INF FAKULTÄT FÜR INFORMATIK

DIPLOMARBEIT

Übersicht, Gegenüberstellung und Bewertung der Analyseregeln von Datenvisualisierungen hochdimensionaler Datensätze

Autor:	Marco Kirschke
Fachbereich:	Informatik
Hochschullehrer:	Prof. Dr-Ing.habil. Holger Theisel (*ISG*)
Betreuer:	Dipl.-Ing. Dirk Joachim Lehmann (*ISG*)
Abgabedatum:	21. Dezember 2011

Danksagung

Ich bedanke mich in erster Linie bei meinem Betreuer und Mentor Dirk Joachim Lehmann für die Motivation, gute Ratschläge und vor allem die konstruktive Kritik, sowie den weitsichtigen Ideen, mit dem er bei mir den nötigen Ansporn zur Bewältigung dieser Aufgabe vermitteln konnte. Desweiteren möchte ich mich bei meinen Mitbewohnern Fred und Linda bedanken, die mir bei den kleinen Fragen zwischendurch stets hilfsbereit zur Seite standen und gerade in der Endphase zu einem ausgesprochen sympathischem Arbeitsklima in der WG beigetragen haben. Und auch vielen Dank für die mit Sicherheit nicht einfache Hilfe beim Korrekturlesen der Diplomarbeit an meine Freunde Raimund und Maren, sowie an meine Mutter Karin. Letztlich nicht vergessen möchte ich auch die großzügige Unterstützung und Geduld während der gesamten Studienzeit durch meine Eltern.

Inhaltsverzeichnis

Abkürzungsverzeichnis

KS-Test	Kolmogorow-Smirnov-Test
LOF	local outlier factor
MDS	multidimensionale Skalierung
MUR	minimal umgebendes Rechteck
PCP	parallel coordinates plot, parallele Koordinaten Plot
RVP	Radviz Plot
SP	Scatterplot

1 Einleitung

Die grundlegende Größe in der Informatik ist die Datenmenge als Träger von Informationen über beliebige Domänen. Eine Information wird hierbei durch eine Anordnung bzw. Struktur der Datenmenge abgebildet. Möchte man solchen Strukturen die Informationen aus der Datenmenge ermitteln, folgt u.a. die Möglichkeit Strukturen zuerst zu detektieren, um sie im Folgenden informationsgewinnend interpretieren zu können. Dieser Problematik widmen sich zwei Teilgebiete der Informatik: zum einen *Data-Mining*, welches auf statistische Verfahren setzt, und zum anderen die *visuelle Analyse*, die bildgebende Techniken (Visualisierungen) verwendet.

1.1 Visualisierung

Unter der Visualisierung von Daten versteht man einen Prozess, dessen Ergebnis die visuelle Repräsentation einer Datenmenge ist. Die Grundidee dabei ist, die visuelle Wahrnehmung des Menschen auszunutzen, und ihm abstrakte Daten visuell leicht zugänglich zu machen. Dafür ausschlaggebend ist die ausgeprägte Fähigkeit des menschlichen visuellen Systems Strukturen, Formen und Farben intuitiv und kontextbezogen wahrzunehmen. Zusätzlich zu der dadurch ermöglichten Mustererkennung unterstützt die visuelle Wahrnehmung die Informationsverarbeitung in weiteren Aspekten. Zunächst können ableitbare Informationen aus der Datenmenge in einer Visualisierung direkt dargestellt und parallel verarbeitet werden. Weiterhin verringert sich der Suchaufwand, da bestimmte Visualisierungen große Datenmengen kompakt und entsprechend eines Ähnlichkeitsmaßes gruppiert darstellen. Zuletzt ermöglicht die Parametrisierung der Visualisierungsverfahren eine interaktive Suche durch die Datenmenge [CMS99].

Die Abbildung der Datenmenge erfolgt dabei zumeist als orthogonale Abbildung auf ein zweidimensionales Ausgabemedium (z.b.Monitor, Ausdruck auf Papier). Hier zeigt sich die grundlegende Problemstellung für die Visualisierung mehrdimensionaler Datensätze: Bei einer direkten orthogonalen Darstellung von Räumen mit mehr als 2 Dimensionen können Verdeckungen auftreten und bei mehr als 3 Dimensionen versagt die intuitive räumliche Wahrnehmung.

Das Ziel des Fachgebiets der Visualisierung ist, solche Verfahren zu erforschen, die entsprechend ausdrucksstarke und leicht verständliche Visualisierungen performant produzieren. Die visuelle Repräsentation der Daten bildet eine Schnittstelle zwischen Daten und Betrachter und dient damit als Grundlage in Entscheidungs- und Auswertungsprozessen. Sie vereinfacht die Analyse der Daten und ermöglich prinzipiell die Trennung

von relevanten Informationen. Dies beschleunigt die Entdeckung innerer Zusammenhänge, Konzepte und Modelle sowie die Überprüfung von Hypothesen bezüglich des Datensatzes. Ebenso wird das Verständnis bei der Kommunikation und Präsentation der ermittelten Fakten erleichtert.

Visualisierungsgestützte Prozesse treten in vielen Gebieten auf, beispielsweise:

- in der Meteorologie die visuelle Repräsentation der Wettervorhersage als dreidimensionalen (Längengrad, Breitengrad, Zeit) und multivariaten (Temperatur, Regenwahrscheinlichkeit, Windstärke, etc.) Beobachtungsraum,
- in den Geowissenschaften die räumliche Verteilung von Bevölkerungs- und Wirtschaftsfaktoren,
- im Finanzsektor die zeitliche Entwicklung von Aktienwerten oder der Verteilung von Ausgaben,
- im Ingenieurwesen die Darstellung technischer Attribute eines Bauteils oder die Überwachung kritischer Systemparameter,

Im folgenden Abschnitt werden die Probleme erläutert, aus denen sich die Aufgabenstellung motiviert.

1.2 Motivation

Ausgehend von den Rohdaten gibt es eine Fülle von Einflussfaktoren, die auf den Visualisierungsprozess einwirken. Daraus ergibt sich, dass trotz Anwendung der gleichen Visualisierungsmethode verschiedene Ergebnisse bei der visuellen Analyse zutage kommen. Im ungünstigsten Fall führen die Einflussfaktoren zum Informationsverlust oder zur Informationsverfälschung der den Daten zugrunde liegenden Konzepte und Modelle, und damit zu Fehleinschätzungen bzw.-entscheidungen im konkreten Sachverhalt.

Für solche in [RH94] definierten Einflussgrößen, gibt es beispielhafte negative Ausprägungen. Zunächst können die Bearbeitungsziele unvollständig oder falsch spezifiziert werden, sodass entsprechende nicht-zielführende Visualisierungsmethoden verwendet werden. Ähnlich verhält es sich bei der Verwendung einer Symbolik, die dem Anwendungsgebiet nicht gerecht wird. Während im Ingenieurwesen und in der Wirtschaft die Farben Rot als kritisch, Blau als neutral und Grün als positiv empfunden werden, steht Rot in der Medizin für (gesundes) Leben und Blau sowie Grün u.a.für infektiöse Erkrankungen [Nem93].

Ist die Visualisierung hinsichtlich der Bearbeitungsziele und dem Anwendungsgebiet korrekt gewählt, können weiterhin die Datenvearbeitungsressourcen sowie die Fähigkeiten des Menschen die visuelle Analyse negativ beeinflussen. Eine mögliche Ressourcenbeschränkung, bspw. eine zu geringe Auflösung des Darstellungsmediums, kann zu

Detail- und somit Informationsverlusten führen. Ebenso können die menschlichen Wahrnehmungsfähigkeiten optischen Täuschungen unterlegen sein, was zusätzlich auch zu Informationsverfälschungen führen kann. Dazu gehören u.a. Kontrasteffekte wie Nachbilder, die durch helle Reize ausgelöst werden (bspw. beim Blick in die Sonne) oder der simultane Einfluss nebeneinanderliegender farbiger Flächen, der die Farbwahrnehmung verfälscht. Zuletzt variiert der Wissensstand des Menschen im Bezug auf eine korrekte Analyse der visualisierten Daten, was unter anderem dem autodidaktischen Lernen des Umgangs mit verschiedenen Visualisierungen geschuldet ist.

1.3 Ziele und Aufgabenstellung

Mit der Diplomarbeit wird erforscht, inwiefern sich die visuelle Analyse auf Basis verschiedener Visualisierungsmethoden für mehrdimensionale Datensätze objektiv korrekt durchführen lässt. Dabei soll ein Satz von Analyse- und Interpretationsregeln entstehen, die mit einer hohen Signifikanz allgemeingültig im Visualisierungsprozess sind. Die Menge von Regeln soll den Nutzer auf Widersprüche zwischen seiner Interpretation, der Visualisierung und den fußenden Daten aufmerksam machen. Damit stellt die Arbeit ein Hilfsmittel in Form eines Leitfadens zur zielgerichteten Heranführung unerfahrener Nutzern an verschiedene Visualisierungsmethoden, als Nachschlagewerk für Visualisierungsexperten sowie als Regelbasis für automatische Verfahren im Bereich der visuellen Analyse dar.

Dazu wird zunächst systematisch eine Menge bekannter Visualisierungsmethoden zusammengetragen, welche dann unter Betrachtung ihrer Ziele auf markante Darstellungen wichtiger statistischer Merkmale hin untersucht werden. Aus den Ergebnissen der Untersuchung werden anschließend solche Analyse- und Interpretationsregeln erarbeitet und gegenübergestellt, die geeignet sind, Fehlinterpretation zu reduzieren. Die zu untersuchenden Visualisierungstechniken sind Scatterplot, parallele Koordinaten, Radviz und multidimensionale Skalierung. Zum Auffinden der Analyseregeln werden verschiedene Datensätze generiert, für die bestimmte statistische Eigenschaften bekannt sind. Eine Regel ist dann eine Relation zwischen dieser Eigenschaft und einer entsprechenden charakteristischen Struktur innerhalb der Visualisierung und ist genau dann evident, wenn das Entfernen der Eigenschaft zum Verschwinden der Struktur führt. Die Validierung der Regeln erfolgt, indem sie auf folgende Datensätze, deren Eigenschaften aus der Literatur bereits bekannt sind, angewendet und die Ergebnisse diskutiert werden: olives, iris und yeast. Abschließend wird anhand der Ergebnisse ein Fazit erteilt, inwieweit das aufgestellte Regelwerk dem Anwender oder automatischen Analyseverfahren als Hilfsmittel dienen kann.

1.4 Gliederung

Die Diplomarbeit gliedert sich in fünf Kapitel. Dieser Einleitung folgt im zweiten Kapitel eine einführende Definition hochdimensionaler Datensätze. Weiterhin werden relevante charakteristische Merkmale von Datensätzen erläutert, die zu untersuchenden Visualisierungstechniken vorgestellt, sowie ein Überblick über den aktuellen Stand der Forschung bezüglich verwandter Arbeiten gegeben. Das dritte Kapitel stellt das Hauptkapitel der Diplomarbeit dar. Hier werden besondere Zusammenhänge zwischen den Visualisierungen und ihren Ursprungsdaten beispielhaft zusammengetragen, aus denen anschließend objektive Analyseregeln abgeleitet werden. Die Relevanz und Allgemeingültigkeit der aufgestellten Regeln wird daraufhin im vierten Kapitel evaluiert. Das geschieht durch die Anwendung der Regeln auf Standarddatensätze der statistischen Literatur und der Gegenüberstellung von ermittelten und tatsächlichen Eigenschaften. Eine abschließende Zusammenfassung der Diplomarbeit, mögliche Verbesserungen und Erweiterungen sowie ein Ausblick auf weitere Anwendungsmöglichkeiten sind im fünften Kapitel nachzulesen.

2 Grundlagen

2.1 Hochdimensionale Datensätze

Die Akquirierung von Daten ist entweder physikalischer oder synthetischer Natur. Physikalische Daten werden durch Sensoren ermittelt oder Beobachtungen erfasst, während synthetische Daten das Ergebnis von Berechnungen auf Grundlage mathematischer Modelle sind. Zur Verallgemeinerung beider Datenquellen wird in [SM00] ein Beobachtungsraum definiert, welcher die gewonnenen Daten von ihrer Erhebungsart und dem physikalischen Raum abstrahiert.

Die unabhängigen Variablen der Datenmengen entsprechen hierbei den Dimensionen, die den Beobachtungsraum aufspannen. Dazu gehören bspw.die Koordinatenachsen der Sensorpositionen sowie die Zeitachse der Beobachtungen, oder im theoretischen Fall die Eingabedaten der Berechnung. Innerhalb des Raums existieren Beobachtungspunkte, deren Merkmale durch die an diesem Punkt physikalisch erfassten oder theoretisch berechneten Daten charakterisiert sind (siehe Abb. 2.1). Die Merkmale unterscheiden sich in Datentyp (skalar, vektoriell, tensoriell) und Wertebereich (qualitativ nominal oder ordinal, sowie quantitativ diskret oder kontinuierlich).

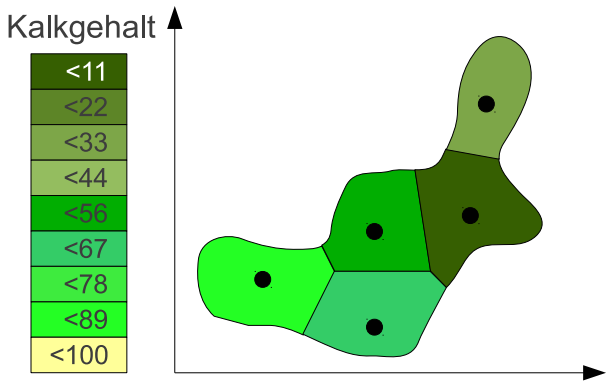


Abbildung 2.1: Beispiel eines Datensatz über den Kalkgehalt (abhängige Variable) eines Sees an 5 Beobachtungspunkten (unabhängige Variablen Breiten- und Längengrad). [SM00, nach Abb. 3.1]

Im einfachsten Fall besteht die Datenmenge nur aus Beobachtungspunkten in einem Beobachtungsraum. Erfasst man an den Beobachtungspunkten die Ausprägung genau eines Merkmals, wir dieses Merkmal zur von den unabhängigen Variablen abhängige Va-

riable. Beträgt die Anzahl abhängiger Variablen zwei oder mehr, spricht man von Multiparameter oder auch multivariaten Daten. Unterschlägt man explizit einen funktionalen Zusammenhang, lässt sich jede abhängige Variable auch als unabhängige Variable betrachten. Andersherum kann jedoch keine unabhängige Variable als abhängig betrachtet werden, wenn die implizite Kausalität nicht explizit gegeben ist.

In dieser Arbeit sollen allgemeingültige und anwendungsfreie Analyseregeln erforscht werden. Diese Regeln stellen eine Relation zwischen visueller Struktur und Datensatz dar, sodass auch der Datensatz allgemeingültig, anwendungsfrei und somit frei von abhängigen Variablen sein muss. Dementsprechend werden alle in dieser Arbeit untersuchten Datensätze als hochdimensionale nicht-multivariate Datensätze betrachtet.

Im nächsten Abschnitt werden statistische Eigenschaften, die in Datensätzen auftreten können, eingeführt.

2.2 statistische Merkmale

Bei der visuellen Analyse steht der Anwender vor einer hohen Anzahl verschiedener Vorgehensweisen, den Datensatz nach Informationen zu durchforschen. In [AES05] wurden die dabei aufkommenden Fragen untersucht und in zehn grundlegende Aufgaben der visuellen Analyse zusammengefasst, von denen vier relevant für die diese Arbeit sind:

- Wie sind die Daten verteilt?
- Gibt es Ausreisserwerte?
- Liegt eine Clusterbildung vor?
- Besteht eine Korrelation zwischen den Dimensionen?

Um einschätzen zu können, ob und inwiefern eine Visualisierungmethode diese Fragen beantworten kann, müssen die Antworten zunächst mathematisch bestimmt sein. Für die genannten 4 Probleme gibt es verschiedene Lösungswege. Die folgenden Abschnitte beschreiben die jeweils ausgewählte Methode zur Bestimmung der Verteilung, Ausreißer, Clusterbildung und Korrelation.

2.2.1 Verteilungstest

Der Kolmogorow-Smirnov-Test [Lov11, S. 718](KS-Test) beantwortet die Frage, ob zwei gegebene Datenreihen derselben Verteilungsfunktion entstammen. Entsprechend wird bei gegebener Datenreihe X eine gegebenen Verteilungsfunktion $P(x)$ als zweite Datenreihe verwendet, sodass als Nullhypothese H_0 getestet wird, ob P die Verteilungsfunktion von X ist.

Dazu wird zunächst die empirische Verteilungsfunktion $F_{X,n}(x)$ aus der Stichprobe X gebildet. Sie ergibt sich aus dem Verhältnis der Anzahl der Elemente kleiner gleich x zur

Gesamtzahl an Elementen.

$$F_{X,n}(x) = \frac{\#i : x_i \leq x}{n}$$

Die Abweichung der Datenreihe X von der Verteilung $P(x)$ wird dann anhand der kleinsten oberen Schranke der Distanz zwischen $F_{X,n}(x)$ und $P(x)$ gemessen.

$$D = \max x : -\infty < x < \infty : F_{X,n}(x) - P(x)$$

Überschreitet D einen kritischen Wert [Mas51] bezüglich eines Signifikanzniveaus α, wird H_0 verworfen und die Alternativhypothese angenommen, dass X nicht $P(x)$ verteilt ist. Dabei ist zu beachten, dass die kritischen Werte der Kolmogorow-Smirnov-Statistik nur für Verteilungsfunktionen mit bekannten Parametern gelten. Schätzt man die Parameter der Verteilungsfunktion $P(x)$ aus der Stichprobe X, muss man entsprechend auf angepasste kritische Werte zurückgreifen. Diese wurden für den Test auf Normalverteilung in [Lil67] und für den Test auf Exponentialverteilung in [Lil69] veröffentlicht.

2.2.2 Clusteranalyse

Einer der vielfältigen Algorithmen zur Clusteranalyse ist DBSCAN [EKSX96]. Hierbei handelt es sich um ein dichtebasiertes Verfahren, welches eine gegebene Punktmenge in Punkt-Cluster (Gebiete mit hoher Punktdichte) und Rauschen (Restfläche mit geringer Punktdichte) unterteilt. Die Grundidee ist die Untersuchung, ob in der näheren Umgebung eines Punktes eine bestimmte Mindestanzahl weiterer Punkte existiert. Hierbei entfällt die vorherige Festlegung auf die Anzahl der Cluster.

Zur formalen Beschreibung eines Clusters C in einer Punktmenge D werden folgenden Relationen und Funktionen definiert:

Epsilonumgebung, $N_{Eps}(p)$ Die Epsilonumgebung eines Punktes p ist eine Teilmenge von D. Sie enthält alle weiteren Punkte, deren Distanz zu p kleiner oder gleich Eps ist.

$$N_{Eps}(p) = \{q \in D \mid \text{dist}(p, q) \leq Eps\}$$

direkt benachbart, $db(p, q)$ Ein Punkt p ist direkt Dichte-verbunden mit einem Punkt q, wenn p in der Epsilonumgebung von q liegt und q eine Mindestanzahl an Nachbarpunkten aufweist: $p \in N_{Eps}(q) \wedge |N_{Eps}(q)| \geq minPunkte$. Hierbei handelt es sich um eine asymmetrische Relation, da die Randpunkte eines Cluster nicht die Mindestanzahl an Nachbarpunkten aufweisen.

indirekt benachbart, $ib(p, q)$ Ein Punkt p ist indirekt benachbart mit einem Punkt q, wenn eine Folge direkt benachbarter Punkte mit p als Start- und q als Endpunkt existiert. Als transitive Hülle der Relation *direkt benachbart* ist diese Relation ebenfalls asymmetrisch.

Dichte-verbunden, $dv(p, q)$ Ein Punkt p ist Dichte-verbunden mit einem Punkt q, wenn ein Punkt o existiert, welcher jeweils mit p und q indirekt benachbart ist. Die Relation ist symmetrisch, womit eine mengentheoretische Definition eines Cluster in einer Menge aus Punkten gegeben werden kann.

Damit ergibt eine nicht leere Untermenge C der gegebenen Punktmenge D bezüglich einer Distanz Eps und einer Mindestanzahl von Punkten $minPunkte$ genau dann einen Cluster, wenn die folgenden Bedingungen erfüllt sind:

$$\forall p, q : p \in C \wedge ib(q, p) \rightarrow q \in C$$

$$\forall p, q \in C : dv(p, q)$$

Das Punktrauschen ist dann eine Untermenge N in einer Punktmenge D mit den bekannten Clustern $C_1, \ldots, C_k$ bezüglich der Parameter Eps_i und $minPunkte_i$ mit $i, \ldots k$, die alle Punkte enthält, welche zu keinem Cluster C_i gehören.

$$N = \{p \in D \mid \forall i : p \notin C_i\}$$

2.2.3 Ausreißertest

Der *local outlier factor* (LOF) [BKNS00] ist ein Maß, das angibt, wie stark ein Objekt von anderen Objekten isoliert ist. Die Berechnung des LOF ist wie bei DBSCAN dichtebasiert, sodass teilweise gleiche Konzepte verwendet werden. Die Grundidee ist hierbei der Vergleich der lokalen Nachbarschaftsdichte eines Objekts zu den lokalen Dichten seiner Nachbarn. Der LOF wird durch folgende Definitionen schrittweise hergeleitet:

k-Distanz, $k\text{-}d(p)$ Die k-Distanz ist die maximale Distanz d eines Objekts p zu einem weiteren Objekt q, welche die k nächsten Nachbarn einschließt. D.h., es müssen mindestens k Objekte eine gleiche oder eine geringere Distanz, sowie maximal k-1 Objekte eine geringe Distanz aufweisen.

k-Distanz-Nachbarschaft, $N_{k\text{-}d(p)}$ Die k-Distanz-Nachbarschaft ist die Untermenge aller Objekte, deren Distanz zu p kleiner gleich der k-Distanz von p ist, also der k nächsten Nachbarn von p. Die Kardinalität von $N_{k\text{-}d(p)}$ kann größer k sein, wenn mehrere Objekte aus $D \setminus p$ die gleiche Distanz zu p aufweisen.

$$N_{k\text{-}d(p)} = \{q \in D \setminus \{p\} \mid d(p, q) \leq k\text{-}d(p)\}$$

Nachbarschaftsdistanz, $nd_k(p, o)$ Die Nachbarschaftsdistanz zwischen p und o entspricht der tatsächlichen Distanz mit der k-Distanz von o als Mindestwert.

$$nd_k(p, o) = \max(k\text{-}d(o), d(p, 0))$$

lokale Nachbarschaftsdichte, $\mathrm{LRD}_m(p)$ Die lokale Nachbarschaftsdichte ist die Inverse der durchschnittlichen Nachbarschaftsdistanz bezüglich der m nächsten Nachbarn.

$$\mathrm{LRD}_m(p) = \left(\frac{\sum_{o \in N_m(p)} n d_m(p, o)}{|N_m(p)|} \right)^{-1}$$

local outlier factor, $\mathrm{LOF}_m(p)$ Der LOF ist das durchschnittliche Verhältnis zwischen den lokalen Nachbarschaftsdichten der m nächsten Nachbarn von p zur lokalen Nachbarschaftsdichte von p selbst.

$$\mathrm{LOF}_m(p) = \frac{\sum_{o \in N_m(p)} \frac{\mathrm{LRD}_m(o)}{\mathrm{LRD}_m(p)}}{|N_{m(p)}|}$$

2.2.4 Korrelation

Der Korrelationskoeffizient nach Pearson [Lov11, S. 315] gibt den Grad des linearen Zusammenhangs zweier Merkmale an. Sein Wert liegt innerhalb von $[-1, 1]$, wobei er bei einem fehlenden linearen Zusammenhang 0, sowie bei vollständiger positiver bzw. negativer Korrelation 1 bzw. -1 beträgt. Er berechnet sich aus dem Verhältnis zwischen der Kovarianz beider Merkmale zum Produkt ihrer Standardabweichungen:

$$\varrho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)} \cdot \sqrt{\mathrm{Var}(Y)}}$$

$$\mathrm{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})$$

$$\mathrm{Var}(X) = \mathrm{Cov}(X, X) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

Berechnet man den Korrelationskoeffizienten nicht über die gegebene Merkmale X, Y, sondern über deren Rangreihenfolgen R_{X_i}, R_{Y_i}, erhält man Spearmans Rangkorrelationskoeffizient. Die Rangreihenfolge ist hierbei die Folge der Ordnungszahlen der Elemente bei aufsteigender Sortierung der Merkmale. D.h., wenn $R_{X_i} = 1$, dann hat X_i den kleinsten Wert von X und wenn $R_{X_i} = n$, dann hat X_i den größten Wert von X. Tritt jeder Rang nur einmal auf, gilt folgenden vereinfachte Formel [Lov11, S. 502]:

$$\varrho(X, Y) = 1 - \frac{6 \sum_{i=1}^{n} (R_{X_i} - R_{Y_i})^2}{n(n^2 - 1)}$$

d_1	d_2	d_3	Farbe
1.0	1.0	1.0	Rot
1.0	1.0	2.0	Blau
2.0	1.0	1.0	Gelb
2.0	1.0	2.0	Grün
1.5	2.0	1.5	Braun

Tabelle 2.1: 3-dimensionaler Beispieldatensatz

2.3 Techniken zur Visualisierung hochdimensionaler Datensätze

Die Darstellungsform der verschiedener Visualisierungsmethoden wird durch drei visuelle Kriterien charakterisiert. Es wird unterschieden, ob die Darstellung 2- oder 3-dimensional erfolgt, der Datensatz vollständig oder unvollständig abgebildet wird und ob es sich um eine statische oder dynamische Visualisierung handelt. Eine Visualisierung bezeichnet man dann als vollständig, wenn sie alle Werte des Datensatzes mit allen Dimensionen gleichzeitig darstellt. Eine dynamische Visualisierung stellt zu verschiedenen Zeitpunkten verschiedene Ansichten auf die Werte dar, während eine statische Visualisierung ein einziges Abbild erzeugt. Insgesamt lassen sich Visualisierungstechniken in vier Klassen unterteilen [SM00]:

- geometrische Transformation oder Projektion
- Ikonisierung und Glyphendarstellung
- pixelbasierte Techniken
- graphbasierte und hierarchische Darstellung

Laut der Aufgabenstellung in Abschnitt 1.3 beschränken sich die Untersuchungen in dieser Diplomarbeit auf Methoden der geometrischen Visualisierungen. Dazu zählen zunächst die beiden populärsten Techniken: Scatterplot (SP) und parallele Koordinaten Plot (PCP). Beide Visualisierungen stellen grundlegend nur Zusammenhänge zwischen zwei Dimensionen dar. Erst durch die parallele Darstellung verschiedener Dimensionspaare erhält man einen multidimensionale Ansicht. Dementsprechend werden weiterhin zwei grundlegend multidimensionale Techniken betrachtet: Radviz (RVP) und die multidimensionale Skalierung (MDS).

Im Folgenden werden die verwendeten Visualisierungstechniken vorgestellt und ihre visuelle Charakteristik anhand eines Beispieldatensatzes (siehe Tabelle 2.1) erläutert.

2.3.1 Scatterplot-Matrix

Ein Scatterplot [WB97] ist die direkte Darstellung der durch den n-dimensionalen Datensatz definierten Struktur im n-dimensionalem Koordinatensystem. Jeder Punkt ist

entsprechend seiner Koordinaten im mehrdimensionalen Raum platziert (siehe Abb. 2.2). Hierbei können mehr als drei Dimensionen naturgemäß nicht mehr intuitiv erfasst werden.

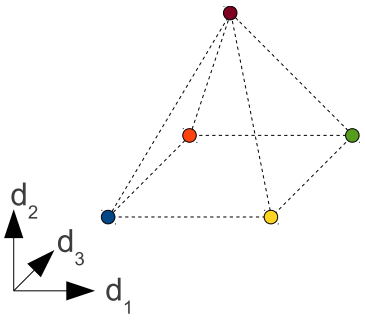


Abbildung 2.2: 3D-Scatterplot des Beispieldatensatzes

Um dennoch hochdimensionale Datensätze zu visualisieren, stellt die Scatterplot-Matrix jede paarweise Kombination aller Dimensionen als $n \times n$-Matrix aus 2-dimensionalen Scatterplots dar. Eine Permutation der Dimensionsmenge $1, \ldots, n$ definiert dazu die vertikale Achse der Zeilen, sowie gleichermaßen die horizontale Achse der Spalten (siehe Abb. 2.3). Auf der Hauptdiagonalen ergeben sich dadurch 2D-Scatterplots jeder Dimension mit sich selbst, sodass dort die Dimensionsbezeichnungen oder weitere Informationen dargestellt werden. Ebenso sind die Scatterplots unterhalb der Hauptdiagonalen Achsen-invertierte Pendants der Scatterplots oberhalb der Hauptdiagonale, was wiederum Platz für weitere Informationen lässt. Die Scatterplot-Matrix ist eine 2-dimensionale, vollständige und statische Visualisierung.

