Analyse & Klassifikation von Blickmustern

Auf Basis einer Datenerhebung der Universität Hamburg

David Klehr 757914, Jörn Malich 752312

26. November 2013

Inhaltsverzeichnis

1 Gegenstand

In dieser Arbeit werden von Psychologen der Universität Hamburg erhobene Daten ausgewertet. In den Untersuchungen wurden die Blickmuster von Mutter Kind Paaren jeweils für eine stressfreie als auch für eine Situation unter äußerem Stresseinfluss dokumentiert. Die Blicke zu einen bestimmten Zeitpunkt wurden bereits in jeweils vier Zustände für die Mutter und das Kind eingeteilt. Bei den Müttern lag jeweils ein uns nicht bekanntes Krankheitsbild vor. Zur Einteilung der Dyaden in Gruppen wurden Kenngrößen wie Aufenthaltswahrscheinlichkeit, Korrelation und Zustandsraten für unterschiedliche Zeitreihen untersucht. Hierzu wurde unter anderem die Methode der hierarchischen Clusterung eingeführt.

2 Dyadenklassifikation anhand der Aufenthaltswahrscheinlichkeit

2.1 Grundlagen

Kombiniert man die jeweils vier möglichen Mutter und Kind Zustände, so erhält man 16 (4×4) mögliche Zustände, die den Gesamtzustand des Systems zu einem bestimmten Zeitpunkt beschreiben.

$$Z(t) \in \{z_1, z_2, ..., z_{15}, z_{16}\}$$

Anhand einer Zeitreihe lässt sich für jede Dyade eine empirische Aufenthaltswahrscheinlichkeit $P(z_i)$ für den Gesamtzustand bestimmen. Die Aufenthaltswahrscheinlichkeiten werden auf 1 normiert. $\sum_{i=1}^{16} P(z_i) = 1$

Die Aufenthaltswahrscheinlichkeiten einer Zeitreihe können einen Ortsvektor $\vec{P}$ im 16-dimensionalen P-Raum definieren $P_i = P(z_i)$. Nun wird die Annahme aufgestellt, dass der $\vec{P}$ Vektor eine relevante Aussage über die Dyade trifft.

Analog zu dieser Herangehensweise lässt sich aus einer Kombination von gestressten und nicht gestressten Aufenthaltswahrscheinlichkeiten ein 32-dimensionaler P'-Raum definieren.

2.2 Methode der Hierarchischen Clusterung

Ziel der hier definierten Methode ist es, die Dyaden in hierarchisch ineinander gestaffelte Gruppen einzuteilen, d.h. eine Gruppe kann jeweils eine oder mehrere Untergruppen enthalten. Um die Komplexität des Ergebnisses zu begrenzen, soll eine Überlappung verschiedener Gruppen vermieden werden. Es gibt verschiedene Methoden der Clusterung[2]. Diese sind bei der bestehenden Problematik jedoch nicht ausreichend geeignet, da die oben genannten Forderung nur teilweise erfüllt werden. Aus diesem Grund wurde hier die Methode der hierarchischen Clusterung eingeführt.

Als Kriterium für die Ähnlichkeit zweier $\vec{P}$-Vektoren wurde ihre Nähe im P-Raum gewählt. Der Abstand d zweier Vektoren $\vec{P}(x)$ und $\vec{P}(y)$ wird hier mit der euklidischen

Metrik definiert.

$$d = \sqrt{\sum_i [P_i(x) - P_i(y)]^2}$$

Nun wird für den $\vec{P}$ -Vektor jeder Dyade eine nach dem Abstand sortierte Liste der Nachbarn aufgestellt. Jetzt werden die Dyaden nach folgendem Verfahren in Gruppen eingeteilt: Zunächst wird eine beliebige Dyade (hier als x bezeichnet) gewählt. Der Abstand dieser Dyade zu ihrem nächsten Nachbarn (hier als y bezeichnet) wird als Maßstab m zur Gruppeneinteilung verwendet.

$$m = d(x, y)$$

Die hier gebildete Gruppe besteht zunächst nur aus den Dyaden x und y. Nun wird überprüft, ob die Dyade y Nachbarn mit einem geringeren Abstand als m aufweist. Sollte dies der Fall sein, so werden diese Nachbarn mit in die Gruppe aufgenommen. Auch die neu aufgenommenen Gruppenmitglieder werden auf Nachbarn mit einem Abstand kleiner m überprüft. Kann kein Nachbar eines der Gruppenmitglieder mit einem Abstand kleiner m ermittelt werden, so wird die Gruppe geschlossen. Dieses Procedere muss für jede Dyade durchgeführt werden.

