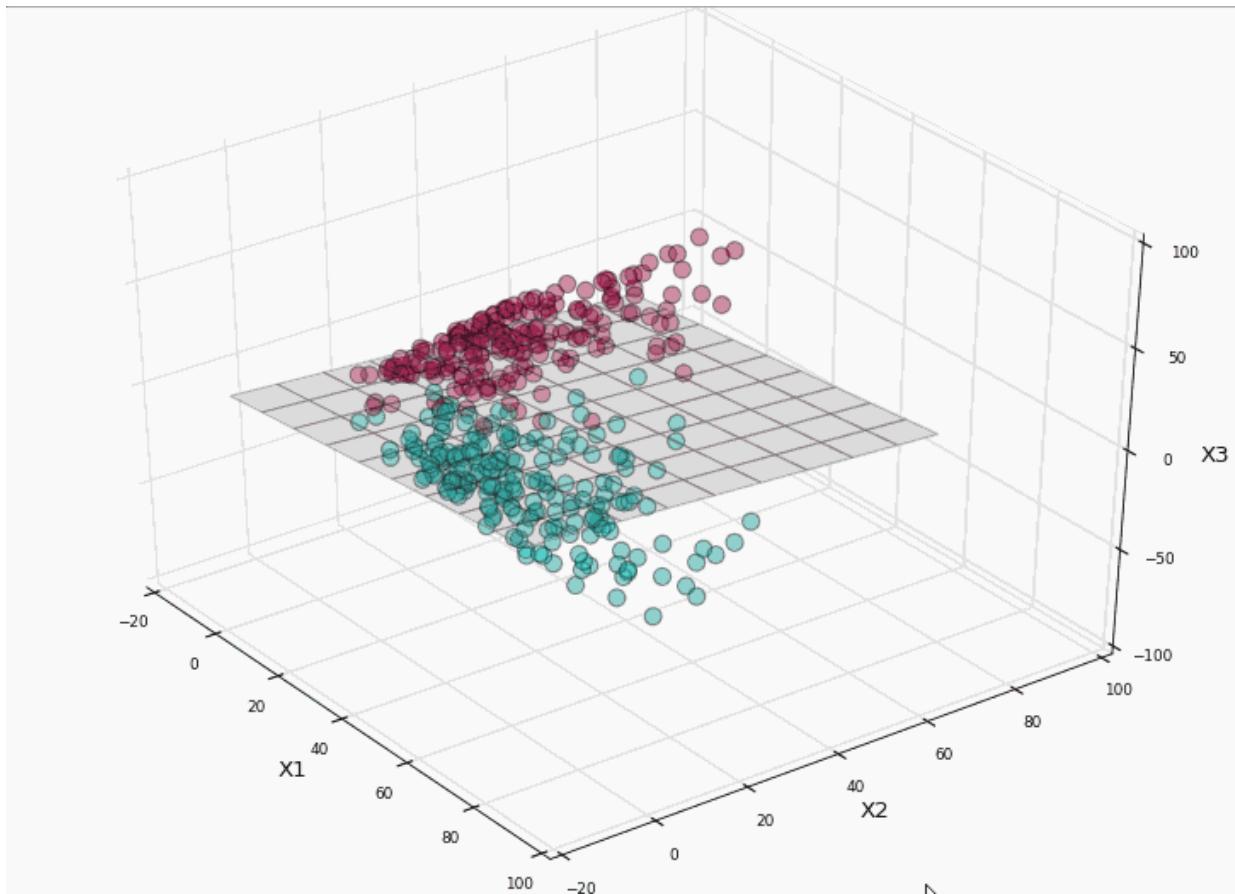


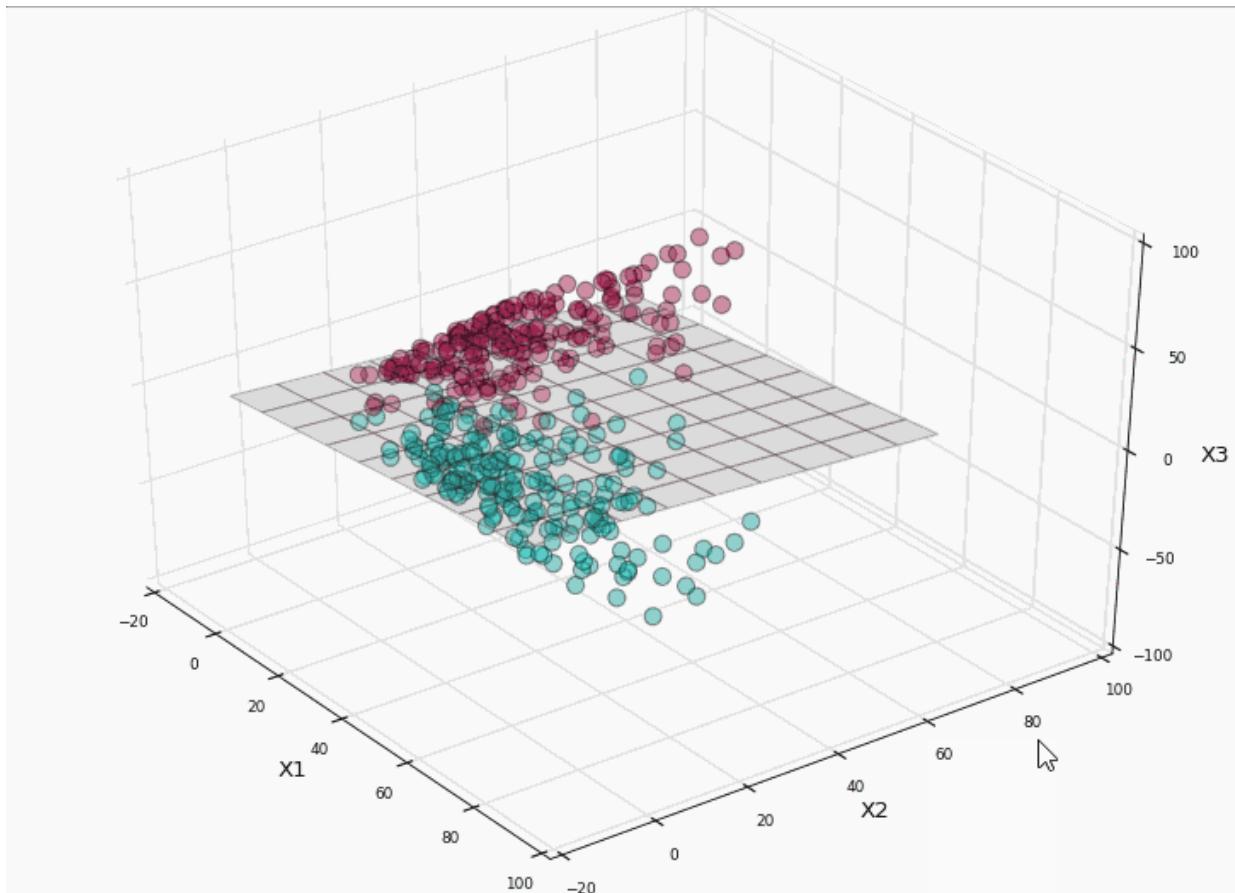


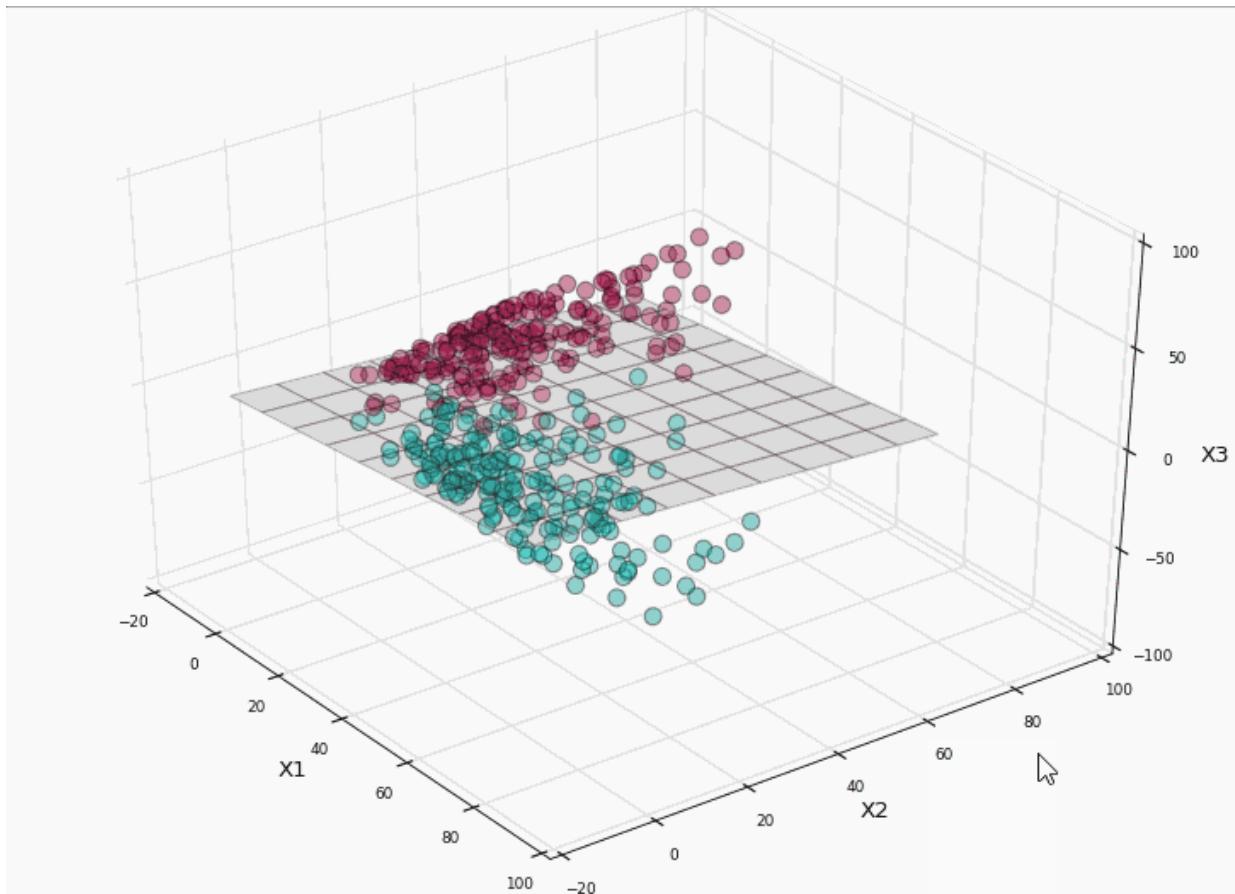


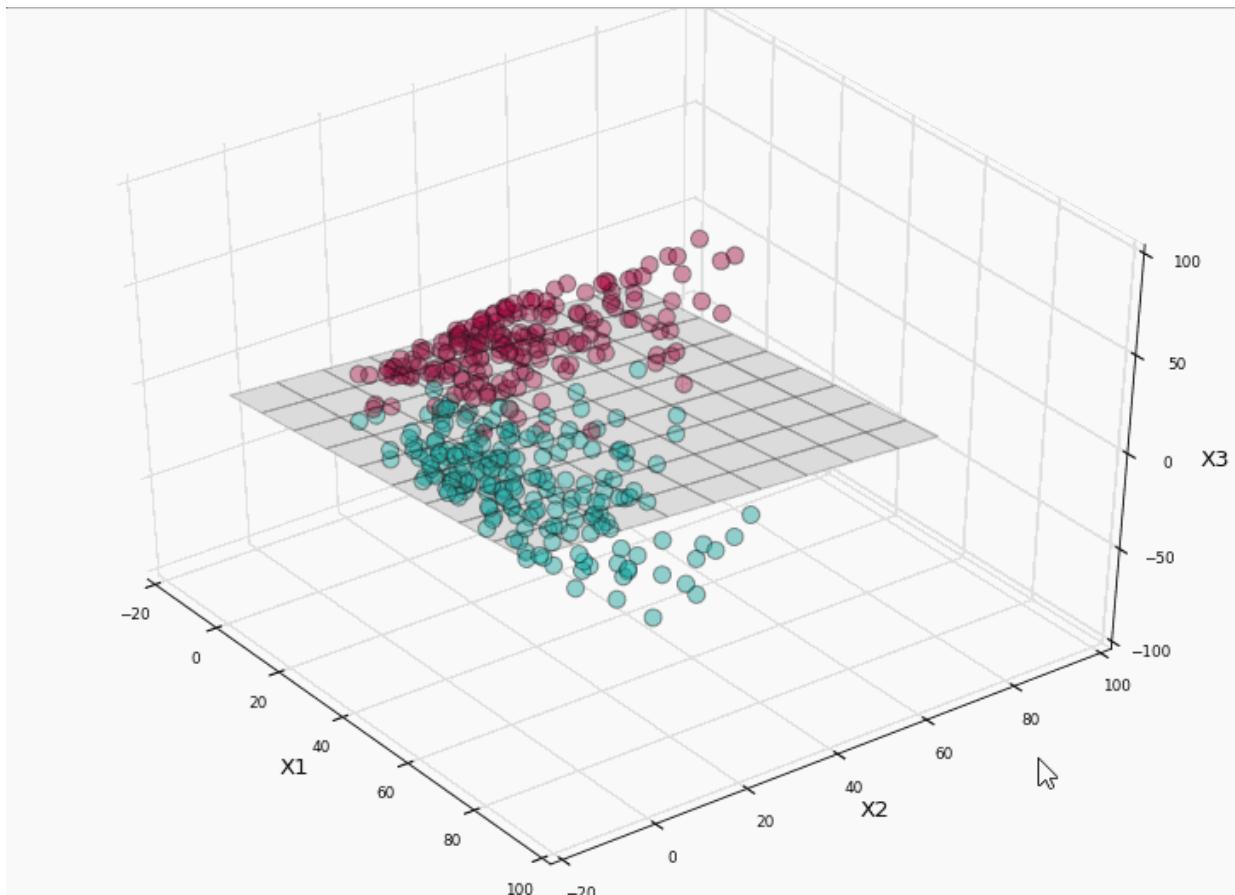


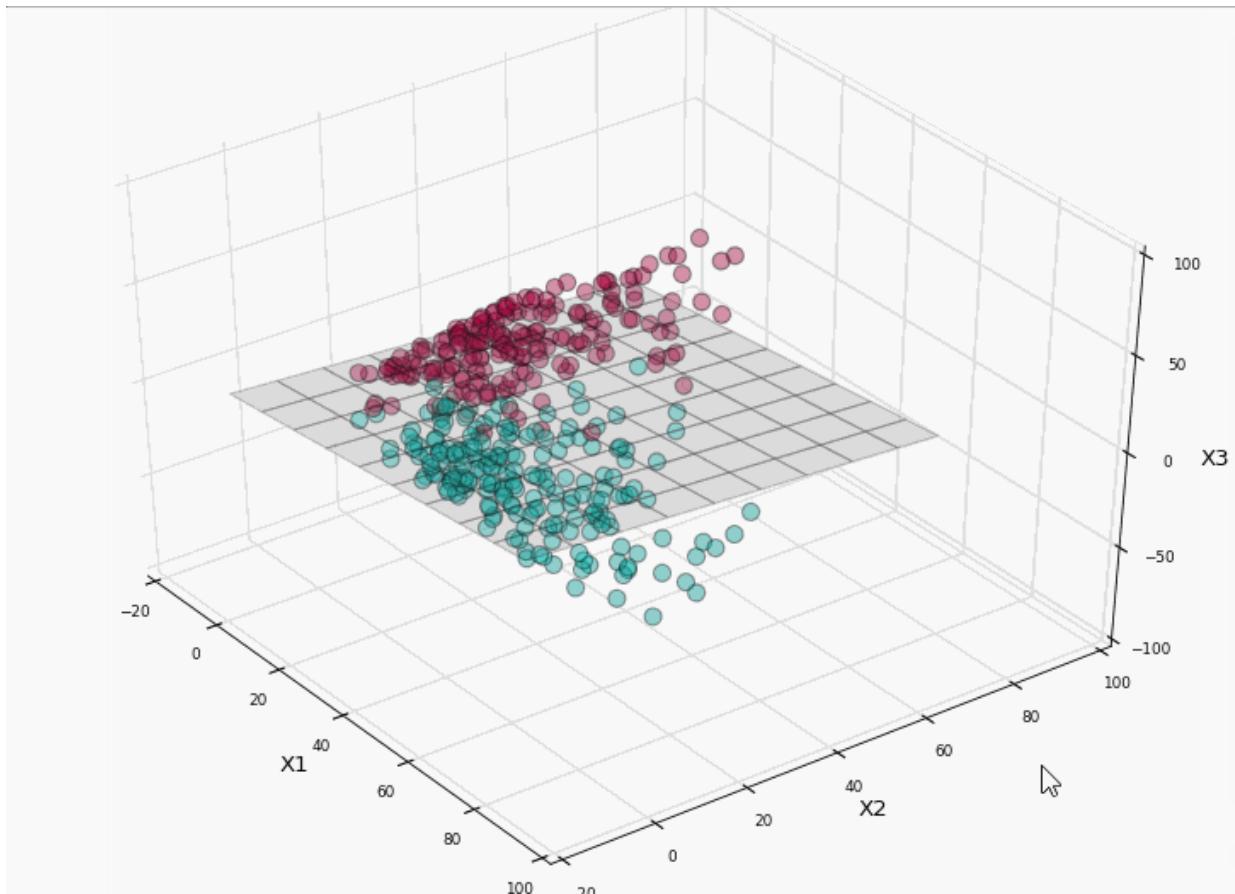


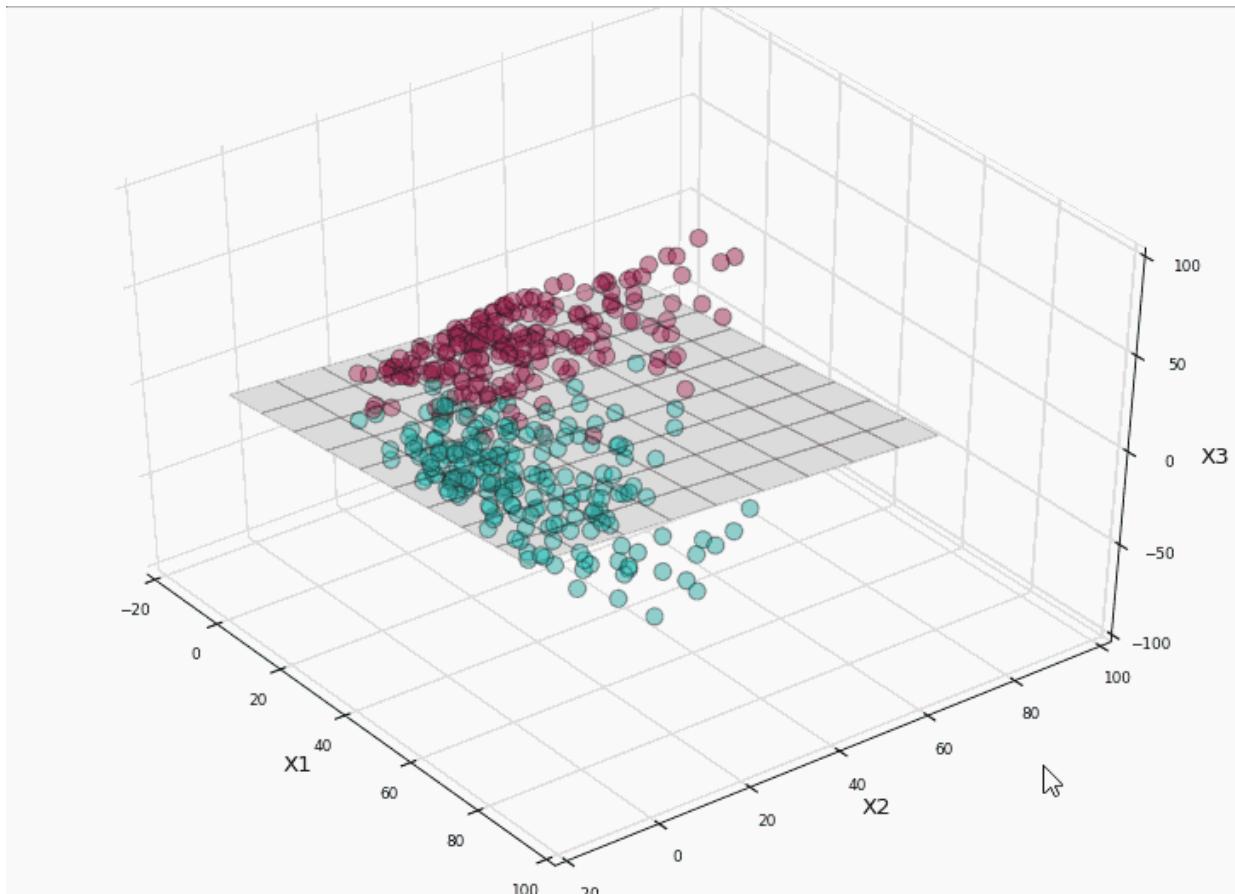


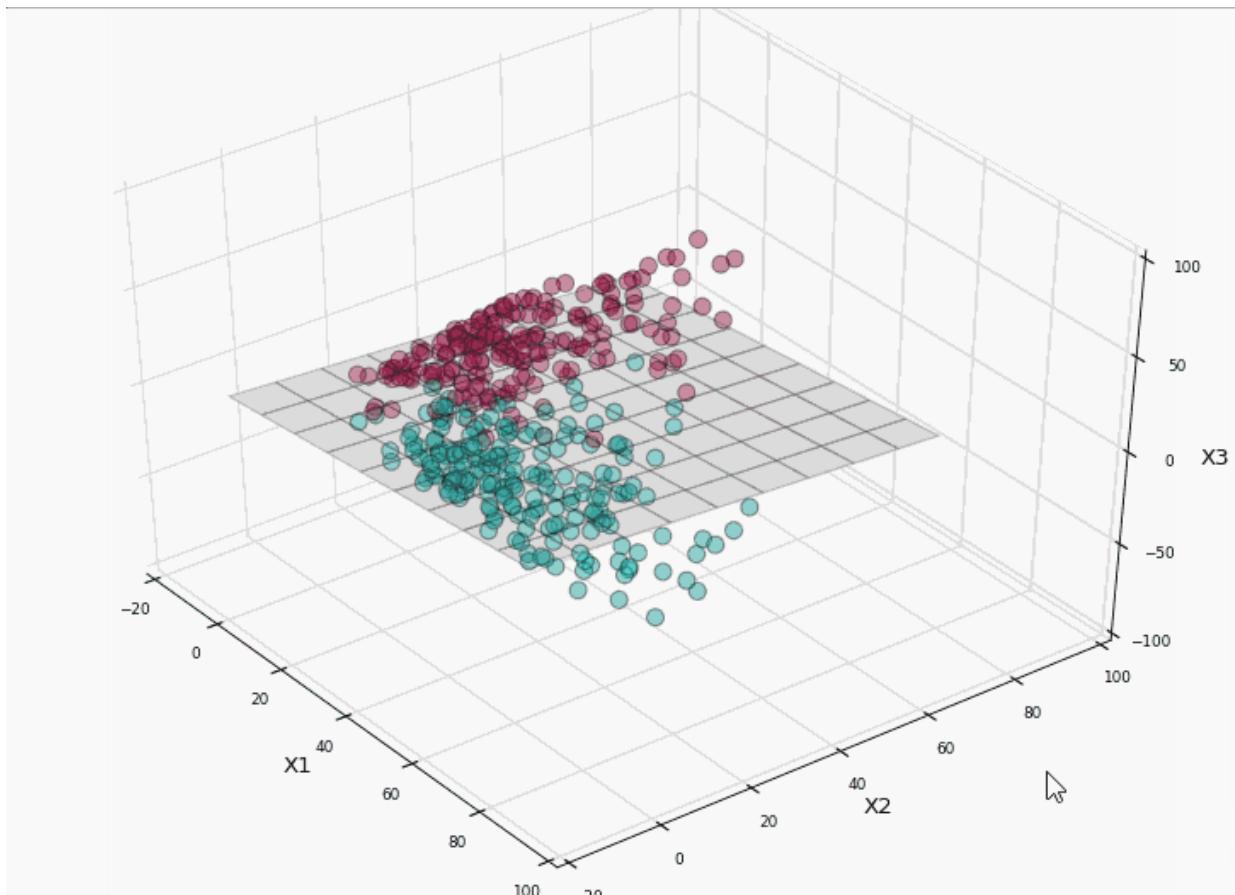


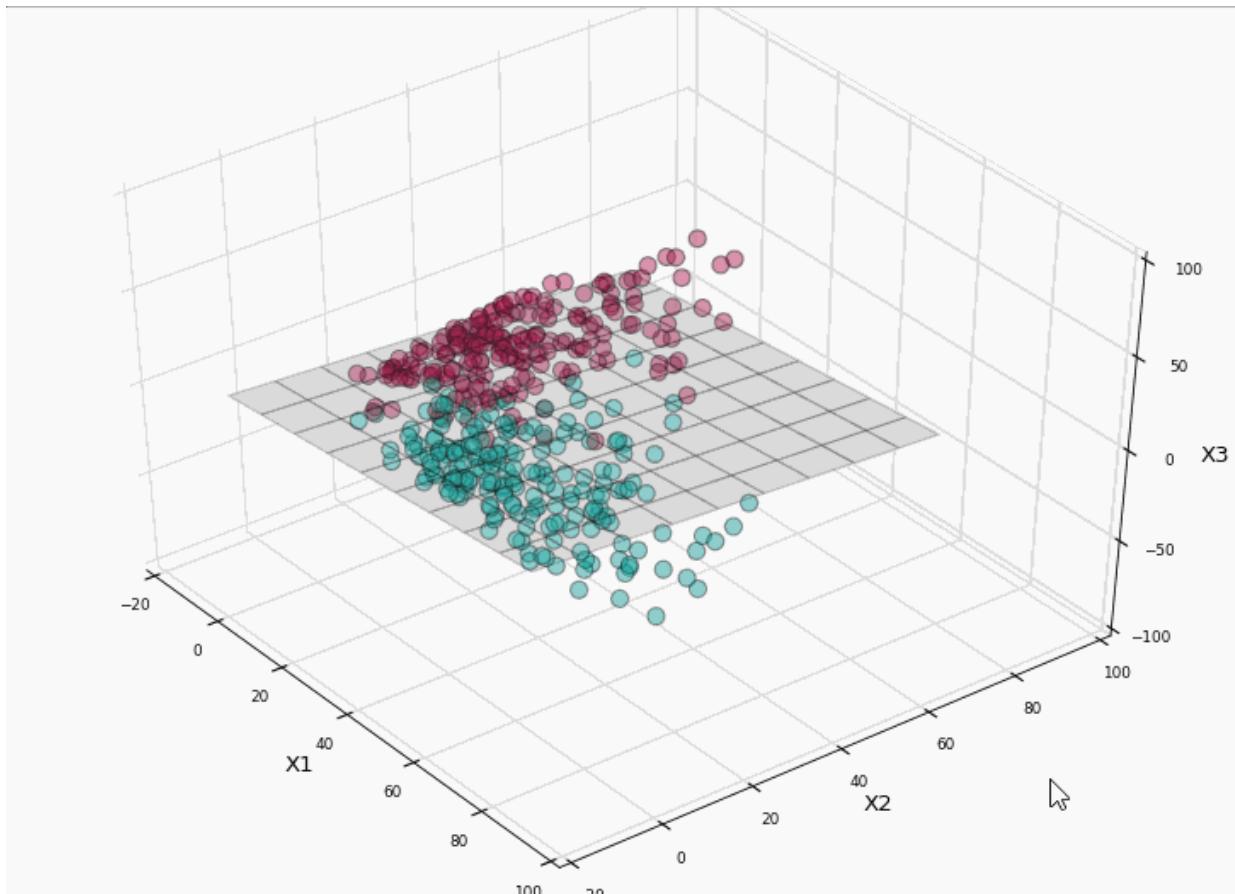


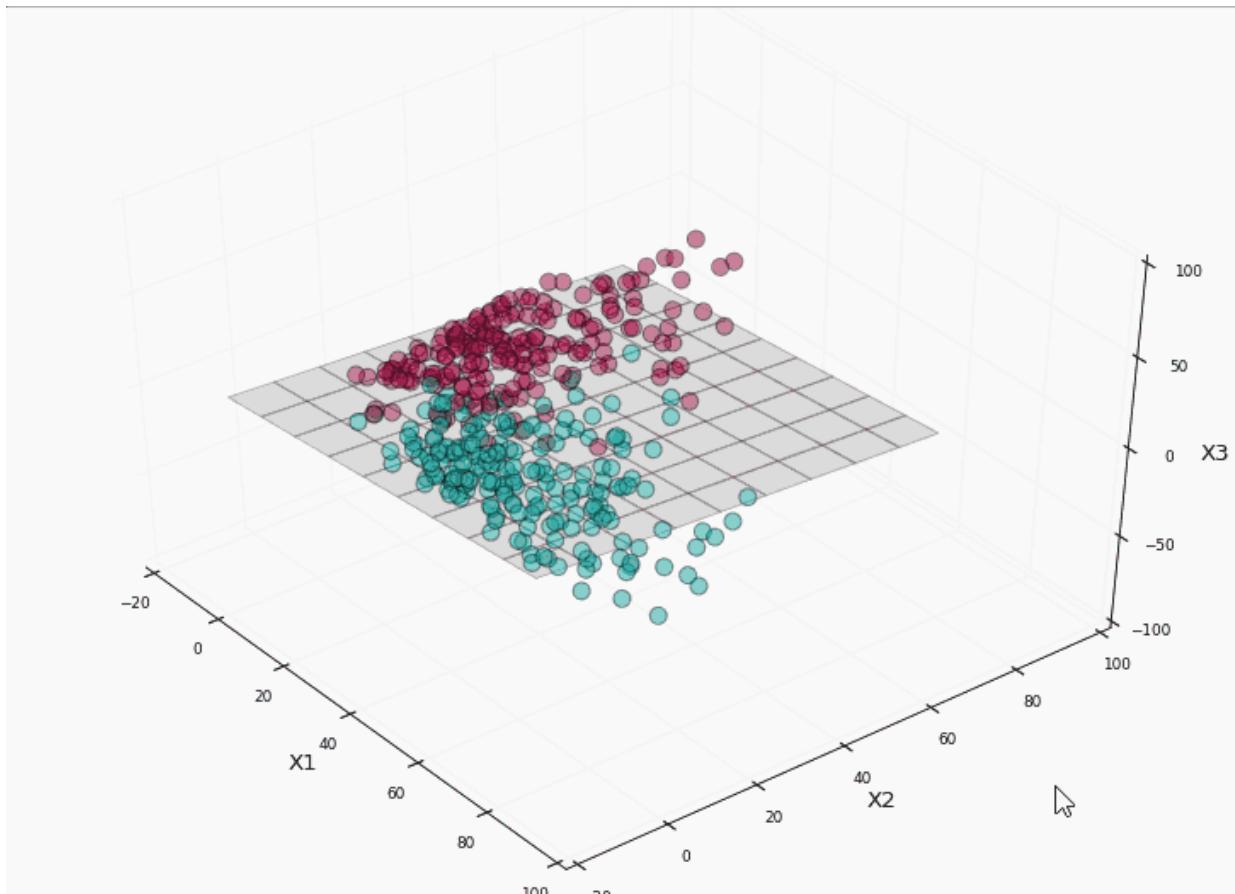


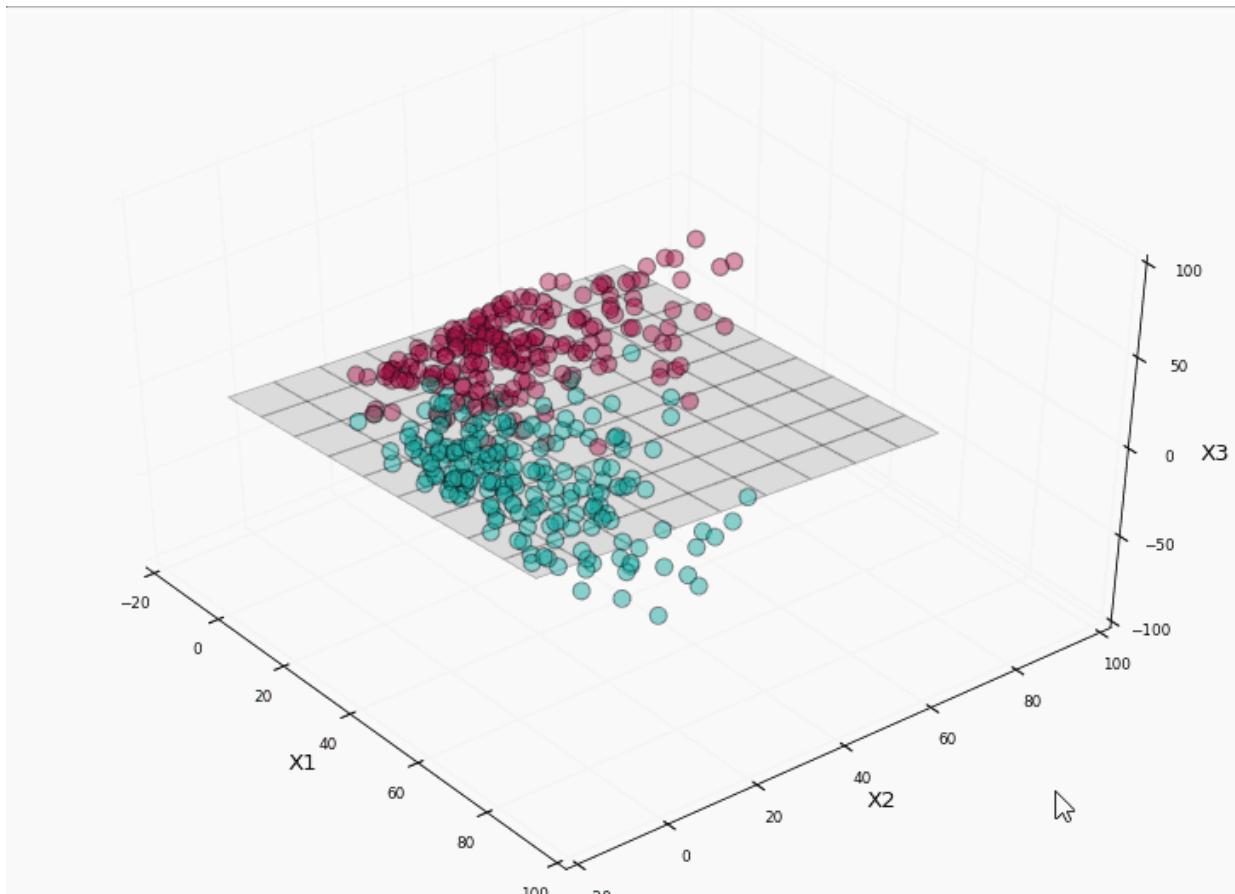


