**Training Fiche Template**

|  |  |  |
| --- | --- | --- |
| **Titel** | Hauptkomponentenanalyse (Principal Components Analysis, PCA) | |
| **Schlagworte (meta tag)** | PCA, Korrelation, quantitative Variablen, erklärte Varianz, Eigenwerte. | |
| **Sprache** | Deutsch | |
| **Zielsetzungen / Lernziele / Lernergebnisse** | **Dieses Modul zielt darauf ab, die Hauptkomponentenanalyse (HKA) einzuführen und zu erklären.**  **In diesem Modul lernen wir:**  **- Die Logik der PCA zu verstehen.**  **- Die Anforderungen zu kennen.**  **- Eine PCA durchzuführen.**  **- Eine PCA in R mit dem Paket FactorMineR durchzuführen.** | |
| **Training Kurs:** | | |
| **Datenwissenschaftliche Kompetenz** | |  |
| **Data Visualisierung und visuelle Datenanalyse** | | **X** |
| **Einführung in die Datenwissenschaft für Geistes- und Sozialwissenschaften** | |  |
| **Data Science for good** | |  |
| **Data Journalism and Storytelling** | |  |
| **Beschreibung** | In diesem Modul lernen wir die multidimensionale Analysetechnik namens Hauptkomponentenanalyse (HKA oder PCA) kennen. Das Ziel der PCA ist es, die Dimensionalität eines untersuchten Phänomens zu reduzieren, während gleichzeitig die darin enthaltene Information bewahrt wird. Die Technik ist anwendbar auf Phänomene, die mit **quantitativen** **Variablen** gemessen werden, und unterscheidet sich damit von anderen Techniken zur Dimensionsreduktion wie der einfachen Korrespondenzanalyse (CA) oder der multiplen Korrespondenzanalyse (MCA), die für die Analyse qualitativer Variablen entwickelt wurden.  Der letzte Teil des Moduls beschäftigt sich mit der Anwendung von PCA mit der Programmiersprache R. | |
| **Contents arranged in 3 levels** | **1. Einleitung**  Die Hauptkomponentenanalyse (PCA) ist eine statistische multivariate Analysetechnik zur Dimensionsreduktion. In der Praxis wird sie eingesetzt, wenn in einem Datensatz viele korrelierte Variablen vorhanden sind, um deren Anzahl zu reduzieren und dabei so wenig wie möglich an Information zu verlieren.  Die PCA hat genau das Ziel, die Varianz zu maximieren, indem sie das Gewicht berechnet, das jeder Ausgangsvariable zugeordnet werden soll, um sie in eine oder mehrere neue Variablen (genannt Hauptkomponenten) zu konzentrieren. Diese Hauptkomponenten werden als lineare Kombination der Ausgangsvariablen gebildet.  **2. Anforderungen für die Durchführung der Hauptkomponentenanalyse (PCA)**  Um zu verstehen, ob es sinnvoll ist, eine Hauptkomponentenanalyse (PCA) durchzuführen, ist es wichtig, die zu verwendenden Variablen zu analysieren, um einige ihrer Eigenschaften klar zu haben. Insbesondere müssen die Variablen folgende Anforderungen erfüllen:  *- Die Variablen müssen quantitativ sein*  PCA ist nur gültig, wenn die Variablen numerisch sind. Falls unterschiedliche Maßeinheiten vorliegen, müssen die Variablen vor der Durchführung standardisiert werden. In einigen Fällen wird PCA jedoch auch für "Likert-Skalen"-Variablen und "binäre Variablen" verwendet. Obwohl die Ergebnisse numerisch ähnlich sind, ist es in diesen Fällen empfehlenswert, alternative Methoden zu verwenden.  *- Es muss eine lineare Korrelation zwischen den Variablen vorliegen*  Bei der Durchführung einer PCA ist der erste Schritt die Berechnung der Varianz-/Kovarianzmatrix oder der Pearson-Korrelationsmatrix. PCA ist eine Technik, die verwendet werden kann, wenn die Annahmen des Pearson-Korrelationskoeffizienten erfüllt sind. Die Pearson-Korrelationskoeffizienten geben Auskunft über die Richtung und Stärke der linearen Beziehung zwischen Phänomenen. Zur Interpretation gilt: Je näher der Koeffizient bei Null liegt, desto schwächer ist die Beziehung. Je näher er -1 oder +1 ist, desto stärker ist die Beziehung. Bei PCA sind akzeptable Werte für diesen Indikator R>0,3 oder R<-0,3. Wenn eine Variable Korrelationskoeffizienten hat, die sehr nahe bei 0 für alle anderen Variablen liegen, sollte diese Variable nicht in die PCA einbezogen werden. Eine erzwungene Kombination dieser Variable mit anderen führt zu einem sehr hohen Informationsverlust, den es zu vermeiden gilt.  *- Es sollten keine Ausreißer (outliers) vorhanden sein*  Wie bei allen varianzbasierten Analysen können einzelne Ausreißer vor allem dann die Ergebnisse beeinflussen, wenn sie sehr groß sind und die Stichprobengröße klein ist.  Um diesem Problem entgegenzuwirken, ist es sinnvoll, Boxplots oder Streudiagramme (scatterplots) zu erstellen, aus denen lineare Beziehungen zwischen Variablenpaaren abgeleitet werden können  *- Relativ große Stichprobengröße*  Es gibt keinen eindeutigen Schwellenwert, aber im Allgemeinen ist es ratsam, mindestens 5-10 statistische Einheiten für jede Variable zu haben, die in die PCA einbezogen werden soll. Wenn du zum Beispiel versuchst, 10 Variablen mit neuen Komponenten zusammenzufassen, wäre es ratsam, eine Stichprobe von mindestens 150 Beobachtungen zu haben.  **3. Durchführung einer Hauptkomponentenanalyse (PCA)**  3.1 Nachdem wir die Anforderungen des Datensatzes überprüft und festgestellt haben, dass die Variablen die richtigen Merkmale für die Durchführung der Hauptkomponentenanalyse aufweisen, können wir die eigentliche Analyse durchführen. Dazu müssen wir folgende Schritte ausführen:  3.2 Wir überprüfen die Angemessenheit der Stichprobe durch zwei Tests:  - Der *Kaiser-Meyer-Olkin-Test (KMO)* überprüft, ob die betrachteten Variablen tatsächlich konsistent für die Verwendung einer Hauptkomponentenanalyse sind. Dieser Index kann Werte zwischen 0 und 1 annehmen. Damit eine Hauptkomponentenanalyse sinnvoll ist, sollte der Wert mindestens **größer als 0.5** sein*.*  Dieser Index kann als Gesamtwert für alle in der PCA enthaltenen Variablen berechnet werden.  - Der *Bartlett-Test auf Sphärizität* ist ein Hypothesentest, bei dem die Nullhypothese besagt, dass die Korrelationsmatrix mit der Identitätsmatrix übereinstimmt. Wenn dies der Fall ist, würde es keinen Sinn machen, eine PCA durchzuführen, denn das würde bedeuten, dass die Variablen überhaupt nicht linear miteinander zusammenhängen. Wie bei allen Hypothesentests ist der entscheidende Wert, um zu bestimmen, ob die Nullhypothese abgelehnt wird oder nicht, der p-Wert. In diesem Fall muss für das Modell als gültig betrachtet werden ein **p-Wert kleiner als 0.05** erreicht werden. In diesem Fall kann die Nullhypothese tatsächlich mit einem Signifikanzniveau von 5% abgelehnt werden.  3.3 Wir extrahieren die Hauptkomponenten:  Der entscheidende Teil der PCA besteht darin, die angemessene Anzahl der Faktoren festzulegen, die die Ausgangsvariablen am besten repräsentieren können.  Um dieses Konzept besser zu verstehen, stellen wir uns vor, der Datensatz sei eine uns unbekannte Stadt und jede Hauptkomponente sei eine Straße in dieser Stadt. Wenn wir diese Stadt kennenlernen möchten, wie viele Straßen würden wir besuchen? Wir würden wahrscheinlich mit der zentralen Straße (der ersten Hauptkomponente) beginnen und dann andere Straßen erkunden. Aber wie viele Straßen sollten es sein?  