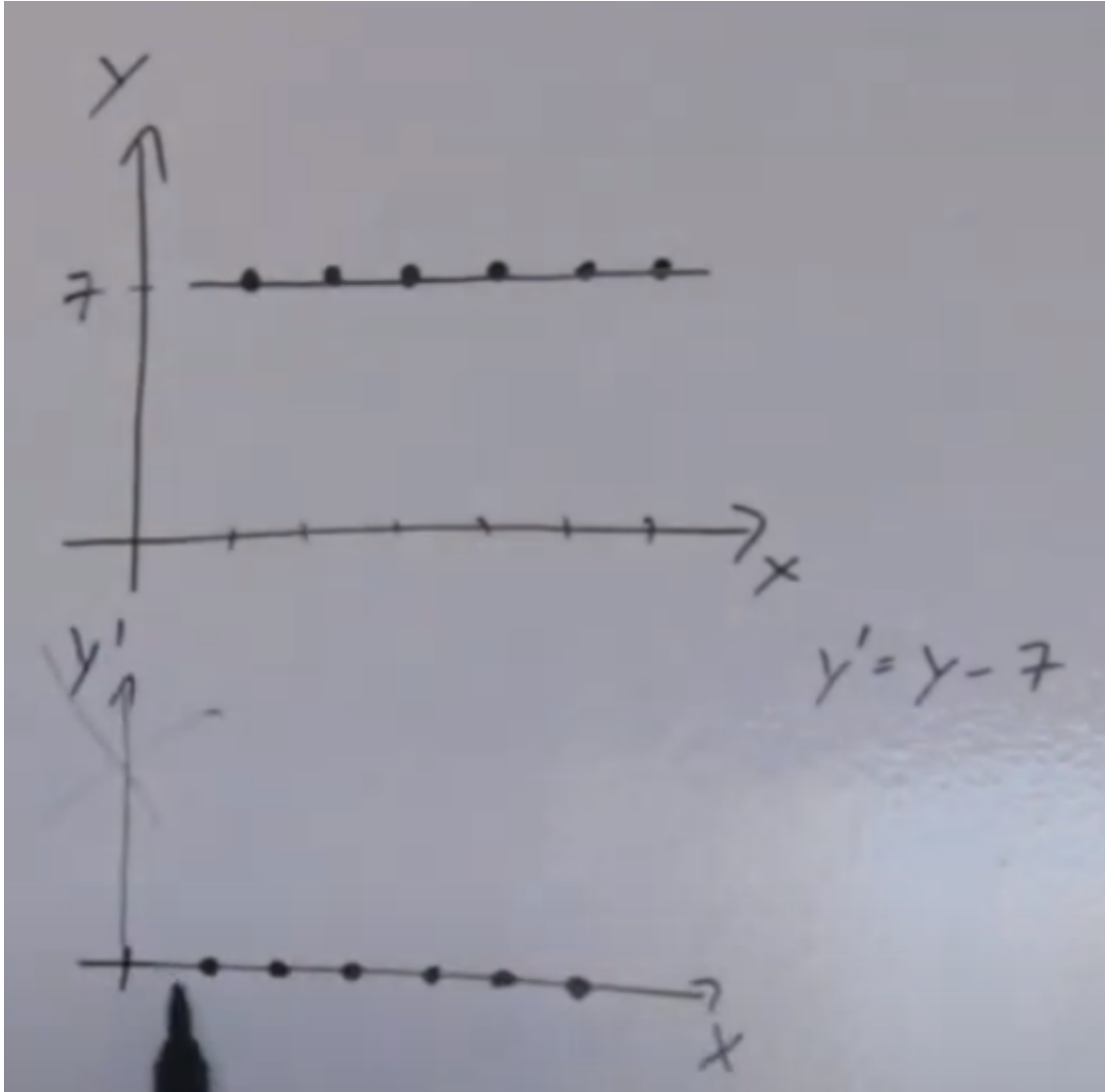


Principal Component Analysis - PCA

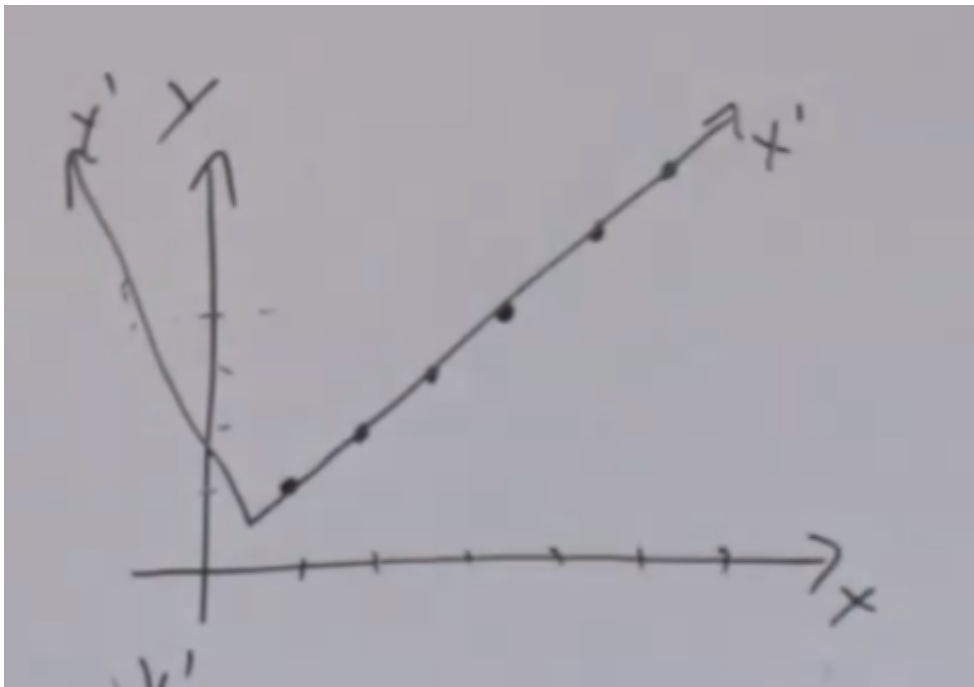
O objetivo dessa técnica é a redução de dimensionalidade, ou seja, tentar diminuir o número de características (features) do conjunto de dados. Para isso ela tenta preservar a variância dos dados.

Um exemplo disso:



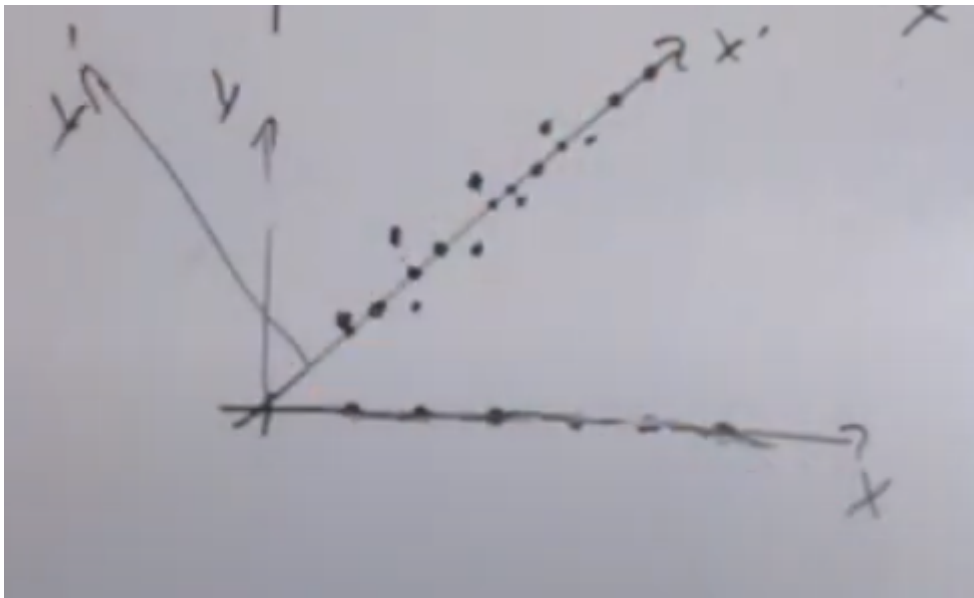
Basicamente temos o primeiro gráfico "y" que depende de y e x, sendo y fixado em 7. Podemos transpor o gráfico para depender apenas de X, com a nova função "y'", onde diminui a dimensionalidade, mantendo a variância dos dados deixando apenas somente a variável X.

Outro exemplo:



Nesse caso, temos um novo gráfico onde y e x variam, porém ao criar uma reta, podemos rotacionar o gráfico e transformar em x' e y' , mantendo assim os dados em uma única variável.

