

Principal Component Analyse (Dimensionsreduktion)

Nehmen wir das Beispiel Nahrungsmittel: Kohlenhydrate, Eiweiß, Fett usw.

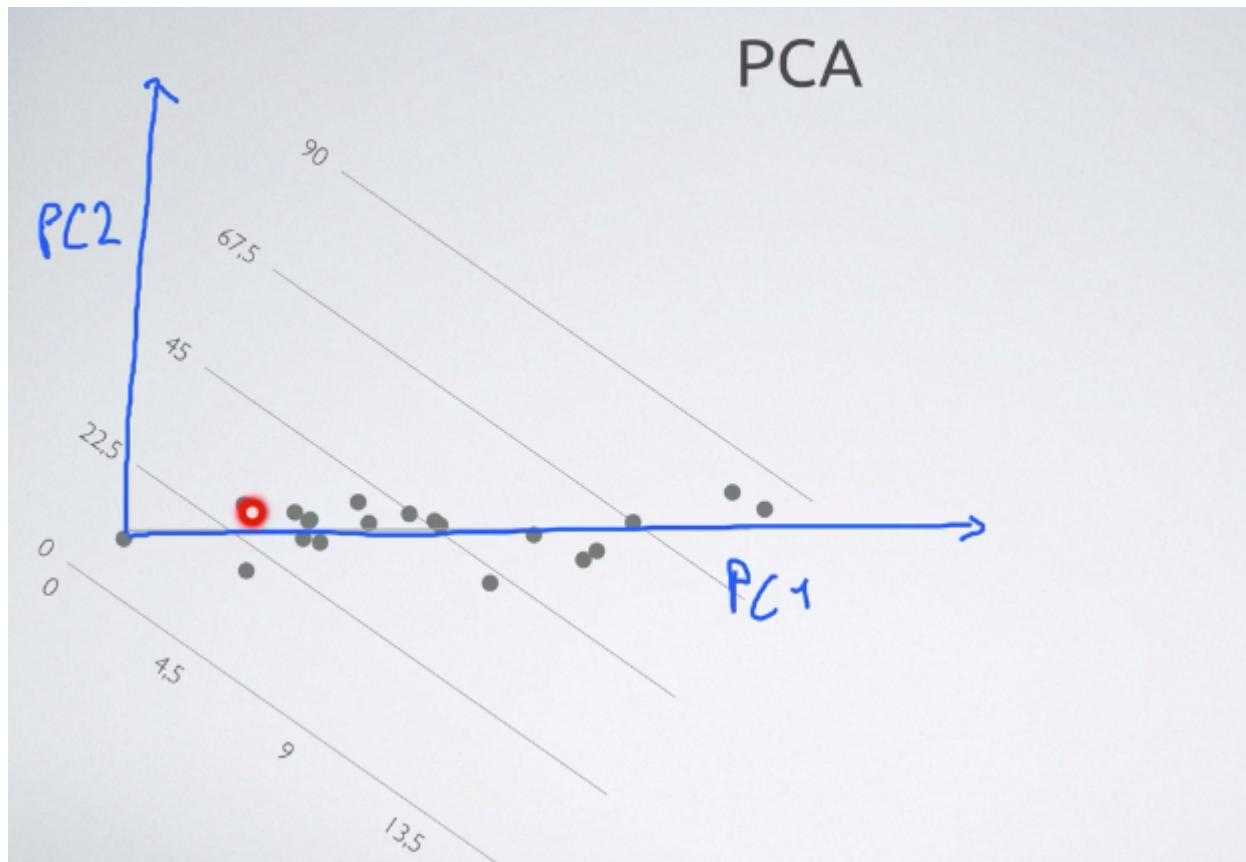
Um zu entscheiden ob man wirklich alle Eigenschaften / Dimensionen braucht, macht man eine PCA.

Brennwert wird zum Beispiel aus mehreren der anderen Variablen berechnet. D.h. herrscht eine Abhängigkeit und man kann die entsprechende Dimension weglassen.

Mathematisch: Wir haben k Dimensionen und wollen $n < k$ Dimensionen

Grundsätzlich besitzt PC1 die größte Varianz des gesamten Datensatzes. PC2 die zweitgrößte Varianz etc.

