Principal Component Analysis

Moritz Sperrfechter[1], Maximilian Nutz[2]

[1] cas349062@cas.dhbw.de,
[2] cas348521@cas.dhbw.de

Abstract: Die Principal Component Analysis (PCA) stellt eine wichtige, parameterlose Methode zur Dimensionsreduktion dar. In dieser Arbeit wird einerseits die Funktionsweise dieser Methode anschaulich vorgestellt. Andererseits wird, in Verbindung mit dem begleitendem Jupyter Notebook, gezeigt, wie mithilfe der PCA eine Gesichtserkennung realisiert werden kann. Weiterhin wird die Kompression von Bilddaten durch die PCA in der Theorie beschrieben und in dem Jupyter Notebook implementiert.

1 Einleitung

Ein bekanntes Problem im Bereich des Machine Learnings (ML) und von AI ist die hohe Dimensionalität des zu analysierenden Datensatzes. Die Principal Component Analysis (PCA) ist eine Methode, um vor dem Training die Dimensionalität des Datensatzes und somit dessen Komplexität zu reduzieren [1]. Dazu wird der Datensatz auf lineare Abhängigkeiten zwischen einzelnen Dimensionen untersucht und anschließend in einen Raum mit weniger Dimensionen transformiert [2], [3]. Wie in [4] gezeigt, kann durch diese Dimensionsreduktion die Trainingszeit erheblich verkürzt werden. [5] zeigt, wie durch die Verwendung der PCA vor dem Training die Modellgenauigkeit verbessert wird. Weiter kann eine Reduktion der Dimensionalität, wie in Kapitel 5 gezeigt, zur Bild-Kompression genutzt werden [3].
Neben der PCA werden in der Literatur noch weitere Methoden zur Dimensionsreduktion vorgestellt. So vergleicht z.B. [6] die PCA mit der Independent Component Analysis (ICA). In dieser Arbeit wird jedoch ausschließlich auf die PCA, deren Funktionsweise und zwei Beispielanwendungen eingegangen.
Zur mathematischen und anschaulichen Erklärung der Funktionsweise, werden zu Beginn der Arbeit die mathematischen Grundlagen erläutert, die für die Durchführung der PCA notwendig sind. Anschließend sind die einzelnen Schritte der PCA an einfachen zweidimensionalen Daten visualisiert und erläutert. Auf den mathematischen Beweis für die Funktion der PCA durch

diese Schritte wird jedoch verzichtet. In Kapitel 4 werden anschließend einige Probleme der PCA identifiziert und Vermeidungsstrategien für diese gegeben. Kapitel 5 führt verschiedene Anwendungen der PCA, unter anderem die implementierten Beispiele zur Gesichtserkennung und zur Datenkompression, auf.
Begleitend zu dieser Arbeit ist ein Jupyter Notebook erhältlich (verfügbar unter https://github.com/MoritzMSP/DHBW_AML), welches die einzelnen Beispiele dieser Arbeit implementiert. Die Überschriften in dem Jupyter Notebook stimmen dabei mit denen dieser Arbeit überein.

2 Mathematische Grundlagen

Dieses Kapitel erläutert die notwendigen mathematischen Grundlagen zur Durchführung einer PCA. Dazu werden im ersten Schritt die Bedeutung sowie die Berechnung der Standardabweichung und der Varianz aufgezeigt. Im zweiten Schritt wird vorgestellt, wie die Eigenvektoren und Eigenwerte einer gegebenen Matrix berechnet werden können.

2.1 Standardabweichung und Varianz

In der deskriptiven Statistik werden die Standardabweichung sowie die Varianz als Maß für die Streuung von Werten innerhalb einer Stichprobe verwendet [7]. Die Varianz gibt dabei an, wie weit die Werte der Stichprobe vom Mittelwert dieser entfernt sind. Je größer die Varianz, je größer ist die Streuung der Werte und je weiter sind sie vom Mittelwert entfernt (vgl. Jupyter Notebook).
Ist x_i ein Wert der Stichprobe X mit n Werten und dem Mittelwert $\bar{x}$, so wird die Varianz s^2 mit Gleichung (2.1) berechnet. Die Standardabweichung s entspricht der Wurzel der Varianz (vgl. Gleichung (2.2)) [7].

$$s^2(X) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2.1}$$

$$s(X) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.2}$$

Zur Berechnung der totalen Varianz T eines p-dimensionalen Datensatzes, werden die einzelnen Varianzen der p Stichproben bzw. Variablen X_i des Datensatzes addiert (vgl. Gleichung (2.3)).

$$T = \sum_{i=1}^{p} s^2(X_i) \tag{2.3}$$

Soll der lineare Zusammenhang zwischen zwei Stichproben X und Y untersucht werden, ist dies durch die Berechnung der Kovarianz möglich (vgl. Gleichung (2.4)) [1]. Dabei ist es unerheblich, ob die Kovarianz zwischen X und Y oder zwischen Y und X bestimmt wird. Beide Fälle führen zu demselben Ergebnis. Zu beachten ist jedoch, dass die Anzahl der Werte in beiden Stichproben identisch sein muss. Wird die Kovarianz zwischen einer Stichprobe und sich selbst berechnet, so ergibt sich die Varianz der Stichprobe (vgl. Gleichung (2.5)).

