# Analyse und Berechnung niedrigdimensionaler Darstellungen von Lösungsmengen zur nichtnegativen Matrixfaktorisierung

Habilitationsschrift

zur

Erlangung des akademischen Grades

doctor rerum naturalium habilitatus (Dr. rer. nat. habil.)

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität Rostock

vorgelegt von

Mathias Sawall

Rostock, 21. Juni 2018

https://doi.org/10.18453/rosdok_id00002437

| | |
|---|---|
| Gutachter: | Prof. Dr. K. Neymeyr, Universität Rostock<br>Prof. Dr. D. Langemann, Technische Universität Braunschweig<br>Prof. Dr. M. Maeder, University of Newcastle |
| Tag des wiss. Kolloquiums: | 01. April 2019 |

# Inhaltsverzeichnis

# Nomenklatur

Eine Matrix wird als nichtnegativ bezeichnet, wenn alle Einträge nichtnegativ sind.
Um Untermatrizen oder zusammenhängende Teile von Vektoren zu extrahieren, wird die Doppelpunktnotation [60] genutzt:

- Für $M \in \mathbb{R}^{m\times n}$ bezeichnen $M(i,:) \in \mathbb{R}^{1\times n}$ die $i$-te Zeile und $M(:,j) \in \mathbb{R}^m$ die $j$-te Spalte von $M$.
- Für $1 \leq i_1 \leq i_2 \leq m$ wird die Untermatrix von $M$, die durch das Extrahieren der Zeilen $i_1$ bis $i_2$ entsteht, mit $M(i_1 : i_2,:) \in \mathbb{R}^{(i_2-i_1+1)\times n}$ bezeichnet. Für $1 \leq j_1 \leq j_2 \leq n$ wird die Untermatrix von $M$, die durch das Extrahieren der Spalten $j_1$ bis $j_2$ entsteht, mit $M(:, j_1 : j_2) \in \mathbb{R}^{m\times(j_2-j_1+1)}$ bezeichnet. Die kombinierte Anwendung führt auf

$$M(i_1 : i_2, j_1 : j_2) = \begin{pmatrix} M_{i_1 j_1} & \dots & M_{i_1 j_2} \\ \vdots & \ddots & \vdots \\ M_{i_2 j_1} & \dots & M_{i_2 j_2} \end{pmatrix} \in \mathbb{R}^{(i_2-i_1+1)\times(j_2-j_1+1)}.$$

  Insbesondere ist $M(i,j) = M_{ij}$.
- Für einen Zeilenvektor $v \in \mathbb{R}^{1\times n}$ und $1 \leq i_1 \leq i_2 \leq n$ ist

$$v(i_1 : i_2) = (v_{i_1}, \dots, v_{i_2}) \in \mathbb{R}^{1\times(i_2-i_1+1)}$$

  und für einen Spaltenvektor $u \in \mathbb{R}^n$ und $1 \leq i_1 \leq i_2 \leq n$ ist

$$u(i_1 : i_2) = (u_{i_1}, \dots, u_{i_2})^{\mathbf{T}} \in \mathbb{R}^{i_2-i_1+1}.$$

Häufig in dieser Arbeit verwendete Symbole sind:

| | |
|---|---|
| $D$ | nichtnegative $k \times n$-Matrix, die faktorisiert werden soll, siehe (1.1), |
| $C, A$ | Faktoren einer nichtnegativen Matrixfaktorisierung, wobei $C \in \mathbb{R}^{k\times s}$, $A \in \mathbb{R}^{s\times n}$, siehe (1.5), mitunter werden auch $X$ und $Y$ als Faktoren genutzt, |
| $U\Sigma V^{\mathbf{T}}$ | abgeschnittene Singulärwertzerlegung von $D$, wobei $U \in \mathbb{R}^{k\times s}$, $\Sigma \in \mathbb{R}^{s\times s}$, $V \in \mathbb{R}^{n\times s}$, siehe (2.1), |
| $T$ | $s \times s$-Matrix zur Bestimmung von $C$ und $A$ mittels $U$, $\Sigma$ und $V$, siehe (2.2), |
| $\mathcal{M}_A$ | Menge zulässiger Lösungen in Bezug auf den zweiten Faktor einer nichtnegativen Matrixfaktorisierung, siehe (2.6), |
| $\mathcal{M}_C$ | Menge zulässiger Lösungen in Bezug auf den ersten Faktor einer nichtnegativen Matrixfaktorisierung, siehe (2.7), |
| $\mathcal{F}_A$ | Obermenge von $\mathcal{M}_A$, welche nur die Nichtnegativität der ersten Zeile des zweiten Faktors berücksichtigt, siehe (2.8), |
| $\mathcal{F}_C$ | Obermenge von $\mathcal{M}_C$, welche nur die Nichtnegativität der ersten Spalte des ersten Faktors berücksichtigt, siehe (2.9), |
| $w(:,i)$ | niedrigdimensionale Darstellungen der Zeilen von $D$, siehe (2.10), |
| $u(:,j)$ | niedrigdimensionale Darstellungen der Spalten von $D$, siehe (2.11), |
| $\mathcal{I}_A$ | konvexe Hülle der $w(:,i)$, $i = 1, \dots, k$, siehe (2.12), |
| $\mathcal{I}_C$ | konvexe Hülle der $u(:,j)$, $j = 1, \dots, n$, siehe (2.13), |
| $\mathbb{R}_+$ | Menge der nichtnegativen reellen Zahlen; $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$. |

# 1 Einleitung

In dieser Schrift wird die Uneindeutigkeit von nichtnegativen Faktorisierungen nichtnegativer Matrizen untersucht. Solche Faktorisierungen ordnen sich in die mathematischen Gebiete der multivariaten Analysis/Faktoranalyse und der Tensorfaktorisierungen ein und werden etwa für die Datenanalyse und die Signalverarbeitung angewendet. Im Sinne des *blind source separation* Problems wird mittels geeigneter Tensorfaktorisierungen das Ziel verfolgt, von hochdimensionalen überlagerten Daten auf die latenten niedrigdimensionalen Quellterme zu schließen und deren Strukturen zu entschlüsseln. Solche Zerlegungen gelingen insbesondere dann gut, wenn zusätzliche Informationen über die zugrunde liegenden Faktoren bekannt sind. Um für eine konkrete Anwendung aussagekräftige oder etwa physikalisch sinnvolle Faktorisierungen zu extrahieren, sind beispielsweise Nichtnegativitätsrestriktionen, Annahmen über sparsity (compressed sensing), Glattheit, Monotonie oder statistische Unabhängigkeit sowie Kenntnisse, inwiefern manche Quellen nur lokal Beiträge liefern, von besonderer Bedeutung. Mitunter kann auch die Konsistenz von Teilen der Faktorisierung mit zeitlichen Modellen etwa in Form eines Anfangswertproblems gewöhnlicher Differentialgleichungen gefordert werden. Die Dynamik und Tiefe des Arbeitsgebiets belegen exemplarisch das Werk von Cichocki, Zdunek, Phan, und Amari über nichtnegative Matrix- und Tensorfaktorisierungen [26] sowie der Sammelband von Dahlke, Dahmen, Griebel et al. über die Extraktion quantifizierbarer Informationen aus komplexen Systemen [29].

Konkret wird in dieser Schrift das Problem betrachtet, dass zu $D \in \mathbb{R}_+^{k\times n}$ eine Faktorisierung der Form

$$D = XY \tag{1.1}$$

mit $X \in \mathbb{R}_+^{k\times s}$ und $Y \in \mathbb{R}_+^{s\times n}$ zu kleinstmöglichem $s$ gesucht ist. Dieses $s$ wird als *nichtnegativer Rang*, $s = \operatorname{rank}_+(D)$, bezeichnet [27] und die Minimalität zielt darauf ab, triviale nichtnegative Faktorisierungen der Form $D = I_k D = D I_n$ auszuschließen. Dabei bezeichnen $I_k$ und $I_n$ Einheitsmatrizen geeigneter Dimension.

Eine nichtnegative Faktorisierung vom Typ (1.1) zu bestimmen, ist ein inverses Problem, welches keine eindeutige Lösung besitzt. Es ist somit ein *schlecht gestelltes* inverses Problem. Triviale Mehrdeutigkeiten ergeben sich durch Permutationen oder/und Umskalierungen in den Spalten des ersten und den Zeilen des zweiten Faktors: Ist $XY$ eine nichtnegative Faktorisierung von $D$ und ist $\Delta \in \mathbb{R}_+^{s\times s}$ eine *verallgemeinerte Permutationsmatrix* (Produkt aus einer Permutations- und einer Diagonalmatrix [121]) mit $\operatorname{rank}(\Delta) = s$ und $\Delta \neq I_s$, so sind $XY$ und

$$(X\Delta^{-1})(\Delta Y) = X'Y' \tag{1.2}$$

zwei verschiedene nichtnegative Faktorisierungen von $D$. Solche Formen der Mehrdeutigkeit sind in der Regel nicht von Interesse und werden durch Äquivalenzklassenbildung [143] im Folgenden nicht weiter betrachtet. Die eigentliche Schwierigkeit, dass das Faktorisierungsproblem für $s \geq 2$ im Normalfall auch unter Vernachlässigung trivialer Mehrdeutigkeiten keine eindeutige Lösung besitzt, zeigt

$$D = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 & 3 \\ 3 & 3 & 3 \end{pmatrix}, \tag{1.3}$$

wobei sich die beiden Produkte nicht mittels einer verallgemeinerten Permutationsmatrix ineinander überführen lassen.

In der vorliegenden Schrift wird das Problem der Mehrdeutigkeit der Faktorisierung systematisch untersucht. Aufgrund der Struktur der Aufgabe ist eine geschlossene Angabe der Menge möglicher Faktorenpaare für $s \geq 2$ im Allgemeinen weder in übersichtlicher Weise möglich, noch sinnvoll. Stattdessen wird der Ansatz verfolgt, die Mengen aller Spalten des ersten und aller Zeilen des zweiten Faktors zu bestimmen, die sich zu nichtnegativen Matrixfaktorisierungen ergänzen lassen. Für diese Mengen werden niedrigdimensionale Darstellungen eingeführt und *Mengen zulässiger Lösungen* definiert. Für Matrizen vom Rang $s$ sind die Mengen zulässiger Lösungen Teilmengen des $\mathbb{R}^{s-1}$. Die Idee der niedrigdimensionalen Darstellungen basiert zentral auf der Perron-Frobenius-Theorie nichtnegativer Matrizen [8, 10, 121]. Die Mengen zulässiger Lösungen werden untersucht und ihre grundlegenden Eigenschaften werden nachgewiesen. Weiter werden Methoden (geometrisch konstruktive und numerisch approximative) zur effektiven Bestimmung der Mengen zulässiger Lösungen für kleine $s$ (etwa $s = 2, 3, 4$) beschrieben und analysiert.

Das Problem der nichtnegativen Matrixfaktorisierung wird in Veröffentlichungen unter verschiedenen Aspekten untersucht [26, 36, 71, 147, 174]. In einigen Veröffentlichungen wird die nichtnegative Matrixfaktorisierung auch als *Positive matrix factorization*[1] bezeichnet [126, 177]. Als zur Klasse der *blind source separation* Probleme gehörig, ist die nichtnegative Matrixfaktorisierung artverwandt zur Unabhängigkeitsanalyse (*independent component analysis*, ICA, [28, 75, 76]) und zur Hauptkomponentenanalyse (*pricipal component analysis*, PCA, [21, 66, 84]). Weiter wird auch auf die Unterklasse der symmetrischen nichtnegativen Matrixfaktorisierung (*symmetric nonnegative matrix factorization*, SNMF, [67, 74]) sowie Methoden zur Bestimmung einzelner Faktorisierungen [95, 106, 107, 143] verwiesen. Die Verallgemeinerung der nichtnegativen Matrixfaktorisierung ist die nichtnegative Tensorfaktorisierung, wobei für einen Tensor der Ordnung $m$ Zerlegungen mit $m$ Faktoren gesucht werden (*PARAFAC model* oder auch *nonnegative tensor factorization*, NTF, [26, 39, 94, 98, 99, 125, 165, 177]).

In der vorliegenden Schrift werden insbesondere Matrizen $D$ mit $\operatorname{rank}_+(D) = \operatorname{rank}(D)$ untersucht und einige der entwickelten Methoden funktionieren nur unter dieser Voraussetzung. Oft wird diese Form der nichtnegativen Matrixfaktorisierung auch als *nonnegative rank factorization* [10, 22, 23, 61, 130, 170] oder als *Vollrangfaktorisierung* bezeichnet. Für die Faktorisierung von Matrizen mit $\operatorname{rank}_+(D) = \operatorname{rank}(D)$ ergibt sich die nützliche Besonderheit, dass sich $X$ und $Y$ mittels einer abgeschnittenen Singulärwertzerlegung konstruieren lassen [105, 112, 117, 124]. Eine solche Vereinfachung ist für Probleme mit $k, n \gg s$ effektiv. Das Problem, den nichtnegativen Rang von $D$ zu bestimmen, wird im Folgenden mit Verweis auf [27, 36, 54, 61] nicht weiter behandelt. Ein Beispiel für eine Matrix $D$ mit $\operatorname{rank}(D) < \operatorname{rank}_+(D)$ ist [27, 170]

$$D = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}. \tag{1.4}$$

Die Matrix $D$ aus (1.4) ist nichtnegativ und vom Rang drei, ihr nichtnegativer Rang beträgt vier. Weiter ist sie nicht eindeutig faktorisierbar, denn es gelten $D = I_4 D = D I_4$. Für gestörte Daten ist in der Regel keine Faktorisierung, sondern eine approximativ nichtnegative Zerlegung, die auf einer Niedrigrangapproximation beruht, gesucht. Für eine solche Aufgabe ist die Wahl eines geeigneten $s$ entscheidend, sodass eine Approximation mit sinnvollen Faktoren und akzeptablen Fehler gelingen kann.[2]

[1] Auch hier werden Faktoren mit $X_{ij} \geq 0$ und $Y_{lm} \geq 0$ gesucht.

[2] Für Messdaten gilt aufgrund von Störungen in der Regel $\operatorname{rank}(D) = \min(k, n)$ und mitunter auch $D \ngeq 0$. Für solche Fälle sind zu gegebenen $s$ approximativ nichtnegative Faktoren $X \in \mathbb{R}^{k \times s}$ und $Y \in \mathbb{R}^{s \times n}$ mit $\|D - XY\|_F^2 \to \min$ gesucht.

Das in dieser Arbeit untersuchte mathematische Problem der nichtnegativen Matrixfaktorisierung ist eng mit dem Anwendungsproblem der Zerlegung spektroskopischer Messdaten zu chemischen Reaktionssystemen in die Beiträge der Reinkomponenten verbunden. Daher sind die Analysen in der vorliegenden Arbeit in einigen Aspekten durch Fragestellungen dieses Anwendungsproblems beeinflusst. Bei der Reinkomponentenrekonstruktion sind spektroskopische Daten etwa auf einem Zeit- und Frequenzraster in Matrixform gegeben. Zu einer Datenmatrix sind die Anzahl der beteiligten Substanzen sowie deren Reinkomponentenspektren und die dazugehörigen Konzentrationsprofile gesucht. Den Zusammenhang zwischen der (zugänglichen) Matrix $D$ der überlagerten Spektren sowie den (gesuchten) Matrizen $C$ der Konzentrationsprofile und $A$ der Reinkomponentenspektren beschreibt das Gesetz von Lambert-Beer in Matrixform [90, 112, 117]

$$D = CA \tag{1.5}$$

näherungsweise/idealisiert. Die Elemente von $D$ sind nichtnegativ und wegen der (natürlichen) Nichtnegativitätsrestriktionen für beide Faktoren führt dies auf das Problem der nichtnegativen Matrixfaktorisierung. Aus diesem Kontext heraus wird mitunter auch von der *Rekonstruktion der Faktoren* gesprochen.

## Aufbau der Arbeit

Neben der Einleitung und einem Ausblick ist die vorliegende Arbeit in vier Kapitel gegliedert. In dem ersten hiervon, Kapitel 2, werden das Problem der nichtnegativen Matrixfaktorisierung vorgestellt und wesentliche Eigenschaften der Aufgabenstellung analysiert. Auf die Schwierigkeit der Mehrdeutigkeit der Faktorisierung wird eingegangen und der zentrale Ansatz, die Faktorisierung mittels einer abgeschnittenen Singulärwertzerlegung zu konstruieren, wird vorgestellt. Weiter werden die Mengen zulässiger Lösungen und, in Vorbereitung auf deren Analyse, einige wesentliche Objekte in Bezug auf diese eingeführt. Die Anwendung der nichtnegativen Matrixfaktorisierung zur Aufklärung spektroskopischer Daten wird näher erläutert. Abschließend werden vier Datensätze vorgestellt, die in dieser Arbeit zu Erläuterungen und vergleichenden Tests genutzt werden.

Kapitel 3 ist einer ausführlichen Analyse der Mengen zulässiger Lösungen gewidmet. Dazu wird zunächst untersucht, unter welchen Voraussetzungen an $D$ der niedrigdimensionale Ansatz genutzt werden kann. Anschließend werden wichtige Eigenschaften der Mengen zulässiger Lösungen unter (schwachen) Voraussetzungen nachgewiesen. Die Voraussetzungen werden gesondert analysiert und lassen sich zu nur einer nichttrivialen zusammenfassen: Das Faktorisierungsproblem darf sich nicht blockweise in untereinander unabhängige Unterfaktorisierungsprobleme aufspalten lassen. Andernfalls lassen sich diese Subsysteme, sofern sie nicht weiter zerlegbar sind, einzeln betrachten und die jeweiligen Mengen zulässiger Lösungen bestimmen. Betrachtet werden die Zusammenhänge, über die die einzelnen Spalten des ersten und die einzelnen Zeilen des zweiten Faktors durch das Dualitätsprinzip verbunden sind. In einem weiteren Abschnitt werden Untersuchungsergebnisse (Eigenschaften und geometrische Interpretation) zu den verallgemeinerten Mengen zulässiger Lösungen für Probleme mit $\operatorname{rank}(D) < \operatorname{rank}_+(D)$ vorgestellt. Abgeschlossen wird das Kapitel mit Untersuchungen zu unterschiedlichen Aspekten des Einflusses von Störungen.

In Kapitel 4 werden Algorithmen zur schnellen und präzisen Approximation der Mengen zulässiger Lösungen detailliert erläutert und kritisch analysiert. Es werden geometrisch konstruktive und numerisch approximative Methoden vorgestellt. Diese bauen auf den Untersuchungen zu verschiedenen Eigenschaften der Mengen zulässiger Lösungen des vorherigen Kapitels auf. Für die Methoden gilt es stets, geeignete Kompromisse zwischen der Approximationsgüte und dem Rechenaufwand zu finden. Ein Schwerpunkt liegt auf der Einbindung und Berücksichtigung von Störungen. Die Einbindung erfolgt über Klassifizierungsroutinen. Diese werden genutzt, um zu

entscheiden, ob eine niedrigdimensionale Darstellung als *zulässig* eingestuft wird oder nicht. Wichtig ist dabei, dass für gestörte Daten betragskleine negative Einträge in den Faktoren zugelassen werden können. Abschließend wird die Anbindung speziell berechneter, eingrenzender Lösungen an die Mengen zulässiger Lösungen untersucht.

Die herausgearbeiteten Methoden werden in Kapitel 5 an Beispieldatensätzen in Bezug auf die Approximationsgüte und den Rechenaufwand in Abhängigkeit von den jeweiligen Steuerparametern getestet und verglichen. Einige Unterroutinen der Polygon inflation Methoden werden separat und detailliert untersucht. Weiterhin werden die unterschiedlichen Ansätze zur Behandlung gestörter Daten und die Abhängigkeit der Approximationen von den Steuerparametern beleuchtet.

Das abschließende Kapitel ist einer Zusammenfassung und einem Ausblick auf geplante Forschungsaktivitäten zu noch offenen oder bislang noch nicht endgültig geklärten Fragestellungen im Umfeld der Mengen zulässiger Lösungen gewidmet.

## Bereitstellung einiger Implementierungen

Im Zuge der Forschungsarbeiten zur Mehrdeutigkeit der nichtnegativen Matrixfaktorisierung wurde vom Verfasser dieser Arbeit und Prof. K. Neymeyr um die neu entwickelten *Polygon inflation Algorithmen* das MatLab-Softwarepaket FacPack erstellt. Die Polygon inflation Algorithmen wurden als erste Methoden zur Berechnung der Mengen zulässiger Lösungen für $s = 3$ frei zugänglich gemacht. Das Softwarepaket wurde inzwischen, auch in Zusammenarbeit mit Dr. A. Moog (geb. Jürß), erweitert. Der Bezug zum Anwendungsbereich der Reinkomponentenrekonstruktion spektroskopischer Daten ist durch die Anbindung an das Leibniz Institut für Katalyse (LIKAT) in Rostock und die Evonik Industries AG über die gemeinsame Forschungsplattform Hydroformylierung begründet [42, 100–102, 148, 151–153, 158, 161].

## Danksagung

An dieser Stelle möchte ich mich bei einigen Personen bedanken, die direkt oder indirekt zur Erzielung von Forschungsergebnissen, die diese Arbeit betreffen, beigetragen haben. Vom Leibniz Institut für Katalyse (LIKAT) sind dies unter anderem Prof. A. Börner, Prof. D. Heller, Dr. C. Kubis und Dr. D. Selent. Vom Projektpartner Evonik Industries AG sind explizit Prof. R. Franke, Dr. K.-D. Wiese, Dr. A. Brächer und Dr. D. Hess zu nennen. In diesem Zusammenhang möchte ich mich auch bei Prof. R. Ludwig (Universität Rostock) und Dr. C. Fischer (ehemals LIKAT, aktuell Universität Greifswald) bedanken.

Den Mitarbeitern der Arbeitsgruppe Numerische Mathematik an der Universität Rostock, insbesondere Prof. K. Frischmuth, Dr. A. Moog und M.Sc. H. Schröder, gilt mein Dank für die vielen hilfreichen Diskussionen. Speziell möchte ich mich beim Leiter dieser Arbeitsgruppe, Herrn Prof. K. Neymeyr, für die stets konstruktiven und fruchtbaren Diskussionen und Hinweise bedanken.

Zudem möchte ich mich bei der Deutschen Forschungsgemeinschaft (DFG) für die Förderung des Forschungsprojekts (Projekttitel: Numerische Verfahren zur Berechnung von Multikomponentenzerlegungen für spektroskopische Anwendungen, Förderzeit: zweimal 3 Jahre) bedanken.

# 2 Nichtnegative Matrixfaktorisierung

In diesem Kapitel werden die im Folgenden behandelten Faktorisierungsaufgaben, der zur Konstruktion einer Faktorisierung genutzte Ansatz, die Mengen zulässiger Lösungen und wichtige Objekte für deren Analyse und Berechnung eingeführt. Das in dieser Arbeit behandelte inverse Problem der Bestimmung einer nichtnegativen Matrixfaktorisierung unter der Bedingung, dass die Faktoren vollen Ranges sein sollen, wird nach Hadamard charakterisiert. Zudem wird die Analyse spektroskopischer Daten als eine Anwendung der nichtnegativen Matrixfaktorisierung erläutert und Test- und Anwendungsdatensätze werden vorgestellt.

## 2.1 Faktorisierungsprobleme

Im Folgenden werden die drei in dieser Arbeit untersuchten Faktorisierungsaufgaben eingeführt: Die allgemeine Aufgabe der nichtnegativen Matrixfaktorisierung, die nichtnegative Vollrangfaktorisierung sowie die approximativ nichtnegative Matrixfaktorisierung, die auf einer Niedrigrangapproximation beruht. Für den störungsfreien Fall lautet das Faktorisierungsproblem:

**Faktorisierungsproblem 2.1.** *Sei $D \in \mathbb{R}_+^{k\times n}$. Gesucht ist eine Faktorisierung $D = CA$ mit $C \in \mathbb{R}_+^{k\times s}$ und $A \in \mathbb{R}_+^{s\times n}$ zu kleinstmöglichem $s$.*

Die Verbindung zwischen $D$ und $s$ ergibt sich durch den nichtnegativen Rang einer nichtnegativen Matrix [27, 54, 61]:

**Definition 2.2.** *Sei $D \in \mathbb{R}_+^{k\times n}$. Der nichtnegative Rang $r := \operatorname{rank}_+(D)$ von $D$ ist die kleinste Zahl $r \in \mathbb{N}$, sodass nichtnegative Faktoren $C \in \mathbb{R}^{k\times r}$, $A \in \mathbb{R}^{r\times n}$ mit $D = CA$ existieren.*

**Bemerkung 2.3.**

1. *Der nichtnegative Rang ist die kleinste Zahl $r$, sodass $D$ als Summe von $r$ dyadischen Produkten mit jeweils komponentenweise nichtnegativen Vektoren dargestellt werden kann.*
2. *Sei $D \in \mathbb{R}_+^{k\times n}$. Es gelten [27, 170]:*
   a) $\operatorname{rank}(D) \leq \operatorname{rank}_+(D) \leq \min(k, n)$ *und*
   b) $\operatorname{rank}(D) = \operatorname{rank}_+(D)$, *falls* $\operatorname{rank}(D) \leq 2$.

Das Faktorisierungsproblem 2.1 ist allgemein gehalten. In dieser Arbeit liegt der Fokus auf Faktorisierungen zu der Wahl $s = \operatorname{rank}(D)$, sodass die Faktoren $C$ und $A$ vollen Ranges sind. Inwiefern dies die Berechnung/Konstruktion einer Faktorisierung vereinfacht, wird in Abschnitt 2.3 deutlich. Faktorisierungen mit $s > \operatorname{rank}(D)$ werden gesondert in Abschnitt 3.7 untersucht.

Die Aufgabe der nichtnegativen Vollrangfaktorisierung lautet:

**Faktorisierungsproblem 2.4.** *Sei $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix vom Rang $s$. Gesucht ist eine Faktorisierung $D = CA$ mit nichtnegativen Faktoren $C \in \mathbb{R}^{k\times s}$ und $A \in \mathbb{R}^{s\times n}$.*

In dieser Arbeit wird für die Analyse der in Abschnitt 2.4 eingeführten Mengen zulässiger Lösungen in der Regel die idealisierte, weil störungsfreie, Faktorisierungsaufgabe 2.4 betrachtet. Unter der Berücksichtigung von Störungen und der Vorgabe der Anzahl der Spalten in $C$ und der Zeilen in $A$ lautet die Faktorisierungsaufgabe:

**Faktorisierungsproblem 2.5.** *Sei $D \in \mathbb{R}^{k\times n}$ mit Einträgen, die nichtnegativ oder betragsklein negativ im Sinne von $D_{ij} \gg -\max_{l,\ell} |D_{l\ell}|$, $i = 1,\dots,k$, $j = 1\dots,n$, sind. Weiter seien $s \in \mathbb{N}$ mit $s \leq \operatorname{rank}(D)$ und $\varepsilon \geq 0$ gegeben. Gesucht sind Faktoren $C \in \mathbb{R}^{k\times s}$ und $A \in \mathbb{R}^{s\times n}$ mit $\min_{i=1,\dots,k} C_{i\ell} / \max_{i=1,\dots,k} |C_{i\ell}| \geq -\varepsilon$ und $\min_{j=1,\dots,n} A_{\ell j} / \max_{j=1,\dots,n} |A_{\ell j}| \geq -\varepsilon$ für $\ell = 1,\dots,s$, sodass $\|D - CA\|_F^2$ minimal wird.*

## 2.2 Charakterisierung

Die Faktorisierungsaufgabe 2.4 ist ein inverses Problem, von dem es zunächst zu klären gilt, ob es korrekt oder schlecht gestellt ist. Sei zu einer Operatorgleichung

$$F(x) = y, \quad x \in \mathcal{X},\ y \in \mathcal{Y},$$

das inverse Problem betrachtet, dass zu einem gegebenen $y \in \mathcal{Y}$ ein $x \in \mathcal{X}$ mit $F(x) = y$ gesucht wird. Das inverse Problem heißt nach Hadamard *korrekt gestellt* [40, 65, 72], falls

1. es zu jedem $y \in \mathcal{Y}$ eine Lösung $x \in \mathcal{X}$ mit $F(x) = y$ gibt (Existenzbedingung),
2. die aus $F(x) = y$ erhaltene Lösung $x$ in $\mathcal{X}$ eindeutig bestimmt ist (Eindeutigkeit) und
3. die Lösung $x$ stetig von der rechten Seite $y$ abhängt (Stabilitätsbedingung).

Ist eines der drei Kriterien nicht erfüllt, so heißt die Aufgabe *schlecht gestellt*. Dabei ist bezüglich des dritten Kriteriums die genutzte Topologie zu spezifizieren [40].

In Bezug auf die Faktorisierungsaufgabe 2.4 ist ein Faktorenpaar $(C, A)$ eine Lösung. Für die Charakterisierung des inversen Problems nach Hadamard wird im Folgenden diesbezüglich mitunter von Lösungen gesprochen. Eine solche Lösung (Faktorenpaar) steht mit den später eingeführten Mengen zulässiger Lösungen nicht direkt in Verbindung.

### 2.2.1 Analyse der Existenzbedingung

Bei der Überprüfung der ersten Bedingung an ein korrekt gestelltes inverses Problem entsteht sofort ein Bezug zum nichtnegativen Rang einer nichtnegativen Matrix $D$. Für ein Beispiel mit $\operatorname{rank}(D) < \operatorname{rank}_+(D)$ müssen nach Bemerkung 2.3 die Matrixdimensionen und der Rang als $n, k \geq 4$ sowie $\min(k, n) > \operatorname{rank}(D) \geq 3$ gewählt sein. Das klassische Beispiel [27, 170] für einen Fall mit $\operatorname{rank}(D) = 3 < 4 = \operatorname{rank}_+(D)$ ist die Matrix $D$ aus (1.4). Somit ist für die Faktorisierungsaufgabe 2.4 bereits die erste Bedingung an ein korrekt gestelltes inverses Problem nicht erfüllt, da es für ein allgemeines $D \in \mathbb{R}_+^{k\times n}$ keine Vollrangfaktorisierung geben muss.[1]

Matrizen, die keine nichtnegativen Faktorisierungen mit Faktoren vollen Ranges besitzen, sind in Bezug auf die Berechnung solcher Faktorisierungen und der Analyse deren Mehrdeutigkeit nicht von Interesse. Daher werden in dieser Schrift (mit Ausnahme von Abschnitt 3.7) nur Matrizen untersucht, für die es eine Faktorisierung mit Faktoren vollen Ranges gibt. Es werden also nur Matrizen $D$ mit $\operatorname{rank}(D) = \operatorname{rank}_+(D)$ behandelt.

### 2.2.2 Analyse der Eindeutigkeitsbedingung

In Bezug auf die zweite Bedingung an ein korrekt gestelltes inverses Problem fällt zunächst auf, dass es bei der Bestimmung von $C$ und $A$ triviale Mehrdeutigkeiten der Form (1.2) gibt. Um diese außer Acht zu lassen, wird eine Äquivalenzrelation eingeführt [143]: Zwei Faktorisierungen $CA$ und $C'A'$ sind äquivalent, falls es eine verallgemeinerte Permutationsmatrix $\Delta \in \mathbb{R}^{s\times s}$

[1] Inwiefern die Existenz mittels simplizialen Kegeln geklärt kann, wird in [23, 170] erläutert.

mit $C' = C\Delta^{-1}$ und $A' = \Delta A$ gibt. Somit werden alle Faktorisierungen, die sich nur um Umskalierungen- und/oder Umsortierungen in den Spalten von $C$ und den Zeilen von $A$ unterscheiden, in eine Äquivalenzklasse eingeteilt. Jedoch besitzt die Faktorisierungsaufgabe 2.4 auch unter Vernachlässigung trivialer Mehrdeutigkeiten in der Regel keine eindeutige Lösung.[2] In (1.3) ist ein einfaches Beispiel für eine Matrix vorgestellt, welche gemäß Faktorisierungsaufgabe 2.4 auch bei einer Einteilung in Äquivalenzklassen nicht eindeutig faktorisierbar ist. Somit ist Faktorisierungsaufgabe 2.4 ein schlecht gestelltes inverses Problem.

## 2.3 Berechnung einer Faktorisierung

Die klassischen Zugänge zur Bestimmung nichtnegativer Matrixfaktorisierungen lassen sich in drei Typklassen unterteilen: multiplikative Korrekturformeln, Gradientenverfahren und alternierende kleinste Quadrate Methoden [11, 93, 95, 96, 107, 109, 110, 140, 141]. Diese Methoden sind allgemein zur Berechnung von nichtnegativen Matrixfaktorisierungen beziehungsweise von nichtnegativen Niedrigrangapproximationen ausgelegt.

In dieser Arbeit wird ein anderer Ansatz zur Bestimmung einer Faktorisierung verfolgt. Die Faktorisierungsaufgabe 2.4 ist ein Spezialfall der nichtnegativen Matrixfaktorisierung (Faktorisierungsproblem 2.1). Da $\operatorname{rank}(D) = \operatorname{rank}_+(D)$ vorausgesetzt wird, lässt sich eine Faktorisierung mittels der Faktoren einer abgeschnittenen Singulärwertzerlegung von $D$ bestimmen [17, 105, 112, 117, 124]. Dies reduziert für $\operatorname{rank}(D) \ll \min(k, n)$ den Rechenaufwand stark, falls $C$ und $A$ mittels einer Optimierung bestimmt werden.

### 2.3.1 Reduktion der Freiheitsgrade

Seien $\tilde{U} \in \mathbb{R}^{k\times k}$, $\tilde{\Sigma} \in \mathbb{R}^{k\times n}$ und $\tilde{V} \in \mathbb{R}^{n\times n}$ die Faktoren einer Singulärwertzerlegung [60] von $D$ und $s = \operatorname{rank}(D)$. Dabei sind $\tilde{U}$ und $\tilde{V}$ orthogonal und für $\Sigma$ gilt $\Sigma_{ii} = \sigma_i$, $i = 1, \ldots, \min(k, n)$, und $\Sigma_{ij} = 0$ für $i = 1, \ldots, k$, $j = 1, \ldots, n$, $i \neq j$ mit $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_s > 0$ und $\sigma_{s+1} = \cdots = \sigma_{\min(k,n)} = 0$ den Singulärwerten von $D$. Zu $s$ seien

$$U = \tilde{U}(:, 1:s), \quad \Sigma = \tilde{\Sigma}(1:s, 1:s), \quad V = \tilde{V}(:, 1:s) \tag{2.1}$$

die Faktoren einer abgeschnittenen Singulärwertzerlegung von $D$. Die Faktoren $C$ und $A$ einer Faktorisierung von $D$ lassen sich mittels einer regulären Matrix $T \in \mathbb{R}^{s\times s}$ als

$$C = U\Sigma T^{-1}, \qquad A = TV^{\mathbf{T}} \tag{2.2}$$

bestimmen [105, 124]. Enthielten vorher die Matrizen $C$ und $A$ zusammen $(k+n)s$ Freiheitsgrade, so sind es unter Nutzung des Ansatzes aus (2.2) nur $s^2$ und somit in der Regel deutlich weniger. Für eine störungsbehaftete Matrix $D$ werden $C$ und $A$ ebenfalls wie in (2.2) konstruiert, wobei $s$ sinnvoll zu wählen ist ($s \leq \operatorname{rank}(D)$, in der Regel $s \ll \operatorname{rank}(D)$).

Später wird im Zuge der Vorstellung der Mengen zulässiger Lösungen eine weitere Reduktion der Freiheitsgrade vorgenommen, indem eine spezielle Skalierung der Faktoren festgelegt wird. Für die Bestimmung eines Faktorenpaars reduziert sich die Anzahl der Parameter auf $s(s-1)$. Diese zusätzliche Möglichkeit basiert auf schwachen Voraussetzungen, die an $D$ gestellt werden.

Die Konstruktion einer Faktorisierung mittels den Faktoren einer abgeschnittenen Singulärwertzerlegung ist auch im Hinblick auf gestörte Daten und Faktorisierungsaufgabe 2.5 sinnvoll. Denn

[2] Triviale Ausnahmefälle sind zum Beispiel Matrizen $D$, welche sich mittels Zeilen- und Spaltenpermutationen auf die Gestalt bringen lassen, dass oben links eine reguläre $s \times s$-Diagonalmatrix mit $s = \operatorname{rank}(D)$ steht; siehe auch [23, 36, 37, 103, 137, 143] im Kontext eindeutiger Faktorisierungen.

nach Eckart und Young [38], siehe auch [59, 166], gilt für ein $\widetilde{D} \in \mathbb{R}^{k \times n}$ und $r \leq \operatorname{rank}(\widetilde{D})$

$$\min_{B \in \mathbb{R}^{k \times n},\, \operatorname{rank}(B)=r} \left\| \widetilde{D} - B \right\|_F^2 = \left\| \widetilde{D} - (\tilde{U}\tilde{\Sigma}(:, 1:r))(\tilde{V}(:, 1:r))^{\mathbf{T}} \right\|_F^2 = \sum_{i=r+1}^{\min(k,n)} \sigma_i^2$$

mit $\tilde{U}$, $\tilde{\Sigma}$ und $\tilde{V}$ den Faktoren einer Singulärwertzerlegung von $\widetilde{D}$. Die abgeschnittene Singulärwertzerlegung $(\tilde{U}\tilde{\Sigma}(:, 1:r))\tilde{V}(:, 1:r)^{\mathbf{T}}$ ist also unter allen Rang-$r$-Matrizen im Sinne der Frobeniusnorm eine Bestapproximation an $\widetilde{D}$. Für $\sigma_r > \sigma_{r+1}$ ist die Bestapproximation eindeutig [38]. Somit führen auch die mittels (2.2) konstruierten Faktoren $C$ und $A$ auf eine Bestapproximation $CA$ an $D$ im Sinne der Frobeniusnorm.

### 2.3.2 Bestimmung von Faktorisierungen

Wird der Ansatz (2.2) zur Berechnung einer nichtnegativen Faktorisierung genutzt, so kann ein konkretes $T \in \mathbb{R}^{s \times s}$ mittels der Minimierung einer geeigneten Zielfunktion bestimmt werden [25, 31, 48, 51, 79–81, 108, 112, 117, 124, 144, 149, 171, 176]. Wird eine beliebige Faktorisierung gesucht, so setzt sich die Zielfunktion nur aus Straffunktionen zusammen. Wird eine Faktorisierung von spezieller Struktur gesucht, so werden zusätzlich Regularisierungsfunktionen eingesetzt. Gesteuert wird die Berechnung oft durch eine Gewichtung der Straf- und Regularisierungsterme. Ob eine geeignete Faktorisierung bestimmt werden kann, hängt von der Wahl und der Gewichtung der einzelnen Zielfunktionen sowie von der Leistungsstärke der eingesetzten Optimierungsroutine ab.

Als Straffunktionen fungieren solche, die zur Einhaltung der Nichtnegativitätsrestriktionen $C, A \geq 0$ sowie der Rekonstruktionsforderung $D = CA$ eingesetzt werden. Diese werden im Vergleich zu Regularisierungsfunktionen höher gewichtet. Der Einsatz von Straffunktionen ist nur eine Möglichkeit nichtnegative Faktoren zu forcieren. Oft werden bei den Faktoren negative Einträge auf null gesetzt und die Abweichung des Produkts der modifizierten Faktoren von den Ausgangsdaten fließt in die Zielfunktion ein [1, 56, 58, 173]. Eine Minimierung dieser Abweichung führt auf eine nichtnegative Faktorisierung oder Approximation. Die Regularisierungsfunktionen werden speziell gewählt und dienen dem Ziel, geeignete Faktorisierungen zu begünstigen.

Methoden zur Berechnung von (regularisierten) nichtnegativen Matrixfaktorisierungen, die nicht auf dem Ansatz der Transformation der Faktoren einer abgeschnittenen Singulärwertzerlegung beruhen, werden in [11, 93, 95, 96, 107, 109, 110, 140, 141] vorgestellt.

## 2.4 Mengen zulässiger Lösungen

Die geschlossene Angabe der Menge möglicher Faktorenpaare des Faktorisierungsproblems 2.4 ist weder in übersichtlicher Weise möglich, noch sinnvoll. Um die Mehrdeutigkeit der Faktorisierung trotzdem systematisch zu untersuchen, werden die Menge möglicher Spalten für $C$ und die Menge möglicher Zeilen für $A$ bestimmt. Es werden also alle Spalten für $C$ und alle Zeilen für $A$ gesucht, die zu nichtnegativen Matrixfaktorisierungen erweitert werden können.

### 2.4.1 Mögliche Spalten des ersten und mögliche Zeilen des zweiten Faktors

Wegen der Permutationsmehrdeutigkeit der Zeilen von $A$ und der Spalten von $C$ genügt es, nur mögliche erste Zeilen in $A$ und mögliche erste Spalten in $C$ zu bestimmen. Denn mittels einer geeigneten Permutationsmatrix $P \in \mathbb{R}^{s \times s}$ lässt sich durch $C'A' = CP^{\mathbf{T}}PA$ jede Zeile von $A$ mit

$A(1,:)$ oder jede Spalte von $C$ mit $C(:,1)$ tauschen. Die Mengen $\mathcal{A} \subset \mathbb{R}^{1\times n}$ möglicher Zeilen von $A$ und $\mathcal{C} \subset \mathbb{R}^k$ möglicher Spalten von $C$ bezüglich des Faktorisierungsproblems 2.4 sind

$$\mathcal{A} = \{a \in \mathbb{R}^{1\times n} : \exists C \in \mathbb{R}_+^{k\times s},\, A \in \mathbb{R}_+^{s\times n} \text{ mit } D = CA \text{ und } A(1,:) = a\}, \tag{2.3}$$

$$\mathcal{C} = \{c \in \mathbb{R}^k : \exists C \in \mathbb{R}_+^{k\times s},\, A \in \mathbb{R}_+^{s\times n} \text{ mit } D = CA \text{ und } C(:,1) = c\}. \tag{2.4}$$

Somit umfassen $\mathcal{A}$ alle Zeilen die im zweiten Faktor und $\mathcal{C}$ alle Spalten die im ersten Faktor einer nichtnegativen Matrixfaktorisierungen von $D$ auftreten können.

### 2.4.2 Niedrigdimensionale Darstellungen

Mittels (2.2) lässt sich die Berechnung eines Faktorenpaars auf die Bestimmung eines $T \in \mathbb{R}^{s\times s}$ vereinfachen und die Anzahl der Freiheitsgrade zur Bestimmung einer nichtnegativen Matrixfaktorisierung reduziert sich unter der Voraussetzung $\operatorname{rank}(D) = \operatorname{rank}_+(D)$ von $(n+k)s$ auf $s^2$. Wird $T$ zu einem $x \in \mathbb{R}^{s-1}$ und einem $S \in \mathbb{R}^{(s-1)\times(s-1)}$ in der Form

$$T = \begin{pmatrix} 1 & x_1 & \cdots & x_{s-1} \\ 1 & & & \\ \vdots & & S & \\ 1 & & & \end{pmatrix} \tag{2.5}$$

gewählt [1, 56, 152], so führt dies auf eine weitere Reduktion der Freiheitsgrade. Die Mengen zulässiger Lösungen ergeben sich wie folgt:

**Definition 2.6.** *Seien $D \in \mathbb{R}_+^{k\times n}$ und $s = \operatorname{rank}(D) = \operatorname{rank}_+(D) \geq 2$. Ferner seien $D^{\mathbf{T}}D$ und $DD^{\mathbf{T}}$ irreduzibel und es seien $U, \Sigma$ und $V$ die Faktoren einer abgeschnittenen Singulärwertzerlegung von $D$ mit $V(:,1) > 0$. Die Menge zulässiger Lösungen $\mathcal{M}_A$, $\mathcal{M}_A \subset \mathbb{R}^{s-1}$, wird als*

$$\mathcal{M}_A = \{x \in \mathbb{R}^{s-1} : \exists S \in \mathbb{R}^{(s-1)\times(s-1)} \textit{ mit } \operatorname{rank}(T) = s,\, U\Sigma T^{-1} \geq 0,\, TV^{\mathbf{T}} \geq 0\} \tag{2.6}$$

*definiert, wobei $T = T(x,S)$ von der Form (2.5) ist. Analog wird die Menge zulässiger Lösungen $\mathcal{M}_C$, $\mathcal{M}_C \subset \mathbb{R}^{s-1}$, als*

$$\begin{aligned} \mathcal{M}_C = \{y \in \mathbb{R}^{s-1} : \exists T \in \mathbb{R}^{s\times s} \textit{ mit } \operatorname{rank}(T) = s,\, (T^{-1})(:,1) = (1, y^{\mathbf{T}})^{\mathbf{T}}, \\ (T^{-1})(1,:) = (1,\ldots,1),\, U\Sigma T^{-1} \geq 0,\, TV^{\mathbf{T}} \geq 0\} \end{aligned} \tag{2.7}$$

*definiert, wobei $T$ nicht von der Form (2.5) ist.*

Jedes Element von $\mathcal{A}$ lässt sich bei entsprechender Skalierung niedrigdimensional durch ein $x \in \mathcal{M}_A$ darstellen und jedes Element von $\mathcal{C}$ lässt sich bei entsprechender Skalierung niedrigdimensional durch ein $y \in \mathcal{M}_C$ erzeugen. Welche Bedingungen an $D$ zu stellen sind, damit der Ansatz (2.5) gerechtfertigt und die Mengen zulässiger Lösungen definiert sind, wird in Abschnitt 3.1 näher untersucht. Als entscheidend stellt sich heraus, dass die Matrizen $D^{\mathbf{T}}D$ und $DD^{\mathbf{T}}$ irreduzibel sind. In dem Fall existiert nach dem Satz von Perron-Frobenius [8, 10, 121, 163, 172] eine Singulärwertzerlegung von $D$, bei der der erste rechtsseitige und der erste linksseitige Singulärvektor komponentenweise strikt positiv sind.

In Abbildung 2.1 wird der Ansatz dieser niedrigdimensionalen Darstellung für den in Abschnitt 2.6 vorgestellten Datensatz 2 mit $\operatorname{rank}(D) = 3$ demonstriert.

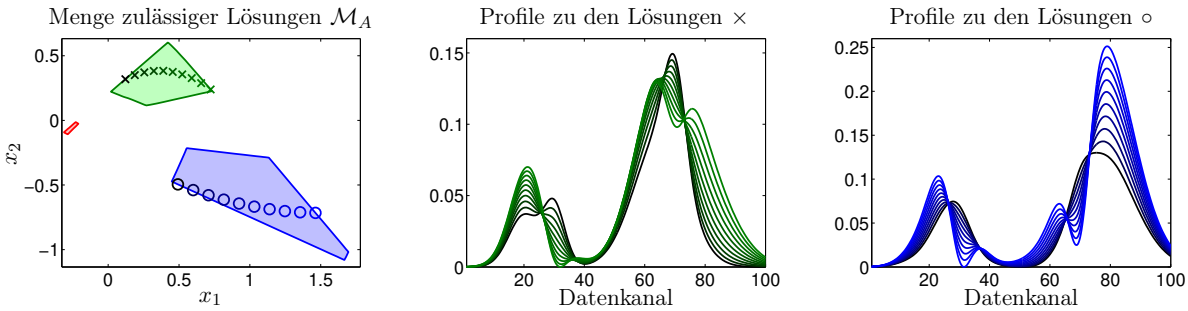


**Abbildung 2.1:** Niedrigdimensionale Darstellung möglicher Zeilen für $A$ am Beispiel des Datensatzes 2, siehe Abschnitt 2.6. Links: Dargestellt sind die in (2.6) definierte Menge zulässiger Lösungen $\mathcal{M}_A$ sowie zwei Serien zulässiger Lösungen. Mitte: Dargestellt sind die Profile (mögliche Zeilen für $A$) zu den Lösungen ×. Rechts: Dargestellt sind die Profile zu den Lösungen ∘. Die Kolorierungen stimmen jeweils überein.

**Bemerkung 2.7.**

1. *Bei der Analyse der Mengen zulässiger Lösungen und der Entwicklung von Methoden zu deren Berechnung genügt es, sich auf $\mathcal{M}_A$ zu konzentrieren. Die Eigenschaften von $\mathcal{M}_C$ sind analog zu denen von $\mathcal{M}_A$ und die Algorithmen zur Berechnung von $\mathcal{M}_A$ können nach leichten Modifikationen auch zur Bestimmung von $\mathcal{M}_C$ eingesetzt werden, denn die Menge $\mathcal{M}_C$ zu $D$ entspricht bei Berücksichtigung der Skalierungen $\sigma_i/\sigma_1$, $i = 2, \ldots, s$, beziehungsweise $\sigma_1/\sigma_i$, $i = 2, \ldots, s$, der Menge $\mathcal{M}_A$ zu $D^{\mathbf{T}}$. In Veröffentlichungen wird $\mathcal{M}_A$ oft als area of feasible solutions (kurz AFS) bezeichnet.*
2. *Die in $\mathcal{M}_A$ und $\mathcal{M}_C$ genutzten Skalierungen sind jeweils in Bezug auf den betrachteten Faktor gewählt und das in der Definition von $\mathcal{M}_C$ genutzte $T$ ist nicht von der Form (2.5). Insbesondere sind für $s \geq 2$ nicht gleichzeitig die Einträge der ersten Spalte von $T$ sowie die der ersten Zeile von $T^{-1}$ alle gleich 1. Dieser Umstand der speziellen Skalierung beeinträchtigt nicht deren später in Abschnitt 3.6 vorgestellte Verknüpfung über das Dualitätsprinzip.*

### 2.4.3 Wichtige Obermengen

Eine wichtige Obermenge von $\mathcal{M}_A$ ist

$$\mathcal{F}_A = \{x \in \mathbb{R}^{s-1} : \ (1, x^{\mathbf{T}})V^{\mathbf{T}} \geq 0\}, \tag{2.8}$$

welche unter anderem für die Anwendung geometrischer Argumente benötigt wird. Analog ist

$$\mathcal{F}_C = \left\{y \in \mathbb{R}^{s-1} : \ U\Sigma(1, y^{\mathbf{T}})^{\mathbf{T}} \geq 0\right\} \tag{2.9}$$

eine wichtige Obermenge von $\mathcal{M}_C$. Oft werden $\mathcal{F}_A$ und $\mathcal{F}_C$ auch mit *FIRPOL* (*first polygon*) oder *outer polygon* bezeichnet [16, 17, 87, 134, 137, 138].

### 2.4.4 Niedrigdimensionale Darstellungen der Daten

Zwei weitere wichtige Mengen für das geometrische Verständnis der Mengen zulässiger Lösungen sind die konvexen Hüllen der niedrigdimensionalen Darstellungen der Daten (Zeilen beziehungsweise Spalten von $D$). Die Zeilen von $D$ sind nichtnegativ und bei einer Faktorisierung $D = CA$ Linearkombinationen der Zeilen von $A$. Somit lassen sie sich in analoger Weise niedrigdimensional unter Nutzung der rechtsseitigen Singulärvektoren darstellen. Für die Untersuchung von $\mathcal{M}_C$ sind die niedrigdimensionalen Darstellungen der Spalten von $D$ entscheidend. Zu $D \in \mathbb{R}_+^{k \times n}$ und einer Singulärwertzerlegung von $D$ mit $V(:, 1) > 0$ und $U(:, 1) > 0$ sind die niedrigdimensionalen

Darstellungen $w(:,i) \in \mathbb{R}^{s-1}$, $i = 1,\ldots,k$, der Zeilen von $D$ und $u(:,j) \in \mathbb{R}^{s-1}$, $j = 1,\ldots,n$, der Spalten von $D$ als

$$w(:,i) = \frac{(D(i,:)V(:,2:s))^{\mathbf{T}}}{D(i,:)V(:,1)} = \frac{(U(i,:)\Sigma(:,2:s))^{\mathbf{T}}}{U_{i1}\sigma_1}, \tag{2.10}$$

$$u(:,j) = \frac{(\Sigma^{-1}(2:s,:))U^{\mathbf{T}}D(:,j)}{\sigma_1^{-1}(U(:,1))^{\mathbf{T}}D(:,j)} = \frac{(V(j,2:s))^{\mathbf{T}}}{V_{j1}} \tag{2.11}$$

definiert. Die Voraussetzungen, dass es eine Singulärwertzerlegung von $D$ mit $V(:,1) > 0$ und $U(:,1) > 0$ gibt, sind gerechtfertigt, sofern $D^{\mathbf{T}}D$ und $DD^{\mathbf{T}}$ irreduzibel sind, siehe Abschnitt 3.1.

Im Hinblick auf die Mengen zulässiger Lösungen spielen $w(:,i)$, $i = 1,\ldots,k$, und $u(:,j)$, $j = 1,\ldots,n$, bei der Anwendung geometrischer Argumente sowie für die Analyse des Einflusses von Störungen eine wichtige Rolle. Entscheidend sind die konvexe Hülle

$$\mathcal{I}_A = \left\{ x \in \mathbb{R}^{s-1} : \exists z \in [0,1]^k \text{ mit } \sum_{i=1}^{k} z_i = 1, \text{ sodass } x = \sum_{i=1}^{k} z_i w(:,i) \right\} \tag{2.12}$$

der $w(:,i)$, $i = 1,\ldots,k$, sowie die konvexe Hülle

$$\mathcal{I}_C = \left\{ y \in \mathbb{R}^{s-1} : \exists z \in [0,1]^n \text{ mit } \sum_{j=1}^{n} z_j = 1, \text{ sodass } y = \sum_{j=1}^{n} z_j u(:,j)^{\mathbf{T}} \right\} \tag{2.13}$$

der $u(:,j) \in \mathbb{R}^{s-1}$, $j = 1,\ldots,n$. Neben $\mathcal{F}_A$ und $\mathcal{F}_C$ sind $\mathcal{I}_A$ und $\mathcal{I}_C$ die anderen wichtigen Mengen zur Analyse und Berechnung von $\mathcal{M}_A$ und $\mathcal{M}_C$. Die Mengen $\mathcal{I}_A$ und $\mathcal{I}_C$ werden auch mit *INNPOL* (*inner polygon*) oder *inner polygon* bezeichnet [16, 17, 87, 134, 137, 138].

In Abbildung 2.2 sind $\mathcal{F}_A$ sowie die aus $w(:,i)$, $i = 1,\ldots,k$, konstruierte Menge $\mathcal{I}_A$ für den in Abschnitt 2.6 vorgestellten Datensatz 2 mit $\operatorname{rank}(D) = 3$ dargestellt.

**Bemerkung 2.8.** *Eine direkte Bestimmung der Menge $\mathcal{F}_A$ kann für große $n$ schon bei $s \in \{3,4\}$ aufwendig sein. Später wird in Abschnitt 3.6.3 unter Nutzung des Dualitätsprinzips [122, 145, 151, 156] sowie des Randes von $\mathcal{I}_C$ eine effiziente Möglichkeit zur Bestimmung von $\mathcal{F}_A$ aufgezeigt, vergleiche auch [12, 69, 133, 148].*

### 2.4.5 Festlegung zur Wahl des Richtungssinns der Singulärvektoren

Die (normierten) links- und rechtsseitigen Singulärvektoren einer Matrix $D$ sind bis auf ihren Richtungssinn eindeutig, wenn sich der zugehörige Singulärwert von allen anderen Singulärwerten unterscheidet. Die Festlegung des Richtungssinns von $V(:,i)$ entscheidet automatisch den Richtungssinn von $U(:,i)$ und andersherum. Die Achsenorientierungen der Mengen zulässiger Lösungen im $\mathbb{R}^{s-1}$ hängen von den Richtungssinnen der Singulärvektoren $U(:,i)$, $V(:,i)$, $i = 2,\ldots,s$, ab. Um für Vergleiche nicht gegebenenfalls auf unterschiedliche Richtungssinne Rücksicht nehmen zu müssen, ist eine einheitliche Vorgehensweise sinnvoll. Dazu werden die Singulärvektoren $U(:,i)$ und $V(:,i)$, $i = 2,\ldots,s$, zu Singulärwerten $\sigma_i$ mit $\sigma_i \neq \sigma_j \; \forall j \neq i$ im Folgenden so gewählt, dass $-\min V(:,i) \leq \max V(:,i)$ gelten. Für den Fall, dass $-\min V(:,i) = \max V(:,i)$ für ein $i$ gilt, wird der Richtungssinn so gewählt, dass die erste betragsmaximale Komponente von $V(:,i)$ positiv ist. Diese Vorgehensweise wird auch in FacPack angewendet.

## 2.5 Anwendungen in der Spektroskopie

Für spektroskopische Daten ist das Gesetz von Lambert-Beer [62, 90] von zentraler Bedeutung. Für eine lichtabsorbierende Probe beschreibt es den Zusammenhang zwischen ihrer Absorption

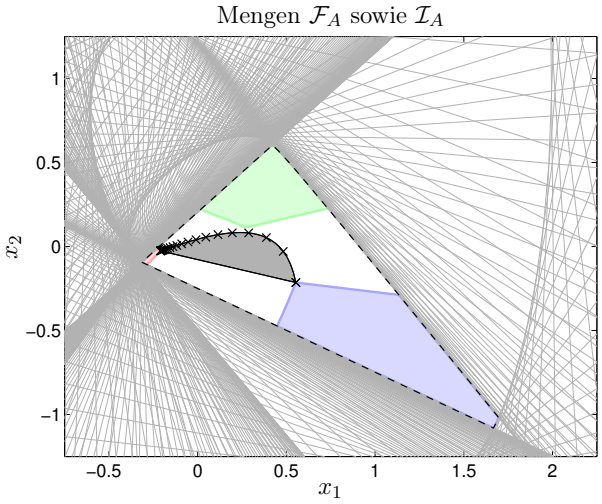


**Abbildung 2.2:** Die Mengen $\mathcal{F}_A$ und $\mathcal{I}_A$ sowie deren Konstruktion für den Datensatz 2, siehe Abschnitt 2.6, mit $\operatorname{rank}(D) = 3$. Die grauen Linien begrenzen die affinen Halbräume $x_1 V_{i2} + x_2 V_{i3} \geq -V_{i1}$, $i = 1, \ldots, n$, wobei der Nullpunkt in allen Halbräumen liegt. Die sich als Schnitt der Halbräume ergebende Menge $\mathcal{F}_A$ ist schwarz gestrichelt dargestellt. Die $w(:, i)$ sind mittels $\times$ dargestellt (aus Übersichtlichkeitsgründen sind nicht alle eingezeichnet). Deren konvexe Hülle $\mathcal{I}_A$ ist grau unterlegt. Die drei Segmente der Menge zulässiger Lösungen $\mathcal{M}_A$ sind farblich transparent dargestellt.

$\alpha$, ihrer molaren Konzentration $c$, der Extinktion $a = a(\lambda)$ des untersuchten Stoffes bei der Wellenlänge $\lambda$ sowie der Schichtdicke $l$ des Strahlendurchgangs. Die Absorption $\alpha$ ergibt sich als

$$\alpha(\lambda) = \log_{10} \frac{I_0}{I_1}$$

mit $I_0$ der Intensität der einfallenden Strahlung sowie $I_1$ der Intensität der austretenden Strahlung. Der im Gesetz von Lambert-Beer idealisiert formulierte Zusammenhang zwischen der Absorption, der molaren Konzentration, der Extinktion und der Schichtdicke lautet

$$\alpha = \alpha(\lambda) = ca(\lambda)l. \tag{2.14}$$

Für die Analyse einer Probe mit $s$ Substanzen ergibt sich die Absorption als Superposition der Einzelabsorptionen

$$\alpha = l \sum_{i=1}^{s} c_i a_i(\lambda).$$

Die Schichtdicke $l$ kann ohne Beschränkung der Allgemeinheit als $l = 1$ angenommen werden. Werden nun Absorptionen etwa auf einem Zeit- und Frequenzraster gemessen, so ergibt sich das Gesetz von Lambert-Beer in Matrixform idealisiert als $D = CA$. Dabei sind $D \in \mathbb{R}^{k \times n}$ die Matrix der Absorptionen, $C \in \mathbb{R}^{k \times s}$ die Matrix der Konzentrationsprofile und $A \in \mathbb{R}^{s \times n}$ die Matrix der Reinkomponentenspektren mit $k$ der Anzahl gemessener Spektren, $n$ der Anzahl untersuchter Frequenzen sowie $s$ der Anzahl beteiligter Komponenten. In $C$ sind die Konzentrationsprofile spaltenweise und in $A$ sind die Reinkomponentenspektren zeilenweise eingetragen. Aus physikalischen Gründen sind $C$ und $A$ nichtnegative Matrizen und $D$ ist es somit auch. Unter der Berücksichtigung von nichtlinearen Termen und von Fehlerquellen erweitert sich das Modell zu

$$D = CA + E \tag{2.15}$$

mit den in $E \in \mathbb{R}^{k \times n}$ zusammengefassten Abweichungen.

In der Anwendung ist nur die Matrix $D$ zugänglich und es ergibt sich die Aufgabe, die Anzahl der beteiligten Komponenten sowie deren Konzentrationsprofile und Reinkomponentenspektren

zu bestimmen. Für den störungsfreien Fall und unter der Annahme, dass es eine Faktorisierung mit Faktoren vollen Ranges gibt, führt dies zu der Faktorisierungsaufgabe 2.4. Unter der Berücksichtigung von Störungen und der Vorgabe von $s$ liegt die Faktorisierungsaufgabe 2.5 vor.

Die Reinkomponentenrekonstruktion ist eine Aufgabenstellung der Faktoranalyse und der Chemometrie [18, 19, 112, 117, 119, 139]. Da moderne spektroskopische Messtechniken zeitlich enge Spektrenfolgen und hohe Frequenzauflösungen ermöglichen, ergibt sich für spektroskopische Daten bei der Reinkomponentenzerlegung die Besonderheit einer Datenkompression. So sind oft die Spalten- und die Zeilendimension der Ausgangsmatrix wesentlich höher als die Spaltenanzahl des ersten und die Zeilenanzahl des zweiten Faktors der gesuchten Zerlegung. Dieser Aspekt ist für die Entwicklung effizienter Methoden zur Berechnung der niedrigdimensionalen Darstellungen der Mengen möglicher Spalten des ersten und möglicher Zeilen des zweiten Faktors entscheidend.

Für Übersichtsarbeiten zum Problem der Reinkomponentenrekonstruktion im Hinblick auf einen spektroskopischen Hintergrund und die fehlende Eindeutigkeit der Faktorisierung wird auf [49, 64, 97, 112, 115, 117, 127, 128, 147, 169] verwiesen. Generell teilen sich die Forschungsarbeiten zur Reinkomponentenrekonstruktion in zwei Hauptstränge. In dem einen geht es um die Entwicklung von Methoden zur Berechnung einzelner Faktorisierungen. Um etwa für spektroskopische Daten nicht nur eine beliebige Faktorisierung zu ermitteln, sondern eine mit bestimmten Eigenschaften, werden Zielfunktionen mit speziellen Regularisierungsfunktionen aufgestellt (in Veröffentlichungen *soft modeling* genannt) [79–81, 124, 175]. Ein spezieller Fall liegt vor, wenn zu dem Datensatz zusätzlich eine Reaktionskinetik bekannt ist (*hard modeling*). Hierbei lässt sich die Berechnung einer Faktorisierung mit der Anpassung der kinetischen Konstanten kombinieren und die Faktorisierung ergibt sich als ein Nebenresultat [30, 63, 82, 112, 144, 162]. Jedoch bleibt auch bei der Nutzung solcher spezialisierter Ansätze (soft- oder hard modeling) unter Umständen unklar, ob die berechnete Faktorisierung den Daten tatsächlich zugrunde liegt [162]. In dem anderen Hauptzweig der Forschungsarbeiten zum Faktorisierungsproblem 2.4 liegt der Fokus auf den Mengen zulässiger Lösungen.

## 2.6 Test- und Anwendungsdatensätze

Zum Abschluss dieses einführenden Kapitels werden vier Datensätze vorgestellt. Anhand dieser werden die in der vorliegenden Arbeit vorgestellten und entwickelten Methoden demonstriert und analysiert. Es handelt sich um

- einen Datensatz zur Analyse von Butiphane Liganden und Hydrieraktivität [41] mit zwei signifikant absorbierenden Komponenten,
- einen Modelldatensatz mit drei Komponenten,
- einen Datensatz zur Rhodium-katalysierten Hydroformylierung [101], wobei im untersuchten Frequenzbereich nur drei Komponenten signifikant absorbieren, sowie
- einen Modelldatensatz zu einem Vierkomponentensystem.

Zu den beiden Modelldatensätzen stehen die originalen Faktoren zur Verfügung und können zur Verifizierung der berechneten Resultate genutzt werden. Bei den beiden Datensätzen spektroskopischer Messdaten ist dies nicht der Fall. Für beide Datensätze ist jedoch je eine Kinetik bekannt, die den Daten zugrunde liegt. Die jeweiligen Kinetiken enthalten Reaktionsgeschwindigkeiten als unbekannte und anzupassende Parameter. Die Anpassung und die Berechnung der mutmaßlich korrekten Faktoren erfolgt simultan wie etwa in [30, 63, 112, 144, 162] beschrieben. Jedoch enthalten beide Kinetiken reversible Reaktionen. Solche führen für Reaktionen erster Ordnung oft zu Uneindeutigkeiten bei der Anpassung und bei der Faktorisierung [160, 162]. Für experimentelle Daten ist es auch für Kinetiken mit Beteiligung von Reaktionen höherer Ordnung unter bestimm-

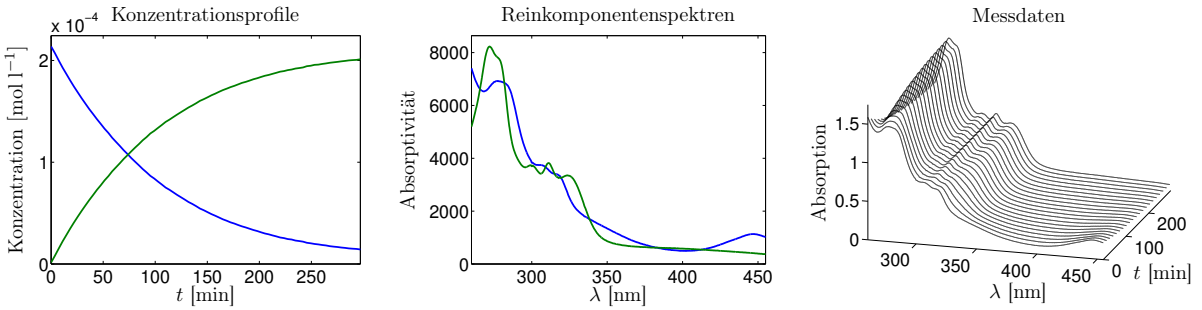


**Abbildung 2.3:** Die Konzentrationsprofile (links), die Reinkomponentenspektren (mitte) sowie die überlagerten Daten (rechts) für Datensatz 1. Der Faktor $A$ ist entsprechend $C$ und $D$ sowie einer Schichtdicke $l = 1$ skaliert. Blau: [Rh((R,R)-*i*Pr–ButiPhane)(COD)]$BF_4$, grün: [Rh((R,R)-*i*Pr–ButiPhane)($MeOH_2$)]$BF_4$.

ten Umständen, die für Datensatz 3 vorliegen, möglich, dass es keine eindeutige Faktorisierung gibt. Speziell weist die Kinetik zu Datensatz 3 unter den vorliegenden Rahmenbedingungen für zwei der drei kinetischen Parameter einen affin linearen Zusammenhang auf. Der Wert der aufgestellten Zielfunktion variiert für kinetische Parameter, die diesem Zusammenhang folgen, nur sehr wenig. Die Auswirkungen der Uneindeutigkeit von zwei Geschwindigkeitskonstanten auf die Faktoren $C$ und $A$ sind sehr gering, da die wesentlichen Unterschiede zwischen möglichen Faktorisierungen für das gewählte Zeitfenster (erste Messung bei $t = 4.73\,\text{min}$) nicht erfasst werden können.

Unter Nutzung der zugrunde liegenden Kinetiken werden für die beiden spektroskopischen Datensätze Faktorisierungen bestimmt. Diese Faktorisierungen werden für die weiteren Untersuchungen in dieser Arbeit als korrekt angenommen. Für den Datensatz 1 wird eine zusätzliche Annahme getroffen. Die bestimmten Faktorisierungen sind aus [41] und [101] bekannt.

**Datensatz 1.** *Der Datensatz wurde zur Analyse von Butiphane Liganden sowie der Hydrieraktivität und der Selektivität von Rhodiumkomplexen aufgenommen [41]. Es handelt sich um UV/Vis Spektren. Dabei wurden* $k = 82$ *Spektren im Intervall* $t \in [0, 297]\text{min}$ *zu je* $n = 1951$ *Wellenlängen im Bereich* $\lambda \in [260, 455]\text{nm}$ *gemessen. Beide Diskretisierungen sind äquidistant.*

*Für die Messdaten wird die Kinetik*

$$X \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} Y$$

*mit den Geschwindigkeitskonstanten* $k_1$ *und* $k_{-1}$ *angenommen. Die Anfangskonzentrationen sind* $c_X(0) = 2.15 \cdot 10^{-4}\,\text{mol}\,\text{l}^{-1}$ *sowie* $c_Y(0) = 0\,\text{mol}\,\text{l}^{-1}$.

*In Abbildung 2.3 sind die Profile der Faktoren* $C$ *und* $A$ *unter der Annahme* $k_{-1} = 0\,\text{min}^{-1}$ *sowie die Spektren aus* $D$ *dargestellt. Zu* $k_{-1} = 0\,\text{min}^{-1}$ *wurde, im Sinne der Anpassung eines kinetischen Modells an* $C$*, numerisch der optimale Wert* $k_1 = 9.494 \cdot 10^{-3}\,\text{min}^{-1}$ *bestimmt. Um einen Eindruck bezüglich der Störungen bei diesem Datensatz zu erhalten, sind in Abbildung 2.4 jeweils die ersten vier links- und rechtsseitigen Singulärvektoren sowie die ersten 20 Singulärwerte von* $D$ *grafisch dargestellt.*

*Die Menge zulässiger Lösungen* $\mathcal{M}_A$ *ist in den Abbildungen 4.2, 4.5 und 5.16 (jeweils rechts) dargestellt, die Menge zulässiger Lösungen* $\mathcal{M}_C$ *in den Abbildungen 4.5 und 5.16 (jeweils links).*

**Datensatz 2.** *Dem Modelldatensatz mit den drei Komponenten* $X$*,* $Y$ *und* $Z$ *liegt die Kinetik*

$$X \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} Y \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} Z$$

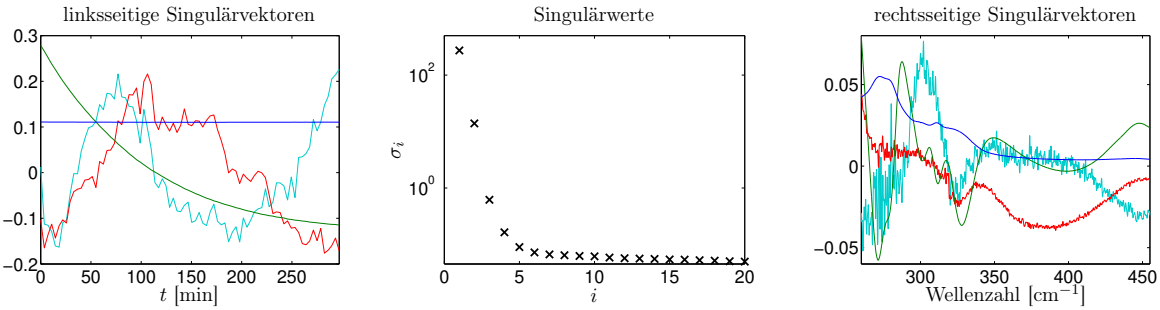


**Abbildung 2.4:** Teile der Faktoren der Singulärwertzerlegung von $D$ bezüglich des Datensatzes 1. Links: die ersten vier linksseitigen Singulärvektoren. Mitte: die ersten 20 Singulärwerte. Rechts: die ersten vier rechtsseitigen Singulärvektoren. (Farbreihenfolge aufsteigend: blau, grün, rot, türkis.)

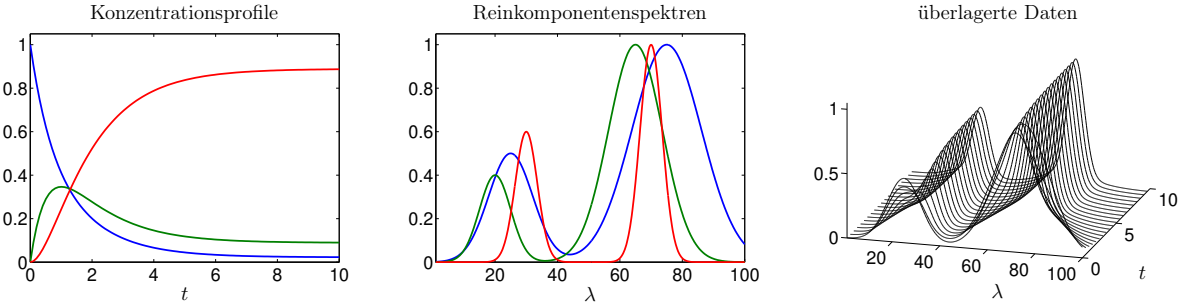


**Abbildung 2.5:** Die Konzentrationsprofile (links), die Reinkomponentenspektren (mitte) sowie die überlagerten Daten (rechts, nur jedes fünfte Spektrum eingezeichnet) für das Modellproblem aus Datensatz 2.

*mit den Geschwindigkeitskonstanten* $k_1 = 1$, $k_{-1} = 0.25$, $k_2 = 1$ *und* $k_{-2} = 0.1$ *zugrunde. Der Faktor* $C$ *wird über dem Zeitintervall* $t \in [0, 10]$ *mit* $k = 100$ *äquidistanten Stützstellen bestimmt. Die Berechnung von* $C$ *erfolgt numerisch (unter Nutzung der* MATLAB*-Routine* ode45 *mit den Einstellungen* RelTol$= 10^{-10}$ *und* AbsTol$= 10^{-10}$*) zu den Anfangswerten* $c_X(0) = 1$, $c_Y(0) = 0$ *und* $c_Z(0) = 0$.

*Die Bestimmung des Faktors* $A$ *erfolgt mittels*

$$a_X(\lambda) = 0.5 \exp\left(-\frac{(\lambda - 25)^2}{100}\right) + \exp\left(-\frac{(\lambda - 75)^2}{250}\right),$$
$$a_Y(\lambda) = 0.4 \exp\left(-\frac{(\lambda - 20)^2}{50}\right) + \exp\left(-\frac{(\lambda - 65)^2}{150}\right),$$
$$a_Z(\lambda) = 0.6 \exp\left(-\frac{(\lambda - 30)^2}{25}\right) + \exp\left(-\frac{(\lambda - 70)^2}{25}\right)$$

*sowie einer äquidistanten Diskretisierung des Intervalls* $\lambda \in [1, 100]$ *mit* $n = 400$ *Stützstellen als* $A_{1i} = a_X(\lambda_i)$, $A_{2i} = a_Y(\lambda_i)$ *und* $A_{3i} = a_Z(\lambda_i)$ *für* $i = 1, \ldots, n$. *Es ist* $D = CA \in \mathbb{R}^{100 \times 400}$. *In Abbildung 2.5 sind die Profile der Faktoren* $C$ *und* $A$ *sowie die überlagerten Daten aus* $D$ *dargestellt. Die Mengen zulässiger Lösungen sind unter anderem in den Abbildungen 2.1, 2.2 und 4.6 (jeweils* $\mathcal{M}_A$*) und 3.1, 4.7 und 5.18 (jeweils* $\mathcal{M}_C$*) dargestellt.*

**Datensatz 3.** *Der Datensatz wurde in einer Messreihe zur Analyse des Einflusses des CO-Drucks bei der Rhodium-katalysierten Hydroformylierung im Zuge der Untersuchungen zu [101] aufgenommen. Es handelt sich um FT-IR Spektren. Der hier zur Methodenanalyse genutzte Ausschnitt umfasst* $k = 1611$ *Spektren zu je* $n = 650$ *Datenkanälen im Wellenzahlenbereich* $\nu \in [1962.1, 2118.5]\,\mathrm{cm}^{-1}$. *Die vorliegenden Messungen sind im Zeitraum* $t \in [4.73, 1731]\,\mathrm{min}$ *aufgenommen, wobei* $t = 0$ *dem Reaktionsbeginn entspricht.*

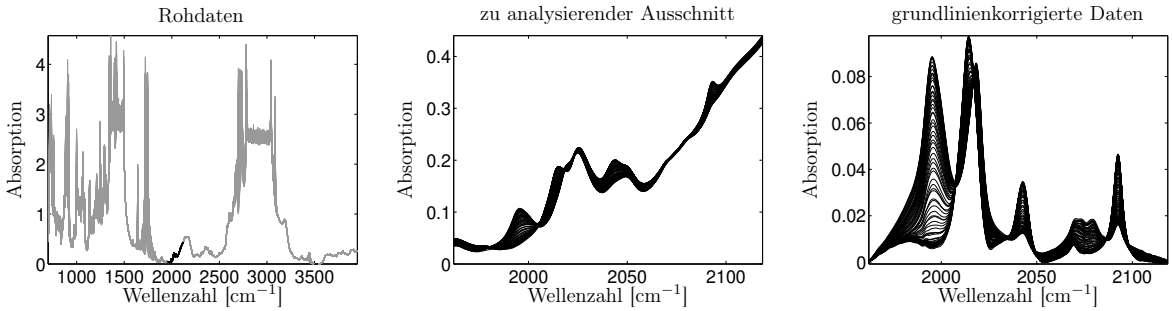


**Abbildung 2.6:** Die Rohdaten, ein Ausschnitt dieser sowie die Daten für $D$ nach Abzug des Lösemittelspektrums und einer Korrektur der Grundlinie zu Datensatz 3 (Hydroformylierung). Links: die Rohdaten (grau) sowie darin schwarz hervorgehoben der zu untersuchende Ausschnitt. Mitte: vergrößerte Darstellung des zu untersuchenden Ausschnitts. Rechts: die, nach Abzug des Lösemittelspektrums, grundlinienkorrigierten Daten, welche in den Zeilen von $D$ eingetragen sind. Es ist jeweils nur jedes 20. Spektrum abgebildet. Gut zu erkennen ist, wie klein die Absorptionen des ausgewählten und zu untersuchenden Wellenzahlenbereichs im Vergleich zu den Absorptionen der restlichen Bereiche sind.

*In Abbildung 2.6 sind die Rohdaten, deren zu analysierender Ausschnitt sowie die Spektren nach Abzug des Lösemittelspektrums und einer Korrektur der Grundlinie dargestellt. Diese Spektren sind in den Zeilen der Matrix $D$ eingetragen.*

*Den Messdaten liegt vereinfacht eine Michaelis-Menten-Kinetik*

$$S + K \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} [S-K] \xrightarrow{k_2} P + K \tag{2.16}$$

*mit den zu bestimmenden Geschwindigkeitskonstanten $k_1$, $k_{-1}$ und $k_2$ zugrunde. Die Anfangskonzentrationen sind $c_S(0) = 0.44877\,\mathrm{mol\,l^{-1}}$, $c_K(0) = 3.0091 \cdot 10^{-4}\,\mathrm{mol\,l^{-1}}$ sowie $c_{S-K}(0) = c_P(0) = 0\,\mathrm{mol\,l^{-1}}$. In dem ausgewählten Wellenzahlenbereich absorbieren nur die Komponenten $S$ (Olefin), $K$ (Hydridokomplex) und $S-K$ (Acylkomplex) signifikant, die Komponente $P$ (Aldehyd) jedoch nicht.*

*Vorteilhaft bei diesem Datensatz ist, dass das kinetische Modell aus (2.16) zur Bestimmung einer Faktorisierung genutzt werden kann [30, 63, 112, 144, 162]. Die so zugängliche Faktorisierung wird als korrekt angesehen und in dieser Arbeit zur Verifizierung von berechneten Resultaten herangezogen. In Abbildung 2.7 sind die, mittels kinetischer Modellierung bestimmten, Profile der Faktoren $C$ und $A$ sowie die Spektren aus $D$ dargestellt.*

*Die ersten vier linksseitigen und die ersten fünf rechtsseitigen Singulärvektoren sowie die ersten 20 Singulärwerte von $D$ sind in Abbildung 2.8 dargestellt und vermitteln einen Eindruck bezüglich der Störungen.*

*Die Mengen zulässiger Lösungen sind unter anderem in den Abbildungen 3.4, 5.6 und 5.7 ($\mathcal{M}_A$ und $\mathcal{M}_C$) dargestellt.*

**Bemerkung 2.9.** *Die mittels kinetischer Modellierung bestimmte Faktorisierung zum Datensatz 3 wird unter Nutzung von $z = 5$ links- und rechtsseitigen Singulärvektoren bestimmt. Diese, als korrekt angesehene, Faktorisierung kann mit nur $s = 3$ Singulärvektoren nicht dargestellt, sondern nur approximiert werden. Dies erschwert die Bestimmung der Mengen zulässiger Lösungen, für welche $s = 3$ Singulärvektoren genutzt werden.*

**Datensatz 4.** *Dem Modelldatensatz mit den vier Komponenten $W$, $X$, $Y$ und $Z$ liegt die Kinetik*

$$W \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} X \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} Y \xrightarrow{k_3} Z$$

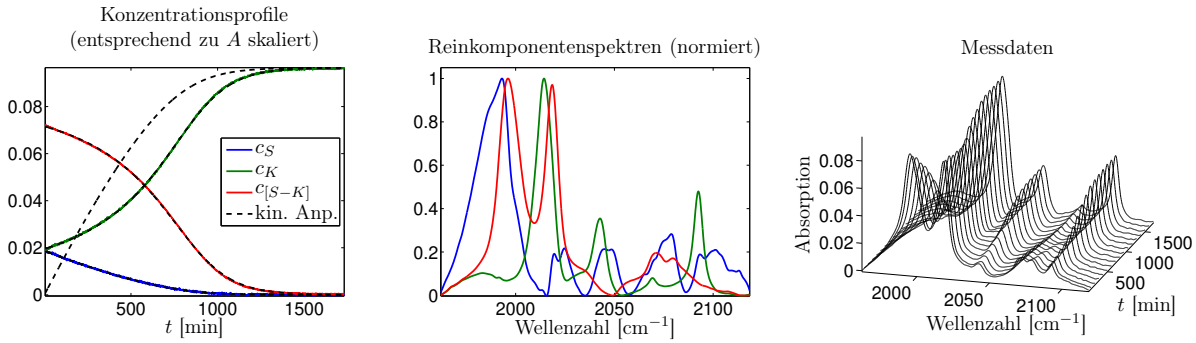


**Abbildung 2.7:** Die Konzentrationsprofile (links, farbige Linien), die Reinkomponentenspektren (mitte) sowie die Messdaten (rechts, nur jedes 100. Spektrum dargestellt) für Datensatz 3 (Hydroformylierung). Die Reinkomponentenspektren sind normiert und die Konzentrationsprofile dementsprechend skaliert. Zusätzlich sind links die Profile (inklusive Komponente $P$) der Anpassung des kinetischen Modells (schwarz gestrichelt) eingezeichnet.