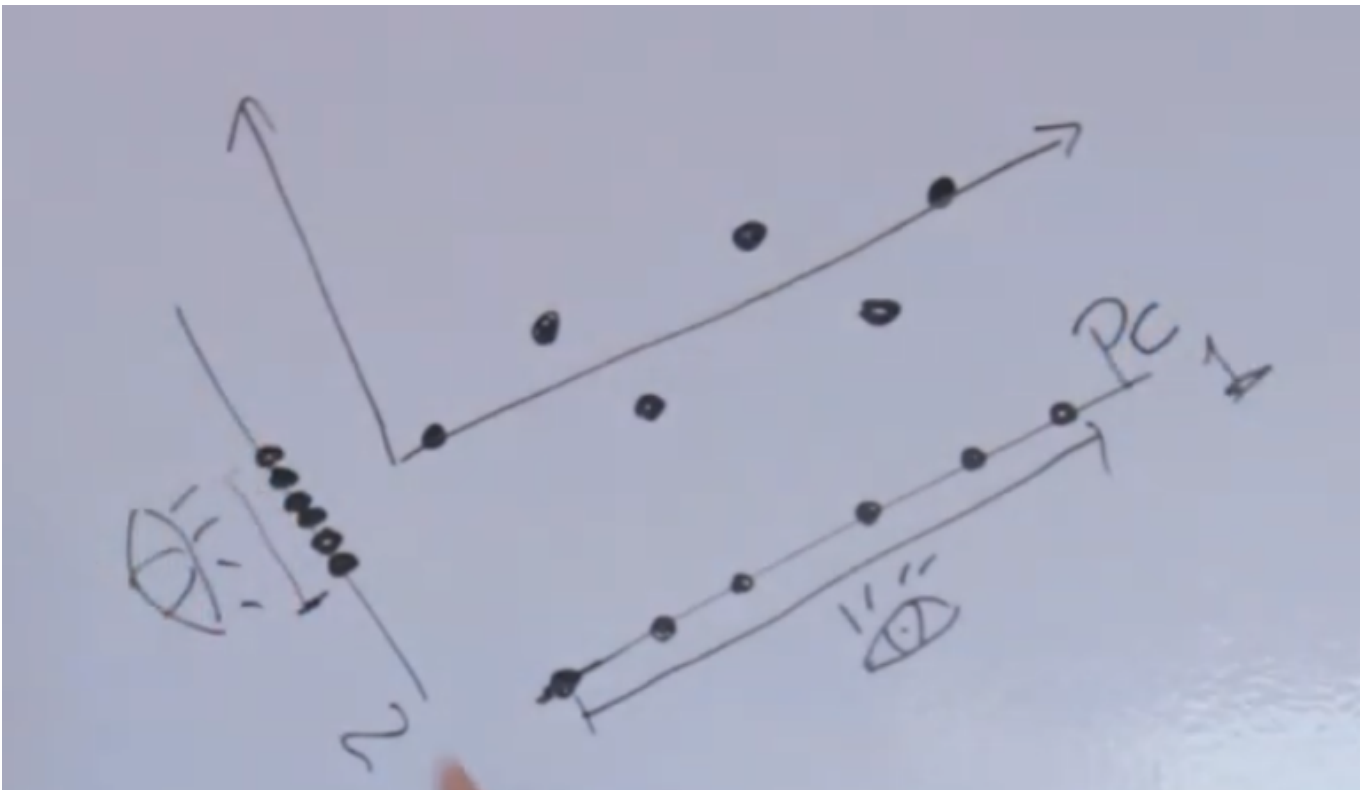
Outro exemplo utilizando o conceito de rotação, podemos ver que ao rotacionar o gráfico, a reta não consegue capturar todos os pontos, porém mesmo assim, a variável X' tem um erro muito pequeno, ou seja, a variância ainda sim é pequena retendo o máximo de informação possível:



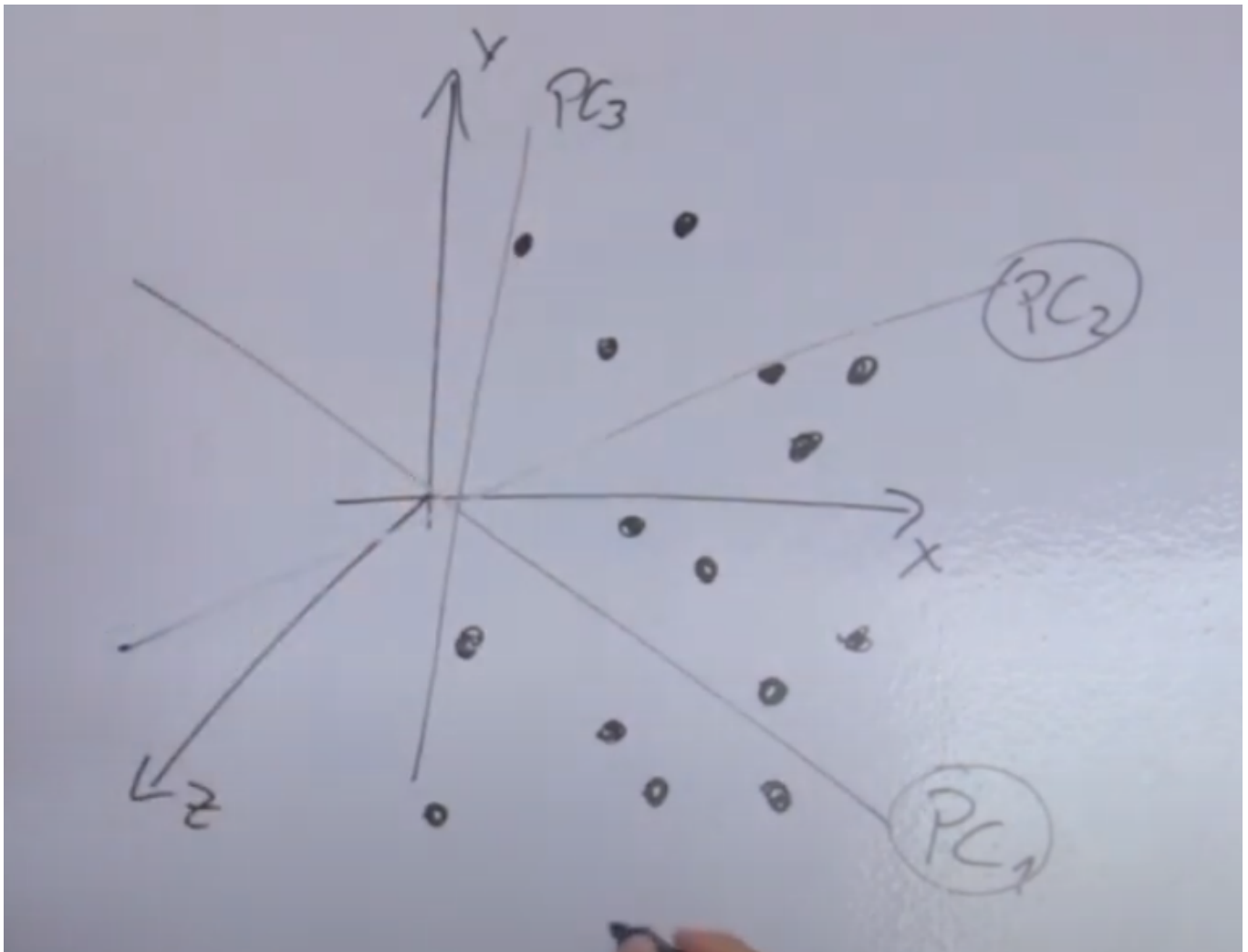
Esse conceito de variação, pode ser visualizado da seguinte forma:



Dado duas pessoas observando um conjunto de dados, a que está a esquerda irá vê-los muito juntos, não conseguindo separá-los ou classificá-los, já o da direita, consegue visualizar os dados mais espaçados, conseguindo separá-los adequadamente, capturando a maior parte da variabilidade dos dados.



