LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR STATISTIK

Bestimmung der Anzahl von Clustern mittels einer modifizierten GAP-Funktion

Bachelorarbeit

Name:	Le
Vorname:	Khac Phuoc
Studiengang:	Statistik
Semesterzahl:	6
Betreuer:	Prof. Dr. Volker Schmid
Abgabedatum:	02. August 2011

Eidesstattliche Erklärung zur Bachelorarbeit

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

München, den 02. August 2011

(Khac Phuoc Le)

Inhaltsverzeichnis

1 Einleitung

Clustering-Verfahren spielen bei der Auswertung von Daten eine große Rolle. Das Ziel ist dabei die heterogenen Daten möglichst in homogene Gruppen (Cluster) zu teilen. Dabei sollen die Elemente in den einzelnen Cluster untereinander möglichst ähnlich sein, die Cluster selber jedoch zueinander möglichst unterschiedlich(vgl. Fahrmeir et al. (1996)). Beispielsweise lassen sich in der Praxis (vgl. Eckey et al. (2002)) anhand von Merkmalen wie z.B. Geschlecht, Beruf, Bildung, Einkaufsgewohnheiten oder Lebensstil Konsumentengruppen bilden um geeigente Zielgruppen zu identifizieren. Gleichermaßen wird dies in der Medizin und Psychologie dadurch umgesetzt, dass anhand von Krankheitsbildern und Persönlichkeitsstrukturen versucht wird Patientengruppen zu klassifizieren.

Die Herausforderung liegt nun darin, die optimale Anzahl von Cluster zu bestimmen. Eine Faustregel, die oft angewendet wird um die Clusteranzahl zu finden, ist das sogenannte „Elbow-Kriterium“ (vgl. Backhaus et al. (2008) und Tibshirani et al. (2001)). Hierbei werden die Fehlerquadratsummen bzw. Streuung der Cluster W_k in Abhängigkeit der Anzahl von Clustern betrachtet. Dabei fällt W_k monoton mit der Steigung von der Clusteranzahl k ab, aber ab einem bestimmten k aus nimmt der monotone Abfall deutlich ab. Dieses k, das bei dem dieser „Knick“ zu beobachten ist, wird als optimale Clusterzahl hergenommen.

Tibshirani et al. (2001) hatten das Ziel aus dieser Heuristik ein statistisches Verfahren zu entwickeln und schlagen die sogenannte Gap Statistik vor. Dabei werden die Differenzen von $log(W_k)$, mit der dazugehörigen Referenzverteilung betrachtet. Die optimale Clusterzahl ist dann gegeben, wenn der Abstand zwischen beiden Werten maximal ist. Mohajer et al. (2010) schlugen vor, bei der Berechnung der Gap-Statistik anstatt $log(W_k)$ direkt die Streung der Cluster W_k zu verwenden. Diese wird anschließend mit der Erwartung von W_k unter einer Null Referenzverteilung verglichen. In der Arbeit wurde bei der Berechnung ihrer modifizierten Gap*-Statistik bzw. W_k das hierarchisch agglomerative average linkage Verfahren verwendet. Sie konnten zeigen, dass ihre modifizierte Gap* Funktion in einigen Fällen bessere Ergebnisse liefert als die Gap-Funktion. Im Umfang meiner Bachelorarbeit habe ich untersucht, inwiefern sich die Gap-Funktionen unterscheiden, falls bei der Berechnung das weitgehend bekannte k-means Verfahren verwendet wird.

2 Theorie

2.1 Gap Statistik

Sei $\{x_{ij}\}$ ein Datensatz mit $i = 1, 2, ..., n$ unabhängigen Beobachtungen und $j = 1, 2, ...p$ Variablen, welche in $k \in \mathbb{N}$ Cluster $C_1, C_2, ..., C_k$ geteilt wird. Dabei bezeichnet C_r das r-te Cluster und $n_r = |C_r|$ die Anzahl der darin enthaltenen Beobachtungen. Ferner sei $d_{ii'}$ die Distanz zwischen Beobachtung i und i'. Ein mögliches Distanzmaß wäre die quadratische Euklidische Distanz

$$d_{ii'} = \sum_j (x_{ij} - x_{i'j'})^2 \qquad .$$

Die Summe der paarweisen Distanzen D_r für alle Punkte im Cluster r lautet

$$D_r = \sum_{i,i' \in C_r} d_{ii'} \qquad . \tag{1}$$

Mit (1) können wir W_k folgendermaßen definieren:

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r \qquad . \tag{2}$$

Falls d die quadratische Euklidische Distanz ist, dann ist W_k die within-cluster dispersion. Diese lässt sich dadurch berechnen, indem von jedem Cluster die mittleren quadratischen Abstände berechnet und diese anschließend summiert werden. Die Idee von Tibshirani et al. (2001) war, sich die Differenz von $log(W_k)$ mit einer geeigneten Referenzverteilung von $log(W_k)$ anzuschauen. Die Schätzung für die optimale Anzahl von Clustern k ist gegeben, bei dem $log(W_k)$ den größten Abstand zu dessen Referenzverteilung hat. Darum wird die Gap-Statistik wie folgt definiert:

$$Gap_n(k) = E_n^* log(W_k) - log(W_k) \qquad . \tag{3}$$

E_n^* bezeichnet dabei die Erwartung einer Stichprobe der Größe n von der Referenzverteilung. Für das Erzielen der wahrscheinlich besten Ergebnisse für die Gap Statistik schlägt Tibshirani et al. (2001) als Referenzverteilung erzeugte Datenpunkte vor, die auf die Originaldaten basierend gleichverteilt sind. Für die Berechnung der Referenzverteilung wurden folgende Möglichkeiten vorgeschlagen:

1. Sei $n \times p$ die Dimension des Originaldatensatzes mit $i = 1, 2, ..., n$ und $j = 1, 2, ...p$. Ferner sei W_j^O der Wertebereich der j-ten Variable des Originaldatensatzes. Erstelle die Refererenzdatensätze der gleichen Dimension und

die p Variablen wie folgt: Für die Werte der j-ten Variable des Referenzdatensatzes ziehe n gleichverteilte Werte aus dem entsprechenden Wertebereich des Originaldatensatzes $\mathbb{W}_j^O$.

2. Erzeuge die Referenzdatensätze aus einer Gleichverteilung über einer Box mit den abgeglichenen Hauptkomponenten des Datensatzes. Falls X unsere $n \times p$ Datenmatrix ist, gehe davon, dass die Variablen jeweils den Mittelwert von 0 besitzen und erzeuge mit der Singulärwertzerlegung $X = UDV^\top$. Transformiere $X' = XV$ und ziehe anschließend gleichverteilte Werte Z' aus dem Wertebereich der Variablen von X' wie in der ersten Methode. Schließlich transformiere diese zurück durch $Z = Z'T^\top$, um den Referenzdatensatz Z zu erhalten.

Die erste Methode hat den Vorteil, dass sie einfach ist. Die zweite Methode berücksichtigt die Verteilung der Daten und ermöglicht somit eine invariante Prozedur, sofern die Clusterverfahren selbst invariant sind. Für die Schätzung von $E_n^* log(W_k)$ werden B Monte Carlo Stichproben generiert, die jeweils aus der Referenzverteilung gezogen werden. Für jeden erzeugten Datensatz wird $log(W_k^*)$ berechnet und anschließend aus den B Werten der Mittelwert gebildet. Somit erhalten wir für die Schätzung:

$$E_n^* log(W_k) = \frac{1}{B} \sum_b log(W_{kb}^*) \tag{4}$$

Somit ergibt sich für die manuelle Berechnung der Gap-Statistik folgende Gleichung:

$$Gap_n(k) = \frac{1}{B} \sum_b log(W_{kb}^*) - log(W_k) \tag{5}$$

Die optimale Clusterzahl für den gegebenen Datensatz ist das kleinste k, so dass

$$Gap_n(k) \geq Gap_n(k+1) - s_{k+1} \tag{6}$$

gelten muss, wobei s_k der Simulationsfehler ist, welcher aus der Standardabweichung $sd(k)$ von den B Monte Carlo Simulationen berechnet wurde. Für

$$sd(k) = [\frac{1}{B} \sum_b \{log(W_{kb}^*) - \frac{1}{B} \sum_b log(W_{kb}^*)\}^2]^{1/2}$$

ist

$$s_k = \sqrt{1 + \frac{1}{B}} sd(k) \qquad .$$

Mithilfe eines Beispiels soll die bisherige Theorie dargestellt werden (Abbildung (1)). Oben links wurden 3 Clustern zu je 50 Daten aus $N((2,2)', \mathbf{I})$ (rot),