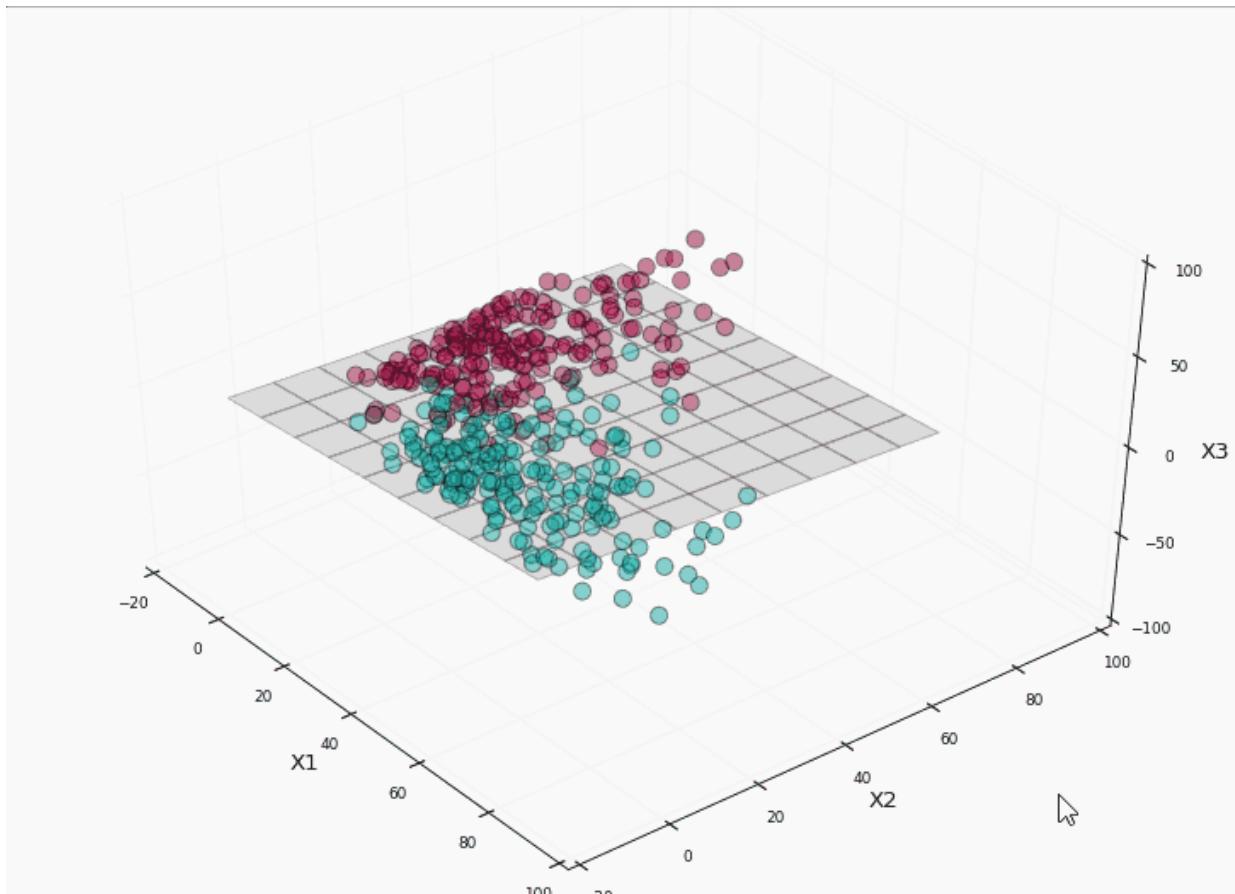


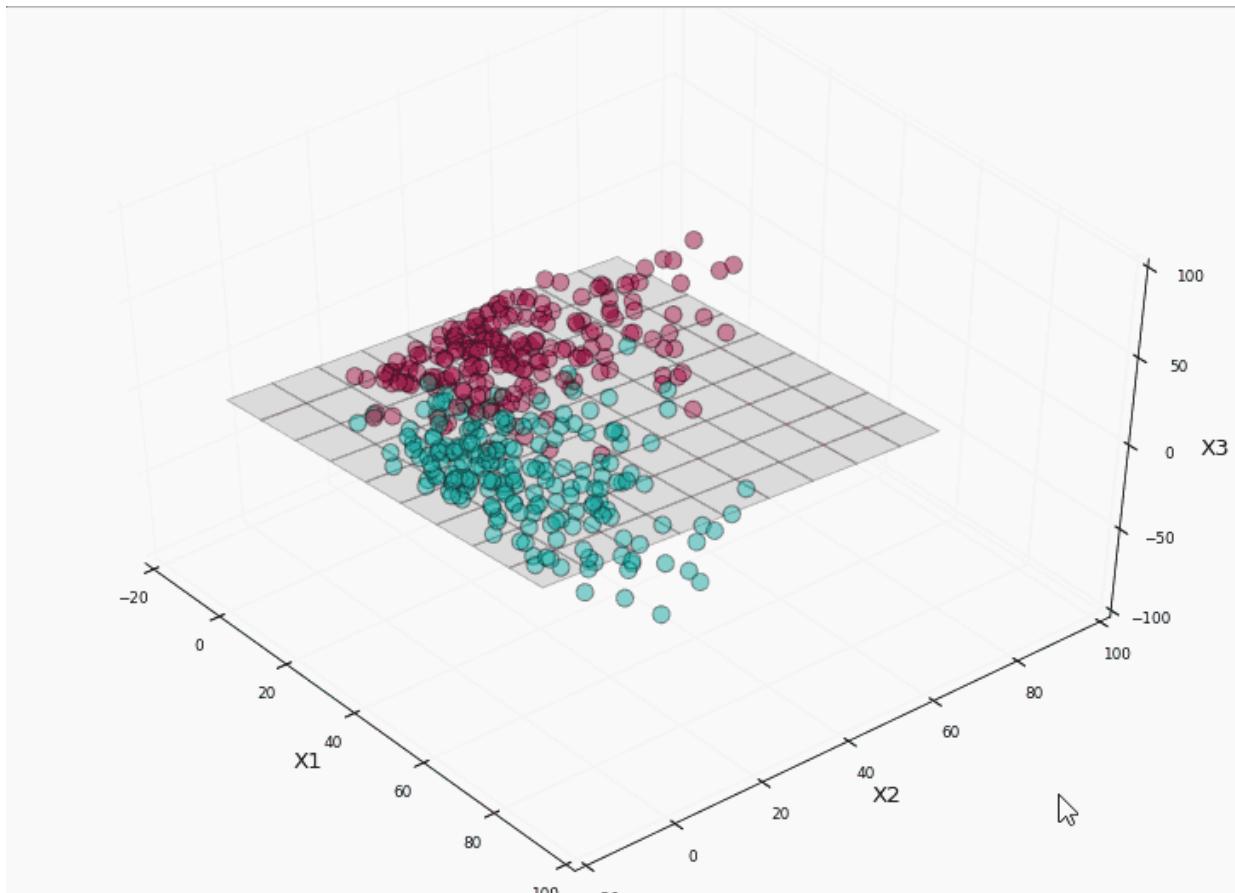


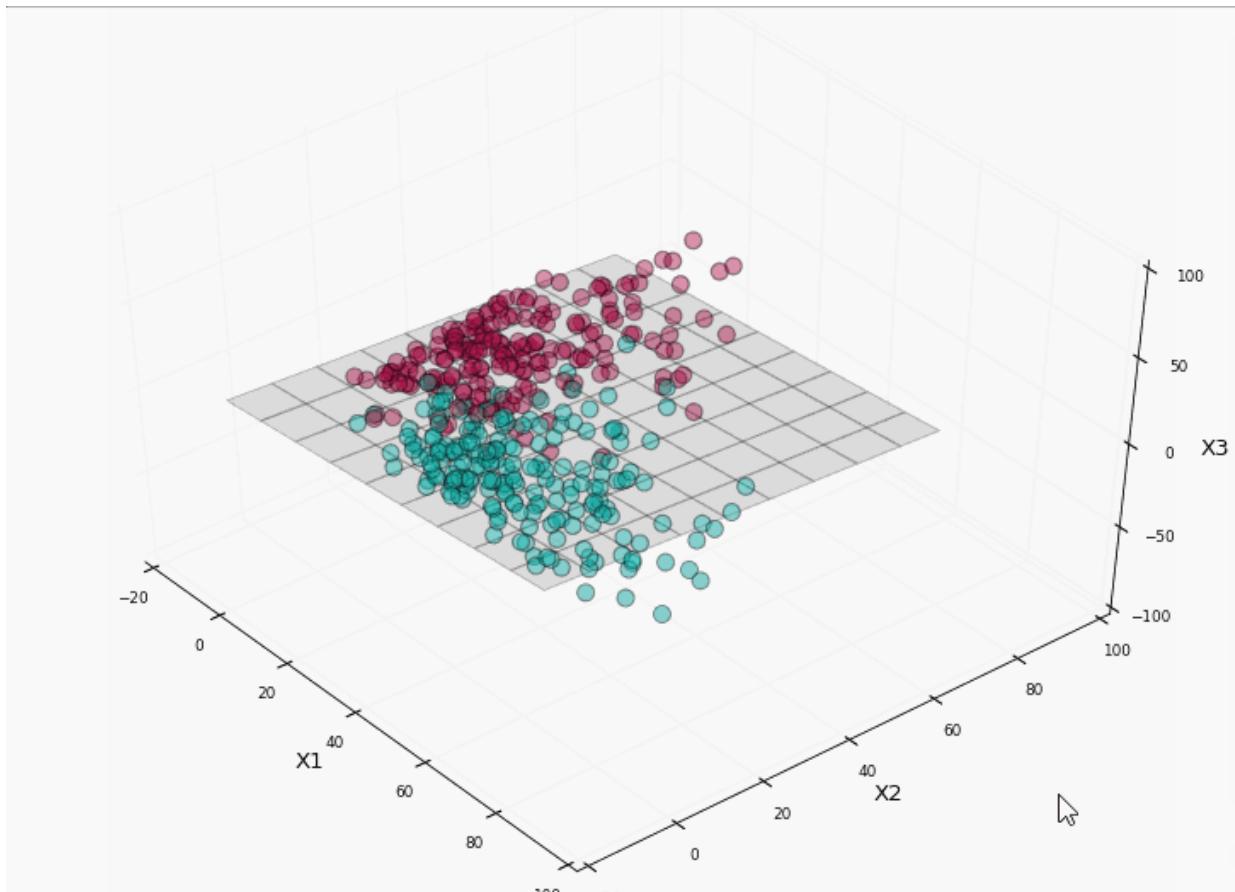


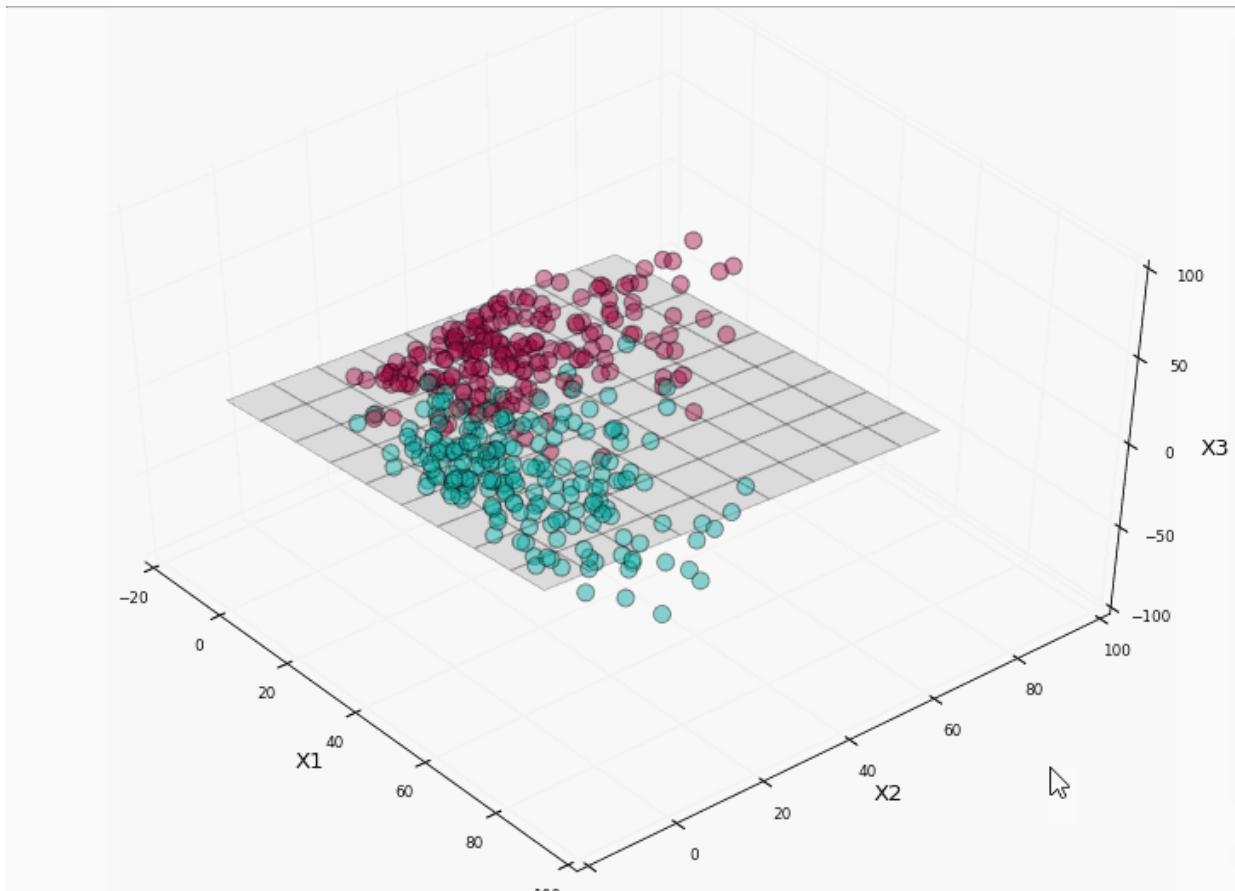


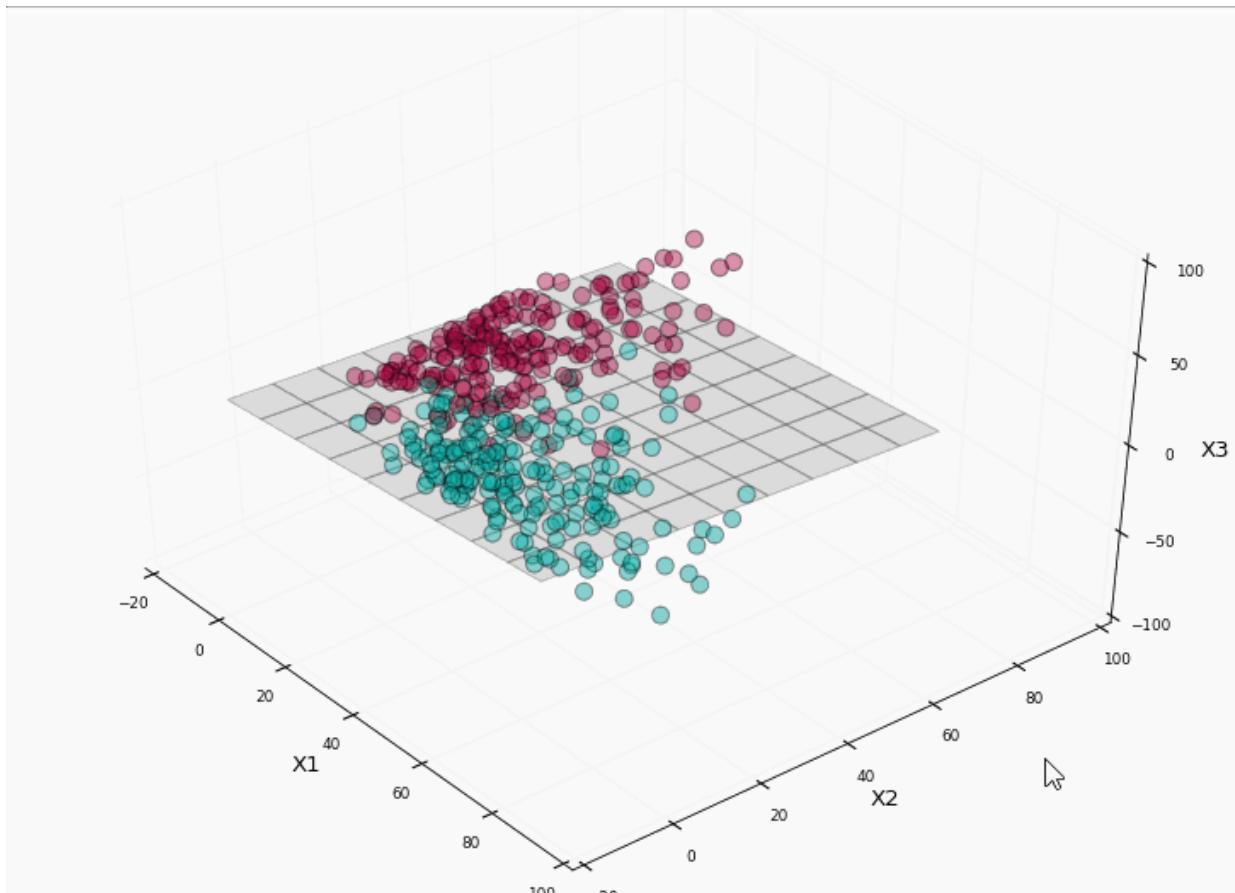


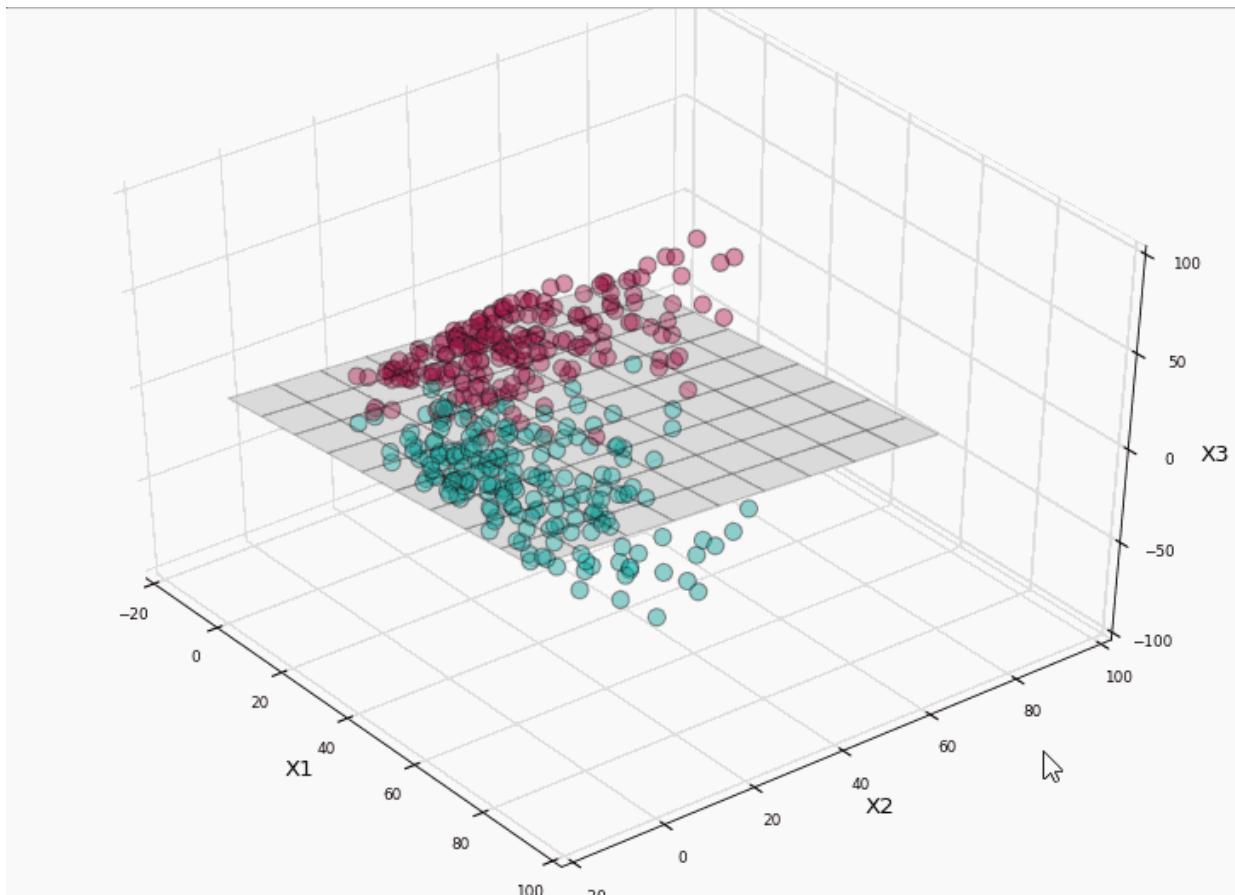


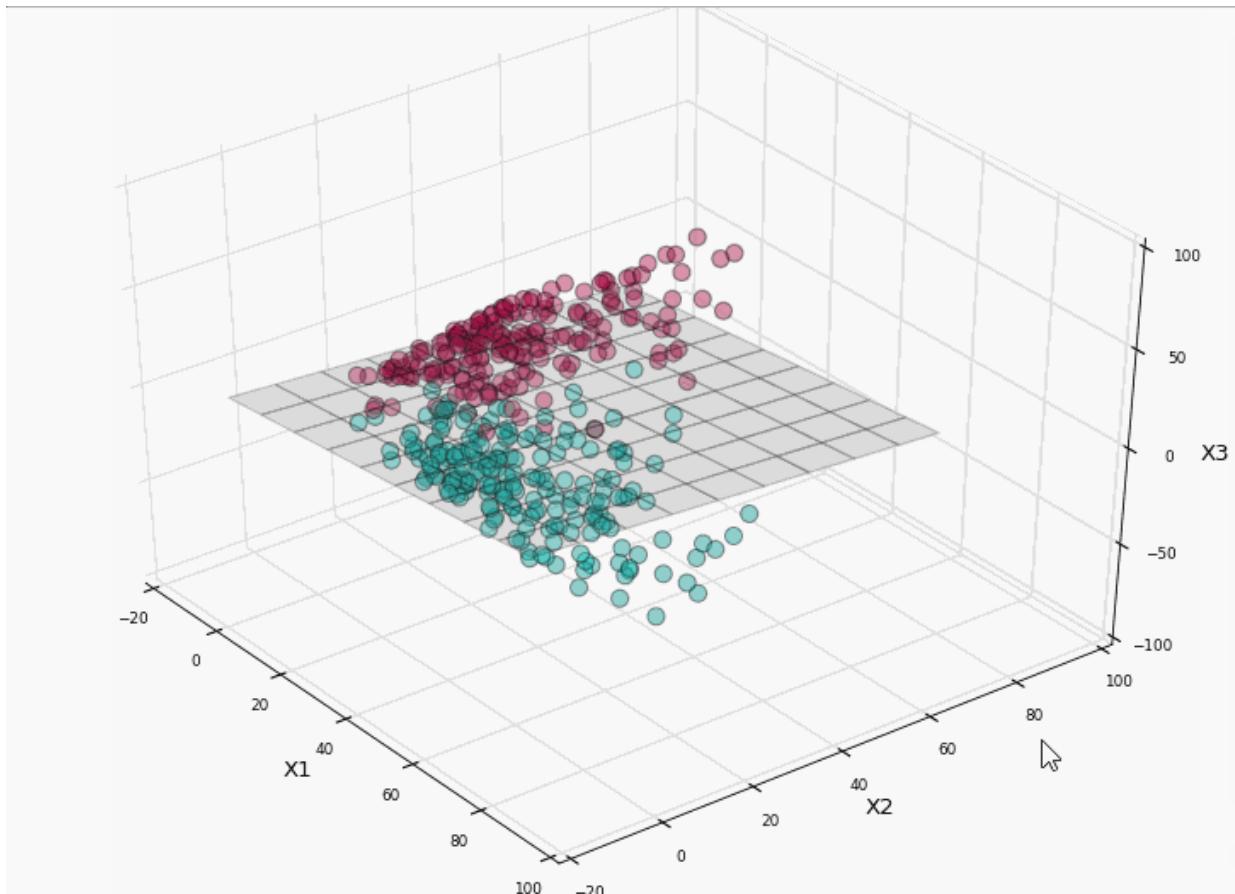


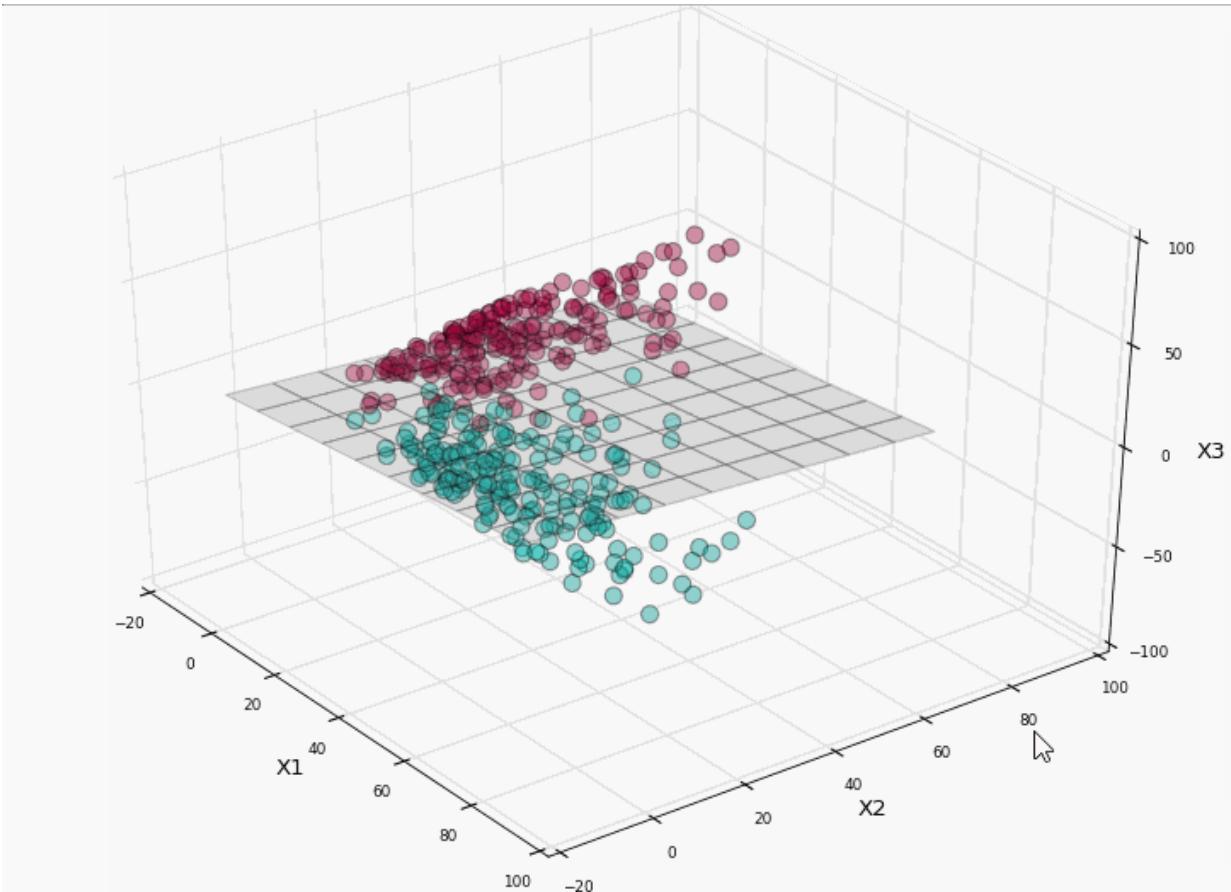


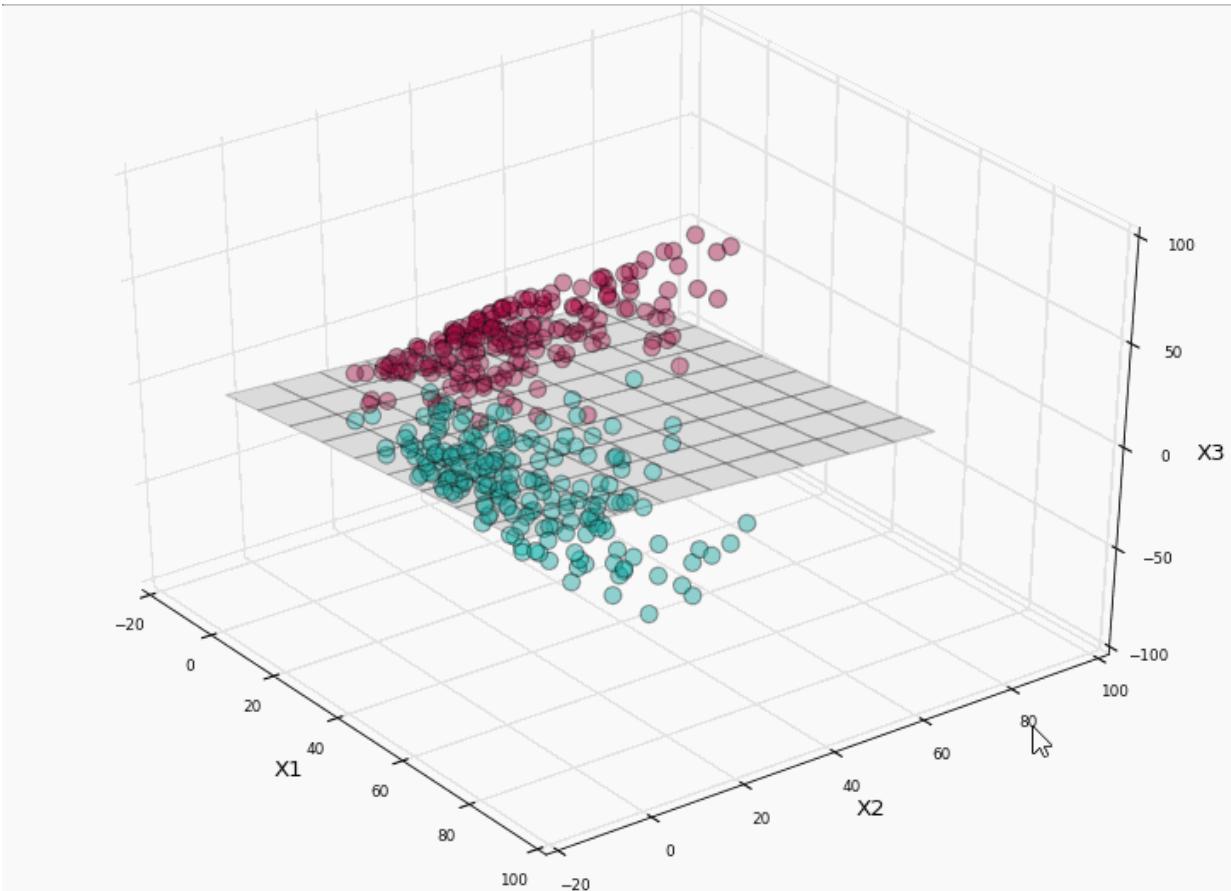