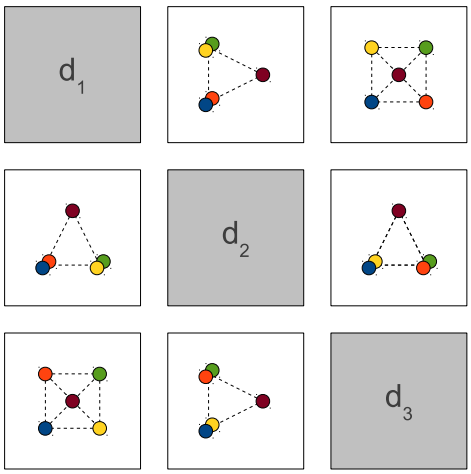


Abbildung 2.3: Scatterplot-Matrix des Beispieldatensatzes

Ein Nachteil der Scatterplot-Matrix ist, dass die einzelnen Projektionen nur die Ebenen der Basisvektoren zeigen. Nicht-orthogonale Strukturen können aufgrund dessen bei

der visuellen Analyse unentdeckt bleiben. Eine Lösung dafür ist die Grand Tour [Asi85], welche sukzessive jedes Dimensionspaar aus allen Betrachtungswinkeln als Scatterplot darstellt. Eine weitere Möglichkeit ist die automatische Bestimmung der interessantesten Projektionen (starke Abweichung von der Normalverteilung) mittels dem Projection Pursuit-Verfahren [FT74].

2.3.2 Parallele Koordinaten

Bei parallelen Koordinaten [ID90] wird jede Dimension durch eine eigene vertikale Achse repräsentiert. Die Achsen sind parallel zueinander auf einer horizontalen Ebene angeordnet. Ein Punkt des Datensatzes entspricht dann einem Streckenzug entlang seiner auf $[0, 1]$-normierten Koordinaten. Die Visualisierung des Beispieldatensatzes aus Tabelle 2.1 in Form paralleler Koordinaten ist in Abb. 2.4 dargestellt. Ein erheblicher Nachteil ist der Verlust der strukturellen Informationen bei großen Datensätzen aufgrund der vielen sich überschneidenden Streckenzügen. Neben dem naiven Ansatz der Verwendung von Transparenz, wurden verschiedene Techniken entwickelt, um dieser visuellen Überladung entgegenzuwirken: Kurven statt Streckenzüge [The00] entwirren Kreuzungspunkte an den Achsen und Cluster-Bündelung [MM08] schafft Platz zwischen den Achsen.

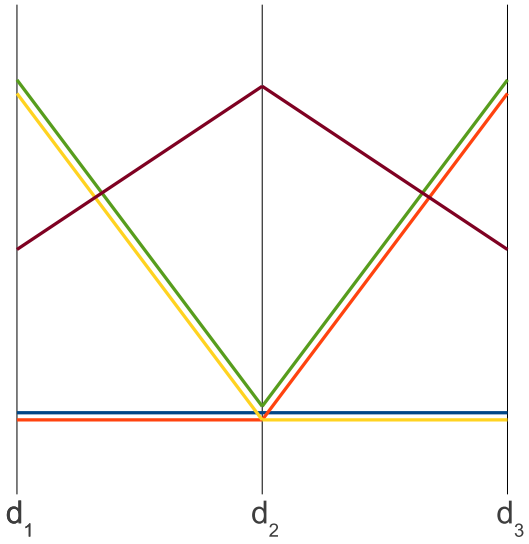


Abbildung 2.4: Parallele Koordinaten des Beispieldatensatzes

Es existiert eine Dualität zwischen Scatterplot und parallelen Koordinaten. Eine Linie im Scatterplot entspricht einem Linienschnittpunkt bei den parallelen Koordinaten.

$$l_{SP} : y = mx + n \;\leftrightarrow\; l_{PCP} : (\frac{d}{1-m}, \frac{n}{1-m})$$

d ist der Abstand zwischen den parallelen Achsen im PCP. Das macht Schnittpunkte und Parallelitäten (als Sonderfall $m = 1$: Schnittpunkt im Unendlichen) zu inter-

essanten Strukturen, hinsichtlich linearer Zusammenhänge zwischen den Dimensionen. Auch univariate Ausreißer lassen sich auf den Achsen leicht erkennen, und Trends durch den Streckenzug entlang der Koordinaten intuitiv ablesen. Zusätzliche Informationen bringen Erweiterungen wie Parahistogramme[OL96], die an den Achsen zusätzlich die Verteilungsdichte der entsprechenden Dimension als Histogramm darstellt. Ordnet man die Achsen nicht parallel, sondern radial um einen Mittelpunkt an, erhält man einen Starplot[Ric95], welcher der Visualisierung einen ikonischen Charakter verleiht.

2.3.3 Multidimensionale Skalierung

Bei der multidimensionalen Skalierung handelt es sich um eine Ähnlichkeitsanalyse [Lov11, S. 875]. Das Ziel ist, Objekte in einem 2- oder 3-dimensionalen Raum so anzuordnen, dass die euklidischen Distanzen im niederdimensionalen Raum den Ähnlichkeits- oder Distanzmaßen zwischen den Objekten weitestgehend gleichen. D.h., ähnliche Objekte sind nah beieinander und unähnliche Objekte weit entfernt platziert.

Man unterscheidet die metrische und nicht-metrische MDS. Die metrische MDS verwendet als Eingabe eine Distanzmatrix, in der die Distanzen zwischen allen Objekten zueinander gegeben sind. Diese Distanzen des hochdimensionalen Raums beeinflussen direkt die euklidischen Distanzen im niederdimensionalen Raum. In [Tor52] ist eine analytischer Lösung gegeben, der aus der Eingangs-Distanzmatrix die euklidischen Distanzen berechnet. Die Lösung wird als Konfiguration bezeichnet und ist bis auf Rotation und Skalierung eindeutig. Abb. 2.5 zeigt auf der linken Seite die Lösung der metrischen-MDS für den Beispieldatensatz.

Die nicht-metrische MDS ist ein iteratives Verfahren zur Bestimmung der optimalen Konfiguration, ohne auf metrische Distanzen als Eingabe angewiesen zu sein. Es reicht eine gegebene Ordnung der (Un-)Ähnlichkeiten zwischen den Objekten, sodass auch ordinale oder nominale Maße verwendet werden können. Nach [Kru64] muss eine Monotoniebedingung zwischen den Ähnlichkeiten δ_{ij} und den euklidischen Distanzen d_{ij} in der Konfiguration erfüllt sein, d.h. wenn $\delta_{ij} < \delta_{i'j'}$ dann muss auch $d_{ij} \leq d_{i'j'}$ sein. Es wird ebenfalls ein Algorithmus vorgestellt, der ausgehend von einer zufälligen Startkonfiguration die Monotoniebedingung prüft und bei einer auftretenden Verletzung entsprechende Disparitäten $\hat{d}_{ij}$ schätzt, aus denen dann die neuen Positionen x_{ij} berechnet werden. Ziel ist die Minimierung der Kenngröße $STRESS$, welches die Abweichung der Disparitäten von den Distanzen angibt. Da der Algorithmus mit Erfüllung der Monotoniebedingung abbricht, können verschiedene Startkonfigurationen aufgrund lokaler Minima verschiedene Konfigurationen ergeben. Die nicht-metrische MDS des Beispieldatensatzes führt deshalb neben der Konfiguration der metrischen MDS auch zu weiteren Lösungen (siehe Abb. 2.5, rechts).

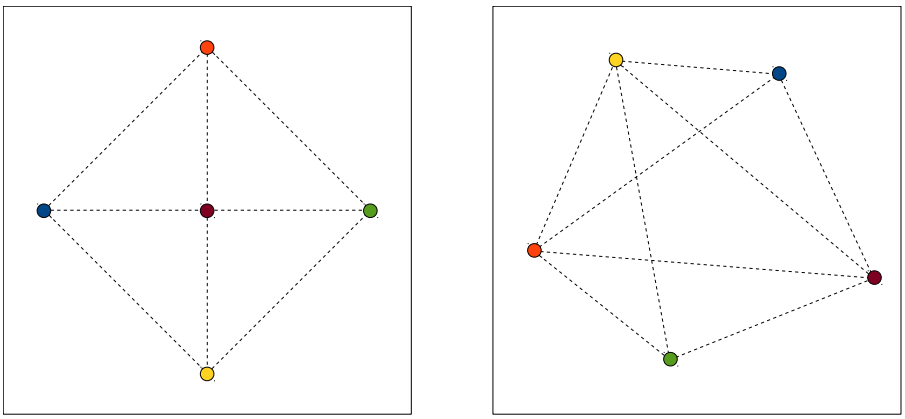


Abbildung 2.5: Multidimensionale Skalierung des Beispieldatensatzes. Die metrische MDS ergibt stets die gleiche Konfiguration (links). Die nicht-metrische MDS kann zur linken und rechten Konfiguration führen.

2.3.4 Radviz

Radviz [HGM+97] ist eine kreisförmige Visualisierung. Die n Dimensionen werden durch n Ankerpunkte repräsentiert, welche gleichmäßig auf dem Rand eines Kreises mit gegebenem Mittelpunkt verteilt sind. Sinngemäß verbinden n Federn einen Punkt des Datensatzes mit jeweils einem Ankerpunkt. Die Federkonstante K_i jeder Feder entspricht dabei dem $[0, 1]$-normierten Wertes der i-ten Koordinate des Punktes. Damit ergibt sich als Position des Punktes der Ort, an der die Summe der Federkräfte gleich 0 ist (siehe Abb. 2.6).

$$\mathbf{p}_i = \frac{\sum_{j=1}^{n} \mathbf{d}_j x_{i,j}}{\sum_{j=1}^{n} x_{i,j}}$$

d_j ist der Vektor vom Kreismittelpunkt zum entsprechenden Ankerpunkt. [AEL+10] Die Radviz-Visualisierung des Beispieldatensatzes ist in Abb. 2.7 zu sehen.

2.4 Stand der Forschung

Der folgende Abschnitt gibt eine Übersicht von relevanten Arbeiten im Bereich der Datenerzeugung und der Analyse konkreter Eigenschaften von Visualisierungen. Zudem werden Softwaretools hinsichtlich ihrer Eigenschaften verglichen, die eingeführten statistischen Merkmale und Visualisierungsmethoden zu berechnen bzw. anzuwenden.

2.4.1 Synthetische Datengenerierung

Für die Untersuchungen der Zusammenhänge zwischen Daten und besonderer Merkmale in den Visualisierungen müssen solche Daten zunächst vorhanden sein. Einerseits kann

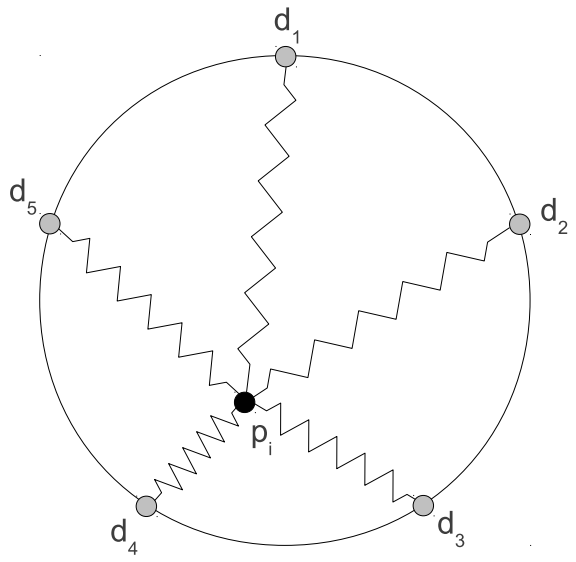


Abbildung 2.6: Die Position p_i eines Datenpunkts x_i ergibt sich durch den Ausgleich der Federkräfte, die entsprechend der Ausprägung seiner Koordinaten $x_{i,j}$ zwischen den Dimensionsankerpunkten- d_j und p_i wirken (nach [AEL+10, Abb. 1].

man sich bekannte Standarddatensätze aus der statistischen Literatur besorgen, andererseits ist die Modellierung eigener Datensätze hinsichtlich der Ausnutzung besonderer Eigenschaften des Visualisierungsalgorithmus' zielführender. Dementsprechend bietet es sich an, vorhandene Tools zur Datengenerierung zusammenzutragen.

Ein Framework zur Generierung synthetischer hochdimensionaler Datensätze wird in [ALM11] beschrieben. Es erlaubt dem Anwender, innerhalb von 3 Hauptschritten eigene Datensätze zu erstellen. Zunächst werden die Rahmenbedingungen festgelegt. Dazu gehören die Anzahl der Dimensionen, Datenpunkte und Klassen, sowie die Verteilungsfunktion des Zufallsgenerators. Anschließend erfolgt die Modellierung der gewünschten Strukturen mittels vorgegebener Generator-Objekte. Es können pro Dimension univariate Wahrscheinlichkeitsverteilungen parametrisiert werden, pro Dimensionspaar 2-dimensionale Verteilungen gezeichnet werden oder eine Generatorebene im 3-dimensionalen Raum frei platziert werden. Auf der Generatorebene kann dann ebenfalls eine 2-dimensionale Verteilungsfunktion gezeichnet werden. Die stellt eine wichtige Besonderheit dar, mit der achsenunabhängige Strukturen erzeugt werden können (siehe Abb. 2.8). Der letzte Schritt ist dann die Ausführung der Datengenerierung auf Basis der gegebenen Bedingungen und Objekte.

Ein weiteres Programm ist PreDO [Vej00] und wurde u.a. in [Nv06] zur Datengenerierung eingesetzt. Das Tool verfolgt einen mengentheoretischen Ansatz, bei dem die Dimensionen eine gleichverteilte Gesamtmenge aufspannen und Strukturen durch Klassen als Teilmengen definiert werden. D.h., die Elemente der Teilmengen werden durch

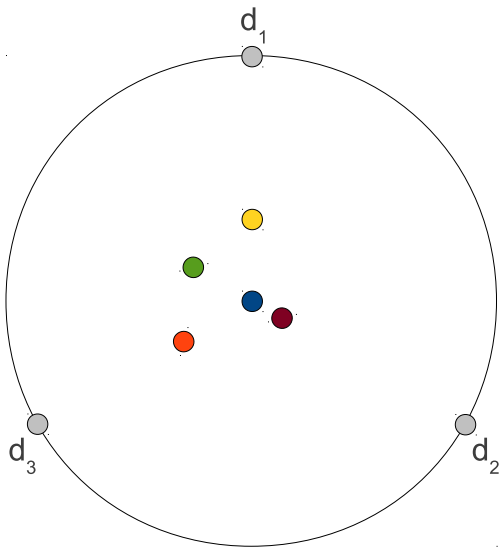


Abbildung 2.7: Radviz-Visualisierung des Beispieldatensatzes

Eigenschaften der Klassen festgelegt. Beispielsweise kann eine Einheitskreisscheibe innerhalb der Dimensionen d_1 und d_2 durch folgende Klasse k_1 beschrieben.

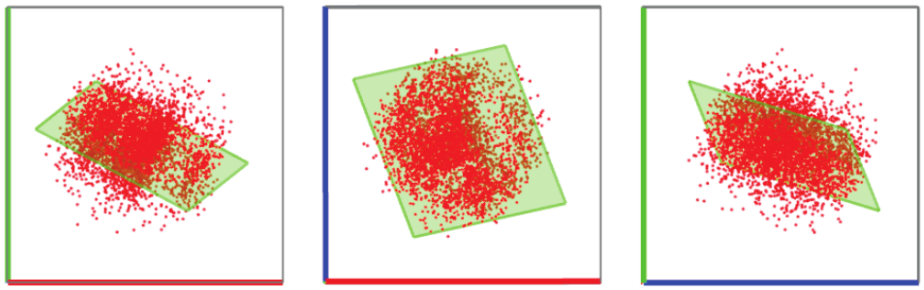
$$k_1 = \{(x, y) \in d_1 \times d_2 \mid x^2 + y^2 < 1\}$$

Ein Zufallsgenerator produziert schließlich so lange gleichverteilte Daten, bis eine vorgegebene Anzahl von Datenwerten erreicht ist, welche die Eigenschaften aller Klassen erfüllt. Die Klassendichte kann zusätzlich jeweils als prozentualer Anteil der Gesamtmenge definiert werden.

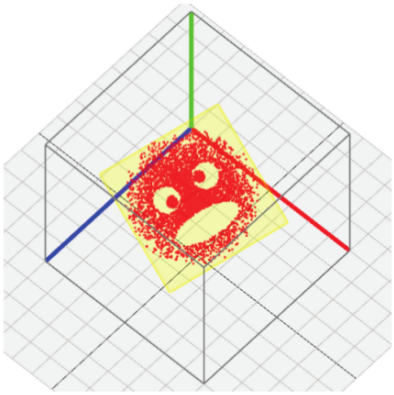
2.4.2 Analyse von Visualisierungstechniken

Untersuchungen bezüglich der Darstellung von Trends und Ausreißern mittels Radviz wurden in [Nv09] veröffentlicht. Trends beschreiben die Entwicklung der Datenwerte entlang einer Dimension. Dabei können drei Verhaltensweisen benachbarter Datenwerte beobachtet werden: wachsen, fallen und gleich bleiben. Um diese Attribute in Radviz identifizieren zu können, wurde zunächst bestimmt, welche Teilmengen der Daten in welchem Bereich der Visualisierung platziert sind. Die Abb. 2.9 zeigt die Aufteilung einer 4-dimensionalen Radviz-Visualisierung in ihre vier Quadranten und der Teilmengen entsprechenden Relationspaare.

Damit lassen sich Trends in den Dimensionen zwischen drei aufeinanderfolgenden Punkten t_0, t_1, t_2 visualisieren, indem die Ankerpunkte als gegenüberliegende Relationspaare $t_0 \leftrightarrow t_1$ und $t_1 \leftrightarrow t_2$ gesetzt werden (siehe Abb. 2.10). Der erste Quadrant beinhaltet dann beispielsweise alle Dimensionen mit durchweg fallender Tendenz ($t_0 > t_1$ und $t_1 > t_2$), während der dritte Quadrant jene mit durchweg steigender Tendenz ($t_0 < t_1$



(a) Dimensions 3 and 8 (b) Dimensions 3 and 9 (c) Dimensions 9 and 8



(d) Hidden structure

Abbildung 2.8: Eine nicht-orthogonale Struktur ist in den Scatterplots nicht erkennbar [ALM11, Abb. 16].

und $t_1 < t_2$) enthält. Um mit dieser Methode Trends über mehr als drei Punkte zu verfolgen, kann man die Visualisierung rekursiv für jeden Quadranten fortsetzen.

Des Weiteren behandelt die Veröffentlichung eine Methode zur Verstärkung univariater Ausreißer nach Hawkins [Haw80]. Dazu erhält jede Dimension einen neuen entgegengesetzten Ankerpunkt, der für jeden Punkt einen konstanten Wert aufweist. Dieser Wert ist das arithmetische Mittel aller Wert der gegenüberliegenden Dimension. Damit entspricht die Anordnung einem Paar Federn mit jeweils fester und variabler Federkonstante. Das hat zur Folge, dass Werte, die nahe dem Mittelwert der Dimension liegen, ebenfalls nahe dem Mittelpunkt der Radviz-Visualisierung platziert werden. Univariate Ausreißer liegen dann entsprechend weit entfernt vom Mittelpunkt.

Hierbei ist es gegebenenfalls notwendig, die Dimensionspaare zu spiegeln, damit die Ausreißer tatsächlich zum Vorschein kommen. Das liegt daran, dass sich die Position aus der Summe der durch die Federkonstanten (Koordinaten) des Datenwerts gewichteten Positionsvektoren der Ankerpunkte ergibt. Konträre Federkräfte (niedrige und hohe) können sich aber ausgleichen und somit zu einer Vektorsumme von 0 führen. Der Vorzeichenwechsel durch die Spiegelung erwirkt im Zweifelsfall das Zusammenspiel der

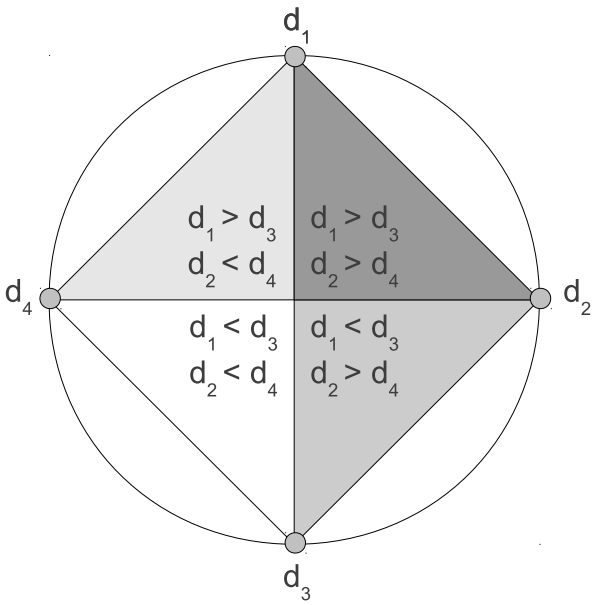


Abbildung 2.9: Eigenschaften der Punkte der Quadranten einer 4-dimensionalen Radviz-Visualisierung [Nv09, nach Abb. 1]

Federkräfte und somit die gewünschte vom Mittelpunkt weit entfernte Platzierung der Ausreißer.

Bei der Untersuchung der Darstellung von Cluster mittels Radviz [Nv06] sind zwei wesentliche Beobachtungen hervor gegangen. Die erste Beobachtung besagt, dass alle Punkte des $[0, 1]$-normalisierten n-dimensionalen Raumes, die auf einer Linie liegen, welche den Ursprung kreuzt, an der gleichen Stelle der Radviz-Ebene positioniert werden. Dies wirkt sich im folgenden Beispiel negativ aus. Es existiert eine Hohlkugel mit Mittelpunkt im Ursprung und einem bestimmten Radius, sowie eine Kugel ebenfalls mit Mittelpunkt im Ursprung aber mit kleinerem Radius (siehe Abb. 2.11. Trotz klarer Trennung der Cluster im Datensatz stellt Radviz die Punkte beider Strukturen gleichermaßen verteilt und überlappend dar. Die vorgeschlagene Lösung dieses Problems ist, Radviz um eine z-Achse im Mittelpunkt zu erweitern, die den Abstand vom Ursprung darstellt.

Die zweite Beobachtung ist eine Herleitung, dass ein kleiner Abstand zwischen 2 Punkten im Datensatz, bei denen die $[0, 1]$-normalisierten Koordinaten größer 0.5 sind, in der Radviz-Visualisierung ebenfalls klein bleibt. Andersherum verzerrt sich die Darstellung der Abstände, umso kleiner die Koordinaten des Datensatzes sind. Existieren beispielsweise zwei Kugeln mit gleichem Radius, wobei eine Kugel nahe dem Ursprung und die andere entfernt davon liegt, überdeckt die Radviz-Darstellung der entfernten Kugel trotz gleicher Clustergröße, die der nahen (siehe Abb. 2.12). Der Lösungsvorschlag ist dementsprechend eine Normalisierung des Datensatzes auf $[0.5, 1]$ statt $[0, 1]$.

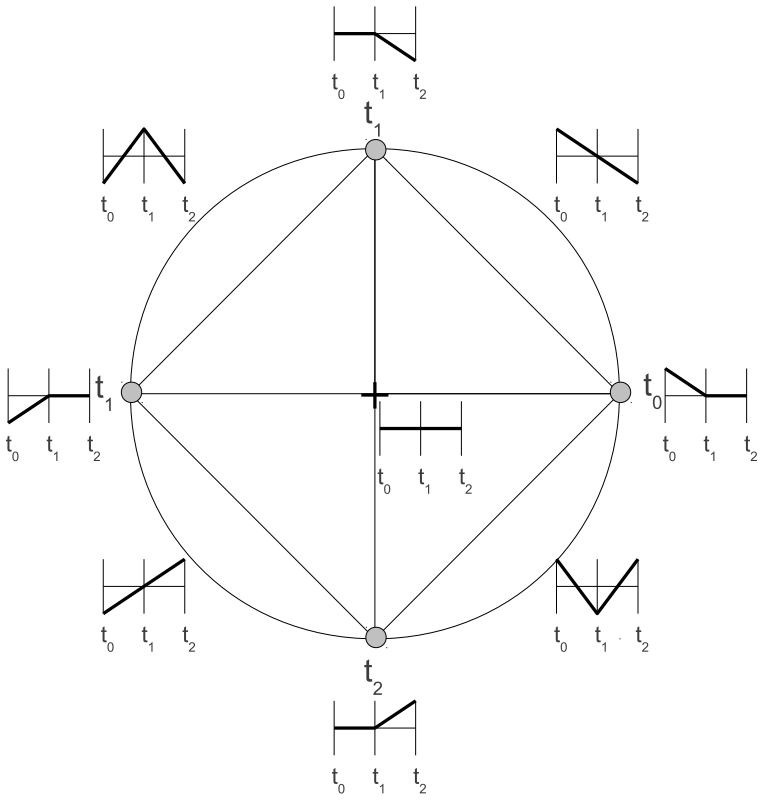


Abbildung 2.10: Eine angepasste 4-dimensionale Radviz Visualisierung ermöglicht die Darstellung von 8 verschiedenen Trends über 3 Zeitpunkte (nach Abbildung 6 aus [Nv09])

2.4.3 Statistik- und Visualisierungstools

Eine Vorauswahl von kostenfreier Statistiksoftware ergab 4 populäre Vertreter: *R*, *SciPy*, *Octave* und *PSPP*. R [SC11] ist eine alternative Implementierung der Programmiersprache und Arbeitsumgebung S für statistische Berechnungen und Visualisierungen. SciPy [Com11] ist eine Bibliothek für die Programmiersprache Python und stellt Algorithmen und Datenstrukturen für das wissenschaftliche Rechnen bereit. Octave [Eat11] ist eine Arbeitsumgebung für mathematische Probleme, deren Skriptsprache weitgehend mit MATLAB kompatibel ist. PSPP [FSF11] ist schließlich eine freie Alternative für die bekannte proprietäre Statistiksoftware SPSS. Der Vergleich der 4 genannten Statistik-Tools hinsichtlich ihrer Möglichkeiten zur Berechnung der in Abschnitt 2.2 genannten statistischen Merkmale (siehe Tabelle 2.2) hat ergeben, dass R als einzige Software von vornherein alle geforderten Funktionen bietet und entsprechend zur Verwendung innerhalb der Diplomarbeit ausgewählt wird.

Gleichermaßen wurde eine Vorauswahl kostenfreier Visualisierungs-Tools getroffen. Zusätzlich zu den Visualisierungsmöglichkeiten von R sind *orange*, *GGobi*, *XmdvTool*

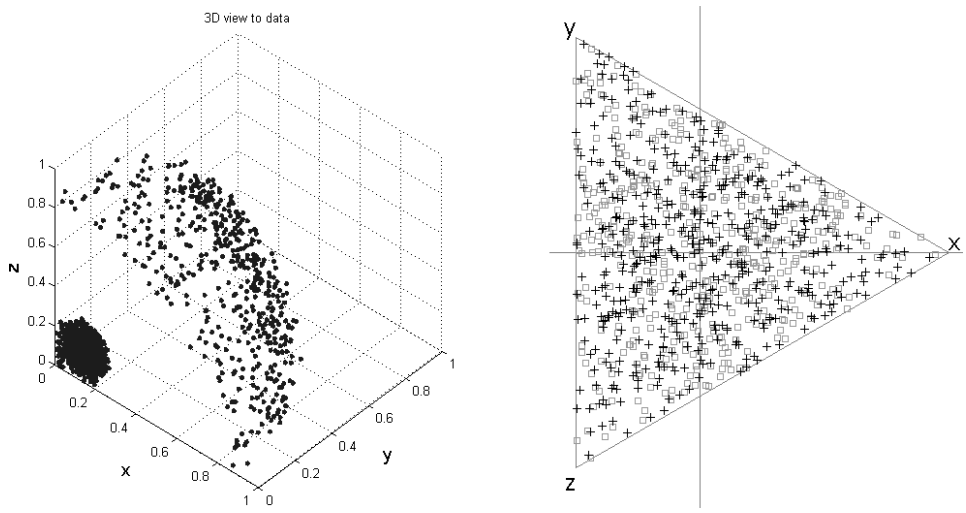


Abbildung 2.11: Klar getrennte Punktmengen im Datensatz (links) sind in der Radviz-Visualisieurng (rechts) nicht unterscheidbar. [Nv06, Abb. 3, Abb. 4]

	Verteilungstest	Korrelation	Clusteranalyse	Ausreißertest
R	√	√	√	√
SciPy	√	√	√	×
Octave	√	√	×	×
PSPP	√	√	×	×

Tabelle 2.2: Statistikfunktionen verschiedener Tools

und *HCE* bekannte Alternativen. Orange [Ls11] ist ein Python-basiertes Data-Mining- und Visualisierungs-Framework, die den Prozess der Datenanalyse als visuelle Programmierung nach dem Datenstrom-Paradigma umsetzt. GGobi [TGF11] und XmdvTool [War11] sind beides Komplettpakete zur interaktiven Exploration und zur dynamischen Darstellung von hochdimensionalen Datensätzen. Der Hierarchical Clustering Explorer (HCE) [SS11] ist ein Tool, das speziell mit Fokus auf die Exploration von Cluster in Datensätzen entworfen wurde. Im direkten Vergleich zeigt sich, dass das orange-Framework die meisten geforderten Visualisierungstechniken unterstützt (siehe Tabelle 2.3) und somit innerhalb der Diplomarbeit für die Visualisierungen der Datensätze verwendet wird. Die fehlende Matrix-Anordnung der Scatterplots wird hierbei mit Hilfe eines eigenen Python-Skripts umgesetzt.

