Elisabeth KIRSCHBAUM

Entwicklung einer statistischen Methodik zur Charakterisierung von Lastkollektiven

MASTERARBEIT

zur Erlangung des akademischen Grades einer Diplom-Ingenieurin

Finanz- und Versicherungsmathematik

Technische Universität Graz

Betreuer:
Univ.-Prof. Dipl.-Ing. Dr.techn. Ernst STADLOBER

Institut für Statistik

Graz, im September 2013

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am ....
(Unterschrift)

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

....
date (signature)

Danksagung

Zuerst möchte ich Dr. Johannes Schauer danken, der es mir ermöglichte diese Masterarbeit auf diesem interessanten Gebiet der Statistik zu schreiben und Fragen bezüglich dieses Themas trotz eines engen Terminkalenders immer beantwortete.

Des Weiteren möchte ich Hr. Prof. Dr. Ernst Stadlober danken für die Betreuung der Arbeit und dabei vor allem für den fachlichen Input während dieser Zeit.

Elisabeth Kirschbaum
Graz, September 2013

Zusammenfassung

Für den Vergleich der klassischen und robusten Hauptkomponentenanalyse (PCA) und der robust and sparse PCA wurden aus realen Messzyklen von Fahrzeugen Merkmale anhand von Quantilen, Interquartilbereichen und Differenzen definiert. Aufbauend auf diesen Merkmalen wurden eine Basisauswahl und 2 Selektionen mit Einschränkungen definiert, welche mit jeder Reduktion weniger Merkmale enthielten. Von den Basisdaten wurden 10 Stichproben gezogen, woraus sich mit den Variablenauswahlen Basisauswahl und den 2 Szenarien 10×3 Datensets ergaben.

Die robuste Hauptkomponentenanalyse zeigte gegenüber der klassischen Methode die Vorteile, (i) dass sich die Variablen leichter den Hauptkomponenten zuordnen ließen, (ii) die robuste Methode nicht sensitiv auf Variablen mit vielen Ausreißern reagierte und (iii) dass die signifikantesten Variablen nicht korrelierten. Die robust and sparse PCA zeigte zwar den Vorteil, dass die Zuordnung der signifikanten Variablen zu den Hauptkomponenten eindeutig war, jedoch hatte diese eine sehr lange Berechnungszeit aufgrund des Tuningparameters, und die optische Darstellung der Hauptkomponenten anhand der Ladungsmatrix war nicht leicht zu interpretieren.

Es stellte sich heraus, dass die Skalierung der Datensets einen großen Einfluss hatte. Da sich Standardabweichung und Q_n-Schätzer bei einigen Variablen stark unterschieden, lieferten die klassische und die robuste Methode andere signifikante Variablen. Der Q_n-Schätzer zeigte sich hier als für die Skalierung besser geeignet.

Die Auswertung der Ergebnisse zeigte zudem, dass die robuste Hauptkomponentenanalyse für die Praxis plausible Ergebnisse lieferte.

Abstract

This thesis deals with the comparison of the classical, robust and robust and sparse principal component analysis (PCA). The data base for empirical evaluations consistet of road measurements from truck cycles, of which we defined statistical characteristics, quantiles, interquartile ranges and differences (for dynamical aspects). From these so defined characteristics three different groups has been selected as follows. There was a basic selection with all characteristics and two reduced sets. The first reduction contained a subset from the basic selection, and the second reduction contained a subset of the first reduction. For each reduction ten samples from the data base were taken so we had 10×3 different data sets to evaluate.

The benefits of the robust method were, (i) that variable selection from the loadings matrix was simpler, (ii) it was less sensitive to outliers in certain variables and (iii) there was no correlation between significant variables.

Although the robust and sparse method had additional benefits like a distinct assignment of variables to principal components, no sensitivity to outliers and almost no correlation between significant variables. However, there are at least two drawbacks of this method as high calculation time, because of repeated calculation of tuning parameters, and

the loadings plots are very difficult to interpret.

The scaling of the data sets had a high impact on the outcome as is highlighted. Standard deviation and the Q_n-estimator were remarkably different on certain variables. The methods delivered also different significant variables. Overall it turned out that the Q_n-estimator was more appropriate than the standard deviation.

As conclusion we mention that the robust method provided plausible values on the real data sets.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einführung

1.1 Ziele

Diese Arbeit beinhaltet zwei Ziele. Zum einen wird ein theoretischer und empirischer Vergleich der *robusten*, *robust and sparse* und *klassischen* Hauptkomponentenanalyse (PCA) in Bezug auf Interpretierbarkeit, Berechnung und Ergebnisse gezeigt. Es soll weiters ermittelt werden, ob die *robuste* bzw. die *robust and sparse* Methode Vorteile gegenüber der *klassischen* Methode liefert und welche Unterschiede bei der Berechnung auftreten. Aus diesen Ergebnissen resultiert, ob sich die *robuste* und die *robust and sparse* Methode für praktische Anwendungen eignen.

Zum anderen ist es ein Ziel dieser Arbeit, durch die oben genannten statistischen Methoden Lastkollektive zu charakterisieren. Dabei sollten einerseits Merkmale erhalten werden, die einen hohen Erklärungswert für die Daten haben, andererseits sollten die Ergebnisse nicht von Ausreißern beeinflusst sein und die statistisch signifikanten Merkmale sollten auch einen realen Zusammenhang mit dem Lastkollektiv haben. Da reale Messdaten oft sehr anfällig für Fehler sind und man trotz deren Vermeidung im Vorfeld diese in den Berechnungsdaten nicht ausschließen kann, wird der Schwerpunkt bei den Methoden auf Robustheit gelegt. Somit werden beide Ziele in dieser Arbeit parallel behandelt.

1.2 Datenbasis und Merkmalsdefinition

1.2.1 Daten

Für die Datenbasis wurden Daten von 37 Fahrzeugen über den Zeitraum von 2 Jahren herangezogen, welche sich unter anderem im Motortyp und der Nutzung unterschieden. Diese Daten wurden ausgewählt, da man hier über 2 Jahre hinweg durchwegs vollständige Messfahrten hatte und die aufgezeichneten Kanäle sich nicht nur auf Fahrzeug und Motor bezogen, sondern auch Messungen im Bereich AGN (Abgasnachbehandlung) beinhalteten.

Eine Messung umfasst die Länge einer Schicht oder einer Tagesfahrt (hier wurden aufeinander folgende Messungen zusammen gespeichert). Es wurden nur Messungen in die

Berechnung miteinbezogen, die eine Fahrzeuglaufzeit von mindestens 4 Stunden hatten, da man bei Messungen mit geringerer Laufzeit die Erfahrung gemacht hat, dass diese möglicherweise kein gewöhnliches Fahrverhalten aufweisen.

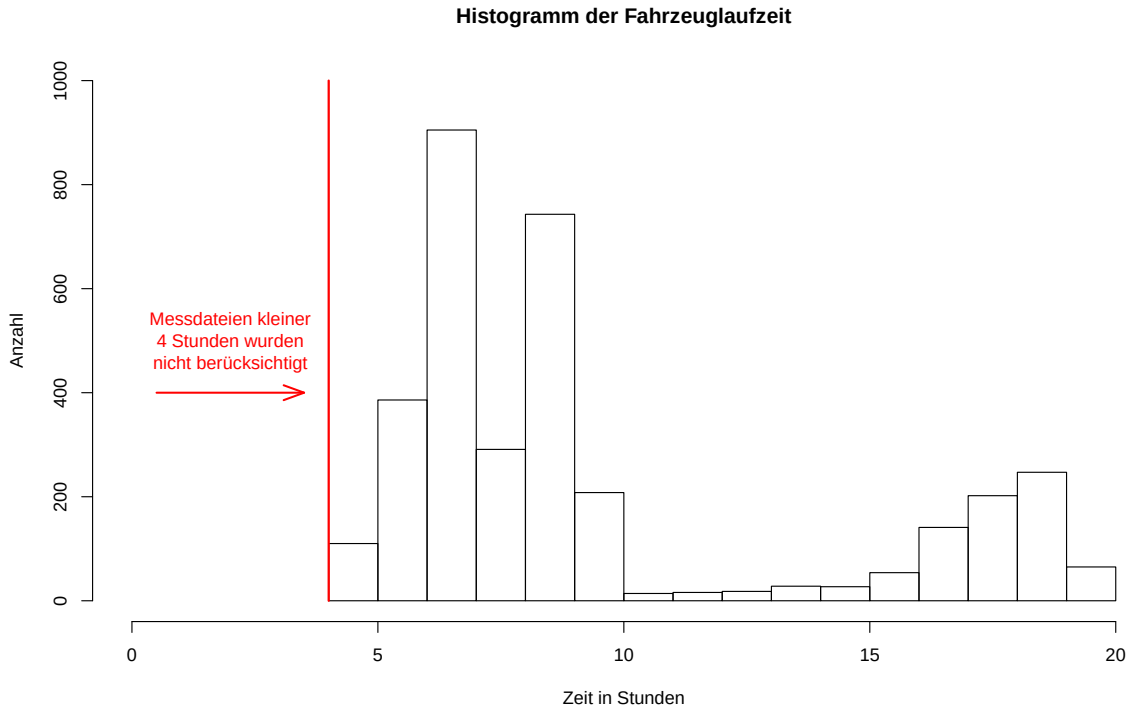


Abbildung 1.1: Histogramm der Fahrzeuglaufzeit

Die Länge der Messungen ist sehr heterogen, was in Abbildung 1.1 deutlich wird. Nicht alle Fahrzeuge sind in diesen zwei Jahren regelmäßig gefahren, jedoch erhält man mit den ausgewählten Fahrzeugen in Summe einen sehr guten Querschnitt über die Nutzung innerhalb eines Jahres. Abbildung 1.2 stellt die Summe der Fahrzyklen pro Monat über die ausgewählten zwei Jahre dar. Basis der Abbildung waren die von Messfehlern bereinigten Daten.

Ein Augenmerk sollte nicht nur auf die Variabilität in den Motorklassen gelegt werden, sondern auch bezüglich der Nutzung, da durch die unterschiedliche Nutzung auch unterschiedliche Bereiche eines Fahrzeugs beansprucht werden. Durch die zugrunde liegenden Informationen konnte unterschieden werden zwischen

- Nutzung (2 Stufen): Testbetrieb/Kundenbetrieb
- Länder
- Beladung in Tonnen (Metrisch)
- Schichtanzahl (2 Stufen): 1 Schicht /2 Schichten

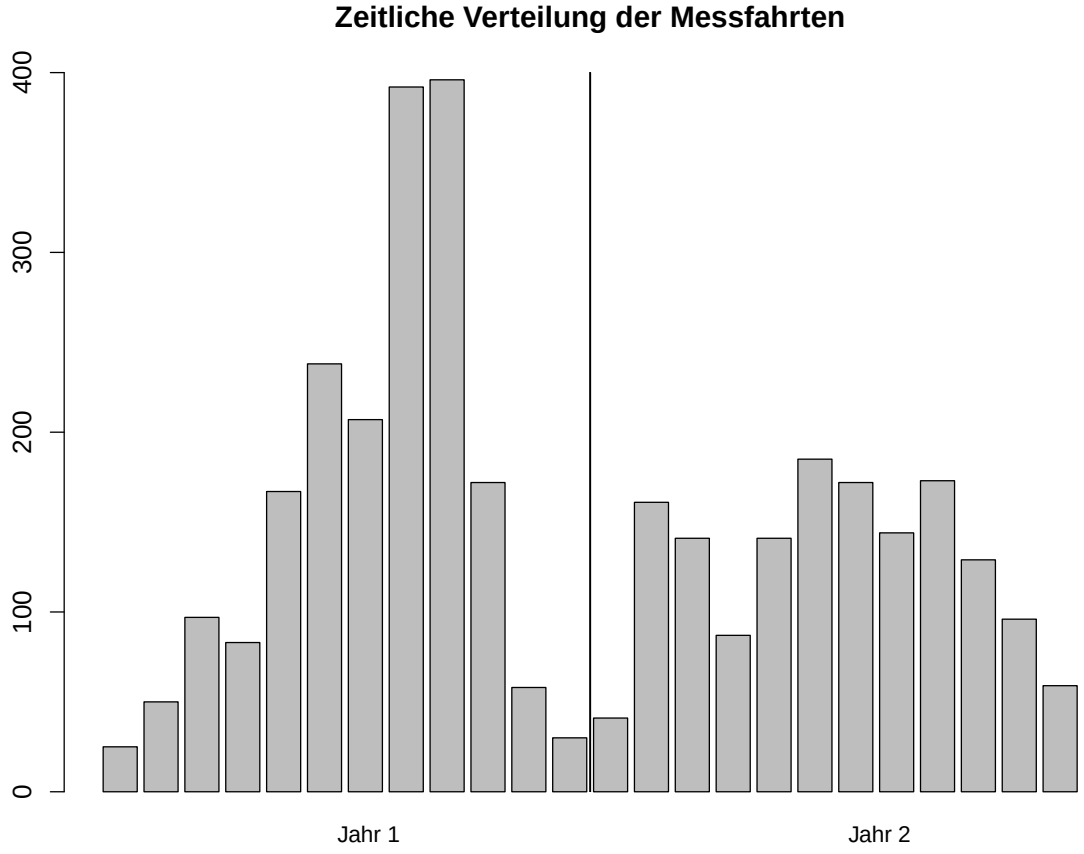


Abbildung 1.2: Verteilung der Messdaten in den Jahren 2010 und 2011

- Fahrzeugtyp
- Motortyp
- Betriebsart (7 Stufen): Fernverkehr Passstrecke, Fernverkehr, Fernverkehr hohe Beladung, Mischbetrieb, Kundenbetrieb, Stadt Verteilerbetrieb und Verteilerbetrieb

Die Betriebsart *Fernverkehr Passstrecke* beschreibt einen Hochlastbetrieb mit anspruchsvoller Topographie. In der Betriebsart *Mischbetrieb* sind *Stadt Verteilerbetrieb* und *Verteilerbetrieb* zusammengefasst. Alle weiteren Betriebsarten erklären sich durch ihre Bezeichnung.

Insgesamt wurden in die Auswertung 37 Fahrzeuge miteinbezogen, welche sich bezüglich Motorklassen wie in Tabelle 1.1 aufteilen.

	Testbetrieb	**Kundenbetrieb**	**Gesamtergebnis**
Motortyp 1	9	0	9
Motortyp 2	14	4	18
Motortyp 3	8	2	10
Gesamtergebnis	**31**	**6**	**37**

Tabelle 1.1: Übersicht der Fahrzeuge bezüglich Motortyp

1.2.2 Merkmalsdefinition

Nachdem die Fahrzeuge ausgewählt wurden, mussten die Kanäle für die Auswertung festgelegt werden. Die Kanäle lassen sich in die 4 Gruppen *Fahrzeug*, *AGN*, *Motor* und *Umgebung* einteilen. Tabelle 1.2 zeigt eine Selektion der ausgewählten Kanäle.

Fahrzeug	**Umgebung**
Fahrzeuggeschwindigkeit Leistung	Umgebungslufttemperatur C° Umgebungsdruck in bar
AGN (Abgasnachbehandlung)	**Motor**
Abgastemperatur (Eintritt / Austritt) Temperatur Harnstoff	Motordrehzahl Drehmoment Kühlwasser Motoröltemperatur

Tabelle 1.2: Beispielhafte Kanalauflistung

Die Messkanäle beinhalteten Aufzeichnungen in Form von Zeitreihen in 1Hz Auflösung, die kontinuierliche Verläufe beinhalteten, wie z.B. die Fahrzeuggeschwindigkeit oder Drehzahl, als auch Dosiermengen und Temperaturen. Ebenso gab es Kanäle mit Statusbits. Diese verschiedenen Aufzeichnungsarten mussten nun in geeignete statistische Merkmale übersetzt werden. Dazu wurden die folgenden statistischen Kenngrößen verwendet:

- Quantile berechnet von Messkanälen, Eventlängen oder Differenzen (um Dynamiken zu beschreiben).
- IQR berechnet von Messkanälen
- Anzahl und Zeitanteile von Events pro Stunde

Es wurden Quantile und der Interquartilbereich gewählt, um die Daten bereits hier zu robustifizieren. Dadurch konnten bereits im ersten Schritt die Daten von extremen Werten und Messfehlern bereinigt werden. Das Thema Messfehler wird später in Abschnitt 3.1 näher behandelt. Für die Datenbasis von $n = 888$ selektierten Messzyklen wurden aus jedem Messzyklus $p = 376$ quantitative statistische Merkmale berechnet und 25 beschreibende Faktoren festgelegt.

1.3 Festlegung der Merkmalsauswertung

Nach der Definition der Merkmale muss deren Auswertung festgelegt werden. Dabei gibt es für die AGN und Motorkenngrößen meistens die Einschränkung, dass diese nur bei laufendem Motor ausgewertet werden sollten. Daher wurden diese Größen von der Drehzahl abhängig gemacht. Ist diese größer Null, so wird angenommen, dass der Motor läuft und die AGN und Motorkenngrößen werden auch ausgewertet. Die Fahrzeuggeschwindigkeit wurde auch eingeschränkt und nur berechnet, wenn diese größer Null war. Ein weiteres Augenmerk sollte auf den Definitionsbereich der Kanäle gelegt werden. Beim Einlesen der Daten sollte bereits eine Überprüfung der Plausibilität erfolgen, damit Messfehler herausgefiltert werden können. Das Programm, in dem die Berechnung erfolgt, hängt vom Datenformat ab, in welchem die Rohdaten vorliegen. Die Outputmatrix, im Folgenden als $(n \times p)$-Datenmatrix benannt, soll in den n Spalten die statistischen Größen enthalten und in den p Zeilen die Messzyklen. Somit enthält jede Zelle eine statistische Größe für einen bestimmten Messzyklus. Dieses Format wird für die Hauptkomponentenanalyse benötigt.

1.4 Auswertung und Bewertung

Grundlage für die Hautpkomponentenanalyse ist eine $(n \times p)$-Datenmatrix, wie in Abschnitt 1.3 bereits beschrieben. Da die Daten teilweise sehr anfällig für Messfehler oder fehlende Werte sind, muss darauf ein besonderes Augenmerk gelegt werden, und zwar durch eine robuste Definition, durch eine optische Überprüfung und durch eine robuste Auswertung. Das Filtern der Messfehler wird in Unterabschnitt 3.1.1 behandelt. Da man bei der Definition der statistischen Merkmale großzügig war, wurden diese auf lineare Abhängigkeiten überprüft. Anschließend wurden für die Hauptkomponentenanalyse Repräsentanten unter den korrelierten statistischen Merkmalen gewählt.

Um die Stabilität der robusten Methode zu zeigen, wurden aus der $(n \times p)$-Datenmatrix Stichproben mit gleich großer Dimension gezogen. Um Rechenzeit zu sparen, wurden diese Stichproben parallel behandelt. Dies bringt vor allem bei der robusten Methode Vorteile, da die Rechenzeit hier um ein vielfaches höher ist als bei der klassischen Methode. Ist dann noch eine robust and sparse Lösung gewünscht, erhöht sich der Rechenaufwand um den Faktor der zu berechnenden Tuningparameter.

Die Auswahl des Tuningparameters erfolgt mit Betrachtung der erklärten Varianz, definiert in Unterabschnitt 2.2.5, welche über die Tuningparameter aufgetragen wird. Diese Abbildung wird auch *trade off curve* genannt. Der Parameter sollte so gewählt werden, dass der steile Abfall erst nach dem gewählten Parameter eintritt. Aufgrund der langen Berechnungsdauer sollte das Parametertuning iterativ erfolgen. Man beginnt mit einem relativ groben Intervall und betrachtet in weiteren Schritten interessante Bereiche näher, bis man ein zufriedenstellendes Ergebnis erhalten hat. Für die Überprüfung des gewählten Tuningparameters kann der BIC verwendet werden, welcher in Unterabschnitt 2.2.1 definiert wird.

Wenn der Tuningparameter fixiert ist, werden die einflussreichsten Variablen zu den

Hauptkomponenten mit Hilfe der Ladungsmatrix und der erklärten Varianz herausgefiltert. Anhand der Scorematrix und der gewählten Variablen lassen sich Interpretationen über das Verhalten der Daten anstellen. Dafür werden Grafiken erstellt, die die Hauptkomponenten gegeneinander auftragen. In diesen Grafiken sollten sich vorhandene Gruppen finden lassen, die die Nutzung näher beschreiben.

Kapitel 2

Theorie

2.1 Klassische Methode der Hauptkomponentenanalyse

Everit and Hothorn (2011) definieren die Hauptkomponentenanalyse (PCA: Principal Component Analysis) als ein Werkzeug, um die Dimension eines Datensets zu reduzieren und gleichzeitig so viel Variabilität der Grunddaten wie möglich zu erhalten. Dies erreicht man durch Transformation der korrelierten Variablen in ein neues Set von unkorrelierten Variablen der Hauptkomponenten. Die Hauptkomponenten sind Linearkombinationen der ursprünglichen Variablen, absteigend geordnet nach deren erklärter Variabilität.

2.1.1 Definition der Hauptkomponenten

Die Theorie zur klassischen Hauptkomponentenanalyse wurde aus dem Buch „Principal Component Analysis“ von Jollife (2002) entnommen.

Die Hauptkomponentenanalyse ist mathematisch ein Optimierungsproblem das iterativ gelöst wird. Sei nun $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ ein Vektor von Zufallsvariablen der Länge p. Eine Hauptkomponente ist definiert als lineare Funktion, deren Elemente $\boldsymbol{x}$ maximale Varianz haben, mit $\boldsymbol{\alpha_1} = (\alpha_{11}, \alpha_{12}, \cdots, \alpha_{1p})^\top$ als Vektor mit p Konstanten,

$$\boldsymbol{\alpha_1}^\top \boldsymbol{x} = \alpha_{11}x_1 + \cdots + \alpha_{1p}x_p = \sum_{j=1}^{p} \alpha_{1j}x_j. \tag{2.1}$$

Die Gleichung (2.1) stellt die erste Hauptkomponente dar. Im nächsten Schritt wird nach einer linearen Funktion $\boldsymbol{\alpha_2}^\top \boldsymbol{x}$ gesucht, die unkorreliert zu $\boldsymbol{\alpha_1}^\top \boldsymbol{x}$ ist und dabei dennoch maximale Variabilität besitzt. Dieser Schritt ist nun bis zur p-ten Hauptkomponente auszuführen, wobei die im k-ten Schritt, $1 < k \leq p$, gefundene lineare Gleichung zu allen $k-1$ zuvor gefundenen Hauptkomponenten unkorreliert sein muss.

Um dieses Problem als ein mathematisches Optimierungsproblem zu formulieren, nehmen wir an, dass die $(p \times p)$ Kovarianzmatrix des Vektors $\boldsymbol{x}$ bekannt ist und die Form

$\boldsymbol{\Sigma}$ besitzt. Dann kann man zeigen, dass die k-te Hauptkomponente definiert ist durch $z_k = \boldsymbol{\alpha_k}^\top \boldsymbol{x}$ und $\boldsymbol{\alpha_k}$ ein Eigenvektor von $\boldsymbol{\Sigma}$ mit dem k größten Eigenwert λ_k ist. Um die maximale Variation entlang $\boldsymbol{x}$ zu erzielen, betrachtet man nun das Maximierungsproblem für $\text{Var}(\boldsymbol{\alpha_1}^\top \boldsymbol{x}) = \boldsymbol{\alpha_1}^\top \Sigma \boldsymbol{\alpha_1}$,

$$\begin{aligned} &\max \boldsymbol{\alpha_1}^\top \boldsymbol{\Sigma} \boldsymbol{\alpha_1} \\ &\text{so dass } \boldsymbol{\alpha_1}^\top \boldsymbol{\alpha_1} = 1. \end{aligned}$$

Die Restriktion für $\boldsymbol{\alpha_k}$ garantiert, dass $\text{Var}(z_k) = \lambda_k$, wobei $\text{Var}(z_k)$ die Varianz von z_k bezeichnet. Um das Maximierungsproblem zu lösen, wird die Methode der Lagrange Multiplikatoren verwendet,

$$\boldsymbol{\alpha_1}^\top \boldsymbol{\Sigma} \boldsymbol{\alpha_1} - \lambda(\boldsymbol{\alpha_1}^\top \boldsymbol{\alpha_1} - 1),$$

wobei λ den Lagrange Multiplikator darstellt. Nach $\boldsymbol{\alpha_1}^\top$ abgeleitet ergibt sich,

$$(\boldsymbol{\Sigma} - \lambda \boldsymbol{I_p}) \boldsymbol{\alpha_1} = \boldsymbol{0},$$

wobei $\boldsymbol{I_p}$ die $(p \times p)$ Einheitsmatrix und $\boldsymbol{0}$ ein $p \times 1$ Vektor bestehend aus Nulleinträgen ist. Welcher der p Eigenvektoren ergibt nun für $\boldsymbol{\alpha_1}^\top \boldsymbol{x}$ die maximale Varianz? Betrachtet man die Menge, die maximiert wird,

$$\boldsymbol{\alpha_1}^\top \boldsymbol{\Sigma} \boldsymbol{\alpha_1} = \boldsymbol{\alpha_1}^\top \lambda \boldsymbol{\alpha_1} = \lambda \boldsymbol{\alpha_1}^\top \boldsymbol{\alpha_1} = \lambda,$$

erhält man, dass λ maximal sein soll. Da $\boldsymbol{\alpha_1}$ der Eigenvektor ist, der zum größten Eigenwert von $\boldsymbol{\Sigma}$ gehört, und $\text{Var}(\boldsymbol{\alpha_1}^\top \boldsymbol{x}) = \lambda_1$, ist daher $\boldsymbol{\alpha_1}$ der Eigenvektor mit der maximalen Varianz. Allgemein lässt sich schreiben, dass die k-te Hauptkomponente von $\boldsymbol{x}$ als $\boldsymbol{\alpha_k}^\top \boldsymbol{x}$ definiert ist und $\text{Var}(\boldsymbol{\alpha_k}^\top \boldsymbol{x}) = \lambda_k$, wobei hier λ_k der k-größte Eigenwert mit dem Eigenvektor $\boldsymbol{\alpha_k}$ von $\boldsymbol{\Sigma}$ ist. Geht man nun iterativ vor, so erhält man die k-te Hauptkomponente durch Maximierung des um eine Restriktion erweiterten Optimierungsproblems,

$$\begin{aligned} &\max \ \text{Var}(\boldsymbol{\alpha_k}^\top \boldsymbol{x}) = \boldsymbol{\alpha_1}^\top \boldsymbol{\Sigma} \boldsymbol{\alpha_k} \\ &\text{so dass} \quad \boldsymbol{\alpha_k}^\top \boldsymbol{\alpha_k} = 1 \\ &\qquad\quad \boldsymbol{\alpha_k}^\top \boldsymbol{\alpha_i} = 0 \quad \forall i < k. \end{aligned}$$

Durch die zweite Restriktion ist die k-te Hauptkomponente unkorreliert zu den vorhergehenden. Nun ist $\boldsymbol{\alpha_k}$ der Eigenvektor mit dem k-größten Eigenwert λ_k.

Sei $\boldsymbol{X} \in \mathbb{R}^{(n \times p)}$ die Matrix mit den Beobachtungen und $\boldsymbol{Z} \in \mathbb{R}^{(n \times p)}$ die dazugehörige Matrix mit den Hauptkomponenten, dann stehen die Matrizen durch folgende Gleichung in Beziehung:

$$\boldsymbol{Z} = \boldsymbol{X} \boldsymbol{A},$$

wobei $\boldsymbol{A} \in \mathbb{R}^{(p \times p)}$ die orthogonale Matrix mit den Vektoren $\boldsymbol{\alpha_k}$ als Spalten. Die Matrix $\boldsymbol{A}$ wird als Ladungsmatrix bezeichnet und $\boldsymbol{Z}$ als Scorematrix.

2.1.2 Hauptkomponentenanalyse mit Korrelationsmatrix

In der Praxis haben Variablen oft verschiedene Skalierungen, was die Anwendung mit einer Kovarianzmatrix nicht sehr sinnvoll erscheinen lässt. Daher sollte im Fall unterschiedlicher Skalierungen die Korrelationsmatrix verwendet werden. Man definiert daher die Hauptkomponenten durch

$$\boldsymbol{z} = \boldsymbol{A}^\top \boldsymbol{x}^*,$$

wobei $\boldsymbol{A}$ wie zuvor eine Matrix mit den Eigenvektoren in Spalten und der Vektor $\boldsymbol{x}^*$ die standardisierten Komponenten $x_j/\sqrt{\sigma_{jj}}$ von $\boldsymbol{x}$ enthält. Hauptkomponenten, die mit der Korrelationsmatrix berechnet wurden, können von $\boldsymbol{x}^*$ zu $\boldsymbol{x}$ nicht rücktransformiert werden. Diese sind nur invariant unter orthogonalen Transformationen, was die Transformation $\boldsymbol{x}^*$ zu $\boldsymbol{x}$ nicht erfüllt. Das Problem bei der Berechnung mit der Kovarianzmatrix bei unterschiedlichen Skalierungen in den Daten ist, dass die Hauptkomponenten in absteigender Reihenfolge ihrer Variabilität geordnet werden. Dadurch kann es passieren, dass Variablen mit einem großen Skalierungsintervall ohne realen inhaltlichen Zusammenhang unter die ersten Hauptkomponenten geordnet werden. Die Hauptkomponentenanalyse mit Kovarianzmatrix ist daher nur zu empfehlen, wenn alle Variablen die gleiche Skalierung besitzen.

2.1.3 Wahl der Hauptkomponenten

Wenn die Ergebnisse der PCA vorliegen, steht man vor der Frage, wie viele Hauptkomponenten gewählt werden sollen. Dafür gibt es mehrere Methoden wobei im Folgenden eine Auswahl über die Erklärte Varianz und der Scree-Graph vorgestellt wird.

Erklärte Varianz

Das meist genutzte Kriterium, um die Anzahl der Hauptkomponenten zu wählen, ist die Erklärte Varianz. Da die Hauptkomponenten absteigend nach der Größe ihrer Varianz geordnet sind, hat die k-te Hauptkomponente die Varianz λ_k. Somit definiert man die Erklärte Varianz EV_k der ersten k Hauptkomponenten durch

$$EV_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}.$$

EV^* sei nun die minimal zu erreichende Erklärende Varianz, dann ist k der kleinste Index, sodass $EV_k > EV^*$. Die Wahl der Prozentschranke hängt von der Anzahl der Beobachtungen n und der Anzahl der Hauptkomponenten p ab. In Jollife (2002) wird eine Schranke zwischen 70% und 90% empfohlen. Im Allgemeinen wird der geeignetste Wert der Erklärten Varianz kleiner, wenn n und p groß sind. Angenommen p sei groß und man wählt m Hauptkomponenten, damit man einen Wert von 70% erhält, so kann es passieren, dass m für weitere Analysen zu groß gewählt wurde. Im Falle, dass $m = 2$ bereits 90% Erklärte Varianz erreicht, kann es sinnvoll sein, weitere Hauptkomponenten zu betrachten und eine

höhere Erklärte Varianz zuzulassen, da diese interessante Informationen beinhalten könnten. In der Praxis kann es auch vorkommen, dass man durch die PCA nur eine Ordnung erhalten möchte und wählt danach für weitere Analysen die zuvor gesetzte Anzahl m an Hauptkomponenten.

Scree-Graph

Eine weitere Möglichkeit ist eine grafische Darstellung der Eigenwerte λ_j aufgetragen gegen die Nummerierung j der Hauptkomponenten. Man sucht hier grafisch nach einem Knick in der Kurve, auch Ellenbogen genannt. Die Abbildung 2.1 zeigt einen solchen Scree-Graph.

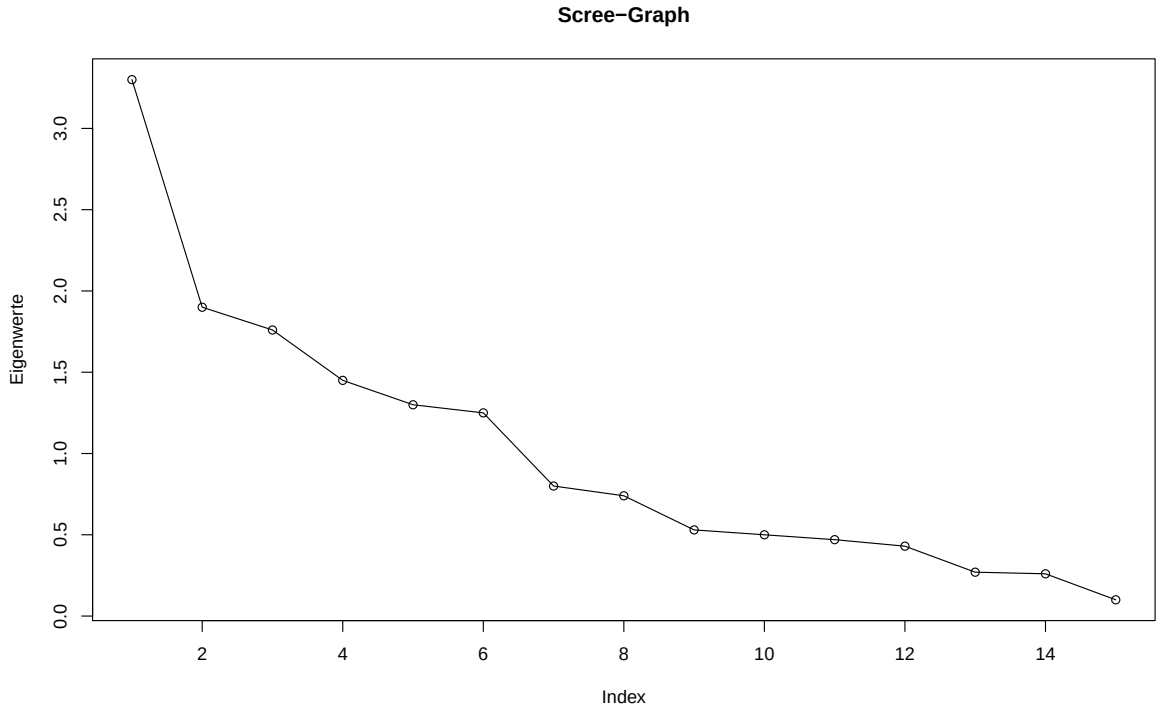


Abbildung 2.1: Scree-Graph

Der erste Punkt auf der Kurve, der einer Geraden horizontal zur x-Achse entspricht, ist die letzte Hauptkomponente, die noch ausgewählt wird. Diese optische Inspektion funktioniert nur dann gut, wenn es auch einen scharfen Knick gibt. Bleibt dieser, wie in Grafik 2.1, aus ist die Interpretation schwierig.

2.2 Robuste Methode der Hauptkomponentenanalyse

In Jollife (2002) wird darauf hingewiesen, dass Ausreißer einen hohen Einfluss auf Hauptkomponenten haben können. Um diesen Einfluss zu vermeiden, soll die Kovarianzmatrix

bzw. Korrelationsmatrix robust berechnet werden.

2.2.1 Robuste Schätzer

Im Folgenden werden robuste Schätzer definiert, die die Streuung messen, welche in der Arbeit von Rousseeuw und Croux (1993) gezeigt wurden.

MAD

Der MAD (Median absolut deviation) ist die mittlere absolute Abweichung vom Median $\tilde{x}$, der definiert ist durch

$$\text{MAD}_n = 1,4826 \text{ median}_i \left|x_i - \tilde{x}\right|, \quad i = 1, \cdots, n.$$

Der Faktor $1,4826 = 1/z_{0,75}$ mit $z_{0,75} = \Phi^{-1}(0,75)$, dem $0,75$-Quantil der $N(0,1)$-Verteilung, wird multipliziert, damit $MAD_n \approx \sigma$, dem Parameter σ der Normalverteilung entspricht. Der MAD_n hat einen Bruchpunkt von 50%, was bedeutet, dass der Schätzer erst bei Störung von mehr als 50% der Daten verzerrt wird. Die Einflussfunktion (diese gibt den maximalen Einfluss einer Beobachtung auf den Schätzer) ist beschränkt und leicht zu berechnen. Die Nachteile des MAD_n sind, dass die Einflussfunktion Unstetigkeitsstellen aufweist, die asymptotische Effizienz gegen eine Normalverteilung (im Englischen *Gaussian Efficiency*) bei nur 37% liegt und die Streuung symmetrisch ist.