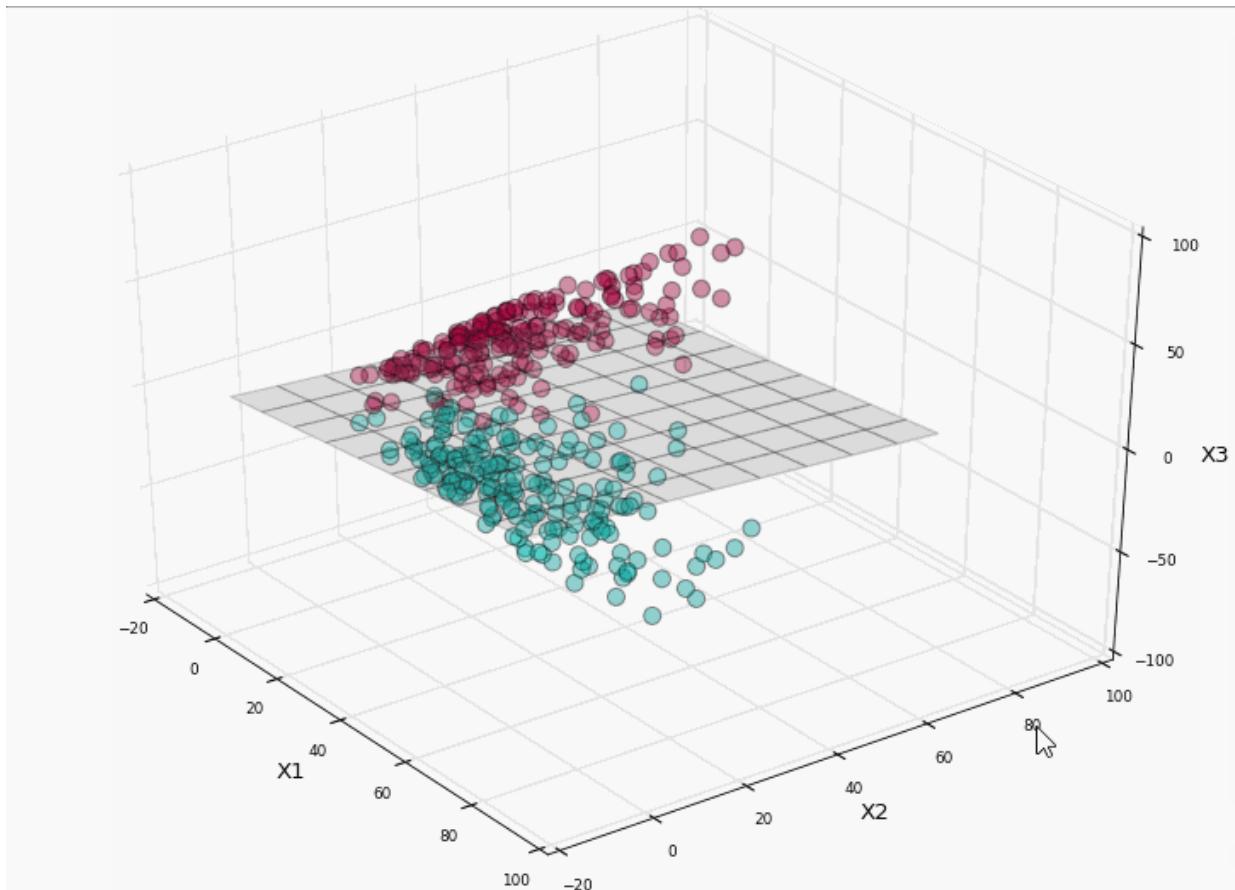


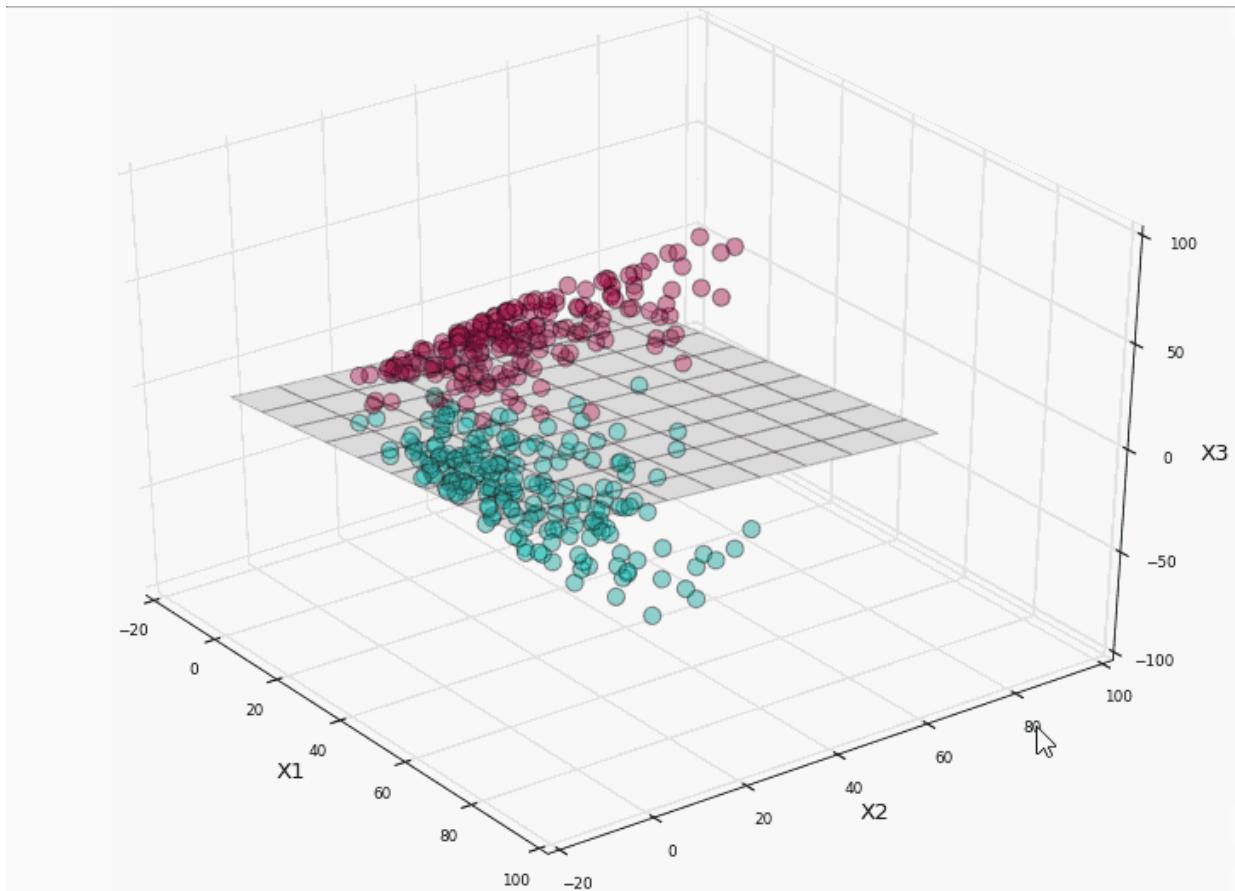


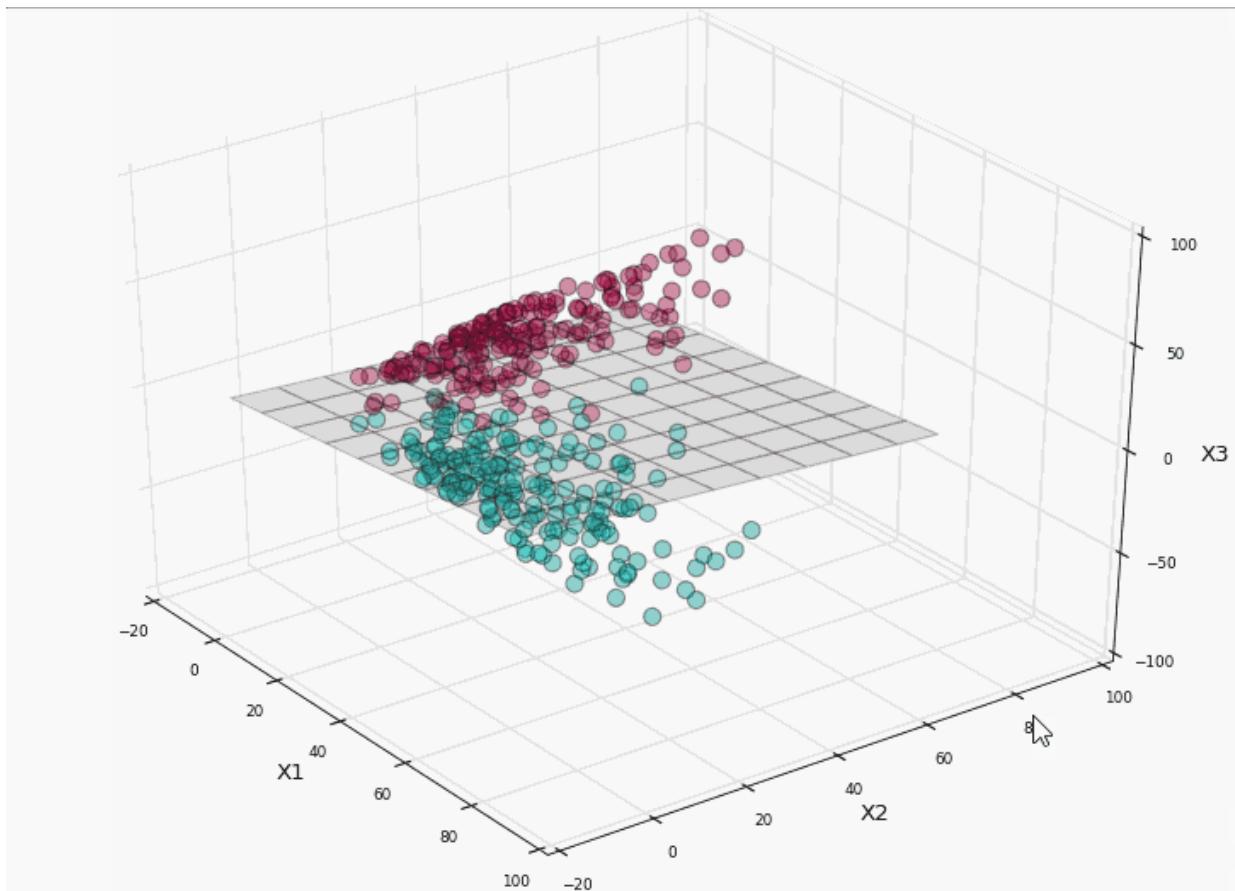


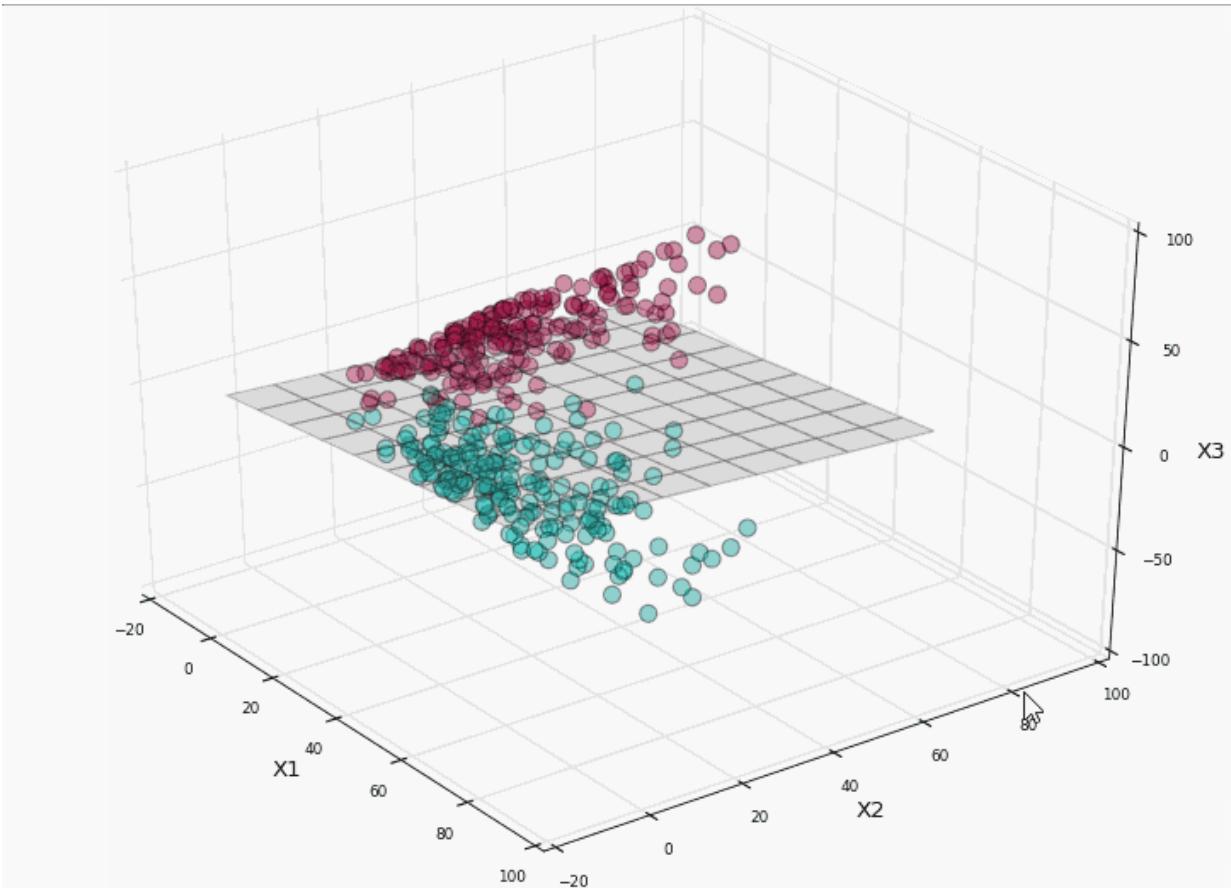


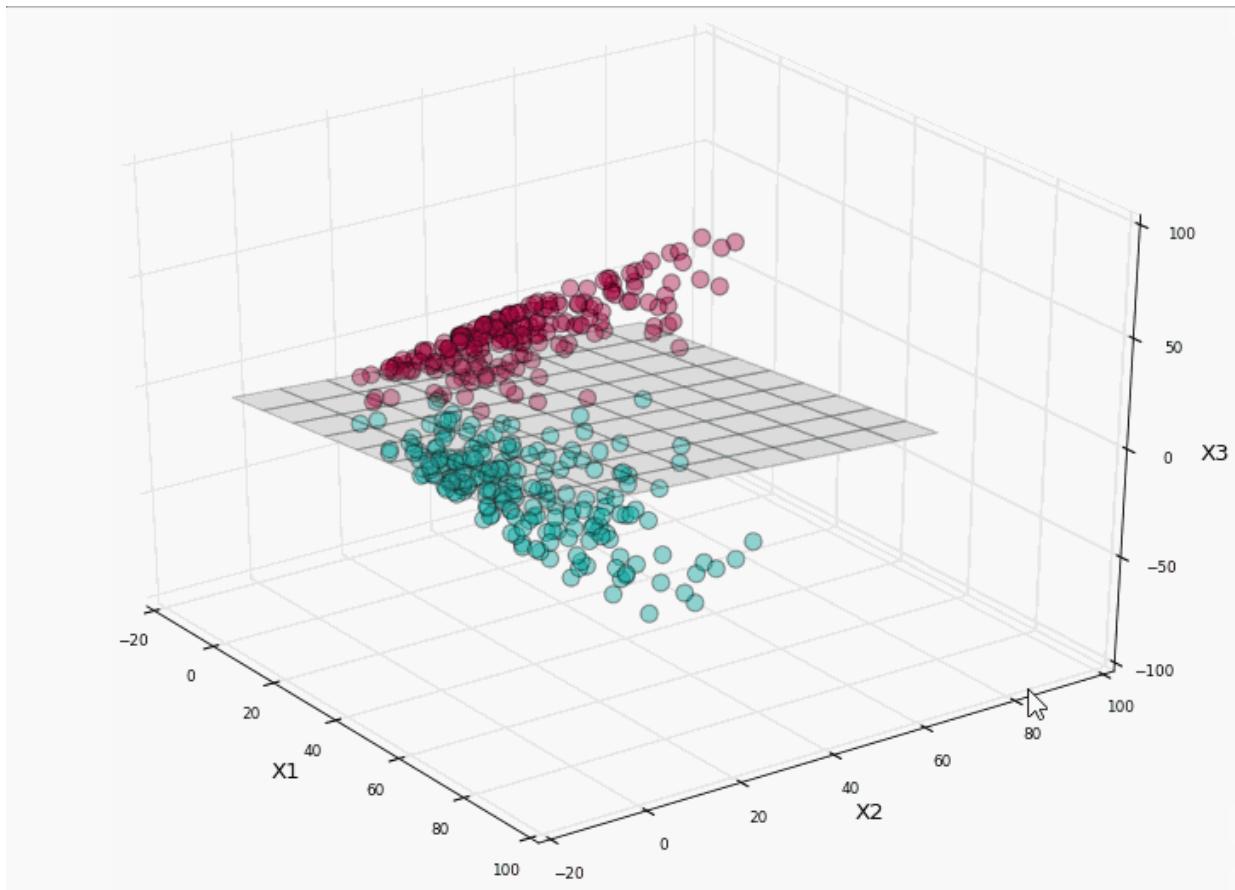


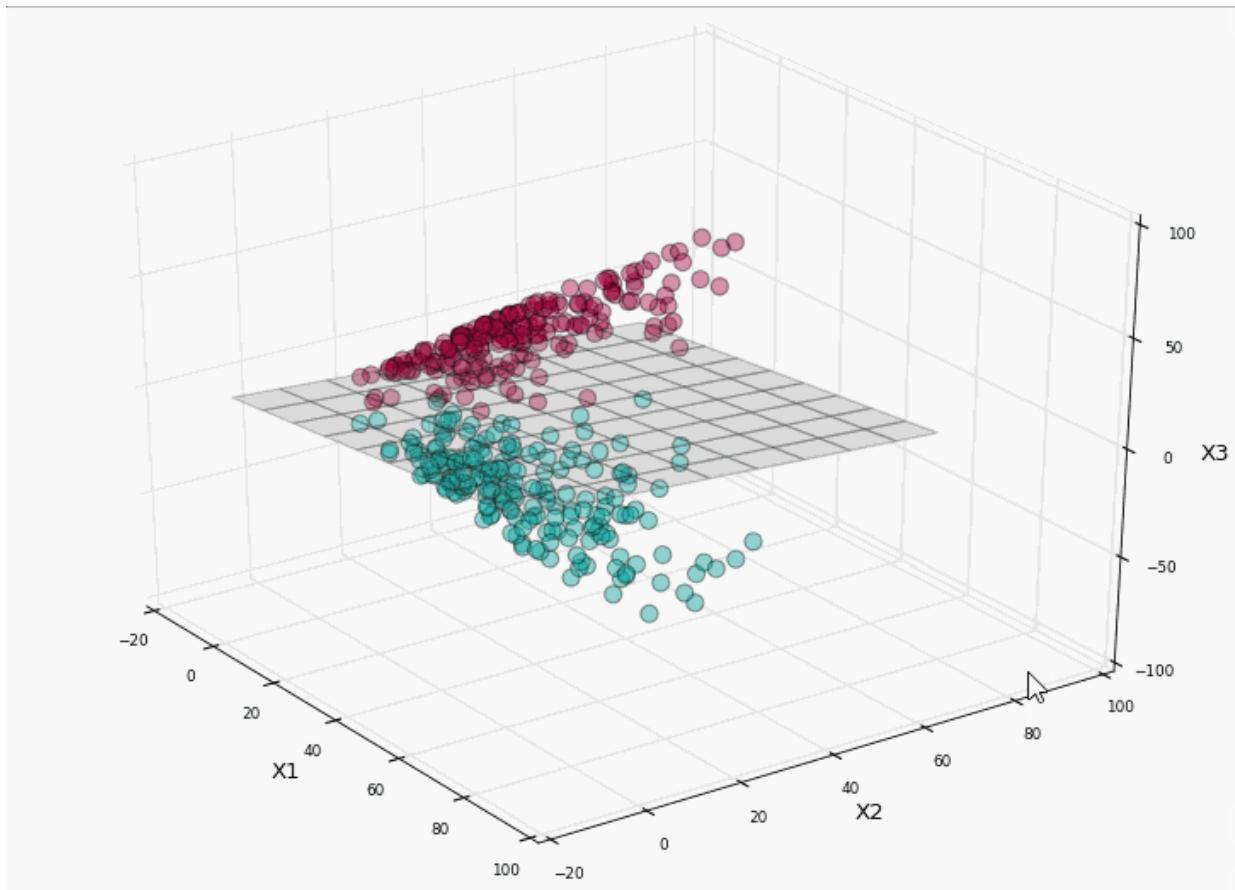


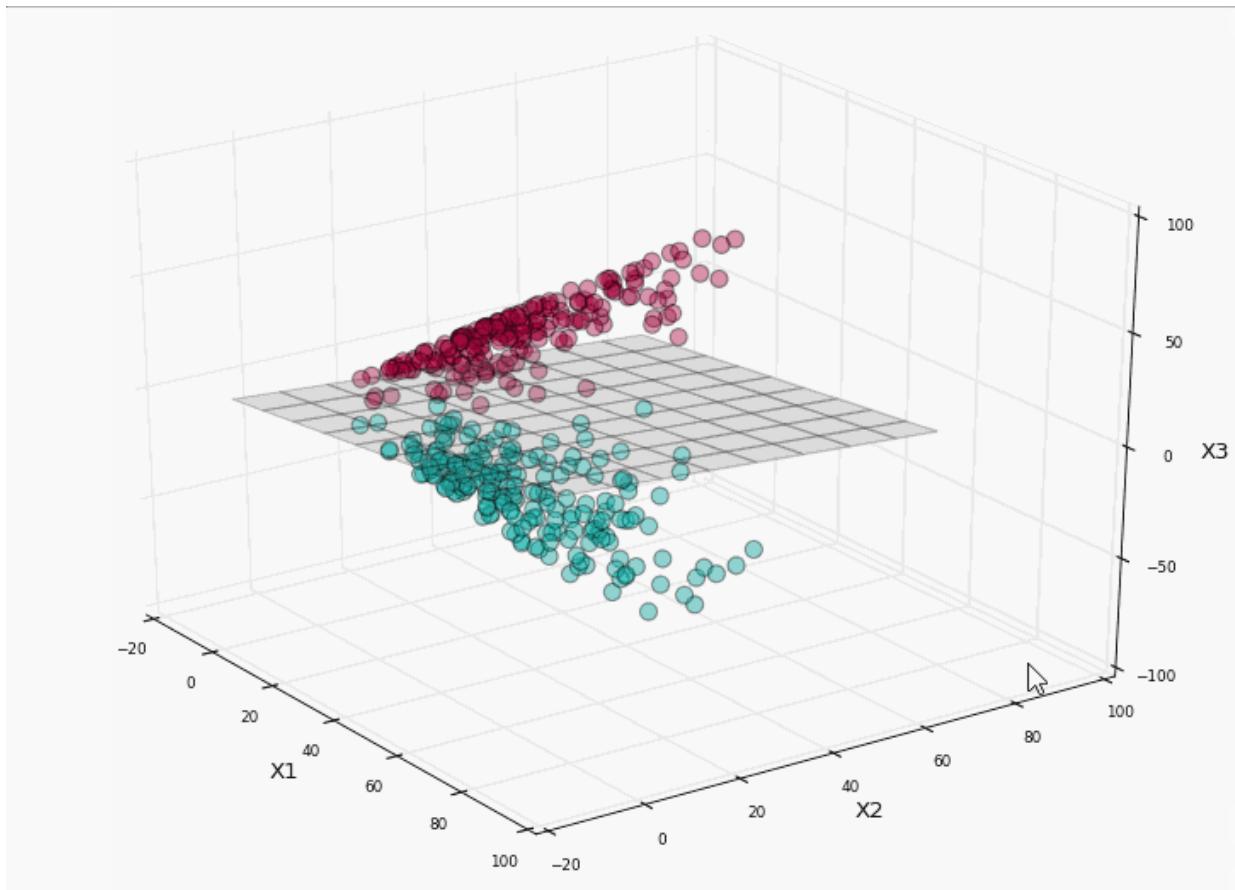


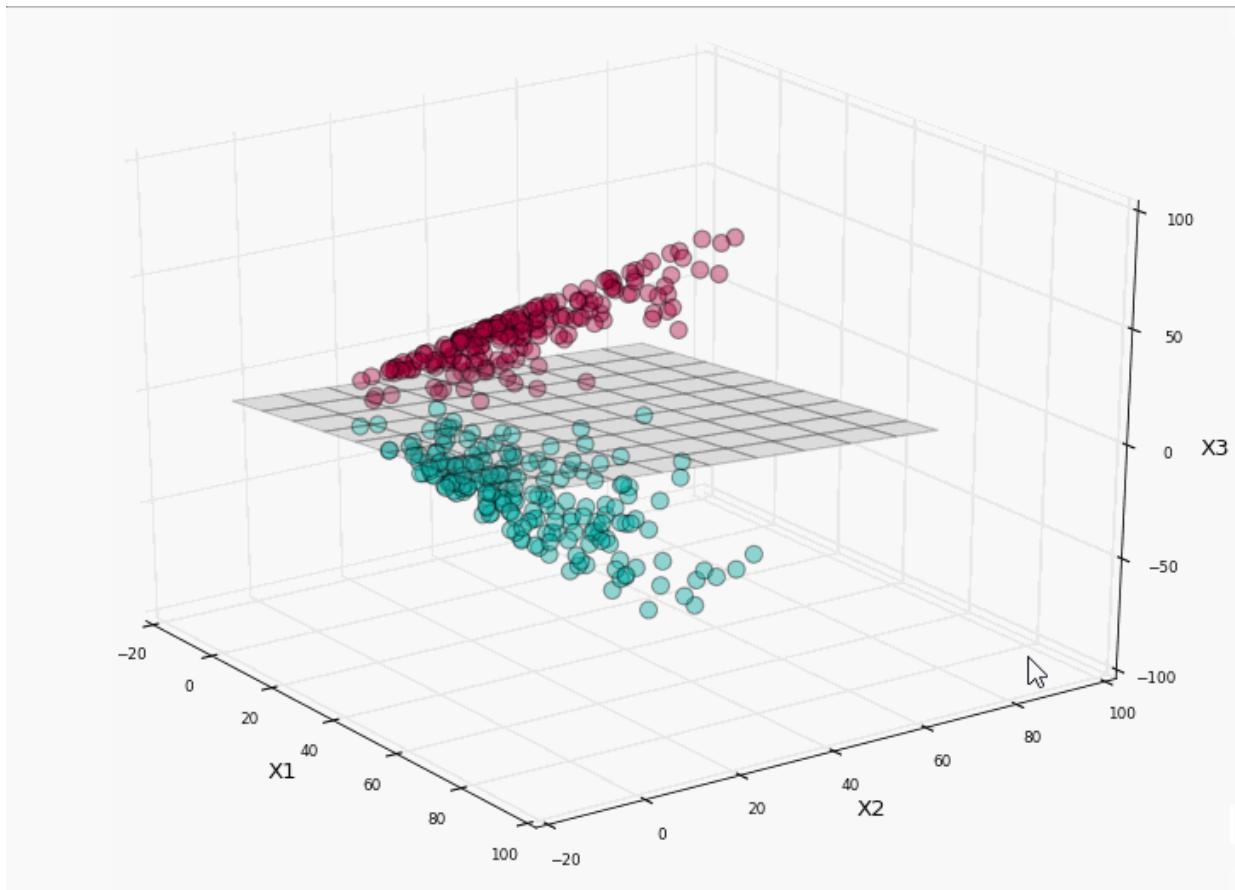


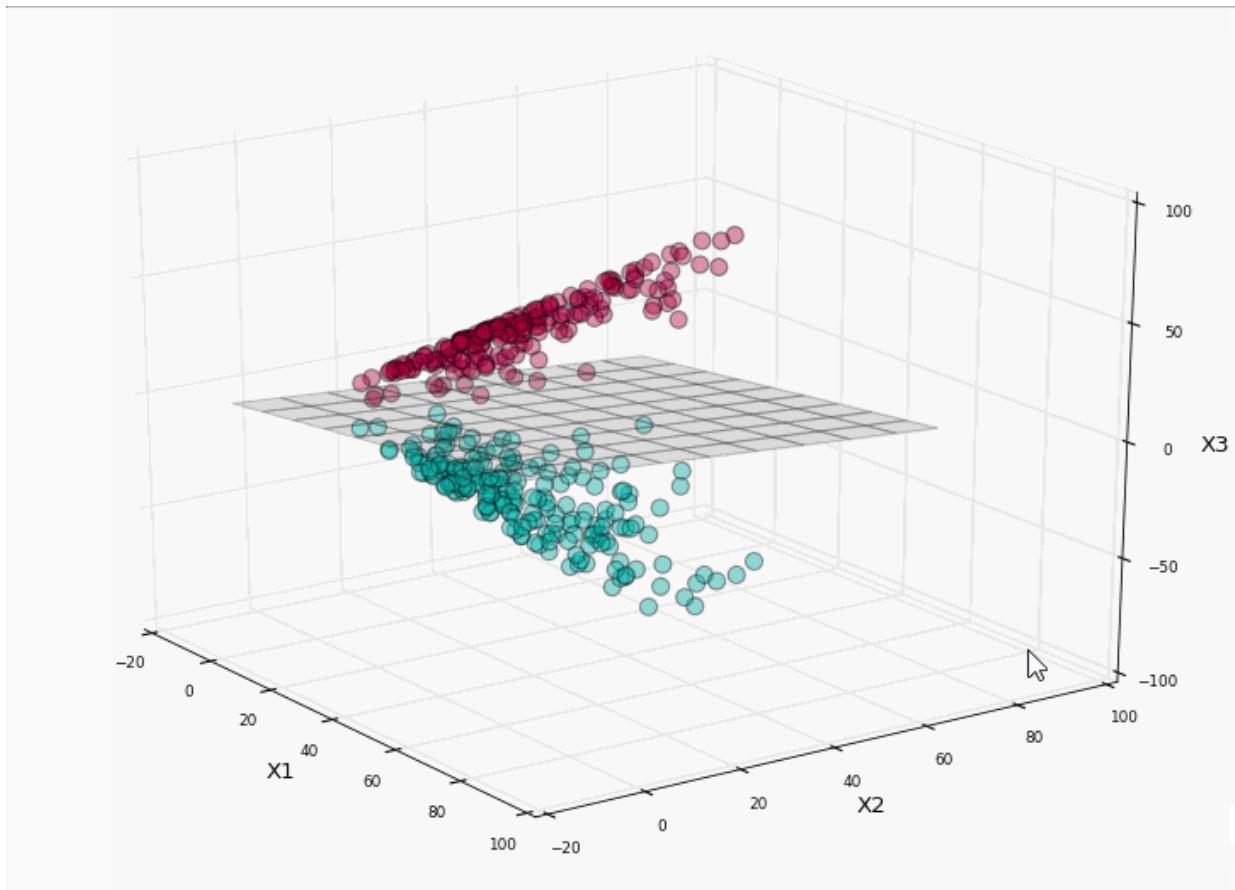


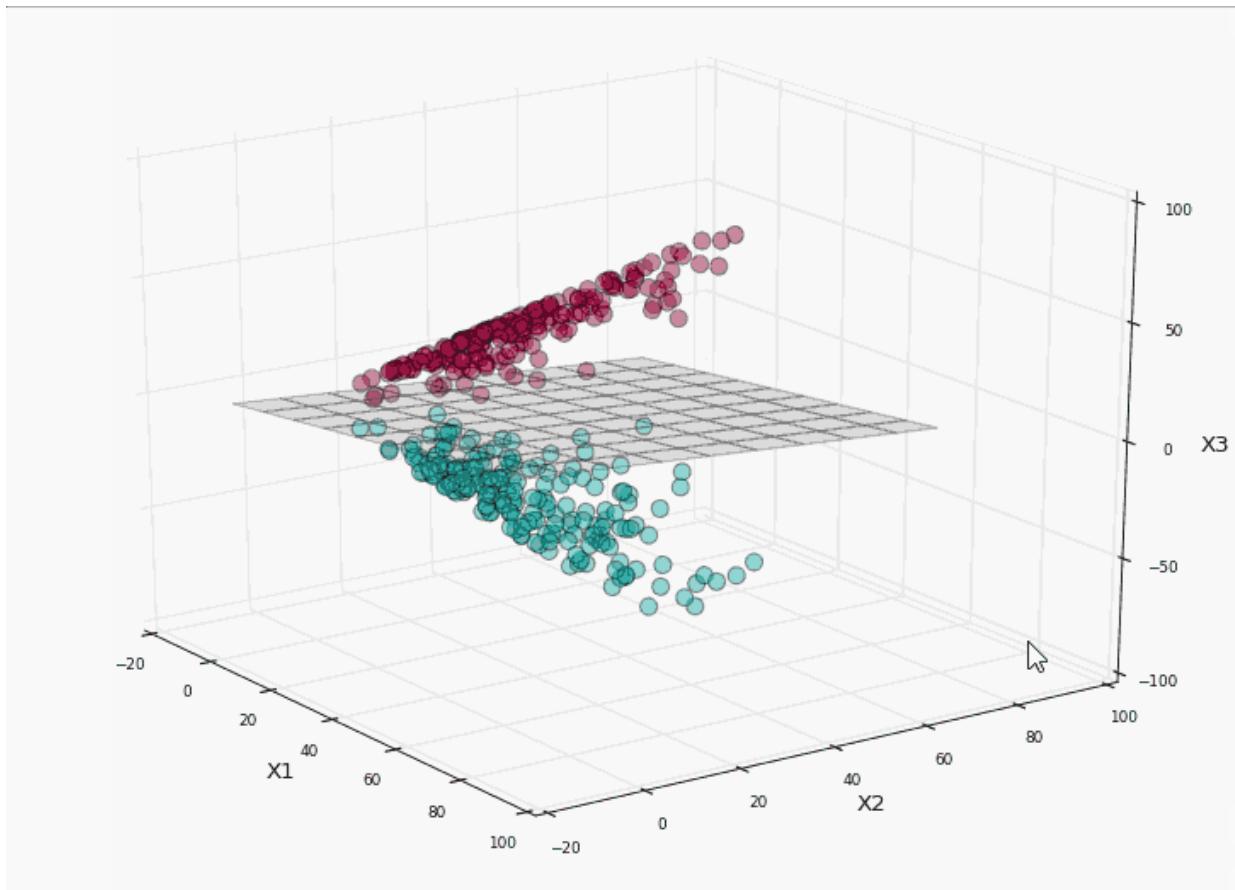


