Masterarbeit

Vergleich von Methoden zur Auswahl von Beobachtungen bei Regression mit fehlenden Y-Werten

Vorgelegt am Lehrstuhl für Statistik

an der Technischen Universität Dortmund

von Niels Lategahn

Dortmund, den 24.03.2016

bei Prof. Dr. Katja Ickstadt

Inhaltsverzeichnis

1 Einleitung

In vielen Bereichen der Wissenschaft werden Daten anhand von linearen Modellen erklärt. Dabei werden Zusammenhänge zwischen unterschiedlichen Variablen genauer untersucht und geschätzt. Mit solchen Modellen können Vorhersagen und Aussagen über die Größen der Einflüsse einzelner Werte auf die abhängige Variable getroffen werden. Ein statistisches Verfahren, das dazu häufig genutzt wird, ist das Regressionsverfahren. Hierbei wird ein Modell aufgestellt und anschließend anhand von vorher erhobenen Daten geschätzt. Eine umfangreiche Datengrundlage ist für die genaue Bestimmung der Faktoren des Modells hilfreich.

In vielen Fällen kann jedoch die Erhebung solcher Daten, aufgrund vieler Faktoren, sehr kostspielig sein. Wird etwa die Probe eines zu untersuchenden Gegenstands bei der Erhebung zerstört oder verbraucht, wie etwa bei Crash-Tests in der Autoindustrie, kann es bei umfangreichen Erhebungen zu sehr hohen Kosten kommen. Ebenfalls ist aus ethischen Gründen eine Erhebung mit hoher Stichprobenzahl etwa bei Tierversuchen problematisch. Eine vollständige Datenerhebung ist daher in diesen Fällen nicht zu empfehlen bzw. nicht möglich. Die Auswahl von wichtigen Datenpunkten bei begrenztem Stichprobenumfang ist daher von besonderer Bedeutung.

In dieser Masterarbeit werden daher zunächst Einflussfaktoren, die auf ein Modell wirken, erläutert und aus diesen theoretischen Ansätzen mehrere Methoden zur Auswahl geeigneter Datenpunkte, aus der Gesamtzahl möglicher Datenpunkte, vorgestellt und entwickelt. Damit soll eine möglichst genaue Annäherung an das lineare Modell, das mit sämtlichen Daten geschätzt wird, erreichen. Anschließend werden die Verfahren weitergehend untersucht und die Ergebnisse unter verschiedenen Voraussetzungen verglichen. Damit die benötigte Stichprobengröße und damit die benötigte Menge an Datenpunkte, die erhoben werden sollen, eingeschränkt wird und so die Kosten einer Erhebung gesenkt werden, werden bei der Auswahl zunächst nur die mögliche Datenpunkte der unabhängigen Variablen, verwendet. Die Anzahl der im Anschluss erhobenen abhängigen Werte werden dabei auf einen feste Stichprobengröße gesetzt und nach der Wahl

der Teilstichprobe verwendet, um sich den Vorhersagen und Modellen, die man mit der gesamten Datenmenge mit der Datenauswahl möglichst gut annähern zu können.

In Kapitel 2 wird zunächst die Problemstellung weiter ausgeführt. Kapitel 3 stellt daraufhin das lineare Modell und mögliche Einflüsse auf dieses dar und erläutert einige im weiteren Verlauf verwendete Verfahren. Die daraus abgeleiteten Methoden und Algorithmen werden dann im folgenden Kapitel 4 vorgestellt und es wird weiter eine Gewichtung von Beobachtungen eingeführt. Zudem werden in diesem Abschnitt zwei Kriterien zur Bewertung der Ergebnisse vorgestellt. Die in der Arbeit zugrundeliegende Simulationsstudie mit einzelnen verschiedenen Einstellungen und Grundlagen wird eingehend in Kapitel 5 erklärt und es wird zudem ein Parameter zur Beurteilung der Schwierigkeit einer Auswahl an Beobachtungen eingeführt. Die Ergebnisse der Untersuchungen und der Vergleich der Methoden folgen dann in Kapitel 6. Das letzte Kapitel 7 fasst diese dann zusammen und gibt einen Ausblick auf weitere mögliche Forschungsansätze.

2 Problemstellung

Bei der Erhebung von großen Datenmengen kann es in vielen Fällen zu hohen Kosten kommen. Oftmals ist eine vollständige Erhebung aller möglichen Datenpunkte nicht zu bewältigen. Daher ist häufig eine Begrenzung der Stichprobengröße aufgrund dieser Kosten notwendig. Eine Auswahl von Punkten, aus einer Vielzahl möglicher Punkte, bilden eine Teilmenge. Ziel dieser Auswahl ist es, eine möglichst gute Annäherung an das lineare Modell zu erreichen, das durch sämtliche Datenpunkte geschätzt wird. Diese Teilstichprobe repräsentiert die Gesamtdaten und wird Kernmenge genannt. Dabei ist zu beachten, dass bei den Verfahren zunächst nur die Variablen, die unabhängig von den zu erhebenden Daten sind, für die Wahl der Kernmenge hinzugezogen werden, die abhängige zu erhebende Variable jedoch zunächst nicht vorliegt. Die unabhängigen

Variablen können etwa den Einstellungsmöglichkeiten bei einem Test entsprechen. So könnten bei Crash-Tests von Fahrzeugen, die Geschwindigkeit oder die Beladung mögliche, variierende Einstellung sein. Das Prinzip geeigneter Kernmengen zur Repräsentation von Daten bei der linearen Regression wurde bereits in dem Seminar „Fallstudien 2“ an der TU Dortmund im Wintersemester 2014/15 behandelt und wird in dieser Masterarbeit fortgeführt und vertieft. Dabei werden unterschiedliche Auswahlverfahren und Verzerrungen, die eine Abweichung von dem Modell in Datensätzen darstellen und die Auswirkungen auf das lineare Modell eingehend untersucht.

3 Methoden und Theorie

In diesem Kapitel wird zunächst das lineare Modell und die allgemeinen Notationen vorgestellt. Daraufhin wird nähergehend auf Einflüsse, die auf ein solches Modell einwirken können, erläutert und anschließend zwei Kostenkriterien für die Auswahl der Datenpunkte vorgestellt, anhand derer die Kosten bzw. die Güte einer Kernmenge zu erfassen sind, um so die Verfahren vergleichbar zu machen. Im Anschluss werden dann die zu vergleichenden Auswahlmethoden aufgrund der theoretischen Überlegungen motiviert und vorgestellt. Diese wurden zum Teil im Verlaufe dieser Arbeit entwickelt und werden daher erstmals in einer wissenschaftlichen Arbeit untersucht.

3.1 Lineare Modelle

Das lineare Modell ist die Untersuchungsgrundlage dieser Masterarbeit und wird in Fahrmeir et al. (1996) beschrieben. Dabei wird eine abhängige Zielvariable (Y-Variable) durch ein oder mehrere unabhängige Einflussvariablen (X- Variablen) erklärt und modelliert. Es wird ein Zusammenhang dieser Variablen vorausgesetzt. Die Werte der d Einflussfaktoren werden zunächst in einer sogenannten Designmatrix $X \in R^{n\times(d+1)}$

folgendermaßen zusammengefasst. Die erste Spalte der Matrix besteht aus einem Einservektor, um einen y-Abschnitt zu berücksichtigen. Die weiteren Spalten der Matrix enthalten die Werte der Einflüsse der einzelnen unabhängigen Variablen. Zusätzlich liegen die Beobachtungen der abhängigen Variable, die modelliert werden soll, in einem Vektor $y \in R^n$. n entspricht dabei der Stichprobengröße. Ziel des Verfahrens ist es, den Parametervektor $\beta \in R^{(d+1)}$ folgenden Modells mit Hilfe der vorliegenden Daten zu schätzen:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_d x_d + e = X\beta + e.$$

Dabei entspricht $e \in R^n$ einen um Null unabhängig, identisch normalverteilten Fehlervektor, was bei der Modellierung vorausgesetzt wird. Die x_i sind die jeweiligen Spalten der Designmatrix mit den Werten der i-ten Einflussvariablen.

Bei der Regression wird der geschätzte Vektor der Einflussgrößen mit Hilfe der Kleinsten-Quadrate-Methode ermittelt. Diese minimiert die Summe der quadratischen Abweichungen der Zielvariablevektor $y \in R^n$ zu dem geschätzten Modell $X\beta$ und somit folgenden Term, in Hinsicht auf β.

$$\hat{\beta} = argmin_\beta((X\beta - y)^T(X\beta - y)).$$

Diese Methode wird in der Wissenschaft häufig genutzt, da sie die beste lineare erwartungstreue Schätzung für das Schätzproblem liefert.

Einfluss einzelner Beobachtungen auf ein lineares Modell

Ist die Stichprobengröße und somit die Anzahl der Beobachtungen, die man für das Schätzen eines linearen Modells zur Verfügung hat, begrenzt, ist es wichtig eine geeignete Datenauswahl aus allen Beobachtungen anhand der unabhängigen Variablen zu treffen. Der Einfluss, den ein oder mehrere Datenpunkte auf das Modell haben, ist bei der Wahl einer solchen Kernmenge von Bedeutung, da sie das Modell entsprechend stark

bzw. geringfügig verändern. Chatterjee et al. (1986) als auch Penã et al. (1995) haben sich genauer mit den Einflüssen von Datenpunkten auf lineare Modelle beschäftigt und haben diese in geeignete Kennzahlen gefasst.

Leverage Scores

In einem linearen Modell haben einige Datenpunkte, abhängig von ihren x-Werten, einen unterschiedlich hohen Einfluss auf die Schätzung. Beobachtungen, die nahe bei den durchschnittlichen Werten liegen haben eine kleinere Bedeutung und sind weniger einflussreich. Punkte, die weit außerhalb liegen, haben eine höhere Hebelwirkung für das Modell und sind entsprechend entscheidender für die Schätzung des Modells und können dies leichter beeinflussen.
Abbildung 1 verdeutlicht dieses Prinzip anhand eines Beispiels im Falle eines Modells mit nur einer Einflussvariablen.

Dabei ist zu erkennen, dass das Modell, das aus allen schwarzen Punkten geschätzt wird deutlich von der Schätzung mit dem zusätzlichen grünen Punkt abweicht, da dieser eine starke Hebelwirkung besitzt.
Mit Hilfe der Levarage Scores kann die Stärke des Einflusses einer Beobachtung anhand ihres x-Wertes ermittelt werden. Sie ergeben sich durch folgende Berechnung und sind die Diagonalelemente der sogenannten Hat-Matrix $X^T(X^TX)^{-1}X$, die zum schätzen der Parameter benutzt wird:

$$p_i = x_i^T(X^TX)^{-1}x_i,$$

wobei p_i den Scorewert der i-ten Beobachtung angibt, X entspricht der Desginmatrix des Modells und x_i der entsprechenden i-ten Spalte dieser Matrix.
Die Berechnung ist ebenfalls durch eine QR-Zerlegung möglich und aus numerischer Sicht häufig sogar sinnvoll. Dies wird in Hoaglin und Welsch (1978) genauer beschrieben.
Ist der Wert des Scores für eine Beobachtung sehr hoch, hat diese einen großen Einfluss.

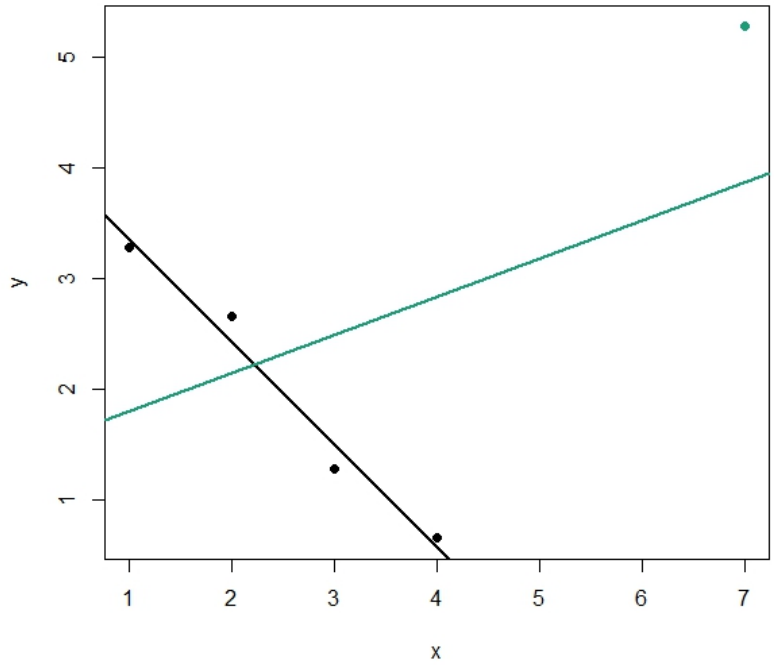

Abbildung 1: Beispiel für geschätzte Modelle mit (grün) und ohne (schwarz) einen Punkt mit hoher Hebelwirkung.

Bei kleineren Werten ist die Hebelwirkung entsprechend gering. Der Vorteil der Leverage Scores ist, dass bei der Berechnung ausschließlich die x-Werte hinzugezogen werden. Somit ist eine Erhebung der y-Werte für diese Bestimmung nicht notwendig. Leverage Scores werden in vielen Veröffentlichungen zur Regressionsanalyse, wie Montgomery et al. (2006), beschrieben.

Cooks Distance

Eine weitere Kennzahl für den Einfluss eines Datenpunktes auf die Schätzung eines linearen Modells ist die Cooks Distance. Diese wird in Chatterjee et al. (1986) vorgestellt. Für den Einfluss, den ein Wert auf das zu schätzdende Modell hat, ist nicht nur die Position im Raum und somit die Hebelwirkung der unabhängigen Variablen

von Bedeutung. Es kann vorkommen, dass die y-Werte der Beobachtungen stark von dem geschätzten Modell abweichen und diese Abweichung, auch Residuum genannt, zu einer deutlich unterschiedlichen Schätzung für den Vektor β führt. Je nach Position im Raum der x-Werte können die entsprechenden hohen Residuen zu starken bzw. geringen Änderungen des Schätzwertes führen. Ist die Hebelwirkung einer Beobachtung besonders groß, ist die Höhe der Abweichung bedeutender. Andererseits ist auch bei kleiner Abweichung der Einfluss auf das geschätzte Modell gering. Somit ist nicht nur die Hebelwirkung sondern auch die Abweichung vom Modell für die Schätzung entscheidend. Ein Beispiel für diese Einflüsse ist grafisch in Abbildung 2 dargestellt.

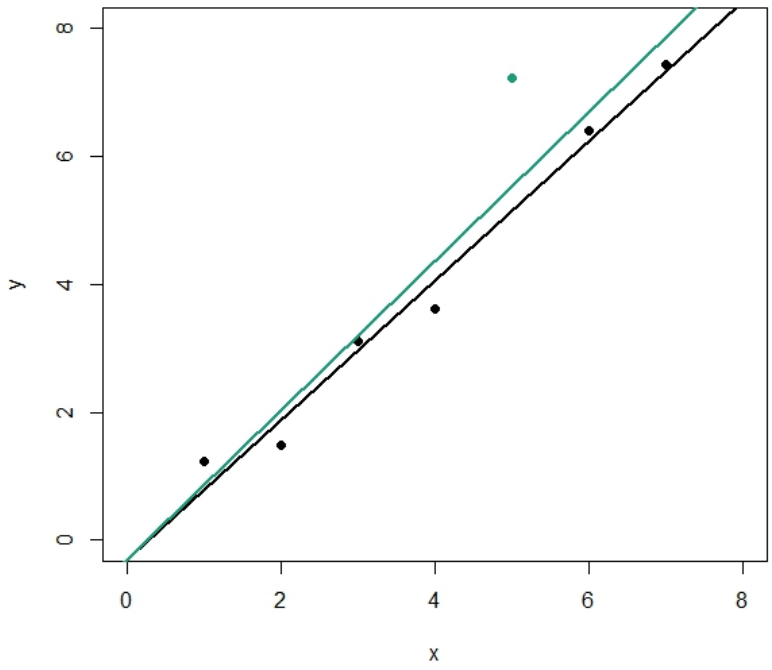

Abbildung 2: Beispiel für einen Punkt (grün) mit geringer Hebelwirkung und hoher Abweichung. Resultierende Schätzung mit diesem Punkt (grün) und ohne diesen Punkt (schwarz)

Dabei ist zu sehen, dass das Modell, das aufgrund aller Beobachtungen (schwarze Punkte) geschätzt wird, sich durch die rechten Beobachtungen mit hohen Hebelwirkungen kaum

verändern würden, da diese nur kleine Abweichungen vom Modell beinhalten. Wird jedoch ein Punkt mit hohem Residuum zusätzlich zur Schätzung des Modells hinzugezogen, ändert sich die Modellschätzung (grüne Gerade) trotz kleinerer Hebelwirkung deutlich. Die Cooks Distance berücksichtigt bei der Bestimmung des Einflusses dieses Phänomen und verbindet so die Hebelwirkung und zusätzlich die Größe der Abweichung eines Punktes vom Modell. Sie berechnet sich über die Formel:

$$C_i = \frac{(\hat{Y} - \hat{Y}_{(i)})^T(\hat{Y} - \hat{Y}_{(i)})}{(d+1)\hat{\sigma}^2} = \frac{(X\hat{\beta} - X\hat{\beta}_{(i)})^T(X\hat{\beta} - X\hat{\beta}_{(i)})}{(d+1)\hat{\sigma}^2}.$$

$\hat{Y}$ entspricht hierbei dem i-ten Vorhersagewert der abhängigen Variable unter Schätzung des Modells mit sämtlichen Beobachtungen, $\hat{Y}_{(i)}$ ist der Wert, der sich ergibt, falls die i-te Beobachtung bei der Vorhersage ausgelassen wird. Ebenso ist $\hat{\beta}$ bzw. $\hat{\beta_{(i)}}$ der Parametervektor der Einflussgrößen, der aus allen bzw. allen außer der i-ten Beobachtungen geschätzt wird. $\hat{\sigma}$ ist die geschätzte Varianz des Modells mit allen Beobachtungen.
Dabei ist zu beachten, dass für die Bestimmung dieses Wertes hier die y-Werte mit einbezogen werden müssen. Die Cooks Distanz wird in der vorliegenden Arbeit zur Verbesserung einer Vorauswahl von Beobachtungen und einer Bewertung für die Schwierigkeit einer Auswahl in bestimmten Situationen benutzt.

Gruppeneinfluss

Bei der Schätzung des linearen Modells kann es dazu kommen, dass zwei oder mehrere Beobachtungspunkte nahe beieinander liegen. Dabei ist es möglich, dass der Einfluss eines der Punkte, z.B durch eine positive Abweichung von geschätzten Modell, durch den Einfluss eines anderen Punktes in der Nähe verstärkt oder aber geschwächt wird. In diesem Falle spricht man von einer sogenannten Gruppeneinfluss, da der Einfluss auf das Modell durch die Werte einer gesamten Gruppe von Daten bestimmt wird (vgl. Chatterjee et al., 1986). Die Grafik 3 zeigt ein Beispiel für solche Einflüsse.

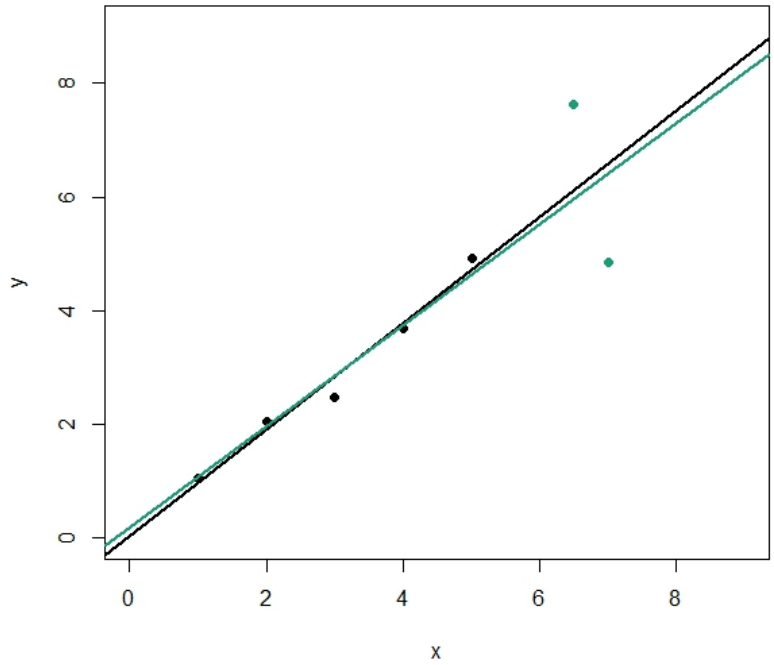

Abbildung 3: Beispiel für Einflüsse von Gruppen von Beobachtungen

Dabei ist zu sehen, dass sich das Modell bei Hinzunahme der zusätzlichen Punkte, (grüne Werte) trotz der relativen hohen Abweichungen und Hebelwirkungen der einzelnen Punkte kaum ändert. Die beiden Beobachtungen wirken entgegen des jeweils andern Punktes und haben daher eine (ausgleichende) geringen Gruppeneinfluss. Dabei würde sich die Schätzung des Modells jeweils stark ändern, falls nur eine dieser Beobachtung für die Schätzung hinzugezogen wird.
Die unterschiedliche Stärke der Einflüsse einzelner oder mehrerer Beobachtungen sind die Ansatzgrundlage und Motivation der später vorgestellten Auswahlmethoden. Des weiteren werden bei den im Kapitel 4 vorgestellten Methoden zusätzlich wie das Clusterverfahren verwendet, die im folgenden Abschnitt erläutert werden.

3.2 Clusterverfahren

Clusterverfahren werden häufig benutzt, um Werte von Objekten bzw. Beobachtungen in Klassen, sogenannte Cluster, zu unterteilen und sie aufgrund ihrer Eigenschaften so zu klassifizieren. Das Ziel dieser Methoden ist es, Gruppen mit möglichst homogenen Objekten zu finden, wobei die Gruppen jeweils heterogen voneinander sind.

k-means-Verfahren

Ein Clusterverfahren ist das k-means-Verfahren, das im Bacher et al. (2010) beschrieben wird. Die Anzahl der zu bestimmenden Gruppen ist dabei mit k Clustern fest vorgegeben. Im ersten Schritt des Verfahrens wird eine Klassifizierung der Beobachtung aufgrund einer Zufallsauswahl getroffen. Aufgrund von Vorinformationen und Untersuchungen kann jedoch auch eine andere Verteilung benutzt werden.
Ist eine solche Startgruppierung gewählt, werden die Mittelwerte, die sogenannten Klassenzentroide, für jede Klasse berechnet. Im zweiten Schritt wird jede der vorliegenden Beobachtungen der Klasse, die dem Zentroid am nächsten liegt, zugeordnet. Die Abstände zu den Klassenmitten werden dabei mit der euklidischen Norm berechnet. Dadurch entsteht eine neue Zuordnung der Daten. Daraufhin werden wiederum die neuen Klassenzentroide ermittelt und der zweite Schritt wird wiederholt. Dieser Vorgang wird wiederholt bis keine Unterschiede mehr in den Zuordnungen, aufgrund der Distanzen zu den Klassenmitten, vorkommen. In diesem Fall ist jede Beobachtung dem Cluster zugeordnet, dessen Zentroid sie am nächsten liegt und der Vorgang ist konvergiert.
Bei der Methode hängen die Ergebnisse von der Startklassifikation ab. Insbesondere kann eine Anwendung des Verfahrens durch die zufällige Startklassifikation zu unterschiedlichen Zuordnungen führen. In einigen Fällen kann zudem der Algorithmus nicht konvergieren, da die Zuordnung nicht eindeutig ist. Eine Beschränkung der Wiederholungen und Iterationsschritten kann hierbei nützlich sein. Deshalb kann es sinnvoll sein,

eine Startklassifikation vorzugeben bzw. eine Klassifizierung aufgrund von vorgegebenen Klassenzentroiden im ersten Schritt zu benutzen.

k-Means ++

Das k-Means++-Verfahren ist ein weiteres Klassifikationsverfahren und nutzt vor dem ersten Iterationsschritt des k-Means ein zufällige, aber nicht gleichverteilte, Auswahl an Clustermittelpunkten. Dabei wird hierbei der erste Wert eines solchen Klassenmittelpunktes zufällig mit gleicher Wahrscheinlichkeit aus den Beobachtungen ausgewählt. Danach wird für die übrige Auswahl der Werte eine Wahrscheinlichkeit benutzt, die proportional zu dem Abstand des jeweiligen Punktes zu den schon gezogenen Clusterzentroiden ist. Der mit dieser Wahrscheinlichkeit gezogene Wert wird wiederum zu der Menge der k Clusterzentren genommen. Der Wahrscheinlichkeitsvektor der nächsten Wahl bestimmt sich proportional zu den kleinsten Abständen der Beobachtungen zu den bereits gewählten Zentroiden. Dieser Vorgang wird wiederholt bis k Klassenmitten ausgewählt sind. Eine Klassifizierung zu den Gruppen wird dann wie im vorherigen Abschnitt bestimmt, indem das k-Means Verfahren weiter angewendet wird.
Diese Initialauswahl der k Clusterzentren führt mit hoher Wahrscheinlichkeit zu einer Auswahl, bei der die Mittelpunkte der Gruppen schon zu Beginn über einen großen Abschnitt des Raumes der Daten verteilt sind. Zudem werden Teile des Raumes, die sehr wenige Beobachtungen enthalten aufgrund ihres Abstandes zu den anderen Werten als eigenes Cluster definiert. Teilräume mit sehr vielen Datenpunkten werden zudem entsprechend ihrer Verteilung häufiger ausgewählt und in mehrere Cluster klassifiziert, da zwar kleine Wahrscheinlichkeiten für das Ziehen eines Punktes vorliegen, die Beobachtungen jedoch zahlreicher sind. Ein weiterer Vorteil dieser Vorauswahl ist die teilweise deutliche Verringerung der notwendigen Iterationsschritte des k-Means und führt somit zu einer geringeren Laufzeit des Algorithmus. Das Verfahren wird in Arthur et al. (2007) vorgestellt und erläutert.

4 Auswahlverfahren

Nachdem die Einflüsse auf ein lineares Modell in den obigen Kapitel dargestellt wurden, werden diese nun als Grundlage für die Auswahl geeigneter Datenpunkte genutzt. Im folgendem Kapitel werden die Methoden zur Auswahl von r Beobachtungen aus einem Datensatz beschrieben. Eine solche Teilstichprobe, die den Datensatz repräsentiert, wird Kernmenge genannt. Die Algorithmen folgen dabei den verschiedenen Ansätzen und werden unterschiedlich motiviert. Einige Ansätze beruhen auf bereits veröffentlichten Arbeiten, andere wurden bisher noch nicht untersucht. Die Prozeduren werden jeweils für die Auswahl einer Kernmenge mit r Beobachtungen angepasst, mit dem Ziel, die Daten im linearen Modell gut repräsentieren zu können. Dabei ist es zum einen wichtig, Daten auszuwählen, die über den ganzen Raum verteilt sind, um die Variation in den unabhängigen Variablen vollständig widerzuspiegeln und so Beobachtungen auszuwählen, die eine starke Hebelwirkung haben. Zum anderen ist es wichtig redundante Daten, die keine neuen Informationen liefern, nicht in die Auswahl einzubeziehen, Gruppeneinflüsse und stark abweichende, „ungewöhnliche“ Daten jedoch zu überprüfen.

4.1 Gewichtungen

In vielen Fällen werden Kernmengen benutzt, bei denen einzelne Datenpunkte viele andere Datenpunkte repräsentieren. Wird etwa eine Beobachtung genutzt, um einen Teil des Beobachtungraumes zu repräsentieren, der besonders viele Beobachtungen enthält, wirkt sich dies auf die Schätzung des linearen Modells aus. Werden nur wenige Beobachtungen repräsentiert, kann der Einfluss wiederum anders sein. Bei der Auswahl der Kernmenge ist daher zu beachten, wie viele Datenpunkte jeweils durch einen Punkt in der Menge repräsentiert werden. Dieser Effekt wird in Abbildung 4 grafisch veranschaulicht.

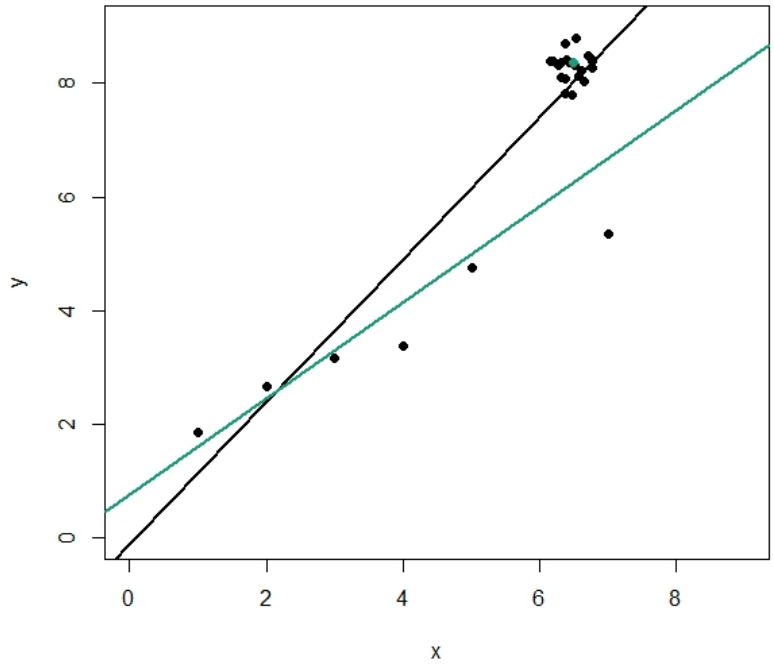

Abbildung 4: Beispiel für eine Schätzung mit allen Beobachtungen (schwarz) und mit Berücksichtigung nur eines einzelnen Punktes aus einem Cluster (grün).

In dieser Grafik ist zu sehen, dass die geschätzten Modelle voneinander abweichen, falls nur ein einzelner Punkt (grüner Punkt) berücksichtigt wird, obwohl er deutlich mehr nahe (ähnliche) Punkte repräsentiert als die übrigen Beobachtungen. Die schwarze Gerade entspricht dem geschätzten Modell mit allen Beobachtungen.

Eine Gewichtung der einzelnen Datenpunkte ist daher sinnvoll und kann im Falle von sehr unterschiedlichen Anzahlen an zu repräsentierenden Daten je Kernmengenpunkt die Ergebnisse der Auswahl verbessern. Ist die Verteilung der Beobachtung jedoch in etwa gleichverteilt sind auch die Gewichte ähnlich. Dies kann jedoch dazu führen, dass einzelne Beobachtungen in der ausgewählten Menge mit besonders hohen Abweichungen und Einflüssen zu starken Einfluss haben, da sie zudem noch vervielfacht werden. Eine Beobachtung in der Nähe, die nicht in die Menge aufgenommen wurde, die Beobachtung jedoch ausgleichen bzw. relativieren kann, würde den Einfluss jedoch verstärken. Eine

Gewichtung führt in diesem Fall zu zu starken Verzerrungen und verringert die Güte der Schätzung. Es werden daher in dieser Arbeit nur Daten gewichtet, bei denen Situation von nicht gleichmäßig im Raum verteilten Daten, vorliegen.
Die Datenwerte aus der Kernmenge und die entsprechend erhobenen zugehörigen y-Werte werden dabei entsprechend ihrer Gewichtung vervielfacht und gehen mehrfach in die Schätzung des Modell ein. Wird eine Gruppe durch mehrere Punkte repräsentiert, werden diese zunächst einzeln berücksichtigt. Zusätzlich wird der Mittelwert der x- und y-Werte dieser Punkte mehrfach hinzugenommen, so dass insgesamt n_i Beobachtungen aus einer Klasse mit n_i Punkten in die Schätzung mit eingehen. Die n_i entsprechen dabei der Anzahl der in der i-ten Klasse repräsentierten Datenpunkte.

