Vergleichende Analyse der Methoden zur Hauptkomponenten-Analyse

Aram Abramov[1], Johann-Friedrich Luy [1,2] *IEEE Fellow*

[1] Department of Electrical Engineering, TUM School of Computation, Information and Technology, Technische Universität München

[2] COREPROG engineering, www.coreprog.de

Einleitung und Problemstellung

Die Hauptkomponentenanalyse (HKA) ist ein statistisches Verfahren der multivariaten Analyse, das in vielen Disziplinen wie der Bildverarbeitung, dem Maschinellen Lernen und der Mustererkennung Anwendung findet. Ihre Bedeutung resultiert aus der Fähigkeit, komplexe Datensätze zu vereinfachen, indem sie die wichtigsten Merkmale bzw. Variablen identifiziert und in neue, unabhängige Variablen transformiert, die als Hauptkomponenten bezeichnet werden. Diese Transformation ermöglicht es, die zugrunde liegenden Muster und Strukturen in den Daten zu erkennen und interpretierbare Einblicke zu gewinnen. Die HKA ist eine verbreitete statistische Methode zur Reduzierung der Dimensionalität von Daten, wobei die wichtigen Informationen erhalten bleiben sollen.

Beispielsweise wird auf dem Bild 1 ein zweidimensionaler Datensatz betrachtet, der in einem x-y-Koordinatensystem gemessen wurde. Die Richtung, entlang der der Datensatz die größte Varianz aufweist, ist durch die u-Achse gekennzeichnet worden. Die „zweitwichtigste" Richtung ist orthogonal dazu und durch die v-Achse definiert. Wenn wir jede (x,y) Koordinate in die (u,v) -Koordinaten transformieren, sind die neuen Datenvariablen dekorreliert, das heißt, die Kovarianz zwischen den (u-v) -Variablen ist 0.

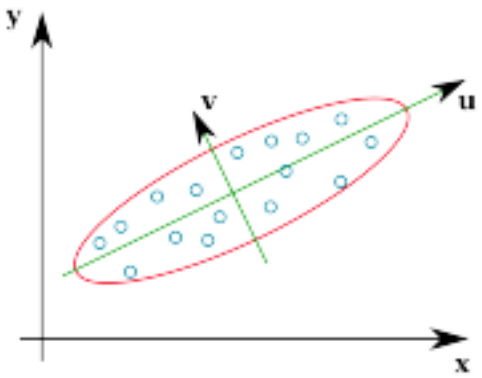

Bild 1: Prinzip der Hauptkomponentenanalyse

Für einen gegebenen Datensatz findet also die Hauptkomponentenanalyse das Achsensystem, das die größte Varianz im Datensatz beschreibt. Man sucht die Hauptkomponenten des Datensatzes. Sie alle sind senkrecht zueinander und ermöglichen somit die Darstellung des Datensatzes durch neue unkorrelierte Variablen. Diese sind allerdings Linearkombinationen der ursprünglichen Variablen und deswegen oft nur schwer zu interpretieren.

Es ist zu beachten, dass die entgegengesetzte Orientierung der u- oder v-Achse genauso möglich ist. Die Wahl der Orientierung der Achsen im neuen Koordinatensystem hat einen Einfluss auf die Interpretierbarkeit der Daten. Im Weiteren wird eine Vorschrift zur Wahl der Orientierung der Hauptkomponenten analysiert.

Das gleiche Prinzip wie im vorliegenden Beispiel wird auch bei den Datensätzen höherer Dimensionen angewendet. Die erste Hauptkomponente beschreibt die größte Varianz im Datensatz, die zweite Hauptkomponente steht senkrecht zur ersten und beschreibt die größte Varianz im verbliebenen Unterraum des Datensatzes, usw.

Zur Reduktion eines jeweiligen Datensatzes auf *k* Dimensionen, betrachtet man nur die ersten *k* Hauptkomponenten, die den *k* Dimensionen entsprechen. Der Datensatz wird auf *k* Hauptkomponenten projiziert und der Informationsverlust ist dabei minimiert. Falls man bis zu drei Hauptkomponenten betrachtet, kann der Datensatz durch die jeweiligen Hauptkomponenten visualisiert werden. Dies ist ein wichtiger Grund zur Anwendung der HKA.

Die mathematische Fragestellung der HKA lautet wie folgt:

Gegeben sei ein Datensatz $\boldsymbol{X}$ bestehend aus *p* numerischen Merkmalen, die an *n* Objekten erhoben wurden. Wie können wir eine lineare Transformation finden, die die Varianz in $\boldsymbol{X}$ maximiert und gleichzeitig die Korrelation zwischen den Variablen minimiert?