$$cov(X,Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = cov(Y,X) \tag{2.4}$$

$$cov(X,X) = s^2(X) \tag{2.5}$$

Je höher die Kovarianz, je größer ist der Zusammenhang zwischen den Stichproben X und Y. Eine negative Kovarianz deutet auf einen inversen Zusammenhang hin. Beträgt die Kovarianz Null, besteht kein linearer Zusammenhang zwischen den Stichproben [8]. Die standardisierte Kovarianz wird Korrelation r genannt. Zur Berechnung dieser wird die Kovarianz durch die Standardabweichung der einzelnen Stichproben dividiert (vgl. Gleichung (2.6)). Dadurch besitzt die Korrelation im Gegensatz zur Kovarianz einen begrenzten Wertebereich von -1 bis 1.

$$r_{X,Y} = \frac{cov(X,Y)}{s(X) \cdot s(Y)} \tag{2.6}$$

2.2 Eigenvektoren und Eigenwerte

Der Eigenvektor $\vec{v}$ einer quadratischen Matrix M der Größe $p \times p$ ist definiert als derjenige Vektor der durch die Multiplikation der Matrix M ausschließlich skaliert und nicht gedreht wird. Der Eigenwert λ entspricht dabei dem Faktor, um welchen der Eigenvektor durch die Multiplikation skaliert wird (vgl. Gleichung (2.7)) [7].

$$M\vec{v} = \lambda\vec{v} \tag{2.7}$$

Durch Umformen der Gleichung (2.7) und durch die Voraussetzung, dass der Eigenvektor nicht dem Nullvektor entspricht, ergibt sich, dass die Determinante der Matrix $M - \lambda E$ Null entsprechen muss (vgl. Gleichung (2.8)). Durch ein Lösen der Gleichung (2.8) werden p Eigenwerte λ erhalten. Dabei ist es auch möglich mehrfache Eigenwerte $\lambda_i = \lambda_{i+1} = \cdots = \lambda_{i+l}$ zu erhalten.

$$\det(M - \lambda E) = 0 \tag{2.8}$$

Zur Berechnung eines Eigenvektors $\vec{v}$ muss nun die Gleichung (2.9) gelöst werden, die sich durch Umformen aus der Gleichung (2.7) ableiten lässt.

$$(M - \lambda E)\vec{v} = \vec{0} \tag{2.9}$$

Um alle Eigenvektoren der Matrix M zu erhalten, muss die Gleichung (2.9) mit jedem der zuvor berechneten Eigenwerte gelöst werden. Folglich werden p Eigenvektoren erhalten (sofern keine mehrfachen Eigenwerte berechnet wurden). Nach [7] sind die berechneten Eigenvektoren orthogonal zueinander. Dies ist jedoch nur der Fall, wenn die Matrix M reell und symmetrisch ist (Beweis hierfür im Anhang).

3 Principal Component Analysis und Durchführung dieser

Ziel der PCA ist es Abhängigkeiten zwischen den p Dimensionen eines Datensatzes zu finden und diesen in einen Datensatz in $q \leq p$ Dimensionen zu transformieren. Dazu werden zunächst die Dimensionen gesucht, in denen die am Daten am stärksten variieren, d.h. die Varianz am höchsten ist. Diese Dimensionen werden als Principal Components (Hauptkomponenten) bezeichnet [2]. Im Anschluss sind die Hauptkomponenten nach ihrem Beitrag zur totalen Varianz zu ordnen. Werden die Hauptkomponenten mit dem geringsten Beitrag zur totalen Varianz vernachlässigt, verringert sich die Dimensionalität des Datensatzes. Die PCA ist dabei unabhängig von den Labels der Daten und parameterlos [1].
In den folgenden Kapiteln werden die einzelnen Schritte zur Durchführung einer PCA vorgestellt. Diese umfassen die Standardisierung des Datensatzes, die Berechnung der Kovarianzmatrix, die Berechnung der Eigenvektoren dieser, die Bildung eines Feature Vektors und die Transformation des Datensatzes [7]. Wie zu Beginn dieser Arbeit angekündigt, wird der Beweis für dieses Vorgehen nicht hergeleitet. Ist dies dennoch von Interesse, führen [1] und [9] einen solchen Beweis auf.

Nachfolgend wird ein Datensatz M der Größe $n \times p$ betrachtet, wobei p der Anzahl der Dimensionen (Variablen) und n der Anzahl an Datenpunkten einer Variablen entspricht. Eine Variable selbst wird mit X bezeichnet, während ein Wert einer Variablen mit x bezeichnet wird.