4.2 k-Means-Auswahl

Verwendet man das k-Means Clusterverfahren auf die vorliegenden x-Werte, erhält man eine Aufteilung der für die Auswahl zur Verfügung stehenden Beobachtung in k Cluster. Die dazugehörigen Zentroide der Gruppen werden nun berechnet. Um ein Cluster gut zu repräsentieren, wird der Punkt in einer Klasse gewählt, der dem Klassenmittelwert am nächsten liegt. Dabei wird wieder die euklidische Norm für die Berechnung der Distanzen benutzt. Durch dieses Vorgehen erhält man eine Kernmenge mit r Beobachtungen. Insbesondere sind diese Punkte so über den gesamten Raum verteilt, dass keine Beobachtung nahe an einer anderen liegt und alle Gruppen, und damit alle Punkte, repräsentiert werden. Das Anwenden des k-Means-Algorithmus auf die behandelte Problemstellung ist (in dieser Form) bisher aus der Literatur nicht bekannt.
Bei diesem Verfahren ist eine Gewichtung bei Gruppen hilfreich, falls diese deutliche Unterschiede in den Gruppengrößen aufweisen. Als entsprechender Faktor für die Gewichtung eines Datenpunktes für die Schätzung, wird die jeweilige Clustergröße gewählt. Ist der Raum der gesamten Stichprobe nicht gleichmäßig verteilt, kann dies jedoch zu Problemen führen. In dem Fall, dass eine Variable über Ausprägungen verfügt, die eine große Spannweite besitzen, eine andere Variable jedoch eine sehr kleine, werden durch

das Klassifizierungesverfahren Gruppen gebildet, die einen Großteil der größeren Spannweite abdecken, jedoch nur einen kleinen Teil der Spannweite der anderen Variablen. Die Abbildung 5 verdeutlicht dies in einem zweidimensional Fall, indem die Variablen über einen sehr unterschiedlichen Wertebereich verteilt sind. Die schwarzen Punkte stellen dabei die jeweiligen Klassenmittelpunkte dar.

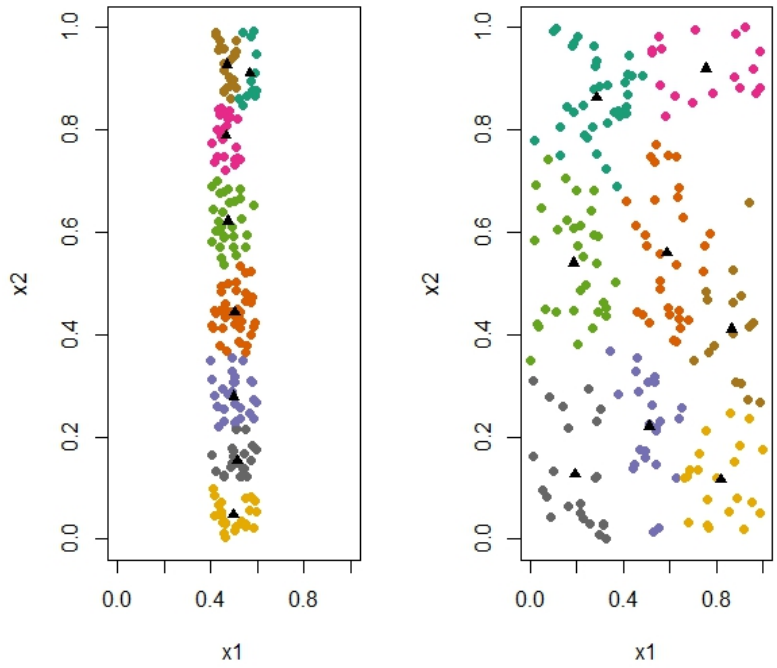

Abbildung 5: Links: Clusterung bei nicht standardisierten Daten.
Rechts: Clusterung nach Standardisierung der Daten.

Eine Auswahl der an den Zentroiden nahegelegensten Daten würde dazu führen, dass diese alle nahezu denselben Wert in der einen Variable besitzen (linke Seite der Abbildung) und nur in der Variable mit größerer Variablitität variieren. Abweichungen im linearen Modell fallen jedoch durch einen großen Abstand der Daten weniger ins Gewicht. Die folgende Formel legt dies anhand der Geradenbestimmung im eindimensionalen Fall nahe:

Steigung einer Geraden:

$$m = \frac{y_1 - y_2}{d(x_1, x_2)}.$$

Steigung einer Geraden mit Abweichung:

$$\tilde{m} = \frac{y_1 - y_2 + e}{d(x_1, x_2)} = \frac{y_1 - y_2}{d(x_1, x_2)} + \frac{e}{d(x_1, x_2)} = m + \frac{e}{d(x_1, x_2)}.$$

Dabei ist zu sehen, dass die Abweichung vom Modell und somit der Fehler in der Schätzung, in die Bestimmung der Geradensteigung eingeht. Dieser Schätzfehler wird jedoch durch die Distanz der Messpunkte geteilt. Eine große Distanz der Punkte führt also zu einer geringeren Abweichung der Steigung. Dies kann auch auf den mehrdimensionalen Fall verallgemeinert werden.
Es ist also von Bedeutung, die gesamte Breite des Raumes aller Einflussvariablen für die Wahl der Kernmenge zu nutzen. Um dies bei der Klassifizierung und der anschließenden Auswahl zu berücksichtigen, ist folgende Transformation sinnvoll. Zunächst wird dabei jeweils von den Spalten der Desginmatrix das jeweilige Minimum abgezogen. Dies führt dazu, dass nun der minimale Wert jeder der transformierten Spalten (Variablen) 0 ist. Daraufhin werden die Spalten durch das Maximum der Spalten, der so entstandenen neuen Matrix geteilt, wodurch sich der neue maximale Wert auf 1 beläuft. Die Standardisierung der Matrixeinträge verläuft also in folgenden zwei Schritten:

Schritt 1:

$$\tilde{x}_{ij} = x_{ij} - min(x_{*j}).$$

Schritt 2:

$$\tilde{\tilde{x}}_{ij} = \frac{\tilde{x}_{ij}}{max(\tilde{x}_{*j})},$$

wobei x_{*j} bzw. $\tilde{x}_{*j}$ die j-te Spalte der Designmatrix bzw. im ersten Schritt transformierten Matrix entspricht. x_{ij} sind jeweils die Einträge der Designmatrix an der i-ten Stelle der j-ten Spalte.

Durch diese zweischrittige Änderung der Daten ist der Raum der geänderten Einflussvariablen auf den n-dimensionalen Hyperwürfel $[0,1] \times \cdots \times [0,1]$ beschränkt.Eine solche Matrix wird auch Ähnlichkeitsmatrix genannt, da bei einer Distanz eines Einflusses von 1 ein maximaler Unterschied in der Variable vorliegt. Eine Auswahl der Kernmenge mit Hilfe des k-Means Verfahren mit diesen Werten führt nun zu einer größeren Variation der Variablenwerte bei Einflüssen mit geringer Spannweite. Die Grafik 5 zeigt dies anhand des zweidimensionalen Raumes (rechte Seite).
Bei der Schätzung des Modells mit der Teilstichprobe werden allerdings wieder die unveränderten Werte genutzt.

4.3 K-Means++-Auswahl

Um eine Auswahl aus Elementen eines Datensatzes zu bestimmen, können ebenso die Initialisierungswerte des k-Means++ verwendet werden. Dabei werden, wie oben beschrieben, die Startwerte des Verfahrens als Kernmenge benutzt. Dies hat den Vorteil, dass die so ausgewählten Punkte bereits im Datensatz enthalten sind und nicht mittels der Distanz zu den Zentroiden und einer Klassifizierung ermittelt werden müssen. Beobachtungen, die nahe zusammen liegen, werden dabei, durch ihre kleine Distanz zueinander, selten gleichzeitig in die Kernmenge aufgenommen, wodurch redundante Beobachtungen mit kleiner Wahrscheinlichkeit auftreten. Ebenso werden in den Iterationsschritten Teilräume der Einflussvariablen, die bisher nicht gut repräsentiert sind, da nur Beobachtung mit großer Distanz zu diesen ausgewählt wurden, bei der Wahl bevorzugt. Eine anschließende Verwendung des k-Means wird im Anschluss nicht mehr durchgeführt, da bereits eine Kernmenge von Datenpunkten vorliegt. Eine Gewichtung der Elemente dieser Kernmenge in diesem Fall ist zudem auch möglich. Da allerdings keine Klassifizierung bei der Initialisierung, wie beim k-Means, vorliegt, werden die Gewichte berechnet, indem zu jeder ausgewählten Beobachtung die Anzahl an weiteren Beobachtungen ermittelt werden, die diesem Wert am nächsten liegt. Ein jedes Element der Kernmenge repräsentiert also die Werte, die am ähnlichsten (in der Lage) sind. Dabei wird auch

hier die transformierte Designmatrix zur Bestimmung der Distanzen hinzugenommen. Diese Methode wird im folgenden als KM++V bezeichnet.

4.4 Sampling Algorithsm for L_p Regression

In einer Veröffentlichung von Dasgupta et al. aus dem Jahr 2009 ist ein Algorithmus vorgestellt, der sich bereits mit dem Ansatz der Kernmenge bei linearer Regression auseinandersetzt. Dabei werden in mehreren Schritten Beobachtungen ausgewählt, die eine vorher definierte Fehlertoleranz nicht unter bzw. überschreitet.
Im ersten Schritt des Verfahrens wird für jede der Beobachtungen eine Wahrscheinlichkeit berechnet, die proportional zu den Leverage Score der Beobachtung ist. Diese Wahrscheinlichkeiten $\tilde{p_i}$ berechnen sich wie folgt:

$$\tilde{p_i} = \min(1, \frac{p_i}{\sum p_i} r).$$

p_i ist dabei der Leverage Score der i-ten Beobachtung, r die erwartete Anzahl so ausgewählter Beobachtungen. Dabei wird bei der Berechnung der Leverage Scores die QR-Zerlegung benutzt, wodurch sich der obige Ausdruck zu folgendem ergibt:

$$\tilde{p_i} = \min(1, \frac{||U_{i*}||_2^2}{||U||_2^2} r),$$

wobei U der orthogonalen Matrix der QR-Zerlegung entspricht. U_{i*} ist die entsprechende Vektornorm der i-ten Zeile und $||U||$ die Matrixnorm.
Jede Beobachtung wird nun im zweiten Schritt mit der Wahrscheinlichkeit $\tilde{p_i}$ in die Kernmenge aufgenommen. Der Erwartungswert der Anzahl der Beobachtungen in der Kernmenge ist dabei der Faktor r. In der Veröffentlichung wird gezeigt, dass der Wert des Faktors auf $r = 96 * d^2(d * ln(96) + ln(200))$gesetzt werden kann, um folgende Schranke für die Abweichung im Modell einhalten zu können:

$$(1-\epsilon)||X\hat{\beta}-y||_2 \leq ||S(X\hat{\beta}_K-y)||_2 \leq (1-\epsilon)||X\hat{\beta}-y||_2.$$

Dabei bezeichnet S eine Diagonalmatrix mit den Diagonalelementen $\frac{1}{\sqrt{p_i}}$ und $\epsilon \in (1; \frac{1}{7})$ und $\hat{\beta}_K$ den aus der Kernmenge geschätzten Parametervektor. Im letzten Schritt wird ein Modell mit den Beobachtungen aus der Kernmenge geschätzt, wobei diese jeweils mit der Gewichtung $\frac{1}{\sqrt{p_i}}$ eingehen.

Der vorgestellte Algorithmus (Salp) benutzt daraufhin bei der weiteren feineren Auswahl der Datenpunkte in den darauffolgenden Schritten die y-Werte der Regression bzw. trifft eine weiterführende Auswahl anhand der Residuen. Dies ist jedoch für die hier vorliegende Problemstellung nicht möglich, da diese Werte nicht zur Auswahl herangezogen werden, sondern nur die x-Werte vollständig für die Bestimmung der Kernmenge genutzt werden. Der oben beschrieben Algorithmus wird nun auf die Problemstellung angepasst, indem ebenfalls die Leverage Scores der Daten benutzt werden, um Datenpunkte mit hoher Hebelwirkung auf das Modell mit hoher Wahrscheinlichkeit in die Kernmenge mit aufzunehmen. Die Auswahl erfolgt indem genau r unterschiedliche Punkte aus den Daten mit einer Wahrscheinlichkeit proportional zu den Leverage Scores gezogen werden. Diese Teilmenge ist dann eine Kernmenge, die genau r Beobachtungen enthält.

Eine Auswahl der r Beobachtungen mit den höchsten Scores ist nicht ratsam, da dies dazu führen kann, dass nur einzelne Bereiche des gesamten Variablenraumes betrachtet werden. Unterscheiden sich etwa die x-Werte einer Gruppe von r oder mehreren Daten sehr von den übrigen Daten, sind die Scores dieser Gruppe besonders hoch. Bei der deterministischen Wahl der Punkte mit höchsten Scores, würde ausschließlich diese Gruppe berücksichtigt. Dies würde dazu führen, dass die übrigen Werte, die einzeln wenig Einfluss haben, jedoch einen Großteil der Daten ausmachen, nicht berücksichtigt werden. Dadurch würde auch die Spannweite der Variablen in der Kernmenge entsprechend der Klasse, gering ausfallen. Die Auswahl mit Hilfe des gewichteten Zufalls ist daher einer solchen Wahl vorzuziehen.

4.5 Clusterscore

Bei dem Salp-Algorithmus werden mit höheren Wahrscheinlichkeiten Werte ausgewählt, die hohe Leverage Scores besitzen. Dies führt dazu, dass die Variablenwerte häufiger am Rande des Wertebereichs liegen und einen großen Abstand zu den Mittelwerten der gesamten Stichprobe aufweisen. Die Nutzung des k-Means-Verfahrens hingegen führt zu Werten, die nahezu gleichmäßig über den Raum verteilt sind und mit nur geringer Wahrscheinlichkeit einzelne Klassen nicht richtig repräsentiert werden. Der Clusterscore-Algorithmus verbindet nun diese beiden Prinzipien. Dabei werden die Daten zunächst mit k-Means klassifiziert, wobei dieselbe oben beschriebene Transformation der Daten benutzt wird. Die Anzahl der Cluster wird dabei jedoch um den Faktor 0.95 verringert (also $k = 0.95r$) und ggf. abgerundet. Ebenfalls werden die Leverage Scores des gesamten Datensatzes berechnet. Im folgenden Schritt wird aus jedem der Klassen die Beobachtung deterministisch ausgewählt, die den höchsten Score aufweist. Somit besteht die vorläufige Kernmenge aus $\lfloor 0.95r \rfloor$ Beobachtungen. Nun werden die entsprechenden y-Werte hinzugezogen und eine Schätzung des Modells vorgenommen. Diese Vorauswahl und vorläufige Schätzung ermöglicht nun die Berechnung der Cooks Distanzen. Diese Distanzen geben Aufschluss darüber, ob einer der Werte das Modell, aufgrund der Abweichung und Position, stark ändern. Es ist sinnvoll, solche Werte genauer zu prüfen, um die Robustheit der Schätzung zu stärken und eventuelle starke Abweichungen zu relativieren und so Gruppeneinfluss aufzudecken. Ist etwa bei einer starken positiven Abweichung ein weiterer Wert mit negativer Abweichung in der Nähe, so gleichen sich diese Residuen in dieser Hinsicht auf den Einfluss auf das Modell aus. Es liegt ein kleiner Gruppeneinfluss vor. Die in diesem Kapitel beschriebene Prozedur wählt jedoch bei der Vorauswahl nur einzelne Werte eines Clusters aus. Deshalb wird in nächsten Schritt des Verfahrens die Beobachtungen identifiziert, die die $\lceil 0.05r \rceil$ größten Cooks Distanzen aufweisen. Zusätzlich werden nun Datenpunkte ausgewählt, die nahe an diesen Beobachtungen liegen und noch nicht in der Vorauswahl gewählt wurden. Dadurch können in einigen Fällen Gruppeneinflüsse identifiziert werden und abweichende

Ergebnisse noch einmal kontrolliert werden. Insgesamt besteht die Kernmenge wiederum aus r Elementen und für die Schätzung sind ebenfalls nur r y-Werte zu erheben.
Eine Gewichtung kann bei dieser Methode zusätzlich hilfreich sein. Die Gewichte entsprechen dabei, wie beim k-Means-Auswahl, den Clustergrößen. Durch die Wahl der letzten $\lceil r * 0.05 \rceil$ Punkten ist dabei zu beachten, dass eine Kernmenge mehrere Elemente aus einem Cluster enthalten kann. In einem solchen Fall wird jedoch wie im obigen Kapitel 4.1 beschrieben verfahren und der Mittelwert der Beobachtungen aus dem Cluster entsprechend den Gewichten, mehrfach hinzugefügt. Diese Methode wurde bisher noch nicht untersucht und entstand im Verlaufe dieser Masterarbeit.

4.6 Deterministisches Epsilonverfahren

Ein weiteres Verfahren zur Auswahl einer geeigneten Kernmenge ist das deterministische Epsilonverfahren (DEV). Es beruht auf einer Methode, die in einer abgeänderten Form in der Veröffentlichung von Felman et al. (2011) beschrieben wird. Das dort dargestellte Verfahren wählt zunächst zufällig mit gleicher Wahrscheinlichkeit eine feste Anzahl an Punkten aus der Menge aller Beobachtungen aus. Daraufhin werden alle weiteren Datenpunkte, die nahe (mit einem Abstand von weniger als einem festen Wert ϵ) an den bereits ausgewählten Punkte liegen, ausgeschlossen. Aus den Elemente, die noch nicht ausgewählt und ausgeschlossen wurden, wird im nächsten Schritt wieder eine zufällige Auswahl getroffen und naheliegende Punkte wieder aus der weiteren Betrachtung entfernt. Diese Vorgehen wiederholt sich bis keine Beobachtungen mehr ausgeschlossen bzw. ausgewählt werden kann.
Dieser Algorithmus wählt so eine Kernmenge aus, die in einer mehrdimensionalen, gemischten Verteilungen die Daten bzw. die Verteilungen repräsentieren. Die Wahl des ϵ bzw. der Umgebung um die jeweiligen Punkte ist dabei entscheidend für die Anzahl der Elemente der Kernmenge. Ebenso führt der Algorithmus bei festem ϵ, durch die zufällige Auswahl in jedem Iterationsschritt, teilweise zu unterschiedlichen Teilstichprobengrößen. Dieses Vorgehen wird nun auf die Problemstellung der Kernmengen im linearen Modell

angepasst und angewendet. Zunächst werden die Einflussstärke der x-Werte mit Hilfe der Leverage Scores ermittelt. Im Folgenden wird die Beobachtung aus der gesamten Menge der n Beobachtungen betrachtet, die die höchsten Leverage Scores aufweisen. Da davon ausgegangen wird, dass die naheliegenden Daten ähnliche Werte für die Zielvariable und somit keine neuen Informationen für das Modell liefern, werden die $\frac{n-r}{r}$ Datenpunkte, die dem ausgewählten Punkt am nächsten liegen, aus dem Datensatz entfernt. Die Anzahl der in jedem Schritt ausgeschlossenen Elemente $\frac{n-r}{r}$ wird ggf. aufgerundet. Im weiteren Verlauf wird diejenige Beobachtung mit dem höchsten Einfluss aus den verbliebenen Werten ausgewählt und zur Kernmenge hinzugefügt und dieses Vorgehen wiederholt, bis die Kernmengengröße r entspricht. Durch diese Anpassung des obigen Verfahrens werden Werte bevorzugt, die hohen Einfluss und eine starke Hebelwirkung auf das Modell haben. Da die Werte der Scores sich erhöhen, je weiter diese vom Zentrum der Daten entfernt liegen, liegt auch in jedem Schritt der ausgewählte Datenpunkt am Rande der übrigen Daten. Durch das Abrunden des Anteils der zu entfernenden Elemente kann es im letzten Schritt vorkommen, dass einige Objekte übrig bleiben. Da diese jedoch zu den mit den niedrigsten r Leverage Scores gehören, ist ihr Einfluss entsprechend geringer.
Außerdem werden Teile des Raumes, die viele Elemente enthalten, durch mehrere Beobachtungen repräsentiert, da jeweils nur eine feste Anzahl ausgeschlossen wird, unabhängig von ihrer euklidischen Distanz. Aus diesem Grund ist eine Gewichtung bei diesem angepassten Algorithmus nicht notwendig, da in jedem Schritt jeweils die feste Anzahl an Elementen als Gewichtung herangezogen werden kann (diese aber fest und somit gleich ist).

4.7 Maximal-Distanz-Algorithmus

Bei der Auswahl von Elementen zu einer Kernmenge ist es sinnvoll, den gesamten Raum aller Daten darzustellen. Die weitere Auswahl eines Wertes in direkter Nähe eines bereits gewählten Wertes bringt daher meist weniger neue Informationen und ist redundant.

Ebenso werden Abweichungen vom Modell bei Daten mit kleinen Distanzen in den x-Werten einen deutlich höheren Einfluss auf die Zielparameter haben (vgl. Überlegung zum k-Means in Kapitel 4.1). Daher macht es Sinn, Werte aus den Teilräumen auszuwählen, die weit entfernt von der bisherigen Auswahl liegen, um diese zusätzlich zu repräsentieren und die Abweichungen durch einen hohen Abstand zueinander zu relativieren. Der Maximal-Distanz-Algorithmus (MDA) geht von diesem Ansatz aus. Ein ähnlicher Ansatz wurde bereits in Agarwal et al. (2005) beschrieben. Dieser wird nun auf die obige Problemstellung angepasst.
Im ersten Schritt der Methode wird eine Beobachtung zufällig ausgewählt und der Kernmenge hinzugefügt. Daraufhin werden die Distanzen sämtlicher anderen Daten zu zur neuen Beobachtung berechnet und der Punkt ebenfalls ausgewählt und hinzugefügt, der die größte Distanz hat. In den folgenden Schritten wird jeweils die Distanzen eines jeden Punktes zu den Kernmengenpunkten ermittelt, die am kleinsten sind, also die jeweils kleinsten Distanzen zu einem der Elementen der Auswahlmenge aufweisen. Die Beobachtung mit der maximalen dieser Minimaldistanzen wird daraufhin ausgewählt und die Schritte werden wiederholt, bis die Teilstichprobe r Elemente beinhaltet.
Insgesamt läuft der Algorithmus also in folgenden Schritten ab:

Schritt 1: zufällige Wahl einer Beobachtung mit gleichen Wahrscheinlichkeiten
Schritt 2: Berechnung der Distanzen der übrigen Punkte zu denen aus der Kernmenge
Schritt 3: Wahl der Beobachtung mit dem maximalen Minimalabstand
Wiederholung von Schritt 2 und 3 bis r Elemente in der Kernmenge vorliegen

Da diese Methode aus den Teilräumen mit vielen Beobachtungen nicht häufiger Daten wählt, ist auch bei dieser Methode eine Gewichtung sinnvoll. In diesem Fall wird die Anzahl der Beobachtungen, die einem Punkt aus der Kernmenge am nächsten liegen, wie beim k-Means++ Auswahlverfahren als Gewichtung für die Teilstichprobe verwendet.

4.8 Zufallsauswahl

Um bei der Untersuchung der Auswahlverfahren die Ergebnisse einordnen zu können, wird zudem eine zufällige Auswahl vorgenommen. Dabei wird jede der Beobachtung mit gleicher Wahrscheinlichkeit in die Teilstichprobe aufgenommen. Insgesamt wird die Anzahl der Elemente dieser Menge auf r beschränkt. Diese Methode ist rein zufällig und dient somit als Vergleichsmethode zu den systematischen Ansätzen der anderen Verfahren.

	verwendete Ansätze	determ.	Gewichtung
Zufalls-Auswahl	zufällig mit gleichen Wahrscheinlichkeiten	nein	nein
K-Means-Auswahl	Nähe zu k-Means-Zentroide	nein	ja
Salp	Zufall prop. zu Leverage Scores	nein	nein
MDA	erste Beob. zufällig, schrittw. Maximale Abstände zur Kernmenge	nein	ja
KM++V	Startklassenmitten des k-Means++	nein	ja
Clusterscore	k-Means, Leverage Scores, Prüfen der „ungewöhnlichen“ Daten	nein	ja
DEV	Leverage Scores, schrittw. Ausschluss naher Punkte	ja	nein

Tabelle 1: Übersicht der Auswahlmethoden

Die Tabelle 1 gibt eine Übersicht über die beschriebenen Auswahlverfahren mit den jeweiligen Ansätzen an. Ebenfalls ist aufgelistet, ob es sich um ein deterministisches Verfahren handelt und eine Gewichtung in Fällen nicht gleichmäßig verteilten Daten benutzt wird.

4.9 Kostenkriterien

Um die vorgestellten Methoden und die Auswahl der Kernmengen, die aufgrund der Anwendungen der Methoden getroffen werden, bewerten zu können, benötigen wir ein geeignetes Kriterium. Im folgenden Kapitel werden zwei Kriterien vorgestellt, die die Kosten (Abweichungen vom Modell anhand der Teilstichprobe) der Teilstichproben und so eine Kennzahl für den Erfolg einer Auswahl darstellen.

Da das Modell, das mit Hilfe des gesamten Datensatzes geschätzt wird, möglichst durch das Modell mit durch Schätzung einer Kernmenge angenähert werden soll, ist es sinnvoll dieses Modell als Grundlage eines solchen Maßes zu benutzen und mit der Schätzung aus der Teilmenge an Daten zu vergleichen. Als Maß zur Bewertung der Ähnlichkeit dieser beiden Modelle kann etwa der Abstand der sich ergebenen Parametervektoren herangezogen werden. Dieser berechnet sich mit der euklidischen Norm wie folgt:

$$K_1 = \sqrt{(\hat{\beta} - \hat{\beta}_K)^T(\hat{\beta} - \hat{\beta}_K)}.$$

Ist dieser Wert hoch, weichen die Parameter des mit der Kernmenge ermittelten Modells deutlicher ab. Ist der Wert klein, liegt entsprechend eine geringe Abweichung der beiden Modelle vor. In diesem Fall sind die Kosten, die man durch die Beschränkung der Stichprobe erhält, niedrig und eine gute Annäherung an das Modell der gesamten Stichprobe wurde mit der Kernmenge erreicht. Das Kriterium wird im Folgenden als Bewertungskriterium 1 bezeichnet (vgl. Eckey et al., 2003).

Bei der Schätzung eines linearen Modells mit der KQ-Methode wird die Summe der quadratischen Abweichungen vom Modell über den Parametervektor minimiert (siehe

obiges Kapitel 3). Dies kann ebenfalls als Grundlage einer Bewertung herangezogen werden. Dazu berechnet man zunächst die Summe der quadratischen Abweichung aller Beobachtungen zu dem mit den Kernmengen geschätzten Modells. Daraufhin dividiert man diesen Wert durch den entsprechenden Wert, der sich aus dem Modell mit allen Beobachtungen ergibt:

$$K_2 = \frac{||(X\hat{\beta}_K - y)||^2}{||(X\hat{\beta} - y)||^2}.$$

Dabei entspricht $\hat{\beta}_K$ dem Vektor, der mit Hilfe der Kernmenge geschätzt wird. Da der Nenner das Minimum über alle Parametervektoren darstellt und der Zähler aus diesem Grund größer oder gleich groß ist, ergibt sich der Wert in jedem Fall zu einem Wert, der höher als 1 ist.
Unterscheidet sich das Modell der Kernmenge deutlich von dem des aus der gesamten Stichprobe geschätzten Modells, so ergibt sich ein höherer Wert für dieses Kriterium und die Anpassung ist somit schlechter. Im Folgenden wird dieser Wert für die Kosten als Kriterium 2 bezeichnet. Eine Bewertung mit diesem Kriterium wird in Boutsidis et al. (2013) beschrieben.

Diese beiden Kriterien dienen im folgenden Verlauf der Arbeit als Maß für die Bewertung der Auswahl der Teilstichproben und ermöglichen es, so die Ergebnisse anhand dieser Kennzahlen vergleichen zu können. Zu beachten ist dabei, dass ein hoher Wert im ersten Kriterium nicht zwangsläufig zu einem großen Wert des zweiten Kriteriums führen muss, da die erste Bewertung den Unterschied im geschätzten Parametervektor misst, während der zweite einen Änderungsfaktor (Vergrößerung) der nicht durch das Modell erklärten Varianz wiederspiegelt.

5 Simulationsstudie

Damit die oben vorgestellten Methoden untersucht werden können, wird nun eine Simulationsstudie aufgestellt. Dabei werden Daten unter verschiedenen Voraussetzungen und mit unterschiedlichen Parametern betrachtet und ausgewertet. Dies ist insbesondere deshalb sinnvoll, da sämtliche Daten simuliert werden können und die Bewertungskriterien, für die man den Gesamtdatensatz und insbesondere die y-Werte benötigt, berechnet werden können. Außerdem können in darauffolgenden Untersuchungen die Voraussetzungen und Einstellungen einer jeder Simulation betrachtet werden und Einflüsse und Zusammenhänge von einzelnen Parametern genauer untersucht werden, da diese in der Studie vorliegen.

Die Simulationsstudie beinhaltet eine Vielzahl von Datensätzen und Parametern. So wird etwa die Stichprobengröße der einzelnen Stichproben auf den Wert $n = 1000$, 2000 bzw. 5000 gesetzt. Zusätzlich wird die Anzahl der Einflussfaktoren, die auf die abhängige Variable Einfluss haben mit $d = 5$, 10 bzw. 20 variiert.
Mit diesen Einstellungen wird nun ein Designmatrix mit zufälligen x-Werten simuliert. Dabei sind die Werte aus einem Intervall $[a_i; b_i]$ mit einer Rechteckverteilung ermittelt. Die untere Intervallgrenze a_i wird für jede der d Einflussvariablen zufällig mit gleicher Wahrscheinlichkeit aus dem Intervall $[-3; 2]$ gezogen. Die obere Grenze b_i wird daraufhin ebenfalls jeweils mit einer Rechtecksverteilung bestimmt, indem der Wert b_i wie folgt aus einer Rechtecksverteilung gezogen wird:

$$b_i = a_i + R[1, 3 - a_i].$$

Die so entstehenden x-Werte liegen für jede Variable in einem Bereich zwischen -3 und 3. Also in dem Hyperwürfel $[-3; 3] \times \cdots \times [-3; 3]$) Zudem gilt, dass $b_i > a_i$ und $b_i - a_i > 1$. Nach der Erzeugung der x-Werte der Studie werden die d Einflussgrößen der Variablen im Modell ermittelt, indem diese zufällig, gleichverteilt aus dem Intervall $[-3; 3]$ gezogen werden. Ein Wert für den y-Achsenabschnitt des simulierten Modells wird ebenfalls

auf dieselbe Weise simuliert. Für jede dieser Einstellungen werden je 100 Durchläufe simuliert, um zufällige Effekte bei der Simulation und der weiteren Untersuchungen weitgehendst ausschließen zu können.
Nach der Erzeugung der Designmatrix und des Modellparametervektors β, werden nun die dazugehörigen y-Werte berechnet. Dabei werden verschiedene Szenarien für mögliche Verzerrungen in den Datensätzen simuliert. Diese werden in folgendem genauer dargestellt.

Als erstes werden Szenarien untersucht, die ein einzelnes lineares Modell betrachten. Zudem ist die Voraussetzungen der Linearität in diesen Modellen gegeben und deren Verzerrungen beruhen auf einem normalverteilten Fehlerterm. Diese Verzerrungen werden im Folgenden als Verzerrungen des Typs A bezeichnet.