Mathematisch ausgedrückt suchen wir also nach einer *p*×*k* Matrix, wobei *k* die Anzahl der gewünschten Hauptkomponenten ist, die die ursprünglichen *p* Variablen in einen *k*-dimensionalen Raum transformiert. Dabei soll die Varianz der Daten maximiert werden.

Die Hauptkomponentenanalyse wird oft verwendet, um die Dimensionalität eines Datensatzes zu reduzieren, indem sie eine kleinere Anzahl von Hauptkomponenten identifiziert, die einen Großteil der Variation in den Daten erklären können. Dies ermöglicht eine kompaktere Darstellung der Daten und kann dazu beitragen, Rauschen zu reduzieren oder Redundanzen zu eliminieren. Jedoch kann die HKA auch genutzt werden, um eine Struktur in den Daten zu erkennen, ohne unbedingt die Dimensionalität zu reduzieren. In manchen Fällen ist es wichtig, alle Hauptkomponenten zu betrachten, um ein umfassendes Verständnis der Daten zu erhalten, auch wenn dies zu keiner Dimensionalitätsreduktion führt.

HKA durch Eigenwertzerlegung der Kovarianzmatrix

Wir gehen in der Problemstellung von einem Datensatz aus, der in einer Datenmatrix $\boldsymbol{X}$ zusammengefasst ist. Es sind *p* numerische Merkmale, die für jedes der *n* Objekte erhoben wurden. Die Objekte können auch Merkmalsträger genannt werden. Der Datensatz besteht aus *n* Vektoren $\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_p}$, die *p*-dimensional sind und bildet eine $n \times p$-Datenmatrix $\boldsymbol{X}$, deren *j*-te Spalte ein Vektor $\boldsymbol{x_j}$ der Werte des *j*-ten Merkmals ist.

$$\begin{matrix} & Merkmale\ [Spalten] \\ Objekte\ (Merkmalsträger)[Zeilen] & \end{matrix} \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Nach der Grundidee der HKA suchen wir nach einer linearen Kombination der Spaltenvektoren von $\boldsymbol{X}$, die maximale Varianz aufweist. Diese lineare Kombination ist gegeben durch

$$\sum_{j=1}^{p} a_j \boldsymbol{x}_j = \boldsymbol{Xa}$$

mit einem Vektor $\boldsymbol{a}$ der Konstanten $a_1, \dots, a_p$.

Für die Varianz dieser linearen Kombination gilt: $var(\boldsymbol{Xa}) = \boldsymbol{a^T S a}$, wobei $\boldsymbol{S}$ die Kovarianzmatrix des ursprünglichen, durch Matrix $\boldsymbol{X}$ gegebenen Datensatzes ist [4].
Das Herausfinden dieser linearen Kombination mit größter Varianz kann somit als Maximierungsproblem der quadratischen Form $\boldsymbol{a^T S a}$ aufgefasst werden.
Damit dieses Problem eine eindeutige Lösung hat, muss eine Nebenbedingung definiert werden. Die häufigste Nebenbedingung besteht darin, mit Einheitsnormvektoren zu arbeiten, d. h. $\boldsymbol{a^T a} = 1$ ist die Nebenbedingung.
Es liegt somit ein Optimierungsproblem vor und die Lagrange-Funktion dazu sieht wie folgt aus:

$\boldsymbol{L(a) = a^T S a - \lambda(a^T a - 1)}$, wobei λ der Lagrange-Multiplikator ist.

Differenzieren der $\boldsymbol{L(a)}$ nach Vektor $\boldsymbol{a}$ und Gleichsetzen mit dem Nullvektor liefert die Gleichung $\boldsymbol{Sa - \lambda a = 0}$ bzw.: $\boldsymbol{Sa = \lambda a.}$ Also muss $\boldsymbol{a}$ ein (normierter) Eigenvektor und λ der entsprechende Eigenwert der Kovarianzmatrix $\boldsymbol{S}$ sein. [Die Normierungsbedingung erhält man auch aus der partiellen Ableitung der Lagrange-Funktion nach λ]. Wir sind am größten Eigenwert λ interessiert, da er der Varianz entspricht, denn: $var(Xa) = a^T S a = \lambda\, a^T a = \lambda$.

Die letzte Gleichung gilt auch wenn wir den Eigenvektor mit -1 multiplizieren. Somit ist die Orientierung des Eigenvektors nicht eindeutig. Diese Orientierung legt die Orientierung der entsprechenden Hauptkomponente fest. Im Beispiel aus der Einleitung wird eine andere Wahl des Vorzeichens des Eigenvektors in entgegengesetzter Richtung einer jeweiligen Achse resultieren.