$\mathbf{Q_n}$

Eine Alternative zum MAD_n ist der Q_n Schätzer, spatial Median oder geometrischer Median genannt, welcher das erste Quartil der paarweisen Distanzen der Daten angibt. Der Q_n Schätzer ist definiert durch

$$Q_n = 2,219\{|x_i - x_j| ; i < j\}_{(k)},$$

wobei $k = \binom{h}{2} \approx \binom{n}{2}/4$ und $h = [n/2] + 1$ ca. die Hälfte der Beobachtungen sind. Der Faktor ist definiert durch $2,2219 = 1/(\sqrt{2} \cdot z_{0,625})$ mit $z_{0,625} = \Phi^{-1}(0,625)$. Der Schätzer hat einen Bruchpunkt von 50% und ist somit gleich effizient wie der MAD_n. Die asymptotische Effizienz gegen eine Normalverteilung erzielt jedoch einen weit besseren Wert mit 82%. Er lässt sich leicht berechnen durch eine explizite Formel und hat eine Laufzeit von $O(n \log n)$. Der Schätzer ist auch für asymmetrische Verteilungen geeignet. Der Q_n Schätzer besitzt eine glatte Einflussfunktion.

L_1-Median

Für die Zentrierung der Daten wurde der robuste L_1-Median verwendet. Dieser multivariate Median beschreibt den Punkt, zu welchem alle Punkte des Datensets minimalen Abstand

haben. Der L_1-Median ist für ein Datenset $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})$ mit $\boldsymbol{x_i} \in \mathbb{R}^p$ definiert als Vektor $\hat{\boldsymbol{\mu}}$,

$$\hat{\boldsymbol{\mu}}(\boldsymbol{X}) = \underset{\boldsymbol{\mu}}{\operatorname{argmin}} \sum_{i=1}^{n} \|\boldsymbol{x_i} - \boldsymbol{\mu}\|,$$

wobei $\|.\|$ die Euklidsche Norm ist. Der L_1-Median ist eindeutig, wenn die Vektoren von $\boldsymbol{X}$ linear unabhängig sind. Der Bruchpunkt liegt bei 50% und der Median hat orthogonale Äquivalenz, das heißt, für einen Vektor $\boldsymbol{a} \in \mathbb{R}^p$ und eine orthogonale Matrix $\boldsymbol{Y} \in \mathbb{R}^{p \times p}$ gilt

$$\hat{\boldsymbol{\mu}}(\boldsymbol{Y}\mathbf{X} + \boldsymbol{a}) = \boldsymbol{Y}\hat{\boldsymbol{\mu}}(\boldsymbol{X}) + \boldsymbol{a}.$$

In der Hauptkomponentenanalyse wird aufgrund der orthogonalen Äquivalenz und der geringen Berechnungszeit der L_1-Median für die Zentrierung empfohlen (unter anderem von Croux und Ruiz-Gazen (2005)). Alternativ wird die Skalierung mit komponentenweisen Medianen vorgeschlagen, welche eine schlechtere Approximation für $\hat{\boldsymbol{\mu}}$ liefert. Für die PCA ist jedoch der L_1-Median zu bevorzugen.

2.2.2 Robuste PCA

Als theoretische Grundlage wurde die Arbeit von Croux, Filzmoser und Fritz (2013) verwendet. Seien n Beobachtungsvektoren $\boldsymbol{x_1}, \cdots, \boldsymbol{x_n}$ als Zeilenvektoren der Matrix $\boldsymbol{X}$ gegeben, so ist die erste Hauptkomponente definiert als

$$\boldsymbol{a}_1 = \underset{\|\boldsymbol{a}\|=1}{\operatorname{argmax}} \mathrm{V}(\boldsymbol{a}^\top \boldsymbol{x_1}, \cdots, \boldsymbol{a}^\top \boldsymbol{x_n}),$$

wobei V ein Varianzmaß darstellt. Angenommen die ersten $j-1$ Hauptkomponenten sind bereits berechnet, dann ist die j-te Hauptkomponente definiert als,

$$\boldsymbol{a_j} = \underset{\|\boldsymbol{a}\|=1, \boldsymbol{a} \perp \boldsymbol{a_1}, \cdots, \boldsymbol{a} \perp \boldsymbol{a_{j-1}}}{\operatorname{argmax}} \mathrm{V}(\boldsymbol{a}^\top \boldsymbol{x_1}, \cdots, \boldsymbol{a}^\top \boldsymbol{x_n}),$$

welche orthogonal zu allen bereits gefundenen Hauptkomponenten ist. Im Gegensatz zur klassischen PCA wird bei der robusten PCA als Streuungsmaß ein robuster Schätzer wie z.B. MAD_n oder Q_n verwendet. Aufgrund der guten Eigenschaften des Q_n, Schätzers die in Unterabschnitt 2.2.1 angegeben wurden, wurde Q_n^2 als robuster Schätzer für die Varianz bei den PCA Berechnungen verwendet. Die Score Vektoren berechnen sich für die j-te Hauptkomponente als

$$z_{ij} = \boldsymbol{a_j}^\top \boldsymbol{x_i} \quad \text{für } i = 1, \cdots, n$$

und die Scorematrix berechnet sich somit als

$$\boldsymbol{Z_k} = \boldsymbol{X}\boldsymbol{A_k},$$

wobei $\boldsymbol{A_k}$ hier wie zuvor für die Ladungsmatrix der ersten k Hauptkomponenten steht. Die Berechnung der Hauptkomponenten erfolgt mit einem Grid Algorithmus, auf den in Unterabschnitt 2.2.4 näher eingegangen wird.

2.2.3 Robust and sparse PCA

Da anhand der Elemente in der Ladungsmatrix interpretiert wird, welche Merkmale die prägnantesten für die Hauptkomponente sind, und die Auswahl nicht immer eindeutig getroffen werden kann, wurde von Croux, Filzmoser und Fritz (2013) eine zusätzliche Restriktion eingeführt, damit weniger relevante Ladungen gegen Null streben. Unter der robust and sparse Hauptkomponentenanalyse erhält man somit die Ausdünnung der Ladungsmatrix. Das Optimierungsproblem ist definiert als

$$\max_{\|\boldsymbol{a}\|=1, \boldsymbol{a}\perp\boldsymbol{a_1},\cdots,\boldsymbol{a}\perp\boldsymbol{a_{j-1}}} \boldsymbol{a}^\top \hat{\Sigma} \boldsymbol{a} \tag{2.2}$$

$$\text{so dass } \|\boldsymbol{a}\|_1 \leq t, \quad \text{für } 1 \leq j \leq p. \tag{2.3}$$

Die L_1-Norm $\|\boldsymbol{a}\|_1 = \sum_{j=1}^{p} |\boldsymbol{a_j}|$ ist definiert als Summe der Beträge des Vektors $\boldsymbol{a}$, wobei $\hat{\Sigma}$ die empirische Kovarianzmatrix ist. Das Problem lässt sich auch dual formulieren und ist in dieser Form einfacher in der Handhabung,

$$\max_{\|\boldsymbol{a}\|=1, \boldsymbol{a}\perp\boldsymbol{a_1},\cdots,\boldsymbol{a}\perp\boldsymbol{a_{j-1}}} \boldsymbol{a}^\top \hat{\Sigma} \boldsymbol{a} - \lambda_1 \|\boldsymbol{a}\|_1. \tag{2.4}$$

Der Parameter λ_1 beschreibt einen Tuningparameter. Dies ist in der Gleichung (2.4) leicht zu erkennen: Je größer λ_1, desto mehr Ladungen werden gegen Null gedrückt. Fügt man nun die Restriktion in Gleichung (2.3) ein, so erhält man für die erste Hauptkomponente

$$\boldsymbol{a}_1 = \underset{\|\boldsymbol{a}\|=1}{\operatorname{argmax}} \operatorname{V}(\boldsymbol{a}^t \boldsymbol{x}_1, \cdots, \boldsymbol{a}^t \boldsymbol{x}_n) - \lambda_1 \|\boldsymbol{a}\|_1.$$

Die j-te sparse Hauptkomponente ist definiert als

$$\boldsymbol{a_j} = \underset{\|\boldsymbol{a}\|=1, \boldsymbol{a}\perp\boldsymbol{a_1},\cdots,\boldsymbol{a}\perp\boldsymbol{a_{j-1}}}{\operatorname{argmax}} \operatorname{V}(\boldsymbol{a}^\top \boldsymbol{x_1}, \cdots, \boldsymbol{a}^\top \boldsymbol{x_n}) - \lambda_j \|\boldsymbol{a}\|_1$$

mit dem Tuningparameter λ_j. Setzt man $\lambda_j = 0$ erhält man das robuste Optimierungsproblem.

2.2.4 GRID-Algorithmus

Der GRID-Algorithmus ist eine Projection-Pursuit (PP) Methode. Das Ziel dieser Methode ist es, in multivariaten Daten Strukturen durch Projektion in kleinere Dimensionen zu finden. Der GRID-Algorithmus hat Vorteile gegenüber Algorithmen, die mit Gradienten arbeiten, da das Varianzmaß bei letzterem differenzierbar sein muss, was der GRID-Algorithmus nicht erfordert. In der Arbeit von Croux, Filzmoser und Fritz (2013) wurde der GRID-Algorithmus von Croux und Ruiz-Gazen (2007) für eine sparse Lösung angepasst. Zuerst soll ein Auszug aus dem GRID-Algorithmus von Croux, Filzmoser und Olivera (2007) gezeigt werden:

Ziel des GRID-Algorithmus ist es, für $\operatorname{V}(\boldsymbol{a_k})$ das maximale $\boldsymbol{a_k}$ zu finden, wobei $\|\boldsymbol{a_k}\| = 1$ sein muss. Für $p = 2$ ist das Problem einfach, da die Funktion $\theta \to \operatorname{V}(\cos(\theta), \sin(\theta))^\top$ im

Intervall $[-\frac{\pi}{2}, \frac{\pi}{2})$ zu maximieren ist. Dazu zerlegt man das Intervall in (N_g-1) Teilintervalle und wertet die Funktion an den Stützstellen $(-\frac{1}{2} + \frac{j}{N_g})\pi$ für $j = 0, \cdots, N_g - 1$ aus. Für $p > 2$ muss man hingegen iterativ vorgehen:

1. Ordne die Spalten der Datenmatrix $\boldsymbol{X}$ absteigend nach dem Varianzmaß V, sodass $\mathrm{V}(\boldsymbol{e_1}) \geq \cdots \geq \mathrm{V}(\boldsymbol{e_p})$, wobei $\boldsymbol{e_1}, \cdots, \boldsymbol{e_p}$ die orthogonalen Einheitsvektoren sind. Starte mit $\hat{\boldsymbol{a}} = \boldsymbol{e_1}$.

2. Für $i = 1, \cdots, N_g$
   Für $j = 1, \cdots, p$

   - Maximiere die Funktion $\theta \to \mathrm{V}(\cos(\theta\hat{\boldsymbol{a}}), \sin(\theta\boldsymbol{e}_j))^\top$ im Raum, der aufgespannt wird von den beiden Vektoren $\hat{\boldsymbol{a}}$ und $\boldsymbol{e_j}$. Der Winkel θ ist beschränkt durch das Intervall $[-\frac{\pi}{2^i}, \frac{\pi}{2^i})$. Der Winkel, für den das Maximum angenommen ist, wird mit θ_0 bezeichnet.
   - Aktualisiere $\hat{\boldsymbol{a}} \leftarrow \cos(\theta_0\hat{\boldsymbol{a}}) + \sin(\theta_0\boldsymbol{e_j})$

Während eines Zyklus werden $p-1$ Suchsequenzen durchgeführt und in der j-ten GRID-Suche wird die j-te Koordinate von $\hat{\boldsymbol{a}}$ aktualisiert. Die GRID-Suche wird bei jedem Durchlauf aufgrund des definierten Intervalls $[-\frac{\pi}{2^i}, \frac{\pi}{2^i})$ bei gleich bleibenden N_g verkleinert. Die ersten Durchläufe dienen dazu, den Bereich des Maximums zu finden und die späteren Durchläufe, um die Lösung zu präzisieren.

Für die robust and sparse PCA muss der Algorithmus adaptiert werden. Angenommen die ersten $j-1$ Richtungen der Hauptkomponenten wurden bereits gefunden und stehen als Spalten in der Ladungsmatrix $\boldsymbol{X_i^{(j-1)}}$. Man möchte nun die $\tilde{\boldsymbol{x}}_{\boldsymbol{j}}$ berechnen. Dafür sei $\tilde{\boldsymbol{A}}_{\boldsymbol{0}}^{\perp}$ äquivalent zur Einheitsmatrix, für $j > 1$ sei $\tilde{\boldsymbol{A}}_{\boldsymbol{j-1}}^{\perp}$ die Matrix, welche eine orthogonale Basis für den Unterraum in den Spalten enthält, die orthogonal zu den ersten $j-1$ sparsen Hauptkomponenten stehen. Zu maximieren ist nun die Funktion

$$f(\boldsymbol{a}) = \mathrm{V}(\boldsymbol{a}^\top \boldsymbol{x_1^{(j-1)}}, \cdots, \boldsymbol{a}^\top \boldsymbol{x_n^{(j-1)}}) - \lambda_j \|\tilde{\boldsymbol{A}}_{\boldsymbol{j-1}}^{\perp}\|_1,$$

unter der Restriktion $\|\boldsymbol{a}\| = 1$, hierbei sei $\boldsymbol{x_i^{(j-1)}} = (\boldsymbol{A}_{\boldsymbol{j-1}}^{\perp})^\top \boldsymbol{x_i}$ für $i = 1, \cdots, n$, welche zum Unterraum $\mathbb{R}^{p-j+1}$ gehört. Bevor man die L_1-Norm von $\boldsymbol{a}$ berechnen kann, muss man den Vektor zurücktransformieren, da die Sparseness im Originalraum und nicht im Unterraum gesucht wird. Der erste Schritt des Algorithmus ist analog zur Standard-Version.

1. Die nach ihrer Variabilität absteigend geordneten Spalten der Eingabematrix seien Bestandteil der Matrix $\boldsymbol{X}^j$. Die erste Variable hat nun die größte Variabilität mit dem Ladungsvektor $\boldsymbol{a} = (1, 0, \cdots, 0)$, welcher als erste Approximation dient.

2. Für $l = 1, \cdots, N_c$

   - Für $l \leq i \leq p - j + 1$ maximiere die Funktion

$$f(a^1 b(\gamma), \cdots, a^{i-1} b(\gamma), \cos\gamma, a^{i+1} b(\gamma), \cdots, a^{p-j+1} b(\gamma)),$$

für die Variable γ im Intervall $[\arccos(a^i) - \frac{\pi}{2^{l-1}}, \arccos(a^i) + \frac{\pi}{2^{l-1}}]$, wobei γ^* der maximale Winkel ist und $b(\gamma) = \frac{\sin(\gamma)}{\sqrt{1-(a^i)^2}}$. Der aktualisierte Wert für a^i ist dann gegeben durch $\cos(\gamma^*)$.

Die Prozedur stoppt, wenn die N_c Iterationsschritte ausgeführt sind. Der Algorithmus konvergiert, wenn die absoluten Veränderungen zwischen der optimalen Richtung $\boldsymbol{a}$ und zwei Iterationsschritten kleiner sind als in dem zuvor gesetztem Toleranzlevel $\epsilon = 10^{-k}$. Die optimale sparse Richtung $\boldsymbol{a}$ ist noch in den ursprünglichen Raum zurück zu transformieren als $\tilde{\boldsymbol{a}}_{\boldsymbol{j}} = \tilde{\boldsymbol{A}}^{\perp}_{\boldsymbol{j-1}}\boldsymbol{a}$.

2.2.5 Wahl des Tuningparameters

Croux, Filzmoser und Fritz (2013) haben 2 Kriterien für die Wahl des Parameters λ vorgestellt, zum Einen das BIC-Kriterium und zum Anderen die erklärte Varianz EV_k berechnet für die ersten k Komponenten.

BIC-Kriterium

Der Tuningparameter λ sollte so gewählt werden, dass der BIC (Baysian Information Criterion) minimal ist. Das BIC-Kriterium ist definiert als

$$BIC(\lambda) = \frac{\widetilde{RV}}{RV} + \mathrm{df}(\lambda)\frac{\log(n)}{n},$$

$\mathrm{df}(\lambda)$ ist die Anzahl der nicht Null Ladungen für den Parameter λ, $\widetilde{RV}$ bzw. RV sind die totalen robusten Varianzen der Residuen Matrix der sparse bzw. nicht sparse PCA. $\widetilde{RV}$ und RV sind definiert durch

$$\begin{aligned}\widetilde{RV} =& V(\boldsymbol{X} - \boldsymbol{X}\tilde{\boldsymbol{A}}_k\boldsymbol{A}_k^\top)\\ RV =& V(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}_k\boldsymbol{A}_k^\top)\\ V(\boldsymbol{X}) =& \sum_{i=1}^{p} V(x_i),\end{aligned}$$

wobei hier x_i die i-te Spalte von $\boldsymbol{X}$ und V das robuste Varianzmaß ist, und $\tilde{\boldsymbol{A}}_{\boldsymbol{k}}$ bzw. $\boldsymbol{A}_k$ die Ladungsmatrizen für die ersten k Hauptkomponenten der sparsen bzw. nicht sparsen PCA sind. Man erhält das geeignete λ dadurch, dass man den BIC über ein Intervall $[0, \lambda_{max}]$ berechnet, wobei für λ_{max} die Ladungsmatrix in jeder Spalte nur noch ein Eins Element enthält und ansonsten nur Nulleinträge.

Erklärte Varianz

Ein weiteres Hilfsmittel ist die Erklärte Varianz

$$EV_k = \frac{V(\boldsymbol{Z}_{\boldsymbol{k}})}{V(\boldsymbol{X})}$$

für die ersten k Hauptkomponenten, aufgetragen über ein Intervall der berechneten λ. Hier sollte λ so gewählt werden, dass der steile Abfall der erklärten Varianz nach dem gewählten λ kommt. Diese Wahl des Parameters λ sollte zu der des BIC konsistent sein.

Kapitel 3

Datenauswertung und -analyse

Die Berechnungen dieser Arbeit wurden im Statistik Programm `R` durchgeführt. Dabei wurde das Package `pcaPP` verwendet, dass die Hauptkomponentenanalyse mit dem Gridsearch Algorithmus berechnet. Die verwendeten Funktionen waren `qn` und `l1median` für die Skalierung und `PCAgrid` und `sPCAgrid` für die Berechnung der Hauptkomponentenanalyse mit und ohne Restriktion. Grafische Datenaufbereitung wurde sowohl in `R` als auch in `Excel` erstellt.

3.1 Aufbereitung der Datenbasis

In Abschnitt 1.2 wurde die Auswahl der Daten und der Merkmale beschrieben. Es soll nun näher auf die Berechnung und Aufarbeitung von Daten und Merkmalen eingegangen werden. Im ersten Schritt war das Ziel eine Matrix für die PCA aufzustellen, welche die Variablen in den Zeilen und die Beobachtungen in den Spalten beinhaltete. Bevor man die PCA anwenden konnte, wurden mittels Korrelationsanalyse unabhängige Variablen ausgewählt. Ein großes Problem am Anfang der Berechnung der Datenmatrix war der Umgang mit Messfehlern. Die Daten lagen in kontinuierlichen Messaufzeichnungen vor. Abbildung 3.1 zeigt die Aufzeichnungen der Fahrzeuggeschwindigkeit, der Drehzahl und des Drehmoments von Fahrzeug 2.

3.1.1 Messfehlerbehandlung

Ein Problem, das sich hier besonders zeigte, da es sich um reale Messaufzeichnungen handelte, waren Messfehler. Dabei ist zu unterscheiden zwischen einem Messfehler, der einen unplausiblen Wert beschreibt oder einem ausgefallenem Signal. Im schlimmsten Fall war ein Messkanal bei Totalausfall gar nicht vorhanden, was bei essentiellen Kanälen zur Folge hatte, dass dieser Messzyklus für die PCA unbrauchbar war. Für eine Plausibilitätsprüfung bezüglich der Messfehler wurden Intervallgrenzen festgelegt. Lag eine Messaufzeichnung eines Kanals außerhalb dieses Normbereichs, wurde dieser für die Merkmalsberechnung auf „NaN“ („not a number“) gesetzt.

Ausgefallene Signale waren im Gegensatz zu Messfehlern nicht so einfach herauszufiltern. Diese zeichneten sich dadurch aus, dass sie oft innerhalb des plausiblen Normbereichs Werte annahmen, die über einen gewissen Zeitraum konstant waren.

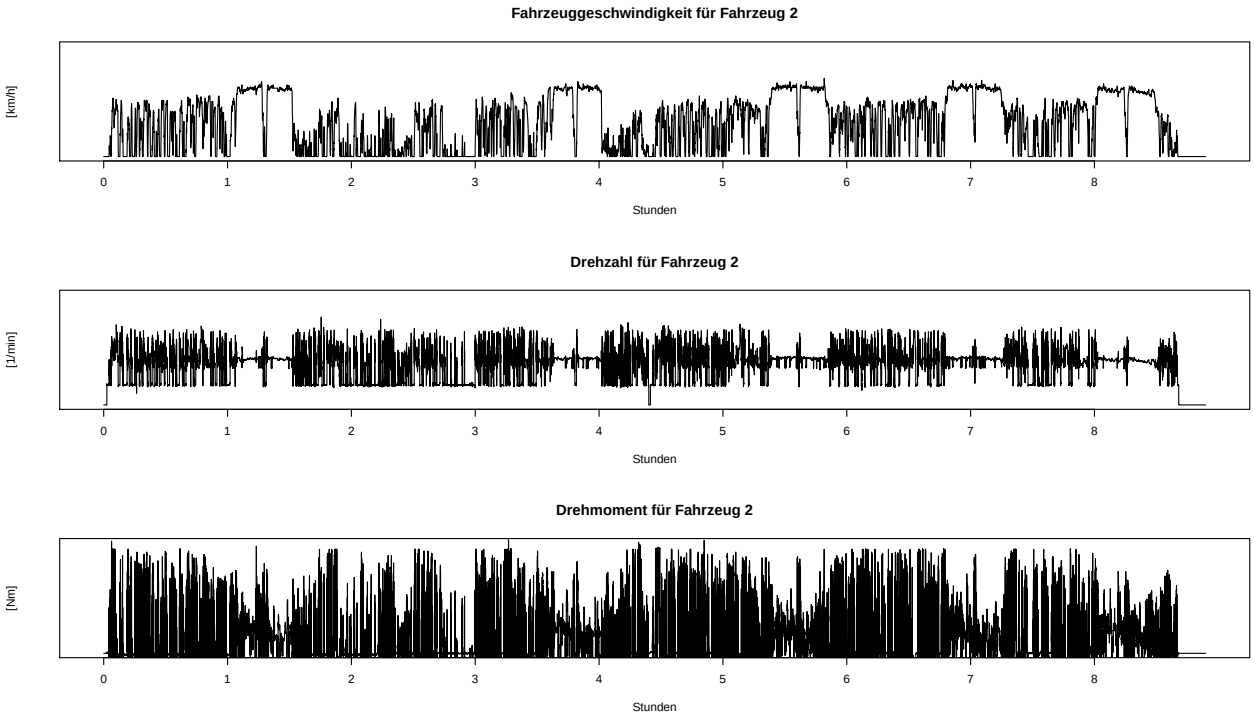


Abbildung 3.1: Zeitreihen der Fahrzeuggeschwindigkeit, der Drehzahl und des Drehmoments von Fahrzeug 2

Da eine optische Prüfung der Kanäle auf ausgefallene Signale aller Fahrzeuge vor Berechnung der Merkmale nur mit einem sehr hohen Zeitaufwand realisierbar gewesen wäre, haben sich ausgefallene Signale teilweise erst bei der Korrelationsanalyse und bei Betrachtung von Scatterplots gezeigt. Als Beispiel ist in Abbildung 3.2 die *Außentemperatur* mit den Quantilen 10%, 50% und 90% angeführt, deren ausgefallenes Signal einen Wert knapp unter 0 C° hat, welcher in der Grafik sehr gut sichtbar ist. Beobachtungen mit ausgefallenen Signalen wurden nachträglich entfernt.

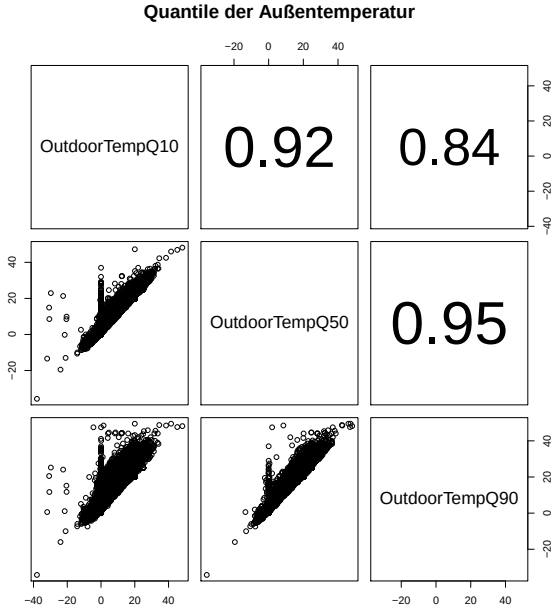


Abbildung 3.2: Scatterplot der Quantile 10%, 50% und 90% der Außentemperatur

3.1.2 Aufbereitung der Datenmatrix für PCA

Im vorherigen Schritt wurden die Daten durch Filtern der Messfehler und ausgefallener Signale auf eine Datenmatrix der Größe $n \times p = 3.455 \times 363$ (3.455 Messzyklen und 363 Merkmale) reduziert. Da die jeweiligen Fahrzeuge verschiedene Nutzungsprofile, Einsatzgebiete und Motorenklassen hatten und sich die Messzyklen nicht gleichmäßig auf diese Größen verteilten, wurde bei dem Ziehen der Stichproben auf Ausgewogenheit geachtet. Als Beispiel lässt sich anführen, dass Stadtzyklen im Gesamtverhältnis proportional öfter gezogen wurden als Fernverkehrszyklen und dadurch eine Ausgewogenheit zwischen den Betriebsarten erzielt werden konnte. Fahrzeuge von Kundenerprobungen wurden sehr oft zur Gänze herangezogen, da diese eher selten waren. Durch diese Auswahl konnte man die Kategorien ausgleichen. Tabelle 3.1 zeigt die Verteilung der Zyklen nach der Auswahl.

Es wurden nach der Auswahl der Zyklen zehn Datensets durch Ziehen ohne Zurücklegen unter einer Gleichverteilung gezogen. Diese hatten eine Größe mit $p = 363$ Merkmalen und $n = 888$ Messzyklen.

Um die Merkmalsanzahl zu reduzieren, betrachtete man die Korrelationen innerhalb der Merkmalsgruppen. Bei Merkmalen innerhalb einer Gruppe, die hoch miteinander korrelieren, einen hohen linearen Zusammenhang zeigen, ist es nicht notwendig alle in die Auswertung mit einzubeziehen. Als Beispiel betrachte man die Quantile der *OxiCatInTemp* und *OxiCatOutTemp* in Abbildung 3.3.

Hier sieht man eine hohe Korrelation der *OxiCatInTemp* mit den *OxiCatOutTemp* Variablen für die jeweiligen Quantile. Die Variable *OxiCatInTempQ10* korreliert mit *OxiCatOutTempQ10* mit $r = 0.98$. Für weitere Auswertungen wurden daher nur Quantile der

Betriebsart	Mot. 1	Mot. 2	Mot. 3	Gesamt
Fernverkehr Passstrecke	60	60	22	142
Mischbetrieb	60	60	0	120
Kundenbetrieb	0	79	60	139
Fernverkehr	45	30	75	150
Stadt Verteilerbetrieb	0	45	0	45
Verteilerbetrieb	44	30	0	74
Fernverkehr hohe Beladung	61	59	98	218
Gesamt	270	363	255	888

Tabelle 3.1: Übersicht für Auswahl der Fahrzyklen

OxiCatInTemp Variablen verwendet. Da die *DOCInTemp* Variablen untereinander auch noch stark korrelierten, konnte man hier auch Variablen vernachlässigen. Aufgrund der Analyse der Korrelationen wurde für die Auswertung eine Anzahl von $p = 128$ Merkmalen festgesetzt.

Da bei der robust and sparse PCA der Q_n-Schätzer als Varianzmaß verwendet wurde und dieser bei 16 Merkmalen den Wert Null ergab, wurden diese Variablen aus der Auswahl entfernt. Es zeigte sich, wenn bei der robust and sparse PCA der Q_n-Schätzer bei Merkmalen Null war, dass die PCA eine degenerierte Lösung lieferte, was bei der Skalierung durch eine Division durch Null hervorgerufen wurde. Daher reduzierten sich die ausgewählten 128 Merkmale auf 105. Die 105 Merkmale enthielten viele Größen der Abgasnachbehandlung, welche sich als ausschlaggebend herausstellten. Da Größen der Abgasnachbehandlung jedoch sehr applikationsspezifisch sind, sollten diese teilweise aus den Analysen ausgeschlossen werden. Dazu wurden zwei Einschränkungen definiert, die nur Fahrzeug-, Umgebungs- und Motorgrößen behinhalteten. Daher lässt sich zusammenfassend ausführen:

- Die Basisauswahl enhält 105 Variablen des Fahrzeugs, des Motors und der Abgasnachbehandlung.
- Die Auswahl 1. Einschränkung enthält 76 Variablen des Fahrzeugs und des Motors.
- Die Auswahl 2. Einschränkung enthält 28 Variablen des Fahrzeugs.

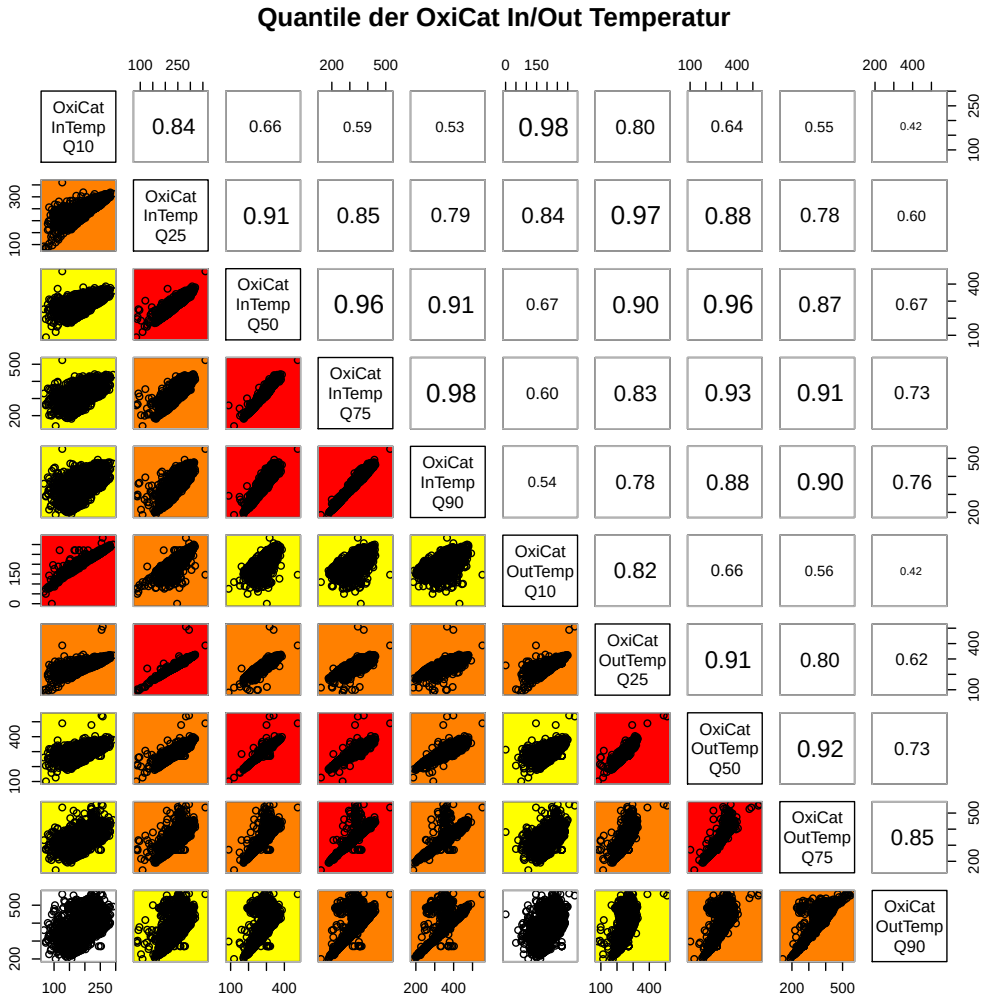


Abbildung 3.3: Scatterplot der Korrelation von Oxidationskatalysatoren Temperaturen

3.2 Klassische PCA

Um eine Basis für Vergleiche der robusten und der robust and sparse Hauptkomponentenanalyse zu erhalten, wurde zuerst eine klassische Hauptkomponentenanalyse durchgeführt. Hierfür wurden die Variablen der Eingabematrizen mit dem Mittelwert zentriert und der Standardabweichung skaliert. Für die Basisauswahl wurden 11, für die 1. Einschränkung 8 und für die 2. Einschränkung 4 Hauptkomponenten ausgewählt. Daraus ergab sich die in Tabelle 3.2 dargestellte erklärte Varianz für die jeweiligen Einschränkungen und Datensets. Die erklärte Varianz ist hier durchwegs über 70%, wobei sie im Mittel über die jeweiligen Einschränkungen an Prozentpunkten abnimmt. Innerhalb der Einschränkungen gibt es keine wesentlichen Schwankungen.

	Erklärte Varianz für PCA klassisch		
	Basisauswahl	1. Einschränkung	2. Einschränkung
Datenset 1	77,69 %	76,40 %	73,15 %
Datenset 2	77,08 %	75,39 %	72,29 %
Datenset 3	77,84 %	76,29 %	73,28 %
Datenset 4	77,17 %	75,78 %	72,94 %
Datenset 5	76,81 %	74,89 %	71,14 %
Datenset 6	77,49 %	76,22 %	72,55 %
Datenset 7	77,97 %	76,68 %	74,06 %
Datenset 8	77,63 %	76,15 %	72,79 %
Datenset 9	78,04 %	76,56 %	73,28 %
Datenset 10	77,46 %	76,02 %	72,75 %

Tabelle 3.2: Erklärte Varianz für klassische PCA

In Abbildung 3.4 sind die absoluten Ladungen für die ausgewählten Hauptkomponenten dargestellt. Diese wurden anhand der Einschränkungen geordnet, denn für die Variablen der jeweiligen Einschränkunge gilt: 2. Einschränkung $\subset$ 1. Einschränkung $\subset$ Basisauswahl. Diese Teilmengenhierarchie wurde von links nach rechts entlang der x-Achse aufgetragen. Diese Ordnung der Variablen bringt den Vorteil, dass man das Verhalten der Variablen über die Einschränkungen hinweg besser beobachten kann.

Stellvertretend für alle anderen Datensets wurde hier das Datenset 1 ausgewählt, da es keine signifikanten Unterschiede zu den weiteren 9 Datensets gab. Der erste Blick auf Abbildung 3.4 zeigt eindeutig Rauschen. Da Variablen mit hohem Erklärungswert in absoluten Zahlen auch einen hohen Ladungswert haben, ist die Interpretation, welche Variablen die Daten am besten beschreiben, schwierig. Bei einer detaillierten Betrachtung kann man Tendenzen erkennen, wie zum Beispiel, dass die Hauptkomponenten 9 und 11 in der Basisauswahl bei den Variablen 35 und 36 ausschlagen. Eine eindeutige Tendenz ist jedoch nicht erkennbar. Abbildung 3.4 zeigt daher die Motivation für eine sparse Lösung, da bei dieser die unbedeutenden Variablen durch die Restriktion gegen null gedrückt werden und die Interpretation dadurch sehr erleichtert wird.