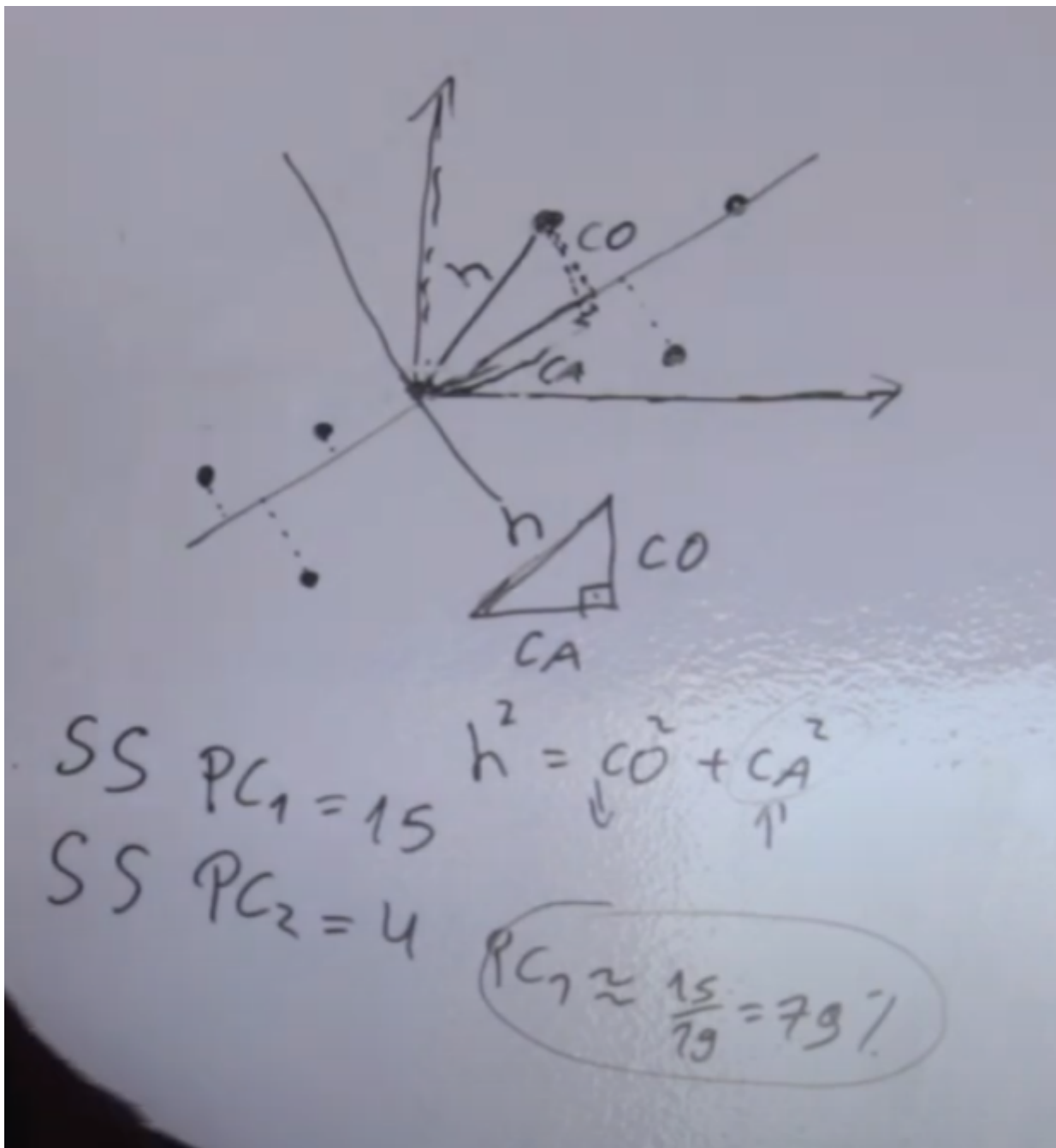
O objetivo do PCA é transformar um conjunto de dimensões em outros, mantendo a maior variabilidade dos dados possível.



Esse exemplo com 3 dimensões, temos 3 novos eixos criados PC1, PC2 e PC3 para capturar a maior variabilidade, lembrando que devem ser criados perpendiculares uns aos outros. Mas se quisermos diminuir a dimensionalidade, podemos selecionar os que capturam a maior parte, nesse caso, se escolhermos PC1 e PC2, reduzimos a dimensionalidade mantendo a maior variabilidade, já que o PC3 só captura uma pequena parte dos dados. Caso quisermos apenas uma variável, nenhum eixo X, Y, ou Z seria tão bom quanto selecionar a PC1 que captura boa parte da variância dos dados.

Como é calculado?

Dado um problema com dois eixos X, Y.



Devemos calcular a soma dos quadrados para cada eixo novo criado e tentar diminuir ao máximo os catetos opostos e aumentar os catetos adjacentes, assim o "erro" de cada ponto será pequeno comparado com o novo eixo.

Por fim é verificado quanto da variância dos dados cada PC está capturando, nesse exemplo PC1 captura 79% dos dados e pode ser escolhido para redução de dimensionalidade.

Mas por que perder variabilidade dos dados selecionando somente alguns PC's?

Basicamente o algoritmo de PCA tenta ganhar performance computacional, diminuindo o número de features, tentando manter o máximo de variabilidade possível no menor número de features, melhorando a velocidade de treinamento.