3.1 Standardisierung des Datensatzes

Im ersten Schritt empfiehlt es sich den zugrundeliegenden Datensatz M zu standardisieren [8]. Dazu werden die Variablen des Datensatzes zunächst zentriert, d.h. der Mittelwert der Variablen wird von den Werten dieser abgezogen (vgl. Gleichung (3.1)) [7], [3]. Dies führt dazu, dass der Mittelwert der Variablen Null ist (vgl. Abbildung 1, standardisierte Daten). Daraus folgt wiederum eine Vereinfachung der späteren Berechnung der Kovarianz und der Standardabweichung.

$$x_i' = x_i - \bar{x} \text{ [4]} \tag{3.1}$$

Der Abzug des Mittelwerts verschiebt eine Variable lediglich und verändert nicht die Varianz von dieser. Nach [3] und [7] wird nach dieser Zentrierung der Daten zum nächsten Schritt übergegangen.
[2] und [8] zeigen jedoch, dass falls die einzelnen Variablen unterschiedlich skaliert sind, dieser Skalierungsfaktor die Varianz beeinflusst (Beweis hierfür im Anhang). Dies ist auch anschaulich verständlich: Angenommen es sollen Personen anhand ihrer Größe und ihres Gewichts voneinander unterschieden werden. Dazu wurden die Daten von mehreren Personen untersucht, die zufällig eine ähnliche Größe besitzen. Hingegen variiert das Gewicht dieser Personen in der Stichprobe stark. Ohne ausreichende Standardisierung der Variablen, wird die PCA daher das Gewicht einer Person als Hauptkomponenten mit dem größten Beitrag zur totalen Varianz identifizieren. Somit würden die Personen ausschließlich anhand ihres Gewichts unterschieden werden [10].
Daher empfiehlt es sich die Variable weiterhin durch die Standardabweichung der Variable selbst zu dividieren (vgl. Gleichung (3.2)). Dadurch werden die Auswirkungen von Skalierungsfaktoren auf die Kovarianz negiert (Beweis hierfür im Anhang).

$$x_i' = \frac{1}{s(X)}(x_i - \bar{x}) \tag{3.2}$$

Durch diese Standardisierung ergibt sich im weiteren Verlauf der PCA nicht die Kovarianzmatrix, sondern die Korrelationsmatrix [9]. Daher ist es auch möglich die Auswirkung von Skalierungsfaktoren auf die PCA zu verhindern, indem während der Standardisierung ausschließlich zentriert wird und im nächsten Schritt anstatt der Kovarianzmatrix die Korrelationsmatrix gebildet wird. Die Berechnung der Korrelationsmatrix anstatt der Kovarianzmatrix im nächsten Schritt birgt jedoch einen erhöhten Speicher- bzw. Rechenaufwand: Während der Standardisierung muss die Standardabweichung bei p Variablen p-mal berechnet werden. Nach der Standardisierung einer Variablen wird die Standardabweichung für diese Variable nicht mehr benötigt. Auch muss bei der Berechnung der Korrelationsmatrix die Standardabweichung p-mal berechnet werden. Jedoch müssen alle p Standardabweichungen für die Zeit der Berechnung der Korrelationsmatrix gespeichert werden. Eine speichereffizientere Methode besteht in der Neuberechnung der Standardabweichung für jede Zelle der Korrelationsmatrix. In diesem Fall muss die Standardabweichung jedoch $(2p^2 - 2p)$-mal berechnet werden.

3.2 Berechnung der Kovarianzmatrix

Nachdem der Datensatz im ersten Schritt standardisiert wurde, wird im zweiten Schritt der PCA der Zusammenhang zwischen den einzelnen Variablen des Datensatzes untersucht. Dazu werden die Kovarianzen zwischen jeweils zwei Variablen berechnet und in eine Matrix eingetragen. Dies wird für alle Variablenkombinationen wiederholt. Dadurch ergibt sich bei p Variablen im Datensatz eine quadratische Matrix der Größe $p \times p$ (vgl. Gleichung (3.3): Beispielhaft mit $p = 3$ (X, Y und Z)) [7]. Diese Matrix ist symmetrisch und wird als Kovarianzmatrix S bezeichnet.

$$S = \begin{bmatrix} s^2(X') & cov(X', Y') & cov(X', Z') \\ cov(Y', X') & s^2(Y') & cov(Y', Z') \\ cov(Z', X') & cov(Z', Y') & s^2(Z') \end{bmatrix} \quad \in \mathbb{R}^{p \times p} \qquad (3.3)$$

Wie im vorherigen Kapitel beschrieben und im Anhang gezeigt, entspricht, bei der Verwendung der Gleichung (3.2) zur Standardisierung, die berechnete Kovarianzmatrix der Korrelationsmatrix. Ist dies der Fall, liegen alle Elemente der Matrix zwischen -1 und 1. Die Elemente der Hauptdiagonalen entsprechen 1.