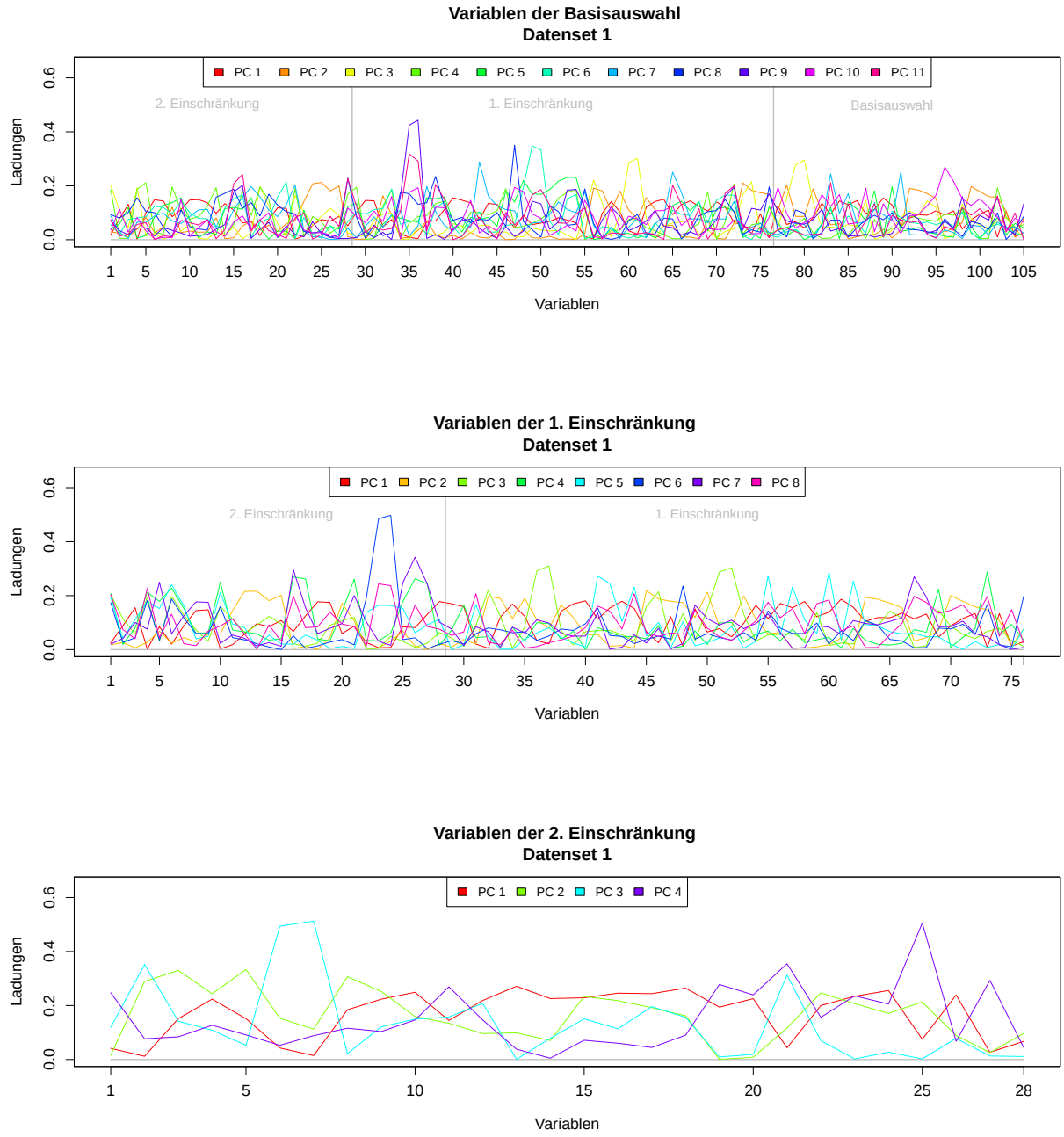


Abbildung 3.4: Darstellung der klassischen Ladungen anhand der Daten

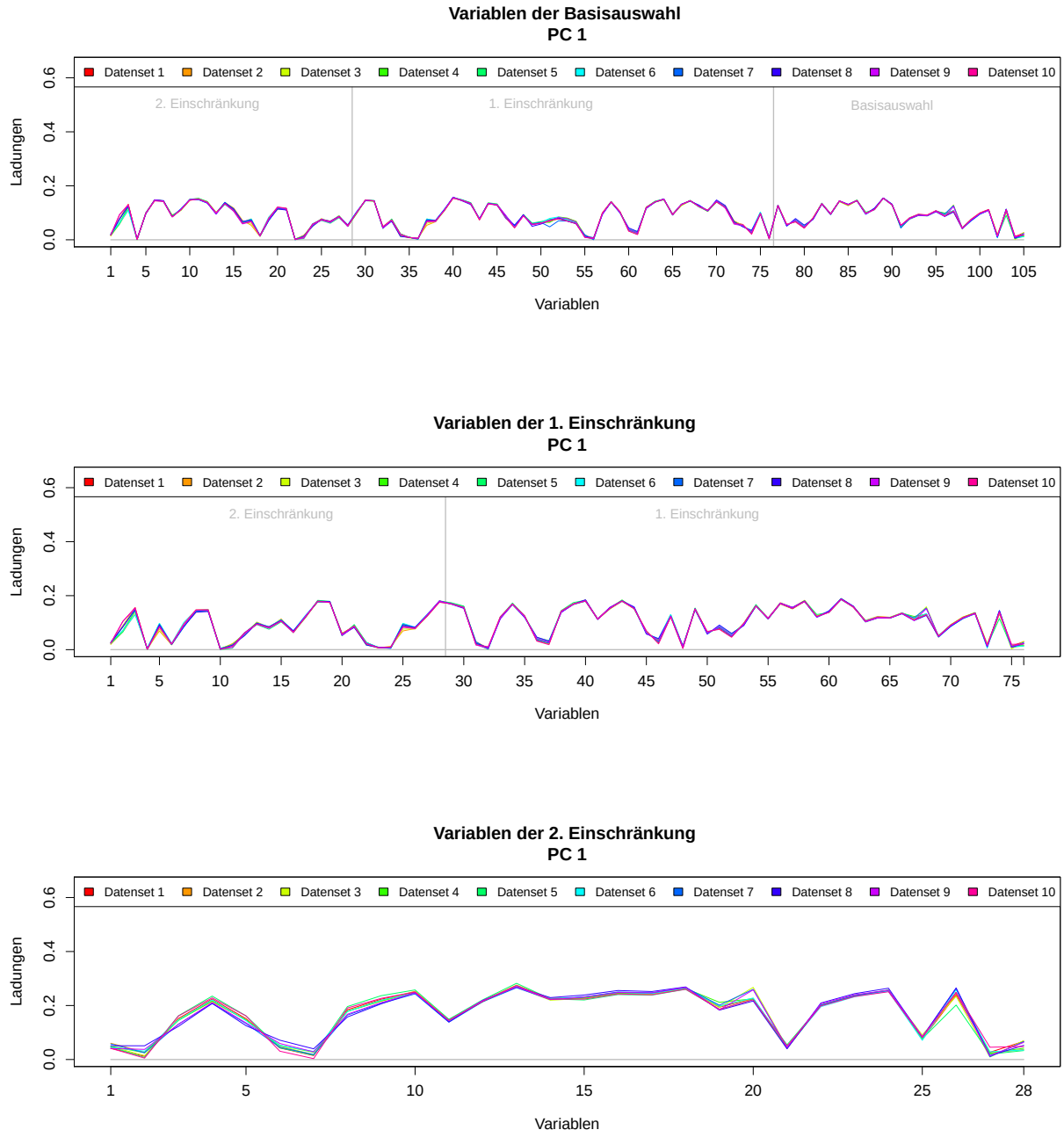


Abbildung 3.5: Darstellung der klassischen Ladungen anhand der Hauptkomponenten

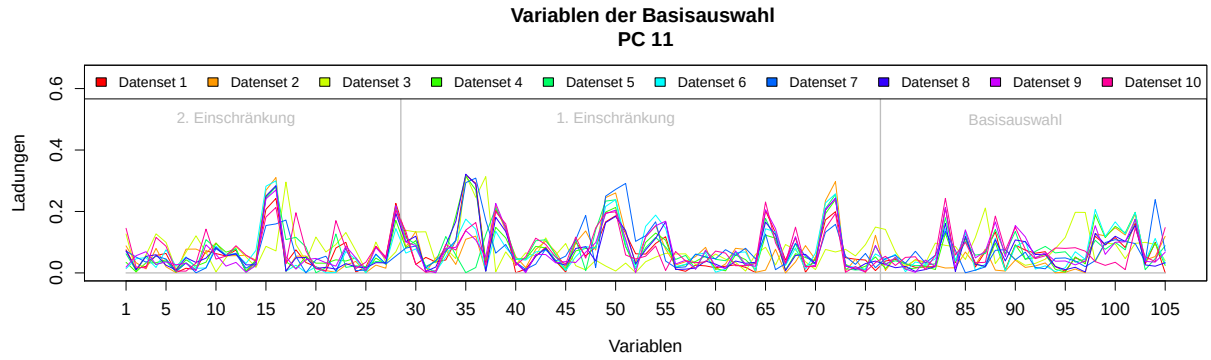


Abbildung 3.6: Darstellung der klassischen Ladungen anhand der Hauptkomponenten für Basisauswahl und Hauptkomponente 11

Eine Beobachtung bezüglich der Einschränkungen kann jedoch gemacht werden. So schlagen in der Basisauswahl wie zuvor erwähnt die Variablen 35 und 36 der Hauptkomponenten 9 und 11 stark aus, dieses Bild gibt jedoch die 1. Einschränkung nicht wieder. Hier ist die Variable 35 eher unauffällig und die Variable 36 hat nur in der Hauptkomponente 3 einen hohen Ladungswert. Hauptkomponente 6 zeigt in der 1. Einschränkung hohe Ladungswerte für die Variablen 23 und 24 und die Hauptkomponenten 5 und 7 haben hohe Ladungen bei den Variablen 26 und 27. Es zeigt sich, dass nicht jede Variable in jedem Datenset als gleich signifikant erachtet wird.

Abbildung 3.5 liefert die Darstellung der ersten Hauptkomponente für die 10 Datensets entlang der Variablen. In der ersten Haupkomponente kann man über alle Einschränkungen hinweg eine gute Übereinstimmung erkennen. Diese nimmt jedoch mit der numerischen Höhe der Hauptkomponenten ab, wie in der als Beispiel angeführten Abbildung 3.6 für die Basisauswahl und die Hauptkomponente 11.

3.3 Robuste PCA

In der robusten Hauptkomponentenanalyse wurde als Varianzmaß für die Berechnungen der Q_n^2-Schätzer gewählt, da dieser Vorteile gegenüber dem MAD_n^2-Schätzer hat, welcher auch als alternatives Varianzmaß verwendet werden kann. Diese Vorteile wurden bereits in Unterabschnitt 2.2.1 diskutiert. Anzumerken ist, dass die robuste Hauptkomponentenanalyse zwei robuste Schritte beinhaltet, da die Variablen bereits mit robusten Schätzern berechnet wurden. Zuerst wird ein Blick auf die erklärte Varianz geworfen. Die erklärte

	Erklärte Varianz PCA robust		
	Basis Auswahl	1. Einschränkung	2. Einschränkung
Datenset 1	82,91 %	81,29 %	68,99 %
Datenset 2	83,57 %	80,85 %	70,66 %
Datenset 3	82,35 %	80,20 %	68,22 %
Datenset 4	82,69 %	80,10 %	69,15 %
Datenset 5	82,33 %	79,79 %	68,76 %
Datenset 6	82,15 %	80,72 %	68,10 %
Datenset 7	84,68 %	82,05 %	69,26 %
Datenset 8	82,70 %	80,75 %	67,60 %
Datenset 9	81,97 %	79,88 %	67,07 %
Datenset 10	81,56 %	79,18 %	70,32 %

Tabelle 3.3: Erklärte Varianz für robuste PCA

Varianz der Datensets innerhalb der Einschränkungen scheint stärker zu schwanken als im Vergleich zu der erklärten Varianz der klassischen Hauptkomponentenanalyse in Tabelle 3.2. Vergleicht man die Interquartilbereiche der Einschränkungen so wird dies auch deutlich sichtbar, so ist der Interquartilbereich bei der klassischen Hauptkomponentenanalyse für die Basisauswahl 0,56%, für die 1. Einschränkung 0,53% und für die 2. Einschränkung 0,65%. Für die robuste Hauptkomponentenanalyse ergeben sich die Werte 0,66%, 0,89% und 1,1%. Weiters fällt auf, dass die Prozentsätze der Basisauswahl und der 1. Einschränkung im Median höher sind als bei der klassischen Hauptkomponentenanalyse. Der Median für die Basisauswahl bei der klassischen Hauptkomponentenanalyse lautet 77,56% und bei der 1. Einschränkung 76,18%. Im Vergleich liegen die Mediane der robusten Hauptkomponentenanalyse bei 82,52% und 80,46%. Das Gegenteil ist bei der 2. Einschränkung der Fall, hier liegt der Median der klassische Hauptkomponentenanlyse bei 72,87% und der Median der robusten bei 68,88%. In Abbildung 3.7 findet sich die Erklärung für den höheren Wert der erklärten Varianz.

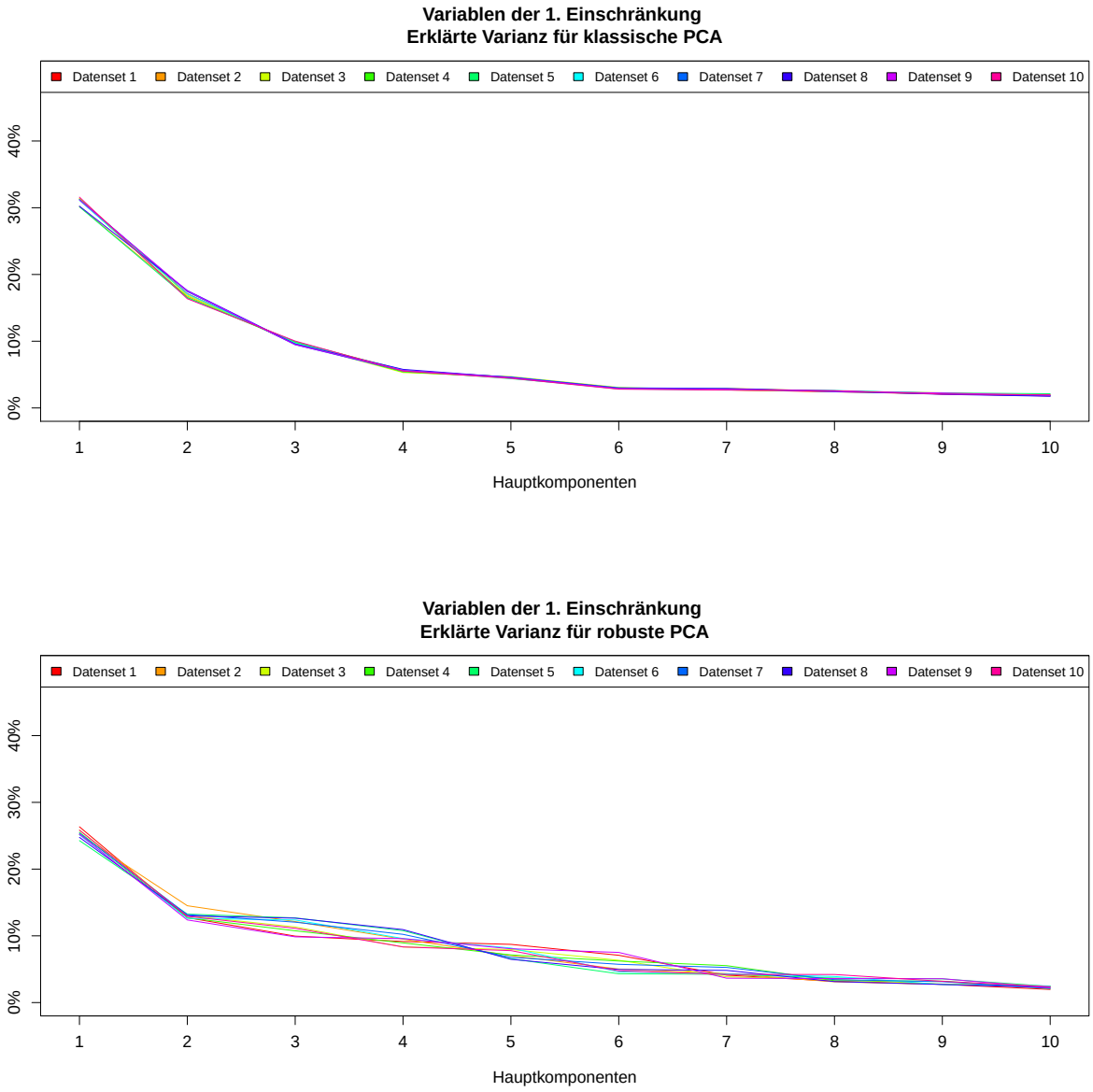


Abbildung 3.7: Darstellung der erklärten Varianz für die klassische und robuste Hauptkomponentenanalyse für die 1. Einschränkung

Abbildung 3.7 zeigt den Verlauf der erklärten Varianz über die ersten 10 Hauptkomponenten hinweg für die 1. Einschränkung anhand der klassischen Hauptkomponentenanalyse und der robusten Hauptkomponenetenanalyse. Man kann gut erkennen, dass die ersten Hauptkomponenten der klassischen Methode prozentuell mehr erklären und die Kurve einen steileren Abfall hat als die robuste Methode. Da in der 1. Einschränkung 8 Hauptkomponenten gewählt wurden, liegt die erklärte Varianz bei der robusten Methode aufgrund der

weniger steilen Kurve höher. Dies erklärt auch den Unterschied im Prozentniveau der 2. Einschränkung zwischen klassischer und robuster Hauptkomponentenanalyse. Wären hier mehr Komponenten gewählt worden, wäre auch hier der Prozentsatz der robusten Methode höher. Weiters kann man beobachten, dass die erklärte Varianz zwischen den Komponenten 2 und 9 schwankt. Dies ist in der klassischen Methode nicht der Fall, wo es keine signifikanten Unterschiede zwischen den Varianzen der einzelnen Datensets gibt.

Abbildung 3.8 zeigt die Ladungen der robusten Hauptkomponenten entlang der Variablen wieder für das Datenset 1. Im Vergleich mit Abbildung 3.4 kann man hier besser erkennen, welche Variablen Einfluss haben. Es zeigen sich mehrere Variablen mit hohen Ladungswerten und es ist bereits besser ersichtlich, dass die Hauptkomponente von speziellen Variablen beschrieben wird. Eine eindeutige Zuweisung der Variablen zu den Hauptkomponenten fällt hier jedoch auch noch sehr schwer. In Abbildung 3.9 wird dies ebenso deutlich. Hier ist die erste Hauptkomponente anhand der Variablen für die 10 Datensets aufgetragen. Im Vergleich zu Abbildung 3.5 sind hier eindeutig Spitzen zu erkennen, die auf Variablen mit Erklärungswert für das Modell hinweisen. Jedoch zeigt sich, dass die Datensets untereinander in den Werten der Ladungen bereits sichtbare Abweichungen aufweisen. Das Rauschen zwischen den Datensets nimmt auch hier mit steigendem Wert der Hauptkomponenten zu.

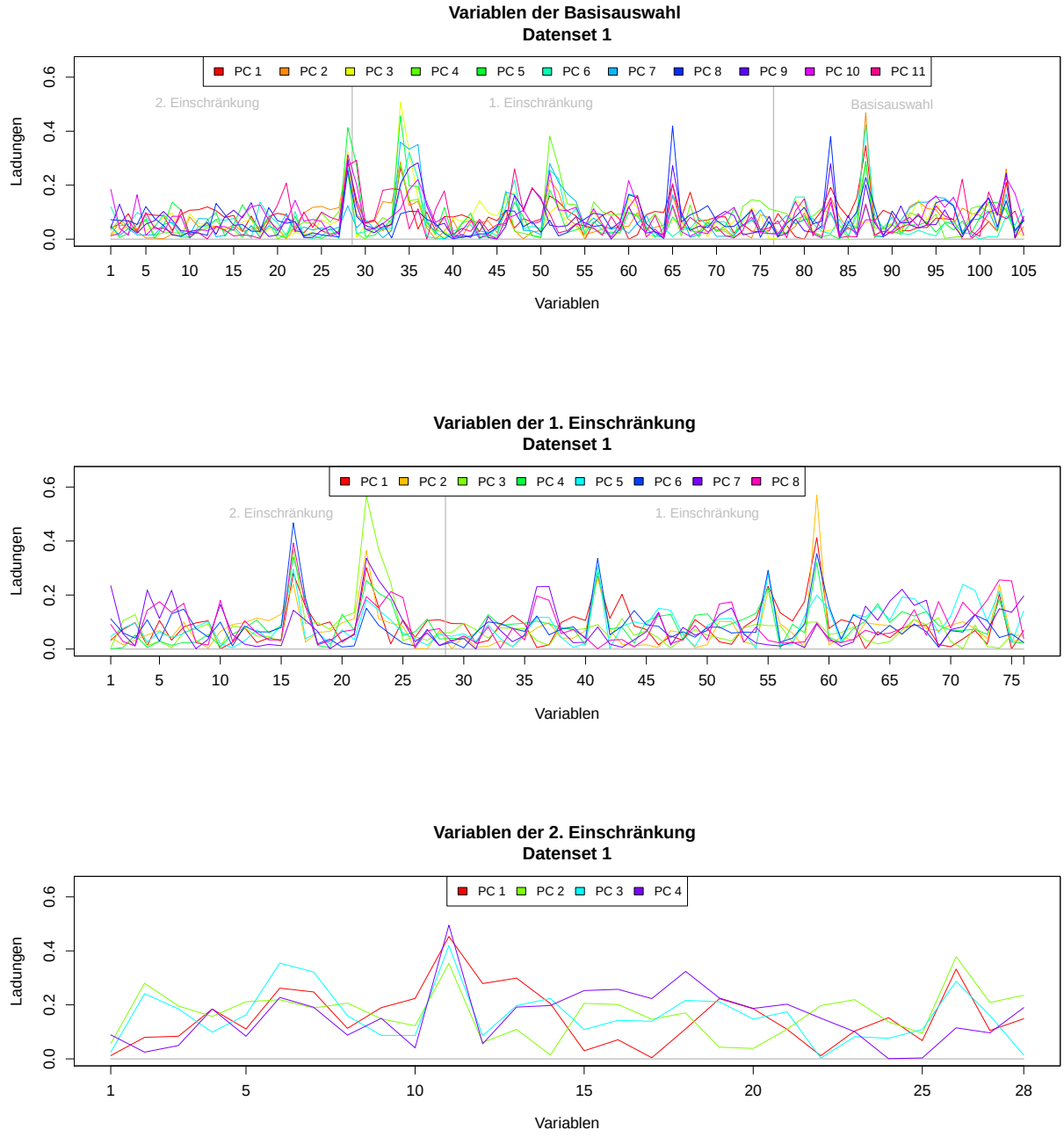


Abbildung 3.8: Darstellung der robusten Ladungen anhand der Daten

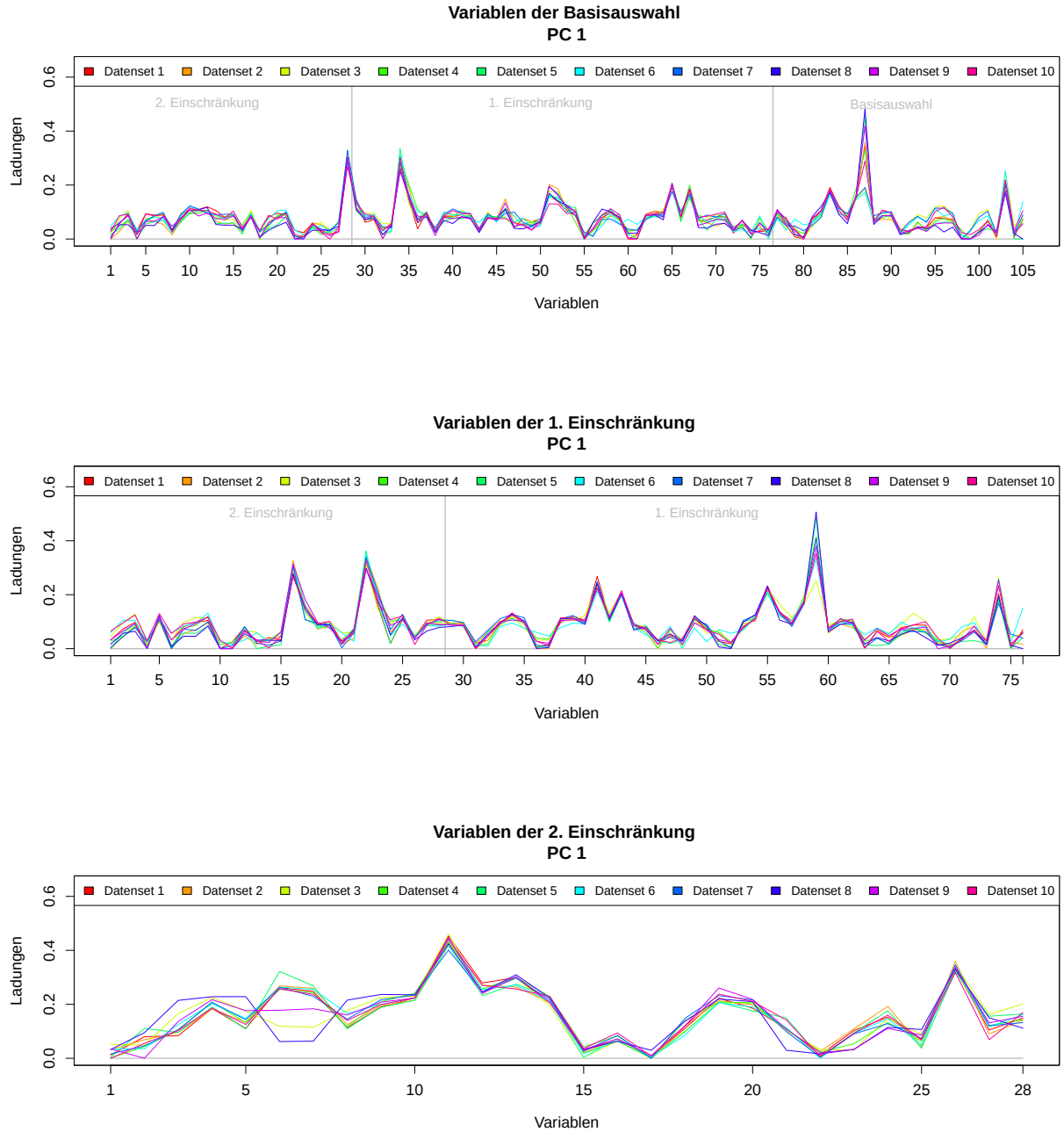


Abbildung 3.9: Darstellung der robusten Ladungen anhand der Hauptkomponenten

3.4 Robust and sparse PCA

In Abbildung 3.4 und Abbildung 3.8 ist die Problematik der Wahl geeigneter Variablen zu den jeweiligen Hauptkomponenten gut erkennbar. Dies war auch die Motivation für die robust and sparse Hauptkomponentenanalyse, die in Unterabschnitt 2.2.3 bereits theoretisch vorgestellt wurde. Um nun eine schwach besetzte Ladungsmatrix zu erhalten, muss ein geeigneter Parameter λ aus Gleichung (2.4) gefunden werden. Je höher der Wert des Parameters ist, desto dünner ist die Ladungsmatrix besetzt, woraus eine bessere Interpretierbarkeit der Variablen bezüglich der Hauptkomponenten folgt. Man kann den Parameter so weit erhöhen, dass jede Spalte nur noch mit einem Element besetzt ist. Im Gegenzug muss man berücksichtigen, dass ein hohes λ auf Kosten der erklärten Varianz geht. Um hier ein ausgewogenes Verhältnis zu erhalten, wurden die Methoden aus Unterabschnitt 2.2.5 verwendet.

Die robust and sparse Hauptkomponentenanalyse wurde in dieser Arbeit für den robusten Fall berechnet, da wir an einer robusten Lösung interessiert sind. Einen Vergleich der robust and sparse Methode im klassischen und robusten Fall findet man in der Arbeit von Filzmoser and Todorov (2013). Der wichtigste Schritt bei der robust and sparse Hauptkomponentenanalyse ist das Anpassen des Parameters . Da für jedes λ eine Hauptkomponentenanalyse berechnet werden muss, ist hier auch ein Augenmerk auf die Laufzeit zu richten. Die Laufzeiten für die robuste Methode dauerten für die Basisauswahl und die 2 Einschränkungen jeweils: Basisdatenset 16,95 Minuten, 1. Einschränkung 9,52 Minuten und 2. Einschränkung 1,47 Minuten. Verwendet wurde ein Rechner mit einem Intel(R) Core(TM) i5-2520M CPU, 2,50 GHz mit 4 GB Arbeitsspeicher. Aufgrund der langen Rechenzeiten, besonders bei dem Basisdatenset, wurde anfangs ein sehr grobes Intervall für die Berechnung gewählt, welches in mehreren Durchläufen verkleinert wurde, um Rechenzeiten zu sparen. Dabei wurde in jedem Schritt die erklärte Varianz berechnet, um zu entscheiden, welcher Bereich näher betrachtet werden soll.

Abbildung 3.10 zeigt den Verlauf der erklärten Varianz für alle berechneten λ. Das nach unseren und theoretischen Kriterien optimale λ ist mit einem Punkt in Abbildung 3.10 eingezeichnet mit dessen Wert und erklärter Varianz.

Ein Auswahlkriterium diente der praktischen Interpretation. Es wurde daher festgesetzt, dass die erklärte Varianz den Wert von 60% nicht unterschreiten soll. Diese untere Schranke ist in Abbildung 3.10 mit einer grauen Linie eingezeichnet. Die Theorie besagt, dass sich der optimale Punkt kurz vor einem steilen Abfall der erklärten Varianz befindet. Stellvertretend für alle anderen Datensets wurde hier wieder das Datenset 1 ausgewählt. Für die Basisauswahl ergibt sich für das Datenset 1 ein optimales Bild, welches nicht bei allen Datensets so eindeutig ist wie in Abbildung 3.10.

Da die optische Auswahl des Tuningparameters sehr subjektiv ausfallen kann, ist es sinnvoll, den BIC Wert für jeden Tuningparameter zu berechnen. Der minimale BIC über die Tuningparameter eines Datensets ergibt dann das optimale λ. Es stellt sich nun natürlich die Frage, warum man überhaupt eine optische Auswahl macht, wenn es doch ein eindeutiges Entscheidungskriterium gibt.

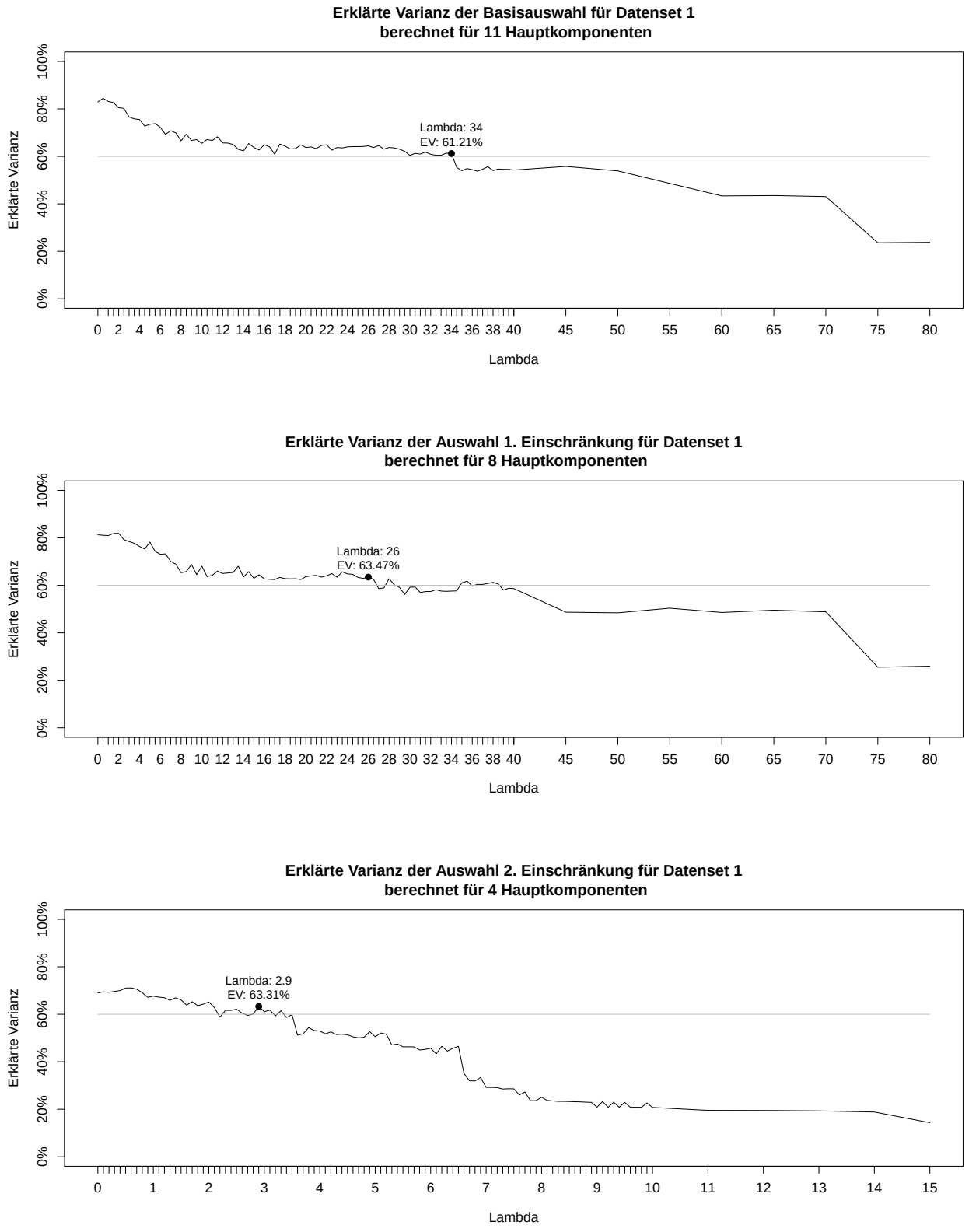


Abbildung 3.10: Darstellung der erklärten Varianz für den Parameter Lambda

Abbildung 3.11 liefert hier eine Antwort. Dargestellt ist die erklärte Varianz bzgl. der Basisauswahl für das Datenset 8 mit dem optimalen λ für den BIC-Wert und dem optisch gewählten λ. Zu sehen ist, dass das λ des BIC mit dem Wert 39 unterhalb der 60% Grenze für die erklärte Varianz liegt, die nur 50,53% beträgt. Da diese Grenze aus zuvor erwähntem Grund nicht unterschritten werden soll, wurde hier mit 27 ein λ gewählt, dass zwar noch nicht optimal ist, jedoch den Ansprüchen für weitere praktische Interpretationen erfüllt. Das BIC-Kriterium wurde daher zur Überprüfung des optisch gewählten λ's verwendet.

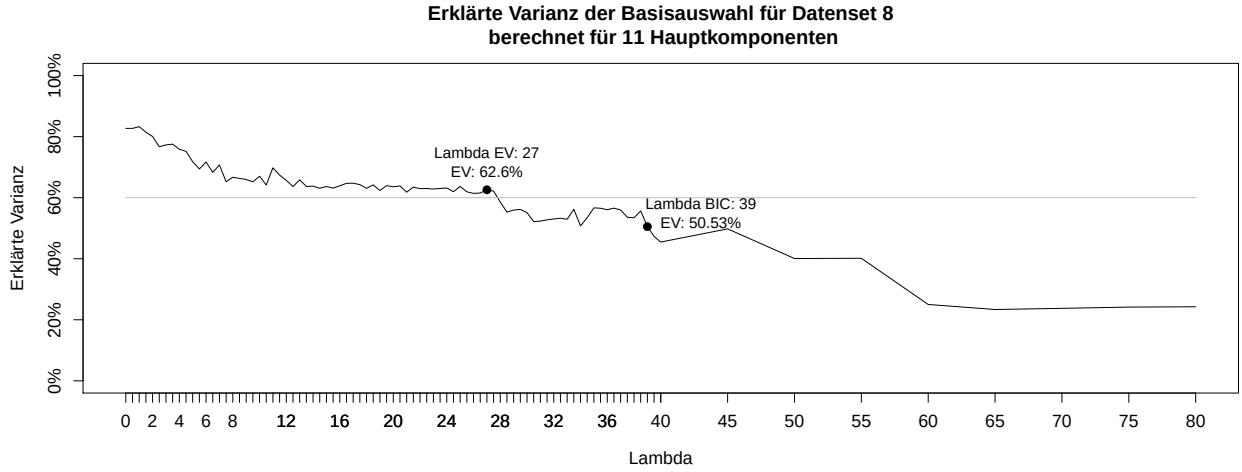


Abbildung 3.11: Darstellung der erklärten Varianz und dem optimalen BIC-Wert für die 1. Einschränkung des Datenset 8

Die BIC-Werte sind in Abbildung 3.12 dargestellt. Zusätzlich wurde die 60% Schranke der erklärten Varianz eingezeichnet.

Tabelle 3.4 zeigt eine Gegenüberstellung der λ's für die jeweiligen Datensets und Einschränkungen. Die Basisauswahl und die 1. Einschränkung haben einige optimale λ, die gleich sind. Dies wurde nicht im Nachhinein optimiert, damit sie übereinstimmen, sondern die Werte wurden anhand des optischen Kriteriums so gewählt. Dies zeigt, dass die empirische Wahl der Tuningparameter gut funktioniert. Da bei der 2. Einschränkung alle minimalen BIC-Werte eine erklärte Varianz über 60% hatten und die Auswahl eines Tuningparameters sich hier schwieriger gestaltete, wurden für die Auswahl die BIC-Werte herangezogen.