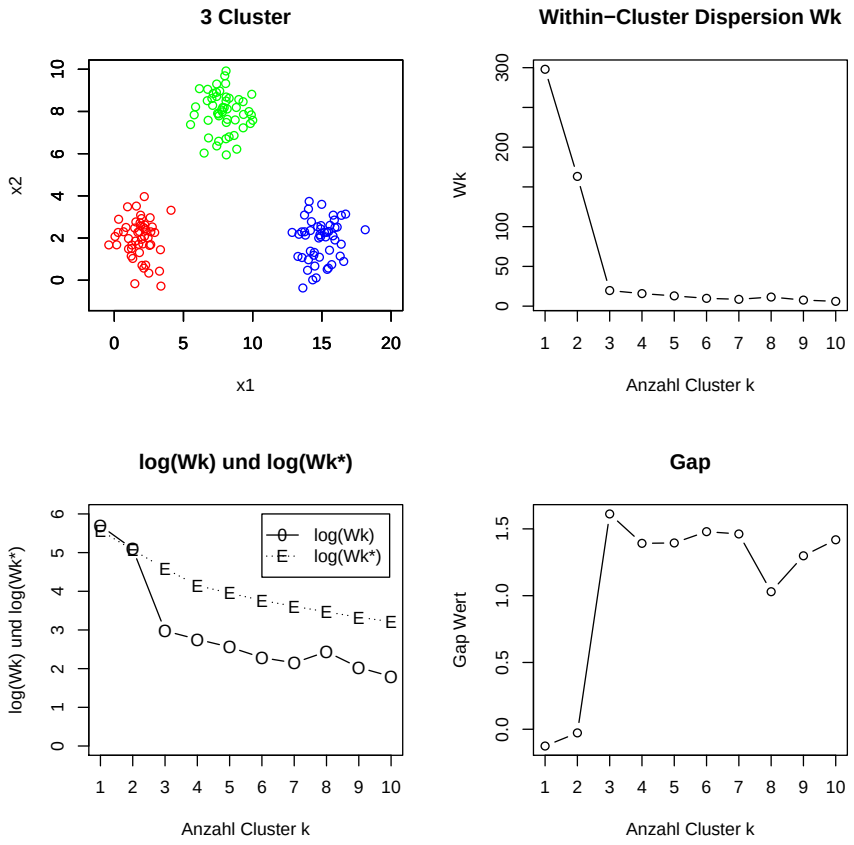

Abbildung 1: Beispiel mit 3 Clustern und dazugehörigem W_k, $log(W_k)$, $log(W_k^*)$ und Gap

$N((8,8)', \mathbf{I})$ (grün) und $N((15,2)', \mathbf{I})$ (rot) erzeugt, wobei $\mathbf{I}$ die Identitätsmatrix ist. Rechts daneben ist W_k in Abhängigkeit von der Clusteranzahl k abgebildet. Es ist deutlich ein „Knick“ bei der Clusterzahl $k = 3$ zu sehen. Somit ist nach dem „Elbow-Kriterium“ eine 3-Cluster Lösung zu wählen. Unten links werden $log(W_k)$ und die dazugehörige Referenzverteilung $log(W_k^*)$ gegen k aufgetragen. Daneben wird die Gap Funktion, die sich aus der Differenz von $log(W_k)$ und $log(W_k^*)$ berechnen lässt, dargestellt. Die Gap Kurve hat ein deutliches Maximum bei $k = 3$. Das Ergebnis der Bedingung, formuliert in Gleichung (6), bestätigt die grafische Auswertung und liefert ebenfalls das Ergebnis $k = 3$.

2.2 Modifizierte Gap Statistik

Die von Mohajer et al. (2010) modifizierte Gap-Statistik verwendet bei der Berechnung W_k anstatt $log(W_k)$. Somit folgt für Gleichung (3)

$$Gap_n^*(k) = E_n^*(W_k) - W_k \qquad , \tag{7}$$

wobei die Referenzverteilung nun durch

$$E_n^*(W_k) = \frac{1}{B} \sum_b W_{kb}^* \tag{8}$$

gegeben ist. Tibshirani et al. (2001) hinterließen in ihrer Arbeit die Anmerkung, dass in dem Fall einer speziellen Gaußmischverteilung $log(W_k)$ als log-likelihood interpretiert werden kann (Scott and Symons (1971)). Für die Berechnung des Maximum-Likelihood Schätzers ist es vom rechnerischen Vorteil, die Maximum-Likelihood Funktion zu Logarithmieren, um aus Produkten Summen zu erhalten. Bei der Berechnung der Gap Statistik jedoch bringt es keinen rechnerischen Vorteil $log(W_k)$ anstatt W_k zu verwenden, da bei diesem Verfahren eben keine Produkte vorkommen.

Mohajer et al. (2010) zeigten, dass das originale Gap_n eine hinreichende Bedingung für ihre modifierte Gap_n^* ist, jedoch nicht umgekehrt. Das heißt: Liefert Gap_n ein Ergebnis mit dem optimalen Werten an Clustern k, ist diese auch in Gap_n^* möglich. Andererseits besteht die Möglichkeit, dass Gap_n^* Ergebnisse liefert, bei dem Gap_n erfolglos blieb.

2.3 Clusterverfahren: hierarchisch agglomeratives average linkage vs. k-means

Wie bereits erwähnt, ist die optimale Clusterzahl k so zu wählen, dass die Gleichung $Gap_n(k) \geq Gap_n(k+1) - s_{k+1}$ bzw. mit Gap_n^* erfüllt ist. Das heißt, es muss bei schrittweiser Erhöhung von k jedes Mal die Gap-Statistik und somit auch W_k berechnet werden. Um W_k jedoch Berechnen zu können, müssen zunächst Cluster gebildet werden.

Mohajer et al. (2010) verwendeten für die Clusterbildung das hierarchische agglomerative average linkage Verfahren. Bei diesem Verfahren werden sukzessiv immer mehr Objekte zu größeren Clustern zusammengefasst. Ausgehend von der größtmöglichsten Clusterzahl, das bedeutet jedes Objekt bildet zu Beginn ein eigenes Cluster, werden Schritt für Schritt die ähnlichsten Objekte bzw. Cluster vereint. Dabei lautet die Zuordnungsregel, dass diejenigen Cluster fusioniert werden, die die kleinste durchschnittliche Distanz zueinander haben. Auf diese Weise entsteht eine Hierarchie der Gruppen, die bei jedem Schritt in die nächste Ebene

durch Vereinigung disjunkter Gruppen enstehen. Die Hierarchie stellt also eine geordnete Darstellung der Schrittweisen Vereinigung der Daten dar (vgl. Eckey et al. (2002) und Mohajer et al. (2010)).

Das in meiner Arbeit verwende k-means Verfahren ist hingegen ein partitionierendes Verfahren. Hierbei werden k zufällige Startpunkte als Clusterzentren festgelegt. Die einzelnen Objekte werden dem Clusterzentrum zugeordnet, zu dem sie die geringste Distanz besitzen. Anschließend wird in jedem der k Cluster der Mittweltwert ermittelt, die nun als neue Clusterzentren gelten. Ist die Distanz eines Elements zu einem neu berechneten Clusterzentrum geringer, so wird diese umgruppiert. Dieser Algorithmus wird solange fortgesetzt, bis entweder eine vorgegebene maximale Iterationszahl an Wiederholungen erreicht wurde oder die Schwerpunkte sich nicht mehr verschieben und somit ein Objekt keinem anderem Cluster mehr zugeteilt werden kann (vgl. Litz (2000) und Eckey et al. (2002)).

3 Anwendung auf simulierte und reale Daten

Im vorigen Kapitel wurde der theoretische Teil der Gap Statistik und der modifizierten Gap Statistik vorgestellt. Ziel dieses Kapitels ist die beiden Verfahren miteinander zu vergleichen, indem sie auf simulierte sowie reale Daten angewendet werden. Mohajer et al. (2010) haben diese Untersuchung mit dem hierarchischen agglomerativen average linkage Verfahren bereits durchgeführt. In dieser Arbeit werden die gleichen Simulationen und Datensätzen mit dem k-means Verfahren durchgeführt.

3.1 Klassische Testdatensätze

Zuerst werden die Gap Funktionen auf folgende zwei bekannten Datensätzen angewendet: den „Fisher's Iris data set“ (Fisher (1963)) und den „Breast Cancer Wisconsin data set“ (Wolberg (1992)). Der Datensatz von Fisher besteht aus 50 Beobachtungen zu je 3 verschiedenen Blumentypen. Jede Beobachtung wird von vier Variablen beschrieben. Wolbergs Brustkrebs Datensatz besteht aus 699 Beobachtungen und neun Variablen. Dieser Datensatz wird in zwei Hauptgruppen un-

Gap Statistik	Anzahl Cluster	
	Iris	Breast
Gap	3	2
Gap*	3	2

Tabelle 1: Ergebnisse Mohajer et al. (2010)

Gap Statistik	Anzahl Cluster	
	Iris	Breast
Gap	3	2
Gap*	1	1

Tabelle 2: Ergebnisse mit k-means Algortihmus

terteilt: 458 Beobachtungen gehören zu den gutmütigen und 241 zu den bösartigen Tumoren.

Tabelle 1 zeigt die Ergebnisse von Mohajer et al. (2010) für Schätzungen der Clusterzahl der Gap Funktionen mit den hierarchisch agglomeratives average linkage Verfahren. Man sieht, dass sowohl Gap als auch Gap* die richtige Anzahl an Cluster in beiden Datensätzen korrekt geschätzt wurden.