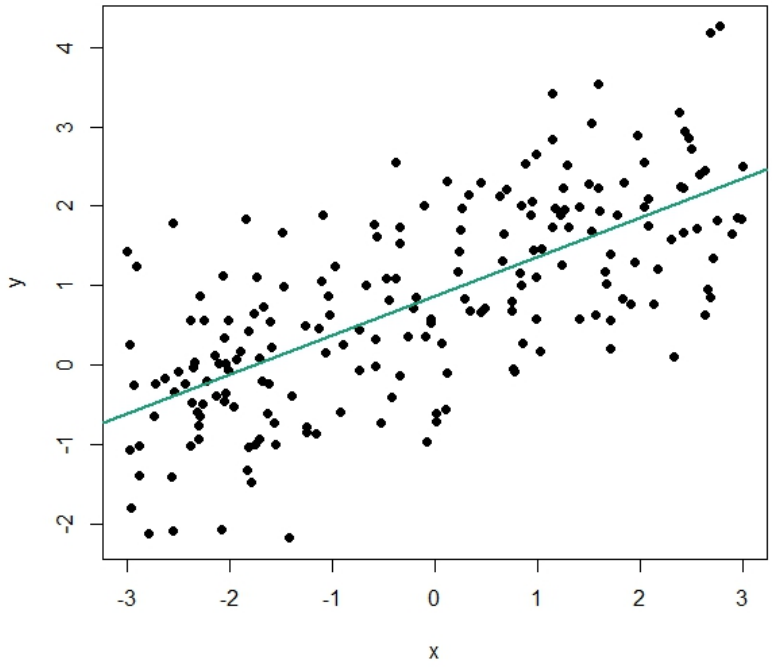

Abbildung 6: Verzerrung A0 im eindimensionalen Fall

Zunächst wird in der Verzerrung A0 ein Modell mit einer geringen Varianz vom linearen

Modell erzeugt. Dabei werden die y-Werte des Datensatzes erst mit $X\beta$ berechnet und daraufhin mit einer Abweichung verzerrt. Die Residuen werden aus einer Standardnormalverteilung zufällig gezogen. Diese Datensätze bilden eine eher gutmütige Situation für die Daten ab, da diese mit nur einer geringen Fehlervarianz verzerrt werden und dient daher als nahezu optimale Situation. In Abbildung 6 ist dieser Fall im eindimensionalen dargestellt.

Eine weitere Einstellung beinhaltet die Verzerrung A1. Bei dieser Abweichung vom Optimalfall, werden zunächst wieder die y-Werte des Modells, wie zuvor, berechnet. Die Residuen des Modells werden jedoch aus einer Verteilung mit deutlich höherer Varianz ermittelt. Diese entspricht einer Normalverteilung mit einer Varianz von 100. Dies führt zu weitaus höheren Abweichungen der Werte zu den Modellwerten. Die Grafik 7 zeigt dies zur Veranschaulichung im eindimensionalen Fall.

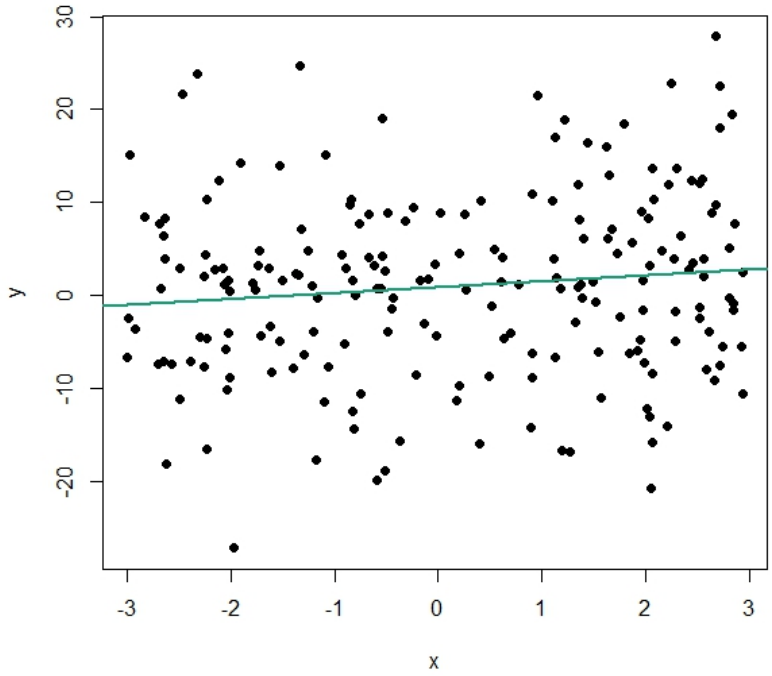

Abbildung 7: Verzerrung A1 im eindimensionalen Fall

Bei der Verzerrung A2 liegen wieder mit einer Varianz von 100 verzerrte y-Werte vor. Außerdem werden zuvor 5 Prozent der x-Werte so geändert, dass sie als Gruppe weit außerhalb der übrigen x-Werte liegen. Dazu werden diese 5 Prozent neu ermittelt, indem zu den höchsten möglichen Grenzen b_i des jeweiligen Einflussfaktors ein Wert von 10 hinzugefügt wird und zu dem nochmal ein Wert hinzugefügt wird, der aus $R[-0.25, 0.25]$ gezogen wird. Dadurch entsteht in dem Datensatz eine Gruppe von Beobachtungen, die in den x-Werten starke Abweichungen zu den übrigen Punkten haben. Alle diese Daten liegen jedoch in einem kleinen Bereich zusammen.
Die zugehörigen y-Werte werden wieder wie in Verzerrung A1 ermittelt. Diese Art der Verzerrung ist für ein Modell mit nur einem Einflussfaktor in Abbildung 8 dargestellt.

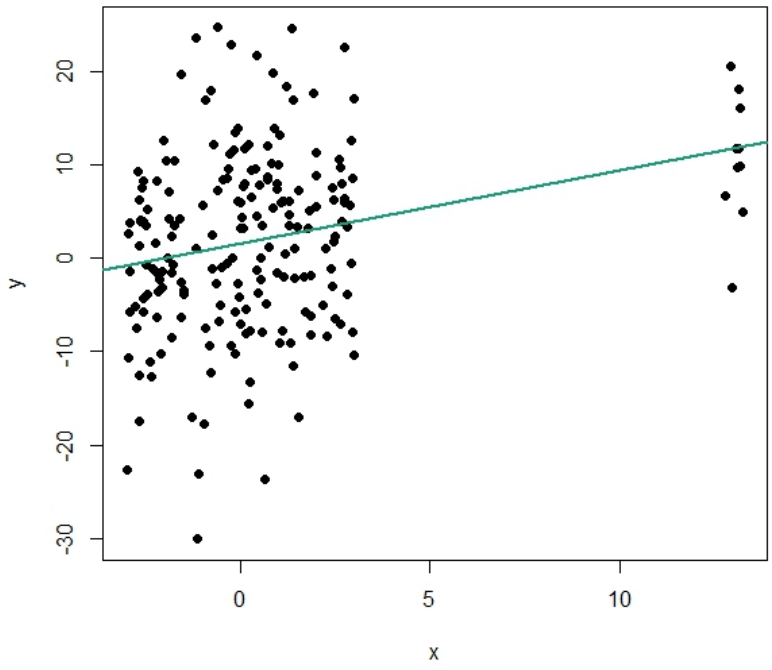

Abbildung 8: Verzerrung A2 im eindimensionalen Fall

Ein weiteres Szenario (A3), das betrachtet wird, ist das von heteroskedastischen Daten. Dabei handelt es sich um Datenpunkte, die unterschiedlich hohe Streuungen zum Modell aufweisen. Die y-Werte werden hier mit einem Meßfehler e versehen, der abhängig von der Höhe der y-Werte ist. Die Abweichungen werden aus einer Normalverteilung $N(0, v_i)$ gezogen, wobei sich v_i aus dem Term $1 + 99\frac{y_i - min(y)}{max(y) - min(y)}$ ergibt. Sind die y-Werte also klein, ist auch die Varianz, mit der diese vom Modell abweichen, entsprechend gering. Für die kleinste abhängige Beobachtung gilt eine Varianz von 1, für das Maximum eine Varianz von 100. Für den eindimensionalen Fall wird dies in der Grafik 9 verdeutlicht.

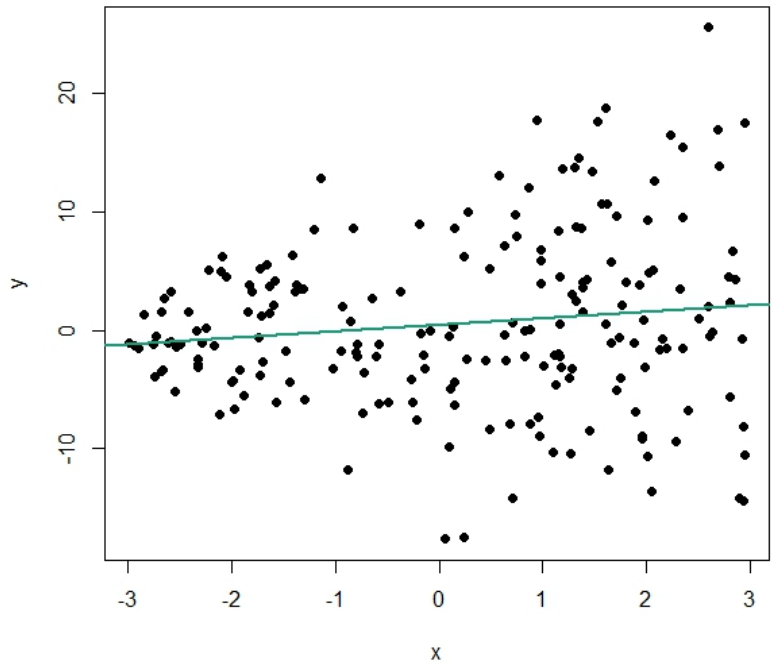

Abbildung 9: Verzerrung A3 im eindimensionalen Fall

In linearen Modellen kann es in manchen Fällen zu starken Ausreißern kommen. Diese Form der Verzerrung führt zu deutlich unterschiedlichen Ergebnissen bei der Schätzung. Dabei ist zu beachten, dass bei der Auswahl auch das Fehlen solcher Ausreißer proble-

matisch werden kann, da dadurch Verzerrung im Gesamtdatensatz nicht vollständig abgebildet werden können. Verzerrungen, die in der Simulationsstudie erzeugt werden und solche starke Ausreißer in der abhängigen Variable enthalten, werden mit dem Buchstaben B gekennzeichnet.

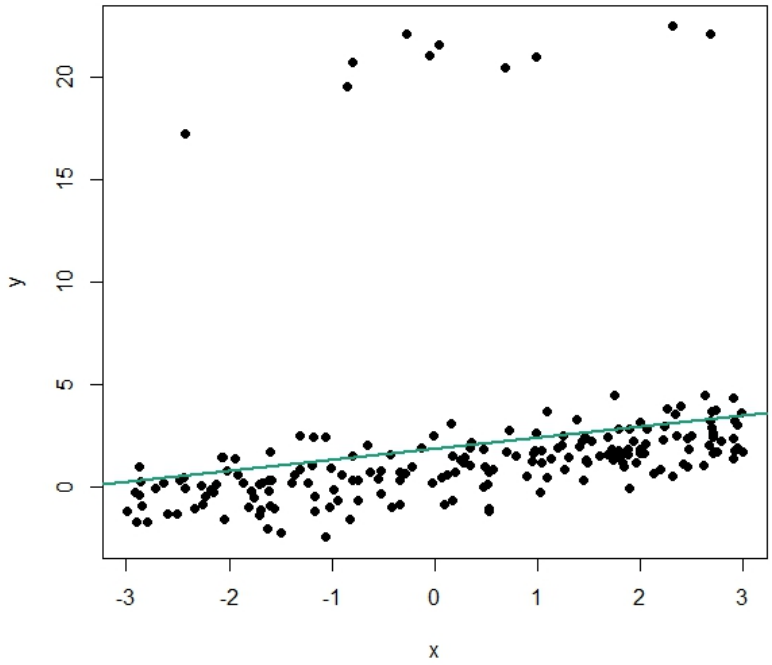

Abbildung 10: Verzerrung B1 im eindimensionalen Fall

Eine solche Verzerrung (B1) wird erzeugt, indem zufällig mit gleicher Wahrscheinlichkeit Beobachtungen aus den n Datenpunkten ausgewählt werden. Bei der Erzeugung der entsprechenden y-Werte und der anschließenden Verzerrung mit einer geringen Varianz von 1, werden auf diese zusätzlich je ein Wert aus $|N(20, 1)|$ addiert. Der Anteil der Daten im Datensatz, bei dem solche Ausreißer vorkommen, wird auf 5 Prozent festgesetzt. Durch den Absolutbetrag wird zudem sichergestellt, dass nur starke Ausreißer nach oben vorkommen. Dies ist sinnvoll, da starke Abweichungen nach oben und unten lediglich zu unterschiedlichen Varianzen in der Simulation führen und so das Modell nicht sehr viel

anders verzerrt als bei einer hohen Varianz. Eine Darstellung der Verzerrung B1 für den eindimensionalen Fall ist in der Abbildung 10 zu sehen.

Bei der Verzerrung B2 handelt es sich ebenfalls um Abweichungen, die mit Hilfe von starken Ausreißern simuliert werden. Um jedoch den Einfluss einer ganzen Gruppe von Ausreißern in einem kleinen Teilraum der Daten untersuchen zu können, werden zunächst 95 Prozent des Datensatzes, wie zuvor mit der Verzerrung A0 (geringe Varianz) erzeugt. Die übrigen Beobachtungen werden um einen zufälligen Datenpunkt gebildet, indem zunächst eine Beobachtung zufällig mit gleicher Wahrscheinlichkeit ausgewählt wird. Um diesen Punkt werden nun die übrigen x-Werte erzeugt, indem der ausgewählte mit einem Wert aus $R[0, 0.5]$ variiert wird. Jede dieser Variationen wird nun mit demselben Absolutwert aus B1 addiert, um so eine Gruppe zu erzeugen, die alle in einem kleinen Teil des Raumes als Ausreißer gewertet werden.

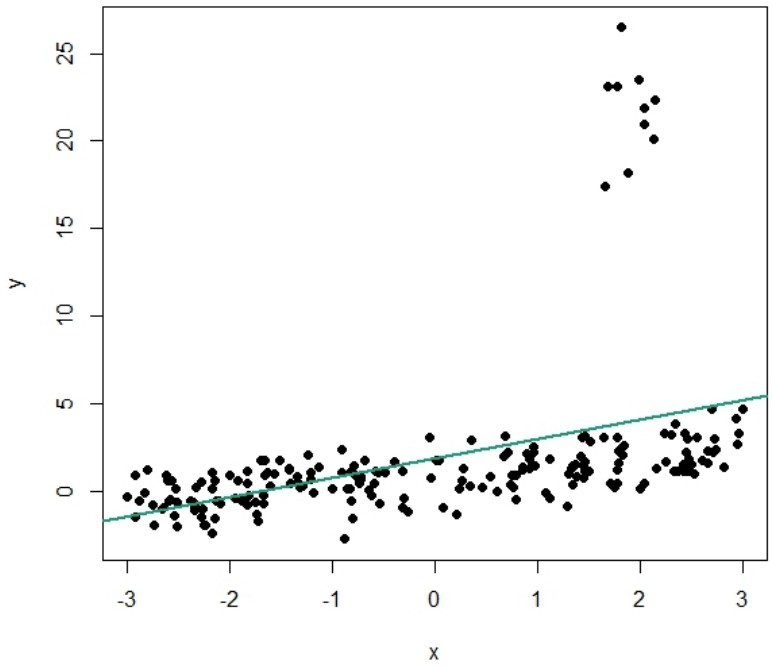

Abbildung 11: Verzerrung B2 im eindimensionalen Fall

Die Abbildung 11 stellt diese Situation grafisch für den eindimensionalen Fall dar. Dabei ist in diesem Fall zu berücksichtigen, dass der Bereich des Datenraumes, der Ausreißer enthält, teilweise deutlich mehr Beobachtungspunkte enthalten kann, da diese zusätzlich zu den vorher gleichverteilten Daten gezogen werden. Dies berücksichtigt also, durch die unterschiedliche Verteilung der x-Werte, noch eine weitere Variation in der Simulationsstudie.

Eine weitere problematische Situation, die in einem Datensatz auftreten kann, ist, dass die Daten aus unterschiedlichen Populationen stammen und die Einflüsse auf diese Beobachtungen unterschiedlich sind. Um diese Problemstellung ebenfalls zu untersuchen, werden nun Datensätze simuliert, die aus zwei unterschiedlichen Modellen ermittelt werden. Dabei werden die x-Werte für alle Daten aus demselben Wertebereich gezogen. Die y-Werte werden jedoch in 33 Prozent der Fälle mit einem Parametervektor β_1, wie zuvor, ermittelt. Die übrigen Beobachtungen der abhängigen Variable werden mit einem anderen Vektor β_2 berechnet. Dieser ergibt sich auf dieselbe Weise mit einer zufälligen Auswahl von Werten aus der $R[-3, 3]$ Gleichverteilung, unterscheidet sich allerdings von dem des ersten Modells, da unterschiedliche Parameter ermittelt werden. Es werden also Daten aus zwei unterschiedlichen Modellen gezogen, mit gleicher Anzahl an Einflüssen. Im Anschluss werden die Werte wiederum mit einer Fehlervarianz von 1 verzerrt. Diese Problematik wird mit Verzerrung C bezeichnet und ist in Abbildung 12 für Modelle mit je nur einem Einfluss dargestellt.

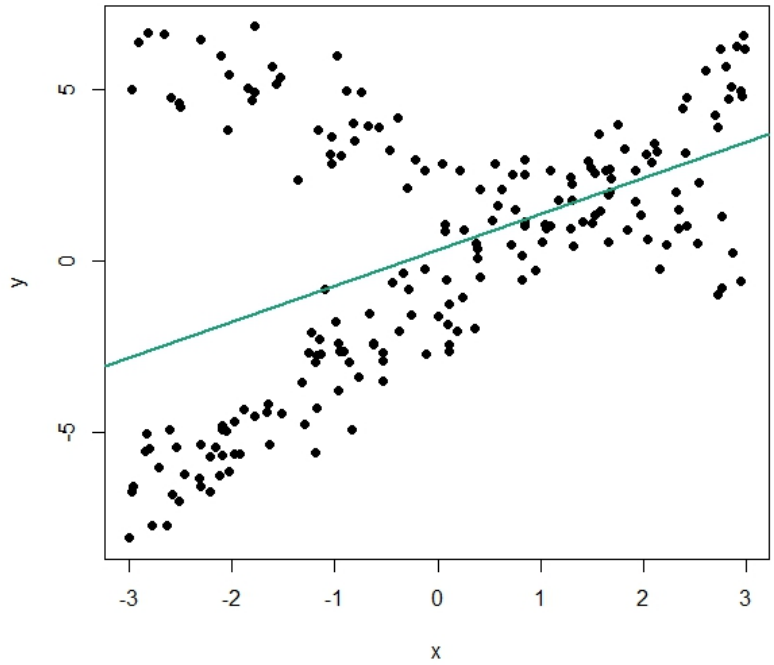

Abbildung 12: Verzerrung C im eindimensionalen Fall

Liegt ein nicht linearer Zusammenhang zwischen der abhängigen und den unabhängigen Variablen vor, kann dies zu fehlerhaften Schätzungen und Vorhersagen des Modells führen. Dieser Fall wird in der Simulationsstudie mit Verzerrung D abgedeckt. Dabei wird ein nicht linearer Zusammenhang für die Berechnung der y-Werte herangezogen, der bei der Schätzung jedoch linear geschätz wird. (Es wird also angenommen, dass nicht erkannt wird, dass ein nicht linearer Zusammenhang in den Daten vorliegt.) Als nicht lineare Funktion bietet sich hierbei die Sinusfunktion an, da diese, je nach Lage der Werte, sowohl einen konkaven als auch einen konvexen Verlauf beschreibt. Vor der Berechnung der y-Werte wird daher zunächst der Sinuswert jeder Einflussvariable ermittelt und ein lineares Modell anhand dieser transformierten Daten aufgestellt. Das transformierte Modell ergibt sich zum Ausdruck:

$$y = sin(1)\beta_0 + sin(x_1)\beta_1 + \cdots + sin(x_d)\beta_d + e,$$

mit den entsprechenden obigen Bezeichnungen. Der Fehlerterm e ist dabei wiederum mit Varianz 1 simuliert. Anschaulich ist dies für den eindimensionalen Fall in Grafik 13 abgebildet.

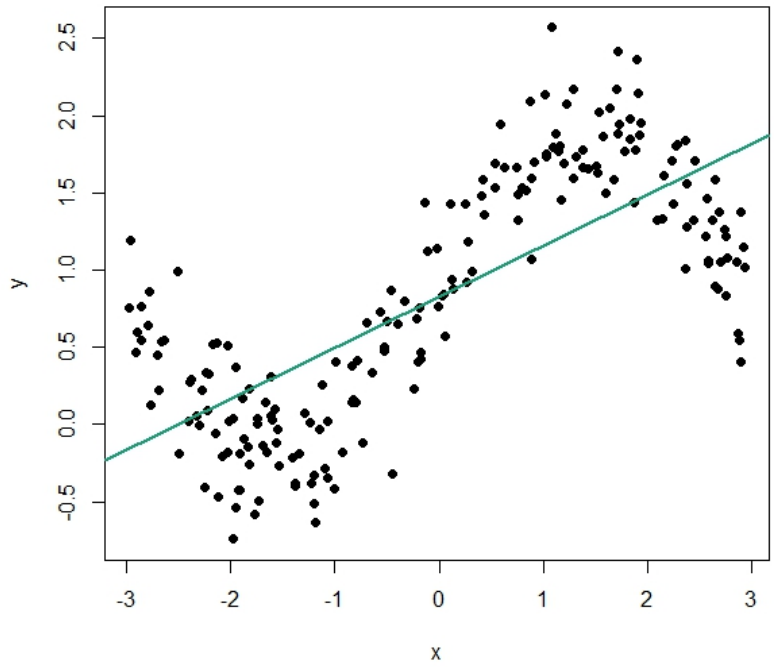

Abbildung 13: Verzerrung D im eindimensionalen Fall

Zuletzt werden noch Wechselwirkungen berücksichtigt. Hierbei werden jedoch Wechselwirkung simuliert, die in das Modell bzw. die abhängige Variable eingehen, jedoch nicht im linearen Modell geschätzt werden. Dadurch entsteht eine Verzerrung E der Werte, die bei dieser Simulationsstudie ebenfalls untersucht werden soll. Zur Simulation der Wechselwirkungen wird zunächst die Anzahl an solchen Wechselwirkungen aus der diskreten Gleichverteilung zwischen fünf und zehn gezogen. Für jede Wechselwirkung werden nun wieder zwei Einflüsse jeweils zufällig aus den Einflussparametern ermittelt, die eine Wechselwirkung darstellen sollen. Zu beachten ist, dass hierbei derselbe Einfluss mehrfach gezogen werden kann. In einem solchen Fall liegt dann ein quadratischer Zusammenhang der Variable vor. Das simulierte Modell besitzt dann folgende

Form:

$$y_i = \beta_0 + x_{1;i}\beta_1 + \cdots + x_{d,i}\beta_d + x_{v_1;i}x_{w_1;i}\beta_{1;1} + \cdots + x_{v_u;i}x_{w_u;i}\beta_{u;u} + e,$$

wobei u der Anzahl der Wechselwirkungen, $\beta_{j,j}$ der Parametergröße der j-ten Wechselwirkung und $x_{v_j,i}$ und $x_{w_j,i}$ den entsprechenden i-ten Wert der v_j bzw. w_j Variable ermittelten Wechselwirkung entspricht. Die Parameter der Wechselwirkungen werden dabei auf dieselbe Weise zufällig, wie die übrigen Einflussparameter erstellt.

Schwierigkeitsparameter

Enthält ein Datensatz ausschließlich Beobachtungen, die keine Abweichung von einem linearen Modell beinhalten, also mit einem Fehlerterm, der sich jeweils zu Null ergibt, ist eine optimale Schätzung bereits mit der minimal benötigten Anzahl an Beobachtungen $d + 1$ möglich. Eine systematische Auswahl an geeigneten Beobachtungen, die die Daten gut repräsentieren, ist hierbei unnötig, da keine der Beobachtungspunkte, durch Austausch einer anderen Beobachtung, das Modell bzw. die Schätzung der Parameter ändern würde. Eine optimale Schätzung würde in jedem Fall erreicht werden. Deshalb ist eine Kennzahl für die einzelnen Datensätze sinnvoll, die angibt in welchem Maße sich ein Modell bei Hinzu- bzw. Hinfortnahme eines Elements des Datensatzes ändert. Da die Cooks Distanz ein Maß liefert, das die Änderung der Varianz beim Entfernen einer einzelnen Beobachtung angibt, kann diese als Grundlage für ein solches Schwierigkeitsmaß für das Auffinden von geeigneten Kernmengen herangezogen werden. Sind sämtliche Cooks Distanzen aus einem Datensatz gleich, ist somit auch die Änderung an der Modellvarianz gleich und das Bewertungskriterium 2 liefert in diesem Fall für jede Teilmenge mit $n - 1$ Elementen denselben Wert. Dies bedeutet jedoch nicht, dass die Schätzung der Modellparameter dieselben Schätzwerte ergibt. Unterscheiden sich die berechneten Cooks Distanzen jedoch stark und haben eine hohe Streuung, variiert der Einfluss der einzelnen Beobachtungen. In einem solchen Fall ist die Auswahl einzelner

Beobachtungen entscheidender und die Bildung einer Kernmenge daher schwieriger. Als Maß für die Schwierigkeit wird daher die Varianz bzw. die Standardabweichung der Cooks Distanzen und somit die Variation einzelner Einflüsse von Beobachtungen herangezogen.

Ist dieser Wert für einen Datensatz besonders hoch, unterscheiden sich die Ergebnisse der Kernmengen stark, da einige Werte hohen, andere geringen Einfluss haben und diese in der Teilmenge enthalten oder nicht enthalten sein können. Bei einem kleinen Wert gibt es, auch bei unterschiedlichen Teilmengen des Datensatzes, keine großen Abweichungen der Schätzungen der Modellvarianz und somit der Ähnlichkeit der Güte der Schätzungen, da der Einfluss in etwa bei jeder Beobachtung gleich groß ist. Bei der Berechnung dieser Distanzen wird zudem durch die Varianz des Gesamtmodells geteilt. Deshalb ist bei unterschiedlich hohen Streuungen in den Daten nur der Faktor der Änderung entscheidend und nicht die Höhe der Streuung. Dieses Maß ist deshalb geeignet, um die Schwierigkeit der Auswahl einer Kernmenge für einen der simulierten Datensätze, darzustellen.

Das Maß spiegelt jedoch nur die Änderung anhand jeder einzelnen Beobachtung wieder. In einem Datensatz mit großem Stichprobenumfang ist diese jedoch deutlich geringer, da einzelne Datenpunkte weniger ins Gewicht fallen als in einer Stichprobe mit geringerem Umfang. Ebenso ist zu beachten, dass eine Streuung der Cooks Distanzen von 0 nicht dazu führt, dass jede Kernmenge dasselbe Ergebnis erzielt, sondern nur im Falle von $n - 1$ Elementen in der Teilmenge und nur in Bezug auf das Bewertungskriterium 2. Das Maß bezieht sich nur auf den Gesamtdatensatz. Eine geeignete Auswahl muss auch hier getroffen werden und die Einflüsse variieren je nach Kernmenge ebenfalls.

Um den unterschiedlich großen Distanzen gerecht zu werden, kann etwa die berechnete Standardabweichung durch den Mittelwert der Distanzen geteilt werden, um den Variationskoeffizienten zu ermitteln. Der Variationskoeffizient liefert auch bei verschiedenen Stichprobenumfängen ein geeignetes Maß.

In dem Sonderfall, dass keine Abweichungen der Daten vom Modell vorliegen, wäre der Wert der Einflüsse konstant, da die gleiche Streuung von 0 für jede Kernmenge

erreicht wird. Hierbei wäre eine Auswahl einer systematischen Kernmenge sogar komplett überflüssig und eine Annäherung an das Modell würde optimal erreicht werden, unabhängig von der Wahl der Teilstichprobe. Eine Mindestanzahl an $d+1$ Elementen in der Kernmenge wird allerdings auch hier benötigt, um sämtliche Einflussparameter zu schätzen. Weicht hingegen nur eine einzelne Beobachtung besonders stark vom Modell ab, ändert sich durch diese auch die geschätzten Modellparameter in einem hohen Maß. Die Cooks Distanz dieses Elementes ist hierbei besonders hoch im Gegensatz zur Cooks Distanz der anderen Werte. Dadurch wird die Verzerrung im Modell nur bei einer Kernmenge eingehen, die diese Beobachtung beinhaltet. Eine Auswahl in diesem Fall ist besonders schwierig, der Wert des Variationskoeffizienten ist zudem ebenfalls höher als im vorherigen Fall.

Die oben beschriebenen Verfahren zielen darauf ab, auch bei schwierig zu bildenden Kernmengen, diese durch systematische Ansätze und Auswahlverfahren geeignet zu wählen.

6 Auswertung

Sämtliche Verfahren, Simulationen und Analysen, sowie Abbildungen werden in dieser Arbeit mit `R` durchgeführt (R Development Core Team, 2015). Dabei werden zusätzlich die R-Pakete „RColorBrewer“ (vgl. Neuwirth, 2014), „rlecuyer“ (vgl. Sevcikova et al., 2015) und „xtable“ (vgl. Dahl, 2016) genutzt.

Die simulierten Datensätze spiegeln unterschiedliche Problemstellungen wieder und unterliegen verschiedenen Verzerrungen. Das oben beschriebene Schwierigkeitsmaß variiert daher auch bei den einzelnen Szenarien. Der Durchschnitt der Maße ist in den Tabellen 2 bis 10 im Anhang dargestellt. Dabei erkennen wir direkt, dass in jedem der Fälle die Stichprobengröße des Gesamtdatensatzes keine bedeutende Rolle spielt. Liegt etwa eine große Stichprobengröße vor, ist der Wert jeweils in etwa so groß wie bei kleineren Datensätzen. Da einzelne Beobachtungen weniger ins Gewicht fallen, sind dabei meist ein fester Anteil an stärkeren Ausreißern, die das Modell verzerren, bedeutend. (In einer Kernmenge mit nur wenigen Elementen, führen diese jedoch wiederum zu größeren Verzerrungen.)

Die Werte der ersten beiden Verzerrungen A0 und A1 sind hier nahezu identisch. Dies ist durch die Normierung durch die Varianz des Gesamtmodells zu erklären. Die Wahl der Kernmengen sollte in diesen Verzerrungen etwa denselben Einfluss haben.

Bei der Verzerrung A2 hingegen haben die von den übrigen x-Werten abweichenden Daten deutlich höhere Einflüsse auf die Schätzungen. Hierdurch wird die Bedeutung für die Kernmenge ebenfalls erhöht und somit auch für die Änderung des Faktors der Modellvarianz (der Modelle der Teil- bzw. Gesamtmenge an Beobachtungen).

Liegt Heteroskedastie im Modell vor, sind die Daten im Bereichen mit hohen Schwankungen einflussreicher. Die Höhe der Einflüsse unterscheidet sich also und die Wichtigkeit in der Schätzung variiert daher je nach Beobachtung etwas stärker. Eine sinnvolle Auswahl ist hierbei ebenfalls etwas bedeutender für die Schätzung. Die Datensätze mit hohen Ausreißern, B1 und B2, stellen eine besonders hohe Herausforderung für die Auswahlmethoden dar. Die Maße sind entsprechend deutlich höher, da die jeweiligen Ausreißer

im Gegensatz zu den übrigen Beobachtungen größere Einflüsse aufweisen. Dadurch ist es wichtig, diese in der Kernmengen gut wiederzuspiegeln. Ein gänzliches Fehlen dieser Daten würde zu schlechteren Annäherungen führen, eine Überrepräsentierung jedoch ebenfalls.

Zusätzlich ist zu erkennen, dass bei der Verzerrung B2 der Parameter mit steigender Dimension zu geringeren Werten führt. Da das Gebiet, auf dem die Ausreißer auftreten, relativ klein und die Ausdehnung fest ist, der Gesamtraum der Werte jedoch sehr viel größer, werden bei einer größeren Dimension anteilig jeweils mehr verzerrende Beobachtung dort hinzugenommen. Der Anteil von Ausreißern innerhalb des Gebietes steigt dadurch. Anteil an Ausreißern entspricht 5 Prozent der Gesamtbeobachtungen, das Teilgebiet des Raumes entspricht allerdings nicht dem Anteil der Größe des Gesamtraumes und dieser Anteil fällt mit steigender Dimension. Dadurch wird insbesondere das Modell stärker den Ausreißern in diesem Teilraum angepasst, wodurch die Abweichung dieser Beobachtungen geringer als zuvor ausfällt. Dies hat ebenfalls Einfluss auf die Auswahl, da die Cooks Distanzen ebenfalls kleiner ausfallen.