Jede reelle symmetrische $p \times p$ Matrix, wie die Kovarianzmatrix $\boldsymbol{S}$, hat p reelle Eigenwerte und zu ihnen korrespondierende Eigenvektoren. Diese bilden eine orthogonale Basis, die zu einer orthonormalen Basis umgeformt werden kann, sodass gilt: $a_k a_{k'} = 1$, falls k=k', oder 0 sonst. Der Ansatz mit Lagrange-Multiplikator und der zusätzlichen Nebenbedingung der Normierung kann weiterhin verwendet werden, um zu zeigen, dass alle Eigenvektoren der Matrix $\boldsymbol{S}$ die Lösung des Problems zur Gewinnung neuer linearen Kombinationen $\boldsymbol{Xa}_k = \sum_{j=1}^{p} a_{jk} \boldsymbol{x}_j$ sind, und sukzessive die Varianz maximieren und die Unkorreliertheit der ursprünglichen Daten herbeiführen.

Diese lineare Kombination $\boldsymbol{Xa}_k$ heißt Hauptkomponente des Datensatzes. Alle Hauptkomponenten stehen senkrecht zueinander und ermöglichen somit die Darstellung des Datensatzes durch neue unkorrelierte Variablen.

Die Kovarianzmatrix des Datensatzes kann bestimmt werden durch die Formel $\boldsymbol{S} = \frac{\mathbf{1}}{(n-1)} \boldsymbol{X}^{*\boldsymbol{T}} \boldsymbol{X}^{*}$, wobei $\boldsymbol{X}^{*}$ die zentrierte Datenmatrix (also mit nullzentrierten Spaltenvektoren) ist. (S. 24, [4])

Falls $\boldsymbol{V}$ die Matrix der Eigenvektoren der Matrix $\boldsymbol{S}$ ist, gilt: $\boldsymbol{SV} = \boldsymbol{VL}$, wobei $\boldsymbol{L}$ die Diagonalmatrix mit den Eigenwerten ist. Da die Eigenvektoren normiert sind und senkrecht zu einander stehen, gilt: $\boldsymbol{V^T V} = \boldsymbol{VV^T} = \boldsymbol{I}$. Es folgt $\boldsymbol{V^T SV} = \boldsymbol{L}$ und folglich ist $\boldsymbol{S} = \boldsymbol{VLV^T}$

Es wird also für die Hauptkomponentenanalyse die Eigenwertzerlegung der Kovarianzmatrix vollzogen, wobei die Eigenwerte in der absteigenden Reihenfolge geordnet werden. Man kann die Eigenwertmatrix als Basiswechsel des ursprünglichen Koordinatensystems der Daten in das neue Koordinatensystem der Daten sehen. Falls man bei der Hauptkomponentenanalyse die Dimensionalität des Datensatzes auf k reduzieren will, betrachtet man die ersten k Eigenvektoren und projiziert den Datensatz auf diese, indem man die Matrixmultiplikation der jeweiligen Eigenvektormatrix mit der Datenmatrix durchführt.

HKA durch Singulärwertzerlegung der zentrierten Datenmatrix

Es ist üblich, mit zentrierten Daten zu arbeiten. Wir betrachten die $n \times p$ Matrix $\boldsymbol{X}^{*}$, deren Spalten nullzentriert sind.

Es gilt: $(n-1)\boldsymbol{S} = \boldsymbol{X}^{*\boldsymbol{T}} \boldsymbol{X}^{*}$**.** Diese Gleichung führt auf die Idee, die Eigenwertzerlegung der Kovarianzmatrix $\boldsymbol{S}$ durch die Singulärwertzerlegung der Spalten-zentrierten Datenmatrix $\boldsymbol{X}^{*}$ zu erreichen. Im Allgemeinen ändern sich Eigenvektoren und Eigenwerte einer Matrix bei Multiplikation mit einer Konstanten nicht, deswegen ist die Eigenwertzerlegung für $(n-1)\boldsymbol{S}$ und $\boldsymbol{S}$ Matrizen gleich.