Tabelle 2 gibt die von mir erzeugten Ergebnisse mit dem k-means Algorithmus wieder. Die ursprüngliche Gap Statistik liefert die korrekte Anzahl an Clustern von den Datensätzen, was bei der modifizierten Version jedoch nicht der Fall ist. Betrachtet man die Gap* Funktion von „Fisher's Iris data set“ in Abbildung 2, sieht man ein Maximum bei $k = 2$, was auch die falsche Clusteranzahl wiederspiegelt. Die Bedingung nach Gleichung (6) liefert ein geschätztes Optimum für $k = 1$. An der Gap* Funktion des „Breast Cancer Wisconsin data set“ ist nicht einmal ein Optimum zu sehen und Bedingung Gleichung (6) liefert ebenfalls die falsche Clusteranzahl von $k = 1$.

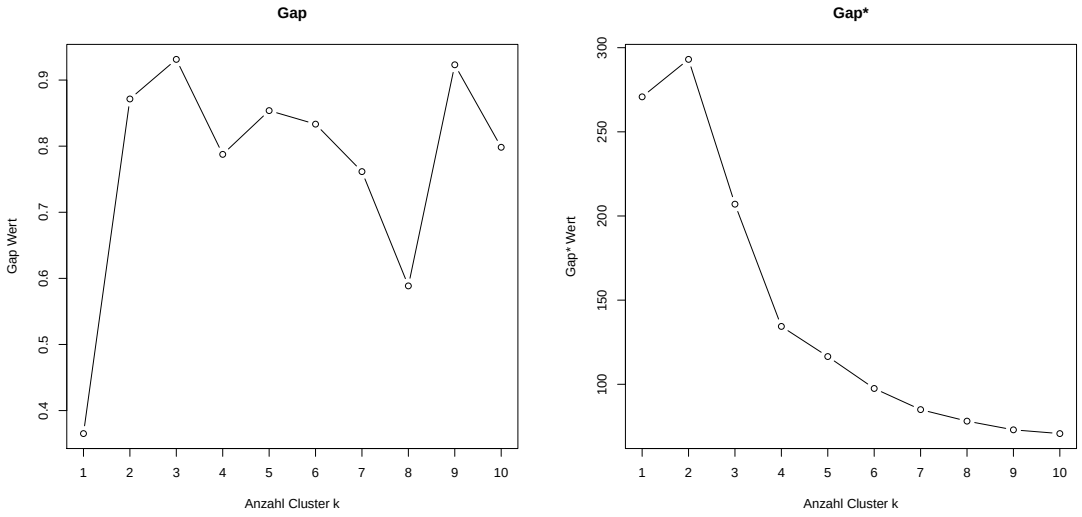

Abbildung 2: Gap und Gap* für Fisher's Iris data set

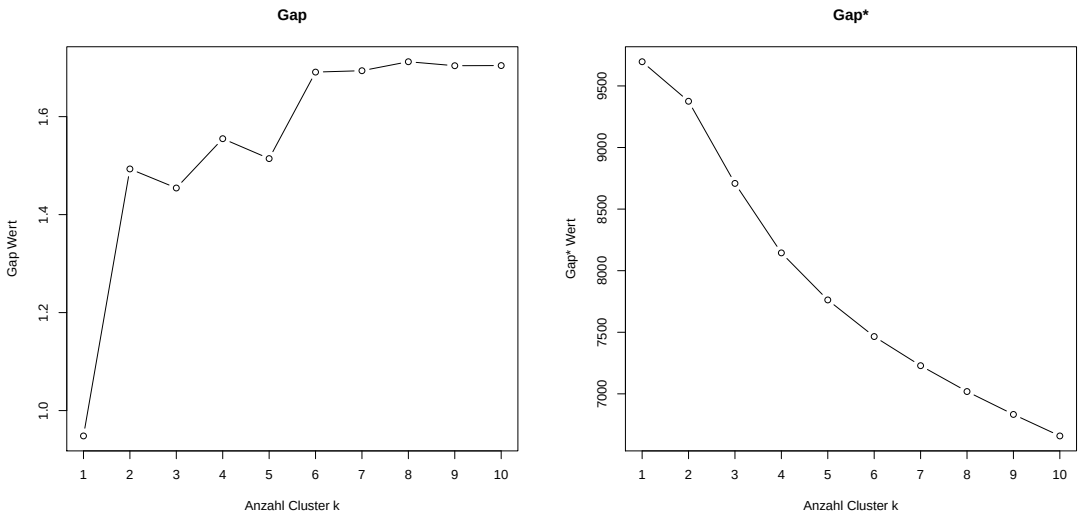

Abbildung 3: Gap und Gap* für Breast Cancer Wisconsin data set

3.2 Überlappende Cluster

Anschließend wurden Daten simuliert, bei denen sich die Cluster überlappen und somit nicht mehr eindeutig voneinander getrennt werden können. Hier wurden insgesamt 1000 Datensätze mit je zwei Cluster erzeugt. Jeder Cluster besteht aus 50 Beobachtungen mit je zwei Variablen, die unabhängig voneinander normalverteilt sind. Der erste Cluster hat den Erwartungswert 0 und die Standardabweichung 1. Für den zweiten Cluster wurden normalverteilte Werte generiert, die die Erwartungswerte $\Delta = 0.5, 1, 1, 5, ..., 5.0$ und die Standardabweichung von 1 besitzen. Für jeden Wert von Δ wurden 100 Datensätze erzeugt. Anschließend wurde untersucht, bei welchen Differenzen der Erwartungswerte die zwei separaten Cluster noch von den Gap Statistiken aufgefasst werden können.

Meine Ergebnisse mit dem k-means Verfahren werden in Abbildung (4) dargestellt. Man sieht, dass das originale Gap zwei überlappende Cluster besser erkennt als Gap*. Beträgt die Differenz der Erwartungswerte mindestens 5, so erkennen beide Verfahren die richtige Anzahl von zwei Clustern. Mohajer et al. (2010) beobachteten mit dem hierarchisch agglomerativen average linkage Verfahren dasselbe. Dieses Ergebnis wurde erwartet, da in dem Artikel von Fridlyand and Dudoit (2002) berichtet wurde, dass Gap die Tendenz hat die Anzahl von Clustern zu überschätzen.

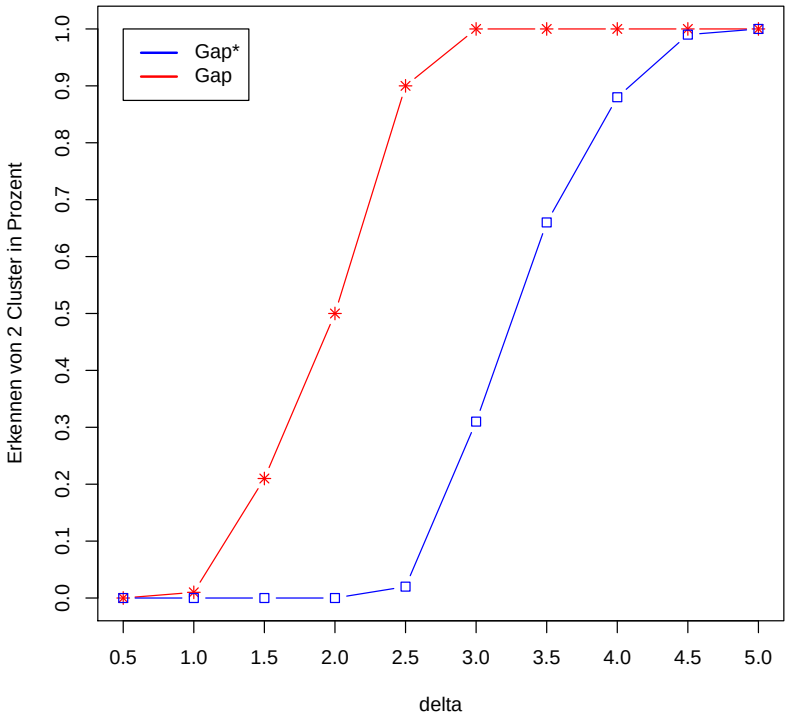

Abbildung 4: Überlappende Cluster Ergebnisse mit k-means

3.3 Ungleichgroße Cluster

Des Weiteren haben Mohajer et al. (2010) untersucht, wie gut die Gap Funktionen Cluster erkennen, wenn diese von unterschiedliche Größe sind. Yin et al. (2008) berichteten in ihrer Arbeit, dass wenn die Beobachtungen in einem Cluster mindestens 6-mal so groß ist als andere Cluster, dies zur Folge hat, dass die Gap Statistik die korrekte Clusterzahl nicht mehr schätzen kann.

Für die Untersuchung wurden zwei Cluster erzeugt, welche zweidimensional Normalverteilt $N(\boldsymbol{\mu}, \mathbf{I})$ und $N(\boldsymbol{\mu}', \mathbf{I})$, wobei $\boldsymbol{\mu}$ und $\boldsymbol{\mu}'$ zwei voneinander unterschiedliche Erwartungswerte sind und $\mathbf{I}$ die Identitätsmatrix.