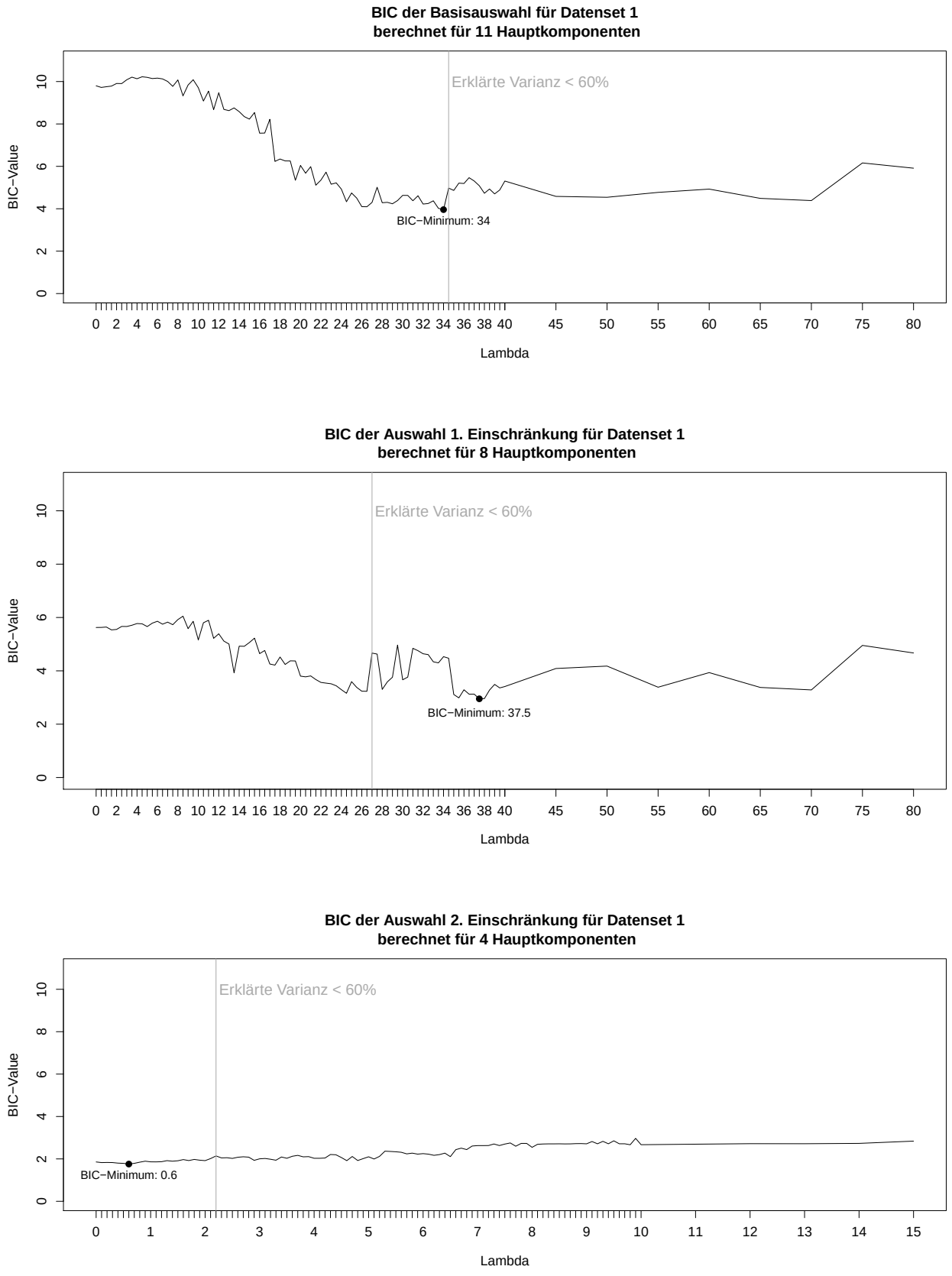


Abbildung 3.12: Darstellung des BIC für den Parameter Lambda

Einschränkung	**Basisauswahl**		**1. Einschränkung**		**2. Einschränkung**	
Entscheidungskriterium	EV	BIC	EV	BIC	EV	BIC
Datenset 1	34,0	34,0	26,0	37,5	2,9	2,9
Datenset 2	30,5	30,5	29,0	29,0	1,6	1,6
Datenset 3	20,0	39,0	26,0	37,0	2,0	2,0
Datenset 4	16,0	60,0	17,5	21,0	2,3	2,3
Datenset 5	29,5	35,0	34,0	34,0	2,1	2,1
Datenset 6	20,5	65,0	27,5	26,0	1,7	1,7
Datenset 7	25,0	39,0	29,0	29,0	2,9	2,9
Datenset 8	27,0	39,0	36,5	33,0	2,6	2,6
Datenset 9	27,0	39,0	26,5	26,5	3,1	3,1
Datenset 10	19,5	30,0	25,0	25,5	2,3	2,3

Tabelle 3.4: Vergleichende Darstellung der Entscheidungskriterien mit erklärter Varianz und BIC für die Wahl von Lambda

	Erklärte Varianz PCA robust and sparse		
	Basisauswahl	1. Einschränkung	2. Einschränkung
Datenset 1	61,21 %	63,47 %	63,31 %
Datenset 2	63,35 %	68,03 %	71,80 %
Datenset 3	62,37 %	63,93 %	65,62 %
Datenset 4	62,52 %	62,91 %	64,73 %
Datenset 5	62,60 %	63,99 %	64,02 %
Datenset 6	65,14 %	63,33 %	64,21 %
Datenset 7	62,54 %	66,55 %	62,38 %
Datenset 8	62,60 %	62,39 %	62,44 %
Datenset 9	65,97 %	64,96 %	62,23 %
Datenset 10	61,60 %	63,00 %	65,40 %

Tabelle 3.5: Erklärte Varianz für robust and sparse PCA

In Tabelle 3.5 sind die erklärte Varianz, die Einschränkungen und die jeweiligen Datensets dargestellt. Sofort fällt auf, dass die erklärte Varianz für alle Einschränkungen geringere Prozentsätze aufweist als für die klassische und die robuste Methode. Mit einem Median für die Basisauswahl von 62,57%, 63,7% für die 1. Einschränkung und 64,11% für die 2. Einschränkung liegen diese weit unter den Medianen der robusten und klassischen Methode. Somit zeigt sich hier, dass eine schwach besetzte Ladungsmatrix auf Kosten der erklärten Varianz geht. Die Schwankung, auch hier berechnet mit dem Interquartilbereich, innerhalb der Basisauswahl und der Einschränkungen liegt auch höher: 0,75% für die Basisauswahl, 1,64% für die 1. Einschränkung und 2,58% für die 2. Einschränkung. Dies folgt aus der Tatsache, dass die Tuningparameter nicht in jedem Datenset innerhalb einer Einschränkung gleich gewählt wurden. Vergleicht man nun den Verlauf der erklärten Varianz

über die ersten 10 Hauptkomponenten in Abbildung 3.13 mit Abbildung 3.7, so fällt auf, dass die ersten Hauptkomponenten einen hohen Anteil erklären und die Kurve sehr steil fällt. In der Kurve schwanken die Werte über die Datensets hinweg. Bereits in Abbildung 3.13 zeigt sich, dass es wenige signifikante Variablen gibt.

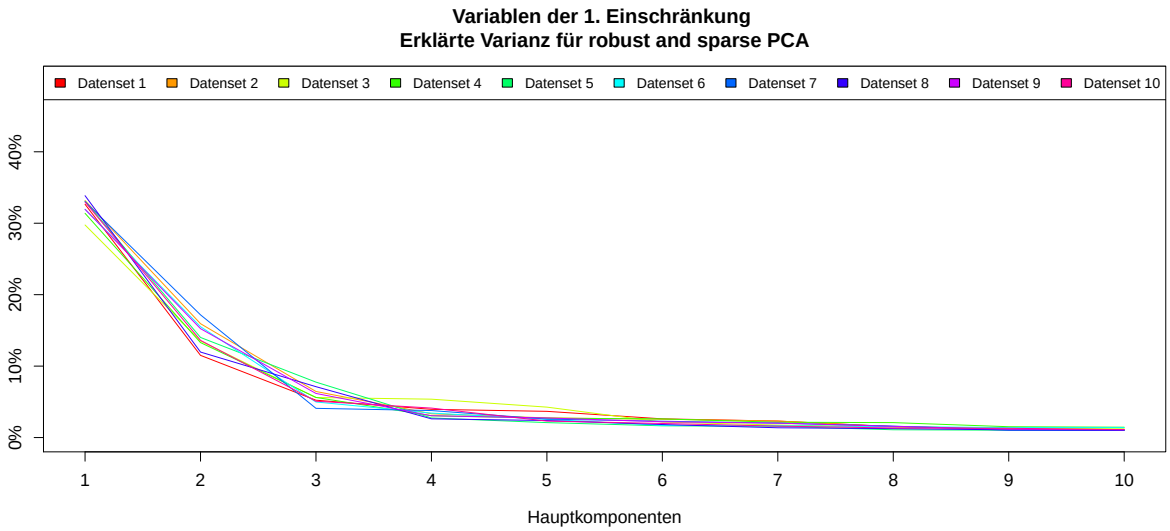


Abbildung 3.13: Darstellung der erklärten Varianz für robust and sparse Hauptkomponentenanalyse für die 1. Einschränkung

Abbildung 3.14 zeigt nun überaus deutlich den wesentlichen Vorteil der robust and sparse Methode. Im Vergleich zu Abbildung 3.4 und Abbildung 3.8 ist hier klar erkennbar, welche Variablen für welche Hauptkomponente relevant sind. Vor allem bei den großen Einschränkungen, nämlich der Basisauswahl und der 1. Einschränkung, können nun sehr viele irrelevante Variablen vernachlässigt werden, denen bei den Methoden zuvor noch eine Bedeutung zugemessen worden wäre. Das eindeutige Filtern der relevanten Variablen kann bei der 2. Einschränkung optisch noch optimiert werden. Im Vergleich mit Abbildung 3.4 und Abbildung 3.8 gibt es bereits eine wesentliche Verbesserung, jedoch haben die Hauptkomponenten 1 und 4 im Vergleich zu den Einschränkungen mit einer größeren Anzahl an Variablen überproportional viele Variablen als relevant markiert. Die Auswahl der Variablen für die Hauptkomponenten innerhalb der Einschränkungen für die Datensets ist bis auf wenige Ausnahmen konsistent, wie in Abbildung 3.15 gezeigt wird. Die 2. Einschränkung bildet hier wieder eine Ausnahme, da hier die Variablen 6 und 7 von unterschiedlichen Datensets als relevant und irrelevant ausgewählt wurden. Das Rauschen nimmt bei der 2. Einschränkung mit steigender Zahl der Hauptkomponenten zu. Anders bei der Basisauswahl und der 1. Einschränkung, hier ist auch bei den weiteren Hauptkomponenten kaum Rauschen erkennbar.

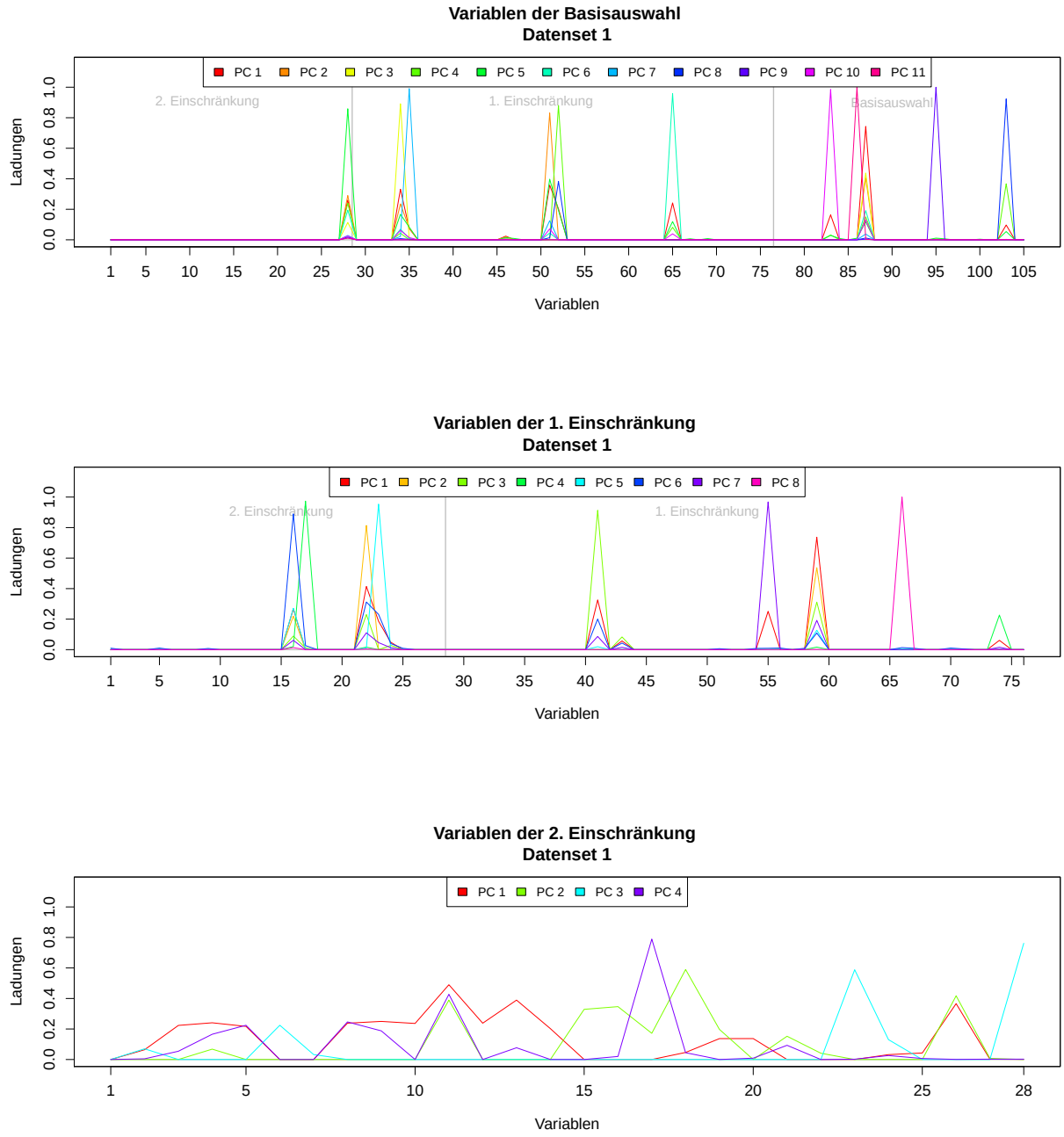


Abbildung 3.14: Darstellung der robust and sparse Ladungen anhand der Daten

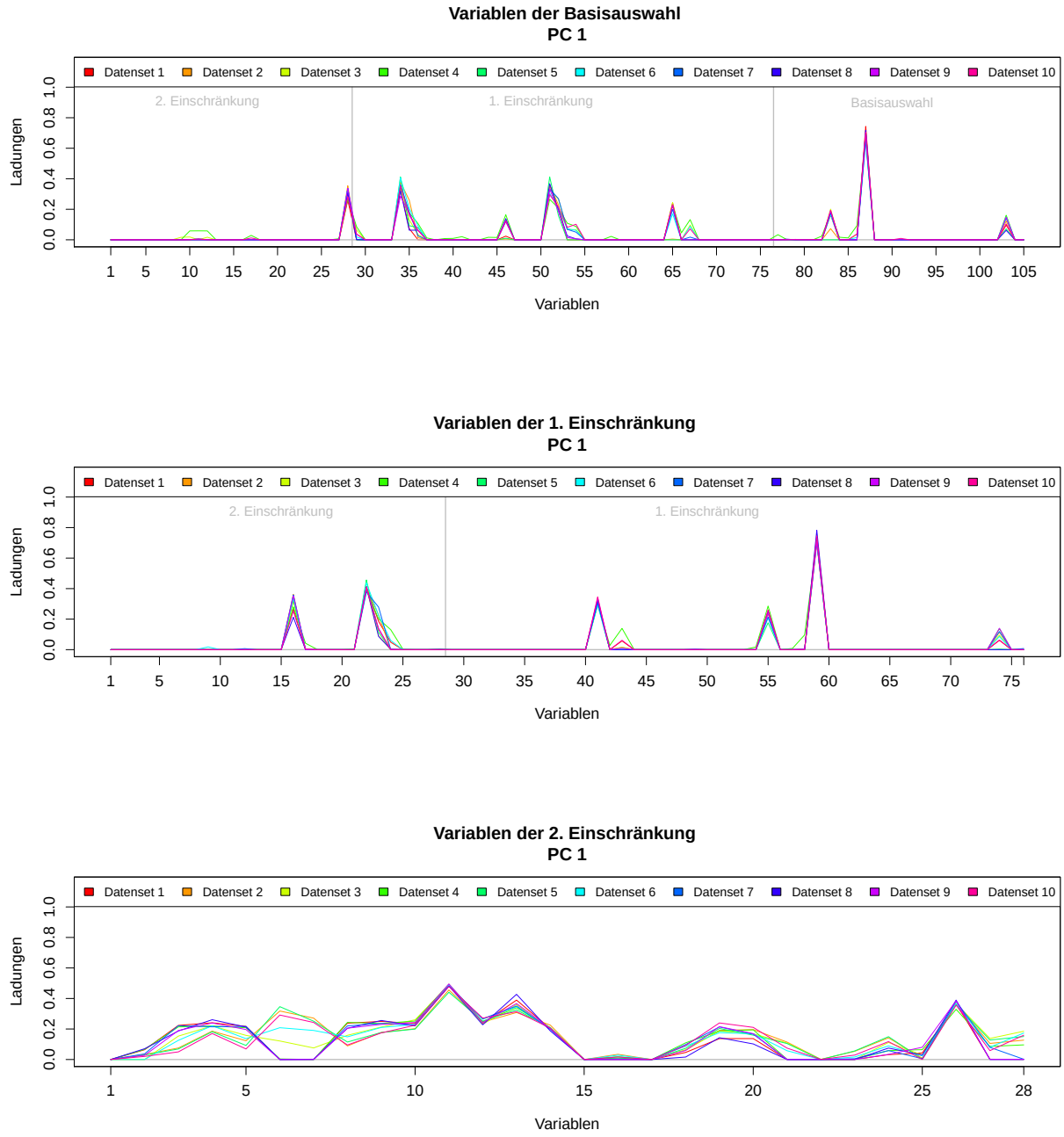


Abbildung 3.15: Darstellung der robust and sparse Ladungen anhand der Hauptkomponenten

3.5 Einfluss der Skalierung

In den vorhergehenden Abschnitten wurden die Datensets jeweils ihrer Berechnungsmethode entsprechend skaliert, das heißt, bei der klassischen Methode wurden die Datensets mit den klassischen Schätzern Mittelwert und Standardabweichung skaliert und bei der robusten Methode mit den robusten Schätzern L_1-Median und Q_n-Schätzer. Zum einen ist es interessant zu erfahren, wieviel Einfluss die Skalierung auf die Ergebnisse hat, und zum anderen hat dies auch eine praxisrelevante Motivation, da die robuste Methode, vor allem wenn man sie sparse rechnet, eine lange Laufzeit hat, die bei einer klassischen Methode mit robuster Skalierung minimiert werden könnte.

Die erklärte Varianz für die robuste Methode und die klassische Skalierung in Tabelle 3.6 liefert keine großen Auffälligkeiten, die Prozentwerte sind in einem unauffälligen Bereich und zeigen auch keine starken Schwankungen innerhalb der Datensets auf. Das Prozentniveau der 2. Einschränkung ist wieder etwas geringer als bei der robusten Methode in Tabelle 3.3.

	Erklärte Varianz PCA robust (klassisch skaliert)		
	Basisauswahl	1. Einschränkung	2. Einschränkung
Datenset 1	78,57 %	77,36 %	70,93 %
Datenset 2	79,24 %	77,49 %	72,41 %
Datenset 3	78,99 %	77,77 %	71,66 %
Datenset 4	79,13 %	78,07 %	72,08 %
Datenset 5	78,69 %	77,28 %	71,49 %
Datenset 6	78,61 %	77,01 %	70,33 %
Datenset 7	79,14 %	77,63 %	71,03 %
Datenset 8	79,29 %	77,43 %	71,07 %
Datenset 9	79,42 %	77,42 %	71,27 %
Datenset 10	79,02 %	77,53 %	71,07 %

Tabelle 3.6: Erklärte Varianz für robuste PCA mit klassischer Skalierung der Daten

Ein ganz anderes Bild zeigt sich hier in Abbildung 3.7. Die Prozenwerte der Basisauswahl und der 1. Einschränkung haben ein sehr hohes Niveau. Dieses Ergebnis würde bedeuten, dass die ersten 11 bzw. 8 Hauptkomponenten fast alles erklären und alle weiteren irrelevant wären. Die 2. Einschränkung hingegen liefert Prozentwerte in einem Bereich, den man erwarten würde. Es stellt sich nun die Frage, durch welchen Effekt ein derartiges Ergebnis hervorgerufen wird.

Zunächst wird in Abbildung 3.16 ein Blick auf den Verlauf der erklärten Varianz über die ersten 10 Hauptkomponenten geworfen. Dazu wurde wieder die 1. Einschränkung ausgewählt, um zu den Abbildungen 3.7 und 3.13 vergleichbar zu bleiben. Man kann bei der robusten Methode mit klassischer Skalierung nichts Auffälliges erkennen. Die ersten Hauptkomponenten haben einen hohen Erklärungswert, der in einer flachen Kurve abnimmt. Ganz im Gegensatz zur klassischen Methode mit robuster Skalierung, hier erklärt

	Erklärte Varianz PCA klassisch (robust skaliert)		
	Basisauswahl	1. Einschränkung	2. Einschränkung
Datenset 1	99,20 %	99,34 %	74,41 %
Datenset 2	99,18 %	99,31 %	75,86 %
Datenset 3	99,08 %	99,23 %	73,42 %
Datenset 4	99,18 %	99,31 %	76,94 %
Datenset 5	99,00 %	99,10 %	74,71 %
Datenset 6	99,21 %	99,36 %	79,16 %
Datenset 7	99,05 %	99,23 %	75,79 %
Datenset 8	99,25 %	99,39 %	73,50 %
Datenset 9	99,12 %	99,27 %	74,24 %
Datenset 10	99,15 %	99,29 %	76,22 %

Tabelle 3.7: Erklärte Varianz für klassische PCA mit robuster Skalierung der Daten

die erste Hauptkomponente an die 90%, die zweite und dritte Hauptkomponente erklären noch mehr als 1%, alle weiteren liegen darunter.

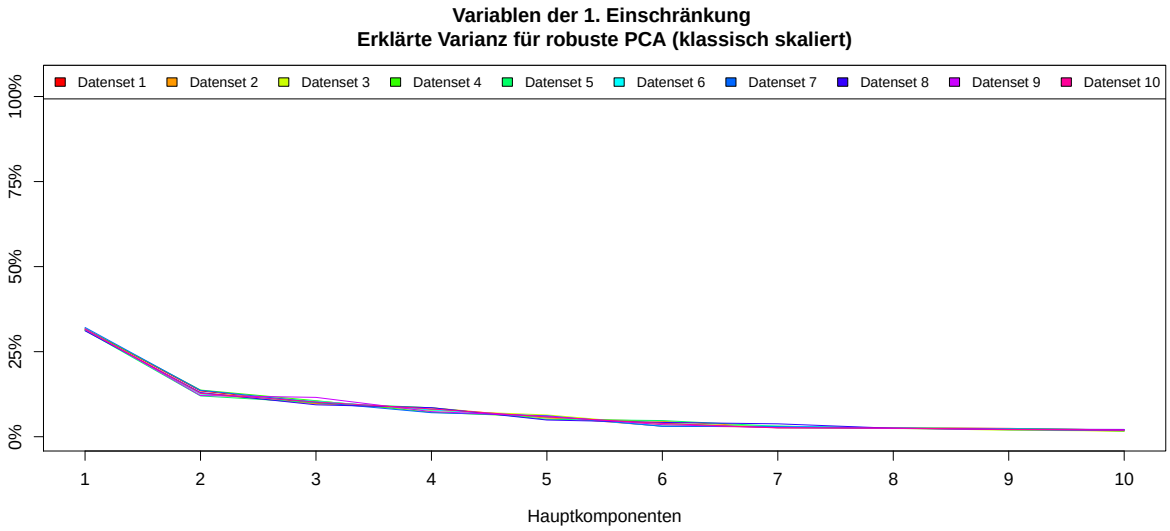
Die Variable, die hier bei der Basisauswahl und der 1. Einschränkung diesen großen Einfluss haben soll, ist das 5%-Quantil der Drehzahl. Diese ist bei der robusten und der robust and sparse Methode auch unter den einflussreichen Variablen, jedoch nicht in diesem Ausmaß. Dass die 2. Einschränkung nicht dieses abweichende Verhalten hat, wird nun auch klar, da die 2. Einschränkung nur fahrzeugbezogene Größen enthält, wie in Abschnitt 3.1 definiert und folglich diese Variable hier nicht vorkommen kann.

Kommen wir zurück zur Frage, warum diese Variable in diesem Setting soviel Einfluss hat? Die Antwort liegt hier in der Skalierung mit dem L_1-Median. Der in Unterabschnitt 2.2.1 definierte robuste Schätzer ist ein multivariater Schätzer, der im Gegensatz zu einem spaltenweisen eindimensionalen Median eine Matrix nicht auf dieselbe Weise zentriert. Wendet man den Median spaltenweise auf eine Matrix an, ist diese um den Wert null zentriert. Die spaltenweise Zentrierung einer robusten Skalierung mit L_1-Median und Q_n-Schätzer zeigt sich in Abbildung 3.17.

Der Median der Variablen ist extrem hoch im Vergleich zu den anderen Variablen im Datenset, worauf die klassische Methode überaus sensibel reagiert. Die Ladungen werden in diesem Kapitel nur für das erste Datenset und die 1. Einschränkung dargestellt, da weitere Abbildungen keinen Mehrwert liefern würden. In Abbildung 3.18 kann man nicht so eindeutige Tendenzen erkennen wie in Abbildung 3.9.

Anders in Abbildung 3.19, hier ist sehr gut sichtbar, dass die klassische Methode auf wenige robust skalierte Variablen sehr sensibel reagiert.

Auf die Darstellung der Ladungen anhand der Hauptkomponenten über die Datensets wurde auch verzichtet, da sich hier keine Auffälligkeiten oder Abweichungen gezeigt haben.



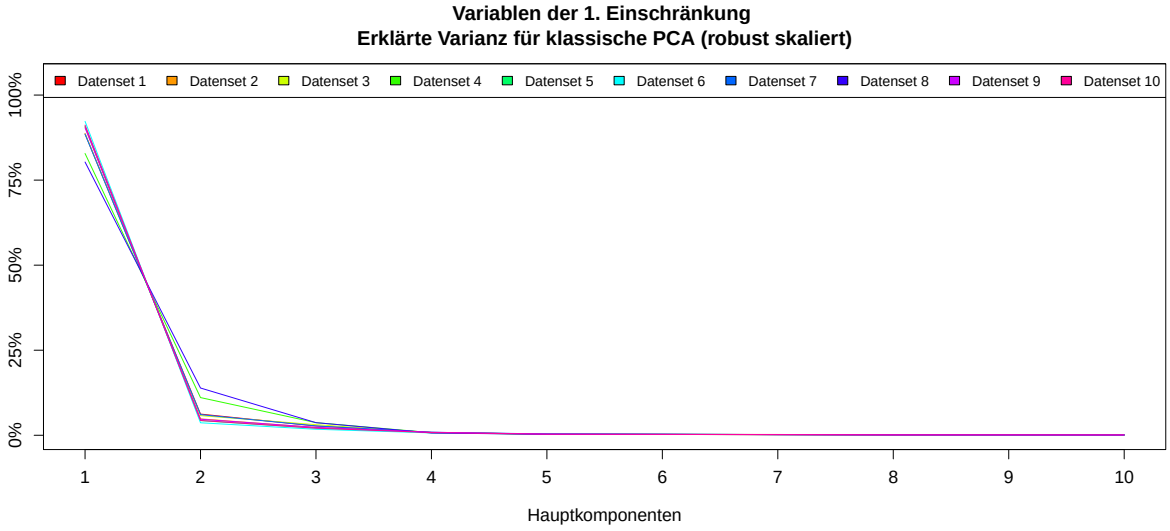


Abbildung 3.16: Darstellung der erklärten Varianz für die klassische Hauptkomponentenanalyse mit robuster Skalierung und für die robuste Hauptkomponentenanalye mit klassischer Skalierung für die 1. Einschränkung

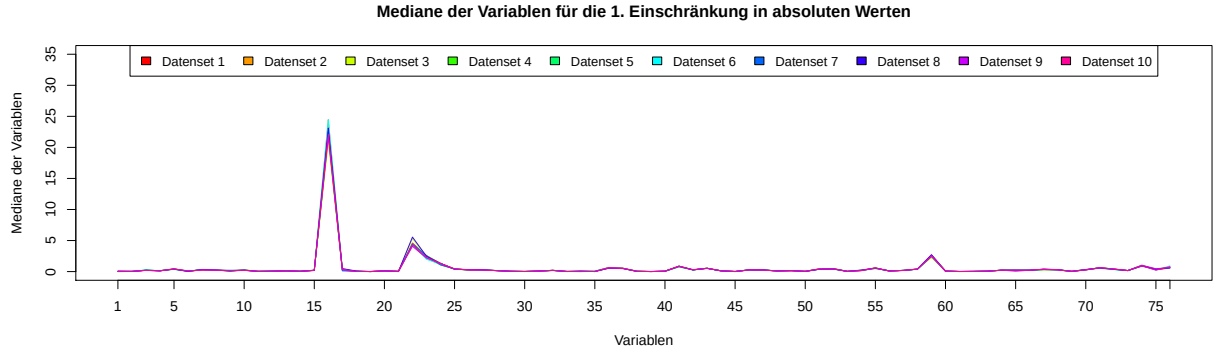


Abbildung 3.17: Darstellung der absoluten spaltenweisen Mediane der robust zentrierten Matrix des Datensets 1 für die 1. Einschränkung

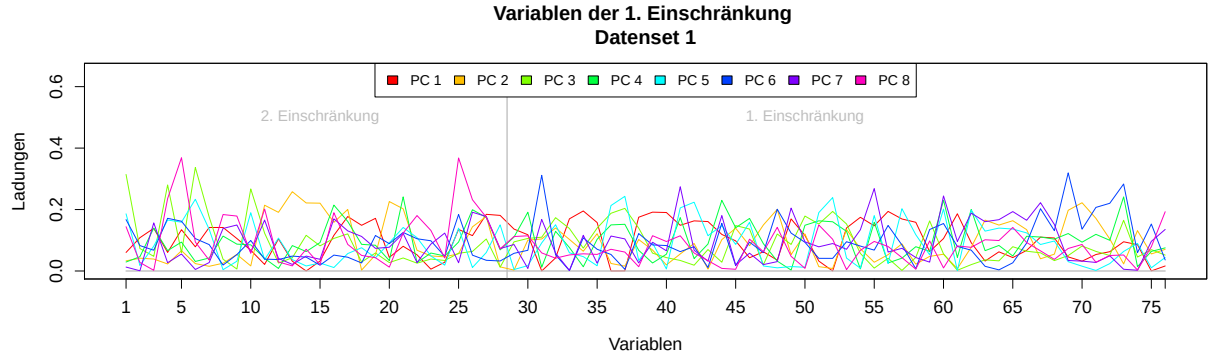


Abbildung 3.18: Darstellung der Ladungen der 1. Einschränkung für klassische Skalierung mit robuster Methode anhand der Daten

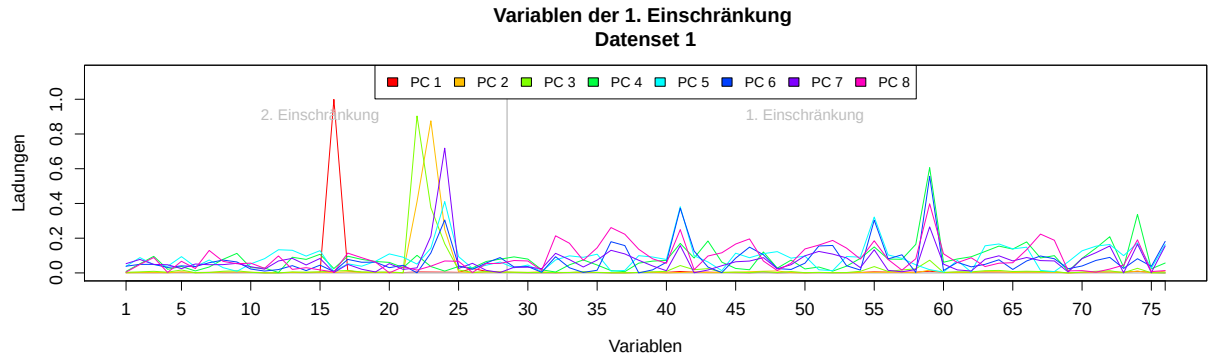


Abbildung 3.19: Darstellung der Ladungen der 1. Einschränkung für robuste Skalierung mit klassischer Methode anhand der Daten

3.6 Gegenüberstellung der Methoden

In den vorherigen Abschnitten wurden bereits Vergleiche zwischen den Methoden gezogen, welche nun intensiver behandelt und zusammengefasst werden. Abbildung 3.20 zeigt die berechneten Kombinationen bezüglich der Skalierung und der Methoden. Wenn im Folgenden von klassisch, robust und robust and sparse Methode gesprochen wird, stimmen die Skalierungsmethode und die Berechnungsmethode bezüglich Robustheit überein. Andernfalls wird dies gesondert erwähnt.

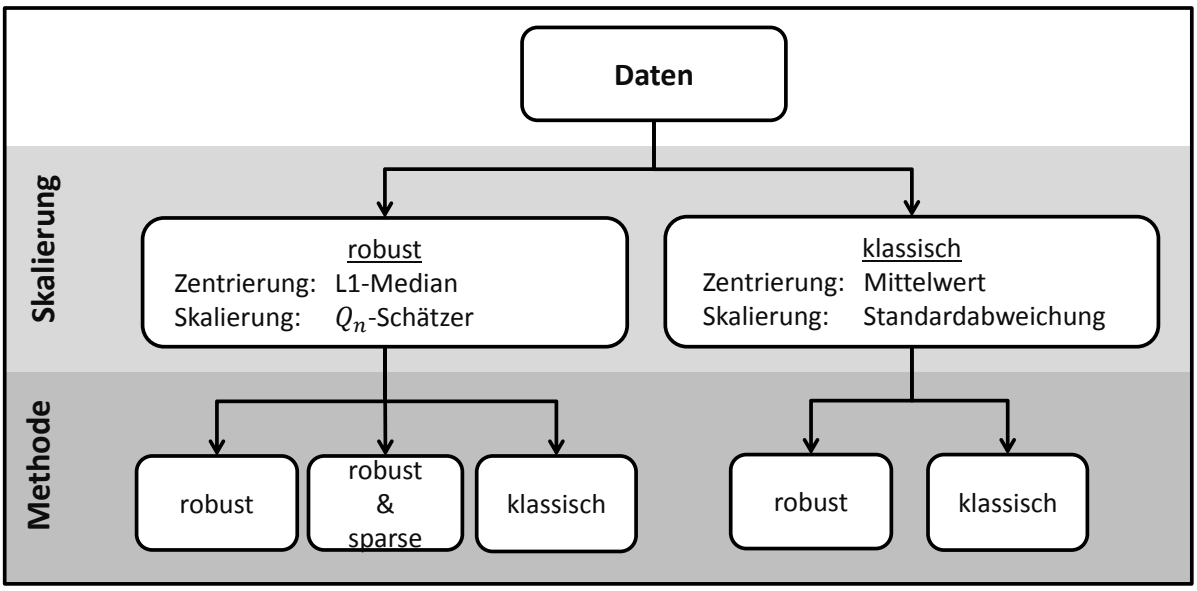


Abbildung 3.20: Diagramm der Skalierungen und Berechnungsmethoden

Als Resümee aus den obigen Abschnitten kann gezogen werden, dass es von der klassischen zur robusten Methode bereits möglich wird, die Variablen mit großem Erklärungswert herauszufiltern, und in der robust and sparse Methode die relevanten Variablen klar erkennbar sind. Durch das Ausdünnen der Ladungsmatrix in der robust and sparse Methode leidet jedoch die erklärte Varianz, welche geringer ist als bei den Methoden robust und klassisch.

