Hauptkomponentenanalyse

Einführung

Die Hauptkomponentenanalyse ist ein Verfahren um Muster in Daten zu erkennen und diese in Form von Ähnlichkeiten und Unterschieden auszudrücken. Da es insbesondere in höherdimensionalen Datensätzen schwierig ist die Muster zu entdecken und auch eine graphische Darstellung nicht möglich ist, ist die HKA ein wichtiges und mächtiges Werkzeug der Datenanalyse.

Ein wichtiger Vorteil der HKA ist die Möglichkeit nach der Erkennung der Muster die Daten zu komprimieren durch eine Reduktion der Dimensionen ohne signifikante Informationen zu verlieren.

In diesem Kapitel werden die einzelnen Schritte dargestellt, die erforderlich sind um eine HKA mit einem gegebenen Datensatz durchzuführen.

Die Methode

Schritt 1: Datensatz

In einem einfachen Beispiel wird ein 2-dimensionaler Datensatz verwendet. Damit ist eine graphische Darstellung möglich und eine Veranschaulichung der einzelnen Schritte der HKA Analyse.

Die verwendeten Daten sind in *HKA Beispiel Daten. Links die Originaldaten, rechts die zentrierten Daten* dargestellt.

Schritt 2: Zentrierung

Für die Durchführung der HKA muss der Mittelwert von jeder Daten-Dimension subtrahiert werden. Der abzuziehende Mittelwert ist der Durchschnitt in jeder Dimension. Das bedeutet, dass von allen x-Werten der Wert $\bar{x}$ (das ist das Mittel über alle x-Werte der Datenpunkte) abgezogen wird und von allen y-Werten wird der Wert $\bar{y}$ abgezogen. Dadurch wird der Datensatz um den Nullpunkt zentriert.

Originaler Datensatz	
x	**y**
2,5	2,4
0,5	0,7
2,2	2,9
1,9	2,2
3,1	3,0
2,3	2,7
2	1,6
1	1,1
1,5	1,6
1,2	0,9

Zentrierter Datensatz	
x	**y**
0,69	0,49
-1,31	-1,21
0,39	0,99
0,09	0,29
1,29	1,09
0,49	0,79
0,19	-0,31
-0,81	-0,81
-0,31	-0,31
-0,71	-1,01

Tabelle 1: HKA Beispiel Daten. Links die Originaldaten, rechts die zentrierten Daten

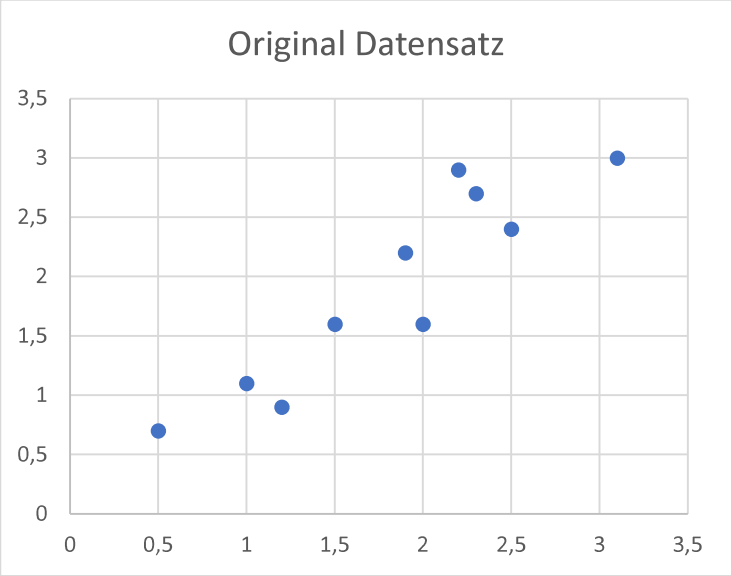

Abbildung 1: Graphische Darstellung der Originaldaten

Schritt 3: Berechnung der Kovarianz-Matrix

Dieser Schritt ist in *Grundlagen der Statistik* beschrieben. Die Kovarianz-Matrix ist vom Typ 2×2 , da die Daten 2-dimensional sind. Die Kovarianz-Matrix lautet

$$cov = \begin{pmatrix} 0{,}616555556 & 0{,}615444444 \\ 0{,}615444444 & 0{,}716555556 \end{pmatrix}$$

Da die nicht in der Hauptdiagonale stehenden Elemente in dieser Kovarianz-Matrix positiv sind, kann erwartet werden, dass die Variablen x und y gleichsinnig anwachsen.