Die Datensätze der Verzerrung C liefern ebenfalls höhere Maßzahlen als etwa die der Verzerrungen A0. In dem Fall, dass Daten zwei unterschiedlichen Modellen unterliegen, wird ein Modell zwischen diesen beiden Modellen geschätzt. Dies führt dazu, dass einige Punkte, je nach Modellen, eine hohe Abweichung darstellen, andere, insbesondere an eventuellen Schnittstellen der Modelle, geringe Residuen ergeben. Die Einflüsse variieren daher auch etwas stärker und das Maß ergibt so einen höheren Wert. Dies ist somit mit dem Fall der Heteroskedastie zu vergleichen, der ähnliche Maßgrößen liefert. Bei einem sinusförmigen Zusammenhang sind die Abweichungen ebenfalls in einigen Abschnitten gering und in anderen höher. Da jedoch die Einflüsse und Abweichungen systematisch variieren und somit keine deutlichen Ausreißer erzeugt werden, ist der Wert des Schwierigkeitsmaßes ähnlich hoch, wie das der ersten Szenarien (A0 und A1).

Zuletzt wird die Situation eines Modelles mit nicht erkannten Wechselwirkungen betrachtet. Liegen x-Werte in einem hohen oder sehr niedrigen Bereich der Einflussvariable und tragen zur Verzerrung mit der Wechselwirkung bei, ist deren Einfluss besonders

hoch. Ein Wertebereich nahe des Nullpunktes bewirkt hingegen nur geringeren Einfluss auf das geschätzte Modell, da die Verzerrung und somit die Abweichung vom Modell in diesem Teilraum gering ist. Die Variation der Einflüsse ist in diesem Fall ist mit einem durchschnittlichen Wert des Maßes von 1.85 bei einer Stichprobe von $n = 1000$ und einer Anzahl an Einflussvariablen von $d = 5$ im Vergleich relativ hoch. Mit steigender Dimension sind jedoch die Wechselwirkungen im Schnitt nicht mehr so unterschiedlich einflussreich, da die Anzahl der Wechselwirkungen im Schnitt gleich bleibt. Daher verringert sich der Parameter etwas bei steigender Dimension. Feste Anzahl an Wechselwirkungen bei steigenden Dimension bedeutet weniger Verzerrung im Schnitt, da mehr einzelne Parameter unverzerrt, ohne Wechselwirkung sind.

Im Folgenden werden die Ergebnisse, der Auswahl der Kernmengen der einzelnen Algorithmen, anhand der Bewertungskriterien berechnet und miteinander verglichen. Die Anzahl der Elemente r, die jeweils in einer Kernmenge auftreten, werden hier zusätzlich variiert. Eine Auswahlmenge r von 50, 100, 200 und 500 Beobachtungen werden jeweils mit Hilfe der Verfahren ausgewählt. Im Falle eines Gesamtdatensatzes von $n = 5000$ Beobachtungen, werden zudem Teilstichproben mit 1000 Elementen gezogen.

Zunächst werden die simulierten Daten, die mit der Verzerrung A0 simuliert werden, betrachtet. Die Ergebnisse werden dabei in allen der simulierten Fällen über alle 100 Wiederholungen (100 erzeugte Datensätze je Verzerrung) gemittelt, um den Zufall in den Ergebnissen zu minimieren und eine robusteren Wert zu erhalten. Diese Werte sind in Tabelle 11 bis 17 auf Seite 71f. im Anhang dargestellt. Man erkennt direkt, dass die Werte bei größeren Kernmengen abfallen, also eine bessere Anpassung liefern. Dies ist zu erwarten, da ein größerer Anteil am Gesamtdatensatzes, diesen besser widerspiegelt, wichtige Datenpunkte mit höherer Wahrscheinlichkeit hinzugezogen werden und Gruppeneinflüsse sowie einzelne starke Abweichungen besser abgebildet bzw. relativiert werden.
Dabei ist weiter zu erkennen, dass der Clusterscore-Algorithmus die niedrigsten Kos-

tenkriterien erzielt und somit die höchste Güte und Anpassung zum Modell mit dem Gesamtdatensatz besitzt. Die Kernmengen, die durch dieses Verfahren ermittelt werden, sind also in diesem Szenario am besten. Ebenso erkennen wir direkt, dass auch die MDA- und die DEV- Methode Werte erzielen, die sehr gering sind und lassen somit auf eine gute Annäherung zum Modell schließen lassen. Diese Auswahlverfahren erzielen deutlich bessere Ergebnisse als die Zufallsauswahl.

Der Salp- und KM++ Initialisierungs-Algorithmus hingegen schließen dabei, je nach Gesamtstichprobengröße und Anzahl an Einflussfaktoren im Modell, nur geringfügig besser als die zufällige Auswahl ab, in Einzelfällen sogar etwas schlechter. Die Abweichung der durchschnittlichen Kriterien sind dabei allerdings nahezu gleich.

Die k-Means-Auswahl liefert in diesem Fall eine vergleichsweise geringe Anpassung an das Modell, das aufgrund des Gesamtdatensatzes geschätzt wird. Besonders bei niedriger Anzahl an Beobachtungen in der Teildatenmenge (r = 50,100), sind hohe Kosten zu beobachten. Diese Methoden liefern in der Einstellung sogar ungünstigere Kernmengen als eine zufällige Auswahl an Beobachtungen. Eine Erklärung hierfür ist, dass die Variationsbreite der einzelnen Einflussvariablen hierbei nicht vollständig erreicht wird. Da jeweils ein Wert in der Mitte eines Clusters nahe dem Mittelwert einer Gruppe gewählt wird, sind Randpunkte bzw. Eckpunkte aus dem Gesamtdatensatz mit geringerer Wahrscheinlichkeit in der Teilstichprobe vorhanden.

Die gesamte Spannweite der Einflussvariablen wird daher, mit hoher Wahrscheinlichkeit, nicht in der Auswahl der Datenpunkte erreicht. Ist die Anzahl der Beobachtungen in der Teilstichprobe zudem sehr gering, werden auch weniger Cluster mit dem k-Means gebildet. Es liegen somit auch mehr Punkte innerhalb einer Gruppierung und die Gruppe wird in ihrer Ausdehnung, größer. Dies führt weiter dazu, dass die Mittelwerte weiter in der Mitte des Gesamtdatensatzes liegen und somit der oben beschriebene Effekt weiter verstärkt wird. Das Verfahren liefert daher bei kleinem r deutlich schlechtere Werte, die sogar höher als die der Zufallsauswahl liegen.

Da das KM++V eine initiale Annäherung an Klassenmittelpunkte bei Clusterverfahren darstellt, ist dieser Effekt ebenso bei diesem von Bedeutung. Jedoch wird durch das

zufällige Ziehen anhand einer proportionalen zur Distanz der schon gewählten Beobachtungen ermittelten Wahrscheinlichkeit, Punkte am Rand etwas begünstigt und die Auswahl so, besonders bei kleinen Teilmengen, etwas verbessert. Abbildung 14 stellt die Ergebnisse der Auswahlverfahren beispielhaft für den Fall $n = 1000$ und $d = 5$ für das Kriterium 2 dar.

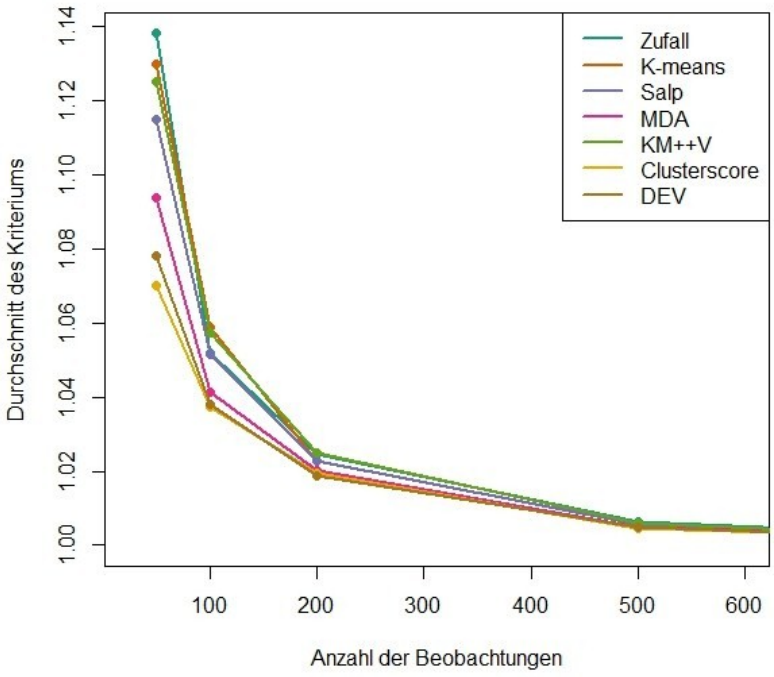

Abbildung 14: Durchschnittliche Ergebnisse des Kriteriums 2 der Auswahlverfahren bei Verzerrung A0. Die Anzahl Elemente in der Kernmenge ist jeweils $r =$ 50, 100, 200 und 500.

Liegt eine Verzerrung der Form A1 vor, zeigt sich bei den Ergebnisse ein ähnliches Bild wie zuvor. Da der Schwierigkeitsparameter in diesem Fall ähnliche Werte annimmt, ergeben sich für das Kriterium 2 ebenso nahezu gleiche Wertebereiche für die Zufallsauswahl wie in der vorherigen Untersuchung.

Der clusterscore- und DEV-Algorithmus schneiden hierbei wieder am besten ab. Auch bei kleiner Anzahl an gewählten Beobachtungen liefern diese Algorithmen deutlich bessere Ergebnisse als eine zufällige Auswahl. Da wieder eine Verzerrung aufgrund von

einer Varianz zum Modell vorliegt und so die Modellgleichung aus Kapitel 3 erfüllt ist, diese Verzerrung jedoch durch die Distanzen der x-Werte der Beobachtungen zueinander relativiert werden, begünstigt dies ebenfalls Methoden, die die Spannweite somit vollständig wiederspiegeln. Aus diesem Grund schneidet der MDA hier wiederum relativ erfolgreich ab.
Ebenfalls liefert der Salp niedrigere Werte und somit eine bessere Anpassung als die Zufallsmethode. Diese liegen allerdings wiederum nur geringfügig unterhalb der Werte des Zufalls. Zu beachten ist auch hier, dass Leverage Score Werte am Rand des Raumes, außerhalb der übrigen Werte, höher sind und somit bei der Auswahl begünstigt werden. Durch das Ziehen mit entsprechender Wahrscheinlichkeit wird der Effekt der Spannweite auf die Abweichungen eingehen, jedoch nur mit höherer Wahrscheinlichkeit, nicht deterministisch wie etwa bei der Maximal-Distanz-Methode.
Sowohl die Werte für das KM++V als auch der k-Means-Algorithmus lassen auf eine vergleichsweise schlechte Anpassung der Auswahlen schließen. Diese Werte liegen in einigen der Fällen sogar oberhalb derer der Zufallsauswahl. Aus oben genannten Überlegungen ist dies wieder zu erklären und der Effekt kommt bei kleiner Teilstichprobengröße deutlicher zu tragen. Die Rangreihenfolge der Werte der einzelnen Auswahlverfahren ändern sich bei unterschiedlichen Gesamtstichprobenumfängen bzw. bei verschiedenen Dimension kaum. Die sehr geringfügigen Änderungen, die dabei auftreten, sind auf zufällige Schwankungen innerhalb der simulierten Daten zurückzuführen.
Die Ergebnisse der Verfahren unter diesen Voraussetzungen sind in Tabelle 18 bis 17 auf Seite 71f. des Anhanges zu erkennen.

Die Datensätze, die mit der Verzerrung A2 simuliert werden, liefern hingegen ein etwas anderes Bild. Zunächst ist zu beachten, dass hierbei die Daten einer kleinen Gruppe von x-Werten deutlich größer sind (siehe oben). Dies führt zu entsprechend sehr hohen Werten für die Leveragescores in dieser Gruppe von Beobachtungen.
Anhand der Ergebnisse des ersten Kriteriums erkennt man, dass die Auswahlverfahren im Vergleich zum Zufall besser abschneiden. Lediglich der Algorithmus des k-means und das KM++V fallen in kleinen Bereichen höher aus. Das Clusterscore- und das

deterministische Epsilonverfahren liefern wiederum die niedrigsten Durchschnittswerte.

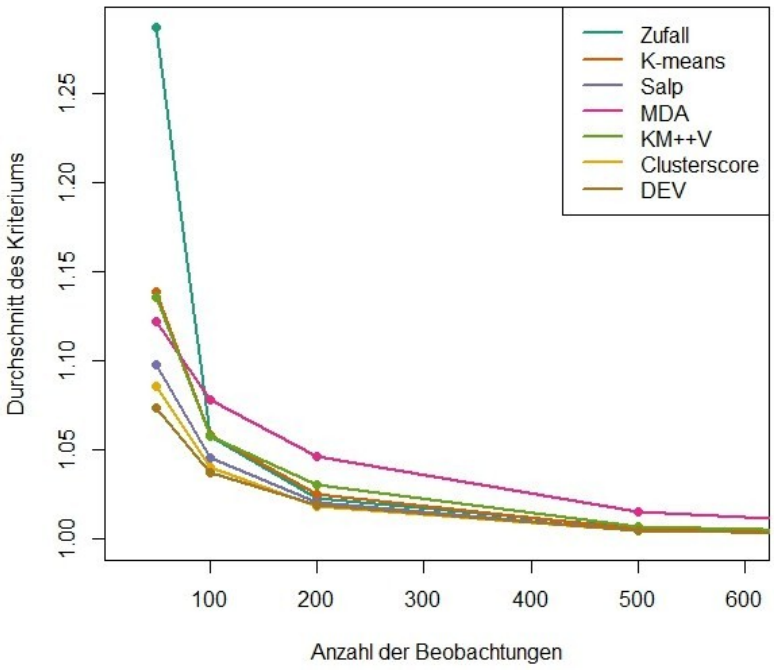

Abbildung 15: Durchschnittliche Ergebnisse des Kriteriums 2 der Auswahlverfahren bei Verzerrung A2. Die Anzahl Elemente in der Kernmenge ist jeweils $r = 50$, 100, 200 und 500.

Betrachtet man nun jedoch das zweite Kriterium, das in Abbildung 15 repräsentativ für die Einstellung $n = 1000$ und $d = 5$ für die Verzerrung A2 abgebildet ist, ergibt sich hingegen ein etwas anderes Bild. Insbesondere der MDA liefert zunächst bei kleiner Auswahlmenge noch relativ gute Werte. Diese fallen allerdings im Vergleich zur Zufallsauswahl, bei höherer Anzahl an Elementen in der Kernmenge, deutlich schlechter aus. Die Zufallsauswahl selbst liefert Kernmengen, die bei geringer Anzahl an Elementen zu sehr schlechten Anpassungen führen.

Die Ergebnisse lassen daher rückschließen, dass die angewendeten Kriterien nicht immer zu denselben Rangfolgen der Güte der Verfahren führen. Dies ist plausibel, da im ersten Kriterium die Abweichungen der Einflussgrößen jeweils quadratisch eingehen, bevor die

Wurzel der Summe dieser berechnet wird. Eine hoher Wert kann daher auch von der Abweichung eines einzelnen geschätzten Parameters entstehen. Das zweite Kriterium berücksichtigt hingegen nur den Faktor für die Änderung in den Modellvarianzen. Dies führt in diesem Fall zu unterschiedlichen Rangfolgen, da die Einflussvariablen, trotz stärkerer Änderungen der Streuungen, trotzdem gleichmäßiger geschätzt werden und die Abweichungen vom Gesamtmodell nicht vorwiegend aus einer einzelnen, schlecht geschätzten Variable stammen.

Bei dem Maximal-Distanz-Verfahren wird jeweils die Beobachtung gewählt, die weit von den bereits gezogenen Beobachtungen entfernt liegt. In den Datensätzen liegt eine Gruppe an Daten weit entfernt von den übrigen. Jedoch liegen die Beobachtungen in dieser Gruppe deutlich näher zusammen und verteilen sich daher über einen kleineren Teilraum. Die Distanzen innerhalb der Gruppe sind entsprechend kleiner. Durch diese Auswahl wird deshalb nur ein einzelner bzw. sehr wenige Werte aus diesem Cluster gewählt. Da diese Beobachtung jedoch sehr große Hebelwirkung und damit einen hohen Einfluss auf die Schätzung aufweist, hängt die Schätzung besonders von diesem Wert ab und somit von der Abweichung, die in diesem Punkt vorliegt. Auch bei einer größeren Anzahl an auszuwählenden Datenpunkten hängt die Schätzung ebenfalls stark von der Abweichung dieser weniger einzelner Punkte ab. Somit verbessert sich die Anpassung an das Modell auch mit mehr Beobachtungen in der Kernmenge nur wenig. Da die Zufallsauswahl alle Punkte mit gleicher Wahrscheinlichkeit auswählt, kann es vorkommen, dass kein Punkt aus dem abweichenden Cluster ausgewählt wird. Dies macht die Schätzung daher deutlich unsicherer, da die Punkte den höchsten Einfluss auf das geschätzte Modell haben. Bei der Auswertung der Kernmengen der Zufallsauswahl sind diese einzelnen Fälle deutliche Ausreißer in den Ergebnissen und führen zu einem durchschnittlich sehr schlechtem Wert besonders bei kleiner Anzahl an ausgewählten Elementen. Die Abbildung 16 zeigt für die Zufallsauswahl die Ergebnisse des Kriteriums aller 100 simulierter Durchläufe anhand eines Boxplots. Dabei sind deutlich Ausreißer in den Werten bei einem r von 50 zu erkennen. Die Anzahl der Ausreißer wird jedoch mit steigender Anzahl an Elementen in der Kernmenge seltener. Zur besseren Darstellung wurde ein zusätzlicher Ausreißer

mit dem Wert 13.08 bei einer Auswahl von 50 Elementen in der Kernmenge aus der Grafik entfernt.

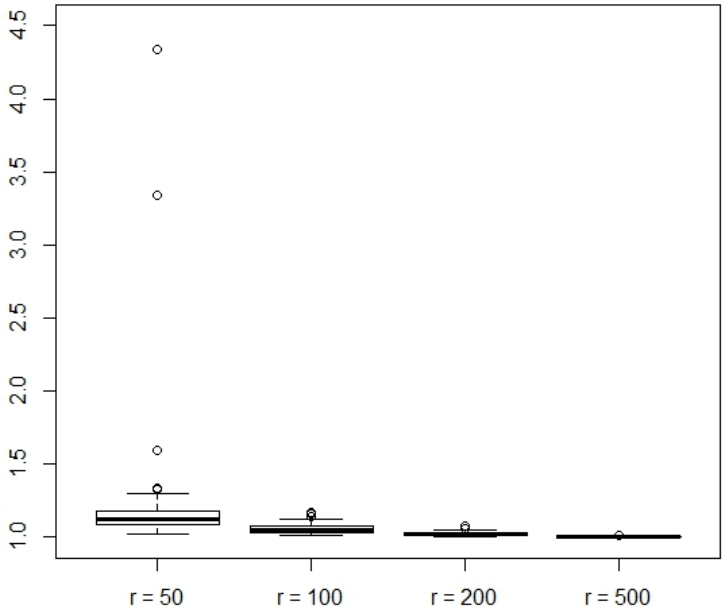

Abbildung 16: Boxplot der einzelnen Ergebnisse der Kriteriums 2 der Zufallsauswahl im Falle A2.

Zudem wird die Spannweite der Einflussvariablen ebenfalls nicht vollständig wiedergespiegelt. Werden jedoch viele Daten für die Teilstichprobe ausgewählt, tritt dieser Fall mit geringerer Wahrscheinlichkeit auf. Die Güte der Anpassung verbessert sich daher und die Werte der Kriterien sind geringer.

Beim Klassifizieren der Daten mit Hilfe des k-Meansverfahrens ist mindestens eine der Klassen in den Bereich der abweichenden Daten angesiedelt. Jedoch ist es durch die in Kapitel 4 beschriebene Iterationen möglich, dass auch ein Wert aus diesem Bereich einer anderen Klasse zugeordnet wird. Dies hängt von der Startverteilung des k-Means ab und führt zu einer weiteren Klasse in diesem Wertebereich. Es ist also immer mindestens ein Cluster, häufig aber auch mehrere, in diesem Teilraum. Da jeweils eine Beobachtung

pro Klasse ausgewählt wird, liefert der Algorithmus deshalb eine Kernmenge mit entsprechenden Beobachtungen aus diesen Regionen, was zu einer besseren Annäherung an das Modell führt, das mit Hilfe aller Daten geschätzt wird.
Die übrigen Verfahren, insbesondere das Clusterscore und das deterministische Epsilonverfahren, liefern Werte die wiederum auf eine gute Auswahl schließen lassen. Beim Clusterscore ist dies ähnlich wie beim k-Means zu erklären, jedoch werden zusätzlich noch die bedeutenderen Beobachtung je Gruppe gewählt und somit zusätzlich die komplette Variation der Werte der Einflussvariablen ausgenutzt.
Speziell das DEV liefert hier die besten Ergebnisse, da zunächst aus der oben beschrieben Gruppe, aufgrund ihrer hohen Hebelwirkungen, immer Daten zuerst ausgewählt werden. Zudem liegt der Anteil der ausgewählten Daten aus dieser Gruppe in der Kernmenge, aufgrund des anteiligen Ausschlusses der übrigen Daten, in etwa bei dem der Gruppe im Gesamtdatensatz. Der Salp hat bei der Auswahl ebenfalls einige dieser Daten in die Kernmenge ausgewählt. Mit hoher Wahrscheinlichkeit werden aus der genannten Klasse mehrfach Beobachtungen ausgewählt. Außerdem sind aufgrund der Vielzahl der Elemente in der anderen Gruppe, ebenfalls Beobachtungen aus dieser ausgewählt, wodurch eine bessere Anpassung vorgenommen werden kann.

Weichen die Beobachtungen in den Datensätzen heteroskedastisch ab, liegt die Verzerrung A3 vor. In diesem Fall ähneln die Ergebnisse der untersuchten Verfahren stark denen der Verzerrung A0 und A1. Die Rangfolge ändert sich dabei auch bei unterschiedlichen Einstellungen für die Dimension und die Gesamtstichprobengröße kaum für das zweite Bewertungskriterium.
Es liegen unterschiedliche Schwankungen, die sich auch je Einflussvariable unterscheiden vor. Je höher der Einfluss einee Variable auf das Modell ist, desto höher werden die simulierten y-Werte und somit auch die Schwankung. Die Schätzung der einzelnen Einflussgrößen ist hierbei also unterschiedlich schwer und es kann zu Abweichungen dieser Größen kommen, die dadurch mal stark und mal weniger stark ausfallen. Dadurch schwanken allerdings auch die Werte des Kriteriums 1 und die Rangreihenfolge fällt hier deshalb in einzelnen Fällen geringfügig verschieden aus. Die durchschnittlichen

Ergebnisse der ausgewählten Kernmengen sind in Tabellen 25 bis 31 auf Seite 75f. im Anhang festgehalten.

Bei der Verzerrung B1 ergeben die Bewertungskriterien ebenfalls eine ähnliche Rangreihenfolge. Die Werte des zweiten Kriteriums liegen allerdings aufgrund der größeren Schwierigkeit für die Zufallsauswahl etwas höher. Die Auswahlen ergeben also insgesamt schlechtere Anpassungen. Die Rangreihenfolge der Werte der Methoden ändert sich über die unterschiedlichen Dimensionen und die maximale Stichprobengröße, sowie für die unterschiedlichen Bewertungskriterien dabei nur wenig und nur bei einzelnen unterschiedlichen Größen der Kernmengen. Diese sind aufgrund der sehr geringfügigen Änderungen in den Werten diese auf den Zufall zurückzuführen. Die Abbildung 17 zeigt das durchschnittliche Kriterium 2 beispielhaft für den Fall $n = 1000$ und $d = 10$.

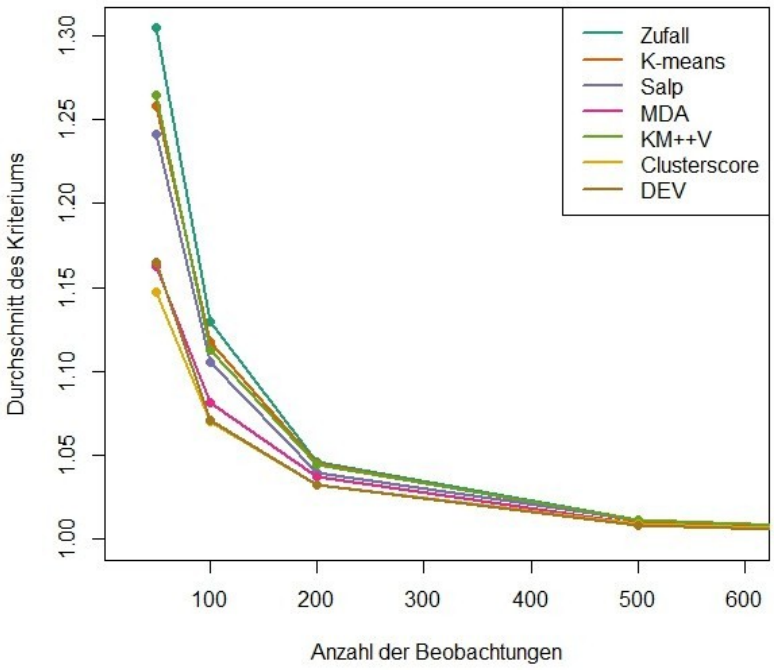

Abbildung 17: Durchschnittliche Ergebnisse des Kriteriums 2 der Auswahlverfahren bei Verzerrung B1. Die Anzahl Elemente in der Kernmenge ist jeweils $r =$ 50, 100, 200 und 500.

Bei der Betrachtung ist sofort zu erkennen, dass der Clusterscorealgorithmus und das DEV die besten Anpassungen liefern. Auch der MDA erbringt sehr gute Ergebnisse, die nur wenig schlechter sind. Alle diese Verfahren nutzen dabei die gesamte Bandbreite an Variation der einzelnen Variablen. Dies ist auch in diesem Fall von Bedeutung, da simulierte Ausreißer an jeder Stelle des Raumes auftreten können. Durch maximale Distanzen werden diese jedoch wieder relativiert und eine entsprechend kleinerer Fehler tritt in der Schätzung der Parameter auf. Fehlen die Ausreißer in der Kernmenge, so wird die Verzerrung jedoch nicht in der Teilstichprobe wiedergespiegelt und geht so nicht in die Schätzung mit ein, was dazu führt, dass das Modell, das mit allen Beobachtungen geschätzt wird, nicht vollständig abgebildet wird. Dies ist jedoch für jede der Methoden der Fall, da die Ausreißer nicht systematisch, durch falsche Modellannahmen oder etwa an besonders wichtigen/ besonderen Bereichen des Wertebereichs auftreten, sondern zufällig verteilt sind.
Insbesondere das Clusterscoreverfahren schneidet im Vergleich hierbei gut ab, da die Auswahl eines vermutlichen Ausreißers, der besonders hohen Einfluss auf die Schätzung hat, nach Vorauswahl (vgl. Kapitel 4) nochmal überprüft wird. Dadurch wird ggf. durch die Hinzunahme eines Nicht-Ausreißers der Einfluss dieser einzelnen Beobachtung gemindert und evtl. Gruppeneinflüsse, die so auftreten können, berücksichtigt. Die weitere Wahl eines Ausreißers könnte hier die Schätzung verschlechtern. Diese ist jedoch aufgrund des relativ geringen Anteils an Ausreißern entsprechend unwahrscheinlich. Die übrigen Auswahlverfahren, der Salp, KM++V und k-means, liefern Werte, die auf eine deutlich schlechtere Annäherung schließen lassen. Sie liefern Kernmengen, die in diesem Szenario kaum bessere, teilweise sogar geringe Güte der Anpassungen erzielen. Die vollständigen durchschnittlichen Werte der Kriterien für diese Verzerrungsform ist auf Seite 39 bis 45 im Anhang in Form einer Tabelle dargestellt.

Unterliegen die simulierten Datensätze der Verzerrungsart B2, so ergibt sich ein interessantes Bild. Die Grafik 18 stellt diese besonderen Ergebnisse für den Fall $n = 1000$ und $d = 5$ dar. Eine gesamte Darstellung der Werte ist in Tabelle 46 bis 52 auf Seite 81f. im Anhang zu sehen.

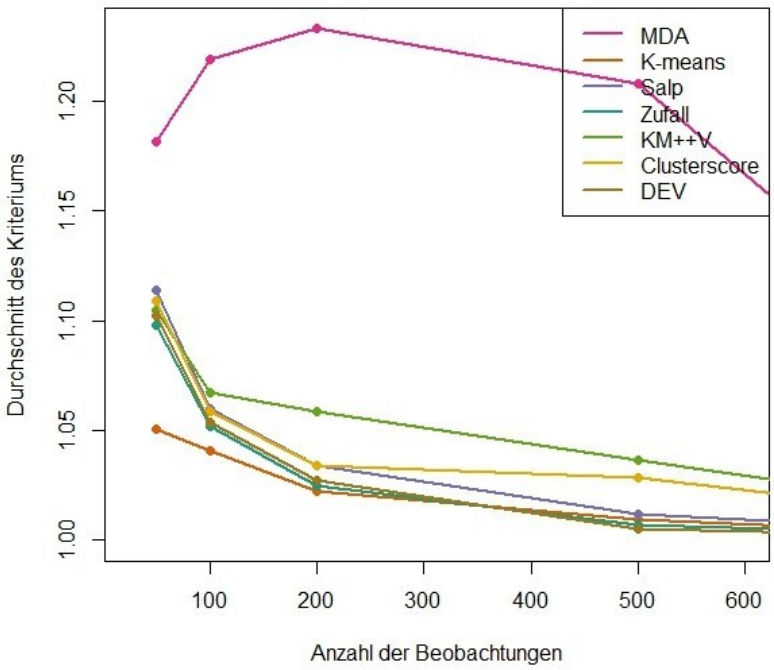

Abbildung 18: Durchschnittliche Ergebnisse des Kriteriums 2 der Auswahlverfahren bei Verzerrung B2. Die Anzahl Elemente in der Kernmenge ist jeweils $r = 50$, 100, 200 und 500.

Man sieht sofort, dass hier deutlich höhere Werte für das Kostenkriterium bei größeren Kernmengen erzielt werden. Zudem fällt besonders auf, dass einige der Verfahren, bei einer höheren Anzahl an Elementen der Kernmenge, zu schlechteren Anpassungen führen als bei einer niedrigen Anzahl.

Auch schließt hier der k-Means im Falle weniger Beobachtungen in der Auswahlmenge im Vergleich relativ gut ab. In diesem Szenario werden Datenpunkte mit hohen (positiven) Abweichungen in einem relativ kleinen Wertebereich, zusätzlich zu den übrigen Daten erzeugt, um die Verzerrung zu simulieren. Dadurch entstehen, besonders viele Punkte in diesem Bereich, wodurch mit hoher Wahrscheinlichkeit mindestens eines der Klassenzentren des k-Means in diesem Bereich liegt. Dadurch wird auch mit hoher Wahrscheinlichkeit mindestens einer der Ausreißer dort in die Kernmenge aufgenommen.