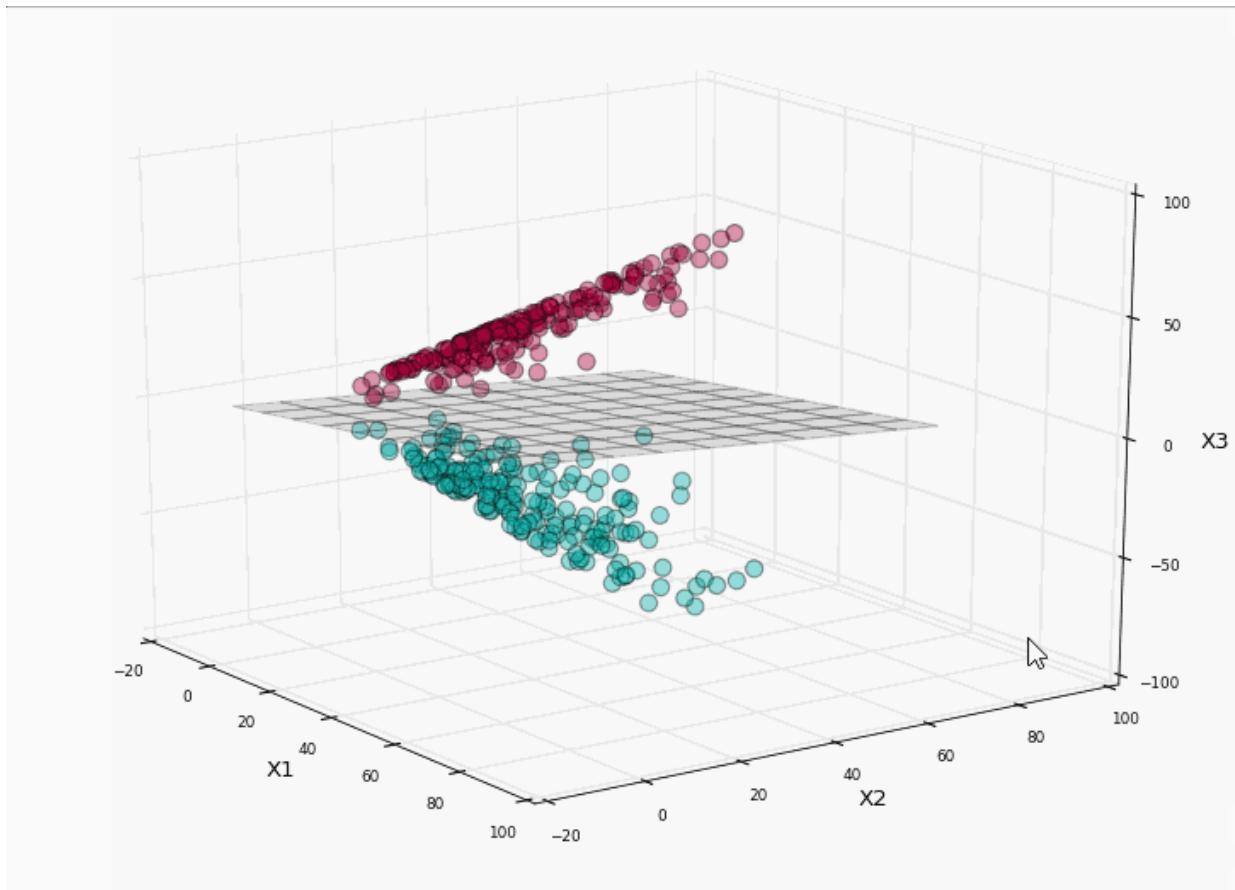


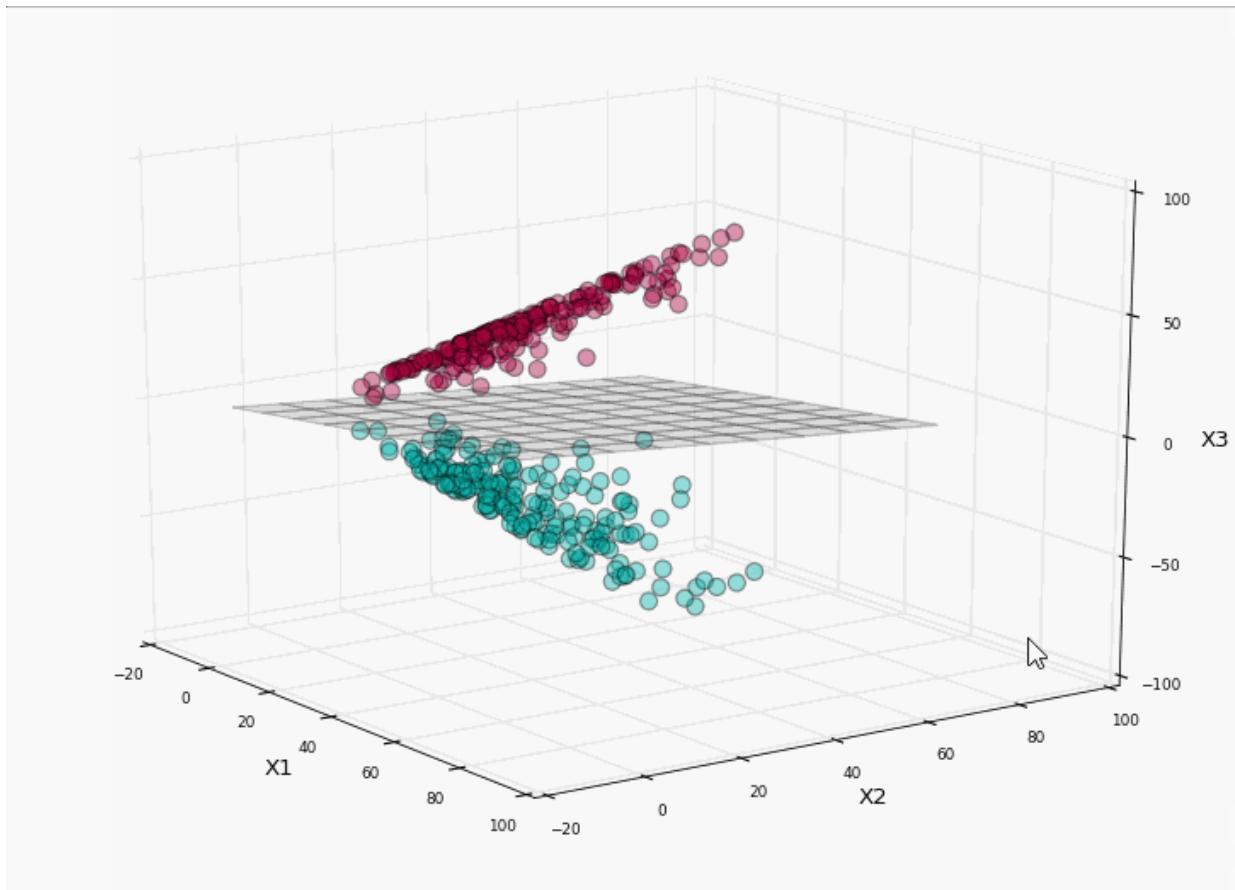


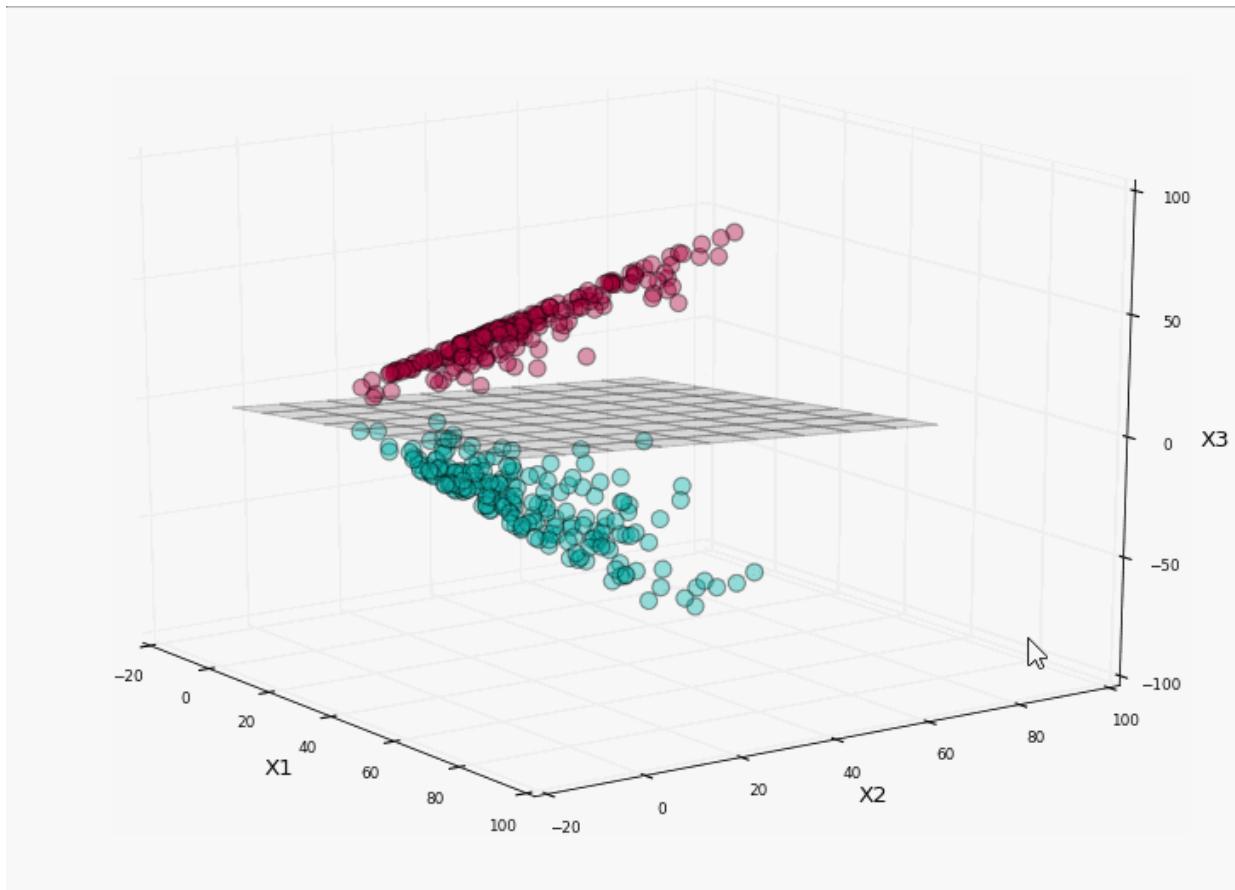


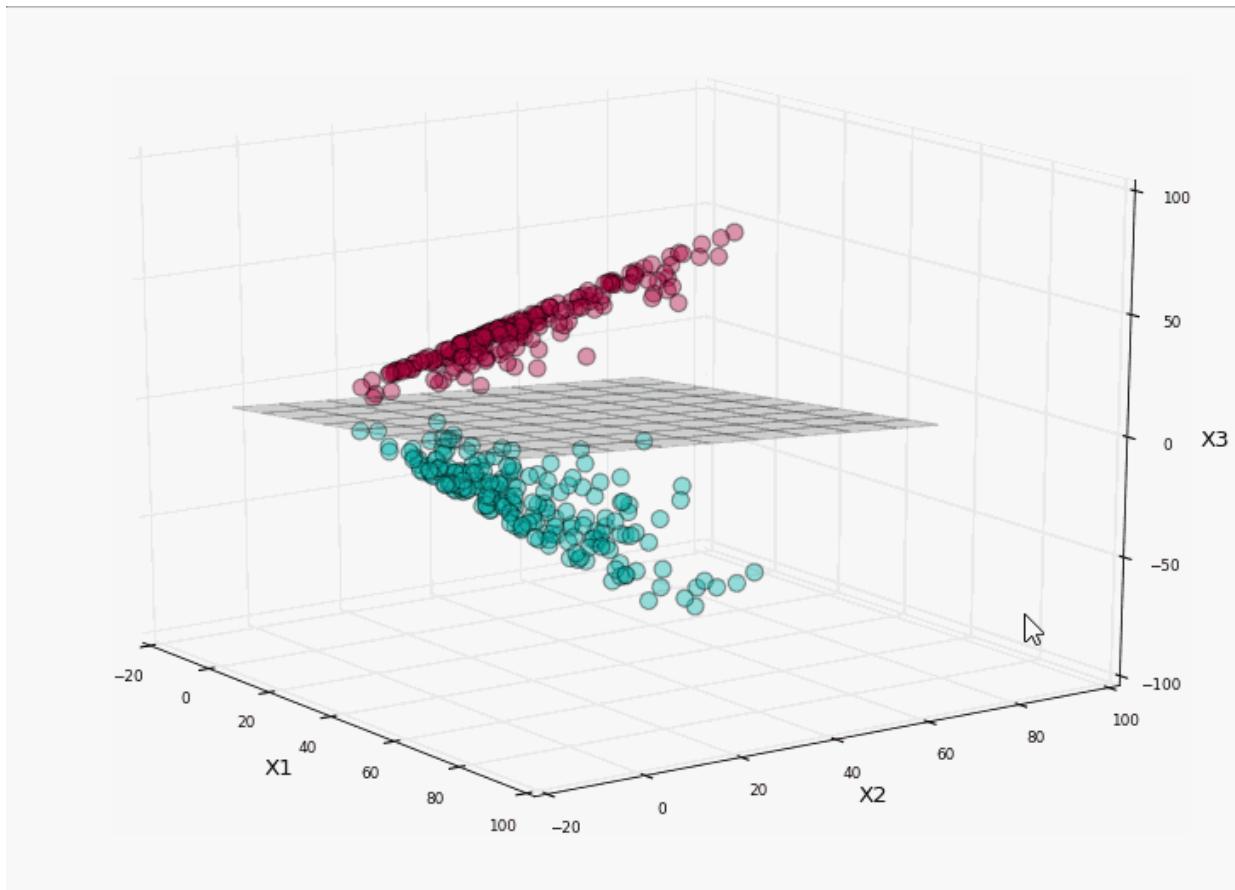


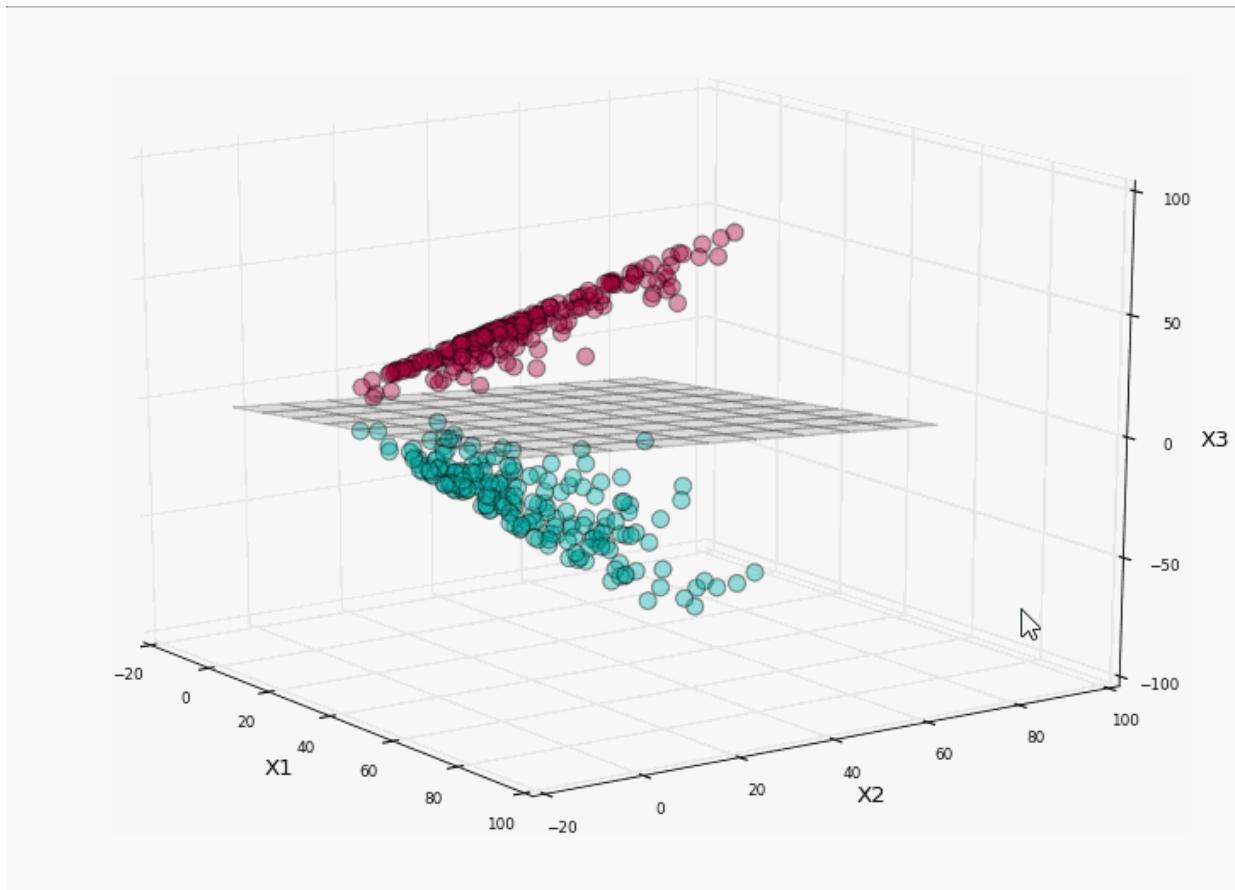


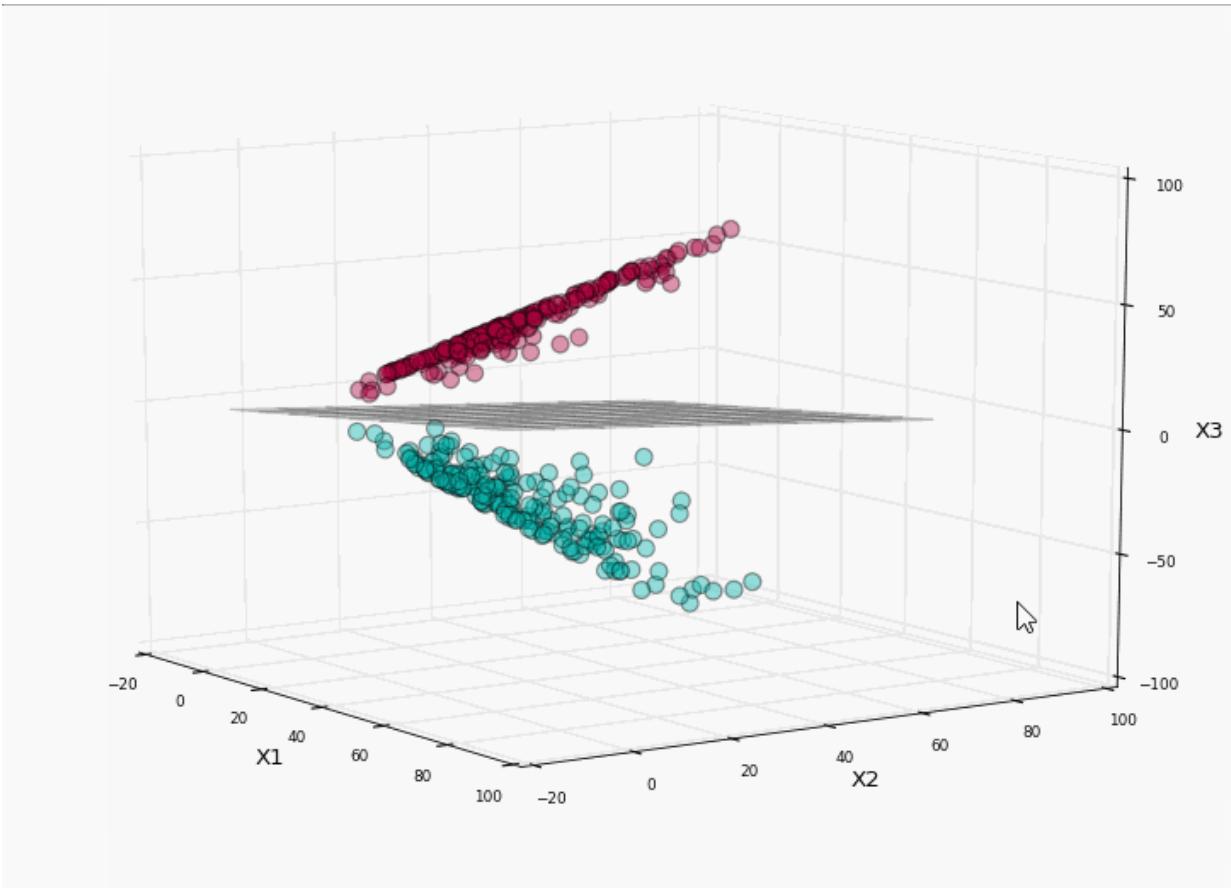


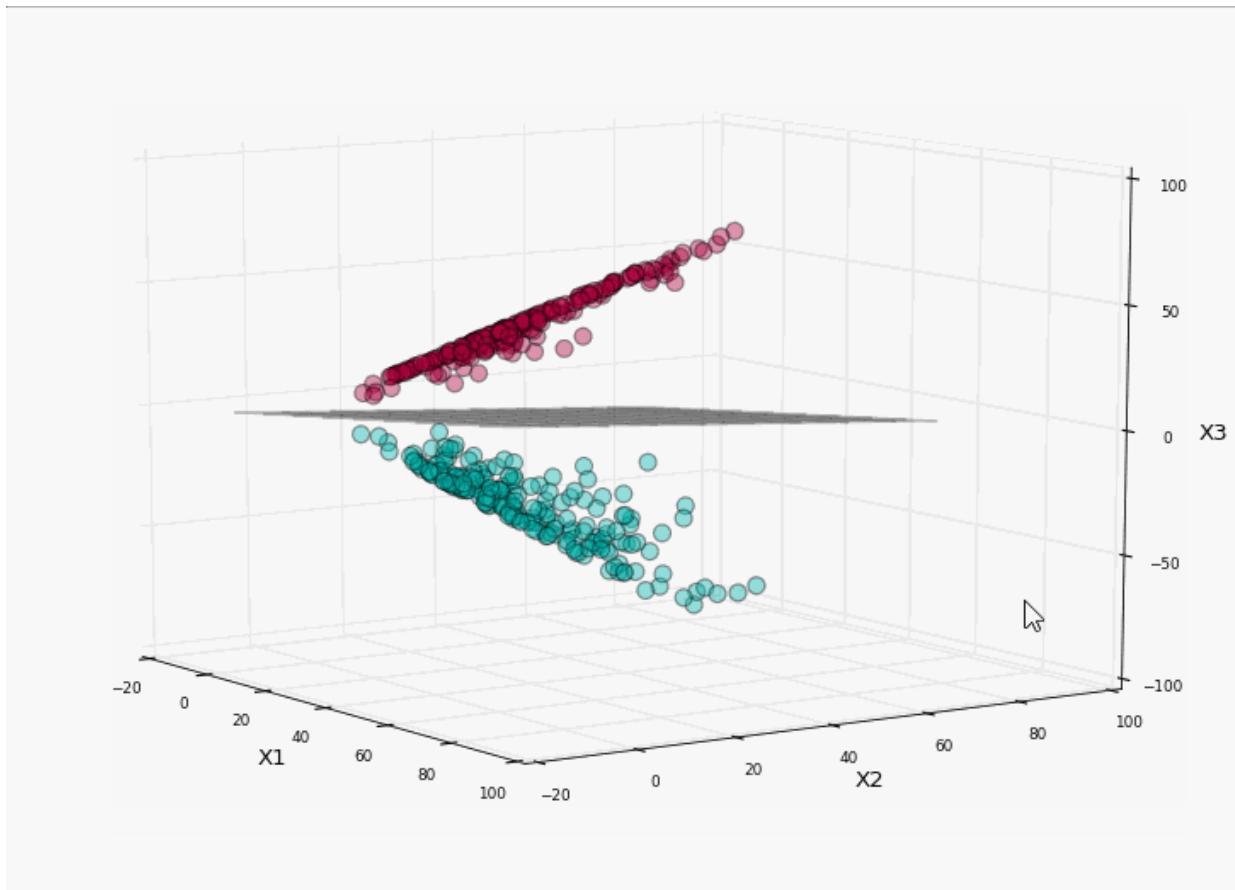


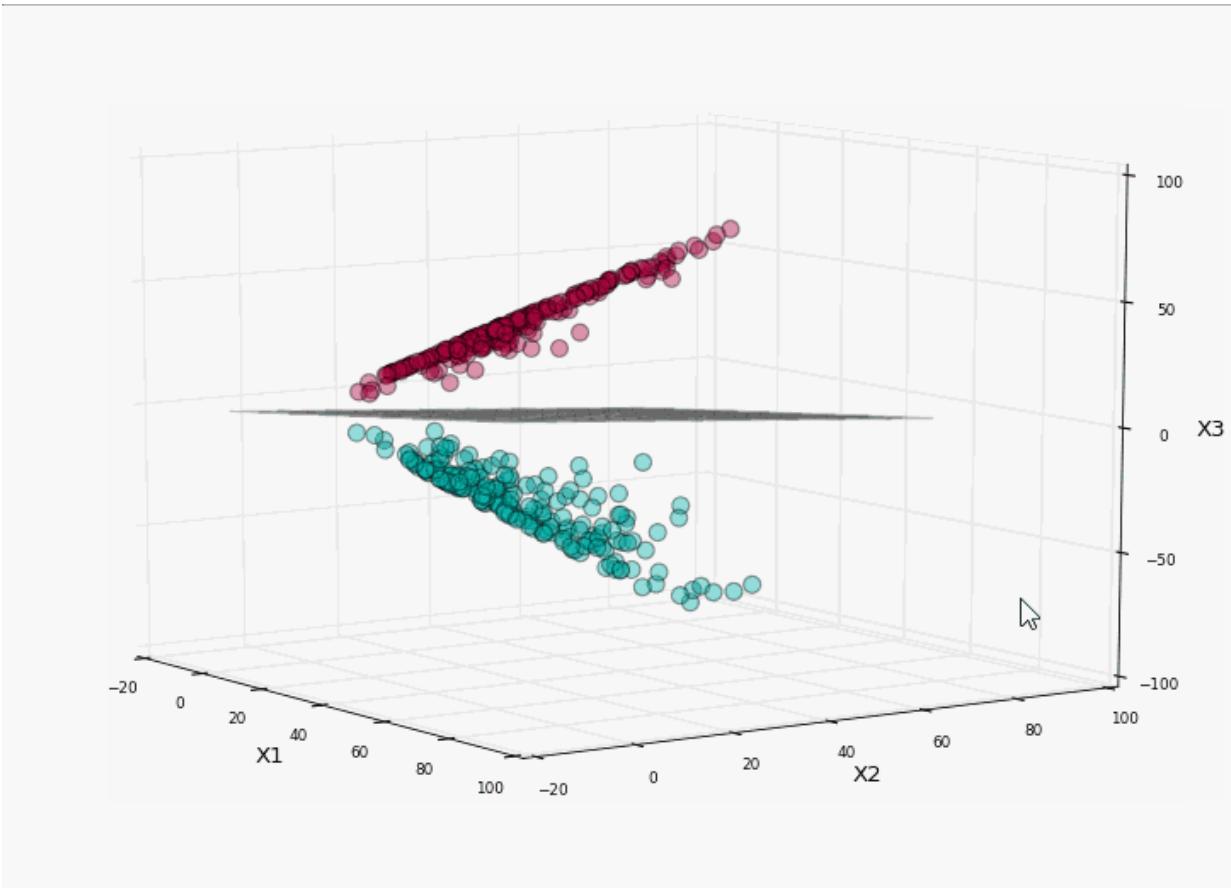


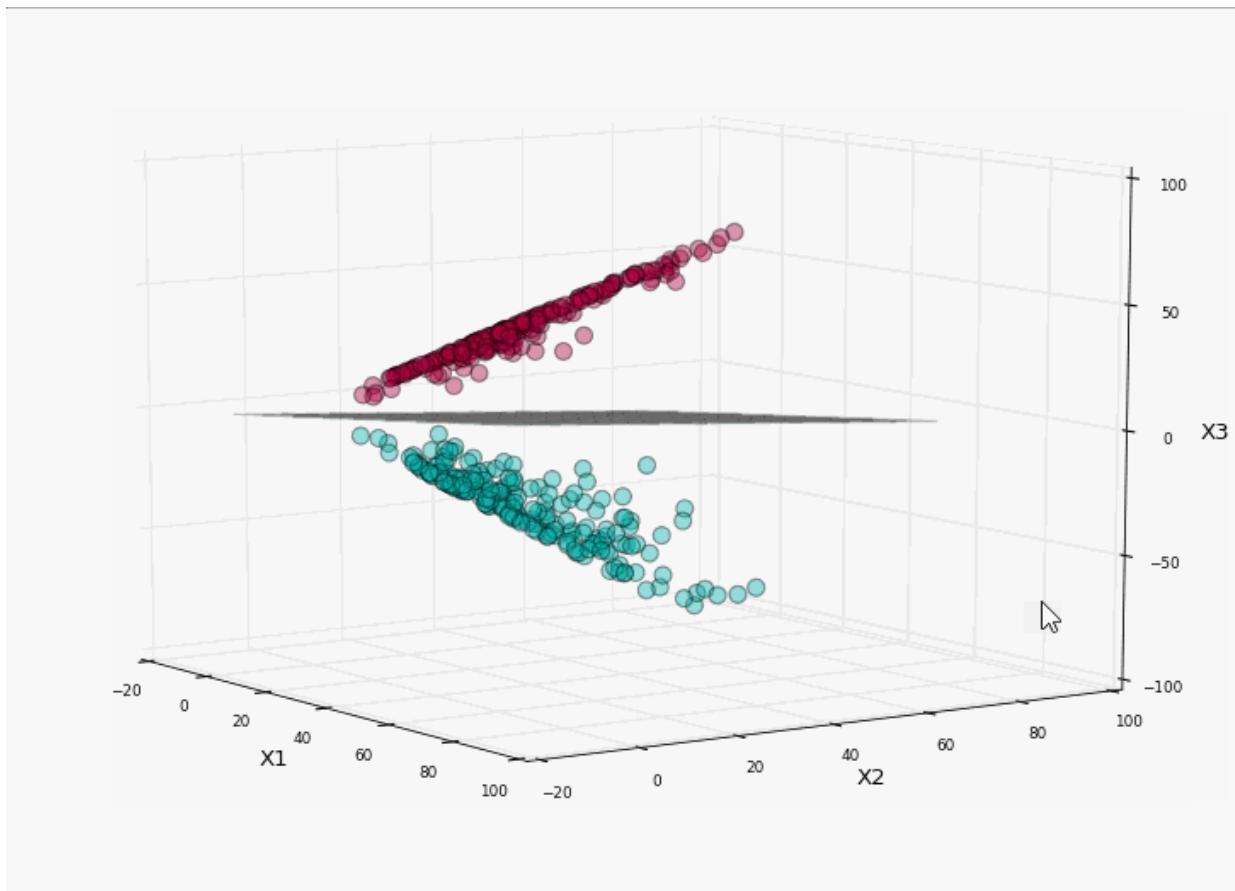


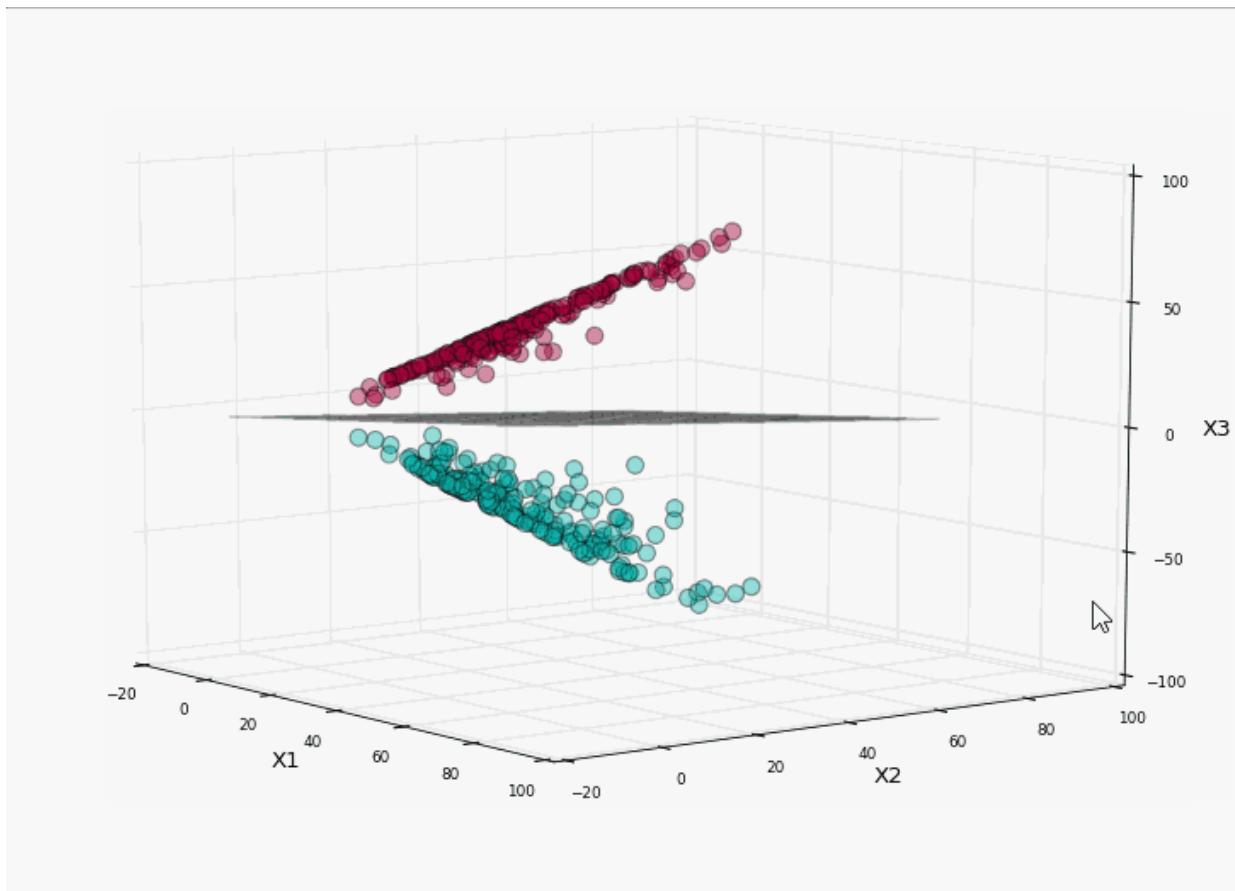