Schritt 4: Berechnung der Eigenvektoren und Eigenwerte der Kovarianz-Matrix

Da die Kovarianz-Matrix quadratisch ist, können die Eigenvektoren und Eigenwerte dieser Matrix berechnet werden. Diese sind wichtig, da sie, wie weiter unten ausgeführt wird, wichtige Informationen über die Daten wiedergeben. Die Eigenvektoren und Eigenwerte sind:

$$Eigenwerte = \begin{pmatrix} 0{,}0490833989 \\ 1{,}28402771 \end{pmatrix}$$

$$Eigenvektoren = \begin{pmatrix} -0{,}735178656 & -0{,}677873399 \\ 0{,}677873399 & -0{,}735178656 \end{pmatrix}$$

Hier ist wicht zu vermerken, dass beide Eigenvektoren Einheitseigenvektoren sind, dass also ihre Länge jeweils 1 beträgt. Dies ist für die HKA wichtig, und die meisten Software-Pakete, mit denen eine Eigenvektor-Berechnung durchgeführt werden kann, liefern auch Einheitseigenvektoren.

In *Graphik des zentrierten Datensatzes mit den Eigenvektoren der Kovarianz-Matrix gestrichelt* ist zu sehen, dass die Daten ein deutliches Muster aufweisen. Und wie aus der Kovarianz-Matrix zu erwarten ist, wachsen die beiden Variablen zusammen an. In dieser Graphik sind auch die Eigenvektoren eingetragen. Es sind die gestrichelten Linien. Wie im Abschnitt über Eigenvektoren aufgeführt, stehen sie senkrecht aufeinander. Und sie liefern wichtige Informationen über die Muster in den Daten. Einer der Eigenvektoren geht mittig durch die Punkte hindurch, wie eine optimal angepasste Gerade. Dieser Eigenvektor zeigt wie sich diese zwei Datensätze in Bezug zu dieser Linie verhalten. Der zweite Eigenvektor liefert das andere, weniger wichtige, Muster in den Daten, dass alle Punkte der Hauptlinie mit einer gewissen Abweichung folgen.

Durch den Prozess der Bestimmung der Eigenvektoren der Kovarianz-Matrix können Linien bzw. Achsen ermittelt werden, die die Daten charakterisieren. Die weiteren Schritte behandeln eine Transformation der Daten, so dass diese abhängig von den gefundenen Achsen dargestellt werden können.

Schritt 5: Auswahl der Komponenten und Aufstellung des Merkmalsvektors

Nun erfolgt die Datenkompression und Reduzierung der Dimensionalität. Der Blick auf die im vorigen Abschnitt ermittelten Eigenvektoren und Eigenwerte zeigt, dass die Eigenvektoren recht unterschiedliche Werte haben. Tatsächlich zeigt sich, dass der Eigenvektor mit dem *größten* Eigenwert die *Hauptkomponente* des Datensatzes darstellt. Im Beispiel geht der Eigenvektor mit dem größten Eigenwert durch die Mitte der Daten. Dies ist die wichtigste Beziehung zwischen den Dimensionen der Daten.

Im Allgemeinen wird nach der Berechnung der Eigenvektoren aus der Kovarianz-Matrix eine Reihenfolge der Eigenvektoren nach den zugehörigen Eigenwerten gebildet, vom größten zum kleinsten. Damit sind die Komponenten in eine Bedeutungs-Reihenfolge gebracht. Nun können die Komponenten mit geringerer Bedeutung *vernachlässigt* werden. Dabei geht Information verloren, aber wenn die Eigenwerte klein sind, ist der Informationsverlust gering. Wenn Komponenten weggelassen werden, so hat der neue Datensatz weniger Dimensionen als der Original-Datensatz. Wenn ursprünglich *n* Dimensionen in den Daten

vorhanden waren und *n* Eigenvektoren und Eigenwerte berechnet wurden und dann nur die ersten *p* Eigenvektoren ausgewählt werden, dann hat der neue Datensatz nur *p* Dimensionen.