Sei nun N_1 die Anzahl der Beobachtungen im ersten Cluster und N_2 die Anzahl der Beobachtungen im zweiten Cluster, mit $N_1 = mN_2$ und $n = N_1 + N_2$. Für eine feste Anzahl an Beobachtungen n, ergibt sich bei einer Erhöhung von m die Abnahme von W_1. Folglich sinkt Gap_1, wobei Gap_2 unverändert bleibt. Falls m groß genug wird, wird Gap_1 größer als Gap_2 und die geschätzte Anzahl an Clustern lautet $k = 1$. Die Daten wurden wir folgt erzeugt (vgl. Mohajer et al. (2010) S.9 :

Simulation	N_1	N_2	$m = N_1/N_2$	Gap	Gap*
1	765	765	1	2	1
2	1020	510	2	1	1
3	1224	306	4	1	1
4	1360	170	8	1	1
5	1440	90	16	1	1

Tabelle 3: Übersicht Ungleichgroßer Cluster (siehe Mohajer et al. (2010) S. 8) und Ergebnisse mit k-means

1. Wähle N_1^{max} als die Maximale Anzahl an Objekten im ersten Cluster für alle fünf Datensätze.

2. Wähle N_2^{max} als die Maximale Anzahl an Objekten im zweiten Cluster für alle fünf Datensätze.

3. Erzeuge N_1^{max} Objekte aus einer Bivariaten Normalverteilung $N(\boldsymbol{\mu}, \mathbf{I})$, mit $\boldsymbol{\mu} = (0, 0)$

4. Erzeuge N_2^{max} Objekte aus einer Bivariaten Normalverteilung $N(\boldsymbol{\mu}', \mathbf{I})$, mit $\boldsymbol{\mu}' = (5, 0)$

5. Wähle für jeden Datensatz die ersten N_1 Elemente von N_1^{max} entsprechend der Tabelle (3)

6. Wähle für jeden Datensatz die ersten N_2 Elemente von N_1^{max} entsprechend der Tabelle (3)

Mohajer et al. (2010) zeigten, dass für dieses Beispiel, das originale Gap bei $m < 6$ und die modifizierte Gap* bei $m < 2$ noch die richtige Schätzung von 2 Clustern abgibt. Tabelle (3) liefert eine Übersicht der fünf Datensätzen mit den steigendem m und den dazugehörigen Clustergrößen. Zusätzlich sind daneben die geschätzte Anzahl der Clustern mit dem k-means Verfahren angegeben. Sobald also ein Cluster doppelt so groß ist wie der andere, können weder Gap noch Gap* zwei Cluster erkennen. Demnach konnte nicht gezeigt werden, dass Gap bei $m < 6$ und Gap* bei $m < 2$ noch zwei separate Cluster auffassen.

Die Abweichung kann unter anderem darauf zurückgeführt werden, dass k-means keine optimale Clusterbildung vornimmt. Durch die zufällige Wahl der Clusterzentren können natürliche Cluster getrennt oder zwei unterschiedliche Cluster als eines zusammengefügt werden. K-means liefert also je nach Wahl der Startpositionen nur lokal optimale und keine global optimalen Lösungen(Eckey et al. (2002)). Folglich ist W_k im Gegensatz zum hierarchisch agglomerativen average

linkage Verfahren nicht für jedes k eindeutig definiert. Diese Folgerung ist letztendlich der Grund dafür, dass falsche Schätzungen der Clusteranzahl abgegeben werden..

3.4 Dynamische kontrastmittelbasierte Magnetresonanz-Bilder (DCE-MRI)

Zum Schluss werden die Gap Funktionen auf sieben realen Datensätze der Dynamische kontrastmittelbasierte Magnetresonanz-Bilder (DCE-MRI) von Brustkrebs Tumoren angewendet(German Cancer Research Center (2004)). Für jeden Datensatz wurden 6.9 Minuten lang in einem Abstand von 3.25s ein ausgewählter Schnitt eines Tumors mit Dicke $TH = 6mm$ und Sichtfeld $FOV = 320mm \times 320mm$ gemessen. Jedes Voxel in einem Datensatz wird durch eine Signal-Zeit Kurve der Längte $T = 128$ beschrieben während durch den Tumor ein Kontrastmittel fließt. Diese Kurven liefern wertvolle Informationen über den Blutkreislauf und die Durchlässigkeit des Tumorgewebes. Für die Medizin ist es wichtig Voxel zu finden, die eine Ähnlichkeit in den Signal-Zeit Kurven aufweisen(vgl. Brix et al. (2004) und Mohajer et al. (2010)).

Aus diesem Grund werden die Gap Statistiken auf die DCE-MRI Daten angewendet. Die Stichproben bestehen aus den Signalkurven der Voxel, die durch 128 Variablen, hier 128 Zeitpunkte, beschrieben werden. Des Weiteren stammen die Tumore in allen sieben Datensätzen bzw. Bildern von der gleichen Art. Tabelle (4) gibt die Ergebnisse von Mohajer et al. (2010) sowie meine mit dem k-means Verfahren wieder. Gap* konnte mit dem hierarchisch agglomerativen average linkage Verfahren in fünf der sieben Datensätzen eine Schätzung von fünf Clustern, hier Regionen abgeben. Das Gap hingegen erzielte keine konsistente Schätzung für die Anzahl von Clustern.

Mit k-means konnte Gap* in sechs von den sieben Datensätzen eine Clusteran-

Datensatz	Anzahl Voxel	Ergebnisse Mohajer et al. (2010)		Ergebnisse mit k-means	
		Gap	Gap*	Gap	Gap*
1	1260	7	7	1	1
2	207	9	5	6	1
3	116	9	5	4	1
4	262	nicht definiert	5	5	1
5	141	11	5	1	1
6	277	nicht definiert	5	3	1
7	151	13	4	5	2

Tabelle 4: DCE-MRI Übersicht und Ergebnisse

zahl von eins entdecken und somit keine Gruppierung der Voxel entdeckt werden konnten. Gap lieferte für fünf Datensätze Schätzungen ungleich eins, die jedoch genau wie bei Mohajer et al. (2010) nicht konsistent waren. Wie dem auch sei gibt es bislang keinerlei Informationen über die Anzahl der Regionen

4 Zusammenfassung

Für die Schätzung von Clustern in einem Datensatz schlugen Tibshirani et al. (2001) die Gap Statistik vor. Die Methodik dahinter ist nicht sehr kompliziert. Es werden dabei die Ergebnisse von Clusteralgorithmen, wie zum Beispiel von hierarchischen oder partitionierenden Verfahren, verwendet. Anschließend werden die Streuungen der Cluster W_k mit dessen Referenzverteilung verglichen. Die optimale Clusteranzahl ist gegeben, wenn die Differenz von $log(W_k)$ mit der dazugehörigen Referenzverteilung maximal ist. Obwohl die Gap Statistik, laut Tibshirani et al. (2001), besser ist als andere Verfahren, gibt es einige Fälle bei denen die Schätzungen der Cluster fehlerhaft sind. Yin et al. (2008) berichteten, dass falls ein Cluster mindestens 6-mal so groß ist wie ein anderer Cluster, die Gap Statistik versagt. Auch teilten Fridlyand and Dudoit (2002) mit, dass Gap die Tendenz besitzt die Anzahl an Clustern zu überschätzen.

Aus diesem Grund schlugen Mohajer et al. (2010) vor, den Logarithmus bei der Berechnung der Gap Statistik wegzulassen, da dieser im Gegensatz zur Maximum-Likelihood Schätzung keinen rechnerischen Vorteil bringt. Sie zeigten, dass Gap eine hinreichende Bedingung für ihre modifizierte Gap* ist, jedoch nicht umgekehrt. Allerdings kann Gap* Ergebnisse liefern, bei der Gap scheitert.

Bei den klassischen Testdatensätzen von „Fisher's Iris data set“ (Fisher (1963)) und „Breast Cancer Wisconsin data set“ (Wolberg (1992)) konnte mit dem k-means Algortihmus nur die ursprüngliche Gap Statistik die richtige Anzahl beider Datensätzen bestimmen. Verwendet man hingegen das hierarchisch agglomerative average linkage Verfahren, können Gap sowie Gap* die richtige Gruppierung, für beide Datensätze, durchführen (vgl.Mohajer et al. (2010)).

Bei der Simulation mit Clustern, die sich überlappen, kamen Mohajer et al. (2010) zu dem Ergebnis, dass das originale Gap eine bessere Arbeit verrichtet als die modifizierte Version Gap*, welche auf die Überschätzung von Gap zurückzuführen ist. Das Ergebnis konnte in meiner Arbeit bestätigt werden.

In der zweiten Simulation wurde untersucht, bis zu welchem Verhältnis der Clustergrößen die Gap Funktionen zwei separate Cluster erkennen können. Anhand des Beispiels von Mohajer et al. (2010), das im Abschnitt (3.3) vorgestellt wurde, wurde die Aussage von Yin et al. (2008) untersucht. Diese berichtete, dass falls ein Cluster mindestens 6-mal so groß ist wie ein anderer Cluster, die Gap Statistik nicht mehr richtig schätzen kann. Dabei gilt für diese Simulation, dass die Gap

Statistik Cluster unterscheiden kann, solange ein Cluster nicht mindestens 6-mal so groß ist wie ein Anderer. Ferner prognostizierten sie, dass Gap* Cluster nicht mehr voneinander unterscheiden kann, wenn ein Cluster mindestens doppelt so viele Objekte besitzt wie ein Anderer. Mohajer et al. (2010) konnten nach der Durchführung der Simulation, ihre Schätzungen bestätigen. Dieses Ergebnis kann nach meiner Durchführung mit k-means nicht bestätigt werden. Gap konnte in dem Fall, dass ein Cluster doppelt so groß ist wie ein Anderer, die Cluster nicht mehr voneinander trennen. Gap* konnte bereits bei gleichgroßen Clustern keine Gruppierung durchführen.