Hat man zum Beispiel eine 'normale' Regression zwischen einer X und Y Achse, dreht man einfach die Regressionsgerade zur neuen X-Achse. Somit beschreibt nun die neue X-Achse (PC1) die größte Varianz. Die alte Y-Achse gibt jetzt wieder nur noch die Höhe an, jedoch wird diese nun als PC2 beschrieben. Und wie man in dem Schaubild dann sieht, beschreibt diese Achse eine geringere Varianz (hoch / runter) als die PC1 (links / rechts)



Merksätze:

Hierbei werden neue Dimensionen gefunden, die absteigend möglichst viel Varianz erfassen

Die erste Dimension enthält die meiste Varianz, die zweite dann etwas weniger etc.

Dadurch können wir die Daten "automatisch" komprimieren

Neue Dimensionen werden automatisch gefunden

Können uns angeben lassen, wie viel Varianz enthalten geblieben ist

Achtung: Neue Dimensionen haben keine direkte Bedeutung mehr! PC's sind Kombinationen aus den Daten

PCA und PCA-Regression

Die PCA braucht im Vergleich zur PCA-Regression eine Vorgabe auf wie viele Hauptkomponenten der Datensatz reduziert werden soll.

Die PCA-Regression dagegen hilft dabei herauszufinden welche Anzahl an Hauptkomponenten am besten geeignet ist.

Josh Starmer

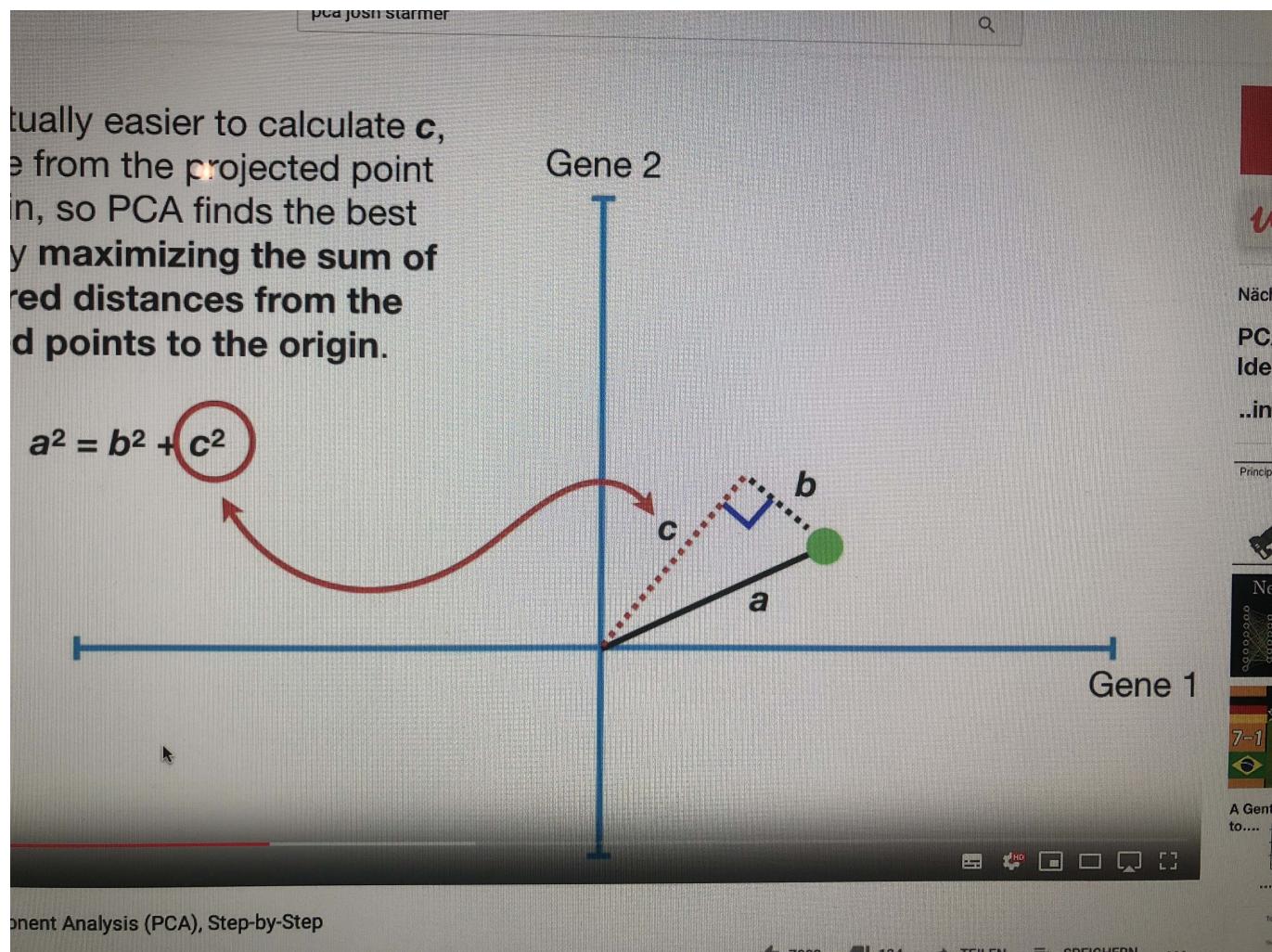
PCA mit SVD (Single Value Decomposition)