In Abbildung 1 ist die hierarchische Clusterung (HC) anhand von Punkten im zweidimensionalen Raum veranschaulicht. Die Clusterbildung findet hier für den Punkt 2 statt. Zur vollständigen Einteilung in Cluster muss das Verfahren auch auf die Punkte 1, 3, und 4 angewandt werden. In Abbildung 1.2 wird der charakteristische Abstand für die Clusterbildung anhand von Punkt 2 abgebildet. In den Abbildungen 1.3 und 1.4 gelangen die Punkte 1 und 3 in den Cluster. Punkt 4 befindet sich zu keinem der Clusterpunkte in einem Abstand geringer dem des charakteristischen Anstandes und ist somit nicht Teil des Clusters. Wendet man die hierarchische Clusterung auch auf die Punkte 1,3 und 4 an, so erhält man folgende Cluster:

$$C_1 = \{1, 3\}$$

$$C_2 = \{1, 2, 3\}$$

$$C_3 = \{1, 3\}$$

$$C_4 = \{1, 2, 3, 4\}$$

C_1entspricht C_3 und somit würde man für die vier Punkte 3 unterschiedliche Cluster erhalten, wobei C_1 eine Untermenge von C_2 ist, welcher wiederum eine Untermenge von C_4 ist.

Die hierarchische Clusterung im höherdimensionalen Raum verläuft analog.

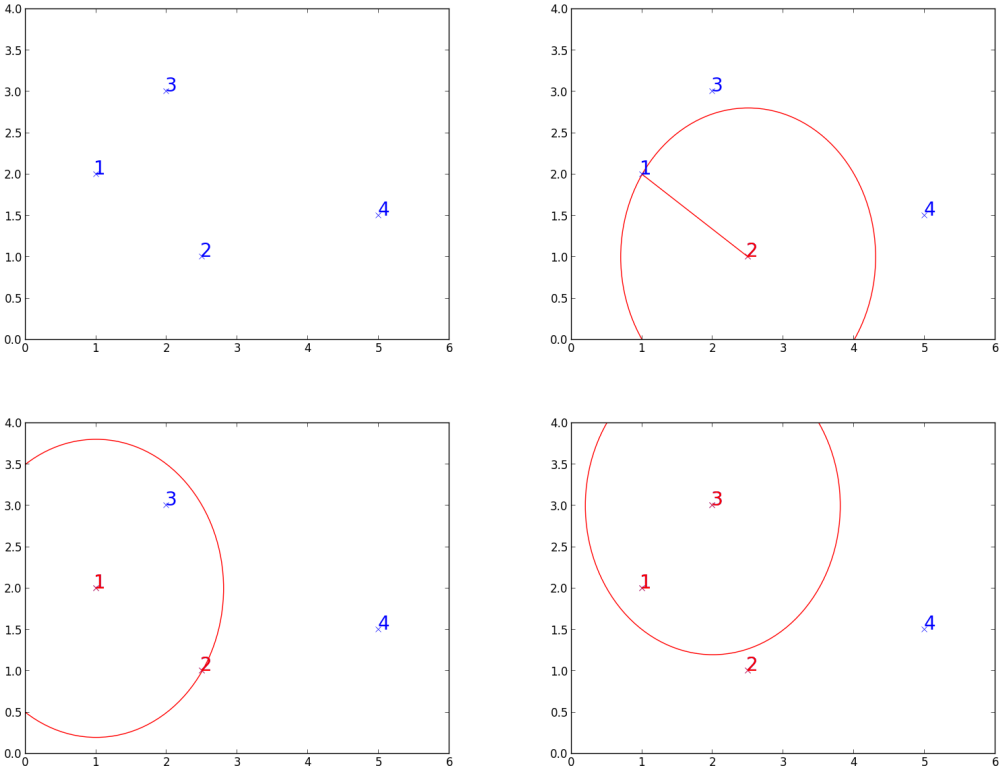

Abbildung 1: Beispiel einer hierarchischen Clusterung für den Punkt 2

2.3 Anwendung

In den folgenden Abbildungen sind die Ergebnisse der hierarchischen Clusterung grafisch dargestellt. Zu unterst sind die großen Obergruppen dargestellt. Oberhalb einer Obergruppe befinden sich jeweils die kleineren Teilgruppen.

Abbildung 2: Dyaden Gruppen ohne Stress

Abbildung 3: Dyaden Gruppen unter Stress

An dieser Stelle kann die Hypothese, dass die Cluster medizinische Muster widerspiegeln von Psychologen überprüft werden.

3 Dyadenklassifikation nach Zustandsänderungsraten

Die Häufigkeiten des Wechsels zwischen einzelnen Zuständen der Mütter sind möglicherweise charakteristisch für das vorliegende Krankheitsbild. Aus diesem Grund wurden die Zustandsänderungsraten der Mütter sowohl im stressfreien Zustand, als auch unter Stresseinwirkung ermittelt. Weiterhin wurde ein Histogramm (Abb. 4) für die absolute Änderung der Änderungsrate erzeugt.