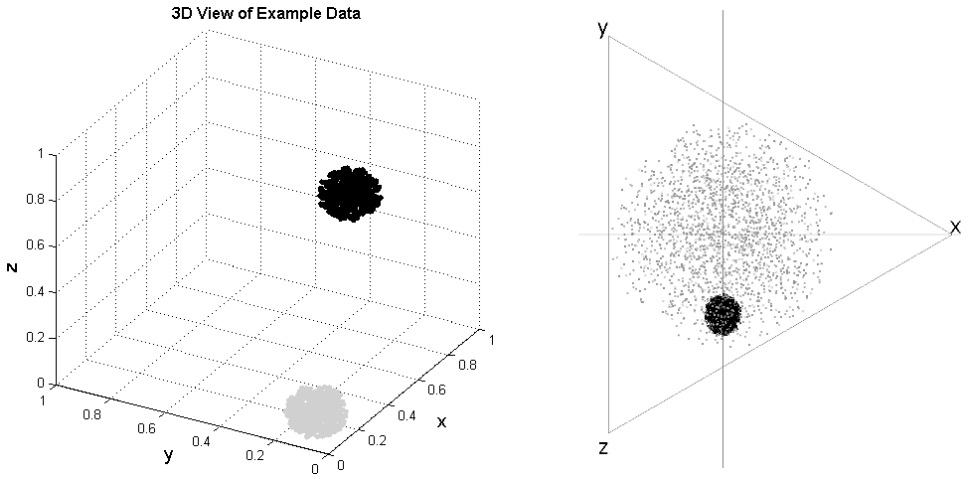


Abbildung 2.12: Gleichgroße und klar getrennte Punktmengen im Datensatz (links) können aufgrund der wertabhängigen Skalierung zu einer Überlappung in der Radviz-Visualisierung (rechts) führen. [Nv06, Abb. 5, Abb. 6]

	Scatterplot/-Matrix	parallele Koordinaten	Radviz	MDS
R	√/√	√	×	√
orange	√/×	√	√	√
GGobi	√/√	√	×	√
XmdvTool	√/√	√	×	×
HCE	√/√	√	×	×

Tabelle 2.3: Visualisierungstechniken verschiedener Tools

3 Analyseregeln

Das Hauptanliegen der Diplomarbeit liegt in der Suche nach generischen Regeln für Visualisierungen, anhand derer Eigenschaften von hochdimensionalen Datesätzen analysiert werden können, unbeeinflusst von subjektiven Einflüssen des Betrachters. Dazu wird im Folgenden die Forschungsmethodik definiert.

Es wird für jedes Forschungsbeispiel ein synthetischer Datensatz D mit n-Dimensionen V_j generiert: $D = \{V_1, V_2, \ldots, V_n\}$. Eine Dimension ist eine Menge von reellen Werten aus dem Intervall $[0, 1]$. Um vergleichbare Ergebnisse verschiedener Datensätze zu erhalten, werden alle verwendeten Datensätze durch eine Intervallskalierung standardisiert. Dies erfolgt mit der Umrechnung aller Werte einer Dimension mit: $V_j = \frac{V_j - \min(V_j)}{\max(V_j) - \min(V_j)}$. Zur Modellierung von Strukturen in den synthetischen Daten werden Stichproben der Gleich-, Normal- oder Exponentialverteilung entnommen. Diese Verteilungen wurden ausgewählt, da sich die Gleichverteilung ($\mathcal{U}(a, b)$) mit ihren scharfen Intervallgrenzen am genauesten und leichtesten zur Modellierung eignet und die Normal- ($\mathcal{N}(\mu, \sigma)$) und Exponentialverteilung ($\mathrm{Exp}(g)$) eine sehr hohe Relevanz als Modelle für naturwissenschaftliche Vorgänge aufweisen. Die Strukturen werden so modelliert, dass sie Informationen in Form von Korrelation, Clusterbildung und Ausreisserfaktor, sowie Ausdehnung und geometrische Form beinhalten. Ein synthetischer Datensatz setzt sich dabei aus mehreren Strukturen mit verschiedenen Ausprägungen dieser Informationen zusammen. Anschliessend wird der Datensatz mittels Visualisierungstechniken dargestellt.

Zunächst werden aufgrund ihrer Dualität die beiden 2-dimensionalen Techniken Scatterplot und parallele Koordinaten zusammen untersucht, gefolgt von der Radviz-Technik, sowie der multidimensionalen Skalierung. Es erfolgt die Erkennung besonderer Merkmale in den Plots, deren Bedeutung in Relation zu den modellierten Eigenschaften des jeweiligen Beispieldatensatzes interpretiert werden. Diese Interpretationen dienen dann der Formulierung der Analyseregeln, welche im Abschluß diskutiert werden. Der letzte Schritt ist die Validierung der Regeln. Dazu erfolgt im folgenden Kapitel eine Anwendung der Regeln auf reale Datensätze mit anschließender Gegenüberstellung der Ergebnisse zu den tatsächlich enthaltenen Informationen. Vorerst werden in den folgenden Abschnitten Beispiele erläutert, interpretiert und entsprechende Regeln abgeleitet.

3.1 Bivariate Visualisierungen - Scatterplot und Parallele Koordinaten

Die Linie-Punkt-Dualität zwischen den beiden ausgewählten 2-dimensionalen Visualisierungstechniken Scatterplot und parallelle Koordinaten legt nahe, beide Methoden nicht getrennt, sondern gegenüberstellend zu betrachten. Damit wird es möglich, solche Regeln zu finden, die die Zusammenhänge zwischen beiden Plots verdeutlichen und somit das Verständnis der Dualität stärken. Ein Anwender, welcher noch keine Erfahrungen mit parallelen Koordinaten gemacht hat, kann somit der Einstieg erleichtert werden, indem er sich dadurch die duale Abbildung im zumeist bekannten Scatterplot vorstellen kann.

Entsprechend der Anzahl der dargestellten Dimensionen des SP und PCP besteht jeder folgende Datensatz aus 2 Dimensionen $D = \{V_1, V_2\}$. Die erste Dimension wird im SP auf der X-Achse und im PCP auf der linken Achse abgetragen und die zweite Dimension entsprechend auf der Y-Achse, bzw. auf der rechten Achse. Weil alle Datensätze auf $[0, 1]$ skaliert sind, entfällt die Achsenbeschriftung zur besseren Übersicht. Im SP befindet sich der $(0, 0)$ links unten und $(1, 1)$ rechts oben. Bei den Achsen im PCP ist der Nullpunkt ebenfalls unten und die 1 oben. Die Punktgröße im SP beträgt 3pt und die Linienbreite im PCP beträgt 1pt. Zur annähernden Darstellung der Dichten der Punktewolken bzw. Linienbündel werden die einzelnen Punkte bzw. Linien transparent geplottet. Wenn nicht anders angegeben, entspricht der PCP in allen Abbildungen dem darüberliegenden SP.

3.1.1 Untersuchung

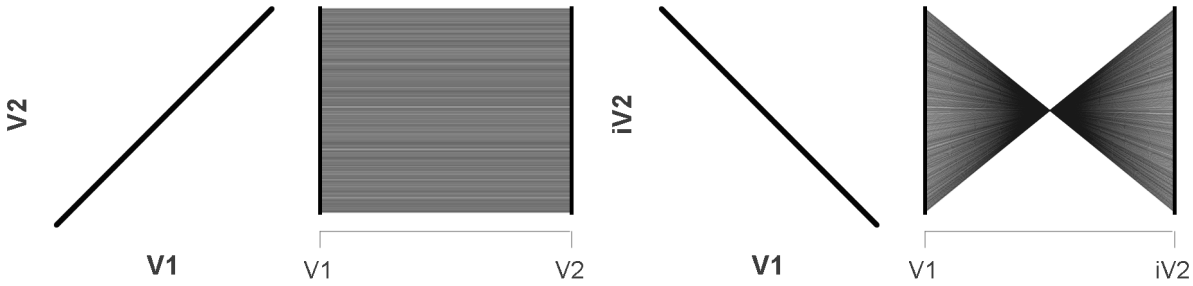
Zur Einführung wird die Linie-Punkt-Dualität verdeutlicht. Hierzu wird eine komplett Gerade modelliert, welche mit positiven Anstieg gleichermaßen eine vollständige lineare Abhängigkeit zwischen beiden Dimensionen erzeugt.

Beispiel 1. *Der Datensatz besteht aus gleichverteilten Werten, wobei* $D = \{p_{i,j} \in [0, 1]^2 \mid \forall i : p_{i,1} = p_{i,2}\}$.

Dazu wird für die ersten Dimension eine Stichprobe von $n = 1000$ Werten der Gleichverteilung entnommen, sodass $V_1 \sim \mathcal{U}(0, 1)$. Die zweite Dimension ergibt sich aus exakt den gleichen Werten: $V_2 = V_1$. Der Korrelationskoeffizient zwischen beiden Dimensionen beträgt dann $\varrho(V_1, V_2) = 1$(siehe Anhang: `pcp/cor.csv`).

Der SP zeigt erwartungsgemäß eine aufsteigende Linie mit gleichmäßiger Dichte (s. Abb. 3.1a). Der korrespondierende PCP (Abb. 3.1b) zeigt eine zu den Achsen orthogonale rechteckige Form mit gleichmäßiger Dichte und paralleler Linienstruktur. Das lässt sich aus der Tatsache ableiten, dass die einzelnen Einträge im Datensatz für beide Dimension die gleichen Werte aufweisen. Es handelt sich um den Sonderfall der Linie-Punkt-Dualität, da eine Linie im SP mit Anstieg von $m = 1$ im PCP aufgrund der resultierenden Division durch Null nicht definiert ist (siehe Abschnitt 2.3.2).

Invertiert man die zweite Dimension durch $V_{2i} = \max(V_2) - V_2 + \min(V_2)$ wechselt im SP das Vorzeichen des Linienanstiegs und der Schnittpunkt mit der Y-Achse erhält den



(a) SP (b) PCP (c) SP, V2 invertiert (d) PCP, V2 invertiert

Abbildung 3.1: Visualisierung von Beispiel 1

Wert des vorigen Schnittpunkts mit der Intervallgrenze. D.h. im SP entsteht eine nun absteigende Linie bei gleichbleibender Dichte (Abb. 3.1c). Im PCP ändert sich sowohl Form, als auch Dichte (Abb. 3.1d). Das Rechteck wird zum verschränkten Trapez mit Kreuzungspunkt im Zentrum des Plots. Die Dichte nimmt Richtung Zentrum parallel zur Vertikalen zu und hat ihr Maximum im Zentrum. Ausgehend vom Zentrum ist eine Linienstruktur erkennbar, die sich nach außen hin auffächert. Das Maximum ist die duale Repräsentation der Linie und liegt wegen $m = -1$ und $n = 1$ genau im Zentrum des Plots.

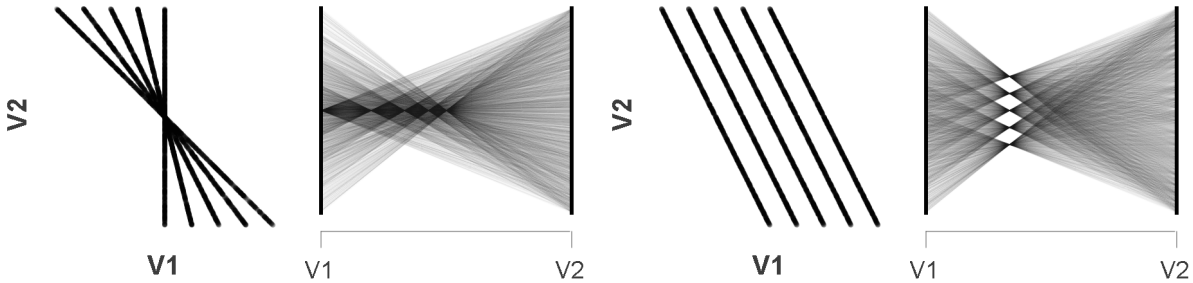
Eine Struktur mit negativen Anstieg, und damit mit negativer Korrelation, erzeugt also einen charakteristischen Kreuzungspunkt mit maximaler Dichter im PCP, der bei positiver Korrelation nicht vorhanden ist. Dieser Kreuzungspunkt ist aufgrund seiner höheren Dichte und scharfkantigem Abstieg zu den Formgrenzen weitaus markanter und auffälliger als die rechteckige gleichverteilte Dichte mit paralleler Linenstruktur. Durch Invertierung einer Dimensionsachse, lassen sich die Parallelitäten jedoch stets in einen Kreuzungspunkt formen, weshalb sich die weiteren Betrachtungen auf den Krezungspunkt als markante Eigenschaft konzentrieren.

Als weiterer Schritt zum Verständnis der Auswirkungen der Linie-Punkt-Dualität werden im nächsten Beispiel die Auswirkung verschiedener Anstiege und Positionen linearer Strukturen verdeutlicht. Hierfür wird eine absteigende Linie modelliert und im Anstieg, sowie in der Verschiebung variiert.

Beispiel 2. *Der Datensatz besteht aus gleichverteilten Werten, wobei* $D = \{p_{i,j} \in [0,1]^2 \mid \forall i : p_{i,1} = mp_{i,2} + \frac{n}{-m}\}$ *mit* $-\infty \leq m \leq -1$ *und* $2 \leq n \leq 4$.

Für die Variation des Anstiegs m wird jeweils eine Stichprobe von $n = 1000$ Werten der Gleichverteilung entnommen, sodass $V_1 \sim \mathcal{U}(\frac{1}{2} + \frac{1}{2m}, \frac{1}{2} - \frac{1}{2m})$. Die zweite Dimension ergibt sich aus: $V_2 = mV_1 + \frac{1}{2} * (-m + 1)$. Der Datensatz besteht aus 5 Variationen mit $m = \{-1, -\frac{4}{3}, -2, -4\}$, sowie dem Sonderfall $m = -\infty$, wobei dann $V_1 = \frac{1}{2}$ und $V_2 \sim \mathcal{U}(0, 1)$ (siehe Anhang `pcp/lines_m.csv`).

Die Verschiebung der Linie bezüglich der V_1 wird mit folgenden Werten durchgeführt: $, d_{V_1} = \{-\frac{1}{4}, -\frac{1}{8}, 0, \frac{1}{8}, \frac{1}{4}\}$. Dazu werden widerum Stichproben von $n = 1000$ gleichver-



(a) SP (b) PCP (c) SP (d) SP

Abbildung 3.2: Visualisierung von Beispiel 2

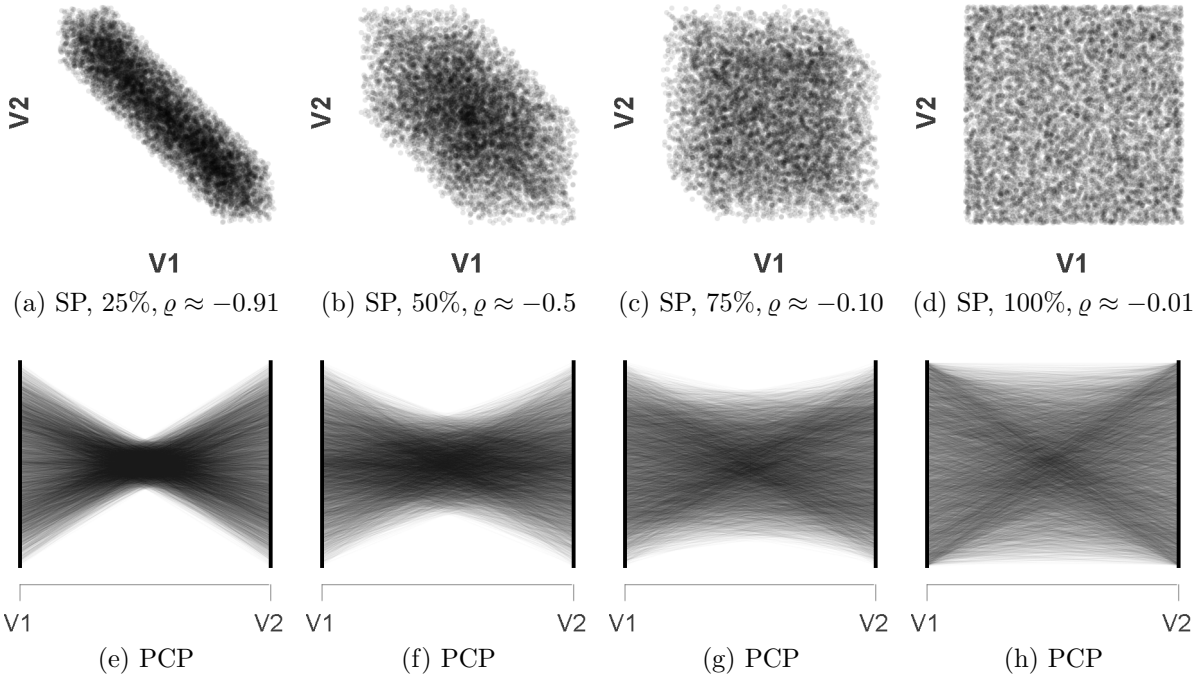
teilten Werten entnommen, $V_1 \sim \mathcal{U}(\frac{1}{4} + d_{V_1}, \frac{3}{4} + d_{V_1})$ und die zweite Dimension mit $V_2 = -2V_1 + 1.5 + 2d_x$ berechnet (siehe Anhang: `pcp/lines_n.csv`).

Der SP des variierenden Anstiegs in Abb. 3.2a zeigt die modellierten absteigenden Linien, die sich im Zentrum kreuzen. In der Visualisierung mittels PCP (Abb. 3.2b) sieht man die entsprechenden Kreuzungspunkte der Doppeltrichter, welche sich bei gleichbleibender vertikaler Position, horizontal zur V_1-Achse verschoben haben. Die Parallelverschiebung der Linien um d_{V_1} ist in Abb. 3.2c zu erkennen . Die duale Repräsentation (Abb. 3.2d) zeigt, dass sich die Kreuzungspunkte bei gleichbleibender horizontaler Position, vertikal verschieben.

Im SP kann man somit den Anstieg der Linie mittels 2 Punkten dieser Linie berechnen: $m = \frac{\Delta y}{\Delta x}$. Die horizontale Verschiebung lässt sich direkt an der horizontalen Achse ablesen. Beim PCP ergibt sich der Anstieg aus der horizontalen Position des Kreuzungspunktes durch $m = 1 - \frac{d}{x}$, wobei d der Referenzabstand zwischen den Dimensionsachsen ist. Die horizontale Verschiebung berechnet sich hier aus dem vertikalen Abstand zwischen 2 Kreuzungspunkten mit $d_x = \frac{\Delta y(1-m)}{m}$. Daraus ergibt sich, dass lineare Strukturen mit gleichem Schnittpunkt im PCP ein Maximum entlang einer horizontalen Linie bilden, wohingegen parallele lineare Strukturen ein Maximum entlang einer vertikalen Linie bilden. Dieses Zusammenspiel ist die Grundlage zum Anwendungszweck des PCP zur Liniendetektion in Bildern, wie sie in [DHH11] als Alternative zur Hough-Transformation vorgestellt wird.

Die in den vorhergehenden Beispielen modellierte vollständige lineare Abhängigkeit zwischen zwei Dimensionen, also 100% Korrelation, ist in realen Datensätzen praktisch nicht vorhanden. Rauschen verschlechtert die Korrelation, da es naturgemäß vollständig unkorreliert ist. Die folgenden 3 Beispiele untersuchen den Einfluss von Rauschen und damit Strukturbreite auf die Korrelation, unter einer zusätzlichen Abhängigkeit vom Anstieg. Als Erstes wird eine Linie mit negativen Anstieg modelliert, bei der der Einfluss durch gleichverteiltes Rauschen variiert wird.

Beispiel 3. *Der Datensatz besteht aus gleichverteilten Werten, wobei* $D = \{p_{i,j} \in [0,1]^2 \mid \forall i : p_{i,1} = 1 - p_{i,2}\}$. *Der Datensatz wird in verschiedenen Ausprägungen von 0% bis 100%*



(a) SP, 25%, $\varrho \approx -0.91$ (b) SP, 50%, $\varrho \approx -0.5$ (c) SP, 75%, $\varrho \approx -0.10$ (d) SP, 100%, $\varrho \approx -0.01$

(e) PCP (f) PCP (g) PCP (h) PCP

Abbildung 3.3: Visualisierung von Beispiel 3

bezüglich des Wertebereichs $[0, 1]$ *gleichverteilt verrauscht.*

Die erste Dimension besteht aus einer Stichprobe von $n = 10000$ gleichverteilten Werten, $V_1 \sim \mathcal{U}(0, 1)$. Die zweite Dimension ergibt sich mit $V_2 = 1 - V_1$. Beide Dimensionen werden mit einer prozentualen Ausprägung p verrauscht: $V_1 = noise(V_1, p)$, $V_2 = noise(V_2, p)$. Die Rauschfunktion skaliert die Ursprungsdaten um den Faktor p und addiert gleichverteiltes Rauschen: $noise_{\mathcal{U}}(V, p) = scale(V, 1 - p) + N_p$ mit $N_p \sim \mathcal{U}(-\frac{p}{2}, \frac{p}{2})$ und $scale(V, f) = \bar{V}(1 - f) + fV$. Die verschiedenen Ausprägungen des Rauschens sind $p = \{0\%, 25\%, 50\%, 75\%, 100\%\}$ und ergeben nach der Anwendung folgende Korrelationskoeffizienten $\varrho_p(V_1, V_2) \approx \{-1, -0.91, -0.5, -0.1, -0.01\}$(siehe Anhang: `pcp/g_cor_noise.csv`).

Man erkennt in den SP von Abb. 3.3, dass die absteigende lineare Form mit steigendem Rauscheinfluss direkt proportional breiter wird, während die Dichteverteilung entlang der Diagonalen abflacht, bis bei 100% Rauscheinfluss eine gleichverteilte Dichte mit quadratischer Form erreicht wird. In den PCP steigt die vertikale Ausdehnung des Kreuzungsbereich ebenfalls direkt proportional zur Rauschausprägung. Dabei wird die Dichteverteilung um den Kreuzungspunkt flacher. Bei 100% Rauschen ergibt sich eine rechteckige Form, deren Dichte annähernd gleichverteilt ist, wobei sich eine Häufung entlang der Diagonalen x-förmig abzeichnet.

Die Korrelation zwischen den beiden Dimensionen steigt also beim steigenden absoluten Verhältnis der Ausdehnungen der beiden Hauptkomponenten einer symmetrischen

Struktur im SP. Hat diese Struktur also ein Ausdehnungsverhältnis von 1 und weist eine quadratische Form mit gleichmäßiger Dichte auf, kann man beide Dimensionen als gleichverteilt einstufen. Im PCP sinkt die Korrelation hingegen bei steigender Ausdehnung des Kreuzungsbreiches des verschränkten Trapez'. Steigt die Ausdehnung auf die gesamte Höhe des Plots ergibt sich ein Rechteck mit gleichmäßiger Dichte und markanter X-Struktur, was auf die Gleichverteilung beider Dimensionen hinweist. Betrachtet man SP und PCP zusammen, zeigt sich, dass die Ausdehnung der zweiten Hauptkomponenten einer symmetrischen Struktur in SP der Ausdehnung des Kreuzungspunktes im PCP entspricht.

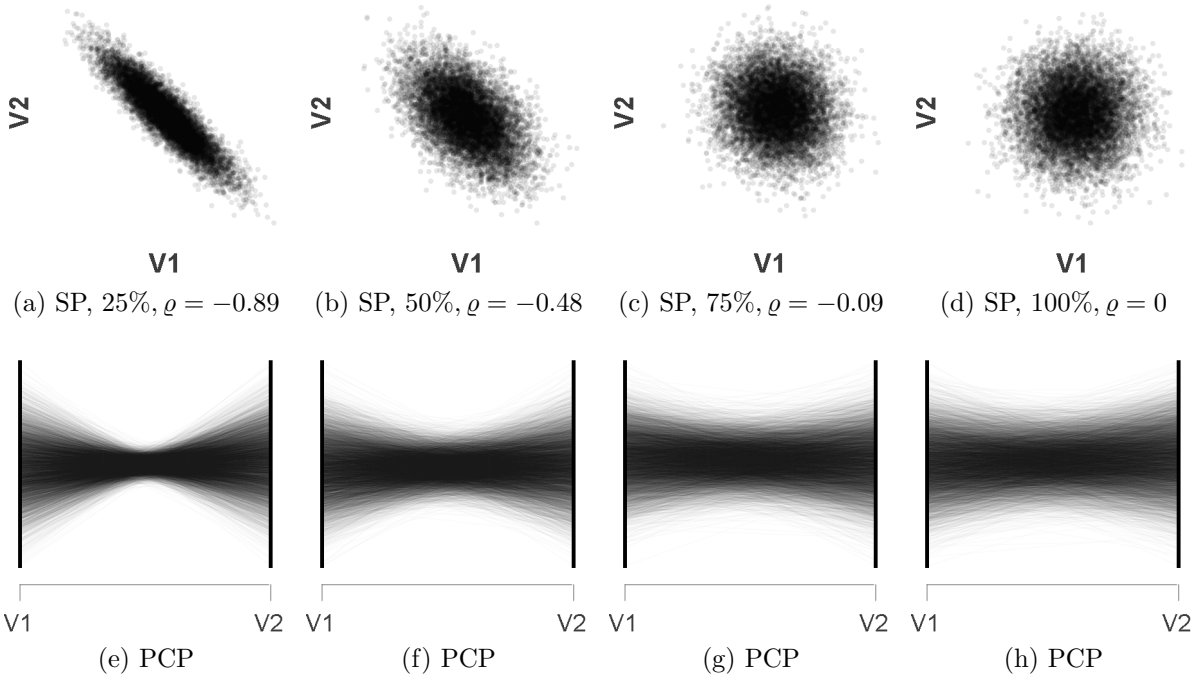
Beispiel 4. *Der Datensatz besteht aus normalverteilten Werten, wobei* $D = \{p_{i,j} \in [0,1]^2 \mid \forall i : p_{i,1} = 1 - p_{i,2}\}$. *Der Datensatz wird in verschiedenen Ausprägungen von 0% bis 100% bezüglich des Wertebereichs* $[0,1]$ *normalverteilt verrauscht.*

Die Vorgehensweise zur Generierung der Daten entspricht weitesgehend Beispiel 3. Die Stichproben werden allerdings aus der Standardnormalverteilung entnommen, sodass $V_1, N_p \sim \mathcal{N}(0,1)$. Da dadurch Werte ausserhalb der geforderten Intervallgrenzen von $[0,1]$ entstehen, werden V_1 und N_p nach der Stichprobenentnahme entsprechned intervallskaliert. Für die verschiedenen Rauschausprägungen ergeben sich ähnliche Korrelationskoeffzienten wie im vorigen Beispiel: $\varrho_p(V_1, V_2) \approx \{-0.89, -0.48, -0.09, 0\}$ (siehe Anhang: `pcp/n_cor_noise.csv`).

Der Einfluss von normalverteiltem Rauschen ergibt im SP eine absteigende ellipsenförmige Struktur mit Dichtehäufung im Zentrum (Abb. 3.4. Die Ellipse wird mit steigendem Rauschen breiter, bis hin zu einer kreisförmigen Struktur. Dabei fällt auf, dass das Verhältnis der Ausdehnung der Hauptkomponenten im Gegensatz zu Beispiel 3 nicht proportional zum Rauschen steigt (vergleiche Abb. 3.4b zu 3.4c mit 3.4c zu 3.4d). Die Form der erzeugten Struktur im PCP nimmt im Kreuzungsbreich ebenfalls nicht proportional mit steigendem Rauschen an Ausdehnung zu. Zudem zeichnet sich eine Abrundung der zuvor linearen Seiten des verschränkten Trapez' oben und unten ab. Die Dichteverteilung weist eine längliche Häufung auf der Horizontalen des Krezungspunkt auf, die mit steigender Rauschausprägung nur leicht abflacht.

Abb. 3.4d zeigt, dass man aus einer kreisförmigen Punktwolke mit zentraler Dichte im SP und ein Rechteck mit nach innen gewölbter oberer und unterer Seite, sowie einer horizontalen Dichtehäufung im vertikalen Zentrum auf 100% unkorrelierte und normalverteilte Dimensionen schließen kann. Umso mehr Korrelation zwischen den Dimension vorhanden ist, umso ellipsenförmiger wird die Punktwolke und umso eckiger das verschränkte Trapez.