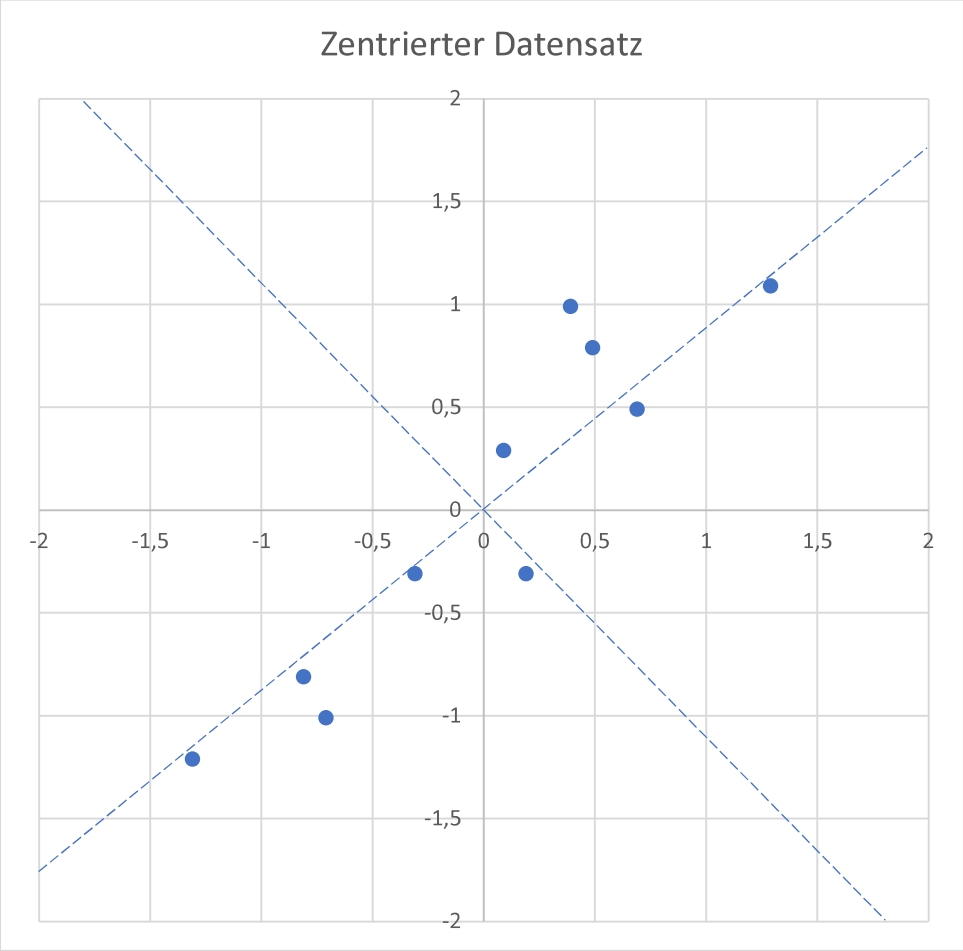

Abbildung 2: Graphik des zentrierten Datensatzes mit den Eigenvektoren der Kovarianz-Matrix gestrichelt

Nun muss ein *Merkmalsvektor* ermittelt werden, wobei es sich um eine Matrix von Vektoren handelt. Diese entsteht aus den ausgewählten Eigenvektoren, die in einer Matrix als Spalten angeordnet werden.

$$Merkmalsvektor = (eig_1\ eig_2\ eig_3 \dots eig_p)$$

Im Beispiel mit zwei Eigenvektoren gibt es zwei Möglichkeiten. Einmal kann ein Merkmalsvektor mit beiden Eigenvektoren gebildet werden

$$\begin{pmatrix} -0{,}677873399 & -0{,}735178656 \\ -0{,}735178656 & 0{,}677873399 \end{pmatrix}$$

oder es kann die kleinere, weniger signifikante Komponente weggelassen werden und nur eine Spalte gebildet werden

$$\begin{pmatrix} -0{,}677873399 \\ -0{,}735178656 \end{pmatrix}$$

Die Auswirkungen dieser beiden Varianten werden im nächsten Abschnitt dargestellt.