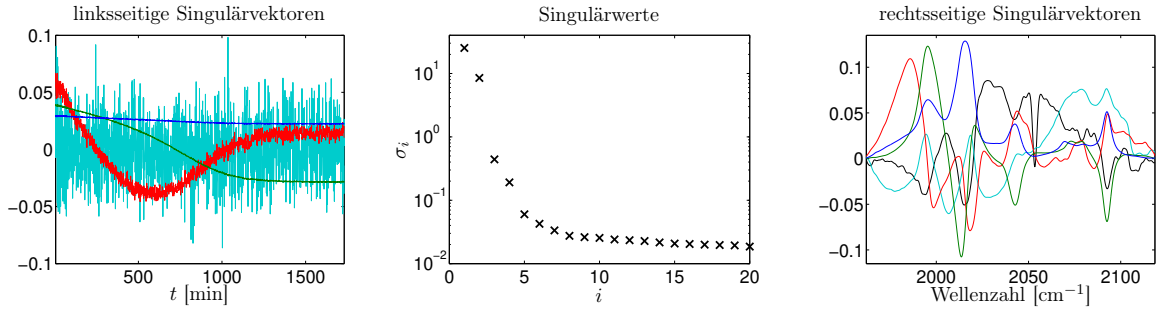


**Abbildung 2.8:** Teile der Faktoren der Singulärwertzerlegung von $D$ bezüglich des Datensatzes 3. Links: die ersten vier linksseitigen Singulärvektoren. Mitte: die ersten 20 Singulärwerte. Rechts: die ersten fünf rechtsseitigen Singulärvektoren. Der vierte Singulärwert hebt sich zwar von den folgenden ab, jedoch sind für diesen Wellenzahlausschnitt nur die drei Komponenten $S$, $K$ und $S-K$ rekonstruierbar. Der Hauptgrund für diesen zweiten signifikanten Bruch in den Singulärwerten zwischen $\sigma_4$ und $\sigma_5$ ist möglicherweise das Rangdefizit aufgrund der Michaelis-Menten-Kinetik. (Farbreihenfolge aufsteigend: blau, grün, rot, türkis, schwarz.)

*mit den Geschwindigkeitskonstanten* $k_1 = 1$, $k_{-1} = 0.75$, $k_2 = 1$, $k_{-2} = 0.25$ *und* $k_3 = 0.5$ *zugrunde. Der Faktor* $C$ *wird über dem Zeitintervall* $t \in [0, 10]$ *mit* $k = 70$ *äquidistanten Stützstellen bestimmt. Die Berechnung von* $C$ *erfolgt numerisch (unter Nutzung der* MatLab*-Routine* ode45 *mit den Einstellungen* RelTol$= 10^{-10}$ *und* AbsTol$= 10^{-10}$*) zu den Anfangswerten* $c_W(0) = 1$, $c_X(0) = c_Z(0) = c_Z(0) = 0$.

*Die Bestimmung des Faktors* $A$ *erfolgt mittels*

$$
\begin{aligned}
a_W(\lambda) &= 0.2\exp\left(-\frac{(\lambda-15)^2}{20}\right) + 0.4\exp\left(-\frac{(\lambda-45)^2}{100}\right) + \exp\left(-\frac{(\lambda-85)^2}{100}\right),\\
a_X(\lambda) &= 0.3\exp\left(-\frac{(\lambda-17.5)^2}{30}\right) + 0.3\exp\left(-\frac{(\lambda-42.5)^2}{50}\right) + \exp\left(-\frac{(\lambda-70)^2}{75}\right),\\
a_Y(\lambda) &= 0.4\exp\left(-\frac{(\lambda-20)^2}{40}\right) + 0.2\exp\left(-\frac{(\lambda-47.5)^2}{50}\right) + \exp\left(-\frac{(\lambda-80)^2}{50}\right),\\
a_Z(\lambda) &= 0.5\exp\left(-\frac{(\lambda-22.5)^2}{50}\right) + 0.25\exp\left(-\frac{(\lambda-50)^2}{100}\right) + \exp\left(-\frac{(\lambda-75)^2}{50}\right)
\end{aligned}
$$

*sowie einer äquidistanten Diskretisierung des Intervalls* $\lambda \in [1, 100]$ *mit* $n = 100$ *Stützstellen als* $A_{1i} = a_W(\lambda_i)$, $A_{2i} = a_X(\lambda_i)$, $A_{3i} = a_Y(\lambda_i)$ *und* $A_{4i} = a_Z(\lambda_i)$ *für* $i = 1, \ldots, n$. *Damit ist* $D = CA \in \mathbb{R}^{70\times 100}$. *In Abbildung 2.9 sind die Profile der Faktoren* $C$ *und* $A$ *sowie die überlagerten Daten aus* $D$ *dargestellt. Die Mengen zulässiger Lösungen sind in den Abbildungen 4.15 und 4.17, (jeweils* $\mathcal{M}_C$*) sowie 4.19 (*$\mathcal{M}_A$*, mittlere Grafik) dargestellt.*

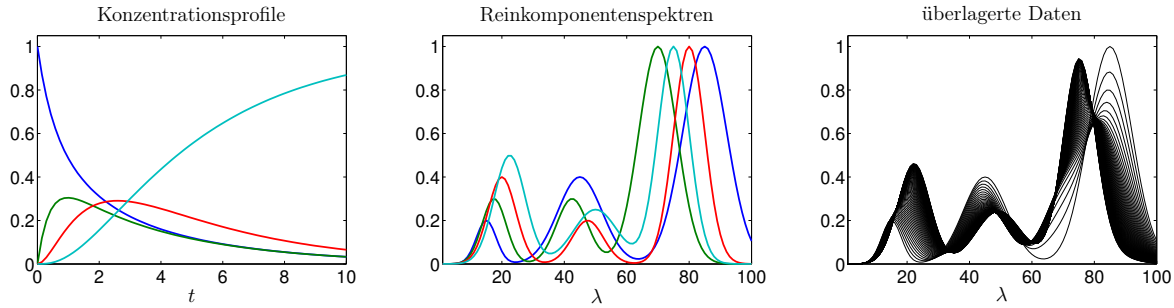


**Abbildung 2.9:** Die Konzentrationsprofile (links), die Reinkomponentenspektren (mitte) sowie die überlagerten Daten (rechts) für das Modellproblem aus Datensatz 4.

# 3 Analyse der Mengen zulässiger Lösungen

In Kapitel 2 sind die Mengen zulässiger Lösungen in Bezug auf die in Faktorisierungsaufgabe 2.4 formulierte Form der nichtnegativen Matrixfaktorisierung eingeführt. In diesem Kapitel wird die Menge zulässiger Lösungen $\mathcal{M}_A$ analysiert.[1] Im Fokus stehen Eigenschaften, die für die Entwicklung numerischer Verfahren zur Approximation der Mengen zulässiger Lösungen entscheidend sind. So wird unter anderem gezeigt, dass $\mathcal{M}_A$ unter schwachen Voraussetzungen beschränkt ist und dass die Obermenge $\mathcal{F}_A$ den Nullpunkt enthält, dieser aber nicht zu $\mathcal{M}_A$ gehört. Weiter wird für $\mathcal{M}_A$ die Eigenschaft nachgewiesen, dass der Schnitt mit einem Strahl vom Ursprung aus entweder leer oder ein Geradenabschnitt ist. Ferner wird die geometrische Bewertung eines $x$ für $s \geq 2$ gezeigt und untersucht.

Die Idee der niedrigdimensionalen Darstellung beruht auf dem Ansatz, nur einen Faktor (hier $A$) und von diesem nur eine Zeile/Spalte (die erste) zu betrachten und die spezielle Skalierung aus (2.5) zu nutzen. In Abschnitt 3.1 wird untersucht, unter welchen Voraussetzungen diese Skalierung (basierend auf dem Satz von Perron-Frobenius) genutzt werden kann.

Anschließend werden in den Abschnitten 3.2, 3.3 und 3.4 wichtige Eigenschaften von $\mathcal{M}_A$ untersucht und nachgewiesen. Dabei stehen in Abschnitt 3.2 allgemeine Eigenschaften im Vordergrund, der Abschnitt 3.3 ist dem geometrischen Zusammenhang zwischen $\mathcal{M}_A$, $\mathcal{F}_A$ und $\mathcal{I}_A$ gewidmet und unter dessen Nutzung wird anschließend in Abschnitt 3.4 die Kompaktheit von $\mathcal{M}_A$ unter schwachen Voraussetzungen nachgewiesen.

Eine wichtige Voraussetzung für den Ansatz aus (2.5) ist die Irreduzibilität der Matrix $D^{\mathbf{T}}D$. Die unter schwachen Voraussetzungen geltende Äquivalenz der Irreduzibilität von $D^{\mathbf{T}}D$ und der Irreduzibilität von $DD^{\mathbf{T}}$ wird in Abschnitt 3.5 in Zusammenhang mit der Aufspaltung des Faktorisierungsproblems in eine Blockstruktur thematisiert.

In Abschnitt 3.6 wird die Situation angenommen, dass einzelne Teile der Faktoren $C$ und/oder $A$ (einzelne Spalten von $C$ und/oder einzelne Zeilen von $A$) a-priori bekannt sind. Dazu werden Möglichkeiten untersucht, wie sich die Mengen $\mathcal{M}_A$ und $\mathcal{M}_C$ durch solche Zusatzkenntnisse reduzieren lassen. Solche Ansätze sind für das Verständnis der Zusammenhänge zwischen den einzelnen Teilen der Faktoren wichtig. Solche Situationen, dass einzelne Teile der Faktoren zugänglich sind, treten etwa bei der Analyse spektroskopischer Daten auf und die Reduktionsmöglichkeiten sind von Interesse.

In dem weiterführenden Abschnitt 3.7 wird eine Verallgemeinerung der Menge zulässiger Lösungen $\mathcal{M}_A$ für Probleme mit Rangdefizit eingeführt. Dies bedeutet, dass die allgemeine Faktorisierungsaufgabe 2.1 zugrunde gelegt wird. Weiter werden speziell Probleme untersucht, die zwar kein Rangdefizit haben, für die es aus Sicht einer konkreten Anwendung aber keine sinnvolle Faktorisierung gibt. Es bedarf einer Erhöhung der Spaltendimension von $C$ und der Zeilendimension von $A$, um auf sinnvolle Faktoren zu schließen.

Abschließend werden in Abschnitt 3.8 die Sensitivitäten von $\mathcal{M}_A$ und $\mathcal{M}_C$ sowie der $w(:,i)$, $i = 1,\ldots,k$, und der $u(:,j)$, $j = 1,\ldots,n$, in Bezug auf Störungen analysiert.

[1] Aufgrund der bilinearen Struktur des Modells $D = CA$ besitzt $\mathcal{M}_C$ analoge Eigenschaften; die Analyse erfolgt aber meistens nur für $\mathcal{M}_A$.

## 3.1 Vorbetrachtungen

Bevor die Menge $\mathcal{M}_A$ untersucht werden kann, ist es nötig die Konstruktion aus (2.5) mit $T(i,1) = 1$ für $i = 1, \ldots, s$ zu rechtfertigen. Die Herangehensweise ist gleichbedeutend damit, dass der erste rechtsseitige Singulärvektor (mit $V(:,1) > 0$) für jede mögliche nichtnegative Zeile von $A$ einen positiven Beitrag hat. Für die Rechtfertigung des Ansatzes unter bestimmten Voraussetzungen werden der Begriff einer irreduziblen Matrix sowie zwei Aussagen des Satzes von Perron-Frobenius benötigt.

**Definition 3.1** (Vergleiche etwa [10, 121])**.** *Eine $n \times n$ Matrix $M$ mit $n \geq 2$ wird reduzibel oder zerlegbar genannt, falls eine $n \times n$ Permutationsmatrix $P$ existiert, sodass*

$$PMP^{\mathbf{T}} = \begin{pmatrix} M_{1,1} & M_{1,2} \\ 0 & M_{2,2} \end{pmatrix}$$

*mit einer $m \times m$ Matrix $M_{1,1}$ und einer $m \times (n-m)$ Matrix $M_{1,2}$ mit $1 \leq m < n$. Andernfalls wird $M$ irreduzibel oder unzerlegbar genannt.*

Der Satz von Perron-Frobenius [8, 10, 121, 163, 172] fasst grundlegende spektrale Eigenschaften positiver quadratischer beziehungsweise nichtnegativer irreduzibler Matrizen zusammen. Perron zeigte diese 1907 in [129] für quadratische Matrizen $M$ mit $M_{ij} > 0 \ \forall i, j$, und Frobenius verallgemeinerte diese 1912 in [43] für irreduzible Matrizen $M$ mit $M_{ij} \geq 0 \ \forall i, j$.

**Satz 3.2** (Perron-Frobenius, siehe etwa [172])**.** *Sei $M \in \mathbb{R}^{n \times n}$ eine nichtnegative und irreduzible Matrix.*

*Es gelten:*

*(i) Der Spektralradius $\rho(M)$ ist ein einfacher Eigenwert von $M$.*

*(ii) Zu $\rho(M)$ gehört ein komponentenweise positiver Eigenvektor.*

*(iii) Die Funktion $\rho(M)$ ist strikt monoton wachsend in jedem Matrixelement von $M$.*

*Beweis.* Siehe etwa [172]. □

Die Aussagen aus Satz 3.2 werden in der vorliegenden Schrift häufig genutzt. Für ein konkretes $D$ garantiert die Irreduzibilität von $D^{\mathbf{T}}D$, dass der größte Singulärwert von $D$ (die positive Wurzel des größten Eigenwerts von $D^{\mathbf{T}}D$) einfach ist und es eine Singulärwertzerlegung von $D$ gibt, bei der der erste rechtsseitige Singulärvektor (der Eigenvektor zum größten Eigenwert von $D^{\mathbf{T}}D$) komponentenweise positiv ist.

Zur Konstruktion einer Faktorisierung wird im Folgenden in vielen Aussagen mit einer abgeschnittenen Singulärwertzerlegung mit $V(:,1) > 0$ gearbeitet. Dass es eine solche gibt, ist im Hinblick auf Satz 3.2 klar. Aus Übersichtlichkeitsgründen wird im Folgenden in der Regel nicht stets zusätzlich auf Satz 3.2 verwiesen.

### 3.1.1 Rechtfertigung des niedrigdimensionalen Ansatzes

Aufbauend auf Satz 3.2 lässt sich eine Bedingung angeben, unter welcher der Ansatz der niedrigdimensionalen Darstellung mittels der speziellen Form von $T$ aus (2.5) gerechtfertigt ist.

**Satz 3.3.** *Seien $D \in \mathbb{R}^{k \times n}$ eine nichtnegative Matrix, $s = \operatorname{rank}(D) \geq 2$ und $D^{\mathbf{T}}D$ irreduzibel. Weiter sei $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$ von der Form (2.1).*

*Es existiert kein $x \in \mathbb{R}^{s-1} \setminus \{0\}$ mit*

$$(0, x^{\mathbf{T}})V^{\mathbf{T}} \geq 0.$$

*Beweis.* Sei $D^{\mathbf{T}}D$ irreduzibel und sei angenommen, es gäbe ein $x \in \mathbb{R}^{s-1} \setminus \{0\}$ mit $(0, x^{\mathbf{T}})V^{\mathbf{T}} \geq 0$. Satz 3.2 garantiert, dass der erste rechtsseitige Singulärvektor echt positiv oder echt negativ ist. Weiter gilt mit $x \neq 0$ und da $\operatorname{rank}(V) = s$, dass $\|(0, x^{\mathbf{T}})V^{\mathbf{T}}\| > 0$. Somit gibt es ein $\ell \in \{1, \ldots, n\}$ mit $V(\ell, 1:s)(0, x^{\mathbf{T}})^{\mathbf{T}} \neq 0$. Für den Fall $V(:,1) > 0$ führt dies aufgrund der Orthogonalität von $V$ und unter der Annahme $(0, x^{\mathbf{T}})V^{\mathbf{T}} \geq 0$ auf den Widerspruch

$$0 < \underbrace{V(\ell,1)}_{>0}\underbrace{V(\ell,1:s)(0,x^{\mathbf{T}})^{\mathbf{T}}}_{>0} \leq \underbrace{(V(:,1))^{\mathbf{T}}V(:,1:s)}_{(1,0,\ldots,0)}(0,x^{\mathbf{T}})^{\mathbf{T}} = 0.$$

Für den Fall $V(:,1) < 0$ ist die Argumentation analog. □

Somit ist der Ansatz aus (2.5) gerechtfertigt und die Menge zulässiger Lösungen $\mathcal{M}_A$ kann definiert werden, falls $D^{\mathbf{T}}D$ irreduzibel ist und $V$ mit $V(:,1) > 0$ gewählt wurde.[2] Die zusätzlich angeführte Bedingung $\operatorname{rank}(D) \geq 2$ ist dahingehend formal, als dass ansonsten $\mathcal{M}_A \subset \mathbb{R}^0$ wäre und zudem die Faktorisierung einer Rang-1-Matrix im Sinne von Äquivalenzklassen eindeutig ist. Bezüglich der Rechtfertigung des Ansatzes zur niedrigdimensionalen Darstellung der Spalten von $C$ und zur Definition der Menge $\mathcal{M}_C$ lässt sich eine analoge Aussage treffen.

**Korollar 3.4.** *Seien $D \in \mathbb{R}^{k \times n}$ eine nichtnegative Matrix, $s = \operatorname{rank}(D) \geq 2$ und $DD^{\mathbf{T}}$ irreduzibel. Weiter sei $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$.*

*Es existiert kein $y \in \mathbb{R}^{s-1} \setminus \{0\}$ mit*

$$U\Sigma \begin{pmatrix} 0 \\ y \end{pmatrix} \geq 0.$$

*Beweis.* Für $D^{\mathbf{T}} = V\Sigma^{\mathbf{T}}U^{\mathbf{T}}$ existiert kein $y \in \mathbb{R}^{s-1} \setminus \{0\}$ mit $(0, y^{\mathbf{T}})U^{\mathbf{T}} \geq 0$, wenn $DD^{\mathbf{T}}$ irreduzibel ist. Da $\Sigma$ eine Diagonalmatrix mit positiven Diagonalelementen ist, gilt dies auch für $(0, y^{\mathbf{T}})\Sigma^{\mathbf{T}}U^{\mathbf{T}} \geq 0$ und nach Transposition folgt die Behauptung. □

### 3.1.2 Erweiterung zu einer notwendigen und hinreichenden Bedingung

In Satz 3.3 ist eine hinreichende Bedingung dafür angegeben, dass der niedrigdimensionale Ansatz aus (2.5) anwendbar ist. Sofern zwei kleine Zusätze eingebracht werden, lässt sich diese Bedingung zu einer notwendigen und hinreichenden erweitern. Diese Zusätze werden nur an dieser Stelle angeführt und für die Erweiterung der Aussage von Satz 3.3 gezeigt. Für den weiteren Verlauf dieser Schrift wird von den Zusätzen abgesehen, um den Lesefluss nicht durch ständige Untersuchungen etwaiger Spezialfälle zu beeinträchtigen.

**Lemma 3.5.** *Seien $D \in \mathbb{R}^{k \times n}$ eine nichtnegative Matrix, welche keine Nullspalte besitzt, und $s = \operatorname{rank}(D) \geq 2$.*

*Es ist $D^{\mathbf{T}}D$ genau dann eine irreduzible Matrix, wenn für keine abgeschnittene Singulärwertzerlegung $U\Sigma V^{\mathbf{T}}$ von $D$ (natürlich mit absteigend geordneten Singulärwerten in $\Sigma$) ein $x \in \mathbb{R}^{s-1} \setminus \{0\}$ mit $(0, x^{\mathbf{T}})V^{\mathbf{T}} \geq 0$ existiert.*

*Beweis.* Die eine Richtung entspricht der Aussage in Satz 3.3. Sei für den Beweis der anderen Richtung $D^{\mathbf{T}}D$ reduzibel, das heißt nach Definition 3.1 gibt es eine Permutationsmatrix $P$ mit

$$PD^{\mathbf{T}}DP^{\mathbf{T}} = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}.$$

[2] Dass der Ansatz $(-1, x)V^{\mathbf{T}}$ für $V(:,1) > 0$ nicht auf eine nichtnegative Zeile für $A$ führen kann, ist klar.

(Wegen der Symmetrie von $D^{\mathbf{T}}D$ ist die rechtsobere Matrix ebenfalls eine Nullmatrix.) Da $D$ keine Nullspalte enthält, sind die Diagonalelemente von $D^{\mathbf{T}}D$ positiv. Somit enthält $PD^{\mathbf{T}}DP^{\mathbf{T}}$ keine Nullspalte und $D_1$ und $D_2$ sind keine Nullmatrizen. Ohne Beschränkung der Allgemeinheit können $D_1$ und $D_2$ als irreduzibel angenommen werden; andernfalls erfolgt der Nachweis rekursiv zu irreduziblen Untermatrizen. Seien $\lambda_1$ und $\lambda_2$ die (betrags)größten Eigenwerte von $D_1$ beziehungsweise $D_2$. Nach Satz 3.2 sind $\lambda_1$ und $\lambda_2$ einfach und positiv. Die zugehörigen normierten Eigenwerte $u_1$ und $u_2$ sind (gegebenenfalls Multiplikation mit $-1$) komponentenweise positiv. Weiter lassen sie sich durch ein Auffüllen mit Nullen zu rechtsseitigen Singulärvektoren von $D$ erweitern. Das heißt, es gibt Indizes $i_1$ und $i_2$ sowie eine abgeschnittene Singulärwertzerlegung $U\Sigma V^{\mathbf{T}} = D$ mit

$$P(V(:,i_1)) = \begin{pmatrix} \pm u_1 \\ 0 \end{pmatrix}, \quad P(V(:,i_2)) = \begin{pmatrix} 0 \\ \pm u_2 \end{pmatrix}. \tag{3.1}$$

Ohne Beschränkung der Allgemeinheit sei $i_2 \neq 1$ (andernfalls wäre $i_1 \neq 1$). Dazu sei $(0, x^{\mathbf{T}}) := \pm e_{i_2}^{\mathbf{T}}$ mit $e_{i_2}$ dem $i_2$-ten Einheitsvektor. Wegen (3.1) und dem nicht eindeutigen Vorzeichen der rechtsseitigen Singulärvektoren von $D$ ergibt sich

$$PV(\pm e_{i_2}) = \begin{pmatrix} 0 \\ u_2 \end{pmatrix} \geq 0,$$

woraus

$$(0, x^{\mathbf{T}})V^{\mathbf{T}} = \pm e_{i_2}^{\mathbf{T}} V^{\mathbf{T}} P^{\mathbf{T}}\, P = (0, u_2^{\mathbf{T}})\, P \geq 0$$

folgt. □

Inwiefern die spezielle Zusatzbedingung „für keine abgeschnittene Singulärwertzerlegung“ in der Formulierung von Lemma 3.5 wichtig ist, zeigt die folgende Bemerkung sowie das darauf folgende Beispiel.

**Bemerkung 3.6.** *Die Zusatzbedingung, dass es für keine abgeschnittene Singulärwertzerlegung von $D$ ein $x \in \mathbb{R}^{s-1} \setminus \{0\}$ mit $(0, x^{\mathbf{T}})V^{\mathbf{T}} \geq 0$ gibt, taucht im Beweis von Lemma 3.5 nur indirekt auf. Sie kann aber nicht weggelassen werden. Im Beweis ist sie insofern präsent, als dass für den Fall $\lambda_1 = \lambda_2$ die in (3.1) auftretenden Singulärvektoren auch unabhängig vom Vorzeichen nicht eindeutig bestimmt sind. Die spezielle Wahl von $V(:,i_1)$ und $V(:,i_2)$ im Beweis von Lemma 3.5 wird bewusst genutzt, sodass (3.1) erfüllt ist. Es gibt unter Umständen aber auch andere Möglichkeiten, die zu $\lambda_1 = \lambda_2$ gehörigen Singulärvektoren zu wählen, welche den darauf folgenden Schluss nicht zulassen, siehe Beispiel 3.7.*

**Beispiel 3.7.** *Sei $D$ die $2 \times 2$-Einheitsmatrix. Eine mögliche Singulärwertzerlegung dieser ist*

$$U = V = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

*mit $\Sigma = D$. Nun gibt es kein $x \in \mathbb{R}$, sodass komponentenweise $(0, x)V^{\mathbf{T}} \geq 0$ gilt, aber trotzdem ist $D^{\mathbf{T}}D$ reduzibel.*

*Andererseits gehören aber auch die Faktoren $U = \Sigma = V = D$ zu einer Singulärwertzerlegung von $D$. Für diese gibt es ein $x \in \mathbb{R}$, sodass komponentenweise $(0, x)V^{\mathbf{T}} \geq 0$ gilt, nämlich etwa $x = 1$.*

Ist $D^{\mathbf{T}}D$ irreduzibel, so lässt sich die Menge zulässiger Lösungen $\mathcal{M}_A$ definieren, und ist $DD^{\mathbf{T}}$ irreduzibel, so lässt sich die Menge zulässiger Lösungen $\mathcal{M}_C$ definieren. Dabei implizieren sich die Irreduzibilitäten von $D^{\mathbf{T}}D$ und $DD^{\mathbf{T}}$ gegenseitig, sofern $D$ keine Nullzeile- und keine Nullspalte enthält, siehe später Satz 3.38. Somit muss nur eines der beiden Matrixprodukte diesbezüglich untersucht werden. Weiter lässt sich ein Zusammenhang zwischen den Irreduzibilitäten von $D^{\mathbf{T}}D$ und $DD^{\mathbf{T}}$ und einer Blockstruktur der Faktorisierung nachweisen, siehe Abschnitt 3.5.

## 3.2 Wichtige Eigenschaften

In diesem Abschnitt werden wichtige Eigenschaften der Menge zulässiger Lösungen $\mathcal{M}_A$ untersucht und nachgewiesen. Ein vertieftes Verständnis der Mengen zulässiger Lösungen ist unter anderem im Hinblick auf die Entwicklung effizienter Methoden zu deren Berechnung beziehungsweise Approximation unerlässlich. Eine wichtige Eigenschaft ist die Kompaktheit von $\mathcal{M}_A$. Zu deren Beweis bedarf es der geometrischen Interpretation einer zulässigen Lösung. Da diese erst in Abschnitt 3.3 behandelt wird, erfolgt der Nachweis der Kompaktheit später in Abschnitt 3.4.

### 3.2.1 Beschränktheit

Basierend auf Satz 3.3 wird gezeigt, dass $\mathcal{F}_A$ beschränkt ist. Wegen $\mathcal{M}_A \subset \mathcal{F}_A$ folgt daraus, dass auch $\mathcal{M}_A \subset \mathcal{F}_A$ beschränkt ist. Die Obermenge $\mathcal{F}_A$ aus (2.8) ist der Schnitt der affinen Halbräume

$$\left\{x \in \mathbb{R}^{s-1} :\ V(i, 2:s)x \geq -V(i,1)\right\}, \quad i = 1, \ldots, n, \tag{3.2}$$

welche sich aus den Bedingungen für eine nichtnegative erste Zeile von $A$ ergeben.

**Satz 3.8.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix, $s = \operatorname{rank}(D) = \operatorname{rank}_+(D) \geq 2$, $D^{\mathbf{T}}D$ irreduzibel und $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$ mit $V(:,1) > 0$.*

*Die Mengen $\mathcal{F}_A$ und $\mathcal{M}_A$ sind beschränkt.*

*Beweis.* Es ist $\mathcal{F}_A$ die Schnittmenge der $n$ affinen Halbräume aus (3.2) und somit konvex. Weiter gilt offensichtlich $(0, \ldots, 0)^{\mathbf{T}} \in \mathcal{F}_A$. Demzufolge ist $\mathcal{F}_A$ beschränkt, wenn es kein $x \in \mathbb{R}^{s-1} \setminus \{0\}$ gibt, sodass für alle $\gamma \geq 0$

$$(1, \gamma x^{\mathbf{T}})V^{\mathbf{T}} \geq 0 \tag{3.3}$$

gilt. Nach Satz 3.3 gibt es kein $x \in \mathbb{R}^{s-1}\setminus\{0\}$ mit $(0, x^{\mathbf{T}})V^{\mathbf{T}} \geq 0$. Somit kann es wegen $V(:,1) \geq 0$ kein $x \in \mathbb{R}^{s-1} \setminus \{0\}$ geben, welches für alle $\gamma \geq 0$ die zu (3.3) äquivalente Forderung

$$\gamma V(:, 2:s)x \geq -V(:,1)$$

erfüllt. Somit ist $\mathcal{F}_A$ beschränkt und deren Teilmenge $\mathcal{M}_A$ ist es ebenfalls. □

Die Beschränktheit von $\mathcal{F}_C$ lässt sich analog zeigen und wegen $\mathcal{M}_C \subset \mathcal{F}_C$ ist auch $\mathcal{M}_C$ beschränkt. Im Beweis treten die Verhältnisse $\sigma_i/\sigma_1$, $i = 2, \ldots, s$, beziehungsweise $\sigma_1/\sigma_i$, $i = 2, \ldots, s$, als positive Skalierungsfaktoren auf.

### 3.2.2 Der Ursprung gehört nicht zu den Mengen zulässiger Lösungen

Als nächstes wird mittels Korollar 3.4 gezeigt, dass der Ursprung $o = (0, \ldots, 0)^{\mathbf{T}} \in \mathbb{R}^{s-1}$ zwar in $\mathcal{F}_A$ enthalten ist, aber nicht in $\mathcal{M}_A$. Die Menge zulässiger Lösungen $\mathcal{M}_A$ umlagert den Ursprung somit in gewisser Weise, enthält ihn aber nicht. Um dies zusammengefasst mit $o \in \mathcal{F}_C$ und $o \notin \mathcal{M}_C$ in einem Satz zu zeigen, wird zunächst folgende, sehr einfache Aussage nachgewiesen.

**Lemma 3.9.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix mit $\operatorname{rank}(D) \geq 1$ und $D^{\mathbf{T}}D$ sowie $DD^{\mathbf{T}}$ irreduzibel. Weiter sei $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$ mit $V(:,1) > 0$.*

*Es gilt komponentenweise $U(:,1) > 0$.*

*Beweis.* Wegen $\operatorname{rank}(D) \geq 1$ ist $\sigma_1 > 0$. Weiter gilt $DV(:,1) = \sigma_1 U(:,1)$ und es folgt für alle $i = 1, \ldots, k$, dass

$$U_{i1} = \sigma_1^{-1} D(i,:)V(:,1) > 0,$$

da $D$ keine Nullzeile enthält (wegen Irreduzibilität von $DD^{\mathbf{T}}$) sowie nichtnegativ ist und $V(:,1)$ komponentenweise positiv ist. □

**Satz 3.10.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix, $s = \operatorname{rank}(D) = \operatorname{rank}_+(D) \geq 2$ und $D^{\mathbf{T}}D$ sowie $DD^{\mathbf{T}}$ irreduzibel. Sei $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$ mit $V(:,1) > 0$.*

*Es gelten $o = (0,\ldots,0)^{\mathbf{T}} \in \mathcal{F}_A$ und $o \notin \mathcal{M}_A$ sowie $o \in \mathcal{F}_C$ und $o \notin \mathcal{M}_C$.*

*Beweis.* Dass $o \in \mathcal{F}_A$ gilt, folgt aus der Nichtnegativität von $V(:,1)$. Sei nun $o \in \mathcal{M}_A$ angenommen. Das hieße, es gäbe eine Faktorisierung $D = CA$ mit $A(1,:) = (1, o^{\mathbf{T}})V^{\mathbf{T}}$ und $C \in \mathbb{R}_+^{k\times s}$ sowie $A \in \mathbb{R}_+^{s\times n}$. Sei $T$ die mit (2.2) zu $C$ und $A$ gehörige Matrix. Dann wäre

$$0 = T(1,:)((T^{-1})(:,2)) = (1, o^{\mathbf{T}})((T^{-1})(:,2)) = (T^{-1})_{12}.$$

Dies wäre aber ein Widerspruch, da nach Korollar 3.4 aus $C(:,2) \geq 0$ mit $U(:,1) > 0$ (Lemma 3.9) folgt, dass $(T^{-1})_{12} > 0$ ist. Somit muss $o \notin \mathcal{M}_A$ gelten.

Analog folgt aus der Nichtnegativität von $U(:,1)$, dass $o \in \mathcal{F}_C$ gilt. Sei nun $o \in \mathcal{M}_C$ angenommen. In analoger Weise hieße dies, es gäbe eine Faktorisierung $D = CA$ mit $C(:,1) = U\Sigma(1, y^{\mathbf{T}})^{\mathbf{T}}$ und $C \in \mathbb{R}_+^{k\times s}$ sowie $A \in \mathbb{R}_+^{s\times n}$. Sei $T$ die mit (2.2) zu $C$ und $A$ gehörige Matrix. Dann wäre

$$0 = T(2,:)((T^{-1})(:,1)) = T(2,:)(1, o^{\mathbf{T}})^{\mathbf{T}} = T_{21}.$$

Dies wäre aber ein Widerspruch, da nach Satz 3.3 aus $A(2,:) \geq 0$ folgt, dass $T_{21} > 0$ ist. Somit muss $o \notin \mathcal{M}_C$ gelten. □

Zwei einfache Resultate aus Satz 3.10 sind, dass $V(:,1) \notin \mathcal{A}$ und $U(:,1) \notin \mathcal{C}$, falls $s \geq 2$.

**Korollar 3.11.** *Seien die Voraussetzungen aus Satz 3.10 gegeben.*

*Es gibt keine nichtnegative Faktorisierung von $D$ mit $A(1,:) = V(:,1)^{\mathbf{T}}$ und es gibt auch keine nichtnegative Faktorisierung von $D$ mit $C(:,1) = U(:,1)$.*

*Beweis.* Nach Satz 3.10 ist $o = (0,\ldots,0)^{\mathbf{T}} \notin \mathcal{M}_A$. Somit gibt es keine nichtnegative Faktorisierung von $D$ mit $A(1,:) = (1,0,\ldots,0)V^{\mathbf{T}}$. Anwendung von Satz 3.10 auf $D^{\mathbf{T}}$ ergibt, dass es keine nichtnegative Faktorisierung von $D$ mit $C(:,1) = U(:,1)$ gibt. □

Als nächstes wird gezeigt, dass der Nullpunkt stets im Inneren von $\mathcal{I}_A$ liegt. Dies wird später unter anderem im Rahmen des Nachweises der Kompaktheit von $\mathcal{M}_A$ benötigt, siehe Satz 3.32.

**Definition 3.12.** *Sei $K \subset \mathbb{R}^d$. Die Menge*

$$\operatorname{int}(K) = \left\{ x \in \mathbb{R}^d : \exists \varepsilon > 0 \text{ mit } B_\varepsilon(x) \subset K \right\} \quad \text{mit} \quad B_\varepsilon(x) = \{u \in \mathbb{R}^d : \|u - x\| < \varepsilon\}$$

*wird als das Innere von $K$ bezeichnet. Weiter ist $\partial K = K \setminus \operatorname{int}(K)$ der Rand von $K$.*

**Lemma 3.13.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix, $s = \operatorname{rank}(D) \geq 2$ und $D^{\mathbf{T}}D$ sowie $DD^{\mathbf{T}}$ irreduzibel. Sei $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$ mit $V(:,1) > 0$.*

*Der Nullpunkt $o = (0,\ldots,0)^{\mathbf{T}} \in \mathbb{R}^{s-1}$ liegt im Inneren von $\mathcal{I}_A$, das heißt es gelten $o \in \mathcal{I}_A$ und $o \notin \partial\mathcal{I}_A$.*

*Beweis.* Sei angenommen, es würde $o \in \partial\mathcal{I}_A$ oder gar $o \notin \mathcal{I}_A$ gelten. Da $\mathcal{I}_A$ die konvexe Hülle von $w(:,i)$, $i = 1,\ldots,k$, ist, gäbe es somit eine Hyperebene, die den Nullpunkt enthält und für die alle $w(:,i)$, $i = 1,\ldots,k$, auf einer Seite der Hyperebene liegen. Das heißt, es gibt ein $x \in \mathbb{R}^{s-1} \setminus \{0\}$ mit

$$(w(:,i))^{\mathbf{T}} x \geq 0 \qquad \forall i = 1,\ldots,k. \tag{3.4}$$

Nach Lemma 3.9 ist $U(:,1)$ komponentenweise positiv. Somit ist (3.4) äquivalent zu

$$U(i,:)\Sigma(:,2:s)x \geq 0 \quad \forall i = 1,\ldots,k,$$

und dies ist wiederum äquivalent zu

$$U\Sigma \begin{pmatrix} 0 \\ x \end{pmatrix} \geq 0.$$

Nach Korollar 3.4 kann es ein solches $x \in \mathbb{R}^{s-1} \setminus \{0\}$ aber nicht geben und demzufolge gibt es auch keine Hyperebene, die den Nullpunkt enthält und für die alle $w(:,i)$, $i = 1,\ldots,k$, auf einer Seite der Hyperebene liegen. Damit ist der Beweis vollständig. □

### 3.2.3 Unterbrechungsfreie Schnitte mit Strahlen vom Ursprung

In diesem Teilabschnitt wird gezeigt, dass der Schnitt der Menge zulässiger Lösungen $\mathcal{M}_A$ mit einem Strahl, der vom Ursprung ausgeht, entweder leer oder ein Geradenabschnitt ist. In Satz 3.10 wurde gezeigt, dass der Ursprung in der Obermenge $\mathcal{F}_A$ aber nicht in $\mathcal{M}_A$ liegt. Zusammen mit den Sätzen 3.8 und 3.10 eröffnet dies die Möglichkeit, einen Strahlenalgorithmus zur Approximation der Mengen zulässiger Lösungen für $s \geq 2$ zu entwickeln. Weiterhin wird diese Eigenschaft auch für den später vorgestellten inversen Polyhedron inflation Algorithmus genutzt.

**Lemma 3.14.** *Seien $C \in \mathbb{R}^{k\times s}$ und $A \in \mathbb{R}^{s\times n}$ nichtnegative Matrizen mit* $\operatorname{rank}(C) = \operatorname{rank}(A) = s$. *Ferner sei $X \in \mathbb{R}^{s\times s}$ eine Matrix mit $X_{11} \geq 1$, $X_{1j} \leq 0$ und*

$$X_{ij} = \begin{cases} 1 & \text{für } i = j, \\ 0 & \text{für } i \neq j. \end{cases}, \quad \text{für } i = 2,\ldots,s,\ j = 1,\ldots,s. \tag{3.5}$$

*Unter diesen Voraussetzungen ist $X \in \mathbb{R}^{s\times s}$ regulär und falls $X(1,:)A$ nichtnegativ ist, so sind*

$$\widetilde{C} = CX^{-1}, \qquad \widetilde{A} = XA \tag{3.6}$$

*ebenfalls nichtnegativ.*

*Beweis.* Wegen der speziellen Struktur von $X$ gilt $\det(X) = X_{11} \geq 1$ und der erste Teil der Behauptung ist nachgewiesen. Zum Nachweis des zweiten Teils: Aus der Struktur von $X$ und $\sum_{j=1}^{s}(X^{-1})_{1j}X_{j1} = 1$ folgt zunächst $(X^{-1})_{11} > 0$. Weiter folgt damit aus $\sum_{j=1}^{s}(X^{-1})_{1j}X_{ji} = 0$, $i = 2,\ldots,s$, dass $(X^{-1})_{1j} \geq 0$, $j = 2,\ldots,s$. Einfaches Nachrechnen ergibt, dass $(X^{-1})_{ij} = X_{ij}$ für $i = 2,\ldots,s$ und $j = 1,\ldots,s$. Somit ist die Matrix $X^{-1}$ nichtnegativ und $\widetilde{C} = CX^{-1}$ ist es damit auch. Nach Voraussetzung ist $\widetilde{A}(1,:) = X(1,:)A$ nichtnegativ. Für die weiteren Zeilen von $\widetilde{A}$ gilt

$$\widetilde{A}(2:s,:) = X(2:s,:)A = A(2:s,:)$$

und auch $\widetilde{A} \geq 0$ ist gezeigt. □

**Satz 3.15.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix, $s = \operatorname{rank}(D) = \operatorname{rank}_+(D) \geq 2$ und $D^{\mathbf{T}}D$ sowie $DD^{\mathbf{T}}$ irreduzibel. Sei $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$ mit $V(:,1) > 0$. Weiter sei $x \in \mathcal{M}_A$. Zu $x$ sei der Wert $\gamma^* \geq 1$ so definiert, dass $\gamma^* x$ auf dem Rand von $\mathcal{F}_A$ liegt.*

*Für alle $\widetilde{x} = \gamma x$ mit $\gamma \in [1, \gamma^*]$ gilt $\widetilde{x} \in \mathcal{M}_A$.*

*Beweis.* Die Beweisidee ist wie folgt: Zu $x$ existiert eine Untermatrix $S \in \mathbb{R}^{(s-1)\times(s-1)}$ wie in (2.5), welche auf eine nichtnegative Faktorisierung mit $C = U\Sigma T^{-1}$ und $A = TV^{\mathbf{T}}$ führt. Diese Matrix führt in derselben Weise auch für $\widetilde{x}$ zu einer nichtnegativen Faktorisierung $D = \widetilde{C}\widetilde{A}$. Um dies zu zeigen, werden die Faktoren $\widetilde{C}$ und $\widetilde{A}$ zu $\widetilde{x}$ mittels $C$ und $A$ konstruiert. Diese Transformation ist von der Form wie in Lemma 3.14 und somit sind $\widetilde{C}$ und $\widetilde{A}$ nichtnegativ.

Es ist $x$ zulässig und somit existiert eine Matrix $S$, sodass die zusammengesetzte Matrix $T$ aus (2.5) eine reguläre Transformation mit $C = UST^{-1} \geq 0$ und $A = TV^{\mathbf{T}} \geq 0$ ist. Für das neue $\widetilde{x} = \gamma x$ wird die Transformation

$$\widetilde{T} = \begin{pmatrix} 1 & \gamma x_1 & \cdots & \gamma x_{s-1} \\ 1 & & & \\ \vdots & & S & \\ 1 & & & \end{pmatrix}$$

genutzt und es bleiben $\operatorname{rank}(\widetilde{T}) = s$ sowie $\widetilde{C} = U\Sigma\widetilde{T}^{-1} \geq 0$, $\widetilde{A} = \widetilde{T}V^{\mathbf{T}} \geq 0$ zu zeigen.

Die Transformation zur Berechnung von $\widetilde{C}$ und $\widetilde{A}$ unter Nutzung von $C$ und $A$ wie in (3.6) ist $X = \widetilde{T}T^{-1}$. Nun wird gezeigt, dass $X$ die Voraussetzungen von Lemma 3.14 erfüllt. Die Zeilen $X(2:s,:)$ erfüllen die Bedingung aus (3.5). Für $j = 2, \ldots, s$ lauten die Einträge in der ersten Zeile von $X$

$$X_{1j} = (1, \gamma x_1, \ldots, \gamma x_{s-1})T^{-1}(:,j) = (1-\gamma)\left(e_1 + \frac{\gamma}{1-\gamma}(1, x^{\mathbf{T}})\right)T^{-1}(:,j)$$

mit $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^{1\times s}$. Nach Korollar 3.4 gilt $T^{-1}(1,:) > 0$. Mit $\gamma \geq 1$ folgt

$$X_{1j} = (1-\gamma)e_1T^{-1}(:,j) = (1-\gamma)(T^{-1})_{1j} \leq 0, \qquad j = 2, \ldots, s.$$

Für $X_{11}$ gilt

$$X_{11} = \widetilde{T}(1,:)T^{-1}(:,1) = (\gamma T(1,:) + (1-\gamma, 0, \ldots, 0))T^{-1}(:,1) = \gamma + (1-\gamma)(T^{-1})_{11}.$$

Zur Bestimmung von $(T^{-1})_{11}$ lässt sich $1 = T^{-1}(1,:)T(:,1) = \sum_{j=1}^{s}(T^{-1})_{1j}$ nutzen und es folgt $(T^{-1})_{11} = 1 - \sum_{j=2}^{s}(T^{-1})_{1j}$. Somit ist

$$X_{11} = \gamma + (1-\gamma)\left(1 - \sum_{j=2}^{s}(T^{-1})_{1j}\right) = 1 - \sum_{j=2}^{s} X_{1j} \geq 1.$$

Die Matrix $X$ besitzt demzufolge eine Form wie in Lemma 3.14 vorausgesetzt und es gelten $\operatorname{rank}(\widetilde{T}) = s$ sowie $\widetilde{C}\widetilde{A} = D$. Wegen $\widetilde{x} \in \mathcal{F}_A$ gilt $(1, \widetilde{x}^{\mathbf{T}})V^{\mathbf{T}} = \widetilde{A}(1,:) = X(1,:)A \geq 0$ und die resultierenden Faktoren $\widetilde{C}$ und $\widetilde{A}$ sind nichtnegativ. Somit ist $\widetilde{x} \in \mathcal{M}_A$. □

**Bemerkung 3.16.** *Der Beweis zu Satz 3.15 lässt sich auch sehr einfach geometrisch führen. Später wird dies in dem Beweis von Satz 3.73 in ähnlicher Weise gemacht.*

### Lochfreie Segmente

Zwei einfache Folgerungen aus Satz 3.15 sind in den Korollaren 3.18 und 3.19 formuliert. Zuvor werden konvexe und strikt konvexe Linearkombinationen eingeführt.

**Definition 3.17.** *Seien $b_i \in \mathbb{R}^n$ für $i = 1, \dots, s$. Eine Linearkombination*

$$v = \sum_{i=1}^{s} \alpha_i b_i$$

*wird Konvexkombination genannt, falls $0 \leq \alpha_i \leq 1$ für alle $i = 1, \dots, s$ sowie $\sum_{i=1}^{s} \alpha_i = 1$ gelten, und strikte Konvexkombination genannt, falls $0 < \alpha_i < 1$ für alle $i = 1, \dots, s$ und $\sum_{i=1}^{s} \alpha_i = 1$ gelten.*

Mitunter wird eine Linearkombination auch (strikt) konvex genannt, wenn sie eine (strikte) Konvexkombination ist.

**Korollar 3.18.** *Falls ein $x$ auf dem Rand von $\mathcal{F}_A$ nicht zu $\mathcal{M}_A$ gehört, so gibt es kein $\gamma \geq 0$, sodass $\gamma x$ zu $\mathcal{M}_A$ gehört.*

*Beweis.* Sei angenommen, es gäbe ein solches $\gamma \geq 0$ mit $\widetilde{x} = \gamma x \in \mathcal{M}_A$. Dann wären nach Satz 3.15 auch alle $(1-\delta)\widetilde{x} + \delta x$ für $0 \leq \delta \leq 1$ zulässig und insbesondere $x$ selbst. Dies ist ein Widerspruch zur Annahme und es kann ein solches $\gamma$ nicht geben. □

**Korollar 3.19.** *Seien die Voraussetzungen aus Satz 3.15 erfüllt und $\mathcal{M}_A$ bestehe aus mehr als einem Segment.*[3]

*Die einzelnen Segmente sind frei von $(s-1)$-dimensionalen Löchern. Das heißt, es gibt keine von $\mathcal{M}_A$ vollständig umschlossene Menge $\mathcal{L}$ mit $\mathcal{L} \cap \mathcal{M}_A = \emptyset$ und $o = (0, \dots, 0)^{\mathbf{T}} \notin \mathcal{L}$.*

*Beweis.* Angenommen $\mathcal{T}$ sei ein Segment, welches ein Loch enthält. Somit gäbe es $x, \widetilde{x} \in \mathcal{T}$, sodass die Gerade durch $x$ und $\widetilde{x}$ sowohl Teile des Lochs enthält als auch durch den Nullpunkt geht. Weiter ist nach Voraussetzung der Nullpunkt keine Konvexkombination von $x$ und $\widetilde{x}$. Somit gibt es ein $\gamma > 1$ mit entweder $x = \gamma\widetilde{x}$ oder $\widetilde{x} = \gamma x$. Dann gehört nach Satz 3.15 die gesamte Strecke zwischen $x$ und $\widetilde{x}$ zu $\mathcal{T}$, was im Widerspruch zu den Voraussetzungen steht. □

### Gültigkeit auch für bereits fixierte Lösungen

Abschließend wird gezeigt, dass die Eigenschaft der unterbrechungsfreien Schnitte mit Strahlen vom Ursprung aus auch für eine oder mehrere fixierte zulässige Lösung(en) bestehen bleibt. Seien $x^{(i)} \in \mathcal{M}_A$, $i = 1, \dots, s_0$, mit $s_0 < s$ fixiert. Die restringierte Menge zulässiger Lösungen sei

$$\begin{aligned}\mathcal{M}_A^{x^{(1)},\dots,x^{(s_0)}} = \big\{ & x \in \mathcal{M}_A : \exists S \in \mathbb{R}^{(s-1)\times(s-1)},\ S(s-i,:) = (x^{(i)})^{\mathbf{T}} \text{ für } i = 1, \dots, s_0, \\ & \operatorname{rank}(T) = s,\ C = U\Sigma T^{-1} \geq 0,\ A = TV^{\mathbf{T}} \geq 0 \big\} \cup \{x^{(1)}, \dots, x^{(s_0)}\}.\end{aligned}$$

**Korollar 3.20.** *Seien die Voraussetzungen aus Satz 3.15 gegeben und seien $x^{(i)} \in \mathcal{M}_A$, $i = 1, \dots, s_0$, mit $s_0 < s$ fixiert. Weiter seien $x \in \mathcal{M}_A^{x^{(1)},\dots,x^{(s_0)}}$ mit $x \neq x^{(i)}$, $i = 1, \dots, s_0$, und zu $x$ der Wert $\gamma^* \geq 1$ so definiert, dass $\gamma^* x$ auf dem Rand von $\mathcal{F}_A$ liegt.*

*Für alle $\widetilde{x} = \gamma x$ mit $\gamma \in [1, \gamma^*]$ gilt $\widetilde{x} \in \mathcal{M}_A^{x^{(1)},\dots,x^{(s_0)}}$.*

[3] Die einzelnen Zusammenhangskomponenten der Mengen zulässiger Lösungen werden in dieser Schrift auch Segmente genannt.

*Beweis.* Sei $S \in \mathbb{R}^{(s-1)\times(s-1)}$ eine Untermatrix mit $S(s-i,:) = (x^{(i)})^{\mathbf{T}}$ für $i = 1,\dots,s_0$, welche mit $x$ wie in (2.5) auf eine Transformation $T$ mit $U\Sigma T^{-1} \geq 0$ und $TV^{\mathbf{T}} \geq 0$ führt. Dazu ergibt sich der Rest analog zum Beweis des Satzes 3.15 und die Matrix $S$ führt auch hier für $\widetilde{x} = \gamma^* x$ auf eine nichtnegative Zerlegung. □

## 3.3 Geometrische Zusammenhänge

Ein wichtiger Teilschritt bei der Approximation von $\mathcal{M}_A$ ist die Klassifizierungen einzelner $x \in \mathbb{R}^{s-1}$ als *zulässig* oder *nicht zulässig*. Für die meisten, der später in Kapitel 4 vorgestellten, Algorithmen zur Bestimmung von $\mathcal{M}_A$ werden numerische Ansätze zur Klassifizierung einzelner $x \in \mathbb{R}^{s-1}$ genutzt. Diese basieren auf Optimierungen und Bewertungsfunktionen.

In diesem Abschnitt steht die Klassifizierung von $x \in \mathbb{R}^{s-1}$ mittels geometrischer Argumente im Mittelpunkt. Die Vorgehensweise geht auf [17,120,135,138] zurück, wobei in diesen Arbeiten eine andere Skalierung für $D$ und $A$ genutzt wird. Die Idee ist es, die niedrigdimensionale Darstellung des gesamten Faktors $A$ zu betrachten und eine Bedingung aufzustellen, sodass ein solcher Faktor zu einer nichtnegativen Matrixfaktorisierung führt. Daraus lässt sich eine notwendige und hinreichende Bedingung für eine zulässige Lösung $x \in \mathbb{R}^{s-1}$ ableiten. Entscheidend sind neben der niedrigdimensionalen Darstellung des gesamten Faktors $A$ mittels eines Simplex $\mathcal{S}$ die Mengen $\mathcal{F}_A$ und $\mathcal{I}_A$. Zuvor wird der Zusammenhang zwischen Faktoren $C$ und $A$ einer Faktorisierung und deren niedrigdimensionalen Blockdarstellung durch ein Simplex eingeführt.

**Definition 3.21.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix und $s = \operatorname{rank}(D) = \operatorname{rank}_+(D) \geq 2$. Weiter seien $D^{\mathbf{T}}D$ und $DD^{\mathbf{T}}$ irreduzibel und $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$ mit $V(:,1) > 0$. Seien $C \in \mathbb{R}^{k\times s}$ und $A \in \mathbb{R}^{s\times n}$ Faktoren mit $D = CA$, welche sich als $C = U\Sigma T^{-1}$, $A = TV^{\mathbf{T}}$ mittels eines regulären $T$ mit der Struktur aus (2.5) bilden lassen.[4] Die Faktoren $C$ und $A$ und ein Simplex $\mathcal{S} \subset \mathbb{R}^{s-1}$ werden einander zugehörig genannt, wenn $T(i,2:s)^{\mathbf{T}}$, $i = 1,\dots,s$, die Ecken von $\mathcal{S}$ sind (mit $T = AV$). Die Ecken von $\mathcal{S}$ werden spaltenweise in $\mathsf{S} \in \mathbb{R}^{(s-1)\times s}$ eingetragen, es gilt also*

$$T(i,2:s)^{\mathbf{T}} = \mathsf{S}(:,i), \qquad i = 1,\dots,s.$$

**Bemerkung 3.22.** *Definition 3.21 schließt unter Berücksichtigung der Skalierungsmehrdeutigkeit alle nichtnegativen Matrixfaktorisierungen $D = CA$ mit ein. Eine geeignete Umskalierung der Zeilen von $A$ erfolgt durch* $\operatorname{diag}(AV(:,1))^{-1}$.

### 3.3.1 Geometrische Interpretation einer zulässigen Lösung

Eingeschränkt auf $s = 3$ und unter Nutzung einer anderen Skalierung als der in (2.5) geht die geometrische Klassifizierung eines $x \in \mathbb{R}^{s-1}$ auf [17, 120, 135, 138] zurück, vergleiche auch [87, 148]. Zunächst wird in Satz 3.23 die geometrische Interpretation einer kompletten Faktorisierung behandelt und in Satz 3.26 wird die geometrische Interpretation einer zulässigen Lösung in der Art angegeben, wie sie aus Veröffentlichungen bekannt ist. Zusätzlich wird eine weitere sinnvolle Formulierung in Satz 3.31 angeführt.

**Satz 3.23.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix und $s = \operatorname{rank}(D) = \operatorname{rank}_+(D) \geq 2$. Weiter seien $D^{\mathbf{T}}D$ sowie $DD^{\mathbf{T}}$ irreduzibel und $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$ mit $V(:,1) > 0$. Es seien $C \in \mathbb{R}^{k\times s}$ und $A \in \mathbb{R}^{s\times n}$ Faktoren mit $CA = D$, zu denen es gemäß Definition 3.21 ein zugehöriges Simplex $\mathcal{S} \subset \mathbb{R}^{s-1}$ gebe. Sei $\mathcal{S}$ dieses Simplex.*

[4] An $C$ und $A$ werden hier ausnahmsweise keine Nichtnegativitätsforderungen gestellt.

*Die beiden folgenden Aussagen sind äquivalent:*

*a) Die Faktoren $C$ und $A$ sind nichtnegativ.*

*b) Das Simplex $\mathcal{S}$ erfüllt folgende zwei Bedingungen:*

*i) die Obermenge $\mathcal{F}_A$ enthält $\mathcal{S}$, das heißt $\mathsf{S}(:,i) \in \mathcal{F}_A$ für alle $i = 1,\dots,s$, und*

*ii) das Simplex $\mathcal{S}$ enthält die Menge $\mathcal{I}_A$, es enthält also alle $w(:,i)$, $i = 1,\dots,k$.*

*Beweis.* Zunächst wird a) ⇒ b) gezeigt: Aus a) folgt direkt i). Seien für $i \in \{1,\dots,k\}$ die Skalare $\gamma_i \in \mathbb{R}$ und die Vektoren $z_i \in \mathbb{R}^s$ definiert als

$$\gamma_i = D(i,:)V(:,1), \qquad z_i^{\mathbf{T}} = \frac{U(i,:)\Sigma T^{-1}}{\gamma_i} = \frac{C(i,:)}{\gamma_i}.$$

Es sind $\gamma_i > 0$ und $z_i \geq 0$. Dann ist

$$z_i^{\mathbf{T}} T = \frac{U(i,:)\Sigma}{\gamma_i} = \frac{D(i,:)V}{\gamma_i} = (1, w(:,i)^{\mathbf{T}})$$

und mit $T(i,2:s)^{\mathbf{T}} = \mathsf{S}(:,i)$ sowie $T(:,1) = (1,\dots,1)^{\mathbf{T}}$ folgen $\sum_{\ell=1}^{s}(z_i)_\ell = 1$ und $z_i^{\mathbf{T}}\mathsf{S}^{\mathbf{T}} = w(:,i)^{\mathbf{T}}$. Da dies für alle $i = 1,\dots,k$ gilt, folgt die Behauptung.
Nun wird b) ⇒ a) gezeigt: Bedingung i) impliziert $A \geq 0$. Da alle $w(:,i)$ in $\mathcal{S}$ liegen, gibt es für jedes $i \in \{1,\dots,k\}$ ein $z_i \in \mathbb{R}^s_+$, sodass $z_i^{\mathbf{T}} T = (1, w(:,i)^{\mathbf{T}})$ gilt. Damit ist

$$\begin{aligned} C(i,:) &= U(i,:)\Sigma T^{-1} \\ &= D(i,:)VT^{-1} \\ &= D(i,:)V(:,1)\ \underbrace{\frac{D(i,:)V}{D(i,:)V(:,1)}}_{=(1,w(:,i)^{\mathbf{T}})=z_i^{\mathbf{T}}T}\ T^{-1} \\ &= D(i,:)V(:,1)z_i^{\mathbf{T}}TT^{-1} \\ &= \gamma_i z_i^{\mathbf{T}} \geq 0 \end{aligned}$$

mit $\gamma_i = D(i,:)V(:,1) > 0$. Somit ist der Beweis vollständig und die Aussagen a) und b) sind äquivalent. □

**Definition 3.24.** *Erfüllt ein Simplex die Bedingung b) aus Satz 3.23, so wird es ein zulässiges Simplex genannt.*

**Bemerkung 3.25.** *Die Bedingung ii) aus Satz 3.23 ist äquivalent dazu, dass sich alle $w(:,i)$ als Konvexkombinationen aus den Ecken von $\mathcal{S}$ darstellen lassen.*

Die in Veröffentlichungen oft genutzte Klassifizierung eines $x \in \mathbb{R}^{s-1}$ als *zulässig* oder *nicht zulässig*, lässt sich direkt aus Satz 3.23 ableiten:

**Satz 3.26.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix und $s = \operatorname{rank}(D) = \operatorname{rank}_+(D) \geq 2$. Weiter seien $D^{\mathbf{T}}D$ sowie $DD^{\mathbf{T}}$ irreduzibel und $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$ mit $V(:,1) > 0$.*

*Ein $x \in \mathbb{R}^{s-1}$ ist genau dann zulässig, wenn es zu $\mathcal{F}_A$ gehört und es $x^{(i)} \in \mathcal{F}_A$, $i = 1\dots,s-1$, gibt, sodass das daraus gebildete Simplex $\mathcal{S} \subset \mathbb{R}^{s-1}$ mit den Ecken $\mathsf{S}(:,1) = x$ und $\mathsf{S}(:,i+1) = x^{(i)}$, $i = 1\dots,s-1$, alle $w(:,i)$, $i = 1\dots,k$, enthält.*

*Beweis.* Sei $x \in \mathbb{R}^{s-1}$ zulässig. Nach der Definition von $\mathcal{M}_A$ aus (2.6) gibt es eine Matrix $S \in \mathbb{R}^{(s-1)\times(s-1)}$, sodass die dazu wie in (2.5) gebildete Matrix $T$ regulär ist und auf nichtnegative Faktoren $C = U\Sigma T^{-1}$ und $A = TV^{\mathbf{T}}$ führt. Sei $\mathcal{S}$ das, zu einer solchen nichtnegativen Faktorisierung, gehörige Simplex, sodass $\mathsf{S}(:,1) = x$ und $\mathsf{S}(:,i) = S(i-1,:)^{\mathbf{T}}$, $i = 2,\ldots,s$. Nach Satz 3.23 liegen alle Ecken von $\mathcal{S}$ in $\mathcal{F}_A$ und $\mathcal{S}$ enthält die Menge $\mathcal{I}_A$. Mit der Wahl $x^{(i)} = \mathsf{S}(:,i+1)$, $i = 1\ldots,s-1$, ist die eine Richtung nachgewiesen.

Seien andersherum zu $x \in \mathcal{F}_A$ die $x^{(i)} \in \mathcal{F}_A$, $i = 1\ldots,s-1$, so gewählt, dass das daraus gebildete Simplex $\mathcal{S}$ alle $w(:,i)$, $i = 1\ldots,k$, enthält. Sei $T$ wie in (2.5) mit $S(i,:)^{\mathbf{T}} = x^{(i)}$, $i = 1,\ldots,s-1$, gebildet. Da $\mathcal{S}$ die Menge $\mathcal{I}_A$ enthält, für welche wiederum $(0,\ldots,0)^{\mathbf{T}} \in \operatorname{int}(\mathcal{I}_A)$ gilt (siehe Lemma 3.13), hat $\mathcal{S}$ ein positives Volumen und es gilt

$$|\det(T)| = \left|\det\left(\begin{pmatrix} 1 & x^{\mathbf{T}} \\ 0 & S(1,:) - x^{\mathbf{T}} \\ \vdots & \vdots \\ 0 & S(s-1,:) - x^{\mathbf{T}} \end{pmatrix}\right)\right| = \left|\det\left(\begin{pmatrix} S(1,:) - x^{\mathbf{T}} \\ \vdots \\ S(s-1,:) - x^{\mathbf{T}} \end{pmatrix}\right)\right| = (s-1)!\,\mathrm{vol}(\mathcal{S}) > 0$$

mit $\mathrm{vol}(\mathcal{S})$ dem Volumen von $\mathcal{S}$. Somit ist $T$ regulär. Weiter sind die mittels $T$ gebildeten Faktoren $C$ und $A$ nach Satz 3.23 nichtnegativ und der Beweis ist vollständig. □

In analoger Weise lässt sich die Aussage von Satz 3.26 auch für die Interpretation einer, im Sinne der Menge $\mathcal{M}_C$, zulässigen Lösung formulieren.

**Korollar 3.27.** *Unter den in Satz 3.26 gemachten Voraussetzungen ist ein $y \in \mathbb{R}^{s-1}$ genau dann zulässig, wenn es in $\mathcal{F}_C$ liegt und es $y^{(i)} \in \mathcal{F}_C$, $i = 1\ldots,s-1$, gibt, sodass das aus diesen $s$ Ecken gebildete Simplex, alle $u(:,j)$, $j = 1\ldots,n$, enthält.*

### 3.3.2 Zulässige Lösungen in den Daten

In Satz 3.26 ist das geometrische Argument für eine zulässige Lösung formuliert, wonach ein $x \in \mathbb{R}^{s-1}$ genau dann zulässig ist, wenn es ein Simplex gibt, dessen Ecken in $\mathcal{F}_A$ liegen und das alle $w(:,i)$, $i = 1,\ldots,k$, enthält. Für $s = 2$ sind stets mindestens zwei (in der Regel genau zwei) der $w(:,i)$ in $\mathcal{M}_A$ enthalten [2, 27, 143, 154]. Für $s \geq 3$ gilt nur unter Umständen $w(:,i) \in \mathcal{M}_A$ für einzelne $i = 1,\ldots,k$, vergleiche Bemerkung 3.29. Es ergibt sich aus den Sätzen 3.15 und 3.23 folgendes einfache Resultat:

**Korollar 3.28.** *Gilt $w(:,i_0) \in \mathcal{M}_A$ für ein $i_0 \in \{1,\ldots,k\}$, so ist $w(:,i_0)$ auf dem Rand von $\mathcal{M}_A$.*

*Beweis.* Angenommen, es sei $w(:,i_0) \in \mathcal{M}_A$, jedoch liege $w(:,i_0)$ nicht auf dem Rand von $\mathcal{M}_A$. Dann gäbe es ein $\gamma$ mit $0 < \gamma < 1$, sodass $\{x \in \mathbb{R}^{s-1} : \ x = \widetilde{\gamma} w(:,i_0)$ für ein $\widetilde{\gamma} \in [\gamma,1]\} \subset \mathcal{M}_A$. Da somit $\gamma w(:,i_0) \in \mathcal{M}_A$ gelten würde, gäbe es ein zulässiges Simplex $\mathcal{S} \subset \mathbb{R}^{s-1}$ mit $\mathsf{S}(:,1) = \gamma w(:,i_0)$. Das Simplex $\mathcal{S}$ enthielte aber nicht $w(:,i_0)$, was in Bezug auf eine nichtnegative Faktorisierung von $D$ im Widerspruch zu Satz 3.23 steht. □

**Bemerkung 3.29.** *Für spektroskopische Daten ist das Auftreten eines $w(:,i_0)$ in $\mathcal{M}_A$ häufig ein Indiz dafür, dass es sich bei dem dazugehörigen Profil tatsächlich um ein gesuchtes Reinspektrum handelt und $C(i_0,:)$ genau einen Eintrag ungleich Null enthält. In der Chemometrie wird ein solches $w(:,i_0)$ auch „pure variable“ genannt [2].*

### 3.3.3 Alternative geometrische Bedingung an eine zulässige Lösung

In diesem Teilabschnitt wird eine weitere geometrische Bedingung, welche ein Simplex und somit insbesondere dessen einzelne Ecken als *zulässig* charakterisiert, vorgestellt. Diese kommt ohne die

$w(:,i)$ und $u(:,j)$ aus und basiert stattdessen auf $\mathcal{F}_A$ und $\mathcal{F}_C$. Zunächst wird folgendes Lemma formuliert:

**Lemma 3.30.** *Seien $T \in \mathbb{R}^{s\times s}$ regulär und von der Form (2.5) sowie $\mathcal{S}$ das zu $C = U\Sigma T^{-1}$ und $A = TV^{\mathbf{T}}$ gehörige Simplex mit den Ecken $\mathsf{S}(:,i) \in \mathbb{R}^{s-1}$, $i = 1,\dots,s$. (Die Forderung $C, A \geq 0$ ist hier nicht gestellt.)*

*Die Einträge der ersten Zeile von $T^{-1}$ sind genau dann positiv, wenn $o = (0,\dots,0)^{\mathbf{T}} \in \mathbb{R}^{s-1}$ im Inneren von $\mathcal{S}$ liegt, das heißt $o$ lässt sich als strikte Konvexkombination von $\mathsf{S}(:,i)$, $i = 1,\dots,s$, darstellen.*

*Beweis.* Der Koeffizientenvektor $z \in \mathbb{R}^s$ sei (die eindeutige) Lösung von $T^{\mathbf{T}}z = e_1$ mit $e_1 = (1,0,\dots,0)^{\mathbf{T}}$. Somit gilt

$$z = (T^{\mathbf{T}})^{-1}e_1 = ((T^{-1})^{\mathbf{T}})(:,1) = (T^{-1}(1,:))^{\mathbf{T}}.$$

Wegen $T^{\mathbf{T}}z = e_1$ gilt weiter $\mathsf{S}z = o$. Damit ist $\mathsf{S}z = o$ genau dann eine strikte Konvexkombination ($z_i > 0\ \forall i$ und $\sum z_i = 1$), wenn $(T^{-1})(1,:)$ komponentenweise positiv ist. □

Eine alternative geometrische Klassifizierung eines $x$ als *zulässig* oder *nicht zulässig* lautet:

**Satz 3.31.** *Seien die Voraussetzungen aus Satz 3.23 erfüllt. Sei weiter das Simplex $\overline{\mathcal{S}} \subset \mathbb{R}^{s-1}$ zu $C$ durch die Ecken $\overline{\mathsf{S}}(:,i)$, $i = 1,\dots,s$, definiert mit*

$$\overline{\mathsf{S}}(:,i) = \frac{(T^{-1})(2:s,i)}{(T^{-1})_{1i}}, \qquad i = 1,\dots,s.$$

*Folgende Aussagen sind äquivalent:*

*a) Die Faktoren $C = U\Sigma T^{-1}$ und $A = TV^{\mathbf{T}}$ sind nichtnegativ.*

*b) Für $\mathcal{S}$ und $\overline{\mathcal{S}}$ sind folgende drei Bedingungen erfüllt:*

*i) die Obermenge $\mathcal{F}_A$ enthält $\mathcal{S}$, das heißt $\mathsf{S}(:,i) \in \mathcal{F}_A$ für alle $i = 1,\dots,s$,*

*ii) die Obermenge $\mathcal{F}_C$ enthält $\overline{\mathcal{S}}$, das heißt $\overline{\mathsf{S}}(:,i) \in \mathcal{F}_C$ für alle $i = 1,\dots,s$, und*

*iii) der Ursprung liegt im Inneren von $\mathcal{S}$, er lässt sich also als strikte Konvexkombination aus $\mathsf{S}(:,i)$, $i = 1,\dots,s$, darstellen.*

*Beweis.* Zunächst wird a) ⇒ b) gezeigt. Die Bedingungen i) und ii) folgen direkt aus $A \geq 0$ und $C \geq 0$. Weiter ist wegen $C \geq 0$ die Zeile $T^{-1}(1,:)$ komponentenweise positiv, siehe Korollar 3.4. Mit Lemma 3.30 folgt iii).
Nun wird b) ⇒ a) gezeigt. Aus i) folgt $A \geq 0$. Weiter folgt nach Lemma 3.30 aus iii), dass $T^{-1}(1,:)$ komponentenweise positiv ist. In Kombination mit ii) folgt $C \geq 0$. □

## 3.4 Nachweis der Kompaktheit

In diesem Abschnitt wird die Kompaktheit (und damit auch die Abgeschlossenheit) der Menge zulässiger Lösungen $\mathcal{M}_A$ unter schwachen Voraussetzungen nachgewiesen. Der Beweis geht auf [123] zurück und es werden die Sätze 3.8 und 3.23 sowie das Lemma 3.13 genutzt.

**Satz 3.32** (Siehe [123])**.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix, $s = \operatorname{rank}(D) = \operatorname{rank}_+(D) \geq 2$ und $D^{\mathbf{T}}D$ sowie $DD^{\mathbf{T}}$ irreduzibel. Sei $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$ mit $V(:,1) > 0$.*

*Zu $D$ gibt es ein $\sigma > 0$, sodass für jede reguläre Matrix $T \in \mathbb{R}^{s\times s}$ der Form (2.5), die mittels (2.2) auf eine nichtnegative Matrixfaktorisierung $D = CA$ führt, gilt, dass*

$$\sigma_s(T) \geq \sigma \tag{3.7}$$

*mit $\sigma_s(T)$ dem kleinsten Singulärwert von $T$.*

*Beweis.* Zum Nachweis der Behauptung wird die Annahme, es gelte (3.7) nur für $\sigma = 0$, zu einem Widerspruch geführt. Falls (3.7) nur für $\sigma = 0$ gelten würde, so gäbe es eine Folge von nichtnegativen Faktorisierungen $D = C^{(i)}A^{(i)}$, $i = 1, 2, \ldots$, mit $A^{(i)} = T^{(i)}V^{\mathbf{T}}$ sowie $T^{(i)}$ jeweils wie in (2.5) und

$$\lim_{i\to\infty} \sigma_s(T^{(i)}) = 0.$$

Es wird nun gezeigt, dass $\sigma_s(T^{(i)})$ unabhängig von $i$ nach unten beschränkt bleibt. Aus Übersichtlichkeitsgründen sei dazu $T := T^{(i)}$ gesetzt. Analog wie im Beweis von Satz 3.26 gilt

$$|\det(T)| = \left|\det\left(\begin{pmatrix} 1 & x^{\mathbf{T}} \\ 0 & S(1,:) - x^{\mathbf{T}} \\ \vdots & \vdots \\ 0 & S(s-1,:) - x^{\mathbf{T}} \end{pmatrix}\right)\right| = \left|\det\left(\begin{pmatrix} S(1,:) - x^{\mathbf{T}} \\ \vdots \\ S(s-1,:) - x^{\mathbf{T}} \end{pmatrix}\right)\right| = (s-1)!\,\mathrm{vol}(\mathcal{S}(T)),$$

mit dem $(s-1)$-dimensionalen Simplex $\mathcal{S}(T) = \mathcal{S}$, das durch die Ecken $x$ und $S(j,:)^{\mathbf{T}}$, $j = 1, \ldots, s-1$, mit $S$ aus $T$ wie in (2.5) definiert ist, und $\mathrm{vol}(\mathcal{S}(T))$ dessen Volumen. Da $T$ zu einer nichtnegativen Faktorisierung von $D$ gehört, enthält das Simplex $\mathcal{S}(T)$ nach Satz 3.23 die Menge $\mathcal{I}_A$. Das Volumen von $\mathcal{I}_A$ ist unabhängig von $i$ und positiv, da nach Lemma 3.13 der Nullpunkt in $\mathcal{I}_A$ enthalten ist und nicht auf dem Rand liegt. Somit enthält $\mathcal{I}_A$ eine Kugel $B_\varepsilon(0)$ mit Radius $\varepsilon > 0$ um den Ursprung. Es gilt

$$\frac{1}{(s-1)!}|\det(T)| = \mathrm{vol}(\mathcal{S}(T)) \geq \mathrm{vol}(\mathcal{I}_A) \geq \mathrm{vol}(B_\varepsilon(0)) > 0.$$

Der Absolutbetrag der Determinante von $T$ ist gleich dem Produkt aller Singulärwerte $\sigma_j(T)$, $j = 1, \ldots, s$, von $T$ und es gilt

$$\frac{1}{(s-1)!}\,|\det(T)| = \frac{1}{(s-1)!}\prod_{j=1}^{s}\sigma_j(T) = \mathrm{vol}(\mathcal{S}(T)) \geq \mathrm{vol}(B_\varepsilon(0)) > 0. \tag{3.8}$$

Um den Widerspruch abzuschließen, bleibt zu zeigen, dass die Singulärwerte auch nach oben und unabhängig von $i$ beschränkt sind. Wegen

$$\|T\|_2 = \sigma_1(T) \geq \sigma_2(T) \geq \cdots \geq \sigma_s(T) \tag{3.9}$$

reicht es zu zeigen, dass $\|T\|_2$ unabhängig von $i$ nach oben beschränkt ist. Für jede Zeile $T(j,:)$ von $T$ gilt wegen $T(j,1) = 1$ und $T(j, 2:s) \in \mathcal{F}_A$, dass

$$\|T(j,:)\|_2 \leq \sqrt{1 + \left(\max_{a\in\mathcal{F}_A}\|a\|\right)^2} =: M.$$

Der Wert $\max_{a\in\mathcal{F}_A}\|a\|$ ist für eine konkrete Matrix $D$ fix und insbesondere endlich, da $\mathcal{F}_A$, unter den gegebenen Voraussetzungen, nach Satz 3.8 eine beschränkte Menge ist. Somit ist auch $M$ eine endliche Konstante und es ergibt sich durch einfache Rechnung

$$\|T\|_2 = \|T^{\mathbf{T}}\|_2 = \max_{b\neq 0}\frac{\|T^{\mathbf{T}}b\|_2}{\|b\|_2} \leq \max_{b\neq 0}\frac{\|b\|_1}{\|b\|_2}M \leq \sqrt{s}M. \tag{3.10}$$

Wegen (3.8) – (3.10) ist somit der kleinste Singulärwert von $T = T^{(i)}$ unabhängig von $i$ beschränkt, was im Widerspruch zur Annahme steht, dass es eine Folge von Zerlegungen mit $\lim_{i\to\infty}\sigma_s(T^{(i)}) = 0$ gibt. Somit ist die Annahme falsch und der Beweis ist vollständig. □

**Bemerkung 3.33.** *Abhängig von $\mathcal{F}_A$ und $\mathcal{I}_A$ lässt sich $\sigma$ aus Satz 3.7 als*

$$\sigma \geq \left( s(1 + \max_{a \in \mathcal{F}_A} \|a\|^2) \right)^{-\frac{s-1}{2}} (s-1)! \operatorname{vol}(\mathcal{I}_A)$$

*mit* $\operatorname{vol}(\mathcal{I}_A)$ *dem Volumen der Menge $\mathcal{I}_A$ angeben.*

Mit Hilfe des Satzes 3.32 lässt sich die Kompaktheit von $\mathcal{M}_A$ zeigen.

**Satz 3.34** (Siehe [123])**.** *Seien die Voraussetzungen aus Satz 3.32 gegeben.*

*Die Menge $\mathcal{M}_A$ ist kompakt (und somit auch abgeschlossen).*

*Beweis.* Betrachtet sei die Menge

$$\mathcal{T} = \left\{ T \in \mathbb{R}^{s \times s} : \ \sigma_s(T) \geq \sigma, \ T(j,:) = (1, (x^{(j)})^{\mathbf{T}}) \text{ mit } x^{(j)} \in \mathcal{F}_A, \ j = 1, \ldots, s \right\}$$

zu einem festen $\sigma > 0$ wie in Satz 3.32. Die Menge $\mathcal{T}$ ist kompakt, da $\mathcal{F}_A$ kompakt ist (Beschränktheit ist in Satz 3.8 gezeigt, Abgeschlossenheit ist wegen Nichtnegativitätsrestriktionen klar) und $\sigma_s(T) \geq \sigma > 0$ gefordert wird. Die Abschätzungen $\sigma_s(T) \geq \sigma > 0$ sind in Satz 3.32 gezeigt und gelten für jede nichtnegative Matrixfaktorisierung von $D$, die der speziellen Skalierungsform von $T$ genügt. Durch die Abschätzungen wird der Fall ausgeschlossen, dass eine Folge von regulären Matrizen $\{T^{(i)}\}_{i=1,2,\ldots}$ gegen eine singuläre Matrix konvergiert. Die Bildmenge von $\mathcal{T}$ bei Anwendung der (stetigen) Abbildung der Matrixinversenbildung

$$\overline{\mathcal{T}} = \{T^{-1} : \ T \in \mathcal{T}\}$$

ist ebenfalls kompakt. Weiter ist die Teilmenge

$$\widehat{\mathcal{T}} = \{T^{-1} \in \overline{\mathcal{T}} : \ U \Sigma T^{-1} \geq 0\}$$

von $\overline{\mathcal{T}}$, der Matrizen, die zusätzlich $U\Sigma T^{-1} \geq 0$ erfüllen, abgeschlossen, da Nichtnegativitätsrestriktionen angewendet werden. Somit ist der Schnitt

$$\mathcal{T}^* = \mathcal{T} \cap \{T : \ T^{-1} \in \widehat{\mathcal{T}}\}$$

kompakt. Weiter lässt sich $\mathcal{M}_A$ schreiben als

$$\mathcal{M}_A = \left\{ x \in \mathbb{R}^{s-1} : \exists T \in \mathcal{T}^* \text{ und } i \in \{1, \ldots, s\} \text{ mit } x^{\mathbf{T}} = T(i, 2:s) \right\}$$

und somit ist auch die Menge $\mathcal{M}_A$ kompakt (und damit auch abgeschlossen). □

## 3.5 Blockstruktur der Faktorisierung

Die Irreduzibilität der Matrix $D^{\mathbf{T}}D$ ist eine wichtige Voraussetzung, dass der Ansatz aus (2.5) anwendbar ist. Weiter folgt daraus die Beschränktheit der Menge zulässiger Lösungen. Für die in Satz 3.10 untersuchte Eigenschaft, dass $(0, \ldots, 0)^{\mathbf{T}} \notin \mathcal{M}_A$ gilt, wird die Irreduzibilität von $DD^{\mathbf{T}}$ gebraucht. In diesem Abschnitt sollen die Zusammenhänge zwischen der Irreduzibilität von $D^{\mathbf{T}}D$, der Irreduzibilität von $DD^{\mathbf{T}}$ und einer Blockstruktur in den Faktoren $C$ und $A$ untersucht werden.