Für jede Matrix $\boldsymbol{Y}$ der Dimension $n \times p$ existiert die folgende Darstellung $\boldsymbol{Y} = \boldsymbol{U\Sigma V^T}$, die Singulärwertzerlegung (kurz SVD, von engl. singular value decomposition) heißt. Dabei sind $\boldsymbol{U}$ und $\boldsymbol{V}$ jeweils $n \times r$ und $r \times p$ Matrizen mit orthonormalen Spaltenvektoren (D. h. $\boldsymbol{U^T U} = \boldsymbol{I}$, $\boldsymbol{V^T V} = \boldsymbol{I}$) und $\boldsymbol{\Sigma}$ ist eine $r \times r$ Diagonalmatrix. Die Spalten von $\boldsymbol{U}$ heißen linke Singulärvektoren von $\boldsymbol{Y}$ und sind Eigenvektoren der Matrix $\boldsymbol{YY^T}$ korrespondierend zu nicht-null Eigenwerten. Die Spalten von $\boldsymbol{V}$ heißen rechte Singulärvektoren von $\boldsymbol{Y}$ und sind Eigenvektoren der Matrix $\boldsymbol{Y^T Y}$ korrespondierend zu nicht-null Eigenwerten. Die Diagonalelemente der Matrix $\boldsymbol{\Sigma}$ heißen Singulärwerte der Matrix $\boldsymbol{Y}$ und sind nichtnegative Wurzeln der Eigenwerte der beiden Matrizen $\boldsymbol{Y^T Y}$ und $\boldsymbol{YY^T}$. Es wird vorausgesetzt, dass die Diagonalelemente der Matrix $\boldsymbol{\Sigma}$ in absteigender Reihenfolge geordnet sind und folglich die Reihenfolge der korrespondierenden Eigenvektoren festgelegt ist. Wie man aus der Definition der Singularwertzerlegung $\boldsymbol{Y} = \boldsymbol{U\Sigma V^T}$ sehen kann, gilt $\boldsymbol{U^T Y} = \boldsymbol{\Sigma V^T}$ bzw. $\boldsymbol{v_k^T} = \frac{\mathbf{1}}{\boldsymbol{\sigma_i}} \boldsymbol{u_i^T y_i}$ für

die einzelnen jeweiligen Vektoren. Mit dieser Formel lässt sich die Matrix $\boldsymbol{V^T}$ bestimmen, wenn die Matrix $\boldsymbol{U}$ bereits berechnet wurde.

Wir setzen $\boldsymbol{Y} = \boldsymbol{X}^*$ ein.

$$(n-1)\boldsymbol{S} = \boldsymbol{X}^{*\boldsymbol{T}}\boldsymbol{X}^* = \left(\boldsymbol{U\Sigma V^T}\right)^{\boldsymbol{T}}\left(\boldsymbol{U\Sigma V^T}\right) = \boldsymbol{V\Sigma U^T U\Sigma V^T} = \boldsymbol{V\Sigma^2 V^T}$$

wobei $\boldsymbol{\Sigma^2}$ die Diagonalmatrix ist, deren Diagonalelemente Quadrate der Singulärwerte der Matrix $(n-1)\boldsymbol{S}$ sind.

Die Eigenschaften der Singulärwertzerlegung haben eine geometrische Interpretation in der HKA. Gegeben sei eine Datenmatrix des Rangs *r* der Dimension $n \times p$, die Matrix $\boldsymbol{Y_q}$ der gleichen Dimension aber des niedrigeren Rangs *q*<*r*, deren Elemente die quadratischen Differenzen korrespondierender Matrizenelemente der beiden Matrizen minimieren ist gegeben mit:

$$\boldsymbol{Y_q} = \boldsymbol{U_q \Sigma_q V_q^T}$$

wobei $\boldsymbol{\Sigma_q}$ die Diagonalmatrix mit den ersten (größten) *q* Elementen von $\boldsymbol{\Sigma}$ ist und $\boldsymbol{U_q}$ und $\boldsymbol{V_q^T}$ $n \times q$ und $q \times p$ Matrizen, erhalten durch Beibehalten nur der ersten *q* Spalten jeweils der Matrizen $\boldsymbol{U}$ und $\boldsymbol{V}$. (S. 4 in [1])

Im Kontext der HKA definieren *n* Zeilen der Spalten-zentrierten Datenmatrix $\boldsymbol{X}^*$ mit Rang r die "Punktwolke" aus *n* Punkten in einem *r*-dimensionalem Unterraum von $\mathbb{R}^p$ mit Ursprung im Schwerpunkt der „Punktwolke". Die oben erhaltene Formel beschreibt eine „beste" Approximation von n Punkten dieser Punkwolke in einem q-dimensionalen, durch die Spalten von $\boldsymbol{X_q^*}$ gegebenen Unterraum [1].