Es liegen allerdings auch Nicht-Ausreißer in diesem Teilraum, die so ausgewählt werden können. Bei einer geringen Anzahl auszuwählender Objekte entspricht dies in etwa dem Anteil an verzerrenden Beobachtungen, der auch im Gesamtdatensatz vorzufinden ist. Bei hoher Auswahlmenge wird dieser Teilbereich jedoch wieder unterrepräsentiert, da nicht unbedingt proportional mehr Cluster in dem kleinen Bereich entstehen.
Die Anzahl der Punkte aus dem Bereich haben bei der Zufallsauswahl erwartungsgemäß ebenfalls denselben Anteil, jedoch wird nicht immer eine dieser Beobachtungen gezogen. Dies führt (mit den gleichen Überlegungen, wie bei Verzerrung A3) zu sehr hohen Bewertungskriterien, wenn kein Ausreißer ausgewählt wird, da die Verzerrung in einem solchen Fall nicht abgebildet wird. Mit hohem r wird die Wahrscheinlichkeit hierfür jedoch sehr viel geringer und die Ergebnisse bessern sich mit steigender Kernmengengröße daher deutlich. Da die Anteile an Ausreißern in der Gesamtstichprobe im Erwartungswert immer dem in der Kernmenge entspricht, liefert das Verfahren im Vergleich relativ gute Werte.
Der Clusterscorealgorithmus nutzt zuerst ebenfalls die Klassifizierung mit Hilfe des k-Means. Die anschließende Auswahl wird allerdings mit Hilfe der Leveragescores in jedem der Cluster gewählt. Da der Bereich, in dem die verzerrenden Beobachtungen liegen, sehr klein ist, werden bei der Klassifizierung mit hoher Wahrscheinlichkeit zusätzlich zu diesen Beobachtungen auch weitere Punkte außerhalb dieses Teilraumes zu der Gruppe klassifiziert. Da hohe Leveragescores eher am Rande des Wertebereiches der Einflussvariablen liegen, werden die Ausreißer nur dann ausgewählt, wenn sie am Rand der Klasse liegen und die Verzerrung nur in diesem Fall berücksichtigt.
Bei steigender Anzahl auszuwählender Punkte werden, wie beim k-Means, wieder mehr Cluster gebildet. Somit steigt zunächst auch die Wahrscheinlichkeit einen Ausreißer und somit einen wichtigen Punkt zur Abbildung der Verzerrung im Modell in die Kernmenge aufzunehmen. Insbesondere werden weniger zusätzliche Nicht-Ausreißer in das interessierende Cluster klassifiziert. Bei hoher Anzahl werden allerdings wieder nicht proportional zu dem Anteil Punkte in dem Bereich viele Cluster in diesem Bereich gebildet, was zu der nur geringen Verbesserung führt. In vielen Fällen liegt der Algorithmus daher mit

seinen Ergebnisse unter der Güte der Zufallsauswahl.
Der Salp erzielt aufgrund der Auswahl mit Hilfe von Leverage Scores relativ gute Werte. Jedoch liegen diese Werte auch bei diesem Verfahren über denen der zufälligen Auswahl, was auf eine durchschnittlich schlechtere Anpassung als die mit einer zufälligen Auswahl an Beobachtungen schließen lässt. Dies liegt besonders an der Position der Gruppe der verzerrenden Beobachtungen. Bei hoher Hebelwirkung in dieser Gruppe werden diese mit hoher Wahrscheinlichkeit häufiger ausgewählt und überrepräsentieren so die Verzerrung innerhalb der Auswahlmenge. Andernfalls werden die Punkte kaum oder gar nicht berücksichtigt, was zum Fehlen der Verzerrung in der Teilstichprobe führt.
Betrachtet man die Werte des DEVs so erkennt man, dass diese teilweise besser, teilweise schlechter als der Zufall abschneiden, jedoch im Vergleich zu den anderen Verfahren recht gute Werte erzielt. Da dieser Algorithmus in jedem Schritt einen festen Wert an Beobachtungen ausschließt, ist die Größe des oben beschriebenen Bereiches nicht so stark von Bedeutung. Da hier sehr viele Punkte liegen, werden so auch entsprechend viele Punkte ausgewählt. Der Anteil verzerrender Beobachtungen in der zu wählenden Teilstichprobe entspricht daher in etwa dem in der Gesamtstichprobe.
Sowohl der Initialschritt des KM++ Algorithmus als auch der MDA verwenden Distanzen der Beobachtungen zueinander. In dem einen Fall ist die Auswahl probabilistisch in dem anderen deterministisch. Dadurch, dass das Teilgebiet mit Ausreißern relativ klein ist, haben diese x-Werte entsprechend kleine Abstände. Dies bedeutet insbesondere, dass, falls ein Wert bereits aus diesem Bereich gezogen wurde, nur mit kleiner Wahrscheinlichkeit ein weiteren Wert aus diesem Bereich ausgewählt wird. Es kann sogar vorkommen, dass keine der Beobachtungen, etwa durch „Überspringen“ des Bereiches oder eine Wahl eines Nicht-Ausreißers in diesem Bereich, in die Kernmenge aufgenommen wird. Dieser Teilbereich, der für die Abbildung der Verzerrung jedoch wichtig ist, wird daher unterrepäsentiert bzw. in einzelnen Fällen gar nicht aufgenommen. Wird bei kleinem Anteil von auszuwählenden Beobachtungen häufig zumindest eine dieser Beobachtung ausgewählt, so entspricht dies bei kleiner Anzahl an Elementen eher dem Anteil Ausreißern in der Gesamtstichprobe. Werden jedoch sehr viele Punkte ausgewählt,

und ebenfalls nur ein Ausreißer berücksichtigt, ist der Anteil entsprechend noch geringer und die Verzerrung wird mit einem geringeren Anteil abgebildet. Dies führt zu einem Ansteigen der Werte der Kriterien, die auch bei kleinem r schon deutlich über denen des Zufalls liegen. Die Rangfolge der Bewertungskriterien ändert sich je nach Dimension und Gesamtstichprobenumfang dabei kaum. Da bei diesen Voraussetzungen die Verteilung der x-Werte stark von einer Gleichverteilung abweicht und einige der Bereiche des Werteraumes mit sehr vielen, andere mit wenigen Beobachtungen simuliert werden, ist eine Gewichtung in diesem Fall sinnvoll. Diese kann dazu führen, dass über- bzw. unterrepräsentierte Bereiche weniger bzw. stärker berücksichtigt werden. So könnten die Ergebnisse der Auswahlverfahren mit zusätzlichen Informationen über die Menge zu repräsentierender Elemente je ausgewähltem Punkt verbessert werden.
Zu beachten ist hier allerdings, dass nicht jeder Algorithmus mit einer Gewichtung versehen ist. Da beispielsweise der deterministische Epsilonalgorithmus jeweils einen festen Anteil an Beobachtungen in jedem Schritt ausschließt, kann ein gewählter Wert als Repräsentant genau dieser Beobachtung verstanden werden und liefert so bereits eine Form der Gewichtung.
Eine Veranschaulichung der Ergebnisse der Verfahren ist die Grafik 19 für den Fall $n = 1000$ und $d = 5$.

Zum besseren Vergleich mit der Zufallsauswahl, des Salp und DEV werden diese Ergebnisse ebenfalls ungewichtet hinzugenommen. Die gesamten Ergebnisse sind in Tabelle 53bis 56 im Anhang auf Seite 53 festgehalten. Man erkennt sofort, dass sich die Werte deutlich durch die Gewichtung verbessern und eine genauere Anpassung in den Fällen erreicht wird, als zuvor ohne Gewichtungen.
Der k-Means erreicht dabei die besten Werte. Dies ist durch die obigen Überlegungen und die zusätzliche Gewichtung der wenigen Punkte innerhalb eines großen Clusters (und damit einer hohen Vervielfachungsrate des Punktes in der Gewichtung) zu erklären. Das Verfahren liefert mit Hilfe der gewichteten Kernmengen nun in jedem Fall eine bessere Anpassung als die zufällige Teilstichprobe.
Das Maximal-Distanz- und das Clusterscoreverfahren ergeben ebenfalls bessere Werte

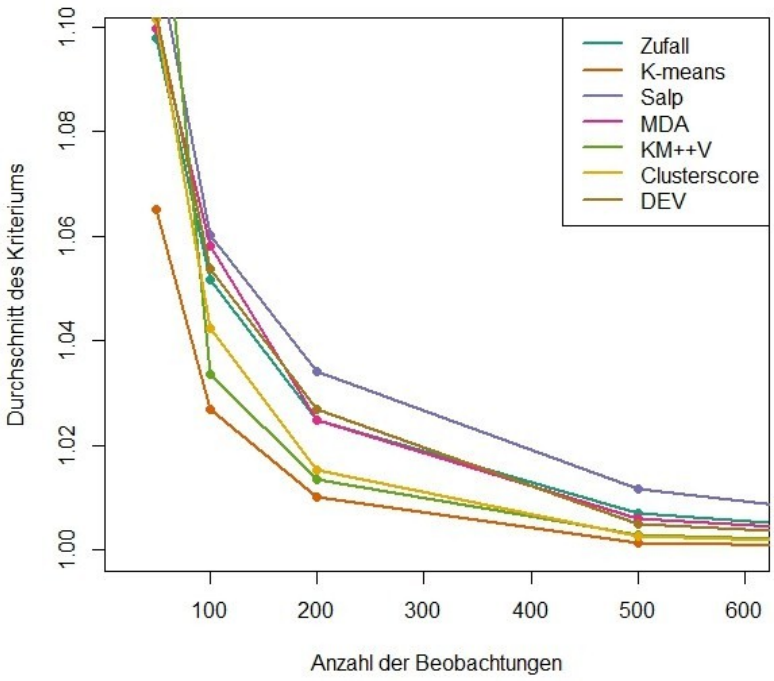

Abbildung 19: Durchschnittliche Ergebnisse des Kriteriums 2 der Auswahlverfahren mit Gewichtung bei Verzerrung B2. Die Anzahl Elemente in der Kernmenge ist jeweils $r = 50$, 100, 200 und 500.

als ohne eine passende Gewichtung der Kernmenge. Bei kleiner Anzahl an auszuwählenden Objekten ist die Auswahl in einigen Fällen jedoch weiterhin schlechter als der Zufall. Dies liegt daran, dass in diesen Fällen die Wahrscheinlichkeit höher ist, den Bereich der Ausreißer zu „überspringen“ bzw. einen Bereich zu klassifizieren, in dem zwar die Ausreißer enthalten sind, jedoch der ausgewählte Punkt am Rand des Clusters liegt und so evtl. nicht zu der Gruppe der Ausreißer gehört. (vgl. obige Überlegungen.) Bei kleinem Teilstichprobenumfang kann es in diesen Fällen dazu kommen, dass die Verzerrung nicht berücksichtigt wird. Insbesondere bei einer hohen Anzahl an Beobachtungen im gesamten Datensatz werden die Klassen mit Hilfe des k-Means nicht klar getrennt, da mehrere zusätzliche Beobachtungen am Rande des Clusters hinzukommen. Ebenso kann es mit höherer Wahrscheinlichkeit dazu führen, dass Punkte beim MDA hinzugenommen werden, die näher am Bereich der verzerrenden Punkte liegt. Dies ist so, da mehrere Punkte im gesamten Raum verteilt sind und die Daten dichter

zusammen liegen. Dadurch wird der Bereich entsprechend mit einer höheren Chance „übersprungen“. Dies führt sowohl beim Clusterscoreverfahren als auch beim MDA bei einem großen Gesamtstichprobenumfang zu schlechteren Werten, die teilweise sogar höhere Werte als der Zufall ergeben. Im Falle von mehreren Einflussvariablen wird dieser Effekt jedoch wieder etwas abgeschwächt. Insbesondere das Clusterscoreverfahren liefert hierbei wiederum bessere Ergebnisse.
Durch die große Anzahl an Beobachtungen in einem sehr kleinen Bereich wird dieser beim probabilistischen Auswählen mit Hilfe proportional zu den Distanzen ermittelten Wahrscheinlichkeiten auch mit höherer Wahrscheinlichkeit berücksichtigt, da sehr viele Distanzen ermittelt werden und diese größer sind, falls noch kein Element aus dem Bereich gezogen wurde. Die Kernmengen des KM++V schließen deshalb bedeutend besser ab als die des MDA und auch häufig des Zufallsalgorithmus. Zudem ist zu erwähnen, dass in jedem Fall eine größere Anzahl an auszuwählenden Datenpunkten zu einer entsprechend besseren Anpassung führt.
Durch den hohen Schwierigkeitsfaktor der simulierten Datensätze sind die Ergebnisse entsprechend höher als in „einfacheren“ Szenarien.

In Datensätzen, die mit Hilfe zweier unterschiedlicher Modelle simuliert werden (Verzerrung C), zeigt sich ein ähnliches Bild wie schon bei vorherigen Szenarien. Der Clusterscorealgorithmus und der DEV schneiden dabei wieder im Kriterium 2 am besten ab. Dies gilt für sämtliche Gesamtstichprobengrößen und Dimensionen. Steigt jedoch die Anzahl an Einflussgrößen, verkleinert sich der Unterschied zum Zufall jedoch etwas. Ebenfalls gute Werte des Kriteriums 2 liefert die Anpassung mit dem Maximal-Distanzverfahren. Diese Unterscheiden sich ab einer Kernmengengröße von 200 kaum von denen der anderen Verfahren.
Bei einer kleinen Auswahlgröße sind die Ergebnisse der k-Meansauswahl und der Salp-Auswahl sowie der Wahl mit dem KM++V jedoch etwas höher.
Betrachtet man nun das Kriterium 1 ergibt sich teilweise hier eine andere Rangfolge. Dabei schneidet der MDA und das DEV meist recht gut ab. Die Anpsssung mit den anderen Verfahren liefert dabei recht unterschiedliche Rangreihenfolgen und keines der

Verfahren ist unter jeder Voraussetzung am geeignetsten. Eine vollständige Aufstellung der durchschnittlichen Werte ist in Tabelle 57 bis 63 auf Seite 84 im Anhang zu betrachten. Beispielhaft zeigt die Darstellung 20 diese Werte des Kriteriums 2 für den Fall $n = 1000$ und $d = 5$.

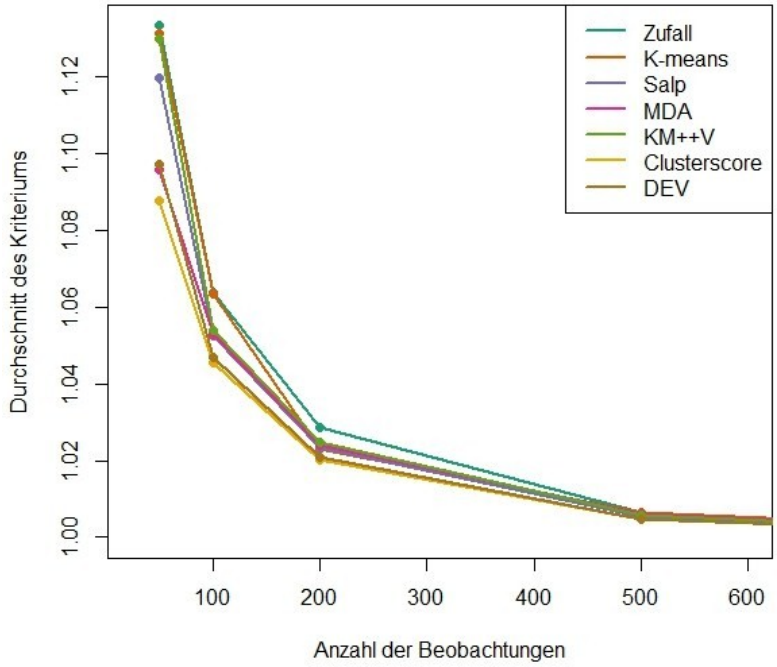

Abbildung 20: Durchschnittliche Ergebnisse des Kriteriums 2 der Auswahlverfahren bei Verzerrung C. Die Anzahl Elemente in der Kernmenge ist jeweils $r = 50$, 100, 200 und 500.

Da die beiden Modelle, denen die Daten zugrunde liegen im Verhältnis 2 zu 1 Daten simulieren, ist in jedem Teilraum ebenfalls die Wahrscheinlichkeit entsprechend vorauszusetzen. Da jeweils ein gezogener Punkt aus beiden Modellen stammen kann, ist die entsprechende Verteilung der Anzahl aus den Modellen, die auch in der Kernmenge enthalten sind, gleich. (Das heißt, dass ebenfalls in der Kernmenge das Verhältnis im Erwartungswert gleich ist.) Betrachtet man jedoch nur eine Einflussvariable, so kann es dabei, besonders bei kleiner Anzahl an gezogenen Daten, vorkommen, dass dieses Verhältnis teilweise abweichend ist. Dadurch wird dieser Einflussvariablenparameter deutlich anders geschätzt und der Fehler der Schätzung dieses Parameters ist entsprechend groß.

Dies führt zu einzelnen großen Abweichungen im Fehler des Parametervektors und somit zu unterschiedlichen Ergebnissen und Rangreihenfolgen der Kriterien.
Dieses Szenario ist dem der Heteroskedastie zudem sehr ähnlich, da die Residuen nicht mit gleicher Streuung in jedem Teilgebiet von dem Modell abweichen. Insbesondere bei Schnittpunkten der beiden Modelle ist diese gering, da sich auch die Modelle dort ähneln. Ansonsten ist die Streuung bimodal normalverteilt.
Zu beachten ist auch, dass in fast allen Fällen die Verfahren besser als eine zufällige Auswahl der Beobachtungen abschneiden.

Liegt ein nichtlinearer Zusammenhang, der mit der Sinusfunktion berechnet wird vor, so sind die y-werte der simulierten Daten mit der Verzerrung D simuliert. In diesem Szenario erbringen sämtliche Verfahren, Werte des Kriteriums 2, die sich nicht stark von denen der Zufallsauswahl unterscheiden. Im Falle des Clusterscoreverfahren und des DEV ergeben sich meist sogar deutliche schlechtere Anpassungen bezüglich des Kriteriums 2. Das Kriterium 1 ergibt jedoch einen Unterschied in der Bewertung der Auswahlen. Dabei schneiden die oben genannten beiden Verfahren am besten, hier sogar besser als der Zufall ab. Dabei liegen die Werte sämtlicher Verfahren unter dem der Zufallsauswahl.
Dies bedeutet, dass die Schätzung zwar deutlich andere Varianzen ergeben, jedoch eine geringe Verzerrung der Schätzung einigermaßen gleichmäßig auf die Einflussvariablen verteilt sind und so das Kriterium 1 einen niedrigen Wert ergibt. Da die Sinusfunktion je nach Wertebereich unterschiedliche Residuen, bei Schätzung eines linearen Zusammenhangs, ergibt, werden Verfahren, die besonders an den Eckpunkten Beobachtungen auswählen, hier besonders Beobachtungen mit hohen bzw. niedrigen Residuen erhalten. Dies führt zu einer starken Abweichung vom optimalen Modell. Da jedoch diese Abweichungen in jeder Variable in etwa gleich groß sind und ein breiter Bereich repräsentiert wird, ergibt sich entsprechend ein kleinerer Wert im Kriterium 1. Einzelne Parameter werden bei der Zufallsauswahl hierbei mit besonders hohem Fehler angepasst. Dies ist durch den mal stärkeren mal schwächeren Sinusverlauf (je nach Wertebereich und damit ein fast linearer Verlauf mal stark konvex bzw. konkav) und die unterschiedlich

hohen Parametern zu erklären. Einzelne Parameter unterliegen dann einem entsprechend hohem Fehlerterm bei der Schätzung, der sich hingegen aber weniger stark auf die Gesamtstreuung auswirkt. Eine Auswahl, die zudem nur ein Teil der Sinusfunktion abbildet, führt ebenfalls zu einem sehr unterschiedlich geschätzten Parametervektor, ohne dabei notwendigerweise die Streuung stark zu erhöhen.
Die Grafik 21 zeigt diese Gegebenheit anhand eines konstruierten Beispiels im eindimensionalen Fall. Die gesamten ermittelten Werte der Kriterien sind im Anhang in Tabelle 64 bis 70 dargestellt.

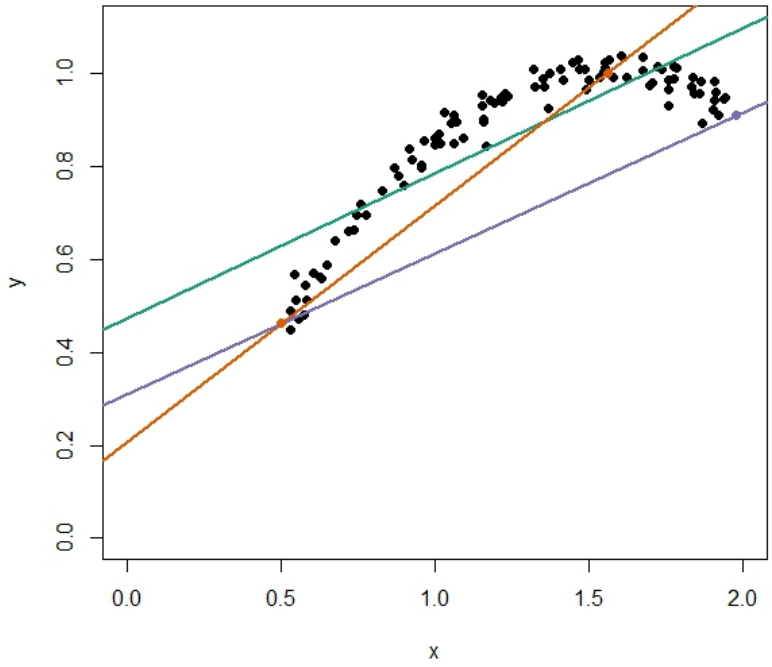

Abbildung 21: Konstruiertes, eindimensionales Beispiel für Auswahlen mit unterschiedlicher Rangreihenfolge in den Bewertungskriterien. Geschätzte Modelle mit allen Werten (grün), mit äußeren Datenpunkten (blau) und einem äußeren und einen mittleren Punkt (orange).

Zu sehen ist, dass die Schätzung des Modells, mit Hilfe der äußeren beiden Punkte (blaue Gerade) zu einer sehr hohen Varianzänderung führt, da besonders in der Mitte des Wertebereichs hohe Abweichungen und somit eine hohe Varianz entsteht. Die geschätzte

Steigung und der geschätzte y-Achsenabschnitt ist jedoch dem des Modells aller Beobachtungen (grüne Gerade) sehr ähnlich. Dies bedeutet in diesem Fall also ein hoher Wert des zweiten Kriteriums und ein nur geringer Wert des ersten Bewertungskriteriums. In dem anderen Fall (orangene Gerade) wird eine Beobachtung gewählt, die nicht am Rand liegt. Dies ergibt zwar eine geringere Varianz, da nun die Residuen bis auf einen kleinen Teil der Beobachtung am rechten Rand des Wertebereichs, sehr klein sind. Allerdings ist auch zu erkennen, dass die Steigung und der y-Achsenabschnitt deutlich mit einem anderen Wert geschätzt wird. Die Rangreihenfolge in diesen beiden Kernmengen ist also für die Bewertungskriterien umgedreht.

Die Tabelle 71 bis 77 im Anhang auf Seite 89 stellt die durchschnittlichen Kriteriumswerte dar. Zur Veranschauung werden die Werte zudem für den Fall $n = 1000$ und $d = 5$ in Abbildung 22 illustriert. Dabei handelt es sich um Kernmengen, die jeweils aus einem Datensatz entnommen werden, der mit der Verzerrung E simuliert wird.

Die Auswahlverfahren liefern dabei in einem Großteil der Fälle ein besseres Ergebnis als das Zufallsverfahren. Besonders die Kernmengen, die mit Hilfe des k-Meansverfahrens ermittelt werden, erzielen die besten Anpassungswerte für das Kriterium 2. Dies ist dadurch zu erklären, dass in diesem Szenario Wechselwirkungen vorliegen, die jedoch mit geschätzt oder berücksichtigt werden. Dies führt zu besonders hohen Verzerrungen an den Ecken des Wertebereichs, an denen hohe Ausprägungen in den x-Werten beider Variablen, die eine Wechselwirkung beinhalten, vorliegen. Dort liegen besonders hohe bzw. niedrige y-Werte (je nachdem ob die Wechselwirkung mit einem positiven oder negativen Parameter simuliert wird) vor. Da das k-Meansauswahlverfahren jedoch bei vielen Elementen in einem Cluster nicht die Beobachtungen am Rand, sondern nahe des Clustermittelpunktes auswählt, werden solche Extremwerte seltener berücksichtigt. Das Modell, das aus sämtlichen Beobachtungen geschätzt wird, liegt zudem zwischen den extrem hohen und niedrigen Werten in der Mitte. Die Auswahl mit Hilfe des k-Means ist hier also besonders geeignet und liefert daher gute Anpassungen.

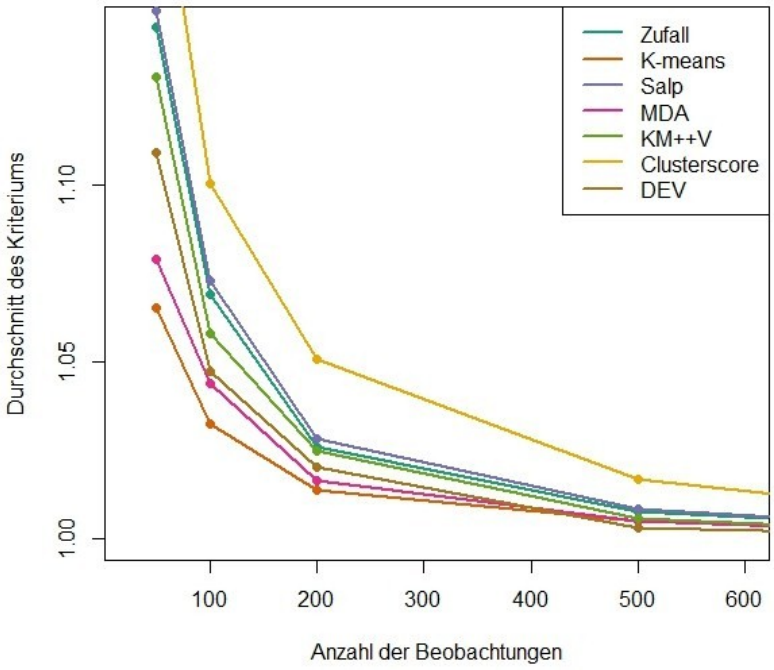

Abbildung 22: Durchschnittliche Ergebnisse des Kriteriums 2 der Auswahlverfahren bei Verzerrung E. Die Anzahl Elemente in der Kernmenge ist jeweils $r = 50$, 100, 200 und 500.

Aus demselben Grund ergeben die Kernmengen des Clusterscorealgorithmus vergleichsweise schlechtere Werte, da dieses Verfahren in einem Cluster einen Randpunkt mit höchstem Leveragescore auswählt. Auch beim Salp werden Beobachtungen mit hoher Hebelwirkung mit höherer Wahrscheinlichkeit gezogen und andere Beobachtungen im Zentrum der Wertebereiche, die zur Relativierung der Verzerrungen wichtig sind, bleiben weitgehendst unberücksichtigt. Diese Verfahren liefern deshalb meist sogar schlechtere Anpassung als eine Zufallsauswahl von Beobachtungen.

Verfahren, die Beobachtungen über den gesamten Raum auswählen, wie der MDA und das DEV, ergeben hierbei wiederum recht gute Werte.

Betrachtet man das erste Bewertungskriterium, schließen Verfahren, die besser als der Zufall bezüglich des Kriteriums 2 sind, auch in diesem Kriterium besser ab. Jedoch ändert sich durch die unterschiedliche Stärke der Wechselwirkung (je nach größer der x-Werte und des Parameters) die Schätzung einzelner Variablen teilweise und eine leicht

unterschiedliche Rangreihenfolge ist zu erkennen. In den meisten Fällen erbringt hier der deterministische Epsilonalgorithmus im Vergleich die besten Werte. Zudem ist häufig auch zu erkennen, dass die Auswahl mit dem MDA auch teilweise zu besseren Werten als die der k-Meansauswahl führen. Dies liegt daran, dass einzelne besonders starke Wechselwirkung, die an den Eckpunkten der beiden Variablen auftreten, nicht ganz durch das Verfahren geschätzt werden können (da die Eckpunkte meist nicht in den Kernmengen vorhanden sind). In diesen Fällen wird das Modell besonderes stark in dieser Variable verzerrt und diese Verzerrung wird nicht komplett abgebildet. Diese einzelnen Parameter werden daher nicht richtig geschätzt und führen zu vergleichsweise hohen Werten des Kriteriums 1. Andere weniger gute Verfahren in Bezug auf das zweite Kriterium liefern auch im ersten Kriterium höhere Werte der Abweichungen vom Gesamtmodell.

7 Zusammenfassung

Im Verlauf dieser Untersuchungen wurden eine Vielzahl an verschiedenen Szenarien und Verzerrungen betrachtet, die in Datensätzen zur Schätzung eines linearen Modells auftreten können. Dazu wurde eine umfangreiche Simulationsstudie erstellt. Die Auswahl geeigneter Kernmengen durch verschiedene Auswahlverfahren mit unterschiedlichen systematischen Ansätzen wurde daraufhin auf die Problemstellung angepasst bzw. neue Verfahren entwickelt und diese auf die jeweiligen Daten angewendet. Zur Bewertung der Resultate wurden Kriterien vorgestellt und die Ergebnisse wurden zudem auch mit einer zufälligen Auswahl verglichen.
Insgesamt schließen dabei meist der Clusterscorealgorithmus und das DEV am erfolgreichsten ab. Jedoch erbringt auch der MDA meistens Ergebnisse, die anderen Verfahren überlegen sind. In einigen wenigen in den Daten vorliegenden Verzerrungsarten liefern andere Methoden teilweise bessere Werte. Die k-Means-Auswahl erbringt etwa in dem Szenario E besonders gute Ergebnisse. In anderen Fällen jedoch meist nur geringfügig bessere Werte als die Zufallsauswahl. Eine plausible Erklärung nach genauerer Betrachtung der gezogenen Kernmengen wurde jeweils erbracht. Die Werte des Salps liegen meist in der Mitte der Werte der Verfahren.

Die Auswahlmethoden und Szenarien sowie die Ergebnisse, die in dieser Arbeit vorgestellt und betrachtet werden, können in weiterführenden Projekten untersucht werden. Verfahren, die in einigen Situationen weniger gut abschneiden, könnten dabei weiter angepasst werden, Einstellungen optimiert und weiterentwickelt werden. Zudem könnten weitere Analysen und Algorithmenschritte ergänzt werden, die frühzeitig eine drohende schlechte Anpassung verhindern. Im Clusterscoreverfahren ist dies bereits erfolgt, indem dort einzelne Elemente mit starken Abweichungen „überprüft“ werden. Dies oder auch das Ziehen mehrerer Elemente je Schritt könnte bei einigen Methoden die Robustheit der Anpassung mit der Auswahlmenge verbessern und daher ein Ansatz weiterer Untersuchungen bzw. Weiterentwicklungen sein. Ebenfalls wären Kombinationen der Ansätze und Algorithmen interessant. Bei Einflussvariablen, die wenig bzw. keinen

Einfluss auf die abhängige Variable haben, könnten die Verfahren etwa mit Methoden zur Dimensionsverkleinerung verbunden werden, um Datensätze weiter zu verringern, ohne die Güte der Anpassung stark zu verkleinern.
In fast allen Fällen wird jedoch der Zufall durch eine Auswahl mit Hilfe systematischer Ansätze deutlich übertroffen und die Alogrithmen führen weitgehendst zu erfolgreichen Anpassungen. Die besonders erfolgreichen Methoden sollten zudem auf weitere Szenarien angewendet werden, da eine komplette Untersuchung aller möglichen Situationen im Rahmen dieser Arbeit unmöglich ist.
Im Falle unterschiedlicher Verteilung der Punkte im Wertebereich wurden bei einigen Methoden Gewichtungen eingeführt. Diese verbesserten die sonst eher schlechteren Ergebnisse in entsprechenden Situationen deutlich. Weitere Simulationsstudien mit vielen unterschiedlichen Verteilungen der x-Werte in den Datensätzen wären aus diesem Grund ein sinnvoller Bestandteil weiterführender Arbeiten.
Die Bewertungskriterien liefern jedoch teilweise unterschiedliche Rangfolgen. Eine weitere Betrachtung dieser ist daher ebenfalls sinnvoll. So könnte ein Kriterium, das eine Kombination aus den hier beschriebenen ist, eingeführt werden. In manchen Bereichen ist zudem die Berechnungszeit, die Algorithmen für die Ergebnisse benötigen, entscheidend. Diese könnte daher ebenfalls untersucht werden und die Methoden beispielsweise auf besonders große Datensätze angewandt werden. Ist in einigen Bereichen die Datenübertragung beschränkt, ist die Verkleinerung des Gesamtdatensatzes von Bedeutung, selbst wenn sämtliche Werte (auch y-Werte) vorliegen. Eine Erweiterung der Algorithmen auf diesen Fall (mit vorliegenden y-Werten) wäre daher auch denkbar.
Die Kennzahl der Schwierigkeit einer guten Anpassung mit Hilfe von Kernmengen ist zudem ein erster Ansatz. Dieser Parameter könnte durch verschiedene Annahmen zu einem Regularitätskriterium weiterentwickelt werden, mit dem man so evtl. sogar Schranken für die Güte eines Verfahrens in einigen Situationen herleiten könnte. Die Streuung der Werte der Kriterien und so die Robustheit der Auswahlverfahren kann ebenfalls Bestandteil weiterer Forschung sein.