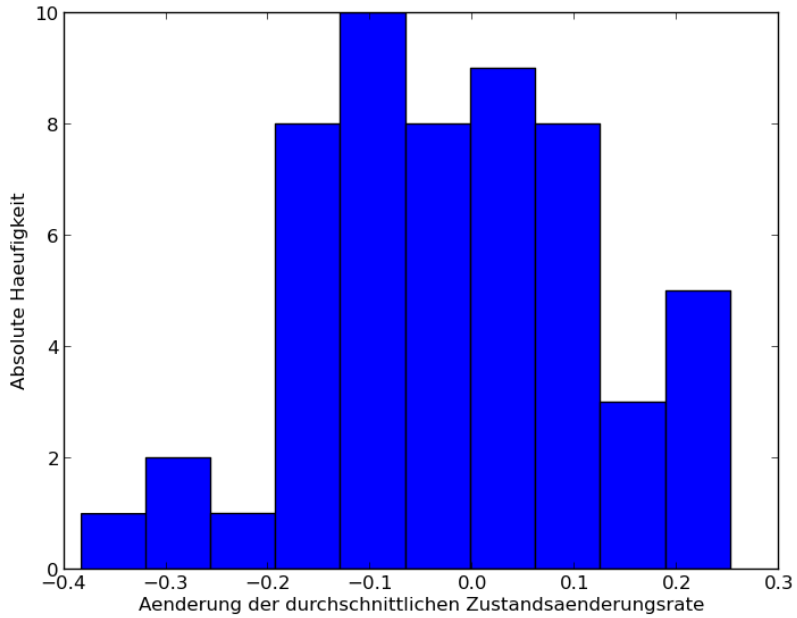

Abbildung 4: Histogramm der Änderung der Zustandsänderungsrate

Erstaunlicherweise hat dies einen Mittelwert bei -0,02, sodass die Änderungsrate sich im Mittel nicht merklich ändert. Das erkennt man auch in der oberen Abbildung, da die Anzahl der Messpunkte im oberen Bereich >0 der im unteren Bereich entspricht. Ein Klassifikationsvorschlag ist in Abbildung 5 dargestellt. Um zu überprüfen, ob die Clusterentstehung zufällig oder begründet ist, werden zufällig generierte Daten hinzugefügt. Diese sind durch Kreuze im Diagramm repräsentiert. Weiterhin wurde die Methode der Hierarchischen Clusterung auf die Messdaten angewandt. Gruppenzugehörigkeiten werden hier durch unterschiedlich gefärbte Kreise um den jeweiligen Dyadenpunkt dargestellt. Das Klassifikationsresultat wird außerdem auf der Konsole ausgegeben. Die Anzahl der Klassen kann im Python Skript correl.py frei eingestellt werden.

