UNIKASSEL
VERSITÄT

mrt
Mess- und Regelungstechnik
Univ.-Prof. Dr.-Ing. Andreas Kroll

Oberseminar

Principal Component Analysis (PCA) und Partial Least Squares (PLS) im Big-Data-Kontext

Alexander Rehmer

Die Methoden der Hauptkomponentenanalyse (Principal Component Analysis, PCA) und der Partial Least Squares (PLS) sind Bestandteile der multivariaten Statistik. Sie dienen dazu, große Datensätze vereinfacht darzustellen und Zusammenhänge erkennbar zu machen. Für ein Forschungsprojekt kommt die Anwendung beider Methoden in Frage, sofern sie auf Big-Data-Probleme anwendbar sind.

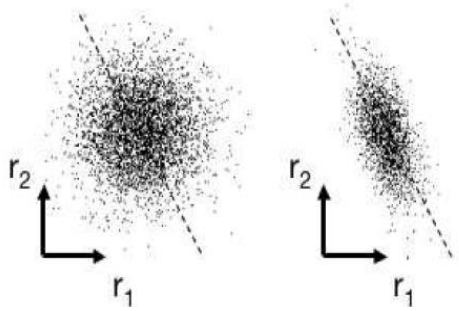

Abb.: Beispieldaten mit versch. Redundanzen

Im Rahmen dieser Arbeit soll zunächst eine Literaturrecherche zu den oben genannten Methoden erfolgen. Dabei soll insbesondere auf den möglichen Einsatz für sehr große Datenmengen eingegangen werden (Big Data).

Der Big-Data-Kontext wirft Fragen bezüglich der Anwendbarkeit von PCA und PLS auf, was beispielsweise die Flexibilität bewährter Algorithmen hinsichtlich der Parallelisierung betrifft. Hier ist zu überprüfen, inwiefern einzelne Algorithmen auf verschiedene Big-Data-Strukturen angewendet werden können oder ob Ausschlusskriterien existieren.

Ferner ist bei sehr großen Datenmengen zu berücksichtigen, dass fehlerhafte Daten vorliegen können oder dass Daten aus anderen Gründen fehlen. Diesbezüglich ist zu untersuchen, wie sich die obigen Methoden in einem solchen Fall verhalten.

Folgende Aufgaben sind im Rahmen der Seminararbeit zu erledigen:

- Einarbeitung in die Grundlagen von PCA und PLS
- Durchführung einer Literaturrecherche hinsichtlich Big Data
- Einordnung der in der Literatur beschriebenen Algorithmen, Überprüfung ob Implementierungen für Matlab oder R vorliegen
- Diskussion der Flexibilität der Algorithmen im Big-Data-Kontext
- Dokumentation der Ergebnisse und Kolloquiumsvortrag

Betreuer: B. Jäschke, M.Sc., Prof. Dr.-Ing. A. Kroll

Beginn: Oktober 2014

Ende: Februar 2015