Literaturverzeichnis

Agarwal, P. K., Har-Peled, S., Varadarajan, K. R. (2005): Geometric Approximation via Coresets, *Mathematic Science Research Institute*, Vol. 52.

Arthur, D., Vassilvitskii S. (2007): k-means++: The Advantages of Careful Seeding, *Society for Industrial and Applied Mathematics*, Philadelphia, PA, USA, S. 1027-1035.

Bacher, J., Pöge, A., Wenzig, K.(2010): *Clusteranalyse: Anwendungsorientierte Einführung in Klassifikationsverfahren*, 3. Auflage, Oldenbourg, München.

Boutsidis, C., Drineas, P., Magdon-Ismail, M. (2013): Near-optimal Coresets For Least-Squares Regression, *IEEE Transactions on Information Theory59:10*, S. 6880-6892.

Chatterjee L., Hadi A. S. (1986): Influential Observations, High Leverage Points, and Outlier in Linear Regression, *Statistical Science*, Vol. 1, Nr. 3, S379-416.

Dahl, D. B. (2016). xtable: Export Tables to LaTeX or HTML. *R package version 1.8-2*, http://CRAN.R-project.org/package=xtable

Dasgupta, A., Drineas, P., Harb, B., Kumar, R., Mahoney, M. W. (2009): Sampling Algorithms and Coresets for l_p-Regression, *SIAM Journal on Computing 38:5*, S. 2060-2078.

Eckey, H.F., Kosfeld, R., Rengers, M.(2002): *Multivariate Statistik*, Auflage, Gabler, Wiesbaden.

Fahrmeir, L., Hamerle, A. und Tutz, G. (1996): *Multivariate statistische Verfahren*, 2. Auflage, de Gruyter, Berlin.

Feldman, D, Faulkner, M., Krause, A. (2011): Scalable Training of Mixture Models via Coresets, *Advances in Neural Information Processing Systems 24*, S. 2142-2150.

Hoaglin, D. C., Welsch, R. E. (1978): The Hat Matrix in Regression and ANOVA, *The American Statistician*, Vol. 32, Nr. 1, S. 17-22.

Montgomery, C., Peck, E. A. und Vining, G. G. (2006): Introduction to Linear Regression Analysis, 4. Auflage, Wiley.

Neuwirth, E. (2014). RColorBrewer: ColorBrewer Palettes. *R package version 1.1-2*, http://CRAN.R-project.org/package=RColorBrewer

Penã D., Yohai V. J. (1995): The Detection of Influential Subsets in Linear Regression by using an Influential Matrix, *Journal of Royal Statistic Society*, S. 145-156.

R Development Core Team (2015). *R 3.2.0: A language and environment for statistical computing*, *r* Foundation for Statistical Computing, Vienna, Austria, `www.R-project.org`.

Sevcikova, H., Rossini, T. (2015). rlecuyer: *r* Interface to RNG with Multiple Streams. *R package version 0.3-4*, http://CRAN.R-project.org/package=rlecuyer

A Anhang

A.1 Zusätzliche Tabellen

Tabelle 2: **Durchschnittliche Schwierigkeitsparameter** für die Verzerrung A0

	$d = 5$	$d = 10$	$d = 20$
$n = 1000$	1.53	1.49	1.47
$n = 2000$	1.54	1.49	1.47
$n = 5000$	1.53	1.48	1.46

Tabelle 3: **Durchschnittliche Schwierigkeitsparameter** für die Verzerrung A1

	$d = 5$	$d = 10$	$d = 20$
$n = 1000$	1.52	1.48	1.46
$n = 2000$	1.52	1.48	1.45
$n = 5000$	1.52	1.48	1.45

Tabelle 4: **Durchschnittliche Schwierigkeitsparameter** für die Verzerrung A2

	$d = 5$	$d = 10$	$d = 20$
$n = 1000$	1.81	1.52	1.45
$n = 2000$	1.80	1.52	1.45
$n = 5000$	1.80	1.53	1.46

Tabelle 5: **Durchschnittliche Schwierigkeitsparameter** für die Verzerrung A3

	$d = 5$	$d = 10$	$d = 20$
$n = 1000$	1.69	1.62	1.55
$n = 2000$	1.69	1.60	1.56
$n = 5000$	1.69	1.60	1.55

Tabelle 6: **Durchschnittliche Schwierigkeitsparameter** für die Verzerrung B1

	$d = 5$	$d = 10$	$d = 20$
$n = 1000$	4.18	4.07	3.97
$n = 2000$	4.19	4.09	4.01
$n = 5000$	4.20	4.10	4.03

Tabelle 7: **Durchschnittliche Schwierigkeitsparameter** für die Verzerrung B2

	$d = 5$	$d = 10$	$d = 20$
$n = 1000$	2.96	2.23	1.55
$n = 2000$	2.99	2.23	1.54
$n = 5000$	2.97	2.24	1.55

Tabelle 8: **Durchschnittliche Schwierigkeitsparameter** für die Verzerrung C

	$d = 5$	$d = 10$	$d = 20$
$n = 1000$	1.60	1.54	1.51
$n = 2000$	1.60	1.53	1.54
$n = 5000$	1.56	1.55	1.56

Tabelle 9: **Durchschnittliche Schwierigkeitsparameter** für die Verzerrung D

	$d = 5$	$d = 10$	$d = 20$
$n = 1000$	1.56	1.49	1.46
$n = 2000$	1.56	1.50	1.46
$n = 5000$	1.55	1.50	1.46

Tabelle 10: **Durchschnittliche Schwierigkeitsparameter** für die Verzerrung E

	$d = 5$	$d = 10$	$d = 20$
$n = 1000$	1.85	1.74	1.62
$n = 2000$	1.87	1.70	1.61
$n = 5000$	1.87	1.74	1.60

Verzerrung A0

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	1.151 / 1.138	0.665 / 1.052	0.488 / 1.025	0.246 / 1.006	-
$n = 1000$, $d = 10$	1.613 / 1.248	1.046 / 1.119	0.666 / 1.045	0.31 / 1.011	-
$n = 1000$, $d = 20$	2.817 / 1.755	1.647 / 1.239	1.104 / 1.092	0.487 / 1.021	-
$n = 2000$, $d = 5$	1.055 / 1.14	0.683 / 1.057	0.529 / 1.031	0.273 / 1.008	-
$n = 2000$, $d = 10$	1.645 / 1.257	1.193 / 1.116	0.69 / 1.051	0.433 / 1.016	-
$n = 2000$, $d = 20$	2.882 / 1.743	1.641 / 1.268	1.166 / 1.104	0.627 / 1.032	-
$n = 5000$, $d = 5$	1.02 / 1.133	0.725 / 1.057	0.504 / 1.03	0.299 / 1.009	0.206 / 1.005
$n = 5000$, $d = 10$	1.792 / 1.27	1.237 / 1.123	0.815 / 1.06	0.476 / 1.022	0.309 / 1.009
$n = 5000$, $d = 20$	2.819 / 1.683	1.745 / 1.264	1.151 / 1.112	0.673 / 1.039	0.478 / 1.017

Tabelle 11: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Zufall Verfahrens bei der Verzerrung A0

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	1.052 / 1.13	0.764 / 1.059	0.487 / 1.023	0.241 / 1.006	-
$n = 1000$, $d = 10$	1.9 / 1.307	1.018 / 1.112	0.656 / 1.047	0.297 / 1.01	-
$n = 1000$, $d = 20$	3.012 / 1.815	1.88 / 1.3	1.074 / 1.102	0.495 / 1.02	-
$n = 2000$, $d = 5$	1.142 / 1.135	0.742 / 1.065	0.526 / 1.03	0.288 / 1.01	-
$n = 2000$, $d = 10$	1.965 / 1.313	1.244 / 1.136	0.808 / 1.059	0.449 / 1.016	-
$n = 2000$, $d = 20$	3.376 / 1.91	1.801 / 1.295	1.301 / 1.122	0.711 / 1.035	-
$n = 5000$, $d = 5$	1.162 / 1.14	0.799 / 1.061	0.507 / 1.029	0.283 / 1.011	0.225 / 1.005
$n = 5000$, $d = 10$	1.96 / 1.37	1.316 / 1.134	0.925 / 1.07	0.51 / 1.022	0.334 / 1.01
$n = 5000$, $d = 20$	3.225 / 1.963	1.987 / 1.356	1.193 / 1.142	0.78 / 1.046	0.46 / 1.018

Tabelle 12: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem K-means Verfahrens bei der Verzerrung A0

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	0.9 / 1.115	0.696 / 1.051	0.469 / 1.023	0.222 / 1.006	-
$n = 1000$, $d = 10$	1.491 / 1.238	0.98 / 1.109	0.755 / 1.049	0.312 / 1.011	-
$n = 1000$, $d = 20$	3.029 / 1.714	1.681 / 1.235	1.03 / 1.086	0.54 / 1.022	-
$n = 2000$, $d = 5$	1.076 / 1.126	0.603 / 1.054	0.491 / 1.027	0.275 / 1.009	-
$n = 2000$, $d = 10$	1.71 / 1.262	1.042 / 1.111	0.784 / 1.051	0.401 / 1.016	-
$n = 2000$, $d = 20$	2.889 / 1.7	1.63 / 1.257	1.071 / 1.105	0.617 / 1.032	-
$n = 5000$, $d = 5$	1.047 / 1.129	0.654 / 1.055	0.469 / 1.027	0.293 / 1.009	0.193 / 1.004
$n = 5000$, $d = 10$	1.692 / 1.258	1.243 / 1.119	0.789 / 1.051	0.463 / 1.02	0.303 / 1.008
$n = 5000$, $d = 20$	2.788 / 1.722	1.661 / 1.247	1.17 / 1.111	0.657 / 1.036	0.442 / 1.017

Tabelle 13: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Salp Verfahrens bei der Verzerrung A0

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	0.844 / 1.094	0.565 / 1.041	0.394 / 1.02	0.186 / 1.005	-
$n = 1000$, $d = 10$	1.288 / 1.175	0.868 / 1.078	0.6 / 1.039	0.276 / 1.01	-
$n = 1000$, $d = 20$	2.549 / 1.455	1.542 / 1.178	0.964 / 1.071	0.434 / 1.017	-
$n = 2000$, $d = 5$	0.802 / 1.087	0.634 / 1.043	0.45 / 1.022	0.252 / 1.008	-
$n = 2000$, $d = 10$	1.54 / 1.185	0.89 / 1.071	0.706 / 1.04	0.357 / 1.013	-
$n = 2000$, $d = 20$	2.776 / 1.478	1.506 / 1.189	1.016 / 1.078	0.562 / 1.026	-
$n = 5000$, $d = 5$	0.828 / 1.083	0.577 / 1.043	0.435 / 1.023	0.3 / 1.01	0.188 / 1.004
$n = 5000$, $d = 10$	1.491 / 1.191	0.979 / 1.085	0.697 / 1.041	0.436 / 1.016	0.281 / 1.007
$n = 5000$, $d = 20$	2.706 / 1.451	1.385 / 1.182	1.011 / 1.079	0.608 / 1.031	0.4 / 1.014

Tabelle 14: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem MDA Verfahrens bei der Verzerrung A0

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	1.027 / 1.125	0.745 / 1.057	0.477 / 1.025	0.24 / 1.006	-
$n = 1000$, $d = 10$	1.635 / 1.276	0.925 / 1.104	0.689 / 1.048	0.339 / 1.011	-
$n = 1000$, $d = 20$	2.904 / 1.673	1.707 / 1.242	1.073 / 1.096	0.527 / 1.022	-
$n = 2000$, $d = 5$	0.935 / 1.114	0.783 / 1.063	0.506 / 1.029	0.272 / 1.009	-
$n = 2000$, $d = 10$	1.687 / 1.269	1.146 / 1.117	0.765 / 1.05	0.411 / 1.017	-
$n = 2000$, $d = 20$	2.85 / 1.733	1.656 / 1.247	1.066 / 1.098	0.623 / 1.032	-
$n = 5000$, $d = 5$	1.05 / 1.126	0.721 / 1.063	0.466 / 1.029	0.299 / 1.009	0.201 / 1.004
$n = 5000$, $d = 10$	1.832 / 1.262	1.202 / 1.121	0.778 / 1.055	0.466 / 1.019	0.311 / 1.009
$n = 5000$, $d = 20$	2.698 / 1.702	1.66 / 1.262	1.14 / 1.108	0.663 / 1.038	0.431 / 1.017

Tabelle 15: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem KM++V Verfahrens bei der Verzerrung A0

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	0.829 / 1.07	0.589 / 1.037	0.408 / 1.019	0.199 / 1.005	-
$n = 1000$, $d = 10$	1.251 / 1.184	0.909 / 1.08	0.538 / 1.033	0.295 / 1.009	-
$n = 1000$, $d = 20$	2.11 / 1.5	1.335 / 1.166	0.918 / 1.073	0.446 / 1.017	-
$n = 2000$, $d = 5$	0.828 / 1.081	0.634 / 1.042	0.429 / 1.02	0.266 / 1.007	-
$n = 2000$, $d = 10$	1.349 / 1.155	0.879 / 1.075	0.614 / 1.036	0.369 / 1.012	-
$n = 2000$, $d = 20$	2.352 / 1.505	1.578 / 1.182	0.986 / 1.078	0.543 / 1.024	-
$n = 5000$, $d = 5$	0.75 / 1.072	0.591 / 1.039	0.419 / 1.021	0.274 / 1.009	0.171 / 1.004
$n = 5000$, $d = 10$	1.236 / 1.145	0.957 / 1.082	0.612 / 1.037	0.436 / 1.016	0.278 / 1.007
$n = 5000$, $d = 20$	2.545 / 1.51	1.41 / 1.189	0.907 / 1.079	0.609 / 1.031	0.411 / 1.014

Tabelle 16: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Clusterscore Verfahrens bei der Verzerrung A0

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	0.794 / 1.078	0.531 / 1.038	0.371 / 1.019	0.194 / 1.005	-
$n = 1000$, $d = 10$	1.373 / 1.166	0.818 / 1.072	0.572 / 1.032	0.266 / 1.008	-
$n = 1000$, $d = 20$	2.335 / 1.458	1.426 / 1.162	0.922 / 1.064	0.458 / 1.017	-
$n = 2000$, $d = 5$	0.762 / 1.068	0.533 / 1.039	0.371 / 1.019	0.236 / 1.007	-
$n = 2000$, $d = 10$	1.236 / 1.157	0.815 / 1.073	0.613 / 1.036	0.366 / 1.013	-
$n = 2000$, $d = 20$	2.295 / 1.436	1.438 / 1.176	0.974 / 1.074	0.557 / 1.024	-
$n = 5000$, $d = 5$	0.663 / 1.073	0.529 / 1.038	0.369 / 1.021	0.244 / 1.008	0.157 / 1.004
$n = 5000$, $d = 10$	1.287 / 1.161	0.924 / 1.075	0.63 / 1.038	0.414 / 1.014	0.28 / 1.006
$n = 5000$, $d = 20$	2.343 / 1.445	1.481 / 1.18	1.018 / 1.081	0.576 / 1.03	0.377 / 1.013

Tabelle 17: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem DEV Verfahrens bei der Verzerrung A0

Verzerrung A1

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	10.719 / 1.129	6.779 / 1.054	4.768 / 1.023	2.186 / 1.006	-
$n = 1000$, $d = 10$	14.918 / 1.279	10.221 / 1.109	6.385 / 1.045	3.341 / 1.011	-
$n = 1000$, $d = 20$	28.227 / 1.738	17.923 / 1.245	10.692 / 1.098	5.169 / 1.021	-
$n = 2000$, $d = 5$	12.492 / 1.133	7.567 / 1.062	5.875 / 1.032	2.933 / 1.009	-
$n = 2000$, $d = 10$	16.373 / 1.263	11.499 / 1.128	7.047 / 1.056	4.213 / 1.017	-
$n = 2000$, $d = 20$	29.609 / 1.71	17.514 / 1.253	11.557 / 1.108	5.778 / 1.032	-
$n = 5000$, $d = 5$	11.978 / 1.142	7.617 / 1.059	5.464 / 1.029	3.244 / 1.011	1.978 / 1.005
$n = 5000$, $d = 10$	17.569 / 1.268	10.893 / 1.12	7.839 / 1.055	4.563 / 1.02	3.042 / 1.008
$n = 5000$, $d = 20$	27.883 / 1.741	17.083 / 1.259	10.925 / 1.11	6.702 / 1.041	4.459 / 1.017

Tabelle 18: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Zufall Verfahrens bei der Verzerrung A1

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	10.972 / 1.131	6.964 / 1.056	4.245 / 1.023	2.3 / 1.006	-
$n = 1000$, $d = 10$	17.574 / 1.309	10.414 / 1.118	7.029 / 1.048	3.113 / 1.011	-
$n = 1000$, $d = 20$	30.913 / 1.847	16.367 / 1.287	11.876 / 1.101	4.5 / 1.02	-
$n = 2000$, $d = 5$	11.453 / 1.129	8.007 / 1.064	5.716 / 1.029	2.953 / 1.009	-
$n = 2000$, $d = 10$	19.532 / 1.329	12.088 / 1.126	7.304 / 1.059	4.136 / 1.017	-
$n = 2000$, $d = 20$	34.078 / 1.928	19.049 / 1.31	12.692 / 1.124	6.566 / 1.036	-
$n = 5000$, $d = 5$	10.707 / 1.131	7.494 / 1.066	5.018 / 1.03	3.01 / 1.01	1.844 / 1.005
$n = 5000$, $d = 10$	19.932 / 1.365	12.515 / 1.145	8.457 / 1.066	4.746 / 1.021	3.056 / 1.009
$n = 5000$, $d = 20$	37.061 / 2.031	22.272 / 1.357	12.232 / 1.141	7.2 / 1.045	4.733 / 1.019

Tabelle 19: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem K-means Verfahrens bei der Verzerrung A1

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	11.351 / 1.118	6.698 / 1.055	4.354 / 1.022	2.198 / 1.005	-
$n = 1000$, $d = 10$	14.645 / 1.248	10.038 / 1.105	6.143 / 1.04	3.181 / 1.01	-
$n = 1000$, $d = 20$	29.144 / 1.763	16.045 / 1.224	9.328 / 1.089	4.964 / 1.021	-
$n = 2000$, $d = 5$	11.671 / 1.125	7.335 / 1.057	5.22 / 1.024	2.712 / 1.007	-
$n = 2000$, $d = 10$	16.71 / 1.272	11.283 / 1.117	7.319 / 1.048	4.076 / 1.016	-
$n = 2000$, $d = 20$	28.924 / 1.643	15.766 / 1.232	11.997 / 1.109	5.758 / 1.032	-
$n = 5000$, $d = 5$	10.671 / 1.132	7.177 / 1.06	4.951 / 1.027	3.108 / 1.01	2.064 / 1.004
$n = 5000$, $d = 10$	17.033 / 1.26	11.716 / 1.116	7.472 / 1.051	4.727 / 1.019	2.949 / 1.008
$n = 5000$, $d = 20$	29.225 / 1.681	16.742 / 1.252	12.353 / 1.113	6.625 / 1.038	4.496 / 1.016

Tabelle 20: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Salp Verfahrens bei der Verzerrung A1

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	8.403 / 1.091	6.323 / 1.048	4.206 / 1.021	1.934 / 1.006	-
$n = 1000$, $d = 10$	14.105 / 1.179	9.092 / 1.079	6.014 / 1.037	3.136 / 1.01	-
$n = 1000$, $d = 20$	26.613 / 1.478	14.769 / 1.183	9.447 / 1.075	4.89 / 1.019	-
$n = 2000$, $d = 5$	8.118 / 1.084	6.486 / 1.046	4.844 / 1.023	2.637 / 1.008	-
$n = 2000$, $d = 10$	13.003 / 1.165	9.878 / 1.087	5.687 / 1.038	3.541 / 1.013	-
$n = 2000$, $d = 20$	26.125 / 1.488	15.592 / 1.181	9.222 / 1.08	5.78 / 1.026	-
$n = 5000$, $d = 5$	8.906 / 1.083	6.495 / 1.045	4.641 / 1.023	2.553 / 1.009	1.853 / 1.004
$n = 5000$, $d = 10$	12.777 / 1.162	8.657 / 1.078	6.757 / 1.041	4.267 / 1.017	2.564 / 1.007
$n = 5000$, $d = 20$	24.879 / 1.488	15.809 / 1.189	10.294 / 1.09	6.292 / 1.031	4.367 / 1.014

Tabelle 21: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem MDA Verfahrens bei der Verzerrung A1

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	10.823 / 1.128	6.979 / 1.057	4.416 / 1.024	2.369 / 1.007	-
$n = 1000$, $d = 10$	17.288 / 1.268	11.537 / 1.115	6.933 / 1.046	3.359 / 1.011	-
$n = 1000$, $d = 20$	30.154 / 1.708	16.15 / 1.246	9.939 / 1.094	5.221 / 1.022	-
$n = 2000$, $d = 5$	11.517 / 1.144	6.631 / 1.056	5.259 / 1.028	2.608 / 1.009	-
$n = 2000$, $d = 10$	17.669 / 1.285	10.911 / 1.107	7.984 / 1.053	4.527 / 1.017	-
$n = 2000$, $d = 20$	27.68 / 1.686	16.646 / 1.24	11.04 / 1.109	6.006 / 1.032	-
$n = 5000$, $d = 5$	10.55 / 1.13	6.981 / 1.06	5.441 / 1.03	3.081 / 1.01	1.963 / 1.004
$n = 5000$, $d = 10$	17.565 / 1.276	11.794 / 1.127	8.04 / 1.055	4.269 / 1.019	3.207 / 1.009
$n = 5000$, $d = 20$	26.508 / 1.736	17.331 / 1.26	12.115 / 1.115	7.364 / 1.04	4.271 / 1.017

Tabelle 22: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem KM++V Verfahrens bei der Verzerrung A1

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	7.71 / 1.08	5.314 / 1.041	4.396 / 1.02	2.127 / 1.005	-
$n = 1000$, $d = 10$	13.88 / 1.175	8.318 / 1.077	5.98 / 1.033	3.031 / 1.009	-
$n = 1000$, $d = 20$	23.518 / 1.463	14.276 / 1.17	8.858 / 1.074	4.612 / 1.019	-
$n = 2000$, $d = 5$	8.483 / 1.075	5.99 / 1.037	4.404 / 1.018	2.722 / 1.007	-
$n = 2000$, $d = 10$	12.816 / 1.156	8.21 / 1.075	6.024 / 1.032	3.612 / 1.013	-
$n = 2000$, $d = 20$	21.722 / 1.452	13.732 / 1.17	8.66 / 1.076	5.374 / 1.025	-
$n = 5000$, $d = 5$	7.808 / 1.076	5.701 / 1.039	4.449 / 1.021	2.616 / 1.008	1.86 / 1.004
$n = 5000$, $d = 10$	12.611 / 1.156	8.031 / 1.068	6.314 / 1.038	4.016 / 1.015	2.972 / 1.007
$n = 5000$, $d = 20$	23.919 / 1.48	13.651 / 1.178	8.98 / 1.08	5.885 / 1.03	4.157 / 1.014

Tabelle 23: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Clusterscore Verfahrens bei der Verzerrung A1

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	8.258 / 1.075	5.404 / 1.042	3.428 / 1.019	1.891 / 1.005	-
$n = 1000$, $d = 10$	12.744 / 1.171	8.05 / 1.075	5.304 / 1.034	2.568 / 1.009	-
$n = 1000$, $d = 20$	22.187 / 1.454	12.729 / 1.162	8.195 / 1.067	4.641 / 1.018	-
$n = 2000$, $d = 5$	7.755 / 1.08	6.011 / 1.037	3.937 / 1.02	2.714 / 1.007	-
$n = 2000$, $d = 10$	12.842 / 1.169	8.304 / 1.073	5.801 / 1.035	3.269 / 1.012	-
$n = 2000$, $d = 20$	22.012 / 1.42	13.91 / 1.161	9.804 / 1.076	5.248 / 1.023	-
$n = 5000$, $d = 5$	8.387 / 1.082	5.535 / 1.04	3.929 / 1.021	2.362 / 1.008	1.892 / 1.004
$n = 5000$, $d = 10$	12.915 / 1.157	8.929 / 1.075	5.977 / 1.035	3.64 / 1.014	2.8 / 1.007
$n = 5000$, $d = 20$	22.164 / 1.44	13.566 / 1.166	9.637 / 1.077	6.053 / 1.029	4.042 / 1.014

Tabelle 24: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem DEV Verfahrens bei der Verzerrung A1

Verzerrung A2

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	8.708 / 1.287	5.437 / 1.058	3.498 / 1.023	1.76 / 1.006	-
$n = 1000$, $d = 10$	14.007 / 1.841	9.489 / 1.112	6.411 / 1.05	2.788 / 1.011	-
$n = 1000$, $d = 20$	25.907 / 2.615	14.272 / 1.239	9.676 / 1.098	4.525 / 1.022	-
$n = 2000$, $d = 5$	8.039 / 1.23	5.91 / 1.081	3.86 / 1.029	1.974 / 1.009	-
$n = 2000$, $d = 10$	14.216 / 1.873	8.828 / 1.115	6.818 / 1.055	3.302 / 1.018	-
$n = 2000$, $d = 20$	26.899 / 3.068	14.819 / 1.355	9.492 / 1.106	5.274 / 1.031	-
$n = 5000$, $d = 5$	8.699 / 1.267	5.694 / 1.089	3.8 / 1.027	2.499 / 1.012	1.511 / 1.005
$n = 5000$, $d = 10$	15.219 / 1.714	9.144 / 1.131	6.447 / 1.057	4.132 / 1.02	2.586 / 1.008
$n = 5000$, $d = 20$	22.047 / 3.507	15.674 / 1.274	9.865 / 1.111	5.633 / 1.041	3.741 / 1.018

Tabelle 25: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Zufall Verfahrens bei der Verzerrung A2

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	7.523 / 1.139	5.459 / 1.059	3.516 / 1.025	1.89 / 1.006	-
$n = 1000$, $d = 10$	17.275 / 1.32	10.603 / 1.12	5.733 / 1.048	2.985 / 1.011	-
$n = 1000$, $d = 20$	28.557 / 1.854	16.551 / 1.28	9.337 / 1.103	4.822 / 1.021	-
$n = 2000$, $d = 5$	8.627 / 1.151	6.032 / 1.065	3.842 / 1.028	2.104 / 1.01	-
$n = 2000$, $d = 10$	16.466 / 1.33	11.159 / 1.147	6.568 / 1.059	3.807 / 1.019	-
$n = 2000$, $d = 20$	29.438 / 1.961	16.62 / 1.325	9.76 / 1.113	5.249 / 1.034	-
$n = 5000$, $d = 5$	9.91 / 1.152	6.068 / 1.069	4.546 / 1.033	2.209 / 1.011	1.454 / 1.004
$n = 5000$, $d = 10$	18.691 / 1.405	12.819 / 1.178	7.706 / 1.068	3.998 / 1.023	2.739 / 1.01
$n = 5000$, $d = 20$	30.782 / 2.016	17.836 / 1.355	10.843 / 1.142	6.724 / 1.049	3.827 / 1.018

Tabelle 26: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem K-means Verfahrens bei der Verzerrung A2

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	7.031 / 1.098	4.904 / 1.046	3.374 / 1.02	1.599 / 1.005	-
$n = 1000$, $d = 10$	12.627 / 1.247	9.644 / 1.107	5.877 / 1.041	2.55 / 1.01	-
$n = 1000$, $d = 20$	26.532 / 2.678	15.482 / 1.236	9.078 / 1.092	4.13 / 1.02	-
$n = 2000$, $d = 5$	8.094 / 1.12	5.498 / 1.052	3.773 / 1.022	2.113 / 1.007	-
$n = 2000$, $d = 10$	13.596 / 1.256	9.221 / 1.111	5.753 / 1.047	3.29 / 1.015	-
$n = 2000$, $d = 20$	22.231 / 1.941	15.614 / 1.269	9.131 / 1.105	5.552 / 1.032	-
$n = 5000$, $d = 5$	8.799 / 1.11	5.539 / 1.058	4.026 / 1.024	2.305 / 1.009	1.498 / 1.004
$n = 5000$, $d = 10$	13.525 / 1.471	8.707 / 1.112	6.047 / 1.051	3.555 / 1.019	2.35 / 1.008
$n = 5000$, $d = 20$	24.474 / 4.068	15.382 / 1.265	9.509 / 1.111	5.562 / 1.04	3.661 / 1.017

Tabelle 27: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Salp Verfahrens bei der Verzerrung A2

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	6.435 / 1.122	4.545 / 1.078	3.235 / 1.046	1.576 / 1.015	-
$n = 1000$, $d = 10$	10.852 / 1.198	7.554 / 1.108	5.544 / 1.078	2.646 / 1.049	-
$n = 1000$, $d = 20$	21.607 / 1.499	12.522 / 1.22	8.462 / 1.131	4.255 / 1.078	-
$n = 2000$, $d = 5$	6.789 / 1.123	4.603 / 1.078	3.387 / 1.047	1.992 / 1.02	-
$n = 2000$, $d = 10$	10.74 / 1.197	7.853 / 1.12	5.561 / 1.083	3.405 / 1.053	-
$n = 2000$, $d = 20$	21.339 / 1.493	13.957 / 1.233	8.211 / 1.12	4.657 / 1.073	-
$n = 5000$, $d = 5$	6.039 / 1.116	4.433 / 1.078	3.999 / 1.054	2.071 / 1.02	1.494 / 1.008
$n = 5000$, $d = 10$	11.234 / 1.195	7.19 / 1.114	5.475 / 1.082	3.23 / 1.053	2.202 / 1.043
$n = 5000$, $d = 20$	22.477 / 1.504	13.289 / 1.221	8.602 / 1.135	5.314 / 1.079	3.412 / 1.06

Tabelle 28: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem MDA Verfahrens bei der Verzerrung A2

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	8.01 / 1.136	5.392 / 1.058	3.518 / 1.031	1.739 / 1.007	-
$n = 1000$, $d = 10$	15.126 / 1.279	9.593 / 1.12	6.252 / 1.054	3.049 / 1.012	-
$n = 1000$, $d = 20$	23.522 / 1.666	13.577 / 1.233	9.511 / 1.105	4.591 / 1.023	-
$n = 2000$, $d = 5$	7.703 / 1.118	5.206 / 1.066	3.862 / 1.033	2.269 / 1.011	-
$n = 2000$, $d = 10$	12.576 / 1.278	9.021 / 1.133	6.31 / 1.055	3.404 / 1.019	-
$n = 2000$, $d = 20$	25.953 / 1.765	16.307 / 1.281	9.724 / 1.116	5.437 / 1.034	-
$n = 5000$, $d = 5$	8.042 / 1.13	5.626 / 1.071	4.13 / 1.033	2.305 / 1.012	1.616 / 1.005
$n = 5000$, $d = 10$	13.968 / 1.281	8.73 / 1.133	6.409 / 1.064	3.958 / 1.023	2.498 / 1.01
$n = 5000$, $d = 20$	25.304 / 1.715	14.769 / 1.261	9.7 / 1.117	5.857 / 1.042	3.484 / 1.018

Tabelle 29: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem KM++V Verfahrens bei der Verzerrung A2

