Nichtlineare Projektion mit dem Sammon-Verfahren

Datenvisualisierung in der Statistik

1. Zielsetzung / Nutzen

2-, 3- oder auch 4-dimensionale Datensätze können für den Menschen noch anschaulich z.B. in Streudiagrammen dargestellt werden. Kommen wir aber zu höher dimensionalen Datensätzen, benötigen wir Projektionsverfahren, um diese für den Menschen entsprechend zu veranschaulichen. Die Sammon-Projektion ist eine hierfür gängige nichtlineare Methode, welche einen höher in einen niedriger dimensionalen Datensatz transformiert.
Im Gegensatz zu linearen Methoden der Dimensionsreduzierung, wie z.B. die Hauptkomponentenanalyse, repräsentiert die Sammon-Abbildung keine explizite Transformationsfunktion, sondern gibt einen Vergleichswert an, wie gut ein transformierter niedriger dimensionaler Datensatz die Struktur des originalen Datensatzes repräsentiert.
Das Sammon-Verfahren versucht also nicht eine optimale Abbildung des Original-Datensatzes zu finden, sondern einen neuen niedriger dimensionalen Datensatz, dessen Struktur möglichst ähnlich der Struktur des originalen Datensatzes ist.

2. Methodik

Idee: Abbildung eines Datensatzes X auf einen anderen Datensatz Y, sodass die Abstände zwischen den Punkten angenähert werden.

➔ **Initialisierung** durch $Y = (1, 2, 3, 4)$

1. Aufstellung der Distanzmatrizen D^x und D^y

$$D^y = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{pmatrix}$$

2. Minimierung des Fehlers zwischen D^x und D^y mithilfe der *Sammon-Fehlerfunktion*

$$E_3 = \frac{1}{3 \cdot 1 + 2 \cdot \sqrt{2} + \sqrt{5}} \cdot \left(2 \cdot \frac{\left(2 - \sqrt{2}\right)^2}{\sqrt{2}} + \frac{\left(3 - \sqrt{5}\right)^2}{\sqrt{5}} \right) \approx 0.0925$$

3. Anwendung des Gradientenverfahrens zur Bestimmung des Minimums von E_3

$$\frac{\partial E_3}{\partial y_1} = \frac{2}{3 \cdot 1 + 2 \cdot \sqrt{2} + \sqrt{5}} \cdot \left(-\frac{2 - \sqrt{2}}{\sqrt{2}} - \frac{3 - \sqrt{5}}{\sqrt{5}} \right) \approx -0.1875$$

   3.1. Minimaler Fehler liefert nächste Annäherung an die Projektion

4. Wiederhole die Schritte 1 bis 3

Projektion: nach einer Iteration: $Y \approx (1.1875, 2.1027, 2.8973, 3.8125), E_3 \approx 0.0925$
nach zehn Iterationen: $Y \approx (1.3058, 2.1359, 2.8641, 3.6942), E_3 \approx 0.0212$

3. Anwendung an Beispielen

Bouquet of Circles
In folgendem 6-dimensionalen Beispiel schneiden sich 3 Ebenen, wobei jede Ebene senkrecht zu den anderen beiden Ebenen ist, an einem gemeinsamen Punkt. Auf jeder der 3 Ebenen befinden sich Datenpunkte, welche jeweils einen Kreis bilden

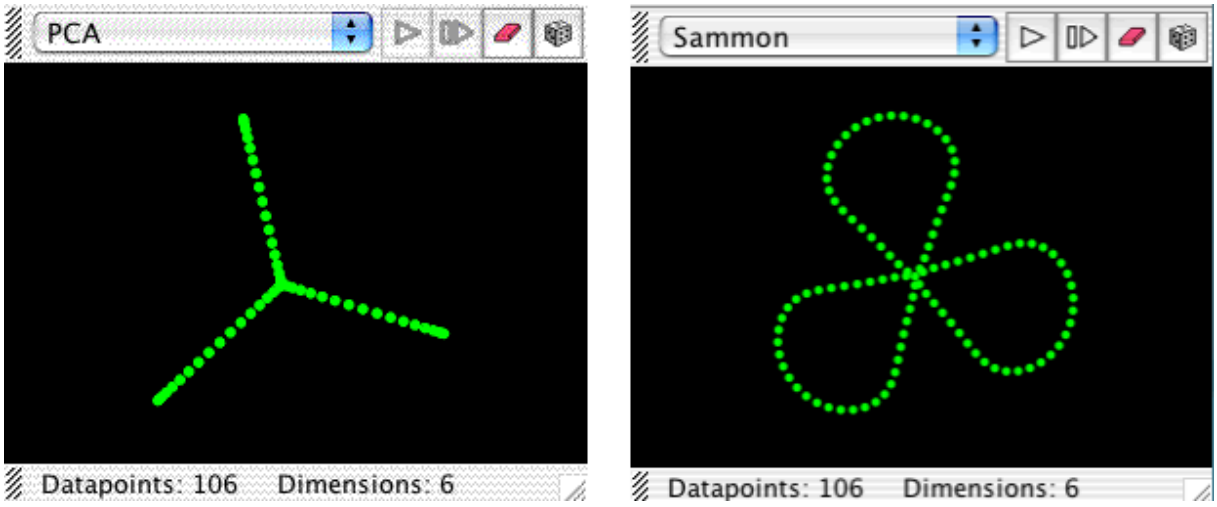

Abbildung 1: Hauptkomponentenanalyse und Sammon-Abbildung (Quelle: HiSee)

Die Abbildung links zeigt den Datensatz, welcher mit der Hauptkomponentenanalyse auf seine Achsen projiziert wird, um die sich die Datenpunkte am stärksten häufen. Deshalb werden alle 3 Ebenen bzw. Kreise sozusagen von der Seite dargestellt, wodurch sie als eine Linie in der Abbildung zu erkennen sind. Dadurch wird die Symmetrie des Datensatzes erkennbar, da alle drei Linien den gleichen Winkel zueinander haben.
Auf der rechten Seite ist der Datensatz mit dem Sammon-Verfahren projiziert, mit welchem man nun zusätzlich zur Symmetrie auch die Struktur des Datensatzes erkennen kann. So werden auch die Kreise auf den Ebenen sichtbar.

4. Literatur

Ausführlichere Informationen und weitere Beispiele finden Sie hier:

- J.W. Sammon (1969). *„A nonlinear mapping for data structure analysis"*. IEEE Transactions on Computers. 18: 401, 402,403–409.
- Thomas A. Runkler (2015). „*Data Mining – Modelle und Algorithmen intelligenter Datenanalyse*“. In Computational Intelligence, S. 47-51. 2. Auflage. Springer Vieweg. Berlin 2015.
- Elzbieta Pękalska, Dick de Ridder, Robert P.W. Duin and Martin A. Kraaijveld (1999). *„A new method of generalizing Sammon mapping with application to algorithm speed-up“.* Proc. 5th Annual Conference of the Advanced School for Computing and Imaging (ASCI1999).
- Scott Hotton and Jeff Hoshimi. „*HiSee – Projection Algorithms*“. Online: http://hisee.sourceforge.net (Abrufdatum: 08.03.2018).
- Paul Henderson (2012). „*Sammon Mapping*“. Online: http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/AV0910/henderson.pdf (Abrufdatum: 08.03.2018).