Spätestens ab diesem Zeitpunkt wurde klar, dass das k-means Verfahren nicht für die Gap bzw. Gap* Statistik hergenommen werden sollte. Das partitionierende k-means Verfahren wählt zu Beginn zufällige Clusterzentren aus, denen anschließend Objekte zugeteilt werden, welche die geringste Distanz zu den Zentren aufweisen. Abhängig von der Wahl der Startzentren, können unterschiedliche Ergebnisse entstehen. Dies hat zur Folge, dass im Gegensatz zum hierarchisch agglomerativen average linkage Verfahren, W_k nun nicht mehr für jedes k eindeutig definiert ist und somit auch die Gap Funktionen invariant werden. Diese Invarianz ist letztendlich verantwortlich für die falsche Schätzung der Clusteranzahl.

Nichts desto trotz, wurden zum Abschluss noch die Gap Statistiken an realen Datensätzen angewendet. Sieben Datensätze der dynamischen kontrastmittelbasierten Magnetresonanz-Bildern (DCE-MRI) von Brustkrebs - Tumoren wurden verwendet. Die Anzahl der Cluster war hierbei nicht bekannt. Mohajer et al. (2010) gaben mit Gap* eine Schätzung von fünf Clustern ab. Das originale Gap konnte dabei keine eindeutige Clusteranzahl festlegen. Die Schätzungen mit k-means lauten bei Gap* in sechs von sieben Fällen eins. Somit konnten mit Gap* die Daten nicht in mehrere Cluster aufgeteilt werden. Gap hingegen erzielte Schätzungen ungleich eins, die jedoch wie bei Mohajer et al. (2010) variieren.

5 Anhang

Beispiel (1) mit Gap*

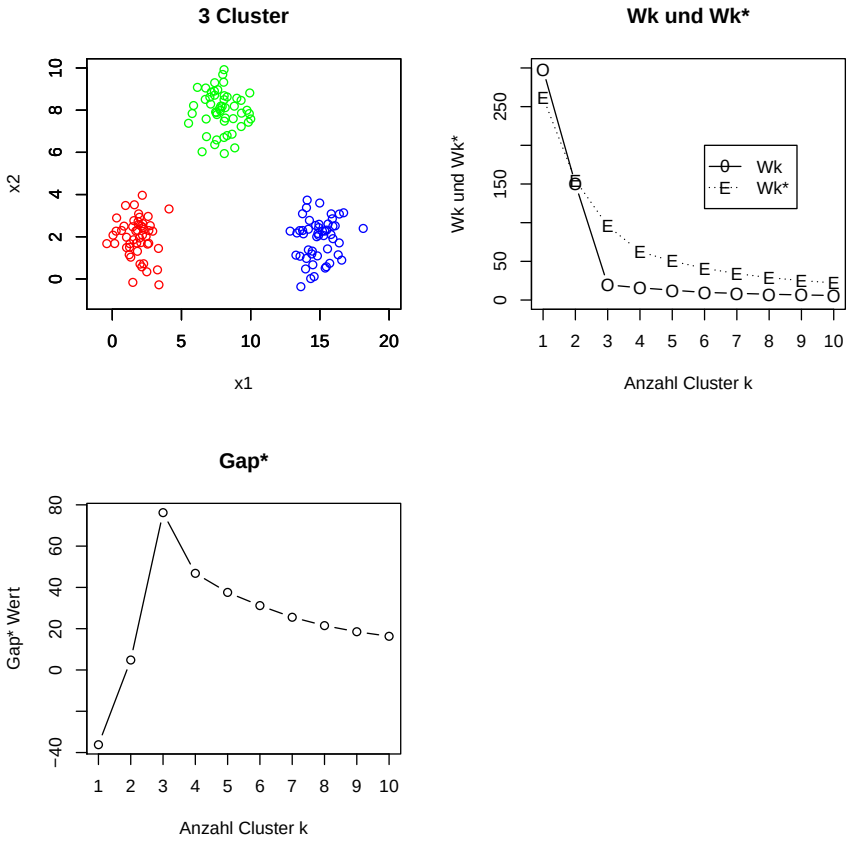

Abbildung 5: Gap* für Beispiel mit 3 Clustern und dazugehörigem W_k, W_k^* und Gap*

Grafiken für „Fisher's Iris data set“ mit Gap

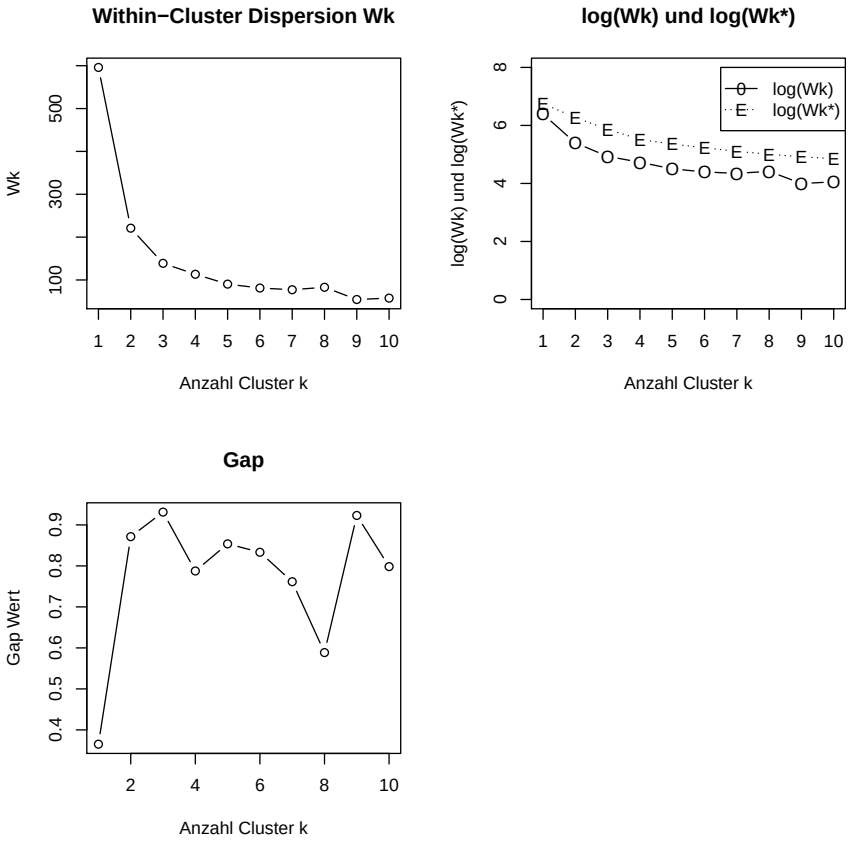

Abbildung 6: „Fisher's Iris data set“ mit dazugehörigem W_k, $log(W_k)$, $log(W_k^*)$ und Gap