**Bemerkung 3.35.** *In diesem Abschnitt werden an einigen Stellen Nullmatrizen geeigneter Dimensionen genutzt. Aus Übersichtlichkeitsgründen werden diese durch eine* 0 *dargestellt, wobei sich die jeweilige Dimension eindeutig aus dem Zusammenhang ergibt.*

### 3.5.1 Vorüberlegungen

Um die Beweise zu vereinfachen, werden zunächst zwei Lemmata aufgestellt.

**Lemma 3.36.** *Sei $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix, die keine Nullzeile enthält und zu der es einen Index $k_1 \in \{1,\dots,k-1\}$ gibt mit*

$$\underbrace{D(1:k_1,:)}_{\in\mathbb{R}^{k_1\times n}}\underbrace{(D(k_1+1:k,:))}_{\in\mathbb{R}^{(k-k_1)\times n}}^{\mathbf{T}} = 0 \in \mathbb{R}^{k_1\times(k-k_1)}.$$

*Es gibt eine Zerlegung der Indexmenge $\{1,\dots,n\}$ in nichtleere disjunkte Teilmengen $I_1$ und $I_2$ mit $D(1:k_1,I_1)=0$ und $D(k_1+1:k,I_2)=0$.*

*Beweis.* Sei angenommen, es gäbe keine Zerlegung von $\{1,\dots,n\}$ in disjunkte Mengen $I_1$ und $I_2$. Das heißt es gäbe Indizes $i_1,\in\{1,\dots,n\}$, $j_1\in\{1,\dots,k_1\}$ und $j_2\in\{k_1+1,\dots,k\}$ mit $D_{j_1i_1}>0$ und $D_{j_2i_1}>0$. Dies würde bedeuten, dass

$$D(j_1,:)(D(j_2,:))^{\mathbf{T}} \geq D_{j_1i_1}D_{j_2i_1} > 0,$$

was der Voraussetzung widerspricht. Bleibt $I_1, I_2 \neq \emptyset$ zu zeigen. Ohne Beschränkung der Allgemeinheit sei $I_2=\emptyset$ angenommen, dann wäre jedoch $I_1=\{1,\dots,n\}$ und $D(1:k_1,:)=0$, was der Voraussetzung widerspricht, dass $D$ keine Nullzeile enthält. □

Mittels Lemma 3.36 lässt sich das folgende Lemma 3.37 über den Zusammenhang zwischen der Struktur von $D$ und der Irreduzibilität von $DD^{\mathbf{T}}$ zeigen.

**Lemma 3.37.** *Seien $D\in\mathbb{R}^{k\times n}$ eine nichtnegative Matrix, welche keine Nullzeile enthält, und $k\geq 2$.*

*Es ist $DD^{\mathbf{T}}$ genau dann reduzibel, wenn es Permutationsmatrizen $P\in\mathbb{R}^{k\times k}$ und $Q\in\mathbb{R}^{n\times n}$ gibt, sodass*

$$PDQ = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} \tag{3.11}$$

*mit $D_1\in\mathbb{R}^{k_1\times n_1}$ und $D_2\in\mathbb{R}^{(k-k_1)\times(n-n_1)}$ sowie $0<k_1<k$ und $0<n_1<n$.*

*Beweis.* Sei zunächst angenommen, es gibt solche $P,Q$, dass (3.11) gilt. Dann ist

$$PDD^{\mathbf{T}}P^{\mathbf{T}} = \begin{pmatrix} D_1D_1^{\mathbf{T}} & 0 \\ 0 & D_2D_2^{\mathbf{T}} \end{pmatrix}$$

und $DD^{\mathbf{T}}$ ist reduzibel.

Sei andersherum $DD^{\mathbf{T}}$ reduzibel, d.h. es gibt eine Permutationsmatrix $P$, sodass

$$PDD^{\mathbf{T}}P^{\mathbf{T}} = \begin{pmatrix} \widetilde{D}_1 & 0 \\ 0 & \widetilde{D}_2 \end{pmatrix}$$

mit $\widetilde{D}_1\in\mathbb{R}^{k_1\times k_1}$ sowie $\widetilde{D}_2\in\mathbb{R}^{(k-k_1)\times(k-k_1)}$ und $0<k_1<k$. Dabei sind $\|\widetilde{D}_1\|,\|\widetilde{D}_2\|>0$, da $D$ keine Nullzeile enthält. Weiter folgt

$$(PD)(1:k_1,:)((PD)(k_1+1:k,:))^{\mathbf{T}} = 0 \in \mathbb{R}^{k_1\times(k-k_1)}$$

und nach Lemma 3.36 gilt, dass es eine Zerlegung von $\{1,\dots,n\}$ in nichtleere disjunkte Mengen $I_1$ und $I_2$ gibt, sodass $(PD)(1:k_1,I_1)=0$ und $(PD)(k_1+1:k,I_2)=0$. (Weiter ist damit klar,

dass sich $\widetilde{D}_1$ mithilfe einer nichtnegativen Matrix $D_1$ als $\widetilde{D}_1 = D_1 D_1^{\mathbf{T}}$ faktorisieren lässt.) Seien $n_1 = |I_1|$ sowie $Q \in \mathbb{R}^{n \times n}$ eine Permutationsmatrix mit

$$Q(I_1, n - n_1 + 1 : n) = \operatorname{diag}(1, \ldots, 1), \qquad Q(I_2, 1 : n - n_1) = \operatorname{diag}(1, \ldots, 1).$$

Daraus folgt

$$PDQ = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}$$

sowie $D_i D_i^{\mathbf{T}} = \widetilde{D}_i$ für $i \in \{1, 2\}$ und der Beweis ist vollständig. □

### 3.5.2 Irreduzibilitäten von $D^{\mathbf{T}}D$ und $DD^{\mathbf{T}}$

Als wichtiges Resultat aus Lemma 3.37 lässt sich eine Aussage über den Zusammenhang zwischen den Irreduzibilitäten von $D^{\mathbf{T}}D$ und $DD^{\mathbf{T}}$ treffen.

**Satz 3.38.** *Sei $D \in \mathbb{R}^{k \times n}$ eine nichtnegative Matrix, die keine Nullzeile und keine Nullspalte enthält.*

*Es ist $D^{\mathbf{T}}D$ genau dann reduzibel, wenn $DD^{\mathbf{T}}$ reduzibel ist.*

*Beweis.* Nach Lemma 3.37 ist $DD^{\mathbf{T}}$ genau dann reduzibel, wenn es Permutationsmatrizen $P \in \mathbb{R}^{k \times k}$ und $Q \in \mathbb{R}^{n \times n}$ gibt, sodass (3.11) gilt. Damit gilt weiter

$$Q^{\mathbf{T}} D^{\mathbf{T}} P^{\mathbf{T}} = \begin{pmatrix} D_1^{\mathbf{T}} & 0 \\ 0 & D_2^{\mathbf{T}} \end{pmatrix}.$$

Dies ist äquivalent dazu, dass $D^{\mathbf{T}}D$ reduzibel ist (Anwendung von Satz 3.37 auf $D^{\mathbf{T}}$). □

### 3.5.3 Irreduzibilität von $D^{\mathbf{T}}D$ und Blockstruktur in der Faktorisierung

Inwiefern die Eigenschaft, dass es Permutationsmatrizen $P$ und $Q$ gibt, die auf die Darstellung (3.11) führen, von den Strukturen der Faktoren $C$ und $A$ abhängt, wird in Satz 3.40 gezeigt.

**Lemma 3.39.** *Seien $D \in \mathbb{R}^{k \times n}$ eine nichtnegative Matrix, welche keine Nullzeile und keine Nullspalte enthält, und $s = \operatorname{rank}(D) = \operatorname{rank}_+(D) \geq 2$. Weiter gelte (3.11) mit den zwei Permutationsmatrizen $P \in \mathbb{R}^{k \times k}$ und $Q \in \mathbb{R}^{n \times n}$ und es sei mit $C \in \mathbb{R}^{k \times s}$ und $A \in \mathbb{R}^{s \times n}$ eine nichtnegative Matrixfaktorisierung von $D$ gegeben.*

*Es gibt eine Zerlegung der Menge $\{1, \ldots, s\}$ in nichtleere disjunkte Teilmengen $I_1$ und $I_2$, sodass*

$$\begin{aligned} &(PC)(1 : k_1, I_2) = 0, && (AQ)(I_2, 1 : n_1) = 0, \\ &(PC)(k_1 + 1 : k, I_1) = 0, && (AQ)(I_1, n_1 + 1 : n) = 0. \end{aligned} \tag{3.12}$$

*Beweis.* Der Beweis wird in zwei Schritten vollzogen. Zunächst wird gezeigt, dass es eine solche Zerlegung in disjunkte $I_1$ und $I_2$ geben muss, danach wird gezeigt, dass die Mengen nicht leer sind.

Angenommen, es gäbe keine solche Zerlegung in $I_1$ und $I_2$. Dann gäbe es vier mögliche Fälle zu betrachten, wovon jeweils zwei analog zueinander sind. Grundsätzlich sind zu unterscheiden:

a) es gibt Indizes $i \in \{1, \ldots, s\}$, $j_1 \in \{1, \ldots, k_1\}$ und $j_2 \in \{k_1 + 1, \ldots, k\}$ mit $(PC)_{j_1 i} > 0$ und $(PC)_{j_2 i} > 0$ oder

b) es gibt Indizes $i \in \{1,\dots,s\}$, $j_1 \in \{1,\dots,k_1\}$ und $j_2 \in \{n_1+1,\dots,n\}$ mit $(PC)_{j_1 i} > 0$ und $(AQ)_{ij_2} > 0$.

Im Fall a) würde gelten, dass

$$\underbrace{(PC)_{j_1 i}}_{>0}(AQ)(i, n_1+1:n) \le (PDQ)(j_1, n_1+1:n) = 0,$$
$$\underbrace{(PC)_{j_2 i}}_{>0}(AQ)(i, 1:n_1) \le (PDQ)(j_2, 1:n_1) = 0,$$

was zur Folge hätte, dass $(AQ)(i,:) = 0$ gelten würde und somit $A$ nicht den vollen Rang hätte. Der Fall b) führt auf den Widerspruch

$$0 < (PC)_{j_1 i}(AQ)_{ij_2} \le (PCAQ)_{j_1 j_2} = 0.$$

Damit sind beide Fälle ausgeschlossen und die Annahme muss falsch sein.

Bleibt $I_1, I_2 \neq \emptyset$ zu zeigen. Angenommen $I_1 = \emptyset$. Dann ist jedoch $(PC)(1:k_1,:) = 0$ und es gilt $(PD)(1:k_1,:) = 0$. Dies ist ein Widerspruch zur Annahme, dass $D$ keine Nullzeile enthält. Analog zeigt sich $I_2 \neq \emptyset$ und der Beweis ist vollständig. □

Mit Lemma 3.39 lässt sich der folgende wichtige Satz 3.40 über den Zusammenhang zwischen der Irreduzibilität von $D^{\mathbf{T}}D$ und den Zerfall der Faktorisierung in eine Blockstruktur formulieren.

**Satz 3.40.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix, welche keine Nullzeile und keine Nullspalte enthält, und $s = \operatorname{rank}(D) = \operatorname{rank}_+(D) \ge 2$. Weiter seien $C \in \mathbb{R}^{k\times s}$ und $A \in \mathbb{R}^{s\times n}$ Faktoren einer nichtnegativen Matrixfaktorisierung von $D$.*

*Äquivalent sind:*

*a) Es gibt Permutationsmatrizen $P \in \mathbb{R}^{k\times k}$, $Q \in \mathbb{R}^{n\times n}$, sodass (3.11) gilt.*

*b) Es ist $DD^{\mathbf{T}}$ reduzibel.*

*c) Es ist $D^{\mathbf{T}}D$ reduzibel.*

*d) Es gibt Permutationsmatrizen $P \in \mathbb{R}^{k\times k}$, $Q \in \mathbb{R}^{n\times n}$ und $R \in \mathbb{R}^{s\times s}$ sowie Indizes $s_1, k_1, n_1$ mit $0 < s_1 < s$, $0 < k_1 < k$ und $0 < n_1 < n$, sodass*

$$PCR = \begin{pmatrix} C_1 & 0 \\ 0 & C_2 \end{pmatrix}, \quad R^{\mathbf{T}}AQ = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} \tag{3.13}$$

*mit $C_1 \in \mathbb{R}^{k_1\times s_1}$, $C_2 \in \mathbb{R}^{(k-k_1)\times(s-s_1)}$, $A_1 \in \mathbb{R}^{s_1\times n_1}$ und $A_2 \in \mathbb{R}^{(s-s_1)\times(n-n_1)}$ gelten.*

*Beweis.* Die Äquivalenzen von a), b) und c) sind in Lemma 3.37 und Satz 3.38 gezeigt. Dass a) aus d) folgt, ist trivial. Aus der Existenz einer Permutationsmatrix $R$, sodass (3.13) erfüllt ist, folgt, dass mit $D = CA$ (3.11) gilt.

Sei nun a) erfüllt. Nach Lemma 3.39 gibt es eine Zerlegung von $\{1,\dots,s\}$ in nichtleere disjunkte Teilmengen $I_1$ und $I_2$, für die (3.12) gilt. Die durch $R(I_1, 1:s_1) = \operatorname{diag}(1,\dots,1)$ und $R(I_2, s_1+1:s) = \operatorname{diag}(1,\dots,1)$ mit $s_1 = |I_1|$ definierte Permutationsmatrix $R \in \mathbb{R}^{s\times s}$ führt zu (3.13). □

**Bemerkung 3.41.** *Das Auftreten von Nullspalten oder/und von Nullzeilen in $D$ ist unproblematisch.[5] Für eine nichtnegative Faktorisierung $D = CA$ folgt aus $D(:, j_0) = (0,\dots,0)^{\mathbf{T}}$, dass $A(:, j_0) = (0,\dots,0)^{\mathbf{T}}$ gilt, siehe folgendes Lemma. Analog folgt $C(i_0, :) = (0,\dots,0)$ aus $D(i_0, :) = (0,\dots,0)$. Sofern $D$ Nullspalten oder/und -zeilen enthält, lassen sich diese streichen und die Menge zulässiger Lösungen wird zu der verkleinerten Matrix berechnet.*

[5] Enthält $D$ eine Nullspalte, so ist $D^{\mathbf{T}}D$ reduzibel. Enthält $D$ eine Nullzeile, so ist $DD^{\mathbf{T}}$ reduzibel.

**Lemma 3.42.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix, $s = \operatorname{rank}(D) = \operatorname{rank}_+(D) \geq 2$ und $C \in \mathbb{R}_+^{k\times s}$, $A \in \mathbb{R}_+^{s\times n}$ mit $D = CA$.*

1. *Ist $D(:, j_0) = (0, \ldots, 0)^{\mathbf{T}}$ für ein $j_0 \in \{1, \ldots, n\}$, so gilt $A(:, j_0) = (0, \ldots, 0)^{\mathbf{T}}$.*
2. *Analog gilt $C(i_0, :) = (0, \ldots, 0)$, falls $D(i_0, :) = (0, \ldots, 0)$ für ein $i_0 \in \{1, \ldots, k\}$.*

*Beweis.* Zum Beweis der ersten Aussage sei angenommen, es gelte $A_{\ell_0 j_0} > 0$ für ein $\ell_0 \in \{1, \ldots, s\}$. Wegen $A, C \geq 0$ und $0 = D_{ij_0} = \sum_{\ell=1}^{s} C_{i\ell} A_{\ell j_0} \geq C_{i\ell_0} A_{\ell_0 j_0}$, $i = 1, \ldots, k$, würde daraus $C(:, \ell_0) = (0, \ldots, 0)^{\mathbf{T}}$ folgen. Dies widerspricht $\operatorname{rank}(C) = \operatorname{rank}(D) = s$ und es muss $A(:, j_0) = (0, \ldots, 0)^{\mathbf{T}}$ gelten. Analog lässt sich zeigen, dass die $i_0$te Zeile von $C$ eine Nullspalte ist, falls die $i_0$te Zeile von $D$ eine Nullspalte ist. □

**Bemerkung 3.43.**

1. *Falls es für eine konkrete Faktorisierung $CA$ keine Permutationsmatrizen $P, Q$ und $R$ gibt, sodass $C$ und $A$ in die Form aus (3.13) gebracht werden können, so gibt es auch keine andere nichtnegative Faktorisierung von $D$ mit Faktoren für die dies so wäre.*
2. *Gibt es Permutationsmatrizen $P$ und $Q$, sodass (3.11) gilt, so lassen sich die jeweiligen Blöcke der Faktorisierungsaufgabe zu $PDQ$ einzeln betrachten. Sofern diese nicht in weitere Blöcke zerlegbar sind, lassen sich diese Subsysteme einzeln untersuchen und es ergeben sich dazu jeweils Mengen zulässiger Lösungen.*
3. *Die Voraussetzung der Nullzeilen- und Nullspaltenfreiheit ist nicht einschränkend, vergleiche Bemerkung 3.41 und Lemma 3.42.*
4. *Weiter ist auch die Bedingung $s \geq 2$ nicht einschränkend, da die Faktorisierung einer Rang-1-Matrix bei Vernachlässigung der Skalierungsmehrdeutigkeit eindeutig ist.*

## 3.6 Reduktionen mittels des Dualitätsprinzips

Bei der Komplementaritäts- und Kopplungstheorie [122, 143, 145, 148, 151, 156], auch Dualitätsprinzip oder -theorie genannt [12, 69, 133], geht es um die Zusammenhänge zwischen einzelnen Teilen des einen und einzelnen Teilen des anderen Faktors. Das Ziel ist es dabei, mögliche Zusatzinformationen in die Berechnung einer Faktorisierung einfließen zu lassen und so die Freiheitsgrade der Zerlegung maximal einzuschränken. Im Fokus stehen die Beziehungen zwischen den Spalten von $C$ und den Zeilen von $A$ in dem Sinne, welche Restriktionen sich aus der Kenntnis einer Spalte von $C$ für die Zeilen von $A$ und welche Restriktionen sich aus der Kenntnis einer Zeile von $A$ für die Spalten von $C$ ergeben.

**Bemerkung 3.44.** *Die in [122, 143, 145, 148, 151, 156] eingeführten Begriffe „komplementär“ und „Komplementaritätstheorie“ werden im weiteren Verlauf dieser Schrift nicht genutzt. Stattdessen werden die Bezeichnungen „dual“ und „Dualitätsprinzip“ aus [12, 69, 133, 148, 153] verwendet. Hinter beiden Bezeichnungen stehen dieselben Zusammenhänge.*

In [145] ist das Dualitätsprinzip ohne Bezug zu den Mengen zulässiger Lösungen beschrieben. Inwiefern sich $\mathcal{M}_A$ und $\mathcal{M}_C$ mittels dualer Beziehungen reduzieren lassen, wird für $s = 3$ in [13, 69, 133] sowie allgemein für $s \geq 3$ in [148, 156] untersucht. Das Hauptresultat aus [156] liefert einen Zusammenhang zwischen Elementen von $\mathcal{M}_A$ und $(s-2)$-dimensionalen affinen Unterräumen, welche $\mathcal{M}_C$ schneiden, beziehungsweise zwischen Elementen aus $\mathcal{M}_C$ und $(s-2)$-dimensionalen affinen Unterräumen, welche $\mathcal{M}_A$ schneiden.

### 3.6.1 Allgemeine Einschränkungen für die Faktoren

Das Dualitätsprinzip wird zusammen mit dem Kopplungsargument in [143,145] untersucht, vergleiche auch [13, 68, 118, 122]. Bezüglich der sich durch teilweise Kenntnis des einen Faktors ergebenden Restriktionen des anderen Faktors wird zwischen zwei Arten unterschieden. Ist etwa für die Berechnung einer Faktorisierung eine Zeile von $A$ fixiert, so ergeben sich Einschränkungen sowohl für die dazugehörige Spalte in $C$ als auch für die verbleibenden anderen Spalten von $C$. Im Folgenden werden nur die Restriktionen für die verbleibenden Spalten von $C$ betrachtet. Diese lassen sich über die Konstruktion von $C$ und $A$ mittels einer abgeschnittenen Singulärwertzerlegung $C = U\Sigma T^{-1}$, $A = TV^{\mathbf{T}}$ herleiten. Die Kenntnis einer Zeile von $A$ legt die entsprechende Zeile von $T$ fest, was die verbleibenden Spalten deren Inversen eingeschränkt. So sind auch die entsprechenden Spalten in $C$ eingeschränkt. Analog ergeben sich zu einer fixierten Spalte von $C$ sowohl Einschränkungen für die dazugehörige Zeile von $A$ als auch für die verbleibenden Zeilen von $A$.

**Satz 3.45.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix und $s = \operatorname{rank}(D) = \operatorname{rank}_+(D)$. Ferner seien von der gesuchten Faktorisierung bereits $A(1:s_0,:)$ mit $1 \leq s_0 < s$ bekannt.*

*Die Spalten $C(:,i)$, $i = s_0+1,\ldots,s$, liegen in dem $(s-s_0)$-dimensionalen linearen Unterraum*

$$\left\{c \in \mathbb{R}^k :\ \exists t \in \mathbb{R}^s \setminus \{0\} \text{ mit } c = U\Sigma t \text{ und } A(j,:)Vt = 0,\ j = 1,\ldots,s_0\right\}. \tag{3.14}$$

*Beweis.* Siehe [143,145]. □

### 3.6.2 Reduktionen der Mengen zulässiger Lösungen

Die Aussage aus Satz 3.45 ist unabhängig von den Restriktionen der Mengen zulässiger Lösungen und im Folgenden wird untersucht, inwiefern die Restriktionen aus Satz 3.45 auch die Mengen zulässiger Lösungen einschränken. In Bezug auf $\mathcal{M}_A$ und $\mathcal{M}_C$ überträgt sich die Aussage aus Satz 3.45 auf die niedrigdimensionalen Darstellungen der Zeilen von $A$ und der Spalten von $C$.

Sei dazu $D = CA$ eine nichtnegative Faktorisierung, wobei es ein reguläres $T \in \mathbb{R}^{s\times s}$ gibt mit $A = TV^{\mathbf{T}}$ und $T(:,1) = (1,\ldots,1)^{\mathbf{T}}$. Dazu seien

$$x^{(i)} = T(i,2:s)^{\mathbf{T}}, \qquad i = 1,\ldots,s, \tag{3.15}$$

die niedrigdimensionalen Darstellungen der Zeilen von $A$ mit $x^{(i)} \in \mathcal{M}_A$ und

$$y^{(j)} = \frac{(T^{-1})(2:s,j)}{(T^{-1})_{1j}}, \qquad j = 1,\ldots,s, \tag{3.16}$$

die niedrigdimensionalen Darstellungen der Spalten von $C$ mit $y^{(j)} \in \mathcal{M}_C$. Mit der Annahme $T(:,1) = (1,\ldots,1)^{\mathbf{T}}$ werden mit Rücksicht auf die Skalierungsmehrdeutigkeit qualitativ keine anderen nichtnegativen Faktorisierungen $D = \widetilde{C}\widetilde{A}$ ausgeschlossen. Für eine allgemeine nichtnegative Faktorisierung $D = \widetilde{C}\widetilde{A}$ führt eine Umskalierung der Zeilen von $\widetilde{A}$ mittels $R = \operatorname{diag}(\widetilde{A}V(:,1))^{-1}$ und der Spalten von $\widetilde{C}$ mittels $R^{-1}$ auf eine Faktorisierung $CA = \widetilde{C}R^{-1}R\widetilde{A}$, für welche $T$ von entsprechender Form ist, vergleiche Bemerkung 3.22. Zentral ist:

**Satz 3.46.** *Seien die Voraussetzungen aus Satz 3.10 erfüllt und $D = CA$ eine nichtnegative Matrixfaktorisierung mit $A = TV^{\mathbf{T}}$ für ein reguläres $T \in \mathbb{R}^{s\times s}$ mit $T(:,1) = (1,\ldots,1)^{\mathbf{T}}$. Weiter seien $x^{(i)}$, $i = 1,\ldots,s$, und $y^{(j)}$, $j = 1,\ldots,s$, wie in (3.15) und (3.16) definiert.*

*Für $i \neq j$ gilt $(x^{(i)})^{\mathbf{T}}y^{(j)} = -1$.*

*Beweis.* Sei $T = AV$. Für $i \neq j$ gilt $T(i,:)(T^{-1})(:,j) = 0$. Mit $T(i,:) = (1,(x^{(i)})^{\mathbf{T}})$ und $(T^{-1})(:,j) = (T^{-1})_{1j}(1,(y^{(j)})^{\mathbf{T}})^{\mathbf{T}}$ folgt

$$0 = (1,(x^{(i)})^{\mathbf{T}})((T^{-1})_{1j}(1,(y^{(j)})^{\mathbf{T}})^{\mathbf{T}}) = (1,(x^{(i)})^{\mathbf{T}})(1,(y^{(j)})^{\mathbf{T}})^{\mathbf{T}} = 1 + (x^{(i)})^{\mathbf{T}}y^{(j)}.$$

□

**Definition 3.47.** *Ein $x \in \mathbb{R}^{s-1} \setminus \{0\}$ und eine durch $x_E \in \mathbb{R}^{s-1} \setminus \{0\}$ und $\delta \in \mathbb{R} \setminus \{0\}$ definierte affine Hyperebene $E = \{z \in \mathbb{R}^{s-1} : x_E^{\mathbf{T}} z = \delta\}$ heißen dual zueinander, wenn $x_E = -\delta x$ gilt.*

Wird der Faktor $A$ (beziehungsweise $C$) einer Faktorisierung sequenziell zeilenweise (spaltenweise) konstruiert, so werden sukzessive die niedrigdimensionalen Darstellungen der einzelnen Zeilen (Spalten) fixiert. Inwiefern sich durch ein fixiertes $x^{(i_1)}$ Einschränkungen für $y^{(j)}$, $j = 1,\ldots,s$, $j \neq i_1$, und durch ein fixiertes $y^{(j_1)}$ Einschränkungen für $x^{(i)}$, $i = 1,\ldots,s$, $i \neq j_1$, in Form von Teilmengen von affinen Hyperebenen ergeben, zeigt:

**Korollar 3.48.** *Seien die Voraussetzungen aus Satz 3.46 erfüllt, $i_1, j_1 \in \{1,\ldots,s\}$ und*

$$\mathcal{M}_C^{[i_1]} = \left\{ y \in \mathcal{M}_C : \ (x^{(i_1)})^{\mathbf{T}} y = -1 \right\}, \qquad \mathcal{M}_A^{[j_1]} = \left\{ x \in \mathcal{M}_A : \ x^{\mathbf{T}} y^{(j_1)} = -1 \right\}. \tag{3.17}$$

*Es gelten:*

1. *Für $j \in \{1,\ldots,s\} \setminus \{i_1\}$ ist $y^{(j)} \in \mathcal{M}_C^{[i_1]}$. Somit liegen $y^{(j)}$, $j = 1,\ldots,s$, $j \neq i_1$, in dem Schnitt von $\mathcal{M}_C$ mit der zu $x^{(i_1)}$ dualen affinen Hyperebene.*
2. *Für $i \in \{1,\ldots,s\} \setminus \{j_1\}$ ist $x^{(i)} \in \mathcal{M}_A^{[j_1]}$. Somit liegen $x^{(i)}$, $i = 1,\ldots,s$, $i \neq j_1$, in dem Schnitt von $\mathcal{M}_A$ mit der zu $y^{(j_1)}$ dualen affinen Hyperebene.*

Die Mengen $\mathcal{M}_C^{[i_1]}$ beziehungsweise $\mathcal{M}_A^{[j_1]}$ sind Schnitte von $(s-2)$-dimensionalen affinen Unterräumen (affine Hyperebenen im $\mathbb{R}^{s-1}$) mit $\mathcal{M}_C$ beziehungsweise $\mathcal{M}_A$.

**Bemerkung 3.49.**

1. *Die zu $x^{(i_1)}$ duale Hyperebene $\{y \in \mathbb{R}^{s-1} : \sum_{\ell=1}^{s-1} x_\ell^{(i_1)} y_\ell = -1\}$ liefert eine $(s-2)$-dimensionale affine Einschränkung für alle $y^{(j)}$ mit $j \neq i_1$. Deren Schnitt mit $\mathcal{M}_C$ garantiert zudem, dass die $y^{(j)}$ im Sinne der niedrigdimensionalen Darstellung auch zu nichtnegativen Matrixfaktorisierungen gehören. Dies führt auf $\mathcal{M}_C^{[i_1]}$ aus (3.17).*
2. *Korollar 3.48 liefert somit Aussagen, inwiefern Elemente der Menge zulässiger Lösungen des einen Faktors mit Teilmengen der Menge zulässiger Lösungen des anderen Faktors verknüpft sind. Für $s = 3$ sind zulässige Lösungen mit Geradenabschnitten verknüpft [13, 156]. Ein Beispiel für eine solche Verknüpfung ist in Abbildung 3.1 dargestellt. Für $s = 4$ sind beispielsweise zulässige Lösungen mit Teilmengen von Ebenen verknüpft [156].*
3. *Sind mehrere $x^{(i_1)}, x^{(i_2)}, \ldots, x^{(i_{s_0})}$ in $\mathcal{M}_A$ mit $s_0 < s$ fixiert, so ergibt sich für $y^{(j)}$, $j \in \{1,\ldots,s\} \setminus \{i_1,\ldots,i_{s_0}\}$, eine Einschränkung in Form einer Teilmenge einer $(s-s_0-1)$-dimensionalen affinen Menge und für $y^{(j)}$, $j \in \{i_1,\ldots,i_{s_0}\}$ ergeben sich jeweils Einschränkungen in Form von Teilmengen von $(s-s_0)$-dimensionalen affinen Mengen.*
4. *Durch $x^{(i_1)}$ ergeben sich auch für die zugeordnete Lösung $y^{(i_1)}$ Einschränkungen. Diese Restriktionen werden in dieser Schrift nur numerisch untersucht, siehe Abschnitt 4.8.2.*

### 3.6.3 Dualität der Randflächen von $\mathcal{I}_A$ und der Ecken von $\mathcal{F}_A$

Für die Anwendung geometrischer Argumente in Bezug auf $\mathcal{M}_A$ werden die Mengen $\mathcal{F}_A$ und $\mathcal{I}_A$ benötigt. Analog basiert die Nutzung geometrischer Argumente in Bezug auf $\mathcal{M}_C$ auf den

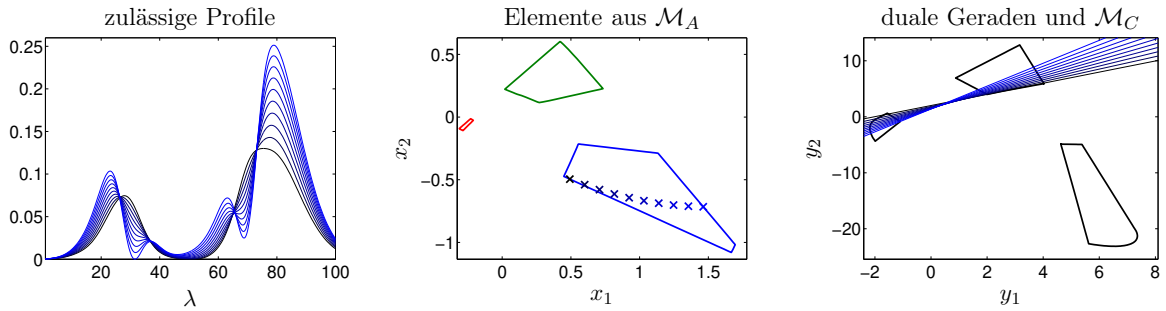


**Abbildung 3.1:** Illustration zu Korollar 3.48 bezüglich des Datensatzes 2. Links: eine Reihe zulässiger Reinkomponentenspektren. Mitte: die zugehörigen Darstellungen in $\mathcal{M}_A$. Rechts: die Menge $\mathcal{M}_C$ und die zu den zulässigen Lösungen aus $\mathcal{M}_A$ (mitte) dualen Geraden.

Mengen $\mathcal{F}_C$ und $\mathcal{I}_C$. Eine direkte Möglichkeit zur Berechnung der Ecken von $\mathcal{F}_A$ funktioniert wie folgt: Es werden alle $\binom{n}{s-1}$ möglichen Schnittpunkte (von jeweils $s-1$) der $n$ affinen Hyperebenen zu den Halbräumen aus (3.2) berechnet. Dazu wird jeweils überprüft, ob der berechnete Schnittpunkt in $\mathcal{F}_A$ liegt oder nicht. So teilt sich die Indexmenge der Schnittpunkte (von jeweils $s-1$ affinen Hyperebenen) in eine Menge *relevanter Punkte* und eine Menge *irrelevanter Punkte* auf, wobei $\mathcal{F}_A$ die konvexe Hülle der relevanten Schnittpunkte ist. Eine andere Variante, siehe [86], ist es, $\mathcal{F}_A$ einzuschachteln. Dabei werden zunächst nur $s$ der affinen Halbräume, aus denen sich $\mathcal{F}_A$ ergibt, betrachtet. Diese sind so gewählt, dass sich ein Polytop ergibt. Dieses enthält $\mathcal{F}_A$. Anschließend werden sukzessive auch die weiteren affinen Halbräume berücksichtigt. Mit jeder Hinzunahme eines weiteren affinen Halbraums verkleinert sich das aktuelle Polytop unter Umständen. Schlussendlich führt dieses Vorgehen auf $\mathcal{F}_A$.

Bei beiden Varianten werden häufig viele Rechnungen unnötigerweise ausgeführt. So gehören etwa bei der erstgenannten Möglichkeit oft viele der untersuchten Schnittpunkte nicht zu $\mathcal{F}_A$ und sind somit unnötig berechnet worden. Im Folgenden wird eine elegante und schnelle Alternative zur Bestimmung der Ecken von $\mathcal{F}_A$ vorgestellt. Dazu seien folgende affine Hyperebenen definiert

$$E_j^{(A)} = \left\{ x \in \mathbb{R}^{s-1} : \frac{V(j, 2:s)x}{V_{j1}} = -1 \right\}, \qquad j = 1, \ldots, n, \tag{3.18}$$

$$E_i^{(C)} = \left\{ y \in \mathbb{R}^{s-1} : \frac{U(i, :)\Sigma(:, 2:s)y}{\sigma_1 U_{i1}} = -1 \right\}, \qquad i = 1, \ldots, k. \tag{3.19}$$

Die affinen Hyperebenen aus (3.18) legen die affinen Halbräume fest, aus denen sich $\mathcal{F}_A$ ergibt und analog legen die affinen Hyperebenen aus (3.19) die affinen Halbräume fest, aus denen sich $\mathcal{F}_C$ ergibt. Häufig begrenzen sowohl nur einige der $n$ affinen Hyperebenen aus (3.18) die Menge $\mathcal{F}_A$ als auch nur einige der $k$ affinen Hyperebenen aus (3.19) die Menge $\mathcal{F}_C$.

**Lemma 3.50.** *Für $j \in \{1, \ldots, n\}$ ist $u(:, j)$ aus (2.11) gemäß Definition 3.47 zu der affinen Hyperebene $E_j^{(A)}$ dual. Für $i \in \{1, \ldots, k\}$ ist analog $w(:, i)$ aus (2.10) zu der affine Hyperebene $E_i^{(C)}$ dual.*

*Beweis.* Einfaches Einsetzen von $x = u(:, j)$ und $x_E^{\mathbf{T}} = V(j, 2:s)/V_{j1}$ sowie $\delta = -1$ in Definition 3.47 zeigt die erste Behauptung und die zweite folgt analog. □

Als nächstes wird in zwei Schritten (Satz 3.51 und Lemma 3.52) gezeigt, dass die affinen Hyperebenen, die $\mathcal{I}_C$ begrenzen, zu den Ecken von $\mathcal{F}_A$ dual sind. Die entscheidende Aussage wird in Satz 3.51 getroffen. Das Lemma 3.52 ist notwendig, um zu zeigen, dass die Fälle, die in Satz 3.51 nicht untersucht werden, nicht von Belang sind.

**Satz 3.51.** *Seien $j_1, \dots, j_{s-1} \in \{1, \dots, n\}$ so gewählt, dass $u(:, j_1), \dots, u(:, j_{s-1})$ eindeutig eine affine Hyperebene festlegen und diese nicht den Nullpunkt enthalte. Diese affine Hyperebene sei mit $E$ bezeichnet. Sei $x$ gemäß Definition 3.47 zu $E$ dual.*

*Es gelten:*

1. *Es ist $x$ der (eindeutige) Schnittpunkt von $E_{j_1}^{(A)}, \dots, E_{j_{s-1}}^{(A)}$.*
2. *Ist $E$ keine Randfläche von $\mathcal{I}_C$, so gehört $x$ auch nicht zu (dem Rand von) $\mathcal{F}_A$, denn es gilt $(1, x^{\mathbf{T}})V^{\mathbf{T}} \ngeq 0$.*

*Beweis.* Zunächst wird die erste Behauptung bewiesen. Seien $\widetilde{e} = (-1, \dots, -1)^{\mathbf{T}} \in \mathbb{R}^{s-1}$ und $B \in \mathbb{R}^{(s-1)\times(s-1)}$ mit

$$B(\ell, :) = \frac{V(j_\ell, 2:s)}{V_{j_\ell 1}}, \qquad \ell = 1, \dots, s-1.$$

Die Matrix $B$ ist regulär, da $B(\ell, :) = u(:, j_\ell)^{\mathbf{T}}$ und durch $u(:, j_1), \dots, u(:, j_{s-1})$ eine affine Hyperebene definiert ist, die nicht den Nullpunkt enthält. Die affine Hyperebene $E$ enthält alle $u(:, j_1), \dots, u(:, j_{s-1})$. Damit, und da $E$ nicht den Nullpunkt enthält, lässt sich $E$ durch $x_E = B^{-1}\widetilde{e}$ als $E = \{y \in \mathbb{R}^{s-1} : \ x_E^{\mathbf{T}} y = -1\}$ darstellen. Es ist $x = B^{-1}\widetilde{e}$ zu $E$ dual. Einfaches Nachrechnen ergibt, dass $x$ der Schnittpunkt der affinen Hyperebenen $E_{j_1}^{(A)}, \dots, E_{j_{s-1}}^{(A)}$ ist. Wegen der Regularität von $B$ ist $x$ eindeutig.

Nun wird die zweite Behauptung bewiesen. Sei $E$ keine Randfläche von $\mathcal{I}_C$. Das heißt, es gibt ein $j_0$, sodass $u(:, j_0) \in \mathcal{I}_C \setminus E$ nicht auf derselben Seite wie der Nullpunkt von $E$ liegt. (Nach Voraussetzung gehört der Nullpunkt selbst nicht zu $E$.) Für den Nullpunkt gilt $x_E^{\mathbf{T}}(0, \dots, 0)^{\mathbf{T}} = 0 > -1$ und da der Nullpunkt und $u(:, j_0) = (V(j_0, 2:s))^{\mathbf{T}}/V_{j_0 1}$ auf verschiedenen Seiten von $E$ liegen, folgt

$$x_E^{\mathbf{T}} \frac{V(j_0, 2:s)^{\mathbf{T}}}{V_{j_0 1}} < -1. \tag{3.20}$$

Somit gilt für das zu $E$ gemäß Definition 3.47 duale $x = x_E$

$$(1, x^{\mathbf{T}})(V(j_0, :))^{\mathbf{T}} = V_{j_0 1} + x^{\mathbf{T}}(V(j_0, 2:s))^{\mathbf{T}} < 0.$$

Also kann $x$ nicht zu $\mathcal{F}_A$ und insbesondere auch nicht zum Rand von $\mathcal{F}_A$ gehören. □

In Satz 3.51 wurden Ebenen, die den Nullpunkt enthalten, ausgeschlossen. Nun wird gezeigt, dass solche Ebenen nicht von Interesse sind, da sie keine Randflächen von $\mathcal{I}_C$ sein können. Das Lemma ist analog zu Lemma 3.13.

**Lemma 3.52.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix, $s = \operatorname{rank}(D) = \operatorname{rank}_+(D) \geq 2$ und $D^{\mathbf{T}}D$ sowie $DD^{\mathbf{T}}$ irreduzibel. Sei $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$ mit $V(:, 1) > 0$.*

*Der Nullpunkt $o = (0, \dots, 0)^{\mathbf{T}} \in \mathbb{R}^{s-1}$ gehört zu $\mathcal{I}_C$ und liegt nicht auf dem Rand von $\mathcal{I}_C$.*

*Beweis.* Der Beweis ist analog zu dem Beweis von Lemma 3.13. Sei angenommen, es würde $o \in \partial\mathcal{I}_C$ oder gar $o \notin \mathcal{I}_C$ gelten. Da $\mathcal{I}_C$ die konvexe Hülle der $u(:, j)$, $j = 1, \dots, n$, ist, hieße dies, dass es ein $y \in \mathbb{R}^{s-1} \setminus \{0\}$ gäbe mit $(u(:, j))^{\mathbf{T}} y \geq 0$ für alle $j = 1, \dots, n$. Einsetzen von $u(:, j)$ führt auf

$$(u(:, j))^{\mathbf{T}} y = \frac{V(j, 2:s)y}{V_{j1}} \geq 0 \qquad \forall j = 1, \dots, n.$$

Dies ist wegen $V_{j1} > 0$, $j = 1, \dots, n$, aber gleichbedeutend mit $V(j, 2:s)y \geq 0$ für alle $j = 1, \dots, n$, was Satz 3.3 widerspricht. □

**Bemerkung 3.53.**

1. *Als Folgerung aus den Betrachtungen von oben ergibt sich eine einfache und indirekte Methode zur Bestimmung von $\mathcal{F}_A$ [12, 148, 153]. Nach Satz 3.51 lässt sich $\mathcal{F}_A$ als die konvexe Hülle der Punkte bestimmen, die zu den affinen Hyperebenen, die $\mathcal{I}_C$ begrenzen, dual sind. Da die Berechnung von $\mathcal{I}_C$, beispielsweise mittels der* MatLab*-Routine* convhull*, schnell möglich ist, gilt gleiches auch für die Berechnung von $\mathcal{F}_A$.*
2. *Analog zu der Aussage von Satz 3.51 lässt sich zeigen, dass die Ecken von $\mathcal{F}_C$ gemäß Definition 3.47 zu den affinen Hyperebenen, die $\mathcal{I}_A$ begrenzen, dual sind. Somit lässt sich $\mathcal{F}_C$ unter Nutzung von $\mathcal{I}_A$ schnell indirekt bestimmen.*

### 3.6.4 Das Dualitätsprinzip im Kontext von Polarität

Die Zusammenhänge zwischen den niedrigdimensionalen Darstellungen der Spalten des ersten und der Zeilen des zweiten Faktors einer nichtnegativen Matrixfaktorisierung, zwischen den Mengen $\mathcal{F}_A$ und $\mathcal{I}_C$ sowie zwischen den Mengen $\mathcal{F}_C$ und $\mathcal{I}_A$ basieren auf dem Dualitätsprinzip. Diese Zusammenhänge sind strukturell sehr ähnlich zu denen zwischen einer Menge $K \subset \mathbb{R}^d$ und dessen Polare beziehungsweise zwischen einem Polytop[6] und dessen Polare [9, 179]. Im Folgenden werden kurz einzelne Parallelen angeführt. Diese Betrachtungen stehen nicht im Fokus dieser Schrift und erfolgen nur in Ansätzen.

**Bemerkung 3.54.** *Für die folgenden Betrachtungen wird an manchen Stellen die Kurzschreibweise $-K = \{b \in \mathbb{R}^d : -b \in K\} \subset \mathbb{R}^d$ zu einer Menge $K \subset \mathbb{R}^d$ verwendet.*

**Definition 3.55.** *Zu einer Menge $K \subset \mathbb{R}^d$ wird die Menge $K^\circ$ als ihre Polare bezeichnet mit*

$$K^\circ = \{b \in \mathbb{R}^d : b^{\mathbf{T}}x \leq 1 \text{ für alle } x \in K\}.$$

**Definition 3.56.** *Die Dimension $\dim(K)$ einer konvexen Menge $K \subset \mathbb{R}^d$ ist die Dimension des kleinsten affinen Unterraums von $\mathbb{R}^d$, welcher $K$ enthält.*

**Definition 3.57.** *Seien $K \subset \mathbb{R}^d$ eine abgeschlossene konvexe Menge und $F$ eine Teilmenge von $K$ mit $0 \leq \dim(F) < d$. Es wird $F$ eine Seite (englisch* face*) von $K$ genannt, falls eine affine Hyperebene $H$, welche an $K$ anliegt (das heißt $H \cap K = H \cap \partial K$), mit $F = K \cap H$ existiert.*

**Lemma 3.58.** *Seien die Voraussetzungen aus Satz 3.10 erfüllt.*

*Es gelten $\mathcal{F}_A = -\mathcal{I}_C^\circ$ und $\mathcal{F}_C = -\mathcal{I}_A^\circ$.*

*Beweis.* Es gilt

$$\mathcal{F}_A = \left\{ x \in \mathbb{R}^{s-1} : -\frac{x^{\mathbf{T}}(V(i,2:s))^{\mathbf{T}}}{V_{i1}} \leq 1,\ i = 1, \ldots, n \right\}.$$

Da $\mathcal{I}_C$ ein Polytop ist, welches von $u(:,i) = (V(i,2:s))^{\mathbf{T}}/V_{i1},\ i = 1, \ldots, n$, aufgespannt wird, und $(0, \ldots, 0)^{\mathbf{T}} \in \operatorname{int}(\mathcal{I}_C)$ gilt (Lemma 3.52), folgt trivialerweise (vergleiche etwa Satz 2.11 aus [179])

$$b \in \mathcal{I}_C^\circ \Leftrightarrow b^{\mathbf{T}} \frac{(V(i,2:s))^{\mathbf{T}}}{V_{i1}} \leq 1,\ i = 1, \ldots, n,$$

und somit ist $\mathcal{F}_A = -\mathcal{I}_C^\circ$. Analog wird $\mathcal{F}_C = -\mathcal{I}_A^\circ$ gezeigt. □

Des Weiteren lässt sich folgende Aussage über den Zusammenhang zwischen den niedrigdimensionalen Darstellungen der Faktoren einer Faktorisierung mittels Simplexen treffen:

[6] In dieser Arbeit wird der Begriff *Polytop* ohne den Zusatz *konvex* genutzt [9,179]. Es werden keine nicht konvexen Polytope betrachtet. Ein Polytop ist die konvexe Hülle einer endlichen Menge von Punkten in $\mathbb{R}^d$.

**Lemma 3.59.** *Seien die Voraussetzungen aus Satz 3.10 erfüllt und $D = CA$ eine nichtnegative Faktorisierung, für die es ein reguläres $T \in \mathbb{R}^{s\times s}$ mit $A = TV^{\mathbf{T}}$ und $T(:,1) = (1,\ldots,1)^{\mathbf{T}}$ gibt. Weiter seien $x^{(i)}$, $i = 1,\ldots,s$, die niedrigdimensionalen Darstellungen der Zeilen von $A$ und $y^{(j)}$, $j = 1,\ldots,s$, die niedrigdimensionalen Darstellungen der Spalten von $C$ wie in (3.15) und (3.16) sowie $\mathcal{S} = \operatorname{conv}(x^{(1)},\ldots,x^{(s)})$ das Simplex mit den Ecken $x^{(i)}$, $i = 1,\ldots,s$, und $\overline{\mathcal{S}} = \operatorname{conv}(y^{(1)},\ldots,y^{(s)})$ das Simplex mit den Ecken $y^{(j)}$, $j = 1,\ldots,s$.*

*Es gelten $\overline{\mathcal{S}}^\circ = -\mathcal{S}$ und $\mathcal{S}^\circ = -\overline{\mathcal{S}}$.*

*Beweis.* Sei $o = (0,\ldots,0)^{\mathbf{T}} \in \mathbb{R}^{s-1}$. Da $\mathcal{S}$ zu einer nichtnegativen Faktorisierung von $D$ gehört, gilt $o \in \operatorname{int}(\mathcal{S})$, siehe Lemma 3.13 und Satz 3.23. Somit gilt auch $o \in \operatorname{int}(-\mathcal{S})$. Weiter ist $o \in \operatorname{int}(\overline{\mathcal{S}})$, da $o \in \operatorname{int}(\mathcal{I}_C)$ und $\mathcal{I}_C \subset \overline{\mathcal{S}}$ (analoge Argumentationen). Es ist $\overline{\mathcal{S}}^\circ = \{b \in \mathbb{R}^{s-1} : b^{\mathbf{T}}y^{(j)} \leq 1,\ j = 1,\ldots,s\}$, vergleiche Satz 2.11 aus [179]. Die Randflächen von $\overline{\mathcal{S}}^\circ$ sind $F^{(j)} = \{b \in \mathbb{R}^{s-1} :\ b^{\mathbf{T}}y^{(j)} = 1\}$, $j = 1,\ldots,s$. Wegen $(x^{(i)})^{\mathbf{T}}y^{(j)} = -1$ für $i \neq j$ (siehe Satz 3.46) sind $F^{(j)}$, $j = 1,\ldots,s$, auch die Randflächen von $-\mathcal{S}$. Trivialerweise gilt $o \in \operatorname{int}(\overline{\mathcal{S}}^\circ)$ und wegen $o \in \operatorname{int}(-\mathcal{S})$ folgt somit $\overline{\mathcal{S}}^\circ = -\mathcal{S}$. Analog lässt sich $\mathcal{S}^\circ = -\overline{\mathcal{S}}$ zeigen. □

Anhand von Bemerkung 3.49 (Punkt 3) lässt sich ableiten, inwiefern einzelne Seiten des Simplex $\mathcal{S}$ der niedrigdimensionalen Darstellung des Faktors $A$ mit den Seiten des Simplex $\overline{\mathcal{S}}$ der niedrigdimensionalen Darstellung des Faktors $C$ zusammenhängen: Sind bereits $s_0$ zulässige Lösungen als Ecken von $\mathcal{S}$ fixiert, so sind eine $(s_0-1)$-dimensionale Seite von $\mathcal{S}$ und die dazu duale $(s-1-s_0)$-dimensionale affine Hyperebene, auf welcher eine Seite des Simplex $\overline{\mathcal{S}}$ eines möglichen Faktors $C$ liegt, festgelegt. (Analoge Zusammenhänge gelten natürlich auch, falls $s_0$ Ecken von $\overline{\mathcal{S}}$ bekannt sind.) Die damit in Verbindung stehende Aussage bezüglich den Seiten eines Polytops und den Seiten dessen Polare ist in Satz 3.60 angegeben. Zu beachten ist, dass $d = s - 1$.

**Satz 3.60** (Siehe [9])**.** *Seien $K \subset \mathbb{R}^d$ ein Polytop (also die konvexe Hülle einer endlichen Menge von Punkten des $\mathbb{R}^d$) mit $(0,\ldots,0)^{\mathbf{T}} \in \operatorname{int}(K)$ und $F$ eine Seite von $K$. Zu $F$ sei $\widehat{F} = \{b \in K^\circ :\ b^{\mathbf{T}}x = 1$ für alle $x \in F\}$.*

*Es ist $K^\circ$ ein Polytop. Weiter ist $\widehat{F}$ eine Seite von $K^\circ$ und es gilt $\dim(F) + \dim(\widehat{F}) = d - 1$.*

## 3.7 Verallgemeinerung für Probleme mit Rangdefizit

Die Faktorisierungsaufgabe 2.4 ist ein Spezialfall der nichtnegativen Matrixfaktorisierung (Faktorisierungsaufgabe 2.1), da $\operatorname{rank}(D) = \operatorname{rank}_+(D)$ vorausgesetzt wird. Dies kann für $s \geq 3$ dazu führen, dass die Faktorisierungsaufgabe 2.4 keine Lösung besitzt [170]. In einem solchen Fall müssen die Spaltendimension von $C$ und die Zeilendimension von $A$ vergrößert werden und der Ansatz, beide Faktoren $C$ und $A$ mittels einer abgeschnittenen Singulärwertzerlegung wie in (2.2) zu bestimmen, ist nicht mehr anwendbar.

Im Hinblick auf die Faktorisierungsaufgabe 2.1 wird in diesem Abschnitt eine verallgemeinerte Menge zulässiger Lösungen definiert und untersucht. Dabei wird vorausgesetzt, dass sich einer der Faktoren, nämlich der, für den das Rangdefizit angenommen wird, mittels des entsprechenden Faktors einer abgeschnittenen Singulärwertzerlegung von $D$ konstruieren lässt. Um auch für die verallgemeinerte Menge zulässiger Lösungen den Fokus auf $A$ zu legen, wird ohne Beschränkung der Allgemeinheit $s = \operatorname{rank}(A) < \operatorname{rank}(C)$ angenommen. Sollte andernfalls der Rangverlust für den ersten Faktor angenommen werden und die verallgemeinerte Menge zulässiger Lösungen dazu gesucht sein, so ist $D^{\mathbf{T}}$ zu betrachten.

Bei diesem Abschnitt handelt es sich um einen weiterführenden, der einen Ausblick über ein zukünftiges Forschungsfeld gibt. Es werden einzelne Eigenschaften der verallgemeinerten Menge zulässiger Lösungen analysiert und nachgewiesen. Darauf aufbauend lassen sich Algorithmen zur

Berechnung der verallgemeinerten Menge zulässiger Lösungen entwickeln. In Kapitel 4 werden zwar nur Methoden zur Berechnung der (klassischen) Menge zulässiger Lösungen behandelt, jedoch lassen sich viele dieser so modifizieren, dass sie sich zur Berechnung der verallgemeinerten Menge zulässiger Lösungen eignen.

### 3.7.1 Verallgemeinerte Menge zulässiger Lösungen

Wird das Rangdefizit für Faktor $A$ angenommen, so führt dies zu:

**Definition 3.61.** *Zu $s = \operatorname{rank}(D)$ und $m = \operatorname{rank}_+(D)$ ist die verallgemeinerte Menge zulässiger Zeilen für $A$ definiert als*

$$\widehat{\mathcal{A}} = \left\{ a \in \mathbb{R}^{1\times n} : \exists C \in \mathbb{R}_+^{k\times m},\, A \in \mathbb{R}_+^{m\times n},\ \textit{mit } \operatorname{rank}(A) = s,\ A(1,:) = a \textit{ sowie } D = CA \right\}.$$

Während sich diese Definition nur geringfügig von der aus (2.3) unterscheidet, ist die niedrigdimensionale Darstellung deutlich unterschiedlich zu (2.6). Das Problem ist, dass $C$ nicht allein aus den ersten $s$ linksseitigen Singulärvektoren konstruiert werden kann.

**Bemerkung 3.62.**

1. *Unverändert und wie in (2.1) festgelegt, sind auch in diesem Abschnitt die Faktoren einer abgeschnittenen Singulärwertzerlegung von den Dimensionen $U \in \mathbb{R}^{k\times s}$, $\Sigma \in \mathbb{R}^{s\times s}$ und $V \in \mathbb{R}^{n\times s}$.*
2. *Die in Definition 2.6 mit Bezug auf die Faktorisierungsaufgabe 2.4 eingeführte Menge zulässiger Lösungen $\mathcal{M}_A$ wird im Folgenden auch als klassische Menge zulässiger Lösungen bezeichnet.*

**Definition 3.63.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix, $s = \operatorname{rank}(D)$, $m = \operatorname{rank}_+(D) \geq s$ und $D^{\mathbf{T}}D$ irreduzibel. Weiter sei $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$ mit $V(:,1) > 0$ und es sei angenommen, dass der Rangverlust in $A$ auftritt, sodass nur Faktorisierungen mit $\operatorname{rank}(A) = s$ gesucht werden. Zu $\widehat{\mathcal{A}}$ lautet die verallgemeinerte Menge zulässiger Lösungen*

$$\begin{aligned}\widehat{\mathcal{M}}_A = \{ x \in \mathbb{R}^{s-1} : \exists C \in \mathbb{R}_+^{k\times m},\, T \in \mathbb{R}^{m\times s} \textit{ mit } T(1,:) = (1, x^{\mathbf{T}}),\\ T(:,1) = (1,\dots,1)^{\mathbf{T}},\, D = CTV^{\mathbf{T}},\, TV^{\mathbf{T}} \geq 0\}.\end{aligned}$$

**Bemerkung 3.64.** *In Kapitel 4 werden einige Algorithmen zur Berechnung von $\mathcal{M}_A$ vorgestellt. Die Klassifizierung, ob ein $x \in \mathbb{R}^{s-1}$ zulässig ist oder nicht, ist eine zentrale Unterroutine der numerischen Methoden. Die Klassifizierung erfolgt durch eine Optimierung über die Teilmatrix $S$ von $T \in \mathbb{R}^{s\times s}$, siehe (2.5), wobei sich $C = U\Sigma T^{-1}$ und $A = TV^{\mathbf{T}}$ in der aufgestellten Zielfunktion direkt berechnen lassen. Für die Klassifizierung, ob ein $x$ zu $\widehat{\mathcal{M}}_A$ gehört oder nicht, ist deutlich mehr Rechenaufwand nötig. Es lässt sich $C$ nur mittels der Lösung eines linearen Ausgleichsproblems unter Nichtnegativitätsrestriktionen bestimmen, siehe etwa [14, 104]. Dies erhöht den Rechenaufwand deutlich.*

### 3.7.2 Wichtige Eigenschaften der verallgemeinerten Menge zulässiger Lösungen

In den Abschnitten 3.2 – 3.4 wurden wichtige Eigenschaften von $\mathcal{M}_A$ vorgestellt. Viele dieser Eigenschaften übertragen sich auf $\widehat{\mathcal{M}}_A$. Die Beschränktheit (unter schwachen Voraussetzungen, siehe Satz 3.8) ist klar, da $\widehat{\mathcal{M}}_A$ eine Teilmenge von $\mathcal{F}_A$ ist und $\mathcal{F}_A$ unter den Voraussetzungen aus Satz 3.8 beschränkt ist. Weitere Eigenschaften wie, dass $\widehat{\mathcal{M}}_A$ den Nullpunkt nicht enthält und dass der Schnitt von $\widehat{\mathcal{M}}_A$ mit einem Strahl vom Ursprung ausgehend unterbrechungsfrei ist, folgen nicht unmittelbar und werden im Folgenden nachgewiesen. Weiterhin ändert sich die geometrische Interpretation einer zulässigen Lösung dahingehend, dass mit Polytopen anstatt von Simplexen gearbeitet wird.

### Geometrische Interpretation einer zulässigen Lösung

Die geometrische Interpretation eines zulässigen Polytops ist sehr ähnlich zu der geometrischen Interpretation eines zulässigen Simplex aus Satz 3.23 in Bezug auf die klassische Menge zulässiger Lösungen $\mathcal{M}_A$. Analog zu Definition 3.21 wird auch für Probleme mit Rangdefizit die niedrigdimensionale Blockdarstellung eines potentiellen Faktors $A$ mittels eines $\mathsf{P} \in \mathbb{R}^{(s-1)\times m}$ definiert. Inwiefern die Blockdarstellung eines Faktors $A$, der letztlich zu einer nichtnegativen Faktorisierung führt, die Ecken eines Polytops induziert, wird in Lemma 3.67 gezeigt.

**Definition 3.65.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix, $s = \operatorname{rank}(D)$, $m = \operatorname{rank}_+(D) > s$ und $D^{\mathbf{T}}D$ irreduzibel. Weiter sei $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$ mit $V(:,1) > 0$. Es seien $C \in \mathbb{R}^{k\times m}$ und $A \in \mathbb{R}^{m\times n}$ Faktoren mit $D = CA$, wobei sich $A$ mittels eines $T \in \mathbb{R}^{m\times s}$ als $A = TV^{\mathbf{T}}$ mit $T_{i1} = 1$ für $i = 1,\dots,m$ darstellen lässt. Der Faktor $A$ und eine Blockdarstellung mittels eines $\mathsf{P} \in \mathbb{R}^{(s-1)\times m}$ werden einander zugehörig genannt, wenn für $T = AV$ gilt:*

$$T(i, 2:s)^{\mathbf{T}} = \mathsf{P}(:,i), \qquad i = 1,\dots,m.$$

**Satz 3.66.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix, $s = \operatorname{rank}(D)$, $m = \operatorname{rank}_+(D) > s$ und $D^{\mathbf{T}}D$ sowie $DD^{\mathbf{T}}$ irreduzibel. Sei $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$ mit $V(:,1) > 0$. Weiter sei $A \in \mathbb{R}^{m\times n}$ mit $\operatorname{rank}(A) = s$ von der Art, dass es ein $T \in \mathbb{R}^{m\times s}$ mit $T_{i1} = 1$, $i = 1,\dots,m$, gibt, sodass $A = TV^{\mathbf{T}}$ gilt. Sei $\mathsf{P} \in \mathbb{R}^{(s-1)\times m}$ die zu $A$ gehörige Blockdarstellung.*

*Äquivalent sind:*

*a) Es ist $A \geq 0$ und es existiert ein $C \in \mathbb{R}_+^{k\times m}$, sodass $CA$ eine nichtnegative Matrixfaktorisierung von $D$ ist.*

*b) Die Blockdarstellung $\mathsf{P}$ erfüllt folgende zwei Bedingungen:*

*i) die Obermenge $\mathcal{F}_A$ enthält alle $\mathsf{P}(:,i)$, $i = 1,\dots,m$, und*

*ii) für alle $i = 1,\dots,k$ existiert jeweils ein $z^{(i)} \in \mathbb{R}^m$ mit $z^{(i)} \geq 0$ und $\sum_{j=1}^m z_j^{(i)} = 1$, sodass $w(:,i) = \mathsf{P}z^{(i)}$ gilt.*

*Beweis.* Zunächst wird a)⇒b) gezeigt. Die Bedingung i) folgt direkt aus $A \geq 0$. Sei weiter $C \in \mathbb{R}_+^{k\times m}$, sodass $D = CA$. Für alle $i = 1,\dots,k$ gilt

$$\begin{aligned} D(i,:) = C(i,:)A = C(i,:)TV^{\mathbf{T}} &\Leftrightarrow D(i,:)V = C(i,:)T \\ &\Leftrightarrow D(i,:)V(:,1)(1, w(:,i)^{\mathbf{T}}) = C(i,:)T \\ &\Leftrightarrow (1, w(:,i)^{\mathbf{T}}) = \frac{C(i,:)}{D(i,:)V(:,1)}T. \end{aligned}$$

Mit $T(j,:) = (1, \mathsf{P}(:,j)^{\mathbf{T}})$ für $j = 1,\dots,m$ werden die $z^{(i)} \in \mathbb{R}^m$ als

$$z_j^{(i)} = \frac{C_{ij}}{D(i,:)V(:,1)}, \qquad j = 1,\dots,m,$$

definiert und Bedingung ii) ist erfüllt.
Nun wird b)⇒a) gezeigt. Aus i) folgt $A \geq 0$. Weiter wird ein geeignetes $C$ zeilenweise wie folgt konstruiert: Für $i = 1,\dots,k$ werden $C(i,:) = D(i,:)V(:,1)(z^{(i)})^{\mathbf{T}}$ gesetzt, wobei $z^{(i)} \in \mathbb{R}^m$ jeweils Bedingung ii) für $i$ erfüllt. Offensichtlich gilt $C \geq 0$. Für alle $i = 1,\dots,k$ gilt wegen

$$D(i,:)V = D(i,:)V(:,1)(1, w(:,i)^{\mathbf{T}})$$

und weiter

$$C(i,:)A = D(i,:)V(:,1)\underbrace{(z^{(i)})^{\mathbf{T}}T}_{(1,w(:,i)^{\mathbf{T}})}V^{\mathbf{T}} = D(i,:)VV^{\mathbf{T}} = U(i,:)\Sigma\underbrace{V^TV}_{I_s}V^T = D(i,:),$$

sodass $D = CA$ gilt. □

**Lemma 3.67** (Vergleiche auch [3])**.** *Seien die Voraussetzungen von Satz 3.66 gegeben und Bedingung b) erfüllt.*

*Es gibt kein $i_0 \in \{1,\dots,m\}$, sodass für $\mathsf{P}(:,i_0)$ ein $z \in \mathbb{R}^{m-1}$ mit $z \geq 0$, $\sum_{i=1}^{m-1} z_i = 1$ und*

$$\mathsf{P}(:,i_0) = \sum_{\ell=1}^{i_0-1} z_\ell \mathsf{P}(:,\ell) + \sum_{\ell=i_0+1}^{m} z_{\ell-1}\mathsf{P}(:,\ell)$$

*existiert. Somit ist $\mathcal{P} = \operatorname{conv}(\mathsf{P}(:,1),\dots,\mathsf{P}(:,m))$ ein Polytop mit den Ecken $\mathsf{P}(:,i)$, $i = 1,\dots,m$.*

*Beweis.* Seien $A$ wie in Definition 3.65 aus $\mathsf{P}$ bestimmt und $C$ wie im Beweis von Satz 3.66 konstruiert. Es ist $CA$ eine nichtnegative Faktorisierung von $D$. Ohne Beschränkung der Allgemeinheit sei angenommen, dass für $i_0 = 1$ ein $z \in \mathbb{R}^{m-1}$ existiert mit $z \geq 0$, $\sum_{i=1}^{m-1} z_i = 1$ und $\mathsf{P}(:,1) = \mathsf{P}(:,2:m)z$. Für dieses $z$ sei die Matrix $B \in \mathbb{R}^{m\times m}$ definiert durch $B_{11} = 1$, $B_{1j} = -z_{j-1}$ für $j = 2,\dots,m$ sowie

$$B_{ij} = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j. \end{cases} \qquad \text{für } i = 2,\dots,m \text{ und } j = 1,\dots,m.$$

Aus $\mathsf{P}(:,1) = \mathsf{P}(:,2:m)z$ folgt $\widetilde{A} = BA \geq 0$ sowie $\widetilde{A}(1,:) = (0,\dots,0)$. Weiter ist aus dem Beweis von Lemma 3.14 bekannt, dass $\operatorname{rank}(B) = m$ (analoger Schluss) und $B^{-1} \geq 0$ gelten und somit auch $\widetilde{C} = CB^{-1} \geq 0$ ist. Wegen $D = CA$ ist $\widetilde{C}\widetilde{A}$ eine nichtnegative Matrixfaktorisierung von $D$. Weiter gilt $D = \widetilde{C}(:,2:m)A(2:s,:)$. Dies ist ein Widerspruch zu $\operatorname{rank}_+(D) = m$, da eine nichtnegative Matrixfaktorisierung von $D$ mittels Faktoren $\widehat{C} \in \mathbb{R}^{k\times(m-1)}$ und $\widehat{A} \in \mathbb{R}^{(m-1)\times n}$ existiert, nämlich gerade die mit den Faktoren $\widehat{C} = \widetilde{C}(:,2:m)$ und $\widehat{A} = A(2:s,:)$. Somit ist die Annahme falsch und $\mathsf{P}(:,1)$ kann keine Linearkombination von $\mathsf{P}(:,2:m)$ mit nichtnegativen Koeffizienten, deren Summe 1 ergibt, sein. □

**Bemerkung 3.68** (Vergleiche auch [36])**.** *Aus Lemma 3.67 folgt für eine nichtnegative Matrix $D$ unter den oben genannten Bedingungen, dass $\operatorname{rank}_+(D)$ die kleinste Zahl ist, sodass es ein $m$-Eckiges Polytop gibt, welches in $\mathcal{F}_A$ liegt und alle $w(:,i)$, $i = 1,\dots,k$, einschließt.*

**Definition 3.69.** *Erfüllt eine Blockdarstellung $\mathsf{P}(:,i)$ die Bedingung b) aus Satz 3.66, so wird $\mathcal{P} = \operatorname{conv}(\mathsf{P}(:,1),\dots,\mathsf{P}(:,m))$ ein zulässiges Polytop genannt.*

In Satz 3.26 ist eine notwendige und hinreichende Bedingung für eine zulässige Lösung in Bezug auf die klassische Menge zulässiger Lösungen $\mathcal{M}_A$ formuliert. Aus Satz 3.66 lässt sich eine analoge Aussage für die Klassifizierung eines $x \in \mathbb{R}^{s-1}$ bezüglich der verallgemeinerten Menge zulässiger Lösungen ableiten.

**Satz 3.70.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix, $s = \operatorname{rank}(D)$, $m = \operatorname{rank}_+(D) > s$ und $D^{\mathbf{T}}D$ sowie $DD^{\mathbf{T}}$ irreduzibel. Sei $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$ mit $V(:,1) > 0$.*

*Ein $x \in \mathbb{R}^{s-1}$ gehört genau dann zu $\widehat{\mathcal{M}}_A$, wenn es zu $\mathcal{F}_A$ gehört und es $x^{(i)} \in \mathcal{F}_A$, $i = 1\dots,m-1$, gibt, sodass das daraus gebildete Polytop $\mathcal{P} = \operatorname{conv}(\mathsf{P}(:,1),\dots,\mathsf{P}(:,m)) \subset \mathbb{R}^{s-1}$ mit den Ecken $\mathsf{P}(:,1) = x$ und $\mathsf{P}(:,i+1) = x^{(i)}$, $i = 1\dots,m-1$, alle $w(:,i)$, $i = 1\dots,k$, enthält.*

*Beweis.* Sei $x \in \widehat{\mathcal{M}}_A$ angenommen, das heißt nach Definition 3.63 gibt es ein $C \in \mathbb{R}_+^{k\times m}$ und ein $T \in \mathbb{R}^{m\times s}$ mit $T(:,1) = (1,\dots,1)^{\mathbf{T}}$ und $T(1,2:s) = x^{\mathbf{T}}$, sodass $A = TV^{\mathbf{T}} \geq 0$ und $D = CA$. Trivialerweise gilt $x \in \mathcal{F}_A$. Nach Satz 3.66 gilt für die zu $A$ gehörige Blockdarstellung $\mathsf{P}$, dass $\mathsf{P}(:,i) \in \mathcal{F}_A$, $i = 1,\dots,m$, und $w(:,i) \in \operatorname{conv}(\mathsf{P}(:,1),\dots,\mathsf{P}(:,m))$, $i = 1,\dots,k$. Somit sind $x^{(i)} = \mathsf{P}(:,i+1) = T(i+1,2:s)^{\mathbf{T}}$, $i = 1\dots,m-1$, geeignet. Nach Lemma 3.67 sind $\mathsf{P}(:,i)$, $i = 1,\dots,m$, die Ecken von $\mathcal{P} = \operatorname{conv}(\mathsf{P}(:,1),\dots,\mathsf{P}(:,m))$.
Gibt es andersherum zu $x \in \mathcal{F}_A$ weitere $x^{(i)} \in \mathcal{F}_A$, $i = 1,\dots,m-1$, sodass $\mathcal{P} = \operatorname{conv}(\mathsf{P}(:,1),\dots,\mathsf{P}(:,m)) \supset \mathcal{I}_A$ gilt, so ist mit $T(1,:) = (1,x^{\mathbf{T}})$, $T(i+1,:) = (1,{x^{(i)}}^{\mathbf{T}})$, $i = 1,\dots,m-1$, nach Satz 3.66 $A = TV^{\mathbf{T}} \geq 0$ und es gibt ein $C \in \mathbb{R}_+^{k\times m}$ mit $D = CA$. Die Matrix $T$ ist von der (für die Definition von $\widehat{\mathcal{M}}_A$) geforderten Form und $x$ liegt in $\widehat{\mathcal{M}}_A$. □

**Korollar 3.71.** *Seien die Voraussetzungen aus Satz 3.66 gegeben, die Bedingung b) erfüllt und* $\mathcal{P} = \operatorname{conv}(\mathsf{P}(:,1),\dots,\mathsf{P}(:,m))$ *das zugehörige zulässige Polytop.*

*Es gilt* $\mathcal{I}_A \subset \mathcal{P}$.

*Beweis.* Nach Voraussetzung gilt $w(:,i) \in \operatorname{conv}(\mathsf{P}(:,1),\dots,\mathsf{P}(:,m))$, $i = 1,\dots,k$. Es sind $\mathcal{P}$ die konvexe Hülle von $\mathsf{P}(:,1),\dots,\mathsf{P}(:,m)$ und $\mathcal{I}_A$ die konvexe Hülle von $w(:,i)$, $i = 1,\dots,k$. Folglich ist $\mathcal{I}_A \subset \mathcal{P}$. □

### Der Ursprung ist nicht enthalten

Eine wichtige Eigenschaft der klassischen Menge zulässiger Lösungen $\mathcal{M}_A$ ist, dass sie nicht den Nullpunkt enthält. Diese überträgt sich auch auf die verallgemeinerte Menge zulässiger Lösungen. Der Beweis wird mittels der geometrischen Aussagen der Lemmata 3.13 und 3.67 sowie des Korollars 3.71 geführt und unterscheidet sich von dem des Satzes 3.10.

**Satz 3.72.** *Seien* $D \in \mathbb{R}^{k\times n}$ *eine nichtnegative Matrix,* $s = \operatorname{rank}(D)$, $m = \operatorname{rank}_+(D) > s$ *und* $D^{\mathbf{T}}D$ *sowie* $DD^{\mathbf{T}}$ *irreduzibel. Weiter sei* $U\Sigma V^{\mathbf{T}}$ *eine abgeschnittene Singulärwertzerlegung von* $D$ *mit* $V(:,1) > 0$.

*Der Nullpunkt* $(0,\dots,0)^{\mathbf{T}} \in \mathbb{R}^{s-1}$ *liegt nicht in* $\widehat{\mathcal{M}}_A$.

*Beweis.* Sei $\mathsf{P}$ eine Blockdarstellung, die zu einem Faktor $A$ gehört, welcher zu einer nichtnegativen Faktorisierung führt. Nach Lemma 3.67 liegt kein $\mathsf{P}(:,i_0)$ im Inneren von $\mathcal{P} = \operatorname{conv}(\mathsf{P}(:,1),\dots,\mathsf{P}(:,m))$. Da nach Korollar 3.71 das Polytop $\mathcal{P}$ die Menge $\mathcal{I}_A$ enthält, kann kein $\mathsf{P}(:,i_0)$ im Inneren von $\mathcal{I}_A$ liegen. Weiter liegt der Nullpunkt nach Lemma 3.13 im Inneren von $\mathcal{I}_A$. Somit gilt $\mathsf{P}(:,i) \neq (0,\dots,0)^{\mathbf{T}}$, $i = 1,\dots,m$. Da dies für jedes zulässige Polytop $\mathcal{P}$ gilt, kann $(0,\dots,0)^{\mathbf{T}}$ wegen Satz 3.70 nicht zu $\widehat{\mathcal{M}}_A$ gehören. □

### Unterbrechungsfreier Schnitt mit Strahlen vom Ursprung aus

Weiter besitzt die verallgemeinerte Menge zulässiger Lösungen auch die Eigenschaft, dass ihre Schnitte mit Strahlen, die vom Ursprung ausgehen, unterbrechungsfrei sind. Der Beweis des folgenden Satzes 3.73 muss jedoch grundsätzlich anders geführt werden, als der zu Satz 3.15. Wie in Bemerkung 3.16 angeführt, ließe sich die Beweisargumentation zu dem nun folgenden Satz analog auch zum Beweis des Satzes 3.15 nutzen.

**Satz 3.73.** *Seien die Voraussetzungen von Satz 3.72 erfüllt und sei* $x \in \widehat{\mathcal{M}}_A$. *Zu* $x$ *sei der Wert* $\gamma^* \geq 1$ *definiert, sodass* $\gamma^* x$ *auf dem Rand von* $\mathcal{F}_A$ *liegt.*

*Für alle* $\widetilde{x} = \gamma x$ *mit* $\gamma \in [1,\gamma^*]$ *gilt* $\widetilde{x} \in \widehat{\mathcal{M}}_A$.

*Beweis.* Zu $x$ existiert eine Matrix $T \in \mathbb{R}^{m\times s}$ mit $T(1,:) = (1, x^{\mathbf{T}})$ und $T(:,1) = (1,\dots,1)^{\mathbf{T}}$, welche auf eine nichtnegative Faktorisierung mit $C \in \mathbb{R}^{k\times m}$ und $A = TV^{\mathbf{T}}$ führt. Sei $\mathsf{P}$ die zu $A$ gehörige Blockdarstellung und $\mathcal{P}$ das entsprechende zulässige Polytop. Nach Satz 3.66 liegen die Ecken $\mathsf{P}(:,i)$, $i = 1,\dots,m$, von $\mathcal{P}$ in $\mathcal{F}_A$ und $\mathcal{P}$ enthält alle $w(:,i)$, $i = 1,\dots,k$. Es ist $\mathsf{P}(:,1) = x$. Ein neues Polytop $\widetilde{\mathcal{P}} \subset \mathbb{R}^{s-1}$ wird durch die Ecken $\widetilde{\mathsf{P}}(:,2:m) = \mathsf{P}(:,2:m)$ sowie $\widetilde{\mathsf{P}}(:,1) = \widetilde{x} = \gamma x$ mit $\gamma \in [1,\gamma^*]$ definiert. Da $\mathcal{P}$ den Nullpunkt enthält, siehe den Beweis des Satzes 3.72, folgt, dass das Polytop $\widetilde{\mathcal{P}}$ das Polytop $\mathcal{P}$ enthält und insbesondere auch alle $w(:,i)$, $i = 1,\dots,k$. Es liegen auch alle Ecken von $\widetilde{\mathcal{P}}$ in $\mathcal{F}_A$ und somit gilt $\widetilde{x} = \gamma x \in \widehat{\mathcal{M}}_A$ □

### 3.7.3 Beispiel einer verallgemeinerten Menge zulässiger Lösungen

Zur Veranschaulichung der verallgemeinerten Menge zulässiger Lösungen wird ein Beispiel untersucht, welches auf der $4\times 4$-Matrix aus (1.4) basiert. Die Matrix wird um einen Parameter $\alpha$ erweitert, wobei es für $\alpha = 0$ bei der ursprünglichen Matrix aus (1.4) bleibt.

**Beispiel 3.74.** *Sei für $\alpha \geq 0$ die Rang-3 Matrix $D^{(\alpha)} \in \mathbb{R}^{4\times 4}$ definiert als*

$$D^{(\alpha)} = \begin{pmatrix} 1+\alpha & 1+\alpha & \alpha & \alpha \\ 1+\alpha & \alpha & 1+\alpha & \alpha \\ \alpha & 1+\alpha & \alpha & 1+\alpha \\ \alpha & \alpha & 1+\alpha & 1+\alpha \end{pmatrix}. \tag{3.21}$$

Es ist bekannt, dass $D^{(0)}$ keine nichtnegative Matrixfaktorisierung mit Matrizen $C^{(0)} \in \mathbb{R}^{4\times 3}$ und $A^{(0)} \in \mathbb{R}^{3\times 4}$ besitzt [27,170]. Für $D^{(\alpha)}$ gibt es eine solche Faktorisierung nur für $\alpha \geq \sqrt{2}/2$, wobei beispielsweise

$$C^{(\alpha)} = \begin{pmatrix} \frac{1}{2}+\alpha & \frac{1}{2}+\alpha & 0 \\ \frac{1}{2}\frac{(1+2\alpha)^2}{1+\alpha} & 0 & \frac{1}{2}\frac{1+2\alpha}{1+\alpha} \\ 0 & \frac{1}{2}\frac{(1+2\alpha)^2}{1+\alpha} & \frac{1}{2}\frac{1+2\alpha}{1+\alpha} \\ \frac{1}{2}\frac{\alpha(1+2\alpha)}{1+\alpha} & \frac{1}{2}\frac{\alpha(1+2\alpha)}{1+\alpha} & \frac{1+2\alpha}{1+\alpha} \end{pmatrix},$$

$$A^{(\alpha)} = \begin{pmatrix} \frac{2(1+\alpha)^2}{(1+2\alpha)^2} & \frac{2(1+\alpha)\alpha}{(1+2\alpha)^2} & \frac{(1+\alpha)^2+\alpha^2}{(1+2\alpha)^2} & \frac{2\alpha^2-1}{(1+2\alpha)^2} \\ \frac{2(1+\alpha)\alpha}{(1+2\alpha)^2} & \frac{2(1+\alpha)^2}{(1+2\alpha)^2} & \frac{2\alpha^2-1}{(1+2\alpha)^2} & \frac{(1+\alpha)^2+\alpha^2}{(1+2\alpha)^2} \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

geeignete Faktoren sind. Für $\alpha < \sqrt{2}/2$ ist nur eine nichtnegative Matrixfaktorisierung mit $C^{(\alpha)} \in \mathbb{R}^{4\times 4}$ und $A^{(\alpha)} \in \mathbb{R}^{4\times 4}$ möglich. Somit ist

$$\operatorname{rank}_+\big(D^{(\alpha)}\big) = \begin{cases} 3 & \text{für } \alpha \geq \sqrt{2}/2, \\ 4 & \text{für } \alpha < \sqrt{2}/2. \end{cases}$$

Die Matrix $D^{(\alpha)}$ besitzt die Singulärwerte $\sigma_1 = 2+4\alpha$, $\sigma_2 = \sigma_3 = \sqrt{2}$ und $\sigma_4 = 0$. Wegen der Gleichheit des zweiten und des dritten Singulärwerts, sind die zugehörigen Singulärvektoren auch über ihren jeweiligen Richtungssinn hinaus nicht eindeutig. Aus dieser Mehrdeutigkeit heraus resultiert ein Freiheitsgrad in der Orientierung der Menge $\mathcal{F}_A$ und somit auch der klassischen/verallgemeinerten Menge zulässiger Lösungen. Die Mengen $\mathcal{A}$ beziehungsweise $\widehat{\mathcal{A}}$ sind davon natürlich nicht betroffen und unabhängig von der gewählten Singulärwertzerlegung. Im Folgenden wird mit einer abgeschnittenen Singulärwertzerlegung mit

$$V(:,1:3) = \begin{pmatrix} \frac{1}{2} & \sqrt{2}/2 & 0 \\ \frac{1}{2} & 0 & -\sqrt{2}/2 \\ \frac{1}{2} & 0 & \sqrt{2}/2 \\ \frac{1}{2} & -\sqrt{2}/2 & 0 \end{pmatrix}$$

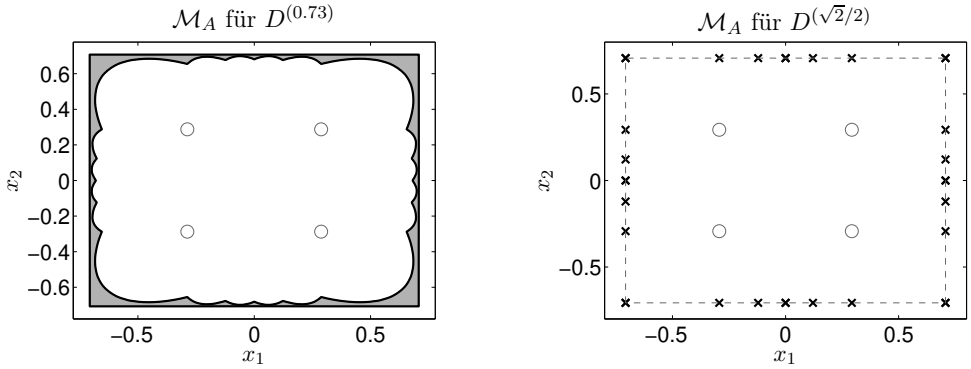


**Abbildung 3.2:** Die (klassischen) Mengen zulässiger Lösungen $\mathcal{M}_A$ für $D^{(0.73)}$ (links, grau dargestellt) und $D^{(\sqrt{2}/2)}$ (rechts, die Punkte $\times$) aus Beispiel 3.74. Die Orientierung ist wegen $\sigma_2 = \sigma_3$ frei wählbar. Die $w(:,i)$, $i = 1,\ldots,4$, sind durch $\circ$ gekennzeichnet. Rechts ist zudem $\mathcal{F}_A$ (gestrichelt) eingezeichnet.