Im Grunde geht es darum mehrdimensionale Daten (8,9, 100 Dimensionen auf 2 herunterzubrechen um diese Darstellen zu können). Man bekommt außerdem Auskunft darüber welche Variable am besten für ein Clustering geeignet ist. Nachdem der durchschnittliche Abstand zur x-Achse und zur y-Achse berechnet wurde, kann man den Mittelpunkt der Daten berechnen. Dieser wird auf den 0, 0 Punkt des Graphen gelegt. (Ursprung) Relative Position der Daten wird nicht verändert!

Nun wird eine Linie durch die Daten gefittet. Um die best Mögliche Linie zu finden wird der Abstand der Punkte zu der entsprechenden Linie gesucht. Alternativ kann versucht werden den Abstand von den Vorhergesagten Punkten auf der gefitteten Gerade zum Ursprung zu maximieren.

Mathematisch passiert folgendes:

Der Abstand von einem beliebigen Punkt zum Ursprung verändert sich nicht. Jedoch entsteht zusammen mit der Linie zu der gefitteten Linie und dem Abstand daszu ein Dreieck mit einem rechten Winkel. Mit dem Satz des Pythagoras kann man somit also **b** versuchen zu minimieren.

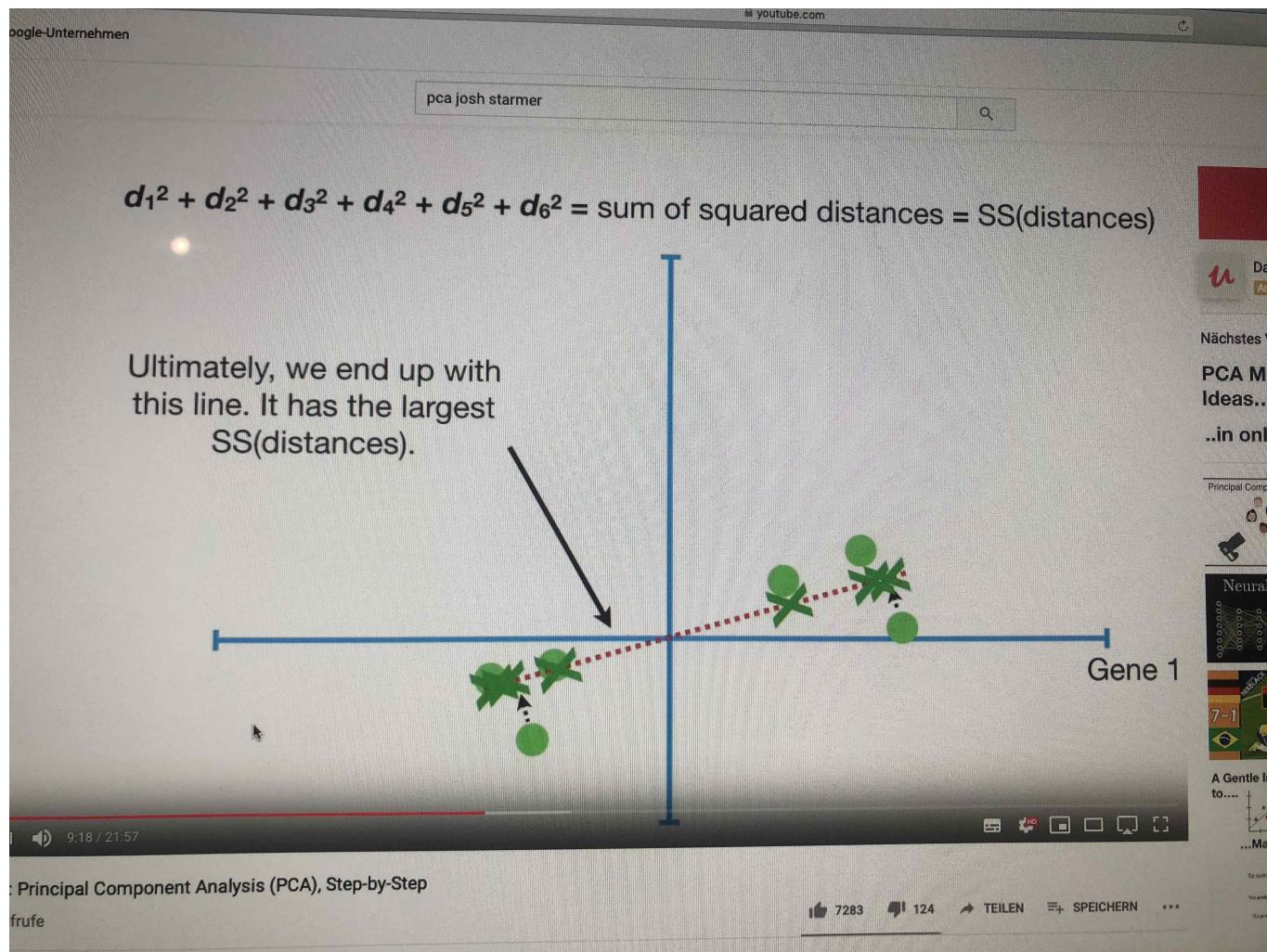


Component Analysis (PCA), Step-by-Step

7283 124 TEILEN SPEICHERN

Nun wird für jeden weiteren Punkt der Abstand zum Ursprung berechnet und versucht diesen Gesamtenabstand zu maximieren. (Zwischen dem Vorhergesagten Punkt und dem Ursprung)