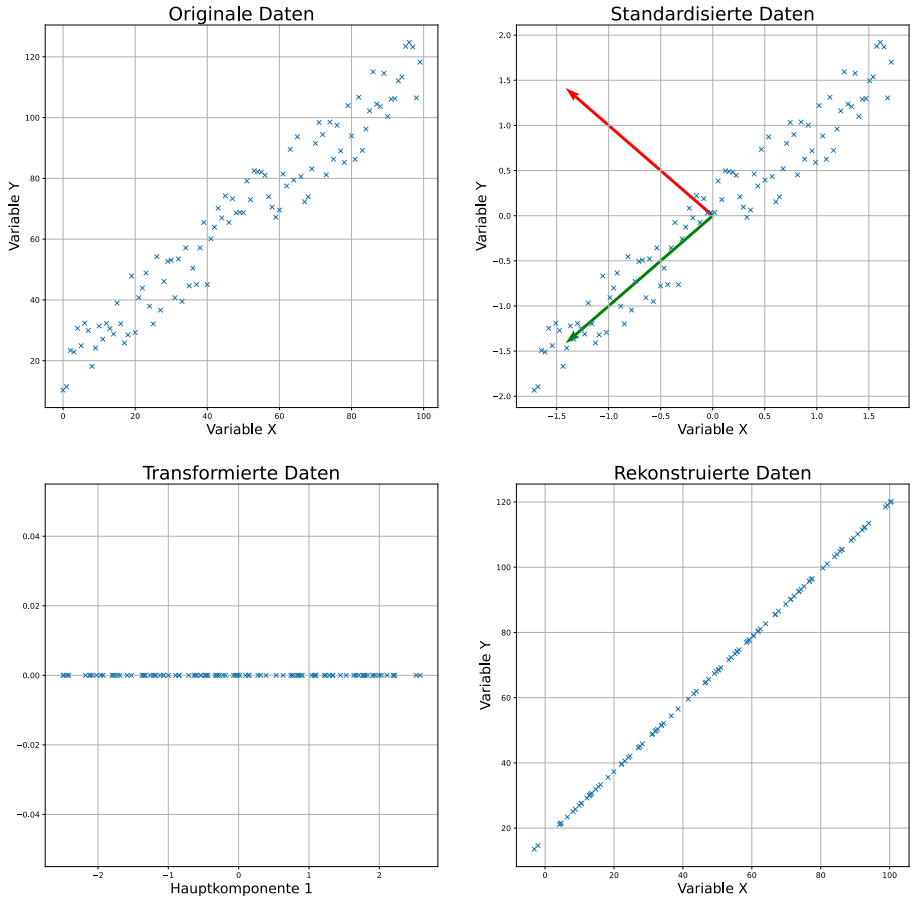

Abbildung 1: Vergleich der Daten während den einzelnen Schritten der PCA. Links oben die originalen Daten, rechts oben die durch Gleichung (3.2) standardisierten Daten mit den berechneten Hauptkomponenten (erste Hauptkomponente grün, Zweite rot), links unten die transformierten Daten (nur abhängig von der ersten Hauptkomponente), rechts unten die rekonstruierten Daten aus einer Hauptkomponente. Eigene Darstellung nach **[2]** und **[7]**.

3.3 Berechnung der Eigenwerte und Eigenvektoren

Zur Ermittlung der Hauptkomponenten und deren Beitrag zur totalen Varianz, müssen die Eigenwerte und die Eigenvektoren der Kovarianzmatrix S berechnet werden. Die Eigenvektoren der Kovarianzmatrix geben die Richtungen der Achsen mit der meisten Varianz also auch der meisten Informationen an. Diese Richtungen stellen die Hauptkomponenten dar [8] (vgl. Abbildung 1, standardisierte Daten: die Richtung der ersten Hauptkomponente in grün, die der zweiten in rot). Der Eigenwert zum

jeweiligen Eigenvektor gibt den Anteil der Hauptkomponente zur totalen Varianz wieder und bestimmt somit die Bedeutung der Hauptkomponente [11]. Dies geschieht, indem die jeweiligen Eigenwerte durch die Summe aller Eigenwerte geteilt werden. Je höher der Eigenwert, je höher ist die Bedeutung dieser Hauptkomponente. Werden die Eigenvektoren in der Reihenfolge ihrer Eigenwerte vom höchsten zum niedrigsten sortiert, werden die Hauptkomponenten somit automatisch in der Reihenfolge ihrer Bedeutung sortiert [7]. Wie in Kapitel 2.2 gezeigt, und da die Kovarianzmatrix quadratisch und symmetrisch ist, liegen die Eigenvektoren und damit alle Hauptkomponenten orthogonal zueinander [2], [8]. Daraus folgt, dass die Hauptkomponenten unkorreliert zueinander sind.

3.4 Bildung eines Feature Vektors und Transformation der Daten

Nachdem alle Hauptkomponenten bestimmt sind, ist im vierten Schritt zu entscheiden, wie viele der Hauptkomponenten verworfen werden. Aus den jeweiligen Eigenvektoren $\vec{v_i}$ der verbleibenden q Hauptkomponenten wird daraufhin eine Matrix gebildet. Diese Matrix der Größe $p \times q$ wird als Feature Vektor F bezeichnet (vgl. Gleichung (3.4)) [7].

$$F = \left[\vec{v}_1, \quad \dots, \quad \vec{v}_q\right] \quad \in \mathbb{R}^{p \times q} \tag{3.4}$$