Da die Eigenwerte der SVD bei der HKA der Varianz entsprechen, kann die Qualität der Approximation der Dimensionalitätsreduktion durch die HKA durch das Verhältnis der verbleibenden Eigenwerte zu allen Eigenwerten beurteilt werden. Die Qualität einer Approximation durch eine einzelne Hauptkomponente kann mit der Formel beurteilt werden: $\frac{\lambda_j}{\sum_{j=1}^{p}\lambda_j} = \frac{\lambda_j}{tr(\boldsymbol{S})}$ wobei $tr(\boldsymbol{S})$ die Spur der Kovarianzmatrix $\boldsymbol{S}$ bezeichnet [1].

Gegenüberstellung der Eigenwertzerlegung der Kovarianzmatrix mit der Singulärwertzerlegung der zentrierten Datenmatrix

Wir fassen hier zuerst die beiden Methoden zusammen.

<u>Eigenwertzerlegung der Kovarianzmatrix:</u>

1. Schritt 1: Berechnung der Kovarianzmatrix: Die Kovarianzmatrix wird aus den Daten berechnet. Sie zeigt die Beziehung zwischen den verschiedenen Variablen und ihre Streuung.

2. Schritt 2: Berechnung der Eigenwerte und Eigenvektoren der Kovarianzmatrix: Die Eigenwerte und Eigenvektoren der Kovarianzmatrix werden berechnet. Die Eigenvektoren repräsentieren die Richtungen im Raum der Variablen, entlang derer die Daten am meisten variieren. Die Eigenwerte geben die Varianz an, die durch jeden Eigenvektor erklärt wird.
3. Schritt 3: Auswahl der Hauptkomponenten: Die Hauptkomponenten werden basierend auf den Eigenwerten ausgewählt, wobei diejenigen mit den größten Eigenwerten bevorzugt werden.

Singulärwertzerlegung (SVD) der zentrierten Datenmatrix:

1. Schritt 1: Zentrierung der Daten: Die Datenmatrix wird zentriert, indem der Mittelwert jeder Spalte von den Daten subtrahiert wird. Dies ist wichtig, um sicherzustellen, dass die HKA nicht von den Mittelwerten der Daten beeinflusst wird.
2. Schritt 2: Berechnung der Singulärwertzerlegung: Die SVD wird auf die zentrierte Datenmatrix angewendet. Sie zerlegt die Datenmatrix in drei Matrizen: eine linke singuläre Vektormatrix, eine Diagonalmatrix mit singulären Werten und eine rechte singuläre Vektormatrix.
3. Schritt 3: Auswahl der Hauptkomponenten: Die Hauptkomponenten werden aus den singulären Werten und singulären Vektoren ausgewählt. Die singulären Werte geben die Streuung der Daten in den entsprechenden Richtungen an, während die singulären Vektoren die Richtungen im Raum der Variablen repräsentieren.

Gegenüberstellung der beiden Methoden kann man in folgenden Punkten zusammenfassen:

- Beide Methoden führen zu denselben Hauptkomponenten.
- Die Eigenwertzerlegung der Kovarianzmatrix ist etwas intuitiver und wird häufiger verwendet, da sie die Varianzstruktur der Daten explizit berücksichtigt.
- Die Singulärwertzerlegung der zentrierten Datenmatrix ist numerisch stabiler, weil sie keine explizite Berechnung der Kovarianzmatrix erfordert, die potenziell numerisch instabil sein kann, insbesondere wenn die Variablen in den Daten stark korreliert sind oder die Kovarianzmatrix singulär ist.
- Die Singulärwertzerlegung der zentrierten Datenmatrix ist effizienter in der Berechnung, weil sie durch verschiedene numerische Algorithmen berechnet werden kann, wie zum Beispiel die Jacobi-SVD, die QR-SVD und die Lanczos-SVD. Diese Algorithmen sind effizient und skalierbar für große Datenmengen [6],[7].
- Die SVD wird oft in großen Datensätzen verwendet, da sie weniger anfällig für numerische Fehler ist.
- Die Eigenwertzerlegung der Kovarianzmatrix erfordert die Berechnung der Kovarianzmatrix, was bei großen Datensätzen speicher- und rechenintensiv sein kann.

- Insgesamt bieten beide Methoden wirksame Möglichkeiten, um die Hauptkomponenten zu identifizieren und für die Dimensionsreduktion in Datenanalysen zu verwenden. Die Wahl zwischen ihnen hängt oft von der Größe des Datensatzes, der verfügbaren Rechenressourcen und den spezifischen Anforderungen des Problems ab.