Im theoretischen Fall von 100% Korrelation hat der Anstieg der Struktur bis auf die Sonderfälle $m = 0$ und $m = \pm\infty$ keinen Einfluss auf den Korrelationskoeffizienten. Der Einfluss von Rauschen erhöht jedoch das Ausdehnungsverhältnis der Struktur, was widerum bei der Berechnung des Korrelationskoeffizienten den Einfluss einer Dimension bei einem geringen oder hohen absoluten Anstieg verstärkt. Diesen Zusammenhang verdeutlicht die Berechnung des Korrelationskoeffizienten einer Struktur in Abhängigkeit



(a) SP, 25%, $\varrho = -0.89$ (b) SP, 50%, $\varrho = -0.48$ (c) SP, 75%, $\varrho = -0.09$ (d) SP, 100%, $\varrho = 0$

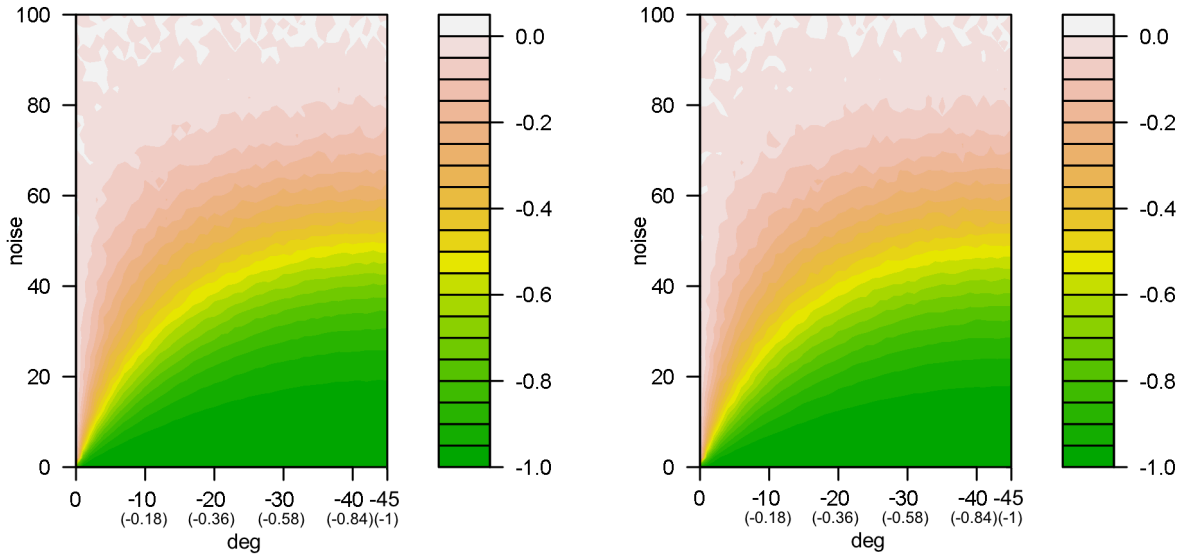
(e) PCP (f) PCP (g) PCP (h) PCP

Abbildung 3.4: Visualisierung von Beispiel 4

von variierenden Rauscheinfluss und Anstieg. Dazu werden die Daten wie in Beispiel 3 und 4 generiert, wobei sich die zweite Dimension aus $V_2 = mV_1 + \frac{1}{2} * (-m + 1)$ mit $-1 \leq m \leq 0$ ergibt. Der Korrelationskoeffizient wird nun für den gesamten Bereich des Rauschens $0 \leq p \leq 1$ mit $\varrho(noise(V_1, p), noise(V_2, p))$ berechnet.

Diese bivariate Funktion ist in Abb. 3.5 als farbiges Höhenfeld dargestellt. Der Wert des Korrelationskooeffizienten ist entsprechend der Legende der Farbskala zugeordnet und ist abhängig vom Anstieg in Grad (X-Achse) und der Ausprägung des Rauschens in Prozent (Y-Achse). Das Höhenfeld zeigt ein gleiches Verhalten im gleichverteilten (Beispiel 3) und normalverteilten (Beispiel 4) Fall. Man erkennt, dass im Bereich bis 10% Rauschen der absolute Anstieg um 80% von $\frac{1}{5}$ bis 1 variieren kann, ohne den Korrelationskoeffizienten zu verringern. D.h., umso weniger Rauschen vorhanden ist, desto schwächer ist der Einfluss des Anstiegs. Weiterhin zeigt sich ab einem absoluten Anstieg von $\frac{2}{5}$, dass dessen Anteil nur noch ca. 10% ausmacht, und der Korrelationskoeffizient somit praktisch nur vom Rauschen abhängt.

Im gleichverteilten Fall verhält sich die Strukturbreite direkt proportional zum Rauschen und im normalverteilten Fall bis 50%. Eine Korrelation wird im Allgemeinen ab einem absoluten Koeffizienten von $\frac{1}{2}$ als gut eingestuft, was laut dem Höhenfeldplot auch nur von Rauschen bis 50% erreicht wird. Dementsprechend eignet sich der Plot um den Korrelationskoeffizienten einer symmetrischen Struktur durch visuelles Abschätzen der Breite und des Anstiegs zu bestimmen.



(a) $\mathcal{U}(0,1)$ (b) $\mathcal{N}(0,1)$

Abbildung 3.5: Korrelationskoeffizient abhängig von Anstieg und Ausdehnung der Struktur

Die Beispiele 3 und 4 haben u.a. einen Zusammenhang zwischen der Form einer Struktur im SP (quadratisch, bzw. kreisförmig) und im PCP (rechteckig, bzw. abgerundete Seiten) gezeigt. Entsprechend soll das folgende Beispiel zeigen, welchen Einfluss Formänderungen auf eine absteigende lineare Struktur haben.

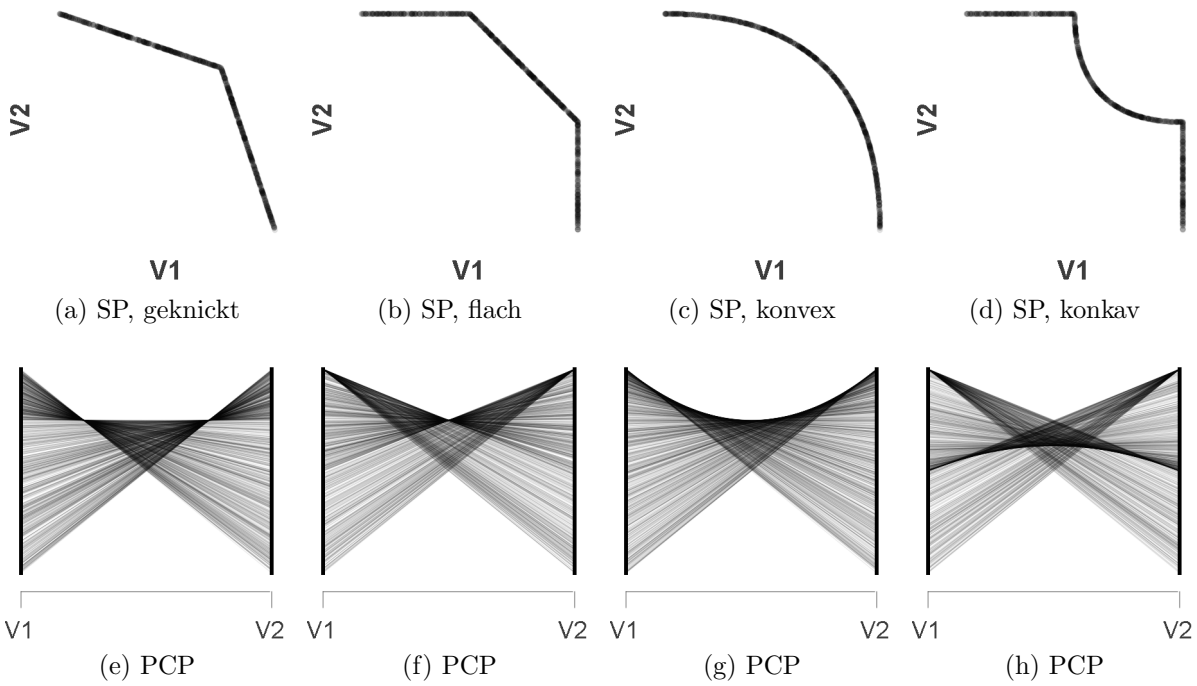
Beispiel 5. *Der Datensatz besteht aus mehreren nicht-linearen Strukturen, deren Start- und Endpunkte sich jeweils an* $(1,0)$ *bzw.* $(0,1)$ *befinden. Ausgehend von der Diagonalen wird eine geknickte, flache, konvexe und konkave Form Richtung (1,1) modelliert.*

Die geknickte Struktur besteht aus 2 Linien a und b für die jeweils $n_a = 500$ und $n_b = 500$ gleichverteilte Werte als Stichprobe genommen werden: $V_{1a} \sim \mathcal{U}(0, \frac{3}{4})$ und $V_{1b} \sim \mathcal{U}(\frac{3}{4}, 1)$. Die zweite Dimension ergibt sich aus $V_{2a} = -\frac{1}{3}V_{1a}+1$ und $V_{2b} = -3V_{1b}+3$. Mit der Vereinigung der Teilstrukturen ergeben sich die Dimensionen $V_1 = V_{1a} \cup V_{1b}$ und $V_2 = V_{2a} \cup V_{2b}$.

Die flache Struktur besteht aus 3 Linien a, b und c mit den Stichprobenumfängen von $n_a, n_b = 333$ und $n_c = 334$. Die Dimensionen setzen sich aus den einzelnen Linienabschnitten zusammen: $V_1 = V_{1a} \cup V_{1b} \cup V_{1c}$ und $V_2 = V_{1a} \cup V_{1b} \cup V_{1c}$. Linie a hat ergibt sich aus $V_{1a} \sim \mathcal{U}(0, \frac{1}{2})$ und $V_{2a} = 1$. Linie b ergibt sich aus $V_{1b} \sim \mathcal{U}(\frac{1}{2}, 1)$, $V_{2b} = -V_{1b} + \frac{3}{2}$. Linie c ergibt sich aus $V_{1c} = 1$ und $V_{2c} \sim \mathcal{U}(0, \frac{1}{2})$.

Für die konvexe Struktur wird eine Stichprobe von $n = 1000$ gleichverteilten Werten entnommen, $V_1 \sim \mathcal{U}(0,1)$. Die zweite Dimension berechnet sich mit $V_2 = \sqrt{(1 - V_1^2)}$.

Die konkave Struktur besteht aus 3 Teilstrukturen mit den Stichprobenumfängen



(a) SP, geknickt (b) SP, flach (c) SP, konvex (d) SP, konkav

(e) PCP (f) PCP (g) PCP (h) PCP

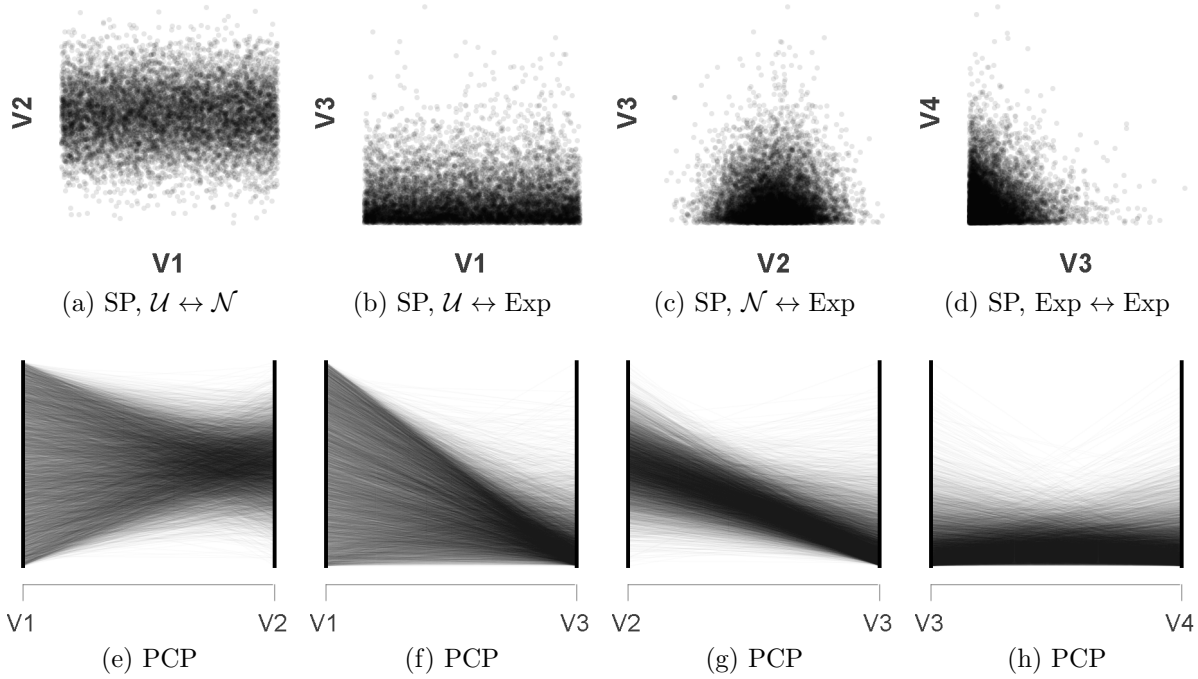
Abbildung 3.6: Visualisierung von Beispiel 5

$n_a, n_b = 333$ und $n_c = 334$. Es gilt $V_1 = V_{1a} \cup V_{1b} \cup V_{1c}$ und $V_2 = V_{2a} \cup V_{2b} \cup V_{2c}$. Die erste Teilstruktur ist eine Linie mit $V_{1a} \sim \mathcal{U}(0, \frac{1}{2})$ und $V_{2a} = 1$. Die Dimensionen der zweiten Teilstruktur ergeben sich aus $V_{1b} \sim \mathcal{U}(\frac{1}{2}, 1)$ und $V_{2b} = \frac{1}{2} + \sqrt{(\frac{1}{4} - (x - \frac{1}{2})^2)}$. Die dritte Teilstruktur ist wieder eine Linie mit $V_{1c} = 1$ und $V_{2c} \sim \mathcal{U}(0, \frac{1}{2})$.

Die Datensätze der Strukturen befinden sich im Anhang unter `pcp/lin.csv`, `pcp/flat.csv`, `pcp/roundhull.csv` und `pcp/lin.csv`.

In Abb. 3.6a zeigt sich der Eckpunkt im SP als horizontale Linie im PCP, im gleichen Abstand oberhalb vom Kreuzungspunkt, wie der Eckpunkt im SP zur Diagonalen. Der PCP weist 2 lokale Maxima am Schnittpunkt der horizontalen Linie mit den Diagonalen des verschränkten Trapez' auf. Die flache Struktur im SP der Abb. 3.6b erzeugt einen neuen Kreuzungspunkt oberhalb des Zentrums. Es entstehen 3 lokale Maxima, die sich oberhalb des Zentrums, links auf der Achse, horizontal mittig, und rechts auf der Achse befinden. Das Verhalten entspricht der Linie-Punkt-Dualität. Betrachtet man die Strukturen als Grenzen symmetrischer Strukturen, kann man schlussfolgern, dass eine kantige Ausdehnung des Kreuzungspunkts im PCP auf eine eckige Form im SP hinweist und eine eckige Ausdehnung im PCP auf eine kantige Form im SP.

Die bezüglich der Diagonalen konvexe Form im SP von Abb. 3.6c ergibt eine bezüglich der Horizontalen im vertikalen Zentrum konkave Rundung der betreffenden oberen Seite des verschränkten Trapez'. Da keine Linie im SP vorhanden ist, ergibt sich kein einzelnes lokales Maximum im PCP, sondern eine Dichtehäufung entlang der gekrümmten oberen



(a) SP, $\mathcal{U} \leftrightarrow \mathcal{N}$ (b) SP, $\mathcal{U} \leftrightarrow$ Exp (c) SP, $\mathcal{N} \leftrightarrow$ Exp (d) SP, Exp $\leftrightarrow$ Exp

(e) PCP (f) PCP (g) PCP (h) PCP

Abbildung 3.7: Visualisierung von Beispiel 6

Seite. Die konkave Form im SP von Abb. 3.6d erzeugt als obere Seite im PCP die gleiche eckige Form wie in Abb. 3.6b. Die Krümmung zeigt sich nur als Dichtehäufung innerhalb des verschränkten Trapez' und ist konvex bezüglich der Horizontalen. Daraus folgt zum einen, dass eine konkave Ausdehnung im PCP auf eine konvexe Form im SP hinweist. Zum anderen lässt sich aus der oberen und unteren Seitenform im PCP nur auf die konvexe Hülle der Struktur im SP schließen.

Das folgende Beispiel ergänzt die in Beispiel 3 und 4 dargestellte charakteristische Form und Dichte von 100% unkorrelierten Dimensionen mit gleich- bzw. normalverteilt Werten. Es zeigt, welche Form und Dichte die Plots bei Dimensionen mit jeweils verschiedener Verteilung aufweisen und betrachtet neben der Gleich- und Normalverteilung, die ebenfalls wichtige Exponentialverteilung.

Beispiel 6. *Der Datensatz hat 4 Dimensionen. Die ersten 3 Dimensionen sind Stichproben der Gleich-, Normal- und Exponentialverteilung. Die 4. Dimension ist eine weitere Stichprobe der Exponentialverteilung.*

Der Stichprobenumfang beträgt jeweils $n = 10000$. Der höhere Umfang dient einer genaueren Formgebung und feineren Dichtestruktur in den Plots. Die Werte der Dimensionen entstammen folgenden Verteilungen: $V_1 \sim \mathcal{U}(0,1)$, $V_2 \sim \mathcal{N}(0,1)$, $V_3 \sim \text{Exp}(1)$ und $V_4 \sim \text{Exp}(1)$.

Der SP von Abb. 3.7a weist eine rechteckige Form mit einer horizontalen Häufung im vertikalen Zentrum auf. Die Häufung flacht symmetrisch nach oben und unten ab. Die

Form in Abb. 3.7b ist ebenfalls rechteckig. Die horizontale Häufung befindet sich jedoch am unteren Ende und hat einen steileren Abstieg nach oben. In Abb. 3.7c zeichnet sich ein gleichschenkliges, spitzwinkliges Dreieck ab, dessen Höhe sich orthogonal zur Vertikalen im horizontalen Zentrum befindet. Die Häufung ist horizontal mittig unten und flacht nach oben, links und rechts ab. Zuletzt zeigt Abb. 3.7d ein gleichschenkliges, rechtwinkliges Dreieck, dessen Spitze sich links unten befindet, während die Höhe auf der Diagonalen mit positiven Anstieg liegt. Die Häufung befindet sich links unten, und flacht nach oben und rechts gleichermaßen ab.

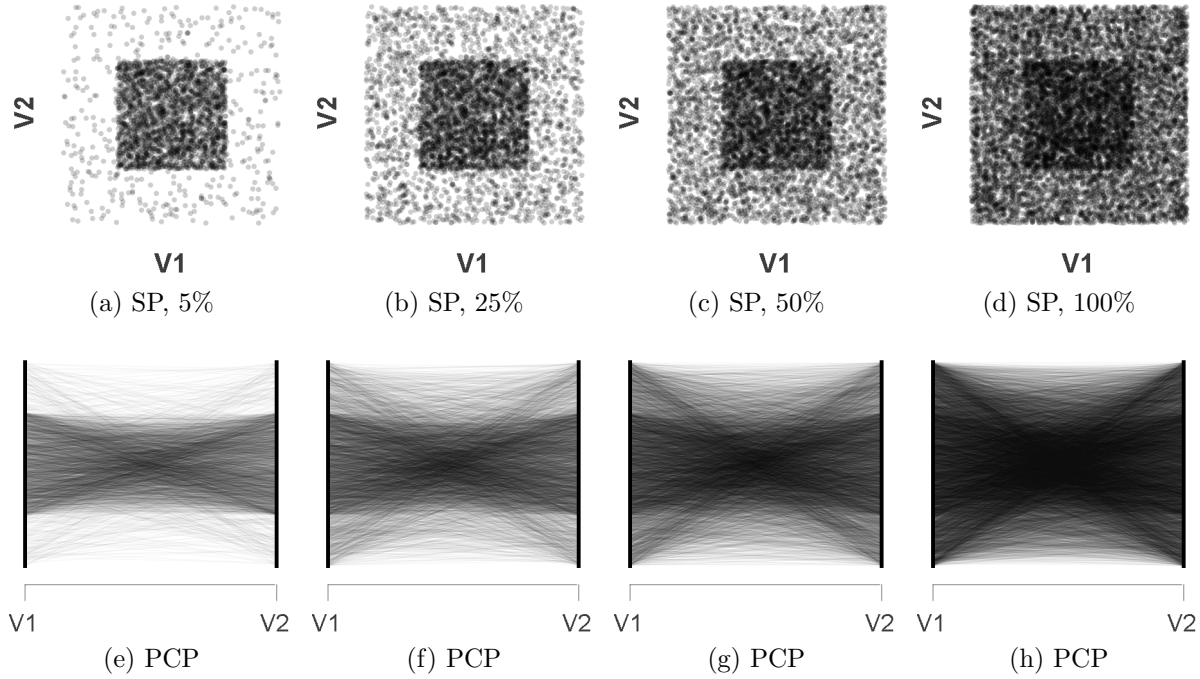
Der PCP von Abb. 3.7a zeigt ein symmetrisches Trapez mit der langen Seite links und der kurzen Seite rechts. Die Dichte nimmt strahlenförmig zulaufend zum breitem Maximum zu, welches sich vertikal mittig auf der rechten Seite befindet. Die Dichte verläuft auf der linken Seite gleichmäßig und auf der rechten Seite nach oben und unten abflachend. Der Plot Abb. 3.7b hat die Form eines rechtwinklinges Trapez', bei dem sich die lange Seite ebenfalls links, und die kurze Seite rechts unten befindet. Die Dichte nimmt strahlenförmig zulaufend zum Maximum rechts unten zu, wobei die linke Seite gleichmäßig verteilt ist, und die rechte Seite nach oben abflacht. In Abb. 3.7c entspricht die Form einem rechtwinkligen Trapez mit oben stark und unten leicht gekrümmten Schenkeln. Die linke Seite ist länger als die rechte Seite. Die Dichtehäufung verläuft von links Mitte, nach rechts unten zum Maximum. Dabei ist die linke Seite nach oben sowie unten, und die rechte Seite nur nach oben abflachend. Der PCP von 3.7d hat die Form eines Rechtecks mit gekrümmter oberer Seite in Höhe der vertikalen Mitte. Die Dichte weist eine horizontale Häufung an der unteren Seite auf, die nach oben abflacht.

Die Art der Verteilung entlang einer Hauptkomponente lässt sich im SP entsprechend wie folgt abschätzen. Gibt es keine Strukturänderung, handelt es sich um eine Gleichverteilung. Eine symmetrische starke Dichteabnahme ausgehend vom zentralen Maximum weist auf eine Normalverteilung hin. Die Exponentialverteilung ist charakterisiert durch eine steile Dichteabnahme in eine Richtung ausgehend von einem seitlichen Maximum. Die gleichen Merkmale gelten eingeschränkt auch für den PCP. Hier können nur Verteilungen orthogonal zu den Dimensionen geschätzt werden.

Die bisherigen Beispiele beziehen sich auf die Untersuchung einer globalen Struktur. In einem Datensatz können aber auch mehrere lokale Strukturen auftreten. Das folgende Beispiel veranschaulicht die Identifizierung solcher Cluster. Dazu wird zunächst als einfachster Fall ein Cluster mit lokalen Ausmaßen erzeugt, der mit variierenden globalen Rauschen überlagert wird.

Beispiel 7. *Im Datensatz existieren 2 sich überlagernde Stukturen: zum einen gleichverteilte Werte in* $[\frac{1}{4}, \frac{3}{4}]^2$*, zum anderen gleichverteilte Wete in* $[0, 1]^2$*.*

Der Cluster wird mit einer Stichprobe von $n_a = 2500$ Werten aus der Gleichverteilung generiert, sodass $V_{1a}, V_{2a} \sim \mathcal{U}(\frac{1}{4}, \frac{3}{4})$. Die verschiedene Stärke des Rauschens wird jeweils durch einen verschiedenen Stichprobenumfang modelliert: $n_b = 500$, $n_c = 2500$, $n_d = 5000$, $n_e = 10000$ und $V_{1b}, V_{1c}, V_{1d}, V_{1e}, V_{2b}, V_{2c}, V_{2d}, V_{2e} \sim \mathcal{U}(0, 1)$. Die Dimensionen des



(a) SP, 5% (b) SP, 25% (c) SP, 50% (d) SP, 100%

(e) PCP (f) PCP (g) PCP (h) PCP

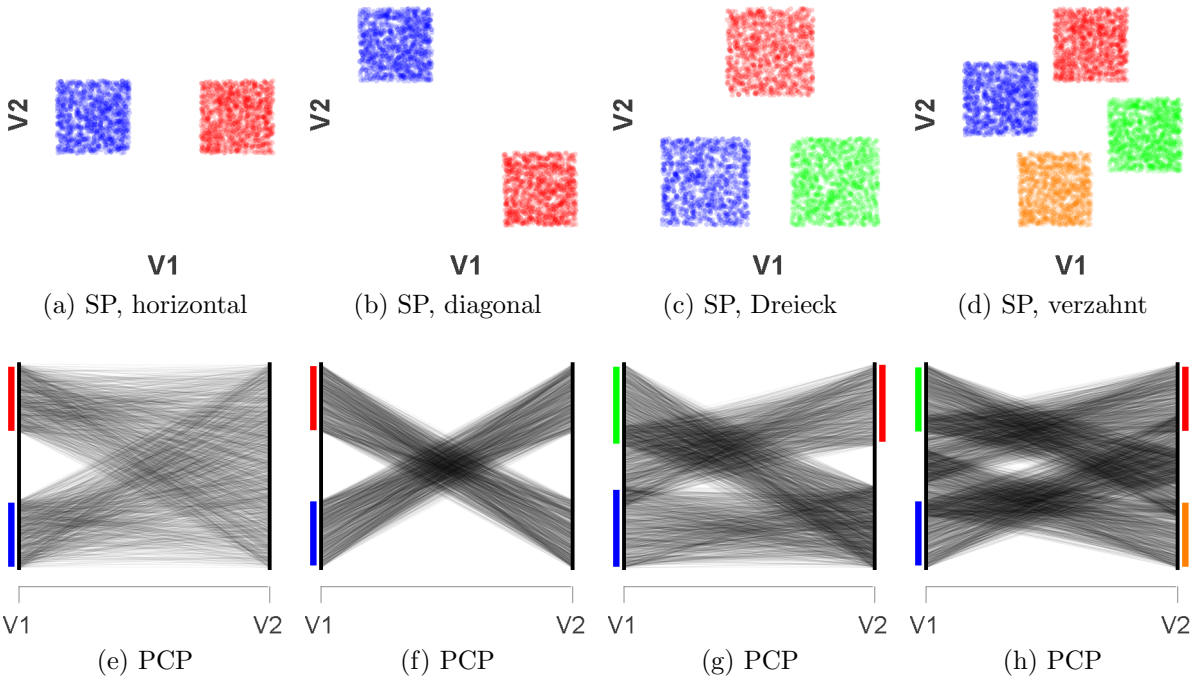
Abbildung 3.8: Visualisierung von Beispiel 7

Datensatzes setzen sich dann aus den Teilstrukturen zusammen, wobei $V_1 = V_{1a} \cup V_{1b} \cup V_{1c} \cup V_{1d} \cup V_{1e}$ und $V_2 = V_{2a} \cup V_{2b} \cup V_{2c} \cup V_{2d} \cup V_{2e}$ (siehe Anhang: `pcp/cl_noise.csv`).

In jedem SP der Abb. 3.8 erkennt man die Form eines kleinen Quadrats im Zentrum eines größeren Quadrats. Das kleinere Quadrat hat eine größere Dichte gegenüber dem globalen Bereich. Mit wachsender Annäherung der Dichten (von a nach d) beider Bereiche wird es schwerer, sie visuell zu unterscheiden. In den PCP zeigt sich ein Rechteck mit geringerer Höhe im Zentrum eines Rechtecks mit größerer Höhe. Der kleinere lokale Bereich ist ebenfalls dichter als der globale Bereich. Die für die Gleichverteilung charakteristische X-Struktur der Dichte, ist im lokalen Bereich entsprechend seiner Höhe gestaucht.

Man erkennt, dass die Form und Struktur des Cluster im SP relativ zum globalen Rauschen horizontal und vertikal skalieren. Im PCP geschieht dies vertikal, aber nicht horizontal. Zudem lassen sich die überlappenden Cluster mit gleicher Dichte (100%) im SP noch visuell unterscheiden, während die Grenzen im PCP kaum zu erkennen sind (vergleiche Abb. 3.8d).

Das vorige Beispiel zeigt wie die Dichte von SP und PCP die Identifizierung eines Clusters ermöglicht. Da mehrere Cluster in einem Datensatz vorhanden sein können, verdeutlicht das nächste Beispiel, inwiefern diese Cluster unterschieden werden können. Dies ist eine Grundlage für die Exploration eines Datensatz durch sukzessive Auswahl, Betrachtung und Ausblendung einzelner Cluster. Hierfür wird eine verschiede Anzahl



(a) SP, horizontal (b) SP, diagonal (c) SP, Dreieck (d) SP, verzahnt

(e) PCP (f) PCP (g) PCP (h) PCP

Abbildung 3.9: Visualisierung von Beispiel 8

von Cluster variierend angeordnet.