Schritt 5: Aufstellung des neuen Datensatzes

Schließlich folgt der vergleichsweise einfache letzte Schritt der HKA. Nach dem die Komponenten (Eigenvektoren), die beibehalten werden sollen, ausgewählt wurden und die Merkmalsvektoren gebildet wurden, wird der transponierte Vektor auf der linken Seite mit dem originalen transponierten Datensatz multipliziert.

$$Finale\ Daten = Zeilen\ Merkmals - Vektor\ \times Zeilen\ Datensatz\ Angepasst$$

Hier stellt der $Zeilen\ Merkmals - Vektor$ die Eigenvektoren aus den Spalten transponiert dar, so dass die Eigenvektoren nun in Zeilen organisiert sind, wobei die signifikantesten Eigenvektoren ganz oben stehen. Der $Zeilen\ Datensatz\ Angepasst$ stellt die zentrierten Daten transponiert dar, d.h., die Daten sind jeweils in einer Spalte vorhanden und jede Zeile steht für eine separate Dimension.

Der Vorteil der Verwendung von transponierten Vektoren und des transponierten Datensatzes ist, dass der finale Datensatz gleich die Daten in Spalten und die Dimensionen in Zeilen enthält.

Was ist jetzt das Ergebnis? Die Originaldaten stehen jetzt in Abhängigkeit von den gewählten Vektoren zur Verfügung. Der Originaldatensatz des Beispiels hatte zwei Achsen, $x \text{ und } y$, und die Daten waren in diesen Koordinaten ausgedrückt. Die Daten können abhängig von beliebig gewählten Achsen bzw. Koordinaten ausgedrückt werden. Am effizientesten können die Daten in Bezug auf senkrecht stehende Achsen ausgedrückt werden. Daher war (und ist) es wichtig, dass die Eigenvektoren immer senkrecht aufeinander stehen. Die Daten werden jetzt nicht mehr als Koordinaten der Achsen $x \text{ und } y$ ausgedrückt, sondern abhängig von 2 Eigenvektoren. Für den Fall, dass der neue Datensatz eine reduzierte Dimensionalität aufweist, wurden Eigenvektoren weggelassen und der neue Datensatz ist nur noch eine Funktion der ausgewählten und beibehaltenen Eigenvektoren.

Diese finale Transformation wird nun mit den beiden möglichen Merkmalsvektoren des Beispiels gemacht. Die transponierte Form der Ergebnisse findet sich in Form einer Tabelle wieder. In einer Graphik wird deutlich, wie sich die finalen Punkte in Bezug auf die Komponenten anordnen.

Wenn beide Eigenvektoren für die Transformation beibehalten werden, so entstehen die Daten und die Graphik wie in *Tabelle 2: Tabelle mit den Daten nach Durchführung der HKA mit zwei Eigenvektoren* und *Abbildung 3: Graphik nach Durchführung der HKA mit zwei Eigenvektoren mit der neuen Datenpunkte.* dargestellt. Die Graphik enthält die Originaldaten, die so rotiert wurden, dass die beiden Eigenvektoren die neuen Achsen darstellen, da in diesem Beispiel kein Informationsverlust durch Weglassen von Eigenvektoren entstanden ist.