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	5.707 / 1.086	3.931 / 1.04	2.891 / 1.018	1.469 / 1.005	-
$n = 1000$, $d = 10$	10.357 / 1.171	6.876 / 1.07	4.507 / 1.033	2.417 / 1.009	-
$n = 1000$, $d = 20$	21.612 / 1.49	12.065 / 1.172	7.841 / 1.073	4.136 / 1.019	-
$n = 2000$, $d = 5$	5.598 / 1.082	4.171 / 1.042	3.162 / 1.022	1.863 / 1.008	-
$n = 2000$, $d = 10$	11.251 / 1.17	6.888 / 1.073	5.041 / 1.038	3.21 / 1.014	-
$n = 2000$, $d = 20$	21.073 / 1.511	12.62 / 1.187	7.996 / 1.077	4.521 / 1.025	-
$n = 5000$, $d = 5$	5.24 / 1.075	3.736 / 1.037	2.921 / 1.022	1.898 / 1.008	1.325 / 1.004
$n = 5000$, $d = 10$	9.777 / 1.154	6.859 / 1.072	4.803 / 1.038	2.929 / 1.014	2.072 / 1.006
$n = 5000$, $d = 20$	20.061 / 1.499	12.403 / 1.186	7.767 / 1.082	4.734 / 1.029	3.088 / 1.013

Tabelle 30: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Clusterscore Verfahrens bei der Verzerrung A2

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	5.335 / 1.074	3.703 / 1.037	2.852 / 1.019	1.463 / 1.005	-
$n = 1000$, $d = 10$	9.97 / 1.145	6.766 / 1.073	4.339 / 1.031	2.444 / 1.009	-
$n = 1000$, $d = 20$	19.938 / 1.433	11.477 / 1.165	7.739 / 1.068	3.78 / 1.018	-
$n = 2000$, $d = 5$	5.634 / 1.079	4.069 / 1.04	3.037 / 1.02	1.816 / 1.008	-
$n = 2000$, $d = 10$	9.591 / 1.153	7.053 / 1.071	4.997 / 1.036	3.119 / 1.013	-
$n = 2000$, $d = 20$	19.971 / 1.447	11.455 / 1.16	7.561 / 1.072	4.697 / 1.025	-
$n = 5000$, $d = 5$	5.745 / 1.079	4.049 / 1.044	2.997 / 1.021	1.88 / 1.009	1.427 / 1.004
$n = 5000$, $d = 10$	9.883 / 1.153	6.524 / 1.077	4.874 / 1.039	2.912 / 1.013	2.087 / 1.006
$n = 5000$, $d = 20$	20.224 / 1.458	12.029 / 1.176	8.143 / 1.08	4.771 / 1.029	3.144 / 1.013

Tabelle 31: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem DEV Verfahrens bei der Verzerrung A2

Verzerrung A3

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	7.397 / 1.128	5.42 / 1.06	3.198 / 1.026	1.812 / 1.007	-
$n = 1000$, $d = 10$	12.204 / 1.274	7.942 / 1.116	4.95 / 1.051	2.558 / 1.012	-
$n = 1000$, $d = 20$	19.206 / 1.734	12.273 / 1.247	7.496 / 1.096	3.637 / 1.024	-
$n = 2000$, $d = 5$	7.176 / 1.119	5.736 / 1.066	3.941 / 1.028	2.244 / 1.009	-
$n = 2000$, $d = 10$	12.017 / 1.272	7.264 / 1.114	4.903 / 1.051	2.91 / 1.017	-
$n = 2000$, $d = 20$	21.2 / 1.692	12.649 / 1.267	7.72 / 1.106	4.454 / 1.033	-
$n = 5000$, $d = 5$	8.204 / 1.137	5.538 / 1.056	4.21 / 1.03	2.534 / 1.012	1.551 / 1.004
$n = 5000$, $d = 10$	12.787 / 1.268	8.356 / 1.121	5.825 / 1.055	3.284 / 1.02	2.037 / 1.008
$n = 5000$, $d = 20$	18.831 / 1.704	13.048 / 1.262	8.169 / 1.111	5.053 / 1.039	3.146 / 1.017

Tabelle 32: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Zufall Verfahrens bei der Verzerrung A3

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	8.372 / 1.139	5.105 / 1.064	3.437 / 1.023	1.406 / 1.005	-
$n = 1000$, $d = 10$	13.966 / 1.312	8.53 / 1.128	5.253 / 1.045	2.355 / 1.011	-
$n = 1000$, $d = 20$	24.497 / 1.819	14.105 / 1.263	8.114 / 1.098	3.747 / 1.021	-
$n = 2000$, $d = 5$	8.477 / 1.126	5.7 / 1.058	4.014 / 1.028	1.975 / 1.008	-
$n = 2000$, $d = 10$	13.404 / 1.333	7.703 / 1.13	5.225 / 1.055	2.949 / 1.017	-
$n = 2000$, $d = 20$	22.629 / 1.875	12.827 / 1.301	8.335 / 1.12	4.479 / 1.036	-
$n = 5000$, $d = 5$	9.279 / 1.142	6.738 / 1.065	4.058 / 1.031	2.123 / 1.01	1.512 / 1.005
$n = 5000$, $d = 10$	13.78 / 1.33	9.007 / 1.143	6.099 / 1.064	3.643 / 1.02	2.222 / 1.009
$n = 5000$, $d = 20$	25.505 / 1.969	14.536 / 1.343	9.325 / 1.146	5.082 / 1.045	3.532 / 1.018

Tabelle 33: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem K-means Verfahrens bei der Verzerrung A3

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	7.376 / 1.11	4.633 / 1.052	3.51 / 1.024	1.474 / 1.005	-
$n = 1000$, $d = 10$	11.972 / 1.26	8.034 / 1.109	4.517 / 1.042	2.127 / 1.01	-
$n = 1000$, $d = 20$	19.154 / 1.671	12.32 / 1.235	6.687 / 1.089	3.585 / 1.021	-
$n = 2000$, $d = 5$	8.329 / 1.127	4.876 / 1.05	3.725 / 1.026	2.139 / 1.008	-
$n = 2000$, $d = 10$	10.945 / 1.264	7.38 / 1.11	5.544 / 1.052	2.954 / 1.017	-
$n = 2000$, $d = 20$	21.823 / 1.732	11.621 / 1.238	7.369 / 1.098	4.015 / 1.032	-
$n = 5000$, $d = 5$	7.516 / 1.117	4.456 / 1.056	3.774 / 1.026	2.155 / 1.009	1.571 / 1.004
$n = 5000$, $d = 10$	12.3 / 1.258	7.385 / 1.111	5.475 / 1.051	3.016 / 1.019	2.186 / 1.009
$n = 5000$, $d = 20$	19.806 / 1.685	12.496 / 1.243	7.474 / 1.111	4.948 / 1.039	3.048 / 1.015

Tabelle 34: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Salp Verfahrens bei der Verzerrung A3

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	5.866 / 1.078	3.91 / 1.039	3.057 / 1.022	1.477 / 1.005	-
$n = 1000$, $d = 10$	9.98 / 1.176	6.622 / 1.083	4.892 / 1.038	2.333 / 1.009	-
$n = 1000$, $d = 20$	17.619 / 1.486	10.383 / 1.174	7.159 / 1.075	3.453 / 1.019	-
$n = 2000$, $d = 5$	6.583 / 1.089	4.758 / 1.047	3.693 / 1.023	1.935 / 1.008	-
$n = 2000$, $d = 10$	10.596 / 1.18	6.22 / 1.083	4.216 / 1.04	2.485 / 1.013	-
$n = 2000$, $d = 20$	18.551 / 1.485	9.657 / 1.196	6.63 / 1.078	3.992 / 1.027	-
$n = 5000$, $d = 5$	6.077 / 1.083	4.541 / 1.047	3.506 / 1.023	1.994 / 1.01	1.332 / 1.004
$n = 5000$, $d = 10$	9.97 / 1.158	7.32 / 1.079	4.805 / 1.042	3.021 / 1.015	2.025 / 1.007
$n = 5000$, $d = 20$	17.836 / 1.478	10.715 / 1.184	8.038 / 1.087	4.313 / 1.031	2.847 / 1.014

Tabelle 35: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem MDA Verfahrens bei der Verzerrung A3

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	7.383 / 1.12	5.132 / 1.054	2.925 / 1.023	1.699 / 1.006	-
$n = 1000$, $d = 10$	12.287 / 1.248	7.586 / 1.101	5.141 / 1.043	2.535 / 1.011	-
$n = 1000$, $d = 20$	21.591 / 1.68	13.529 / 1.252	7.944 / 1.1	4.298 / 1.023	-
$n = 2000$, $d = 5$	8.157 / 1.126	5.998 / 1.066	4.164 / 1.027	2.303 / 1.009	-
$n = 2000$, $d = 10$	11.008 / 1.262	7.863 / 1.123	4.953 / 1.05	2.753 / 1.017	-
$n = 2000$, $d = 20$	20.845 / 1.71	11.519 / 1.255	8.067 / 1.105	4.244 / 1.033	-
$n = 5000$, $d = 5$	8.135 / 1.124	5.402 / 1.057	3.583 / 1.029	2.33 / 1.011	1.618 / 1.005
$n = 5000$, $d = 10$	12.602 / 1.273	8.398 / 1.113	5.412 / 1.056	3.038 / 1.019	2.395 / 1.009
$n = 5000$, $d = 20$	21.744 / 1.695	12.159 / 1.245	7.924 / 1.109	4.632 / 1.038	3.176 / 1.017

Tabelle 36: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem KM++V Verfahrens bei der Verzerrung A3

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	5.204 / 1.07	3.918 / 1.036	2.803 / 1.018	1.387 / 1.005	-
$n = 1000$, $d = 10$	9.146 / 1.158	6.051 / 1.074	4.164 / 1.032	2.316 / 1.009	-
$n = 1000$, $d = 20$	16.392 / 1.494	10.549 / 1.176	6.63 / 1.072	3.194 / 1.018	-
$n = 2000$, $d = 5$	6.014 / 1.075	4.148 / 1.038	3.124 / 1.019	1.886 / 1.007	-
$n = 2000$, $d = 10$	8.996 / 1.17	5.772 / 1.073	4.37 / 1.037	2.495 / 1.013	-
$n = 2000$, $d = 20$	14.636 / 1.432	9.641 / 1.184	6.92 / 1.074	3.672 / 1.025	-
$n = 5000$, $d = 5$	5.171 / 1.069	3.88 / 1.039	2.945 / 1.02	2.088 / 1.008	1.323 / 1.004
$n = 5000$, $d = 10$	8.852 / 1.157	6.697 / 1.071	4.67 / 1.033	2.939 / 1.014	1.849 / 1.006
$n = 5000$, $d = 20$	16.552 / 1.497	10.478 / 1.179	6.696 / 1.081	3.908 / 1.028	2.847 / 1.013

Tabelle 37: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Clusterscore Verfahrens bei der Verzerrung A3

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	5.127 / 1.073	3.724 / 1.037	2.698 / 1.019	1.467 / 1.005	-
$n = 1000$, $d = 10$	8.842 / 1.165	5.653 / 1.073	3.811 / 1.033	2.144 / 1.008	-
$n = 1000$, $d = 20$	16.14 / 1.492	10.031 / 1.18	6.313 / 1.071	3.099 / 1.017	-
$n = 2000$, $d = 5$	5.687 / 1.079	4.326 / 1.041	3.263 / 1.023	1.612 / 1.007	-
$n = 2000$, $d = 10$	8.383 / 1.158	5.897 / 1.077	4.012 / 1.035	2.307 / 1.012	-
$n = 2000$, $d = 20$	17.383 / 1.451	10.031 / 1.163	6.644 / 1.074	3.728 / 1.025	-
$n = 5000$, $d = 5$	5.321 / 1.076	4.191 / 1.042	2.808 / 1.021	2.023 / 1.009	1.346 / 1.004
$n = 5000$, $d = 10$	9.029 / 1.148	6.66 / 1.072	4.482 / 1.036	2.704 / 1.014	1.958 / 1.007
$n = 5000$, $d = 20$	15.985 / 1.456	10.107 / 1.18	6.713 / 1.08	4.304 / 1.03	2.775 / 1.013

Tabelle 38: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem DEV Verfahrens bei der Verzerrung A3

Verzerrung B1

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	4.631 / 1.136	3.462 / 1.057	1.92 / 1.024	1.026 / 1.006	-
$n = 1000$, $d = 10$	7.628 / 1.304	5.095 / 1.13	3.342 / 1.046	1.596 / 1.012	-
$n = 1000$, $d = 20$	14.214 / 1.792	7.272 / 1.228	4.696 / 1.095	2.392 / 1.023	-
$n = 2000$, $d = 5$	5.046 / 1.14	3.463 / 1.058	2.251 / 1.028	1.276 / 1.009	-
$n = 2000$, $d = 10$	7.426 / 1.267	4.981 / 1.122	3.139 / 1.051	1.999 / 1.018	-
$n = 2000$, $d = 20$	11.573 / 1.685	7.654 / 1.247	5.478 / 1.115	2.614 / 1.032	-
$n = 5000$, $d = 5$	4.865 / 1.133	3.437 / 1.062	2.336 / 1.029	1.436 / 1.012	0.993 / 1.005
$n = 5000$, $d = 10$	7.254 / 1.269	4.926 / 1.118	3.75 / 1.054	2.355 / 1.022	1.472 / 1.009
$n = 5000$, $d = 20$	13.3 / 1.742	7.557 / 1.274	5.571 / 1.114	2.807 / 1.04	2.136 / 1.017

Tabelle 39: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Zufall Verfahrens bei der Verzerrung B1

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	5 / 1.128	3.176 / 1.061	1.998 / 1.025	1.039 / 1.006	-
$n = 1000$, $d = 10$	7.066 / 1.258	4.837 / 1.117	3.104 / 1.045	1.5 / 1.011	-
$n = 1000$, $d = 20$	12.729 / 1.731	7.889 / 1.262	5.187 / 1.102	2.456 / 1.023	-
$n = 2000$, $d = 5$	4.702 / 1.123	3.402 / 1.057	2.524 / 1.031	1.333 / 1.009	-
$n = 2000$, $d = 10$	7.845 / 1.313	5.497 / 1.14	3.49 / 1.057	1.819 / 1.017	-
$n = 2000$, $d = 20$	12.822 / 1.868	8.629 / 1.305	5.858 / 1.121	2.912 / 1.036	-
$n = 5000$, $d = 5$	5.373 / 1.138	3.769 / 1.064	2.465 / 1.028	1.459 / 1.011	0.893 / 1.005
$n = 5000$, $d = 10$	8.868 / 1.32	5.428 / 1.137	3.971 / 1.066	2.293 / 1.021	1.592 / 1.01
$n = 5000$, $d = 20$	12.885 / 1.946	8.904 / 1.329	6.079 / 1.144	3.139 / 1.046	2.148 / 1.018

Tabelle 40: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem K-means Verfahrens bei der Verzerrung B1

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	4.502 / 1.134	3.129 / 1.054	2.077 / 1.022	0.955 / 1.005	-
$n = 1000$, $d = 10$	7.508 / 1.241	4.814 / 1.105	2.95 / 1.04	1.606 / 1.011	-
$n = 1000$, $d = 20$	11.774 / 1.638	7.124 / 1.244	4.275 / 1.089	2.175 / 1.021	-
$n = 2000$, $d = 5$	4.637 / 1.13	3.367 / 1.057	2.266 / 1.027	1.183 / 1.008	-
$n = 2000$, $d = 10$	7.621 / 1.316	4.642 / 1.106	3.128 / 1.046	1.779 / 1.016	-
$n = 2000$, $d = 20$	12.621 / 1.656	8.392 / 1.246	4.782 / 1.104	3.142 / 1.033	-
$n = 5000$, $d = 5$	5.108 / 1.122	3.406 / 1.059	2.293 / 1.026	1.389 / 1.01	0.911 / 1.004
$n = 5000$, $d = 10$	7.346 / 1.249	5.018 / 1.113	3.547 / 1.055	2.143 / 1.019	1.344 / 1.009
$n = 5000$, $d = 20$	14.221 / 1.709	7.898 / 1.262	4.881 / 1.113	3.109 / 1.038	1.972 / 1.017

Tabelle 41: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Salp Verfahrens bei der Verzerrung B1

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	3.628 / 1.08	2.362 / 1.04	1.752 / 1.018	0.887 / 1.005	-
$n = 1000$, $d = 10$	5.983 / 1.163	4.357 / 1.081	2.554 / 1.037	1.417 / 1.009	-
$n = 1000$, $d = 20$	10.518 / 1.45	7.244 / 1.188	4.413 / 1.075	2.171 / 1.018	-
$n = 2000$, $d = 5$	3.774 / 1.092	2.703 / 1.049	2.1 / 1.024	1.085 / 1.008	-
$n = 2000$, $d = 10$	5.841 / 1.174	4.273 / 1.087	2.898 / 1.042	1.535 / 1.014	-
$n = 2000$, $d = 20$	10.818 / 1.435	6.864 / 1.193	4.633 / 1.079	2.873 / 1.03	-
$n = 5000$, $d = 5$	3.721 / 1.086	2.745 / 1.042	2.085 / 1.023	1.325 / 1.01	0.902 / 1.004
$n = 5000$, $d = 10$	6.46 / 1.186	4.358 / 1.082	2.97 / 1.043	1.853 / 1.016	1.339 / 1.007
$n = 5000$, $d = 20$	11.267 / 1.49	6.726 / 1.19	4.516 / 1.088	2.825 / 1.031	1.85 / 1.014

Tabelle 42: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem MDA Verfahrens bei der Verzerrung B1

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	4.944 / 1.124	3.316 / 1.054	2.007 / 1.024	1.114 / 1.007	-
$n = 1000$, $d = 10$	8.463 / 1.264	4.915 / 1.112	3.427 / 1.045	1.623 / 1.011	-
$n = 1000$, $d = 20$	11.036 / 1.743	7.449 / 1.239	4.529 / 1.089	2.18 / 1.023	-
$n = 2000$, $d = 5$	4.307 / 1.108	3.15 / 1.062	2.342 / 1.028	1.402 / 1.009	-
$n = 2000$, $d = 10$	7.545 / 1.306	4.922 / 1.115	3.123 / 1.051	1.889 / 1.016	-
$n = 2000$, $d = 20$	12.291 / 1.666	7.455 / 1.261	4.854 / 1.105	2.852 / 1.032	-
$n = 5000$, $d = 5$	5.124 / 1.14	3.631 / 1.063	2.546 / 1.031	1.464 / 1.01	0.98 / 1.005
$n = 5000$, $d = 10$	7.083 / 1.275	4.943 / 1.107	3.806 / 1.054	2.202 / 1.019	1.399 / 1.008
$n = 5000$, $d = 20$	12.803 / 1.716	8.413 / 1.268	5.574 / 1.11	2.949 / 1.038	1.894 / 1.016

Tabelle 43: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem KM++V Verfahrens bei der Verzerrung B1

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	3.459 / 1.06	2.425 / 1.036	1.785 / 1.017	0.892 / 1.004	-
$n = 1000$, $d = 10$	5.35 / 1.148	3.847 / 1.07	2.694 / 1.033	1.405 / 1.009	-
$n = 1000$, $d = 20$	9.974 / 1.439	6.204 / 1.171	3.753 / 1.067	2.065 / 1.018	-
$n = 2000$, $d = 5$	3.135 / 1.067	2.401 / 1.035	2.023 / 1.019	1.069 / 1.007	-
$n = 2000$, $d = 10$	5.966 / 1.154	3.912 / 1.075	2.727 / 1.036	1.729 / 1.013	-
$n = 2000$, $d = 20$	9.619 / 1.477	6.823 / 1.17	4.183 / 1.077	2.479 / 1.026	-
$n = 5000$, $d = 5$	3.195 / 1.055	2.238 / 1.032	1.715 / 1.016	1.309 / 1.008	0.869 / 1.004
$n = 5000$, $d = 10$	5.038 / 1.126	3.608 / 1.064	2.733 / 1.035	1.728 / 1.013	1.196 / 1.007
$n = 5000$, $d = 20$	9.042 / 1.403	6.151 / 1.177	4.052 / 1.074	2.665 / 1.028	1.776 / 1.014

Tabelle 44: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Clusterscore Verfahrens bei der Verzerrung B1

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	3.492 / 1.068	2.422 / 1.036	1.7 / 1.017	0.854 / 1.005	-
$n = 1000$, $d = 10$	5.771 / 1.165	3.876 / 1.071	2.632 / 1.033	1.238 / 1.008	-
$n = 1000$, $d = 20$	8.96 / 1.408	5.898 / 1.166	4.065 / 1.068	1.981 / 1.016	-
$n = 2000$, $d = 5$	3.373 / 1.079	2.372 / 1.038	1.785 / 1.019	1.003 / 1.007	-
$n = 2000$, $d = 10$	5.448 / 1.152	3.669 / 1.073	2.637 / 1.036	1.477 / 1.012	-
$n = 2000$, $d = 20$	9.828 / 1.431	6.78 / 1.19	4.671 / 1.081	2.478 / 1.027	-
$n = 5000$, $d = 5$	3.147 / 1.061	2.39 / 1.036	1.604 / 1.019	1.129 / 1.007	0.801 / 1.004
$n = 5000$, $d = 10$	5.608 / 1.143	3.921 / 1.069	2.889 / 1.037	1.756 / 1.014	1.111 / 1.006
$n = 5000$, $d = 20$	10.048 / 1.434	6.54 / 1.178	4.545 / 1.081	2.597 / 1.028	1.753 / 1.013

Tabelle 45: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem DEV Verfahrens bei der Verzerrung B1

Verzerrung B2

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	3.366 / 1.098	2.438 / 1.052	1.667 / 1.025	0.921 / 1.007	-
$n = 1000$, $d = 10$	5.883 / 1.302	3.943 / 1.097	2.189 / 1.042	1.047 / 1.01	-
$n = 1000$, $d = 20$	9.392 / 1.843	4.902 / 1.232	3.242 / 1.088	1.361 / 1.018	-
$n = 2000$, $d = 5$	4.148 / 1.135	2.642 / 1.057	1.783 / 1.023	1.03 / 1.009	-
$n = 2000$, $d = 10$	5.56 / 1.276	3.585 / 1.107	2.604 / 1.046	1.362 / 1.016	-
$n = 2000$, $d = 20$	9.752 / 1.892	5.156 / 1.241	2.753 / 1.092	1.658 / 1.028	-
$n = 5000$, $d = 5$	3.809 / 1.107	2.761 / 1.06	1.789 / 1.028	1.096 / 1.009	0.759 / 1.004
$n = 5000$, $d = 10$	5.499 / 1.261	3.903 / 1.114	2.416 / 1.051	1.446 / 1.02	1.005 / 1.008
$n = 5000$, $d = 20$	8.626 / 1.75	4.897 / 1.226	2.974 / 1.097	1.795 / 1.035	1.194 / 1.016

Tabelle 46: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Zufall Verfahrens bei der Verzerrung B2

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	2.594 / 1.051	2.018 / 1.041	1.502 / 1.022	0.994 / 1.009	-
$n = 1000$, $d = 10$	3.865 / 1.136	3.062 / 1.083	2.421 / 1.067	1.685 / 1.029	-
$n = 1000$, $d = 20$	7.063 / 1.514	4.329 / 1.187	2.885 / 1.088	2.053 / 1.049	-
$n = 2000$, $d = 5$	2.698 / 1.047	1.957 / 1.036	1.582 / 1.023	1.159 / 1.014	-
$n = 2000$, $d = 10$	4.427 / 1.141	3.004 / 1.075	2.32 / 1.053	1.85 / 1.04	-
$n = 2000$, $d = 20$	7.374 / 1.495	4.293 / 1.178	2.876 / 1.089	1.795 / 1.044	-
$n = 5000$, $d = 5$	2.392 / 1.045	1.779 / 1.033	1.493 / 1.025	1.047 / 1.012	0.912 / 1.009
$n = 5000$, $d = 10$	4.906 / 1.166	3.075 / 1.081	2.026 / 1.046	1.483 / 1.03	1.311 / 1.021
$n = 5000$, $d = 20$	7.7 / 1.576	4.533 / 1.204	3.059 / 1.083	1.8 / 1.042	1.387 / 1.027

Tabelle 47: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem K-means Verfahrens bei der Verzerrung B2

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	3.477 / 1.114	2.519 / 1.06	1.932 / 1.034	1.119 / 1.012	-
$n = 1000$, $d = 10$	5.676 / 1.27	3.447 / 1.101	2.403 / 1.047	1.241 / 1.016	-
$n = 1000$, $d = 20$	8.975 / 1.759	5.013 / 1.262	3.448 / 1.111	1.986 / 1.045	-
$n = 2000$, $d = 5$	3.817 / 1.114	2.96 / 1.079	1.864 / 1.034	1.378 / 1.016	-
$n = 2000$, $d = 10$	5.531 / 1.233	3.307 / 1.111	2.533 / 1.063	1.477 / 1.026	-
$n = 2000$, $d = 20$	8.458 / 1.687	4.907 / 1.255	3.447 / 1.124	2.408 / 1.067	-
$n = 5000$, $d = 5$	3.848 / 1.116	2.835 / 1.072	2.069 / 1.038	1.43 / 1.019	1.215 / 1.015
$n = 5000$, $d = 10$	5.769 / 1.276	3.485 / 1.125	2.7 / 1.069	1.779 / 1.029	1.285 / 1.02
$n = 5000$, $d = 20$	7.247 / 1.72	5.015 / 1.264	3.658 / 1.139	2.368 / 1.069	1.918 / 1.051

Tabelle 48: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Salp Verfahrens bei der Verzerrung B2

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	4.964 / 1.182	5.365 / 1.219	5.573 / 1.233	5.266 / 1.208	-
$n = 1000$, $d = 10$	5.88 / 1.327	6.926 / 1.403	7.872 / 1.505	8.452 / 1.588	-
$n = 1000$, $d = 20$	7.886 / 1.665	8.101 / 1.641	9.438 / 1.847	10.682 / 2.048	-
$n = 2000$, $d = 5$	5.167 / 1.188	5.612 / 1.221	5.853 / 1.245	5.697 / 1.235	-
$n = 2000$, $d = 10$	5.863 / 1.335	6.804 / 1.418	7.842 / 1.518	8.36 / 1.606	-
$n = 2000$, $d = 20$	8.707 / 1.832	7.719 / 1.705	8.771 / 1.878	9.829 / 2.066	-
$n = 5000$, $d = 5$	5.078 / 1.19	5.82 / 1.231	6.132 / 1.255	5.868 / 1.252	5.799 / 1.243
$n = 5000$, $d = 10$	6.096 / 1.407	6.703 / 1.441	7.384 / 1.528	7.928 / 1.604	8.074 / 1.631
$n = 5000$, $d = 20$	8.227 / 1.888	7.979 / 1.782	8.322 / 1.852	9.419 / 2.033	9.749 / 2.114

Tabelle 49: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem MDA Verfahrens bei der Verzerrung B2

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	3.411 / 1.105	2.741 / 1.067	2.421 / 1.058	2.069 / 1.037	-
$n = 1000$, $d = 10$	5.089 / 1.185	3.863 / 1.139	3.843 / 1.13	3.099 / 1.093	-
$n = 1000$, $d = 20$	8.894 / 1.689	4.83 / 1.252	5.009 / 1.236	4.464 / 1.191	-
$n = 2000$, $d = 5$	3.874 / 1.112	2.873 / 1.065	2.836 / 1.065	2.164 / 1.043	-
$n = 2000$, $d = 10$	5.048 / 1.184	4.196 / 1.149	4.096 / 1.148	3.535 / 1.116	-
$n = 2000$, $d = 20$	7.291 / 1.489	4.82 / 1.267	5.054 / 1.283	4.253 / 1.231	-
$n = 5000$, $d = 5$	3.381 / 1.094	2.688 / 1.065	2.887 / 1.064	2.338 / 1.044	2.369 / 1.049
$n = 5000$, $d = 10$	5.144 / 1.222	4.002 / 1.148	3.805 / 1.142	3.654 / 1.145	3.551 / 1.128
$n = 5000$, $d = 20$	7.204 / 1.516	4.668 / 1.272	5.208 / 1.275	4.517 / 1.254	4.232 / 1.229

Tabelle 50: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem KM++V Verfahrens bei der Verzerrung B2

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	3.146 / 1.109	2.207 / 1.058	1.677 / 1.034	1.771 / 1.028	-
$n = 1000$, $d = 10$	5.276 / 1.298	3.523 / 1.139	2.248 / 1.054	1.597 / 1.024	-
$n = 1000$, $d = 20$	8.272 / 1.578	4.211 / 1.181	2.878 / 1.077	1.734 / 1.031	-
$n = 2000$, $d = 5$	3.395 / 1.12	2.845 / 1.087	2.025 / 1.043	1.573 / 1.022	-
$n = 2000$, $d = 10$	5.508 / 1.309	3.38 / 1.147	2.482 / 1.073	1.257 / 1.023	-
$n = 2000$, $d = 20$	8.819 / 1.842	4.916 / 1.289	2.966 / 1.107	1.907 / 1.042	-
$n = 5000$, $d = 5$	3.506 / 1.138	3.031 / 1.1	2.169 / 1.064	1.438 / 1.025	1.422 / 1.026
$n = 5000$, $d = 10$	5.802 / 1.408	4.471 / 1.253	3.083 / 1.122	1.616 / 1.035	1.407 / 1.023
$n = 5000$, $d = 20$	8.525 / 1.902	5.777 / 1.481	3.445 / 1.166	2.032 / 1.051	1.75 / 1.038

Tabelle 51: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Clusterscore Verfahrens bei der Verzerrung B2

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	2.971 / 1.102	2.403 / 1.054	1.778 / 1.027	0.907 / 1.005	-
$n = 1000$, $d = 10$	5.685 / 1.34	3.472 / 1.168	2.029 / 1.061	0.945 / 1.007	-
$n = 1000$, $d = 20$	10.811 / 2.104	6.208 / 1.387	2.978 / 1.103	1.165 / 1.013	-
$n = 2000$, $d = 5$	3.577 / 1.117	2.894 / 1.076	2.392 / 1.041	1.597 / 1.014	-
$n = 2000$, $d = 10$	6.707 / 1.456	4.388 / 1.209	2.591 / 1.086	1.193 / 1.018	-
$n = 2000$, $d = 20$	9.878 / 2.167	6.816 / 1.604	2.933 / 1.137	1.323 / 1.022	-
$n = 5000$, $d = 5$	3.651 / 1.141	2.81 / 1.081	2.584 / 1.051	2.002 / 1.024	1.444 / 1.011
$n = 5000$, $d = 10$	7.331 / 1.545	4.864 / 1.31	3.198 / 1.136	1.666 / 1.035	1.15 / 1.013
$n = 5000$, $d = 20$	10.298 / 2.206	7.916 / 1.752	4.137 / 1.234	1.527 / 1.037	0.961 / 1.013

Tabelle 52: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem DEV Verfahrens bei der Verzerrung B2

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	3.09 / 1.065	1.863 / 1.027	1.152 / 1.01	0.458 / 1.001	-
$n = 1000$, $d = 10$	5.545 / 1.209	3.332 / 1.065	1.63 / 1.019	0.698 / 1.003	-
$n = 1000$, $d = 20$	9.116 / 1.732	4.846 / 1.189	2.721 / 1.061	1.174 / 1.012	-
$n = 2000$, $d = 5$	4.158 / 1.1	2.287 / 1.039	1.452 / 1.014	0.72 / 1.004	-
$n = 2000$, $d = 10$	7.095 / 1.285	3.872 / 1.087	2.057 / 1.029	0.945 / 1.006	-
$n = 2000$, $d = 20$	8.883 / 1.696	5.242 / 1.223	3.18 / 1.078	1.548 / 1.02	-
$n = 5000$, $d = 5$	3.362 / 1.084	2.295 / 1.035	1.512 / 1.016	0.843 / 1.004	0.502 / 1.002
$n = 5000$, $d = 10$	7.84 / 1.352	4.409 / 1.125	2.411 / 1.041	1.286 / 1.011	0.727 / 1.004
$n = 5000$, $d = 20$	9.823 / 1.887	5.652 / 1.262	3.449 / 1.104	1.895 / 1.03	1.116 / 1.011

Tabelle 53: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem K-means Verfahrens mit Gewichtungen bei der Verzerrung B2