Beispiel 8. *Im Datensatz existieren 4 Cluster mit gleichverteilten Werten, die folgendermaßen angeordnet sind: 2 Cluster horizontal, 2 Cluster diagonal, 3 Cluster als Dreieck und 4 Cluster als verzahntes Viereck.*

Die Cluster a, b, c und d werden jeweils aus einer Stichprobe von $n = 1000$ gleichverteilten Werten generiert: $V_{1a}, V_{1b}, V_{1c}, V_{1d}, V_{2a}, V_{2b}, V_{2c}, V_{2d} \sim \mathcal{U}(0, \frac{1}{3})$. Die Dimensionen der horizontalen Anordnung ergeben sich dann aus $V_1 = V_{1a} \cup V_{1b} + \frac{2}{3}$ und $V_2 = V_{2a} + \frac{1}{2} \cup V_{2b} + \frac{1}{2}$. Die diagonale Anordnung erfolgt durch $V_1 = V_{1a} \cup V_{1b} + \frac{2}{3}$ und $V_2 = V_{2a} + \frac{2}{3} \cup V_{2b}$. Für die Anordnung im Dreieck gelten folgende Berechnungen: $V_1 = V_{1a} + \frac{1}{12} \cup V_{1b} + \frac{1}{3} \cup V_{1c} + \frac{7}{12}$ und $V_2 = V_{2a} + \frac{1}{12} \cup V_{2b} + \frac{7}{12} \cup V_{2c} + \frac{1}{12}$. Zuletzt erfolgt die verzahnte Anordnung durch $V_1 = V_{1a} \cup V_{1b} + \frac{5}{12} \cup V_{1c} + \frac{1}{3} \cup V_{1d} + \frac{1}{4}$ und $V_2 = \{V_{2a} + \frac{5}{12} \cup V_{2b} + \frac{2}{3} \cup V_{2c} + \frac{1}{4} \cup V_{2d}\}$.

Der SP von Abb. 3.9a zeigt 2 vertikal zentrierte, horizontal angeordnete Quadrate, im Abstand von einer Clusterbreite. Der PCP ergibt 2 sich überlappende rechtwinklige Trapeze, mit den kurzen Seiten links und den langen Seiten rechts. Die linke Achse zeigt eine Lücke von einer Clusterbreite. Aufgrund der Überlappung erkennt man eine dreiecksförmige Hervorhebung mit der rechten Achse als Grundseite. In Abb. 3.9b sieht man im SP 2 Quadrate, die diagonal absteigend in einem Abstand von einer Clusterdiagonalen angeordnet sind. Die Repräsentation als PCP zeigt eine X-Form mit einer Linienstärke von einer Clusterbreite, die sich wegen der Überlappung beider Cluster bildet. Daraus

ergibt sich jeweils eine Lücke auf den Dimensionsachsen, sowie eine rhombusförmige Hervorhebung des Überlappungsbreichs.

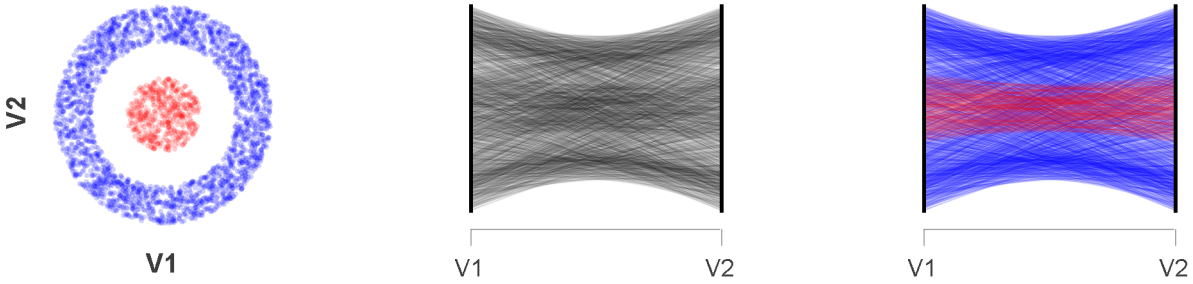
Die Dreieckanordnung zeigt sich im SP von Abb. 3.9c als 2 Quadrate, die unten auf einer Horizontalen liegen und einem Quadrat, das sich oberhalb, horizontal zentriert befindet. Der Abstand zwischen den Quadraten beträgt $\frac{1}{2}$ Clusterbreite. Die Anordnung erzeugt eine Form im PCP, die auf der linken Achse lückenfrei, oben eingeknickt, und unten gerade ist, sowie auf der rechten Achse eine Lücke im Abstand von $\frac{1}{2}$ Clusterbreite aufweist. Die Dichte zeigt eine rhombusförmige Hervorhebung im Überlappungsbereich vom roten und grünen Cluster, sowie eine dreicksförmige Überlappung an den Achsen im Überlappungsbereich des blauen und grünen, sowie des roten und blauen Clusters. In Abb. 3.9d sind die 4 Cluster im SP ringförmig um das Zentrum angeordnet. Der Abstand zwischen den benachbarten Quadraten beträgt $\frac{1}{4}$ Clusterbreite und zwischen den gegenüberliegenden Quadraten eine Clusterbreite. Die Form im PCP ist links und rechts lückenfrei, sowie oben und unten eingeknickt. Rhombusförmige Dichtehäufungen ergeben sich zwischen dem roten und grünen, blauen und orangen, sowie blauen und grünen Cluster. Die dreicksförmigen Häufungen treten bei den roten und orangen, roten und blauen, sowie orangen und grünen Clustern auf.

Klar getrennte zweidimensionale Cluster in den Daten lassen sich im SP klar unterscheiden und damit zur Einzelbetrachtung oder Ausblendung selektieren. Im PCP zeichnen sich die einzelnen Cluster durch Dichtelücken und sowie Dichtehäufungen ab. Dichtehäufungen, die aus der Überlappung von Clustern resultieren, sind rhombusförmig und liegen zwischen den Achsen, wenn die Cluster diagonal absteigend angeordnet sind. Sonst weisen sie eine Dreiecksform auf, bei der eine Seite auf den Achsen liegt. Damit sind die Cluster im PCP trotz visueller Überladung noch erkennbar.

Die Selektion der Cluster kann zum einen erfolgen, wenn sich die Intervalle der Cluster entlang mindestens einer Dimension nicht überschneiden. In diesem Fall kann man mittels schrittweiser Dekomposition der Struktur zuvor nicht selektierbare Cluster erreichen. Der Vorgang ist in Abb. 3.9c nachvollziehbar. Der rote Cluster ist selektierbar und erlaubt nach Ausblendung den Zugriff auf den blauen und grünen Cluster. Sind beide Achsen lückenfrei, ist dies nicht mehr möglich (vergleiche Abb. 3.9d).

In diesem Fall, können zum anderen Dichtelücken zwischen den Achsen zur Selektion benutzt werden. Dazu selektiert man in Abb. 3.9d ausgehend von der Lücke im PCP nach oben oder unten. Der Vorgang trennt die Cluster Rot und Grün von Blau und Orange. Einzeln betrachtet weisen Rot und Grün, bzw. Blau und Orange widerum jeweils eine Lücke an den Achsen auf, womit zuletzt alle 4 Cluster selektierbar sind. Sind im PCP keine Dichte

Die Abhängigkeit der Form des PCP von der konvexen Hülle im SP (siehe Beispiel 5d und 8d) deuten darauf hin, dass mittels PCP keine umschlossenen eindeutigen Strukturen identifiziert werden können. Angelehnt an dem Beispiel aus Abb. 2.11, soll dieses Problem im nächsten Beispiel veranschaulicht werden. Dazu wird eine Ringstruktur generiert, die eine Kreisscheibe mit geringerem Radius umschliesst.



(a) SP, Ring um Kreis (b) PCP, Schwarz/Weiß (c) PCP, Farbe

Abbildung 3.10: Visualisierung von Beispiel 9

Beispiel 9. *Im Datensatz existiert eine Ringstruktur mit Mittelpunkt* $(\frac{1}{2}, \frac{1}{2})$*, Außenradius* $\frac{1}{2}$ *sowie Innenradius* $\frac{1}{3}$ *und eine Kreisscheibe mit gleichen Mittelpunkt und Radius* $\frac{1}{6}$*.*

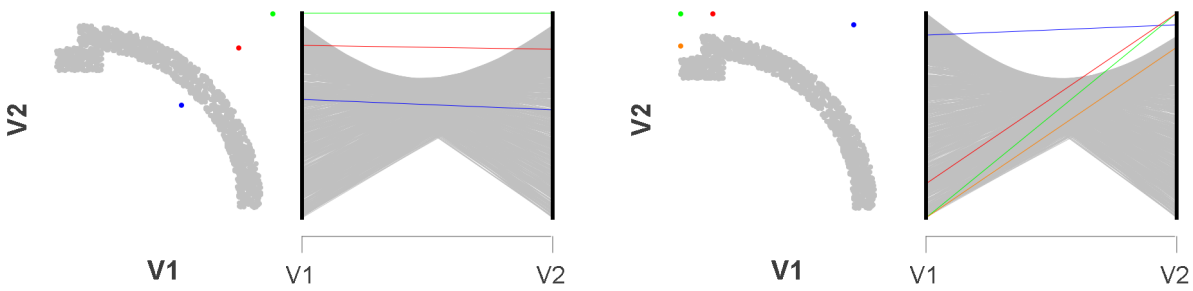
Der Datensatz besteht aus der Ringstruktur a und der Kreisscheibe b: $D = (D_a \cup D_b) + (\frac{1}{2}, \frac{1}{2})$. Die Stichprobe umfasst $n = 10000$ gleichverteilte Werte, wobei $D_a = \{v \in X \sim \mathcal{U}(0,1)^2 | \forall i : \frac{1}{3} \leq \sqrt{v_{i,1}^2 + v_{i,2}^2} \leq \frac{1}{2}\}$ und $D_b = \{v \in X \sim \mathcal{U}(0,1)^2 | \forall i : \sqrt{(}v_{i,1}^2 + v_{i,2}^2) \leq \frac{1}{6}\}$.

In Abb. 3.10a erkennt man die Ringstruktur mit umschlossener Kreisscheibe als SP. Abb. 3.10b zeigt den entsprechenden PCP, dessen Form links und rechts lückenfrei, sowie oben und unten konkav zum Zentrum ist. Die Dichte weist am oberen und unteren Rand eine Häufung auf, deren Ausdehnung der Ringbreite entspricht. Im Zentrum zeigt sich ein Rechteck mit leicht konkaver oberer und unterer Seite, sowie einer Höhe, die dem Kreisdurchmesser entspricht. Der SP lässt eine eindeutige Trennung der Cluster zu, während die Form des PCP durch die Lückenfreiheit auf keinen Cluster hinweist. Die Dichte deutet zwar auf den zweiten Cluster hin, eignet sich aber nicht zur überschneidungsfreien Selektion.

Das letzte Beispiel bezüglich der Untersuchung der zweidimensionalen Visualisierungen bezieht sich auf die Identifizierung von Ausreißer einer Struktur. Diese Abweichung vom Standard können entweder Messfehler oder besonders interessante Ereignisse darstellen, was sie zu einem wichtigen Bestandteil der Datenanalyse macht. Zur Veranschaulichung werden mehrere Ausreisserwerte um eine absteigende gekrümmte Struktur positioniert.

Beispiel 10. *Der Datensatz besteht aus einem Kreisringviertel innerhalb* $[\frac{1}{10}, \frac{9}{10}]^2$ *mit einer Ringbreite von* $\frac{1}{10}$ *und ist mit einem gleichverteilten Bereich in* $[0, \frac{1}{5}] \times [\frac{7}{10}, \frac{4}{5}]$ *verbunden. Um die Struktur sind 7 Ausreisser platziert.*

Der Kreisringsektor a und die Ausreisserwerte V_1 und V_2 vereinen den Datensatz: $D = (D_a + (\frac{1}{10}, \frac{1}{10})) \cup \{V_{1a}, V_{2a}\} \cup \{V_{1b}, V_{2b}\}$. Zur Generierung des Kreisringsektors wird eine Stichprobe von $n = 10000$ gleichverteilten Werten entnommen, sodass $D_a = \{v \in$



(a) SP (b) PCP (c) SP (d) PCP

Abbildung 3.11: Visualisierung von Beispiel 10

$X \sim \mathcal{U}(0,1)^2 | \forall i : \frac{7}{10} \leq \sqrt{v_{i,1}^2 + v_{i,2}^2} \leq \frac{4}{5}\}$. Die Stichprobe des gleichverteilten Bereich umfasst $n_a = 200$ Werte, wobei $V_{1a} \sim \mathcal{U}(0, \frac{1}{5})$ und $V_{2a} \sim \mathcal{U}(\frac{7}{10}, \frac{4}{5})$. Die Ausreisser sind mit den Dimensionen $V_{1b} = 0, 0, \frac{3}{20}, \frac{11}{20}, \frac{4}{5}, \frac{4}{5}, 1$ und $V_{2b} = 1, \frac{3}{20}, 1\frac{11}{20}, \frac{4}{5}, \frac{19}{20}, 1$ definiert.

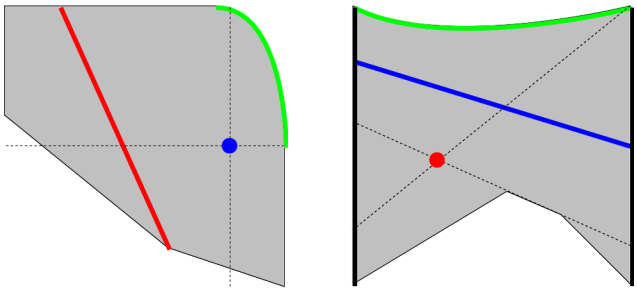
Die Abb. 3.11a und c zeigen, dass im SP alle Ausreisser eindeutig erkennbar sind. Im PCP sind die Schnittpunkte der dualen Linien der Ausreisser entweder beidseitig sichtbar, einseitig sichtbar, segmentweise sichtbar oder nicht sichtbar (siehe Abb. 3.11b und d). Beidseitige Sichtbarkeit haben Ausreisser, die ausserhalb der Intervallgrenzen der Struktur liegen (Abb. 3.11b und d, jeweils grün). Liegt der Ausreisser innerhalb der Intervallgrenzen von einer Dimension, ist er im PCP nur einseitig sichtbar (Abb. 3.11d, blau und rot). Liegt er innerhalb der Intervallgrenzen beider Dimension, aber ausserhalb der konvexen Hülle der Struktur zuzüglich $(0, 1)$ und $(1, 0)$, ist zumindest noch ein Liniensegment sichtbar (Abb. 3.11b, rot). Nicht sichtbar sind Ausreisser, die sich innerhalb der konvexen Hülle der Struktur zuzüglich $(0, 1)$ und $(1, 0)$ befinden (Abb. 3.11b, blau; 3.11d, orange).

3.1.2 Regeln

Aus den im vorigen Abschnitt beschriebenen Beispielen ergeben sich 5 Regeln, die im folgenden formuliert und erläutert werden.

Regel 1. *Eine punktförmige Dichtehäufung im PCP entspricht einer linearen Struktur mit negativem Anstieg im SP. Eine lineare Dichtehäufung im PCP entspricht einer punktförmigen Struktur im SP. Eine kurvige Dichtehäufung im PCP entspricht einer ebenfalls kurvigen Struktur im SP. Dabei ist die Krümmung bezüglich der absteigenden Diagonalen im SP invers zur Krümmung bezüglich der zentralen Horizontalen im PCP.*

Die Regel leitet sich aus der Punkt-Linie-Dualität zwischen SP und PCP ab und zeigt sich zum einen in den Beispielen 1, 2 und 5. Hier zeigen die Abb. 3.1c und 3.1d, sowie 3.2, 3.6a und 3.6c eine oder mehrere punktförmige Dichtehäufungen im PCP und die entsprechenden linearen Strukturen gleicher Anzahl im SP. Die Einschränkung auf



(a) SP (b) PCP

Abbildung 3.12: Schematische Darstellung von Regel 1 bezüglich des Zusammenhangs zwischen Punkt und Linie (blau), Linie und Punkt (rot), sowie Kurve und Kurve (grün) im SP und PCP

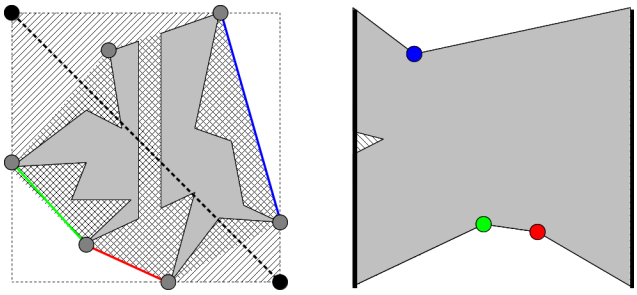
negative Anstiege ergibt sich dabei aus Abb. 3.1a und 3.1c, da die lineare Struktur im SP in diesem Fall keine punktförmige Dichtehäufung im PCP aufweist. Der Übergang vom positiven zum negativen Anstieg liegt bei $m = \pm\infty$ und $m = 0$. Die Abb. 3.6b und 3.6d zeigen, dass in diesen Spezialfällen die punktförmigen Dichtehäufungen im PCP auf den Dimensionsachsen liegen. Die Erhaltung kurviger Strukturen bei inverser Krümmung ist in Abb. 3.6c und Abb. 3.6d zu sehen.

Zum anderen ist aus aus den Abb. 3.8 und 3.9 der Beispiele 7 und 8 ersichtlich, dass die linearen Dichtehäufungen im PCP dem Cluster (quadratischer Punkt) im SP entsprechen. Überlagern sich lineare Strukturen im PCP, ergibt sich ein Überlappungsbreich höherer Dichte, welcher widerum auf einen linearen Zusammenhang mit negativem Anstieg hinweist. In Abb. 3.9d zeigt sich dies durch die diagonal absteigende Anordnung der roten und grünen, sowie blauen und orangenen Cluster. Zuletzt ergibt sich aus Beispiel 3 und Beispiel 4, dass mit zunehmendem Übergang von einer linearen zu einer punktförmigen Struktur im SP, ein Übergang von einer punkförmiger zu einer linearen Dichtehäufung im PCP erfolgt (siehe Abb. 3.3 und 3.4).

Die Regel ist in 3.12 schematisch dargestellt. Der blaue Punkt im SP entspricht der blauen Linie im PCP. Hierbei sind die Koordinaten des Punktes die Schnittpunkte der Linie mit den Dimensionsachsen. Der rote Punkt im PCP entspricht der roten Geraden im SP. Die Schnittpunkte der Seiten eines in diesem Punkt verschränkten Trapez' mit den Dimensionsachsen sind die Koordinaten der Start- und Endpunktes der roten Gerade. Die grüne Kurve im SP ist konvex zur absteigenden Diagonalen und entspricht der grünen Kurve im PCP, dessen Krümmung konkav zur zentralen Horizontalen ist.

Regel 2. *Die Eckpunkte der konvexen Hülle der Struktur im PCP entsprechen den Kanten der konvexen Hülle der Vereinigung der Struktur im SP mit der absteigenden Diagonalen ihres minimal umgebenden Rechtecks. Ebenso entsprechen die Kanten der konvexen Hülle im PCP den Eckpunkten der konvexen Hülle im SP.*

Diese Regel ergibt sich zunächst aus Beispiel 5. Die Abbildungen 3.6a, 3.6b und 3.6c



(a) SP (b) PCP

Abbildung 3.13: Schematische Darstellung von Regel 2 bezüglich des Zusammenhangs zwischen den konvexen Hüllen der Stukturen im SP und PCP

zeigen, dass mit jeder Kante im SP, die nicht auf dem minimal umgebenden Rechteck (MUR) liegt, ein Eckpunkt im PCP hinzukommt. Eine Ausnahme ist Abb. 3.6d, bei der die Kanten der kurvigen Struktur im SP keinen Einfluss auf die Form des PCP haben. Diese gleicht der Form des PCP von Abb. 3.6b, deren Kante im SP genau der nicht sichtbaren Kante der konvexen Hülle der Struktur im SP von Abb. 3.6d entspricht.

Verdeutlicht wird dies in Beispiel 9. Die rote Kreisscheibe liegt innerhalb der konvexen Hülle der gesamten Struktur von Abb. 3.10a und hat keinen Einfluss auf die konvexe Hülle des PCP (siehe Abb. 3.10c). Weiterhin zeigt Beispiel 10 in Abb. 3.11a und 3.11b, dass der blaue Ausreißer, der im SP außerhalb der grauen Struktur, aber innerhalb ihrer konvexer Hülle liegt, als blaue Linie im PCP innerhalb der Struktur verschwindet. In Abb. 3.11c und 3.11d liegt der orangene Ausreißer außerhalb der konvexen Hülle der grauen Struktur, dennoch verschwindet er als orangene Linie im PCP innerhalb der Struktur. Daher gilt die Regel nur nach Vereinigung der Struktur im SP mit der absteigenden Diagonalen des MUR.

Abb. 3.13 stellt die Regel schematisch dar. Der SP zeigt eine Struktur mit ihrer konvexen Hülle (gekreuzte Schraffur, graue Eckpunkte) und der konvexen Hülle nach Vereinigung mit der absteigenden Diagonalen des MUR (parallele Schraffur, graue und schwarze Eckpunkte). Die Kanten der konvexen Hülle, die nicht auf dem MUR liegen, sind farblich markiert. Der PCP zeigt die duale Struktur mit ihrer konvexen Hülle (parallele Struktur) und die dualen Eckpunkte der Kanten des SP entsprechend ihrer farblichen Markierung.

Regel 3. *Die Form und Dichte eines SP oder PCP, weist auf die Verteilung der Daten hin. Dabei können anhand von jeweils 6 Schemata für SP und PCP die Kombinationen* $\{\mathcal{U}, \mathcal{N}, Exp\}^2$ *unterschieden werden.*

Die Schemata ergeben sich aus den Plots der Beispiele 3, 4 und 6. Die Beschreibung der Form und der Dichte beschränkt sich auf ein geometrisches Primitiv bzw. eine Art des Helligkeitsverlaufs. Dabei sind Bereiche mit geringer Dichte heller als Bereiche mit hoher Dichte. Das Seitenverhältnis der Schemata beträgt 1:1, was den auf $[0, 1]^2$ normierten,

untersuchten Daten entspricht. Zunächst werden die Schemata mit gleicher Verteilung der beiden Dimensionen und anschließend die Schemata mit unterschiedlicher Verteilung vorgestellt. Hierbei erfolgt die Zuweisung der Achsen nach $X \leftrightarrow Y$, sodass die Dimension vor dem Pfeil der X-Achse im SP und der linken Achse im PCP entspricht.

$\mathcal{U} \leftrightarrow \mathcal{U}$ Der SP hat die Form eines Quadrats mit gleichmäßiger Helligkeit. Der PCP ist ebenfalls ein Quadrat mit gleichmäßiger Helligkeit, auf dem sich allerdings eine leicht dunklere x-förmige Struktur entlang der Diagonalen abzeichnet. Das Schema ist in Abb. 3.14a dargestellt.

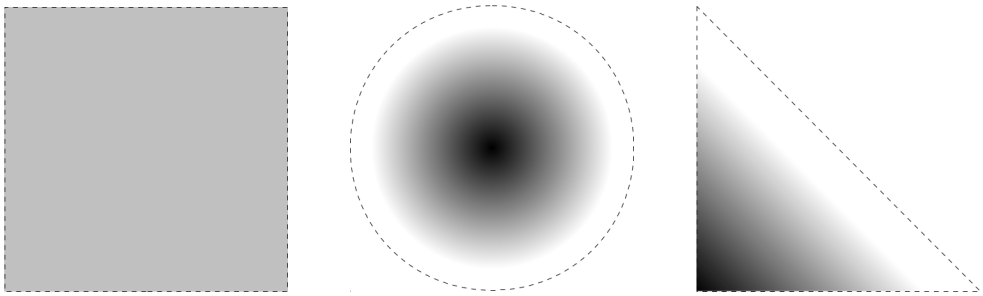
$\mathcal{N} \leftrightarrow \mathcal{N}$ Der SP ist kreisförmig mit einer radial zum Zentrum zunehmenden Helligkeit. Die Form des PCP ist ein oben und unten gleichermaßen mit konkaver Krümmung rund geschnittenes Quadrat. Die Helligkeit nimmt linear und parallel zur zentralen Horizontalen zu. Das Schema ist in Abb. 3.14b dargestellt.

Exp $\leftrightarrow$ **Exp** Der SP hat die Form eines rechtwinkligen, gleichseitigen Dreiecks, dessen rechter Winkel im Nullpunkt liegt. Die Helligkeit nimmt ausgehend von und parallel zur Hypothenuse linear in Richtung des Nullpunkts zu. Der PCP ist ein oben mit konkaver Krümmung geschnittenes Quadrat, bei dem die Helligkeit linear und parallel zur Horizontalen nach unten zunimmt. Das Schema ist in Abb. 3.14c dargestellt.

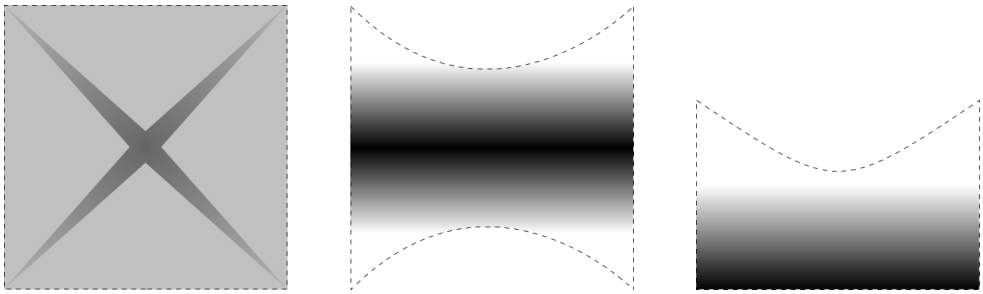
$\mathcal{U} \leftrightarrow \mathcal{N}$ Der SP ist ein Quadrat, dessen Helligkeit linear und parallel zur $\mathcal{U}$-Achse zum Zentrum hin zunimmt. Der PCP ist ein oben und unten gleichermaßen mit konkaver Krümmung rund geschnittenes Quadrat. Die Krümmung der Kurve nimmt in Richtung $\mathcal{N}$-Achse zu. Ebenso verhält es sich mit der Helligkeit, die linear, parallel und in Richtung der $\mathcal{N}$-Achse zunimmt. Das Schema ist in Abb. 3.15a dargestellt.

$\mathcal{U} \leftrightarrow$ **Exp** Der SP hat die Form eines Quadrats, dessen Helligkeit linear sowie parallel zur $\mathcal{U}$-Achse zum Nullpunkt hin zunimmt. Der PCP ist ein oben mit konkaver Krümmung rund geschnittenes Quadrat, bei der die Krümmung Richtung Exp-Achse zunimmt. Die Helligkeit nimmt linear und parallel zur und in Richtung der Exp-Achse zu. Das Schema ist in 3.15b dargestellt.

$\mathcal{N} \leftrightarrow$ **Exp** Der SP ist ein gleichschenkliges, spitzwinkliges Dreieck, dessen Grundseite auf der $\mathcal{N}$-Achse liegt. Die Helligkeit nimmt elliptisch zum Mittelpunkt der $\mathcal{N}$-Achse und zum Nullpunkt der Exp-Achse zu. Die Hauptachse der Ellipse ist parallel zur Exp-Achse. Der PCP ist ein oben stark und unten schwach mit konkaver Krümmung rund geschnittenes Quadrat. Die Krümmung nimmt in Richtung Exp-Achse aufer oberen Seite zu und aufer unteren Seite ab. Die Helligkeit steigt kaum wahrnehmbar, parallel zur und in Richtung der Exp-Achse und bildet damit einen linearen, dunklen Bereich vom Zentrum der $\mathcal{N}$-Achse zum Nullpunkt der Exp-Achse. Das Schema ist in 3.15c dargestellt.