Grafiken für „Fisher's Iris data set“ mit Gap*

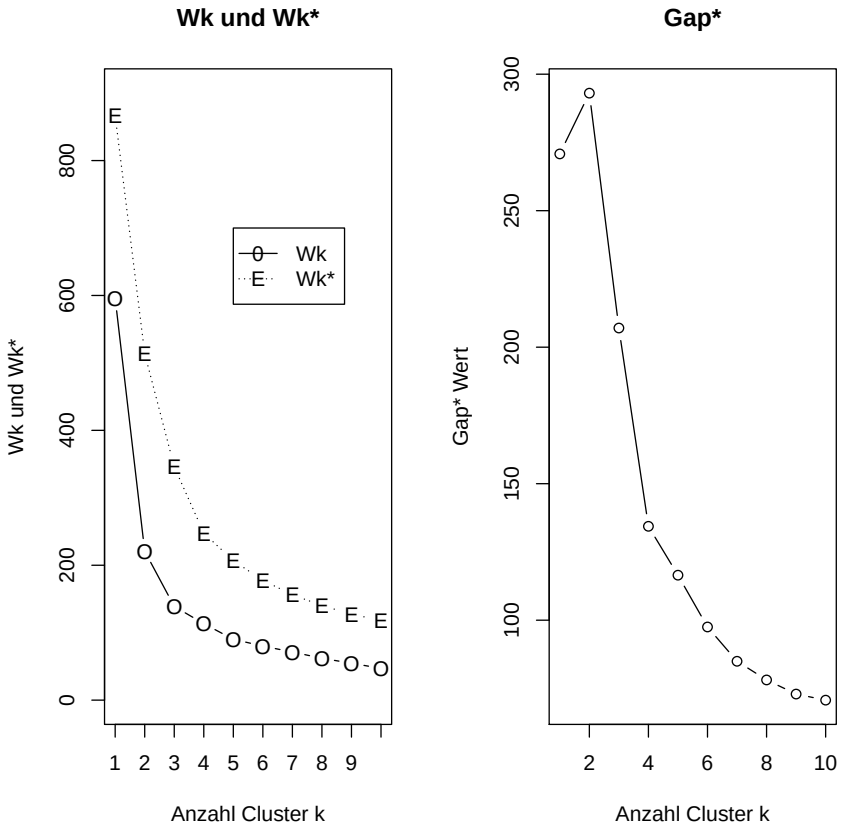

Abbildung 7: „Fisher's Iris data set“ mit dazugehörigem W_k, W_k^* und Gap*

Grafiken für „Breast Cancer Wisconsin data set“ mit Gap

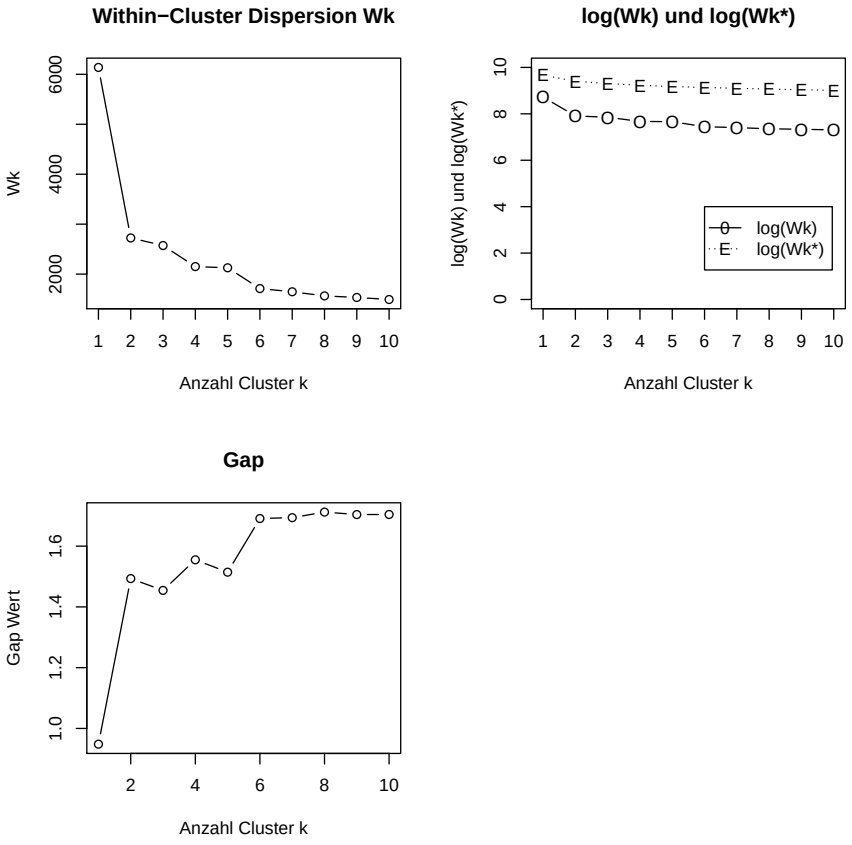

Abbildung 8: „Breast Cancer Wisconsin data set“ mit dazugehörigem W_k, $log(W_k)$, $log(W_k^*)$ und Gap

Grafiken für „Breast Cancer Wisconsin data set“ mit Gap*

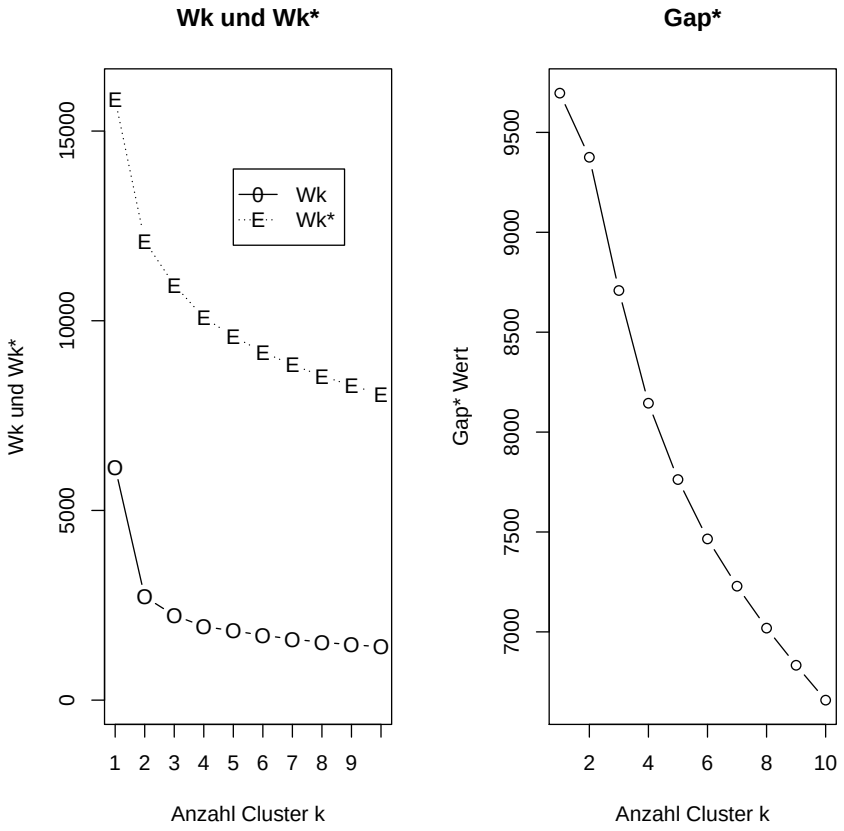

Abbildung 9: „Breast Cancer Wisconsin data set“ mit dazugehörigem W_k, W_k^* und Gap*

Grafiken für ungleichgroße Cluster (m=1) mit Gap

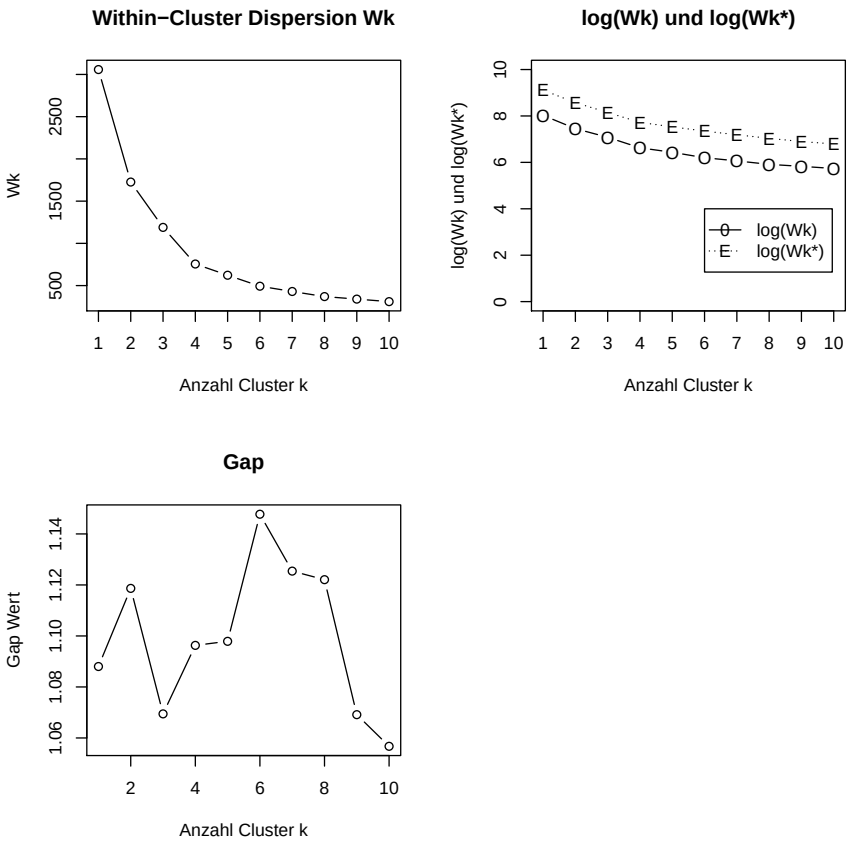

Abbildung 10: ungleichgroße Cluster: m=1 mit dazugehörigem W_k, $log(W_k)$, $log(W_k^*)$ und Gap