Aufgrund der hohen Anzahl an Daten wurden anfangs 10 Datensets mit geringer Anzahl an Daten für die Auswertung gezogen. Da der Aufwand, diese Datensets einzeln auszuwerten wenig zielführend und sinnvoll gewesen wäre, wurden die Ergebnisse mit Hilfe der Ladungsmatrix und der erklärten Varianz zusammengeführt. Von Interesse sind nach wie vor die Variablen, die die Daten am besten erklären, welche sich durch einen hohen Ladungswert gepaart mit der erklärten Varianz der Hauptkomponenten für den Ladungswert auszeichnen. Darum wurden die Spalten der Hauptkomponenten der Ladungsmatrizen mit deren erklärter Varianz gewichtet und anschließend die Werte für jede Variable addiert. Diese Überlegung scheint besonders sinnvoll für die robust and sparse Methode, bei welcher die Ladungsmatrix sehr dünn besetzt ist und das Finden der signifikanten Variablen

somit erleichtert wird. Eine Addition über alle 10 Datensets ergab eine Reihung für die Variablen.

In Abbildung 3.21 sind die absteigend sortierten Variablen für die Basisauswahl zu sehen, wobei anzumerken ist, dass die beiden Grafiken nicht dieselbe Ordnung haben. Der unterschiedliche Kurvenverlauf sticht hier sofort ins Auge. Die Kurve zur klassischen Methode verläuft sehr flach und fällt erst sehr spät bei Variable 85, im Gegensatz zur robusten Kurve, die zwischen den Variablen 1 und 15 stark fällt und dann anähernd stagniert. Zieht man hier Schlüsse über die Auswahl der Anzahl der Variablen, stellt sich diese Aufgabe bei der klassischen Methode als schwierig heraus. Für die robust and sparse Methode fällt die Kurve noch steiler ab als bei der robusten Methode und verweilt ab Variable 15 nahe bei null. Abbildung A.1 im Anhang zeigt dieses Verhalten.

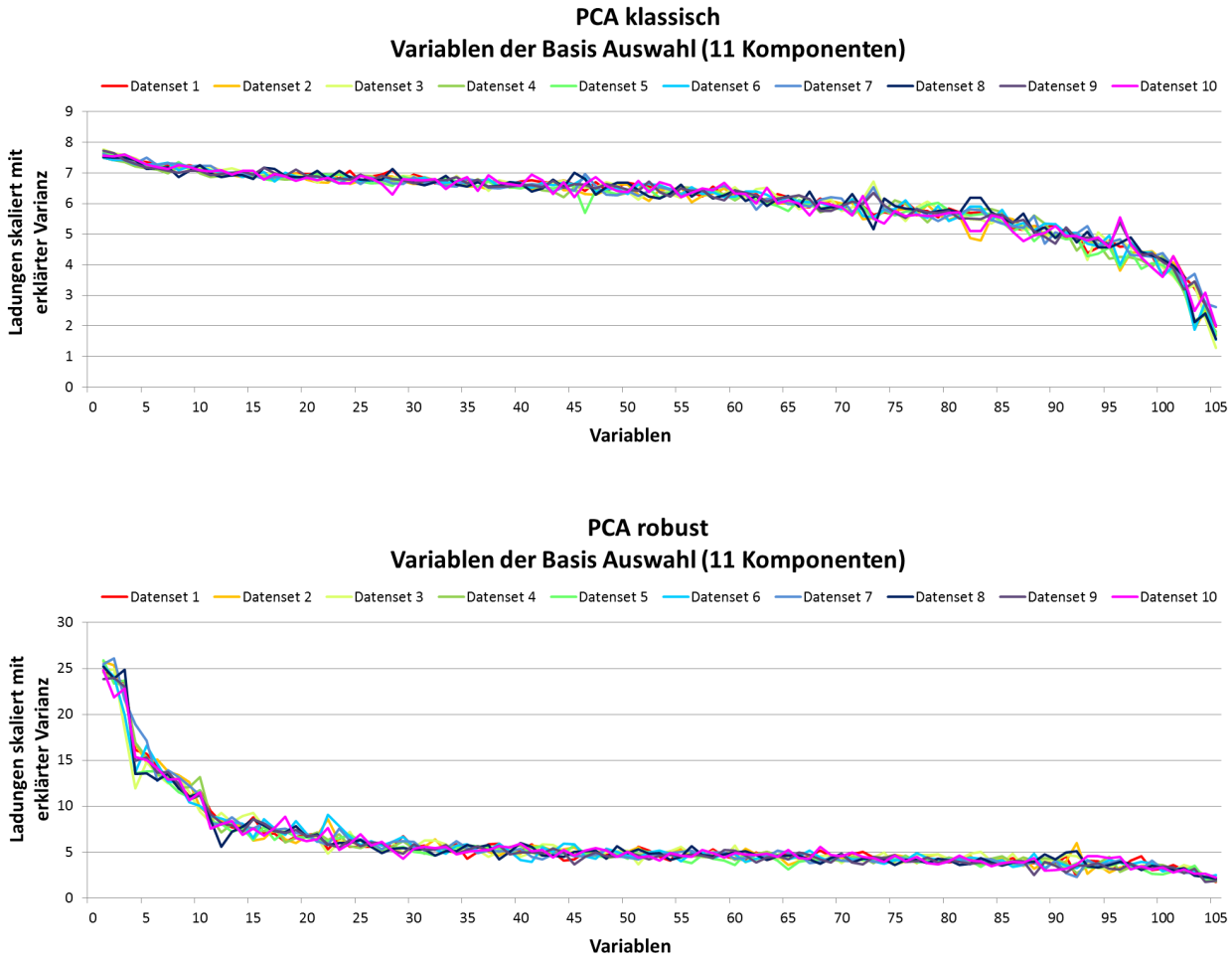


Abbildung 3.21: Absteigende Reihung der Variablen

Die nächsten Abbildungen werden zeigen, welche Merkmalsgruppen die verschiedenen

Methoden und Skalierungskombinationen als relevant erachtet haben. Die ersten beiden Spalten in Abbildung 3.22 haben eine identische Verteilung der Variablen, was auch plausibel erscheint, da die robust and sparse Methode mit einer zusätzlichen Restriktion im Vergleich zur robusten Methode arbeitet. Es sind hier nicht nur die Merkmalsgruppen identisch, sondern auch die Variablen. Nur bei der Ordnung gibt es Unterschiede. Eine Spalte weiter rechts ergibt die klassische Methode mit robuster Skalierung ein ähnliches Bild. Die Gruppen *Drehmoment Niveau* und *AGN Dosierung/Verbrauch/Konversion* haben jeweils eine Variable weniger und die Gruppe *Drehmoment Niveau* eine Variable mehr zugeteilt bekommen. Aus der Merkmalsgruppe *Abgasmassenstrom* wurde auch eine Variable gewählt. Die Methoden robust und klassisch unterscheiden sich nur in 2 Variablen, haben daher ein sehr ähnliches Verhalten. Besonders einflussreiche Gruppen sind hier die *Motorstopps Anzahl* und die *AGN Dosierung/Verbrauch/Konversion*.

Betrachtet man die Spalte der klassischen Skalierung ergibt sich bereits ein anderes Bild. Während die Berechnungen mit der robusten Skalierung durchwegs Merkmalsgruppen auswählte, die Niveaus, Anzahl und Dosierungen beschrieben, wählte die klassische Methode mit klassischer Skalierung fast ausschließlich Gruppen, die Dynamiken beschrieben, aus. Hier sind vor allem die Gruppen *Fahrzeuggeschwindigkeit Dynamik* und *Leistung Dynamik* stark vertreten. Die robuste Methode mit klassischer Skalierung wählte Gruppen, die Temperaturen und Niveaus beschrieben und zeigte eine geringe Schnittmenge mit den Methoden der robusten Skalierung. Innerhalb der klassischen Skalierung gibt es bis auf eine Variable in der Gruppe der AGN Größen keine Überschneidung. Diese Abbildung zeigt, dass in der Basisauswahl die Art der Skalierung der Daten bereits eine große Rolle spielt. Innerhalb der robusten Skalierung zeigten sich die Ergebnisse, die Gruppen und Variablen betrafen, relativ stabil, hingegen liefern die Methoden innerhalb der klassischen Skalierung völlig unterschiedliche Ergebnisse.

Die 1. Einschränkung enthält keine *AGN Größen*, daher ist diese Gruppe in Abbildung 3.23 nicht aufgeführt. In den Spalten der robusten Skalierung gibt es nur wenige Änderungen zu den Ergebnissen der Basisauswahl. Innerhalb der robust and sparse Methode hat die Merkmalsgruppe *Motorstopps Anzahl* eine Variable verloren und *Leistung Niveau* eine dazugewonnen. Die Merkmalsgruppe *Leistung Niveau* ist in der klassischen Methode nicht mehr vertreten. Innerhalb der klassischen Skalierung bei der robusten Methode gab es mehr Bewegung. Hier wurden erstmals Variablen in den Gruppen *Leistung Dynamik*, *Drehzahl Niveau* und *Abgasmassenstrom* ausgewählt. Die klassische Methode bleibt stark vertreten bei den Dynamiken und wie zuvor wieder bei den Gruppen *Fahrzeuggeschwindigkeit Dynamik* und *Leistung Dynamik*. Die Ergebnisse der Methoden innerhalb der robusten Skalierung bleiben noch vergleichbar.

Die 2. Einschränkung besteht nur noch aus fahrzeugbezogenen und Umgebungsgrößen und enthält daher nur noch eine kleine Teilmenge aller definierten Gruppen. Innerhalb der robusten Skalierung gibt es Überlappungen bei den Gruppen *Fahrzeugstopps Niveau/Anzahl* und *Leistung Niveau*. Innerhalb der klassischen Skalierung ist die Merkmalsgruppe *Leistung Dynamik* mit drei Variablen bei der robusten Methode stark vertreten. Die Merkmalsgruppe *Fahrzeuggeschwindigkeit Dynamik* ist bei der klassischen Methode mit der Anzahl drei die Gruppe mit den meisten Variablen.

Basisauswahl					
Skalierung	**robust**			**nicht robust**	
Methode	**robust & sparse**	**robust**	**nicht robust**	**robust**	**nicht robust**
Fahrzeuggeschwindigkeit Niveau					
Fahrzeuggeschwindigkeit Dynamik					4
Fahrzeugstopps Niveau/Anzahl	1	1	1		
Leistung Niveau	1	1	1	1	
Leistung Dynamik					3
Drehzahl Niveau	1	1	2		1
Drehzahl Dynamik					
Motorstopps Anzahl	3	3	3		
Motortemperaturen				3	
Drehmoment Niveau	2	2	1	2	
Drehmoment Dynamik					2
Umgebung					
Abgasmassenstrom			1		
AGN Temperaturen				3	
AGN Dosierungen/Verbrauch/Konversion	3	3	2	2	1

Abbildung 3.22: Verteilung der ersten 11 Variablen auf die Merkmalsgruppen

Fasst man die letzten Erkenntnisse zusammen, stellt man fest, dass die verschiedenen Methoden unterschiedliche Ergebnisse liefern. Diese Unterschiede scheinen ihren Ursprung nicht nur in den Methoden, sondern auch in der Skalierung der Datenmatrix zu haben. Um diese These empirisch zu überprüfen, wurden die besten 11 Variablen der robusten und klassischen Methode skaliert und in den Abbildungen 3.25 und 3.26 dargestellt. Die Abbildung 3.25 zeigt Boxplots für die besten 11 Variablen skaliert mit der Standardabweichung und dem Mittelwert. Die rot und blau gefärbten Werte in der Abbildung geben den Prozentsatz der Ausreißer an, die nicht mehr innerhalb der Grafik liegen. Die ersten 11 ausgewählten Variablen gehören zur klassischen Methode und die letzten 11 Variablen sind jene für die robust und robust and sparse Methode. Stellvertretend für die anderen Datensets wurde hier das Basisdatenset ausgewählt. Die ausgewählten Variablen für die klassische Methode unterscheiden sich von den anderen Variablen optisch dadurch, dass der Interquartilbereich größer ist, was auf eine größere Variabilität in den Daten schließen lässt.

In Abbildung 3.26 ist das Datenset 1 mit robuster Skalierung dargestellt. Hier zeigt sich ein gegenteiliges Verhalten: die Variablen, die für die robusten Methoden ausgewählt wurden, haben hier bis auf wenige Ausnahmen einen hohen Interquartilbereich. Die Variable *TorquePosQ99* sticht hier mit einer sehr großen Box hervor. Diese Variable wurde in der robust and sparse Methode auf den ersten Platz gereiht. Betrachtet man die Varia-

ble *EngineSpeedQ05* wird nun auch deutlich, warum die klassische Methode mit robuster Skalierung diese Variable als äußerst wichtig erachtet. Diese Variable hat in der robusten Skalierung viele Ausreißer nach oben, worauf die klassische Methode sensibel reagiert.

1. Einschränkung					
Skalierung	**robust**			**nicht robust**	
Methode	**robust & sparse**	**robust**	**nicht robust**	**robust**	**nicht robust**
Fahrzeuggeschwindigkeit Niveau					1
Fahrzeuggeschwindigkeit Dynamik					4
Fahrzeugstopps Niveau/Anzahl	1	1	1		
Leistung Niveau	2	1		2	
Leistung Dynamik				1	3
Drehzahl Niveau	1	1	2	2	
Drehzahl Dynamik					
Motorstopps Anzahl	2	3	3		
Motortemperaturen				1	
Drehmoment Niveau	2	2	1	1	
Drehmoment Dynamik					
Umgebung					
Abgasmassenstrom			1	1	

Abbildung 3.23: Verteilung der ersten 8 Variablen auf die Merkmalsgruppen

2. Einschränkung					
Skalierung	**robust**			**nicht robust**	
Methode	**robust & sparse**	**robust**	**nicht robust**	**robust**	**nicht robust**
Fahrzeuggeschwindigkeit Niveau			1		
Fahrzeuggeschwindigkeit Dynamik	1				3
Fahrzeugstopps Niveau/Anzahl	1	2	2		
Leistung Niveau	2	1	1	1	
Leistung Dynamik		1		3	1
Umgebung					

Abbildung 3.24: Verteilung der ersten 4 Variablen auf die Merkmalsgruppen

Dieser deutliche Unterschied liegt bei den verwendeten Skalierungsschätzern Standardabweichung und Q_n-Schätzer und deren Eigenschaften. Für die Variable *EngineSpeedQ05* ergibt die Standardabweichung *sd* einen Wert von 145,52 1/min und der Q_n-Schätzer einen

viel geringeren Wert von 1,66 1/min. Da die Standardabweichung durch ihre Sensitivität gegenüber wenigen großen Werten diese nun zu kleinen Werten skaliert, ergibt sich, dass diese Variable in der klassischen Skalierung unbedeutend wird. Im Gegensatz zum Q_n-Schätzer, welcher die Werte der Variablen *EngineSpeedQ05* mit einem Wert von 1,66 nur im geringen Ausmaß skaliert. Abbildung 3.27 zeigt, dass die Standardabweichung für die Skalierung hier ungeeignet ist, da sie das Verhalten der *Fernverkehr hohe Beladung* Fahrzeuge wegskaliert. Ganz anders liegt der Sachverhalt bei *EngineSpeedStd*, diese Variable zeigt keine extremen Unterschiede im Verhalten zwischen den Fahrzeugen, daher liegen hier die Standardabweichung mit 63,62 1/min und der Q_n-Schätzer mit 63,2 1/min nah beieinander. Die Effekte, die sich dadurch ergeben, sieht man in den Ergebnissen. Die klassische Methode ignoriert die Variable *EngineSpeedQ05*, die die *Fernverkehr hohe Beladung* Fahrzeuge von den anderen Betriebsarten trennt und wählt hingegen die Variable *EngineSpeedStd*, die zwar Variabilität zeigt jedoch keine derart klare Trennung.

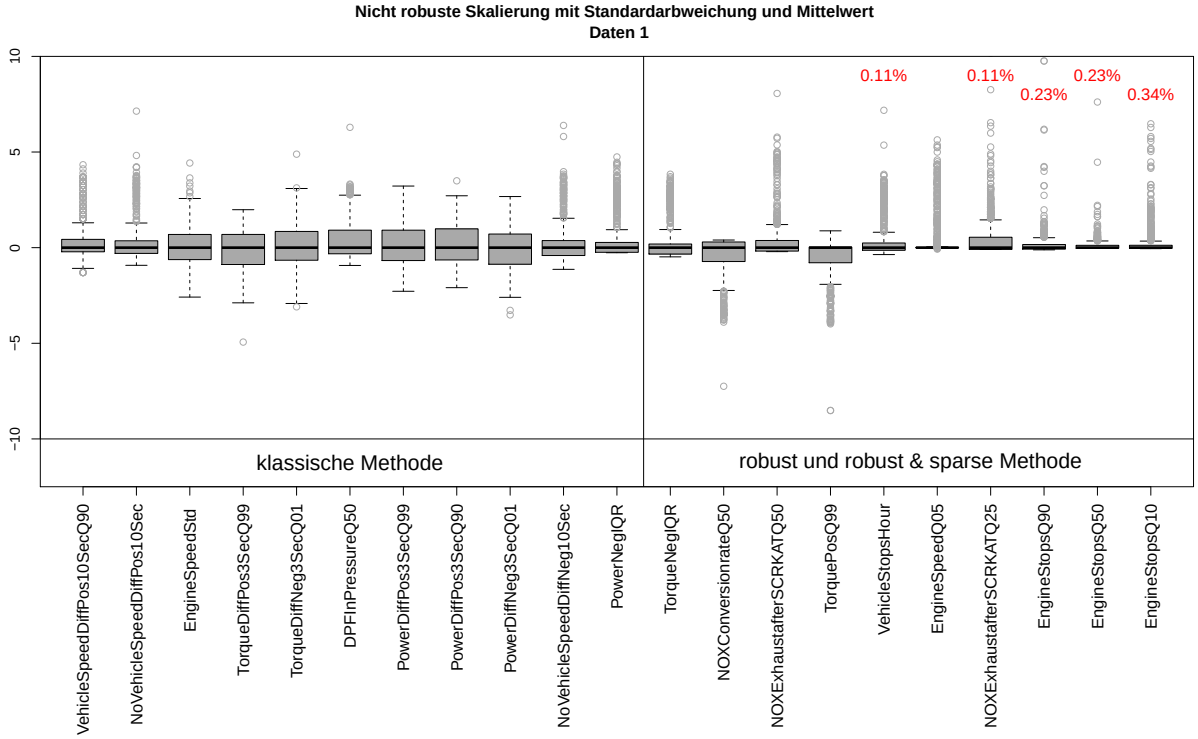


Abbildung 3.25: Für Basismatrix Ausgewählte Variablen skaliert mit Standardabweichung und Mittelwert

Will man nun bereits vor der Hauptkomponentenanalyse wissen, welche Variablen sich im Ergebnis bei der klassischen und der robusten Methode anders verhalten könnten, kann man einen Quotienten aus den Q_n Schätzern und der Standardabweichung bilden. Alle Variablen, die stark von Eins abweichen, sind Kandidaten für ein differentes Verhalten. Abschließend lässt sich also sagen, dass die Art der Skalierung bereits einen großen Einfluss

hat.

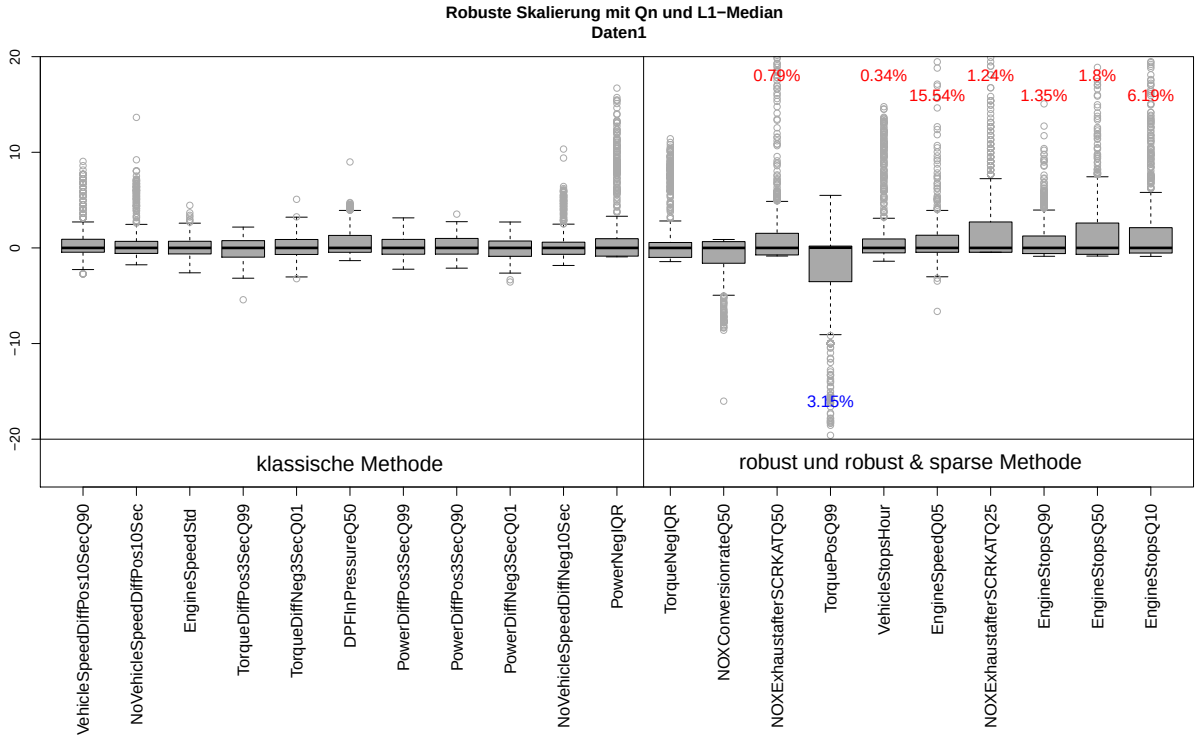


Abbildung 3.26: Für Basismatrix Ausgewählte Variablen skaliert mit Q_n und L1-Median

Die robuste Methodik ergab einen weiteren Vorteil, der in Abbildung 3.28 gezeigt wird. In Abbildung 3.28 sind die Korrelationen der 11 ausgewählten Variablen für die robuste Methode dargestellt, gefärbt nach dem Grad ihrer Abhängigkeit. Zwischen den Variablen herrschen kaum lineare Zusammenhänge, ausgenommen *TorqueNegIQR* und *PowerNegIQR*, die mit $r = 0,9778$ stark korrelieren. Ansonsten gibt es nur noch innerhalb der AGN-Größen Gruppe und den Fahrzeugstopps nennenswerte Korrelationen. Dieses Ergebnis zeigt, dass die Variablen untereinander nicht redundant sind und jede Variable zusätzliche Informationen liefert.

Ein ganz anderes Bild zeigt sich in Abbildung 3.29: Hier korrelieren die ersten 4 Variablen sehr stark miteinander, sind jedoch fast unkorreliert zu den restlichen 7 Variablen. Die restlichen 7 Variablen korrelieren schwach bis sehr stark. Die Variable *EngineSpeedStd* ist die einzige, die nur schwach mit den weiteren Variablen korreliert. Dieser starke Zusammenhang zwischen den Variablen bringt Redundanz mit sich, woraus sich ergibt, dass auf Variablen verzichtet werden könnte, da diese kaum zusätzliche Information liefern. Das Verhalten, dass die Variablen der robusten Methode kaum korrelieren und die der klassischen aber stark, konnte in bei der 1. Einschränkung und der 2. Einschränkung auch beobachtet werden.

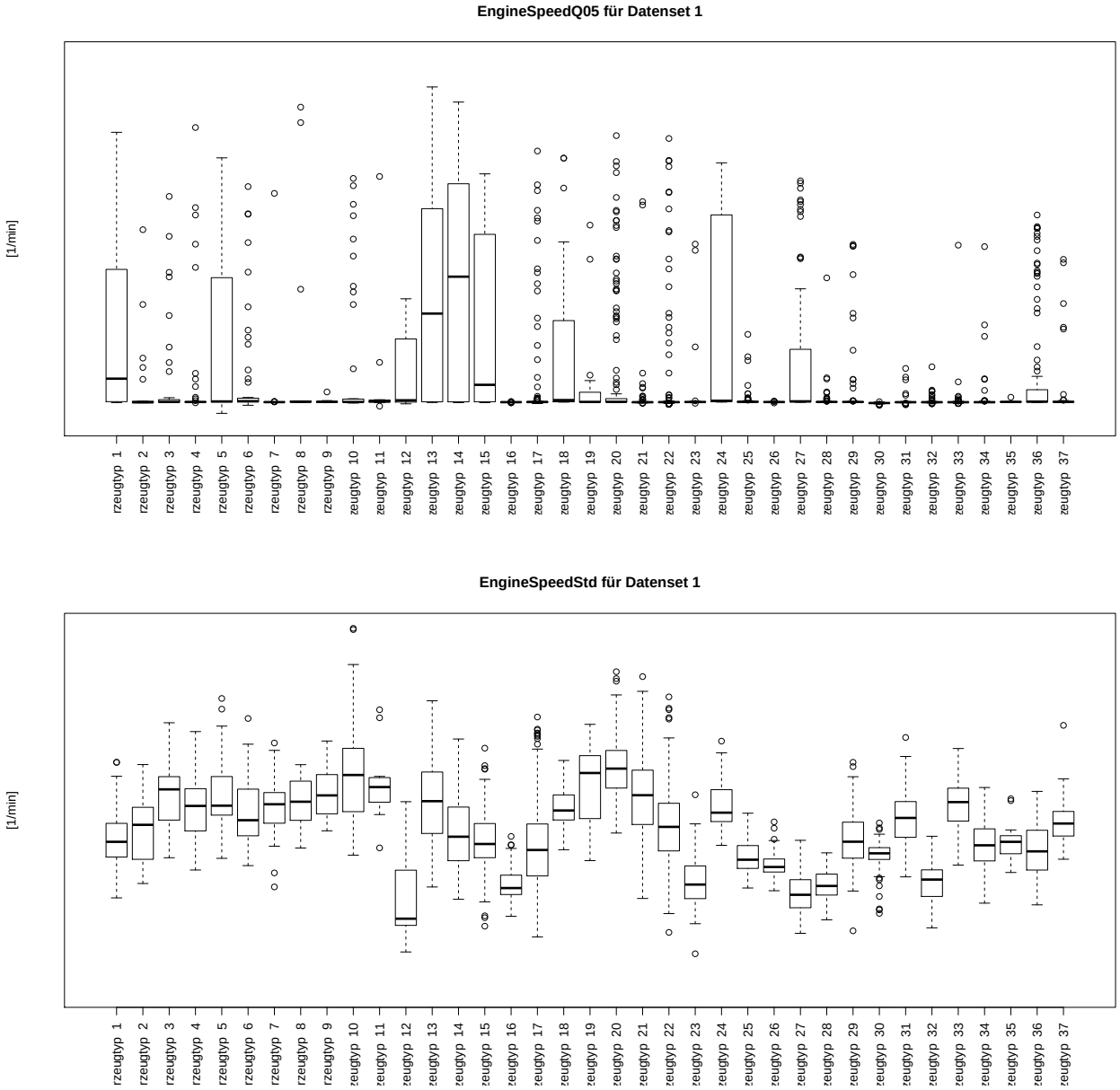


Abbildung 3.27: Darstellung der Variablen EngineSpeedQ05 und EngineSpeedStd getrennt nach Fahrzeugtypen

Ausgewählte Variablen der robusten Methode	EngineStopsQ10	EngineSpeedQ05	TorquePosQ99	EngineStopsQ50	NOXExhaustafterSCRKATQ25	VehicleStopsHour	PowerNegIQR	NOXExhaustafterSCRKATQ50	TorqueNegIQR	EngineStopsQ90	NOXConversionrateQ50
EngineStopsQ10	1,0000	0,0208	0,0583	0,3906	0,0263	-0,0480	0,1211	0,0382	0,1255	0,1526	-0,0275
EngineSpeedQ05	0,0208	1,0000	0,1990	-0,0011	0,1568	-0,2147	0,2796	0,1161	0,2731	-0,0562	-0,1048
TorquePosQ99	0,0583	0,1990	1,0000	-0,0212	0,0234	-0,5408	0,2788	0,0213	0,3031	-0,0229	-0,1201
EngineStopsQ50	0,3906	-0,0011	-0,0212	1,0000	0,0073	-0,0219	-0,0026	0,0697	0,0245	0,7285	-0,0644
NOXExhaustafterSCRKATQ25	0,0263	0,1568	0,0234	0,0073	1,0000	-0,1273	0,1459	0,7097	0,1522	-0,0032	-0,6129
VehicleStopsHour	-0,0480	-0,2147	-0,5408	-0,0219	-0,1273	1,0000	-0,2133	-0,1304	-0,2316	0,0334	0,1538
PowerNegIQR	0,1211	0,2796	0,2788	-0,0026	0,1459	-0,2133	1,0000	0,1522	0,9778	-0,1013	-0,2202
NOXExhaustafterSCRKATQ50	0,0382	0,1161	0,0213	0,0697	0,7097	-0,1304	0,1522	1,0000	0,1538	0,0328	-0,8101
TorqueNegIQR	0,1255	0,2731	0,3031	0,0245	0,1522	-0,2316	0,9778	0,1538	1,0000	-0,0783	-0,2215
EngineStopsQ90	0,1526	-0,0562	-0,0229	0,7285	-0,0032	0,0334	-0,1013	0,0328	-0,0783	1,0000	-0,0668
NOXConversionrateQ50	-0,0275	-0,1048	-0,1201	-0,0644	-0,6129	0,1538	-0,2202	-0,8101	-0,2215	-0,0668	1,0000

Abbildung 3.28: Korrelationen der 11 Variablen für die robuste Methode

Um die leichtere Interpretation der Ladungen einer robust and sparse Ladungsmatrix zu zeigen, wurde der prozentuelle Anteil der Nullladungen berechnet. Eine Ladung wird als null angenommen, wenn sie kleiner als 10^{-16} ist. Die robust and sparse Methode erreichte somit im Median über die Datensets bei der Basisauswahl einen Anteil an Nullladungen von 68,83%, die 1. Einschränkung 76,07% und die 2. Einschränkung 7,14% innerhalb der betrachteten Hauptkomponenten. Im Vergleich dazu erreichte die robuste Methode bei der Basisauswahl 0,35%, bei der 1. Einschränkung 0,41% und bei der 2. Einschränkung 0,45%. Die klassische Methode hatte durchwegs einen Median von 0%. Die Spannweite bei der robust and sparse Methode reichte bei den Basisdatensets von 45,45% für das Datenset 4 mit einem Lambda von 16 und 80,87% für das Datenset 1 mit einem Lambda von 34. Hier zeigt sich auch empirisch, dass die Restriktion die Ladungen gegen Null drückt und dass ein hoher Tuningparameter eine Ladungsmatrix produziert, die sehr dünn besetzt ist.

Auch wenn die Interpretation anhand der Ladungen mit der robust and sparse Methode einfacher ist, so ist die optische Interpretation dieser Methode oftmals schwierig, wie man in Abbildung A.3 im Anhang sehen kann. Hier wurde die erste Hauptkomponente gegen die 2. Hauptkomponente aufgetragen und nach der Betriebsart eingefärbt. Dabei kann man

eine gute Trennung zwischen *Fernverkehr hohe Beladung* und den weiteren Betriebsarten erkennen, jedoch ist eine Trennung des *Kundenbetriebs*, *Stadt Verteilerbetriebs* und *Verteilerbetriebs* kaum möglich. Die Erklärung dafür ist, dass hier bei diesen Hauptkomponenten jeweils eine Variable einen hohen Ladungswert hat, was im Sinne der Methode ist, und diese Variablen in ihrem Verhalten den *Fernverkehr hohe Beladung* von den restlichen Zyklen trennen. Es ist hier daher sinnvoll, Kombinationen von Hauptkomponenten zu betrachten, welche unterschiedliche Betriebsarten trennen.

Um eine Reihung der besten 11 Variablen aus dem Basisdatenset zu bekommen, wurde nochmals eine robust and sparse Hautkomponentenanalyse gerechnet. Dabei zeigte sich bei der Auswahl des Parameter Lambda ein abweichendes Verhalten zu den bisher gemachten Berechnungen. Abbildung 3.30 zeigt, dass die erklärte Varianz mit immer höher werdenden Lambda ansteigt und erst sehr spät nach dem Lambdawert 70 fällt. Dies zeigt, dass die Höhe des Tuningparameters nicht durch die Anzahl der Variablen vorgegeben ist.

Ausgewählte Variablen der nicht robusten Methode	SPEEDDiffPos10SecQ90	SPEEDDiffPos10SecQ75	NoVehicleSpeedDiffPos10Sec	NoVehicleSpeedDiffNeg10Sec	TorqueDiffPos3SecQ99	PowerDiffPos3SecQ90	PowerDiffNeg3SecQ01	EngineSpeedStd	DPFInPressureQ50	PowerDiffPos3SecQ99	TorqueDiffNeg3SecQ01
SPEEDDiffPos10SecQ90	1,0000	0,9554	0,9352	0,9206	-0,3347	-0,2054	0,3271	-0,0977	-0,4662	-0,3201	0,2351
SPEEDDiffPos10SecQ75	0,9554	1,0000	0,9463	0,9435	-0,3101	-0,1368	0,2930	-0,0381	-0,3848	-0,2717	0,2117
NoVehicleSpeedDiffPos10Sec	0,9352	0,9463	1,0000	0,9876	-0,2492	-0,0576	0,2327	-0,0212	-0,3695	-0,2110	0,1647
NoVehicleSpeedDiffNeg10Sec	0,9206	0,9435	0,9876	1,0000	-0,2060	-0,0118	0,1843	0,0044	-0,3095	-0,1640	0,1192
TorqueDiffPos3SecQ99	-0,3347	-0,3101	-0,2492	-0,2060	1,0000	0,7917	-0,9155	0,3618	0,6697	0,9218	-0,9290
PowerDiffPos3SecQ90	-0,2054	-0,1368	-0,0576	-0,0118	0,7917	1,0000	-0,8390	0,4355	0,6952	0,8792	-0,7564
PowerDiffNeg3SecQ01	0,3271	0,2930	0,2327	0,1843	-0,9155	-0,8390	1,0000	-0,4723	-0,7406	-0,9481	0,9186
EngineSpeedStd	-0,0977	-0,0381	-0,0212	0,0044	0,3618	0,4355	-0,4723	1,0000	0,4156	0,5151	-0,3103
DPFInPressureQ50	-0,4662	-0,3848	-0,3695	-0,3095	0,6697	0,6952	-0,7406	0,4156	1,0000	0,7759	-0,6200
PowerDiffPos3SecQ99	-0,3201	-0,2717	-0,2110	-0,1640	0,9218	0,8792	-0,9481	0,5151	0,7759	1,0000	-0,8384
TorqueDiffNeg3SecQ01	0,2351	0,2117	0,1647	0,1192	-0,9290	-0,7564	0,9186	-0,3103	-0,6200	-0,8384	1,0000

Abbildung 3.29: Korrelationen der 11 Variablen für die klassische Methode

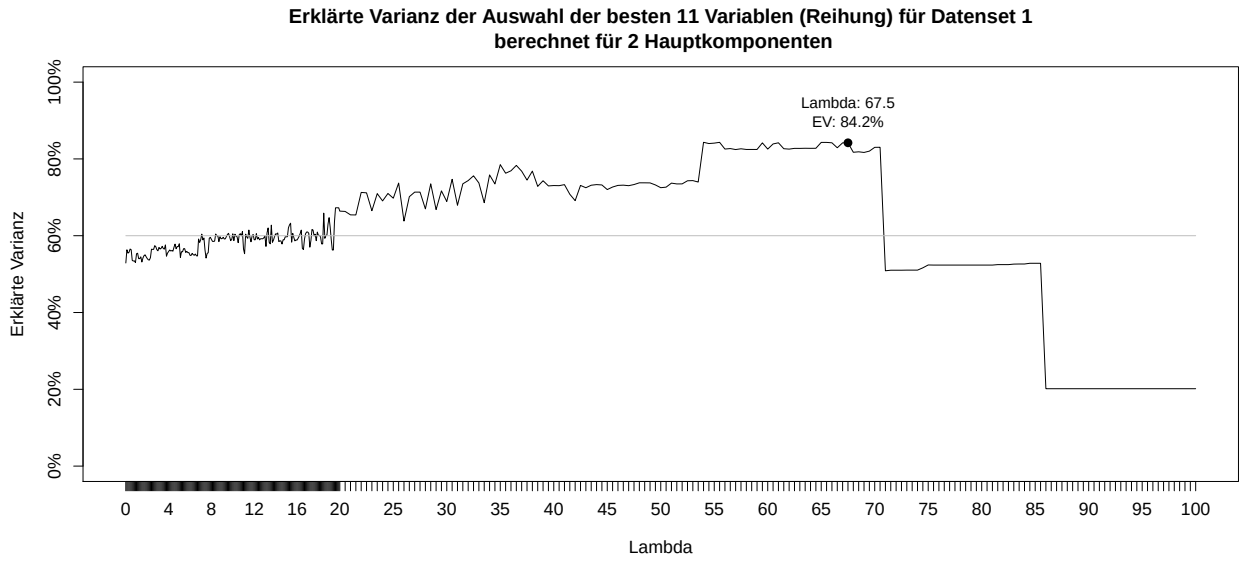


Abbildung 3.30: Darstellung der erklärten Varianz für den Parameter Lambda

Auf die Laufzeit wurde in Abschnitt 3.4 bereits kurz eingegangen. Um den Unterschied zwischen den Methoden deutlich zu machen, geben wir hier in Tabelle 3.8 die realen Laufzeiten an.