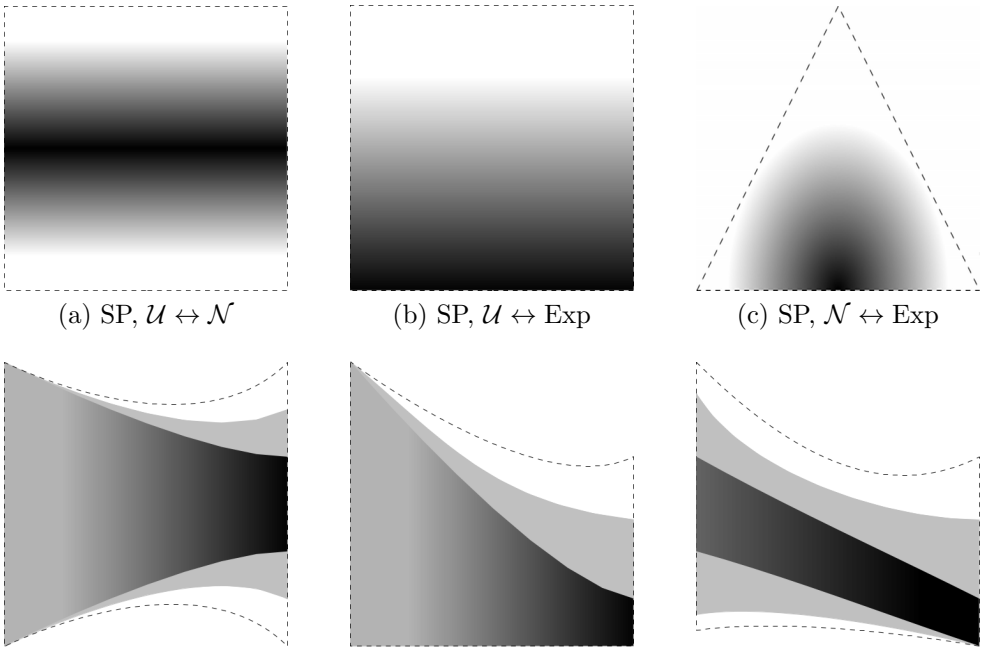


(a) SP, $\mathcal{U} \leftrightarrow \mathcal{U}$ (b) SP, $\mathcal{N} \leftrightarrow \mathcal{N}$ (c) SP, Exp $\leftrightarrow$ Exp



(d) PCP (e) PCP (f) PCP

Abbildung 3.14: Schematische Darstellung von Regel 3 bezüglich der Abhängigkeit der Form und Dichte im SP (oben) und PCP (unten) von der Art der Verteilung der Daten



(a) SP, $\mathcal{U} \leftrightarrow \mathcal{N}$ (b) SP, $\mathcal{U} \leftrightarrow$ Exp (c) SP, $\mathcal{N} \leftrightarrow$ Exp

(d) PCP (e) PCP (f) PCP

Abbildung 3.15: Schematische Darstellung von Regel 3 bezüglich der Abhängigkeit der Form und Dichte im SP (obere Reihe) und PCP (untere Reihe) von der Art der Verteilung der Daten

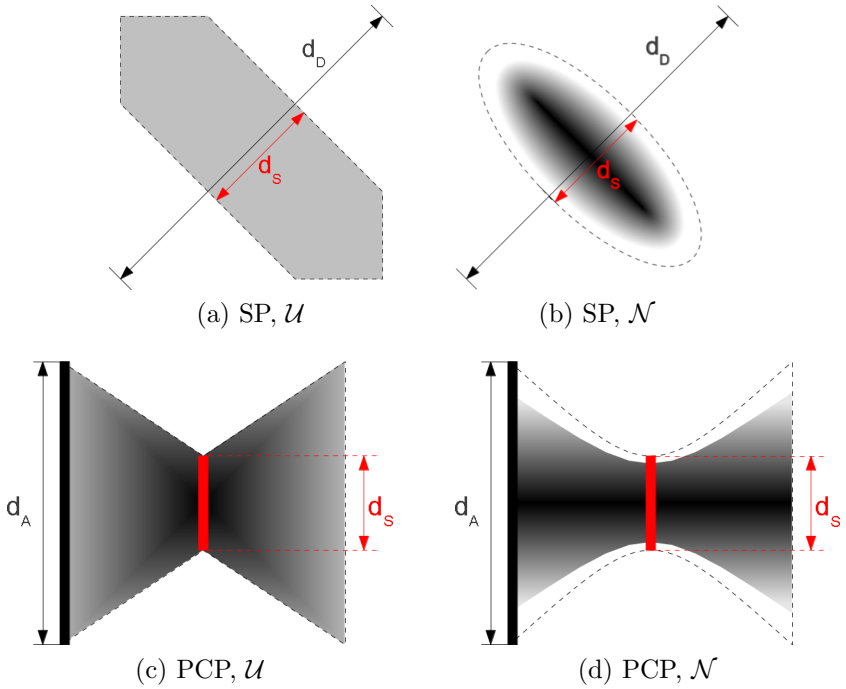
Regel 4. *Der Korrelationskoeffizient einer symmetrischen Struktur kann im SP aus ihrer Breite entlang der aufsteigenden Diagonalen und im PCP aus der Breite entlang der Vertikalen im Zentrum geschätzt werden.*

Die Regel ist das Ergebnis der Untersuchung aus den Beispielen 3 und 4. Hier zeigen die Abb. 3.3 und Abb. 3.4, dass die Verbreiterung der linearen Strukturen den Absolutwert des Korrelationskoeffizientens zwischen den Dimensionen senkt. Genauer zeigt sich nach Abb. 3.5 die Abhängigkeit des Koeffizienten von dem Anstieg und der Rauschausprägung der Struktur, wobei der Anstieg hier einen geringeren Einfluss als das Rauschen hat. Ob die Daten gleich- oder normalverteilt vorliegen hat wiederum keinen wahrnehmbaren Einfluss. Aufgrund der in den Rahmenbedingungen vereinbarten Skalierung der einzelnen Dimensionen auf $[0, 1]$ erhält man stets einen Anstieg von $0, \pm 1$ oder $\pm\infty$ und die Strukturbreite variiert abhängig vom vorhandenen Rauschen und ursprünglichen Anstieg.

In der Abb. 3.17 sieht man den Verlauf des Korrelationskoeffizienten ϱ in Abhängigkeit des Verhältnis $d = \frac{d_S}{d_D}$ der Länge d_S des Schnitts der aufsteigenden Diagonalen mit einer symmetrischen Struktur im SP zur Länge d_D der aufsteigenden Diagonalen des MUR (im skalierten Fall ist $d_D = \sqrt{2}$). Das Verhältnis ist äquivalent zum Verhältnis der Länge d_S des Schnitts der zentralen Vertikalen durch die Struktur im PCP zur Länge d_A der Dimensionsachsen. 3.17. Der Plot ergibt sich aus der Generierung von $n = 10000$ gleichverteilten Werten, die mit 65 Werten für $0 \leq d \leq 1$ modelliert werden. Damit lässt sich der Korrelationskoeffizient in Abhängigkeit von der relativen Strukturbreite linear annähernd durch folgende Funktion bestimmen:

$$\varrho^*(d) \approx \begin{cases} \frac{1}{2}d - 1, & [0, \frac{1}{3}) \\ -2d - \frac{3}{2}, & [\frac{1}{3}, \frac{2}{3}) \\ \frac{1}{2}d - \frac{1}{2}, & [\frac{2}{3}, 1] \end{cases}$$

Diese Funktion kann bei normalverteilten Daten nur eingeschränkt verwendet werden, da dort die Strukturbreite ab einem Verhältnis von $\approx \frac{1}{2}$ zur Diagonalen nicht direkt proportional mit der Rauschausprägung skaliert. Dieser Umstand ist in Abb. 3.18 erkennbar. Dementsprechend sollte die Schätzung von normalverteilten Strukturen nur im Bereich von 0 bis 50% relativer Strukturbreite erfolgen. Die Abb. 3.16 zeigt schematisch, wie die einzelnen Längen im SP und PCP abgemessen werden, um das Verhältnis d zur Schätzung des Korrelationskoeffizienten zu gewinnen. Achtet man bei der Visualisierung auf ein Seitenverhältnis von 1:1, kann man im SP auch die horizontalen bzw. vertikalen anstatt der diagonalen Abstände ins Verhältnis setzen.



(a) SP, $\mathcal{U}$ (b) SP, $\mathcal{N}$

(c) PCP, $\mathcal{U}$ (d) PCP, $\mathcal{N}$

Abbildung 3.16: Schematische Darstellung der Ermittlung von der relativen Strukturbreite d zur Schätzung von ϱ im SP und PCP nach Regel 4

$$d = \frac{d_S}{d_D} = \frac{\sqrt{(\Delta_x^2 + \Delta_y^2)}}{\sqrt{(l_x^2 + l_y^2)}}$$

$$\Delta x = \Delta y \text{ und } l_x = l_y \qquad \text{wegen der Skalierung auf [0,1]}$$

$$d = \frac{\sqrt{(2\Delta_x^2)}}{\sqrt{(2l_x^2)}} = \frac{\sqrt{(2)}\Delta_x}{\sqrt{(2)}l_x} = \frac{\Delta_x}{l_x}$$

Regel 5. *Ein punktförmiger Cluster im SP, entspricht einer linearen Dichtehäufung im PCP. Sind zwei Cluster diagonal absteigend angeordnet überlagern sich ihre Linienbündel auf oder zwischen den Dimensionsachsen. Das Vorhandensein mehrerer Cluster unterteilt den PCP in überlagerte und überlagerungsfreie Bereiche, wodurch dieser wiederum in seine einzelnen Strukturen zerlegt werden kann.*

Die Regel leitet sich aus Beispiel 8 ab. Abb. 3.9 zeigt im PCP die linearen Dichtehäufungen und im SP die entsprechenden quadratischen Cluster. Im PCP lassen sich drei Bereiche definieren: Hintergrund, Cluster und rhombusförmige Überlagerung. Es ergeben sich drei Kombinationen von Anordnungen dieser Bereich entlang einer beliebigen Gerade im PCP, die zur systematischen Dekomposition der Struktur verwendet werden

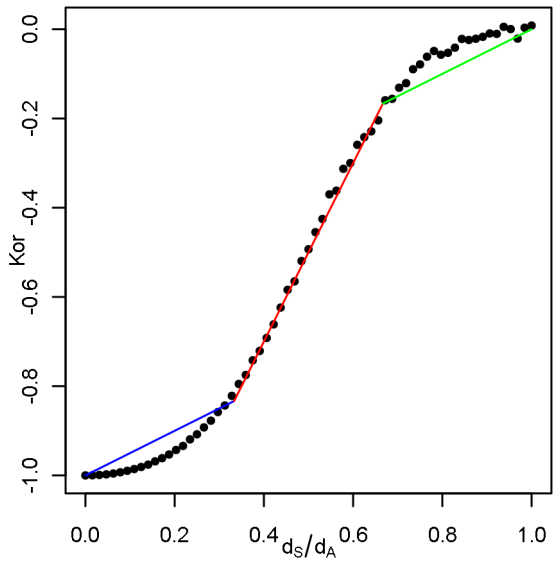


Abbildung 3.17: Plot des Korrelationskoeffizienten in Abhängigkeit von der relativen Strukturbreite d mit Annäherung durch lineare Funktionen (blau, rot, grün)

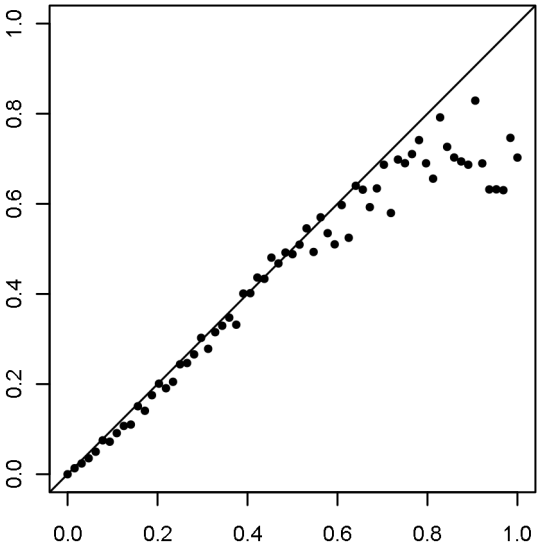


Abbildung 3.18: Plot der relativen Strukturbreite in Abhängigkeit der Ausprägung normalverteilten Rauschens

können. Dabei wird angenommen, dass sich die Cluster mindestens in einer Dimension nicht überlagern.

Schneidet die Gerade den zunächst den Hintergrund, damm einen Clusterbereich und wieder den Hintergrund (HCH-Kombination), selektiert sie den gesamten Cluster und keinen weiteren. Der Cluster wird gekennzeichnet und zur weiteren Exploration ausgeblendet. Schneidet die Gerade einen Cluster, einen Überlagerungsbereich und wieder den gleichen Cluster (CUC-Kombination), hat man einen ganzen Cluster und einen Teilcluster selektiert. Da sich die Cluster in mindestens einer Dimension nicht überlagern, muss hier oder nach Invertieren einer Dimensionsachse eine Position der Geraden existieren, die nur den Cluster selektiert und somit vom Teilcluster trennt. Schneidet die Gerade den Hintergrund, eine Überlagerung und wieder den Hintergrund (HUH-Kombination), hat man eine Clustergruppe selektiert. Eine Gruppe wird separat weiter zerlegt, bis all ihre einzelnen Cluster gekennzeichnet sind.

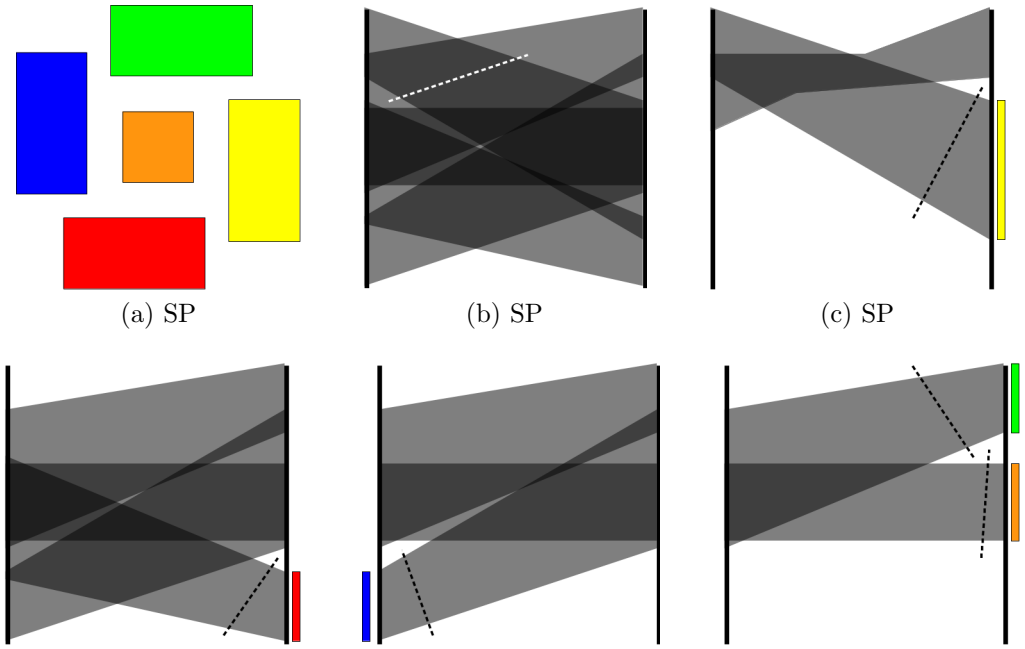
Der Vorgang ist in Abb. 3.19 für die Struktur aus 3.19a schematisch veranschaulicht. Es findet sich in Abb. 3.19b keine Möglichkeit für die HCH-Kombination und keine sinnvolle HUH-Kombination, weshalb zunächst eine Gerade mit CUC-Kombination gesucht wird. Abb. 3.19c zeigt die Einzelbetrachtung der selektierten Daten, bei der die Trennung des Clusters vom Teilcluster und somit die Kennzeichnung mit Gelb möglich ist. Der gelbe Cluster wird für die weitere Exploration ausgeblendet. In den Abb. 3.19d bis 3.19f könnnen jeweils die HCH-Kombinationen gefunden und die Cluster somit einzeln gekennzeichnet und ausgeblendet werden.

3.2 Multivariate Visualisierungen - Radviz und multidimensionale Skalierung

Multivariate Viusalisierungen haben den Vorteil, dass man zur Erkennung von univariater oder bivariater Merkmale von Strukturen eines hochdimensionalen Datensatzes theoretisch nur ein Plot betrachten brauch. Mit bivariaten Visualisierungen würde man die Plots aller Kombinationen der einzelnen Dimensionspaare benötigen. Dies motiviert den Ansatz für die Untersuchung von Radviz und MDS zunächst solche Datensätze zu generieren, die über die gleichen Merkmale, wie die Beispiele aus Abschnitt 3.1.1. Dazu werden die bereits vorhanden Datensätze verwendet und die restlichen Dimensionen mit Zufallsdaten, festen Werten oder Kopien (auch invertiert) der Ursprungsdaten gefüllt. In den Plots der Visualisierungen sind die generierten Daten mit einem kleine Anfangsbuchstaben gekennzeichnet. Dabei steht „g“ für gleichverteilte, „n“ für normalverteilte, „e“ für exponentialverteilte, „c“ für kopierte und „i“ für invertierte Daten. Die weiteren Rahmenbedingungen des vorigen Abschnitts bleiben weiterhin gültig.

3.2.1 Untersuchung

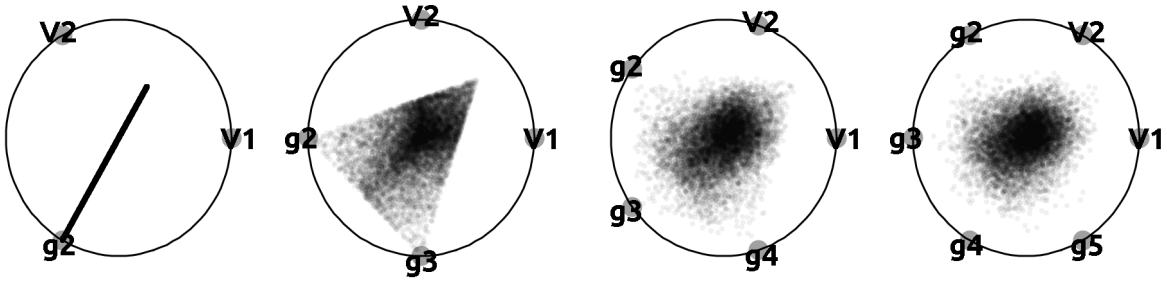
Als erstes wird das Beispiel 1 verwendet und um gleichverteilte Dimensionen erweitert.



(a) SP (b) SP (c) SP

(d) PCP (e) PCP (f) PCP

Abbildung 3.19: Schematische Darstellung von Regel 5 bezüglich der Zerlegung eines PCP in einzelne Strukturen



(a) RVP (b) RVP (c) RVP (d) RVP

Abbildung 3.20: Visualisierung des Beispiels 11

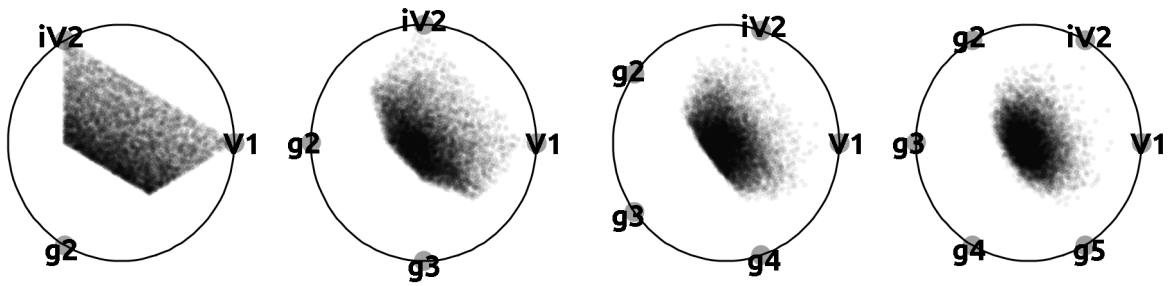
Beispiel 11. *Der Datensatz besteht aus gleichverteilten Werten, wobei* $D = \{p_{i,j} \in [0,1]^d \mid \forall i : p_{i,2} = p_{i,1}\}$ *für* $d \in \{2, \ldots, 8\}$.

Dazu werden V_1, V_2 aus `pcp/cor10k.csv` geladen. Dieser Datensatz entspricht dem Datensatz des Beispiels 1 mit der Ausnahme, dass er $n = 10000$ Werte umfasst. D.h., $\varrho(V_1, V_2) = 1$ Dementsprechend wird für die restlichen $d - 2$ Dimensionen jeweils eine Stichprobe von $n = 10000$ Werten der Gleichverteilung entnommen, sodass $V_{2\ldots d} \sim \mathcal{U}(0,1)$.

Für die Untersuchung der Visualisierung werden V_1 und V_2 adjazent zueinander platziert. Man erkennt im 3-dimensionalen Fall eine Gerade mit einem Endpunkt in g_2 und dem anderen Endpunkt auf dem Mittelpunkt der Geraden von V_1 zu V_2 (siehe Abb. 3.20a). Die Abb. 3.20b zeigt im 4-dimensionalen Fall ein Dreieck mit den 2 Eckpunkten in g_2 und g_3 und den dritten Eckpunkt wieder zwischen V_1 und V_2. Die Dichte des Dreiecks nimmt zum Radviz-Mittelpunkt zu. Im 5-dimensionalen Fall (siehe Abb. 3.20c) hat die Struktur eine viereckige Form. 3 Eckpunkte liegen wiederum auf den gleichverteilten Dimensionen g_2, g_3 und g_4 und der 4. zwischen V_1 und V_2. Die Dichte nimmt stärker zum RVP-Mittelpunkt zu. Zuletzt zeigt sich im 6-dimensionalen Fall ein Fünfeck, dessen Form durch die schwache Dichte ausserhalb des Zentrums schwer zu erkennen ist. Ein Eckpunkt liegt zwischen V_1 und V_2, die anderen Eckpunkte liegen auf den Dimensionen $g_2 \ldots g_5$(siehe Abb. 3.20d).

Daraus lässt sich ableiten, dass aus zwei miteinander positiv korrelierende Dimensionen bei n-Dimensionen ein n-1-eckiges Polygon mit spitzwinkligen Eckpunkt auf dem Mittelpunkt der Gerade zwischen den Dimensionen abgebildet wird. Die restlichen Eckpunkte liegen auf den Ankerpunkten, der nicht korrelierenden Dimensionen. Dabei ist die Dichte stets zentral und der spitzeste Winkel der Form zeigt in Richtung der korrelierenden Dimensionen. Da die Dichte stets zentral liegt, enthält sie keine Information und man kan zur besseren Formerkennung ohne Transparenz plotten. Mit zunehmender Dimensionalität ordnen sich die Werte der Struktur zentral an.

Das nächste Beispiel zeigt das Verhalten bei 2 negativ zueinander korrelierenden Dimensionen.



(a) RVP (b) RVP (c) RVP (d) RVP

Abbildung 3.21: Visualisierung des Beispiels 12

Beispiel 12. *Der Datensatz besteht aus gleichverteilten Werten, wobei* $D = \{p_{i,j} \in [0,1]^d \mid \forall i : p_{i,2} = 1 - p_{i,1}\}$ *für* $d \in \{2, \ldots, 8\}$.

Es entspricht somit Beispiel 11, außer dass gilt: $iV_2 = 1 - V_2$. Die Ankerpunkte von V_1 und iV_2 sind wieder benachbart. Im 3-dimensionalen Fall ergibt sich ein gleichschenkliges Trapez mit 2 Eckpunkten an den korrelierenden Dimensionen. Die Dichte nimmt linear zum Zentrum und parallel zur Gerade zwischen V_1 und iV_2 (siehe Abb. 3.21a). Die Struktur ist im 4-dimensionalen Fall 6-eckig und 2 Eckpunkte befinden sich wieder auf den korrelierenden Dimensionen. Die Dichte nimmt wieder parallel zur Gerade der korrelierenden Dimensionen zum Zentrum hin zu. Die Abb. 3.21c und 3.21d zeigen, dass die Form mit zunehmender Dimension ellipsenförmig.

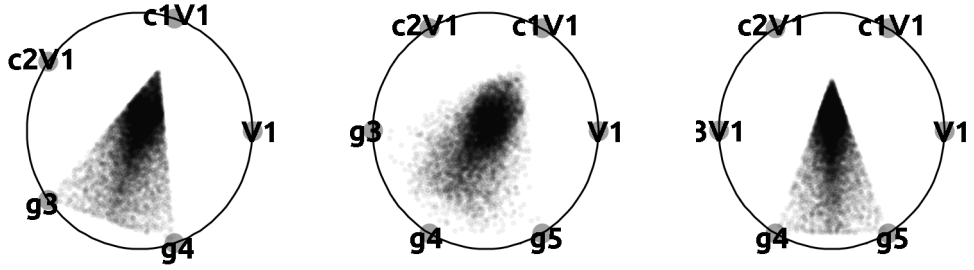
Das Beispiel zeigt, dass 2 negative korrelierende Dimensionen eine symmetrische Struktur erzeugt, mit zum Zentrum zunehmender Dichte. Die Symmetrieachse liegt orthoganal zur Geraden der korrelierenden Dimensionen und schneidet deren Mittelpunkt. Damit ist auch die Hauptachse der Struktur stets parallel zu der Gerade zwischen V_1 und $iV2$.

Das folgeden Beispiel zeigt ds Verhalten bei einer variablen Anzahl miteinander korrelierender Dimensionen.

Beispiel 13. *Der Datensatz besteht aus gleichverteilten Werten, wobei* $D = \{p_{i,j} \in [0,1]^d \mid \forall i : p_{i,2\ldots c} = m p_{i,1}\}$ *für* $d \in \{2, \ldots, 6\}$, $m \in \{-1, 1\}$ *und* $c \in \{2, 3, 4\}$.

Hier entstammt V_1 aus `pcp/cor10k.csv` und wird um c Dimensionen ergänzt, sodass $V_{2\ldots c} = mV_1$. Die restlichen Dimensionen entsprechen einer Stichprobe von $n = 10000$ gleichverteilten Werten, sodass $V_{c+1\ldots d} \sim \mathcal{U}(0,1)$.

Die Abb. 3.22a und 3.22b zeigen die Plots für $m = 1$ und $c = 3$. Im 5-dimensionalen entsprechen Form und Dichte der Struktur dem 4-dimensionalen Fall aus Beispiel 11. Allerdings liegt hier die Spitze im Mittelpunkt des durch die korrelierende Dimensionsankerpunkte aufgespannten Dreiecks. Der 6-dimensionale Fall entspricht in Form und Dichte dem 5-dimensionalen Fall aus Beispiel 11. Auch hier befindet sich die Spitze im Dreieck der korrelierenden Dimensionen.



(a) RVP (b) RVP (c) RVP

Abbildung 3.22: Visualisierung des Beispiels 13

Für $m = 1$,$c = 4$ zeigt die Abb. 3.22c im 6-dimensionael Fall, dass die Form und Dichte, der des 4-dimensionalen Falles aus Beispiel 11 entspricht. Die Spitze liegt jetzt im Mittelpunkt des durch die korrelierendem Dimensionen aufgespannten Trapez.

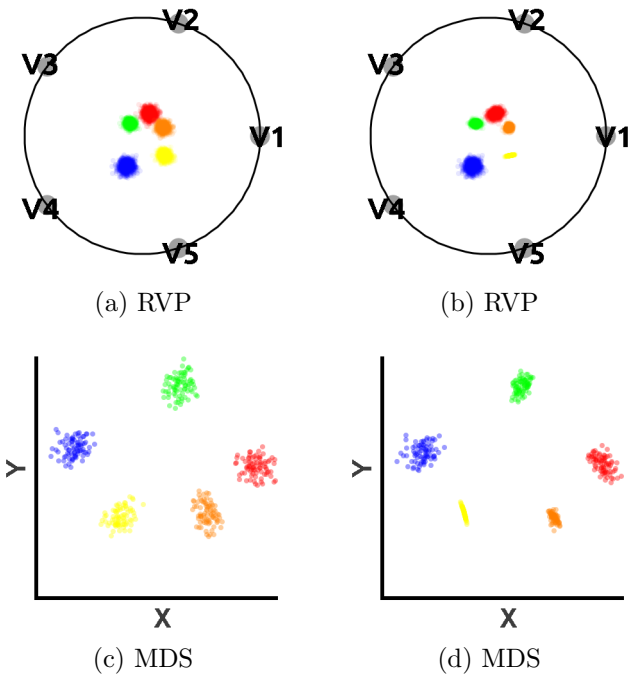
Damit wird deutlich, dass im n-dimensionalen Fall bei c korrelierenden Dimensionen mit adjazenten Ankerpunkten eine $n - c + 1$ eckige Form entsteht. Der Eckpunkt mit dem spitzesten Winkel liegt dabei stets auf dem Mittelpunkt der konvexen Hülle der korrlierenden Dimensionen.

Das nächste Beispiel stellt den mögliche Einflüsse auf die Darstellung von Clusterbildungen in Radviz und MDS gegenüber.