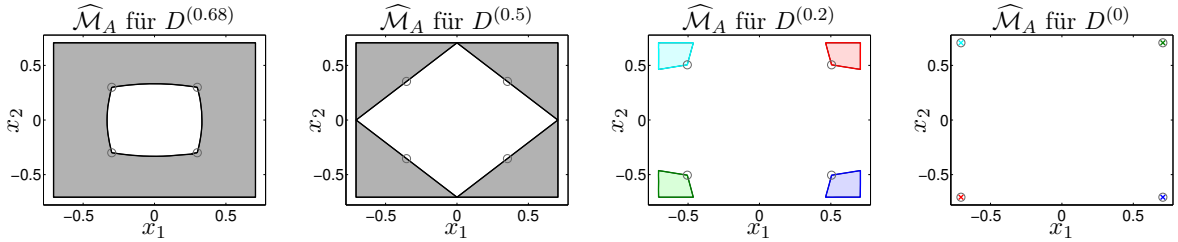


**Abbildung 3.3:** Die verallgemeinerten Mengen zulässiger Lösungen $\widehat{\mathcal{M}}_A$ mit $m = 4$ für $D^{(0.68)}$, $D^{(0.5)}$, $D^{(0.2)}$ und $D^{(0)}$ aus Beispiel 3.74. Für $D^{(0)}$ besteht die Menge aus vier Punkten; dies bedeutet, dass es unter der Vorgabe $\operatorname{rank}(C) = 4$ und $\operatorname{rank}(A) = 3$ unter Vernachlässigung von Skalierungs- und Permutationsmehrdeutigkeiten nur eine Lösung gibt. Die $w(:,i)$, $i = 1,\ldots,4$, sind durch $\circ$ markiert.

gearbeitet.

Für $\alpha = \sqrt{2}/2$ besteht die Menge zulässiger Lösungen $\mathcal{M}_A$ aus 24 isolierten Punkten. Es gilt

$$\mathcal{M}_A = \left\{ \left( \pm\gamma, \pm\frac{\sqrt{2}}{2} \right) : \gamma \in \Gamma \right\} \cup \left\{ \left( \pm\frac{\sqrt{2}}{2}, \pm\gamma \right) : \gamma \in \Gamma \right\}$$

mit

$$\Gamma = \left\{ 0,\ \frac{1}{2}\frac{\sqrt{2}}{(1+\sqrt{2})^2},\ \frac{1}{2}\frac{\sqrt{2}}{1+\sqrt{2}},\ \frac{\sqrt{2}}{2} \right\}.$$

In Abbildung 3.2 sind die (klassischen) Mengen zulässiger Lösungen $\mathcal{M}_A$ zu $\alpha = 0.73$ sowie $\alpha = \sqrt{2}/2$ dargestellt. Abbildung 3.3 zeigt die verallgemeinerten Mengen zulässiger Lösungen $\widehat{\mathcal{M}}_A$ für $\alpha \in \{0.68, 0.5, 0.2, 0\}$.

### 3.7.4 Probleme mit verstecktem Rangdefizit

Die klassische Menge zulässiger Lösungen $\mathcal{M}_A$ ist für Matrizen $D$ mit $\operatorname{rank}(D) = \operatorname{rank}_+(D)$ definiert. In diesem Abschnitt ist dieser Ansatz auf sogenannte Rangdefizitprobleme, das heißt auf Matrizen $D$ mit $m = \operatorname{rank}_+(D) \geq \operatorname{rank}(D)$, ausgeweitet. Dabei enthält die erstgenannte Klasse von Problemen eine Unterklasse, die für Anwendungen auftreten kann und von großem Interesse ist. Dies sind Probleme, für die zwar $\operatorname{rank}(D) = \operatorname{rank}_+(D)$ gilt, für die die Menge $\mathcal{M}_A$ aber nur auf abstrakte Lösungen führt, welche für eine konkrete Anwendung nicht sinnvoll sind. Um für $D$ sinnvolle Ergebnisse zu ermöglichen, muss die Anzahl der Teilfaktoren erhöht werden, sodass $C \in \mathbb{R}^{k\times r}$ und $A \in \mathbb{R}^{r\times n}$ mit $r > \operatorname{rank}(D)$. Eine solche Aufgabenstellung wird im Folgenden als *Problem mit verstecktem Rangdefizit* bezeichnet.

### Der erweiterte Rang und die entsprechende Menge zulässiger Lösungen

Um für oben beschriebene Fälle die Anzahl an nötigen Faktoren $r$ zu definieren, sodass eine für die Anwendung sinnvolle Faktorisierung existiert, wird der Begriff des *erweiterten Ranges* eingeführt:

**Definition 3.75.** *Sei $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix. Der erweiterte Rang $\operatorname{rank}_\dagger(D)$ ist die minimale Anzahl $r$, sodass in Bezug auf die konkrete (zugrunde liegende) Anwendung sinnvolle nichtnegative Faktoren $C \in \mathbb{R}^{k\times r}$ und $A \in \mathbb{R}^{r\times n}$ mit $D = CA$ für Modelldaten beziehungsweise $D \approx CA$ für gestörte Daten existieren.*

Ob die Definition des erweiterten Ranges bei $\operatorname{rank}_\dagger(D) > \operatorname{rank}_+(D)$ für eine konkrete Anwendung stichhaltig ist, hängt davon ab, inwiefern die Klassifizierung in *sinnvolle* und *nicht sinnvolle* Faktoren vorgenommen werden kann. Werden beispielsweise spektroskopische Daten untersucht und sind dazu Monotonierestriktionen für $C$ gefordert und un- oder nur schwach gestörte Daten liegen vor, so ist die Definition durchaus stichhaltig. Werden jedoch Monotonierestriktionen für mittelstark oder stark störungsbehaftete Daten angewendet, so bewegt sich die Definition des erweiterten Ranges in einer Grauzone.

Entsprechend ergibt sich für $D$ mit $\operatorname{rank}_\dagger(D) > \operatorname{rank}_+(D)$:

**Definition 3.76.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix, $s = \operatorname{rank}(D)$ und $D^{\mathbf{T}}D$ irreduzibel. Seien $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$ mit $V(:,1) > 0$ und $r = \operatorname{rank}_\dagger(D) > \operatorname{rank}_+(D)$. Falls für $A$ der Rang $s$ gefordert wird, so ist*

$$\mathcal{R} = \big\{x \in \mathbb{R}^{s-1} : \exists T \in \mathbb{R}^{r\times s},\, C \in \mathbb{R}_+^{k\times r} \text{ mit } T(1,:) = (1, x^{\mathbf{T}}),\, TV^{\mathbf{T}} \geq 0 \text{ sowie } D = CTV^{\mathbf{T}} \\ \text{und } C \text{ und } A = TV^{\mathbf{T}} \text{ sind für die konkrete Anwendung sinnvoll}\big\}$$

*die Menge zulässiger Lösungen unter Berücksichtigung von $\operatorname{rank}_\dagger(D)$.*

Eine (analog zu $\mathcal{F}_A$ von $\mathcal{M}_A$) wichtige Obermenge von $\mathcal{R}$ ist

$$\mathcal{R}^+ = \{x \in \mathbb{R}^{s-1} : (1, x^{\mathbf{T}})V^{\mathbf{T}} \geq 0 \text{ und } (1, x^{\mathbf{T}})V^{\mathbf{T}} \text{ ist für die konkrete Anwendung sinnvoll}\}.$$

Da es sich bei diesem Abschnitt nur um einen weiterführenden handelt, werden zwar wichtige Eigenschaften der Menge $\mathcal{R}$ und Begründungen für deren Gültigkeit genannt, aber keine vollständigen Beweise angeführt. Drei wichtige Eigenschaften von $\mathcal{R}$ sind:

1. Die Menge $\mathcal{R}$ ist beschränkt, sofern $D^{\mathbf{T}}D$ irreduzibel ist (vergleiche Satz 3.8). Es ist $\mathcal{F}_A$ beschränkt und da $\mathcal{R} \subset \mathcal{R}^+ \subset \mathcal{F}_A$ gilt, ist es somit auch $\mathcal{R}$.
2. Die geometrische Argumentation aus Satz 3.26 gilt analog, falls $D^{\mathbf{T}}D$ und $DD^{\mathbf{T}}$ irreduzibel sind. Ein $x \in \mathbb{R}^{s-1}$ ist genau dann zulässig, wenn $x \in \mathcal{R}^+$ gilt und es $r-1$ weitere Elemente in $\mathcal{R}$ gibt, sodass das daraus gebildete Polytop alle $w(:,i)$, $i = 1,\ldots,k$, einschließt.
3. Der Nullpunkt gehört nicht zu $\mathcal{R}$, falls $D^{\mathbf{T}}D$ irreduzibel ist (vergleiche Satz 3.10). Die Beweisidee ist ähnlich zu der aus dem Beweis des Satzes 3.72: Wäre der Nullpunkt in $\mathcal{R}$ enthalten, so gäbe es eine Faktorisierung mit, bezüglich der konkreten Anwendung, sinnvollen Faktoren $C \in \mathbb{R}_+^{k\times(r-1)}$ und $A \in \mathbb{R}_+^{(r-1)\times n}$, was ein Widerspruch zu $\operatorname{rank}_\dagger(D) = r$ wäre.

### Algorithmische Umsetzung des erweiterten Ranges

Die algorithmische Umsetzung der Forderung nach, bezüglich einer konkreten Anwendung, sinnvollen Faktoren kann über den Einsatz von Regularisierungsfunktionen erfolgen. Beispielsweise

lassen sich Unimodalitäts- oder Monotonierestriktionen einsetzen. Inwiefern Restriktionen für $C$ und $A$ zur Reduktion der (klassischen) Mengen zulässiger Lösungen eingesetzt werden, ist in Abschnitt 4.9 und beispielsweise in [12,13,57,132,158] erläutert. Um für störungsbehaftete Daten geeignete Ergebnisse zu generieren, wird auf Steuerparameter zurückgegriffen. Deren geeigneter Einsatz ist für die Berechnung verwertbarer Ergebnisse entscheidend.

## 3.8 Einflüsse von Störungen

Bei den bisherigen Untersuchungen zu den Mengen zulässiger Lösungen wurde stets das idealisierte Problem $D = CA$ betrachtet. Im Hinblick auf die Anwendung für störungsbehaftete Daten werden in diesem Abschnitt einige Details bezüglich $\mathcal{M}_A$ unter dem Aspekt des Einflusses von Störungen diskutiert. Untersucht wird, wie sich Störungen auf $w(:,i)$, $i = 1,\ldots,k$, und $u(:,j)$, $j = 1,\ldots,n$, auswirken. Somit rückt auch die geometrische Konstruktion in den Fokus, da diese direkt von $w(:,i)$ und $u(:,j)$ abhängt. Inwiefern die numerische Klassifizierung unter der Berücksichtigung von Störungen algorithmisch umgesetzt werden kann, wird in Abschnitt 4.1 untersucht.

Der Einfluss von Störungen hängt insbesondere von deren Struktur, aber auch von deren Größe ab. Für die geometrisch konstruktive Klassifizierung eines $x \in \mathbb{R}^{s-1}$, siehe Satz 3.26, können sich gravierende Schwierigkeiten ergeben. So liegen zwischen den ersten Veröffentlichungen [17, 120] zur geometrischen Bewertung beziehungsweise zur geometrischen Konstruktion von $\mathcal{M}_A$ für $s = 3$ und der Erweiterung des Ansatzes für gestörte Daten in [87, 88] etwa drei Jahrzehnte.

Eine Schwierigkeit bei der Berücksichtigung von Störungen ist, dass ihr Einfluss für die Berechnungen von $\mathcal{M}_A$ und $\mathcal{M}_C$ sehr unterschiedlich sein kann. Begründet ist dies in der Lage der $w(:,i)$, $i = 1,\ldots,k$, zu $\mathcal{F}_A$ beziehungsweise der $u(:,j)$, $j = 1,\ldots,n$, zu $\mathcal{F}_C$. Im Hinblick auf die Mengen zulässiger Lösungen ist für Matrizen, deren Zerlegungen in einem der beiden Faktoren viele Einträge nahe Null besitzen, folgendes wichtig: Einerseits sind Einträge nahe an Null in etwa $A$ wünschenswert, da sie die Menge $\mathcal{M}_A$ stark einschränken. Andererseits ist es für diese Einträge problematisch, dass (negative) Störungen die Mengen $\mathcal{I}_A$ und $\mathcal{I}_C$ unter Umständen stark vergrößern. Dies führt oft dazu, dass einzelne $w(:,i)$ außerhalb von $\mathcal{F}_A$, beziehungsweise einzelne $u(:,j)$ außerhalb von $\mathcal{F}_C$, liegen. Für ungestörte Daten ist dies ausgeschlossen.

### 3.8.1 Schwierigkeit bei der geometrischen Klassifizierung unter Störungen

Für die geometrischen Interpretation einer zulässigen Lösung $x \in \mathcal{M}_A$ sind $w(:,i)$, $i = 1,\ldots,k$, von zentraler Bedeutung. Dabei liegt ein $x$ genau dann in $\mathcal{M}_A$, wenn es in $\mathcal{F}_A$ liegt und es $s-1$ andere Elemente in $\mathcal{F}_A$ gibt, sodass das daraus konstruierte Simplex alle $w(:,i)$, $i = 1,\ldots,k$, enthält. Für die geometrische Interpretation eines $y \in \mathcal{M}_C$ sind analog $u(:,j)$, $j = 1,\ldots,n$, zentral.

Für störungsbehaftete Daten ist diese Forderung in der Regel nicht aufrecht zu erhalten. Im Allgemeinen muss $\mathcal{F}_A$ erweitert, das heißt vergrößert, werden. Zusätzlich muss die Forderung, dass alle $w(:,i)$ von einem Simplex eingeschlossen werden, gelockert werden, da mitunter einige dieser außerhalb von $\mathcal{F}_A$ liegen. Diese beiden Schwierigkeiten sowie die Vorgehensweisen um ihnen entgegenzuwirken sind in Abbildung 3.4 für den Datensatz 3 dargestellt. Dabei sind die Mengen $\mathcal{F}_C$ und $\mathcal{F}_A$ sowie die $u(:,j)$ und $w(:,i)$ abgebildet und es ist offensichtlich, dass einige der $u(:,j)$ außerhalb von $\mathcal{F}_C$ und einige der $w(:,i)$ außerhalb von $\mathcal{F}_A$ liegen, wodurch nach Satz 3.26 und Korollar 3.27 die Mengen $\mathcal{M}_C$ und $\mathcal{M}_A$ leer sein müssten. Zusätzlich sind Approximationen an $\mathcal{M}_C$ und $\mathcal{M}_A$ unter der Berücksichtigung von Störungen abgebildet. Diese wurden mit der Polygon inflation Methode unter Verwendung der in Abschnitt 4.1 beschriebenen Klassifizierung

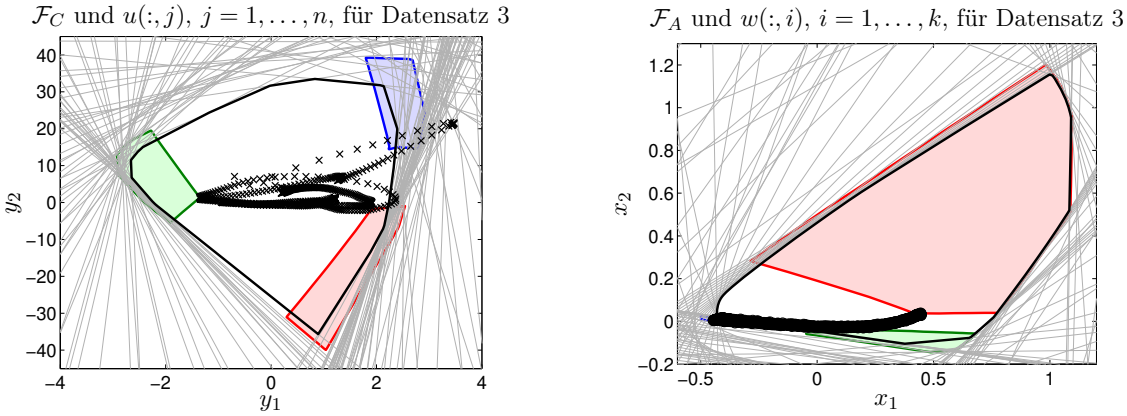


**Abbildung 3.4:** Der Einfluss von Störungen auf die Mengen $\mathcal{F}_A$ und $\mathcal{F}_C$ sowie auf die $w(:,i)$, $i=1,\ldots,k$, und $u(:,j)$, $j=1,\ldots,n$, für Datensatz 3. Links: Dargestellt sind einige der Grenzen für die affinen Halbräume $(1,x_1,x_2)(U\Sigma(i,1:3))^{\mathbf{T}} \geq 0$ (grau), welche $\mathcal{F}_C$ (schwarz) bilden und die $u(:,j)$, $j=1,\ldots,n$, (×). Zusätzlich ist eine Approximation an die Menge $\mathcal{M}_C$ unter der Berücksichtigung von Störungen (farbig), siehe zum Vergleich auch Abbildung 5.6 und die Erläuterungen dazu, eingezeichnet. Rechts: Dargestellt sind einige der Grenzen für die affinen Halbräume $(1,x_1,x_2)V(j,1:3)^{\mathbf{T}} \geq 0$ (grau), welche $\mathcal{F}_A$ (schwarz) bilden und die $w(:,i)$, $i=1,\ldots,k$, (○). Analog ist zusätzlich eine Approximation an die Menge $\mathcal{M}_A$ unter der Berücksichtigung von Störungen (farbig), siehe zum Vergleich ebenfalls auch Abbildung 5.6 und die Erläuterungen dazu, eingezeichnet. Für diese Daten würde die strikte Anwendung des Satzes 3.26 auf leere Mengen zulässiger Lösungen führen. Demgegenüber ergeben sich durch die Berücksichtigung von Störungen sinnvolle Resultate.

mittels $f(x,S)$ berechnet.

### 3.8.2 Sensitivität der $w(:,i)$ und $u(:,j)$

Die Sensitivität eines $u(:,j_1)$ ist damit verbunden, ob die zugehörige Spalte $D(:,j_1)$ ausschließlich Werte enthält, welche, verglichen mit $\max_{i,j} D_{ij}$, sehr klein sind. In solchen Situationen ergeben sich für $V_{j_1 1}$ betragskleine Einträge. Rückgekoppelt wirken sich diese über die niedrigdimensionale Darstellung auf $u(:,j_1)$ aus und ein solches $u(:,j_1)$ reagiert sensitiv auf den Einfluss von Störungen. Dieser Zusammenhang ist analog auch für ein $w(:,i_1)$ sowie die dazugehörige Zeile $D(i_1,:)$ zu beobachten. Beide Effekte werden nun folgend untersucht.

Seien zunächst $u(:,j)$, $j=1,\ldots,n$, betrachtet. Deren Normen ergeben sich als

$$\|u(:,j)\|_2 = \frac{\|V(j,2:s)\|_2}{V_{j1}}, \qquad j=1,\ldots,n.$$

Ist für ein bestimmtes $j_1$ der Wert $V_{j_1 1}$ sehr klein und zumindest ein Eintrag in $V(j_1,2:s)$ betragsmäßig nicht sehr klein, so ist $\|u(:,j_1)\|_2$ groß und es besteht die Gefahr, dass $u(:,j_1)$ nicht in $\mathcal{F}_C$ liegt. Um einen solchen Zusammenhang für die $u(:,j)$ zu quantifizieren, wird der Sensitivitätsschätzer $e^{(u)} \in \mathbb{R}^n$,

$$e_j^{(u)} = \max_{i=2,\ldots,s} \left| \frac{V_{ji}}{V_{j1}} \right|, \qquad j=1,\ldots,n, \tag{3.22}$$

eingeführt. Analog gilt für $w(:,i)$, $i=1,\ldots,k$, dass

$$\|w(:,i)\|_2 = \frac{\|U(i,:)\Sigma(:,2:s)\|_2}{U_{i1}\sigma_1}, \qquad i=1,\ldots,k.$$

Um hier den Effekt bereinigt von den Singulärwerten zu belassen, wird für die $w(:,i)$ der Sensitivitätsschätzer $e^{(w)} \in \mathbb{R}^k$ in der Form

$$e_i^{(w)} = \max_{j=2,\ldots,s} \left| \frac{U_{ij}}{U_{i1}} \right|, \qquad i=1,\ldots,k, \tag{3.23}$$

definiert.

**Bemerkung 3.77.**

1. *Der Schätzer $e^{(u)}$ ist ein guter Indikator, wie sensitiv die $u(:,j)$ auf Störungen reagieren. Sind alle Werte in $e^{(u)}$ gering, so ist auch die Sensitivität für Störungen gering. Sind jedoch einzelne Einträge hoch, so sind jene $u(:,j)$ anfällig für Störungen, welche zu den hohen Einträgen in $e^{(u)}$ gehören. Analog gilt dieser Zusammenhang auch für die Werte in $e^{(w)}$ und die $w(:,i)$. In den Abbildungen 5.12 und 5.13 sind die Sensitivitätsschätzer für Datensatz 3 dargestellt.*
2. *Der Effekt, dass für ein bestimmtes $j_1$ der Wert $V_{j_1 1}$ sehr klein ist und es andererseits mindestens einen Index $i$ mit $2 \leq i \leq s$ gibt, sodass $V_{j_1 i}$ nicht sehr klein ist, tritt etwa bei der Analyse IR-spektroskopischer Daten auf. Für solche besteht die Gefahr, dass einige der $u(:,j)$, $j = 1, \ldots, n$, nicht in $\mathcal{F}_C$ liegen, siehe Abbildung 3.4.*

### 3.8.3 Sensitivität bei der Anwendung des Dualitätsprinzips

Mittels Satz 3.46 und Korollar 3.48 lässt sich zeigen, inwiefern sich eine Abweichung in der niedrigdimensionalen Darstellung einer Zeile von $A$ auf die duale affine Hyperebene für die niedrigdimensionalen Darstellungen der verbleibenden Spalten in $C$ auswirkt.

**Lemma 3.78** (Siehe [156])**.** *Seien $D \in \mathbb{R}_+^{k \times n}$, $s = \operatorname{rank}(D) = \operatorname{rank}_+(D) \geq 2$ und $D^{\mathbf{T}}D$ sowie $DD^{\mathbf{T}}$ irreduzibel. Seien $x \in \mathcal{M}_A$ und $\delta_x \in \mathbb{R}^{s-1}$ eine Störung für $x$. Weiter sei $y \in \mathcal{M}_C$ zulässig und liege in der zu $x$ dualen affinen Hyperebene.*

*Die Norm der induzierten Störung $\delta_y \in \mathbb{R}^{s-1}$ für $y$ ist in der Form*

$$\|\delta_y\| \geq \frac{|\delta_x^{\mathbf{T}} y|}{\|x\|}$$

*nach unten beschränkt, sofern der quadratische Term $\delta_x^{\mathbf{T}}\delta_y$ vernachlässigt wird.*

*Beweis.* Da $y$ in der zu $x$ dualen affinen Hyperebene liegt, gilt $x^{\mathbf{T}}y = -1$. Aus $(x+\delta_x)^{\mathbf{T}}(y+\delta_y) = -1$ folgt nach Subtraktion von $x^{\mathbf{T}}y = -1$, dass

$$\delta_x^{\mathbf{T}} y + x^{\mathbf{T}}\delta_y = -\delta_x^{\mathbf{T}}\delta_y = \mathcal{O}(\|\delta_x\|\|\delta_y\|)$$

für $\delta_x \to 0$ und $\delta_y \to 0$. Das Vernachlässigen des Terms $-\delta_x^{\mathbf{T}}\delta_y$ führt mittels Cauchy-Schwarzscher-Ungleichung auf

$$\|x\|\|\delta_y\| \geq |x^{\mathbf{T}}\delta_y| = |\delta_x^{\mathbf{T}} y|$$

und es gilt $\|\delta_y\| \geq |\delta_x^{\mathbf{T}} y|/\|x\|$. □

Die Abschätzung aus Lemma 3.78 hat für die Anwendung des Dualitätsprinzips für ein fixiertes $x^{(i_1)} = x$ folgenden Einfluss: Die untere Abschätzung von $\|\delta_y\|$ verhält sich reziprok zu $\|x^{(i_1)}\|$. Somit beeinflusst insbesondere $\|x^{(i_1)}\|$ und damit der Abstand von $x^{(i_1)}$ zum Ursprung die Anfälligkeit des Unterraums $\mathcal{M}_C^{[i_1]}$ aus (3.17) für Störungen. Die Berechnung von $\mathcal{M}_C^{[i_1]}$ ist für ein nahe am Ursprung liegendes $x^{(i_1)}$ vergleichsweise sensitiver als für eines, welches einen größeren Abstand zum Ursprung hat.

In Lemma 3.78 wird eine untere Abschätzung für den Einfluss von Störungen zwischen zwei zulässigen Lösungen angegeben. Im nächsten Lemma wird gezeigt, dass der Winkel zwischen $x$ und $x + \delta_x$ gleich dem Winkel zwischen deren dualen affinen Hyperebenen ist.

**Lemma 3.79** (Siehe [156] für $s = 3$)**.** *Seien $D \in \mathbb{R}_+^{k\times n}$, $s = \operatorname{rank}(D) = \operatorname{rank}_+(D) \geq 2$ und $DD^{\mathbf{T}}$ sowie $D^{\mathbf{T}}D$ irreduzibel. Seien $x^{(i_1)}$ und $x^{(i'_1)} = x^{(i_1)} + \delta_x$ aus $\mathcal{M}_A$ gegeben und dazu $\mathcal{M}_C^{[i_1]}$ und $\mathcal{M}_C^{[i'_1]}$ wie in (3.17).*

*Es gilt*

$$\measuredangle(x, x+\delta_x) = \measuredangle(\mathcal{M}_C^{[i_1]}, \mathcal{M}_C^{[i'_1]}).$$

*Beweis.* Die Hesse-Normalform der zu $x = x^{(i_1)}$ dualen affinen Hyperebene ist

$$\left(-\frac{x}{\|x\|}, y\right) = \frac{1}{\|x\|}.$$

Daraus lässt sich ableiten, dass $\mathcal{M}_C^{[i_1]}$ eine Teilmenge der affinen Hyperebene im $\mathbb{R}^{s-1}$ ist, welche senkrecht auf $x = x^{(i_1)}$ steht und den euklidischen Abstand $\|x\|^{-1}$ zum Ursprung besitzt. Analog hat die zu $x^{(i'_1)} = x + \delta_x$ duale affine Hyperebene die Hesse-Normalform

$$\left(-\frac{x+\delta_x}{\|x+\delta_x\|}, y\right) = \frac{1}{\|x+\delta_x\|}$$

und es ist $\mathcal{M}_C^{[i'_1]}$ eine Teilmenge der zu $x + \delta_x = x^{(i'_1)}$ orthogonalen affinen Hyperebene mit dem Abstand $\|x+\delta_x\|^{-1}$ zum Ursprung. Somit sind die Winkel $\measuredangle(x, x+\delta_x)$ und $\measuredangle(\mathcal{M}_C^{[i_1]}, \mathcal{M}_C^{[i'_1]})$ gleich groß. □

## 3.9 Zusammenfassung und Perspektiven

In diesem Abschnitt wurde die Menge zulässiger Lösungen $\mathcal{M}_A$ unter verschiedenen Gesichtspunkten analysiert. Dabei wurden, unter schwachen Voraussetzung an $D$, eine Reihe interessanter Eigenschaften für $\mathcal{M}_A$ nachgewiesen, die im nächsten Kapitel als Ausgangspunkte für die Entwicklung von Methoden zur Approximation von $\mathcal{M}_A$ genutzt werden. Insgesamt sind die Eigenschaften von $\mathcal{M}_A$ untereinander stark verknüpft und basieren oft auf den gleichen Grundprinzipien. Eines dieser Grundprinzipien ist die geometrische Verknüpfung über Konvexkombinationen zwischen den Mengen $\mathcal{F}_A$ und $\mathcal{I}_A$ sowie zulässigen Lösungen als Ecken von Simplexen, welche in Satz 3.23 dargelegt ist. Ein anderes Grundprinzip, das Dualitätsprinzip, beschreibt den Zusammenhang zwischen fixierten zulässigen Lösungen des einen Faktors und affinen Hyperebenen, die die Menge zulässiger Lösungen des anderen Faktors schneiden. So ist in Korollar 3.48 etwa die Schnittmenge einer, zu einem $x \in \mathcal{M}_A$, dualen affinen Hyperebene mit der Menge zulässiger Lösungen $\mathcal{M}_C$ geklärt. Weiter wurden Untersuchung für den Fall, dass die Voraussetzung $\operatorname{rank}(D) = \operatorname{rank}_+(D)$ weggelassen wird, durchgeführt.

Was die Analyse der Mengen zulässiger Lösungen angeht, sind für die Zukunft einige Fragen offen. Beispielsweise ist die Topologie für $s \geq 4$ nicht geklärt. (Für $s = 2$ ist es offensichtlich, für $s = 3$ ist die Topologie in [86] geklärt.) Aus Ergebnissen ist für Vierkomponentensysteme bekannt, dass $\mathcal{M}_A$ aus vier separaten Segmenten oder einem zusammenhängenden Segment bestehen kann. Inwiefern andere Situationen vorliegen können, ist noch nicht untersucht. Auch ist die verallgemeinerte Menge zulässiger Lösungen ($\operatorname{rank}(D) < \operatorname{rank}_+(D)$) noch nicht tiefer analysiert. Hier sind die Fragen nach den möglichen Anzahlen von Segmenten für verschiedene Kombinationen $s = \operatorname{rank}(D)$ und $m = \operatorname{rank}_+(D)$ sowie, wie ein geometrisch konstruktiver Algorithmus zu deren Berechnung aussehen kann, naheliegend. Weiter wird auch die, in dieser Schrift nur in Ansätzen vorgenommene, Analyse des Einflusses von Störungen auf einzelne Eigenschaften von $\mathcal{M}_A$ voranzutreiben sein. Ein anderes großes Forschungsfeld für die mittlere Zukunft wird die Ausweitung der Mengen zulässiger Lösungen bezüglich der nichtnegativen Matrixfaktorisierung auf Tensorprodukte sein [26, 39, 94, 98, 99, 125, 165, 177].

# 4 Berechnungsmethoden

In Kapitel 3 wurden die Mengen zulässiger Lösungen detailliert untersucht. Unter der Annahme schwacher Voraussetzungen, konnten wichtige Eigenschaften nachgewiesen werden. Darauf aufbauend werden in diesem Kapitel verschiedene Methoden zur Bestimmung von Approximationen an die Mengen zulässiger Lösungen herausgearbeitet und analysiert. Alle im Folgenden vorgestellten Methoden funktionieren nach dem Prinzip, dass nur die Ränder von $\mathcal{M}_A$ und $\mathcal{M}_C$ bestimmt werden. Bei einer Methode erfolgt dies geometrisch konstruktiv, bei den anderen numerisch approximativ. Wenngleich in Kapitel 3 stets der allgemeine Fall $s \geq 2$ untersucht wurde, liegt bei den in diesem Kapitel vorgestellten Methoden ein Schwerpunkt auf $s = 3$. Für zwei Methoden werden Erweiterungen für $s = 4$ kurz vorgestellt und eine Methode ist für $s \geq 2$ anwendbar. Analog zum vorherigen Kapitel werden die Methoden zur Approximation von $\mathcal{M}_A$ entwickelt; deren Anwendung zur Approximationen von $\mathcal{M}_C$ ist analog.[1]

Insgesamt werden eine geometrisch konstruktive und drei numerisch approximative Algorithmen vorgestellt und untersucht. Bei der geometrisch konstruktiven Methode lassen sich für $s = 3$ Randpunkte von $\mathcal{M}_A \subset \mathbb{R}^2$ direkt konstruieren. Dies ist bei den numerischen Methoden anders, denn es werden nur Approximationen für Randpunkte bestimmt. Dabei haben alle numerischen Methoden gemein, dass sie eine Unterroutine zur Klassifizierung, ob ein $x$ zu $\mathcal{M}_A$ gehört oder nicht, benötigen. Diesbezüglich muss während der Iteration in der Regel über sehr viele $x$ entschieden werden. Für die numerischen Methoden wird dafür auf den impliziten Ansatz mit der Optimierung einer Zielfunktion zurückgegriffen, um auch für störungsbehaftete Daten sinnvolle Ergebnisse generieren zu können. Ein Nachteil dieser impliziten Herangehensweise ist, dass die Einstufung $x \notin \mathcal{M}_A$ mitunter nicht sicher getroffen werden kann, da sie von einer erfolgreichen numerischen Minimierung abhängt. Die Klassifizierungsroutine beeinflusst somit sowohl die Stabilität der gesamten Approximationsmethode als auch ihren Rechenaufwand entscheidend. Bei den vorgestellten Prozeduren zur Approximation von $\mathcal{M}_A$ liegen die Schwerpunkte auf den Polygon inflation Methoden (direkter und inverser Typ) und dem Strahlenalgorithmus.

Dieses Kapitel ist folgendermaßen gegliedert: Zunächst wird in Abschnitt 4.1 der in den Polygon inflation Methoden und dem Strahlenalgorithmus genutzte Ansatz zur Klassifizierung eines $x \in \mathbb{R}^{s-1}$ vorgestellt [147,152,154]. Auf die Einbindung von Störungen und Details zur Implementierung wird eingegangen. In Abschnitt 4.2 wird die Bestimmung der Menge zulässiger Lösungen $\mathcal{M}_A$ für $s = 2$ erläutert. Die geometrische Konstruktion von $\mathcal{M}_A$ für $s = 3$ wird in 4.3 erläutert und analysiert [12, 16, 17, 85–88, 138]. Anschließend wird in 4.4 die numerische Methode der Dreieckseinschließung behandelt [56, 58]. In 4.5 werden die beiden Polygon inflation Algorithmen (direkter und indirekter Typ[2]) präsentiert [53, 55, 147, 152, 154]. Der darauf aufbauende inverse Polyhedron inflation Algorithmus wird in 4.6 in seinen Grundzügen vorgestellt [131]. In 4.7 wird die Strahlenmethode [157] als ein universelles Werkzeug zur Berechnung von $\mathcal{M}_A$ eingeführt, welche weder auf eine bestimmte Anzahl $s$ an Spalten von $C$ und Zeilen von $A$, noch auf Mengen $\mathcal{M}_A$ bestimmter Topologien beschränkt ist. In den weiterführenden Abschnitten 4.8, 4.9 und 4.10 werden die Einbindung von Restriktionen in Form von Zusatzinformationen über $C$ oder $A$ bei

[1] In den Schlüsselpublikationen zu den Mengen zulässiger Lösungen werden ebenso nur Methoden zur Approximation von $\mathcal{M}_A$ vorgestellt. Die Anwendung auf $D^{\mathbf{T}}$ führt unter Berücksichtigung der Singulärwerte von $D$ auf $\mathcal{M}_C$.

[2] Beim direkten Typ werden die einzelnen Segmente von $\mathcal{M}_A$ direkt approximiert und beim indirekten Typ mittels des Schnitts von zwei Obermengen von $\mathcal{M}_A$.

der Berechnung von $\mathcal{M}_A$ [12, 13, 57, 156], die Hinzunahme von Regularisierungen zur Reduktion von $\mathcal{M}_A$ [12, 132, 158] sowie alternative Berechnungsmethoden im Umfeld der Mengen zulässiger Lösungen behandelt [50, 55, 167, 168, 178].

Ein, in dieser Schrift nicht untersuchter, numerischer Ansatz zur Approximation der Menge zulässiger Lösungen ist es, einen bestimmten Bereich des $\mathbb{R}^{s-1}$ mittels eines (äquidistanten) Gitters zu diskretisieren und die einzelnen Gitterpunkte bezüglich ihrer Zugehörigkeit zu $\mathcal{M}_A$ zu überprüfen [1, 2, 13, 57, 173]. Dieser *grid search*-Ansatz ist einfach aufgebaut und brachial. Die in diesem Kapitel vorgestellten Algorithmen sind deutlich leistungsstärker.

## 4.1 Numerische Klassifizierung

Ein wichtiger Teilschritt zur Approximation von $\mathcal{M}_A$ ist die Entwicklung einer geeigneten (numerischen) Prozedur zur Klassifizierung eines $x \in \mathbb{R}^{s-1}$ als *zulässig* oder *nicht zulässig*. Das geometrische Argument aus Satz 3.26, welches in analoger Form in [17, 138] genutzt wird, ist eine elegante Möglichkeit dafür.[3] Neben der geometrischen Klassifizierung sind zwei weitere numerische Möglichkeiten bekannt [56, 152, 154], welche vorgestellt werden. Diese unterscheiden sich hauptsächlich bezüglich des Rechenaufwands und der Behandlung störungsbehafteter Daten.

Um die Vergleiche der einzelnen Methoden nicht zu verfälschen, wird der folgend vorgestellte Ansatz (in FacPack genutzt) bei den späteren Vergleichsrechnungen für alle numerischen Algorithmen zur Bestimmung von $\mathcal{M}_A$ verwendet. Die Klassifizierung basiert auf der Anwendung von Straffunktionen und der Minimierung (im Sinne der kleinsten Quadrate) einer daraus aufgestellten Zielfunktion. Wird eine vorgegebene Schranke nicht überschritten, so wird ein $x$ als *zulässig* klassifiziert, andernfalls als *nicht zulässig*.

### 4.1.1 Anforderungen an eine zulässige Lösung

Die im Folgenden vorgestellte Prozedur ist für $s \geq 2$ anwendbar. Für die Anwendung im Polygon inflation Algorithmus ist $s = 3$. Für die beiden Typen der Polygon inflation Klasse werden unterschiedliche Bewertungen benötigt. Die direkte Variante erfordert ein Testverfahren, ob $x \in \mathcal{M}_A$ gilt, wohingegen bei der inversen Variante einmal $x \in \mathcal{F}_A$ und einmal $x \in \mathcal{M}_A^*$, mit einer in (4.23) definierten Menge $\mathcal{M}_A^*$, getestet wird.

Um die Menge zulässiger Lösungen $\mathcal{M}_A$ auch für störungsbehaftete Daten stabil bestimmen zu können, müssen unter Umständen betragskleine negative Einträge in $C$ und $A$ zugelassen werden. Diesbezüglich werden zwei Faktoren $C$ und $A$ als *zulässig* klassifiziert, falls

$$\frac{\min\limits_{j=1,\dots,k} C_{ji}}{\max\limits_{j=1,\dots,k} |C_{ji}|} \geq -\varepsilon_c, \qquad \frac{\min\limits_{j=1,\dots,n} A_{ij}}{\max\limits_{j=1,\dots,n} |A_{ij}|} \geq -\varepsilon_a, \quad i = 1, \dots, s, \tag{4.1}$$

mit zwei Steuerparametern $\varepsilon_c \geq 0$ und $\varepsilon_a \geq 0$.

[3] Die in Abschnitt 4.3 vorgestellte geometrische Konstruktion von $\mathcal{M}_A$ für $s = 3$ kommt methodenbedingt ohne Klassifizierungen einzelner $x \in \mathbb{R}^2$ aus.

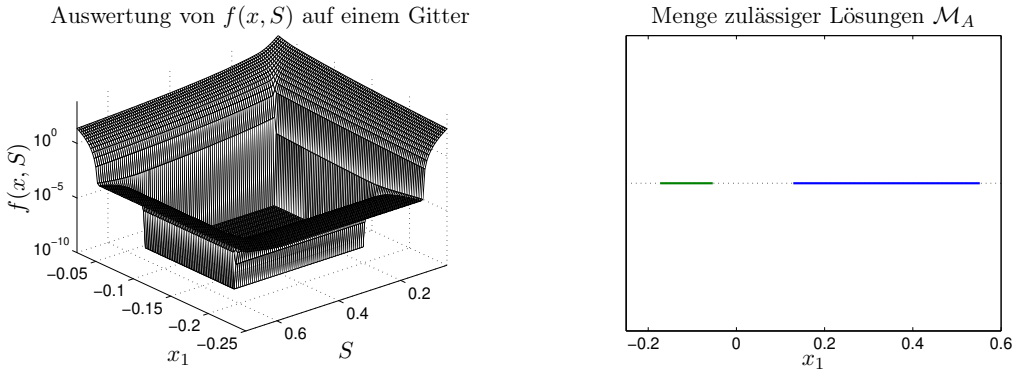


**Abbildung 4.1:** Auswertung der Zielfunktion $f(x,S)$ aus (4.2) und die daraus resultierende Menge zulässiger Lösungen $\mathcal{M}_A$ für Datensatz 1. Links: Dargestellt ist die Zielfunktion über dem Intervall $(x,S) \in [0.05,\ 0.7] \times [-0.25,\ 0.6]$. Rechts: Dargestellt ist $\mathcal{M}_A$ für die Abbruchschranke $\varepsilon_f = 10^{-10}$ aus (4.4). Für $s = 2$ ist $S \in \mathbb{R}$ und es ist genau dann $f(x,S) = 0$, wenn $x$ in einem Segment von $\mathcal{M}_A$ liegt und $S$ in dem anderen.

### 4.1.2 Umsetzung in einer Zielfunktion

Die Umsetzung der Forderungen aus (4.1) erfolgt numerisch in Form eines nichtlinearen Ausgleichsproblems. Als Zielfunktion fungiert $f : \mathbb{R}^{s-1} \times \mathbb{R}^{(s-1)\times(s-1)} \to \mathbb{R}$,

$$\begin{aligned} f(x,S) = \frac{1}{2}\Bigg( \sum_{i=1}^{s}\sum_{j=1}^{k} \left( \min\left(0, \frac{C_{ji}}{\|C(:,i)\|_\infty} + \varepsilon_c\right)\right)^2 + \\ \sum_{i=1}^{s}\sum_{j=1}^{n} \left( \min\left(0, \frac{A_{ij}}{\|A(i,:)\|_\infty} + \varepsilon_a\right)\right)^2 + \|I_s - TT^+\|_F^2 \Bigg), \end{aligned} \tag{4.2}$$

sodass mit einem Steuerparameter $\varepsilon_f \geq 0$ und der Funktion $F : \mathbb{R}^{s-1} \to \mathbb{R}$,

$$F(x) = \min_{S \in \mathbb{R}^{(s-1)\times(s-1)}} f(x,S) \tag{4.3}$$

ein $x$ als *zulässig* klassifiziert wird, wenn

$$F(x) \leq \varepsilon_f \tag{4.4}$$

gilt und andernfalls als *nicht zulässig*. In (4.2) bezeichnet $I_s$ die $s$-dimensionale Einheitsmatrix und die Faktoren $C$ und $A$ werden mittels (2.2) und (2.5) aus $x$ und $S$ bestimmt. In der Anwendung ist beispielsweise $\varepsilon_f = 10^{-12}$ sinnvoll.

**Bemerkung 4.1.** *In (2.6) und (2.7) sind die Mengen zulässiger Lösungen rigoros, das heißt ohne die Berücksichtigung von Störungen, definiert. Die Mengen zulässiger Lösungen unter Berücksichtigung von betragskleinen negativen Einträgen in $C$ oder $A$ werden im Zuge dieser Schrift ebenfalls mit den Symbolen $\mathcal{M}_A$ und $\mathcal{M}_C$ bezeichnet, auch wenn dies formal (2.6) und (2.7) widerspricht. Diese Bezeichnungen erfolgen aus Übersichtlichkeitsgründen, da es verschiedene Ansätze zum Einbeziehen von Störungen gibt und nicht zwischen allen mit jeweils eigenen Symbolen unterschieden werden soll.*

### 4.1.3 Teil 1 der Klassifizierung: Schnelltest mittels einer Obermenge

Um den Rechenaufwand möglichst gering zu halten, wird die Klassifizierung für die direkte Variante des später in Abschnitt 4.5 vorgestellten Polygon inflation Algorithmus zweigeteilt

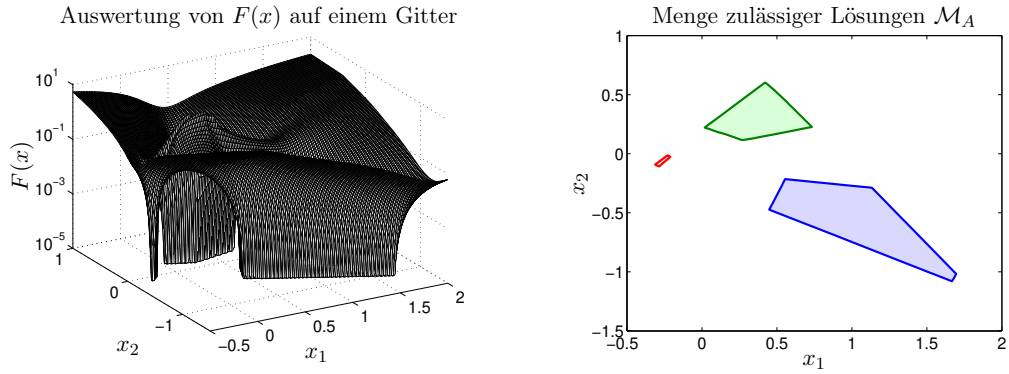


**Abbildung 4.2:** Die Zielfunktion $F(x) = \min_S f(x, S)$ aus (4.3) mit $f(x, S)$ aus (4.2) und die daraus resultierende Menge zulässiger Lösungen $\mathcal{M}_A$ für Datensatz 2. Links: Dargestellt ist die Auswertung der Zielfunktion $F(x)$ über dem Intervall $[-0.5,\ 2] \times [-1.5,\ 1]$. Es wurde ein $150 \times 151$-Gitter gewählt und die Grafik bei $z = 10^{-5}$ abgeschnitten. Rechts: Dargestellt ist $\mathcal{M}_A$ zum gleichen Intervall und für die Abbruchschranke $\varepsilon_f = 10^{-5}$.

durchgeführt. Wegen der Konstruktion von $A$ mittels $T$ aus (2.2) wird zunächst der Schnelltest ausgeführt, ob $x$ die Ungleichung

$$\frac{1}{2}\sum_{i=1}^{n}\left(\min\left(0,\ \frac{(1, x^{\mathbf{T}})(V(i,:))^{\mathbf{T}}}{\|(1, x^{\mathbf{T}})V^{\mathbf{T}}\|_\infty} + \varepsilon_a\right)\right)^2 \leq \varepsilon_f \tag{4.5}$$

erfüllt. Ist dies nicht der Fall, so ist $x \notin \mathcal{M}_A$. Ist (4.5) erfüllt, so wird die weit rechenintensivere Überprüfung, ob es ein $S \in \mathbb{R}^{(s-1)\times(s-1)}$ gibt, sodass $C$ und $A$ die restlichen Bedingungen erfüllen, durchgeführt. Dazu wird die oben genannte Zielfunktion $f(x, S)$ aus (4.2) genutzt. Dabei wird der für (4.5) bereits berechnete Teil, der für die Auswertung von $f(x, S)$ bei konstantem $x$ ebenfalls konstant bleibt, nicht neu berechnet.

In Abbildung 4.1 ist die Zielfunktion $f(x, S)$ auf dem Intervall $(x, S) \in [0.05,\ 0.7] \times [-0.25,\ 0.6]$ für Datensatz 1 ausgewertet und zusammen mit der Menge $\mathcal{M}_A$ dargestellt. Weiter ist in Abbildung 4.2 die Zielfunktionen $F(x)$ auf dem Intervall $[-0.5,\ 2] \times [-1.5,\ 1]$ für Datensatz 2 ausgewertet und zusammen mit der daraus resultierenden Menge $\mathcal{M}_A$ dargestellt.

### 4.1.4 Teil 2 der Klassifizierung: Minimierung der Zielfunktion

Sofern Ungleichung (4.5) für $x$ erfüllt ist, gilt $x \in \mathcal{F}_A$. Anschließend wird überprüft, ob sogar $x \in \mathcal{M}_A$ gilt. Zur Bestimmung von $F(x)$ wird das in $F$ enthaltene nichtlineare Ausgleichsproblem gelöst. Es ist $x \in \mathcal{M}_A$, falls $F(x) \leq \varepsilon_f$ gilt und $x \notin \mathcal{M}_A$ andernfalls. Das Ausgleichsproblem garantiert, dass $x$, sofern $F(x) \leq \varepsilon_f$ gilt, alle Anforderungen an eine zulässige Lösung mit Rücksicht auf den Steuerparameter $\varepsilon_f$ erfüllt.

In FacPack wird zur Optimierung von $f(x, S)$ und damit zur Auswertung von $F(x)$ die leistungsstarke ACM Routine NL2SOL [34,35] genutzt. Die sich ergebende Prozedur für ein $x \in \mathbb{R}^{s-1}$ ist in Klassifizierung 4.2 angegeben und wird beispielsweise in dem später vorgestellten Polygon inflation Algorithmus (direkter Typ) genutzt. Für die inverse Variante des Polygon inflation Algorithmus wird die Einstufung modifiziert (siehe die Klassifizierungen 4.24 und 4.25), worauf später auf Seite 76 eingegangen wird.

**Klassifizierung 4.2** (Direkter Typ)**.** *Für ein $x \in \mathbb{R}^{s-1}$ wird zunächst der Schnelltest, ob (4.5) erfüllt ist, durchgeführt. Sollte dies der Fall sein, wird überprüft, ob zusätzlich (4.4) gilt. Ist auch dies der Fall, so wird $x$ als* zulässig *klassifiziert. Ist (4.5) nicht erfüllt oder gilt (4.4) nicht, so wird $x$ als* nicht zulässig *klassifiziert. In Abbildung 4.3 ist der Entscheidungsbaum dargestellt.*

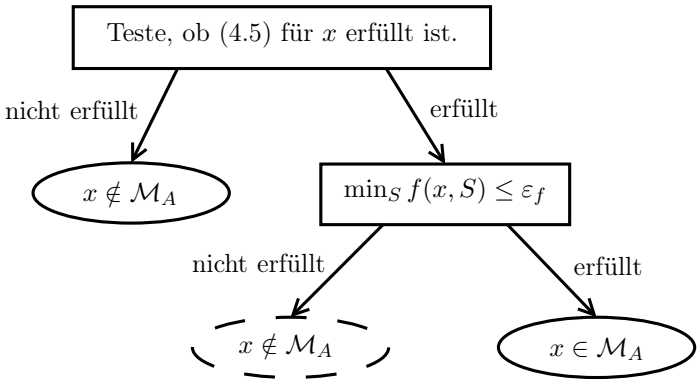


**Abbildung 4.3:** Entscheidungsbaum zur numerischen Klassifizierung. Die beiden sicheren Entscheidung sind mittels durchgezogenen Linien gekennzeichnet. Die Entscheidung, die nicht mit Sicherheit getroffen werden kann, da sie auf einer numerischen Minimierung von (4.4) beruht, ist durch eine gestrichelte Kurve gekennzeichnet.

In Bezug auf die Zuverlässigkeit der Klassifizierung sind drei Fälle zu unterscheiden: Sofern $x$ die Bedingung des Schnelltests erfüllt und die Minimierung zudem auf $F(x) \le \varepsilon_f$ führt, gilt sicher $x \in \mathcal{M}_A$. Ebenso gilt mit Sicherheit $x \notin \mathcal{M}_A$, falls $x$ die Bedingung des Schnelltests nicht erfüllt. Jedoch kann ein $x$, das die Bedingung des Schnelltests erfüllt, für den aber die numerische Minimierung auf ein $\widetilde{S}$ mit $f(x,\widetilde{S}) > \varepsilon_f$ führt, nicht mit Sicherheit als $x \notin \mathcal{M}_A$ klassifiziert werden. Eine falsche Einstufung erfolgt, falls es statt $\widetilde{S}$ ein $S^{(\mathrm{opt})}$ gibt, für das tatsächlich $f(x,S^{(\mathrm{opt})}) \le \varepsilon_f$ gilt. Da falsch klassifizierte $x$ die Genauigkeit der Approximation des Randes von $\mathcal{M}_A$ beeinträchtigen können, liegt bei der Implementierung ein Fokus darauf, falsche Klassifizierungen von vornherein zu vermeiden oder zumindest im Nachhinein zu detektieren. Dazu sind einige Details zur algorithmischen Umsetzung in der folgenden Bemerkung aufgeführt.

**Bemerkung 4.3.** *Bei den* FacPack*-Implementierungen der später vorgestellten Polygon inflation Methoden und dem später vorgestellten Strahlenalgorithmus werden die Optimierungen im Hinblick auf korrekte Klassifizierungen $x \notin \mathcal{M}_A$ in der Regel mehrfach und mit anderen Startwerten durchgeführt. Automatisiert kommt an bestimmten Stellen auch ein genetischer Algorithmus zum Einsatz. Zudem wird beim Polygon inflation Algorithmus der Rand nach jeder Berechnung eines neuen Randpunktes auf mögliche „Ausreißer“ untersucht. Diese werden anschließend mittels erneuter Optimierung, unter Nutzung verbesserter Startwerte, auf Unregelmäßigkeiten untersucht und gegebenenfalls modifiziert oder eliminiert.*

### 4.1.5 Andere Ansätze zur Klassifizierung

Die Funktion $f(x,S)$ ist leicht unterschiedlich zu der in [1, 56, 58, 173] genutzten Funktion ssq. Dies betrifft den Rechenaufwand ($\mathcal{O}(k+n)$ je Funktionsaufruf für $f$ aus (4.2) im Vergleich zu $\mathcal{O}(kn)$ je Funktionsaufruf für ssq) und die Berücksichtigung von Störungen beziehungsweise betragskleiner negativer Einträge, siehe [154] für einen ausführlicheren Vergleich.

Die Funktion $\mathsf{ssq} : \mathbb{R}^{s-1} \times \mathbb{R}^{(s-1)\times(s-1)} \to \mathbb{R}$ ist definiert als

$$\mathsf{ssq}(x,S) = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(D_{ij} - \sum_{\ell=1}^{s}\max(0,C_{i\ell})\max(0,A_{\ell j})\right)^2, \tag{4.6}$$

wobei $C$ und $A$ mittels (2.2) aus $T$ bestimmt werden und sich $T$ mittels (2.5) aus $x$ und $S$ zusammensetzt.

**Klassifizierung 4.4.** *Mittels der Funktion* $\mathsf{ssq}(x,S)$ *wird ein* $x$ *als* zulässig *klassifiziert, sofern*

$$G : \mathbb{R}^{s-1} \to \mathbb{R}, \qquad G(x) = \min_{S\in\mathbb{R}^{(s-1)\times(s-1)}} \mathsf{ssq}(x,S) \le \tilde{\varepsilon}_f$$

*gilt mit einem geeigneten $\tilde{\varepsilon}_f \geq 0$. Andernfalls wird $x$ als* nicht zulässig *klassifiziert. Im Vergleich zur Klassifizierung mit $f(x, S)$ aus (4.2) wie in (4.4) gilt für die Anwendung auf störungsbehaftete Daten in der Regel die Relation $\tilde{\varepsilon}_f \gg \varepsilon_f$.*

## 4.2 Menge zulässiger Lösungen für $s = 2$

Für $s = 2$ ist die Menge zulässiger Lösungen $\mathcal{M}_A$ eine Teilmenge der reellen Zahlen. Aufgrund der Eigenschaften von $\mathcal{M}_A$ ergeben sich für die Berechnung, sowohl für störungsfreie als auch für störungsbehaftete Daten, einige Vereinfachungen. Insbesondere werden für $s = 2$ deutlich weniger Rechenressourcen beansprucht. Entscheidend ist die Kombination aus den Eigenschaften, dass $\mathcal{M}_A$ nicht den Nullpunkt enthält, siehe Satz 3.10, dass der Schnitt eines, vom Ursprung ausgehenden, Strahls mit $\mathcal{M}_A$ unterbrechnungsfrei ist, siehe Satz 3.15, und dass $\mathcal{M}_A$ beschränkt ist, siehe Satz 3.8. Dies führt darauf, dass $\mathcal{M}_A$ aus zwei disjunkten Teilintervallen besteht, eines befindet sich komplett auf dem negativen Teil des reellen Zahlenstrahls und eines komplett auf dem positiven Teil. Ohne die Berücksichtigung von Störungen, etwa wie im Sinne von (4.1), lassen sich die Intervallgrenzen direkt bestimmen. Andernfalls gelingt dies mittels des Einsatzes einer numerischen Bewertungsroutine. Die Berechnung von $\mathcal{M}_A$ für $s = 2$ ist gut untersucht [1, 2, 105, 147, 173] und unproblematisch, weshalb sie in dieser Schrift nur kurz erläutert und diskutiert wird.

In [105] wird zwar nicht direkt die Menge zulässiger Lösungen $\mathcal{M}_A$, wie sie in (2.6) definiert ist, untersucht, jedoch führen die dortigen Untersuchungen auf gleichwertige Ergebnisse. Im Unterschied zu der hier genutzten Form, wird nicht die niedrigdimensionale Darstellung genutzt. Da die grundlegenden Ideen gleichbedeutend sind, werden zunächst die Ansätze daraus erläutert und anschließend wird der Übergang zu $\mathcal{M}_A$ vollzogen.

### 4.2.1 Ansatz von Lawton und Sylvestre

In der wegweisenden Arbeit [105] werden Zweikomponentensysteme untersucht und die Grundlagen moderner Methoden zur Reinkomponentenzerlegung (spektroskopischer Daten) gelegt. Das Problem der uneindeutigen Faktorisierung wird diskutiert. Die niedrigdimensionale Darstellung (2.5) wird nicht genutzt, auch weil es die Dimension $s = 2$ für eine übersichtliche Visualisierung nicht erfordert. Stattdessen wird eine Bestimmung von $A$ in der Form

$$A(i,:) = T_{i1}V(:,1)^{\mathbf{T}} + T_{i2}V(:,2)^{\mathbf{T}}, \quad i = 1, 2,$$

gewählt mit $T \in \mathbb{R}^{2\times 2}$, jedoch ohne die spezielle Struktur aus (2.5). Für eine Analyse der ersten Zeile von $A$ führt dies ohne Berücksichtigung einer speziellen Skalierung auf

$$A(1,:) = \xi_1 V(:,1)^{\mathbf{T}} + \xi_2 V(:,2)^{\mathbf{T}}.$$

Somit ist die Menge $\mathcal{X}$ von Punkten $(\xi_1, \xi_2)$ gesucht, die auf zulässige Zeilen für $A$ führen. Im Unterschied zur Menge $\mathcal{M}_A$ sind die Punkte bei diesem Ansatz nicht speziell skaliert und die Menge $\mathcal{X}$ ist nicht beschränkt (sofern $D \geq 0$ und $\operatorname{rank}(D) = 2$ gibt es eine Lösung und wegen der Skalierungsmehrdeutigkeit auch unendlich viele). Unter den Annahmen, die auch zu der Menge $\mathcal{M}_A$ bekannt sind ($D^{\mathbf{T}}D$ irreduzibel, $V(:,1) > 0$), ergibt sich für $\xi_1$ und $\xi_2$, dass

$$\xi_1 \geq -\xi_2 \frac{V_{i2}}{V_{i1}}, \quad i = 1, \ldots, n, \tag{4.7}$$

sein muss. Dies ist mit den Bedingungen für $\mathcal{F}_A$ vergleichbar. Mit der notwendigen Wahl $\xi_1 > 0$ führt dies auf die Grenzen

$$\xi_2 \geq -\xi_1 \min_{i:\, V_{i2}>0} \frac{V_{i1}}{V_{i2}} \qquad \text{und} \qquad \xi_2 \leq -\xi_1 \max_{i:\, V_{i2}<0} \frac{V_{i1}}{V_{i2}}. \tag{4.8}$$

Weiter wird, analog zur Bedingung aus Satz 3.26, von $\xi$ gefordert, dass

$$\xi_2 \geq \xi_1 \max_i \frac{\sigma_2 U_{i2}}{\sigma_1 U_{i1}} \qquad \text{oder} \qquad \xi_2 \leq \xi_1 \min_i \frac{\sigma_2 U_{i2}}{\sigma_1 U_{i1}}. \tag{4.9}$$

Aus (4.8) und (4.9) ergibt sich schließlich

$$\frac{\xi_2}{\xi_1} \in \left[ -\min_{i:\, V_{i2}>0} \frac{V_{i1}}{V_{i2}},\ \min_i \frac{\sigma_2 U_{i2}}{\sigma_1 U_{i1}} \right] \cup \left[ \max_i \frac{\sigma_2 U_{i2}}{\sigma_1 U_{i1}},\ -\max_{i:\, V_{i2}<0} \frac{V_{i1}}{V_{i2}} \right] \tag{4.10}$$

als Forderung und $\mathcal{X}$ ist die Menge aller $\xi \in \mathbb{R}^2$ mit $\xi_1 > 0$, die (4.10) erfüllen, also

$$\mathcal{X} = \left\{ \xi \in \mathbb{R}^2 :\ -\min_{i:\, V_{i2}>0} \frac{V_{i1}}{V_{i2}} \leq \frac{\xi_2}{\xi_1} \leq \min_i \frac{\sigma_2 U_{i2}}{\sigma_1 U_{i1}} \quad \text{oder} \quad \max_i \frac{\sigma_2 U_{i2}}{\sigma_1 U_{i1}} \leq \frac{\xi_2}{\xi_1} \leq -\max_{i:\, V_{i2}<0} \frac{V_{i1}}{V_{i2}} \right\}. \tag{4.11}$$

Diesbezüglich sei auch auf [154] und [147] verwiesen, wobei dort die Skalierung $\xi_1 = 1$ genutzt wird. In [147] wird dasselbe Resultat unter Nutzung des Dualitätsprinzips hergeleitet. Die Verbindung zu $\mathcal{M}_A$ ergibt sich direkt und ist in Abschnitt 4.2.2 thematisiert.

In Abbildung 4.4 ist die geometrische Konstruktion der Menge $\mathcal{X}$ für Datensatz 1 demonstriert. Die Daten sind zwar gestört, jedoch enthält $D$ keine negativen Einträge. Die beiden Kegel laufen auf den Nullpunkt zu, enthalten diesen aber nicht. Vielmehr besteht $\mathcal{X}$, analog wie die Menge $\mathcal{M}_A$ für $s = 2$, aus zwei Segmenten und es gibt kein zulässiges $\xi$ mit $\xi_1 = 0$ oder $\xi_2 = 0$:

**Bemerkung 4.5.** *Seien die Voraussetzungen aus Satz 3.10 erfüllt. Es gilt zwar, dass es für jedes positive $\varepsilon$ ein $\xi \in \mathcal{X}$ mit $\|\xi\|_2 < \varepsilon$ gibt, jedoch folgt aus $\xi \in \mathcal{X}$ stets $\xi_1, \xi_2 \neq 0$. Es besteht $\mathcal{X}$ somit stets aus zwei getrennten Segmenten. Dass $\xi_1 = 0$ nicht möglich ist, folgt aus Satz 3.3, dass $\xi_2 = 0$ nicht möglich ist, ergibt sich aus Satz 3.10.*

Der oben beschriebene Zugang wird in Publikationen oft als *Lawton-Sylvestre method* oder *Lawton-Sylvestre plot* bezeichnet [6, 12, 57, 58, 78, 178] und ist für $s = 2$ die Vorstufe zur Menge zulässiger Lösungen $\mathcal{M}_A$. Die Anwendung für $s = 3$ ist zwar ebenso möglich (Kegel im $\mathbb{R}^3$ vom Ursprung aus, aber ohne $(0,\ 0,\ 0)^{\mathbf{T}}$), siehe beispielsweise [24, 120, 142]. Eine Vorgehensweise wie für $\mathcal{M}_A$ ist aber in dem Sinne geeigneter, da die Anzahl der Freiheitsgrade um eins reduziert ist. Letztlich ist die Arbeit [105] der Ursprung für viele Arbeiten zur Menge zulässiger Lösungen $\mathcal{M}_A$ und zur, in Abschnitt 4.3 beschriebenen, geometrischen Konstruktion von $\mathcal{M}_A$ für $s = 3$.

### 4.2.2 Niedrigdimensionale Darstellung für $s = 2$

Bei dem oben beschriebenen Ansatz wird die niedrigdimensionale Darstellung aus (2.5) nicht genutzt. Dessen Anwendung führt mit der Skalierung aus (2.5) auf

$$\mathcal{M}_A = \{x \in \mathbb{R} :\ \text{es gibt ein } \xi \in \mathcal{X} \text{ mit } \xi = (1, x)\},$$

sodass gilt

$$\mathcal{M}_A = [a,\ b] \cup [c,\ d] \tag{4.12}$$

mit

$$a = -\min_{i:\, V_{i2}>0} \frac{V_{i1}}{V_{i2}}, \quad b = \min_{j=1,\ldots,k} \frac{\sigma_2 U_{j2}}{\sigma_1 U_{j1}}, \quad c = \max_{j=1,\ldots,k} \frac{\sigma_2 U_{j2}}{\sigma_1 U_{j1}}, \quad d = -\max_{i:\, V_{i2}<0} \frac{V_{i1}}{V_{i2}}. \tag{4.13}$$

Dabei sind auch andere Skalierungen/Normierungen, wie beispielsweise die in [17, 138] genutzte mit $\|T(i,:)\|_1 = 1$, $i = 1, \ldots, s$, möglich. Für alle sinnvollen Skalierungen (das heißt $\xi_1$ ergibt sich

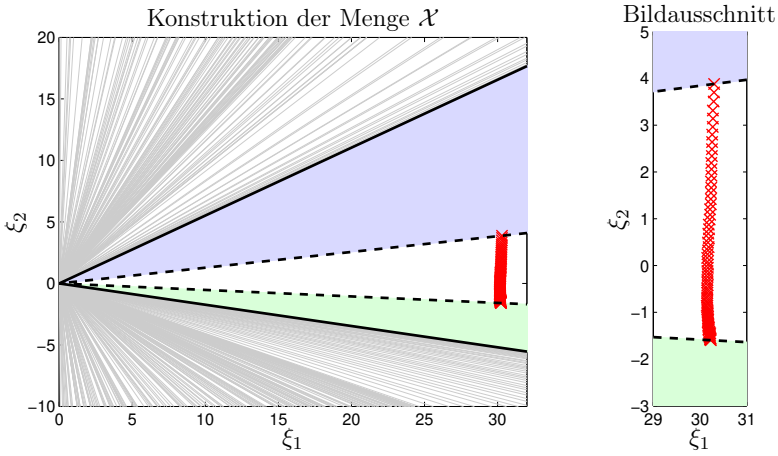


**Abbildung 4.4:** Anwendung der geometrischen Konstruktion für $s = 2$ ohne niedrigdimensionale Darstellung (Lawton-Sylvestre plot) für Datensatz 1. Die grauen Linien sind die jeweiligen Grenzen der affinen Halbräume, die sich zu den Bedingungen $\xi_1 \geq -\xi_2 V_{i2}/V_{i1}$, $i = 1 \ldots, n$, ergeben, siehe (4.7). Nicht alle $n = 1941$ Linien sind eingezeichnet. Die durchgezogenen schwarzen Linien sind die äußeren Grenzen, siehe (4.8). Die roten Markierungen sind die $(\sigma_1 U_{j1}, \sigma_2 U_{j2})$, $j = 1, \ldots, k$. Die gestrichelten schwarzen Linien sind die inneren Grenzen, siehe (4.9). Die beiden Segmente der Menge $\mathcal{X}$ aus (4.11) sind farbig unterlegt. Wie in Abschnitt 4.2.2 beschrieben, ergibt sich $\mathcal{M}_A$ als Schnitt der Linie $\xi_1 = 1$ mit $\mathcal{X}$, vergleiche Abbildung 4.1 (rechts).

eindeutig zu $\xi_2$) lässt sich eine, um einen Freiheitsgrad/eine Dimension reduzierte, Teilmenge von $\mathcal{X}$ abspalten, vergleiche auch Abbildung 5 in [147].

Für störungsfreie Daten lassen sich die Intervallgrenzen aus (4.13) sehr einfach bestimmen. Wegen $s = 2$ sind $w(:, i)$, $i = 1, \ldots, k$, und $u(:, j)$, $j = 1, \ldots, n$, skalare Werte. Somit lassen sich die Intervallgrenzen aus (4.13) auch als

$$a = -\min_{j:\, u(1,j)>0} \frac{1}{u(1,j)}, \quad b = \min_{i=1,\ldots,k} w(1,i), \quad , c = \max_{i=1,\ldots,k} w(1,i), \quad d = -\max_{j:\, u(1,j)<0} \frac{1}{u(1,j)}$$

berechnen.

**Bemerkung 4.6.** *Für $s = 2$ begrenzen die Ränder $b$ und $c$ des Intervalls $\mathcal{I}_A$ die Menge $\mathcal{M}_A$ von innen. Insbesondere gehören sie auch zu $\mathcal{M}_A$. Dies ist gleichbedeutend damit, dass sich für den Fall* $\operatorname{rank}(D) = 2$ *eine nichtnegative Matrixfaktorisierung konstruieren lässt, wobei $A$ aus zwei bestimmten Zeilen von $D$ besteht [27, 143]. Dies sind die zu $b$ und $c$ Gehörigen. Ebenso gehört der aus den zu $a$ und $d$ gehörigen Spalten von $D$ gebildete Faktor $C$ zu einer nichtnegativen Matrixfaktorisierung von $D$.*

### 4.2.3 Anwendung für störungsbehaftete Daten

Es ist bekannt, dass eine nichtnegative Matrix vom Rang $s = 2$ stets eine nichtnegative Matrixfaktorisierung mit Faktoren vollen Ranges besitzt [27, 143, 170]. Somit sind die Mengen zulässiger Lösungen nicht leer. Für störungsbehaftete Daten gilt, sofern $k, n > 2$, in der Regel $\operatorname{rank}(D) > 2$ und mitunter auch $\min_{ij} D_{ij} < 0$. Nichtsdestotrotz stellt die Annahme $s = 2$ für die Berechnung der Menge zulässiger Lösungen $\mathcal{M}_A$ bei störungsbehafteten Daten eine besondere Situation dar. Ist nämlich die Niedrigrangapproximation

$$\widetilde{D} = U\Sigma(:, 1:2)(V(:, 1:2))^{\mathbf{T}}, \tag{4.14}$$

welche mittels einer abgeschnittenen Singulärwertzerlegung bestimmt ist, ebenfalls nichtnegativ, so folgt automatisch, dass $\mathcal{M}_A$ nicht leer ist. Somit können $\varepsilon_a = \varepsilon_c = 0$ gewählt werden und $\mathcal{M}_A$ lässt sich direkt, wie in (4.12) und (4.13) angegeben, bestimmen, ohne dass es der numerischen

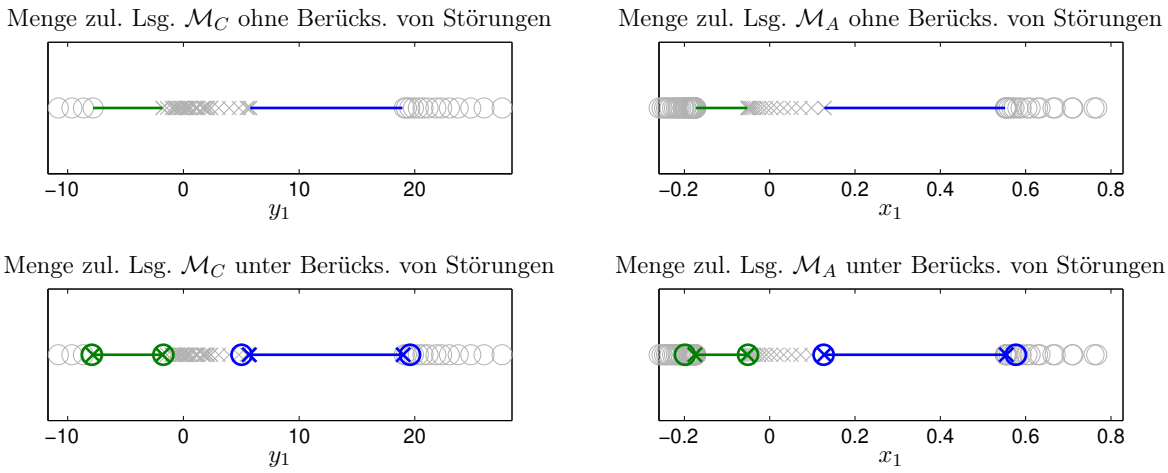


**Abbildung 4.5:** Die Mengen $\mathcal{M}_A$ und $\mathcal{M}_C$ für Datensatz 1 ohne und mit Berücksichtigung von Störungen. Oben: Die durchgezogenen Linien sind die Mengen $\mathcal{M}_A$ und $\mathcal{M}_C$ ohne Berücksichtigung von Störungen (das heißt $\varepsilon_a = \varepsilon_c = 0$). Unten: Zusätzlich zu den durchgezogenen Linie sind die Intervallgrenzen unter Berücksichtigung von Störungen (Intervallgrenzen $\times$ und $\times$ für $\varepsilon_a = \varepsilon_c = 10^{-3}$, Intervallgrenzen $\circ$ und $\circ$ für $\varepsilon_a = \varepsilon_c = 10^{-2}$) eingezeichnet. In den linken Grafiken sind zusätzlich einige der Werte $u(1,j)$, $j = 1,\ldots,n$, (grau $\times$) und einige der Werte $-1/w(1,i)$, $i = 1,\ldots,k$, (grau $\circ$) dargestellt. Analog sind in den rechten Grafiken zusätzlich einige der Werte $w(1,i)$, $i = 1,\ldots,k$, (grau $\times$) und einige der Werte $-1/u(1,j)$, $j = 1,\ldots,n$, (grau $\circ$) dargestellt. Gut zu erkennen: Ohne Berücksichtigung von Störungen sind die Intervallgrenzen der Menge $\mathcal{M}_A$ durch $a = -\min_{j:\,u(1,j)>0} 1/u(1,j)$, $b = \min_{i=1,\ldots,k} w(1,i)$, $c = \max_{i=1,\ldots,k} w(1,i)$ und $d = -\max_{j:\,u(1,j)<0} 1/u(1,j)$, siehe (4.13), bestimmt. Werden betragskleine negative Einträge in $C$ und $A$ zugelassen, so vergrößern sich die Intervalle leicht.

Klassifizierung 4.2 bedarf. Erst um wegen $\varepsilon_a = \varepsilon_c = 0$ ausgeschlossene Lösungen auch mit in $\mathcal{M}_A$ repräsentiert zu wissen, wird eine Anwendung von $\varepsilon_a, \varepsilon_c > 0$ nötig.

Eine Situation mit $U\Sigma(:,1:2)(V(:,1:2))^{\mathbf{T}} \geq 0$ liegt beispielsweise bei Datensatz 1 vor. Für $D$ gilt $\operatorname{rank}(D) = \min(k,n) = 82$, vergleiche auch die Singulärwerte in Abbildung 2.4 (Mitte). Zu $s = 2$ werden die Mengen zulässiger Lösungen bestimmt. Auch bedingt durch die Spektroskopieart (UV/Vis) gilt $\min_{ij} \widetilde{D}_{ij} = 8.9 \cdot 10^{-2}$ mit $\widetilde{D}$ aus (4.14). Somit lässt sich $\mathcal{M}_A$, wie in (4.12) und (4.13) angegeben, bestimmen und insbesondere ist $\mathcal{M}_A \neq \emptyset$. Nichtsdestotrotz lässt sich $\mathcal{M}_A$ auch unter der Berücksichtigung kleiner negativer Einträge bestimmen.

In Abbildung 4.5 sind die Mengen $\mathcal{M}_A$ und $\mathcal{M}_C$ für Datensatz 1 mit und ohne die Berücksichtigung von Störungen dargestellt. Für das Zweikomponentensystem ergeben sich ohne Berücksichtigung von Störungen die Intervallgrenzen wie in (4.13). Unter Berücksichtigung von Störungen $\varepsilon_a = \varepsilon_c = 10^{-3}$ beziehungsweise $\varepsilon_a = \varepsilon_c = 10^{-2}$ vergrößern sich die einzelnen Teilintervalle (Segmente) leicht.

## 4.3 Geometrische Konstruktionen für $s = 3$

Die geometrische Bestimmung des Randes der Menge zulässiger Lösungen $\mathcal{M}_A$ für $s = 3$ ist eine konstruktive Methode, welche auf [17] zurückgeht. Der in [17] vorgestellte Ansatz wurde später unter anderem in [12, 87, 88, 138] aufgegriffen und verfeinert, vergleiche auch [70, 91, 92]. Ausgangspunkt ist Satz 3.26, der unabhängig von $s$ eine Entscheidung, ob $x \in \mathcal{M}_A$ gilt oder nicht, ermöglicht. Basierend darauf wird mittels spezieller Konstruktionen ein Polygonzug von Randpunkten konstruiert. Teile dieses Polygonzugs, oder manchmal auch der gesamte Polygonzug, dienen als eine Diskretisierung der Kurve des inneren Randes von $\mathcal{M}_A$. Die Punkte des Polygonzugs sind entweder außerhalb von $\mathcal{F}_A$ oder sie sind, abgesehen von numerischen Rundungsfehlern, exakt bestimmte innere Randpunkte (siehe Definition 4.8 für innere Randpunkte).

Da sich auch der äußere Rand, ebenfalls bis auf Rundungsfehler, exakt bestimmen lässt, führt dies zu einer hoch genauen Approximation der Menge $\mathcal{M}_A$.

Insgesamt hat die Methode jedoch zwei Nachteile: Aufgrund der Art der Konstruktion ist sie in der Form auf $s = 3$ limitiert und zudem ist die Einbindung von Störungen zwar möglich [86–88, 153], einige entscheidende Nachteile lassen sich aber nur schwer überwinden, siehe Abschnitt 5.3.3.

### 4.3.1 Borgen plots

Als Borgen plots wird die Methode der geometrischen Konstruktion des Randes der Menge zulässiger Lösungen $\mathcal{M}_A$ für $s = 3$ bezeichnet. Die Methode geht auf die fundamentale Arbeit [17] zurück (wenngleich bereits in [120, 142] entscheidende Grundlagen gelegt wurden) und wurde später unter anderem in [12, 85–88, 135, 138] aufgegriffen, analysiert und erweitert. Entscheidend ist die Konstruktion innerer Randpunkte, basierend auf den Zusammenhängen aus Satz 3.26 bezüglich Konvexkombinationen und den Mengen $\mathcal{F}_A$ und $\mathcal{I}_A$. Im Normalfall wird der innere Rand von $\mathcal{M}_A$ nicht vollständig analytisch bestimmt, sondern nur diskretisiert. Der äußere Rand und die Zusammensetzung der Menge $\mathcal{M}_A$ (einzelne Segmente oder zusammenhängende Menge mit Loch) lässt sich leicht mittels $\mathcal{F}_A$ bestimmen.

**Bemerkung 4.7.** *Der innere Rand ließe sich mit der hier vorgestellten Methode, welche aus Veröffentlichungen bekannt ist, zwar analytisch bestimmen [16, 138], dies erfolgt in der Regel aber nicht. Zum einen sind diese Formeln sehr umfangreich und zum anderen sind für $k, n \gg 3$ eine enorme Anzahl an verschiedenen Fällen zu untersuchen, die schlussendlich auf die zusammengesetzte Randkurve führen.*

In diesem Abschnitt wird die Methode der geometrischen Konstruktion (des inneren Randes) der Menge $\mathcal{M}_A$ kurz erläutert und untersucht. Für detailliertere Untersuchungen sei auf die oben genannten Publikationen verwiesen. Weiter wird ein Ansatz vorgestellt, die inneren Ränder der Mengen zulässiger Lösungen $\mathcal{M}_A$ und $\mathcal{M}_C$ simultan zu berechnen [148, 153].