Transformierter Datensatz	
x	**y**
-0,82797	-0,175115
1,7775803	0,1428572
-0,992197	0,384375
-0,27421	0,1304172
-1,675801	-0,209498
-0,912949	0,1752824
0,0991094	-0,349825
1,1445722	0,0464173
0,4380461	0,0177646
1,2238206	-1,626753

Tabelle 2: Tabelle mit den Daten nach Durchführung der HKA mit zwei Eigenvektoren

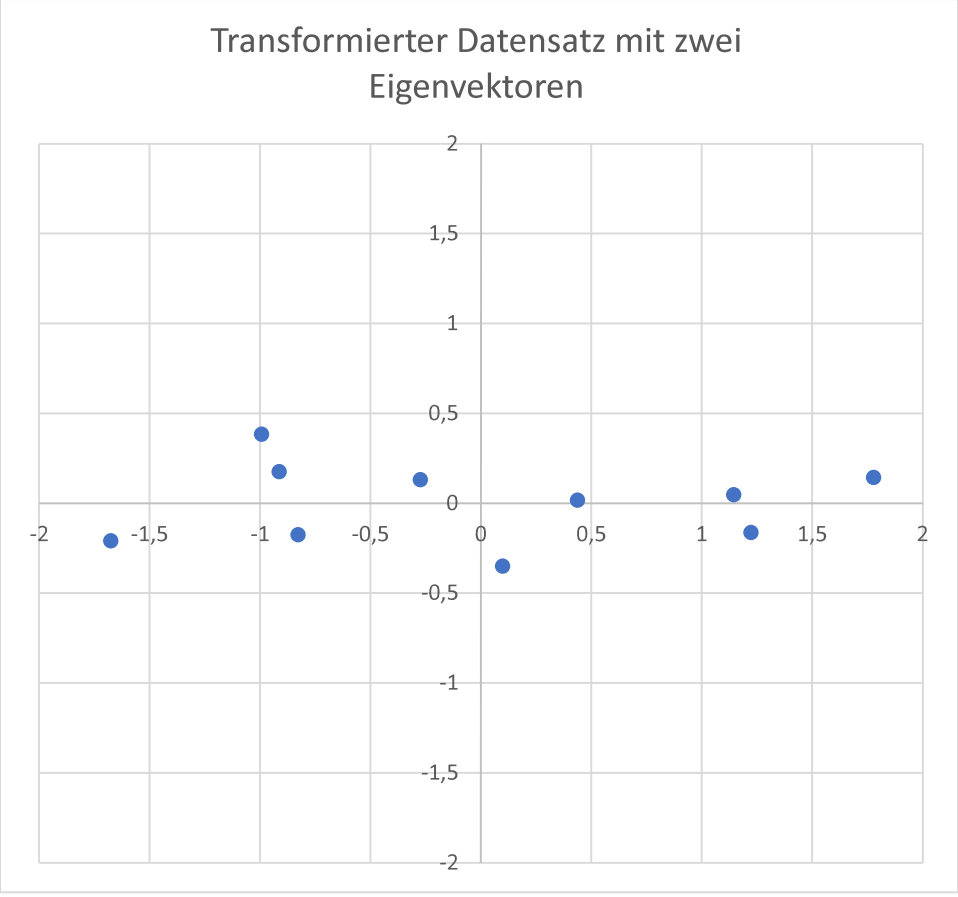

Abbildung 3: Graphik nach Durchführung der HKA mit zwei Eigenvektoren mit der neuen Datenpunkte.

Wird nur der Eigenvektor mit dem größten Eigenwert verwendet, so ergibt sich eine andere Transformation. Die sich dann ergebenden Daten sind in der Tabelle in *Tabelle 3: Transformierte Daten nur unter Verwendung des signifikantesten Eigenvektors* dargestellt. Erwartungsgemäß bleibt nur eine Dimension übrig. Vergleicht man diesen Datensatz mit dem Datensatz für den beide Eigenvektoren verwendet wurden, so ist ersichtlich, dass dieser Datensatz genau der ersten Spalte des anderen Datensatzes entspricht. Wenn diese Daten graphisch dargestellt werden, so sind sie 1-dimensional und sind Punkte auf einer Linie genau an den x-Positionen der Punkte in der Graphik aus *Tabelle 2: Tabelle mit den Daten nach Durchführung der HKA mit zwei Eigenvektoren* und *Abbildung 3: Graphik nach Durchführung der HKA mit zwei Eigenvektoren mit der neuen Datenpunkte.*. Eine komplette Achse, die durch den anderen Eigenvektor dargestellt wird, ist so entfernt worden.