Laufzeiten			
Methode	**Basisauswahl**	**1. Einschränkung**	**2. Einschränkung**
klassisch	5,58 Sekunden	3,01 Sekunden	0,47 Sekunden
robust	16,95 Minuten	9,25 Minuten	1,47 Minuten
robust & sparse (10 Parameter)	169,50 Minuten	92,50 Minuten	14,70 Minuten
robust & sparse (50 Parameter)	14,13 Stunden	7,71 Stunden	1,23 Stunden
robust & sparse (89 Parameter)	25,14 Stunden	13,72 Stunden	2,18 Stunden

Tabelle 3.8: Laufzeiten der drei Hauptkomponenten Methoden

Die Berechnungszeit der klassischen Methode benötigt nur ca. 0,5% von der Zeit der robusten Methode. Nimmt man für die robust and sparse Methode die 89 Berechnungpunkte von zuvor, so ergibt dies für das Basisdatenset mit robust and sparse Variante eine Laufzeit von 25,14 Stunden und für eine classic and sparse Variante eine Laufzeit von 8,28 Minuten. Somit dauert die robust and sparse Variante mehr als 180 mal so lange. An der längeren Laufzeit ist der Q_n-Schätzer maßgeblich beteiligt, der ca. die 27-fache Zeit der Berechnung der Standardabweichung benötigt.

Abschließend lässt sich zusammenfassen, dass die robuste Methode trotz der längeren Laufzeit gegenüber der klassischen Methode wesentliche Vorteile mit sich bringt. Dies wäre zum einen, dass die Variablen mit hohem Erklärungswert kaum korrelieren und die Auswahl der signifikanten Ladungen optisch bereits einfacher fällt. Zum anderen ergibt sich

ein wesentlicher Vorteil daraus, dass sie weniger sensitiv auf Ausreißer reagiert, was bei der klassischen Methode mit robuster Skalierung in Abschnitt 3.5 anschaulich gezeigt wurde. Das letzte Argument ist auch der Grund, warum von der Methode robuste Skalierung mit klassischer Hauptkomponentenanalyse abzusehen ist. Da diese Methode sagt, dass die Variable *EngineSpeedQ05* die Basisauswahl ausreichend erklärt und sich herausstellte, dass *EngineSpeedQ05* nur aufgrund zahlreicher Ausreißer diesen großen Einfluss hat. Auch Abbildung A.2 im Anhang, die die Ladungen mit gewichteter Varianz darstellt, unterstreicht die zuvor geäußerten Behauptungen. In Kapitel 4 wird sich zeigen, dass auch andere Variablen einen Erklärungswert haben.

Dass die Skalierung einen wesentlichen Einfluss auf das Ergebnis hat, wurde auch hinreichend gezeigt. Es ist daher von einer klassischen Skalierung der Datensets abzusehen nicht nur aufgrund der Argumente in Abschnitt 3.5, sondern auch da diese für die klassische und robuste Methode keine konsistenten Outputs lieferte, anders als bei den Berechnungen, denen die robust skalierten Daten zugrunde lagen.

Weil in jedem Berechnungsschritt auf Robustheit geachtet wurde, angefangen bei der Definition der Variablen mit Quantilen und Interquartilbereich, über Skalierung bis hin zur Hauptkomponentenanalyse, kann man davon ausgehen, dass die Ergebnisse durch wenige Ausreißer nicht beeinflusst sind.

Die Ergebnisinterpretation wird durch die robust and sparse Methode deutlich erleichtert, wie man bereits anhand der Abbildung für die Ladungen in den vorangegangenen Abschnitten sehen konnte. Einen Nachteil hat die robust and sparse Methode jedoch in der Darstellung der Ergebnisse, die optisch schwer interpretierbar sind, was in Abbildung A.3 im Anhang gezeigt wird. Die lange Laufzeit kann sich vor allem bei großen Datensets zu einem Nachteil entwickeln. Jedoch ist sie eine gute Möglichkeit auf effiziente Weise relevante Variable zu erhalten.

Zusammenfassend lässt sich nun sagen, dass es empfehlenswert ist, in der Praxis die Daten robust zu skalieren und auch robust zu berechnen, denn obwohl die robuste Methode gegenüber der klassischen eine längere Laufzeit aufweist, überwiegen die zuvor diskutierten Vorteile dieser Methode ganz klar. Die robust and sparse Methode stellte sich für praktische Anwendungen als sehr aufwendig heraus, und bis auf den Vorteil, dass die Variablen den Hauptkomponenten eindeutig zugeordnet werden können, überwiegen hier die Nachteile der langen Laufzeit und der schlechten grafischen Darstellung der Hauptkomponenten.

3.7 Automatisierung der Methoden

Für praktische Anwendungen stellt sich die Frage, wie diese Berechnungsprozeduren automatisiert werden können. Abbildung 3.31 zeigt ein Flussdiagramm, dass die Berechnungsprozedur für die Hauptkomponentenanalyse zusammenfasst. Eine Automatisierung im eigentlichen Sinn ist kaum möglich, da vor allem bei der sparsen Variante in jedem Wiederholungsschritt Enscheidungen aufgrund nicht eindeutiger Kriterien getroffen werden müssen, die eine Interpretation der involvierten Person nötig machen. Darum wird hier eine Anleitung für die Arbeitsschritte gegeben.

Bevor überhaupt an eine Berechnung gedacht werden kann, müssen geeignete Daten ausgewählt werden. Diese sollten möglichst frei von Messfehlern sein und die Zyklen, die hinter den beschreibenden Faktoren stehen, sollten in ihrer Anzahl ausgeglichen sein, damit nicht ein Faktor überproporational vertreten ist. Diese Anforderungen an die Daten sind Idealzustände, die oft nicht gegeben sind, jedoch sollte trotzdem darauf geachtet werden, dass diese so gut wie möglich erfüllt werden.

Die definierten Merkmale legen fest, welche Aussagen später über die Daten getroffen werden können. Dieser Punkt ist daher nicht zu vernachlässigen. Man sollte sich gut überlegen, welche Informationen man erhalten möchte, da man dies bereits in der Merkmalsdefinition berücksichtigen muss. Ein Beispiel wäre das dynamische Verhalten eines Faktors, welches über Quantile der Differenzen dargestellt werden könnte. Fehlen diese Merkmale bezüglich der Differenzen, kann in der späteren Interpretation nicht darauf eingegangen werden. Weiters ist zu entscheiden, ob man die Merkmale mit robusten oder nicht robusten statistischen Schätzern definiert.

Hat man nun eine große Menge von Merkmalen definiert, kann eine Korrelationsanalyse dabei helfen, diese zu reduzieren. Dabei werden von stark korrelierenden Merkmalen Repräsentanten für die Hauptkomponentenanalyse ausgewählt. In der praktischen Berechnung kann es vorkommen, dass es Merkmale gibt, die man auf keinen Fall aussortiert, auch bei starker Korrelation untereinander, da diese zum Beispiel Vergleichswerte mit bereits erfolgten Auswertungen liefern sollen.

Nach dem Schritt der Korrelationsanalyse stehen die Merkmale und Beobachtungen für die Hauptkomponentenanalyse fest. Es muss nun entschieden werden, welche Methode angewandt wird. Die robuste Methode ist zu bevorzugen, da Ausreißer und Extremwerte hier kaum Einfluss nehmen und das tatsächliche Verhalten der Daten in den Ergebnissen wiedergespiegelt wird. Eine Diskussion dieser Vorteile wurde in Abschnitt 3.6 bereits gegeben.

Bevor die robuste Methode angewandt wird, sollte überprüft werden, ob der Q_n-Schätzer bei einer dieser Variablen 0 ist. Variablen, deren Q_n-Schätzer 0 ist, müssen entfernt werden, da ansonsten bei der Skalierung eine Division durch 0 erfolgt. Für die Berechnung der Hauptkomponentenanalyse wurde das R Package `pcaPP` verwendet mit den Funktionen `qn`, `l1median`, `PCAgrid` für die robuste PCA und `sPCAgrid` für die robust and sparse PCA.

In Abbildung 3.31 wird zwischen sparse und nicht sparse getrennt. Dies liegt daran, dass bei der sparsen Hauptkomponentenanalyse ein wiederholtes Berechnen der PCA zur Bestimmung des Tuningparameters erfolgt, bis dieser ein optimales Ergebnis liefert. Als

Entscheidungsgrundlage für die Wahl des Tuningparameter gibt es die Darstellung der erklärten Varianz über die Tuningparameter und das BIC-Kriterium. Die Auswahl eines Tuningparameters mit Hilfe der erklärten Varianz, dargestellt über die Tuningparameter, erfolgt, indem man einen Parameter kurz vor einem starken Abfall der Kurve wählt. Da dieses Kriterium nicht eindeutig ist, sollte es mit dem BIC-Kriterium überprüft werden.

Um zu überprüfen, ob die Ladungsmatrix bereits ausreichend dünn besetzt und damit gut interpretierbar ist, können die Hauptkomponenten entlang der Variablen dargestellt werden. Die Auswahl des Tuningparameters geht mit der Auswahl der Anzahl der Hauptkomponenten einher. Im Falle einer nicht sparsen Berechnung ist nur die Anzahl der Hauptkomponenten zu ermitteln. Liefert dieser Abschnitt ein befriedigendes Ergebniss, kann die Auswertung und Interpretation begonnen werden.

Die relevanten Ladungen können nun mit Gewichtung durch die erklärte Varianz berechnet werden. Hierbei werden die ausgewählten Hauptkomponenten mit ihrer jeweiligen erklärten Varianz gewichtet, summiert und die Variablen absteigend nach ihrer Relevanz geordnet. Dies ist eine relativ schnelle Methode, Variablen, die das Datenset beschreiben, zu erhalten. Es sollte trotzdem noch ein Blick auf die Ladungen der Hauptkomponenten geworfen werden, da es vielleicht Variablen gibt, die nur für eine Komponente relevant sind. Diese Variablen würden dann in der Reihung weiter hinten vorkommen und somit nicht berücksichtigt werden.

Für eine optische Interpretation können die Hauptkomponenten gegeneinander dargestellt werden. Mit Hilfe der Variablen, die zu diesen Hauptkomponenten korrelieren, können Eigenschaften von Faktoren festgestellt werden.

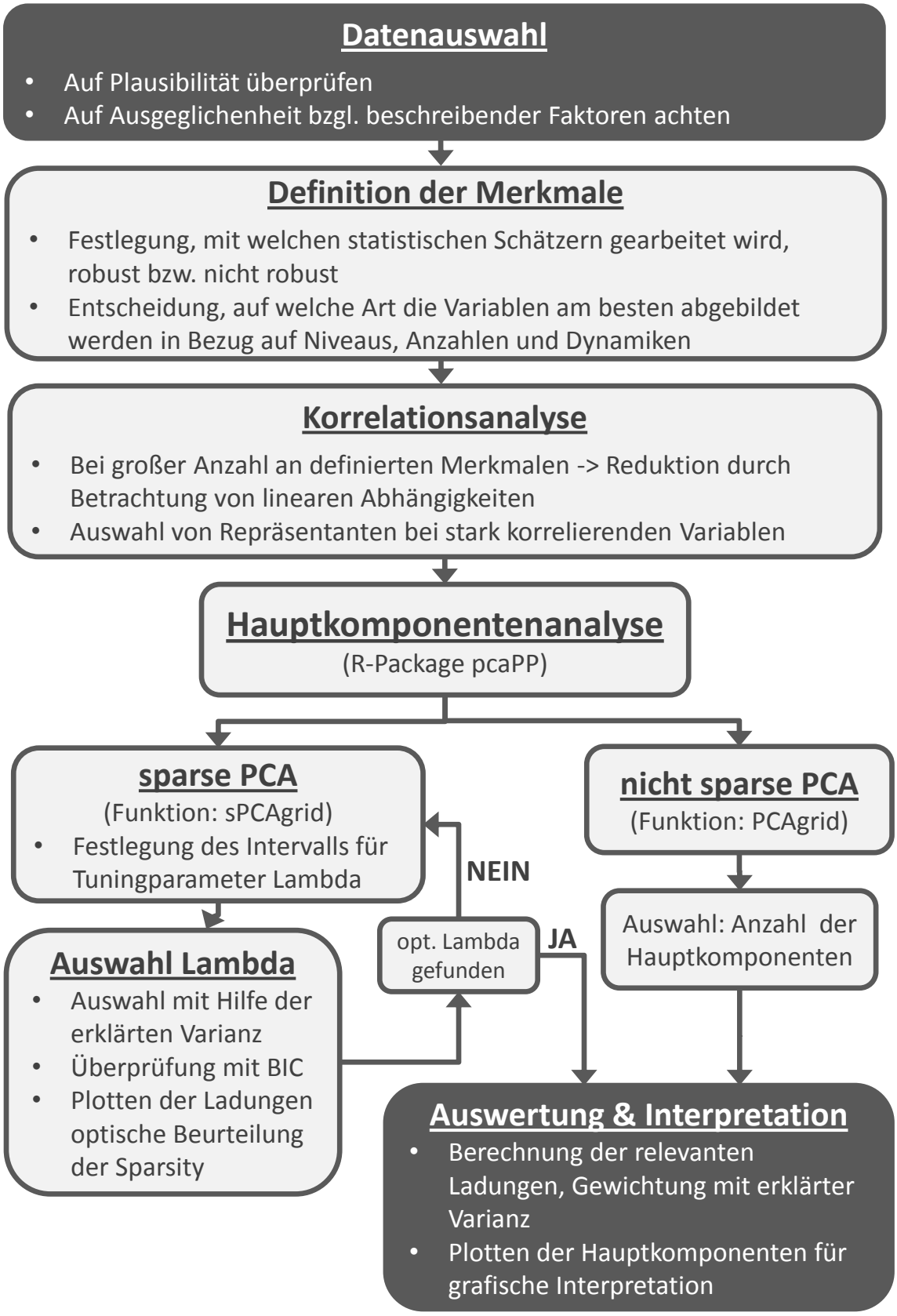


Abbildung 3.31: Flussdiagramm für Auswertungsschema

Kapitel 4

Interpretation der Ergebnisse

4.1 Auswertung einzelner Kenngrößen

Vor der Betrachtung der Ergebnisse der Hauptkomponentenanalysen wird ein Blick auf einzelne ausgesuchte Merkmale geworfen, die bereits in der Hauptkomponentenanalyse großen Einfluss hatten. Die Nutzung und das Verhalten der Variablen stehen in diesem Kapitel im Vordergrund. Die Variablen wurden daher anhand des Faktors *Betriebsart* getrennt dargestellt. Die im Folgenden gezeigten Variablen sollen auch noch einmal klar machen, warum gewisse Variablen von der robusten, jedoch nicht von der klassischen Hauptkomponentenanalyse gewählt wurden.

Die Kenngrößen der Fahrzeuggeschwindigkeit sind unter den ersten Merkmalen, um Informationen über die Betriebsart zu erhalten. Abbildung 4.1 zeigt die 25, 50 und 75% Quantile der Fahrzeuggeschwindigkeit *VehicleSpeedQ25*, *VehicleSpeedQ50* und *VehicleSpeedQ75* dargestellt als Boxplotserie für alle 37 Fahrzeuge. Das heißt, im Konkreten enthält jeder einzelne Boxplot die berechneten Quantile der einzelnen Zyklen für das jeweilige Fahrzeug. Bei der Verteilung der Quantile lassen sich Gemeinsamkeiten feststellen: es haben die Fahrzeuge 14 und 15 optisch eine sehr ähnliche Verteilung der Boxen für die jeweiligen Quantile. Diese Fahrzeuge wurden mit der Betriebsart *Fernverkehr Passstrecke* deklariert. Weitere Fahrzeuge mit Betriebsart *Fernverkehr Passstrecke* sind die Fahrzeuge 1, 2, 20 und 25, welche auch vergleichbar sind. Bei Fahrzeug 30 ist sehr gut zu erkennen, dass es sich um eine Stadtnutzung handelt. Die Boxplots haben im Vergleich zu den anderen Fahrzeugen ein sehr geringes Niveau und sind der Betriebsart *Stadt Verteilerbetrieb* zugeordnet. Das mittlere Niveau von *VehicleSpeedQ25* liegt nur sehr knapp über Null, was bedeutet, dass das Fahrzeug nur sehr langsam fährt. Für das hier nicht eingezeichnete 90% Quantil *VehicleSpeedQ90* liegt der Median bei ca. 50km/h. Das bedeutet, dass maximale Geschwindigkeiten um 50km/h gefahren wurden, wodurch das Bild einer Stadtnutzung zusätzlich unterstrichen wird. Die zwei Fahrzeuge 21 und 22 stechen aufgrund ihrer langen Boxen heraus, was auf eine hohe Streuung innerhalb der Quantil-Klassen hindeutet. Diese beiden Fahrzeuge gehören zum *Stadt Verteilerbetrieb*. Die zwei Fahrzeuge 7 und 12 haben ein sehr hohes Geschwindigkeitsniveau und waren Zyklen aus dem *Kundenbetrieb*.

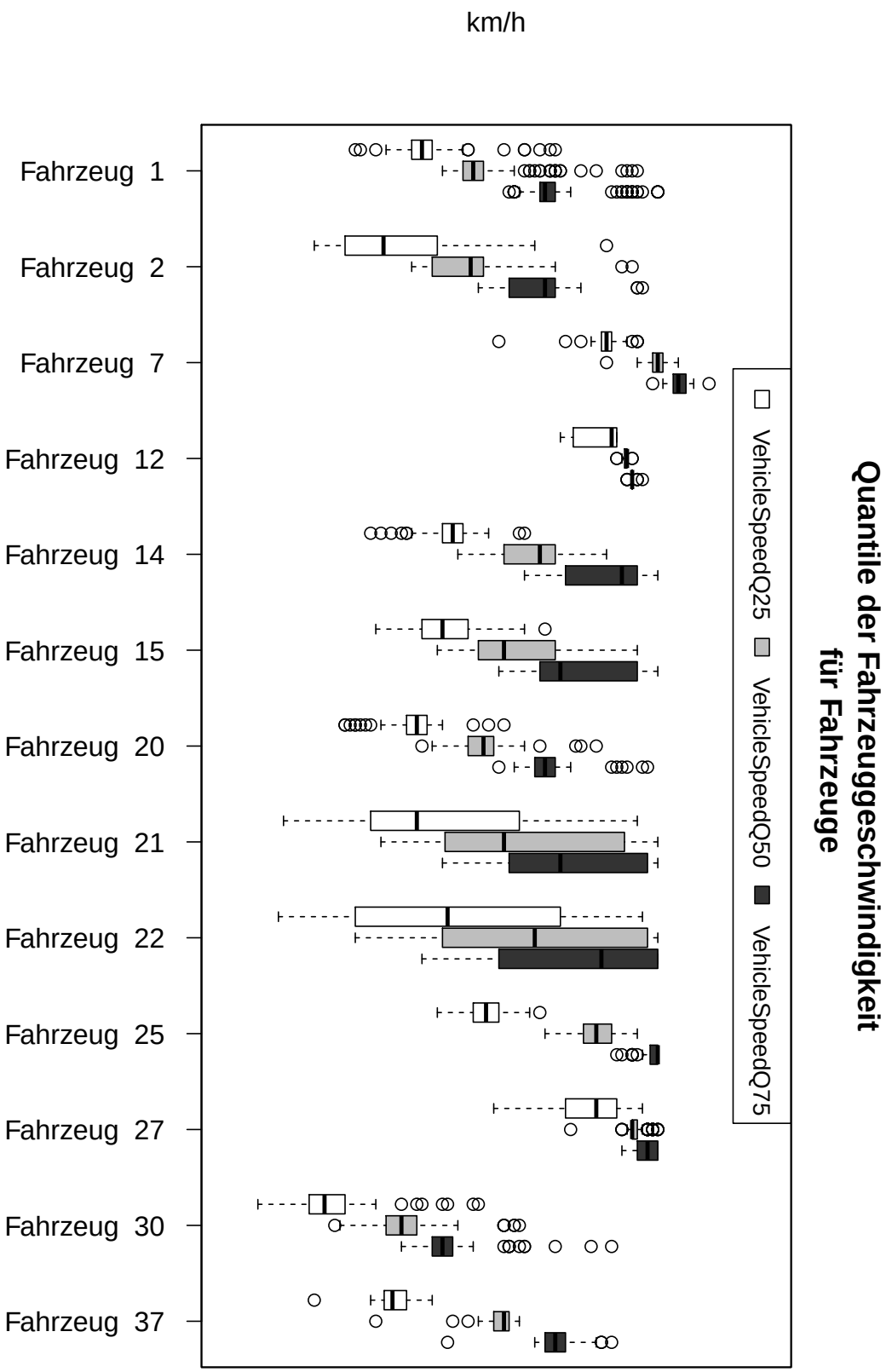


Abbildung 4.1: Boxplotserie der Quantile 25%, 50% und 75% der Fahrzeuggeschwindigkeit für Fahrzeuge

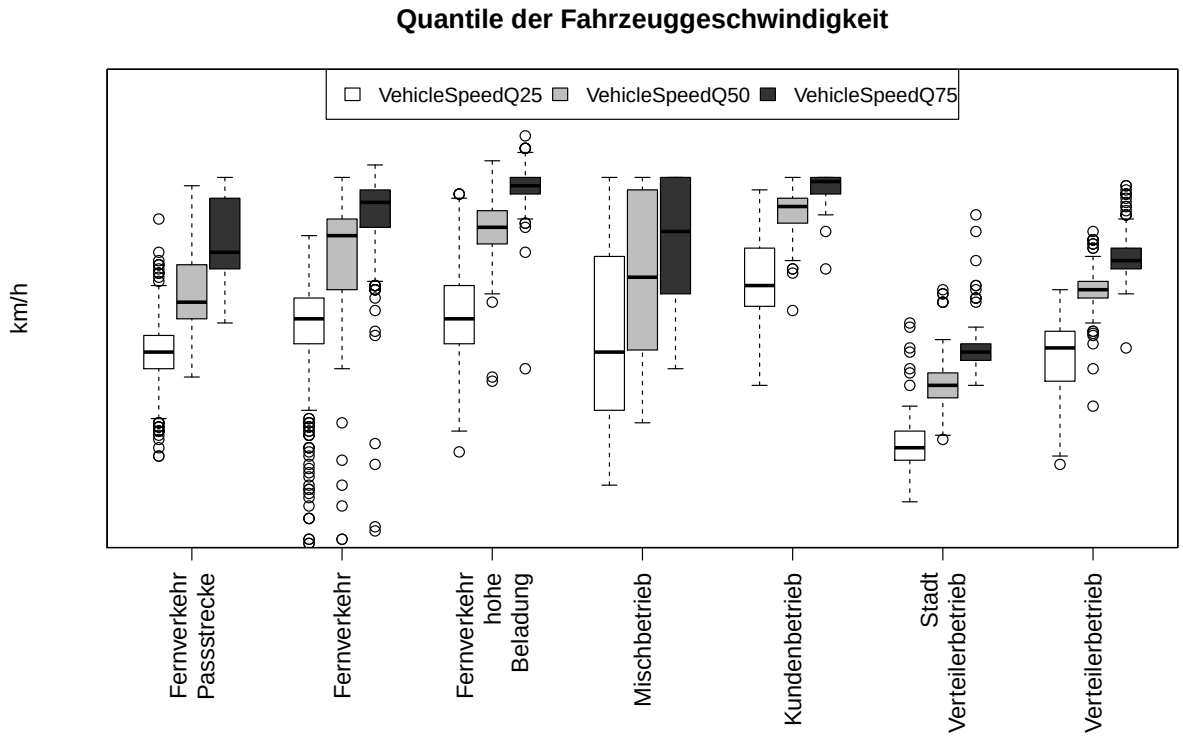


Abbildung 4.2: Boxplotserie der Quantile 25%, 50% und 75% der Fahrzeuggeschwindigkeit für Betriebsarten

Abbildung 4.2 zeigt die Quantile der Fahrzeuggeschwindigkeiten für die Betriebsarten. Bis auf die Betriebsart *Stadt Verteilerbetrieb* enthalten alle Betriebsarten mehr als ein Fahrzeug. Die Fahrzyklen aus dem *Kundenbetrieb* haben hohe Geschwindigkeitsniveaus vermutlich auf Grund von Autobahnfahrten. Die Betriebsart *Stadt Verteilerbetrieb* hat, wie vorhin bereits erwähnt, das geringste Geschwindigkeitsniveau. Zwischen *Fernverkehr* und *Fernverkehr volle Beladung* lassen sich Unterschiede bei *VehicleSpeedQ50* und *VehicleSpeedQ75* erkennen. Die Boxen sind bei *Fernverkehr volle Beladung* kürzer, was auf eine niedrige Streuung hinweist.

Differenzen sind jene Variablen, die unter den Variablenklassen die Dynamiken abbilden. Die positiven Differenzen zeigen Beschleunigung und die negativen Bremsvorgänge. In Abbildung 4.3 sind die positiven Fahrzeuggeschwindigkeitsdifferenzen innerhalb von 10 Sekunden dargestellt. Ein dynamisches Verhalten zeigt hier der *Stadt Verteilerbetrieb*, welcher im Median signifikant höhere Quantile als die restlichen Betriebsarten aufweist. Des Weiteren ist *VehicleSpeedDiffPos10SecQ99* nach oben deutlich beschränkt, was eine Konsequenz der Geschwindigkeitsbegrenzungen innerhalb der Stadt ist. Bei den Betriebsarten *Fernverkehr Passstrecke*, *Fernverkehr volle Beladung* und *Kundenbetrieb* gibt es Ausreißer nach oben mit sehr hoher Geschwindigkeit. Dies ist vermutlich ein Resultat von Auffahrten auf Autobahnen oder Landstraßen.

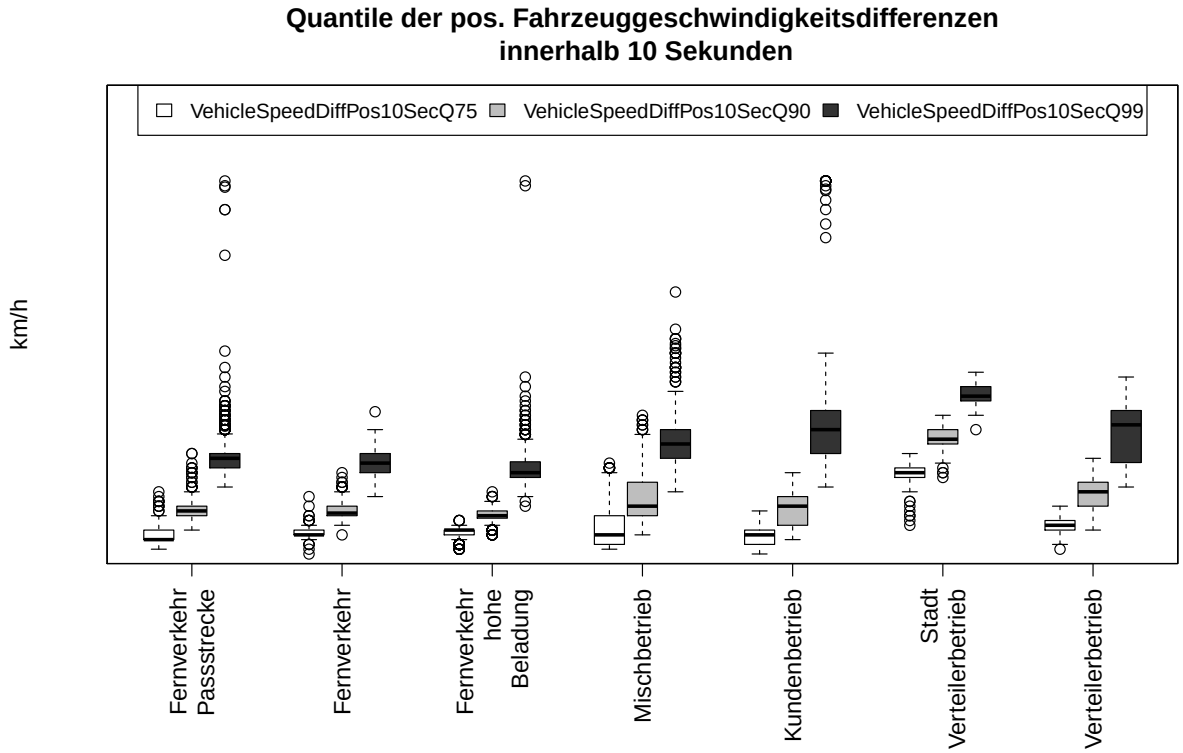


Abbildung 4.3: Positive Fahrzeuggeschwindigkeitsdifferenzen für die Quantile 75%, 90% und 99% der Fahrzeuggeschwindigkeit für Betriebsarten

Dynamiken der Fahrzeuggeschwindigkeit wurden vor allem von der klassischen Hauptkomponentenanalyse gewählt und wurden bei der robusten Hautkomponentenanalyse erst in der 2. Einschränkung relevant, wie in Abschnitt 3.6 bereits gezeigt wurde.

Da 3 Motorklassen unterschiedlicher Leistung und Nenndrehzahl betrachtet wurden, wurden die Variablen der Drehzahl auf einen festen Maximalwert von Umdrehungen pro Minute skaliert. Den gesamten Drehzahlbereich decken *Fernverkehr Passstrecke* und *Fernverkehr volle Beladung* ab. Hervorzuheben ist hier *EngineSpeedQ05* bei der Nutzung *Fernverkehr Passstrecke*, da diese Nutzung die einzige ist, die beim 5% Quantil eine sehr lange Box hat, was auf eine hohe Streuung hinweist. *Fernverkehr* und *Fernverkehr volle Beladung* haben eine sehr kurze Box, was wiederum eine niedrige Streuung bedeutet, und sehr viele Ausreißer nach oben auf. Dies resultiert vermutlich aus der Topografie. Ein ganz anderes Bild liefert die Nutzung *Mischzyklen*, deren 5% Quantil sich auf einen sehr kleinen Bereich konzentriert. In den Ergebnissen der robusten Hauptkomponenten Analyse hat die Variable *EngineSpeedQ05* eine relevante Bedeutung.

Da die Quantile der Motorstopps bei der robusten Methode einen großen Einfluss hatten, sind die Variablen *EngineStopsQ10*, *EngineStopsQ50* und *EngineStopsQ90* in Abbildung 4.5 dargestellt. Die Variable *EngineSpeedQ90* im *Kundenbetrieb* hat eine längere Box im Vergleich zu den anderen Betriebsarten. Die hohen Ausreißer der Betriebsart *Fernver-*

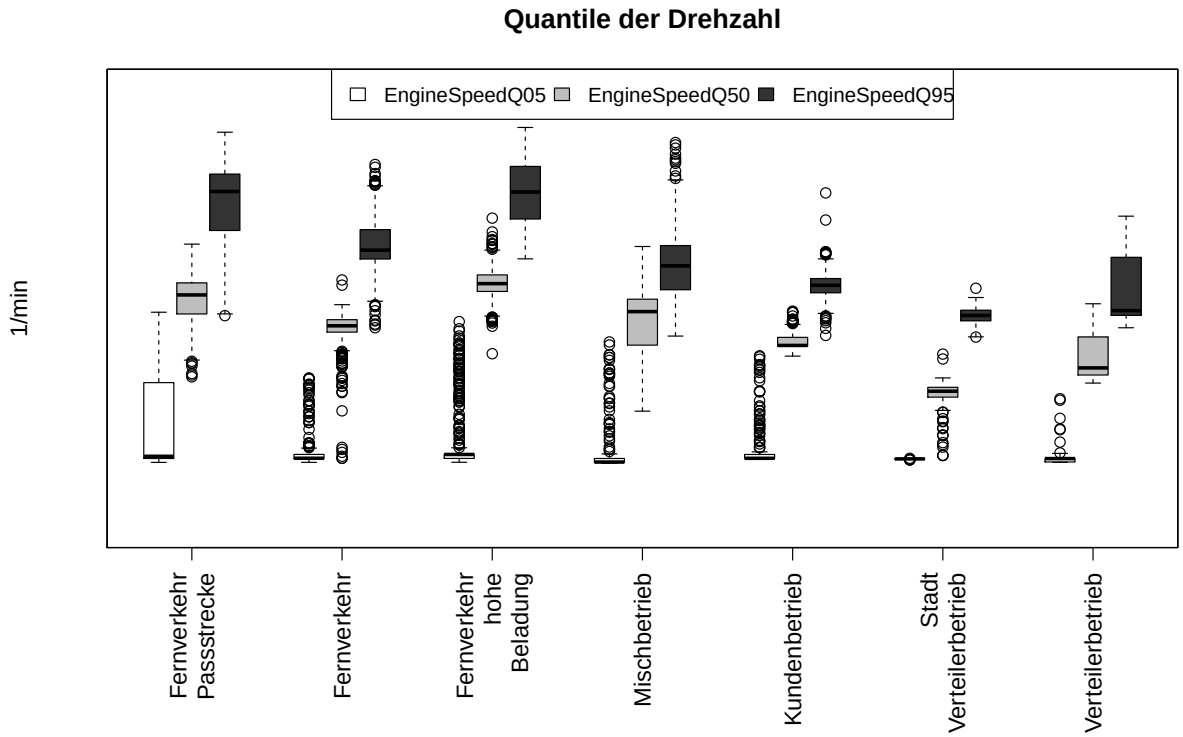


Abbildung 4.4: Boxplotserie für die Quantile 5%, 50% und 95% der Drehzahl für Betriebsarten

kehr volle Beladung lassen sich im Nachhinein nicht mehr erklären, sollten an dieser Stelle aber erwähnt werden. Die Variable *EngineStopsQ90* zeigt gut, warum diese nur für die robust skalierten Datensets gewählt wurde. Hier spielen die Eigenschaften des Q_n-Schätzers und der Standardabweichung eine Rolle, wie bereits in Abschnitt 3.6 erläutert. Die Standardabweichung reagiert viel sensitiver auf die Ausreißer des *Kundenbetriebs* und skaliert sie damit klein. Auf den Q_n-Schätzer wirken die Zyklen des *Kundenbetriebs* geringer, da dieser einen Bruchpunkt von 50% hat, den die Anzahl der Zyklen des *Kundenbetriebs* bei weitem nicht erreichen. Der Q_n-Schätzer bildet daher die große Menge der restlichen Zyklen ab, die ein ähnliches Verhalten zeigen und hat dadurch einen kleineren Wert als die Standardabweichung, wodurch eine andere Skalierung resultiert.

Die Leistung wurde in Prozent der maximal möglichen Leistung, dargestellt, da die Fahrzeuge unterschiedliche Leistungen hatten. In Abbildung 4.6 ist der Interquartilbereich der positiven und der negativen Leistung dargestellt. Die robuste Hauptkomponentenanalyse wählte die Variable *PowerNegIQR* als signifikant. In Abbildung 4.6 zeigt sich der Schubbetrieb bei der Schwerlastnutzung.