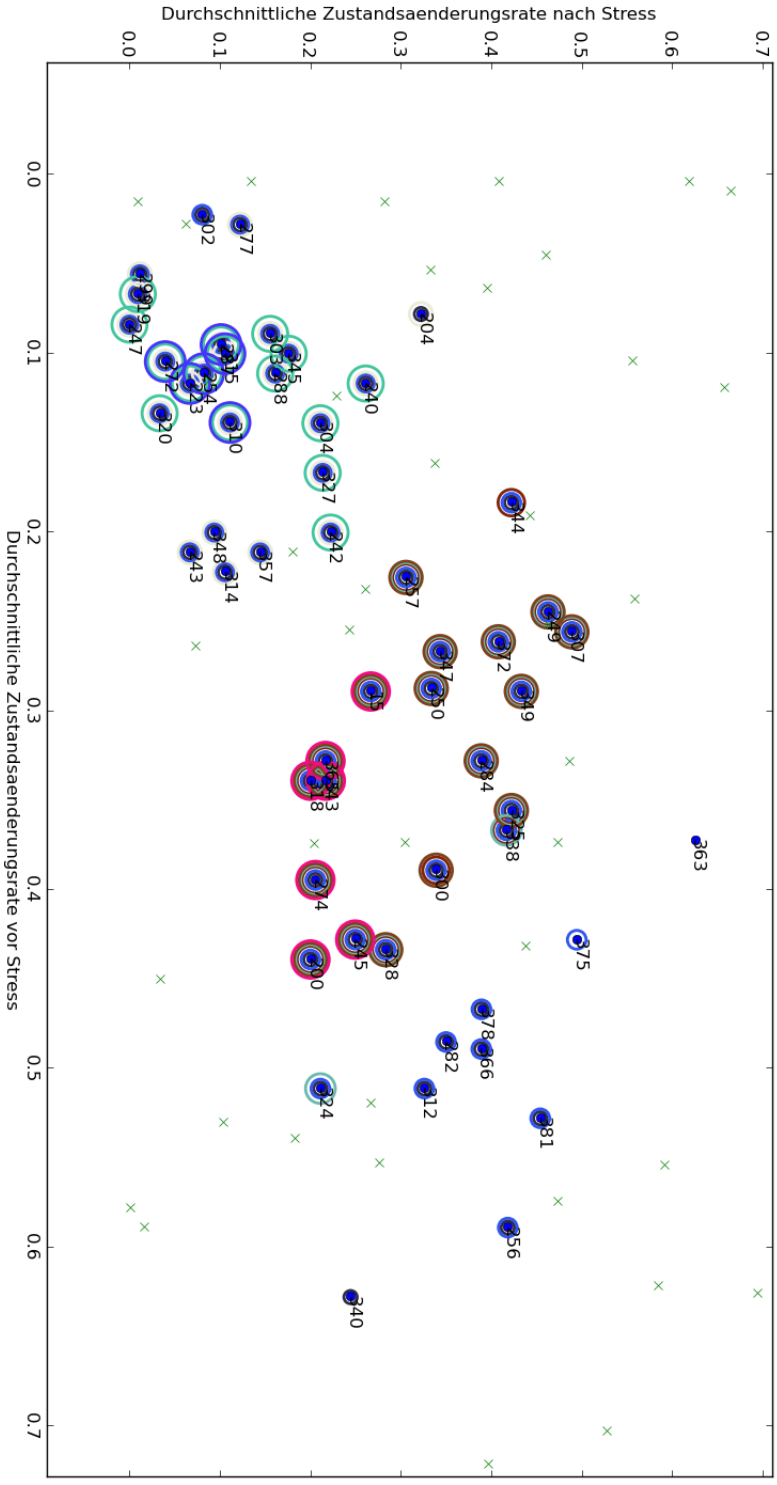

Abbildung 5: Beispielklassifikation der Testdaten nach Zustandsübergangsraten

4 Dyadenklassifikation anhand der Korrelationskoeffizienten der Zeitreihen

Für jede Dyade wurden die Korrelationskoeffizienten zwischen Mutter und Kind bestimmt und in einem Diagramm abgetragen. Der Korrelationskoeffizient stellt dabei ein Maß für den Synchronisationsgrad einer Dyade dar. Das Ergebnis ist in Abbildung 7 zu sehen. Weiterhin wurde ein Histogramm für die Änderung der Korr.koeff. erstellt (Abb. 6). Der Mittelwert befindet sich bei 0,004. Es zeigt sich damit wieder keine deutliche Veränderung der Messgröße. Ein Klassifikationsvorschlag ist nun die Dyaden in 2 Gruppen einzuteilen, je nach dem Vorzeichen des Korr.koeffizienten. Es kann wieder eine Klassifikation mit der HC Methode vorgenommen werden. Das Ergebnis ist in Abbildung 7 dargestellt.

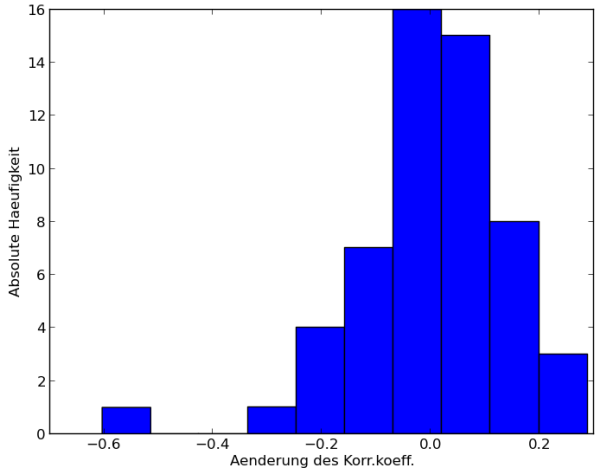

Abbildung 6: Histogramm der Korr.koeff.änderung.