Da der Datensatz nach der Standardisierung nicht weiter verändert wurde und daher noch in seinem ursprünglichen Raum liegt, muss dieser abschließend in den neuen Raum, der durch die Hauptkomponenten dargestellt wird, transformiert werden. Dazu wird der standardisierte Datensatz $M' \in \mathbb{R}^{n \times p}$ zunächst transponiert und dann mit dem transponierten Feature Vektor F multipliziert (vgl. Gleichung (3.5)) [7]. Die daraus entstehenden Daten sind nicht mehr einfach interpretierbar (vgl. Abbildung 1, transformierte Daten).

$$T_{new} = F^T \cdot M'^T \quad \in \mathbb{R}^{q \times n} \tag{3.5}$$

Werden nicht alle Hauptkomponenten bzw. Eigenvektoren zur Bildung von F verwendet, wird dadurch die Dimension des neuen Datensatzes T_{new} verringert. Einher geht damit jedoch auch ein Informationsverlust. Wie groß dieser Informationsverlust ist, ist an der Abweichung der totalen Varianz zu erkennen [11].

3.5 Rekonstruktion der Daten

In manchen Fällen ist es notwendig die transformierten Daten T_{new} wieder in den ursprünglichen Dimensionen des Datensatzes M zurückzuführen. Dies ist zum Beispiel bei der Bildkompression der Fall. In Abbildung 1 wird der Informationsverlust durch die Vernachlässigung einer Hauptkomponente deutlich erkennbar.
Um aus T_{new} wieder M' zu erhalten, wird die Gleichung (3.5) zu Gleichung (3.6) umgeformt [7]. M' wird nachfolgend als $M'_{reduced}$ bezeichnet, da diese nicht mit M' übereinstimmt, wenn nicht alle Hauptkomponenten zur Bildung von T_{new} verwendet wurden.

$$(F^{\,T})^{\,-1} \cdot T_{new} = M'^{\,T}_{reduced} \tag{3.6}$$

Da die Vektoren des Feature Vektors $F^{\,T}$ orthogonal zueinander sind, entspricht die Inverse dieses Vektors dem Vektor selbst [7] (Beweis hierfür im Anhang). Unter der Annahme, dass zu Beginn der PCA die Vektoren standardisiert wurden, muss diese Standardisierung rückgängig gemacht werden [7]. Hierfür werden zwei Matrizen K_1 und K_2 eingeführt (vgl. Gleichung (3.7)). Wurde die Standardisierung mit Gleichung (3.2) durchgeführt, soll K_1 die Skalierung der Vektoren durch die Division mit $s(X_i)$ negieren. Dazu werden auf der Hauptdiagonalen von K_1 jeweils die Standardabweichungen der Variable eingetragen (vgl. Gleichung (3.8)). Wurden die Vektoren während der Standardisierung nicht skaliert, sondern nur zentriert, entspricht K_1 einer Einheitsmatrix. K_2 soll die Zentrierung der Variablen rückgängig machen. Daher stellt K_2 eine Matrix mit den Mittelwerten der Variablen dar (vgl. Gleichung (3.9)).

$$(F \cdot T_{new})^T \cdot \mathrm{K}_1 + \mathrm{K}_2 = M^T_{reduced} \tag{3.7}$$

$$\mathrm{K}_1 = \begin{bmatrix} s(X_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s(X_p) \end{bmatrix} \quad \in \mathbb{R}^{p \times p} \tag{3.8}$$

$$\mathrm{K}_2 = \begin{bmatrix} \bar{x}_1 & \cdots & \bar{x}_p \\ \vdots & \ddots & \vdots \\ \bar{x}_1 & \cdots & \bar{x}_p \end{bmatrix} \quad \in \mathbb{R}^{p \times n} \tag{3.9}$$

Sollen die Daten nach der PCA wieder in ihre ursprünglichen Dimensionen transformiert werden, so ist aus den Gleichungen oben ersichtlich, dass alle Standardabweichungen und Mittelwerte gespeichert werden müssen.