Das positive 99% Quantil des Drehmoments ist in der robust and sparse Methode eine bedeutende Variable in den Ergebnissen. Abbildung 4.7 zeigt die Quantile *TorquePosQ50*, *TorquePosQ90* und *TorquePosQ99*. Es zeigt sich, wie zuvor bei der Drehzahl und

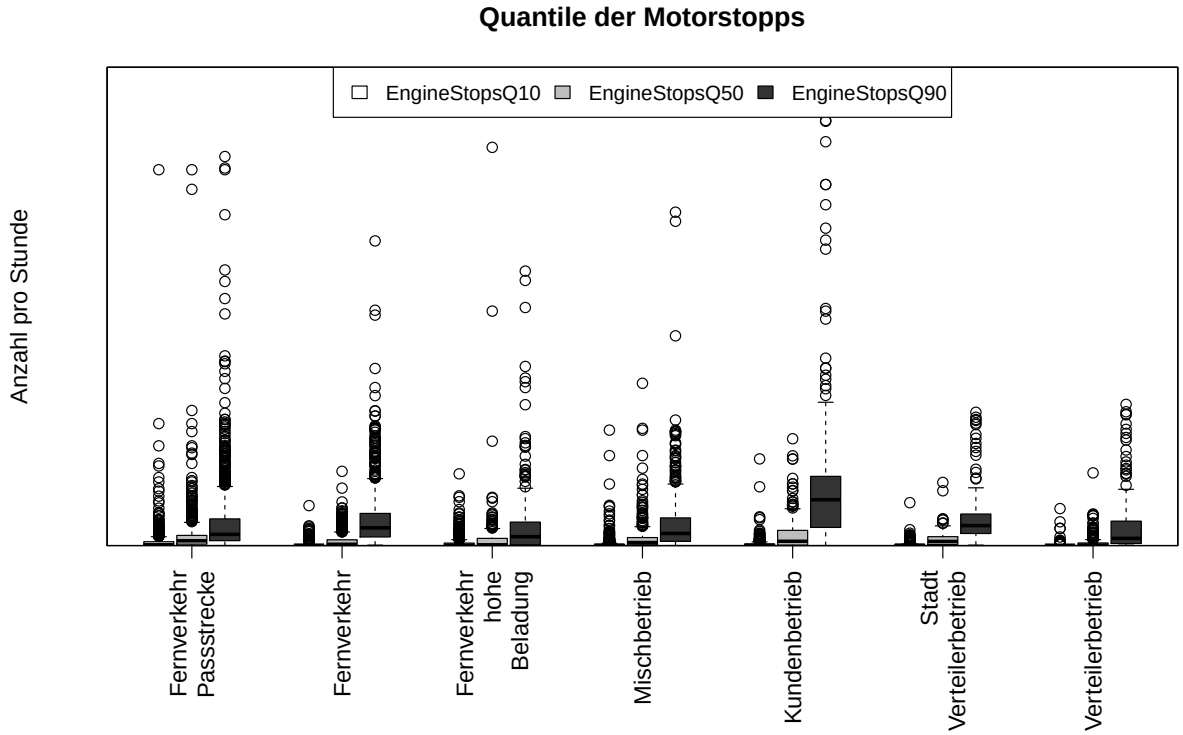


Abbildung 4.5: Boxplotserie der Quantile 10%, 50% und 90% der Motorstopps für Betriebsarten

der Fahrzeuggeschwindigkeit, dass der *Stadt Verteilerbetrieb* geringe mittlere Niveaus hat und Schwerlastfahrten wie *Fernverkehr volle Beladung* hohe mittlere Niveaus haben. Die Variable *TorquePosQ99* zeigt wieder ein Verhalten, das den Q_n-Schätzer viel kleiner werden lässt als die Standardabweichung. Die *Betriebsarten Stadt Verteilerbetrieb* und *Fernverkehr* haben ein geringes mittleres Niveau und *Mischzyklus* und *Verteiler* haben längere Boxen als die übrigen Betriebsarten. Daher konzentriert sich der Q_n-Schätzer auf die große Anzahl an Zyklen, die ein hohes mittleres Niveau mit sehr schmalen Boxen haben.

Die letzten Variablen sind Größen der Abgasnachbehandlung. In Abbildung 4.8 sind die 25% und 50% Quantile der NOx Emission nach SCR-Kat zu sehen. Diese waren bei der robusten Methode relevante Variablen. Hier zeigt vor allem *NOxExhausafterSCRKATQ50*, dass die Schwerlastbetriebsarten *Fernverkehr Passstrecke* und *Fernverkehr volle Beladung* einen hohen Emissionsausstoß haben. Auch hier ist wieder gut zu sehen, warum der Q_n-Schätzer einen viel geringeren Wert als die Standardabweichung liefert und daher für die robuste Hauptkomponentenanalyse gewählt wurde.

Diese relativ einfache Betrachtung der Zyklen, getrennt nach Betriebsarten, zeigt bereits wie sich der reale Betrieb in den Variablen darstellt. Es ist im Umkehrschluss eine Legitimierung für die Auswahl dieser Variablen durch die Hauptkomponentenanalyse, da diese unterschiedliche Betriebsarten darstellen. Ein Beispiel für eine unabhängige Variable ist der

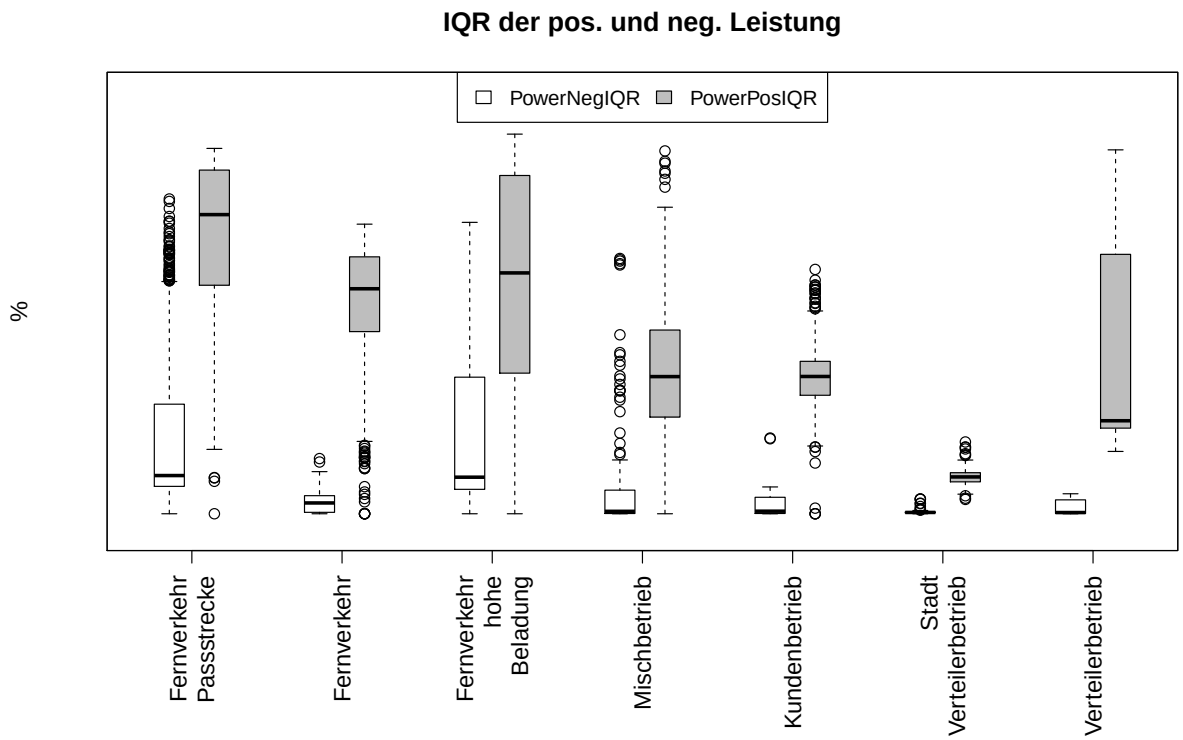


Abbildung 4.6: Boxplotserie des Interquartilbereichs der pos. und neg. Leistung für Betriebsarten

Median der Außentemperatur *OutdoorTempQ50*, der im Anhang in Abbildung A.4 dargestellt ist. Hier kann man keine Variation unter den Betriebsarten erkennen. Diese Variable hatte auch in den Ergebnissen aller Hauptkomponentenanalysen nur geringe Ladungswerte und ist daher als nicht relevant für die hier durchgeführte Analyse zu betrachten.

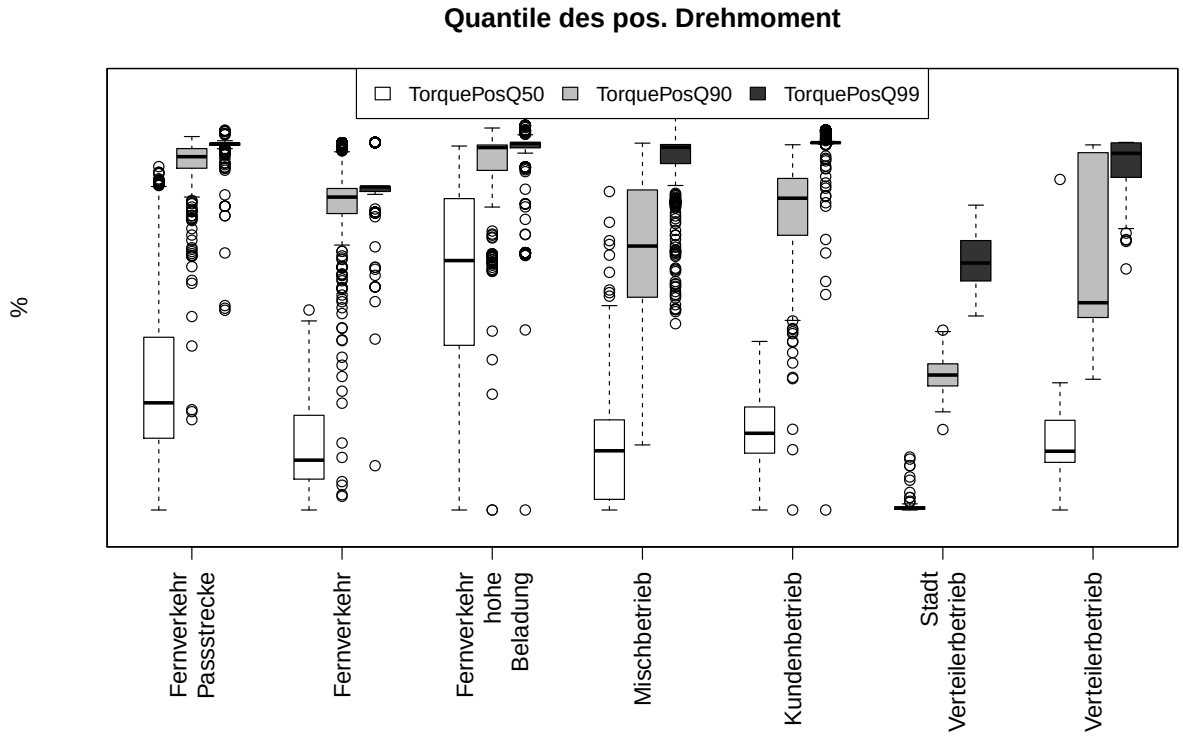


Abbildung 4.7: Boxplotserie der Quantile 50%, 90% und 99% des pos. Drehmoment für Betriebsarten

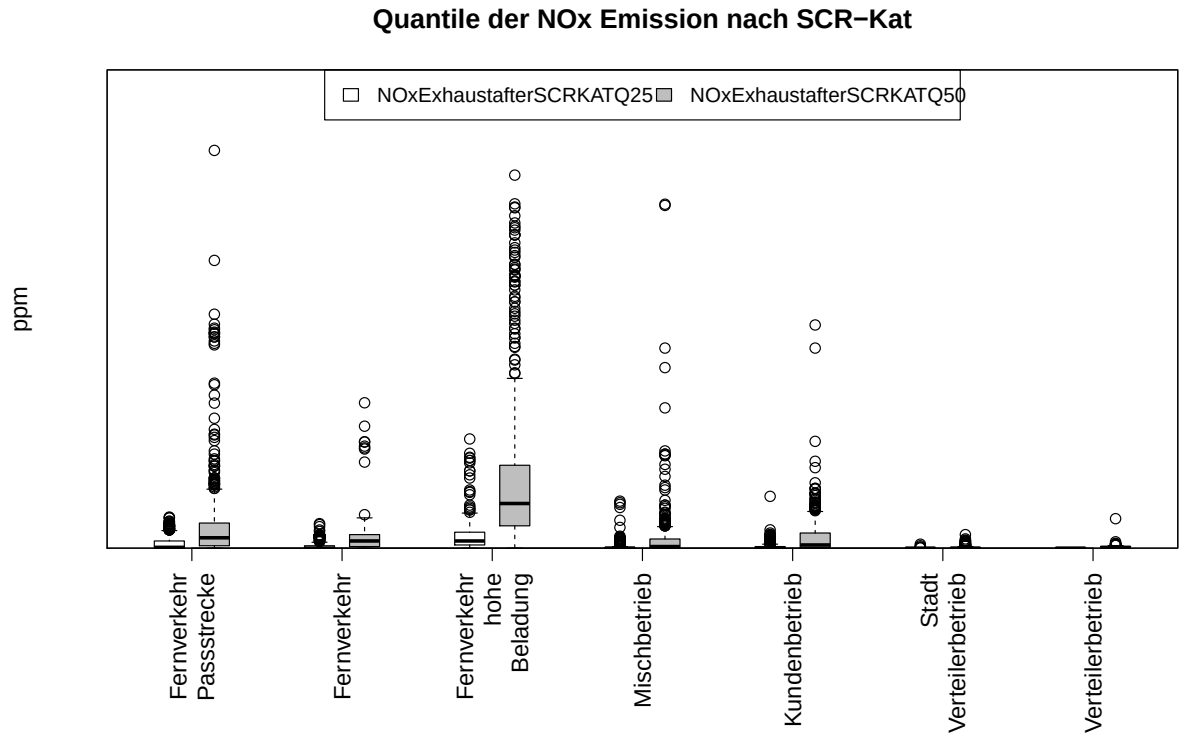


Abbildung 4.8: Boxplotserie der Quantile 25% und 50% der NOx Emission nach SCR-Kat für Betriebsarten

4.2 Auswertung 1. Einschränkung

Für die Interpretation der Ergebnisse wurden die Variablenauswahlen 1. Einschränkung und 2. Einschränkung herangezogen. Der Grund für diese Wahl ist, dass die Basisauswahl Größen der Abgasnachbehandlung beinhaltet, welche in anderen Datenmengen nur sehr selten vorkommen. Darum wurde die Basisauswahl primär für die Auswertung der Methodik herangezogen. Auf diese Variablenreduktionen wurde in Abschnitt 3.1 bereits eingegangen.

Für die erste Einschränkung wurden 8 Hauptkomponenten gewählt, um das Datenset zu beschreiben. Diese 8 Hauptkomponenten erklärten im Median 80,46% der Variabilität über alle 10 Datensets. Eine Auflistung der erklärten Varianz für die einzelnen Datensets findet sich in Tabelle 3.2. In Abbildung 4.9 ist die Reihung der Variablen anhand der Ladungen und der erklärten Varianz für die robuste Hauptkomponentenanalyse zu sehen.

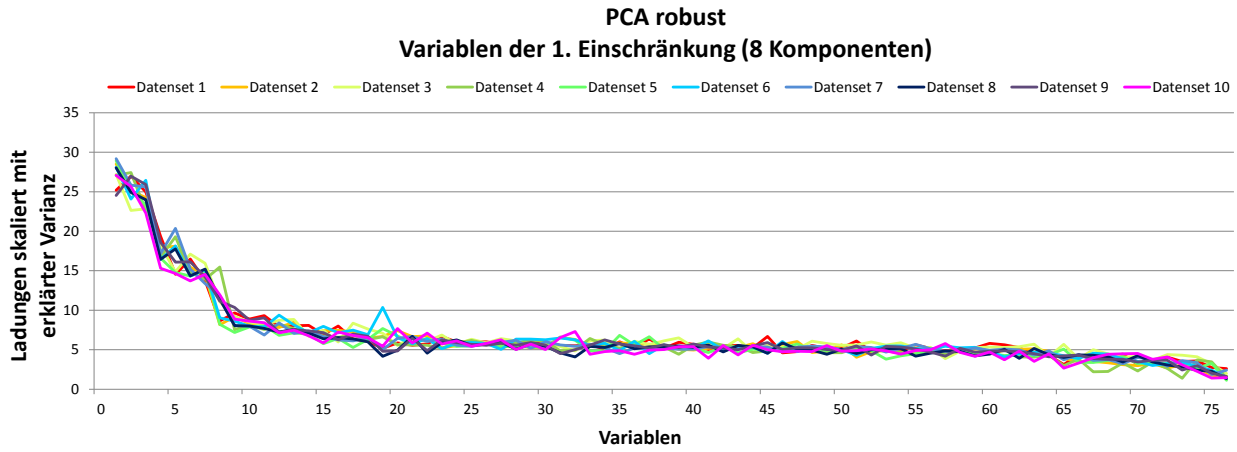


Abbildung 4.9: Absteigende Reihung der Variablen für die 1. Einschränkung und robuste Berechnung

Die Wahl von 8 Variablen erscheint sinnvoll, da nach der achten Variable die Kurve flach weiter läuft. Die Variable 19 *VehicleStopsQ90* bei Datenset 6 zeigt mit seinem hohen Wert eine starke Abweichung zu den anderen Datensets. Bei näherer Betrachtung der Ladungsmatrix von Datenset 6 wird dies auch deutlich, da die Variable *VehicleStopsQ90* in den Hauptkomponenten 3, 6 und 7 hohe positive Ladungen hat. Der Ursprung dieses Unterschieds liegt im Ziehen der Stichproben. Ob die weiteren Variablen nach der achten Variable noch Einfluss auf die Hauptkomponenten, haben wird sich in der weiteren Ergebnisauswertung zeigen.

Tabelle 4.1 zeigt die ersten 8 Variablen aufgrund der Reihung mit Ladungen und erklärter Varianz. Eine Variable erhielt einen Eintrag, wenn die dazugehörige Ladung einen betragsmäßig so großen Wert hatte, dass sie unter den größten positiven beziehungsweise negativen Ladungen war. Dabei bedeutet ein Plus, dass diese Variable die Tendenz zu positiven Ladungen über die zehn Datensets hinweg hatte und ein Minus, dass die Variable eine Tendenz zu negativen Ladungen hatte. War die Tendenz für das Vorzeichen nicht

klar erkennbar, wurde ein Kreis eingetragen. Eine Variable wurde in die Tabelle aufgenommen, wenn sie mindestens 9 mal unter den höchsten positiven bzw. negativen 5 Ladungen war. Variable *TorquePosQ99* bei Hauptkomponente 5 bildet hier eine Ausnahme: Da sie nur 8 mal unter den besten 5 Ladungen vorkam, jedoch hohe Ladungswerte im Vergleich zu anderen Variablen in Hauptkomponente 5 hatte, wurde sie hinzugefügt. Es zeigt sich, dass mit abnehmender Relevanz auch der Einfluss auf die acht Hauptkomponenten schwindet. In Hauptkomponente 8 war keine der ausgewählten 8 Variablen signifikant vertreten. Die Variable *PowerNegIQR* bildet hier eine Ausnahme, da diese oft vertreten war, wenn auch selten bei hohen Ladungen, und auch weit oben gereiht ist, hat sie jedoch nur auf drei Hauptkomponenten wesentlichen Einfluss. Aufgrund dieser Beobachtung wurden die nächsten Variablen gefunden, die nur auf wenige Hauptkomponenten Einfluss haben, bei diesen Hauptkomponenten jedoch hohe Ladungen haben: *SPEEDDiffPos10SecQ99*, *SPEEDDiffPos10SecQ90*, *TorqueDiffNeg3SecQ50*, *PowerDiffNeg3SecQ50*, *TorqueDiffPos3SecQ50* und *PowerDiffPos3SecQ50*.

Robuste Methode für die 1. Einschränkung									
Nr.	Variablen	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8
1	EngineStopsQ10	+	-	+	o	- (9)	+ (9)		
2	TorquePosQ99	+	+ (9)			- (8)			
3	EngineSpeedQ05	+	+	+	o	+		o (9)	
4	PowerNegIQR	+	- (9)				+		
5	EngineStopsQ50			+ (9)					
6	TorqueNegIQR								
7	VehicleStopsHour	-	- (9)	+ (9)					
8	EngineStopsQ90								

Tabelle 4.1: Auswahl der 8 relevanten Variablen für die 1. Auswahl unter der robusten Methode

Die Hauptkomponenten werden mit Hilfe der Scorematrix dargestellt, welche in den Spalten die Hauptkomponenten und in den Zeilen die Beobachtungen beinhaltet. Abbildung 4.10 zeigt die 1. und 2. Hauptkomponente des Datensets 1. Die Punkte sind nach der Betriebsart gefärbt. Auf der 1. Hauptkomponente liegen die Variablen *TorquePosQ99*, *EngineStopsQ10* und *EngineSpeedQ05* mit positiven Ladungen und *VehicleStopsHour* mit negativer Ladung. Auf der 2. Hauptkomponente liegen *TorquePosQ99* und *EngineSpeedQ05* mit positiven Ladungen und *PowerNegIQR* und *EngineStopsQ10* mit negativen Ladungen. Entlang der 1. Hauptkomponente sind die Betriebsarten, bis auf wenige Ausreißer, gut getrennt. Rechts trennen die Variablen *TorquePosQ99*, *EngineStopsQ10* und *EngineSpeedQ05* die Zyklen der *Fernverkehr Passstrecke* und *Fernverkehr volle Beladung* vom *Stadt Verteilerbetrieb* und Teilen der *Verteilerzyklen*. Links auf der 1. Hauptkomponente liegen Zyklen, die viele Fahrzeugstopps hatten, daher ist es legitim, dass der *Stadt Verteilerbetrieb* ganz außen liegt, da hier durch die Nutzungsart viele Fahrzeugstopps entstehen. Im Zentrum liegt eine Gruppe Zyklen der Betriebsart *Verteiler* und die *Fernverkehrzyklen*. In

einem weiteren Radius um null liegen die *Kundenfahrzeuge*, von welchen man kaum Informationen bezüglich der Nutzung hat. Die zweite Hauptkomponente trennt Zyklen mit schwerer Beladung oder einer anspruchsvollen Strecke im positiven Bereich von den Fahrzeugen, die hohe Variabilität im Interquartilbereich der negativen Leistung haben und eine hohe Anzahl an kurzen Motorstopps. Auffallend ist, dass auch die 2. Hauptkomponente die *Verteilerzyklen* in zwei Gruppen trennt.

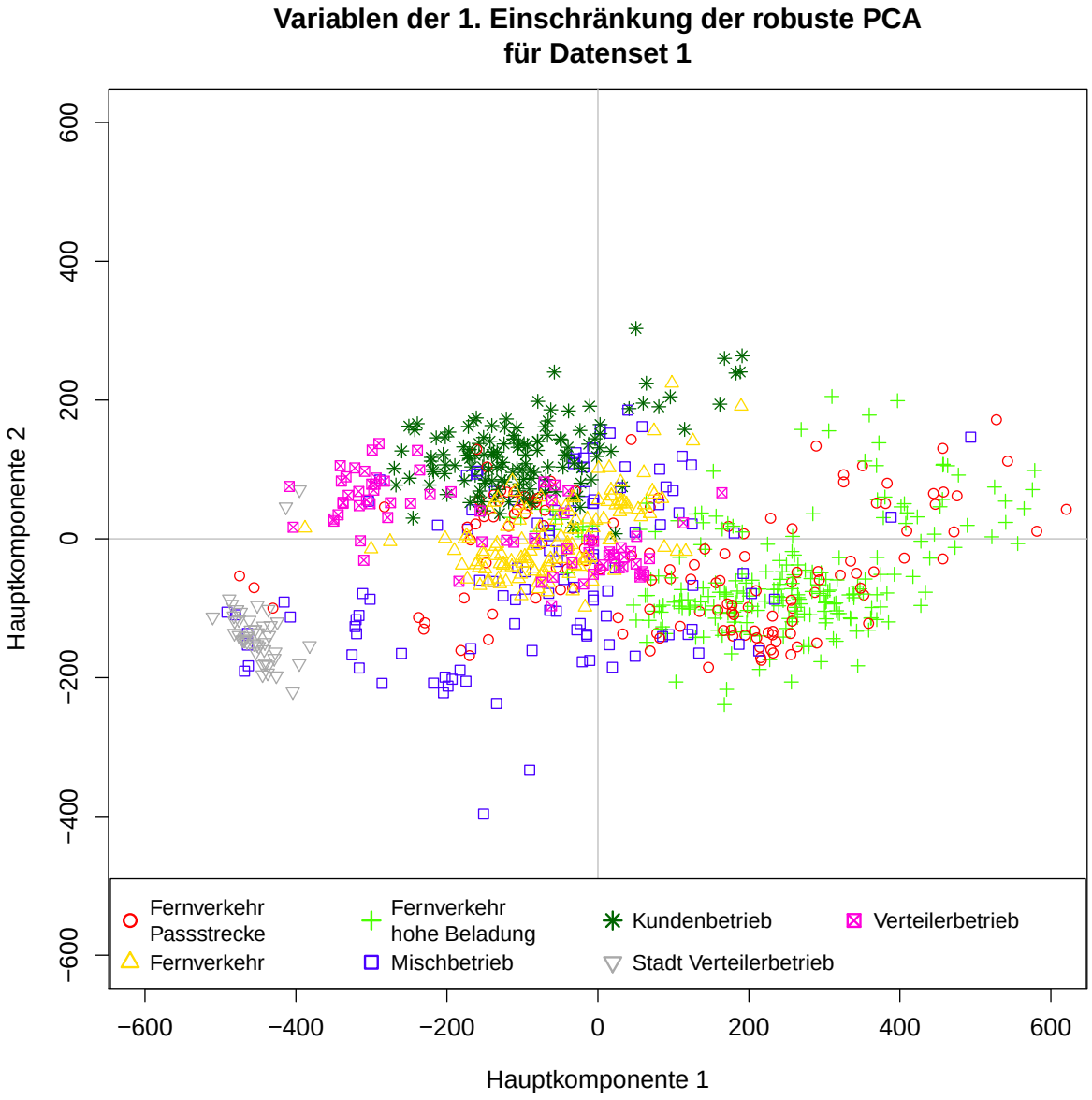


Abbildung 4.10: 1. und 2. Hauptkomponente der 1. Einschränkung für Datenset 1

Die Trennung der Betriebsart *Verteiler* liegt daran, dass hier Zyklen aus drei Fahrzeugen ausgewählt wurden. Dabei konzentrieren sich die Zyklen von zwei Fahrzeugen zu jeweils einer Gruppe und die Zyklen des dritten Fahrzeugs streuen zwischen den beiden Gruppen. Eine Aufstellung, wieviele Fahrzeuge zu jeder Betriebsart gehören, ist in Tabelle A.1 im Anhang dargestellt.

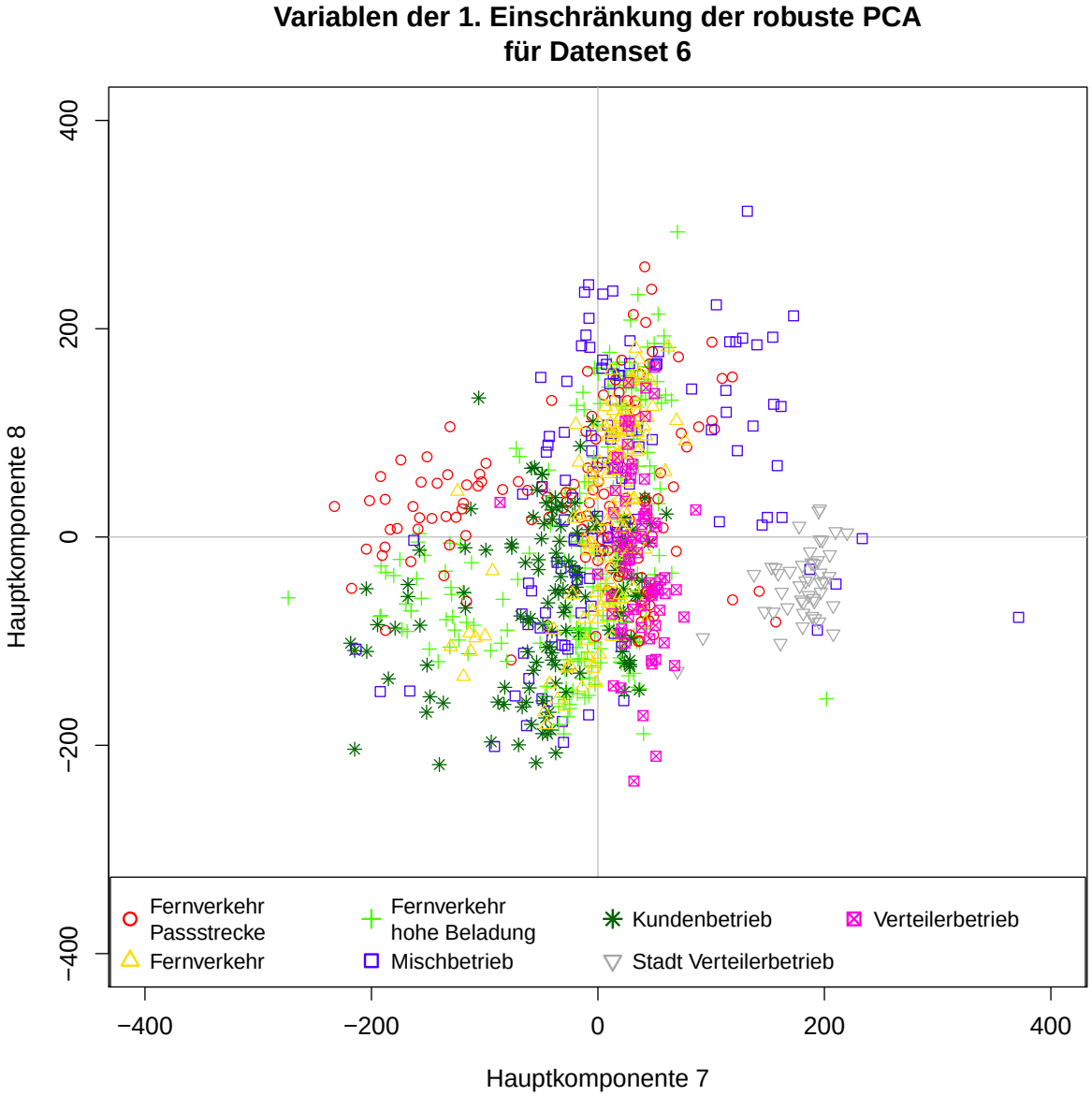


Abbildung 4.11: 1. und 2. Hauptkomponente der 1. Einschränkung für Datenset 6

Abbildung 4.11 zeigt das Datenset 6 mit den Hauptkomponenten 7 und 8. Die 7. Hauptkomponente hatte hohe positive Ladungen bei *EngineStopsQ90* und *EngineStopsQ50* und hohe negative Ladungen bei *TorquePosQ99* und *EngineStopsQ10*. Man kann hier behaupten, die 7. Hauptkomponente trennt zwischen kurzen und langen Motorstopps. Die Variablen *EngineStopsQ10* und *EngineStopsQ50* wurden in Tabelle 4.1 markiert, da diese über die zehn Datensets hinweg häufig in dieser Hauptkomponente vorkamen. Die 8. Hauptkomponente hatte hohe positive Ladungen bei *EngineSpeedQ05* und *VehicleStopsHour*, und hohe negative Ladungen bei *VehicleStopsQ90* und *CoolingWaterTempInQ01*. In der 8. Hauptkomponente kommt nur die Variable *VehicleStopsQ90* vor, die auch in Tabelle 4.1 markiert wurde.

Es zeigte sich, dass je höher der numerische Wert der Hauptkomponente ist, desto höher ist auch das Rauschen unter den Datensets bezüglich der Variablen. Zum Beispiel ist das Rauschen unter den Variablen der 8. Hauptkomponente höher als unter den Variablen der 1. Hauptkomponente. Die Variable *CoolingWaterTempInQ01* kam unter den ersten 8 Hauptkomponenten nur 4 mal vor und hat nur in der 8. Hauptkomponente überhaupt einen Erklärungswert. Diese Variable wurde in der Variablenauflistung auf Platz 67 von 76 gereiht und hat nach dieser Ordnung keinen Einfluss. Jedoch zeigt dieses Beispiel, dass eine genaue Begutachtung der Ladungen der einzelnen Datensets nötig ist, um diese einzelnen Sonderfälle zu berücksichtigen. Abgesehen von Einzelfällen funktioniert die Reihung mit den gewichteten Ladungen sehr gut.

Die 7. Hauptkomponente von Abbildung 4.11 trennt den *Stadt Verteilerbetrieb* gut von den anderen Zyklen. Um dem rechten unteren Quadranten zugeordnet zu werden, muss ein Zyklus lange bis mittlere Motorstopps, eine geringe Kühlwassertemperatur und lange Fahrzeugstopps besitzen, was durchaus auf einen *Stadt Verteilerbetrieb* zutrifft. Im linken oberen Quadranten hat sich eine Gruppe der Betriebsart *Fernverkehr Passstrecke* gebildet. Punkte, die in diesem Quadranten liegen, zeichnen sich durch hohe positive Drehmomentspitzen, viele kurze Motorstopps, ein hohes Geschwindigkeitsniveau im 5% Quantil und häufige Fahrzeugstopps pro Stunde aus. Einige Zyklen der Fernverkehr Passstrecke sind auf der postiven Seite der 7. Hauptkomponente, dies ist wieder darauf zurückzuführen, dass die Betriebsarten mehrere Fahrzeuge beinhalten. Zusammenfassend lässt sich zu den Abbildungen der Hauptkomponenten sagen, dass die Trennungen auch real sinnvoll erscheinen.

Tabelle 4.2 zeigt die ersten 8 Variablen aufgrund der Reihung mit Ladungen und erklärter Varianz nach der robust and sparse Methode. Die ersten 8 Variablen unterscheiden sich nur in der Variable *PowerPosQ95*, die die Variable *EngineStopsQ10* aus der robusten Methode ersetzt. Betrachtet man die Spalten einzeln, fällt auf, dass den Hauptkomponenten im Vergleich zur robusten Methode weniger relevante Variablen zugeordnet wurden. Dies ist eine Folge der dünn besetzten Ladungsmatrix. Hervorzuheben ist die Variable *PowerPosQ95*, die in keiner der acht Hauptkomponenten relevanten Einfluss hat. Hier tritt wieder der Fall ein, dass diese Variable zwar in vielen Komponenten vertreten ist, aber nie einen sehr starken Einfluss hat. Jedoch hat die Variable *SPEEDDiffPos10SecQ90*, die an die neunte Stelle gereiht wurde, bei der 7. und 8. Hautkomponente Einfluss.

Robuste und sparse Methode für die 1. Einschränkung									
Nr.	Variablen	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8
1	TorquePosQ99	+ (9)	- (9)	- (9)		-	- (9)		
2	EngineStopsQ10	+	+	-		- (9)	- (9)		- (9)
3	EngineSpeedQ05	+	+	o	+	o	o	o	o (9)
4	PowerNegIQR	+							
5	EngineStopsQ50			- (9)					
6	TorqueNegIQR								
7	VehicleStopsHour	-							
8	PowerPosQ95								

Tabelle 4.2: Auswahl der 8 relevanten Variablen für die 1. Auswahl unter der Methode robust and sparse

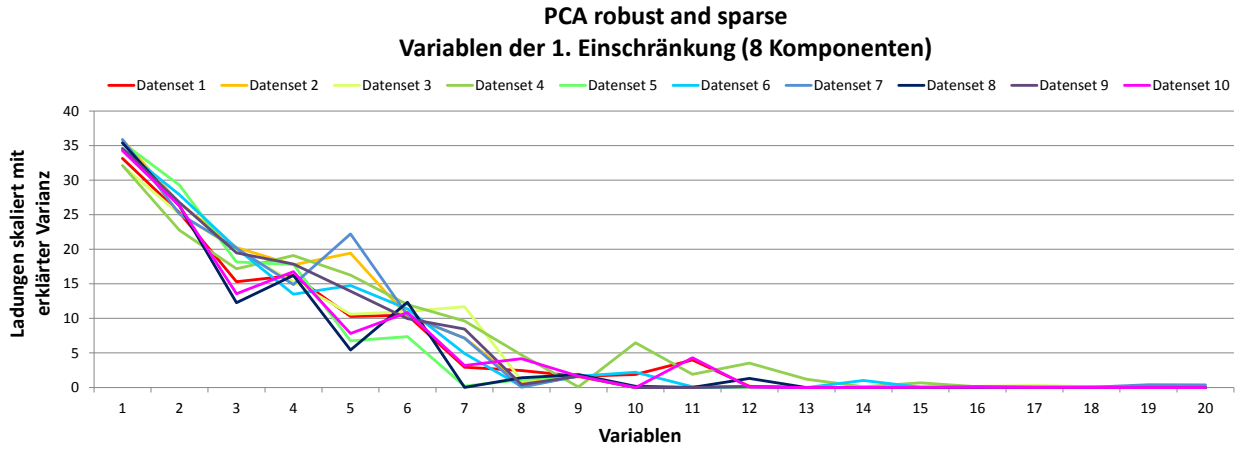


Abbildung 4.12: Absteigende Reihung der Variablen für die 1. Einschränkung und robust and sparse Berechnung

Abbildung 4.12 zeigt die gewichteten ersten 20 Ladungen für die robust and sparse Methode. Die Variablen oszillieren sehr stark zwischen 1 und 13 und fallen in diesem Intervall auch sehr schnell stark ab. Ab Variable 15 sind alle Variablen fast konstant. Abbildung 4.12 zeigt daher auch, wie klar die robust and sparse Methode die relevanten von den irrelevanten Variablen trennt. Jedoch hat die Methode bei der optischen Darstellung große Defizite, was die Interpretation anbelangt, wie bereits in Abschnitt 3.6 erwähnt. Darum werden von der robust and sparse Methode auch keine Abbildungen interpretiert.