Grafiken für ungleichgroße Cluster (m=1) mit Gap*

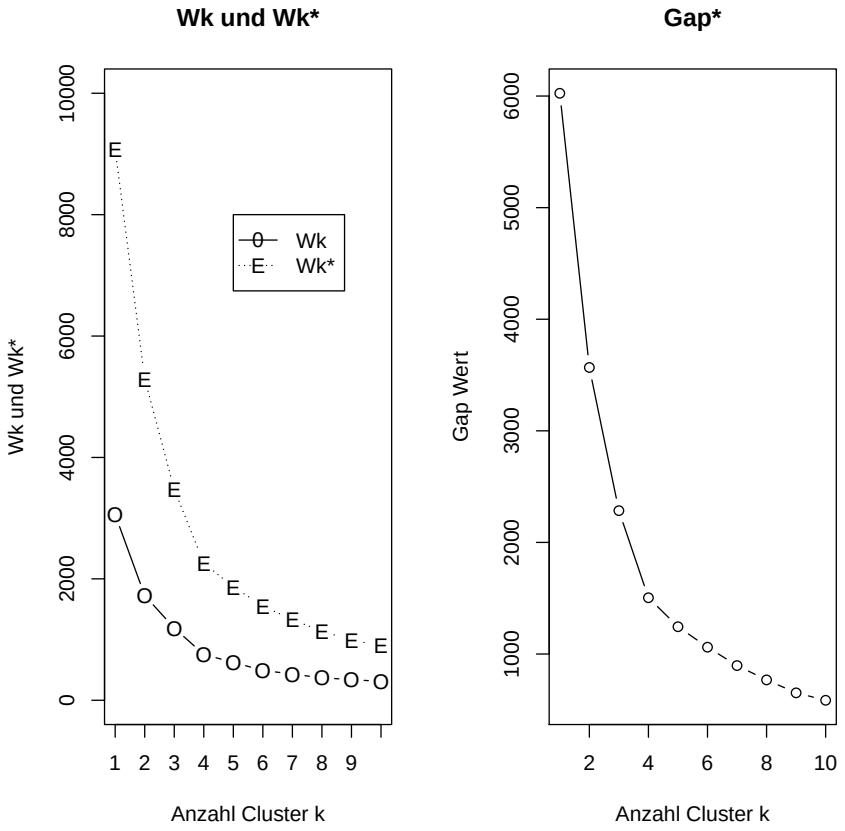

Abbildung 11: ungleichgroße Cluster: m=1 mit dazugehörigem W_k, W_k^* und Gap*

Grafiken für ungleichgroße Cluster (m=2) mit Gap

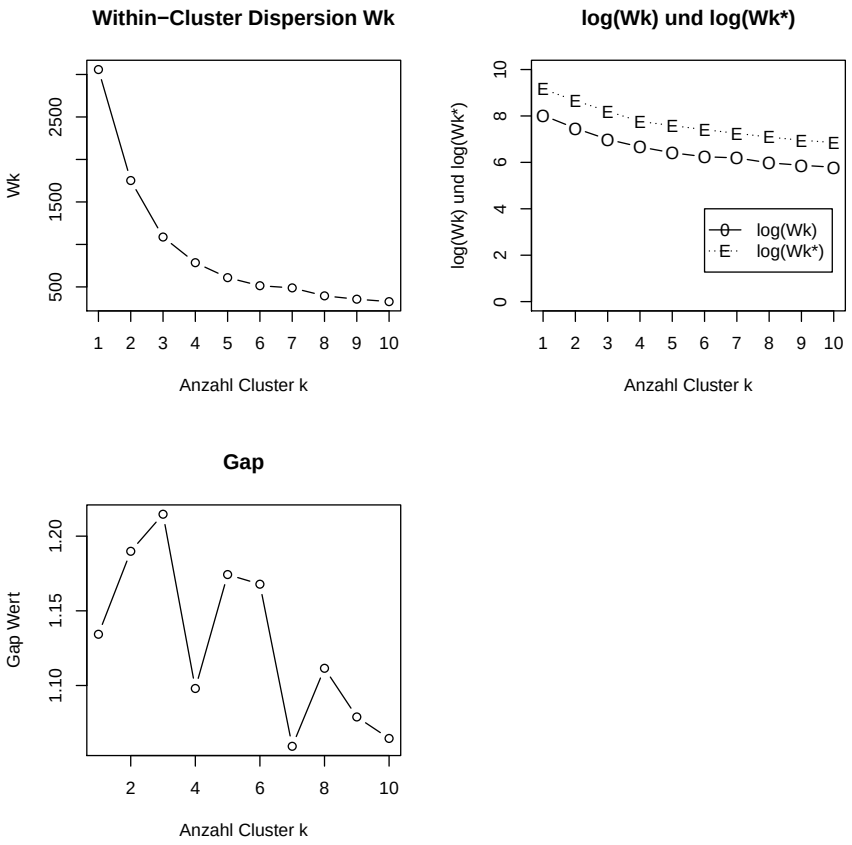

Abbildung 12: ungleichgroße Cluster: m=2 mit dazugehörigem W_k, $log(W_k)$, $log(W_k^*)$ und Gap

Grafiken für ungleichgroße Cluster (m=2) mit Gap*

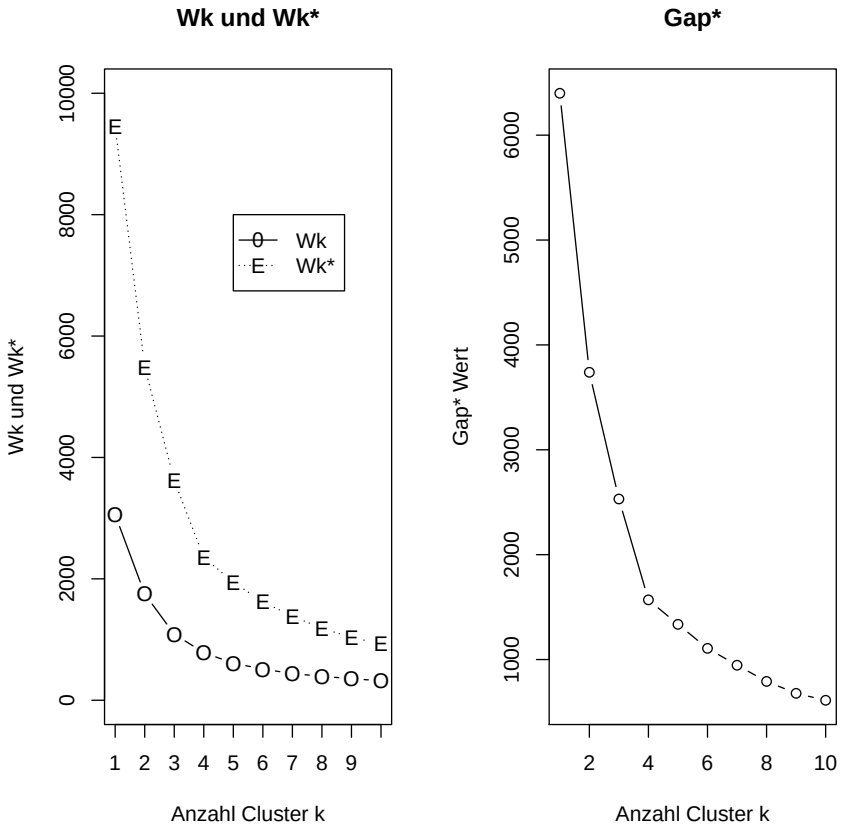

Abbildung 13: ungleichgroße Cluster: m=2 mit dazugehörigem W_k, W_k^* und Gap*

Grafiken für DCE-MRI: Datensatz 1 mit Gap

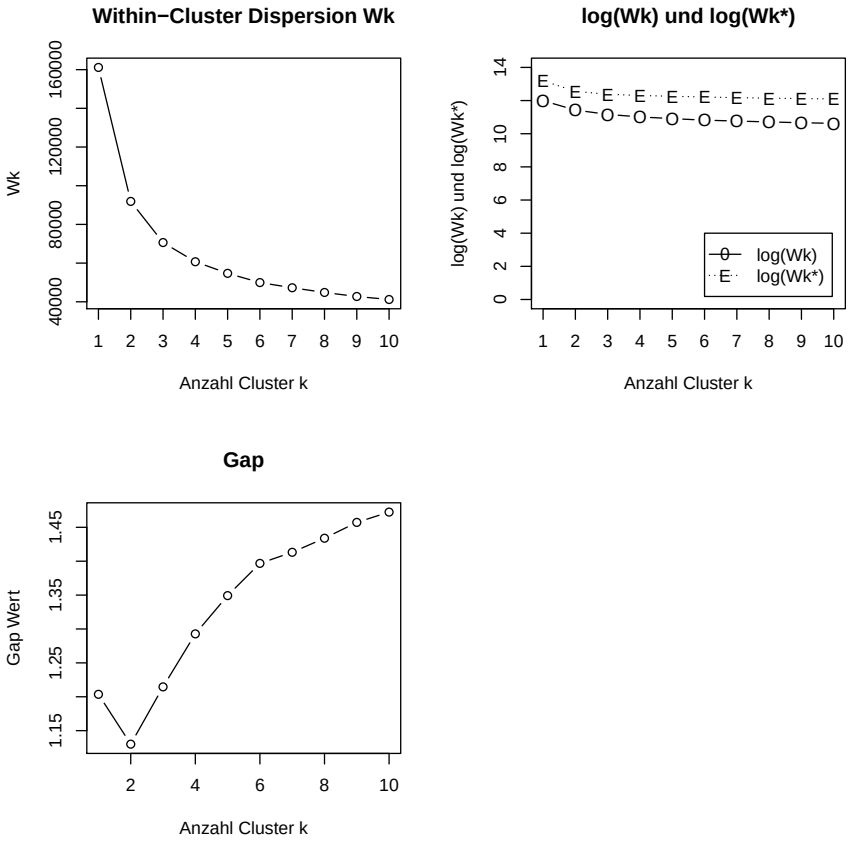

Abbildung 14: DCE-MRI: Datensatz 1 mit dazugehörigem W_k, $log(W_k)$, $log(W_k^*)$ und Gap