4 Probleme und Grenzen der Principal Component Analysis

Die PCA ist eine beliebte Methode zur Dimensionsreduktion. Allerdings dürfen die Schwächen dieser Methode nicht außer Acht gelassen werden. Sind diese Schwächen bekannt, können diese durch entsprechende Methoden behoben werden. Können sie nicht behoben werden, empfiehlt es sich auf eine andere Methode zur Dimensionsreduktion auszuweichen.
Die PCA geht von einer linearen Korrelation zwischen den Variablen aus. Somit werden nichtlineare Zusammenhänge zwischen Variablen schlecht dargestellt. [12] zeigt, dass dies bei unimodalen Verhalten von Variablen zu Problemen führen kann. Im Falle einer ökologischen Analyse kann nach [12] der Fall auftreten, dass bei langen Gradienten und heterogenen Daten die PCA versucht Kurven mit einer Geraden zu approximieren. Dies führt zu keinem guten Ergebnis. Ist es dennoch erwünscht eine PCA durchzuführen, kann der Datensatz in kleinere homogene Teile aufgeteilt werden [12].
Weiterhin wird in [6] die geringe Interpretierbarkeit der Hauptkomponenten als Nachteil der PCA aufgeführt. Da die Hauptkomponenten eine Linearkombination der ursprünglichen Variablen darstellen, ist es nicht einfach die wichtigsten Merkmale zu identifizieren, wenn die Hauptkomponenten bereits berechnet sind.
Verbunden mit einer Dimensionsreduktion ist in der Regel auch immer Informationsverlust. Dies stellt zwangsläufig in einem Kompromiss zwischen Dimensionsreduktion und Informationsverlust dar. Eine Reduktion der Komplexität durch das Vernachlässigen von Hauptkomponenten, ist gleichzeitig eine Entscheidung Informationen zu vernachlässigen. Diese Abwägung muss bei der PCA getroffen und richtig dosiert werden. Auch ist eine Regel, wie viel Informationsverlust in Kauf genommen werden kann, nicht allgemein bestimmbar. In dem Beispiel der Bildkompression aus Kapitel 5.2 ist ein Informationsverlust von 2% bereits erkennbar, während in anderen Fällen ein Informationsverlust von über 10% akzeptierbar ist.
Die PCA reagiert empfindlich auf die Skalierung der Merkmale, die z.B. durch eine Einheitenänderung unbeabsichtigt geändert werden kann. Daher ist der in Kapitel 3.1 vorgestellte Schritt der Standardisierung in den meisten Fällen notwendig, um eine verlässliche Analyse durchzuführen. In gleicher Weise reagiert die PCA sensibel auf Ausreißer. Diese verzerren die Varianz und damit den Algorithmus stark. Es ist daher zu empfehlen diese Ausreißer vor der Analyse zu entfernen.

5 Anwendungen der Principal Component Analysis

Nachdem die Funktion der PCA zu Dimensionsreduktion und die Probleme dieser Analyse in den vorherigen Kapiteln allgemein vorgestellt wurde, werden nun konkrete Anwendungsbeispiele gegeben. Diese sind im begleitendem Jupyter Notebook implementiert. In dieser Arbeit werden ausschließlich die wesentlichen Ergebnisse dieser Beispiele und ihre Eigenheiten aufgezeigt.
Nachfolgend werden zwei Anwendungen der PCA vorgestellt: Eine Gesichtserkennung mithilfe der PCA und eine Anwendung zur Bild-Kompression.
Neben diesen Anwendungen bietet eine Dimensionsreduktion durch die PCA noch weitere Anwendungen. [5] zeigt, wie mithilfe der PCA die Genauigkeit der Vorhersage eines Aktienkurses deutlich verbessert werden kann. Auch wird die PCA gerne verwendet, um hochdimensionale Daten zu visualisieren [7], [8]. Zudem wird die PCA verwendet, um die Trainingsdauer eines Modells zu verkürzen. Dies wird unter anderem von [4] gezeigt.

5.1 PCA zur Gesichtserkennung

Auch im Falle der Gesichtserkennung birgt eine Reduktion der Dimension eine Beschleunigung der Algorithmen. Zudem wird eine Überanpassung durch hochdimensionale Daten bei regressionsbasierten Algorithmen verhindert. Auch in diesem Beispiel der Gesichtserkennung werden die Dimensionen reduziert, um eine Überanpassung zu verhindern. Weiterhin wird dadurch die Berechnung für leistungsschwache Hardware ermöglicht. Dies ist vor allem dann zu berücksichtigen, wenn Algorithmen in Embedded Systemen implementiert werden sollen.
Zur Gesichtserkennung mithilfe der PCA werden sogenannte Eigenfaces aus einem Teil des Datensatz erstellt (vgl. Abbildung 2). Der verbleibende Teil des Datensatzes wird zum Testen des Modells verwendet. In diesen Eigenfaces sind die wichtigsten Merkmale eines Gesichts gespeichert. Somit werden die wichtigsten Informationen von verschiedenen Bildern eines Gesichts in ein Eigenface komprimiert. Anhand dieser Eigenfaces kann anschließend ein unbekanntes Bild eines Gesichts einer Person zu dieser Person zugeordnet werden.

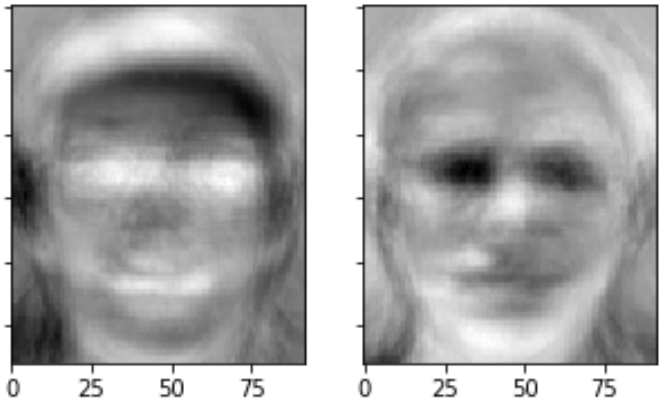

Abbildung 2: Zwei Eigenfaces die durch Dimensionsreduktion mit PCA erhalten wurden.