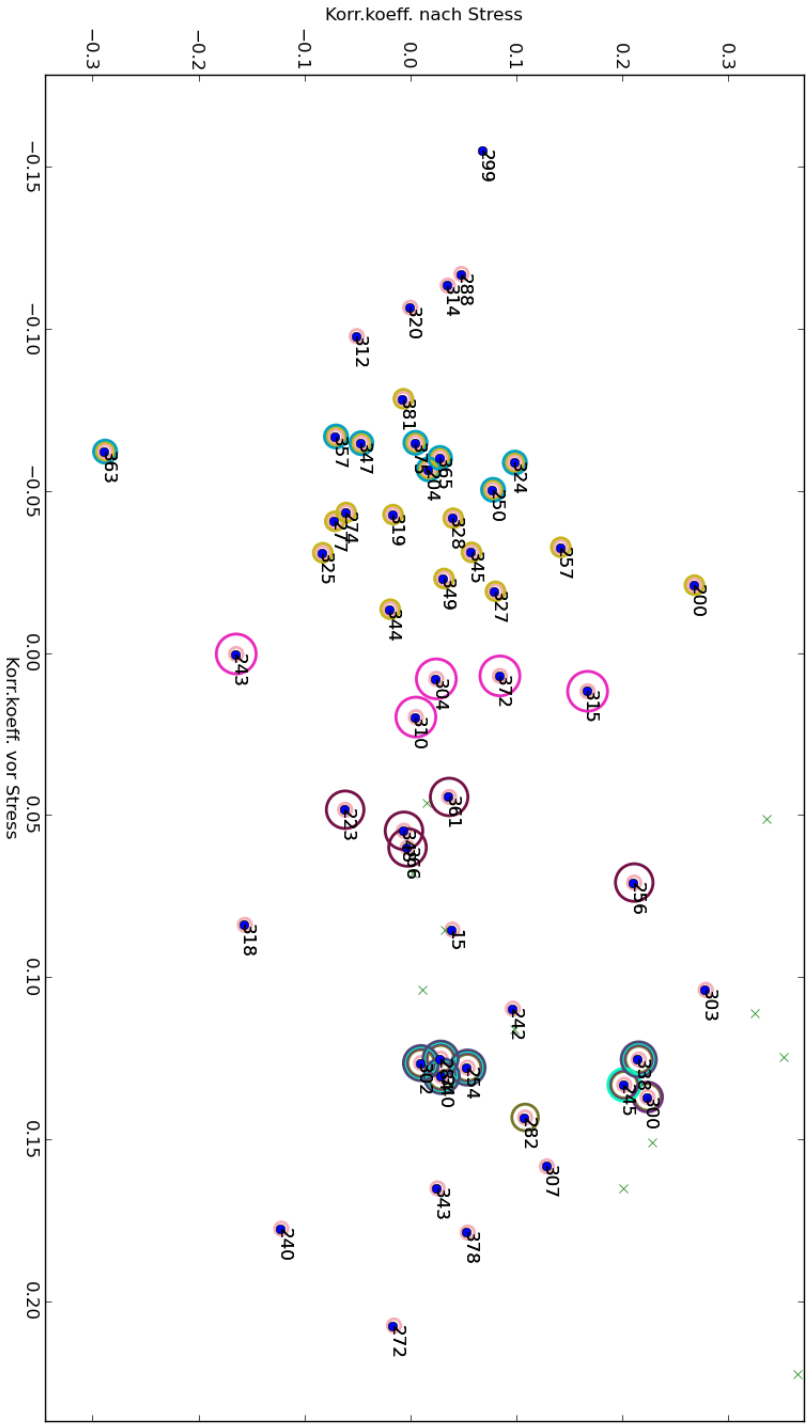

Abbildung 7: Klassifikation nach Korrelation

5 Dyadenklassifikation anhand der Maxima der Kreuzkorrelationsfunktion

Geht man davon aus, dass sich das Verhalten eines Dyadenmitglieds abhängig zum Dyadenpartner erst nach einer bestimmten Reaktionszeit ändert, ist der Korrelationskoeffizient eher ungeeignet. Angebracht ist die Berechnung der Korrelationsfunktion der Zeitreihen. Hiermit lässt sich zum einen die Reaktionszeit τ bestimmen als auch der Fehler aufgrund des zeitlichen Versatzes der Dyaden weitestgehend minimieren. In Abbildung 8 sind die τ Werte für den ungestressten und gestressten Zustand des Korrelationsmaximum im Bereich $[-100, 100] \widehat{=} [-4s, 4s]$ abgetragen. Außerdem sind in Abbildung 9 die Y Werte der Korrelationsmaxima dargestellt. In beiden Fällen wurden die Resultate mit der hierarchischen Clusterung klassifiziert.