Vorzeichen-Unbestimmtheit

Wie bereits angedeutet weist die Hauptkomponentenanalyse eine Vorzeichen-Unbestimmtheit auf, die darin besteht, dass positive und negative Richtungen der Achsen, die die Hauptkomponenten beschreiben, beliebig gewählt werden können. Dies zeigt sich in der Vorzeichenwahl der Eigenvektoren der Kovarianzmatrix bzw. der Vorzeichenwahl der Singularvektoren der zentrierten Datenmatrix bei der Singulärwertzerlegung.

Mathematisch betrachtet stellt jede Vorzeichenwahl eine gültige Lösung zur HKA dar, aber für die praktische Anwendbarkeit hat eine einheitliche Wahlvorschrift der Vorzeichen der Hauptkomponenten viele Vorteile, die in folgenden Stichpunkten zusammengefasst werden können.

- Interpretationserleichterung: Eine konsistente Orientierung der Hauptkomponenten erleichtert die Interpretation der Ergebnisse der HKA und reduziert potenzielle Unstimmigkeiten oder Verwirrungen.
- Vergleichbarkeit: Die Vergleichbarkeit zwischen verschiedenen Analysen oder Datensätzen wird verbessert, da die Richtung der Hauptkomponenten eindeutig festgelegt ist.
- Anwendbarkeit in weiterführenden Analysen: Eine einheitliche Orientierung der Hauptkomponenten kann die Anwendbarkeit in weiterführenden Analysen erhöhen, wie z. B. in der Clusteranalyse oder bei der Feature-Extraktion für maschinelles Lernen. Dadurch wird die Kompatibilität mit anderen Analysetechniken verbessert.

In vielen Anwendungen und Softwarelösungen, die auf einer HKA basieren, ist die Vorzeichen-Unbestimmtheit nicht essenziell und hat keinen signifikanten Einfluss auf das Ergebnis oder die Anwendung, da die HKA in erster Linie darauf abzielt, die Datenstruktur zu erfassen und die wichtigsten Muster oder Variationen in den Daten zu identifizieren. Dabei ist die genaue Richtung der Hauptkomponenten, die durch die Vorzeichen bestimmt wird, oft weniger relevant als ihre relative Bedeutung und Beitrag zur Gesamtvariation der Daten. Aus diesem Grund wird in vielen Anwendungen und Softwarelösungen, die auf Hauptkomponentenanalyse basieren, die Vorzeichen-Unbestimmtheit entweder ignoriert oder durch verschiedene Techniken behandelt, wie

beispielsweise durch die Festlegung von Konventionen für die Vorzeichen oder durch Verfahren wie die Hauptkomponentenrotation, um konsistente Interpretationen zu gewährleisten. Letztendlich liegt der Fokus darauf, die Dimensionalität der Daten zu reduzieren und bedeutende Muster oder Strukturen zu extrahieren, unabhängig von der spezifischen Richtung der Hauptkomponenten.

Um die Vorzeichen-Problematik zu lösen, wird das Verfahren aus [3] behandelt. Es wird hierbei von einer Singulärwertzerlegung mit willkürlich gesetzten Vorzeichen der Singulärvektoren ausgegangen. Das Verfahren beinhaltet eine Vorgehensweise wie man die Vorzeichen aller Eigenvektoren definiert bzw. prüft und gegebenenfalls korrigiert. Der Algorithmus ist im Bild 2 dargestellt. Inputvariablen sind hierbei die Matrix $\boldsymbol{X}$ und ihre Singulärwertzerlegung. Output ist die Singulärwertzerlegung mit „korrekten" Vorzeichen.

Um das Vorzeichen eines Singularvektors zu definieren, wird hierbei angenommen, dass ein Singularvektor das gleiche Vorzeichen haben soll, wie die Vorzeichen der meisten Datenpunkte, die er repräsentiert, bzw. geometrisch betrachtet soll er in die Richtung der meisten Datenvektoren der Datenmatrix gerichtet sein. Das Vorzeichen wird dabei aus den Vorzeichen der Skalarprodukte des Singularvektors mit allen Datenvektoren bestimmt. Die Datenvektoren haben unterschiedliche Orientierung, aber es macht sowohl intuitiv als auch praktisch betrachtet Sinn, die Richtung der meisten Datenvektoren zu wählen.