Beispiel 14. *Der Datensatz hat 5 Dimensionen und besteht aus 5 Clustern. Die Cluster sind bezüglich aller Dimensionen normalverteilt und haben definierte Positionen. Es erfolgt eine Variation der Cluster bezüglich ihrer Größe, Position, sowie Anzahl der Dimension, die Einfluss auf deren Ausmaße haben.*

Der Stichprobenumfang beträgt $n = 1000$. Es existieren 5 normalverteilte Dimensionen $V_{1...5} \sim \mathcal{N}(0, 1)$, eine Positionsmatrix mit 5 Koordinaten für 5 Dimensionen $pos_{1...5,1...5} \sim \mathcal{U}(-7, 7)$, 5 normalverteilte Strukturen unterschiedlicher Dimensionalität $D_1 = \{V_1, V_2, V_3, V_4, V_5\}$ $D_2 = \{V_1, V_2, V_3, V_4, 0\}$ $D_3 = \{V_1, V_2, V_3, 0, 0\}$ $D_4 = \{V_1, V_2, 0, 0, 0\}$ $D_5 = \{V_1, 0, 0, 0, 0\}$, 2 Faktormatrizen $f_2 = (1, 2, 3, 4, 5)$ und $f_1 = \frac{f_2}{5}$. Dabei werden aus $D_a = \{\{D_1\}, \{D_1\}, \{D_1\}, \{D_1\}, \{D_1\}\}$ und $D_b = \{\{D_1\}, \{D_2\}, \{D_3\}, \{D_4\}, \{D_5\}\}$ folgenden Datensätze konstuiert:

- `rv/spheres.csv` mit $D_a + pos$
- `rv/spheres_vardim.csv` mit $D_b + pos$
- `rv/spheres_varsca1.csv` mit $f_1 D_a + pos$
- `rv/spheres_varsca2.csv` mit $f_2 D_a + pos$
- `rv/spheres_varpossca1.csv` mit $D_a + f_1 pos$



(a) RVP (b) RVP

(c) MDS (d) MDS

Abbildung 3.23: Visualisierung des Beispiels 14

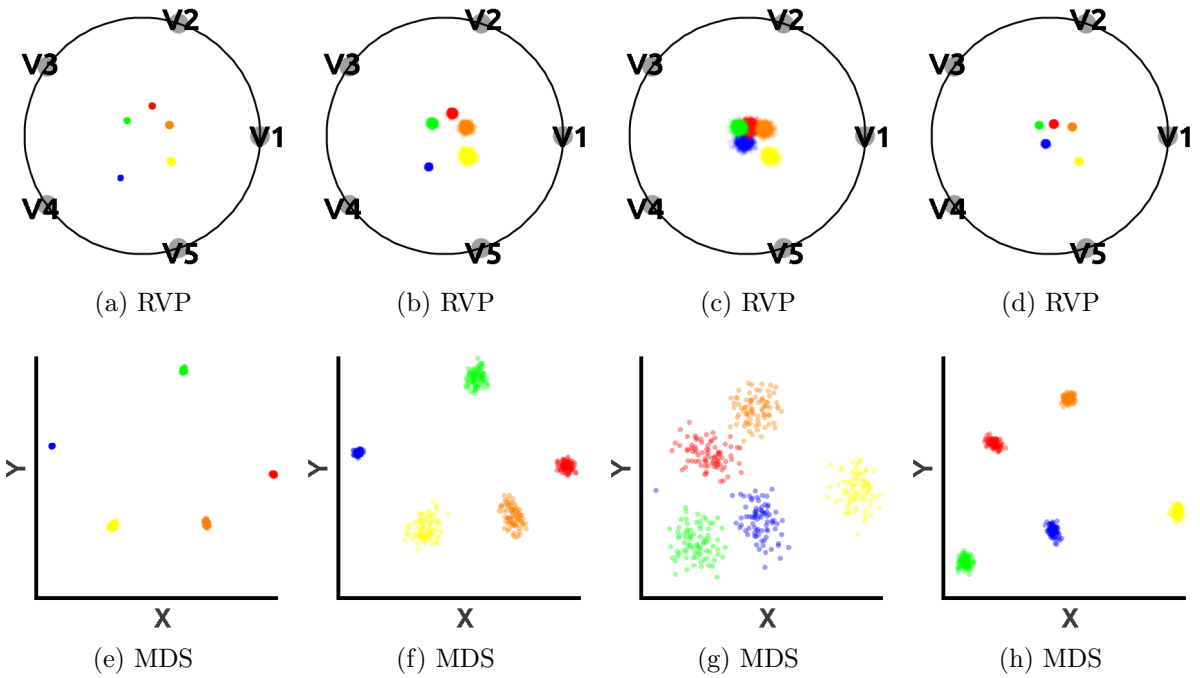
- `rv/spheres_varpossca1.csv` mit $D_a + f_2 pos$

Die Abb. 3.23a und 3.23c zeigen die ausgängliche Positionen, Skalierungen und Dimensionalitäten der Cluster, wobei beide Visualisierungen eine ähnliche Clusteranordnung aufweisen. In Abb. 3.23b und 3.23d erkennt man, dass die Positionen bei veränderter Dimensionalität (der gelbe Cluster bestht bspw. nur noch aus 1 Dimension) gleich bleiben. Dafür ändert sich dieStrukturen in ähnlicher Weise. Der gelbe Cluster hat dieselbe lineare Form.

Die Abb. 3.24a und 3.24e, sowie 3.24b und 3.24f zeigen dass die Auswirkung der Skalierung der Clustergrößen im MDS keine Auswirkungen auf deren Positionen hat. Im RVP bewirkt die Vergrößerung der Cluster eine zentrumsnähere Platzierung der Abbilder. Die Skalierung des Abstandes vom Nullpunkt bewirken bei beiden Visualisierungen eine Positionsänderung gegenüber der anderen Plots (vergleiche Abb. 3.24c,d,g,h mit 3.23a und 3.23c). Dabei bewirkt eine Annäherung an den Nullpunkt eine Vergrößerung der abgebildeten Strukturen (siehe Abb. 3.24c und 3.24g) und die Entfernung von Nullpunkt eine Verkleinerung der Strukturen (3.24d und 3.24h).

3.2.2 Regeln

Regel 6. *Sind die Ankerpunkte von korrelierenden Dimensionen in der Radviz-Visualisierung adjazent zueinander und weisen die restlichen Dimensionen keine ausgeprägte*



(a) RVP (b) RVP (c) RVP (d) RVP

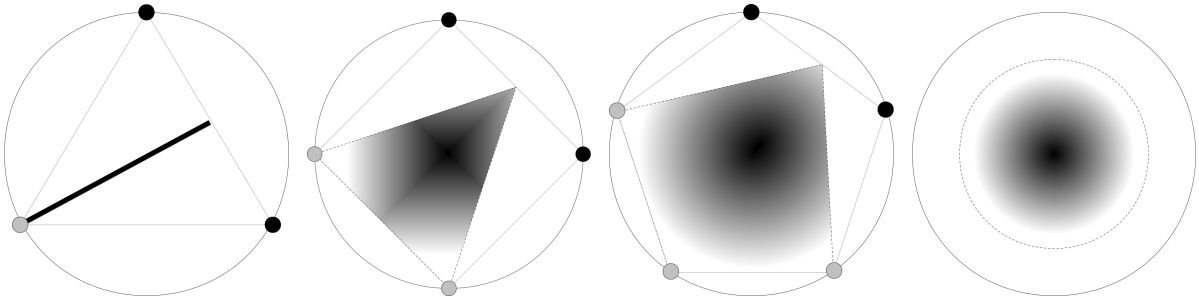
(e) MDS (f) MDS (g) MDS (h) MDS

Abbildung 3.24: Visualisierung des Beispiels 14

Korrelation untereinander und zu den korrelierenden Dimensionen auf, entsteht eine Struktur, deren spitzester Winkel auf den Mittelpunkt der konvexe Hülle der Ankerpunkte der korrelierenden Dimensionen zeigt.

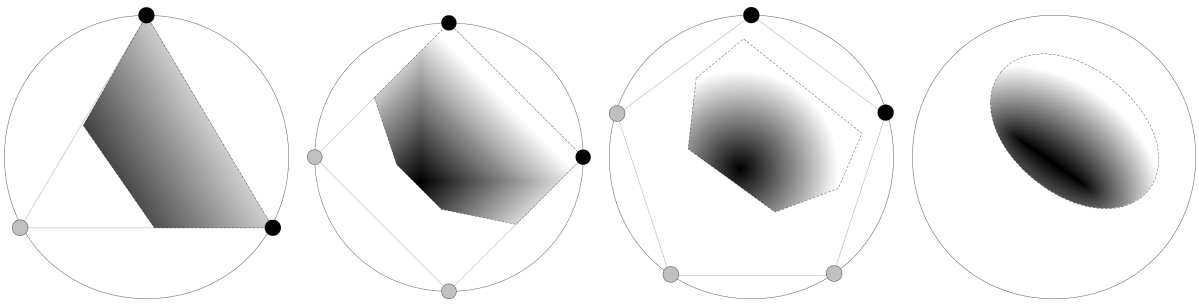
Die Regel kann aus dem Beispiel 11 abgeleitet werden. Hierfür werden Schemata für die Form und Dichte definiert, die bei n Dimensionen und davon c benachbarter korrelierender Dimensionen gilt. Es entsteht eine Polygon mit $n - c + 1$-Eckpunkten, deren Eckpunkte bis auf einen auf den Ankerpunkten der nicht-korrelierten Dimensionen liegen. Die Schemata sind für $n = 3$,$n = 4$,$n = 5$, und $n = \infty$ bei $c = 2$ definiert und in den Abb. 3.25a bis 3.25d zu sehen. Die Ankerpunkte der korrelierenden Dimensionen sind hier schwarz und die restlichen Ankerpunkte grau markiert. Allgemeint lässt sich formulieren, dass eine $n - c + 1$-eckiges Polygon entsteht, von dem $n - c$ Eckpunkte auf den Ankerpunkte der nicht-korrelierenden Dimensionen liegt. Der Eckpunkt mit dem spitzesten Winkel liegt auf Mittelpunkt der konvexen Hülle der c Ankerpunkte. Im Fall von $n = \infty$ ergibt sich eine kreisförmig Struktur mit zum Zentrum zunehmemnder Dichte.

Analog ergeben sich aus dem Beispiel 12 die Schemata für den Fall zweier negativ korrelierender Dimensionen. Die Schemata sind in Abb. 3.26 ersichtlich. Im Fall, dass $n = \infty$ ist die Form elliptisch, und ihre Hauptachse verläuft parallel zur Gerade zwischen den Ankerpunkten der negative korrelierenden Dimensionen. Die Dichte flacht in dieser Richtung flacher und in entgegengesetzter Richtung steiler ab.



(a) RVP (b) RVP (c) RVP (d) RVP

Abbildung 3.25: Visualisierung der Regel 6



(a) RVP (b) RVP (c) RVP (d) RVP

Abbildung 3.26: Schematische Darstellung der Regel 6 bezüglich der Abhängigkeit der Form und Dichte im SP und PCP von der Art der Verteilung der Daten

4 Evaluierung

Die im vorigen Kapitel aufgestellten Analyseregeln sind bereits durch die Untersuchung mehrerer synthetischer Datensätze erörtert worden. Die Daten wurden dabei so modelliert, dass sie bestmöglichst den Sachverhalt der jeweiligen Regel verdeutlichen. Dadurch ist die Aussagekraft der Regeln für diese speziellen Datensätze eindeutig gegeben. Entscheidend für die Anwendbarkeit und damit die Relevanz einer Regel ist jedoch, wie stark sie bei der Betrachtung von allgemeinen Datensätzen greift. Um diesen Aspekt einschätzen zu könnnen, werden die Regeln in diesem Kapitel bezüglich ihrer Allgemeingültigkeit in Bezug auf Standarddatensätze der statistischen Literatur untersucht.

Die Evaluierung erfolgt dabei ebenso wie die Untersuchung der synthetischen Daten getrennt nach bi- und multivariaten Regeln. Dazu werden zunächst verschiedene Plots der Standarddatensätze mit den vier Visualisierungstechniken erzeugt. Aufgrund der großen Anzahl an möglichen Dimensionskombinationen, die im multivariaten Fall durch den Einfluß der Achsenanordnung zusätzlich wächst, werden zielgerichtet nur solche Plots untersucht, die ein hohes Potenzial haben eine Regel entweder zu bestätigen oder zu widerlegen. Nach der Auswahl dieser Plots werden die Regeln angewandt und die Ergebnisse anhand statistischer Berechnungen validiert.

Dazu gehören die in den Grundlagen vorgestellten Verfahren: Spearmans Rho, Kolmogorow-Smirnov-Test, Local Outlier Factor und DBSCAN. Der KS-Test auf Exponentialverteilung wird hierbei durch den Test auf Gammaverteilung ersetzt. Die Gammaverteilung ist eine Verallgemeinerung der Exponentialverteilung und hat den Vorteil, dass einzelne Werte beidseitig vom Maximum der Verteilung liegen können. Dadurch ist der Test robuster gegen Ausreißer bei ähnlicher Charakteristik der Verteilung. Da sich die Gammaverteilung bei der Schätzung ihrer Parameter p und b der Normalverteilung annähern kann, werden Schätzung mit einem Verhältnis von $E(X) = \frac{p}{b} \geq \frac{1}{3}$ verworfen. Dadurch ist die Gammaverteilung innerhalb der Normierung auf $[0, 1]$ stets von der symmetrischen Normalverteilung unterscheidbar, die hier einen Erwartungswert von $\frac{1}{2}$ aufweist. Die Werte der Standarddatensätze sind zumeist auf zwei Stellen hinter dem Komma gerundet. Das gilt ebenso für alle in diesem Kapitel gelisteten Werte. Diese Diskretisierung beinflusst jedoch das Ergebnis eines KS-Test negativ, weshalb dafür die Daten zuvor um $\frac{1}{5}$ ihres kleinsten Abstandes verrauscht werden.

4.1 Anwendung der bivariaten Regeln

Die Untersuchung der bivariaten Regeln erfolgt in den Standarddatensätze *Olives* und *Yeast*. Da die Regel 4 eine visuelle Abmessung der Struktur entlang der aufsteigenden Diagonalen erfordet, werden folgende Rahmenbedingungen für die auf die Regel bezogenen Plots festgelegt: das Seitenverhältnis ist 1:1, die Auflösung beträgt 300dpi und die Gesamtbreite und -höhe des Plots beträgt 7,5cm. Daraus ergibt sich eine Distanz der Horizontalen im SP von $l_x = 692$ Pixel, also 5,86cm, sowie eine Höhe der Dimensionsachsen im PCP von $d_A = 745$ Pixel, also 6,31cm.

4.1.1 Yeast

[1]

Als erstes werden die Dimensionen V_2 und V_3 betrachtet. Mit Hilfe der Schemata von Regel 3 erkennt man in Abb. 4.1a die elliptische Punktwolke mit zum Zentrum abnehmender Helligkeit als $\mathcal{N} \leftrightarrow \mathcal{N}$ verteilen Cluster. Dies bestätigt zudem das entsprechende Schema im PCP, da hier eine rechteckige Form mit linear und parallel zur Horizontalen abnehmenden Helligkeit erkennbar ist. Eine weitere kleinere Struktur mit elliptischer Form überlagert den Cluster, kann aber wegen der geringen Punktanzahl nicht eindeutig klassizifiziert werden. Zur weiteren Analyse werden beide Strukturen entlang ihrer kleinsten Dichte im SP getrennt (siehe Anhang: `eval/yeast_select23a.csv`).

Als geschätzte Parameter der Verteilungen der Daten des Cluster und der Teststatistik D des KS-Test ergeben sich folgenden Werte:

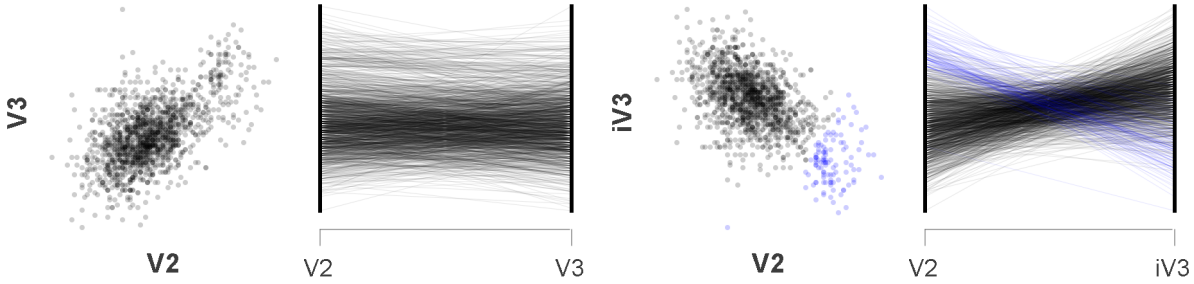
$V_2 \sim$	D	$V_3 \sim$	D
$\mathcal{U}(0,1)$	0.23	$\mathcal{U}(0,1)$	0.22
$\mathcal{N}(0.52, 0.16)$	0.02	$\mathcal{N}(0.5, 0.15)$	0.02
$\gamma(11, 21)$	0.05	$\gamma(11, 22)$	0.06

Bei beiden Dimensionen weist die geschätzte Normalverteilung die geringste Distanz auf, und entspricht somit von den drei Verteilungen am wahrscheinlichsten den Daten. Damit bestätigt der Datensatz das Schema der Normal-zu-Normalverteilung.

Aufgrund der Tatsache, dass der Cluster elliptisch und nicht kreisförmig ist, muss der Korrelationskoeffizient nach Regel 4 einen durch die Strukturbreite bestimmten Wert ungleich 0 aufweisen. Zur Schätzung von ϱ^* wird V_3 invertiert, da die Regel nur für Strukturen mit negativem Anstieg gilt. Die Abb. 4.2 zeigt die Ausmaße der Struktur entlang der aufsteigenden Diagonalen im SP und entlang der zentralen Vertikalen im PCP. Um robuster gegenüber Ausreißer zu sein, wird nur die Strecke vom zweitkleinsten zum zweitgrößten Wert abgetragen. Die Länge der Distanzen beträgt im Verhältnis:

$$d_{\text{SP}} = \frac{377}{692} = 0.54 \text{ und } d_{\text{PCP}} = \frac{403}{745} = 0.54$$

[1] Der Yeast-Datensatz befindet sich mit angepasster Namensgebung und auf $[0, 1]$-normiert im Anhang unter `eval/yeast_n.csv`. Er hat einen Umfang von $n = 1484$ Werte bei 8 Dimensionen.



(a) SP, original (b) PCP. original (c) SP, Cluster (d) PCP, Cluster

Abbildung 4.1: Anwendung von Regel 3 auf die 2. und 3. Dimension des Yeast-Datensatzes

Mit $d > \frac{1}{2}$ liegt der Wert damit geringfügig oberhalb des empfohlenen Bereiches für normalverteilte Daten. Als geschätzter Korrelationskoeffizient errechnet sich daraus:

$$\varrho^*(0.54) = -0.42$$

Der korrekte Wert ergibt sich erst nach einem Vorzeichenwechsel, da die Dimension V_3 invertiert wurde. Der tatsächliche Korrelationskoeffizient nach Spearman beträgt für den Cluster:

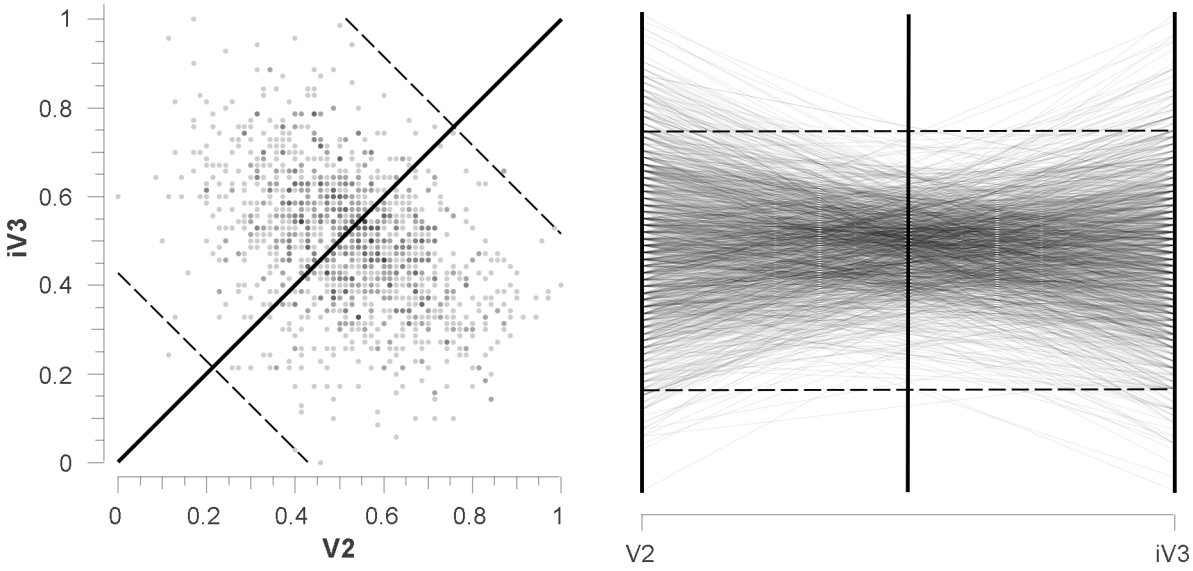
$$\varrho(V_2, V_3) = 0.42$$

Der Fehler der Schätzung bewegt sich in diesem Fall in der Größenordnung von 10^{-2} und bestätigt die Regel 4.

Zur weiteren Validierung der Schemata von Regel 3 eignen sich noch die Dimensionspaare (V_3, V_5) und (V_5, V_9). Im SP der Abb. 4.1a erkennt man eine gleichschenklige, dreieckige Form, mit elliptisch zum Nullpunkt der V_5-Achse und Mittelpunkt der V_3-Achse abnehmender Helligkeit. Dies entspricht dem $\mathcal{N} \leftrightarrow$ Exp-Schema. Der PCP zeigt entsprechend die oben abgerundete, sonst rechteckige Form mit dem linearen dunklen Streifen vom Mittelpunkt der V_3-Achse zum Nullpunkt der V_5-Achse. Daraus folgt, dass die Daten von V_3 normalverteilt und die von V_5 exponentialverteilt sein müssen.

Der SP in Abb. 4.3b zeigt ebenfalls eine dreieckige Form, diesmal mit einem rechten Winkel im Nullpunkt beider Achsen. Die Helligkeit nimmt linear und parallel zur Hypothenuse in Richtung des Nullpunktes ab. Im PCP erkennt man die oben abgerundete, rechteckige Form, die auf der unteren Seite durch strukturfremde Werte leicht verzerrt wird. Die Helligkeit nimmt linear und parallel zum Nullpunkt beider Achsen ab. Der SP und PCP entspricht somit dem Exp $\leftrightarrow$ Exp-Schema, sodass die Daten von V_5 und V_9 exponentialverteilt sein müssen.

Die Ergebnisse der KS-Tests sind:



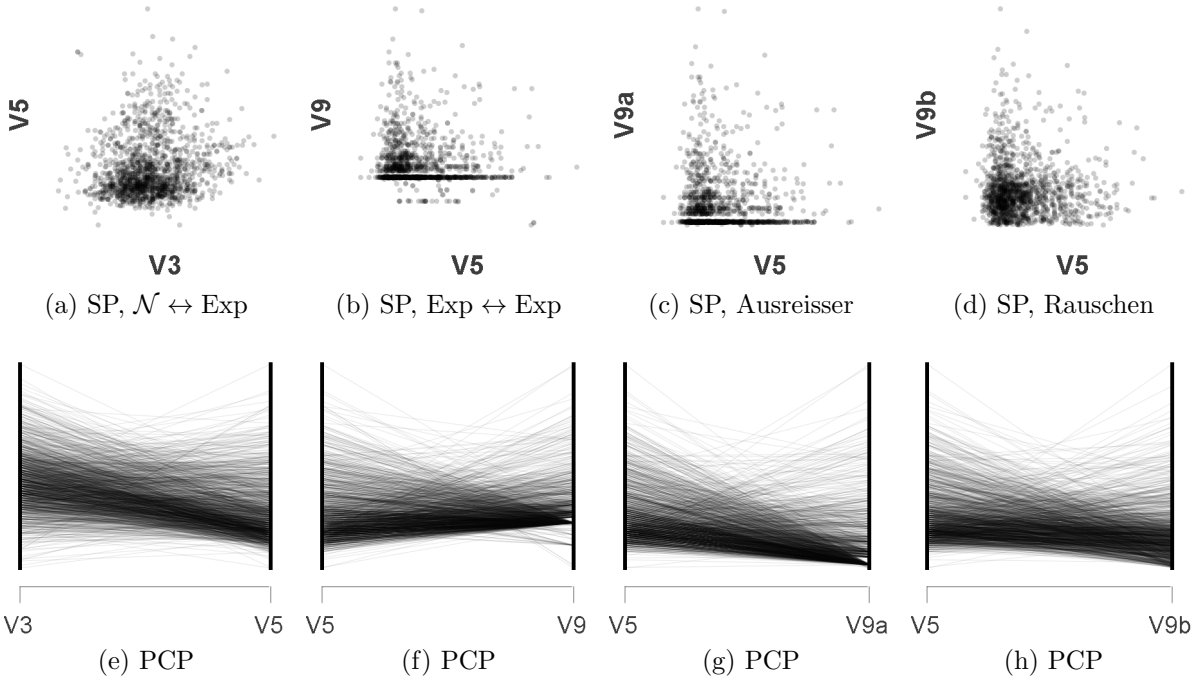
(a) SP, $\frac{d_S}{d_D} = 0.54$, $\varrho^* = -0.42$ (b) PCP, $\frac{d_S}{d_A} = 0.54$, $\varrho^* = -0.42$

Abbildung 4.2: Anwendung von Regel 4 auf die 2. und 3. Dimension des Yeast-Datensatzes

$V_3 \sim$	D	$V_5 \sim$	D	$V_9 \sim$	D
$\mathcal{U}(0,1)$	0.30	$\mathcal{U}(0,1)$	0.45	$\mathcal{U}(0,1)$	0.50
$\mathcal{N}(0.42, 0.14)$	0.05	$\mathcal{N}(0.26, 0.14)$	0.14	$\mathcal{N}(0.28, 0.11)$	0.26
$\gamma(8.9, 21)$	0.04	$\gamma(3.6, 14)$	0.08	$\gamma(6.8, 24)$	0.29

Die Dimension V_3 hat zur geschätzten Gammaverteilung die geringste Distanz. Dieser Wert wird allerdings wegen $\frac{p}{b} > \frac{1}{3}$ verworfen. Damit weist die Normalverteilung die geringste Distanz auf, wodurch ein Teil des $\mathcal{N} \leftrightarrow$ Exp-Schema bestätigt ist. Bei der Dimension V_5 weist die Gammaverteilung ebenfalls die geringste Distanz auf. In diesem Fall beträgt das Verhältnis ihrer Parameter $\frac{p}{b} < \frac{1}{3}$ und wird nicht verworfen, sodass die andere Seite des $\mathcal{N} \leftrightarrow$ Exp-Schemas und eine Seite des Exp $\leftrightarrow$ Exp-Schemas bestätigt wird.

Der KS-Test ergibt für V_9 durchweg hohe Distanzen, wobei die Normalverteilung den geringsten Wert hat. Die Vermutung liegt nahe, dass die strukturfremden Werte unterhalb des Maximum (siehe Abb. 4.1) einen zu starken Einfluss auf den KS-Test haben. Wenn man die Ausreißer durch eine Trennlinie knapp unterhalb des Maximums abtrennt, verschlechtert sich der Wert der Distanz entgegen der Annahme auf $\gamma(7.3, 26) : 0.29$. Das Problem stellt hier die stark ausgeprägte Diskretisierung der Werte dar, die man in der Abb. 4.1 an der horizontalen Linienbildung erkennt. Das Hinzufügen eines Rauschens im Umfang der Distanz der horizontalen Lücke ergibt dann einen entsprechend verbesserten Distanzwert von $\gamma(1.6, 10) : 0.05$(siehe Anhang: `eval/yeast_9_new.csv`).



(a) SP, $\mathcal{N} \leftrightarrow$ Exp (b) SP, Exp $\leftrightarrow$ Exp (c) SP, Ausreisser (d) SP, Rauschen

(e) PCP (f) PCP (g) PCP (h) PCP

Abbildung 4.3: Anwendung von Regel 3 auf die 2. und 3. Dimension des Yeast-Datensatzes

4.1.2 Iris

[2]

Das Dimensionspaar V_1, V_2 weist nach Abb. 4.4a zwei visuell trennbare Strukturen im SP auf. Laut Regel 5 lässt sich jede im SP mögliche Clustertrennung auch im PCP durchführen. Der PCP zeigt jedoch keine Möglichkeit zur Selektion eines Hintergrund-Cluster-Hintergrund-Bereichs. Es existieren nur die Hintergrund-Überlagerung-Hintergrund- sowie Cluster-Überlagerung-Cluster-Bereiche. Dies bestätigt Regel 2, da sich der linksobere Cluster im SP fast vollständig in der konvexen Hülle aus der Vereinigung des rechts-unteren Clusters mit der absteigenden Diagonalen des MUR der gesamten Struktur befindet, und somit im PCP ebenfalls innerhalb der konvexen Hülle seiner Linienbündel verschwindet.

Eine Invertierung der V_2-Achse ändert die Anordnung der Cluster, sodass sich ihre konvexen Hüllen im SP und damit auch im PCP nur noch teilweise überlagern. Dadurch entsteht die Möglichkeit der Selektion eines Hintergrund-Cluster-Hintergrund-Bereiches, sodass die Cluster gekennzeichnet werden können (siehe Abb. 4.4b. Die Abb. 4.4c zeigt gegenüberstellend die Clusteranalyse mittels DBSCAN bei einem Epsilonwert vom dreifachen Mindestabstand der Punkte.