Transformierter Datensatz (Ein Eigenvektor)
x
-0,82797
1,7775803
-0,992197
-0,27421
-1,675801
-0,912949
0,0991094
1,1445722
0,4380461
1,2238206

Tabelle 3: Transformierte Daten nur unter Verwendung des signifikantesten Eigenvektors

Was ist jetzt hier gemacht worden? Im Wesentlichen wurden die Daten so transformiert, dass sie in Abhängigkeit von Mustern der Daten untereinander ausgedrückt werden, wobei diese Muster diejenigen Linien darstellen, die die Beziehungen der Daten untereinander am besten beschreiben. Dadurch wird jeder Datenpunkt jetzt klassifiziert über eine Kombination der Beiträge verschiedener Linien. Der Ausgangspunkt waren die x und y Achsen. Die x und y Werte jedes Datenpunktes geben keine Aussage über die Beziehung eines Datenpunktes zum Rest des Datensatzes. Nach der Transformation kann aus den Datenpunkten genau die Lage bezüglich der Trendlinien (z.B. darüber oder darunter) abgelesen werden. Für den Fall, dass *beide* Eigenvektoren beibehalten werden, werden die Daten so verändert, dass sie in Bezug auf diese Eigenvektoren ausgedrückt werden können anstatt in Bezug zu den üblichen Achsen. Für den Fall der

Zerlegung in nur einen Eigenvektor wird der Beitrag des kleineren Eigenvektors vernachlässigt und die Daten werden ausschließlich abhängig von dem anderen Eigenvektor ausgedrückt.

Wiederherstellung der Originaldaten

Wenn die HKA Transformation verwendet wird um Daten zu komprimieren, spielt es eine große Rolle zu ermitteln, wie die Originaldaten zurückgewonnen werden können.

Wie können die Originaldaten zurückgewonnen werden? Hier ist es wichtig daran zu erinnern, dass nur wenn *alle* Eigenvektoren in der Transformation mitgeführt werden *alle* Originaldaten zurückerhalten werden können. Wenn die Anzahl der Eigenvektoren in der finalen Transformation reduziert wurde, dann werden die wiedergewonnenen Daten einen Informationsverlust aufweisen.

Die finale Transformation war definiert als:

$$Finale\ Daten = Zeilen\ Merkmals - Vektor\ \times Zeilen\ Datensatz\ Angepasst$$

Diese Gleichung kann umgestellt werden, so dass die Originaldaten wiedergewonnen werden,

$$Zeilen\ Datensatz\ Angepasst = Zeilen\ Merkmals\ Vektor^{-1}\ \times Finale\ Daten$$

wobei der $Zeilen\ Merkmals\ Vektor^{-1}$ den inversen $Zeilen\ Merkmals - Vektor$ darstellt. Wenn *alle* Eigenvektoren im Merkmalsvektor mitgeführt werden, dann ergibt sich, dass der invertierte Merkmalsvektor gleich dem transponierten Merkmalsvektor ist. Dies ist deswegen der Fall, da die Elemente der Matrix alle Einheitseigenvektoren des Datensatzes repräsentieren. Dies vereinfacht die Wiederherstellung der Originaldaten, da sich als Gleichung nun ergibt

$$Zeilen\ Datensatz\ Angepasst = Zeilen\ Merkmals\ Vektor^{T}\ \times Finale\ Daten$$

Um nun die tatsächlichen Originaldaten wieder zu erhalten, muss der Mittelwert der Originaldaten (der ja anfänglich subtrahiert wurde) wieder addiert werden. Damit ergibt sich insgesamt:

$$Zeilen\ Original\ Datensatz = (Zeilen\ Merkmals\ Vektor^{T}\ \times Finale\ Daten) + Mittelwert\ Originaldaten$$

Diese Formel trifft auch zu, wenn nicht alle Eigenvektoren im Merkmalsvektor enthalten sind. D.h., auch wenn Eigenvektoren vernachlässigt bzw. entfernt wurden, so gibt diese Formel die korrekte (Rück-) Transformation wieder.