**Definition 4.8.** *Ein $P \in \mathcal{M}_A$ wird als innerer Randpunkt von $\mathcal{M}_A$ bezeichnet, wenn $\alpha P \notin \mathcal{M}_A$ für alle $\alpha$ mit $0 < \alpha < 1$ gilt.*

**Definition 4.9.** *Ein $P \in \mathcal{M}_A$ wird als äußerer Randpunkt von $\mathcal{M}_A$ bezeichnet, wenn $\alpha P \notin \mathcal{M}_A$ für alle $\alpha > 1$ gilt (oder gleichbedeutend $\alpha P \notin \mathcal{F}_A$).*

**Bemerkung 4.10.** *Gemäß den Definitionen 4.8 und 4.9 kann ein $x \in \mathcal{M}_A$ gleichzeitig sowohl innerer als auch äußerer Randpunkt sein.*

Zur Bestimmung des inneren Randes wird ein Polygonzug konstruiert. Dabei sind alle Punkte des Polygonzugs, die in $\mathcal{F}_A$ liegen, innere Randpunkte. Es können aber auch Punkte außerhalb von $\mathcal{F}_A$ liegen, welche gewissermaßen zu viel berechnet wurden. Somit enthält der bestimmte Polygonzug eine Approximation der Kurve des inneren Randes oder ist direkt eine Approximation der Kurve des inneren Randes. Der Rand von $\mathcal{M}_A$ ergibt sich entweder aus Teilen des Randes von $\mathcal{F}_A$ und Teilen des berechneten Polygonzugs oder der Rand von $\mathcal{F}_A$ ist der äußere Rand von $\mathcal{M}_A$ und der berechnete Polygonzug ist der innere Rand von $\mathcal{M}_A$.

#### Geometrische Konstruktion des Randes: der Tangentenalgorithmus

Die Punkte des Polygonzugs zur Bestimmung des inneren Randes können etwa mittels des Tangentenalgorithmus [17, 138] berechnet werden. Die Idee ist es dabei, die Diskretisierungspunkte

mittels einer Serie von Tangenten an $\mathcal{I}_A$ zu konstruieren. Ein anderer Ansatz zur Konstruktion des Randes von $\mathcal{M}_A$ ist es, den Rand von $\mathcal{F}_A$ zu durchlaufen und die Punkte dort als Ausgangsbasis für die Konstruktion innerer Randpunkte zu nutzen.

Im Folgenden wird der Tangentenalgorithmus vorgestellt. Sei $h_i$ eine Tangente an $\mathcal{I}_A$. Zu dieser wird wie folgt ein Punkt des Polygonzugs zur Approximation des inneren Randes bestimmt:

**Iteration 4.11** (Geometrische Konstruktion)**.**

1. *Die Schnittpunkte $S_1$ und $S_2$ von $h_i$ mit $\mathcal{F}_A$ werden bestimmt.*

   *Sollte $h_i$ mit einer Seite von $\mathcal{F}_A$ übereinstimmen, so werden die beiden an dieser Seite angrenzenden Eckpunkte von $\mathcal{F}_A$ als $S_1$ und $S_2$ genommen. Sollte $h_i$ nur einen Punkt mit $\mathcal{F}_A$ gemeinsame haben, so ist dieser (vorausgesetzt $\mathcal{M}_A \neq \emptyset$) ein innerer (und auch äußerer) isolierter Randpunkt (Punktsegment). In diesem Fall wird für dieses $h_i$ die Rechnung gestoppt. Siehe hierzu auch Bemerkung 4.12.*

2. *Ausgehend von $S_1$ wird die Gerade $g_1$ bestimmt, die an $\mathcal{I}_A$ anliegt, jedoch nicht mit $h_i$ übereinstimmt. Analog wird eine Gerade $g_2$ durch $S_2$ bestimmt, die an $\mathcal{I}_A$ anliegt, jedoch nicht mit $h_i$ übereinstimmt.*

3. *Der Schnittpunkt $P_i$ der Geraden $g_1$ und $g_2$ wird bestimmt. Falls das Dreieck mit den Eckpunkten $S_1$, $S_2$ und $P_i$ nicht die Menge $\mathcal{I}_A$ enthält, so wird $P_i$ verworfen. Sofern das Dreieck die Menge $\mathcal{I}_A$ einschließt, wird $P_i$ nicht verworfen. Falls weiter $P_i \in \mathcal{F}_A$ gilt, so ist $P_i$ ein innerer Randpunkt von $\mathcal{M}_A$, siehe Lemma 4.13.*

Die nicht verworfenen $P_i$, $i = 1, \ldots, m$, definieren bei monotonem Umlauf der Tangenten $h_i$ ein Polygon $\mathcal{P}$. Mittels dieses wird die Ebene in zwei Mengen unterteilt. Von diesen sei $\mathcal{N}$ die Menge, die den Ursprung nicht enthält (vergleiche Satz 3.10) und $\mathcal{P}$ selbst gehöre zu $\mathcal{N}$. Dann ist der Rand von $\mathcal{F}_A \cap \mathcal{N}$ eine Diskretisierung des Randes von $\mathcal{M}_A$ und es gilt $\mathcal{F}_A \cap \mathcal{N} \approx \mathcal{M}_A$. Die Feinheit der Diskretisierung des inneren Randes von $\mathcal{F}_A \cap \mathcal{N}$ hängt von der Anzahl der genutzten Tangenten ab. In Abbildung 4.6 ist diese Vorgehensweise illustriert.

Die Tangenten $h_i$ sollten mit einer ausreichend feinen Diskretisierung gewählt werden. Ein Spezialfall liegt vor, wenn $\mathcal{M}_A$ ein Punktsegment (isolierte Lösung) enthält:

**Bemerkung 4.12.** *Für eine endliche Auswahl von Tangenten werden Punktsegmente im Normalfall nicht detektiert. Punktsegmente von $\mathcal{M}_A$ können trivialerweise nur auf dem Rand von $\mathcal{F}_A$ liegen, vergleiche Satz 3.15. Um Punktsegmente nicht zu übergehen, sind zwei Dinge zu beachten:*

1. *Zu allen Kanten von $\mathcal{I}_A$ sind die anliegenden Tangenten ebenfalls für die Konstruktion innerer Randpunkte zu nutzen. Weiter ist zu untersuchen, ob $S_1$ und/oder $S_2$ innere(r) Randpunkt(e) sind/ist.*

2. *Zu allen Ecken von $\mathcal{F}_A$ sind jeweils die zwei Tangenten an $\mathcal{I}_A$ zu nutzen. Auch hier gilt es, zusätzlich die sich daraus ergebenden Schnittpunkte $S_1$ und $S_2$ (einer von beiden ist ein Eckpunkt von $\mathcal{F}_A$) zu untersuchen, ob sie innere Randpunkte sind.*

*Vergleiche diesbezüglich auch Beispiel 3.74 für $x = \sqrt{2}/2$ und Abbildung 3.2 (rechts).*

### Nachweis der Randlage der konstruierten Punkte

In Algorithmus 4.11 wird zu einer Tangente $h = h_i$ ein $P = P_i$ konstruiert. Von diesem wird behauptet, es sei ein innerer Randpunkt, sofern $P \in \mathcal{M}_A$ gilt. Im folgenden Lemma wird dies nachgewiesen.

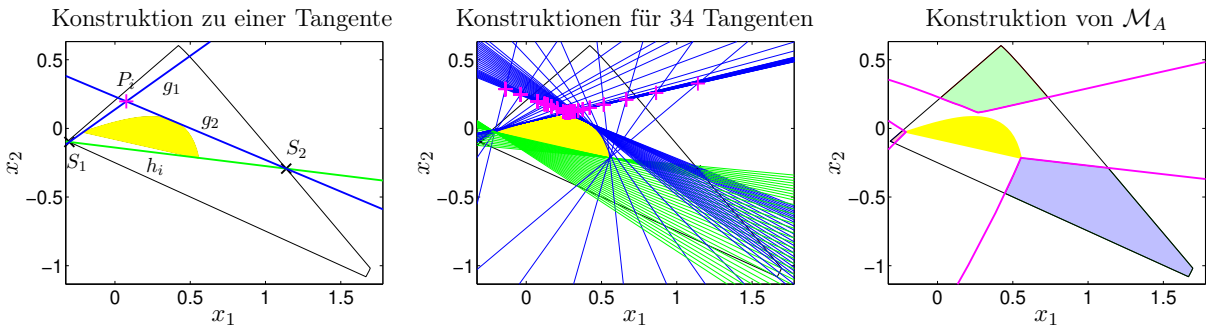


**Abbildung 4.6:** Die Geometrische Konstruktion des Randes von $\mathcal{M}_A$ am Beispiel des Datensatzes 2. Links: Dargestellt ist die geometrische Konstruktion für eine Tangente $h_i$ (grün), welche an $\mathcal{I}_A$ (gelb) anliegt. Zu $h_i$ werden die beiden Schnittpunkte $S_1$ und $S_2$ (×) von $h_i$ und $\partial\mathcal{F}_A$ (schwarz) bestimmt. Ausgehend von diesen, werden die Geraden $g_1$ und $g_2$ (blau) an $\mathcal{I}_A$ „rangeklappt“. Der Schnitt dieser Geraden ist $P_i$ (+). Es liegt $P_i$ in $\mathcal{F}_A$ und ist damit ein innerer Randpunkt von $\mathcal{M}_A$. Mitte: Dargestellt sind die Konstruktionen für 34 Tangenten. Rechts: Die Verbindung der jeweils berechneten $P_i$ ergibt ein Polygon $\mathcal{P}$, welches zusammen mit $\mathcal{F}_A$ zur Konstruktion des Randes von $\mathcal{M}_A$ (hier drei Segmente) genutzt wird.

**Lemma 4.13.** *Sei $s = 3$ und seien weiter $h = h_i$, $g_1$ und $g_2$ paarweise verschiedene Geraden die $\mathcal{I}_A$ berühren. Der Schnittpunkt $S_1$ von $h$ mit $g_1$ liege ebenso auf dem Rand von $\mathcal{F}_A$ wie der Schnittpunkt $S_2$ von $h$ und $g_2$.*

*Liegt der Schnittpunkt $P = P_i$ von $g_1$ und $g_2$ in $\mathcal{F}_A$, so ist er ein innerer Randpunkt von $\mathcal{M}_A$.*

*Beweis.* Sei angenommen, dass $P \in \mathcal{F}_A$ kein innerer Randpunkt sei. Dann gäbe es ein positives $\alpha < 1$ mit $\alpha P \in \mathcal{M}_A$. Nach Satz 3.26 müsste es zu $\alpha P$ zwei Punkte in $\mathcal{F}_A$ geben, sodass das daraus konstruierte Dreieck $\mathcal{I}_A$ einschließt. Nun wird gezeigt, dass es keine zwei solche Punkte geben kann. Seien dazu $g_3$ und $g_4$ die von $\alpha P$ aus an $\mathcal{I}_A$ „rangeklappten“ Geraden. Sofern es überhaupt zwei Punkte gäbe, die zu einer zulässigen Lösungen führen würden, so täten dies die Schnittpunkte $S_3$ und $S_4$ dieser Geraden mit $\mathcal{F}_A$ (bei sinnvoller Wahl, da es jeweils zwei Schnittpunkte gibt). Es liegen jedoch $S_3$ und $S_4$ auf der gleichen Seite von $h$ wie $\alpha P$ und insbesondere nicht auf $h$. Da jedoch bereits $h$ an $\mathcal{I}_A$ anliegt, kann das durch $\alpha P$, $S_3$ und $S_4$ erzeugte Dreieck nicht die gesamte Menge $\mathcal{I}_A$ enthalten. Somit kann $\alpha P$ nicht zulässig sein und folglich ist $P$ ein innerer Randpunkt von $\mathcal{M}_A$. □

### 4.3.2 Simultane Berechnung beider Mengen zulässiger Lösungen

Die Berechnung des inneren Randes von $\mathcal{M}_A$ mittels der geometrischen Konstruktion eröffnet eine Möglichkeit, ohne großen Mehraufwand den inneren Rand von $\mathcal{M}_C$ mit zu bestimmen. In [148, 153] wird diese Methode als *dual Borgen plots* eingeführt. Die Idee ist es, bei der geometrischen Konstruktion zu einer Tangente $h_i$ zwei weitere Geraden so zu bestimmen, dass deren duale Punkte innere Randpunkte von $\mathcal{M}_C$ sind. Somit ergeben die Konstruktionen zu einer Tangente einen (sofern dieser in $\mathcal{F}_A$ liegt) inneren Randpunkt von $\mathcal{M}_A$ sowie zwei (sofern diese in $\mathcal{F}_C$ liegen) innere Randpunkte von $\mathcal{M}_C$. Entscheidend ist dazu der Satz 4.16 zu dessen Beweis die Lemmata 4.13, 4.14 und 4.15 genutzt werden.

**Lemma 4.14.** *Seien $D \in \mathbb{R}^{k\times n}$ eine nichtnegative Matrix, welche keine Nullspalte und keine Nullzeile enthält, und $s = \operatorname{rank}(D) = \operatorname{rank}_+(D) \geq 2$. Seien $DD^{\mathbf{T}}$ irreduzibel und $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$ mit $V(:,1) > 0$. Weiter liege $x_0$ auf dem Rand von $\mathcal{F}_A$ und zwar an der affinen Hyperebene $E^{(A)}_{j_0}$ mit $j_0 \in \{1,\dots,n\}$, sodass also laut Voraussetzungen*

$$x_0^{\mathbf{T}} V(j_0, 2:s)^{\mathbf{T}} = -V_{j_0 1}, \tag{4.15}$$

$$x_0^{\mathbf{T}} V(j, 2:s)^{\mathbf{T}} \geq -V_{j1}, \quad j = 1,\dots,n. \tag{4.16}$$

*Es liegt die zu $x_0$ duale affine Hyperebene $E$ (siehe Definition 3.47) an $\mathcal{I}_C$ an und zwar an $u(:, j_0)$.*

*Beweis.* Nach Definition 3.47 und Lemma 3.50 ergibt sich die zu $x_0$ duale affine Hyperebene als $E = \{z \in \mathbb{R}^{s-1} : x_E^{\mathbf{T}} z = \delta\}$ mit $\delta = -1$ und $x_E = x_0$. Einsetzen von $z = u(:, j_0)$ in die Ebene liefert wegen (4.15), dass $x_E^{\mathbf{T}} z = x_0^{\mathbf{T}} u(:, j_0) = -1$. Wegen (4.16) liefert das Einsetzen aller $u(:, j)$, $j = 1, \ldots, n$, (ob $j \neq j_0$ oder nicht spielt keine Rolle), dass $x_E^{\mathbf{T}} z = x_0^{\mathbf{T}} u(:, j) \geq -1$. Damit ist nachgewiesen, dass $u(:, j_0)$ in der Ebene $E$ enthalten ist und $E$ an $\mathcal{I}_C$ anliegt. □

Weiterhin gilt auch die Umkehrung von Lemma 4.14, was in dieser Arbeit aber nicht extra gezeigt wird. Stattdessen wird im folgenden Lemma die andere Richtung für affine Hyperebenen an $\mathcal{I}_A$ und Randpunkte von $\mathcal{F}_C$ gezeigt.

**Lemma 4.15.** *Seien $D \in \mathbb{R}^{k \times n}$ eine nichtnegative Matrix, welche keine Nullspalte und keine Nullzeile enthält, und $s = \operatorname{rank}(D) = \operatorname{rank}_+(D) \geq 2$. Seien $DD^{\mathbf{T}}$ irreduzibel und $U\Sigma V^{\mathbf{T}}$ eine abgeschnittene Singulärwertzerlegung von $D$ mit $V(:, 1) > 0$. Sei $E = \{z \in \mathbb{R}^{s-1} : x_E^{\mathbf{T}} z = -1\}$ eine affine Hyperebene, die an $\mathcal{I}_A$ anliegt. Konkret liege $E$ an $w(:, i_0)$ mit $i_0 \in \{1, \ldots, k\}$ an.[4] Wegen $(0, \ldots, 0)^{\mathbf{T}} \in \mathcal{I}_A$ gelten also laut Voraussetzungen*

$$x_E^{\mathbf{T}} w(:, i_0) = -1, \tag{4.17}$$
$$x_E^{\mathbf{T}} w(:, i) \geq -1, \quad j = 1, \ldots, n. \tag{4.18}$$

*Das zu $E$ duale $y_0 = x_E$ liegt auf dem Rand von $\mathcal{F}_C$ und zwar in der affinen Hyperebene $E_{i_0}^{(C)}$. Somit gelten*

$$U(i_0, :)\Sigma(:, 2 : s)y_0 = -\sigma_1 U_{i_0 1}, \tag{4.19}$$
$$U(i, :)\Sigma(:, 2 : s)y_0 \geq -\sigma_1 U_{i1}, \quad j = 1, \ldots, k. \tag{4.20}$$

*Beweis.* Einfaches Einsetzen von $w(:, i)$ und $y_0 = x_E$ in (4.17) und (4.18) sowie Transponieren führt auf (4.19) und (4.20). Somit liegt $y_0$ auf dem Rand von $\mathcal{F}_C$ und zwar konkret in der affinen Hyperebene, die sich aus der Bedingung bezüglich $i_0$ ergibt. □

**Satz 4.16.** *Sei $s = 3$ und seien drei Punkte $P, Q, R$ auf dem Rand von $\mathcal{F}_A$ gegeben, welche gemäß Satz 3.23 ein zulässiges Dreieck $\Delta$ bilden.[5] Weiter sei angenommen, dass zwei der drei Seiten von $\Delta$ an $\mathcal{I}_A$ anliegen. Ohne Beschränkung der Allgemeinheit seien es die, welche durch den Punkt $P$ verlaufen. Sei $\Delta'$ das zu $\Delta$ duale Dreieck in der niedrigdimensionalen Darstellung des Faktors $C$, das heißt, dass die Seiten von $\Delta'$ dual zu den Punkten $P, Q$ und $R$ sind.*

*Bezüglich der niedrigdimensionalen Darstellung von $C$ erfüllt $\Delta'$ die Voraussetzungen von Lemma 4.13 und der zur Geraden durch $Q$ und $R$ duale Punkt ist ein innerer Randpunkt von $\mathcal{M}_C$.*

*Beweis.* In Abschnitt 3.6 sind die Zusammenhänge zwischen Punkten in der niedrigdimensionalen Darstellung für $A$ und affinen Hyperebenen in der niedrigdimensionalen Darstellung für $C$ und umgekehrt untersucht. Nach Lemma 4.14 gilt, dass die, zu den Punkten auf dem Rand von $\mathcal{F}_A$, dualen affinen Hyperebenen (im Fall $s = 3$ also Geraden) an $\mathcal{I}_C$ anliegen. Da sich $P$, $Q$ und $R$ auf dem Rand von $\mathcal{F}_A$ befinden, liegen die dazu dualen Geraden an $\mathcal{I}_C$ an. Weiter liegen nach Voraussetzung zwei der drei Seiten von $\Delta$ an $\mathcal{I}_A$ an und somit befinden sich nach Lemma 4.15 zwei der drei Eckpunkte von $\Delta'$ auf dem Rand von $\mathcal{F}_C$. Trivialerweise enthält das Dreieck $\Delta'$ die Menge $\mathcal{I}_C$, vergleiche die Idee zu Satz 3.23. Somit sind die Voraussetzungen von Lemma 4.13 erfüllt und der zur Geraden durch $Q$ und $R$ duale Punkt ist ein innerer Randpunkt von $\mathcal{M}_C$. □

[4] Der Durchschnitt von $E$ und $\mathcal{I}_A$ kann auch weitere Elemente enthalten.
[5] Der Punkt $P$ ist hier beliebig und hat nichts mit den $P_i$ aus Iteration 4.11 zu tun.

Anknüpfend an die drei Schritte des Tangentenalgorithmus zur geometrischen Konstruktion des Randes von $\mathcal{M}_A$ (Iteration 4.11) ergeben sich zusätzlich folgende Schritte zur simultanen Bestimmung des Randes von $\mathcal{M}_C$:

**Iteration 4.17** (Simultane geometrische Konstruktion, aufbauend auf Iteration 4.11)**.**

4. *Zu den Geraden $g_1$ und $g_2$ werden jeweils die (neben $S_1$ beziehungsweise $S_2$) anderen Schnittpunkte mit $\mathcal{F}_A$ bestimmt ($S_3$ und $S_4$).*

   *Sollten $g_1$ oder $g_2$ Kanten von $\mathcal{F}_A$ sein, so wird der (neben $S_1$ beziehungsweise $S_2$) andere Eckpunkt zu der Kante als $S_3$ beziehungsweise $S_4$ genommen.*

5. *Sofern $S_1 \neq S_4$ beziehungsweise $S_2 \neq S_3$ werden die Geraden durch die Punkte $S_1$ und $S_4$ beziehungsweise $S_2$ und $S_3$ bestimmt und mit $g_3$ und $g_4$ bezeichnet.*

6. *Zu $g_3$ und $g_4$ werden die dualen Punkte $Q_i^{(1)}$ und $Q_i^{(2)}$ berechnet.*

   *Mit Rücksicht auf die unten folgende Bemerkung 4.18 (Punkt 2) erfolgt im Fall, dass der Nullpunkt nicht in dem Dreieck mit den Ecken $S_1$, $S_2$ und $S_4$ liegt, ein Vorzeichenwechsel für $Q_i^{(1)}$. Analog erfolgt ein Vorzeichenwechsel für $Q_i^{(2)}$, falls der Nullpunkt nicht in dem von $S_1$, $S_2$ und $S_3$ aufgespannten Dreieck liegt.*

Analog zur Berechnung des Randes von $\mathcal{M}_A$ wird aus allen Punkten $Q_i^{(1)}$ und $Q_i^{(2)}$ ein Polygon $\mathcal{Q}$ gebildet. Dazu ist es nötig, diese zu ordnen. Eine sinnvolle Ordnung nach Winkeln lässt sich etwa mittels deren Polarkoordinaten herstellen. Dies ist möglich und sinnvoll, da nach Satz 3.10 je Richtung nur ein innerer Randpunkt existiert. Dass dies für Punkte $Q_i^{(1)}$ oder $Q_i^{(2)}$, die nicht in $\mathcal{F}_C$ liegen, auch gilt, ist zwar nicht bewiesen, aber für die Bestimmung von $\mathcal{M}_C$ auch nicht relevant.

Die Idee der simultanen Konstruktion eines inneren Randpunktes von $\mathcal{M}_A$ und zweier innerer Randpunkte von $\mathcal{M}_C$ ist in Abbildung 4.7 dargestellt.

**Bemerkung 4.18.**

1. *Bei der Iteration 4.11 ergibt sich pro Tangente ein $P_i$, welches möglicherweise ein innerer Randpunkt von $\mathcal{M}_A$ ist. Im Gegensatz dazu ergeben sich pro Tangente zwei potentielle innere Randpunkte von $\mathcal{M}_C$, da es zwei Geraden ($g_3$ und $g_4$) gibt, die auf $Q_i^{(1)}$ und $Q_i^{(2)}$ führen. Somit ist die Auflösung des inneren Randes von $\mathcal{M}_C$ etwa doppelt so hoch wie die von $\mathcal{M}_A$.*

2. *In Satz 4.16 ist der Fall, dass das Dreieck $\Delta$ mit den Eckpunkten $P$, $Q$, und $R$ nicht den Nullpunkt enthält, ausgeschlossen, da $\Delta$ als zulässig vorausgesetzt ist. Unter Umständen wird die Rechnung jedoch andernfalls trotzdem fortgesetzt, beispielsweise um zusammenhängende Kurven für das Polygon aller Punkte $Q_i^{(1)}$ und $Q_i^{(2)}$ auch außerhalb von $\mathcal{F}_C$ zu erhalten (siehe beispielsweise die rechte Grafik in Abbildung 4.7). Für den Fall, dass das Dreieck mit den Ecken $P$, $Q$, und $R$ nicht den Nullpunkt enthält, ist ein Vorzeichenwechsel für den zu berechnenden dualen Punkt $Q_i^{(1)}$ beziehungsweise $Q_i^{(2)}$ vorzunehmen. Dies ist dadurch begründet, da sich ansonsten eine Rekonstruktion als $c = U\Sigma y$ mit $y_1 = -1$ ergibt. Sofern $DD^{\mathrm{T}}$ irreduzibel ist, folgt nach Korollar 3.4 aber sicher $c \ngeq 0$ und mitunter sogar $c \leq 0$. Der Vorzeichenwechsel behebt diesen Defekt. Nach Satz 3.31 sind solche Dreiecke zwar für die Bestimmung von $\mathcal{M}_A$ beziehungsweise $\mathcal{M}_C$ nicht von Interesse, so (also ohne Vorzeichenwechsel) bestimmte Punkte machen aber in der Regel eine sinnvolle Ordnung der Punkte $Q_i^{(1)}$ und $Q_i^{(2)}$, $i = 1, \ldots, m$, unmöglich.*

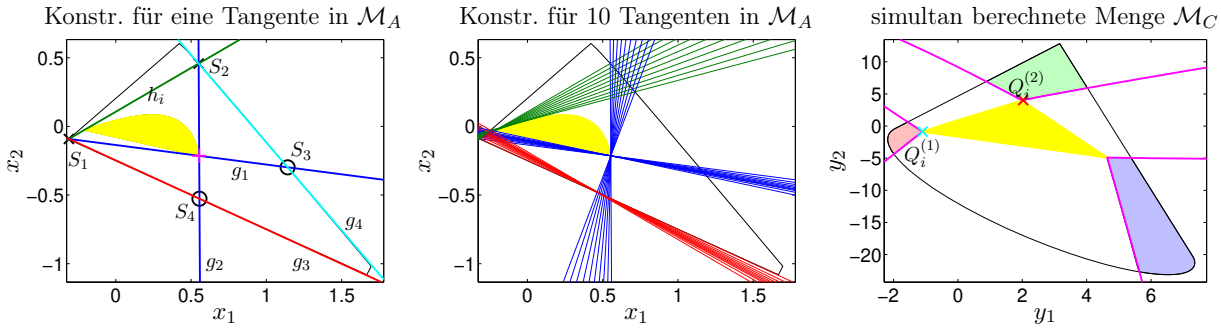


**Abbildung 4.7:** Simultane Konstruktion innerer Randpunkte für $\mathcal{M}_A$ und $\mathcal{M}_C$ am Beispiel des Datensatzes 2. Links: Betrachtet wird die Konstruktion für eine Tangente $h_i$ (grün). Dazu ergeben sich die beiden Schnittpunkte $S_1$ und $S_2$ (×) mit $\mathcal{F}_A$ (schwarz), die an $\mathcal{I}_A$ „rangeklappten" Tangenten $g_1$ und $g_2$ (blau) sowie der innere Randpunkt für $\mathcal{M}_A$ (+) wie in Abbildung 4.6. Die Punkte $S_4$ und $S_3$ (∘) werden zusammen mit $S_1$ und $S_2$ zur Konstruktion der Geraden $g_3$ (rot) und $g_4$ (türkis) genutzt. Zu diesen werden die dualen Punkte $Q_i^{(1)}$ und $Q_i^{(2)}$ (in der rechten Grafik) berechnet. Sofern sie in $\mathcal{F}_C$ liegen, sind es innere Randpunkte von $\mathcal{M}_C$ (× und ×). Mitte: Gezeigt sind die Konstruktionen für zehn Tangenten. Rechts: Dargestellt ist die, zur Berechnung von $\mathcal{M}_A$ simultan durchgeführte, Berechnung des inneren Randes von $\mathcal{M}_C$ (magenta). Zusätzlich sind $\mathcal{F}_C$ (schwarz) und $\mathcal{I}_C$ (gelb) sowie die beiden, bezüglich der linken Grafik berechneten, Punkte (×) und (×) dargestellt.

### 4.3.3 Erweiterungen für gestörte Daten und höhere Dimensionen

Die geometrische Konstruktion der inneren Randpunkte für $s = 3$ hat Vor- und Nachteile. Einerseits ist, anders als bei allen numerischen Methoden, die Bestimmung der inneren Randpunkte für Modelldaten bis auf Rundungsfehler exakt. Zudem ist der Rechenaufwand, natürlich abhängig von der Auflösung, im Normalfall nicht höher als bei numerischen Verfahren, sondern im Gegenteil oft deutlich geringer. Andererseits ist die geometrische Konstruktion auf $s = 3$ begrenzt. Außerdem ist die Methode zwar für gestörte Daten erweiterbar, siehe [86–88], jedoch weist dieser Algorithmus ansatzbedingt nicht die Flexibilität und die Robustheit numerischer Methoden auf. Ein Ausweg ist in [153] aufgezeigt.

Die Schwierigkeit bei der Anwendung auf störungsbehaftete Daten ist, dass auch betragskleine negative Einträge akzeptiert werden müssen. Bei der in Abschnitt 4.1 vorgestellten numerischen Methode zur Klassifizierung eines $x \in \mathbb{R}^{s-1}$ kann dies relativ erfolgen. Bei der etwa aus [1,56,58, 173] bekannten ssq-Funktion ist die Berücksichtigung von Störungen in Form von *Abschneiden* und somit anders umgesetzt. Eine relative Bewertung wie in (4.1) ist bei dem geometrischen Ansatz nicht möglich und es können nur absolute Fehler berücksichtigt werden. Dies führt zu einigen Schwierigkeiten, da die $w(:, i)$, $i = 1, \ldots, k$, für $\mathcal{I}_A$ zu berücksichtigen und mitunter sehr störanfällig sind, siehe Abschnitt 3.8 sowie die Abbildungen 5.6, 5.9 und 5.13.

Das Problem für die Erweiterung auf $s = 4$ ist im Vergleich zu $s = 3$ wie folgt gelagert: Für $s = 3$ ergeben sich zu einer Tangente $h_i$ die Geraden $g_1$ und $g_2$ und damit auch der Punkt $P_i$ eindeutig. Für $s = 4$, also $\mathcal{M}_A, \mathcal{M}_C \in \mathbb{R}^3$, gibt es jedoch im Allgemeinen zu einer an $\mathcal{I}_A$ anliegenden Fläche keine eindeutig bestimmte Fortsetzung mittels dreier weiterer an $\mathcal{I}_A$ anliegender Flächen, sodass das daraus gebildete Tetraeder die Menge $\mathcal{I}_A$ einschließt. In [86] ist eine hybride Herangehensweise mittels geometrischer Konstruktion und numerischer Optimierung zur Bestimmung von $\mathcal{M}_A$ aufgezeigt. Der Ansatz basiert auf der Strahleneigenschaft aus Satz 3.15.

## 4.4 Randeinschließung mittels Dreieckskonstruktionen für $s = 3$

In [56] wird eine Methode zur Einschließung des Randes von $\mathcal{M}_A$ mittels einer Serie gleichseitiger Dreiecke für $s = 3$ vorgestellt, siehe auch [55,147]. Die Dreiecke werden sukzessive berechnet und zwei benachbarte Dreiecke haben stets eine gemeinsame Seite. Von dieser gemeinsamen Seite liegt

stets ein Punkt in $\mathcal{M}_A$ und einer nicht. Ausgehend von einem Startdreieck wird die Iteration solange fortgeführt, bis dieses wieder erreicht wird und so eine geschlossene Serie von Dreiecken bestimmt ist. Wegen der Beschränktheit von $\mathcal{M}_A$ ist diese Vorgehensweise endlich. Die Methode arbeitet auch für störungsbehaftete Daten stabil.

### 4.4.1 Arbeitsweise der Methode

Ausgangspunkt ist eine nichtnegative Faktorisierung von $D$. Durch diese werden drei Elemente aus $\mathcal{M}_A$ generiert. Von diesen ausgehend, werden die zugehörigen Segmente berechnet. Besteht $\mathcal{M}_A$ aus drei separierten Segmenten, so führt dies auf eine Approximation von $\mathcal{M}_A$. Besteht $\mathcal{M}_A$ aus einem Segment mit Loch, so werden der äußere und der innere Rand separat bestimmt. Dafür wird nur eine der drei initialen Elemente von $\mathcal{M}_A$ benötigt. In beiden Fällen ist die Vorgehensweise gleich, weshalb die Randbestimmung nur für den Fall einer Menge $\mathcal{M}_A$ mit drei separierten Segmenten vorgestellt wird.

Für die Approximation eines Segments wird zunächst ein Startdreieck bestimmt, welches den Rand von $\mathcal{M}_A$ überdeckt. Von einer initialen zulässigen Lösungen ausgehend, wird dazu ein gleichseitiges Dreieck bestimmt, welches eine Ecke in $\mathcal{M}_A$ und eine Ecke außerhalb von $\mathcal{M}_A$ hat. Die dritte Ecke ist beliebig. Dieses Dreieck ist das Startdreieck für die eigentliche Iteration.

Zur Bestimmung des Startdreiecks wird zunächst eine der initialen Lösungen um zwei andere Punkte so ergänzt, dass durch die drei Punkte ein gleichseitiges Dreieck zu vorgegebener Seitenlänge gebildet wird. Erfüllt dieses Dreieck $\Delta$ nicht die Bedingungen an ein Startdreieck (folglich gehören alle Ecken von $\Delta$ zu $\mathcal{M}_A$), so wird eine der drei Seiten von $\Delta$ genutzt und durch Hinzufügen eines neuen Punktes ein zweites gleichseitiges Dreieck $\Delta'$ mit $\Delta \neq \Delta'$ gebildet. Erfüllt dieses ebenfalls nicht die Bedingungen an ein Startdreieck (das heißt, der neu bestimmte Punkte gehört auch zu $\mathcal{M}_A$), so wird die Bildung neuer gleichseitiger Dreiecke solange fortgesetzt, bis eine Ecke nicht in $\mathcal{M}_A$ liegt. Die Iteration wird stets in dieselbe Stoßrichtung fortgeführt.

Die eigentliche Dreieckseinschließung beginnt mit dem zuletzt berechneten Dreieck. Deren Ecken seien $x, y, z$ mit (ohne Beschränkung der Allgemeinheit) $x \notin \mathcal{M}_A$ und $y \in \mathcal{M}_A$. Ausgehend von der Seite zwischen $x$ und $y$ wird der neue Punkt $z'$ so konstruiert, dass $z' \neq z$ und $x$, $y$ sowie $z'$ ein gleichseitiges Dreieck bilden. Gilt $z' \in \mathcal{M}_A$, so wird die Rechnung anschließend mit $x$ und $z'$ fortgeführt, andernfalls mit $y$ und $z'$. Die Iteration ist beendet, sofern das Anfangsdreieck wieder erreicht wird. Das Polygon aus den jeweiligen zulässigen Dreiecksecken ergibt die Approximation für den Rand des Segments von $\mathcal{M}_A$.

Für eine Menge $\mathcal{M}_A$ mit drei isolierten Segmenten führt die Anwendung der Methode für alle drei Anfangspunkte (von der initialen nichtnegativen Faktorisierung) auf $\mathcal{M}_A$. Für eine zusammenhängende Menge zulässiger Lösungen muss die Randapproximation einmal für den inneren und einmal für den äußeren Rand durchgeführt werden. Eine Menge $\mathcal{M}_A$ mit mehr als drei Segmenten kann mit der Dreieckseinschließungsmethode nicht vollständig approximiert werden.

**Bemerkung 4.19.** *In [56] ist die Methode der Dreieckseinschließung vorgestellt. Die Entscheidung, ob ein $x$ zu $\mathcal{M}_A$ gehört oder nicht, wird unter Nutzung der Funktion* ssq *aus (4.6) getroffen. Analog kann die Entscheidung jedoch auch mit den Funktionen $f(x, S)$ aus (4.2) und $F(x)$ aus (4.4) erfolgen. Um für den späteren Vergleich diese Stelle als Quelle für unterschiedliche Ergebnisse auszuschließen, wird für die numerischen Methoden stets die Kombination aus $f(x, S)$ und $F(x)$ genutzt.*

In Abbildung 4.8 ist die Anwendung der Methode für Modelldatensatz 2 mit der (recht großen) Seitenlänge $a = 0.03$ dargestellt.

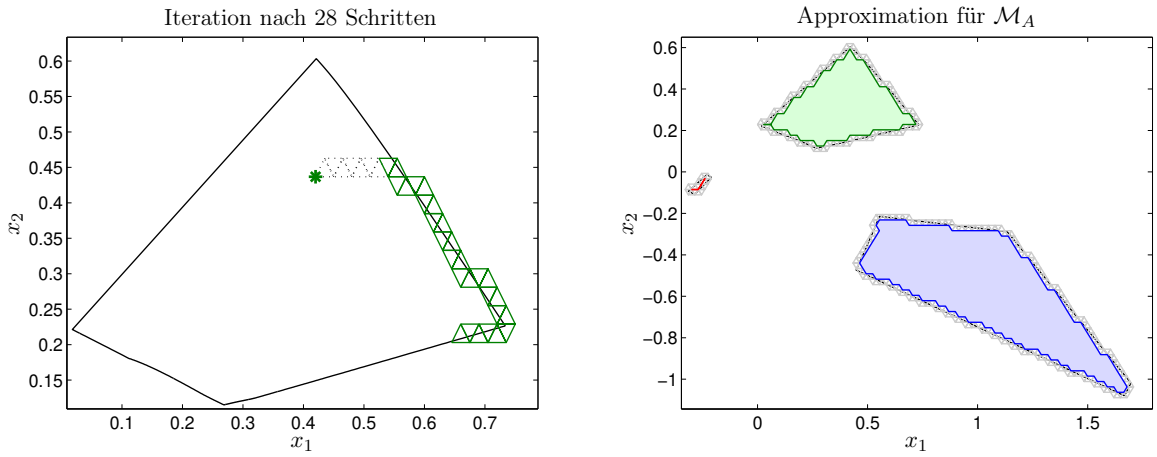


**Abbildung 4.8:** Approximation des Randes der Menge zulässiger Lösungen $\mathcal{M}_A$ mittels Dreieckseinschließungen am Beispiel des Datensatzes 2. Als Seitenlänge ist $a = 0.03$ gewählt. Links: Dargestellt sind die Iterationen zur Berechnung eines Startdreiecks (schwarz gepunktete Linien, ausgehend von einer zulässigen Lösung (✳)) sowie der Fortschritt für ein Segment nach 28 Iterationsschritten (grüne Dreiecke). Rechts: Nach Abschluss der Berechnung werden die zulässigen Punkte zu einem Polygon verbunden. Dies ist die Randapproximation. Der eigentliche Rand (geometrisch konstruiert) ist schwarz eingezeichnet.

### 4.4.2 Genauigkeit der Approximation und fehlende Adaptivität

Die Genauigkeit der Randapproximation wird über die Seitenlänge der Dreiecke gesteuert. Unter der Annahme, dass im Zuge der Randapproximation die Klassifizierungen aller Punkte korrekt ausgeführt wurden, gilt für ein berechnetes $x$ (Randapproximation)

$$\min_{x' \notin \mathcal{M}_A} \|x - x'\|_2 < a. \tag{4.21}$$

Der Nachteil der Randeinschließung mittels Dreiecksketten ist die fehlende Adaptivität und der damit verbundene hohe Rechenaufwand. Für Geradenabschnitte und Abschnitte mit geringer Krümmung werden, ebenso wie auch für Abschnitte mit vergleichsweise hoher Krümmung, eine Vielzahl von Dreiecken bestimmt. Konstruktionsbedingt besteht der äußere Rand aus Geradenabschnitten. Die Approximationen dieser durch viele Dreiecke ist, im Vergleich zur Approximation mit nur wenigen aber gezielt gewählten Punkten, recht aufwendig.

Außerdem ist der Abstand (4.21) zum Rand relativ hoch. So können beispielsweise bei der in Abschnitt 4.5 vorgestellten Methode die Punkte des bestimmten Polygonzugs mittels Bisektionsverfahrens deutlich dichter am Rand gewählt werden. Die euklidischen Abstände der einzelnen Punkte zum Rand belaufen sich ohne hohen Rechenaufwand auf Werte um $10^{-3}$ und weniger.

**Bemerkung 4.20.** *Die Seitenlänge a entspricht prinzipiell dem erwartbaren maximalen Abstand der Approximation des Randes von $\mathcal{M}_A$ zum eigentlichen Rand von $\mathcal{M}_A$. Unter Umständen liegt der Fehler aber auch darüber. Dazu sei folgende Situation angenommen: Es sei eine Ecke mit einem Winkel von etwa $\frac{\pi}{3}$ oder weniger von $\mathcal{M}_A$ zu approximieren. Weiter liege das letzte Dreieck, welches entlang einer Seite zum Winkel hin konstruiert wurde, so, dass es zu einem großen Teil $\mathcal{M}_A$ überdeckt, ein Punkt in $\mathcal{M}_A$ liegt und die beiden anderen außerhalb aber dicht am Rand von $\mathcal{M}_A$ liegen. Bei der Konstruktion des nächsten Dreiecks kommt es nun zu einem Richtungswechsel und die eigentliche Ecke ist nur unzureichend approximiert. In Abbildung 4.9 ist eine solche, nicht ungewöhnliche, Situation dargestellt. Trotz einer Seitenlänge von $a = 0.03$ ergibt sich ein maximaler Fehler von 0.04396. (Hier ließe sich der Fehler auf bis zu $2h - \epsilon$ mit $h = \frac{\sqrt{3}}{2}a$ und $\epsilon > 0$ steigern.) Es sei auch auf die Auswertungen in Tabelle 5.2 hingewiesen. Dort tritt ein ähnlicher Effekt auf.*

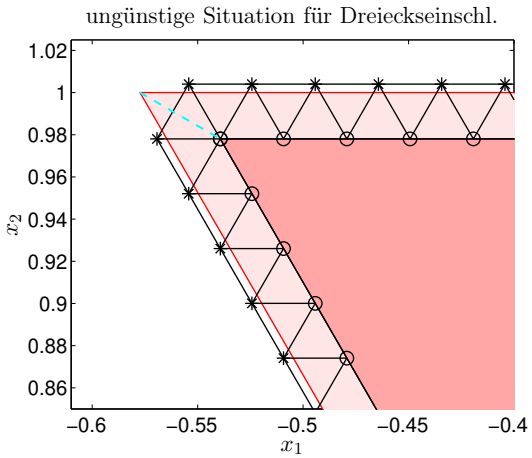


**Abbildung 4.9:** Ungünstige Situation bei der Anwendung der Methode der Dreieckseinschließung, siehe Bemerkung 4.20. Dargestellt sind die Menge $\mathcal{M}_A$ (leichte rote Einfärbung) und deren, mittels Dreieckseinschließung bestimmte, Approximation (tiefe rote Einfärbung). Bei den Dreiecken (schwarz) sind die nicht zulässigen Ecken ($*$) und die zulässigen Ecken ($\circ$) extra markiert. Es entsteht ein maximaler Fehler (türkis gestrichelte Linie), der mit etwa $4.396 \cdot 10^{-2}$ deutlich größer als die Seitenlänge $a = 3 \cdot 10^{-2}$ ist.

### 4.4.3 Erweiterung für $s = 4$

Die Erweiterung der Methode auf $s \geq 4$ ist nicht direkt möglich. Jedoch ist in [58] die Idee vorgestellt, die einzelnen Segmente von $\mathcal{M}_A$ für $s = 4$ scheibenweise zu approximieren und die Dreieckseinschließungsmethode zur Approximation der einzelnen Scheiben zu nutzen, siehe auch [147]. Dieser Ansatz birgt jedoch zwei Schwierigkeiten: Die Approximation ist von der Diskretisierung in der Richtung senkrecht zu den Scheiben abhängig, welche in der Regel, verglichen mit der Seitenlänge der Dreiecke, recht grob ist. Zudem ist die Anwendbarkeit dieser Methode stark von der Topologie der Menge $\mathcal{M}_A$ abhängig. Sofern $\mathcal{M}_A$ in einzelne, klar getrennte Segmente zerfällt, gibt es keine Probleme. Falls $\mathcal{M}_A$ aber nur aus einem Segment besteht und Löcher in der Oberfläche besitzt, kann der Ansatz so nicht verwendet werden. Denn einerseits ist nicht klar, wie die Struktur in der aktuellen Ebene ist. Und andererseits ist für den Fall, dass der Schnitt der Ebene mit $\mathcal{M}_A$ eine zusammenhängende Menge mit Loch ist, nicht klar, wo dieses Loch in der Ebene lokalisiert ist. Dies erschwert eine stabile Implementierung der Methode.

Der Ansatz des inversen Polygon inflation Algorithmus (siehe Abschnitt 4.5), den inneren Rand vom Ursprung ausgehend zu approximieren, ist nicht anwendbar. Dafür sei beispielsweise eine scheibenweise Approximation mit Scheiben parallel zur $x-y$-Ebene angenommen. Nun kann ein $(0, 0, z)$ für $z \neq 0$ aber zu $\mathcal{M}_A$ gehören und es müsste zunächst ein $x$ in der zu untersuchenden Ebene bestimmt werden, das nicht zu $\mathcal{M}_A$ gehört.

**Bemerkung 4.21.**

1. *In [58] ist nur eine scheibenweise Berechnung mit Scheiben parallel zur $x-y$-Ebene beschrieben und angewendet. Generell sollten alle Ebenen zwar parallel zueinander sein, ihre grundsätzliche Richtung ist jedoch frei wählbar. In [147] sind beispielsweise auch Ebenen parallel zur $x-z$- beziehungsweise zur $y-z$-Ebene gewählt.*
2. *Unter Einschränkungen ist die Idee der scheibenweisen Approximation von $\mathcal{M}_A$ auch für höhere Dimensionen (also $s > 4$) erweiterbar, wobei in mehreren Richtungen feste Werte genutzt werden müssten. Es bliebe die Frage zu klären, inwiefern die Topologie von $\mathcal{M}_A$ solche Approximationen erlaubt.*

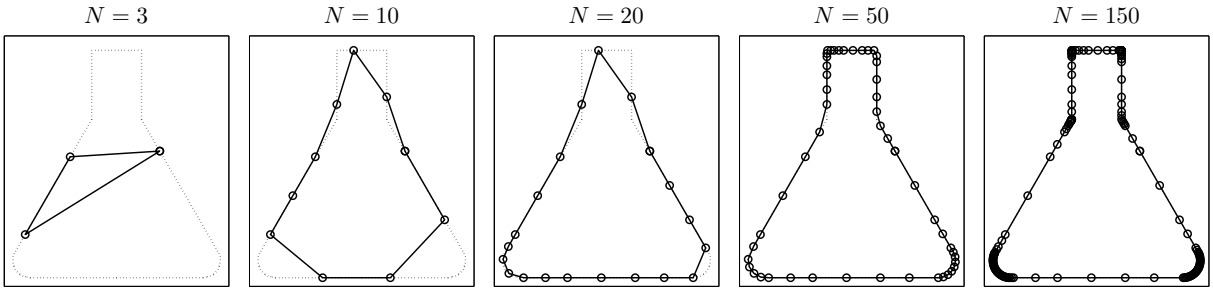


**Abbildung 4.10:** Approximation des Randes eines Erlenmeyerkolbens durch Polygonzüge mit $N \in \{3, 10, 20, 50, 150\}$ Ecken. Zu dem gewählten Startdreieck (linke Grafik) sind mit den Steuerparametern $\varepsilon_b = \varepsilon_d = 10^{-3}$ insgesamt $N = 127$ Ecken nötig, für $\varepsilon_b = \varepsilon_d = 10^{-4}$ sind es $N = 299$ und für $\varepsilon_b = \varepsilon_d = 10^{-5}$ schon $N = 901$. Die Steuerparameter $\varepsilon_b$ und $\varepsilon_d$ sind in Abschnitt 4.5.1 erläutert, siehe (4.26) und (4.29).

## 4.5 Polygon inflation Algorithmen für $s = 3$

Die Polygon inflation Algorithmen (direkter und inverser Typ) sind adaptive Methoden zur Approximation des Randes der Mengen zulässiger Lösungen für $s = 3$. Die direkte Methode wird in [152] und deren inverse Variante in [154] vorgestellt, siehe auch [146, 155] oder [53, 55]. Bei beiden Varianten ist es die Idee, den Rand von $\mathcal{M}_A$ beziehungsweise $\mathcal{M}_C$ durch Polygonzüge zu approximieren. Die direkte Variante ist nur zur Berechnung einer Menge zulässiger Lösungen, welche in drei klar getrennte Segmente zerfällt, geeignet. Andernfalls liefert nur der inverse Typ die korrekte Lösung. Dafür ist der Rechenaufwand für die direkte Variante geringer als der für die allgemein anwendbare inverse Methode. Bei der direkten Variante werden die Ränder der Segmente von innen heraus approximiert. Bei der inversen Variante wird $\mathcal{M}_A$ als Schnitt der Obermenge $\mathcal{F}_A$ aus (2.8) und einer zweiten Obermenge $\mathcal{M}_A^*$, die in (4.23) definiert wird, bestimmt. Da die einzelnen Erweiterungen des Polygons meistens mit Flächenvergrößerungen einhergehen, trägt die Methode den Namen *Polygon inflation*. Mitunter verkleinert sich die aktuelle Flächen mit einer Iteration aber auch. Ebenso wie mit der Dreieckseinschließungsmethode [56, 58], lassen sich mit den Polygon inflation Algorithmen die Mengen zulässiger Lösungen stabil für fehlerbehaftete Daten bestimmen. Die Funktionsweise des Polygon inflation Verfahrens ist in Abbildung 4.10 beispielhaft für die Approximation des Randes eines Erlenmeyerkolbens dargestellt.

Die Erweiterung des Ansatzes auf $s = 4$ ist möglich. Dies wurde umfangreich in [131] untersucht. Der Rand von $\mathcal{M}_A$ wird mittels einer Oberflächentriangulierung approximiert. Die Verfeinerung der Dreiecksstrukturen führt oft zu Schwierigkeiten, sodass sich letztendlich nur der inverse Polyhedron inflation Algorithmus unter Ausnutzung der Strahleneigenschaft aus Satz 3.15 als geeignet erwiesen hat, siehe Abschnitt 4.6 für weitere Erläuterungen.

### Notation für $s = 3$

Die Polygon inflation Algorithmen wurden für die Approximation der Menge zulässiger Lösungen $\mathcal{M}_A$ für $s = 3$ entwickelt. Somit hat die Transformation $T$ aus (2.5) die Form

$$T = \begin{pmatrix} 1 & x_1 & x_2 \\ 1 & S_{11} & S_{12} \\ 1 & S_{21} & S_{22} \end{pmatrix}$$

und analog zu (2.6) ist die (idealisierte) Menge aller zulässigen Lösungen

$$\mathcal{M}_A = \left\{ x \in \mathbb{R}^2 : \exists S \in \mathbb{R}^{2\times 2} \text{ mit } \operatorname{rank}(T) = 3,\, U\Sigma T^{-1} \geq 0,\, TV^{\mathbf{T}} \geq 0 \right\}.$$

Neben der Obermenge $\mathcal{F}_A$ aus (2.8) ist für den inversen Typ die Menge

$$\left\{x \in \mathbb{R}^2 :\ \exists S \in \mathbb{R}^{2\times 2} \text{ mit } \operatorname{rank}(T) = 3,\ U\Sigma T^{-1} \geq 0,\ (TV^{\mathbf{T}})(2:3,:) \geq 0\right\} \tag{4.22}$$

wichtig. In Bezug auf $\mathcal{F}_A$ berücksichtigt die Menge aus (4.22) genau die verbleibenden Restriktionen. Zusammen mit $\mathcal{F}_A$ wird diese genutzt, um $\mathcal{M}_A$ als deren Schnitt zu bestimmen. Um den Rechenaufwand möglichst gering zu halten, wird die Menge aus (4.22) an Stellen, die für die Berechnung von $\mathcal{M}_A$ unwichtig sind, nicht direkt bestimmt. Stattdessen wird die Menge aus (4.22) schlicht durch die Obermenge

$$\begin{aligned}\mathcal{M}_A^* =& \left\{x \in \mathbb{R}^2 :\ \frac{1}{2}\left\| \min\left(0, (1,x^{\mathbf{T}})V^{\mathbf{T}} / \left\|(1,x^{\mathbf{T}})V^{\mathbf{T}}\right\|_\infty + \varepsilon_a\right)\right\|_2^2 \geq \varepsilon_{\text{out}}\right\} \\ &\cup \left\{x \in \mathbb{R}^2 :\ \exists S \in \mathbb{R}^{2\times 2} \text{ mit } \operatorname{rank}(T) = 3,\ U\Sigma T^{-1} \geq 0,\ (TV^{\mathbf{T}})(2:3,:) \geq 0\right\}\end{aligned} \tag{4.23}$$

mit einem $\varepsilon_{\text{out}} > 0$, etwa $\varepsilon_{\text{out}} \in [0.01,\ 0.05]$, ersetzt. Es gilt $\mathcal{M}_A = \mathcal{F}_A \cap \mathcal{M}_A^*$. Die Menge $\mathcal{M}_A^*$ ist unbeschränkt und ihr Rand enthält Teile des inneren Randes von $\mathcal{M}_A$ oder entspricht dem gesamten inneren Rand von $\mathcal{M}_A$.

**Lemma 4.22.** *Unter den Voraussetzungen aus Satz 3.10 gilt* $o = (0,\dots,0)^{\mathbf{T}} \notin \mathcal{M}_A^*$.

*Beweis.* Nach Satz 3.10 gelten $o \in \mathcal{F}_A$ und $o \notin \mathcal{M}_A$. Da trivialerweise $\mathcal{M}_A = \mathcal{F}_A \cap \mathcal{M}_A^*$ gilt, folgt die Behauptung. □

**Bemerkung 4.23.** *Die Menge $\mathcal{M}_A^*$ mag auf den ersten Blick merkwürdig erscheinen, dies ist sie aber nicht. Zu der Menge aus (4.22) werden so lediglich Bereiche mit hinzu genommen, die für $\mathcal{M}_A$ nicht relevant sind, aber den Rechenaufwand mitunter deutlich reduzieren.*

In der Implementierung ist es so organisiert, dass der (innere) Rand von $\mathcal{M}_A^*$ von innen heraus mittels des Polygon inflation Verfahrens mit dem Ursprung als Startpunkt approximiert wird. Die Idee zu $\mathcal{M}_A^*$ wird später in Abbildung 4.12 verdeutlicht.

#### Klassifizierung eines $x \in \mathbb{R}^{s-1}$ für den inversen Polygon inflation Algorithmus

Beim inversen Polygon inflation Algorithmus werden Approximationen für die Mengen $\mathcal{F}_A$ und $\mathcal{M}_A^*$ getrennt berechnet und die Menge zulässiger Lösungen $\mathcal{M}_A$ ergibt sich als deren Schnitt. Somit werden auch zwei verschiedene Routinen zur Klassifizierung eines $x$ eingesetzt. Da im Folgenden auch der inverse Polygon inflation Algorithmus mit vorgestellt und erläutert wird, werden die beiden Klassifizierungen an dieser Stelle eingeführt.

Die eine Klassifizierung ist der Schnelltest aus (4.5). Die andere Klassifizierung ist so aufgebaut, dass mit ihr der Rand der Menge $\mathcal{M}_A^*$ aus (4.23) bestimmt werden kann. Dabei wird die allgemeine Vorgehensweise des Polygon inflation Verfahrens zur Bestimmung von $\mathbb{R}^{s-1} \setminus \mathcal{M}_A^*$ von innen heraus genutzt.

**Klassifizierung 4.24** (Inverser Typ; äußerer Rand)**.** *Für ein $x \in \mathbb{R}^{s-1}$ wird der Schnelltest, ob (4.5) erfüllt ist, durchgeführt. Ist (4.5) erfüllt, so wird $x$ als $x \in \mathcal{F}_A$ klassifiziert, andernfalls als $x \notin \mathcal{F}_A$.*

**Klassifizierung 4.25** (Inverser Typ; innerer Rand)**.** *Für ein $x \in \mathbb{R}^{s-1}$ wird zunächst überprüft, ob*

$$\frac{1}{2}\sum_{i=1}^{n}\left(\min\left(0, \frac{(1,x^{\mathbf{T}})V^{\mathbf{T}}}{\|(1,x^{\mathbf{T}})V^{\mathbf{T}}\|_\infty} + \varepsilon_a\right)\right)^2 \geq \varepsilon_f + \varepsilon_{\text{out}} \tag{4.24}$$

*mit $\varepsilon_{\text{out}} > 0$, beispielsweise $\varepsilon_{\text{out}} = 0.01$, gilt. Ist (4.24) nicht erfüllt, so wird weiter getestet, ob*

$$\widetilde{F}(x) = \min_{S \in \mathbb{R}^{(s-1)\times(s-1)}} \widetilde{f}(x, S) \leq \varepsilon_f$$

*gilt mit*

$$\begin{aligned}\widetilde{f}(x, S) = \frac{1}{2}\Bigg( &\sum_{i=1}^{s}\sum_{j=1}^{k}\left(\min\left(0, \frac{C_{ji}}{\|C(:,i)\|_\infty} + \varepsilon_c\right)\right)^2 + \\ &\sum_{i=2}^{s}\sum_{j=1}^{n}\left(\min\left(0, \frac{A_{ij}}{\|A(i,:)\|_\infty} + \varepsilon_a\right)\right)^2 + \|I_s - TT^+\|_F^2\Bigg).\end{aligned} \tag{4.25}$$

*Ist entweder (4.24) erfüllt oder gilt $\widetilde{F}(x) \leq \varepsilon_f$, so wird $x$ als $x \in \mathcal{M}_A^*$ klassifiziert.*

**Bemerkung 4.26.** *Die Fehlerschranke $\varepsilon_f$ wird damit für beide Tests angewendet und es ist nur $f(x, S) \leq 2\varepsilon_f$ garantiert. Diese Vorgehensweise ist akzeptabel, da $\varepsilon_f$ sehr klein ist.*

### 4.5.1 Folge von Polygonzügen zur Randapproximation

Die beiden Varianten des Polygon inflation Algorithmus basieren auf derselben Idee. Sie erlauben die Approximation der Segmente nicht verschwindender Fläche der Menge zulässiger Lösungen $\mathcal{M}_A$. Dabei sind die Anwendungsbereiche leicht unterschiedlich. Die direkte Variante eignet sich nur zur Approximation der Menge $\mathcal{M}_A$, wenn diese aus drei Segmenten besteht. Dafür ist aber der Übergang zur Approximation von Strecken- und Punktsegmenten offen. Die inverse Variante ermöglicht die Approximation von $\mathcal{M}_A$ unabhängig von der Anzahl deren Segmente. Jedoch ist der Übergang zu Strecken- und Punktsegmenten zwar möglich, aber diese sind deutlich aufwendiger zu detektieren.

Außer für den Fall, dass die Menge $\mathcal{M}_A$ aus einem Segment mit einen Loch um den Nullpunkt besteht, sind die einzelnen Segmente nach Korollar 3.19 frei von Löchern und es reicht jeweils deren Ränder zu bestimmen. Bei den Polygon inflation Methoden werden diese durch eine Folge von Polygonzügen angenähert, wobei in jedem Schritt die jeweils aktuelle Approximation durch die Hinzunahme eines Punktes verfeinert wird. Die Punkte der Polygonzüge werden so bestimmt, dass sie zu $\mathcal{M}_A$ gehören und nicht weiter als eine vorgegebene Genauigkeit vom Rand entfernt sind. Nach einer Initialisierungsphase wird anschließend in jedem Schritt eine Kante zur Verfeinerung ausgewählt und auf der Mittelsenkrechten dieser Kante ein neuer Punkt als Randapproximation bestimmt. Dies wird solange fortgeführt, bis eine hinreichend genaue Approximation erzielt wurde. Zur Initialisierung der Methode wird zunächst ein Dreieck um einen Ausgangspunkt bestimmt. Als Ausgangspunkt zur Bestimmung des Startdreiecks wird beim direkten Typ jeweils ein Punkt aus dem Segment gewählt. Beim inversen Typ fungiert, sowohl für die Approximation von $\mathcal{F}_A$ als auch zur Approximation von $\mathcal{M}_A^*$, der Null- als Ausgangspunkt, wobei zur Approximation des Randes von $\mathcal{M}_A^*$ die Menge $\mathbb{R}^{s-1} \setminus \mathcal{M}_A^*$ angenähert wird.

In diesem Abschnitt wird die Verfahrensweise der Polygon inflation Algorithmen erläutert. Dies erfolgt nur anhand des direkten Typs und der Approximation des Randes eines Segments. Die Funktionsweise der Approximation mittels Polygonzügen ist für den inversen Typ gleich. Bei der direkten Methode werden die Ränder der drei Segmente mittels dreier Polygonzüge bestimmt, beim inversen Typ werden die Ränder von $\mathcal{F}_A$ und $\mathcal{M}_A^*$ bestimmt und anschließend wird $\mathcal{M}_A = \mathcal{F}_A \cap \mathcal{M}_A^*$ gebildet.

### Polygonzug und Lage der Approximationen zum Rand

Die Folge der Polygonzüge wird mit $\{\mathcal{P}^{(q)}\}_{q=0,1,2,\ldots}$ bezeichnet. Dabei setzt sich $\mathcal{P}^{(q)}$ aus den Punkten $\mathcal{P}_1^{(q)}, \ldots, \mathcal{P}_l^{(q)}$ zusammen und $\mathcal{P}^{(0)}$ bezeichnet das Startdreieck. Es gilt $l = q + 3$, für den Fall, dass bis zum $q$-ten Polygonzug kein Punkt eliminiert wurde. Die einzelnen $\mathcal{P}_i^{(q)}$ dienen als Approximationen des Randes und werden so berechnet, dass mit einem zu wählenden Steuerparameter $\varepsilon_b > 0$

$$\mathcal{P}_i^{(q)} \in \mathcal{M}_A, \qquad \min_{x \notin \mathcal{M}_A} \left\| x - \mathcal{P}_i^{(q)} \right\|_2 \leq \varepsilon_b \tag{4.26}$$

für alle $i$, $q$ gelten.

### Bestimmung des Initialisierungsdreiecks

Das Dreieck $\mathcal{P}^{(0)}$ initiiert die Polygonzugfolge. Ausgangspunkt zur Berechnung dieses ist eine nichtnegative Matrixfaktorisierung $D = CA$. Im Fall $s = 3$ ergeben sich aus dieser drei zulässige Lösungen $x^{[1]}, x^{[2]}, x^{[3]}$ als

$$x^{[i]} = \left( \frac{A(i,:)V(:,2:3)}{A(i,:)V(:,1)} \right)^{\mathbf{T}}, \qquad i = 1, 2, 3,$$

wovon, sofern $\mathcal{M}_A$ aus drei Segmenten besteht, jede in einem eigenen Segment liegt. Es sei $x^{[1]}$ mit dem Ziel ausgewählt, den Rand des dazugehörigen Segments zu berechnen. Für die weitere Beschreibung des Verfahrens zur Approximation des zu $x^{[1]}$ gehörigen Segments sei $x^{(0)} := x^{[1]}$. Weiter seien mit $v^{(i)} \in \mathbb{R}^2$, $i = 1, \ldots, m$, eine Reihe von $m$ (in FacPack $m = 50$) verschiedenen Suchrichtungen definiert, mit deren Hilfe das initiale Polygon (Dreieck) bestimmt wird. Sinnvoll ist es beispielsweise, als erste vier Richtungen

$$v^{(1)} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad v^{(2)} = -v^{(1)}, \quad v^{(3)} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \quad v^{(4)} = -v^{(3)}$$

zu nutzen und die weiteren Richtungen äquiangular und in sinnvoller Reihenfolge zu wählen. In dieser Art ist es auch bei der Polygon inflation Implementierung in FacPack umgesetzt. Oft werden nur die ersten drei Richtungen benötigt.

Die Bestimmung des ersten Punktes $\mathcal{P}_1^{(0)}$ erfolgt in der Suchrichtung $v^{(1)}$, sodass

$$\mathcal{P}_1^{(0)} = x^{(0)} + \gamma v^{(1)}$$

mit $\gamma \geq 0$ eine Approximation an einen Randpunkt im Sinne von (4.26) ist. Anschließend wird

$$\mathcal{P}_2^{(0)} = x^{(0)} + \gamma v^{(2)},$$

ebenfalls als eine Approximation eines Randpunktes, mit $\gamma \geq 0$ bestimmt. Dabei sei $\mathcal{P}_1^{(0)} \neq \mathcal{P}_2^{(0)}$ angenommen. Der Fall $\mathcal{P}_1^{(0)} = \mathcal{P}_2^{(0)}$ ist in Bemerkung 4.27 (Punkt 2) behandelt. Zur Konstruktion des Startdreiecks wird $\mathcal{P}_3^{(0)}$ mittels der Suchrichtung $v^{(3)}$ und ausgehend vom Mittelpunkt der Strecke zwischen $\mathcal{P}_1^{(0)}$ und $\mathcal{P}_2^{(0)}$ berechnet. Es ergibt sich

$$\mathcal{P}_3^{(0)} = \frac{1}{2}\left( \mathcal{P}_1^{(0)} + \mathcal{P}_2^{(0)} \right) + \gamma v^{(3)}$$

mit $\gamma \geq 0$. Sollten $\mathcal{P}_1^{(0)}$, $\mathcal{P}_2^{(0)}$ und $\mathcal{P}_3^{(0)}$ auf einer Geraden liegen, so werden zur Berechnung von $\mathcal{P}_3^{(0)}$ die weiteren Suchrichtungen ausprobiert. Sofern eine dieser auf eine Schrittweite $\gamma >$

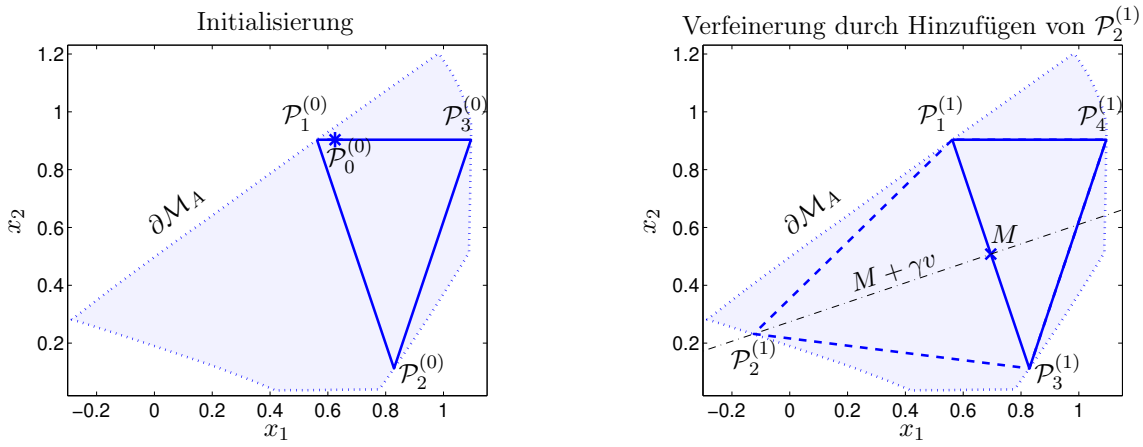


**Abbildung 4.11:** Der Polygon inflation Algorithmus im Detail am Beispiel eines Segments von $\mathcal{M}_A$ zu Datensatz 3. Links: Mittels des initialen $\mathcal{P}_0^{(0)}$ wird das Startdreieck bestimmt. Rechts: Zur Verfeinerung der Approximation von $\partial\mathcal{M}_A$ wird $\mathcal{P}_2^{(1)}$ hinzugefügt und so die Kante zwischen $\mathcal{P}_1^{(0)}$ und $\mathcal{P}_2^{(0)}$ des aktuellen Polygons (Dreieck) durch zwei neue (zwischen $\mathcal{P}_1^{(0)} = \mathcal{P}_1^{(1)}$ und $\mathcal{P}_2^{(1)}$ sowie zwischen $\mathcal{P}_2^{(1)}$ und $\mathcal{P}_2^{(0)} = \mathcal{P}_3^{(1)}$) ersetzt. Die Approximation des Segments von $\mathcal{M}_A$ ist transparent dargestellt und der Rand $\partial\mathcal{M}_A$ ist gepunktet.

0 führt, liegt ein nicht entartetes Dreieck vor, andernfalls deutet dies daraufhin, dass es sich bei dem zu berechnenden Segment um eine Strecke oder einen isolierten Punkt handelt. In diesen Fällen wird die Berechnung mit dem Polygon inflation Algorithmus zunächst abgebrochen und es werden die später in Abschnitt 4.5.5 vorgestellten Methoden zur Bestimmung derartiger Segmente angewendet.

**Bemerkung 4.27.**

1. *In der* FacPack*-Implementierung werden aus organisatorischen Gründen bezüglich der Implementierung, für den Fall, dass die Punkte von $\mathcal{P}^{(0)}$ im Uhrzeigersinn angeordnet sind, $\mathcal{P}_2^{(0)}$ und $\mathcal{P}_3^{(0)}$ getauscht. Die Ecken des Startdreiecks sind also gegen den Uhrzeigersinn angeordnet.*
2. *Sollte bei der Bestimmung von $\mathcal{P}_1^{(0)}$ und $\mathcal{P}_2^{(0)}$ jeweils der Fall $\gamma = 0$ eintreten, so werden zur Berechnung von $\mathcal{P}_2^{(0)}$ nacheinander die weiteren Suchrichtungen $v^{(i)}$, $i = 3, \ldots, m-1$, ausprobiert, bis eine zu einem $\gamma > 0$ führt. Sollte keine davon, trotz genügend großem $m$, auf ein $\mathcal{P}_2^{(0)}$ mit $\gamma > 0$ führen, so deutet dies daraufhin, dass es sich bei dem zu berechnenden Segment um eine Strecke oder einen isolierten Punkt handelt. In diesem Fall wird die Berechnung eines Polygons abgebrochen und zu den in Abschnitt 4.5.5 erläuterten Algorithmen übergegangen.*
3. *Für den Fall, dass ohne Beschränkung der Allgemeinheit $v^{(j)}$, mit $2 < j < m$, eine Suchrichtung mit $\gamma > 0$ ist, $v^{(2)}$ aber nicht, so wird zur Bestimmung von $\mathcal{P}_3^{(0)}$ natürlich zunächst die Suchrichtung $j + 1$ ausprobiert. Falls diese nicht zu einem Dreieck führt, werden alternativ $v^{(\ell)}$, $\ell = j + 2, \ldots, m$, ausprobiert.*

In der linken Grafik von Abbildung 4.11 ist ein Startdreieck zur Approximation eines Segments von $\mathcal{M}_A$ für Datensatz 3 dargestellt.

### Iteration: Bestimmung eines neuen Punktes

Nachdem das Startpolygon (Dreieck) bestimmt wurde, ist es in der Iteration das Ziel, dieses sukzessive um jeweils einen Randpunkt (also eine zusätzliche Ecke) zu erweitern, um die Approximation des Randes von $\mathcal{M}_A$ zu verfeinern. Dazu wird in jedem Iterationsschritt eine Kante

ausgewählt und durch die Hinzunahme eines neuen Randpunktes durch zwei neue Kanten ersetzt. Welche Kante zur Verfeinerung der Approximation von $\partial\mathcal{M}_A$ ausgewählt wird, ist später in Abschnitt 4.5.1 näher thematisiert. Zunächst wird das Hinzufügen eines neuen Punktes erläutert.

Sei $\mathcal{P}^{(q)}$ das $l$-Eck mit den Randpunkten $(\mathcal{P}_1^{(q)}, \ldots, \mathcal{P}_l^{(q)})$, welches zum $(l+1)$-Eck $\mathcal{P}^{(q+1)}$ mit $(\mathcal{P}_1^{(q+1)}, \ldots, \mathcal{P}_{l+1}^{(q+1)})$ erweitert werden soll. Sei die Kante zwischen $\mathcal{P}_i^{(q)}$ und $\mathcal{P}_{i+1}^{(q)}$ zur Verfeinerung ausgewählt. Das neue $\mathcal{P}_{i+1}^{(q+1)}$ wird als Schnittpunkt der Mittelsenkrechten zur Strecke zwischen $\mathcal{P}_i^{(q)}$ sowie $\mathcal{P}_{i+1}^{(q)}$ und dem Rand von $\mathcal{M}_A$ gewählt. Somit wird das neue Vieleck durch die Ecken

$$\left(\mathcal{P}_1^{(q+1)}, \mathcal{P}_2^{(q+1)}, \ldots, \mathcal{P}_{l+1}^{(q+1)}\right) = \left(\mathcal{P}_1^{(q)}, \mathcal{P}_2^{(q)}, \ldots, \mathcal{P}_i^{(q)}, \mathcal{P}_{i+1}^{(q+1)}, \mathcal{P}_{i+1}^{(q)}, \ldots, \mathcal{P}_l^{(q)}\right)$$

mit dem neuen $\mathcal{P}_{i+1}^{(q+1)}$ beschrieben. Die Mittelsenkrechte zur Strecke zwischen $\mathcal{P}_i^{(q)}$ und $\mathcal{P}_{i+1}^{(q)}$ ist

$$g = \{x \in \mathbb{R}^2 : \exists \gamma \in \mathbb{R} \text{ mit } x = M + \gamma v\} \tag{4.27}$$

mit

$$M = \frac{1}{2}\left(\mathcal{P}_i^{(q)} + \mathcal{P}_{i+1}^{(q)}\right), \quad v \perp \left(\mathcal{P}_{i+1}^{(q)} - \mathcal{P}_i^{(q)}\right).$$

**Bemerkung 4.28.**

1. *Im Normalfall gibt es keinen eindeutigen Schnittpunkt von $g$ und $\partial\mathcal{M}_A$. Es ist $\mathcal{P}_{i+1}^{(q+1)}$ so zu wählen, dass das Polygon, in Bezug auf die Approximation des Randes von $\mathcal{M}_A$, nicht unsinnig wird. (Für eine Menge $\mathcal{M}_A$ mit drei Segmenten gibt es für $l > 3$ jeweils nur einen sinnvollen Punkt.)*
2. *Zur Bestimmung von $\mathcal{P}_{i+1}^{(q+1)}$ wird zunächst getestet, ob $M \in \mathcal{M}_A$ gilt. Ist dies der Fall, so wird $\mathcal{P}_{i+1}^{(q+1)}$ auf $g$ in der Richtung nach außen gesucht, andernfalls in der Richtung nach innen.*
3. *An dieser Stelle ist es bei der Implementierung von Vorteil, die Orientierung des Polygons zu kennen und fest zu lassen. So ist bekannt, auf welcher Seite der zu unterteilenden Kante außen ist. Aus diesem Grund wird unter Umständen bei der Initialisierung die in Bemerkung 4.27 (Punkt 1) erwähnte Umordnung vorgenommen.*
4. *Die Schrittweite bei der Suche nach außen sollte nicht zu groß sein. Andernfalls besteht unter Umständen die Gefahr, in ein anderes Segment zu geraten und die Struktur der Randapproximation von $\mathcal{M}_A$ zu zerstören.*

In der rechten Grafik von Abbildung 4.11 ist der Iterationsschritt von einem Dreieck zu einem Viereck dargestellt.

### Umsetzung und Genauigkeit der Randapproximation

Bei der Bestimmung des neuen $\mathcal{P}_{i+1}^{(q+1)}$ als Schnittpunkt von $g$ aus (4.27) mit $\partial\mathcal{M}_A$ handelt es sich nur um eine Approximation. Diese wird mittels des Bisektionsverfahrens bis auf eine vorgegebene Genauigkeit $\varepsilon_b$ so bestimmt, dass (4.26) erfüllt ist. Das Bisektionsverfahren hat zwar lediglich Konvergenzordnung eins, ist für das zugrunde liegende Problem aber ein guter Kompromiss zwischen Aufwand und Stabilität. Die numerische Schwierigkeit des Problems besteht in dem schwer zugänglichen Verhalten der Funktion $F$ in der Umgebung des Randes. Aus der Erfahrung heraus sind dazu, abhängig von der Genauigkeit $\varepsilon_b$ und der Kantenlänge der zu ersetzenden Kante, durchschnittlich zwischen drei und acht Schritte (für etwa $\varepsilon_b \in [10^{-5}, 10^{-2}]$) notwendig. Der neue Punkt wird letztendlich so gewählt, dass er in $\mathcal{M}_A$ liegt und der Abstand zu $\mathbb{R}^2 \setminus \mathcal{M}_A$ maximal $\varepsilon_b$ beträgt, siehe (4.26). Sinnvoll ist etwa $\varepsilon_b \in [10^{-5}, 10^{-3}]$.

#### Auswahl der zu ersetzenden Kante

Ein wichtiger Teil des Polygon inflation Algorithmus ist die Wahl der zu ersetzenden Kante. Die ausgewählte Kante wird durch zwei neue ersetzt, die zu einem Polygon führen sollen, das eine Ecke mehr hat und den Rand vom aktuellen Segment von $\mathcal{M}_A$ feiner oder zumindest nicht schlechter approximiert. Als Referenzwert wird für jede Kante bei ihrer Einführung der Abstand des neu hinzugefügten Punktes zur ersetzten Kante gespeichert. Vom aktuellen Polygon wird die Kante verfeinert, für die der Referenzwert am größten ist.

Die Referenzwerte werden in dem Vektor $\Delta$ gespeichert. Wird zwischen $\mathcal{P}_i^{(q)}$ und $\mathcal{P}_{i+1}^{(q)}$ der neue Punkt $\mathcal{P}_{i+1}^{(q+1)}$ eingefügt, so wird als Referenzwert

$$\Delta' = \left\| \frac{1}{2} \left( \mathcal{P}_i^{(q)} + \mathcal{P}_{i+1}^{(q)} \right) - \mathcal{P}_{i+1}^{(q+1)} \right\|_2 \tag{4.28}$$

berechnet.[6] Sei $l$ die Anzahl der Ecken vor der Verfeinerung. Sind

$$(\Delta_1, \ldots, \Delta_{i-1}, \Delta_i, \Delta_{i+1}, \ldots, \Delta_l) \in \mathbb{R}^l$$

die Referenzwerte vor der Verfeinerung, so ergibt sich danach der Vektor

$$(\Delta_1, \ldots, \Delta_{i-1}, \Delta', \Delta', \Delta_{i+1}, \ldots, \Delta_l) \in \mathbb{R}^{l+1}.$$

Für den Index $i$ der neu zu teilenden Kante gilt

$$\Delta_i = \max_j \Delta_j.$$

Tritt der maximale Wert in $\Delta$ nicht einfach auf (was oft der Fall ist), wird zunächst der kleinste Index $i$ gewählt. Da nach der Initialisierung für das Startdreieck keine Werte $\Delta$ vorhanden sind, wird zunächst jede der drei Startkanten einmal verfeinert.

#### Abbruch der Iteration

Die Iteration wird solange fortgeführt, bis der maximale Eintrag in $\Delta$ kleiner als eine Abbruchschranke $\varepsilon_d > 0$ ist, sodass nach erfolgreichem Durchlauf

$$\max_{i=1,\ldots,l} \Delta_i < \varepsilon_d \tag{4.29}$$

gilt. Eine geeignete Wahl für die Abbruchschranke ist $\varepsilon_d = \varepsilon_b$.

### 4.5.2 Elimination einzelner Punkte

Die einzelnen Eckpunkte der Polygone sind Approximationen an den Rand von $\mathcal{M}_A$ und es ist das Ziel, dass für jede Ecke $\mathcal{P}_i^{(q)}$ die Ungleichung aus (4.26) erfüllt ist. Die Klassifizierung eines $x$ wird zwar unter großem Aufwand betrieben, jedoch kann es durch eine falsche Klassifizierung dazu kommen, dass für ein konkretes $\mathcal{P}_i^{(q)} \in \mathcal{M}_A$ die Ungleichung aus (4.26) nicht erfüllt ist. Dies passiert, wenn ein $x \in \mathcal{M}_A$ aufgrund der numerischen Optimierung fälschlicherweise als $x \notin \mathcal{M}_A$ angenommen wird, siehe Bemerkung 4.3. Zu falschen Klassifizierungen kommt es zwar nur sehr selten, jedoch kann es in solchen Situationen zu Schwierigkeiten bei der Berechnung des Polygonzugs kommen und die Approximationsgüte kann gestört werden. Daher gilt es, Ecken,

[6] In ersten Versionen des Polygon inflation Algorithmus wurde auch mit der Flächenänderungen, die sich durch das Hinzufügen eines neuen Punktes ergab, als Referenzwert $\Delta'$ gearbeitet.

die nicht genügend dicht am Rand von $\mathcal{M}_A$ liegen, zu erkennen und auszusortieren. Die Details zur Umsetzung des Detektierens werden hier nicht weiter thematisiert. Beispielsweise wird ein Punkt aussortiert, wenn er einen Innenwinkel des Polygons von über $1.5\pi$ erzeugt oder einen, der unerwartet klein ist. Die Bewertung *unerwartet klein* ist dabei auch von der aktuellen Anzahl an Ecken abhängig.

### 4.5.3 Inverser Polygon inflation Algorithmus

Der (direkte) Polygon inflation Algorithmus ist eine Methode zur Approximation der Menge $\mathcal{M}_A$ für $s = 3$, die jedoch nur für bestimmte Topologien von $\mathcal{M}_A$ das korrekte Resultat liefert. Besteht $\mathcal{M}_A$ aus mehr als drei Segmenten, so werden jeweils nur drei bestimmt, und zwar eben jene, welche die niedrigdimensionalen Darstellungen der Zeilen von $A$ der initialen nichtnegativen Faktorisierung enthalten. Dass $\mathcal{M}_A$ aus mehr als drei Segmenten besteht, ist selten. Nicht selten hingegen ist der Fall, dass $\mathcal{M}_A$ aus nur einem Segment besteht. In diesem Fall führt die direkte Polygon inflation Methode nicht auf das korrekte Ergebnis. Damit auch diese Arten von Mengen $\mathcal{M}_A$ korrekt approximiert werden können, wurde der inverse Polygon inflation Algorithmus entwickelt. Sofern in $\mathcal{M}_A$ keine Punkt- oder Streckensegmente auftreten, lässt sich $\mathcal{M}_A$ unabhängig von dessen Topologie mit dem inversen Polygon inflation Algorithmus stabil approximieren.