Angenommen, man möchte das Vorzeichen des k-ten linken Singularvektors bestimmen, dann wäre $(\boldsymbol{u}_k)^{\boldsymbol{T}}(\boldsymbol{x}_j)>0$ für j=1,…,J, wenn der Singularvektor richtig zu den Spaltenvektoren der Datenmatrix X orientiert ist. Wir wählen die Vorzeichen so, dass die folgende Summe maximiert wird:

$$s = \sum_{j=1}^{J} sign(\boldsymbol{u_k^T x}_j)(\boldsymbol{u_k^T x}_j)^{\boldsymbol{2}}$$

wobei $\boldsymbol{x}_j$ die j-te Spalte des j-ten Spalte der Matrix $\boldsymbol{X}$ ist.

Auf die gleiche Weise unter Verwendung der Zeilenvektoren der Matrix $\boldsymbol{X}$ werden auch die Vorzeichen der rechten Singularvektoren bestimmt. Falls die beiden optimalen Vorzeichen einander widersprechen, wählt man das Vorzeichen von dem Singularvektor, der betragsmäßig größer ist.

Der dargestellte Algorithmus beinhaltet auch die Subtraktion der zusätzlichen Komponenten vor der Bestimmung des Vorzeichens einer gegebenen Komponente: $\boldsymbol{Y} = \boldsymbol{X} - \sum_{m=1,j\neq K}^{K} \boldsymbol{\sigma}_m \boldsymbol{u}_m \boldsymbol{v_m^T}$. Dies ist im Standard-SVD nicht erforderlich, aber nützlich, wenn die Komponenten korreliert sind. Es kann erwartet werden, dass der Algorithmus funktioniert, wenn die Beträge der Skalarprodukte nicht nahe Null liegen. Wenn die Beträge nahe Null liegen, wird das Vorzeichen beliebig, im Wesentlichen, weil die Vektoren gleichmäßig in alle Richtungen zeigen. Dies wird teilweise im Algorithmus behoben, indem die kombinierte Größe sowohl der linken als auch der rechten singulären Vektoren betrachtet wird, aber im Extremfall wird das Vorzeichen beliebig sein.

SignFlip Function

Input: $\mathbf{X} \in \mathbb{R}^{I \times J}$ and its possibly truncated singular value decomposition (**U**, **V**, **S**)

Output: **U'** and **V'** (left and right singular vectors with appropriate signs)

(Step 1) for each left singular vector, k=1, 2…K and for $\mathbf{y}_j$ being the jth column of **Y**

$$\mathbf{Y} = \mathbf{X} - \sum_{m=1,\ m \neq k}^{K} \sigma_m \mathbf{u}_m \mathbf{v}_m^T$$

$$\text{Let } s_k^{\text{left}} = \sum_{j=1}^{J} \text{sign}(\mathbf{u}_k^T \mathbf{y}_j)\ (\mathbf{u}_k^T \mathbf{y}_j)^2$$

endfor

(Step 2) for each right singular vector, k=1, 2…K and
for $\mathbf{y}_i$ being the ith transposed row of **Y**

$$\mathbf{Y} = \mathbf{X} - \sum_{m=1,\ m \neq k}^{K} \sigma_m \mathbf{u}_m \mathbf{v}_m^T$$

$$\text{Let } s_k^{\text{right}} = \sum_{i=1}^{I} \text{sign}(\mathbf{v}_k^T \mathbf{y}_i)(\mathbf{v}_k^T \mathbf{y}_i)^2$$

endfor

(Step 3) for each singular vector, k=1, 2…K
if $(s_k^{\text{left}})(s_k^{\text{right}}) < 0$ then
if $s_k^{\text{left}} < s_k^{\text{right}}$ then
$s_k^{\text{left}} = -s_k^{\text{left}}$
else
$s_k^{\text{right}} = -s_k^{\text{right}}$
endif
endif
$\mathbf{u'}_k = \text{sign}(s_k^{\text{left}})\mathbf{u}_k$
$\mathbf{v'}_k = \text{sign}(s_k^{\text{right}})\mathbf{v}_k$
endfor

Bild 2: Algorithmus zur Vorzeichenkorrektur bei der SVD nach [2]

Eigenfaces

Ein Beispiel für die Anwendung HKA ist die Gesichtserkennung in der Bildverarbeitung.

Nehmen wir an, wir haben einen Datensatz von Gesichtsbildern, bestehend aus einer großen Anzahl von Bildern von verschiedenen Personen. Jedes Bild ist eine Matrix von Pixelwerten.

Die HKA wird verwendet, um die Hauptkomponenten (Eigenfaces) aus den Gesichtsbildern zu extrahieren. Dazu werden die Bilder in Vektoren umgewandelt und zu einer Datenmatrix zusammengefügt. Die HKA wird dann auf diese Datenmatrix angewendet, um die Hauptkomponenten zu berechnen.