Die Wiedergewinnung der Originaldaten mit dem *kompletten* Merkmalsvektor des Beispiels führt wieder auf die Ausgangsdaten. Die Verwendung des reduzierten Merkmalsvektors zeigt den Informationsverlust auf. Dies ist in *Rekonstruktion der Daten aus nur einem Eigenvektor* dargestellt. Vergleicht man diese Darstellung mit den Originaldaten aus *Abbildung 1: Graphische Darstellung der Originaldaten* so ist ersichtlich, dass die Datenstreuung entlang des ersten Eigenvektors wieder auftritt, dass aber die Streuung entlang der anderen Komponente, das ist der vernachlässigte Eigenvektor, verlorengegangen ist.

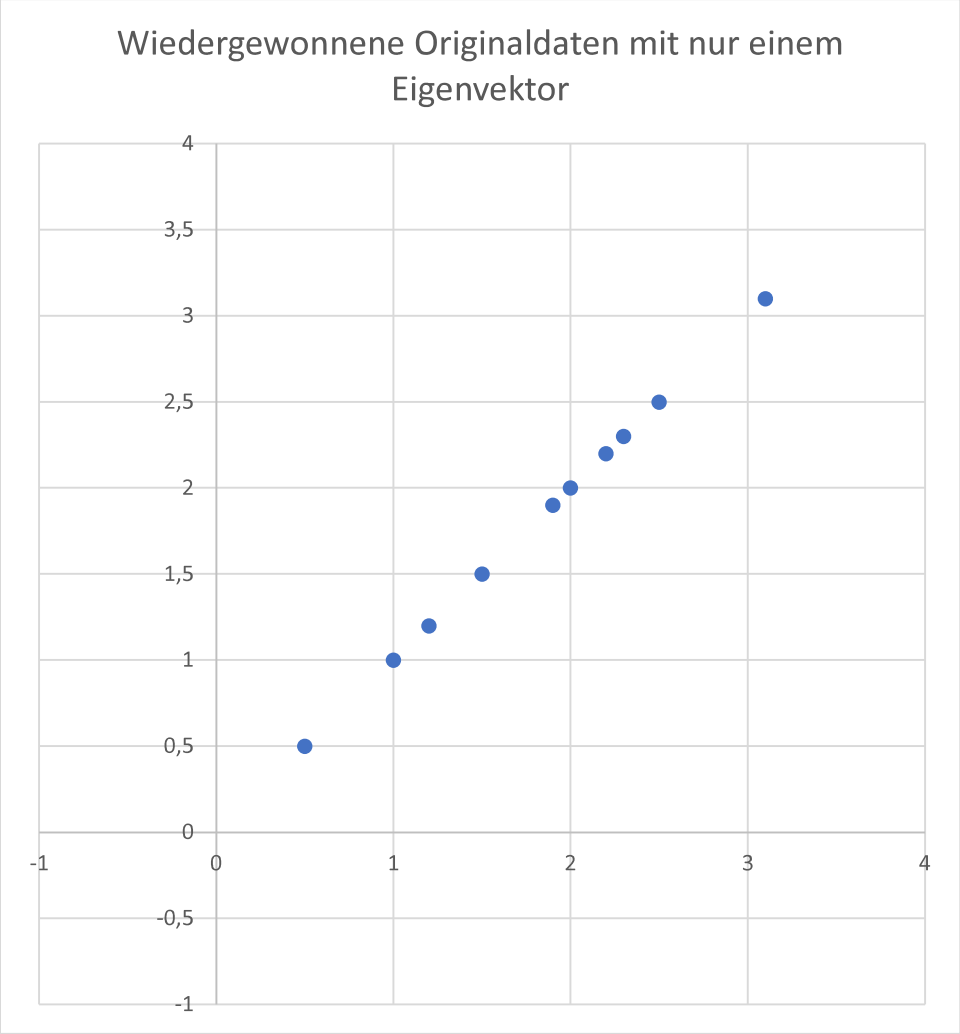

Abbildung 4: Rekonstruktion der Daten aus nur einem Eigenvektor