Um sagen zu können, dass wir eine Stadt gut genug kennen, hängt die Anzahl der zu besuchenden Straßen von der Größe der Stadt und der Ähnlichkeit oder Unterschiedlichkeit der Straßen ab. Ähnlich verhält es sich mit der Anzahl der zu extrahierenden Komponenten, die davon abhängt, wie viele Variablen wir in unsere Hauptkomponentenanalyse einbeziehen und wie ähnlich sie sich sind. Tatsächlich benötigen wir bei stärkerer Korrelation zwischen den Variablen weniger Hauptkomponenten, um eine gute Kenntnis der Ausgangsvariablen zu erlangen. Umgekehrt benötigen wir bei geringerer Korrelation eine größere Anzahl von Hauptkomponenten, um genaue Informationen über den Datensatz zu erhalten.  Es gibt im Wesentlichen zwei Kriterien zur Auswahl der Anzahl der Komponenten: Eigenwerte größer als 1 und parallele Analyse.  *Eigenwerte größer als 1*  Gemäß dieser Regel werden diejenigen Komponenten ausgewählt, denen ein Eigenwert größer als 1 zugeordnet ist. Der Eigenwert ist eine Zahl, die die Varianz angibt, welche durch die Komponente erklärt wird. Da die anfängliche Varianz, die durch jede einzelne Variable erklärt wird, gleich 1 ist, würde es keinen Sinn ergeben, eine Komponente (die eine Kombination von Variablen darstellt) mit einer Varianz von weniger als 1 auszuwählen. Ein hoher Eigenwert entspricht einer größeren Varianz, und Software wie SPSS oder R erstellen dafür eine Tabelle mit abnehmenden Werten. Daher wird der erste Eigenwert immer dem wichtigsten Faktor zugeordnet sein.  *Anteil der erklärten Varianz*  Gemäß diesem Kriterium müssen die extrahierten Komponenten sicherstellen, dass mindestens 70% der Gesamtvarianz der Ausgangsvariablen nicht verloren gehen. Darüber hinaus sollte jede einzelne extrahierte Komponente einen signifikanten Anstieg der Gesamtvarianz bringen (zum Beispiel mindestens 5% oder 10% mehr erklärte Varianz).  *Scree-plot*  Diese Methode basiert auf einem Diagramm, auf dem die Eigenwerte auf der vertikalen Achse und alle möglichen zu extrahierenden Komponenten auf der horizontalen Achse dargestellt werden (die daher in der Anzahl den Ausgangsvariablen entspricht). Durch das Verbinden der Punkte entsteht eine gestufte Linie, die an einigen Stellen eine konkave Form und an anderen Stellen eine konvexe Form aufweisen wird.  Wie Sie auf dem Diagramm sehen können, werden die Komponenten auf der x-Achse (Components number) aufgelistet, während die Eigenwerte (Eigenvalues) auf der y-Achse liegen. Wenn die Kurve auf diesem Diagramm einen "Ellenbogen" macht, zieht man eine Linie und nimmt nur jene Faktoren, die oberhalb liegen.  Aus dem obigen Diagramm können wir zum Beispiel sehen, dass die Anzahl der Punkte über dem Ellenbogen 2 beträgt.  Der letzte Teil der PCA besteht darin, den einzelnen Hauptkomponenten, die gefunden wurden, einen Namen zu geben.  **4. PCA mit R**  Mit statistischer Software (wie SPSS, Jamovi und R) ist die Durchführung einer PCA sehr einfach. Ein paar Klicks genügen, um eine Ausgabe zu erhalten, die interpretiert werden kann. Es gibt daher keine Software, die den anderen vorzuziehen ist, da es sich um eine weit verbreitete Technik handelt und alle statistischen Programme eine einfache Durchführung ohne manuelle Berechnungen ermöglichen.  Der gesamte Prozess zur Implementierung von PCA in R wird in der Powerpoint-Präsentation dargestellt, die diesem Modul beigefügt ist und behandelt folgende Themen:   * Durchführung aller Schritte, die auf Matrizen, geometrischen und statistischen Beweisen basieren; * Verwendung des direkten PCA-Befehls des FactoMineR-Pakets.   In diesem Modul werden wir lediglich das FactoMineR-Paket vorstellen.  