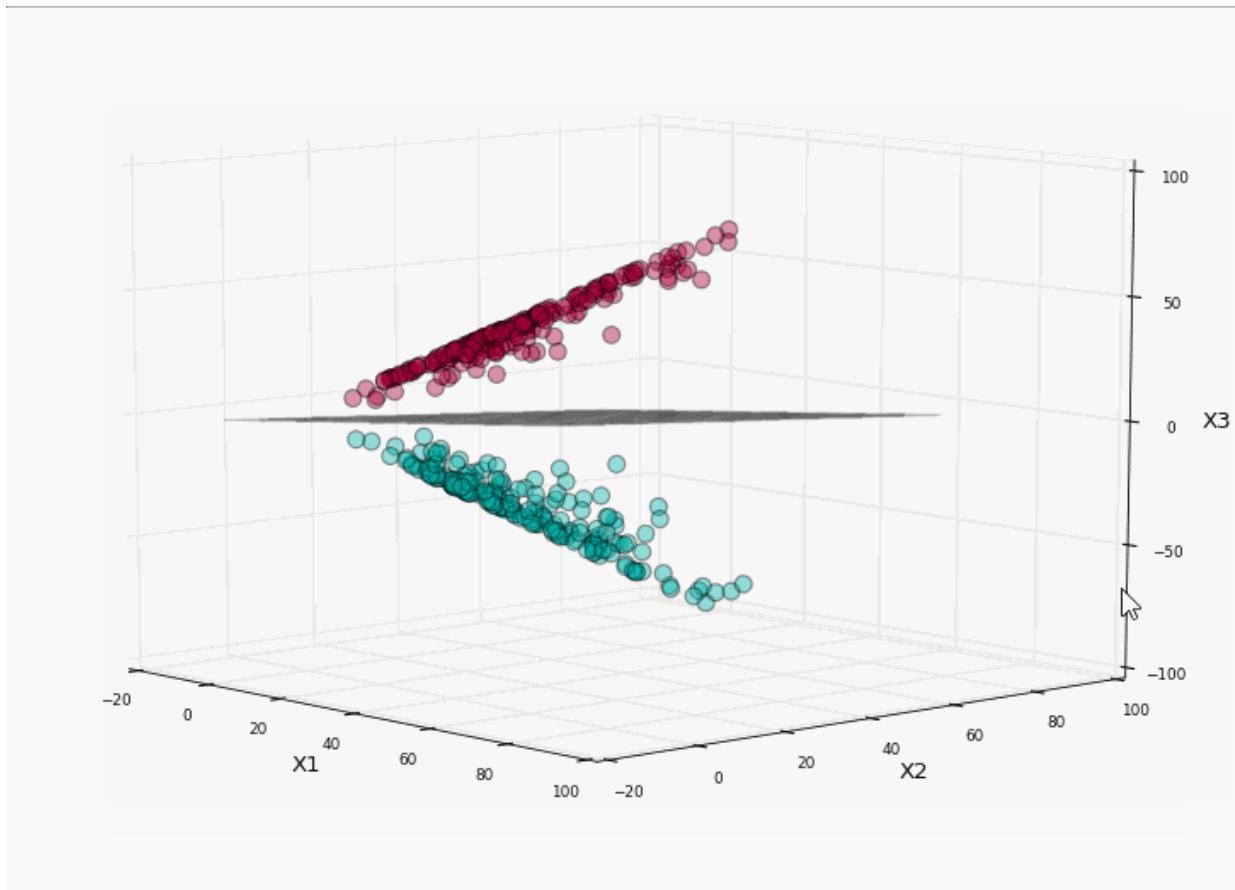


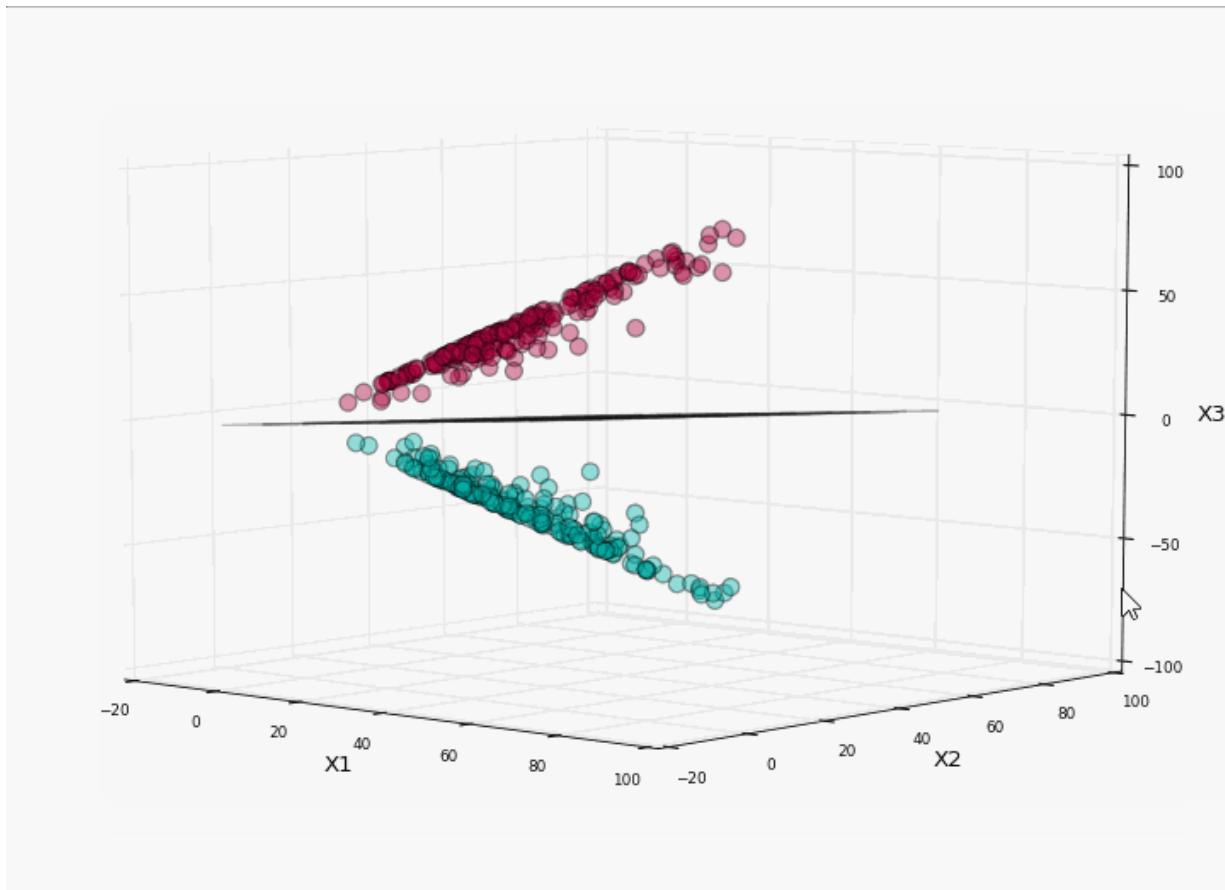


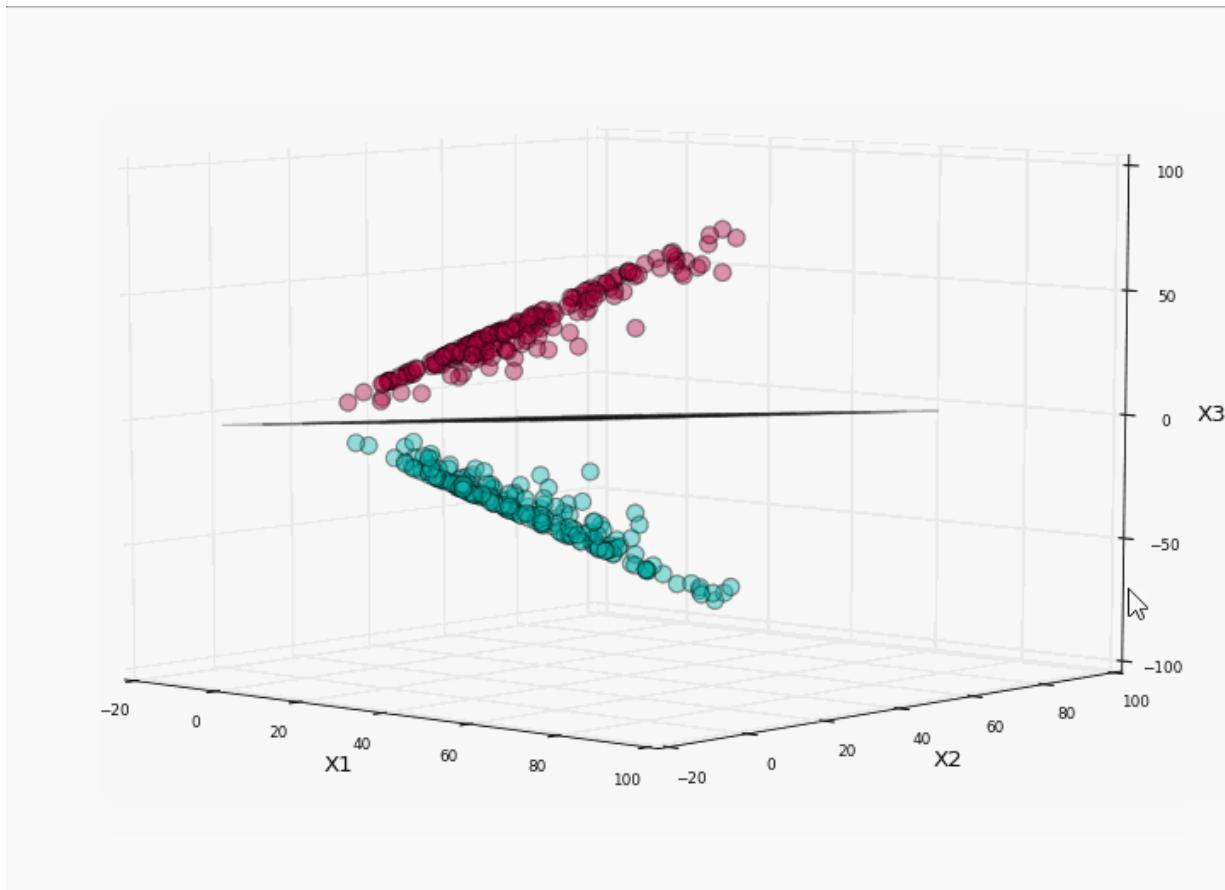


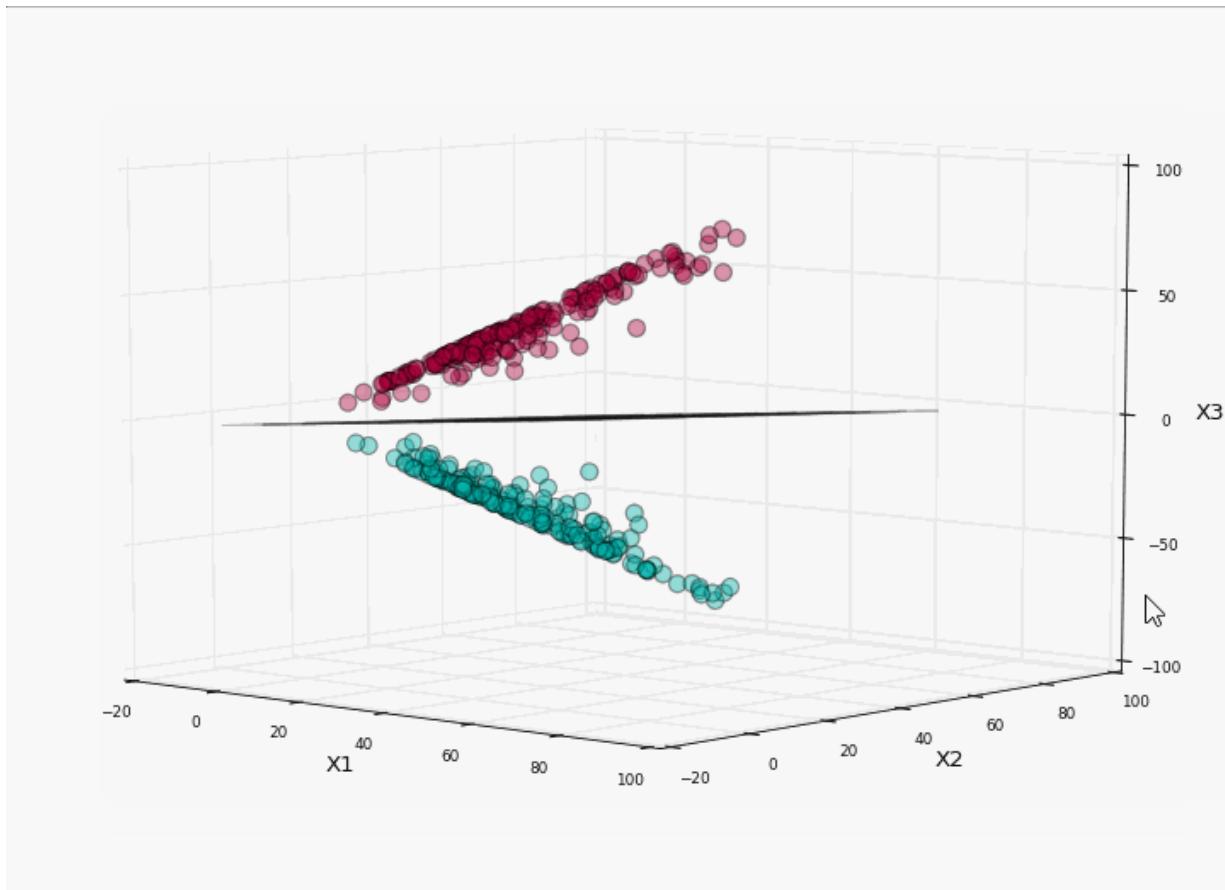


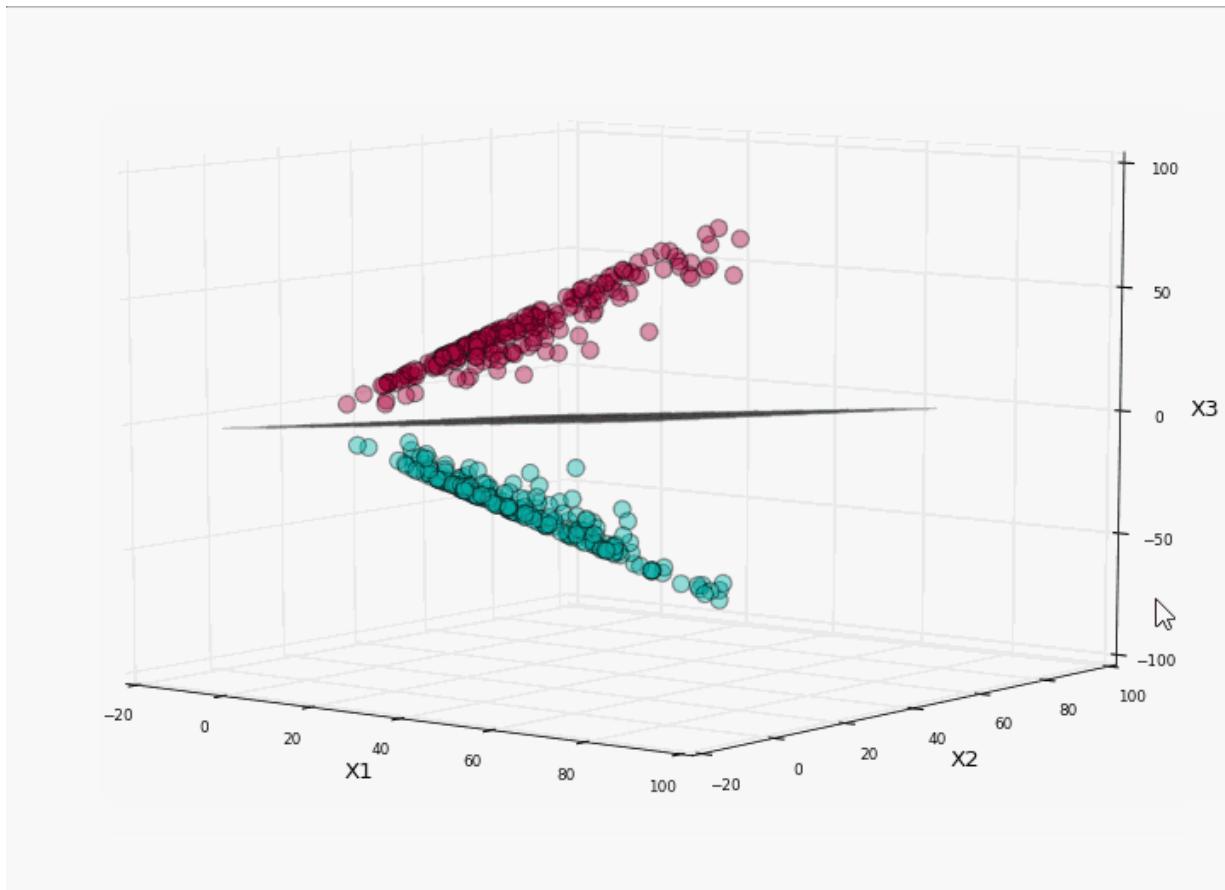


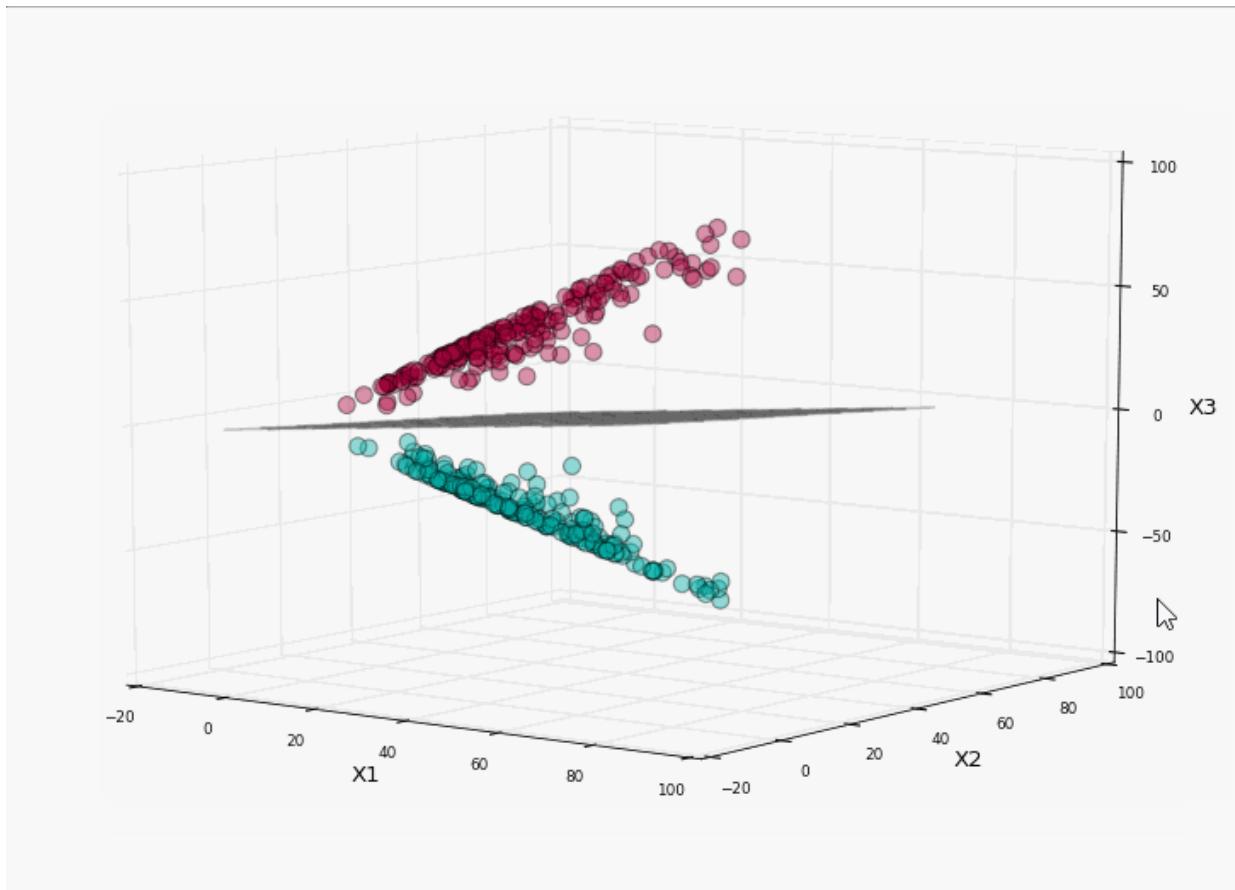


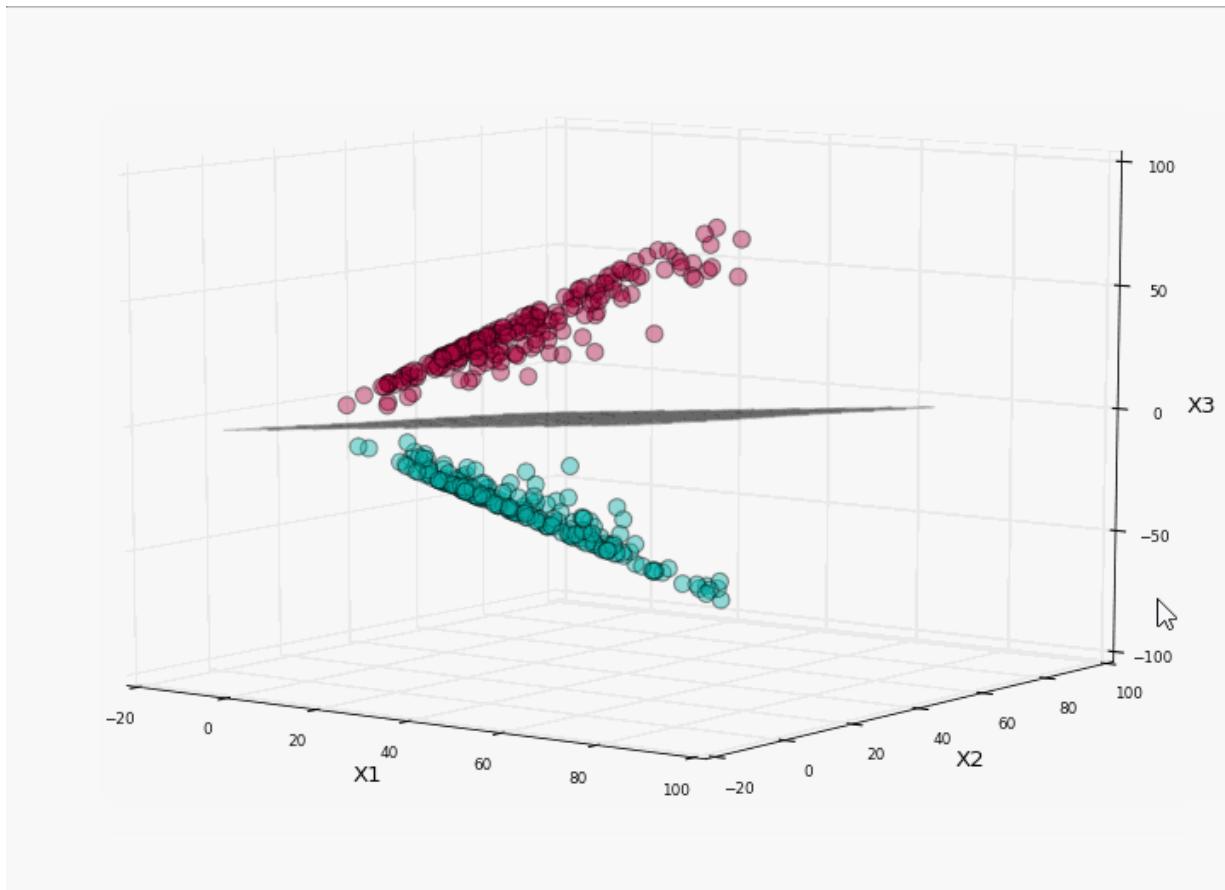


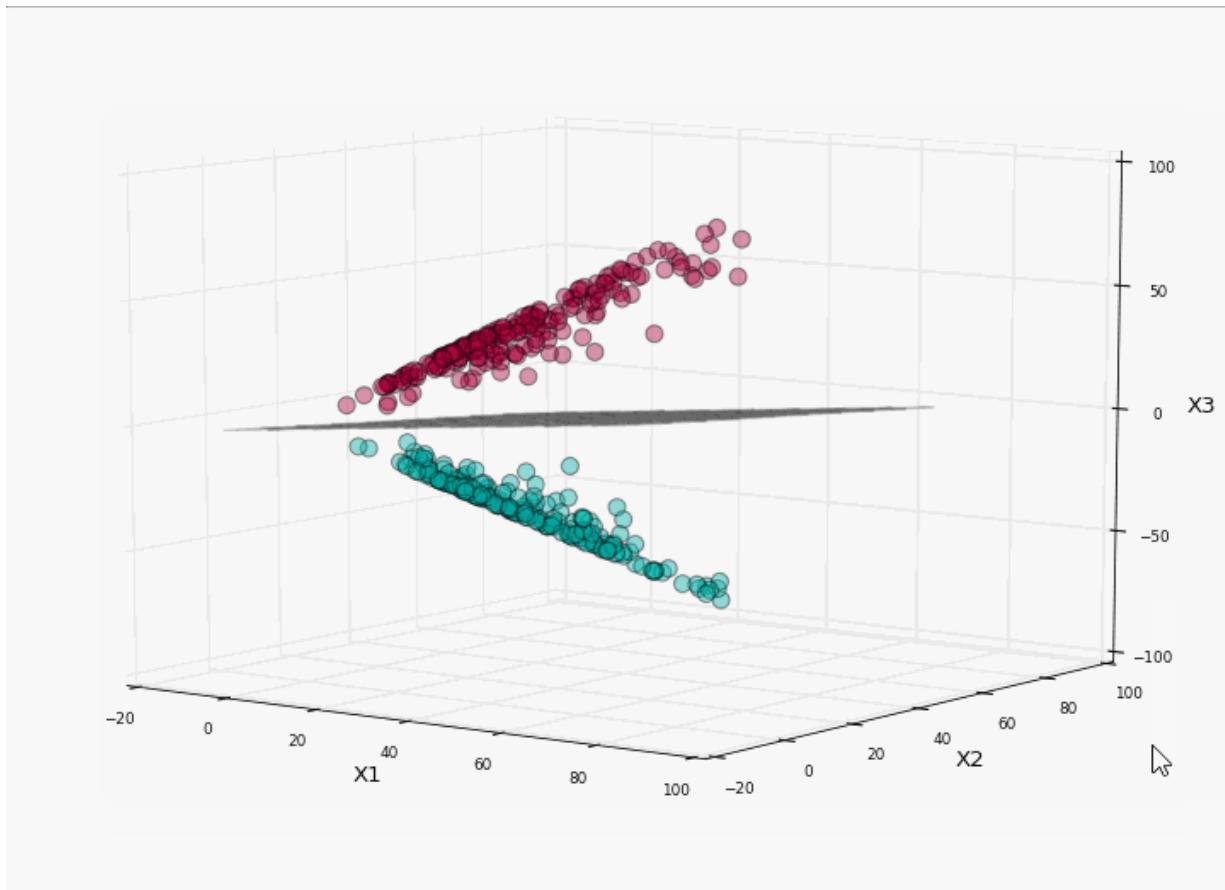


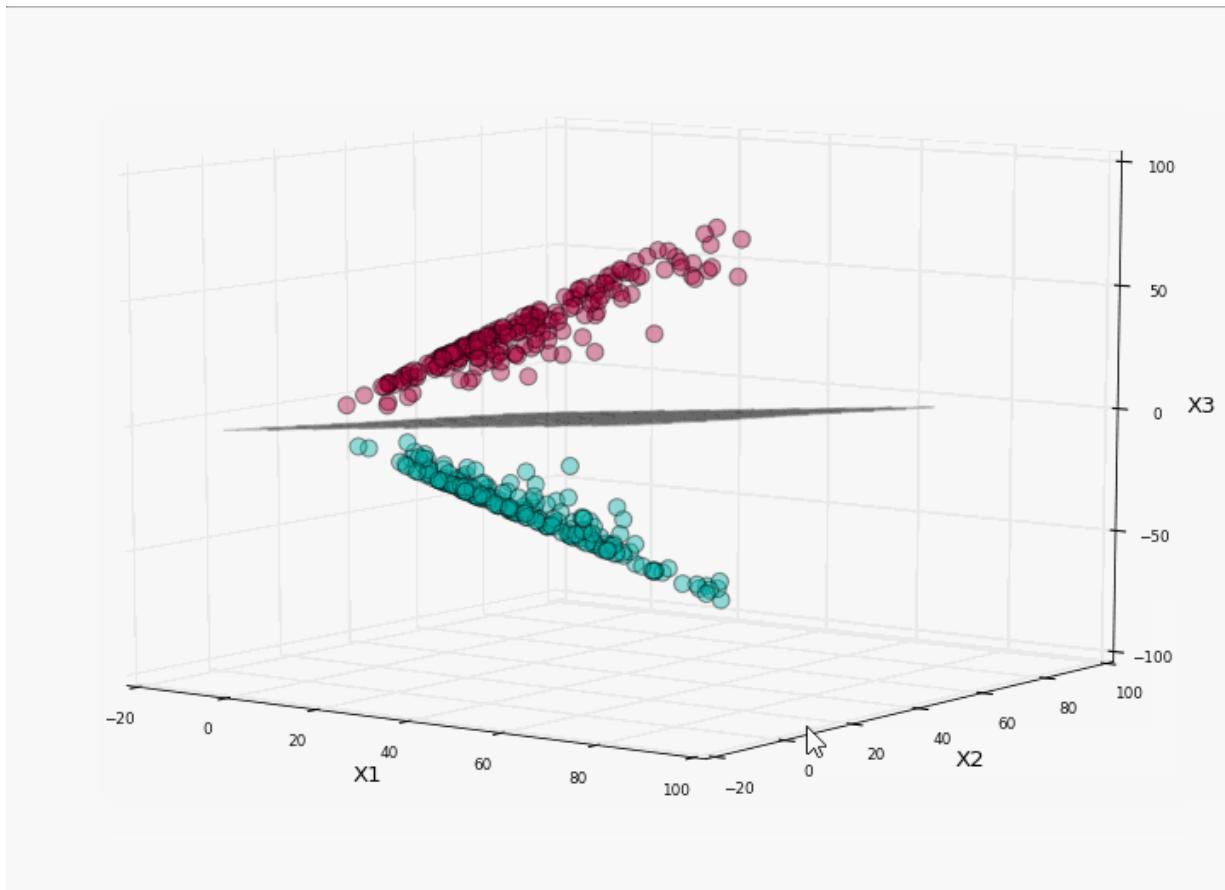