Diese daraus entstehende Linie nennt man **PCA 1**. Die Steigung der entstehenden Linie gibt Auskunft darüber wie wichtig die Variable 1 im Vergleich zu Variable 2 ist. Steigung 0.25 bedeutet also, Variable 1 ist 4-mal so 'wichtig' wie Variable 2.



Dieses Verhältnis 4-mal Variable 1 und 1-mal Variable 2 nennt man **Linear Kombination**.

Verwendet man **SVD** muss man alle Werte skalieren. Das Verhältnis bleibt dabei das Gleiche, nur eben auf eine maximale Länge von 1 skaliert. Bei einem Verhältnis von ursprünglich 4 zu 1 erhält man nach der Skalierung 0.97 (Variable 1) und 0.242 (Variable 2).

Der daraus entstehende Vektor nennt man Eigenvektor (Länge 1)

PC 2

PC2 ist lediglich die senkrecht zu PC1 stehende Gerade durch den Ursprung.

Nimmt man unser oberes Beispiel erhält man folgenden Eigenvektor für PC2:

-0.242 Variable 1 und 0.97 Variable 2 - auch **Loading Scores** genannt

Um nun das finale PCA Plot zu zeichnen wird das ganze einfach rotiert bis PC1 die x-Achse ist und PC2 die y-Achse.

YouTube, ein Google-Unternehmen

pca josh starmer

For the sake of the example, imagine that the Variation for **PC1 = 15**, and the variation for **PC2 = 3**.