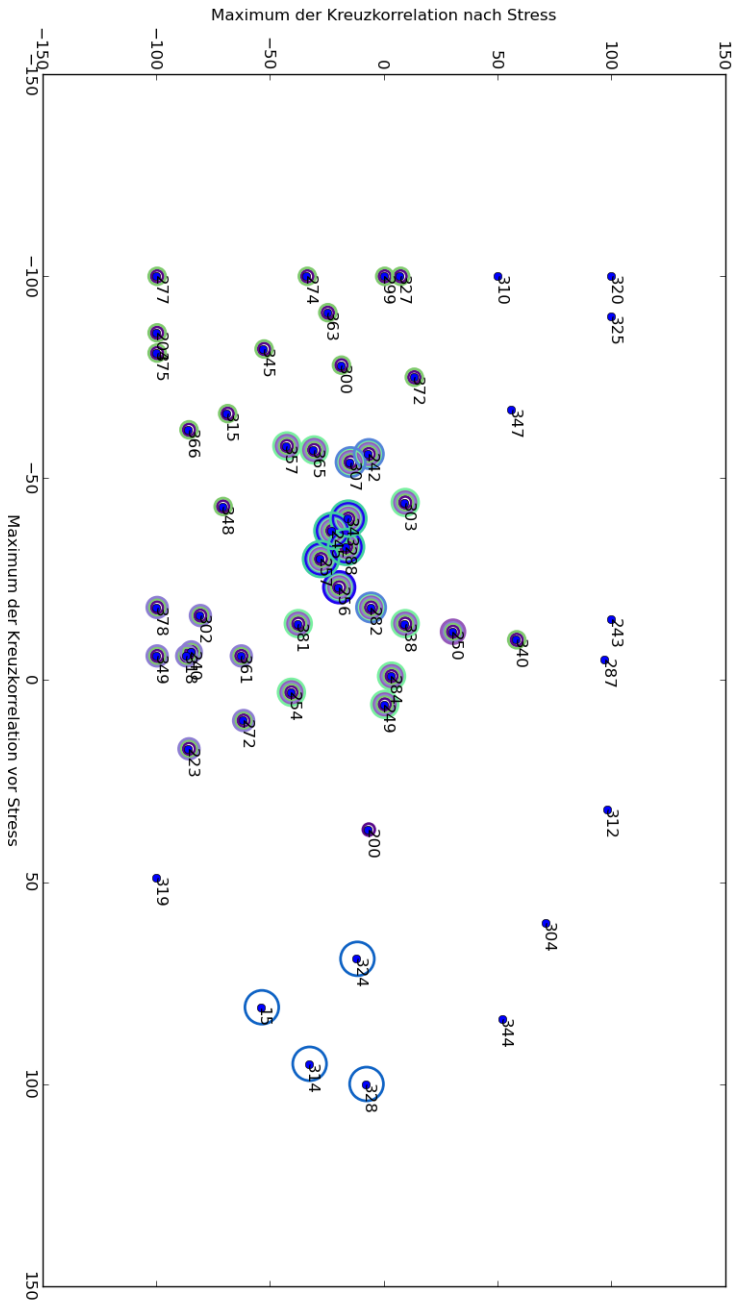

Abbildung 8: Dyadenklassifikation anhand globalem XCorr Maximum

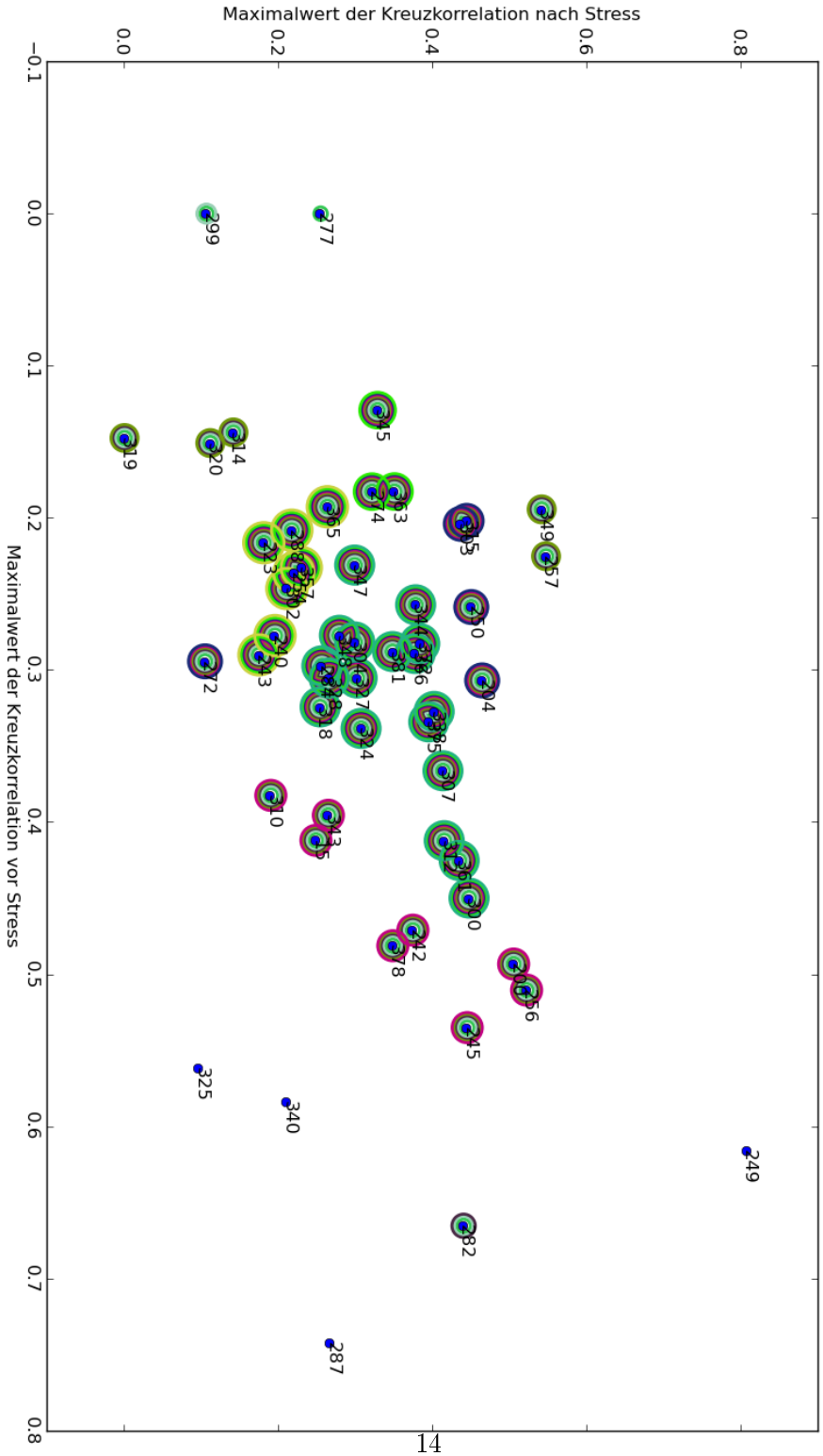

Abbildung 9: Dyadenklassifikation anhand globalem XCorr Maximalwert

6 Simulation des Verhaltens einer Dyade mit Markov Modellierung

Im Markov Modell wird die Annahme getroffen, dass ein Zustand eines Zufallsprozesses X_t jeweils nur von dem vorherigen Zustand abhängig ist. Mithilfe des Markov Modells lassen sich Dyaden lediglich durch Anfangsverteilung $\mu_i = P(X_0 = s_i)$ und Übergangsmatrix (p_{ij}) beschrieben und weitere Zeitreihen für die Auswertung generieren. Dem Programm wird die Zeitreihe einer Mutter Kind Dyade übergeben. Die Anfangsverteilung der Markov Kette wird aus den relativen Aufenthaltswahrscheinlichkeiten gebildet, die Übergangsmatrix durch Abzählen und Einordnen der Zustandsübergänge. Die Markovkette besteht aus 16 Zuständen $S = \{s_1, \ldots, s_{16}\}$. Durch Erzeugung von Zufallszahlen kann damit der Zufallsprozess simuliert werden. Das Ergebnis für die Dyade 288 ist in Abbildung 10 und 11 dargestellt.