5.2 PCA zur Bildkompression

Die PCA kann zur Kompression von mehreren Bildern verwendet werden, wenn diese dieselbe Größe haben. Dazu wird für jeden Pixel eines Bildes ein Vektor mit den Werten dieses Pixels aus allen Bildern gebildet. Ist das Bild p mal q Pixel groß, ergeben sich somit $n = pq$ Vektoren mit je i Elementen, wobei i der Anzahl der Bilder entspricht. Die Dimensionalität dieser Vektoren entspricht daher i. Anschließend wird mit den gebildeten Vektoren die PCA durchgeführt. Da die Vektoren i -dimensional sind, werden auch i Eigenvektoren bzw. Hauptkomponenten gefunden. Werden nur $j < i$ Hauptkomponenten für die anschließende Transformation verwendet, werden die Bilder komprimiert. [7]
Durch diese Kompression werden, wie in Abbildung 3 zu erkennen, Informationen verloren. In dem gezeigten Beispiel wurde auf die Standardisierung der Vektoren zu Beginn der PCA verzichtet. Dies hat den Grund, dass wenn die Standardisierung durchgeführt wird, alle Mittelwerte und Standardabweichungen gespeichert werden müssen, damit nach der PCA die Bilder wiederhergestellt werden können (vgl. Kapitel 3.5). Die Auswirkungen der Standardisierung ohne Speicherung der Mittelwerte und Standardabweichungen kann im Jupyter Notebook getestet werden.

Abbildung 3: Vergleich zwischen einem aus 250 Hauptkomponenten (HKs) wiederhergestellten Bildes (links) und dem Originalbild mit 400 Hauptkomponenten (rechts). Dies entspricht einer Kompression des Bildes um 37,5%, während 98% der Bildinformationen erhalten wurden.

6 Zusammenfassung

Wie in den vorherigen Kapiteln gezeigt, stellt die PCA eine einfache Methode dar, um die Dimensionalität des Datensatzes zu reduzieren. Dadurch ergeben sich viele verschiedene Anwendungsmöglichkeiten, die auch über die einfache Reduktion der Trainingsdauer hinausgehen. Zu beachten ist jedoch, dass die PCA die Performance eines Modells unter bestimmten Umständen verschlechtern kann. Jedoch können hier andere, ähnliche Methoden Abhilfe verschaffen (z.B. ICA, Independent Component Analysis) [6].
In dem Jupyter Notebook ist zu erkennen, dass die Implementierung der PCA einen vernachlässigbaren Mehraufwand mit sich bringt. Daher ist es fast immer zu empfehlen eine PCA zu implementieren und die Auswirkungen dieser auf das Modell zu untersuchen.

Literatur

[1] J. Shlens, „A Tutorial on Principal Component Analysis," ArXiv, Mountain View, CA, 2014.

[2] V. K. Ayyadevara, „Principal Component Analysis," in *Pro Machine Learning Algorithms*, Berkeley, CA, Apress, 2018, pp. 283-297.

[3] L. Battulga, S.-H. Lee, A. Nasridinov und K.-H. Yoo, „Hash-tree PCA: accelerating PCA with hash-based grouping," *The Journal of Supercomputing,* Nr. 76, p. 8248–8264, 2020.

[4] Y.-W. Chen und Y. Iwasaki, „A Robust MR Image Segmentation Technique Using Spatial Information and Principle Component," in *Advances in Neural Networks - ISNN 2006*, Berlin, Heidelberg, Springer Berlin Heidelberg, 2006, pp. 517-522.

[5] R. Deshmukh, P. Jadhav, S. Shelar, U. Nikam, D. Patil und R. Jawale, „Stock Price Prediction Using Principal Component Analysis and Linear Regression," in *Emerging Technologies in Data Mining and Information Security*, Singapore, Springer Nature Singapore, 2023, pp. 269-276.

[6] A. Ng und K. Soo, Data Science - was ist das eigentlich?!, Heidelberg: Springer Berlin, 2018.

[7] L. I. Smith, „A tutorial on Principal Components Analysis," Department of Computer Science, University of Otago, Neuseeland, 2002.

[8] E. Bisong, „Principal Component Analysis (PCA)," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, Berkelay, CA, Apress, 2019, pp. 319-324.

[9] I. Jolliffe, Principal Component Analysis, Bd. 2, New York: Springer New York, 2002.

[10] S. Marukatat, „Tutorial on PCA and approximate PCA and approximate kernel PCA," *Artificial Intelligence Review,* pp. 1573-7462, 31 10 2022.

[11] M. von der Hude, „Dimensionsreduktion - Hauptkomponentenanalyse englissh: principal components (PCA)," in *Predictive Analytics und Data Mining : Eine Einführung mit R*, Wiesbaden, Springer Fachmedien Wiesbaden, 2020, pp. 83-92.

[12] D. I. Leyer und K. Wesche, Multivariate Statistik in der Ökologie, Heidelberg: Springer Berlin, 2007.