Grafiken für DCE-MRI: Datensatz 1 mit Gap*

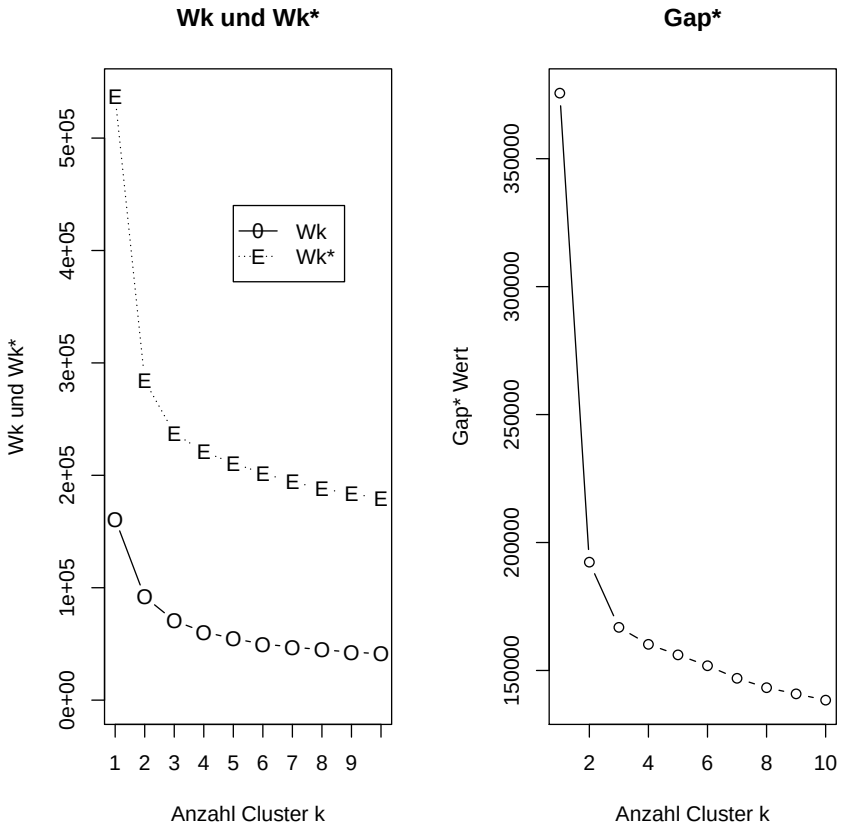

Abbildung 15: DCE-MRI: Datensatz 1 mit dazugehörigem W_k, W_k^* und Gap*

Grafiken für DCE-MRI: Datensatz 4 mit Gap

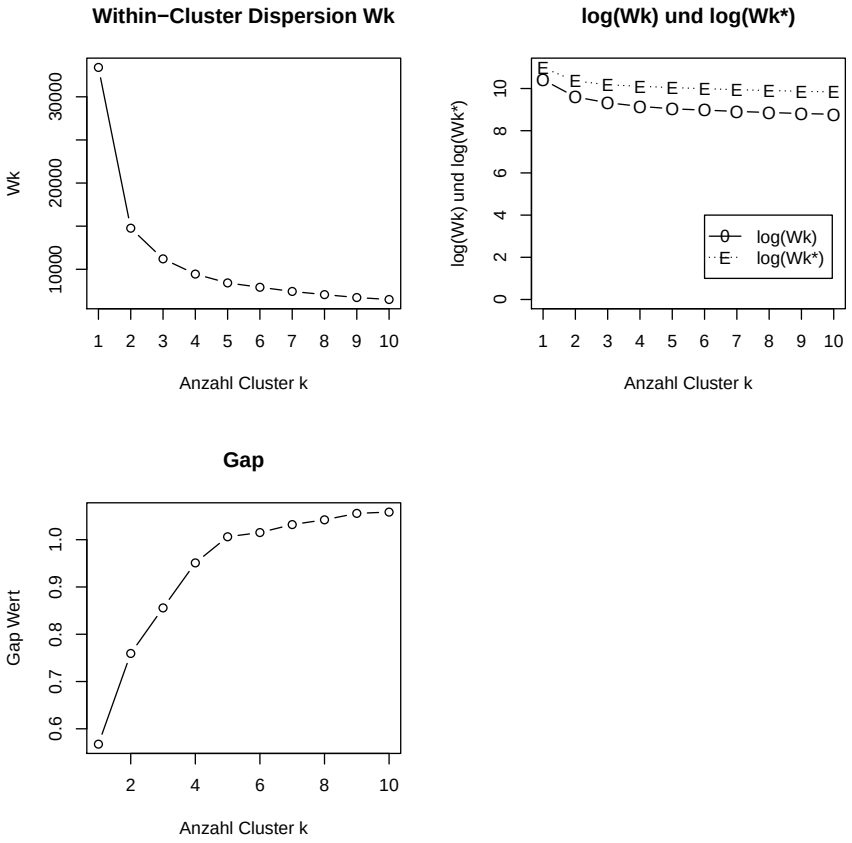

Abbildung 16: DCE-MRI: Datensatz 1 mit dazugehörigem W_k, $log(W_k)$, $log(W_k^*)$ und Gap

Grafiken für DCE-MRI: Datensatz 4 mit Gap*

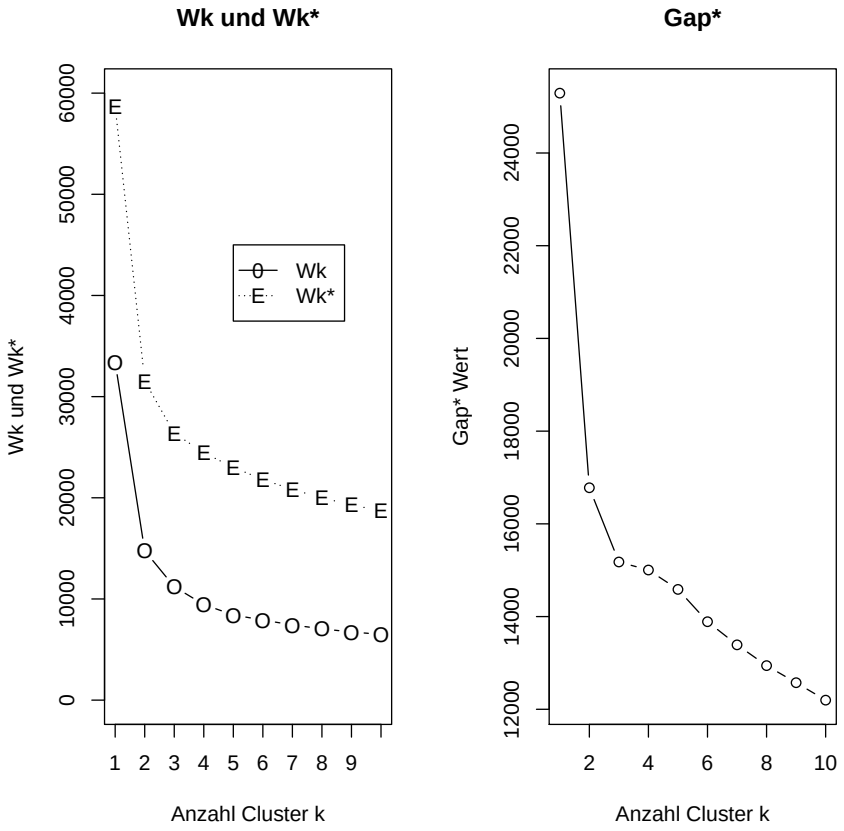

Abbildung 17: DCE-MRI: Datensatz 4 mit dazugehörigem W_k, W_k^* und Gap*

Literatur

Backhaus, K., Erichson, B., Plinke, W. and Weiber, R. (2008). *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*, Springer-Lehrbuch, Springer.

Brix, G., Kiessling, F., Lucht, R., Darai, S., Wasser, K., Delorme, S. and Griebel, J. (2004). Microcirculation and microvasculature in breast tumors: Pharmacokinetic analysis of dynamic mr image series, *Magnetic Resonance in Medicine* **52**: 420–429.

Eckey, H.-F., Kosfeld, R. and Rengers, M. (2002). *Multivariate Statistik*, Gabler.

Fahrmeir, L., Hamerle, A. and Tutz, G. (1996). *Multivariate statistische Verfahren*, 2 edn, Walter de Gruyter, Berlin, New York.

Fisher, R. (1963). UCI machine learning repository.
URL: *http://archive.ics.uci.edu/ml/datasets/Iris*

Fridlyand, J. and Dudoit, S. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset, *Genome Biology* **3**(7).

German Cancer Research Center (2004). Research program „innovative diagnosis and terapy“.

Litz, H. P. (2000). *Multivariate statistische Methoden*, Oldenbourg.

Mohajer, M., Englmeier, K.-H. and Schmid, V. J. (2010). A comparison of gap statistic definitions with and without logarithm function.
URL: *http://epub.ub.uni-muenchen.de/11920/*

Scott, A. J. and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria, *Biometrics* .

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a dataset via the gap statistic, *J.R. Statist. Soc. B* **63**: 411–423.

Wolberg, W. (1992). UCI machine learning repository.
URL: *http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)*

Yin, Zheng, Zhou, Xiaobo, Bakal, Chris, Li, Fuhai, Sun, Youxian, Perrimon, Norbert, Wong and Stephen (2008). Using iterative cluster merging with improved gap statistics to perform online phenotype discovery in the context of high-throughput rnai screens, *BMC Bioinformatics* **9**(1): 264.
URL: *http://www.biomedcentral.com/1471-2105/9/264*