$\frac{\text{SS}(\text{distances for PC1})}{n - 1} = \text{Variation for PC1}$

$\frac{\text{SS}(\text{distances for PC2})}{n - 1} = \text{Variation for PC2}$

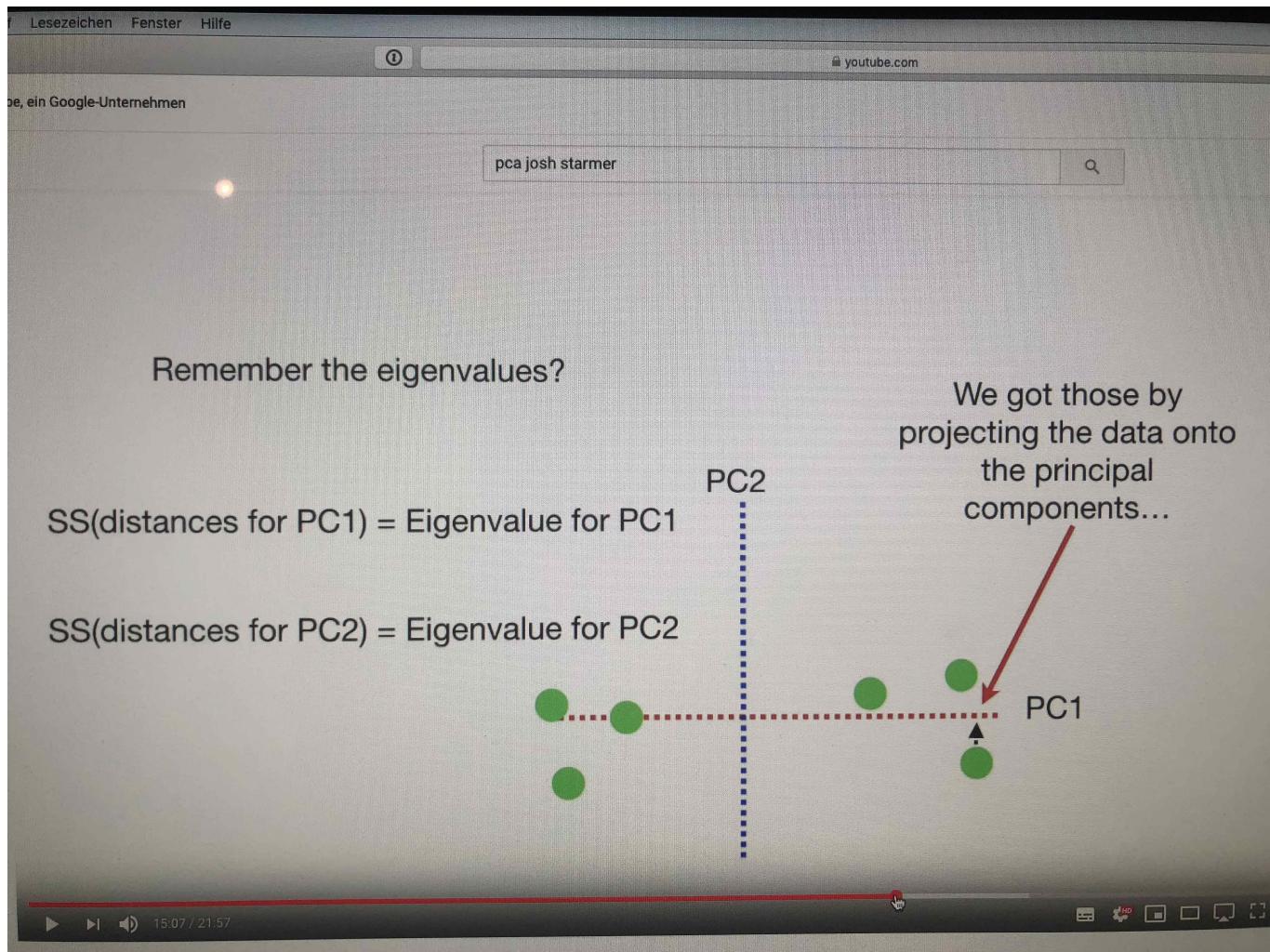
That means that the total variation around both PCs is **15 + 3 = 18...**

PC2 ...and that means PC1 accounts for $15 / 18 = 0.83 = 83\%$ of the total variation around the PCs.

StatQuest: Principal Component Analysis (PCA), Step-by-Step

344.528 Aufrufe

7283 124 TEILEN SPEICHERN ...



Hat man mehr Variablen ist die Vorgehensweise gleich, bei 3 hat man dann eben 3 Loading scores pro Achse. Die Achse mit dem höchsten Loading score auf PC1 ist dann die wichtigste **Zutat**. Wichtig ist, dass alle PC's senkrecht zueinander sind und durch den Ursprung gehen (nach der Skalierung!!!!)

Die PC1 wird immer dadurch bestimmt, dass alle Datenpunkte geplottet werden und dann skaliert werden. Durch diese Daten wird dann eine möglichst passende Linie gezogen - das ist dann PC1.

Long Story Short (Rezept)

Vorgehensweise bei PCA mit Single Value Decomposition

1. Daten plotten
2. Daten zentrieren (Mittelwert jeder Achse berechnen) und auf Ursprung setzen
3. Beste gefittete Linie durch den Ursprung zeichnen (PC1)
4. Senkrecht zur PC1 Linie durch Ursprung zeichnen (PC2)
5. Für alle weiteren Achsen durchführen
6. Werte **Loading Scores** geben Auskunft über die Wichtigkeit der Variablen für jede Achse
7. Scree-Plot zeichnen (Zeigt die Varianz pro Achse)
8. *optional* Auf 2 Achsen unterbrechen (wenn z.B. 95% der Varianz dadurch erklärt werden)