Die Hauptkomponenten, die die größte Varianz in den Gesichtsbildern repräsentieren, werden beibehalten, während die weniger bedeutenden Komponenten weggelassen werden. Dies ermöglicht eine Reduzierung der Dimensionalität der Daten, was die Effizienz der weiteren Verarbeitungsschritte verbessert.

Die Eigenfaces bilden einen Merkmalsraum, in dem jedes Gesicht als eine Kombination von Gewichtungen für die verschiedenen Eigenfaces dargestellt werden kann. Durch die Projektion der Gesichter auf diesen Merkmalsraum können sie effektiv beschrieben und verglichen werden.

Um ein Gesicht zu erkennen, wird das Eingangsbild in den Merkmalsraum projiziert, indem die Gewichtungen der Eigenfaces berechnet werden. Anschließend wird ein Vergleich mit den gespeicherten Gesichtsdaten durchgeführt, um das am besten übereinstimmendes Gesicht zu identifizieren.

Die HKA in der Gesichtserkennung ist ein leistungsstarkes Werkzeug, das eine effiziente Repräsentation von Gesichtsbildern ermöglicht und gleichzeitig die Dimensionalität der Daten reduziert. Es wird oft in verschiedenen Anwendungen wie biometrischer Authentifizierung, Überwachungssystemen und Sicherheitsanwendungen eingesetzt.

In [3] wird die Anwendung der HKA in Eigenfaces mit und ohne Vorzeichen-Korrektur-Funktion (Sign Flip Function) untersucht. Es wird festgestellt, dass die Anwendung der Vorzeichen-Korrektur-Funktion eine bessere Bildqualität in Form von schärferen Eigenfaces liefert.

Bewertung der Methoden zur Festlegung der Vorzeichen

Wie dargestellt, behandelt das Verfahren aus [3] eine Lösung zur Vorzeichen-Problematik, die auf einem heuristischen Ansatz und einer geometrischen Begründung beruht. Am berichteten Beispiel mit Eigenfaces ist erkennbar, dass der Algorithmus zu einer sinnvollen Wahl der Vorzeichen führt und die Qualität der Ergebnisse in der Anwendung verbessert und besser interpretierbar macht.

Wir bewerten die Komplexität dieses Verfahrens. Für jeden der k linken Singularvektoren werden k Skalarprodukte berechnet und genauso für rechte Singularvektoren. Es sind somit jeweils k*n und

k*p Multiplikationen durchzuführen, wobei *k* auch der Anzahl der Hauptkomponenten entspricht. Folglich ist die Komplexität der Vorzeichen-Korrektur-Funktion $O(k*\min\{n,p\})$.

Zusammenfassung

Es wurden zwei Verfahren zur Hauptkomponentenanalyse dargestellt und analysiert. Die Problematik der Vorzeichen-Unbestimmtheit wurde beschrieben und analysiert und ein aus der Literatur bekanntes Verfahren zur Wahl der Vorzeichen erläutert. Die Methode wurde bewertet hinsichtlich ihrer Rechenkomplexität.

Ausblick

Zur besseren Veranschaulichung der HKA können weitere Beispiele aus der Mustererkennung und Bildverarbeitung herangezogen werden. Zur Lösung der Vorzeichen-Unbestimmtheit-Problematik kann eine tiefere Analyse der zugrundeliegenden mathematischen Konzepte wie Koordinatentransformation und Basiswechsel durchgeführt werden, um einen Zusammenhang zwischen der Orientierung der Singulärvektoren und gegebenen Datenvektoren interpretierbar nachzuvollziehen.

Literatur Quellen

[1] Andreas Handl, Torben Kuhlenkasper, "Multivariate Analysemethoden", Springer Verlag, 3. Auflage, 2017

[2] Ian T. Jolliffe, Jorge Cadima, „Principal component analysis: a review and recent developments", Philosophical Transactions: Mathematical, Physical and Engineering Sciences , 13, April 2016, Vol. 374, No. 2065, pp. 1-16

[3] Bro, R.; Acar, E.; Kolda, Tamara G., „Resolving the sign ambiguity in the singular value decomposition", Journal of chemometrics, 2008, Vol.22 (2), p.135-140

[4] Jolliffe, Ian T. "Principal component analysis" Springer-Verlag [1986]

[5] Brunton, Steven L. and Kutz, J. Nathan "Data Driven Science & Engineering Machine Learning, Dynamical Systems, and Control" [2019]

[6] Horn, Roger A. and Johnson, Charles R. "Matrix Analysis" [2012]

[7] Strang, Gilbert "Linear Algebra and Its Applications", fourth edition [2005]