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	2.903 / 1.1	2.317 / 1.058	1.666 / 1.025	0.862 / 1.006	-
$n = 1000$, $d = 10$	5.887 / 1.279	3.999 / 1.142	3.012 / 1.07	2.231 / 1.03	-
$n = 1000$, $d = 20$	9.663 / 1.81	6.192 / 1.366	4.217 / 1.177	2.887 / 1.058	-
$n = 2000$, $d = 5$	3.586 / 1.124	2.888 / 1.087	1.848 / 1.044	1.35 / 1.015	-
$n = 2000$, $d = 10$	6.428 / 1.289	3.86 / 1.146	3.38 / 1.096	2.456 / 1.039	-
$n = 2000$, $d = 20$	9.995 / 1.9	7.189 / 1.449	4.673 / 1.248	3.009 / 1.087	-
$n = 5000$, $d = 5$	3.335 / 1.125	2.802 / 1.073	2.198 / 1.036	1.388 / 1.018	0.919 / 1.009
$n = 5000$, $d = 10$	6.03 / 1.355	4.237 / 1.187	3.388 / 1.118	2.51 / 1.053	2.1 / 1.024
$n = 5000$, $d = 20$	9.11 / 1.936	7.361 / 1.59	4.038 / 1.258	3.057 / 1.12	2.852 / 1.072

Tabelle 54: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem MDA Verfahrens mit Gewichtungen bei der Verzerrung B2

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	4.293 / 1.136	2.281 / 1.034	1.416 / 1.013	0.667 / 1.003	-
$n = 1000$, $d = 10$	7.051 / 1.317	3.794 / 1.104	2.532 / 1.037	1.261 / 1.009	-
$n = 1000$, $d = 20$	11.645 / 2.069	6.046 / 1.271	3.818 / 1.1	1.505 / 1.022	-
$n = 2000$, $d = 5$	4.729 / 1.134	2.661 / 1.046	1.766 / 1.02	0.997 / 1.006	-
$n = 2000$, $d = 10$	6.754 / 1.318	4.093 / 1.115	2.516 / 1.043	1.29 / 1.012	-
$n = 2000$, $d = 20$	9.726 / 1.873	5.607 / 1.279	3.525 / 1.111	1.984 / 1.032	-
$n = 5000$, $d = 5$	3.949 / 1.103	2.708 / 1.046	1.645 / 1.018	1.103 / 1.007	0.625 / 1.003
$n = 5000$, $d = 10$	6.367 / 1.371	3.901 / 1.106	2.792 / 1.045	1.604 / 1.016	0.983 / 1.007
$n = 5000$, $d = 20$	10.107 / 1.966	6.271 / 1.312	3.643 / 1.118	2.132 / 1.037	1.418 / 1.018

Tabelle 55: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem KM++V Verfahrens mit Gewichtungen bei der Verzerrung B2

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	2.891 / 1.101	2.009 / 1.042	1.428 / 1.015	0.63 / 1.003	-
$n = 1000$, $d = 10$	5.202 / 1.288	2.87 / 1.099	1.456 / 1.023	0.673 / 1.004	-
$n = 1000$, $d = 20$	8.718 / 1.627	4.026 / 1.162	2.426 / 1.058	1.137 / 1.011	-
$n = 2000$, $d = 5$	3.118 / 1.109	2.802 / 1.08	2.074 / 1.033	1.175 / 1.008	-
$n = 2000$, $d = 10$	5.421 / 1.302	3.085 / 1.122	1.97 / 1.045	0.961 / 1.007	-
$n = 2000$, $d = 20$	9.102 / 1.844	4.537 / 1.253	2.676 / 1.065	1.455 / 1.017	-
$n = 5000$, $d = 5$	3.206 / 1.124	2.971 / 1.093	2.562 / 1.06	1.6 / 1.02	1.02 / 1.006
$n = 5000$, $d = 10$	5.843 / 1.404	4.159 / 1.237	2.505 / 1.092	1.268 / 1.019	0.835 / 1.006
$n = 5000$, $d = 20$	8.523 / 1.908	5.816 / 1.479	2.993 / 1.139	1.542 / 1.025	0.946 / 1.009

Tabelle 56: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Clusterscore Verfahrens mit Gewichtungen bei der Verzerrung B2

Verzerrung C

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	4.032 / 1.133	2.717 / 1.064	1.907 / 1.029	1.036 / 1.006	-
$n = 1000$, $d = 10$	8.979 / 1.28	5.9 / 1.124	4.137 / 1.047	1.819 / 1.012	-
$n = 1000$, $d = 20$	20.061 / 1.707	12.967 / 1.245	8.31 / 1.089	3.91 / 1.023	-
$n = 2000$, $d = 5$	3.82 / 1.132	2.576 / 1.06	1.918 / 1.031	1.03 / 1.009	-
$n = 2000$, $d = 10$	9.494 / 1.264	7.09 / 1.119	4.138 / 1.054	2.375 / 1.017	-
$n = 2000$, $d = 20$	21.424 / 1.712	12.074 / 1.256	8.195 / 1.103	4.243 / 1.033	-
$n = 5000$, $d = 5$	3.985 / 1.132	3.049 / 1.069	1.804 / 1.031	1.436 / 1.011	0.844 / 1.005
$n = 5000$, $d = 10$	8.414 / 1.291	5.678 / 1.121	4.04 / 1.054	2.333 / 1.021	1.489 / 1.009
$n = 5000$, $d = 20$	20.194 / 1.744	12.155 / 1.272	7.199 / 1.118	4.933 / 1.042	3.033 / 1.017

Tabelle 57: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Zufall Verfahrens bei der Verzerrung C

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	4.285 / 1.131	2.787 / 1.063	1.614 / 1.023	0.869 / 1.006	-
$n = 1000$, $d = 10$	10.198 / 1.321	5.75 / 1.125	3.41 / 1.051	1.767 / 1.017	-
$n = 1000$, $d = 20$	21.264 / 1.768	13.209 / 1.275	8.679 / 1.101	3.818 / 1.042	-
$n = 2000$, $d = 5$	4.144 / 1.137	2.854 / 1.062	1.681 / 1.028	1.063 / 1.01	-
$n = 2000$, $d = 10$	10.573 / 1.3	6.531 / 1.134	3.945 / 1.055	2.204 / 1.018	-
$n = 2000$, $d = 20$	23.612 / 1.825	13.672 / 1.283	8.655 / 1.112	4.672 / 1.037	-
$n = 5000$, $d = 5$	4.534 / 1.14	2.922 / 1.069	2.104 / 1.031	1.176 / 1.012	0.774 / 1.005
$n = 5000$, $d = 10$	9.139 / 1.307	5.995 / 1.131	4.36 / 1.061	2.477 / 1.021	1.595 / 1.01
$n = 5000$, $d = 20$	21.158 / 1.828	12.937 / 1.31	8.204 / 1.125	4.947 / 1.045	3.426 / 1.02

Tabelle 58: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem K-means Verfahrens bei der Verzerrung C

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	4.203 / 1.12	2.331 / 1.053	1.626 / 1.023	0.836 / 1.005	-
$n = 1000$, $d = 10$	9.089 / 1.24	5.591 / 1.114	3.801 / 1.045	1.595 / 1.01	-
$n = 1000$, $d = 20$	21.326 / 1.695	13.792 / 1.235	7.348 / 1.092	3.587 / 1.022	-
$n = 2000$, $d = 5$	3.278 / 1.119	2.282 / 1.058	1.59 / 1.026	0.922 / 1.009	-
$n = 2000$, $d = 10$	9.241 / 1.269	5.864 / 1.123	4.078 / 1.054	2.112 / 1.017	-
$n = 2000$, $d = 20$	19.898 / 1.707	12.227 / 1.25	7.897 / 1.104	4.585 / 1.031	-
$n = 5000$, $d = 5$	3.707 / 1.128	2.458 / 1.054	1.69 / 1.025	1.139 / 1.01	0.675 / 1.005
$n = 5000$, $d = 10$	8.136 / 1.296	5.771 / 1.118	3.583 / 1.059	2.255 / 1.019	1.295 / 1.009
$n = 5000$, $d = 20$	19.85 / 1.684	12.474 / 1.257	8.032 / 1.112	4.506 / 1.041	3.505 / 1.018

Tabelle 59: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Salp Verfahrens bei der Verzerrung C

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	3.225 / 1.096	2.281 / 1.053	1.389 / 1.024	0.822 / 1.006	-
$n = 1000$, $d = 10$	7.305 / 1.217	4.69 / 1.089	3.325 / 1.039	1.684 / 1.01	-
$n = 1000$, $d = 20$	17.321 / 1.56	11.078 / 1.193	7.515 / 1.082	3.512 / 1.019	-
$n = 2000$, $d = 5$	3.032 / 1.109	2.325 / 1.054	1.587 / 1.027	0.869 / 1.009	-
$n = 2000$, $d = 10$	8.539 / 1.205	5.275 / 1.091	3.566 / 1.044	2.071 / 1.014	-
$n = 2000$, $d = 20$	19.154 / 1.562	11.844 / 1.204	6.873 / 1.089	4.437 / 1.029	-
$n = 5000$, $d = 5$	3.118 / 1.105	2.31 / 1.054	1.666 / 1.026	1.184 / 1.01	0.764 / 1.005
$n = 5000$, $d = 10$	7.481 / 1.213	5.158 / 1.1	3.352 / 1.046	1.775 / 1.016	1.376 / 1.008
$n = 5000$, $d = 20$	20.632 / 1.584	11.499 / 1.231	7.154 / 1.094	4.405 / 1.034	2.903 / 1.015

Tabelle 60: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem MDA Verfahrens bei der Verzerrung C

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	3.874 / 1.13	2.469 / 1.054	1.842 / 1.025	0.865 / 1.006	-
$n = 1000$, $d = 10$	8.387 / 1.253	5.371 / 1.114	4.039 / 1.051	1.775 / 1.011	-
$n = 1000$, $d = 20$	20.237 / 1.712	12.806 / 1.256	8.068 / 1.098	3.586 / 1.022	-
$n = 2000$, $d = 5$	3.866 / 1.125	2.477 / 1.056	1.789 / 1.029	1.044 / 1.01	-
$n = 2000$, $d = 10$	9.132 / 1.271	6.437 / 1.116	3.98 / 1.053	2.114 / 1.017	-
$n = 2000$, $d = 20$	20.48 / 1.72	12.215 / 1.259	8.443 / 1.113	4.463 / 1.033	-
$n = 5000$, $d = 5$	3.953 / 1.121	2.699 / 1.065	1.923 / 1.028	1.353 / 1.011	0.82 / 1.005
$n = 5000$, $d = 10$	8.345 / 1.266	6.059 / 1.124	3.858 / 1.06	2.44 / 1.021	1.571 / 1.009
$n = 5000$, $d = 20$	20.708 / 1.728	13.067 / 1.258	8.233 / 1.114	5.388 / 1.042	3.114 / 1.018

Tabelle 61: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem KM++V Verfahrens bei der Verzerrung C

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	3.372 / 1.088	2.632 / 1.045	1.64 / 1.02	0.755 / 1.005	-
$n = 1000$, $d = 10$	7.362 / 1.201	5.167 / 1.092	3.221 / 1.04	1.525 / 1.008	-
$n = 1000$, $d = 20$	18.182 / 1.645	10.94 / 1.2	6.853 / 1.078	3.103 / 1.018	-
$n = 2000$, $d = 5$	3.243 / 1.1	2.058 / 1.046	1.857 / 1.023	0.828 / 1.007	-
$n = 2000$, $d = 10$	7.978 / 1.197	5.48 / 1.093	3.525 / 1.043	2.048 / 1.013	-
$n = 2000$, $d = 20$	18.172 / 1.561	11.695 / 1.223	7.172 / 1.091	4.204 / 1.028	-
$n = 5000$, $d = 5$	3.281 / 1.093	2.332 / 1.043	1.752 / 1.024	1.018 / 1.009	0.727 / 1.005
$n = 5000$, $d = 10$	6.783 / 1.206	4.807 / 1.094	3.346 / 1.047	2.18 / 1.017	1.313 / 1.007
$n = 5000$, $d = 20$	16.789 / 1.597	10.859 / 1.219	7.402 / 1.1	4.636 / 1.037	3.14 / 1.016

Tabelle 62: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Clusterscore Verfahrens bei der Verzerrung C

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	3.289 / 1.097	2.268 / 1.047	1.537 / 1.021	0.787 / 1.005	-
$n = 1000$, $d = 10$	6.52 / 1.171	4.739 / 1.078	2.947 / 1.034	1.446 / 1.009	-
$n = 1000$, $d = 20$	17.618 / 1.55	9.693 / 1.193	6.627 / 1.076	3.393 / 1.018	-
$n = 2000$, $d = 5$	2.852 / 1.097	2.052 / 1.051	1.53 / 1.023	0.681 / 1.007	-
$n = 2000$, $d = 10$	8.487 / 1.201	4.907 / 1.086	3.485 / 1.041	1.936 / 1.014	-
$n = 2000$, $d = 20$	17.396 / 1.528	10.853 / 1.192	6.874 / 1.085	4.151 / 1.028	-
$n = 5000$, $d = 5$	3.262 / 1.088	2.351 / 1.048	1.711 / 1.024	0.921 / 1.01	0.642 / 1.004
$n = 5000$, $d = 10$	6.744 / 1.191	4.48 / 1.088	2.879 / 1.044	1.956 / 1.015	1.396 / 1.007
$n = 5000$, $d = 20$	17.093 / 1.523	10.784 / 1.205	6.674 / 1.092	4.101 / 1.033	2.611 / 1.014

Tabelle 63: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem DEV Verfahrens bei der Verzerrung C

Verzerrung D

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	1.283 / 1.127	0.84 / 1.058	0.575 / 1.024	0.314 / 1.007	-
$n = 1000$, $d = 10$	2.186 / 1.251	1.406 / 1.11	1.005 / 1.049	0.44 / 1.012	-
$n = 1000$, $d = 20$	4.417 / 1.702	2.588 / 1.249	1.581 / 1.097	0.787 / 1.023	-
$n = 2000$, $d = 5$	1.328 / 1.141	0.86 / 1.062	0.608 / 1.029	0.336 / 1.01	-
$n = 2000$, $d = 10$	2.241 / 1.275	1.437 / 1.117	0.943 / 1.052	0.545 / 1.017	-
$n = 2000$, $d = 20$	4.602 / 1.753	2.766 / 1.258	1.85 / 1.105	0.94 / 1.033	-
$n = 5000$, $d = 5$	1.376 / 1.143	0.991 / 1.069	0.645 / 1.031	0.355 / 1.01	0.236 / 1.005
$n = 5000$, $d = 10$	2.213 / 1.256	1.449 / 1.121	0.949 / 1.055	0.598 / 1.021	0.412 / 1.009
$n = 5000$, $d = 20$	4.566 / 1.712	2.436 / 1.254	1.598 / 1.105	1.041 / 1.04	0.689 / 1.017

Tabelle 64: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Zufall Verfahrens bei der Verzerrung D

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	1.339 / 1.129	0.771 / 1.055	0.514 / 1.021	0.256 / 1.006	-
$n = 1000$, $d = 10$	2.173 / 1.298	1.385 / 1.124	0.96 / 1.045	0.433 / 1.011	-
$n = 1000$, $d = 20$	4.277 / 1.759	2.633 / 1.278	1.639 / 1.1	0.759 / 1.021	-
$n = 2000$, $d = 5$	1.265 / 1.149	0.796 / 1.063	0.533 / 1.025	0.288 / 1.007	-
$n = 2000$, $d = 10$	2.137 / 1.321	1.397 / 1.145	1.012 / 1.066	0.568 / 1.016	-
$n = 2000$, $d = 20$	4.779 / 1.847	2.826 / 1.321	1.853 / 1.129	0.945 / 1.035	-
$n = 5000$, $d = 5$	1.271 / 1.139	0.815 / 1.064	0.619 / 1.028	0.352 / 1.009	0.247 / 1.004
$n = 5000$, $d = 10$	2.187 / 1.37	1.596 / 1.164	1.055 / 1.083	0.556 / 1.024	0.377 / 1.009
$n = 5000$, $d = 20$	4.46 / 1.939	2.799 / 1.372	1.946 / 1.167	1.084 / 1.058	0.774 / 1.022

Tabelle 65: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem K-means Verfahrens bei der Verzerrung D

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	1.113 / 1.13	0.835 / 1.054	0.532 / 1.026	0.245 / 1.007	-
$n = 1000$, $d = 10$	2.146 / 1.254	1.508 / 1.107	0.851 / 1.046	0.412 / 1.011	-
$n = 1000$, $d = 20$	4.515 / 1.696	2.666 / 1.245	1.625 / 1.091	0.808 / 1.023	-
$n = 2000$, $d = 5$	1.132 / 1.131	0.776 / 1.057	0.601 / 1.031	0.284 / 1.012	-
$n = 2000$, $d = 10$	2.091 / 1.259	1.444 / 1.114	0.952 / 1.052	0.536 / 1.018	-
$n = 2000$, $d = 20$	4.137 / 1.69	2.44 / 1.256	1.625 / 1.099	1.015 / 1.033	-
$n = 5000$, $d = 5$	1.219 / 1.129	0.928 / 1.056	0.595 / 1.031	0.357 / 1.014	0.242 / 1.008
$n = 5000$, $d = 10$	2.253 / 1.256	1.285 / 1.118	0.984 / 1.057	0.567 / 1.021	0.381 / 1.01
$n = 5000$, $d = 20$	4.603 / 1.711	2.658 / 1.256	1.651 / 1.112	1.008 / 1.04	0.64 / 1.017

Tabelle 66: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Salp Verfahrens bei der Verzerrung D

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	0.95 / 1.109	0.63 / 1.057	0.46 / 1.023	0.247 / 1.006	-
$n = 1000$, $d = 10$	1.871 / 1.195	1.302 / 1.098	0.812 / 1.045	0.354 / 1.011	-
$n = 1000$, $d = 20$	3.932 / 1.502	2.497 / 1.226	1.48 / 1.096	0.749 / 1.023	-
$n = 2000$, $d = 5$	0.909 / 1.111	0.623 / 1.058	0.456 / 1.03	0.266 / 1.011	-
$n = 2000$, $d = 10$	1.976 / 1.255	1.229 / 1.127	0.903 / 1.066	0.43 / 1.022	-
$n = 2000$, $d = 20$	4.22 / 1.625	2.789 / 1.267	1.514 / 1.128	0.805 / 1.046	-
$n = 5000$, $d = 5$	0.933 / 1.124	0.74 / 1.075	0.498 / 1.039	0.333 / 1.017	0.248 / 1.008
$n = 5000$, $d = 10$	1.992 / 1.25	1.307 / 1.138	0.783 / 1.075	0.479 / 1.033	0.327 / 1.017
$n = 5000$, $d = 20$	4.07 / 1.644	2.388 / 1.28	1.64 / 1.153	0.96 / 1.068	0.672 / 1.035

Tabelle 67: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem MDA Verfahrens bei der Verzerrung D

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	1.141 / 1.113	0.833 / 1.055	0.497 / 1.023	0.25 / 1.006	-
$n = 1000$, $d = 10$	2.16 / 1.262	1.517 / 1.107	0.922 / 1.042	0.442 / 1.01	-
$n = 1000$, $d = 20$	4.441 / 1.717	2.371 / 1.232	1.584 / 1.093	0.796 / 1.022	-
$n = 2000$, $d = 5$	1.202 / 1.12	0.839 / 1.053	0.6 / 1.025	0.303 / 1.009	-
$n = 2000$, $d = 10$	2.152 / 1.26	1.429 / 1.112	0.935 / 1.05	0.573 / 1.017	-
$n = 2000$, $d = 20$	4.921 / 1.678	2.589 / 1.236	1.707 / 1.107	0.876 / 1.033	-
$n = 5000$, $d = 5$	1.383 / 1.14	0.908 / 1.061	0.641 / 1.029	0.389 / 1.011	0.233 / 1.004
$n = 5000$, $d = 10$	2.441 / 1.289	1.513 / 1.119	0.992 / 1.056	0.542 / 1.02	0.405 / 1.008
$n = 5000$, $d = 20$	4.475 / 1.692	2.692 / 1.258	1.833 / 1.121	1.016 / 1.041	0.666 / 1.017

Tabelle 68: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem KM++V Verfahrens bei der Verzerrung D

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	1.149 / 1.208	0.725 / 1.1	0.45 / 1.046	0.265 / 1.009	-
$n = 1000$, $d = 10$	1.825 / 1.26	1.213 / 1.133	0.81 / 1.061	0.42 / 1.016	-
$n = 1000$, $d = 20$	3.912 / 1.6	2.354 / 1.229	1.586 / 1.098	0.637 / 1.023	-
$n = 2000$, $d = 5$	1.035 / 1.235	0.727 / 1.129	0.488 / 1.064	0.319 / 1.021	-
$n = 2000$, $d = 10$	1.84 / 1.3	1.278 / 1.182	0.854 / 1.11	0.49 / 1.04	-
$n = 2000$, $d = 20$	3.744 / 1.663	2.41 / 1.276	1.55 / 1.138	0.934 / 1.052	-
$n = 5000$, $d = 5$	1.057 / 1.261	0.772 / 1.181	0.55 / 1.099	0.344 / 1.043	0.247 / 1.018
$n = 5000$, $d = 10$	1.941 / 1.356	1.255 / 1.222	0.883 / 1.152	0.587 / 1.075	0.37 / 1.035
$n = 5000$, $d = 20$	3.523 / 1.658	2.205 / 1.31	1.481 / 1.172	0.906 / 1.082	0.679 / 1.041

Tabelle 69: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Clusterscore Verfahrens bei der Verzerrung D

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	0.852 / 1.134	0.508 / 1.068	0.398 / 1.033	0.188 / 1.006	-
$n = 1000$, $d = 10$	1.676 / 1.233	1.122 / 1.12	0.785 / 1.058	0.375 / 1.013	-
$n = 1000$, $d = 20$	3.604 / 1.536	2.257 / 1.227	1.533 / 1.1	0.692 / 1.023	-
$n = 2000$, $d = 5$	0.835 / 1.168	0.641 / 1.088	0.447 / 1.044	0.26 / 1.014	-
$n = 2000$, $d = 10$	1.631 / 1.274	1.215 / 1.155	0.768 / 1.09	0.419 / 1.032	-
$n = 2000$, $d = 20$	3.75 / 1.584	2.348 / 1.278	1.514 / 1.146	0.84 / 1.058	-
$n = 5000$, $d = 5$	0.929 / 1.204	0.646 / 1.112	0.457 / 1.066	0.321 / 1.029	0.198 / 1.012
$n = 5000$, $d = 10$	1.69 / 1.299	1.058 / 1.19	0.77 / 1.115	0.494 / 1.055	0.337 / 1.025
$n = 5000$, $d = 20$	3.391 / 1.633	2.106 / 1.324	1.488 / 1.18	0.934 / 1.088	0.566 / 1.046

Tabelle 70: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem DEV Verfahrens bei der Verzerrung D

Verzerrung E

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	2.71 / 1.144	1.776 / 1.069	0.973 / 1.026	0.578 / 1.008	-
$n = 1000$, $d = 10$	4.335 / 1.319	2.63 / 1.122	1.588 / 1.053	0.831 / 1.012	-
$n = 1000$, $d = 20$	7.328 / 1.74	4.094 / 1.27	2.607 / 1.103	1.298 / 1.025	-
$n = 2000$, $d = 5$	2.51 / 1.143	1.568 / 1.064	1.265 / 1.034	0.605 / 1.01	-
$n = 2000$, $d = 10$	3.659 / 1.281	2.536 / 1.129	1.847 / 1.057	1.178 / 1.019	-
$n = 2000$, $d = 20$	5.955 / 1.683	4.039 / 1.255	2.485 / 1.11	1.425 / 1.034	-
$n = 5000$, $d = 5$	3.145 / 1.16	1.966 / 1.073	1.191 / 1.033	0.85 / 1.012	0.548 / 1.006
$n = 5000$, $d = 10$	4.033 / 1.279	2.785 / 1.13	1.866 / 1.06	1.138 / 1.021	0.776 / 1.009
$n = 5000$, $d = 20$	7.224 / 1.77	4.281 / 1.273	2.646 / 1.119	1.525 / 1.041	1.11 / 1.019

Tabelle 71: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Zufall Verfahrens bei der Verzerrung E

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	1.65 / 1.065	1.241 / 1.033	0.711 / 1.014	0.436 / 1.005	-
$n = 1000$, $d = 10$	3.253 / 1.179	2.081 / 1.084	1.423 / 1.035	0.71 / 1.009	-
$n = 1000$, $d = 20$	6.648 / 1.537	3.77 / 1.21	2.474 / 1.089	1.172 / 1.018	-
$n = 2000$, $d = 5$	1.569 / 1.067	1.076 / 1.031	0.733 / 1.014	0.48 / 1.006	-
$n = 2000$, $d = 10$	3.397 / 1.194	2.2 / 1.09	1.501 / 1.04	0.921 / 1.012	-
$n = 2000$, $d = 20$	5.703 / 1.576	3.184 / 1.209	2.376 / 1.098	1.24 / 1.027	-
$n = 5000$, $d = 5$	1.584 / 1.057	1.085 / 1.027	0.835 / 1.015	0.471 / 1.006	0.37 / 1.003
$n = 5000$, $d = 10$	3.166 / 1.193	2.033 / 1.087	1.4 / 1.041	0.816 / 1.014	0.567 / 1.006
$n = 5000$, $d = 20$	5.82 / 1.556	4.084 / 1.214	2.555 / 1.097	1.561 / 1.035	1.097 / 1.016

Tabelle 72: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem K-means Verfahrens bei der Verzerrung E

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	2.474 / 1.149	1.777 / 1.073	0.972 / 1.028	0.571 / 1.009	-
$n = 1000$, $d = 10$	4.168 / 1.318	2.582 / 1.123	1.904 / 1.052	0.84 / 1.012	-
$n = 1000$, $d = 20$	7.382 / 1.757	3.962 / 1.252	2.666 / 1.104	1.254 / 1.024	-
$n = 2000$, $d = 5$	2.605 / 1.153	1.509 / 1.071	1.172 / 1.034	0.615 / 1.012	-
$n = 2000$, $d = 10$	4.305 / 1.313	2.782 / 1.132	1.937 / 1.058	1.062 / 1.018	-
$n = 2000$, $d = 20$	6.979 / 1.748	4.51 / 1.271	2.682 / 1.11	1.477 / 1.034	-
$n = 5000$, $d = 5$	2.781 / 1.156	2.074 / 1.072	1.469 / 1.039	0.834 / 1.015	0.532 / 1.008
$n = 5000$, $d = 10$	4.619 / 1.29	2.862 / 1.134	1.812 / 1.061	1.278 / 1.024	0.751 / 1.01
$n = 5000$, $d = 20$	7.031 / 1.752	4.144 / 1.275	2.87 / 1.117	1.649 / 1.042	1.084 / 1.019

Tabelle 73: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Salp Verfahrens bei der Verzerrung E

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	1.567 / 1.079	1.188 / 1.044	0.698 / 1.016	0.4 / 1.005	-
$n = 1000$, $d = 10$	3.75 / 1.196	2.279 / 1.075	1.384 / 1.031	0.655 / 1.008	-
$n = 1000$, $d = 20$	9.349 / 1.707	4.106 / 1.21	2.67 / 1.082	1.117 / 1.017	-
$n = 2000$, $d = 5$	1.542 / 1.074	0.996 / 1.039	0.771 / 1.022	0.451 / 1.008	-
$n = 2000$, $d = 10$	3.888 / 1.206	2.711 / 1.089	1.495 / 1.038	0.789 / 1.013	-
$n = 2000$, $d = 20$	7.731 / 1.674	4.512 / 1.253	2.702 / 1.096	1.356 / 1.028	-
$n = 5000$, $d = 5$	1.642 / 1.084	1.055 / 1.038	0.763 / 1.022	0.533 / 1.009	0.376 / 1.005
$n = 5000$, $d = 10$	4.167 / 1.23	2.236 / 1.085	1.657 / 1.041	0.94 / 1.017	0.553 / 1.008
$n = 5000$, $d = 20$	8.051 / 1.715	4.516 / 1.248	2.766 / 1.1	1.597 / 1.033	1.003 / 1.015

Tabelle 74: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem MDA Verfahrens bei der Verzerrung E

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	2.389 / 1.13	1.548 / 1.058	1.02 / 1.025	0.512 / 1.006	-
$n = 1000$, $d = 10$	3.957 / 1.273	2.469 / 1.119	1.756 / 1.05	0.808 / 1.011	-
$n = 1000$, $d = 20$	6.881 / 1.739	3.869 / 1.247	2.519 / 1.102	1.233 / 1.023	-
$n = 2000$, $d = 5$	2.406 / 1.124	1.641 / 1.061	1.059 / 1.029	0.638 / 1.01	-
$n = 2000$, $d = 10$	4.163 / 1.274	2.691 / 1.126	1.76 / 1.056	1.108 / 1.018	-
$n = 2000$, $d = 20$	7.295 / 1.751	4.162 / 1.274	2.644 / 1.112	1.595 / 1.036	-
$n = 5000$, $d = 5$	2.731 / 1.143	1.997 / 1.064	1.38 / 1.031	0.701 / 1.011	0.447 / 1.005
$n = 5000$, $d = 10$	4.321 / 1.293	2.817 / 1.129	1.816 / 1.055	1.093 / 1.021	0.766 / 1.009
$n = 5000$, $d = 20$	6.828 / 1.744	4.281 / 1.277	2.546 / 1.118	1.5 / 1.04	1.066 / 1.017

Tabelle 75: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem KM++V Verfahrens bei der Verzerrung E

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	2.138 / 1.199	1.388 / 1.1	0.988 / 1.051	0.631 / 1.017	-
$n = 1000$, $d = 10$	4.064 / 1.331	2.557 / 1.12	1.423 / 1.048	0.702 / 1.013	-
$n = 1000$, $d = 20$	7.394 / 1.797	4.237 / 1.259	2.587 / 1.095	1.151 / 1.018	-
$n = 2000$, $d = 5$	2.179 / 1.232	1.466 / 1.132	1.041 / 1.072	0.655 / 1.034	-
$n = 2000$, $d = 10$	4.298 / 1.36	2.78 / 1.163	1.758 / 1.08	1.035 / 1.029	-
$n = 2000$, $d = 20$	7.179 / 1.909	4.683 / 1.301	2.583 / 1.117	1.421 / 1.032	-
$n = 5000$, $d = 5$	2.267 / 1.266	1.842 / 1.168	1.3 / 1.105	0.886 / 1.056	0.758 / 1.033
$n = 5000$, $d = 10$	4.924 / 1.399	2.938 / 1.177	2.007 / 1.092	1.091 / 1.041	0.717 / 1.02
$n = 5000$, $d = 20$	7.331 / 1.903	4.351 / 1.32	2.95 / 1.134	1.693 / 1.049	1.045 / 1.02

Tabelle 76: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem Clusterscore Verfahrens bei der Verzerrung E

	$r = 50$	$r = 100$	$r = 200$	$r = 500$	$r = 1000$
$n = 1000$, $d = 5$	1.606 / 1.109	0.913 / 1.047	0.586 / 1.02	0.287 / 1.003	-
$n = 1000$, $d = 10$	3.731 / 1.232	2.08 / 1.085	1.265 / 1.031	0.576 / 1.007	-
$n = 1000$, $d = 20$	7.077 / 1.755	4.395 / 1.234	2.561 / 1.083	0.972 / 1.016	-
$n = 2000$, $d = 5$	1.676 / 1.114	0.983 / 1.056	0.687 / 1.028	0.339 / 1.009	-
$n = 2000$, $d = 10$	4.082 / 1.284	2.42 / 1.109	1.395 / 1.045	0.8 / 1.013	-
$n = 2000$, $d = 20$	7.375 / 1.849	4.023 / 1.276	2.493 / 1.1	1.24 / 1.027	-
$n = 5000$, $d = 5$	1.508 / 1.114	0.885 / 1.059	0.582 / 1.028	0.39 / 1.011	0.266 / 1.005
$n = 5000$, $d = 10$	4.204 / 1.296	2.454 / 1.125	1.613 / 1.058	0.813 / 1.022	0.478 / 1.009
$n = 5000$, $d = 20$	7.039 / 1.81	4.3 / 1.265	2.8 / 1.112	1.569 / 1.037	0.975 / 1.015

Tabelle 77: Durchschnittlichen Ergebnisse des Bewertungskriterium 1 und 2 der Auswahl mit dem DEV Verfahrens bei der Verzerrung E