#### Funktionsweise des inversen Polygon inflation Algorithmus

Die Menge zulässiger Lösungen $\mathcal{M}_A$ wird als Schnitt der Obermengen $\mathcal{F}_A$ und $\mathcal{M}_A^*$ berechnet. Der Rand von $\mathcal{F}_A$ ist eine Obermenge des äußeren Randes von $\mathcal{M}_A$ oder stimmt komplett mit diesem überein. Die Überprüfung eines $x$, ob es zu $\mathcal{F}_A$ gehört oder nicht, erfolgt über die Klassifizierung 4.24 und mit den Steuerparametern $\varepsilon_a$ und $\varepsilon_f$. Die Klassifizierung ist explizit und fehlerfrei. Der Rand von $\mathcal{M}_A^*$ ist eine Obermenge des inneren Randes von $\mathcal{M}_A$ oder stimmt komplett mit diesem überein. Die Überprüfung eines $x$, ob es zu $\mathcal{M}_A^*$ gehört oder nicht, erfolgt über die Klassifizierung 4.25 und mit den Steuerparametern $\varepsilon_a$, $\varepsilon_c$ und $\varepsilon_f$. Die Klassifizierung ist, da sie mittels der Minimierung von $\widetilde{f}$ aus (4.25) durchgeführt wird, indirekt und somit nicht sicher fehlerfrei, vergleiche Bemerkung 4.3.

Die Berechnung von $\mathcal{M}_A$ mittels des inversen Polygon inflation Verfahrens gliedert sich in folgende drei Schritte, welche anschließend einzeln erläutert werden:

1. Zunächst wird der Rand von $\mathcal{F}_A$ bestimmt.
2. Anschließend wird der (innere) Rand von $\mathcal{M}_A^*$ berechnet.
3. Schlussendlich ergibt sich $\mathcal{M}_A$ als Schnitt von $\mathcal{F}_A$ und $\mathcal{M}_A^*$.

Die Bestimmung des Randes von $\mathcal{F}_A$ erfolgt mit dem Polygon inflation Verfahren und mit dem Ursprung als Startpunkt. Der Ursprung eignet sich als Startpunkt zur Bestimmung eines initialen Dreiecks, da er nach Satz 3.10 (unter schwachen Voraussetzungen) zu $\mathcal{F}_A$ gehört. Im Vergleich zur Approximation von $\mathcal{M}_A^*$ ist die von $\mathcal{F}_A$ wenig rechenintensiv.

Im zweiten Schritt wird erneut das Polygon inflation Verfahren angewendet, diesmal um den (inneren) Rand von $\mathcal{M}_A^*$ zu bestimmen. Die grundsätzliche Idee ist es, die im Hinblick auf $\mathcal{F}_A$ verbleibenden Bedingungen an eine zulässige Lösung zu berücksichtigen, vergleiche die Menge aus (4.22). Da es für die Menge aus (4.22) aber nur auf ihren Schnitt mit $\mathcal{F}_A$ und somit auf den inneren Rand von $\mathcal{M}_A$ ankommt, kann der Rand der Menge aus (4.22) an Stellen genügend weit außerhalb von $\mathcal{F}_A$ abgeschnitten werden. Dies führt auf die Obermenge $\mathcal{M}_A^*$ aus (4.23) und die Klassifizierung 4.25. Ausgangspunkt für die Bestimmung eines Startdreiecks ist erneut der Ursprung. Da dieser nach Lemma 4.22 (unter schwachen Voraussetzungen) nicht zu $\mathcal{M}_A^*$ gehört,

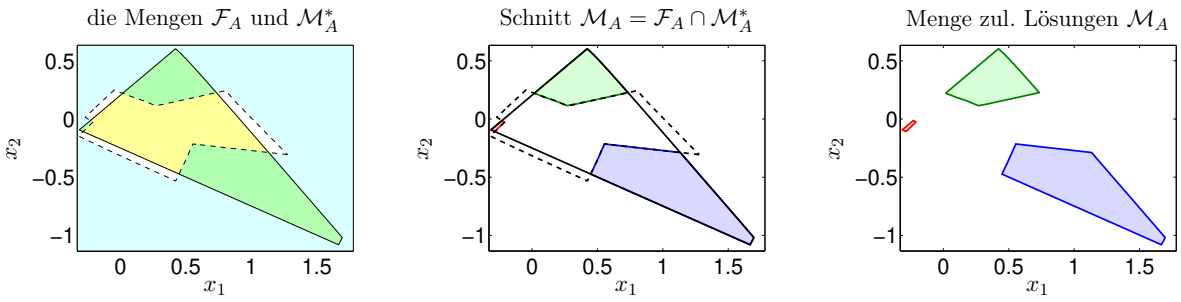


**Abbildung 4.12:** Vorgehensweise beim inversen Polygon inflation Algorithmus am Beispiel des Datensatzes 2. Links: Zunächst werden $\mathcal{F}_A$ (gelb, Rand: durchgezogene Linie) und $\mathcal{M}_A^*$ (türkis, Rand: gestrichelte Linie) berechnet. (Die Überlagerung ist etwas dunkler eingefärbt.) gut zu erkennen ist, inwiefern die spezielle Art der Menge $\mathcal{M}_A^*$ motiviert ist. Anstatt die Menge aus (4.22) zu nutzen, werden zu dieser einfach Teile von $\mathbb{R}^2 \setminus \mathcal{M}_A$ hinzugenommen, die für die Bestimmung von $\mathcal{M}_A$ nicht relevant sind. So wird der Rechenaufwand für eine Menge $\mathcal{M}_A$, die aus drei oder mehr Segmenten besteht, gering gehalten. Mitte: Der Schnitt der berechneten Mengen wird bestimmt. Rechts: die drei Segmente von $\mathcal{M}_A$.

wird der Rand des Komplements von $\mathcal{M}_A^*$, also von $\mathbb{R}^2 \setminus \mathcal{M}_A^*$, bestimmt. Die aufwendigen Teile dieser Randapproximation sind die einzelnen Klassifizierungen von Punkten, welche jeweils die Lösungen von nichtlinearen Ausgleichsproblemen beinhalten.

Im abschließenden Schritt wird der Schnitt der zuvor berechneten Approximationen an $\mathcal{F}_A$ und $\mathcal{M}_A^*$ ermittelt. Dies ergibt $\mathcal{M}_A$. In Abbildung 4.12 ist die Funktionsweise des inversen Polygon inflation Algorithmus am Beispiel des Datensatzes 2 demonstriert.

### Einsatz der Steuerparameter

Ebenso wie beim (direkten) Polygon inflation Algorithmus, werden auch bei der inversen Variante einige Kontrollparameter eingesetzt, um die Berechnung im Sinne der Anforderungen an die Approximation optimal zu steuern. Die Parameter haben größtenteils die gleichen Funktionen wie beim (direkten) Polygon inflation Algorithmus.

Die Einbindung von Störungen durch das Zulassen betragskleiner negativer Einträge erfolgt über die Parameter $\varepsilon_a$ und $\varepsilon_c$ in den Klassifizierungen 4.24 sowie 4.25 und unter Nutzung von (4.5) sowie der Funktion $\widetilde{f}(x, S)$ aus (4.25). An diesen Stellen ist auch die Schranke für akzeptable Funktionswerte $\varepsilon_f$ eingesetzt. Die Genauigkeit für die Randapproximation ist $\varepsilon_b$ und die Abbruchschranke für die beiden Iterationen ist $\delta$.

### Lage der Approximationen zum Rand

Die Eckpunkte der einzelnen Polygone sind Approximationen an die Ränder von $\mathcal{F}_A$ und $\mathcal{M}_A^*$. Der Schnitt $\mathcal{F}_A \cap \mathcal{M}_A^*$ soll auf $\mathcal{M}_A$ führen und es gilt, die Randpunkte dementsprechend jeweils zugehörig zu $\mathcal{F}_A$ und zu $\mathcal{M}_A^*$ zu wählen. Die Punkte sollen nicht weiter als $\varepsilon_b$ von $\mathbb{R}^2 \setminus \mathcal{F}_A$ beziehungsweise von $\mathbb{R}^2 \setminus \mathcal{M}_A^*$ entfernt sein.

Diese Konvention hat den Nachteil, dass mit dem inversen Polygon inflation Algorithmus keine Strecken- oder Punktsegmente bestimmt werden können. Stattdessen kommt es in diesen Regionen in der Regel zu einem leeren Schnitt der Approximationen an $\mathcal{F}_A$ und $\mathcal{M}_A^*$. Bereits leichte, geeignete Modifikationen an den Steuerparametern führen jedoch zu Mengen, die zwar sehr lang gezogen sind, aber eine, wenngleich sehr kleine, nicht verschwindende Fläche haben.

#### Vorteile bei nicht klar separierten Segmenten

Der (direkte) Polygon inflation Algorithmus ist für die Berechnung von Mengen zulässiger Lösungen entwickelt worden, die aus drei klar separierten Segmenten bestehen. Die Berechnungen können instabil werden, falls es zwar drei Segmente sind, diese aber dicht beieinander liegen. Das auftretende Problem ist, dass beim iterativen Hinzufügen von Punkten die Ränder von zwei Segmenten gemischt werden können. Ist dies für ein Segment einmal passiert, so bricht die Struktur der Approximation dieses Segments schnell zusammen.

In solchen Fällen ist der inverse Polygon inflation Algorithmus die stabilere und bessere Wahl. Die Berechnungen von $\mathcal{F}_A$ und $\mathcal{M}_A^*$ sind von den eben genannten Störungen nicht betroffen und unabhängig von der Topologie von $\mathcal{M}_A$. Einzig Punkt- und Streckensegmente lassen sich ohne algorithmische Erweiterungen nicht berechnen.

### 4.5.4 Automatischer Wechsel des Polygon inflation Typs

In der FacPack-Implementierung des Polygon inflation Algorithmus wird zur Berechnung der Menge $\mathcal{M}_A$ für $s = 3$ die Auswahl des Typs angeboten, *direkt* oder *invers*. Standardmäßig ist in FacPack der direkte Polygon inflation Algorithmus eingestellt. Oft ist die Struktur von $\mathcal{M}_A$ jedoch von vornherein nicht bekannt und es ist sinnvoll, gegebenenfalls den Typ automatisiert zu wechseln, sofern Indizien darauf hindeuten, dass $\mathcal{M}_A$ nur aus einem Segment besteht oder die Segmente dicht beieinander liegen. In FacPack wird automatisch auf den inversen Typ umgeschaltet, wenn sich ein Segment über mindestens drei Quadranten der Ebene erstreckt.

Ein anderes Kriterium, Unregelmäßigkeiten bei der Berechnung der Segmente von $\mathcal{M}_A$ zu detektieren und automatisiert vom direkten auf den inversen Polygon inflation Algorithmus umzuschalten, wird im folgenden Unterabschnitt erläutert. Dieses ist jedoch instabil und wird in FacPack nicht eingesetzt.

#### Ein instabiles Umschaltkriterium

Die Idee ist, nach der Berechnung von $\mathcal{M}_A$ mit dem direkten Polygon inflation Algorithmus, jedes der drei Segmente einzeln auf Unstimmigkeiten zu untersuchen. Sei dazu ein einzelnes Segment betrachtet. Zunächst wird dessen Schwerpunkt $S$ berechnet. Anschließend werden zu allen Paaren benachbarter Punkte $\mathcal{P}_i^{(q)}$ und $\mathcal{P}_j^{(q)}$ des finalen Polygons die Winkel $\angle \mathcal{P}_i^{(q)} S \mathcal{P}_j^{(q)}$ bestimmt und aufaddiert. Für ein konvexes Segment ist die Summe der Winkel mit Rücksicht auf Rundungsfehler gleich $2\pi$. Für nicht konvexe Segmente gilt dies ebenso, sofern sich keine zwei (benachbarten) der insgesamt $m$ Dreiecke $\Delta(\mathcal{P}_i^{(q)}, S, \mathcal{P}_{i+1}^{(q)})$, $i = 1 \ldots, m-1$, sowie $\Delta(\mathcal{P}_m^{(q)}, S, \mathcal{P}_1^{(q)})$ echt überschneiden. Dabei ist $m$ die Anzahl der Ecken des Polygons. Sollte sich also für ein Segment ein signifikanter Unterschied zu $2\pi$ ergeben, so würde dies auf eine Struktur von $\mathcal{M}_A$ hindeuten, welche mit dem direkten Polygon inflation Algorithmus schwierig zu berechnen ist. In einem solchen Fall würde auf den inversen Polygon inflation Algorithmus umgeschaltet und die Berechnung erneut gestartet werden.

Dieses Kriterium ist für die Entscheidung, ob vom direkten auf den inversen Polygon inflation Algorithmus umgeschaltet werden sollte oder nicht, zu sensibel. Es ist auch für korrekt bestimmte Lösungen möglich, dass sich zwei von $\mathcal{P}_i^{(q)}$, $S$ sowie $\mathcal{P}_{i+1}^{(q)}$ und von $\mathcal{P}_{i+1}^{(q)}$, $S$ sowie $\mathcal{P}_{i+2}^{(q)}$ aufgespannte Dreiecke echt überschneiden. Inwiefern dieses Kriterium mitunter zu sensibel ist, zeigt das folgende Beispiel:

**Beispiel 4.29.** *Betrachtet werden zwei Matrizen: $D$ aus Datensatz 2 sowie $\widetilde{D} = C(A+0.095)$ mit den ursprünglichen Faktoren $C$ und $A$ aus Datensatz 2. Für beide Matrizen $D$ und $\widetilde{D}$ bestehen*

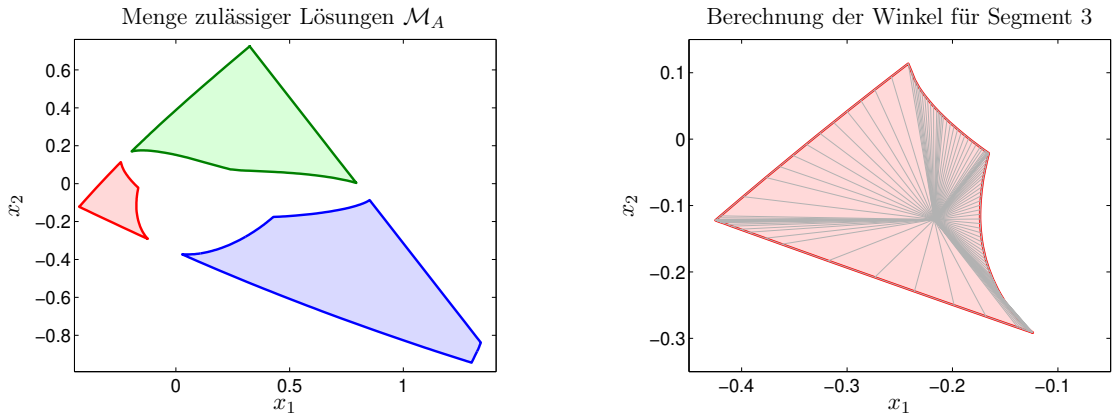


**Abbildung 4.13:** Illustration zum Beispiel 4.29. Links: Die Menge zulässiger Lösungen $\mathcal{M}_A$ für $\widetilde{D} = C(A + 0.095)$ mit $C$ und $A$ den ursprünglichen Faktoren aus Datensatz 2 besteht aus drei getrennten Segmenten. Rechts: Bei der Variante, für alle Paare benachbarter Punkte eines Segments die Winkel $\angle \mathcal{P}_i^{(q)} S \mathcal{P}_i^{(q)}$ mit $S$ dem Schwerpunkt des zu untersuchenden Segments aufzuaddieren, würde unnötigerweise auf den inversen Polygon inflation Algorithmus umgeschaltet werden. Für das rote Segment ergibt sich eine signifikante Abweichung von $2\pi$ in Höhe von $2.5 \cdot 10^{-2}$. Das Problem ist, dass sich einige benachbarte Dreiecke $\Delta(\mathcal{P}_i^{(q)}, S, \mathcal{P}_{i+1}^{(q)})$ und $\Delta(\mathcal{P}_{i+1}^{(q)}, S, \mathcal{P}_{i+2}^{(q)})$ echt überschneiden. Zur Entscheidung, ob die Berechnung von $\mathcal{M}_A$ durch den direkten Polygon inflation Algorithmus erfolgreich war oder nicht, ist ein solcher Ansatz für dieses und ähnliche Beispiele nicht geeignet.

*die Mengen $\mathcal{M}_A$ aus drei separierten Segmenten und sind mit dem direkten Polygon inflation Algorithmus problemlos berechenbar.*

*Für die einzelnen Segmente der Menge $\mathcal{M}_A$ zu $D$ ergeben sich die Abweichungen $8.5 \cdot 10^{-5}$, $2.2 \cdot 10^{-12}$ und $-4.4 \cdot 10^{-12}$ von $2\pi$ für die jeweiligen Summen aller Winkel $\angle QSR$ für Paare $Q$ und $R$ benachbarter Punkte. Der Wert $8.5 \cdot 10^{-5}$ ist fragwürdig aber im Bereich einer numerischen Approximation von $\mathcal{M}_A$ mit dem Polygon inflation Algorithmus akzeptabel (nicht signifikant). Bei der Bildung aller Dreiecke durch $Q$, $S$ und $R$ ($Q$ und $R$ benachbarte Punkte, $S$ der Schwerpunkt des Segments) kommt es tatsächlich zu einer Überschneidung zweier Dreiecke. Die Werte $2.2 \cdot 10^{-12}$ und $-4.4 \cdot 10^{-12}$ sind zweifelsfrei keine signifikanten Abweichungen.*

*Die Abweichungen für $\widetilde{D}$ sind $5.0 \cdot 10^{-5}$, $1.8 \cdot 10^{-2}$ und $2.5 \cdot 10^{-2}$. Die letzten beiden Werte sind offensichtlich signifikant. In diesem Fall würde unnötigerweise auf den inversen Polygon inflation Algorithmus umgeschaltet werden. Das Problem ist die Lage des Schwerpunktes, sodass sich einige benachbarte Dreiecke $\Delta(\mathcal{P}_i^{(q)}, S, \mathcal{P}_{i+1}^{(q)})$ und $\Delta(\mathcal{P}_{i+1}^{(q)}, S, \mathcal{P}_{i+2}^{(q)})$ echt überschneiden. Dieser Effekt ist in Abbildung 4.13 für das rote Segment verdeutlicht.*

### 4.5.5 Berechnung von Streckensegmenten und isolierten Lösungen

Unter Umständen können die Mengen zulässiger Lösungen auch isolierte Punkte und Streckensegmente enthalten. Solche Fälle treten für Messdaten faktisch nicht auf, sehr wohl aber für speziell konstruierte Modelldaten. Eine isolierte Lösung zu berechnen, sofern die Menge $\mathcal{M}_A$ nur aus drei Segmenten besteht, ist kein Problem, da diese durch die initiale nichtnegative Matrixfaktorisierung bekannt ist. Ein Streckensegment lässt sich mit der Polygon inflation Methode nicht direkt bestimmen und es bedarf eines anderen Ansatzes. Zudem ist es wichtig, dass zu einer initialen zulässigen Lösung erkannt wird, ob diese zu einem Streckensegment gehört oder isoliert ist. Beide Fälle lassen sich damit erledigen, dass die Approximation eine Streckensegments gelingt.

Bei der Anwendung des direkten Polygon inflation Algorithmus führt ein degeneriertes Segment zu keinem Startdreieck. Selbst für den Fall eines Streckensegments wird, da nur eine endliche (beispielsweise $m = 50$) Menge an Suchrichtungen durchprobiert wird, fast sicher nicht die richtige getroffen. Insofern kein Startdreieck bestimmt werden konnte, wird ein Test durchgeführt, ob es

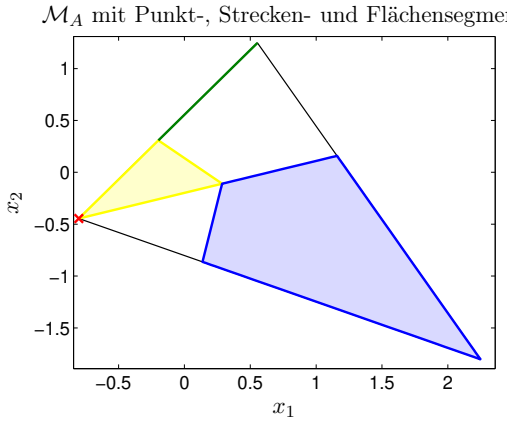


**Abbildung 4.14:** Die Menge zulässiger Lösungen $\mathcal{M}_A$ für die Matrix $D \in \mathbb{R}^{3\times 3}$ mit $D_{ij} = 1$ für $i \geq j$ und $D_{ij} = 0$ für $i < j$ besteht aus je einem Punkt-, Strecken- und Flächensegment. Dazu eingezeichnet sind die Menge $\mathcal{F}_A$ (schwarz) und die Menge $\mathcal{I}_A$ (gelb).

sich um ein Streckensegment handelt. Eine Variante dies umzusetzen ist, direkt nach möglichen Richtungen zu suchen [154].

Zu einem vorgegebenen, kleinen Radius $r > 0$ sowie der initialen Lösung $x^{(0)}$ wird untersucht, ob es einen Winkel $\varphi$ gibt, sodass

$$g_{r,x^{(0)}}(\varphi) = x^{(0)} + r \begin{pmatrix} \sin(\varphi) \\ \cos(\varphi) \end{pmatrix} \tag{4.30}$$

eine zulässige Lösung enthält. Es wird also überprüft, ob der Kreis mit dem Radius $r$ um $x^{(0)}$ einen Schnitt mit $\mathcal{M}_A$ hat. Sofern $r$ klein genug ist, beispielsweise $r = \varepsilon_b$, ist dies für ein Streckensegment, das mindestens die Länge $2r$ hat, der Fall. Kann ein $\varphi_0$ detektiert werden, welches eine geeignete Richtung vorgibt, so werden zu dieser und der entgegengesetzten Richtung die extremalen Punkte auf der Linie zu $x^{(0)}$ und $\varphi_0$ bestimmt. Dies führt auf die zulässige Strecke. Seien $r_l$ in der Richtung zu $\varphi_0 - \pi$ und $r_r$ in der Richtung zu $\varphi_0$ die extremalen Punkte auf der detektierten Linie. Es ergibt sich das Streckensegment

$$\begin{aligned} &\left\{ x \in \mathbb{R}^2 : \; x = x^{(0)} + r \begin{pmatrix} \sin(\varphi_0 - \pi) \\ \cos(\varphi_0 - \pi) \end{pmatrix} \text{ mit } r \in [0,\, r_l] \right\} \\ &\cup \left\{ x \in \mathbb{R}^2 : \; x = x^{(0)} + r \begin{pmatrix} \sin(\varphi_0) \\ \cos(\varphi_0) \end{pmatrix} \text{ mit } r \in [0,\, r_r] \right\}. \end{aligned} \tag{4.31}$$

Führt die Suche mittels (4.30) trotz genügend kleinem $r > 0$ auf kein geeignetes $\varphi$, so ist die initiale Lösung $x^{(0)}$ offenbar in $\mathcal{M}_A$ isoliert.

In Abbildung 4.14 ist $\mathcal{M}_A$ für

$$D = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

dargestellt. Diese besteht aus je einem Punkt-, Strecken- und Flächensegment.

**Bemerkung 4.30.** *Die in Abschnitt 4.4 vorgestellte Methode der Dreieckseinschließung ist in der beschriebenen Form nicht zur Approximation von Streckensegmenten geeignet. Um dies zu ändern muss die Methode jedoch nur leicht modifiziert werden. Analog zu dem oben beschriebenen Ansatz gilt es, eine geeignete Richtung zu finden, wie das Dreieck von der gegebenen Lösung $x^{(0)}$ aus, in Bezug auf das zu approximierende Streckensegment, ausgerichtet sein muss. Bei*

*der Dreieckseinschließungsmethode sind zwei aufeinander folgende Dreiecke stets entgegengesetzt ausgerichtet und jedes Dreieck ist mit seinen beiden übernächsten Nachbarn gleich ausgerichtet. Ist das erste Dreieck geeignet ausgerichtet, so lässt sich ein Streckensegment approximieren. Die Seitenlänge a ist geeignet zu wählen. Ungünstige Fälle wie in Bemerkung 4.20 beschrieben, können nicht auftreten.*

### 4.5.6 Anwendung des inversen Polygon inflation Algorithmus für $s = 2$

Die Berechnung von $\mathcal{M}_A$ für $s = 2$ ist in zahlreichen Veröffentlichungen untersucht [1, 2, 136, 147,154,173]. Einige dieser Methoden arbeiten mit vielen Funktionsauswertungen auf einem Gitter (so genannte *grid search*-Methoden). Aufgrund moderner Computertechnik ist das Problem vieler Funktionsauswertungen nicht schwerwiegend, da, je nach Art der Implementierung, die Klassifizierung eines $x$ direkt geschehen kann und keine rechenintensiven Optimierungen bemüht werden müssen.

Ein numerisch besserer Ansatz basiert auf der Idee des inversen Polygon inflation Algorithmus. Für $s = 2$ gilt $\mathcal{M}_A \subset \mathbb{R}$ und unter den (schwachen) Voraussetzungen aus Satz 3.10 besteht $\mathcal{M}_A$ aus zwei Intervallen (Segmenten). Eines enthält nur negative Einträge, das andere nur positive, denn es gelten $(0, \ldots, 0)^{\mathbf{T}} \in \mathcal{F}_A$ und $(0, \ldots, 0)^{\mathbf{T}} \notin \mathcal{M}_A$, siehe Satz 3.10. Somit lassen sich mit den Suchrichtungen $v_1 = -1$ sowie $v_2 = 1$ Anfangseinschließungen für den Start des Bisektionsverfahrens bestimmen und anschließend Approximationen an die Intervallgrenzen berechnen.

Die beiden Intervalle seien mit $I_1 = [a, b]$ und $I_2 = [c, d]$ bezeichnet. Zur Berechnung des äußeren Randes ($a$ und $d$) wird die Klassifizierung 4.24 genutzt. Da $\mathcal{F}_A$ beschränkt ist, siehe Satz 3.8, sind $a$ und $d$ endlich. Weiter ist mit $x = 0$ ein Element aus $\mathcal{F}_A$ bekannt und es lassen sich jeweils Einschließungen für $a$ beziehungsweise $d$ bestimmen, welche mittels des Bisektionsverfahrens bis zu einer vorgegebenen Genauigkeit präzisiert werden können.

Zur Vereinfachung der Berechnung des inneren Randes, also $b$ und $c$, wird das Dualitätsprinzip angewendet. Aus

$$T = \begin{pmatrix} 1 & x \\ 1 & S \end{pmatrix}$$

ergibt sich mit $xS < 0$, nach Satz 3.23 ist dies für ein zulässiges Paar $(x, S)$ notwendig, dass

$$\frac{1}{S - x} \begin{pmatrix} S & -x \\ -1 & 1 \end{pmatrix}.$$

Somit hat $S$, in Bezug auf das durch $x$ bestimmte Profil $C(:, 2)$, nur skalierenden Einfluss. Sofern wie in (4.1) mit relativen Einträgen gearbeitet wird, entfällt der Einfluss ganz. Zur Berechnung von $b$ und $c$ wird die Zielfunktion

$$\widetilde{f}(x) = \frac{1}{2} \sum_{i=1}^{k} \left( \min \left( 0,\ -\text{sign}(x) \frac{-xU_{i1}\sigma_1 + U_{i2}\sigma_2}{\max\limits_{j=1,\ldots,k} |-xU_{i1}\sigma_1 + U_{i2}\sigma_2|} + \varepsilon_c \right) \right)^2$$

genutzt. Es ist $b$ gleich dem größten Wert $x < 0$, für den $\widetilde{f}(x) \leq \varepsilon_f$ gilt und $c$ ist der kleinste Wert $x > 0$, für den $\widetilde{f}(x) \leq \varepsilon_f$ gilt. (Siehe (4.1) und (4.4) für die Parameter $\varepsilon_c$ und $\varepsilon_f$.) Die Approximationen für die Werte $b$ und $c$ werden, ebenso wie die von $a$ und $d$, mittels des Bisektionsverfahrens und zu einer vorgegebenen Genauigkeit $\varepsilon_b$ berechnet. Als Startpaare sind, nach erfolgreicher Berechnung von $a, d \in \mathcal{M}_A$, beispielsweise $(a, 0)$ beziehungsweise $(0, d)$ geeignet.

**Bemerkung 4.31.** *Auch die Approximation einer isolierten zulässigen Lösung (Punktsegment) gelingt: Da eine isolierte zulässige Lösung $x^*$ nur auf dem Rand von $\mathcal{F}_A$ liegen kann, ist die berechnete Approximation an a beziehungsweise an d gegebenenfalls eine Approximation an $x^*$.*

### 4.5.7 Hybride Methode: Polygon inflation und geometrische Konstruktion

In Abschnitt 4.3.2 ist eine Methode zur simultanen geometrischen Konstruktion von $\mathcal{M}_A$ und $\mathcal{M}_C$ für $s = 3$ erläutert. Diese basiert auf den Mengen $\mathcal{I}_A$, $\mathcal{F}_A$ und $\mathcal{F}_C$ und ist nur für ungestörte Daten ausgelegt. Eine Erweiterung für störungsbehaftete Daten ist in [153] vorgestellt und wird im Folgenden kurz erläutert.

In Bemerkung 3.53 ist erwähnt, dass sich zeigen lässt, dass jeweils eine Ecke von $\mathcal{F}_A$ zu einer Randfläche von $\mathcal{I}_C$ dual ist (Erweiterung zu Satz 3.51). Dies ermöglicht die Konstruktion einer hybriden Methode. Zunächst werden Approximationen für $\mathcal{F}_A$ und $\mathcal{F}_C$ unter Berücksichtigung von Störungen bestimmt. Die Mengen sollten mit hoher Randgenauigkeit, etwa $\varepsilon_b \leq 10^{-6}$, bestimmt werden. Störungen sind in der Form eingebunden, dass (4.5) für alle Eckpunkte gilt. Zu den einzelnen Ecken der Approximation für $\mathcal{F}_C$ lassen sich anschließend die dualen Geraden bestimmen. Mittels derer kann eine Approximation für $\mathcal{I}_A$ unter der Berücksichtigung von Störungen bestimmt werden. Abschließend wird die in Abschnitt 4.3.2 beschriebene simultane geometrische Konstruktion für die berechneten Approximationen an $\mathcal{I}_A$, $\mathcal{F}_A$ und $\mathcal{F}_C$ durchgeführt [153].

In der Approximation für $\mathcal{F}_C$ sind Störungen für den Faktor $C$ direkt berücksichtigt. Durch die Bestimmung einer Approximation für $\mathcal{I}_A$ mittels der Approximation für $\mathcal{F}_C$ werden diese indirekt weitergegeben. So sind die in (4.1) geforderten Bedingungen für $C$ bei der Approximation von $\mathcal{I}_A$ berücksichtigt. Wegen der direkten Approximation von $\mathcal{F}_A$ gelten sie für $A$ ebenso. Unter Nutzung dieser Approximationen an $\mathcal{F}_A$, $\mathcal{I}_A$ und $\mathcal{F}_C$ ist die in Abschnitt 4.3.2 vorgestellte Methode auch für störungsbehaftete Daten einsetzbar und führt qualitativ auf die gleichen Resultate wie die Polygon inflation Methode. Siehe [153] für mehr Details zur Umsetzung dieser hybriden Methode.

**Bemerkung 4.32.**

1. *Für den ungestörten Fall werden somit zunächst $\mathcal{I}_C$ und $\mathcal{I}_A$ berechnet und daraus anschließend mittels des Dualitätsprinzips $\mathcal{F}_A$ und $\mathcal{F}_C$ bestimmt. Für den gestörten Fall ist es genau umgekehrt. Es werden zunächst Approximationen für $\mathcal{F}_A$ und $\mathcal{F}_C$ unter Berücksichtigung von Störungen berechnet und anschließend werden daraus mittels des Dualitätsprinzips Approximationen an $\mathcal{I}_C$ und $\mathcal{I}_A$ bestimmt.*
2. *Nicht konvexe Mengen $\mathcal{F}_A$ oder $\mathcal{F}_C$ können zu Problemen bei der Anwendung der geometrischen Konstruktion von $\mathcal{M}_A$ führen. Daher ist es ratsam, in solchen Fällen die konvexe Hülle der numerisch bestimmten Eckpunkte für $\mathcal{F}_A$ beziehungsweise $\mathcal{F}_C$ zu bilden und damit die Rechnung fortzufahren.*

## 4.6 Inverser Polyhedron inflation Algorithmus für $s = 4$

Die Idee der Polygon inflation Algorithmen ist auf $s = 4$ erweiterbar. Die Menge $\mathcal{M}_A$ ist für $s = 4$ eine Teilmenge von $\mathbb{R}^3$ und die Erweiterung führt auf die Polyhedron inflation Methode. Das Prinzip der Methode ist es, die Oberfläche von $\mathcal{M}_A$ mittels einer geeigneten Triangulierung zu approximieren und die Verfeinerung der Triangulierung adaptiv zu steuern. Beginnend mit einem Starttetraeder, dessen Ecken hinreichend dicht am Rand von $\mathcal{M}_A$ liegen, werden die Kanten der einzelnen Dreiecke adaptiv unterteilt und die Triangulierung wird schrittweise verfeinert. Die algorithmische Umsetzung ist an einigen Stellen problematisch. Aufgrund dessen und da der Fokus dieser Schrift nicht auf dem Polyhedron inflation Algorithmus liegt, wird dieser hier nur kurz erläutert und es wird auf einige Schwierigkeiten hingewiesen. Die Approximation des Randes von $\mathcal{M}_A$ gelingt nur mit der inversen Variante, wobei $\mathcal{F}_A$ und $\mathcal{M}_A^*$ separat bestimmt werden, stabil. Für genaue Erläuterungen und Untersuchungen über die Verfeinerungsschritte, den Umgang mit *regulären* und *irregulären* Unterteilungen sowie die Wahl geeigneter Suchrichtungen wird auf [131] verwiesen.

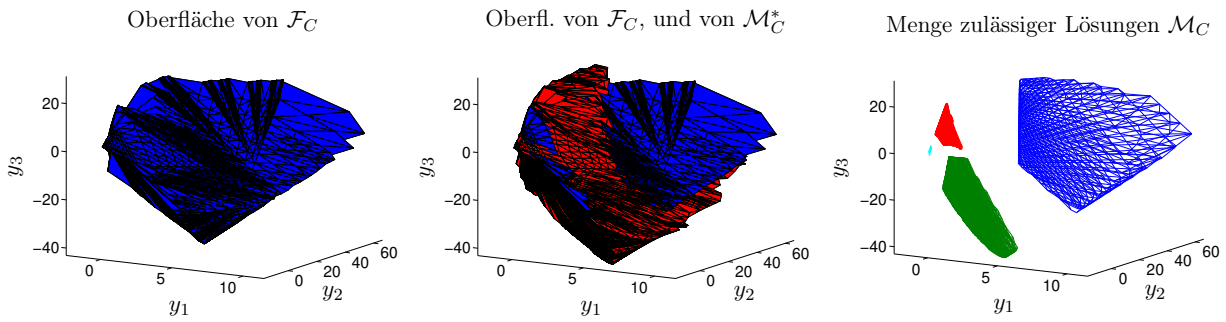


**Abbildung 4.15:** Demonstration der Arbeitsweise des inversen Polyhedron inflation Algorithmus zur Berechnung der dreidimensionalen Menge zulässiger Lösungen $\mathcal{M}_C$ für Datensatz 4. Links: Dargestellt ist eine Oberflächenapproximation von $\mathcal{F}_C$, berechnet mittels des Polyhedron inflation Verfahrens. Mitte: Dargestellt sind die Oberflächenapproximationen von $\mathcal{F}_C$ (blau) und von $\mathcal{M}_C^*$ (rot, die Definition von $\mathcal{M}_C^*$ erfolgt analog zu der von $\mathcal{M}_A^*$ aus (4.23)), beide mittels des Polyhedron inflation Verfahrens berechnet. Rechts: Als letzter Schritt würde $\mathcal{M}_C = \mathcal{F}_C \cap \mathcal{M}_C^*$ berechnet werden. Das hier präsentierte Ergebnis wurde mit der in Abschnitt 4.7 vorgestellten Methode erzeugt.

Hauptschwierigkeit bei der Entwicklung des Polyhedron inflation Algorithmus ist die Wahl der Suchrichtung für die Bestimmung einer neuen Unterteilung einer bestehenden Kante. Sofern die Oberfläche eines isolierten Segments von innen heraus approximiert werden soll, entstehen oft „einspringende Kanten“. Dadurch wird eine adaptive Steuerung und eine genaue Approximation be- oder sogar verhindert. Ein weiteres Problem ist die Approximation von Segmenten, die konkave Oberflächenbereiche enthalten.

Die einzige Möglichkeit diese beiden Schwierigkeiten zu umgehen, ist es, $\mathcal{M}_A$ als Schnitt von $\mathcal{F}_A$ und $\mathcal{M}_A^*$ zu bestimmen [131]. Bei der Verfeinerung einer Kante wird diese durch zwei neue ersetzt indem ihr Mittelpunkt $M$ in sinnvoller Richtung an den Rand von $\mathcal{F}_A$ beziehungsweise $\mathcal{M}_A^*$ verschoben wird. Als Suchrichtung für den neuen Punkt eignet sich nur die Richtung vom Nullpunkt zu $M$. Der Vorteil des Nullpunktes in Bezug auf stabile Berechnungen ist, dass $(0,\ldots,0)^{\mathbf{T}} \in \mathcal{F}_A$ und $(0,\ldots,0)^{\mathbf{T}} \notin \mathcal{M}_A^*$ gelten. Dies hat den willkommenen Nebeneffekt, dass die Menge $\mathcal{M}_A$ unabhängig von deren Topologie stabil und gut approximiert wird.

Die letzte Schwierigkeit, die bisher aufgrund von Zeitmangel die Fertigstellung und Veröffentlichung der Methode in FacPack verhindert, ist die Bestimmung einer Triangulierung der Oberfläche von $\mathcal{M}_A$ anhand von Triangulierungen der Oberflächen von $\mathcal{F}_A$ und $\mathcal{M}_A^*$. Hierzu sei weiterführend auf [73] verwiesen.

In Abbildung 4.15 sind Approximationen für die Oberflächen der Mengen $\mathcal{F}_C$ und $\mathcal{M}_C^*$ (die Definition von $\mathcal{M}_C^*$ erfolgt analog zu der von $\mathcal{M}_A^*$ aus (4.23) nur für den Faktor $C$) für den Datensatz 4 dargestellt. Diese wurden mit dem Polyhedron inflation Algorithmus bestimmt. Der letzte Schritt, die Bestimmung von $\mathcal{F}_A \cap \mathcal{M}_A^*$, ist nicht vollzogen. Die abgebildete Approximation für $\mathcal{M}_C$ wurde mit der in Abschnitt 4.7 vorgestellten Methode berechnet.

## 4.7 Strahlenmethode für $s \geq 2$

Mit der geometrischen Konstruktion (Abschnitt 4.3), der Dreieckseinschließungsmethode (Abschnitt 4.4) sowie den Polygon inflation Algorithmen (Abschnitt 4.5) wurden bereits drei Methoden zur Berechnung der Mengen zulässiger Lösungen für $s = 3$ vorgestellt. In diesem Abschnitt wird eine Methode zur Approximation der Mengen zulässiger Lösungen für beliebiges $s \geq 2$ vorgestellt [157]. Dazu werden vom Ursprung ausgehende Strahlen im $\mathbb{R}^{s-1}$ betrachtet. Für die einzelnen Strahlen wird untersucht, ob Teile dieser zur Menge zulässiger Lösungen $\mathcal{M}_A$ gehören. Für einen einzelnen Strahl gilt, dass, wenn er $\mathcal{M}_A$ schneidet, die Schnittmenge ein Geradenab-

schnitt ist. Für einen Strahl, der $\mathcal{M}_A$ schneidet, sind nur die extremalen zulässigen Lösungen auf diesem Strahl interessant. Sofern die Strahlen geeignet gewählt sind, lässt sich die Oberfläche der Menge zulässiger Lösungen durch die extremalen zulässigen Lösungen gut approximieren.

Die Güte der Approximation wird insbesondere durch die Wahl der Strahlen bestimmt. Diese lassen sich nach einer ersten Approximation adaptiv verfeinern, sodass kritische Bereiche durch zusätzliche Strahlen feiner aufgelöst werden.

Die Methode ist einfach und robust. Es können Mengen zulässiger Lösungen jeglicher Struktur und Anzahl von Segmenten bestimmt werden. Für $q$-dimensionale affine Unterräume mit $q \leq s-3$ müssen zusätzliche Berechnungen durchgeführt werden und für $(s-2)$-dimensionale affine Unterräume verspricht eine zusätzliche Verfeinerung eine bessere Approximation. Eine Implementierung ist in FacPack zugänglich.

### 4.7.1 Idee der Methode

Die Entwicklung der Strahlenmethode zur Approximation der Menge zulässiger Lösungen $\mathcal{M}_A$ basiert auf drei wichtigen Eigenschaften von $\mathcal{M}_A$ beziehungsweise $\mathcal{F}_A$:

- Die Menge $\mathcal{M}_A$ ist beschränkt, siehe Satz 3.8.
- Der Nullpunkt ist in $\mathcal{F}_A$ enthalten aber nicht in $\mathcal{M}_A$, siehe Satz 3.10.
- Für ein $x \in \mathcal{M}_A$ folgt aus $\gamma \geq 1$ und $\gamma x \in \mathcal{F}_A$, dass $\gamma x \in \mathcal{M}_A$, siehe Satz 3.15.

Sei zu einem $v \in \mathbb{R}^{s-1} \setminus \{0\}$ der Strahl

$$\nu = \{\gamma v : \ \gamma \geq 0\}$$

betrachtet. Für $\nu$ gilt, dass der Schnitt $\nu \cap \mathcal{M}_A$ entweder die leere Menge oder ein Geradenabschnitt ist. Sofern $\nu \cap \mathcal{M}_A$ ein Geradenabschnitt ist, sind nur die extremalen zulässigen Lösungen zu berechnen. Eine dieser liegt auf dem Rand von $\mathcal{F}_A$ und ist leicht zugänglich. Weiter ist der Schnitt $\nu \cap \mathcal{M}_A$ genau dann leer, wenn $\nu \cap \partial\mathcal{F}_A$ nicht zu $\mathcal{M}_A$ gehört, siehe Korollar 3.18.

### 4.7.2 Notation

Zunächst wird die genutzt Notation eingeführt. Die Anzahl der Strahlen ist $N$. Festgelegt werden die Strahlen $\nu_i$, $i = 1, \ldots, N$, durch die Vektoren $v_i \in \mathbb{R}^{s-1} \setminus \{0\}$ als

$$\nu_i = \{\gamma v_i : \ \gamma \geq 0\}.$$

Ist der Schnitt $\nu_i \cap \mathcal{M}_A$ nicht leer, so werden die extremalen Radien mit $r_i$ (minimaler Abstand zum Ursprung) und $R_i$ (maximaler Abstand zum Ursprung) bestimmt, sodass

$$r_i = \min\{\gamma > 0 : \ \gamma v_i \in \mathcal{M}_A\}, \qquad R_i = \max\{\gamma > 0 : \ \gamma v_i \in \mathcal{M}_A\}. \tag{4.32}$$

Sofern $\nu_i \cap \mathcal{M}_A \neq \emptyset$ gilt, ist $R_i v_i$ der Schnitt von $\nu_i$ mit dem Rand von $\mathcal{F}_A$. Begrenzt durch die berechneten Stellen ist folglich der Strahlenabschnitt

$$\{\gamma v_i : \ r_i \leq \gamma \leq R_i\} \tag{4.33}$$

eine Teilmenge von $\mathcal{M}_A$. Die Menge aller berechneten inneren Randpunkte wird mit $\mathcal{R}^{\text{in}}$ bezeichnet und die Menge aller ermittelten äußeren Randpunkte mit $\mathcal{R}^{\text{out}}$.

### 4.7.3 Algorithmischer Ablauf

Die Schritte des Strahlenalgorithmus sind:

1. Zunächst werden die $N$ Strahlen $\nu_i$ durch die Richtungen $v_i \in \mathbb{R}^{s-1}$ definiert. Die $v_i$ werden durch Polarkoordinaten festgelegt.
2. Für jeden Strahl wird $\delta_i = \max\{\gamma > 0 : \ \gamma v_i \in \mathcal{F}_A\}$ berechnet.
3. Für jeden Strahl wird getestet, ob $\delta_i v_i \in \mathcal{M}_A$. Falls ja, so werden $R_i = \delta_i$ gesetzt, $r_i = \min\{\gamma > 0 : \ \gamma v_i \in \mathcal{M}_A\}$ berechnet und $\mathcal{R}^{\text{out}}$ um $R_i v_i$ sowie $\mathcal{R}^{\text{in}}$ um $r_i v_i$ erweitert.
4. Abschließend wird/werden die Oberfläche/Oberflächen des/der einzelnen Segments/Segmente von $\mathcal{M}_A$ aus den Elementen der Mengen $\mathcal{R}^{\text{out}}$ und $\mathcal{R}^{\text{in}}$ zusammengesetzt.

Die Strahlen werden systematisch durchlaufen. Die Entscheidung, ob ein Strahl $\mathcal{M}_A$ schneidet, kann sicher mit *ja* aber nicht sicher mit *nein* beantwortet werden, siehe Bemerkung 4.3 und den Entscheidungsbaum aus Abbildung 4.3. Falls ein Strahl $\mathcal{M}_A$ scheidet, kann es vorkommen, dass der berechnete Radius $r_i$ nicht die Extremalbedingung aus (4.32) erfüllt. Das Problem ist hierbei ebenso, dass die Entscheidung, ob ein $x \in \mathcal{F}_A$ nicht zu $\mathcal{M}_A$ gehört, nur implizit unter Nutzung der Optimierung aus (4.4) getroffen werden kann. Beiden beschriebenen Fällen liegt dieselbe Schwierigkeit zugrunde und es gilt, Fehlentscheidungen zu vermeiden. Um die Chancen, eine korrekte Entscheidung getroffen zu haben, zu erhöhen, ist es gerechtfertigt, einen erhöhten Rechenaufwand in Kauf zu nehmen. In der Implementierung werden für alle Strahlen, die $\mathcal{M}_A$ schneiden, die optimalen Untermatrizen $S_i^* \in \mathbb{R}^{(s-1)\times(s-1)}$ gespeichert. Es ist

$$S_i^* = \operatorname*{argmin}_{S \in \mathbb{R}^{(s-1)\times(s-1)}} f(r_i v_i, S)$$

mit $f(x, S)$ aus (4.2). Anschließend werden die einzelnen $S_i^*$ genutzt, um die Approximationen auf den jeweils benachbarten Strahlen dahingehend zu überprüfen, ob Verbesserungen möglich beziehungsweise die Schnitte mit $\mathcal{M}_A$ doch nicht leer sind. Dabei werden neue Optimierungen mit den $S_i^*$ benachbarter Strahlen als Startwerte durchgeführt. Um hier die Möglichkeiten von Fehlentscheidungen zu reduzieren, werden die einzelnen Strahlen der Reihenfolge nach mehrfach und in verschiedenen Richtungen durchlaufen.

### 4.7.4 Berechnung der extremalen zulässigen Lösungen

Der Test, ob $\nu_i \cap \mathcal{M}_A \neq \emptyset$ gilt oder nicht, lässt sich mittels der Funktion $f(x, S)$ aus (4.2) implizit durchführen. Nach Korollar 3.18 enthält ein Strahl genau dann eine zulässige Lösung, wenn sein Schnitt mit dem Rand von $\mathcal{F}_A$ zulässig ist. Somit wird zunächst der Schnittpunkt von $\nu_i$ mit $\partial\mathcal{F}_A$ approximiert. Dazu wird der maximale Radius $\delta_i > 0$ gesucht, sodass $x = \delta_i v_i$ die Ungleichung aus (4.5) erfüllt. Für die Praxis ist es mitunter sinnvoll, als obere Schranke nicht das in (4.5) genutzte $\varepsilon_f$, sondern einen kleineren positiven Wert zu verwenden. Verglichen mit der optionalen Berechnung von $r_i$ ist die Bestimmung von $\delta_i$ mit sehr geringem Rechenaufwand verbunden. Anschließend wird untersucht, ob $\delta_i v_i$ zu $\mathcal{M}_A$ gehört oder nicht, also ob

$$\min_{S \in \mathbb{R}^{(s-1)\times(s-1)}} f(\delta_i v_i, S) \leq \varepsilon_f$$

gilt oder nicht. Sofern $\nu_i \cap \mathcal{M}_A \neq \emptyset$ gilt, wird $R_i = \delta_i$ gesetzt. Anschließend wird

$$r_i = \min\left\{\gamma > 0 : \ \min_{S \in \mathbb{R}^{(s-1)\times(s-1)}} f(\gamma v_i, S) \leq \varepsilon_f\right\}$$

berechnet. Diese Berechnung erfolgt zur Genauigkeit $\varepsilon_b > 0$ und es gelten

$$r_i v_i \in \mathcal{M}_A \quad \text{und} \quad (r_i - \varepsilon_b) v_i \notin \mathcal{M}_A. \tag{4.34}$$

### 4.7.5 Genauigkeit und Auflösung

Die Genauigkeiten der Approximationen $r_i$ und $R_i$ werden durch den Parameter $\varepsilon_b > 0$ gesteuert. Weiter wird die Gitterweite für die Oberfläche der einzelnen Segmente durch die Anzahl $N$ der Strahlen gesteuert. In der Regel ist der seitliche Abstand zweier benachbarter Approximationen $R_i v_i$ und $R_j v_j$ deutlich größer als die Genauigkeit $\varepsilon_b$, weshalb die Wahl von $N$ für die Bestimmung einer fein aufgelösten Oberflächenapproximation entscheidend ist.

Zwischen benachbarten Approximationen $R_i v_i$ und $R_j v_j$ beziehungsweise $r_i v_i$ und $r_j v_j$ erfolgt eine lineare Interpolation. Ebenso wird beim seitlichen Abschluss durch die Verbindung von $R_i v_i$ und $r_i v_i$ linear interpoliert. In der Anwendung passiert es insbesondere bei der letztgenannten Situation oft, dass randnahe Bereiche von $\mathcal{M}_A$ durch die lineare Interpolation „abgeschnitten“ werden. Um diesem Effekt entgegenzuwirken, kann eine adaptive Verfeinerung der Auflösung für kritische Bereiche eingesetzt werden.

### 4.7.6 Die Fälle $s = 2$ und $s = 3$

Die Fälle $s = 2$ und $s = 3$ bieten die Möglichkeit, die Methode anschaulich zu erklären sowie die damit erzielten Approximationsgüten zu verifizieren. Für $s = 2$ gilt $\mathcal{M}_A \subset \mathbb{R}$ und die Approximation der Intervallgrenzen erfolgt mittels der $N = 2$ Strahlen zu $v_1 = -1$ und $v_2 = 1$. Es ergeben sich $\mathcal{R}^{\text{in}} = \{-r_1, r_2\}$ und $\mathcal{R}^{\text{out}} = \{-R_1, R_2\}$. Dies führt auf zwei disjunkte Teilintervalle (Segmente), sodass

$$\mathcal{M}_A = [-R_1, -r_1] \cup [r_2, R_2].$$

**Bemerkung 4.33.** *Die Anwendung des Strahlenalgorithmus für $s = 2$ stimmt mit der in Abschnitt 4.5.6 vorgestellten Variante zur Berechnung von $\mathcal{M}_A$ auf Basis der Idee des inversen Polygon inflation Algorithmus überein. Es sind $a = -R_1$, $b = -r_1$, $c = r_2$ und $d = R_2$.*

Für $s = 3$ ist $\mathcal{M}_A \subset \mathbb{R}^2$. Bei der Anwendung des Strahlenalgorithmus zur Approximation von $\mathcal{M}_A$ unter Nutzung von $N$ äquiangular verteilten Strahlen ergeben sich die Richtungen

$$v_i = \begin{pmatrix} \cos(\phi_i + \phi_0) \\ \sin(\phi_i + \phi_0) \end{pmatrix} \quad \text{mit} \quad \phi_i = 2\pi \frac{i-1}{N}, \quad i = 1, \ldots, N,$$

zu einem Anfangswinkel $\phi_0$. Dieser Anfangswinkel kann so gewählt werden, dass $v_1$ keinen leeren Schnitt mit $\mathcal{M}_A$ hat. Beispielsweise lässt sich dies unter Nutzung einer zuvor berechneten nichtnegativen Matrixfaktorisierung von $D$ erreichen. Außer eines strukturellen Vorteils bei der Implementierung gibt es nur den weiteren Vorteil, dass, falls das zu $v_1$ gehörige Segment ein Punktsegment ist, dieses sicher detektiert wird. In Abbildung 4.16 ist die Anwendung des Strahlenalgorithmus zur Approximation des Randes von $\mathcal{M}_A$ für Datensatz 2 dargestellt.

**Bemerkung 4.34.**

1. *Sofern $\mathcal{M}_A$ nicht nur aus einem Segment besteht, ist es sinnvoll, auch für die beiden nicht zu $\phi_0$ gehörigen zulässigen Lösungen einer initialen nichtnegativen Matrixfaktorisierung zu untersuchen, ob sie isoliert sind.*
2. *Es lassen sich, sofern vorhanden, Streckensegmente im Nachhinein verfeinern, im Regelfall werden diese aber durch die Methode hinreichend gut detektiert.*

### 4.7.7 Der Fall $s = 4$

Für $s = 4$ sind $\mathcal{M}_A, \mathcal{M}_C \subset \mathbb{R}^3$. Die Strahlen sind mittels zweier Winkel bestimmt. Um eine Oberflächentriangulierung zu erleichtern wird für diese kein Rechteckgitter gewählt, sondern

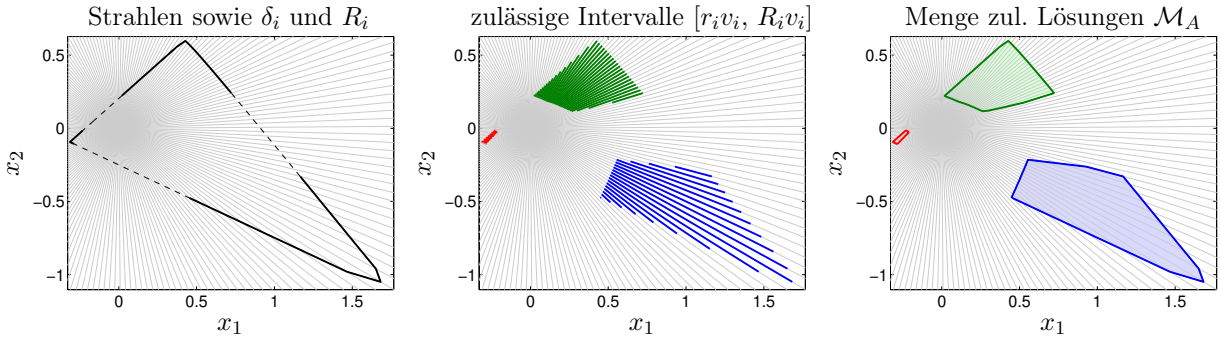


**Abbildung 4.16:** Approximation von $\mathcal{M}_A$ mit dem Strahlenalgorithmus für $N = 200$ Strahlen am Beispiel des Datensatzes 2. Links: Dargestellt sind die äquiangular verteilten Strahlen $\nu_i$ (grau), deren verbundene Schnittpunkte $\delta_i v_i$ mit dem Rand von $\mathcal{F}_A$ (gestrichelte schwarze Linie) und die verbundenen Schnittpunkte $R_i v_i$ mit dem äußeren Rand von $\mathcal{M}_A$ (schwarze Linien). Mitte: Dargestellt sind die berechneten (zulässigen) Intervalle zwischen den äußeren Randpunkten $R_i v_i$ und den inneren Randpunkten $r_i v_i$. Rechts: Das Verbinden der Punkte aus $\mathcal{R}^{\text{out}}$ und $\mathcal{R}^{\text{in}}$ führt auf die Ränder der einzelnen Segmente.

eines, welches von einer Schicht zur nächsten leicht versetzt ist. Weiter werden die vier Richtungen, die sich zu einer zuvor bestimmten, initialen nichtnegativen Faktorisierung $D = C^{(\text{init})} A^{(\text{init})}$ ergeben, ebenfalls untersucht, um mögliche Punkt- oder Streckensegmente zu detektieren.

Mittels $A^{(\text{init})}$ seien $(\varphi_i, \psi_i) \in [0, \pi] \times [-\pi, \pi)$ für $i = 1, \ldots, 4$ so bestimmt, dass

$$\begin{pmatrix} \sin\varphi_i \cos\psi_i \\ \sin\varphi_i \sin\psi_i \\ \cos\varphi_i \end{pmatrix} = \left(T_{i2}^2 + T_{i3}^2 + T_{i4}^2\right)^{-\frac{1}{2}} \begin{pmatrix} T_{i2} \\ T_{i3} \\ T_{i4} \end{pmatrix}$$

mit $T = A^{(\text{init})} V$. Weiter seien $(\varphi_5, \psi_5) = (0, 0)$ und $(\varphi_6, \psi_6) = (0, \pi)$ sowie

$$\varphi_{6+i+j\ell_1} = \begin{cases} (2\frac{i}{\ell_1} - 1)\pi, & \text{falls } j \text{ gerade,} \\ (\frac{2i+1}{\ell_1} - 1)\pi, & \text{sonst} \end{cases}$$

und

$$\psi_{6+i+j\ell_1} = \frac{j+1}{\ell_2 + 1}\pi$$

für $i = 0, \ldots, \ell_1 - 1$ und $j = 0, \ldots, \ell_2 - 1$. Daraus ergeben sich die $N = 6 + \ell_1\ell_2$ Richtungen

$$v_i = \begin{pmatrix} \sin\varphi_i \cos\psi_i \\ \sin\varphi_i \sin\psi_i \\ \cos\varphi_i \end{pmatrix},$$

welche zur Approximation des Randes von $\mathcal{M}_A$ genutzt werden.[7] In Abbildung 4.17 wird die Strahlenmethode zur Berechnung von $\mathcal{M}_C$ für Datensatz 4 anhand von $N = 20\,006$ Strahlen demonstriert.

### 4.7.8 Die Fälle $s \geq 5$

Die uneingeschränkte Anwendbarkeit für $s \geq 2$ ist ein großer Vorteil des Strahlenalgorithmus, wenngleich der Rechenaufwand exponentiell bezüglich $s$ steigt. Bei der Anwendung für $s \geq 5$

[7] In FacPack [146] wird aktuell noch eine andere Vorgehensweise bei der Wahl der $v_i$ genutzt.

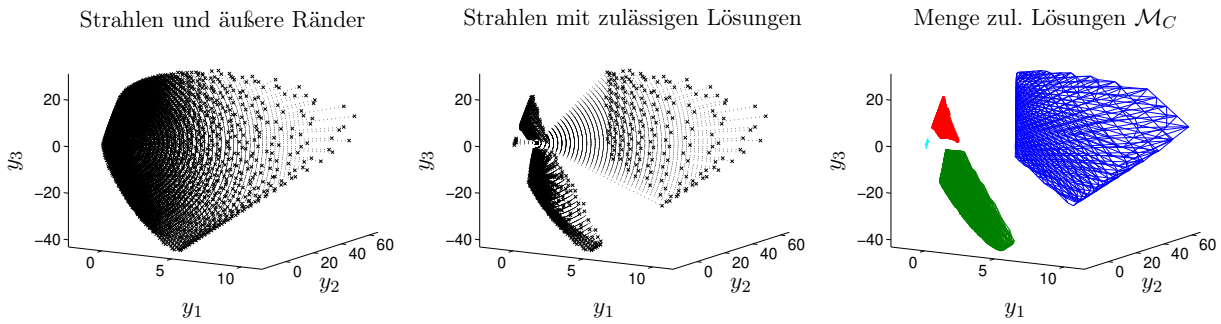


**Abbildung 4.17:** Approximation der Menge $\mathcal{M}_C$ mit dem Strahlenalgorithmus für $N = 20\,006$ Strahlen am Beispiel des Datensatzes 4. Links: Zunächst werden die Werte $\delta_i$ berechnet. Es ist nur jeder vierte der Strahlen $\nu_i$ dargestellt. Mitte: Danach wird überprüft, auf welchen Strahlen zulässige Lösungen liegen. Erneut ist nur jeder vierte der zulässigen Strahlen dargestellt. Rechts: Abschließend werden die inneren Randpunkte berechnet und die Ergebnisse aus $\mathcal{R}^{\text{out}}$ und $\mathcal{R}^{\text{in}}$ zu den Segmentoberflächen verbunden. Bemerkung: Die relativ hohen Werte für die drei Koordinaten in den Darstellungen ergeben sich aufgrund der Singulärwerte, welche für $\mathcal{M}_C$ mit einbezogen werden.

ergeben sich die Richtungen $v_i \in \mathbb{R}^{s-1}$ als

$$v_i = \begin{pmatrix} \cos(\phi_{i,1}) \\ \sin(\phi_{i,1})\cos(\phi_{i,2}) \\ \sin(\phi_{i,1})\sin(\phi_{i,2})\cos(\phi_{i,3}) \\ \vdots \\ \sin(\phi_{i,1}) \cdot \ldots \cdot \sin(\phi_{i,s-3})\cos(\phi_{i,s-2}) \\ \sin(\phi_{i,1}) \cdot \ldots \cdot \sin(\phi_{i,s-3})\sin(\phi_{i,s-2}) \end{pmatrix}$$

mit der Diskretisierung

$$\phi_{i,s-2} = 2\pi \frac{i-1}{\ell_{s-2}}, \qquad i = 1, \ldots, \ell_{s-2},$$

für den letzten Winkel $\phi_{i,s-2}$ sowie den weiteren Winkeldiskretisierungen

$$\phi_{i,j} = \pi \frac{i-1}{\ell_j}, \qquad i = 1, \ldots, \ell_j,$$

für $j = 1, \ldots, s-3$. Sofern auch hier die Richtungen einer initialen nichtnegativen Matrixfaktorisierung genutzt werden, ergeben sich insgesamt $N = s + \prod_{i=1}^{s-2} \ell_i$ zu untersuchende Strahlen.

### 4.7.9 Adaptive Verfeinerung

Die Umsetzung der Strahlenmethode ohne adaptive Steuerung führt im Allgemeinen zu dem Problem, dass eckenähnliche Abschnitte des Randes von $\mathcal{M}_A$ nicht genügend fein aufgelöst werden, sofern die Strahlen gleichmäßig (für $s = 3$ äquiangular) verteilt sind. Die Implementierung einer adaptiven Steuerung für die Strahlenmethode ist für $s = 3$ einfach und wird in diesem Abschnitt vorgestellt. Für $s = 4$ ergibt sich die Schwierigkeit, eine sinnvolle Oberflächentriangulierung ohne spitzwinkelige Dreiecke und einspringende Ecken zu erzeugen um eine gute Darstellung zu ermöglichen.

#### Einfache adaptive Steuerung für $s = 3$

In diesem Teilabschnitt wird eine simple Variante der adaptiven Steuerung für den Strahlenalgorithmus in der Ebene ($s = 3$) vorgestellt. Der Strahlenalgorithmus wird zunächst mit einem

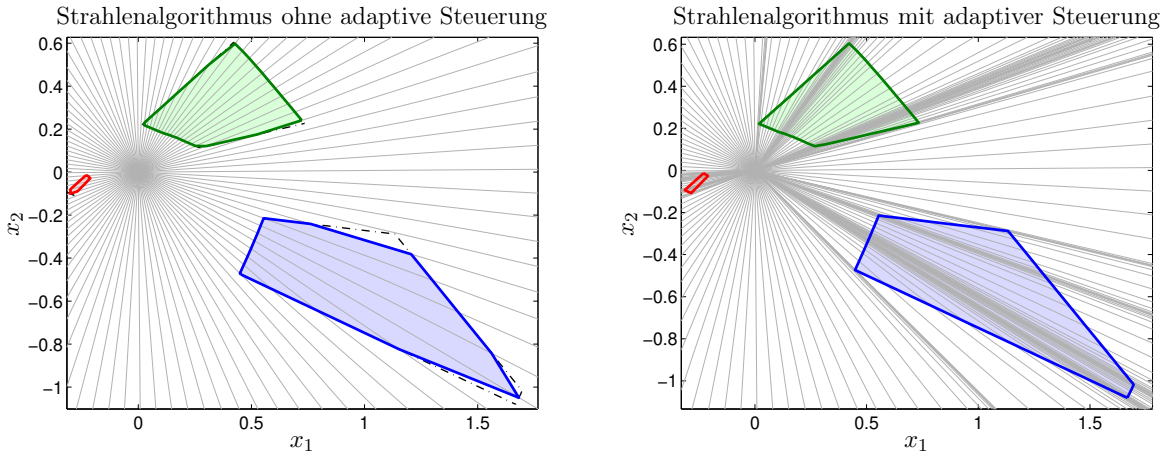


**Abbildung 4.18:** Strahlenalgorithmus mit (rechts) und ohne (links, $N = 100$) adaptive(r) Steuerung bei der Anwendung auf Datensatz 2. Links: Ohne adaptive Steuerung ist die Approximation der eigentlichen Lösung (schwarze gestrichelt-gepunktete Linie) an manchen Stellen verbesserungswürdig. Rechts: Die adaptive Steuerung führt zu einer deutlich verbesserten Approximation. Die Ausgangsbasis ist $N_{\text{basic}} = 100$ und insgesamt werden $N = 243$ Strahlen zur Approximation genutzt. Gut zu erkennen ist, dass jeweils in den Bereichen, in denen der innere oder der äußere Rand eckenähnlich verläuft, die Strahlendichte deutlich erhöht ist.

Grundsatz von $N_{\text{basic}}$ Strahlen ausgeführt und in einem anschließenden Schritt werden einzelne Strahlen zur Verfeinerung hinzugefügt.

Sei $N$ stets die aktuelle Gesamtanzahl von Strahlen. Zwei Kriterien werden zur Entscheidung herangezogen, ob zwischen zwei Nachbarstrahlen $\nu_i$ und $\nu_{\text{mod}(i+1,N)}$ verfeinert wird oder nicht. Dabei gilt es zu beachten, zwischen zwei sehr dicht beieinander liegenden Strahlen nicht zu verfeinern. Konkret wird verfeinert, falls

1. ein innerer oder äußerer Randverlauf mehr als eine leichte Abweichung von einem linearen Verhalten aufweist und $\|R_i v_i - R_{\text{mod}(i+1,N)} v_{\text{mod}(i+1,N)}\|_2$ beziehungsweise $\|r_i v_i - r_{\text{mod}(i+1,N)} v_{\text{mod}(i+1,N)}\|_2$ einen Mindestwert nicht unterschreitet oder
2. nur genau einer der beiden Strahlen $\nu_i$ und $\nu_{\text{mod}(i+1,N)}$ einen nichtleeren Schnitt mit $\mathcal{M}_A$ hat, der dazugehörige Unterschied $R_i - r_i$ beziehungsweise $R_{\text{mod}(i+1,N)} - r_{\text{mod}(i+1,N)}$ einen kritischen Wert übersteigt und $\|\delta_i v_i - \delta_{\text{mod}(i+1,N)} v_{\text{mod}(i+1,N)}\|_2$ einen Mindestwert nicht unterschreitet.

Im algorithmischen Ablauf werden alle Paare benachbarter Strahlen untersucht, ob sie eines der beiden genannten Kriterien erfüllen. Gegebenenfalls wird ein neuer Strahl äquiangular zwischen $\nu_i$ und $\nu_{\text{mod}(i+1,N)}$ eingefügt und die Randapproximation so verfeinert. Es wird stets eine monotone Ordnung (im Sinne des Umlaufs der Strahlen) aufrechterhalten, das heißt der neue Strahl wird zwischen $\nu_i$ und $\nu_{\text{mod}(i+1,N)}$ eingeordnet. Somit ist stets gewährleistet, dass für beliebiges $\ell \in \{1, \ldots, N\}$ die Strahlen $\nu_\ell$ und $\nu_{\text{mod}(\ell+1,N)}$ benachbart sind.

Die Verfeinerung erfolgt mittels zweier Schleifen. In der inneren Schleife werden die Paare benachbarter Strahlen monoton durchlaufen und gegebenenfalls wird zwischen ihnen verfeinert. Weiter sind diese Umläufe in einer äußeren Schleife organisiert. Diese wird ausgeführt, solange es mindestens eine Verfeinerung im letzten Umlauf gab.

In Abbildung 4.18 ist die Anwendung der Strahlenmethode mit und ohne adaptive(r) Steuerung für den Datensatz 2 dargestellt. In der rechten Grafik ist die Verfahrensweise der adaptiven Steuerung deutlich erkennbar. Nahezu geradlinige Abschnitte der Randapproximation werden mit nur wenigen Strahlen approximiert. Demgegenüber werden nicht geradlinige Abschnitte (innen oder außen je Strahl) mittels einer hohen Dichte an Strahlen bestimmt. Die Effizienz belegen auch die Tabellen 5.5 und 5.6.

#### Einfache adaptive Steuerung für $s = 4$

Eine adaptive Steuerung des Strahlenalgorithmus für $s = 4$ zu implementieren ist weit komplexer als für $s = 3$. Die Schwierigkeit liegt in der Erzeugung einer geeigneten Oberflächentriangulierung. Die Umsetzung einer adaptiven Verfeinerung der Triangulierung kann in ähnlicher Form erfolgen, wie in [131] für den inversen Polyhedron inflation Algorithmus beschrieben. Zu der in [131] erläuterten Methode gibt es zwei wichtige Unterschiede. Zum einen sind zur Entscheidung, ob verfeinert werden soll, zwei Werte zu berücksichtigen, nämlich sowohl bezüglich der inneren Randpunkte $r_i v_i$ als auch bezüglich der äußeren Randpunkte $R_i v_i$. Verfeinert wird, sofern von beiden Oberflächentriangulierungen mindestens eine zu verfeinern ist. Zum anderen wirkt sich eine Änderung der einen Triangulierung ebenfalls auf die andere aus, vergleiche Abbildung 4.18 für $s = 3$. Die zweite Verfeinerung ist mitunter nicht nötig. Dieser Ansatz wird in der vorliegenden Schrift nicht weiter untersucht, da die Implementierung aufwändig ist und nicht im Fokus steht.

### 4.7.10 Nachiteration möglicherweise nicht exakt bestimmter innerer Randpunkte

Die Bestimmung der inneren Randpunkte $r_i v_i$ erfolgt indirekt und unter Nutzung der Zielfunktion $f(x, S)$ aus (4.2). Zur Klassifizierung eines $x$ wird diese minimiert. Ebenso wie für den Polygon inflation Algorithmus ist diese Minimierung ein kritischer Schritt, siehe Bemerkung 4.3. Es gilt, sowohl ungenaue Approximationen im Sinne von fälschlicherweise zu groß bestimmten Werten $r_i$, als auch mögliche Falschbewertungen zu Strahlen (numerische Ermittlung von $\nu_i \cap \mathcal{M}_A = \emptyset$, wobei tatsächlich aber $\nu_i \cap \mathcal{M}_A \neq \emptyset$ der Fall ist) zu vermeiden. Dazu werden am Ende der Berechnung aller Werte $r_i$ die folgenden zwei Kriterien abgefragt und gegebenenfalls Nachiterationen durchgeführt:

- Es wird überprüft, ob ein berechnetes $r_i v_i$ tatsächlich eine genügend genaue Approximation an einen inneren Randpunkt auf dem Strahl zu $\nu_i$ ist, sofern $r_i \geq r_j$ für einige oder alle Nachbarstrahlen $\nu_j$ gilt.
- Zudem wird überprüft, ob $\nu_i \cap \mathcal{M}_A = \emptyset$ tatsächlich gilt, falls mindestens einer der Nachbarstrahlen ein zulässiges Intervall beinhaltet.

Für die Nachiterationen werden die Optimalstellen der Zielfunktion für die benachbarten Strahlen als Startwerte für den Strahl $\nu_i$ genutzt. Weiter ist in [86] eine Verfeinerungstrategie basierend auf geometrischen Argumenten eingeführt.

### 4.7.11 Stabile Approximationen für alle Topologien und degenerierte Segmente

Die Menge $\mathcal{M}_A$ kann für $s \geq 3$ eine topologisch zusammenhängende Menge sein aber auch aus mehreren einzelnen Segmenten bestehen. Die Segmente können wiederum zu isolierten Lösungen, Strecken (für $s \geq 3$), Teilmengen von Ebenen mit nichtverschwindender Fläche (für $s \geq 4$), und so weiter degeneriert sein. Einzig für $s = 2$ besteht $\mathcal{M}_A$ immer aus zwei getrennten Segmenten. Eine großer Vorteil des Strahlenalgorithmus ist, dass $\mathcal{M}_A$ unabhängig von dessen Topologie stabil berechnet werden kann. Die korrekte Erkennung der Topologie von $\mathcal{M}_A$ ist einzig von der Auflösung abhängig.

Inwiefern beim Strahlenalgorithmus degenerierte Segmente bestimmt werden können, wird im folgenden Teilabschnitt für $s = 4$ kurz beschrieben. Für mehr Details sei auf [157] verwiesen.

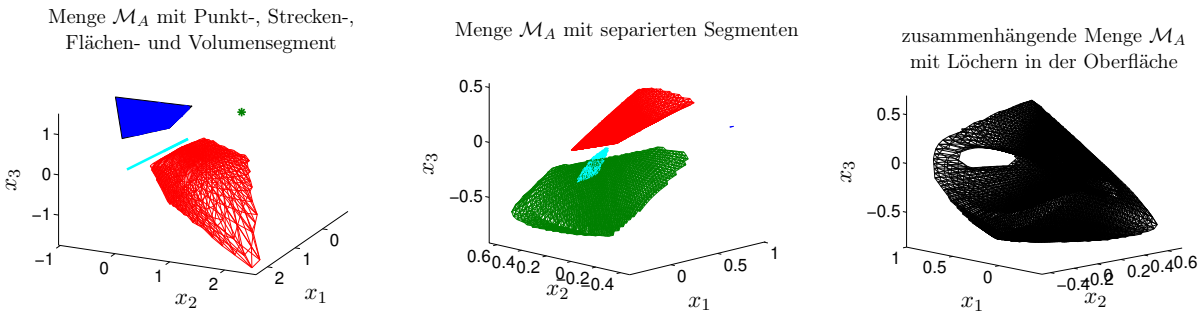


**Abbildung 4.19:** Verschiedene Strukturen für $\mathcal{M}_A$ und $s = 4$. Links: Dargestellt ist $\mathcal{M}_A$ für $D \in \mathbb{R}^{4\times 4}$ mit $D_{ij} = 1$ für $i \geq j$ und $D_{ij} = 0$ sonst. Es besteht $\mathcal{M}_A$ aus je einem Punkt-, Strecken-, Flächen- und Volumensegment. Mitte: Dargestellt ist $\mathcal{M}_A$ für Datensatz 4. Rechts: Für $\widehat{D} = D + 0.03$ mit $D$ aus Datensatz 4 besteht $\mathcal{M}_A$ aus einem Segment mit drei Löchern in der Oberfläche die sich in der Mitte vereinen, wobei in der dargebotenen 3D-Darstellung nicht alle Löcher zu erkennen sind.

### Approximation von degenerierten Segmenten für $s = 4$

Für $s = 4$ gibt es, neben dem Normalfall eines Segments mit positivem Volumen, drei Arten von degenerierten Segmenten: Punkt-, Strecken- und Flächensegmente. In der linken Grafik von Abbildung 4.19 ist eine Menge $\mathcal{M}_A$ dargestellt, die sich aus diesen vier Typen von Segmenten zusammensetzt.

Um eine zufriedenstellende Approximation auch solcher Segmente zu gewährleisten, sind einige Besonderheiten zu beachten. In Abschnitt 4.5.5 ist die Approximation von Streckensegmenten für den Polygon inflation Algorithmus für $s = 3$ erläutert. Der Ansatz ist bei Anwendung des Strahlenalgorithmus ähnlich. Für $s = 4$ wird beim Verdacht auf ein degeneriertes Segment zunächst getestet, ob es sich um ein Flächensegment handelt. Sollte dies nicht der Fall sein, wird überprüft, ob es ein Streckensegment ist. Sollte auch dies nicht zutreffen, so handelt es sich anscheinend um ein Punktsegment (isolierte zulässige Lösung).

Ausgangspunkt zur Detektion von degenerierten Segmenten sind die Richtungen $v_1, \ldots, v_4$ der Initialzerlegung. Degenerierte Segmente sind dadurch gekennzeichnet, dass für einen, dieses Segment schneidenden, Strahl $\nu_i$ der Schnittpunkt $\nu_i \cap \partial\mathcal{F}_A$ nicht nur ein äußerer, sondern auch ein innerer Randpunkt ist. Es gilt $r_i = R_i$. Somit müssen nur solche Strahlen speziell analysiert werden. Es wird zunächst untersucht, ob es sich um einen Abschnitt einer Ebene handelt. Dazu wird überprüft, ob es in der unmittelbaren Nachbarschaft Strahlen gibt, die zulässige Lösungen enthalten. Sofern dies nicht zutrifft, wird überprüft, ob es sich um eine Strecke handelt. Sofern auch das nicht der Fall ist, handelt es sich offenbar um eine isolierte Lösung. Im Detail geschieht dies wie folgt:

- Gibt es in der direkten Nachbarschaft Strahlen die $\mathcal{M}_A$ schneiden, so werden alle zu diesem Segment gehörigen Randpunkte $R_j v_j$ zusammengefasst und im Sinne der kleinsten Quadrate durch eine Ebene approximiert. Liegt der Fehler dieser Approximation unter einer vorgegebenen Schranke, so erfolgt eine detaillierte Approximation dieser Fläche mittels einer Polygon inflation Routine. Diese Routine ist sehr ähnlich zu der in Abschnitt 4.5 beschriebenen, nur dass sie eine beschränkte Teilmenge einer vorgegebenen Ebene im Raum approximiert.[8] Eine Koordinate von $x \in \mathbb{R}^3$ ergibt sich automatisch durch die beiden anderen und die Optimierung läuft für $S \in \mathbb{R}^{3\times 3}$, der Rest bleibt unverändert. Als Startpunkt fungiert der Schwerpunkt aller $R_j v_j$ des Segments.

- Gibt es in der direkten Nachbarschaft keine Strahlen die $\mathcal{M}_A$ schneiden, so wird getestet,

[8] Diese Approximation der Teilmenge einer Fläche ist die Verallgemeinerung der Methode der scheibenweisen Berechnung der Menge $\mathcal{M}_A$ mittels Dreieckseinschließungen [58]. Im Unterschied dazu, ist hier die Ebene im Allgemeinen nicht orthogonal zu einer Achse.

ob es sich bei dem zu $v_i$ gehörigen Segment um eine Linie handelt.[9] Die Vorgehensweise des Testens, ob ein Streckensegment oder eine isolierte zulässige Lösung vorliegt, ist analog zu der in Abschnitt 4.5.5 beschriebenen. Die Funktion $g$ aus (4.30) muss dazu leicht modifiziert werden. Ausgehend von der initialen Lösung $x^{(0)} = r_i v_i = R_i v_i$ wird überprüft, ob es im Abstand von $r = \varepsilon_b$ von $x^{(0)}$ eine weitere zulässige Lösung gibt. Ist dies der Fall, so definieren diese und $x^{(0)}$ die Lage des Streckensegments. Die Grenzen für das zulässige Intervall sind analog zu (4.31) zu bestimmen.

- Gibt es keine zulässige Lösung im Abstand von $r = \varepsilon_b$ um $x^{(0)}$, so wird $x^{(0)}$ als isolierte zulässige Lösung deklariert.

## 4.8 Reduktionen durch fixierte Elemente

In Abschnitt 3.6 sowie [13,151,156] werden die Auswirkungen einer bekannten Zeile von $A$ (beziehungsweise einer bekannten Spalte von $C$) auf die verbleibenden Spalten von $C$ (beziehungsweise Zeilen von $A$) und deren Darstellungen in $\mathcal{M}_C$ (beziehungsweise $\mathcal{M}_A$) vorgestellt und untersucht.

Weiter wirkt sich die Kenntnis einer Zeile $A(i_1, :)$ in der Regel auch auf die Segmente der Menge $\mathcal{M}_A$ sowie das zu $C(:, i_1)$ gehörige Segment von $\mathcal{M}_C$ aus. Inwiefern eine fixierte zulässige Lösung in $\mathcal{M}_A$ die verbleibenden Segmente in restriktiver Form beeinflusst, lässt sich leicht anhand der geometrischen Klassifizierung verdeutlichen. Die Ränder der anderen Segmente werden, siehe Abschnitt 4.3, mittels an $\mathcal{I}_A$ anliegender Tangenten konstruiert. Die genutzten Dreiecke sind extremal und je zwei Punkte werden auf dem Rand von $\mathcal{F}_A$ gewählt. Sofern eine zulässige Lösung fixiert ist, entfällt dieser Freiheitsgrad und die anderen Segmente verringern sich in Bezug auf deren geometrische Ausmaße in der Regel. Eine Ausnahme ergibt sich beispielsweise, falls eine isolierte zulässige Lösung fixiert wird. In diesem Abschnitt werden Möglichkeiten der algorithmischen Umsetzung zur Einbindung derartiger Zusatzinformationen vorgestellt.