4.1.3 Olives

[3]

Die Dimensionspaare (V_8, V_9) und (V_7, V_8) des Olives-Datensatzes zeigen im ersten Fall eine lineare und im zweiten Fall eine punkförmige Struktur. Nach Regel 4 muss der Korrelationskoeffizient der linearen Struktur einen höheren absoluten Wert als die punktförmige Struktur haben. Der Wert der punktförmigen Struktur muss gegen 0 gehen. Zur Untersuchung werden die beiden Cluster jeweils von den restlichen Daten getrennt (siehe Anhang: `eval/olives_89_cl.csv` und `eval/olives_78_cl_1.csv`).

Die Abb. 4.5 zeigt die Strukturausmaße im SP und PCP der für das Dimensionspaar (V_8, V_9). Hierbei ergeben sich als relative Strukturbreiten die Werte:

$$d_{\mathrm{SP}} = \frac{300}{692} = 0.43 \text{ und } d_{\mathrm{PCP}} = \frac{324}{745} = 0.43$$

Daraus ergibt sich als geschätzer Korrelationskoeffizient mit vertauschtem Vorzeichen:

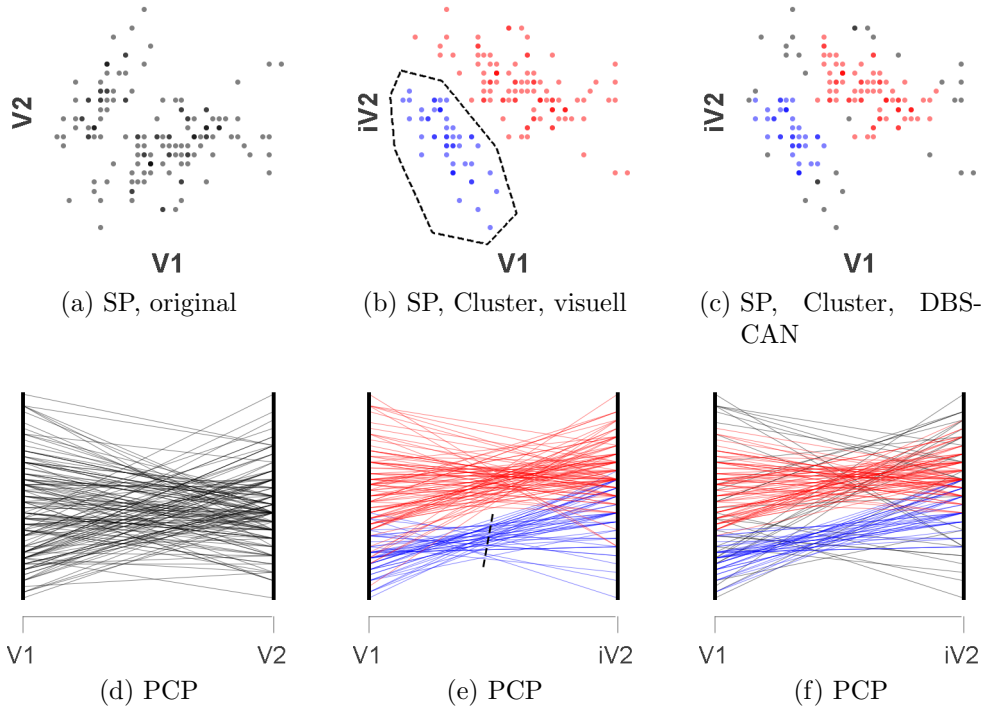
$$\varrho^*(0.43) = -0.63$$

bei einem tatsächlichen Koeffizienten nach Spearman von:

$$\varrho(V_8, V_9) = 0.60$$

[2] Der Iris-Datensatz befindet sich mit angepasster Namensgebung und auf $[0, 1]$-normiert im Anhang unter `eval/iris_n.csv`. Er hat einen Umfang von $n = 150$ Werte bei 4 Dimensionen.

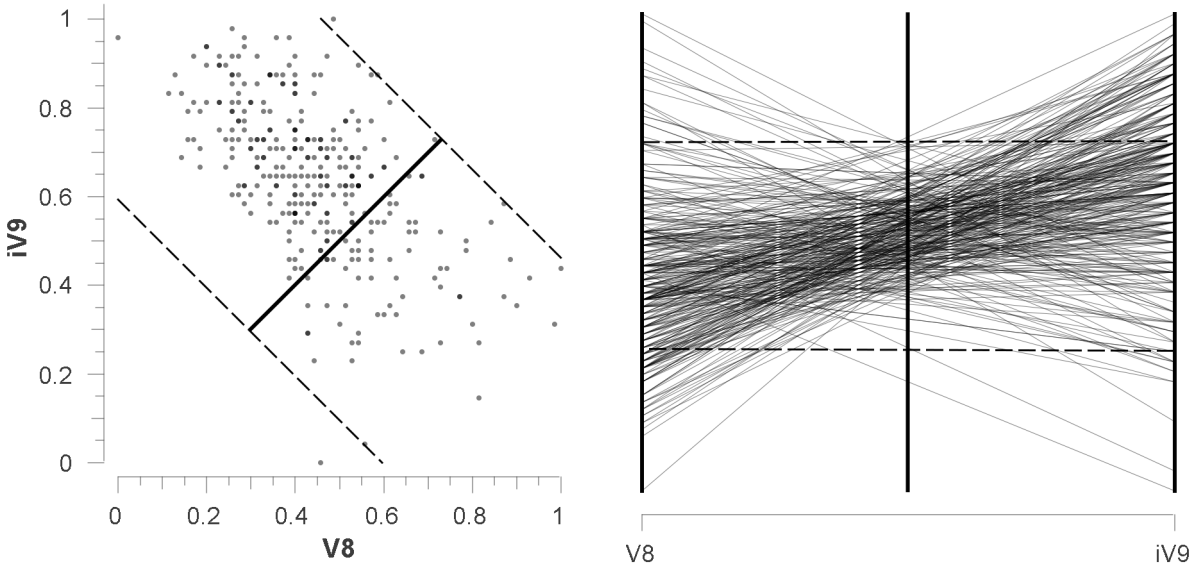
[3] Der Olives-Datensatz befindet sich mit angepasster Namensgebung und auf $[0, 1]$-normiert im Anhang unter `eval/olives_n.csv`. Er hat einen Umfang von $n = 572$ Werte bei 9 Dimensionen.



(a) SP, original (b) SP, Cluster, visuell (c) SP, Cluster, DBS-CAN

(d) PCP (e) PCP (f) PCP

Abbildung 4.4: Anwendung von Regel 2 und Regel 5 auf die 1. und 2. Dimension des Iris-Datensatzes



(a) SP, $\frac{d_S}{d_D} = 0.43$, $\varrho^* = -0.63$ (b) PCP, $\frac{d_S}{d_A} = 0.43$, $\varrho^* = -0.63$

Abbildung 4.5: Anwendung von Regel 4 auf die 2. und 3. Dimension des Yeast-Datensatzes

Damit weicht der geschätzte Wert im Gegensatz zur Berechnung im Yeast-Datensatz vom tatsächlichen Wert leicht ab. Das kann darauf zurückgeführt werden, dass die Struktur nicht komplett dem zugrundeliegenden, symmetrischen Modell der Gleich- bzw. Normalverteilung (quadratisch, bzw. kreisförmig) entspricht.

Für das Dimensionspaar (V_7, V_8) sind die Strukturausmaße in Abb. 4.6 abgetragen. Das Längenverhältnis beträgt dann:

$$d_{\text{SP}} = \frac{482}{692} = 0.7 \text{ und } d_{\text{PCP}} = \frac{516}{745} = 0.69$$

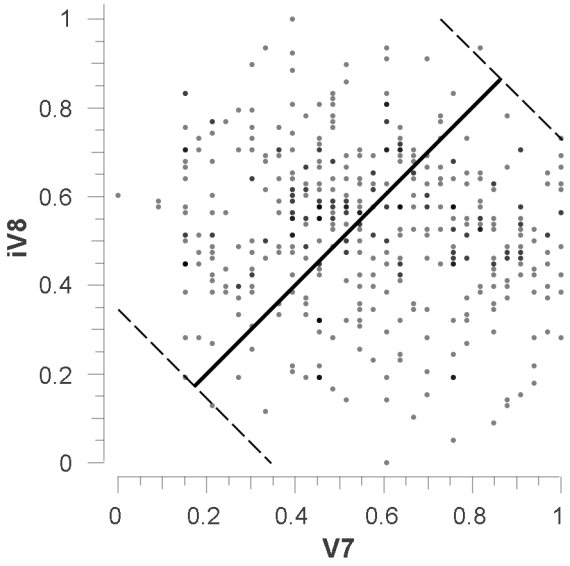
Der Unterschied zwischen beiden Werten ergibt sich aus dem leicht zur V_8-Achse verschobenen Kreuzungsbereich des PCP. Die Annäherung des Korrelationskoeffizienten beträgt

$$\varrho^*(0.7) = -0.15$$

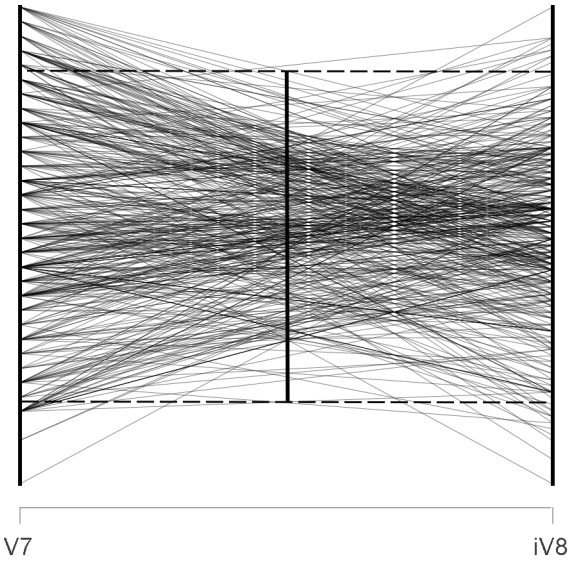
und weicht damit ebenfalls vom tatsächlichen Wert ab

$$\varrho(V_7, V_8) = 0.11$$

Dabei rührt dieser Fehler schon daher rührt, dass die relative Strukturbreite mit 0.7 außerhalb des empfohlenen Bereiches für normalverteilte Daten liegt.



(a) SP, $\frac{d_S}{d_D} = 0.7$, $\varrho^* = -0.15$



(b) PCP, $\frac{d_S}{d_A} = 0.69$, $\varrho^* = -0.16$

Abbildung 4.6: Anwendung von Regel 4 auf die 2. und 3. Dimension des Yeast-Datensatzes

4.2 Anwendung der multivariaten Regeln

4.2.1 Olives

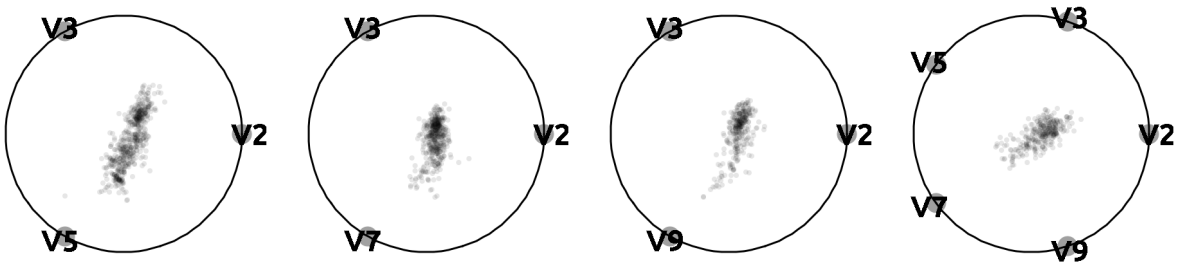
Die Validierung der Regel 6 erfolgt im Olives-Datensatz. Der Regel entsprechend sind solche 3-dimensionale Plots interessant, die eine lineare Struktur aufweisen. In dem Fall korrelieren 2 Dimensionen stark zueinander, und die dritte Dimension schwach zu den beiden anderen. Die Betrachtung aller 3-dimensionalen Kombinationen zeigt diese interessante Struktur für (V_2, V_3, V_5), (V_2, V_3, V_7) und (V_2, V_3, V_9) (siehe Abb. 4.7a, b, c). Hier ist zu beachten, dass für V_5 und V_7 die Cluster *East-Liguria* und *West-Liguria*, sowie für V_9 die Cluster *Coast-Sardinia*, *East-Liguria*, *Inland-Sardinia*, *West-Liguria* und *Umbria* ausgeblendet werden müssen, da diese einen Störeinfluss haben.

Die lineare Struktur hat einen Endpunkt stets auf der Geraden zwischen V_2 und V_3, weshalb diese die beiden stark korrelierenden Dimensionen sein müssen. Dieses Ergebnis wird durch die Berechnung von Spearmans Korrelationskoeffizienten bestätigt.

$\varrho(V_2, V_3) = 0.80$

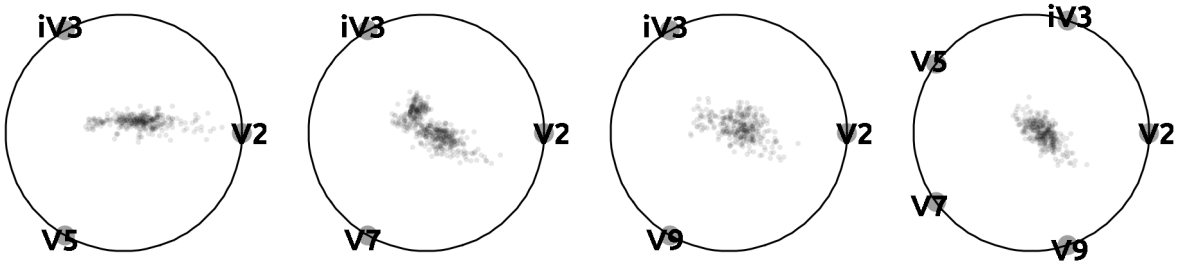
Weiterhin muss nach Regel 6 entsprechend dem Schema für den 5-dimensionalen Fall ein Eckpunkt in Richtung der korrelierenden Dimensionen zeigen. Dies ist auch nach Abb. 4.7d der Fall, wodurch die Regel für positive Korrelation bestätigt wird.

Invertiert man die Dimension V_2 oder V_3 erhält man eine negative Korrelation, für die der zweite Teil von Regel 6 gelten muss. Die Abb. 4.8a,b und c zeigen jedoch vom



(a) RVP (b) RVP (c) RVP (d) RVP

Abbildung 4.7: Validierung der Regel 6 mittels Olives-Datensatz



(a) RVP (b) RVP (c) RVP (d) RVP

Abbildung 4.8: Validierung der Regel 6 mittels Olives-Datensatz

Schema abweichende Formen oder eine fehlende Parallelität zu den negativ korreliernden Dimensionen. Nur die Abb. 4.8d, welche alle 5 Dimensionen darstellt, entspricht dem Schema der Regel 6: die Hauptachse der elliptischen Struktur ist parallel zur Geraden zwischen den Ankerpunkten der negativ korrelierenden Dimensionen und die Dichte fällt zur Gerade flacher und entgegengesetzt steiler ab.

5 Zusammenfassung

In der Diplomarbeit wurden Analyseregeln aufgestellt, die sich durch die Untersuchung modellierter synthetischer Datensätze und dem Zusammentragen besonderer Merkmale abgeleitet haben. Die Regeln umfassen Schemata zur visuellen Abschätzung der Art der Verteilung und des Korrelationskoeffizienten im SP, PCP und RVP, sowie eine regelbasierte visuellen Vorgehensweise zur Clusterung von Daten im PCP. Durch die Anwendung auf reale Standarddatensätze und der Gegenüberstellung der ermittelten Ergebnisse zu den tatsächlichen Eigenschaften, wurden die Analyseregeln für den praktischen Einsatz validiert.

5.1 Schlussfolgerungen

Es hat sich gezeigt, dass solche Zusammenhänge zwischen Daten und auffälligen Merkmalen in verschiedenen Visualisierungstechniken existieren, über welche objektive Aussagen getroffen werden können. Damit kann nicht nur die vorherrschende Autodidaktik in diesem Bereich reduziert werden. Es wird auch allgemein die Datenexploration erleichtert, indem eine rein visuelle Bestimmung von statistischen Eigenschaften erfolgen kann. Das beschleunigt die visuelle Analyse durch den Verzicht auf Zahlen und der Fokussierung auf die intuitive Mustererkennung der erarbeiteten Schemata, bei trotzdem hoher Signifikanz.

Ein weiterer Vorteil ergibt sich durch die Schätzung des Korrelationskoeffizienten auf Basis der Strukturbreite. Hat man zwei sich überlappende Cluster, aber keine Kennzeichnungen der einzelnen Zugehörigkeiten, kann man den struktureigenen Koeffizienten nicht ohne Einfluss durch die in diesem Fall verlustbehafte Trennung der Cluster berechnen. Eine Schätzung anhand der erkennbaren Strukturbreite würde hier einen genaueren Wert liefern.

5.2 Einschränkungen

Nachteilig anzusehen, ist der erforderliche relativ hohe Umfang an Werten um zweifelsfrei erkennbare Formgebung und Dichteverläufe darzustellen. Die Untersuchung der Standarddatensätze hat im Allgemeinen gezeigt, dass die Anzahl der Daten teils zu gering waren, um die Schemata zuzuordnen. Des Weiteren haben auch Rundungen die Wahrnehmung beeinflusst, da so die stetige Komponente für einen lückenfreien Dichteverlauf fehlt. Im Speziellen ist die Abschätzung des Korrelationskoeffizienten bei den

häufig anzutreffenden normalverteilten Daten nicht robust gegenüber der unsauberen Strukturgrenzen aufgrund eines geringen Datenumfangs.

5.3 Gegenmaßnahmen

Bezüglich der visuellen Schätzung des Korrelationskoeffizienten von normalverteilten Daten muss ein besseres Modell, als das von der Gleichverteilung hergeleitete Modell der Strukturbreite. Die nächstliegende Möglichkeit ist, den Strukturmittelpunkt in die Schätzung mit einzubeziehen, da dieser sich durch das dortige Dichtemaximum abzeichnet. Auf der anderen Seite können zum einen die Daten vorbehandelt werden, indem angemessenes Rauschen und Zwischenwerte interpoliert werden. Dies entkräftet die Diskretisierung durch eventuelle Rundungen und erhöhte den Datenumfang für klare Strukturformen und Dichteverläufe. Zum anderen können auch die Visualisierungstechniken bezüglich dieser Problematik weiter entwickelt werden, sodass diese auch bei wenigen Werten ein stetiges Dichtefeld generieren.

5.4 Erweiterungen

Die vorgestellten Schemata können in Datenexplorationsprogrammen als frei bewegliche, skalier- und rotierbare transparente Schablonen implementiert werden. Durch eine manuelle Anpassung der Freiheitsgrade, können die Schablonen vom Anwender interaktiv über den gerade visualisierten Datensatz platziert werden, um somit dem Schema entsprechende Bereiche ausfindig zu machen. Ein im bestimmten Umfang automatisiertes Anpassen der Freiheitsgrade an die unterliegende Struktur kann den Explorationsvorgang dabei vereinfachen.

Die Schemata bezüglich der Korrelation zwischen den Dimensionen in der Radviz-Visualisierung können als Grundlage für ein Quality-Measure dienen. Damit kann die beste Ordnung der Ankerpunkte hinsichtlich der Darstellung korrelierter Dimensionen gefunden werden.

5.5 Ausblick

Die weitere Forschung im Bereich der Analyseregeln für Visualisierungen von hochdimensionalen Datensätzen sollte auf langer Sicht die Erarbeitung eines standardisierten, allgemeingültigen und maschinenverständlichen Regelwerks als Ziel haben. Dadurch könnte sich die Problematik des eigenständigen Erlernens subjektiver Regeln vermindern und es würden sich neue Quality-Measures ergeben. Dazu müssen einerseits die hier genannten Techniken tiefgründiger, aber anderseits auch ein breiteres Spektrum an weiteren Methoden untersucht werden.

Literaturverzeichnis

[AEL+10] Albuquerque, Georgia ; Eisemann, Martin ; Lehmann, Dirk J. ; Theisel, Holger ; Magnor, Marcus: Improving the Visual Analysis of High-dimensional Datasets Using Quality Measures. In: *Proc. IEEE Symposium on Visual Analytics Science and Technology (VAST) 2010*. Salt Lake City, Utah, USA, 2010, S. 19–26

[AES05] Amar, Robert ; Eagan, James ; Stasko, John: Low-Level Components of Analytic Activity in Information Visualization. In: *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*. Washington, DC, USA : IEEE Computer Society, 2005. – ISBN 0–7803–9464–x, 15–

[ALM11] Albuquerque, Georgia ; Löwe, Thomas ; Magnor, Marcus: Synthetic Generation of High-dimensional Datasets. In: *IEEE Transactions on Visualization and Computer Graphics (TVCG, Proc. Visualization / InfoVis)* 17 (2011), Nr. 12

[Asi85] Asimov, Daniel: The grand tour: a tool for viewing multidimensional data. In: *SIAM J. Sci. Stat. Comput.* 6 (1985), January, 128–143. `http://dx.doi.org/10.1137/0906011`. – DOI 10.1137/0906011. – ISSN 0196–5204

[BKNS00] Breunig, Markus ; Kriegel, Hans-Peter ; Ng, Raymond T. ; Sander, Jörg: LOF: Identifying Density-Based Local Outliers. In: *PROCEEDINGS OF THE 2000 ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA*, ACM, 2000, S. 93–104

[CMS99] Card, Stuart K. ; Mackinlay, J. D. ; Shneiderman, Ben: *Readings in Information Visualization: Using Vision to Think*. San Francisco, USA : Morgan Kaufmann Publishers, Inc., 1999

[Com11] Committee, SciPy S.: *SciPy*. `http://www.scipy.org/`. Version: Oktober 2011

[DHH11] Dubská, Markéta ; Herout, Adam ; Havel, Jiří: PClines - Line Detection Using Parallel Coordinates. In: *Proceedings of CVPR 2011*, IEEE Computer Society, 2011. – ISBN 978–1–4577–0393–5, 1489–1494

[Eat11] Eaton, John W.: *Octave*. http://www.gnu.org/software/octave/. Version: Oktober 2011

[EKSX96] Ester, Martin ; Kriegel, Hans-Peter ; Sander, Joerg ; Xu, Xiaowei: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Simoudis, Evangelos (Hrsg.) ; Han, Jiawei (Hrsg.) ; Fayyad, Usama M. (Hrsg.): *Second International Conference on Knowledge Discovery and Data Mining*, AAAI Press, 1996, S. 226–231

[FSF11] Free Software Foundation, Inc.: *GNU PSPP*. http://www.gnu.org/software/pspp/. Version: Oktober 2011

[FT74] Friedman, J. H. ; Tukey, J. W.: A Projection Pursuit Algorithm for Exploratory Data Analysis. In: *IEEE Trans. Comput.* 23 (1974), September, 881–890. http://dx.doi.org/10.1109/T-C.1974.224051. – DOI 10.1109/T–C.1974.224051. – ISSN 0018–9340

[Haw80] Hawkins, Douglas M.: *Identification of outliers*. Chapman and Hall, 1980

[HGM+97] Hoffman, Patrick ; Grinstein, Georges ; Marx, Kenneth ; Grosse, Ivo ; Stanley, Eugene: DNA visual and analytic data mining. In: *Proceedings of the 8th conference on Visualization '97*. Los Alamitos, CA, USA : IEEE Computer Society Press, 1997 (VIS '97). – ISBN 1–58113–011–2, 437–ff.

[ID90] Inselberg, Alfred ; Dimsdale, Bernard: Parallel coordinates: a tool for visualizing multi-dimensional geometry. In: *Proceedings of the 1st conference on Visualization '90*. Los Alamitos, CA, USA : IEEE Computer Society Press, 1990 (VIS '90). – ISBN 0–8186–2083–8, 361–378

[Kru64] Kruskal, J.B.: Nonmetric multidimensional scaling: A numerical method. In: *Psychometrika* 29 (1964), Nr. 2, S. 115–129

[Lil67] Lilliefors, Hubert W.: On the Kolmogorov-Smirnov test for normality with mean and variance unknown. In: *J Amer Statistical Assoc* 62 (1967), Nr. 318, S. 399–402

[Lil69] Lilliefors, H W.: On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. In: *Journal of the American Statistical Association* 64 (1969), Nr. 325, 387– 389. http://www.jstor.org/stable/2283748

[Lov11] Lovric, Miodrag: *International Encyclopedia of Statistical Science*. Springer Verlag, 2011. – ISBN 978–3–642–04897–5

[Ls11] Ljubljana, Faculty of c. o. ; science information: *Orange - Data Mining Fruitful & Fun*. http://orange.biolab.si/. Version: Oktober 2011

[Mas51] Massey, F J.: The Kolmogorov-Smirnov Test for Goodness of Fit. In: *Journal of the American Statistical Association* 46 (1951), Nr. 253, 68–78. http://www.jstor.org/stable/2280095

[MM08] McDonnell, K. T. ; Mueller, K.: Illustrative Parallel Coordinates. In: *Computer Graphics Forum* 27 (2008), Nr. 3, 1031–1038. http://dx.doi.org/10.1111/j.1467-8659.2008.01239.x. – DOI 10.1111/j.1467–8659.2008.01239.x. – ISSN 1467–8659

[Nem93] Nemcsics, Antal: *Farbenlehre und Farbendynamik - Theorie der farbigen Umweltplanung*. Göttingen : Muster-Schmidt Verlag, 1993

[NHM97] Nielson, Gregory M. (Hrsg.) ; Hagen, Hans (Hrsg.) ; Müller, Heinrich (Hrsg.): *Scientific Visualization, Overviews, Methodologies, and Techniques, Dagstuhl, Germany, May 1994*. IEEE Computer Society, 1997 . – ISBN 0–8186–7777–5

[Nv06] Nováková, Lenka ; Štepánková, Olga: Multidimensional clusters in RadViz. In: *Proceedings of the 6th WSEAS International Conference on Simulation, Modelling and Optimization*. Stevens Point, Wisconsin, USA : World Scientific and Engineering Academy and Society (WSEAS), 2006. – ISBN 111–9999–33–X, 470–475

[Nv09] Nováková, Lenka ; Štěpánková, Olga: Visualization of Trends Using RadViz. In: *Proceedings of the 18th International Symposium on Foundations of Intelligent Systems*. Berlin, Heidelberg : Springer-Verlag, 2009 (ISMIS '09). – ISBN 978–3–642–04124–2, 56–65

[OL96] Ong, Hwee-Leng ; Lee, Hing-Yan: Software report: WINVIZ – A Visual Data Analysis Tool. In: *Computers and Graphics* 20 (1996), Nr. 1, 83–84. http://www.elsevier.com/cgi-bin/cas/tree/store/cag/cas_sub/browse/browse.cgi?year=1996&volume=20&issue=1&aid=9688777

[RH94] Robertson, Philip K. ; Hutchins, Matthew A.: An Approach to Intelligent Design of Color Visualizations. In: [NHM97], S. 179–190

[Ric95] Richards, L. G.: Applications of Engineering Visualization to Analysis and Design. Version: 1995. http://books.google.de/books?id=E_kzKYNijEcC. In: *Computer visualization: graphics techniques for scientific and engineering analysis*. CRC Press, 1995. – ISBN 9780849390500, 267 – 289

[SC11] Statistical Computing, The R F.: *The R Project for Statistical Computing*. http://www.r-project.org/. Version: Oktober 2011

[SM00] Schumann, Heidrun ; Müller, Wolfgang: *Visualisierung - Grundlagen und allgemeine Methoden*. Springer, 2000. – I–XI, 1–370 S. – ISBN 978–3–540–64944–1

[SS11] Shneiderman, Ben ; Seo, Jinwook: *HCE - Hierarchical Clustering Explorer*. `http://www.cs.umd.edu/hcil/hce/`. Version: Oktober 2011

[TGF11] The GGobi Foundation, Inc.: *GGobi data visualization system.* `http://www.ggobi.org/`. Version: Oktober 2011

[The00] Theisel, H.: Higher order parallel coordinates. In: Girod, B. (Hrsg.) ; Greiner, G. (Hrsg.) ; Niemann, H. (Hrsg.) ; (Ed.), H.-P. S. (Hrsg.): *Proc. Vision, Modeling and Visualization (VMV)*. Saarbrücken, 2000, S. 119–125

[Tor52] Torgerson, W. S.: Multidimensional scaling: I. Theory and method. In: *Psychometrika* 17 (1952), S. 401–419

[Vej00] Vejmelka, Martin: *Precisely Defined Objects*. `http://gerstner.felk.cvut.cz/machine-learning/sw/predo/`. Version: April 2000

[War11] Ward, Prof. Matthew. O.: *XmdvTool Home Page: Overview*. `http://davis.wpi.edu/xmdv/`. Version: Oktober 2011

[WB97] Wong, Pak C. ; Bergeron, R. D.: 30 Years of Multidimensional Multivariate Visualization. In: [NHM97], S. 3–33

Erklärung

1. Ich versichere hiermit, dass ich die vorliegende Diplomarbeit selbständig, ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Magdeburg, den 21. Dezember 2011