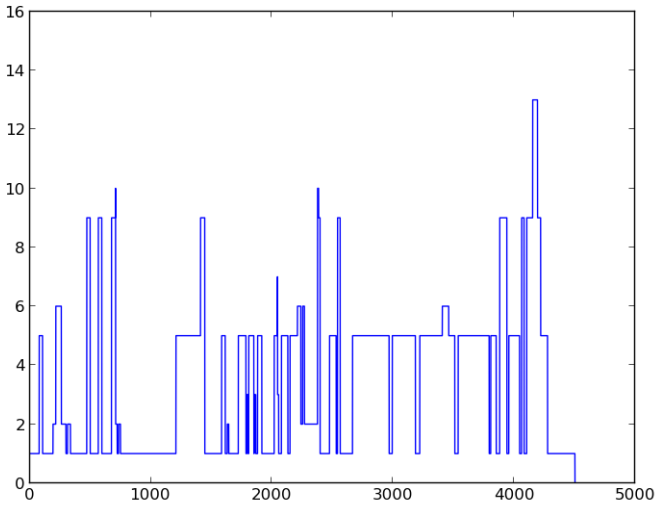

Abbildung 10: Darstellung gemessene Zeitreihe von Dyade 288

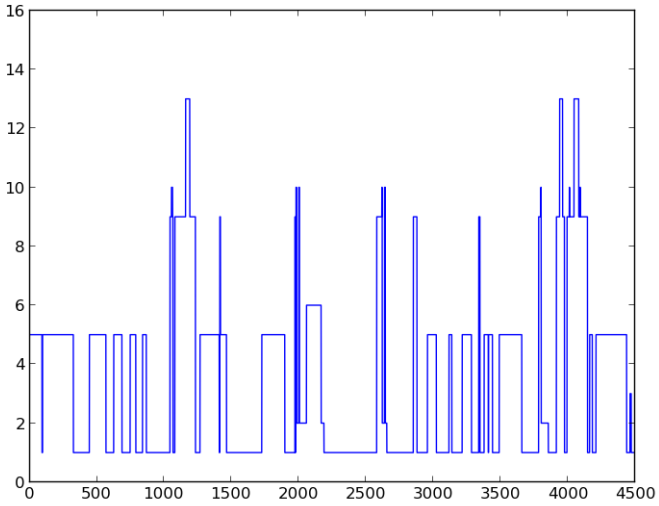

Abbildung 11: Darstellung simulierte Zeitreihe von Dyade 288

Das Programm zur Generierung der Markov Ketten befindet sich in Markov.py. Es wird über python Markov.py <dyaden_id> aufgerufen.

7 Auswertungsautomatisierung

Die Bedienung kann durch die entworfene grafische Benutzeroberfläche erfolgen, mit der sich die vorgestellten Auswertungsmethoden auf eine selbst getroffene Auswahl an Dyaden anwenden lassen. Durch einen Rechtsklick auf eine einzelne Dyade lassen sich die Zeitreihen und Kreuzkorrelationen der Dyaden darstellen. Die Optionen “Stressfrei”, “Gestresst”, “Beide Typen” und “8x8” haben lediglich Einfluss auf die Clusterung nach Aufenthaltswahrscheinlichkeiten. Die Optionen “Stressfrei” und “Gestresst” sind selbsterklärend. Für die Auswahl “Beide Typen” werden die gestressten und ungestressten Dyaden als unterschiedliche Dyaden betrachtet und gemeinsam in Gruppen geclustert. Aus Gründen der Übersicht sollte diese Methode lediglich auf eine kleine Auswahl an Dyaden angewandt werden. Mit der Option “8x8” werden aus einer Kombination aus gestressten und ungestressten Zuständen 8x8 Räume gebildet, auf die die hierarchische Clusterung angewandt wird.

Die Auswahlmöglichkeiten “Clusterung nach Zustandsänderungsrate”, “Clusterung nach Korrelationskoeffizient” und “Clusterung nach Kreuzkorrelation” ermöglichen eine Clusterung der Ausgewählten Dyaden anhand der jeweiligen Merkmale.

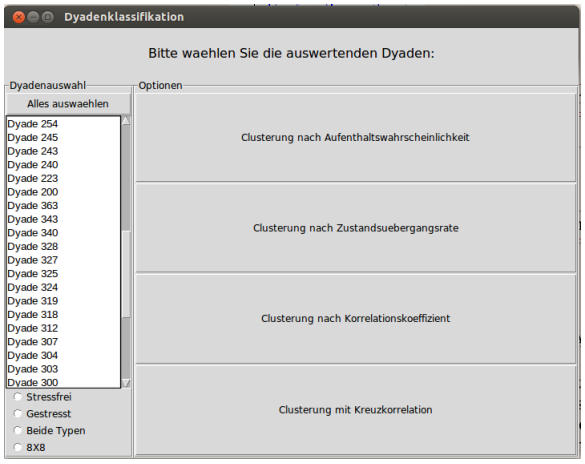

Abbildung 12: Benutzeroberfläche zur Selektion und Auswertung von Dyaden

8 Klassifikation mithilfe der SVM Methode

Hat man mithilfe bereits bestehender Zeitreihen Dyaden in Gruppen eingeteilt, so ist es von Interesse, neue Dyaden den bereits gebildeten Gruppen zuordnen zu können. Eine Methode, die sich hierfür anbietet, ist die der Support Vector Machines(SVM)[1].