### 4.8.1 Ansatz zur Berechnung der restringierten Menge zulässiger Lösungen

Sofern eine Zeile $A(i_1, :)$ gegeben ist, sind die Ideen zur Berechnung der restringierten Menge zulässiger Lösungen grundsätzlich analog zu denen der Berechnung der (nicht restringierten) Menge $\mathcal{M}_A$. Für die geometrische Konstruktion ist ein Eckpunkt fixiert und der Algorithmus ist leicht zu modifizieren. Bei der numerischen Berechnung bieten sich die Dreieckseinschließungsmethode, der Polygon inflation-, und der Strahlenalgorithmus, siehe Korollar 3.20, an. Es ist lediglich die Zielfunktion $f(x, S)$ so zu modifizieren, dass die zusätzliche Information eingebunden wird.

Sind eine Spalte $C(:, i_1)$ gegeben und die Restriktion des dazugehörigen Segments der Menge $\mathcal{M}_A$ gesucht, so ändert sich bei der geometrischen Konstruktion die Kleinigkeit, dass zwei der drei Eckpunkte auf der zur niedrigdimensionalen Darstellung von $C(:, i_1)$ dualen Geraden liegen. Für die numerische Approximation mittels des Polygon inflation Algorithmus ändert sich wiederum die Zielfunktion $f(x, S)$. Für die Anwendung des Strahlenalgorithmus müsste zunächst geklärt werden, ob die Strahleneigenschaft auch in diesem Fall erhalten bleibt.

Da der Schwerpunkt auf den numerischen Methoden liegt, werden in den beiden folgenden Abschnitten die Modifikationen der Zielfunktion $f(x, S)$ aus (4.2) für die Fälle, dass eine Zeile von $A$ gegeben ist und/oder dass die zur aktuellen Berechnung gehörigen Spalte von $C$, also $C(:, 1)$, gegeben ist, untersucht.

[9] Die Fälle, dass ein oder zwei Strahl(en) $\mathcal{M}_A$ schneidet/schneiden und ein Streckensegment vorliegt, laufen auf eine Nullmenge hinaus.

### 4.8.2 Modifikation bei Teilkenntnis des zu untersuchenden Faktors

Angenommen, es sei eine Zeile von $A$ bekannt. Dies führt auf das zu fixierende $x^{(1)} \in \mathbb{R}^{s-1}$ mit

$$x_i^{(1)} = \frac{A(i_0,:)V(:,i+1)}{A(i_0,:)V(:,1)}, \quad i = 1,\ldots,s-1.$$

**Bemerkung 4.35.** *Für gestörte Daten wird das gegebene Profil durch $x^{(1)}$ nur im Sinne der kleinsten quadratischen Abweichung dargestellt und es kann passieren, dass $x^{(1)}$ nicht in $\mathcal{M}_A$ liegt. In solchen Fällen ist möglicherweise die Wahl der Steuerparameter $\varepsilon_a$ und $\varepsilon_c$ nicht geeignet oder die vorgegebene Zeile $A(i_0,:)$ enthält ein, in Bezug auf die Daten $D$, ungünstiges Störungsverhalten und sollte nicht direkt verwendet werden.*

Als eingeschränkte Menge zulässiger Lösungen (Teilmenge von $\mathcal{M}_A$) ergibt sich

$$\mathcal{M}_A^{(x^{(1)})} = \big\{x \in \mathbb{R}^{s-1} : \exists T \in \mathbb{R}^{s\times s}, \text{ mit } T(1,2:s) = x^{\mathbf{T}}, \operatorname{rank}(T) = s, \\ U\Sigma T^{-1}, TV^{\mathbf{T}} \geq 0,\, T(2,2:s) = (x^{(1)})^{\mathbf{T}}\big\} \cup x^{(1)}.$$

Dabei ist in der Umsetzung für $\mathcal{M}_A^{(x^{(1)})}$ ohne Beschränkung der Allgemeinheit die bekannte Zeile von $A$ an Position 2 angenommen. Für die numerischen Methoden erfolgt die Berücksichtigung von $x^{(1)}$ mittels einer Modifikation der Zielfunktion $f(x,S)$ aus (4.2), sodass ein $x$ genau dann als *zulässig* klassifiziert wird, wenn

$$\min_{S\in\mathbb{R}^{(s-2)\times(s-1)}} f_{x^{(1)}}(x,S) \leq \varepsilon_f$$

gilt. Die Funktion $f_{x^{(1)}} : \mathbb{R}^{s-1} \times \mathbb{R}^{(s-2)\times(s-1)}$ ist definiert als

$$f_{x^{(1)}}(x,S) = \frac{1}{2}\Bigg(\sum_{i=1}^{s}\sum_{j=1}^{k}\left(\min\left(0, \frac{C_{ji}}{\|C(:,i)\|_\infty} + \varepsilon_c\right)\right)^2 + \\ \sum_{\substack{i=1\\ i\neq 2}}^{s}\sum_{j=1}^{n}\left(\min\left(0, \frac{A_{ij}}{\|A(i,:)\|_\infty} + \varepsilon_a\right)\right)^2 + \|I_s - TT^+\|_F^2\Bigg)$$

mit $T(1,:) = (1, x^{\mathbf{T}})$, $T(2,:) = (1, (x^{(1)})^{\mathbf{T}})$, $T(i+2,:) = (1, S(i,:))$ für $i = 1,\ldots,s-2$ sowie $C = U\Sigma T^{-1}$ und $A = TV^{\mathbf{T}}$.

Die Einbindung weiterer $x^{(i)}$, $i = 2,\ldots,s_0$, erfolgt analog, sodass letztlich $T(i+1,:) = (1, (x^{(i)})^{\mathbf{T}})$ für $i = 1,\ldots,s_0$ vorgegeben sind. Die Anzahl der Freiheitsgrade in $T$ reduziert sich für jede fixierte zulässige Lösung um $s-1$.

Wie bei der nicht restringierten Variante wird die Klassifizierung eines $x \in \mathbb{R}^{s-1}$ in einen Schnelltest und, sofern dieser die Einordnung $x \in \mathcal{F}_A$ ergibt, einen zweiten, deutlich aufwendigeren, unterteilt.

### 4.8.3 Modifikation bei Teilkenntnis des nicht zu untersuchenden Faktors

Das Fixieren einer zulässigen Lösungen in $\mathcal{M}_A$ schränkt $\mathcal{M}_C$ stark ein. Durch das Dualitätsprinzip ergeben sich affine Hyperebenen die $\mathcal{M}_C$ schneiden. Dabei liegen $s-1$ Ecken eines zulässigen Simplex auf dieser Hyperebene. Weiter ergibt sich für die verbleibende Ecke (gekoppelte Lösung, vergleiche [145]) eine Einschränkung, welche in der Regel weniger gravierend ist. Bezüglich dieser Einschränkung werden in dieser Schrift keine analytischen Untersuchungen vorgenommen (was geometrisch möglich erscheint); es wird nur die numerische Umsetzung erläutert.

Um weiter den Fokus auf der Menge $\mathcal{M}_A$ zu behalten, wird an dieser Stelle der Fall behandelt, dass eine Spalte von $C$ bekannt ist und diese Information in die Rechnung einbezogen werden soll. Sei ohne Beschränkung der Allgemeinheit $C(:,1)$ gegeben und sei $y^{(1)} \in \mathbb{R}^{s-1}$ dessen zugehörige niedrigdimensionale Darstellung in $\mathcal{M}_C$. Die Einbindung von $y^{(1)}$ erfolgt über eine Erweiterung der Zielfunktion $f(x,S)$ aus (4.2) um den Term

$$\frac{1}{2}\sum_{i=2}^{s}\left(\frac{(T^+)_{i1}}{(T^+)_{11}} - y_i^{(1)}\right)^2,$$

wobei sich $T$ wie gehabt aus $x$ und $S$ zusammensetzt. Bei der Berechnung von $\mathcal{M}_A$ wird die erste Zeile von $T$ untersucht. Da die, sich aus der Kenntnis von $C(:,1)$ ergebenden, Restriktionen für $A(1,:)$ gesucht werden, wird die Forderung an die erste Spalte von $T^+$ gestellt.

### 4.8.4 Anwendung für Datensatz 2

Inwiefern sich die genannten Einschränkungen ergeben, wird anhand des Datensatzes 2 untersucht. Dazu werden die Mengen $\mathcal{M}_A$ und $\mathcal{M}_C$ in zwei Schritten eingeschränkt. Zunächst werden die Einschränkungen bestimmt, die sich aus der Vorgabe einer Zeile $A(1,:)$ (berechnet mittels $a_X(\lambda)$ bezüglich der zugrunde liegenden Diskretisierung) ergeben. Anschließend wird zusätzlich die Zeile $A(2,:)$ (mittels $a_Y(\lambda)$) als Extrainformation eingesetzt. Dies schränkt die verbleibenden Segmente von $\mathcal{M}_A$ und auch die Segmente von $\mathcal{M}_C$ ein.

#### Einschränkungen durch eine bekannte Zeile von $A$

Das Einbinden von $A(1,:)$ als $x^{(1)}$ führt auf die reduzierten Mengen zulässiger Lösungen $\mathcal{M}_A^{(x^{(1)})}$ und $\mathcal{M}_C^{(x^{(1)})}$. Bei dem Übergang von $\mathcal{M}_A$ zu $\mathcal{M}_A^{(x^{(1)})}$ reduziert sich ein Segment (blau) auf $x^{(1)}$ und bei den anderen beiden reduzieren sich die Flächen. Für das grüne Segment ist die Reduktion stark, denn es fallen 59.8% der Ursprungsfläche des Segments weg. Für das rote Segment ist die Einschränkung deutlich geringer, nur 11.3% der Ursprungsfläche werden ausgeschlossen. Die Reduktionen beim Übergang von $\mathcal{M}_C$ zu $\mathcal{M}_C^{(x^{(1)})}$ sind von unterschiedlicher Art. Es ergeben sich für die zwei, nicht zu $x^{(1)}$ gehörigen, Segmente wegen des Dualitätsprinzips Einschränkungen in Form einer Geraden, die die Segmente schneidet. Weiter ergibt sich eine Reduktion um 19.2% für das zugehörige blaue Segment. In Abbildung 4.20 sind die restringierten Mengen $\mathcal{M}_A^{(x^{(1)})}$ und $\mathcal{M}_C^{(x^{(1)})}$ dargestellt.

#### Einschränkungen durch zwei bekannte Zeilen von $A$

Als nächstes werden die Mengen $\mathcal{M}_A$ und $\mathcal{M}_C$ nicht nur durch $A(1,:)$ des originalen Faktors, sondern zusätzlich auch durch $A(2,:)$ eingeschränkt. Die durch $x^{(1)}$ und $x^{(2)}$ reduzierten Mengen zulässiger Lösungen seien mit $\mathcal{M}_A^{(x^{(1)},x^{(2)})}$ und $\mathcal{M}_C^{(x^{(1)},x^{(2)})}$ bezeichnet. Beim Übergang von $\mathcal{M}_A^{(x^{(1)})}$ zu $\mathcal{M}_A^{(x^{(1)},x^{(2)})}$ ergibt sich eine weitere (sehr leichte) Reduktion des roten Segments. Etwa 2.0% der Segmentfläche nach der ersten Einschränkung (durch $x^{(1)}$) fallen bei der zweiten (durch $x^{(2)}$) weg. Für $\mathcal{M}_C^{(x^{(1)},x^{(2)})}$ ergeben sich für das blaue und das rote Segment die Einschränkungen durch die zu $x^{(2)}$ duale Gerade. Mittels der zu $x^{(1)}$ dualen Geraden führt dies zu einer eindeutigen Lösung für $C(:,3)$, vergleiche auch [145, 156] sowie Abschnitt 3.6. Weiter ergibt sich eine (ebenfalls nur sehr leichte) Einschränkung für das grüne Segment, welche jedoch nur auf den Schnitt, mit der zu $x^{(1)}$ dualen Geraden, beschränkt ist. In Abbildung 4.21 sind die Einschränkungen dargestellt.

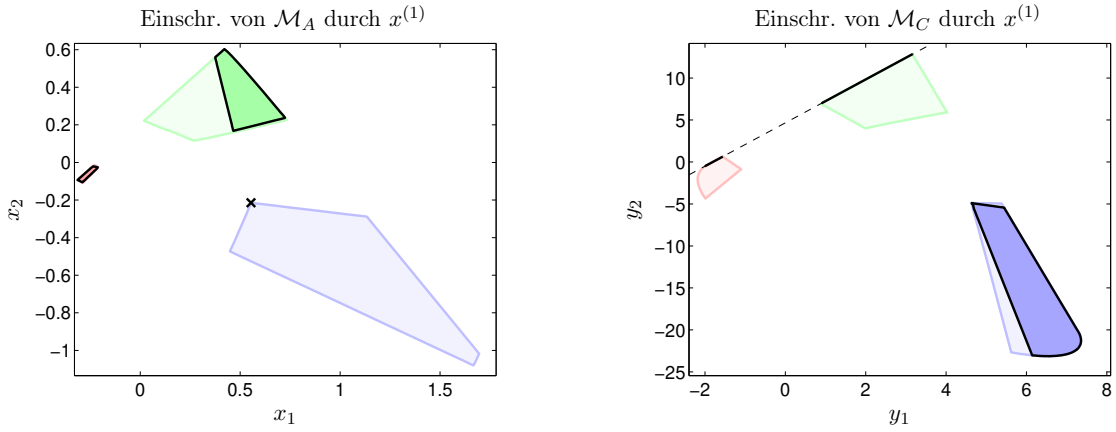


**Abbildung 4.20:** Die Reduktionen von $\mathcal{M}_C$ und $\mathcal{M}_A$ (transparent) zu $\mathcal{M}_A^{(x^{(1)})}$ und $\mathcal{M}_C^{(x^{(1)})}$ (übliche Farbdarstellung, schwarz umrandet) für den Fall, dass eine Zeile von $A$ bekannt ist anhand des Datensatzes 2. Links: Es ist $A(1,:)$ der korrekten Lösung (also $a_X(\lambda)$) in Form der niedrigdimensionalen Darstellung $x^{(1)}$ gegeben ($\times$ im blauen Segment). Daraus ergeben sich für die verbleibenden Segmente von $\mathcal{M}_A$ unterschiedlich starke Einschränkungen. Beim grünen Segment fallen 59.8% der Ursprungsfläche weg, beim roten Segment sind es nur 11.3%. Rechts: Dargestellt sind die Einschränkungen für die Menge $\mathcal{M}_C$. Durch das Dualitätsprinzip ergeben sich für das grüne und das rote Segment die Einschränkungen auf die Schnitte der Segmente mit der zu $x^{(1)}$ dualen Geraden (gestrichelten/durchgezogene Linie). Für das blaue Segment ergibt sich eine leichte Einschränkung.

## 4.9 Reduktionen durch Regularisierungen

Bei der Berechnung der Mengen zulässiger Lösungen werden nur die grundlegenden Bedingungen der Nichtnegativität beider Faktoren sowie der Rekonstruktionsforderung $D = CA$ berücksichtigt. Nun ist das inverse Problem der Bestimmung einer nichtnegativen Matrixfaktorisierung jedoch schlecht gestellt und die Mengen möglicher Lösungen hinsichtlich deren Profilformen oft breit gefächert. Bei vielen Anwendungen sind, was die Rekonstruktion der zugrunde liegende Faktoren betrifft, nur einzelne zulässige Lösungen sinnvoll. Es besteht der Wunsch (der speziellen Anwendung betreffend) nicht relevante Lösungen zu eliminieren.

Eine Möglichkeit die Mengen zulässiger Lösungen $\mathcal{M}_A$ und $\mathcal{M}_C$ auf Mengen zulässiger und, für die vorliegende Anwendung, relevanter Lösungen $\mathcal{M}_A^{(\mathrm{rel})}$ und $\mathcal{M}_C^{(\mathrm{rel})}$ zu reduzieren, ist es, zusätzliche Regularisierungsfunktionen einzusetzen und die Bewertungsfunktion bezüglich dieser zu erweitern. Mit dem Anwendungshintergrund der Analyse spektroskopischer Daten wird in diesem Abschnitt nur auf diesbezüglich sinnvolle Regularisierungen eingegangen. In Publikationen ist die Idee der Reduktionen von $\mathcal{M}_A$ und $\mathcal{M}_C$ durch den Einsatz von Regularisierungen, sowohl unter geometrischen Gesichtspunkten [12, 13] als auch unter numerischen [4, 57, 132, 158] behandelt und erfolgreich an Beispielen getestet worden.

Es werden drei mögliche Formen der Reduktion von $\mathcal{M}_A$ und $\mathcal{M}_C$ durch Regularisierungen vorgestellt: Unimodalitätsforderungen, Monotonieforderungen und Fensterlösungen. Die Regularisierungen werden hier jeweils für Teile des Faktors $C$ angewendet. Für Arbeiten über den Einsatz von Regularisierungen mit Blick auf die Analyse spektroskopischer Daten sei beispielsweise auf [25, 48, 51, 81, 108, 112, 117, 124, 149, 171, 176] verwiesen.

Um die Anwendung der Regularisierungen nicht von den absoluten Einträgen abhängig zu machen, wird für die zusätzlichen Regularisierungen mit normierten Spalten in $C$ gearbeitet. Dazu sei $\widehat{C} \in \mathbb{R}^{k\times s}$,

$$\widehat{C}_{ij} = \frac{C_{ij}}{\|C(:,j)\|_\infty}, \qquad i = 1,\ldots,k,\ j = 1,\ldots,s. \tag{4.35}$$

Die in diesem Abschnitt genutzten Regularisierungsfunktionen sind aus Übersichtlichkeitsgründen direkt für $C$ definiert, in der Anwendung wird jedoch oft mit $\widehat{C}$ gerechnet.

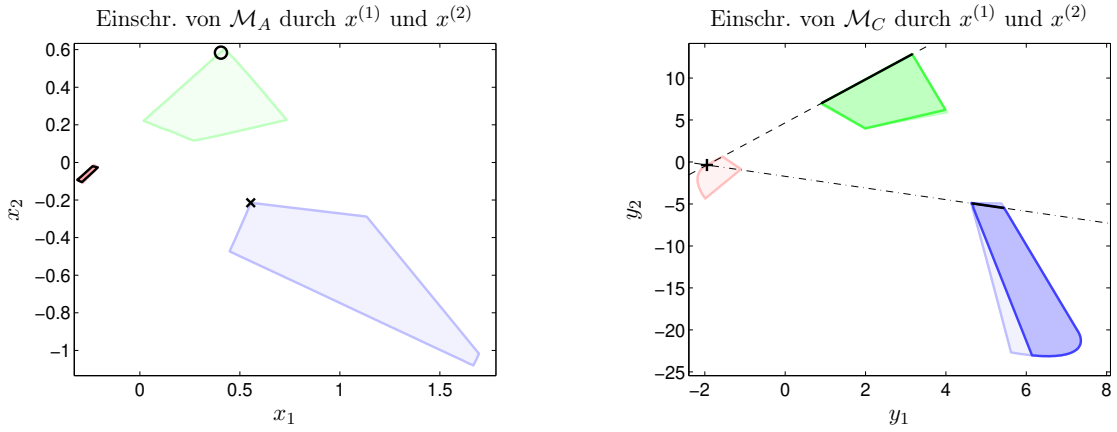


**Abbildung 4.21:** Die Reduktionen von $\mathcal{M}_C$ und $\mathcal{M}_A$ (transparent) zu $\mathcal{M}_A^{(x^{(1)},x^{(2)})}$ und $\mathcal{M}_C^{(x^{(1)},x^{(2)})}$ (übliche Farbdarstellung, schwarz umrandet) für den Fall, dass zwei Zeilen von $A$ bekannt sind anhand des Datensatzes 2. Links: Es sind $A(1,:)$ (×) und $A(2,:)$ (○) der Originallösung gegeben. Dies führt auf die (leichte) Reduktion des roten Segments von $\mathcal{M}_A$. Rechts: Die zu $A(1,:)$ und $A(2,:)$ dualen Linien $(--)$ und $(-\cdot)$ schränken das grüne und das rote beziehungsweise das blaue und das grüne Segment ein: eine Lösung ist in $\mathcal{M}_C$ isoliert (+) und die beiden anderen liegen auf Strecken.

Die Umsetzung der Regularisierungen zur Reduktion von $\mathcal{M}_A$ erfolgt über die Klassifizierungsroutine. Die Funktion $f(x,S)$ aus (4.2) wird um Bewertungen bezüglich Unimodalität, Monotonie und Fensterlösungen erweitert und es ergibt sich

$$\widehat{f}(x,S) = f(x,S) + \frac{1}{2}\big(\gamma_{\text{uni}} f_{\text{uni}}(C) + \gamma_{\text{mono}} f_{\text{mono}}(C) + \gamma_{\text{win}} f_{\text{win}}(C)\big) \tag{4.36}$$

mit Gewichtungsfaktoren $\gamma_{\text{uni}}, \gamma_{\text{mono}}, \gamma_{\text{win}} \geq 0$ sowie Unterfunktionen $f_{\text{uni}}, f_{\text{mono}}, f_{\text{win}} : \mathbb{R}^{k\times s} \to \mathbb{R}$. Diese dienen der Umsetzung der Regularisierungen zur Unimodalität, zur Monotonie und zu Fensterlösungen und werden im Laufe dieses Abschnitts eingeführt und erläutert. Analog zu (4.3) wird ein $x$ als *zulässig* klassifiziert, wenn

$$\min_{S\in\mathbb{R}^{(s-1)\times(s-1)}} \widehat{f}(x,S) \leq \varepsilon_f.$$

**Bemerkung 4.36.** *In diesem Abschnitt ist an manchen Stellen von „negativen Bewertungen“ die Rede. Damit sind keine negativen skalaren Werte gemeint, sondern Strafwerte, die sich zu Profilen ergeben, welche den geforderten Regularisierungen (Unimodalität, Monotonie, begrenzte Träger) nicht entsprechen.*

Für $s = 3$ kann die Berechnung der Menge zulässiger und, für die vorliegende Anwendungsproblematik, relevanter Lösungen $\mathcal{M}_A^{(\text{rel})}$ etwa mit dem (direkten) Polygon inflation Algorithmus erfolgen. Dabei ist es möglich, dass sich einzelne Eigenschaften von $\mathcal{M}_A$ nicht auf $\mathcal{M}_A^{(\text{rel})}$ übertragen. Die Beschränktheit bleibt aber genauso erhalten, wie die Eigenschaft, dass $(0,\dots,0)^{\mathbf{T}} \notin \mathcal{M}_A^{(\text{rel})}$ gilt.

**Bemerkung 4.37.** *Sofern zusätzliche Regularisierungen eingebunden werden, ist nur der direkte, nicht aber der inverse Polygon inflation Algorithmus zur Berechnung von $\mathcal{M}_A^{(rel)}$ geeignet. Dies liegt an den unterschiedlichen Strukturen der Klassifizierungsroutinen. Beim direkten Polygon inflation Algorithmus werden alle Teile von $C$ und $A$ gleichzeitig überwacht, was für die Rückkopplung der Regularisierungen nötig ist. Anders ist dies beim inversen Typ, da hier die Restriktionen zweigeteilt und unabhängig voneinander kontrolliert werden. Die Rückkopplung von beispielsweise Monotonierestriktionen kann nicht für alle Teilfaktoren, sondern entweder nur für $A(1,:)$ oder nur für $C$ und $A(2:3,:)$ beachtet werden.*

```
for i ∈ I_uni do
    i_0 = argmax(C(:,i)) and c_m = C(i_0,i)
    for j = i_0 : −1 : 2 do
        Y(j,i) = min(0, c_m − C(j−1,i) + ω)
        if ((c_m > C(j−1,i)) or (c_m − C(j−1,i) + ω < 0)) then
            c_m = C(j−1,i)
        end
    end
    c_m = C(i_0,i)
    for j = i_0 + 1 : k do
        Y(j,i) = min(0, c_m − C(j,i) + ω)
        if ((c_m > C(j,i)) or (c_m − C(j,i) + ω < 0)) then
            c_m = C(j,i)
        end
    end
end
```

**Pseudocodeelement 1:** Der Pseudocode für die Berechnung der Matrix $Y$, welche in (4.37) zum Einbinden der Unimodalitätsrestriktion in der Bewertung eines $x \in \mathbb{R}^{s-1}$ zum Einsatz kommt. Es werden nur Profile $C(:,i)$ mit $i \in I_{\text{uni}} \subset \{1,\ldots,s\}$ bewertet. Der Steuerparameter $\omega \geq 0$ dient dazu, leichte Störungen zu erkennen und nicht negativ in die Bewertung einfließen zu lassen. Die Matrix $Y$ enthält negative Werte an Stellen, an denen sich das aktuell untersuchte Profil nicht unimodal verhält und Nullen sonst.

### 4.9.1 Unimodalitätsrestriktionen

Unimodalitätsrestriktionen werden oft als Regularisierungen zur Berechnung von Reinkomponentenzerlegung verwendet. Ein unimodales Konzentrationsprofil weist nur ein lokales Maximum auf [20, 32, 167]. Davor verhält es sich monoton steigend, danach monoton fallend. Ein typischer Fall eines unimodalen, aber nicht monotonem, Konzentrationsverlaufs liegt bei einem Intermediat vor, worin auch die häufige Anwendung begründet ist. An dieser Stelle sei angemerkt, dass der Einsatz von Unimodalitätsrestriktionen prinzipiell nicht auf den Faktor $C$ beschränkt ist.

Die Bewertungsfunktion $f_{\text{uni}}(C)$ ist so aufgebaut, dass sie für ein unimodales Profil einen Wert nahe Null (kleiner als $\varepsilon_f$) liefert. Um die Funktion auch für störungsbehaftete Werte anwenden zu können, wird ein Steuerparameter $\omega \geq 0$ eingesetzt. Dieser dient dazu, störungsbedingte Abweichungen von einem unimodalen Verlauf abzufangen. Oft ist die Forderung von Unimodalität für alle Profile nicht sinnvoll und es wird die Indexmenge $I_{\text{uni}} \subset \{1,\ldots,s\}$ für alle Profile, die einem unimodalen Verlauf folgen sollen, eingeführt. Zur Bewertung der einzelnen Profile $C(:,i)$ für $i \in I_{\text{uni}}$ in Bezug auf Unimodalität wird das Pseudocodeelement 1 genutzt. Mit dem darin berechneten $Y$ wird die Zielfunktion

$$f_{\text{uni}}(C) = \sum_{i \in I_{\text{uni}}} \|Y(:,i)\|_2^2 \tag{4.37}$$

ausgewertet. In dem Pseudocodeelement 1 wird die Variable $c_m$ als Referenzwert für die jeweiligen Vergleiche genutzt. Es ist $c_m$ entweder der letzte Wert zu einem unimodalen Verhalten oder der letzte Fehler-/Strafwert für ein fallendes/steigendes Verhalten, der in die Regularisierungsfunktion einging.

### 4.9.2 Monotonierestriktionen

Monotonierestriktionen sind eine Verschärfung der Unimodalitätsrestriktionen, denn es werden nur monoton steigende und monoton fallende Profile zugelassen. Analog zur Implementierung der Unimodalitätsrestriktionen werden die Monotonierestriktionen nur für einzelne Profile gefordert,

```
for i ∈ I_mono do
    z = C(1,i) and w = C(1,i)
    for j = 1 : k − 1 do
        Z_ji = min(0, z − C(j+1,i) + ρ)
        W_ji = max(0, w − C(j+1,i) − ρ)
        if ((z > C(j+1,i)) or (z − C(j+1,i) + ρ < 0)) then
            z = C(j+1,i)
        end
        if ((w < C(j+1,i)) or (w − C(j+1,i) − ρ > 0)) then
            w = C(j+1,i)
        end
    end
end
```

**Pseudocodeelement 2:** Der Pseudocode für die Berechnung der Matrizen $Z$ und $W$, welche in (4.38) zum Einbinden der Monotonierestriktion in die Bewertung eines $x \in \mathbb{R}^{s-1}$ zum Einsatz kommen. Der Steuerparameter $\rho \geq 0$ dient dazu, leichte Störungen als solche auszumachen und nicht negativ in die Bewertung einfließen zu lassen. Die Matrix $Z$ enthält Bewertungen bezüglich eines monoton fallenden Profils und $W$ enthält Bewertungen bezüglich eines monoton steigenden Profils. In die Zielfunktion $f_{\text{mono}}$ fließt pro Profil $C(:,i)$ mit $i \in I_{\text{mono}} \subset \{1,\ldots,s\}$ nur eine der beiden Bewertungen (die mit der geringeren Norm) ein.

nämlich für $C(:,i)$ mit $i \in I_{\text{mono}}$ zu einer geeigneten Indexmenge $I_{\text{mono}}$. Um einen stabilen Umgang mit gestörten Daten zu ermöglichen, kommt ein Steuerparameter $\rho \geq 0$ zum Einsatz. Störungsbedingte Abweichungen werden so nicht negativ bewertet.

In der Implementierung wird für die Analyse des Profils $C(:,i)$ getestet, wie stark $C(:,i)$ von einem monoton steigenden Profil abweicht und wie stark es von einem monoton fallenden Profil abweicht. In die Zielfunktion fließt die normmäßig kleinere negative Bewertung ein. Die Details der Implementierung sind im Pseudocodeelement 2 angegeben und die Zielfunktion ergibt sich als

$$f_{\text{mono}}(C) = \sum_{i \in I_{\text{mono}}} \min\left(\|Z(:,i)\|_2^2,\, \|W(:,i)\|_2^2\right). \tag{4.38}$$

### 4.9.3 Gefensterte Profile

Bei der Analyse spektroskopischer Daten ist es oft der Fall, dass in bestimmten Zeitbereichen nicht alle $s$ Komponenten vorliegen oder dass zu bestimmten Wellenlängen nicht alle Komponenten signifikant absorbieren. Beispielsweise ist bei einer konsekutiven Reaktion, die ausreichend lange beobachtet wird, gegen Ende des Zeitintervalls praktisch nur noch das Produkt vorhanden. Oder andersherum liegt/liegen bei einer langsam startenden Reaktion zu Beginn praktisch nur das Edukt/die Edukte vor. In diesem Zusammenhang wird zu einem Profil $C(:,i)$ oder $A(j,:)$ ein Zeit-/Frequenzbereich *Fenster* genannt, wenn die Komponente nur in diesem Zeitfenster vorliegt/nur in diesem Frequenzfenster absorbiert. Solche Fenster sind in der Anwendung häufig bekannt und lassen sich bei der Analyse der Daten als zusätzliche Informationen einbinden, siehe [33, 116, 159] für die *window factor analysis* (WFA) und [44, 45, 52, 89, 111, 113, 114] für die artverwandte *evolving factor analysis* (EFA).

Ein Einsatz derartiger Fensterinformationen zur Reduktion der Mengen zulässiger Lösungen ist ebenfalls möglich. Vorgestellt wird ein solches Vorgehen nur für den Faktor $C$, wenngleich es analog auch für $A$ anwendbar ist. Seien erneut nur für Profile $C(:,i)$ mit $i \in I_{\text{win}} \subset \{1,\ldots,s\}$ Fensterrestriktionen angewendet. Für jedes $i \in I_{\text{win}}$ wird mit $J^{(i)} \subset \{1,\ldots,k\}$ die Menge an

Indizes $j$ definiert, für welche $C_{ji}$ positiv ist. Es gilt also

$$C_{ji} = 0 \quad \text{für} \quad i \in I_{\text{win}},\ j \in \{1, \ldots, k\} \setminus J^{(i)}.$$

Mit einem Steuerparameter $\varrho \geq 0$ lautet die Regularisierungsfunktion

$$f_{\text{win}}(C) = \sum_{i \in I_{\text{win}}} \sum_{j \notin J^{(i)}} \max\left(C_{ji} - \varrho,\, 0\right)^2.$$

**Bemerkung 4.38.** *Die Parameter $\omega$, $\rho$ und $\varrho$ lassen sich auch für jedes Profil unterschiedlich wählen. Dies ist insbesondere dann nötig, wenn die einzelnen Profile unterschiedlich stark durch Störungen beeinflusst werden. Eine solche Situation liegt für die Anwendung von Monotoniere-striktionen auf Datensatz 3 vor, siehe Abschnitt 5.4.*

### 4.9.4 Bedeutung der Steuerparameter

Auf den stabilen Einsatz von Regularisierungsfunktionen zur Reduktion der Menge $\mathcal{M}_A$ haben die Steuerparameter entscheidenden Einfluss. Die Klassifizierung eines $x \in \mathbb{R}^{s-1}$ erfolgt mittels $\widehat{f}(x, S)$ aus (4.36) in der Form, dass $x$ genau dann als *zulässig* klassifiziert wird, wenn

$$\min_{S \in \mathbb{R}^{(s-1)\times(s-1)}} \widehat{f}(x, S) \leq \varepsilon_f$$

gilt. Nur kleine Werte $\varepsilon_f$, zum Beispiel $\varepsilon_f = 10^{-12}$, ermöglichen sinnvolle Ergebnisse für die Approximation von $\mathcal{M}_A$. Demzufolge sollten die Funktionen $f_{\text{uni}}(C)$, $f_{\text{mono}}(C)$ und $f_{\text{win}}(C)$, für Profile, die die Regularisierungsforderungen erfüllen, Werte kleiner als $\varepsilon_f$ zurückgeben. Dies ist für störungsbehaftete Daten in der Regel für $\omega = 0$, $\rho = 0$ beziehungsweise $\varrho = 0$ nicht der Fall. Um die Regularisierungen effektiv arbeiten zu lassen aber auch keine geeigneten Lösungen auszuschließen, sind die Steuerparameter sinnvoll einzusetzen.

Viele andere Regularisierungen lassen sich nicht zur Reduktion der Mengen zulässiger Lösungen einsetzten. Es können nur Funktionen sinnvoll eingesetzt werden, die für ein Profil, dass den Regularisierungen entspricht, einen Wert nahe Null zurückgibt. Der Einsatz beispielsweise von Norm- oder Glattheitsregularisierungen ist so nur mittels oberer Schranken für die Regularisierungswerte möglich, was die Wirkung der Restriktionen aber stark abschwächt.

So ließe sich beispielsweise eine Glattheitsregularisierung mit den zentralen finiten Differenzen zweiter Ordnung für $C$, siehe etwa [124,149,175,176], für ein $i_0 \in \{1, \ldots, s\}$ und ein äquidistantes Zeitraster mit Schrittweite $\tau$ nur in der Form

$$f_{\text{smooth}}(C) = \sum_{\ell=2}^{k-1} \min\left(0, \left(\frac{C(\ell-1, i_0) - 2C(\ell, i_0) + C(\ell+1, i_0)}{\tau^2}\right)^2 - \delta_C\right)$$

mit einem Steuerparameter $\delta_C > 0$ anwenden. Wäre $\varepsilon_f$ sinnvoll gewählt und $\delta_C = 0$, würde dies zu leeren Mengen zulässiger und relevanter Lösungen führen. Dies gilt unabhängig von Störungen. Um sinnvolle Ergebnisse zu erhalten, muss $\delta_C$ positiv (und nicht zu klein) sein.

## 4.10 Alternativer Zugang über begrenzende Lösungen

Unabhängig von den Mengen zulässiger Lösungen wird in [50, 167, 168] ein alternativer Zugang zur Analyse der Mehrdeutigkeit der Lösung des Faktorisierungsproblems 2.4 genutzt. Ausgehend von einer initialen Zerlegung werden $2s$ spezielle zulässige Zeilen für $A$ und $2s$ spezielle zulässige Spalten für $C$ bestimmt. Diese werden als begrenzende Lösungen für die einzelnen Zeilen von

$A$ und Spalten von $C$ angesehen. Dazu werden Optimierungsprobleme aufgestellt und lokale Extremstellen dieser bestimmt.

Bei dem Ansatz gibt es jedoch zwei Schwierigkeiten. Erstens ist es auch bei Berücksichtigung der Skalierungsuneindeutigkeit im Normalfall für beliebiges $s \geq 3$ nicht möglich, begrenzende Lösungen für die Zeilen von $A$ und die Spalten von $C$ zu bestimmen, welche gleichzeitig auch direkt Elemente aus $\mathcal{A}$ beziehungsweise $\mathcal{C}$ sind (also tatsächlich zulässige Zeilen in $A$ beziehungsweise Spalten in $C$). Bei diesem Problem ist die ungeklärte Frage nach der Skalierung, welche in der Tat ein Problem darstellt, nicht entscheidend. Das zweite Problem ist die Aufstellung der Optimierungsprobleme. Die genutzten Zielfunktionen besizten im Normalfall nur zwei Extremstellen. Dass jedoch jeweils $s$ verschiedene Maximal- und Minimalstellen berechnet werden, liegt an der Nutzung der nur begrenzt leistungsfähigen MATLAB-Routine fmincon. Diese ist nicht in der Lage zwischen den Segmenten von $\mathcal{M}_A$ zu wechseln und es werden je $s$ lokale Extremstellen berechnet, wovon je nur eine auch globale Extremstelle ist.

Nichtsdestotrotz ist die Methode der Bestimmung solcher Kurven, die als Einhüllenden für die Faktoren $C$ und $A$ angesehen werden, ein geeigneter Ansatz zur Analyse der Mehrdeutigkeit der Lösung des Faktorisierungsproblems 2.4. Die Methode ist für störungsbehaftete Daten einsetzbar.

### 4.10.1 Umsetzung mittels einer speziellen Zielfunktion

Basierend auf einer initialen nichtnegativen Faktorisierung aus $C$ und $A$ wird eine Zielfunktion aufgestellt und unter zu wählenden Restriktionen minimiert beziehungsweise maximiert. Verschiedene Restriktionen stehen zur Wahl. Die wichtigsten sind die Nichtnegativitätsbedingungen für die Faktoren sowie die Rekonstruktionsforderung. Weiter können die Unimodalitätsrestriktionen, siehe [20, 32, 167] und Abschnitt 4.9.1, gewählt werden. In der folgenden Beschreibung der Methode wird auf die MATLAB-Implementierung MCR-BANDS eingegangen, welche auf http://www.mcrals.info/ heruntergeladen wurde [83].

**Bemerkung 4.39.** *In der* MCR-BANDS*-Implementierung [83] ist die Rekonstruktionsforderung $D = CA$ als Restriktion weggelassen. Trotzdem ergeben sich in der Regel sinnvolle Ergebnisse beziehungsweise es lassen sich solche, die keinen Sinn ergeben, erkennen.*

Werden nur die elementaren Restriktionen berücksichtigt, so lautet zu einer vorgegebenen (initialen) nichtnegativen Matrixfaktorisierung aus $C$ und $A$ die Zielfunktionen

$$g_i(R) = \frac{\|C(R^{-1})(:,i)R(i,:)A\|_F^2}{\|CR^{-1}RA\|_F^2} \tag{4.39}$$

für $i = 1, \ldots, s$ und jeweils $R \in \mathbb{R}^{s\times s}$. Diese $g_i(R)$ werden unter den zuvor genannten (und nötigen) Nebenbedingungen minimiert beziehungsweise maximiert. Die Optimierungen werden jeweils zu den Startwerten $R = I_s$ und unter Nutzung der MATLAB-Routine fmincon durchgeführt. Es werden $s$ Minimalstellen $R^{(i,\min)} \in \mathbb{R}^{s\times s}$ und $s$ Maximalstellen $R^{(i,\max)} \in \mathbb{R}^{s\times s}$ bestimmt. Aus diesen ergeben sich für $i = 1, \ldots, s$ jeweils

$$a^{(2i-1)} = R^{(i,\min)}(i,:)A \quad \text{beziehungsweise} \quad a^{(2i)} = R^{(i,\max)}(i,:)A$$

sowie

$$c^{(2i-1)} = C(R^{(i,\min)})^{-1}(:,i) \quad \text{beziehungsweise} \quad c^{(2i)} = C(R^{(i,\max)})^{-1}(:,i).$$

Diese Profile werden als begrenzende Lösungen angesehen, wenngleich sie dies nur unter bestimmten Bedingungen tatsächlich sind. Zu einer Minimierung/Maximierung ergeben sich also jeweils eine Spalte für $C$ und eine Zeile für $A$.

**Bemerkung 4.40.** *Prinzipiell besitzen die Funktionen $g_i(R)$, $i = 1, \ldots, s$, alle das gleiche Minimum und das gleiche Maximum. Demzufolge besitzen sie nicht nur jeweils die gleichen Minimal- und Maximalstellen, sondern es gelten prinzipiell $a^{(1)} = a^{(2i-1)}$ und $a^{(2)} = a^{(2i)}$ sowie $c^{(1)} = c^{(2i-1)}$ und $c^{(2)} = c^{(2i)}$ für $i = 2, \ldots, s$. Durch die numerische Lösung der einzelnen Optimierungsprobleme ist dies in der Anwendung jedoch nicht der Fall, was mutmaßlich an der geringen Leistungsstärke der Optimierungsmethode* fmincon *von* MatLab *liegt. Ein Wechsel in der Sortierung der Faktoren (also ein Wechsel zwischen den Segmenten von $\mathcal{M}_A$ beziehungsweise von $\mathcal{M}_C$) wird in der Regel nicht realisiert und es werden nur die jeweiligen lokalen Extremstellen und -werte berechnet.*[10]

### 4.10.2 Anbindung an die Mengen zulässiger Lösungen

Um den Bogen von den begrenzenden Lösungen zu den niedrigdimensionalen Darstellungen sowie den Mengen $\mathcal{M}_A$ und $\mathcal{M}_C$ zu schlagen, werden zu $a^{(i)} \in \mathbb{R}^{1\times n}$ und $c^{(i)} \in \mathbb{R}^k$ deren niedrigdimensionale Darstellungen $x^{(i)}$ für $A$ und $y^{(i)}$ für $C$ für $i = 1, \ldots, 2s$ berechnet. Es ergeben sich

$$x^{(i)} = \left( \frac{a^{(i)} V(:, 2:s)}{a^{(i)} V(:, 1)} \right)^{\mathbf{T}}, \qquad y^{(i)} = \frac{U\Sigma(:, 2:s) c^{(i)}}{U\Sigma(:, 1) c^{(i)}}. \tag{4.40}$$

Die $x^{(i)}$ lassen sich mit $\mathcal{M}_A$ in Verbindung bringen, die $y^{(i)}$ mit $\mathcal{M}_C$. Für störungsfreie Daten sind $x^{(i)} \in \mathcal{M}_A$, $i = 1, \ldots, 2s$. Die Lösung führt für $s \geq 3$ in der Regel zwar nicht zu oberen beziehungsweise unteren Einhüllenden, da es solche für das gestellte Problem und mit einer Darstellung in $\mathcal{M}_A$ beziehungsweise in $\mathcal{M}_C$ im Allgemeinen nicht gibt, vergleiche auch Punkt 1 in Bemerkung 4.41. Oft ergeben sich aber gute Abschätzungen, siehe beispielsweise [115,161,178].

**Bemerkung 4.41.**

1. *Da sich für $s = 2$ die Mengen $\mathcal{M}_A$ und $\mathcal{M}_C$ jeweils mit vier zulässigen Lösungen bestimmen lassen, kann die Berechnung von begrenzenden Lösungen auf die vollständigen Mengen zulässiger Lösungen führen. Eine bislang nicht bewiese Vermutung lautet, dass die $x^{(i)}$ die Intervallgrenzen für $\mathcal{M}_A$ aus (4.13) sind, vergleiche [1,136] und Abschnitt 4.10.3. Für $s \geq 3$ kann die Menge zulässiger Lösungen im Normalfall, das heißt für nicht hinreichend stark degenerierte Segmente, nicht vollständig wiedergegeben werden. Die sinnvolle Darstellung über die begrenzenden Lösungen für $s = 2$ zeigt sich bei der Anwendung auf Datensatz 1 in Abbildung 5.17, siehe Abschnitt 5.5.1.*
2. *Eine Schwierigkeit bei der Bestimmung der begrenzenden Lösungen und deren grafischer Darstellung liegt in der Skalierung/Normierung der Profile, sodass diese zur Analyse der Mehrdeutigkeit der Lösung des Faktorisierungsproblems der nichtnegativen Matrixfaktorisierung geeignet sind. Die in dieser Arbeit genutzte Variante über $A = RV^{\mathbf{T}}$ mit $R_{i1} = 1$ ist ebenso nicht gut geeignet, wie eine Normierung, sodass $\|A(i,:)\|_\infty = 1$ gilt.*

Trotz der einfachen Struktur der Zielfunktion ist deren Analyse schwierig und wird in dieser Arbeit nicht vorgenommen. In erster Linie sind Aussagen zur Lage von $x^{(i)}$ in $\mathcal{M}_A$ und zur Lage von $y^{(i)}$ in $\mathcal{M}_C$ interessant. In der Regel liegen die $x^{(i)}$ auf dem Rand von $\mathcal{M}_A$ und die $y^{(i)}$ auf dem Rand von $\mathcal{M}_C$. Dies gilt für Datensatz 2 jedoch nicht (vorbehaltlich korrekt bestimmter Extrema und einer genügend feinen Auflösung), siehe die Abbildungen 5.18 und 5.21.

[10] Möglich erscheint es hier auch, dass der Autor der vorliegenden Arbeit trotz intensiver Analyse des Quellcodes eine entsprechende Steuerung der Methode übersehen hat.

### 4.10.3 Untersuchungen für $s = 2$

Die Bestimmung der niedrigdimensionalen Darstellungen der begrenzenden Lösungen führt auf $2s$ Elemente $x^{(i)} \in \mathbb{R}^{s-1}$ und $2s$ Elemente $y^{(i)} \in \mathbb{R}^{s-1}$. Die Mengen zulässiger Lösungen sind Teilmengen des $\mathbb{R}^{s-1}$. Somit ist eine Übereinstimmung der Ergebnisse beider Ansätze für nicht degenerierte Segmente nur für $s = 2$ möglich.

Untersucht sei die Optimierungsfunktion $g_i(R)$ aus (4.39) für $s = 2$. Diese soll unter den Restriktionen $\operatorname{rank}(R) = 2$, $CR^{-1} \geq 0$, $RA \geq 0$ minimiert beziehungsweise maximiert werden. Die Menge zulässiger Lösungen $\mathcal{M}_A$ für $s = 2$ hat die Form wie in (4.12) und (4.13). Um die niedrigdimensionalen Darstellungen zu erhalten, lässt sich äquivalent dazu die Zielfunktion

$$\widetilde{g}_i(\alpha, \beta) = \left\| U\Sigma(\widetilde{R}^{-1})(:, i)\widetilde{R}(i, :)V^{\mathbf{T}} \right\|_F^2 \tag{4.41}$$

betrachten mit

$$\widetilde{R} = \begin{pmatrix} 1 & \alpha \\ 1 & \beta \end{pmatrix} \tag{4.42}$$

sowie den Faktoren $U$, $\Sigma$ und $V$ einer abgeschnittenen Singulärwertzerlegung von $D$. Wegen der zu beachtenden Restriktionen müssen $\alpha, \beta \in \mathcal{M}_A$ und $\alpha\beta < 0$ gelten. Einfache Umformungen führen auf

$$\widetilde{g}_1(\alpha, \beta) = \frac{\left(\sigma_1^2\beta^2 + \sigma_2^2\right)\left(1 + \alpha^2\right)}{(\beta - \alpha)^2}, \qquad \widetilde{g}_2(\alpha, \beta) = \frac{\left(\sigma_1^2\alpha^2 + \sigma_2^2\right)\left(1 + \beta^2\right)}{(\beta - \alpha)^2}.$$

Wegen der Permutationsmehrdeutigkeit der Lösung genügt es, nur $\widetilde{g}_1$ weiter zu untersuchen. Dies führt auf die Zielfunktionen

$$\widetilde{G}_1(\alpha) = \max_{\beta \in \mathcal{M}_A,\, \alpha\beta < 0} \widetilde{g}_1(\alpha, \beta), \qquad \widehat{G}_1(\alpha) = \min_{\beta \in \mathcal{M}_A,\, \alpha\beta < 0} \widetilde{g}_1(\alpha, \beta).$$

Es sind die lokalen Maximalstellen von $\widetilde{G}_1(\alpha)$ und die lokalen Minimalstellen von $\widehat{G}_1(\alpha)$ in $\mathcal{M}_A$ gesucht.

Zwei zu $\widetilde{g}_1$ und $\widetilde{g}_2$ analoge Funktionen sind in [1] numerisch für eine konkrete Matrix $D$ sowie in [136] untersucht. In beiden Arbeiten wird für die Transformation $\widetilde{R}$ eine andere Form als in (4.42) genutzt. Die Resultate stimmen aber grundsätzlich überein.

### 4.10.4 Weiterführende Verweise

Die Implementierung von MCR-BANDS geht auf die Gruppe um Prof. R. Tauler, Dr. A. de Juan sowie Dr. J. Jaumot aus Barcelona zurück. Auf diese geht auch die häufig verwendete Methode MCR-ALS [31, 79–81], siehe auch [46, 169], zur Berechnung von Reinkomponentenzerlegungen zurück. Wenngleich auch bei diesem Algorithmus keine Konvergenz gesichert und eine geschlossenen Analyse schwierig ist [134], so liefert die Methode in der Regel sinnvolle Ergebnisse, siehe etwa [7, 47]. Weiter wird im Zuge alternativer Methoden zur Analyse der Mehrdeutigkeit des Faktorisierungsproblems auch auf die in [164] vorgestellte Idee der Nutzung einer Partikelschwarmoptimierung sowie den eingangs erwähnten *grid search*-Ansatz [1,2,13,57,173] verwiesen.

## 4.11 Zusammenfassung und Perspektiven

Zwar wurden schon in [105] und [17] Methoden zur Berechnung der Mengen zulässiger Lösungen vorgestellt, eine hohe Dynamik bei der Entwicklung neuer Approximationsmethoden entstand jedoch erst mit den Arbeiten [56, 138]. Schwerpunkte sind unter anderem:

- die Entwicklung stabiler und schneller Approximationsmethoden, die topologieunabhängig funktionieren,
- die Berücksichtigung von Störungen,
- die Reduktion mittels zusätzlicher Regularisierungen und
- die Erweiterung der Methoden auf $s \geq 4$.

In diesem Kapitel wurden die aktuell wichtigsten und aus Veröffentlichungen bekannten Methoden zur Approximation der Mengen zulässiger Lösungen für $s \in \{2, 3, 4\}$ detailliert vorgestellt und analysiert. Eine Methode ist allgemein für $s \geq 2$ anwendbar. Für $s = 3$ teilen sich die Methoden in zwei Typen auf: geometrisch konstruktive und numerisch approximative. Der Vorteil des geometrisch konstruktiven Ansatzes ist der geringe Rechenaufwand und die exakte Bestimmung von zulässigen Lösungen auf dem Rand der Mengen zulässiger Lösungen. Jedoch ist die Erweiterung der genutzten Konstruktionen auf störungsbehaftete Daten nur unter Abstrichen möglichen [86–88] und die Methode lässt sich nicht direkt für $s \geq 4$ erweitern. Demgegenüber sind die numerischen Methoden problemlos auf gestörte Daten anwendbar. Es wurden adaptiv gesteuerte und nicht adaptiv gesteuerte Methoden vorgestellt, wobei die adaptiv gesteuerten (Polygon inflation und adaptiver Strahlenalgorithmus) von der Theorie her am vielversprechendsten sind. Die Erweiterung der numerischen Methoden für $s \geq 4$ ist auf verschiedene Weisen möglich, wobei es nicht bei allen Methoden problemlos ist. So ist etwa die Dreieckseinschließungsmethode nur indirekt (in Form einer scheibenweisen Approximation und mit Einschränkungen bezüglich der Topologie der Mengen zulässiger Lösungen) erweiterbar. Bei der Polyhedron inflation Methode gilt es, die Implementierung abzuschließen und die Behinderung der adaptiven Steuerung durch einspringende Kanten einzuschränken. Praktisch ist einzig der Strahlenalgorithmus ohne Probleme für beliebiges $s \geq 4$ einsetzbar, wobei eine adaptive Steuerung mitunter nicht einfach zu implementieren ist.

Letztlich sei auch die Möglichkeit der Kombination aus der direkten geometrischen Bewertung eines $x \in \mathbb{R}^2$ mit der adaptiven Steuerung des Polygon inflation Algorithmus als eine vielversprechende Methode zur Berechnung der Mengen zulässiger Lösungen für $s = 3$ angeführt. Dabei würde die numerische Klassifizierung eines $x$ durch die geometrische ersetzt, die adaptive Verfeinerung der Randapproximation aber beibehalten werden. So würde die Auflösung des Randes nicht mehr durch die, vor der Berechnung festgelegte, Wahl der Tangenten bestimmt werden. Stattdessen würde die Auflösung adaptiv durch die Polygonverfeinerung gesteuert werden. Zudem wird der rechenintensive Teil der numerischen Klassifizierung mittels der Lösung vieler Optimierungsprobleme durch die schnelle Klassifizierung mittels konstruierter Dreiecke ersetzt. Die geometrische Konstruktion wird so um die Adaptivität erweitert. Die Anwendung auf gestörte Daten gelingt wie in Abschnitt 4.5.7.

# 5 Numerische Resultate

In Kapitel 4 sind einige Methoden zur Approximation der Mengen zulässiger Lösungen erläutert und analysiert. Der Schwerpunkt liegt auf Methoden zur Berechnung von $\mathcal{M}_A$ und $\mathcal{M}_C$ für $s = 3$. In diesem Kapitel werden die vorgestellten Methoden im Hinblick auf die Qualität der Ergebnisse und den Rechenaufwand untersucht und verglichen. Für die meisten Vergleiche und Analysen werden die Approximationen an $\mathcal{M}_A$ für die Datensätze 2 und 3 herangezogen.

Der Modelldatensatz 2 ist gut geeignet, da sich die Randpunkte von $\mathcal{M}_A$ bis auf Rundungsfehler exakt bestimmen lassen. Eine beliebig feine Diskretisierung von $\mathcal{M}_A$ ist zugänglich und objektive Vergleiche zwischen den Ergebnissen der unterschiedlichen Verfahren sind möglich. Für die Polygon inflation Methoden wird eine detaillierte Analyse vorgenommen. Der Datensatz 3 (IR-Daten zur Hydroformylierung) ist für einen Vergleich der Klassifizierungsmethoden insofern gut geeignet, da er Störungen mittleren Grades enthält und mittels kinetischer Modellierung eine, als korrekt angesehene Lösung, berechnet werden kann.

In Abschnitt 5.1 werden die geometrische Konstruktion, der Dreieckseinschließungsalgorithmus, die Polygon inflation Methoden sowie der Strahlenalgorithmus (starr sowie adaptiv gesteuert) eingehend in Bezug auf die Anwendung auf Datensatz 2 untersucht. In Abschnitt 5.2 werden die Polygon inflation Methoden in Bezug auf ihre Arbeitsweisen detailliert analysiert. Anschließend werden in Abschnitt 5.3 die Herangehensweisen zur Bestimmung von $\mathcal{M}_A$ für störungsbehaftete Daten am Beispiel des Datensatzes 3 untersucht. Zusätzlich werden Effekte analysiert, die die Berechnung von $\mathcal{M}_A$ behindern. In Abschnitt 5.4 wird am Beispiel des Datensatzes 3 gezeigt, wie sich mittels zusätzlicher Regularisierungen $\mathcal{M}_A$ und $\mathcal{M}_C$ zu Mengen zulässiger und relevanter Lösungen reduzieren lassen. Abschließend wird in Abschnitt 5.5 die Methode MCR-BANDS für die Datensätze 1 und 2 angewendet, um spezielle Lösungen als Indikatoren für die Vielfältigkeit möglicher Faktorisierungen zu bestimmen. Dazu werden die Ergebnisse und die jeweiligen Mengen zulässiger Lösungen in Bezug zueinander gebracht. Für die Darstellungen der Segmente von $\mathcal{M}_A$ und $\mathcal{M}_C$ sowie der zugehörigen Profile werden stets dieselben Farbzuordnungen wie in den Abbildungen 2.3 für Datensatz 1, 2.5 für Datensatz 2 und 2.7 für Datensatz 3 verwendet.

## 5.1 Methodenverifikation anhand des Datensatzes 2

Zu dem Datensatz 2 lassen sich die Randpunkte mittels der in Abschnitt 4.3 vorgestellten geometrischen Konstruktion bis auf Rundungsfehler exakt berechnen. Somit ist eine hochauflösende Diskretisierung der Randkurve zugänglich und die Abweichungen, der mit den einzelnen Verfahren berechneten Approximationen, können konkret bestimmt werden. Ein solches Vorgehen ist beispielsweise für Datensatz 3 nicht möglich und die einzelnen Approximationen können nur untereinander vergleichen werden.

### 5.1.1 Referenzlösung, Vergleichskriterien und weitere Details

Eine analytische Bestimmung der Randkurve von $\mathcal{M}_A$ ist wegen $s = 3$ für Datensatz 2 zwar unter großem Rechenaufwand möglich [16, 138], jedoch ist keine Implementierung dazu frei zugänglich. Um nichtsdestotrotz hinsichtlich der Güte der Approximationen sowie den benötigten

Rechenzeiten einen Vergleich zwischen den einzelnen Methoden auch in Abhängigkeit zu den jeweiligen Steuerparametern ziehen zu können, wird eine hoch genaue Approximation an den Rand von $\mathcal{M}_A$ genutzt. Diese Referenzlösung wird mittels der geometrischen Konstruktion (Tangentenalgorithmus) bestimmt. Die genutzten Tangenten (an $\mathcal{I}_A$) setzen sich aus drei verschiedenen Typen zusammen: es werden 10 000 äquiangular verteilte Tangenten an $\mathcal{I}_A$ genutzt, weiter werden die 100 Geraden, die die Kanten von $\mathcal{I}_A$ definieren ($\mathcal{I}_A$ besteht aus 100 Kanten), sowie 100 Tangenten an $\mathcal{I}_A$, die jeweils eine Ecke von $\mathcal{F}_A$ enthalten ($\mathcal{F}_A$ ist für den Datensatz durch 50 Kanten begrenzt), verwendet.

Die geometrische Konstruktion sichert die exakte Lage der berechneten Punkte auf dem Rand und die hohe Auflösung mit insgesamt 10 200 Tangenten sichert eine sehr feine Diskretisierung der Randkurve von $\mathcal{M}_A$. Für die geometrische Konstruktion wird eine vom Verfasser der vorliegenden Schrift in MatLab programmierte und nicht die in FacPack bereitgestellte Version (*Generalized Borgen Plots*-Modul) genutzt. Für die Analyse der geometrischen Konstruktion in Abschnitt 5.1.2 wird die in FacPack bereitgestellte Implementierung verwendet.

Die Methoden werden unter den Gesichtspunkten Approximationsgüte (Abstand zur Referenzlösung) und Rechenzeit zu verschiedenen Wahlen von Steuerparametern verglichen.

Als Abstandsmaß zwischen der hoch aufgelösten (und als Vergleichsergebnis herangezogenen) geometrischen Konstruktion von Randpunkten und den anderen Approximationen wird der Hausdorff-Abstand

$$\delta(X,Y) = \max\left(\max_{x\in X} D(x,Y),\ \max_{y\in Y} D(y,X)\right)$$

zweier Mengen $X$ und $Y$ genutzt mit

$$D(a,B) = \min_{b\in B}\left(\|a-b\|_2\right).$$

Um einen fairen Vergleich zwischen den Approximationsmethoden bezüglich ihrer Rechenzeiten ziehen zu können, werden für die rechenintensiven Teile aller Algorithmen schnelle C-Programme eingesetzt. Für die geometrische Konstruktion sowie die Polygon inflation Methoden erfolgen die Vergleichsrechnungen mit den FacPack-Implementierungen. Die Rechenzeiten wurden auf einem Standard-PC mit einem 2.4GHz Intel Prozessor und 16 GB RAM gemessen, wobei stets nur ein Kern genutzt wurde.

In den Tabellen werden jeweils auch die Anzahlen der Diskretisierungspunkte (kumuliert für alle Segmente) für die Approximationen an $\mathcal{M}_A$ angegeben, um in etwa die Umfänge der Diskretisierungen einschätzen zu können. Bei der Bewertung dieser Anzahlen gilt es zu beachten, dass die adaptiv gesteuerten Methoden im Normalfall bei qualitativ gleichwertigen Approximationen deutlich weniger Diskretisierungspunkte benötigen.

### 5.1.2 Geometrische Konstruktion

Als erstes werden die Ergebnisse der geometrischen Konstruktion, siehe Abschnitt 4.3, für verschiedene Anzahlen von Tangenten, das heißt mit verschiedenen Auflösungen des Rotationswinkels, ausgewertet. In der FacPack-Implementierung wird die Genauigkeit über den Rotationswinkel der Tangenten gesteuert. Begonnen wird mit einer Seite von $\mathcal{I}_A$ als Tangente. Anschließend wird mit einer äquiangularen Diskretisierung mit $\lceil 360/\alpha \rceil$ weiteren Tangenten fortgeführt, wobei $\alpha$ der gewählte Rotationswinkel ist. Zudem werden die verbleibenden Seiten von $\mathcal{I}_A$ ebenfalls als Tangenten genutzt.

Die Abweichungen der Berechnungen von der Referenzlösung sind in Tabelle 5.1 für verschiedene Rotationswinkel (und somit auch verschiedene Anzahlen von Tangenten) aufgelistet. Auffällig ist die hohe Effizienz sowie, dass sich die Rechenzeit bei Verbesserung der Genauigkeit kaum erhöht.

| Rotations-winkel [°] | Anzahl Tangenten | Rechen-zeit [s] | Approximationsgüte Segment 1 | Segment 2 | Segment 3 | Anzahl Punkte |
|---|---|---|---|---|---|---|
| 1.00 | 459 | 1.64 | $2.377 \cdot 10^{-3}$ | $2.166 \cdot 10^{-3}$ | $2.166 \cdot 10^{-3}$ | 73 |
| 0.50 | 819 | 1.64 | $8.785 \cdot 10^{-4}$ | $9.765 \cdot 10^{-4}$ | $9.765 \cdot 10^{-4}$ | 115 |
| 0.10 | 3699 | 1.77 | $2.463 \cdot 10^{-4}$ | $7.732 \cdot 10^{-5}$ | $7.732 \cdot 10^{-5}$ | 449 |
| 0.05 | 7299 | 1.94 | $2.463 \cdot 10^{-4}$ | $2.790 \cdot 10^{-5}$ | $2.790 \cdot 10^{-5}$ | 866 |

**Tabelle 5.1:** Anwendung der geometrischen Konstruktion zur Berechnung von $\mathcal{M}_A$ für Datensatz 2. Die Approximation ist für verschiedene Rotationswinkel, siehe die FacPack-Implementierung, durchgeführt.

| Seitenlänge $a$ | Rechen-zeit [s] | Approximationsgüten Segment 1 | Segment 2 | Segment 3 | Anzahl Dreiecke |
|---|---|---|---|---|---|
| $2.5 \cdot 10^{-2}$ | 0.59 | $2.485 \cdot 10^{-2}$ | $3.077 \cdot 10^{-2}$ | $2.145 \cdot 10^{-2}$ | 467 |
| $1.0 \cdot 10^{-2}$ | 1.23 | $9.937 \cdot 10^{-3}$ | $1.064 \cdot 10^{-2}$ | $9.240 \cdot 10^{-3}$ | 1 175 |
| $5.0 \cdot 10^{-3}$ | 2.39 | $4.985 \cdot 10^{-3}$ | $5.665 \cdot 10^{-3}$ | $4.889 \cdot 10^{-3}$ | 2 341 |
| $1.0 \cdot 10^{-3}$ | 10.67 | $9.977 \cdot 10^{-4}$ | $9.923 \cdot 10^{-4}$ | $1.055 \cdot 10^{-3}$ | 11 721 |
| $5.0 \cdot 10^{-4}$ | 20.47 | $4.990 \cdot 10^{-4}$ | $5.915 \cdot 10^{-4}$ | $4.971 \cdot 10^{-4}$ | 23 445 |
| $1.0 \cdot 10^{-4}$ | 95.12 | $9.988 \cdot 10^{-5}$ | $1.258 \cdot 10^{-4}$ | $1.076 \cdot 10^{-4}$ | 117 211 |

**Tabelle 5.2:** Anwendung der Dreieckseinschließung zur Berechnung von $\mathcal{M}_A$ für Datensatz 2. Die Methode ist für verschiedene Seitenlängen $a$ ausgeführt. Die Werte zeigen, dass die Approximationsgüte etwas schlechter sein kann als die vorgegebene Seitenlänge (=erwartete Randgenauigkeit), siehe Bemerkung 4.20.

### 5.1.3 Algorithmus der Dreieckseinschließung

Der Dreieckseinschließungsalgorithmus ist in Abschnitt 4.4 sowie [55,56,58] vorgestellt und wird auf den Datensatz 2 angewendet. Die Genauigkeit der Approximation wird über die Seitenlänge $a$ der Dreiecke festgelegt. In der Regel entspricht dies auch in etwa dem maximalen Fehler, wobei sich auch größere Fehler ergeben können, siehe Bemerkung 4.20. Um die Vergleiche mit den Polygon inflation Methoden sowie dem Strahlenalgorithmus sinnvoll durchzuführen, wird zur Klassifizierung eines $x$ die Funktion $F(x)$ aus (4.3) mit $f(x, S)$ aus (4.2) genutzt und nicht die in [58] vorgeschlagene Funktion ssq aus (4.6).

Die Ergebnisse der Auswertungen zur Anwendung mit verschiedenen Seitenlängen $a$ sind in Tabelle 5.2 aufgelistet.[1] Die Unterschiede zur geometrischen Konstruktion bezüglich der Rechenzeiten bei vergleichbarer Genauigkeit sind deutlich. Auch tritt der in Bemerkung 4.20 beschriebene Fall auf und es sind einzelne Abweichungen größer als die vorgegebene Genauigkeit (Seitenlänge). Die Methode liefert die Ergebnisse stabil und mit verlässlicher Genauigkeit. Der annähernd proportionale Zusammenhang zwischen der Genauigkeit der Approximation und der Anzahl der berechneten Dreiecke ist ersichtlich.

### 5.1.4 Polygon inflation Algorithmus

Als nächstes wird der in Abschnitt 4.5 und beispielsweise in [55,147,152,154] vorgestellte Polygon inflation Algorithmus (direkter Typ) auf den Datensatz 2 angewendet. Zur Analyse der Methode werden verschiedene Genauigkeiten $\varepsilon_b$ (Genauigkeit der Randapproximation) und $\delta$ (Abbruchkriterium) gewählt. Für die Anwendung ist zu beachten, dass die Rechnungen mit $\varepsilon_f = 10^{-14}$ und nicht mit dem in FacPack vorgewähltem $\varepsilon_f = 10^{-10}$ durchgeführt wurden. Dies ist für

[1] Die zur Initialisierung (Start vom Inneren eines Segments und Suche eines geeigneten Dreiecks, das den Rand überdeckt) notwendigen Funktionsaufrufe und die dafür benötigte Rechenzeit werden nicht berücksichtigt.

| $\varepsilon_b$, $\delta$ | Rechen- | Approximationsgüten | | | Anzahl |
|---|---|---|---|---|---|
| | zeit [s] | Segment 1 | Segment 2 | Segment 3 | Punkte |
| $10^{-2}$ | 0.66 | $1.746 \cdot 10^{-2}$ | $1.269 \cdot 10^{-2}$ | $1.665 \cdot 10^{-2}$ | 86 |
| $10^{-3}$ | 1.83 | $1.533 \cdot 10^{-3}$ | $2.289 \cdot 10^{-3}$ | $3.241 \cdot 10^{-3}$ | 184 |
| $10^{-4}$ | 4.54 | $1.849 \cdot 10^{-4}$ | $2.693 \cdot 10^{-4}$ | $4.079 \cdot 10^{-4}$ | 324 |
| $10^{-5}$ | 12.62 | $3.007 \cdot 10^{-5}$ | $8.140 \cdot 10^{-5}$ | $7.400 \cdot 10^{-5}$ | 650 |
| $10^{-6}$ | 31.29 | $1.958 \cdot 10^{-5}$ | $1.107 \cdot 10^{-5}$ | $2.274 \cdot 10^{-5}$ | 1747 |

**Tabelle 5.3:** Anwendung des Polygon inflation Algorithmus zur Berechnung von $\mathcal{M}_A$ für Datensatz 2. Die Approximation ist für verschiedene Steuerparameter $\delta$ und $\varepsilon_b$ durchgeführt. Deren effiziente Wirkung ist gut erkennbar. Lediglich ab einer Genauigkeit von $\varepsilon_b = \delta = 10^{-6}$ gibt es insofern Schwierigkeiten, als dass sich die Approximationsgüten nicht so stark verbessern, wie es die Wahl der Steuerparameter erwarten lässt.

| $\varepsilon_b$, $\delta$ | Rechen- | Approximationsgüten | | | Anzahl Punkte | | |
|---|---|---|---|---|---|---|---|
| | zeit [s] | Segment 1 | Segment 2 | Segment 3 | für $\mathcal{F}_A$ | für $\mathcal{M}_A^*$ | für $\mathcal{M}_A$ |
| $10^{-2}$ | 2.37 | $1.843 \cdot 10^{-2}$ | $1.330 \cdot 10^{-2}$ | $1.214 \cdot 10^{-2}$ | 48 | 88 | 68 |
| $10^{-3}$ | 4.89 | $1.114 \cdot 10^{-3}$ | $1.729 \cdot 10^{-3}$ | $9.797 \cdot 10^{-4}$ | 76 | 156 | 115 |
| $10^{-4}$ | 11.02 | $1.592 \cdot 10^{-4}$ | $1.551 \cdot 10^{-4}$ | $1.363 \cdot 10^{-4}$ | 124 | 290 | 216 |
| $10^{-5}$ | 25.59 | $1.176 \cdot 10^{-5}$ | $4.864 \cdot 10^{-5}$ | $3.019 \cdot 10^{-5}$ | 212 | 576 | 433 |
| $10^{-6}$ | 59.79 | $2.457 \cdot 10^{-6}$ | $4.313 \cdot 10^{-6}$ | $1.431 \cdot 10^{-6}$ | 342 | 1354 | 928 |

**Tabelle 5.4:** Kennwerte zur Anwendung des inversen Polygon inflation Algorithmus zur Berechnung von $\mathcal{M}_A$ für Datensatz 2. Für dieses Beispiel besteht $\mathcal{M}_A$ aus drei isolierten Segmenten. Daher entspricht jeweils die Summe der Anzahlen der Punkte für die Approximationen von $\mathcal{F}_A$ und von $\mathcal{M}_A^*$ nicht der Anzahl der Punkte für die Approximation von $\mathcal{M}_A$.

ungestörte Modelldaten mitunter entscheidend, sofern, wie es in Datensatz 2 der Fall ist, einige Einträge in möglichen Faktoren $C$ und $A$ sehr klein sind.

Die Kennwerte zu den Berechnungen sind in Tabelle 5.3 aufgelistet. Die Effizienz der adaptiven Steuerung ist erkennbar. Im Vergleich zur geometrischen Konstruktion zeigt sich die Polygon inflation Methode bei der Anwendung auf Datensatz 2 leicht im Nachteil. Dies ist für Modelldaten oft der Fall. Die Genauigkeit lässt sich mit dem Polygon inflation Algorithmus besser steuern, vergleiche Tabelle 5.1 und die Unterschiede in den maximalen Fehlern für die einzelnen Segmente.

### 5.1.5 Inverser Polygon inflation Algorithmus

Der inverse Polygon inflation Algorithmus wird mit verschiedenen Genauigkeiten $\varepsilon_b$ und $\delta$ zur Approximation von $\mathcal{M}_A$ angewendet und die Kennzahlen sind in Tabelle 5.4 angegebenen. Die Rechnungen wurden ebenfalls mit $\varepsilon_f = 10^{-14}$ und nicht mit dem in FacPack vorgewähltem $\varepsilon_f = 10^{-10}$ durchgeführt. Da $\mathcal{M}_A$ als Schnitt von $\mathcal{F}_A$ und $\mathcal{M}_A^*$ bestimmt wird, sind zusätzlich die Anzahlen der Punkte von deren Randdiskretisierungen angegeben. Zur Bestimmung von $\mathcal{M}_A^*$ ist als Steuerparameter $\varepsilon_{\text{out}} = 0.05$ gewählt, siehe (4.23).

Bezüglich der Rechenzeiten sind sowohl deutliche Unterschiede zum direkten Polygon inflation Algorithmus als auch die effektiv arbeitende adaptive Steuerung erkennbar. Der direkte Polygon inflation Algorithmus ist zur Approximation einer Menge $\mathcal{M}_A$ mit drei klar getrennten Segmenten besser geeignet.

| Anzahl Richtungen $N$ | $\varepsilon_b$ | Rechen-zeit [s] | Approximationsgüten Segment 1 | Segment 2 | Segment 3 |
|---|---|---|---|---|---|
| 100 | $10^{-3}$ | 2.20 | $6.751 \cdot 10^{-2}$ | $1.825 \cdot 10^{-2}$ | $1.404 \cdot 10^{-2}$ |
| 200 | $10^{-3}$ | 3.61 | $3.382 \cdot 10^{-2}$ | $1.825 \cdot 10^{-2}$ | $3.219 \cdot 10^{-3}$ |
| 500 | $10^{-4}$ | 9.96 | $9.030 \cdot 10^{-3}$ | $8.038 \cdot 10^{-3}$ | $2.919 \cdot 10^{-3}$ |
| 1000 | $10^{-4}$ | 18.37 | $8.706 \cdot 10^{-3}$ | $2.899 \cdot 10^{-3}$ | $8.032 \cdot 10^{-4}$ |
| 2500 | $10^{-5}$ | 50.49 | $3.680 \cdot 10^{-3}$ | $1.869 \cdot 10^{-3}$ | $7.578 \cdot 10^{-4}$ |
| 4000 | $10^{-6}$ | 91.51 | $2.161 \cdot 10^{-3}$ | $3.207 \cdot 10^{-4}$ | $3.317 \cdot 10^{-4}$ |

**Tabelle 5.5:** Anwendung des Strahlenalgorithmus ohne adaptive Verfeinerung für verschiedene Anzahlen $N$ von Strahlen und bei dazu geeigneter Wahl für die Randgenauigkeit $\varepsilon_b$ bei Anwendung auf Datensatz 2.

| $\varepsilon_b$ | Anzahl Strahlen $N_{\text{basic}}$ | $N$ (gesamt) | Rechen-zeit [s] | Approximationsgüten Segment 1 | Segment 2 | Segment 3 |
|---|---|---|---|---|---|---|
| $1 \cdot 10^{-3}$ | 100 | 254 | 5.58 | $9.577 \cdot 10^{-4}$ | $2.255 \cdot 10^{-3}$ | $2.028 \cdot 10^{-3}$ |
| $5 \cdot 10^{-4}$ | 100 | 316 | 7.29 | $9.577 \cdot 10^{-4}$ | $8.979 \cdot 10^{-4}$ | $7.117 \cdot 10^{-4}$ |
| $1 \cdot 10^{-4}$ | 100 | 466 | 14.92 | $9.654 \cdot 10^{-5}$ | $2.400 \cdot 10^{-4}$ | $2.789 \cdot 10^{-4}$ |
| $1 \cdot 10^{-5}$ | 200 | 1016 | 55.35 | $1.915 \cdot 10^{-5}$ | $3.850 \cdot 10^{-5}$ | $5.047 \cdot 10^{-5}$ |

**Tabelle 5.6:** Anwendung des Strahlenalgorithmus mit einfacher adaptiver Steuerung für verschiedene Anzahlen $N_{\text{basic}}$ von Strahlen vor der adaptiven Verfeinerung (Grundsatz) sowie verschiedene Wahlen für die Randgenauigkeit $\varepsilon_b$ bei Anwendung auf Datensatz 2. Der Vergleich der Approximationsgüten mit denen des Strahlenalgorithmus ohne adaptive Steuerung, siehe Tabelle 5.5, zeigt die Effizienz der zusätzlichen Strahlen.

### 5.1.6 Strahlenmethode (konservativ und adaptiv gesteuert)

Als nächstes wird die in Abschnitt 4.7 und in [157] vorgestellte Strahlenmethode zur Approximation von $\mathcal{M}_A$ für Datensatz 2 angewendet. Zunächst wird die nicht adaptiv gesteuerte Variante für verschiedene Anzahlen an Strahlen genutzt. Die Ergebnisse sind in Tabelle 5.5 aufgelistet. Deutlich wird, dass eine hinreichende Genauigkeit bei der Approximation von Ecken des Randes von $\mathcal{M}_A$ nur mit einer hohen Anzahl an Strahlen gelingt. Nichtsdestotrotz ist die Approximation mit $N = 500$ Strahlen insgesamt ausreichend.

Wie sehr die adaptive Steuerung die Approximationsgüten verbessert, zeigt Tabelle 5.6. Hierbei wird nach einem Grundstock von $N_{\text{basic}} = 100$ beziehungsweise $N_{\text{basic}} = 200$ Strahlen und mit unterschiedlichen Genauigkeiten $\varepsilon_b$ die einfache adaptive Steuerung aus Abschnitt 4.7.9 angewendet. Die maximalen Fehler und die Rechenzeiten sind mit denen des inversen Polygon inflation Algorithmus vergleichbar. In Anbetracht der Flexibilität der Methode ist dies im Hinblick auf $s \geq 4$ bemerkenswert.

### 5.1.7 Vergleiche bezüglich Rechenzeiten und Approximationsgüten

In Abbildung 5.1 sind einige der Kennwerte (maximaler Fehler und Rechenzeit) bei Anwendung der einzelnen Methoden zur Approximation von $\mathcal{M}_A$ für Datensatz 2 für die einzelnen Segmente doppelt logarithmisch dargestellt. Die Grafiken offenbaren den Vorsprung der geometrischen Konstruktion gegenüber allen numerischen Methoden für ungestörte Daten. Weiter wird unter den numerischen Methoden der Vorteil einer adaptiven Steuerung deutlich.

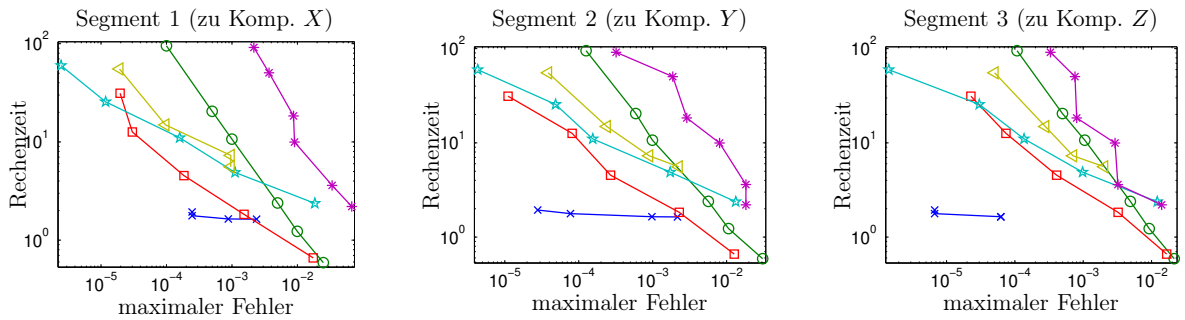


**Abbildung 5.1:** Doppelt logarithmische Darstellung von Rechenzeiten gegen Approximationsgüten zum Vergleich der Berechnungsmethoden bei der Anwendung auf Datensatz 2. Die Zuordnungen sind: geometrische Konstruktion (×), Dreieckseinschließungsalgorithmus (○), Polygon inflation Algorithmus (□), inverser Polygon inflation Algorithmus (☆), Strahlenmethode (✳), adaptiv gesteuerte Strahlenmethode (◁). Die Werte sind den Tabellen 5.1 bis 5.6 entnommen und nach Segmenten unterteilt. Die Dreieckseinschließungs- und die nicht adaptiv gesteuerte Strahlenmethode fallen gegenüber den übrigen Methoden ab. Für ungestörte Modelldaten ist die geometrische Konstruktion unerreicht.

### 5.1.8 Ergebnisse klassischer Faktorisierungsmethoden für Datensatz 2

Abschließend werden nun verschiedene Algorithmen zur Bestimmung einer nichtnegativen Matrixfaktorisierung beziehungsweise -approximation für Datensatz 2 angewendet. Mit den einzelnen Methoden werden nichtnegative Matrixfaktorisierung berechnet. Anschließend werden zu den einzelnen Profilen der berechneten Faktoren die niedrigdimensionalen Darstellungen bestimmt und zusammen mit den Mengen zulässiger Lösungen dargestellt. Die einzelnen Algorithmen werden je 200 mal ausgeführt. Die Einträge der jeweiligen Startiterierten $C^{(0)}$ und $A^{(0)}$ werden als standardnormalverteilte Pseudozufallszahlen in $(0, 1)$ in MatLab bestimmt.[2] Genutzt wird das MatLab-Softwarepaket „Nonnegative Matrix and Tensor Factorization Algorithms Toolbox“ [93, 95, 96]. Aus diesem werden die Routinen

(a) ANLS with block principal pivoting method (mit tol= $10^{-8}$, maxiter= 30 000),

(b) ANLS with active set method and given updating (mit tol= $10^{-8}$, maxiter= 10 000),

(c) ANLS with active set method and column grouping (mit tol= $10^{-8}$, maxiter= 10 000),

(d) alternating least squares method (mit tol= $10^{-8}$, maxiter= 30 000),

(e) hierarchical alternating least squares method (mit tol= $10^{-8}$, maxiter= 30 000),

(f) multiplicative updating method (mit tol= $10^{-8}$, maxiter= 10 000)

angewendet, vergleiche auch [26, 104, 106, 107]. Es ist anzumerken, dass keine der Methoden mit einer abgeschnittenen Singulärwertzerlegung arbeitet. Stattdessen werden die Faktoren frei bestimmt. Dies ist dadurch begründet, dass die Methoden auf die allgemeine Aufgabenstellung der Bestimmung einer nichtnegativen Niedrigrangapproximation ausgerichtet sind. In dieser Schrift wird jedoch das spezielle Faktorisierungsproblem betrachtet, dass es tatsächlich eine Faktorisierung gibt oder zumindest eine Approximation mit relativ kleinem Fehler. Die Resultate sind in Abbildung 5.2 dargestellt. Auffällig ist, dass teilweise einzelne niedrigdimensionale Darstellungen nicht in den Mengen $\mathcal{M}_A$ beziehungsweise $\mathcal{M}_C$ enthalten sind. Zudem sind in Tabelle 5.7 einige Kennwerte zu den Iterationen aufgelistet.