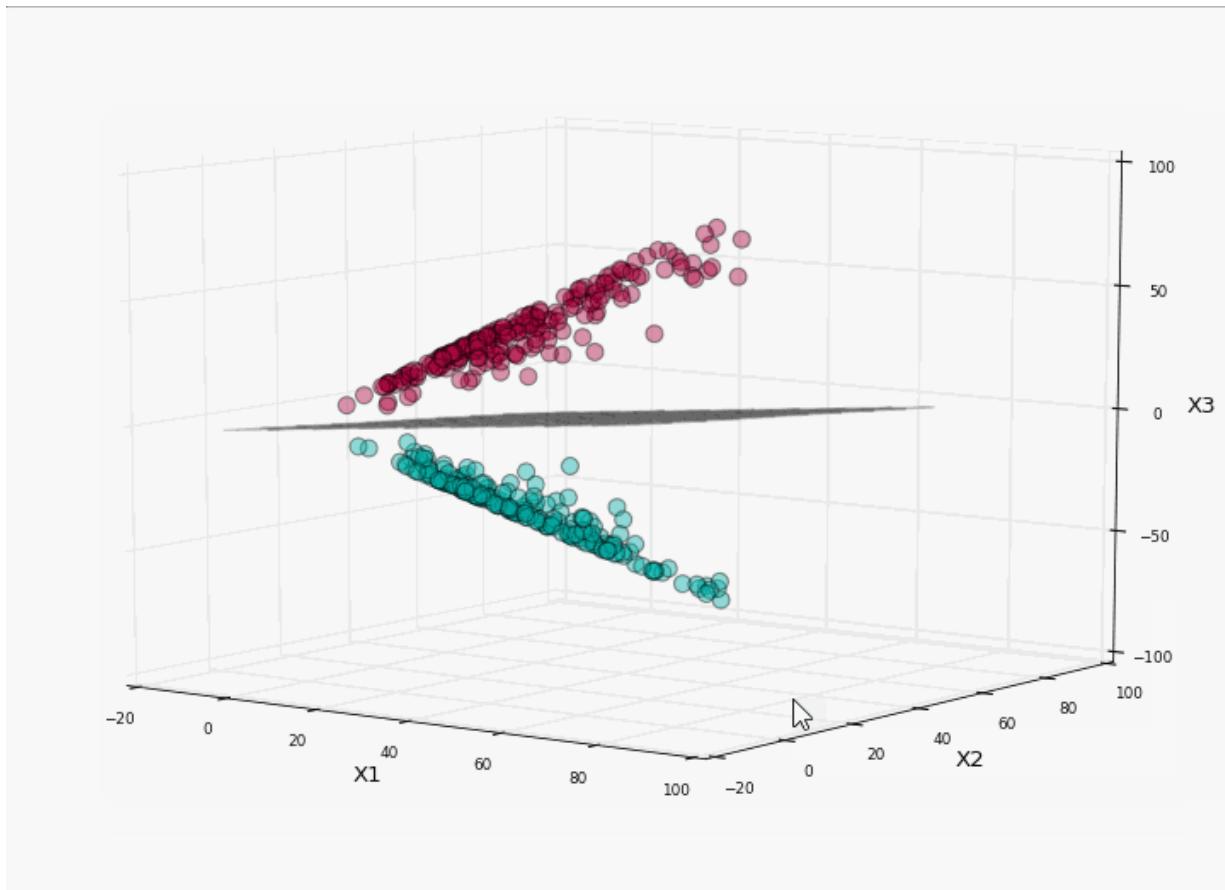


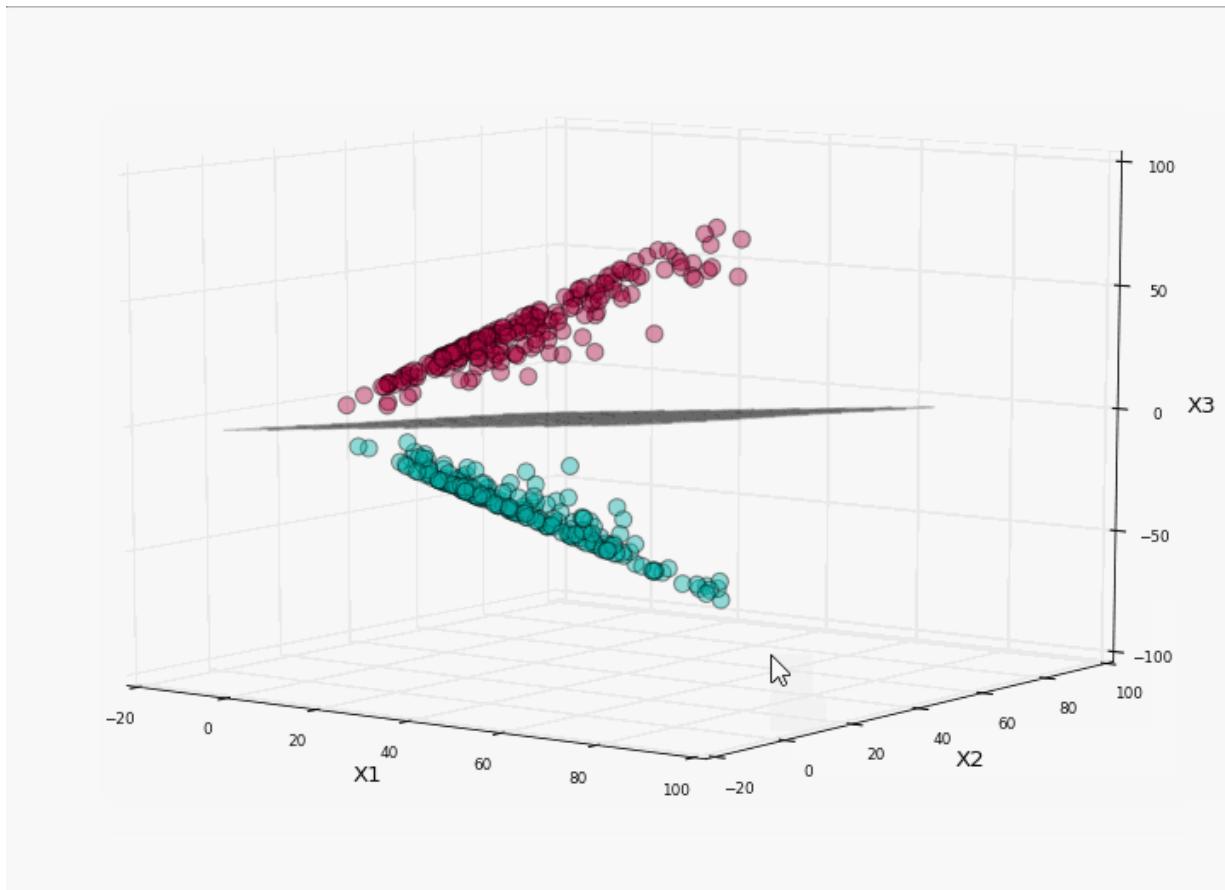


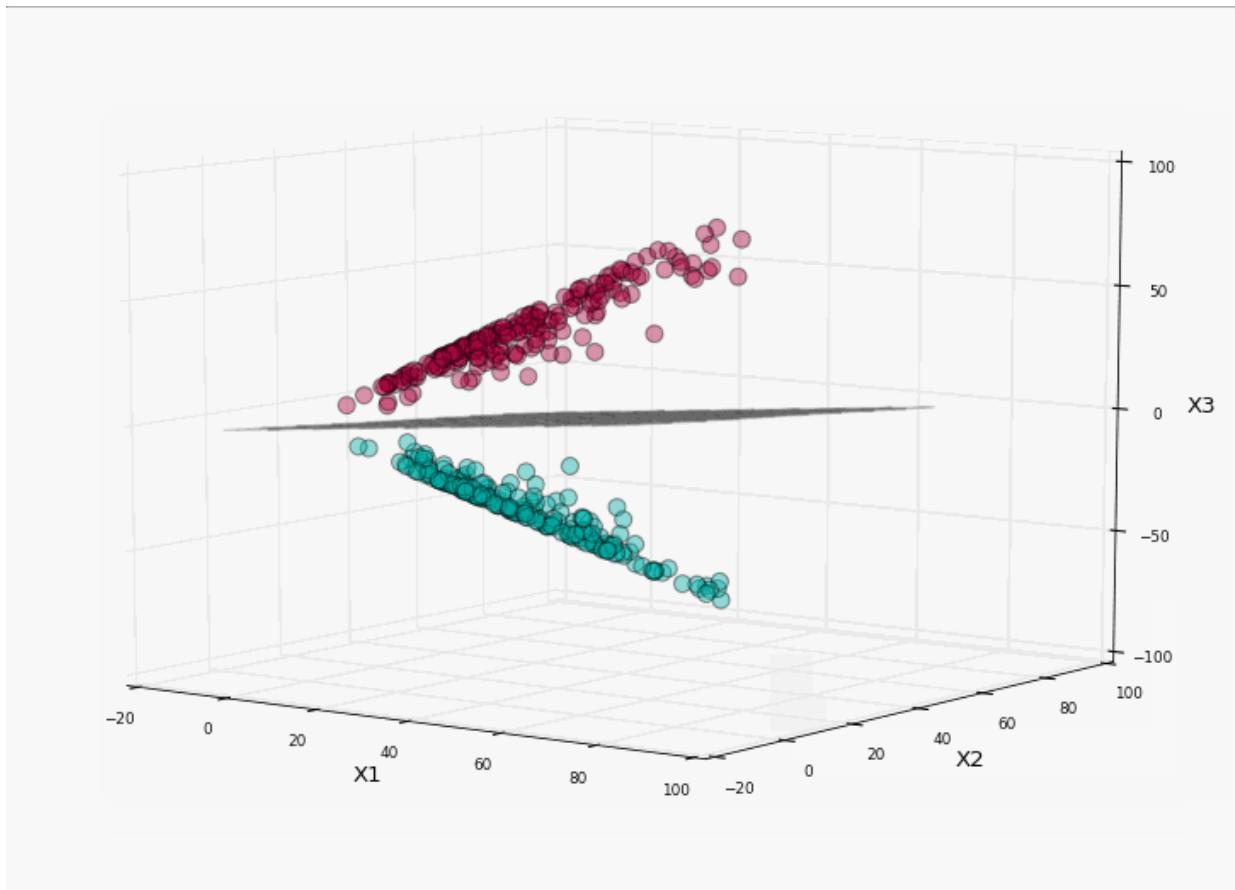


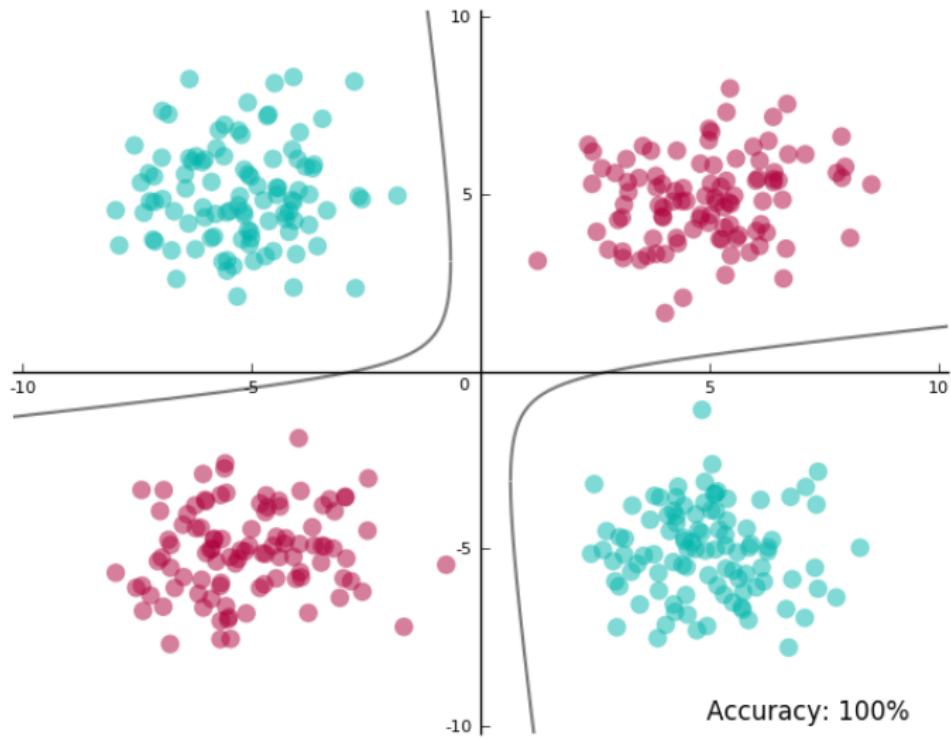












The kernel-based function is exactly equivalent to preprocessing the data by applying similarity function to all inputs, then learning a linear model in the new transformed space.

Commonly used kernels

- Homogeneous polynomials