SVMs werden für die automatisierte Klassifizierung von Messdaten verwendet. Hierfür werden Testdaten mit bereits vorgenommener Klassifizierung eingelesen. Die Einordnung erfolgt durch Erzeugung von Hyperebenen im Zustandsraum in der Weise, dass der Abstand von Punkten mit unterschiedlichen Klassen maximal wird. Durch Verwendung von

entsprechenden Kernelfunktionen im höherdimensionalen Raum als der des Zustandsraums können sogar nichtlineare Trennungsmannigfaltigkeiten erzeugt werden. Für die Umsetzung wurde die Python Bibliothek scikit-learn verwendet. Um sich die Möglichkeiten von SVM vorzustellen, sei hier ein Beispiel dargestellt.

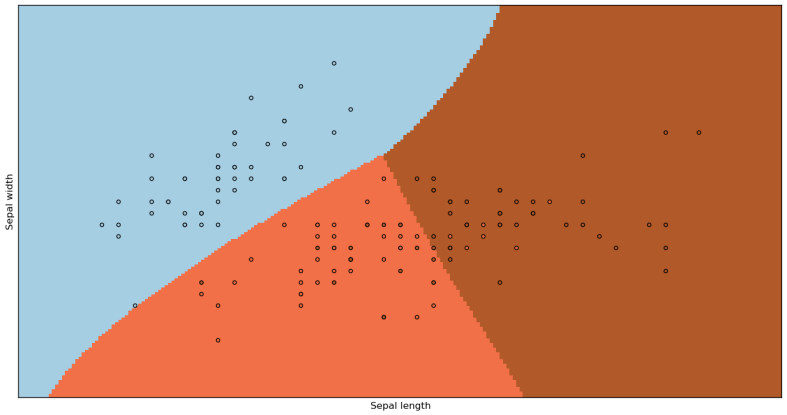

Abbildung 13: Klassifikationsbeispiel mit SVM

Hierbei besteht der Vektorraum aus 2 Komponenten und jeder Datenpunkt kann in eine der 3 farblich dargestellten Klassen eingeordnet werden. Diese Methode wurde auf unsere Daten angewendet:

Aus dem obigen Klassifikationsspektrum (s. Abb. 2) wurde ein Vorschlag entnommen und die SVM dementsprechend trainiert. Das SVM Programm befindet sich in SVM/SVM.py. In Arrays werden die Dyadnnamen den entsprechenden Klassen zugeordnet. Anschließend werden die Lerndaten von der SVM klassifiziert.

Der Klassifikationsvorschlag mit der HC Methode:

[1 1 1 2 2 2 2 2 2 3 3 4 4 5 5 6 6 6 6 6 6 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7]

Die SVM liefert nach entsprechendem Training:

[1 1 1 2 2 6 2 2 2 6 6 7 7 6 7 7 6 6 6 6 6 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7]

Man erkennt deutlich, dass die Klassen mit geringer Mächtigkeit teilweise nicht mehr besetzt sind. Dies ließe sich jedoch mit entsprechender Gewichtung der Klassen optimieren. Der Vorteil zur HC Methode besteht vor allem darin, dass sehr große Datenmengen klassifiziert werden können, da der Lernprozess das Resultat durch die eingeführten Hyperebenen vollständig vorgibt.

9 Zusammenfassung

Die aus den Beobachtungen der Dyaden gewonnen Zeitreihen bieten eine Vielzahl an Möglichkeiten der Clusterbildung. In dieser Arbeit wurde versucht durch Abstraktion wiederkehrende Muster zu entdecken. Dabei hervorzuheben ist das hier entwickelte Verfahren der hierarchischen Clusterung. Von großem Interesse bleibt die Frage, ob die mit diesem Verfahren gefundenen Gruppen reale Krankheitsbilder widerspiegeln. Für den Fall positiver Resultate, lassen sich neue Zeitreihen mit einer recht hohen Genauigkeit wie oben gezeigt mithilfe von SVM Methoden bereits bekannten Krankheitsbildern zuordnen. Letztendlich obliegt es jedoch Psychologen die gefundenen Resultate zu bewerten. Es besteht sowohl die Möglichkeit, dass Dyaden aufgrund nicht relevanter Eigenschaften als auch relevanter Eigenschaften einer Gruppe zugeordnet wurden.

Literatur

[1] Teukolsky S.,W. Vetterling, W. Press, B. Flannery, Numerical Recipes 3rd Edition, Cambridge University Press, New York, 2007

[2] Bishop C., Pattern Recognition and Machine Learning, Springer Science + Business Media, New York, 2006