Anhang

Beweis für die Unabhängigkeit von Skalierungsfaktoren durch die Standardisierung mit der Standardabweichung

Angenommen die Werte der Variablen X und Y sind durch die Faktoren a und b skaliert (vgl. Gleichung (0.1)). Weiterhin seien die Mittelwerte beider Variablen Null.

$$X' = aX\,, \quad Y' = bY \qquad mit\ \bar{x} = 0,\ \bar{y} = 0 \tag{0.1}$$

Die Kovarianz dieser skalierten Variablen ergibt sich dann mit Gleichung (0.2).

$$cov(X', Y') = \frac{1}{n-1} \sum_{i=1}^{n} (ax_i)(by_i) \tag{0.2}$$

$$cov(X', Y') = ab \cdot \frac{1}{n-1} \sum_{i=1}^{n} (x_i)(y_i) \tag{0.3}$$

$$cov(X', Y') = ab \cdot cov(X, Y) \tag{0.4}$$

Gleichung (0.2) kann in Gleichung (0.4) umgeformt werden. Dabei ist zu erkennen, dass die Varianz abhängig von den Skalierungsfaktoren a und b ist.
Werden die Variablen vor der Berechnung der Kovarianz mithilfe der Standardabweichung standardisiert, so wird die Abhängigkeit von den Skalierungsfaktoren entfernt (vgl. Gleichung (0.5) bis (0.8)).

$$X'' = \frac{X'}{s(X')}\,, \quad Y'' = \frac{Y'}{s(Y')} \qquad mit\ \overline{x'} = 0,\ \overline{y'} = 0 \tag{0.5}$$

$$s(X') = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (ax_i)^2} = a \cdot s(X) \tag{0.6}$$

$$X'' = \frac{X'}{s(X')} = \frac{aX}{as(X)} = \frac{X}{s(X)} \tag{0.7}$$

$$cov(X'', Y'') = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i\, y_i}{s(X)s(Y)} = \frac{cov(X,Y)}{s(X)s(Y)} \qquad (0.8)$$

Durch die Berechnung der Kovarianz wird dann die Korrelation zwischen X und Y erhalten (vgl. (0.9)).

$$\frac{cov(X,Y)}{s(X)s(Y)} = r_{X,Y} \qquad (0.9)$$

Beweis: Inverse einer orthogonalen Matrix entspricht transponierter Matrix

Sei die Matrix $A = [\vec{a}_1, \dots, \vec{a}_n]$ eine orthogonale Matrix, d.h. alle Vektoren $\vec{a}_i$ der Matrix sind orthogonal zueinander. Weiterhin sei die Einheitsmatrix I.
Wird $B = A^T A$ berechnet, wird für jedes Element von der Matrix B das Skalarprodukt der Vektoren $\vec{a}_i^T$ und $\vec{a}_j$ berechnet. Unter Berücksichtigung, dass die Vektoren orthogonal zueinander sind, ergibt sich so in allen Fällen für das Skalarprodukt, mit Ausnahme der Elemente auf der Hauptdiagonalen, Null. Auf der Hauptdiagonalen der Matrix B wird das Skalarprodukt des Vektors $\vec{a}_i^T$ mit $\vec{a}_{j=i}$ gebildet. Dieses ergibt Eins. Daher entspricht die Matrix B der Einheitsmatrix I. Zudem ist die Inverse der Matrix A definiert als diejenige Matrix, die Multipliziert mit der Matrix A selbst, die Einheitsmatrix ergibt. Aus Gleichung (0.10) wird daher ersichtlich, dass $A^T = A^{-1}$ gilt.

$$A^T A = B = I = A^{-1} A \qquad (0.10)$$

Beweis für Orthogonalität der Eigenvektoren reeller, symmetrischer Matrizen

Sei die Matrix M reell und symmetrisch, nur dann entspricht die transponierte Matrix M^T der Matrix M. Der Eigenwert von M ist definiert durch:

$$M\vec{v} = \lambda\vec{v} \qquad (0.11)$$

Werden beide Seiten transponiert und mit einem weiteren Eigenvektor $\vec{v}_2$ erweitert, ergibt sich Gleichung (0.14) (einzelne Schritte in Gleichung (0.12) bis (0.14)).

$$(M\vec{v}_1)^T = (\lambda_1\vec{v}_1)^T \tag{0.12}$$

$$\vec{v}_1^T\, M^T = \lambda_1\, \vec{v}_1^T \tag{0.13}$$

$$\vec{v}_1^T\, M^T\, \vec{v}_2 = \lambda_1\, \vec{v}_1^T\, \vec{v}_2 \tag{0.14}$$

Da $\vec{v}_2$ ein Eigenvektor ist und mit der Definition eines Eigenvektors ergibt sich daraus die nachfolgende Gleichung.

$$\vec{v}_1^T\, \lambda_2\, \vec{v}_2 = \lambda_1\, \vec{v}_1^T\, \vec{v}_2 \tag{0.15}$$

Diese Gleichung lässt sich in Gleichung (0.16) umformen. Da die Eigenwerte unterschiedlich sind (verschiedene Eigenvektoren) muss der Term $\vec{v}_1^T\, \vec{v}_2 = 0$ sein. Daraus folgt, dass die Eigenvektoren orthogonal zueinander sind.

$$(\lambda_2 - \lambda_1)\vec{v}_1^T\, \vec{v}_2 = 0 \tag{0.16}$$