FactoMineR kann eine Hauptkomponentenanalyse durchführen, indem es die Dimensionalität der multivariaten Daten auf zwei oder drei reduziert, die dann grafisch mit einem minimalen Informationsverlust dargestellt werden können. Dies kann mit einem einzigen Befehl, nämlich **PCA**, erreicht werden. Dabei wird das Matrixobjekt der Analyse in Klammern eingeschlossen.  X <- as.matrix(DATASET)  library(FactoMineR)  res.pca = PCA(DATASET)  Mit dem Befehl "summary" können wir die Bedeutung der Komponenten in Bezug auf Standardabweichung, Anteil der erklärten Varianz und kumulative erklärte Varianz sowohl für Individuen als auch für Variablen sehen.  summary(res.pca)  Mit der Anweisung *head*  head(ris.pca$eig)  können wir die Bedeutung der Eigenwerte berechnen. Der Befehl gibt uns die Werte der Eigenwerte, den Prozentsatz der erklärten Varianz und die kumulative erklärte Varianz für jede Variable.  Hier ist ein Beispiel dafür, wie die Ausgabe in R aussehen könnte:  Um das Scree-Plot der Eigenwerte zeichnen zu können, verwenden wir den Befehl "barplot" und fügen das Analyseobjekt in Klammern ein:  barplot(res.pca$eig[,1], main="Eigenvalues’ scree-plot")  Der Befehl "main" wird verwendet, um den Titel des Diagramms anzugeben.  Hier ist ein Beispiel dafür, wie die Ausgabe in R aussehen könnte:  Ein weiteres nützliches Paket für PCA (obwohl wir es in diesem Modul nicht behandeln) ist *factoextra*. Es bietet einige benutzerfreundliche Funktionen zum Extrahieren und Visualisieren der Ergebnisse von multivariaten Analysen, einschließlich PCA (Hauptkomponentenanalyse), CA (einfache Entsprechungsanalyse), MCA (multiple Entsprechungsanalyse), MFA (multiple Faktorenanalyse) und HMFA (hierarchische multiple Faktorenanalyse). | |
| **Self-assessment (multiple choice queries and answers)** | 1. Die Hauptkomponentenanalyse zielt darauf ab:  A) statistische Einheiten nach ihrer Distanz zu aggregieren  B) die Dimensionalität eines komplexen Phänomens zu reduzieren  C) einen Datensatz zu beschreiben  2. Die Ausgangsdatenmatrix einer PCA muss auf … basieren:  A) qualitativen Daten  B) standardisierten Daten  C) quantitativen Daten  3. Die extrahierten Komponenten in der Hauptkomponentenanalyse:  A) sind lineare Kombinationen der Ausgangsvariablen  B) haben die Eigenschaft der Gleichverteilung  C) haben alle Eigenwerte größer als 1  4. Mit wie vielen Dimensionen würden wir das folgende Phänomen erklären?   1. Eine 2. Zwei 3. Drei | |
| **Resources (videos, reference link)** | Pozzolo P., *Analisi delle componenti principali: da dove partire*, <https://paolapozzolo.it/analisi-delle-componenti-principali-criteri/>  Gilardone A., *Analisi delle componenti principali: 7 passaggi da eseguire* <https://adrianogilardone.com/analisi-delle-componenti-principali/>  Gilardone A., <https://www.youtube.com/watch?v=OksC-g4K2gY>  Vardanega A., L’Analisi in componenti principali  <https://www.agnesevardanega.eu/wiki/r/analisi_esplorativa/analisi_in_componenti_principali>  Zakaria Jaadi, *A Step-by-Step Explanation of Principal Component Analysis (PCA)*, <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>  Ian T. Jolliffe and Jorge Cadima, *Principal component analysis: a review and recent developments*, <https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202>  Science Snippets Blog, *What Is Principal Component Analysis (PCA) and How It Is Used?*, 2020 <https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186> | |
| **Related material** |  | |
| **Related PPT** |  | |
| **Bibliography** |  | |
| **Provided by** | [UNISALENTO/DEMOSTENE CENTRO STUDI] | |