$$k(x, y) = (\langle x, y \rangle)^d$$

- Inhomogeneous polynomials

$$k(x, y) = (\langle x, y \rangle + 1)^d$$

- Gaussian Kernel

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

- Sigmoid Kernel

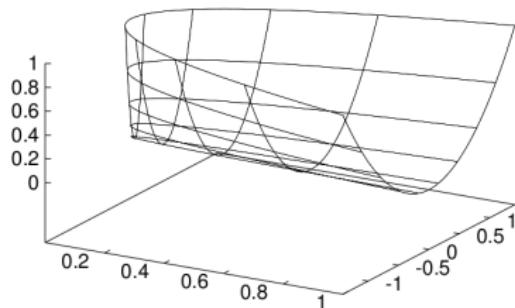
$$k(x, y) = \tanh(\eta \langle x, y \rangle + v)$$

Polynomial kernel

$$k(x, y) = (\langle x, y \rangle)^d$$

Example: $n = 2, d = 2, x = (x_1, x_2)$

- $\Phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$

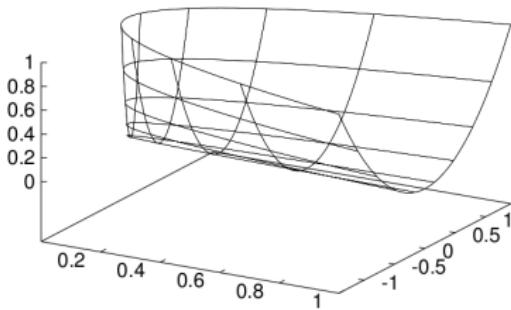


Polynomial kernel

$$k(x, y) = (\langle x, y \rangle)^d$$

Example: $n = 2, d = 2, x = (x_1, x_2)$

- $\Phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$



- Neither the mapping Φ nor the feature space is unique
 - $\Phi(x) = (x_1^2, x_1x_2, x_1x_2, x_2^2)$
 - $\Phi(x) = \frac{1}{\sqrt{2}} (x_1^2 - x_2^2, 2x_1x_2, x_1^2 + x_2^2)$