4.3 Auswertung 2. Einschränkung

Die 2. Einschränkung beinhaltet nur noch fahrzeugs- und umgebungsbezogene Variablen, die in Summe 28 Stück ergeben. Da bei der 1. Einschränkung sehr wenige Variablen aus den fahrzeugbezogenen Variablen ausgewählt wurden, ist es interessant zu erfahren, welche speziellen Variablen nun gewählt werden, da viele Alternativen von zuvor fehlen.

Abbildung 4.13 zeigt die gewichteten Ladungen für die 2. Einschränkung. Es zeigen sich teilweise große Unterschiede in der Wahl der Variablen, zum Beispiel wird Variable 3 von den Datensets 3, 8 und 9 als weniger relevant bewertet als von den anderen Datensets. Auch die weiter hinten gereihten Variablen ab Variable 18 unterscheiden sich über die Datensets hinweg oft stark in der Relevanz.

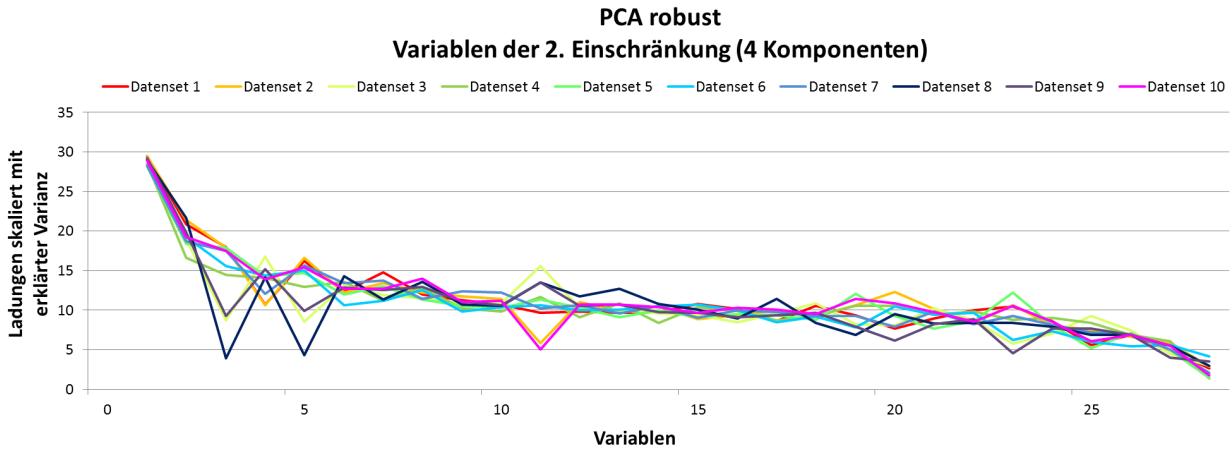


Abbildung 4.13: Absteigende Reihung der Variablen für die 2. Einschränkung und robuste Berechnung

In Tabelle 4.3 sind die ersten 4 Variablen dargestellt. Die erst gereihte Variable *PowerNegIQR* wurde auch in der 2. Einschränkung als relevant bewertet sowie die zweit gereihte Variable *VehicleStopsHour*. Die Variablen *PowerDiffNeg3SecQ50* und *VehicleStopsQ90* spielten in der 2. Einschränkung keine oder nur eine untergeordnete Rolle. Die Variable *PowerDiffPos3SecQ50* auf dem fünften Platz ist zu erwähnen, da sie mit den Hauptkomponenten 3 und 4 korreliert. Die Variable *PowerPosQ95* hat in der 4. Hauptkomponente im Negativen hohe Ladungen und die Variablen *VehicleSpeedQ75* und *VehicleSpeedQ50* haben hohe negative Ladungen in der 2. Hauptkomponente. In der 2. Einschränkung spielen Dynamiken eine größere Rolle als zuvor bei der Basisauswahl und der 1. Einschränkung.

Abbildung 4.14 zeigt die 1. und 2. Hauptkomponente des Datensets 1 für die 2. Einschränkung. Die Variablen *PowerNegIQR* und *PowerPosQ95* haben hohe positive Ladungen auf der 1. Hauptkomponente und *PowerDiffPos3SecQ50* und *VehicleStopsHour* haben hohe negative Ladungen. Hohe positive Ladungen auf der 2. Hauptkomponente haben die Variablen *VehicleStopsHour* und *PowerNegIQR*, eine hohe negative Ladung hat die Variable *VehicleStopsQ90*. Der *Stadt Verteilerbetrieb* wird durch die 1. Hauptkomponente von

den weiteren Betriebsarten gut getrennt. Auf der positiven Achse der 1. Hauptkomponente sieht man die Zyklen des *Fernverkehrs volle Beladung* und Zyklen des *Verteilerbetriebs Passstrecke*. Die Zyklen der Betriebsart *Fernverkehr* bildet hier eine kompaktere Gruppe als zuvor bei der 1. Einschränkung. Auch die Zyklen aus dem *Kundenbetrieb* streuen bei weitem nicht so stark als zuvor bei der 1. Einschränkung und befinden sich zum großen Teil im linken unteren Quadranten.

Robuste Methode für die 2. Einschränkung					
Nr.	Variablen	PC 1	PC 2	PC 3	PC 4
1	PowerNegIQR	+	+	+	+
2	VehicleStopsHour	-	+		
3	PowerDiffNeg3SecQ50				
4	VehicleStopsQ90		-		

Tabelle 4.3: Auswahl der relevanten 4 Variablen für die 2. Auswahl unter der robusten Methode

Die Zyklen im linken unteren Quadranten zeichnen sich durch viele Fahrzeugstopps pro Stunde, eine hohe mittlere Leistung innerhalb von 3 Sekunden und lange Fahrzeugstopps aus. Auch eine Gruppe der *Verteilerzyklen* befindet sich in diesem Quadranten. Im rechten unteren Quadranten befinden sich fast ausschließlich Zyklen der Betriebsart *Fernverkehr volle Beladung*. Diese Zyklen besitzen eine weite Spannbreite in der negativen Leistung, haben hohe positive Leistungspitzen und lange Fahrzeugstopps. Die Zyklen der Betriebsart *Fernverkehr volle Beladungs* im rechten oberen Quadranten zeichnen sich neben einer weiten Spannbreite in der negativen Leistung und hohen positiven Leistungsspitzen auch noch durch viele Fahrzeugstopps pro Stunde aus. Die Zyklen der *Betriebsarte Mischzyklus* zeigen Nutzungen aller Betriebsarten.

Auch in den Hauptkomponenten 3 und 4 spielt die Variable *PowerNegIQR* eine starke Rolle in den positiven Ladungen. Sie ist in beiden Komponenten die Variable mit der höchsten positiven Ladung. Die 3. Hauptkomponente hat noch eine hohe positive Ladung bei Variable *PowerDiffNeg3SecQ50* und eine hohe negative Ladung bei *PowerDiffPos3SecQ50*. Die 4. Hauptkomponente hat eine hohe positive Ladung bei *PowerDiffPos3SecQ50* und eine hohe negative Ladung bei *VehicleSpeedDiffPos10SecQ90*. Abbildung 4.15 zeigt, wie zuvor Abbildung 4.14, dass sich die *Verteilerzyklen* im Zentrum befinden. *Fernverkehr volle Beladung* streut in dieser Kombination um das Zentrum und wird durch die Komponenten 3 und 4 nicht so klar getrennt wie zuvor von 1 und 2. Der *Stadt Verteilerbetrieb* ist wieder weit weg von allen anderen Betriebsarten, streut jedoch nun auch stark. Die 3. Hauptkomponente trennt die *Verteilerzyklen*, welche in der 4. Hauptkomponente auf demselben Niveau liegen. Der *Kundenbetrieb* liegt auf der 4. Hauptkomponente auch auf einem Niveau, was auf die Variable *PowerDiffPos3SecQ50* zurückzuführen ist, da die Variable *PowerNegIQR*, außer im unteren linken Quadranten überall vorkommt. Eine Trennung der Betriebsarten durch die Hauptkomponenten ist also gegeben.

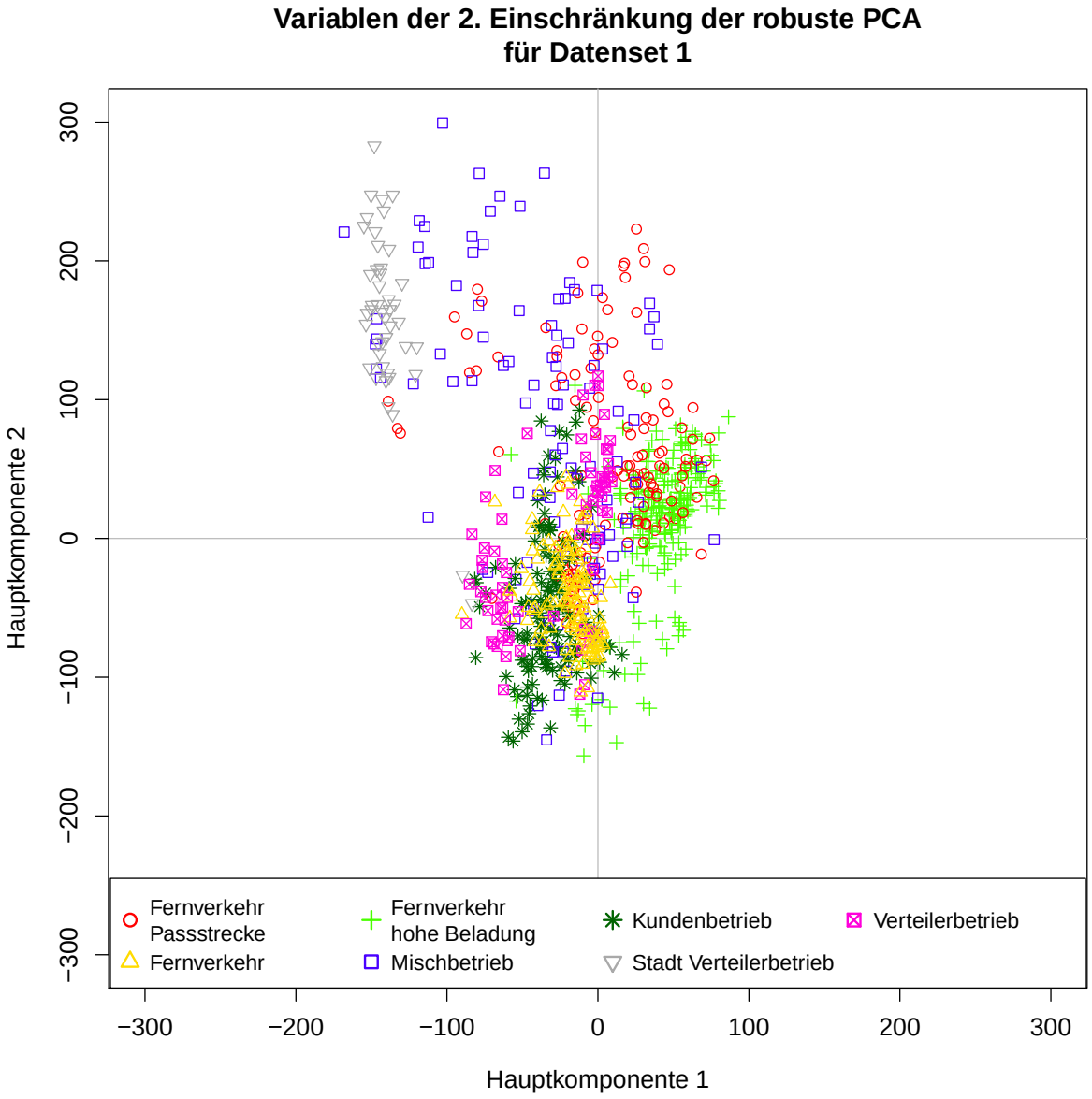


Abbildung 4.14: 1. und 2. Hauptkomponente der 2. Einschränkung für Datenset 1

Zuletzt noch ein Vergleich mit der robust and sparse Methode für die 2. Einschränkung: Hier liefert Abbildung 4.16 keinen eindeutigen Abfall der gewichteten Ladungen wie in Abbildung 4.12. Die letzt gereihten Variablen rauschen sogar sehr stark, hier vor allem die Variablen 18, 21 und 25. Durch das reduzierte Datenset kommen hier die Auswirkungen der Stichprobenziehung stärker zum Vorschein.

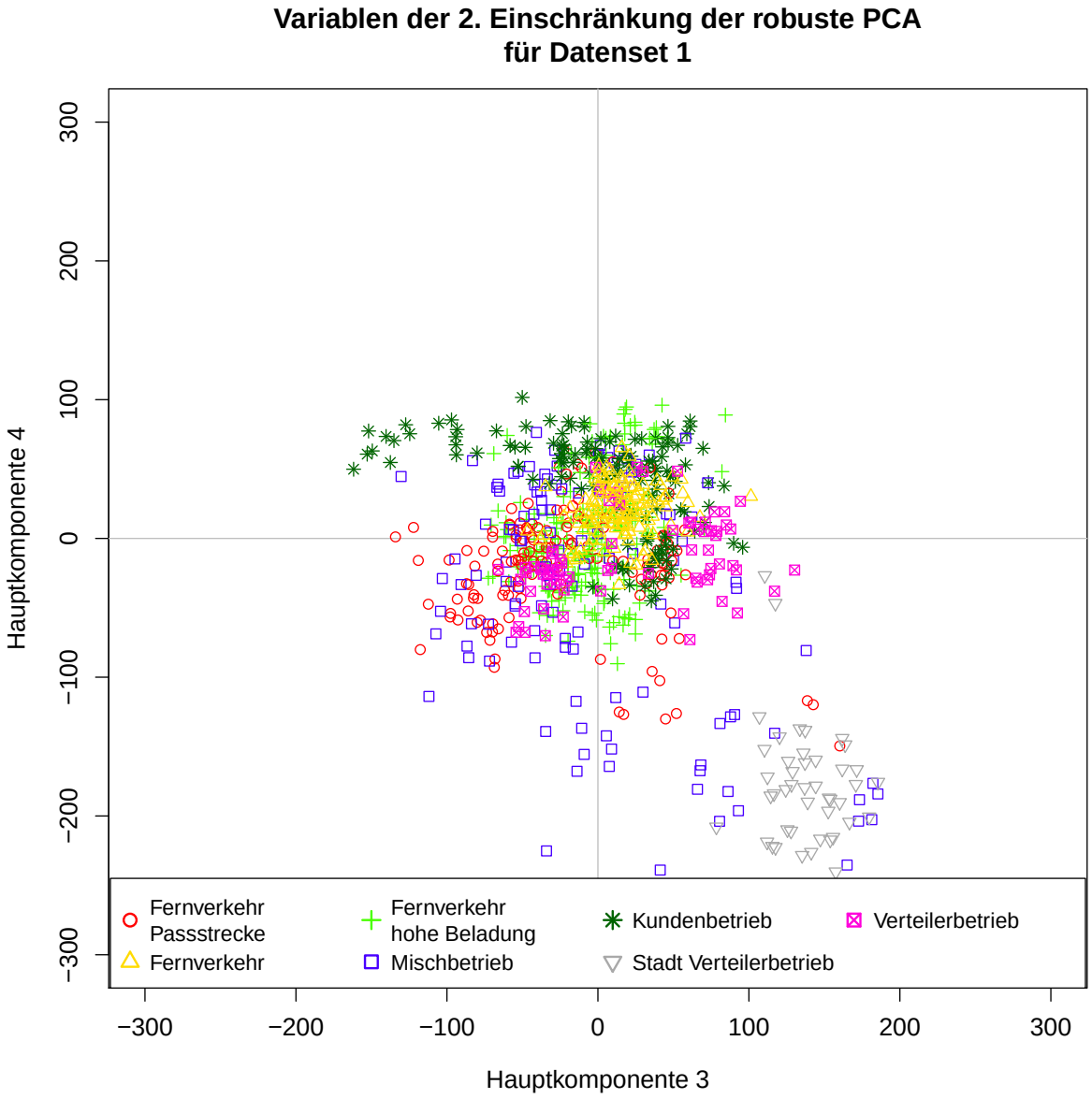


Abbildung 4.15: 3. und 4. Hauptkomponente der 2. Einschränkung für Datenset 1

Die robust and sparse Methode für die 2. Einschränkung zeigt keine nennenswerte Reduktion mehr. Die Variable *PowerNegIQR* ist auch in dieser Methode sehr relevant, und die Variablen *VehicleStopsHour* und *PowerPosQ95* erhalten einen Eintrag in Hauptkomponente 1. Es fällt auf, dass die Relevanz der Variable *PowerNegIQR* in Hauptkomponente 3 und 4 abnimmt, was sich mit der dünn besetzten Ladungsmatrix unter der robust and sparse Methode erklären lässt. Variable *VehicleSpeedDiffPos10SecQ99* wurde nicht markiert, was wieder darauf zurück zu führen ist, dass diese Variable relativ hohe Ladungen hat, jedoch

nie richtig relevant für eine konkrete Hauptkomponente ist.

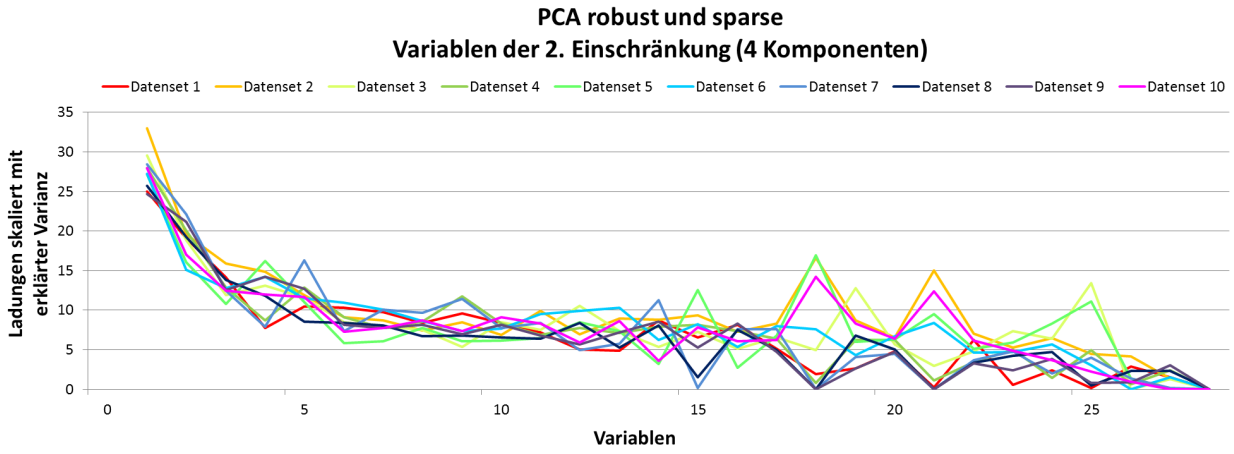


Abbildung 4.16: Absteigende Reihung der Variablen für die 2. Einschränkung und robust and sparse Berechnung

Robuste und sparse Methode für die 2. Einschränkung					
Nr.	Variablen	PC 1	PC 2	PC 3	PC 4
1	PowerNegIQR	+	+	- (8)	o (8)
2	VehicleStopsHour	-			
3	PowerPosQ95	+			
4	SPEEDDiffPos10SecQ99				

Tabelle 4.4: Auswahl der relevanten 4 Variablen für die 2. Auswahl unter der Methode robust and sparse

Zusammenfassung

Das Ziel dieser Arbeit war ein Vergleich der klassischen, robusten und robust and sparse Hauptkomponentenanalyse anhand empririscher Untersuchungen. Zuvor wurden diese beiden Methoden theoretisch dargestellt und anschließend anhand einer Auswahl realer Messdaten ausgewertet. Um die Nutzung praxisnahe zu testen, wurde dabei ein sehr großes Set an Daten definiert, das ca. 4000 Beobachtungen beinhaltete. Da die beschreibenden Faktoren in ihrer Anzahl sehr unausgeglichen vertreten waren, wurde für jede Faktorgruppe eine Menge definiert und anschließend wurden 10 Stichproben gezogen, die in Summe jeweils 888 Beobachtungen enthielten.

Um Informationen über diese Daten zu erhalten, wurden Variablenklassen mit deren untergeordneten Variablen definiert. In diesem Schritt musste bereits auf Messfehler und fehlende Werte geachtet werden. Da im ersten Schritt zahlreiche Variablen für diese Daten definiert wurden, wurde diese Anzahl anhand einer Korrelationsanalyse veringert. Für spätere Interpretationen wurden nun 3 Einschränkungen anhand der Variablenklassen definiert. Das Basisdatenset enthielt Variablen aller Variablenklassen, die 1. Einschränkung enhielt noch Variablen der Motor-, Fahrzeug- und Umgebungskenngrößen und die 2. Einschränkung enthielt nur noch Variablen der Fahrzeug- und Umgebungskenngrößen.

Mit diesen $30 = 3 \times 10$ festgelegten Datensets wurden die Hauptkomponentenanalysen klassisch, robust und robust and sparse durchgeführt. Da es zwischen der klassischen und der robust und robust and sparse Hauptkomponentenanalyse zu deutlichen Unterschieden zwischen den relevanten Daten kam, wurde ein näherer Blick auf die Skalierung der Daten geworfen. Die Daten wurden bei der klassischen Methode mit dem Mittelwert zentriert und mit der Standardabweichung skaliert und bei der robusten Methode mit dem L_1-Median zentriert und mit dem Q_n-Schätzer skaliert. Es stellte sich heraus, dass die beiden Skalierungsschätzer diese Unterschiede in den Outputs verursachten. Eine zentrale Eigenschaft des Q_n-Schätzers ist, dass sich dieser Schätzer erst bei 50% verunreinigter Daten verzerrt, die Standardabweichung jedoch schon bei einer verunreinigten Messung. Dies hatte zur Folge, dass die beiden Schätzer bei einigen Variablen weit auseinander lagen und somit auf eine gänzlich unterschiedliche Spannweite skalierten. Es stellte sich nun die Frage, wie reagiert die klassische Methode auf ein robust skaliertes Datenset und umgekehrt. Es stellte sich heraus, dass die klassische Methode angewandt auf ein robustes Datenset teilweise identische bzw. ähnliche Variablen wählte wie die robuste Methode mit robuster Skalierung. Im Vergleich ließ sich die klassische Methode mit robuster Skalierung eher von Ausreißern beeinflussen, was in der Basisauswahl und der 1. Einschränkung dazu führte, dass die Be-

deutung einer Variable mit vielen Ausreißern für das Datenset überbewertet wurde. Dieses Ergebniss konnte in der Kombination robuste Methode mit klassischer Skalierung nicht reproduziert werden, da diese bis auf wenige Ausnahmen kaum Schnittpunte mit den top Variablen der klassischen Methode mit klassischer Skalierung hatte.

Die robuste Hauptkomponentenanalyse lieferte eine Robustifizierung in allen Berechnungsschritten, da nicht nur die Definition der Variablen sondern auch die Methode und die Skalierung der Daten robust war. Dies führt zu Ergebnissen, die die eigentliche Nutzung darstellen und keine Beeinflussung von Sonderfällen beinhalten.

Für eine leichtere Interpretierbarkeit der Ergebnisse wurde die robust and sparse Methode getestet. Die robust and sparse Methode sollte eine dünn besetzte Ladungsmatrix liefern, die die Zuordnung relevanter Variablen zu den Hauptkomponenten erleichtern soll. Diese Methode machte wiederholte Berechnungen von Hauptkomponentenanalysen nötig, da hier ein Tuningparameter zu optimieren war. Dieser Tuningparameter drückte die Ladungen der irrelevanten Variablen in der Ladungsmatrix gegen Null. Für die Berechnungen wurde zuerst ein sehr grobes Intervall definiert, das nach Überprüfung mit Hilfe der erklärten Varianz immer engmaschiger gewählt wurde. Die Berechnung dieser Hauptkomponentenanalysen brachte einen großen Zeitaufwand mit sich. Positiv an dieser Methode war, dass die Auswahl der Variablen wirklich ein vielfaches leichter war als bei der klassischen und robusten Methode, jedoch gibt es auch einige Nachteile. Einer der für die Praxis relevantesten Nachteile ist die lange Rechenzeit durch wiederholtes Berechnen der Hauptkomponentenanalyse. Hinzu kommt, dass die Wahl des Tuningparameter nicht ganz eindeutig ist und optische Interpretationen der Hauptkomponenten hier schwerer fallen als bei der nicht sparse Methode.

Literaturverzeichnis

[Croux, 2013] Croux, C., Filzmoser, P., and Fritz, H., Robust sparse principal component analysis, Technometrics, 2013, to appear.

[Croux, 2007] Croux, C., Filzmoser, P. and Olivera, M., Algorithms for projection-pursuit robust principal component analysis, Chemometrics and Intelligent Laboratory Systems, 87, 2007, 218-225.

[Croux, 2005] Croux, C. and Ruiz-Gasen, A., High breakdown estimators for principal components: The projection-pursuit approach revisited, Journal of Multivariate Analysis, 95, 2005, 206-226.

[Everit, 2011] Everit, B. and Hothorn, T. (2011), An Introduction to Applied Multivariate Analysis with R, Springer, New York.

[Fritz, 2012] Fritz, H., Filzmoser, P. und Croux, C., A comparison of algorithms for the multivariate L_1-median, Computational Statistics, 27, 2012, 393-410.

[Jollife, 2002] Jollife, I.T.(2002), Principal Component Analysis, 2nd. ed., Springer, New York.

[Rousseeuw, 1993] Rousseeuw, J. and Croux, Ch., Alternatives to the median absolut deviation, Journal of the American Statistical Association, 88, 424, 1993, 1273-1283.

[Todorov, 2013] Todorov, V. and Filzmoser, P., Comparing classical and robust sparse PCA, Advances in Intelligent Systems and Computing, 190, 2013, 283-291.

Anhang A

Grafiken und Tabellen

A.1 Gegenüberstellung der Methoden

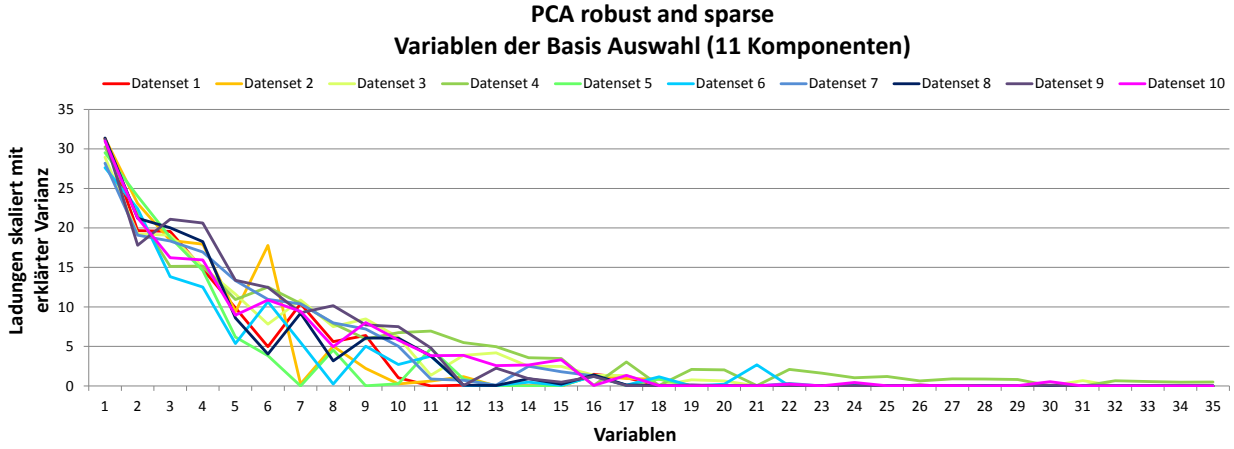


Abbildung A.1: Absteigende Reihung der Variablen für robust and sparse Methode

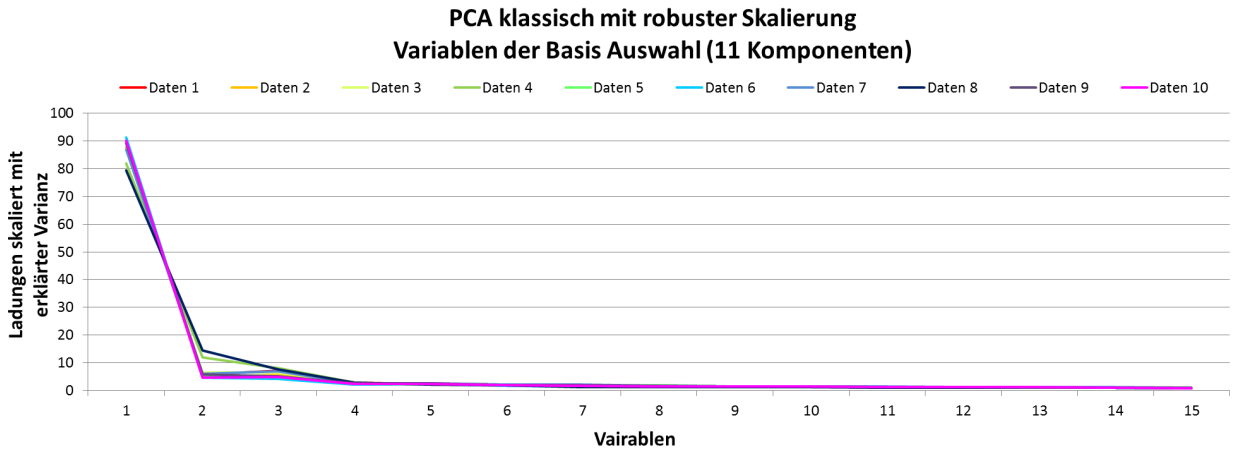


Abbildung A.2: Absteigende Reihung der Variablen für die klassische Methode mit robuster Skalierung

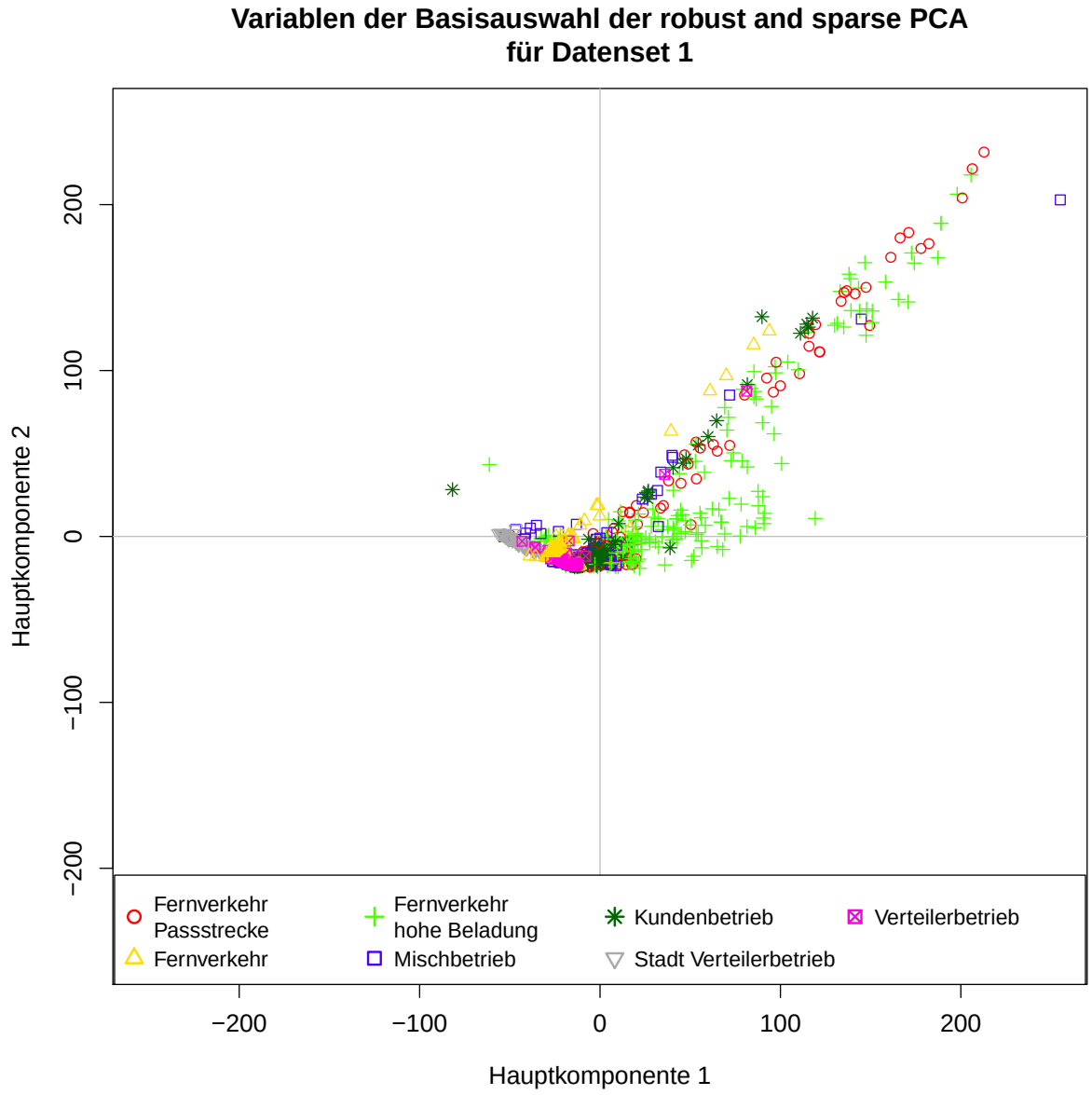


Abbildung A.3: Darstellung der ersten und zweiten Hauptkomponente für die Basisauswahl mit Methode robust and sparse

A.2 Auswertung einzelner Fahrzeug und Motorgrößen

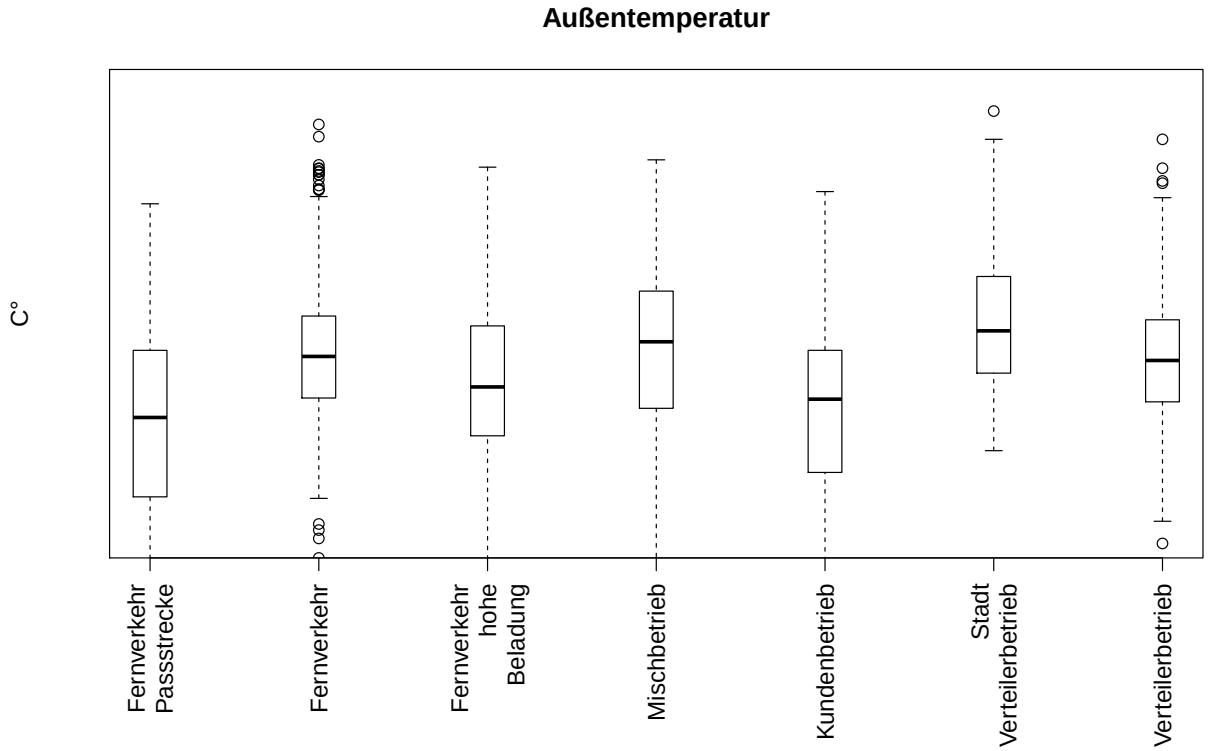


Abbildung A.4: Median der Außentemperatur für Betriebsarten

A.3 Auswertung 1. Einschränkung

Betriebsart	Anzahl der Fahrzeuge
Fernverkehr Passstrecke	7
Fernverkehr	5
Fernverkehr hohe Beladung	12
Mischbetrieb	3
Kundenbetrieb	6
Stadt Verteilerbetrieb	1
Verteilerbetrieb	3

Tabelle A.1: Auflistung der Fahrzeuge pro Betriebsart