[2] Diese Wahl ist für $k = 100$, $n = 400$ und $s = 3$ in Bezug auf eine Konvergenz in wenigen Iterationen sehr ungünstig. Methoden, die auf einer Singulärwertzerlegung von $D$ basieren, benötigen bereits zu einer Startiterierten $T = I_s$ oft weit weniger Schritte.

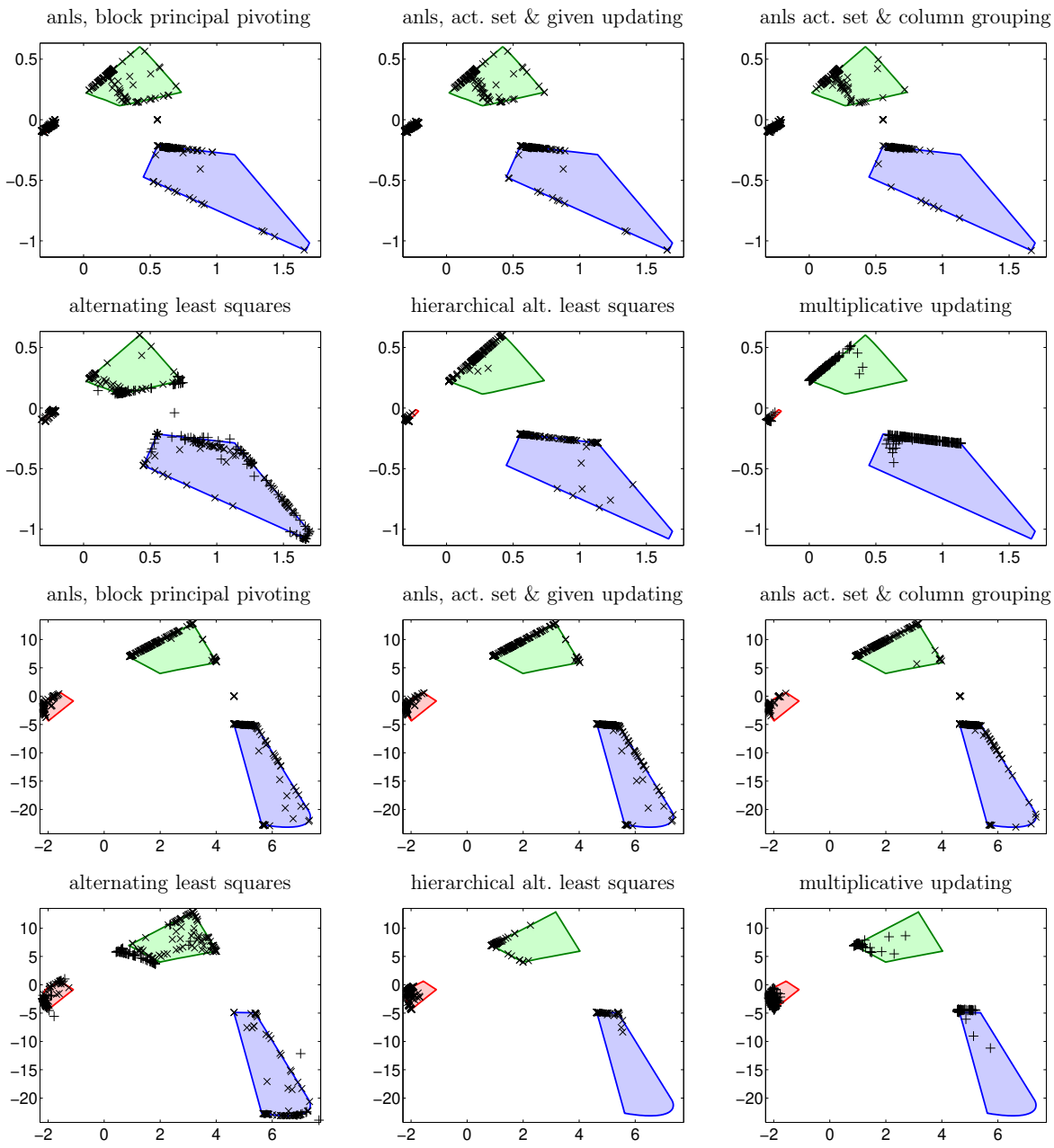


**Abbildung 5.2:** Resultate klassischer Methoden zur Bestimmung einer nichtnegativen Matrixfaktorisierung für Datensatz 2. Die niedrigdimensionalen Darstellungen der einzelnen Zeilen von $A$ und Spalten von $C$ sind in den Mengen zulässiger Lösungen $\mathcal{M}_A$ (obere beiden Reihen) und $\mathcal{M}_C$ (untere beiden Reihen) eingezeichnet ($+$: Abbruch bei Erreichen der maximalen Anzahl an Iterationen, $\times$: Abbruch davor). Jede Methode wurde 200 mal ausgeführt, wobei die Elemente der Startiterationen standardnormalverteilte Pseudozufallszahlen in $(0, 1)$ sind.

## 5.2 Zwischenresultate der Polygon inflation Methoden

Da die Polygon inflation Methoden einen Schwerpunkt dieser Arbeit bilden, erfolgt in diesem Abschnitt eine detaillierte Analyse der Arbeitsweisen der Methoden (direkter und inverser Typ). Zur Auswertung werden die Berechnungen von $\mathcal{M}_A$ für Datensatz 2 mit den Steuerparametern $\varepsilon_b = 10^{-4}$ und $\delta = 10^{-4}$ sowie $\varepsilon_f = 10^{-14}$ herangezogen.

Der Wert $\max_i \Delta_i$ mit $\Delta_i$ aus (4.28) wird als Kennwert für das Abbruchkriterium $\delta = 10^{-4}$ genutzt. Die Iteration der Kantenverfeinerung wird solange fortgeführt, wie $\max_i \Delta_i \geq \delta$ gilt. In Abbildung 5.3 sind die Entwicklungen der Werte $\max_i \Delta_i$ über die Anzahlen der Kanten

| NMF-Alg. | Durchl. mit Abbruch vor Erreichen von maxiter | | | | alle Durchläufe | | |
|---|---|---|---|---|---|---|---|
| | Anzahl d. Durchläufe | Ant. daran mit $\|E\|_F^2 < 10^{-8}$ | durchschn. Anz. Iter. | durchschn. Rechenzeit | Anteil mit $\|E\|_F^2 < 10^{-8}$ | durchschn. Rechenzeit | durchschn. $\|E\|_F^2$ |
| (a) | 200 | 82.0 % | 3039 | 3.7 s | 82.0 % | 3.7 s | $1.4 \cdot 10^{-1}$ |
| (b) | 200 | 100.0 % | 2956 | 52.4 s | 100.0 % | 52.4 s | $1.9 \cdot 10^{-5}$ |
| (c) | 200 | 84.0 % | 2908 | 4.2 s | 84.0 % | 4.2 s | $1.2 \cdot 10^{-1}$ |
| (d) | 121 | 83.5 % | 6898 | 5.8 s | 53.0 % | 17.1 s | $2.5 \cdot 10^{-2}$ |
| (e) | 200 | 100.0 % | 7465 | 5.5 s | 100.0 % | 5.5 s | $5.1 \cdot 10^{-5}$ |
| (f) | 0 | – | – | – | 0.0 % | 7.5 s | $4.9 \cdot 10^{-3}$ |

**Tabelle 5.7:** Ergänzende Informationen zur Anwendung verschiedener klassischer Algorithmen zur Bestimmung von nichtnegativen Matrixfaktorisierungen auf Datensatz 2. Die Zuordnungen der Routinen sowie Details zu deren Anwendung sind in Abschnitt 5.1.8 erläutert. Jede Methode wurde insgesamt 200 mal mit zufälligen Startiterierten ausgeführt. Speziell aufgelistet sind die Werte für die Durchläufe, welche vor Erreichen der maximalen Iterationsanzahl terminierten. Abkürzend ist $E = D - CA$ gewählt, sodass $\|E\|_F^2 = \|D - CA\|_F^2$ angibt.

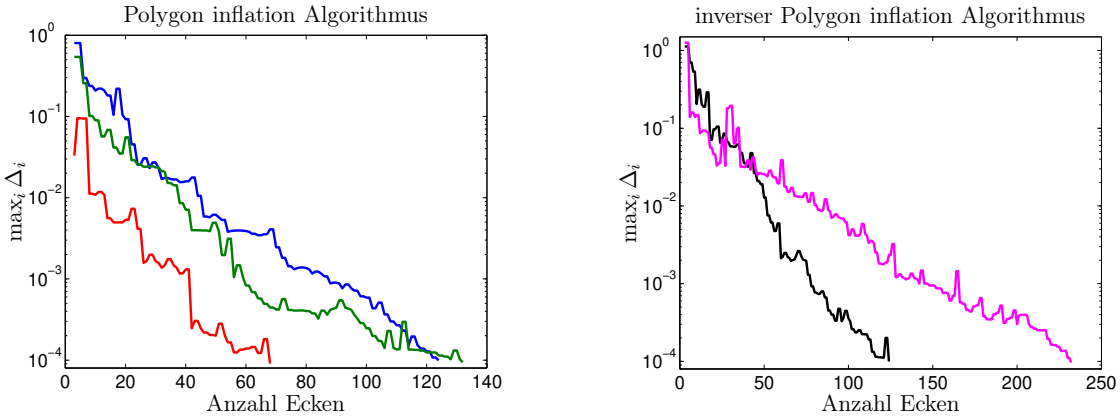


**Abbildung 5.3:** Der Verlauf des Werts $\max_i \Delta_i$ bei den Polygon inflation Algorithmen gegen die Anzahl der berechneten Ecken für Datensatz 2. Als Steuerparameter sind $\varepsilon_b = \delta = 10^{-4}$ gewählt. Der Wert $\max_i \Delta_i$ wird als Kennwert für das Abbruchkriterium genutzt. Die Iteration wird fortgeführt, solange $\max_i \Delta_i \geq \delta$ gilt. Links: für den Polygon inflation Algorithmus (blau: Komponente $X$, grün: Komponente $Y$, rot: Komponente $Z$). Rechts: für den inversen Polygon inflation Algorithmus (schwarz: Berechnung von $\mathcal{F}_A$, magenta: Berechnung von $\mathcal{M}_A^*$).

(Iterationen) für den direkten Polygon inflation Algorithmus (drei Segmente) und für den inversen Polygon inflation Algorithmus (Berechnungen von $\mathcal{F}_A$ und $\mathcal{M}_A^*$) dargestellt. In Abbildung 5.4 sind die, während der einzelnen Verfeinerungen erzeugten, neuen Referenzwerte ($\Delta'$ aus (4.28)) dargestellt. Teilweise tritt der Wert Null auf, der neu bestimmte Punkt liegt folglich auf der zu verfeinernden Kante. Gründe können sein, dass es sich bei dem zu approximierenden Abschnitt tatsächlich um einen Geradenabschnitt handelt oder dass der Abstand zum tatsächlichen Rand kleiner als die Genauigkeit $\varepsilon_b$ ist.

Der größte Anteil des Rechenaufwandes bei der Approximation des Randes von $\mathcal{M}_A$ mittels der Polygon inflation Methoden entfällt auf die Klassifizierungsroutinen. Dies ist bei anderen numerischen Methoden wie der Dreieckseinschließungs- oder der Strahlenmethode nicht anders. Hier liegt der große Vorteil der geometrischen Konstruktion, da die Bewertung konstruktiv ist und nur potentielle Randpunkte konstruiert werden.

Im Zuge der Klassifizierungen wird der Schnelltest zwar am häufigsten aufgerufen, dieser fällt in Bezug auf den Aufwand aber kaum ins Gewicht. Der Test, ob $F(x) < \varepsilon_f$ gilt, beinhaltet die Minimierung der Funktion $f(x, S)$ in Bezug auf $S$. Ein Aufruf von $f(x, S)$ ist zwar nicht sehr teuer, jedoch wird $f(x, S)$ sehr häufig aufgerufen. Dazu sind in Abbildung 5.5 (links) die Aufrufe des Schnelltests sowie die Aufrufe der Funktion $F(x)$ für die Polygon inflation Methoden

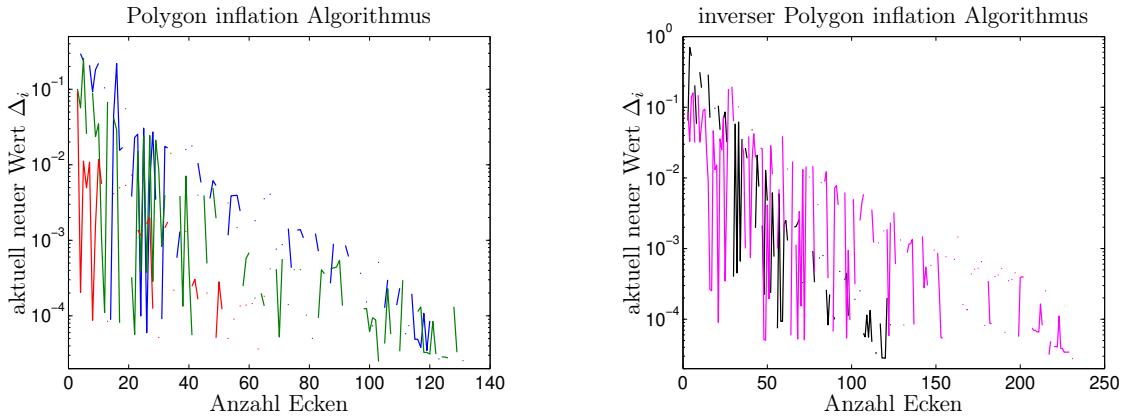


**Abbildung 5.4:** Die aktuell neu hinzugekommenen Einträge in $\Delta$ bei den Anwendungen der Polygon inflation Algorithmen gegen die Anzahl der berechneten Ecken für Datensatz 2. Als Steuerparameter sind $\varepsilon_b = \delta = 10^{-4}$ gewählt. Es wird $\max_i \Delta_i$ als Kennwert für das Abbruchkriterium genutzt, siehe auch Abbildung 5.3. Links: für den Polygon inflation Algorithmus (blau: Komponente $X$, grün: Komponente $Y$, rot: Komponente $Z$). Rechts: für den inversen Polygon inflation Algorithmus (schwarz: Berechnung von $\mathcal{F}_A$, magenta: Berechnung von $\mathcal{M}_A^*$).

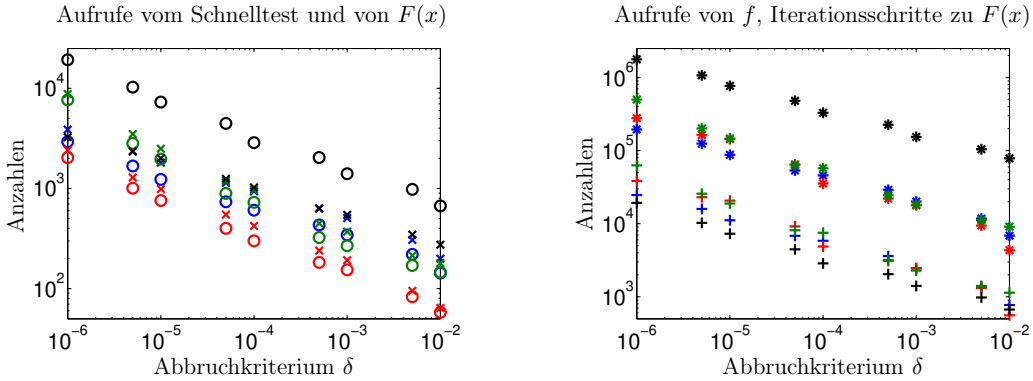


**Abbildung 5.5:** Kennwerte zur Klassifizierung von Punkten bei der Anwendung der Polygon inflation Algorithmen auf Datensatz 2 zu verschiedenen Werten $\delta$. Links: Anzahlen der Aufrufe des Schnelltests, ob ein $x$ in $\mathcal{F}_A$ liegt ($\times$), und die Anzahlen der Aufrufe der Funktion $F(x)$ aus (4.3) ($\circ$). Die Zuordnungen sind wie folgt: Polygon inflation Algorithmus: blau (Komponente $X$), grün ($Y$) und rot($Z$), inverser Polygon inflation Algorithmus: schwarz ($\times$ zur Bestimmung von $\mathcal{F}_A$ und $\circ$ zur Bestimmung von $\mathcal{M}_A^*$). Rechts: Anzahlen der Funktionsaufrufe von $f(x, S)$ aus (4.2) ($+$) und Anzahlen der Iterationsschritte zur Bestimmung von $F(x)$ ($*$). Anmerkung: die Anzahlen der Funktionsaufrufe zur Bestimmung von $\mathcal{M}_A^*$ ($*$) sind so hoch, weil zur Stabilisierung punktuell ein genetischer Algorithmus eingesetzt wird.

(direkter und inverser Typ) für die einzelnen Segmente beziehungsweise für die Berechnungen von $\mathcal{F}_A$ und $\mathcal{M}_A^*$ aufgetragen. In der gleichen Abbildung (rechts) sind jeweils die Aufrufe der Funktion $f(x, S)$ sowie die, bei der Minimierung von $f(x, S)$ mittels der ACM-Routine NL2SOL, ausgeführten Iterationsschritte eingetragen.

## 5.3 Analysen am Beispiel des Datensatzes 3

In diesem Abschnitt wird Datensatz 3 genutzt, um die Berechnung von $\mathcal{M}_A$ für gestörte Daten zu untersuchen. Zunächst werden drei verschiedene Klassifizierungsansätze angewendet. Die Resultate werden in Bezug auf die Auswirkungen der unterschiedlichen Berücksichtigungen von Störungen auf die Approximationen an die Mengen zulässiger Lösungen verglichen. Weiter werden die Approximationen an $\mathcal{M}_A$ und an $\mathcal{M}_C$ für Variationen der Steuerparameter in der Funktion $f(x, S)$ aus (4.2) betrachtet und im Speziellen die Einflüsse der Steuerparameter untersucht. Zudem werden die unterschiedlichen Sensivitäten der $w(:, i)$, $i = 1, \ldots, k$, und der $u(:, j)$,

$j = 1, \ldots, n$, in Bezug auf Störungen analysiert.

Der Datensatz 3 enthält signifikante Absorptionen von $s = 3$ Komponenten. Er ist für die Analysen geeignet, da

- der Datensatz mit $k = 1611$ und $n = 650$ mittleren Umfangs ist,
- die Mengen $\mathcal{M}_A$ und $\mathcal{M}_C$ aus drei, teilweise nicht klar separierten, Segmenten bestehen,
- die Störungen weder sehr klein noch sehr groß aber, aufgrund der Lage der $u(:, j)$, $j = 1, \ldots, n$, in $\mathcal{M}_C$, teilweise problematisch sind und nicht nur die Singulärvektoren $V(:, 1:3)$ sondern auch $V(:, 4)$ und $V(:, 5)$ teilweise signifikante Informationen enthalten, die aber bei der Berechnung von $\mathcal{M}_A$ keine Berücksichtigung finden sowie
- mittels kinetischer Modellierung eine, als korrekt angesehene, Lösung zum Vergleich der Ergebnisse zugänglich ist und zur Verifizierung der Ergebnisse genutzt werden kann.

Die niedrigdimensionalen Darstellungen der Konzentrationsprofile der mittels kinetischer Modellierung bestimmten Lösung werden mit $y_i^{(\mathrm{ode})} \in \mathbb{R}^2$, $i = 1, 2, 3$, bezeichnet. Analog werden die niedrigdimensionalen Darstellungen der Reinkomponentenspektren dieser Lösung mit $x_i^{(\mathrm{ode})} \in \mathbb{R}^2$, $i = 1, 2, 3$, bezeichnet. Die Werte

$$\frac{\min\limits_{i,j} D_{ij}}{\max\limits_{i,j} D_{ij}} = -7.9527 \cdot 10^{-3} \qquad \text{und} \qquad \sqrt{\frac{\sum_{i=4}^{650} \sigma_i^2}{\sum_{i=1}^{650} \sigma_i^2}} = 9.2686 \cdot 10^{-3}$$

sind als zusätzliche Informationen zu dem Datensatz, auch in Bezug auf die Wahl und die Einschätzung der Steuerparameter, von Interesse.

### 5.3.1 Die Mengen zulässiger Lösungen für gestörte Daten

Zunächst wird der Polygon inflation Algorithmus angewendet. Um die Störungen der Daten bei den Berechnungen von $\mathcal{M}_A$ und $\mathcal{M}_C$ zu respektieren, werden $\varepsilon_a = 2.5 \cdot 10^{-3}$ und $\varepsilon_c = 5 \cdot 10^{-2}$ als Steuerparameter gewählt. Weiter wird die Klassifizierungsroutine mit $\varepsilon_f = 10^{-10}$ angewendet. Als Steuerparameter (Genauigkeit der Randapproximation und Abbruchkriterium) werden $\varepsilon_b = \delta = 10^{-3}$ genutzt. In Abbildung 5.6 sind die Ergebnisse dargestellt. Zusätzlich sind rechts $w(:, i)$, $i = 1, \ldots, k$, und links $u(:, j)$, $j = 1, \ldots, n$, eingezeichnet. Die berechnete Approximation an $\mathcal{M}_A$ enthält die, zur kinetischen Lösung gehörigen, Punkte $x_i^{(\mathrm{ode})}$, $i = 1, 2, 3$. Weiter enthält die berechnete Approximation an $\mathcal{M}_C$ die Punkte $y_i^{(\mathrm{ode})}$, $i = 1, 2, 3$. In der linken Grafik ist zu erkennen, dass die geometrische Bedingung an eine zulässige Lösung aus Satz 3.26 für störungsbehaftete Daten auf leere Mengen $\mathcal{M}_A$ und $\mathcal{M}_C$ führen kann.

### 5.3.2 Variation der Steuerparameter

Als nächstes werden die für die Klassifizierungsroutine wichtigen Steuerparameter $\varepsilon_c$ und $\varepsilon_a$ von $f(x, S)$ aus (4.2) variiert. Die Ergebnisse werden mit dem Polygon inflation Algorithmus berechnet. Um die Einflüsse der zwei Steuerparameter erkennbar zu machen, werden beide getrennt voneinander modifiziert und die jeweiligen Approximationen an $\mathcal{M}_A$ und $\mathcal{M}_C$ in separaten Abbildungen dargestellt. In Abbildung 5.7 sind die Ergebnisse bei Variation von $\varepsilon_c$ präsentiert und in Abbildung 5.8 die bei Variation von $\varepsilon_a$.

Deutlich erkennbar sind in den Grafiken beider Abbildungen die jeweiligen Einflüsse von $\varepsilon_c$ und $\varepsilon_a$. Wird $\varepsilon_c$ betragsmäßig vergrößert (beziehungsweise verkleinert), so verschiebt sich hauptsächlich der äußere Rand von $\mathcal{M}_C$ nach außen (nach innen) und der innere Rand von $\mathcal{M}_A$ zum Ursprung (vom Ursprung weg), siehe Abbildung 5.7. Genau umgekehrt verhält es sich bei der

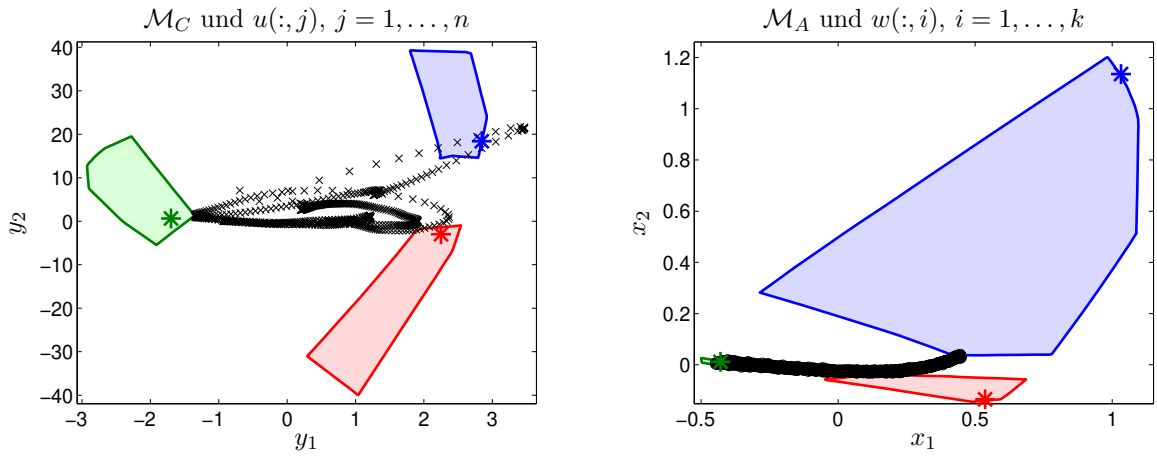


**Abbildung 5.6:** Die Mengen $\mathcal{M}_C$ (links) und $\mathcal{M}_A$ (mitte) für Datensatz 3, berechnet mit dem Polygon inflation Algorithmus und den Parametern $\varepsilon_b = \delta = 10^{-3}$, $\varepsilon_f = 10^{-10}$ sowie $\varepsilon_a = 2.5 \cdot 10^{-3}$ und $\varepsilon_c = 5 \cdot 10^{-2}$. Dazu eingetragen sind $w(:,i)$, $i = 1,\ldots,k$, (×) sowie $u(:,j)$, $j = 1,\ldots,n$, (○) und die niedrigdimensionalen Darstellungen $x_i^{(\mathrm{ode})}$ und $y_i^{(\mathrm{ode})}$, $i = 1,2,3$, der kinetischen Lösungen. Die Zuordnungen sind: Komponente $S$ (✳), Komponente $K$ (✳), Komponente $[S-K]$ (✳).

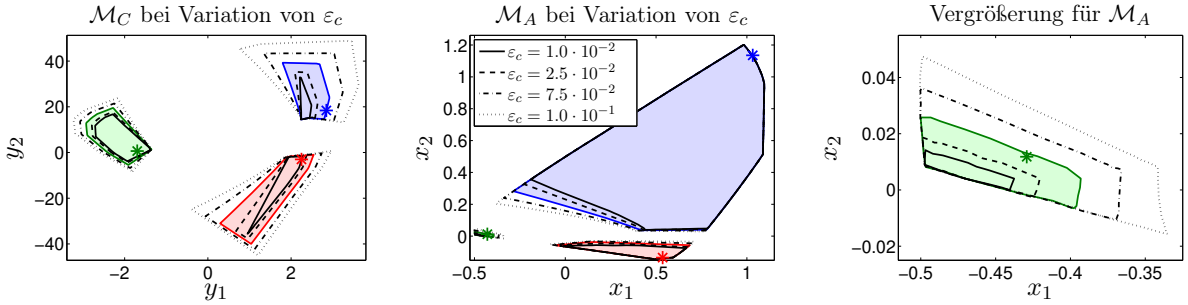


**Abbildung 5.7:** Die Mengen $\mathcal{M}_C$ (links) und $\mathcal{M}_A$ (mitte, rechts) für Datensatz 3 und unterschiedliche Werte $\varepsilon_c$. Die weiteren Parameter sind $\varepsilon_b = \delta = 10^{-3}$, $\varepsilon_f = 10^{-10}$ sowie $\varepsilon_a = 2.5 \cdot 10^{-3}$. Die farbig dargestellten Mengen sind mit $\varepsilon_c = 5 \cdot 10^{-2}$ berechnet. Dargestellt sind mittels ✳ ebenfalls die als korrekt angesehenen $x_i^{(\mathrm{ode})}$ und $y_i^{(\mathrm{ode})}$, $i = 1,2,3$, (kinetische Modellierung). Rechts: Vergrößerung des grünen Segments von $\mathcal{M}_A$.

betragsmäßigen Vergrößerung (beziehungsweise Verkleinerung) von $\varepsilon_a$: hauptsächlich verschiebt sich der innere Rand von $\mathcal{M}_C$ zum Ursprung (vom Ursprung weg) und der äußere Rand von $\mathcal{M}_A$ nach außen (nach innen), siehe Abbildung 5.8. Diese Effekte entsprechen den Erwartungen.

### 5.3.3 Resultate für verschiedene Klassifizierungsansätze

Ein entscheidender Schritt bei der Approximation von $\mathcal{M}_A$ beziehungsweise $\mathcal{M}_C$ ist die Klassifizierung eines $x \in \mathbb{R}^{s-1}$ beziehungsweise eines $y \in \mathbb{R}^{s-1}$ als *zulässig* oder *nicht zulässig*. Dazu sind in dieser Arbeit drei Möglichkeiten diskutiert: das geometrische Argument, siehe die Sätze 3.23 und 3.26, die Bewertung mittels der Funktion $f(x,S)$ aus (4.2) sowie die Bewertung mittels der Funktion $\mathsf{ssq}(x,S)$ aus (4.6). Für ungestörte Daten führen die verschiedenen Klassifizierungen bei geeigneter Wahl der Parameter und unter Vernachlässigung von Rundungsfehlern auf die gleichen Ergebnisse.

Für gestörte Daten liefern die Ansätze in der Regel leicht unterschiedliche Ergebnisse. Die geometrische Bewertung, wie in Satz 3.26 vorgestellt, führt für gestörte Daten oft zu $\mathcal{M}_A = \mathcal{M}_C = \emptyset$. Eine Modifikation des Ansatzes zur verallgemeinerten Geometrischen Konstruktion ist möglich [86–88] und in FacPack implementiert.

Anhand der Approximationen für $\mathcal{M}_A$ und $\mathcal{M}_C$ werden die drei Ansätze zur Klassifizierung eines

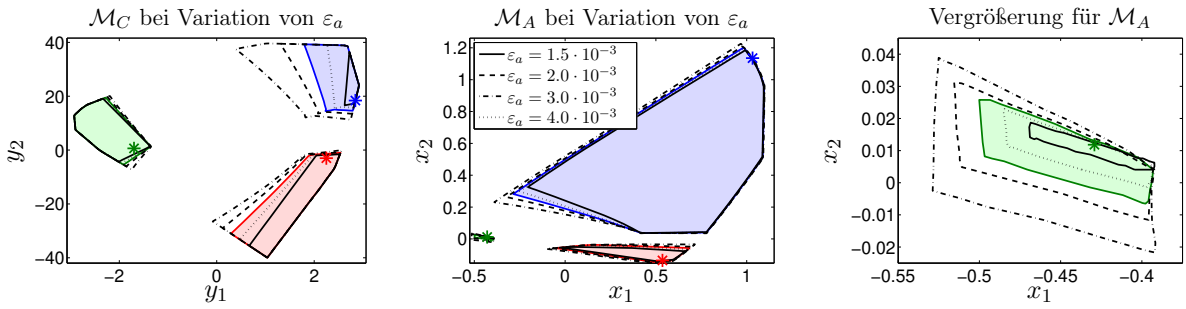


**Abbildung 5.8:** Die Mengen $\mathcal{M}_C$ (links) und $\mathcal{M}_A$ (mitte, rechts) für Datensatz 3 und unterschiedliche Werte $\varepsilon_a$. Die weiteren Parameter sind $\varepsilon_b = \delta = 10^{-3}$, $\varepsilon_f = 10^{-10}$ sowie $\varepsilon_c = 5 \cdot 10^{-2}$. Die farbig dargestellten Mengen sind mit $\varepsilon_a = 2.5 \cdot 10^{-3}$ berechnet. Dargestellt sind mittels $\ast$ ebenfalls die als korrekt angesehenen $x_i^{(\mathrm{ode})}$ und $y_i^{(\mathrm{ode})}$, $i = 1, 2, 3$ (kinetische Modellierung). Rechts: Vergrößerung des grünen Segments von $\mathcal{M}_A$.

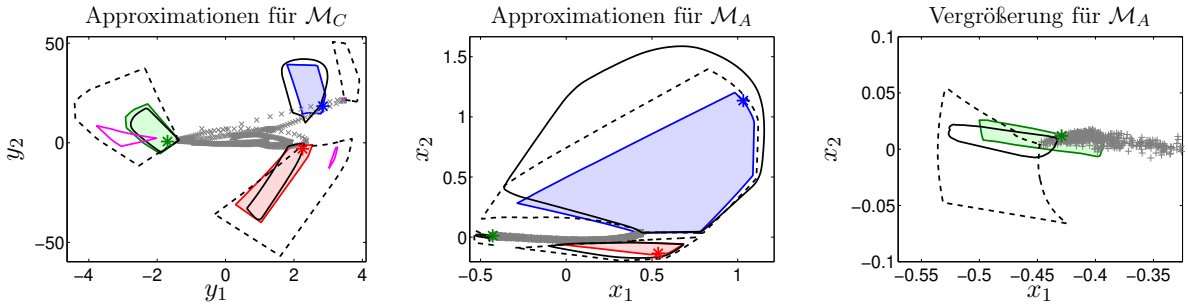


**Abbildung 5.9:** Approximationen an $\mathcal{M}_C$ (links) und $\mathcal{M}_A$ (mitte, rechts) für Datensatz 3 bei der Anwendung verschiedener Klassifizierungsansätze. Berechnet wurden die Approximationen mit dem Polygon inflation Algorithmus unter Nutzung der Funktion $f(x, S)$ (farbig), dem Polygon inflation Algorithmus unter Nutzung der Funktion ssq (schwarze durchgezogene Linien) sowie mit der verallgemeinerten geometrischen Konstruktion (gestrichelte Linien). In $\mathcal{M}_C$ sind zusätzlich $w(:, i)$, $i = 1, \ldots, k$, ($\times$) sowie $y_i^{(\mathrm{ode})}$, $i = 1, 2, 3$, ($\ast$, $\ast$ und $\ast$) eingetragen. Analoges gilt für $\mathcal{M}_A$ und $u(:, j)$, $j = 1, \ldots, n$, ($+$) sowie $x_i^{(\mathrm{ode})}$, $i = 1, 2, 3$ ($\ast$, $\ast$ und $\ast$). In der linken Grafik ist in magenta zusätzlich die Approximation an $\mathcal{M}_C$ für die Steuerparameter (eps_C=$10^{-8}$ und eps_A=$2.91 \cdot 10^{-4}$ in RS-Skalierung) eingezeichnet. Diese Parameter werden durch die FacPack-Implementierung *Generalized Borgen Plots* vorgewählt. Die Approximation enthält kein $y_i^{(\mathrm{ode})}$, $i = 1, 2, 3$.

$x$ verglichen. Die Details zu den Berechnungen der Approximationen sind:

- Die verallgemeinerte geometrische Konstruktion wird mittels der FacPack-Implementierung (Version 1.2, *Generalized Borgen Plots*-Modul) berechnet. Zur Approximation von $\mathcal{M}_A$ wird die RS-Skalierung [87, 146] mit den angebotenen Standardeinstellungen genutzt und anschließend das so erhaltene Ergebnis zu $\mathcal{M}_A$ transformiert. Die Approximation von $\mathcal{M}_C$ erfolgt mittels RS-Skalierung unter Nutzung der Parameter eps_C=$10^{-8}$ und eps_A=$4 \cdot 10^{-4}$ sowie anschließender Transformation.
- Die Funktion ssq$(x, S)$ aus (4.6) wird mit $\tilde{\varepsilon}_f = 6.2 \cdot 10^{-2}$ angewendet. Für die Berechnungen von $\mathcal{M}_A$ und $\mathcal{M}_C$ wird der Polygon inflation Algorithmus mit den Steuerparametern $\varepsilon_b = \delta = 10^{-3}$ verwendet.
- Die in den Implementierungen der Polygon inflation Methode und des Strahlenalgorithmus genutzte Funktion $f(x, S)$ aus (4.2) wird mit $\varepsilon_a = 2.5 \cdot 10^{-3}$, $\varepsilon_c = 5 \cdot 10^{-2}$, $\varepsilon_f = 10^{-10}$ sowie $\varepsilon_b = \delta = 10^{-3}$ angewendet.

In Abbildung 5.9 sind die so berechneten Approximationen an $\mathcal{M}_A$ und $\mathcal{M}_C$ dargestellt. Zusätzlich sind $x_i^{(\mathrm{ode})}$ und $y_i^{(\mathrm{ode})}$, $i = 1, 2, 3$, zur Verifizierung der Ergebnisse eingezeichnet. Weiter sind auch $w(:, i)$, $i = 1, \ldots, k$, in $\mathcal{M}_A$ und $u(:, j)$, $j = 1, \ldots, n$, in $\mathcal{M}_C$ zum besseren Verständnis der Arbeitsweise der verallgemeinerten geometrischen Konstruktion dargestellt.

Für $\mathcal{M}_A$ liefern alle drei Ansätze ähnliche Ergebnisse. Bei der Approximation von $\mathcal{M}_C$ führen

die beiden numerischen Klassifizierungsmethoden zu sehr ähnlichen Resultaten. Das Ergebnis der verallgemeinerten geometrischen Konstruktion ist grundsätzlich verschieden. Für die gewählten Parameter gilt $y_2^{(\mathrm{ode})}$, $y_3^{(\mathrm{ode})} \in \mathcal{M}_C$ und $y_1^{(\mathrm{ode})}$ ist deutlich außerhalb von $\mathcal{M}_C$. Um zu zeigen, dass dies nicht nur an der Parameterwahl liegt, ist die Approximation zusätzlich mit den vorgewählten Parametern ($\mathsf{eps\_C}{=}10^{-8}$ und $\mathsf{eps\_A}{=}2.91 \cdot 10^{-4}$; Berechnung in RS-Skalierung und anschließende Transformation) eingezeichnet. Die Approximation enthält kein $y_i^{(\mathrm{ode})}$, $i = 1, 2, 3$. Die Begründung hierfür liegt in der Konstruktion und in den $u(:, j)$, $j = 1, \ldots, n$. Einige dieser reagieren sehr sensitiv auf Störungen und liegen außerhalb der Approximation an $\mathcal{M}_C$ unter Verwendung der numerischen Klassifizierung. Insbesondere liegen einige der $u(:, j)$, $j = 1, \ldots, n$, außerhalb von $\mathcal{F}_C$. Dies erschwert die geometrische Konstruktion soweit, dass es für das blaue Segment in Abbildung 5.9 (links) einen leeren Schnitt zwischen der geometrisch konstruierten Approximation und den, mittels numerischer Klassifizierungen, berechneten Approximationen gibt.

**Bemerkung 5.1.** *Auch wenn, wie beispielsweise in den Grafiken aus Abbildung 5.9 demonstriert, die Klassifizierungsansätze teils sehr unterschiedliche Ergebnisse für $\mathcal{M}_A$ und $\mathcal{M}_C$ zur Folge haben, so kann rein objektiv keiner der Ansätze als richtig oder falsch eingestuft werden. Einzig kann als Anhaltspunkt dienen, ob die $x_i^{(ode)}$, $i = 1, 2, 3$, in $\mathcal{M}_A$ beziehungsweise die $y_i^{(ode)}$, $i = 1, 2, 3$, in $\mathcal{M}_C$ enthalten sind. Aber auch dies ist schwierig, da die, mittels kinetischer Modellierung bestimmten, Lösungen, Störungen in Form von negativen Einträgen enthalten und so den Mengen $\mathcal{M}_A$ und $\mathcal{M}_C$ widersprechen.*

### 5.3.4 Transformation zulässiger Lösungen zu möglichen Profilen

Letztlich sind die Mengen zulässiger Reinkomponentenspektren und zulässiger Konzentrationsprofile interessant. Dazu werden die Segmente der Mengen zulässiger Lösungen geeignet diskretisiert und die zugehörigen Profile $a = (1, x^{\mathbf{T}})V^{\mathbf{T}}$ sowie $c = U\Sigma(1, y^{\mathbf{T}})^{\mathbf{T}}$ bestimmt. Diese sind Elemente der Mengen $\mathcal{A}$ und $\mathcal{C}$. In den Abbildungen 5.10 und 5.11 sind mögliche Profile zu den Approximationen an die Mengen zulässiger Lösungen für verschiedene Ansätze zur Einbindung von Störungen aus Abbildung 5.9 dargestellt.

### 5.3.5 Sensitivitätsanalyse der Daten

Für die Berechnungen von $\mathcal{M}_A$ und $\mathcal{M}_C$ sind $w(:, i)$, $i = 1, \ldots, k$, und $u(:, j)$, $j = 1, \ldots, n$, wichtig. Insbesondere gilt dies für die geometrische Konstruktion. Dementsprechend ist es sinnvoll, den Einfluss von Störungen und negativen Einträgen in den Daten auf $w(:, i)$, $i = 1, \ldots, k$, und $u(:, j)$, $j = 1, \ldots, n$, zu untersuchen. Die Vorgehensweise der verallgemeinerten geometrischen Konstruktion [86–88] wird in dieser Schrift zwar nicht thematisiert, in Abbildung 5.9 sind jedoch Ergebnisse präsentiert, die mit dieser Methode generierte wurden.

Der Einfluss ergibt sich für die Konstruktion von $\mathcal{M}_A$ über die konvexe Hülle der $w(:, i)$, $i = 1, \ldots, k$, und für die Konstruktion von $\mathcal{M}_C$ über die konvexe Hülle der $u(:, j)$, $j = 1, \ldots, n$. In (3.23), siehe auch Bemerkung 3.77, ist ein Indikator zur Sensitivität der $w(:, i)$, $i = 1, \ldots, k$, vorgestellt. Ebenso ist in (3.22) ein Indikator zur Sensitivität der $u(:, j)$, $j = 1, \ldots, n$, eingeführt. Diese Indikatoren sind in Abbildung 5.12 für Datensatz 3 ausgewertet, wobei $e^{(u)}$ gegen den Vektor der Wellenzahlen dargestellt ist und $e^{(w)}$ gegen den Vektor der Messzeitpunkte. Für $e^{(u)}$ ergeben sich teils große Werte, was auf eine hohe Sensitivität einiger $u(:, j)$, $j = 1, \ldots, n$ hindeutet. Genau diese Sensitivität ist etwa in Abbildung 5.6 (links) erkennbar. Demgegenüber sind die Werte in $e^{(w)}$ eher gering. Dies deutet auf eine geringe Sensitivität der $w(:, i)$, $i = 1, \ldots, k$, hin, vergleiche Abbildung 5.6 (rechts).

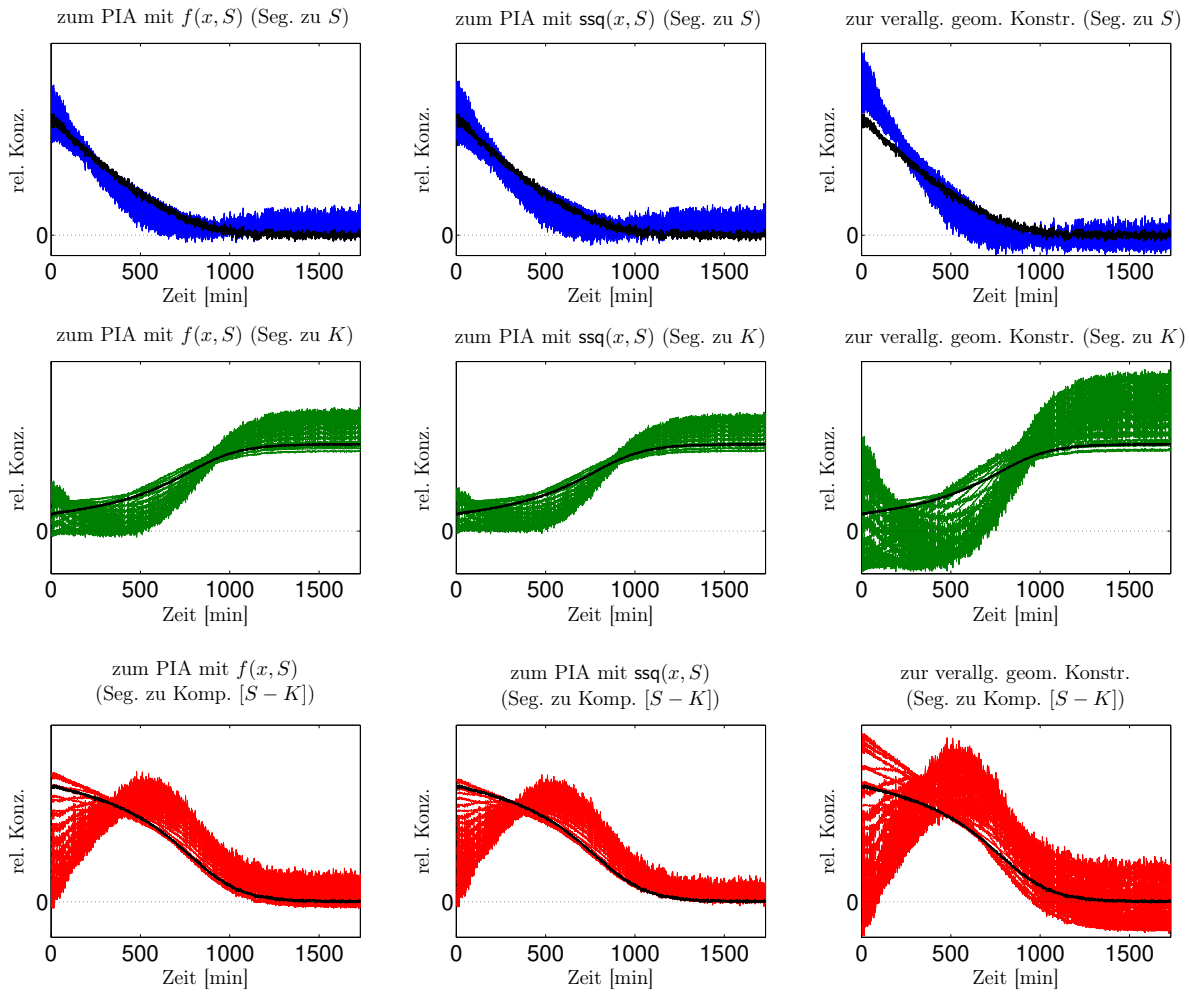


**Abbildung 5.10:** Die transformierten Profile (ausgewählte Elemente der Menge $\mathcal{C}$ aus (2.4)) für die verschiedenen Approximationen von $\mathcal{M}_C$ aus Abbildung 5.9. Links: für den Polygon inflation Algorithmus (PIA) unter Nutzung von $f$ aus (4.2). Mitte: für den Polygon inflation Algorithmus unter Nutzung von ssq aus (4.6). Rechts: mit der verallgemeinerten geometrischen Konstruktion. Auffällig sind die betragsgroßen negativen Werte für $C$ bei der verallgemeinerten geometrischen Konstruktion.

Der Zusammenhang, dass die Indizes $j$ zu hohen Werten $e_j^{(u)}$ zu den $u(:,j)$ gehören, die sensitiv auf Störungen reagieren, ist in Abbildung 5.13 verdeutlicht. In der linken Grafik sind die Werte $e_j^{(u)}$ mit der RGB-Farbkodierung $(e_j^{(u)}/\|e^{(u)}\|_\infty,\ 0,\ 0)$ dargestellt. In der rechten Grafik sind die zugehörigen $u(:,j)$, $j=1,\ldots,n$, mit derselben Farbkodierung sowie die Menge $\mathcal{M}_C$ gezeigt.

## 5.4 Anwendung von Regularisierungen

In [4,12,13,57,132,158] und dem weiterführenden Abschnitt 4.9 sind Möglichkeiten erläutert, die Mengen zulässiger Lösungen mittels zusätzlicher Regularisierungsfunktionen einzuschränken. So lassen sich für eine konkrete Anwendung Mengen zulässiger und relevanter Lösungen $\mathcal{M}_A^{(\mathrm{rel})}$ und $\mathcal{M}_C^{(\mathrm{rel})}$ von $\mathcal{M}_A$ und $\mathcal{M}_C$ abspalten. Die Zielfunktion $f(x,S)$ aus (4.2) wird dazu mittels Regularisierungsfunktionen erweitert. In diesem Abschnitt werden beispielhaft Monotonierestriktionen für $C$ genutzt, um für Datensatz 3 Mengen $\mathcal{M}_A^{(\mathrm{rel})}$ und $\mathcal{M}_C^{(\mathrm{rel})}$ zu bestimmen.

Es werden Monotonierestriktionen für alle drei Konzentrationsprofile angewendet, also ist $I_{\mathrm{mono}} = \{1,2,3\}$. Dies ist mit Rücksicht auf den Zeitpunkt der ersten Messung $t_1 = 4.729\,\mathrm{min}$ möglich.

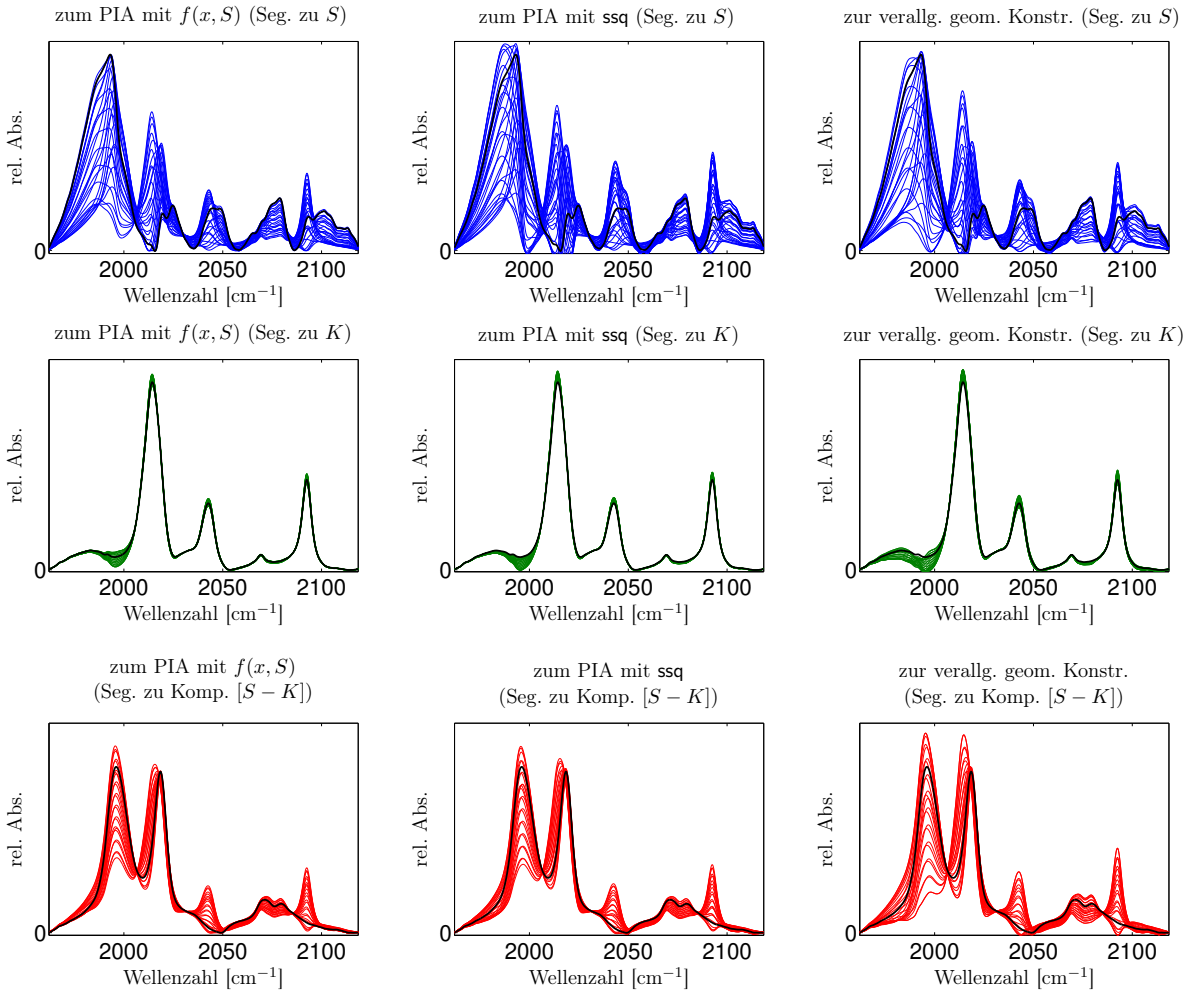


**Abbildung 5.11:** Die transformierten Profile (ausgewählte Elemente der Menge $\mathcal{A}$ aus (2.3)) für die verschiedenen Approximationen von $\mathcal{M}_A$ aus Abbildung 5.9. Links: für den Polygon inflation Algorithmus (PIA) unter Nutzung von $f(x,S)$ aus (4.2). Mitte: für den Polygon inflation Algorithmus unter Nutzung der Funktion $\mathsf{ssq}(x,S)$ aus (4.6). Rechts: mit der verallgemeinerten geometrischen Konstruktion.

Die Bildung der Komponente $[S-K]$ ist zu diesem Zeitpunkt abgeschlossen. Die Details zur Umsetzung von Monotonierestriktionen sind im Pseudocodeelement 2 sowie in (4.38) angegeben. Bei der Wahl der Steuerparameter ist zu beachten, dass das Konzentrationsprofil der ersten Komponente (hier $S$) deutlich stärker durch Störungen beeinflusst ist als es die anderen beiden ($K$ und $[S-K]$) sind. Dies liegt daran, dass der Anteil des dritten Singulärvektors für dieses Konzentrationsprofil deutlich höher ist als bei den anderen beiden und der dritte linksseitige Singulärvektor stark durch Störungen beeinflusst wird, siehe Abbildung 2.8. Um trotzdem eine effiziente Umsetzung der Monotonierestriktionen zu gewährleisten, wird für Komponente $S$ ein größerer Steuerparameter $\rho$ gewählt als für $K$ und $[S-K]$, siehe Bemerkung 4.38.

Für die Berechnungen von $\mathcal{M}_A^{(\mathrm{rel})}$ und $\mathcal{M}_C^{(\mathrm{rel})}$ werden die Steuerparameter $\varepsilon_a = 2.5 \cdot 10^{-3}$, $\varepsilon_c = 5 \cdot 10^{-2}$, $\varepsilon_f = 10^{-10}$ sowie $\gamma_{\mathrm{mono}} = 0.1$ genutzt. Der Steuerparameter $\rho$ wird für Komponente $S$ als $\rho = 0.15$ gewählt und für die anderen beiden als $\rho = 0.06$. Der Polygon inflation Algorithmus wird mit $\varepsilon_b = \delta = 10^{-3}$ ausgeführt. Für die Bewertungen der Monotonie werden die normierten Profile $\widehat{C}$ aus (4.35) genutzt.

Die berechneten Mengen $\mathcal{M}_A^{(\mathrm{rel})}$ und $\mathcal{M}_C^{(\mathrm{rel})}$ sind in Abbildung 5.14 dargestellt. Auffällig ist, dass der Reduktionseffekt (bezogen auf die relative Flächenänderung) für den Faktor $C$ deutlich höher ist als für den Faktor $A$. Dies ist auch für die zulässigen Lösungen für $C$ und für $A$ in

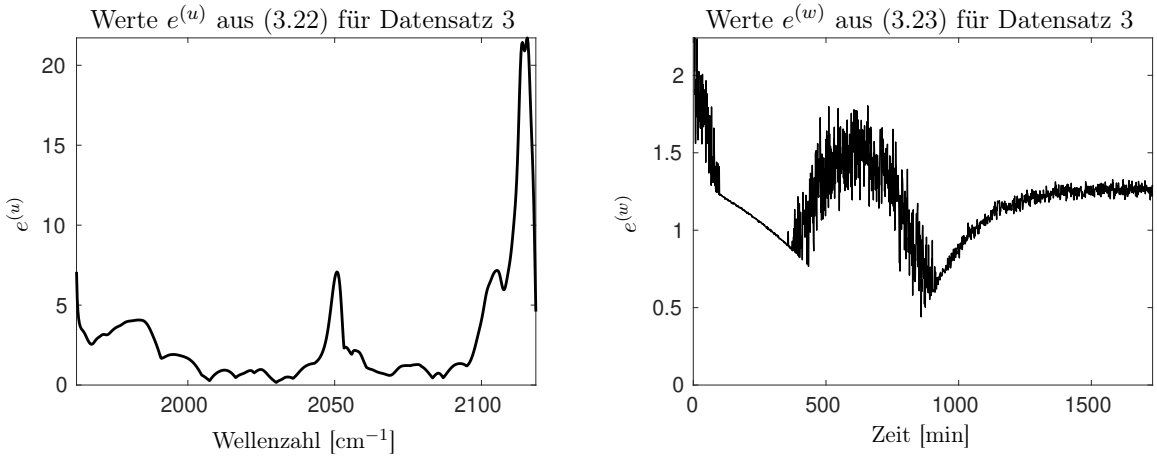


**Abbildung 5.12:** Die Werte $e^{(u)}$ und $e^{(w)}$ aus (3.22) und (3.23) dienen als Sensitivitätsindikatoren für die Approximationen von $\mathcal{M}_A$ und $\mathcal{M}_C$ mittels geometrischer Konstruktionen. In den Grafiken sind sie für Datensatz 3 gegen die Wellenzahlen beziehungsweise gegen die Messzeitpunkte abgebildet. Die hohen Werte in $e^{(u)}$ zeigen die Sensitivität einiger $u(:,j)$, $j = 1,\ldots,n$. Dies erschwert die Berechnung von $\mathcal{M}_C$ mittels verallgemeinerter geometrischer Konstruktionen (linke Grafik in Abbildung 5.9). Die Werte in $e^{(w)}$ sind deutlich niedriger und die geometrische Konstruktion von $\mathcal{M}_A$, unter Nutzung von $w(:,i)$, $i = 1,\ldots,k$, ist weniger anfällig für Störungen (rechte Grafik in Abbildung 5.9). Die Werte für $e^{(u)}$ und $e^{(w)}$ sind nicht von den Singulärwerten beeinflusst.

Abbildung 5.15 ersichtlich. In den oberen Grafiken (Faktor $C$) ist der Einfluss der Monotonierestriktionen klar zu erkennen. Es werden, mit Rücksicht auf Störungen, nicht monotone Profile effizient ausgeschlossen. Die Rückkopplungen für $A$ erfolgen nur aufgrund des Dualitätsprinzips, siehe Abschnitt 3.6 oder [69, 122, 133, 145, 151].

## 5.5 Begrenzende Lösungen für die Datensätze 1 und 2

In Abschnitt 4.10 ist die Methode der Berechnung von Einhüllenden für die Faktoren $C$ und $A$ beschrieben. Der Ansatz entspricht zwar nicht dem der Mengen zulässiger Lösungen und es werden für $s \geq 3$ die Mengen zulässiger Profile für $C$ und $A$ in der Regel auch nicht vollständig abgebildet, die Methode ist aber dennoch ein probates Mittel, die Problematik der Mehrdeutigkeit der Faktorisierung zu untersuchen. In diesem Abschnitt wird die Methode auf die Datensätze 1 und 2 angewendet und die Lösungen werden zu $\mathcal{M}_A$ und $\mathcal{M}_C$ in Bezug gesetzt.[3]

In Bemerkung 4.41 (Punkt 2) ist auf das Problem der Skalierung bei der Darstellung der Lösungen hingewiesen. Zum Vergleich der Lösungen werden in diesem Abschnitt die Skalierungen

$$a = (1, x_1, \ldots, x_{s-1})V^{\mathbf{T}} \tag{5.1}$$

und

$$c = U\Sigma(1, y_1, \ldots, y_s)^{\mathbf{T}} \tag{5.2}$$

mit geeigneten $x \in \mathbb{R}^{s-1}$ und $y \in \mathbb{R}^{s-1}$ genutzt, wie sie auch bei den Berechnungen der Mengen $\mathcal{M}_A$ und $\mathcal{M}_C$ eingesetzt werden. Dass andere Skalierungs-/Normierungsarten auch nicht besser geeignet sind, wird später in Abbildung 5.20 deutlich.

[3] Die Berechnungen erfolgten mit der in [83] vorgestellten MCR-BANDS-Implementierung, heruntergeladen von http://www.mcrals.info/.

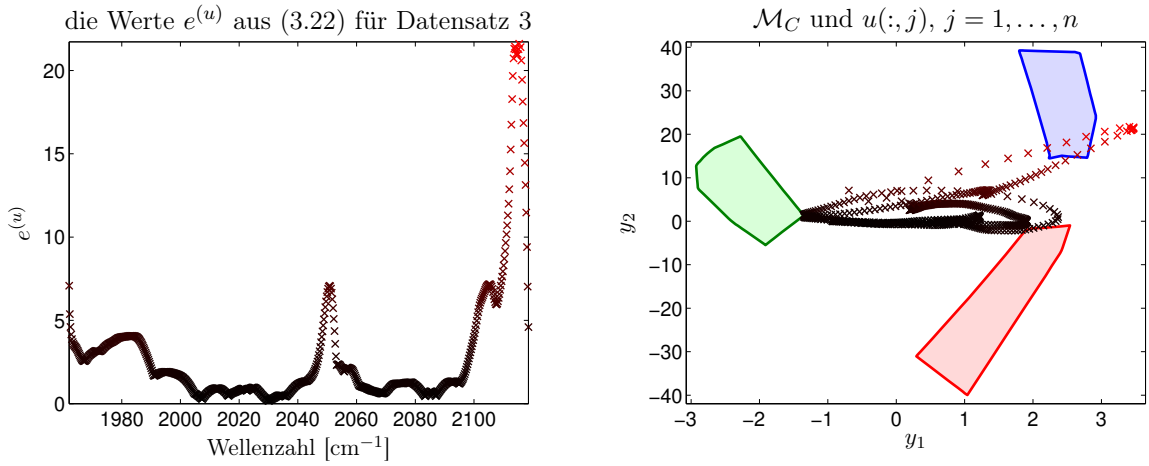


**Abbildung 5.13:** Der Zusammenhang zwischen $u(:,j)$, $j = 1,\ldots,n$, in $\mathcal{M}_C$ und den Werten in $e^{(u)}$ ist durch die farbliche Abhebung gekennzeichnet. Es sind die $u(:,j)$ zu den Indizes $j$, deren Werte $e_j^{(u)}$ hoch sind, welche sich als sensitiv für Störungen und damit als problematisch bei der Berechnung von $\mathcal{M}_C$ erweisen. Sowohl die Werte $e_j^{(u)}$ als auch die $u(:,j)$ sind mit RGB-Farbkodierung $(e_j^{(u)}/\|e^{(u)}\|_\infty,\ 0,\ 0)$ dargestellt.

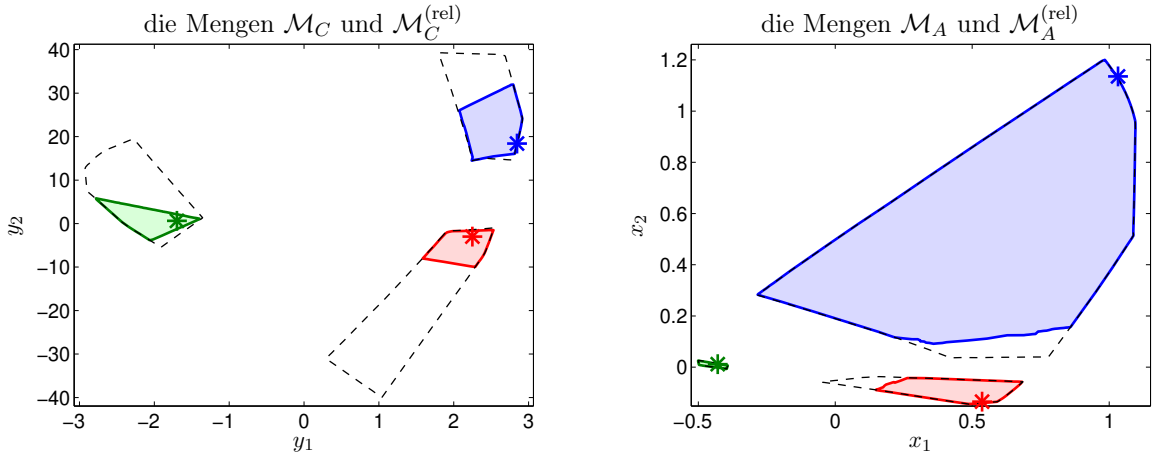


**Abbildung 5.14:** Reduktionen der Mengen zulässiger Lösungen durch Monotonierestriktionen für alle drei Konzentrationsprofile. Links: die Menge zulässiger und relevanter Lösungen $\mathcal{M}_C^{(\mathrm{rel})}$ (farbige Segmente) und die Menge zulässiger Lösungen $\mathcal{M}_C$ (schwarz gestrichelter Rand). Rechts: analog für $\mathcal{M}_A^{(\mathrm{rel})}$ und $\mathcal{M}_A$. Der Reduktionseffekt ist, bezogen auf die relative Flächenänderung, für $\mathcal{M}_C$ größer als für $\mathcal{M}_A$. Zur Verifizierung sind die niedrigdimensionalen Darstellungen der kinetischen Lösungen eingezeichnet ✱.

### 5.5.1 Anwendung für Datensatz 1

Dem Datensatz 1 liegt ein Zweikomponentensystem zugrunde. Die Software MCR-BANDS wird zur Berechnung von Einhüllenden für die Faktoren $C$ und $A$ angewendet. Für die Berechnung werden $C \geq 0$ und $A \geq 0$ als Restriktionen gewählt (die Forderung $D = CA$ ist nicht direkt anwählbar, siehe Bemerkung 4.39). Es wird keine Form der Skalierung genutzt. Als Resultate der einzelnen Optimierungen ergeben sich jeweils vier Profile für $A$ und $C$ als Einhüllende für die Mengen möglicher Profile.

Die niedrigdimensionale Darstellung der Einhüllenden, siehe (4.40), führt auf vier Elemente aus $\mathcal{M}_A$ und vier Elemente aus $\mathcal{M}_C$. Diese sind in Abbildung 5.16 zusammen mit $\mathcal{M}_A$ und $\mathcal{M}_C$ dargestellt und ergeben jeweils die Intervallgrenzen. Sie geben somit die Mengen zulässiger Lösungen vollständig wieder, vergleiche Bemerkung 4.41 (Punkt 1) und Abschnitt 4.10.3.

Die transformierten Lösungen für $C$ und $A$ sowie die Einhüllenden sind in Abbildung 5.17 präsentiert. Die Methode der Bestimmung von Einhüllenden liefert für diesen Datensatz offensichtlich

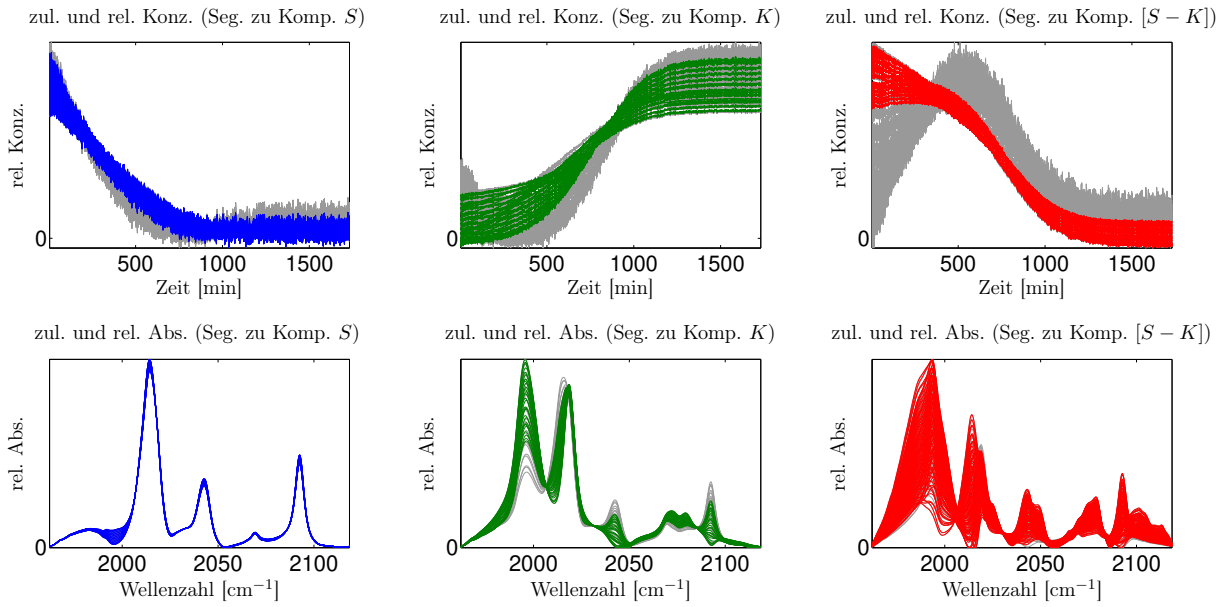


**Abbildung 5.15:** Transformierte Profile zu den Mengen zulässiger Lösungen $\mathcal{M}_A$ und $\mathcal{M}_C$ (grau eingezeichnet) und zu den Mengen zulässiger und relevanter Lösungen $\mathcal{M}_C^{(\mathrm{rel})}$ und $\mathcal{M}_A^{(\mathrm{rel})}$ (farbig) unter Anwendung von Monotonierestriktionen. Die Monotonierestriktionen sind für alle drei Konzentrationsprofile eingesetzt. Die Mengen $\mathcal{M}_C^{(\mathrm{rel})}$ und $\mathcal{M}_A^{(\mathrm{rel})}$ sind in Abbildung 5.14 dargestellt. Oben: die zulässigen Konzentrationsprofile. Unten: die zulässigen Reinkomponentenspektren. Die Effektivität der Monotonierestriktionen ist deutlich erkennbar: Profile, die nicht approximativ monoton sind, werden zufriedenstellend ausgeschlossen.

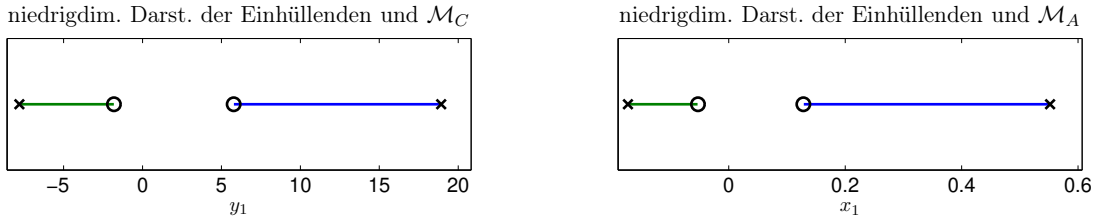


**Abbildung 5.16:** Die Mengen zulässiger Lösungen $\mathcal{M}_C$ (links, farbig) und $\mathcal{M}_A$ (rechts, farbig), vergleiche auch Abbildung 4.5, und die niedrigdimensionalen Darstellungen der Einhüllenden (schwarz markiert) für Datensatz 1. Die niedrigdimensionalen Darstellungen sind offensichtlich die Intervallgrenzen für $\mathcal{M}_C$ und $\mathcal{M}_A$. Bezüglich der aufgestellten Optimierungsprobleme repräsentieren die ∘ die lokalen Maximalstellen und die × die lokalen Minimalstellen.

das korrekte Ergebnis. Dabei ist zu erwähnen, dass es sich nicht um obere und untere Einhüllende im Sinne von Lösungseinschließungen durch Intervalle handelt, was in Abbildung 5.17 auch gut zu erkennen ist.

### 5.5.2 Anwendung für Datensatz 2

Als nächstes werden begrenzende Lösungen für Datensatz 2 berechnet. Dem Datensatz liegt ein Dreikomponentensystem zugrunde und für $C$ und $A$ ergeben sich jeweils sechs begrenzende Lösungen. Anders als für $s = 2$ lässt sich für $s \geq 3$ im Normalfall anhand der berechneten Lösungen nicht auf $\mathcal{M}_A$ beziehungsweise $\mathcal{M}_C$ schließen. In Abbildung 5.18 sind die Mengen $\mathcal{M}_A$ und $\mathcal{M}_C$ sowie die $x^{(i)}$ und $y^{(i)}$, $i = 1, \ldots, 6$, aus (4.40) für Datensatz 2 eingezeichnet. Die $x^{(i)}$ liegen (selbstverständlich) in $\mathcal{M}_A$, sie können $\mathcal{M}_A$ aber nicht vollständig wiedergeben. Analoges gilt für die $y^{(i)}$ und $\mathcal{M}_C$.

In Abbildung 5.19 sind die möglichen Profile für $C$ und $A$ zusammen mit den berechneten Einhüllenden gezeigt. Die Skalierung ist für alle Profile wie in (5.1) beziehungsweise (5.2) gewählt.

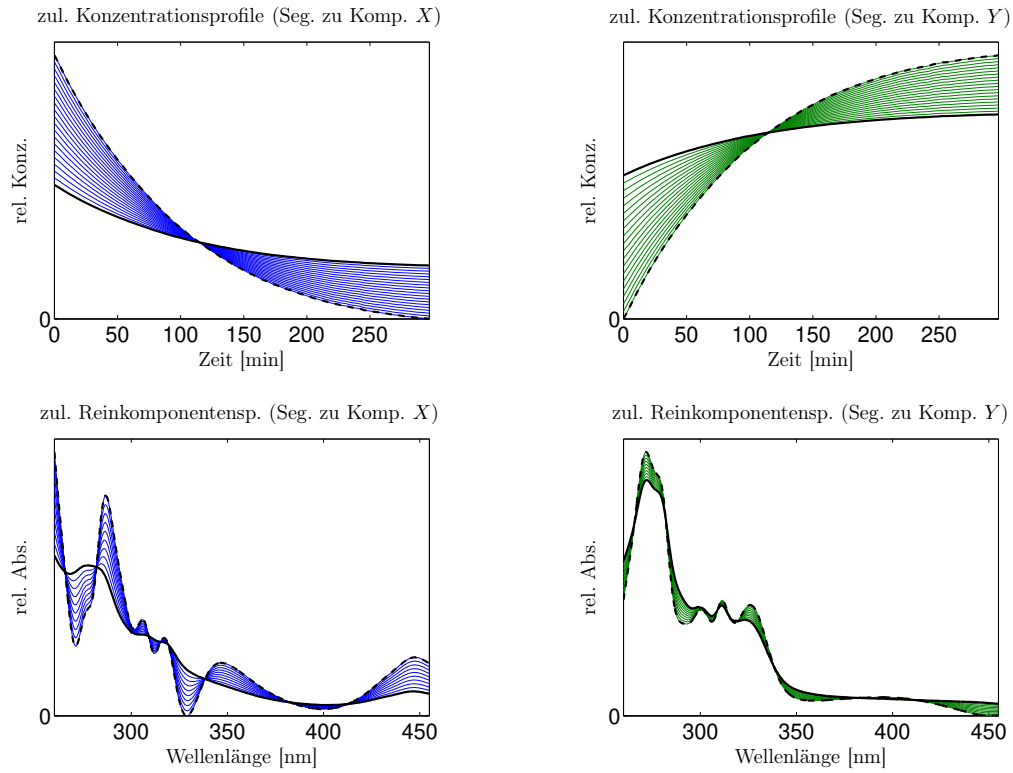


**Abbildung 5.17:** Zulässige Profile (blau und grün) für $C$ und $A$ und die berechneten Einhüllenden (schwarz) für Datensatz 1. Die Skalierung ist wie in (5.1) und (5.2) gewählt. Die gestrichelten schwarzen Linien führen auf die in Abbildung 5.16 mittels × gekennzeichneten Lösungen und die durchgezogenen schwarzen Linien auf die mit ∘ gekennzeichneten. Bezüglich der aufgestellten Optimierungsprobleme repräsentieren die durchgezogenen Linien die lokalen Maximalstellen und die gestichelten Linien die lokalen Minimalstellen.

Das Prinzip der Lösungsbegrenzung funktioniert nur für zwei der sechs Teilfaktoren: bei den Lösungen zum Segment von $\mathcal{M}_C$, dass zum Konzentrationsprofil von $S$ gehört, und bei den Lösungen zum Segment von $\mathcal{M}_A$, dass zum Reinkomponentenspektrum von $[S-K]$ gehört. Bei den anderen Profilen sind teils große Abweichung von einer Einschließung auffällig. Dies ist auch durch eine andere Art der Skalierung/Normierung nicht zu beheben.

### 5.5.3 Schwierigkeit der Darstellung

Eine sinnvolle Darstellung der Ergebnisse zur Berechnung von begrenzenden Lösungen ist schwierig, siehe Bemerkung 4.41 (Punkt 2). Eine verbreitete Variante ist es, ein Spektrum $a = A(i,:)$ in der Form

$$\widehat{a} = \frac{a}{\max\limits_{i=1,\ldots,n} |a_i|}$$

grafisch darzustellen. Aber auch dies ist nicht immer sinnvoll. Beispielhaft sind die Ergebnisse zu Datensatz 2 für den Faktor $A$ in Abbildung 5.20 mit dieser Form der Skalierung dargestellt. Die Darstellung ist nicht besser geeignet als die in Abbildung 5.19 mit der Skalierung aus (5.1).

### 5.5.4 Verhalten der Zielfunktion

Für eine Untersuchung des Verhaltens der Zielfunktion $g_i(T)$ aus (4.39) bei der Anwendung für Datensatz 2 wird $\mathcal{M}_A$ diskretisiert. Anschließend werden zu jedem Stützpunkt $x$ zwei Optimie-

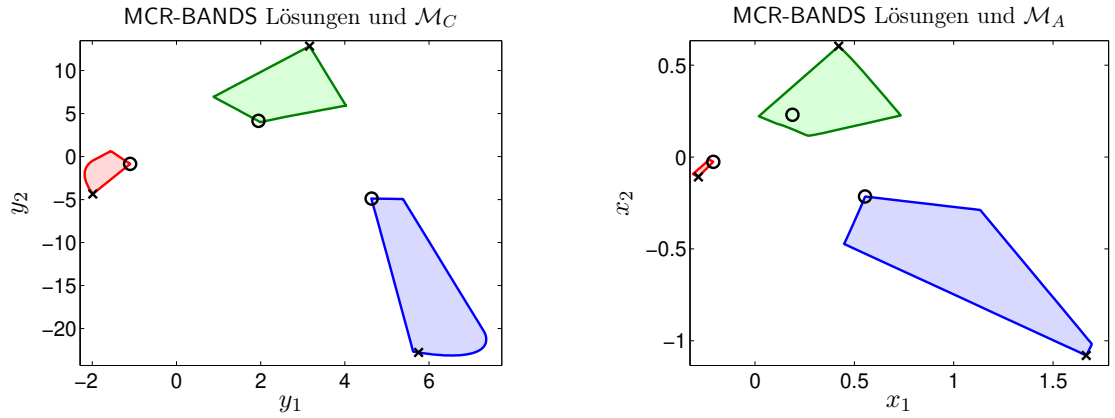


**Abbildung 5.18:** Die Mengen $\mathcal{M}_C$ (links, farbig) und $\mathcal{M}_A$ (rechts, farbig) sowie die niedrigdimensionalen Darstellungen $x^{(i)}$ und $y^{(i)}$ aus (4.40) (schwarz markiert) für Datensatz 2. Die $x^{(i)}$ liegen zwar in $\mathcal{M}_A$, können diese aber nicht komplett wiedergeben. Gleiches gilt für die $y^{(i)}$ und $\mathcal{M}_C$. Die aus $\mathcal{M}_A$ und $\mathcal{M}_C$ abgeleiteten Mengen zulässiger Faktoren $C$ und $A$ sind in Abbildung 5.19 dargestellt. Bezüglich der aufgestellten Optimierungsprobleme repräsentieren die ∘ lokale Maximalstellen und die × lokale Minimalstellen.

rungsprobleme aufgestellt. Die Matrix $T$ wird jeweils in einer Zeile mittels $(1, x^{\mathbf{T}})$ fixiert und $g_i(T)$ wird minimiert beziehungsweise maximiert.

Das $i$-te Segment von $\mathcal{M}_A$ wird mittels der Gitterpunkte $x^{(i,j)}$, $j = 1, \ldots, N_i$, diskretisiert. Es werden für alle Stützstellen des Gitters die Werte

$$w_1^{(i,j)} = \max_T g_i(T), \quad \text{zu} \quad T(i,:) = (1, x_1^{(i,j)}, x_2^{(i,j)}),$$
$$w_2^{(i,j)} = \min_T g_i(T), \quad \text{zu} \quad T(i,:) = (1, x_1^{(i,j)}, x_2^{(i,j)})$$

unter den Nichtnegativitäts- und Rekonstruktionsrestriktionen berechnet. Analog zur MCR-BANDS-Implementierung wird die MATLAB-Routine fmincon verwendet. Zur Verbesserung der Approximationen werden die Startwerte für die einzelnen Optimierungen geeignet gewählt. Unter anderem werden die Minimal- beziehungsweise Maximalstellen von benachbarten zulässigen Lösungen genutzt. Anschließend werden für die einzelnen Segmente $i$ die Flächen zu

$$P_1^{(i,j)} = \left(x_1^{(i,j)}, x_2^{(i,j)}, w_1^{(i,j)}\right), \quad j = 1, \ldots, N_i, \tag{5.3}$$

zur Analyse des Verhaltens der Zielfunktionen $g_i(T)$ aus (4.39) bezüglich deren Maximierung und die Flächen zu

$$P_2^{(i,j)} = \left(x_1^{(i,j)}, x_2^{(i,j)}, w_2^{(i,j)}\right), \quad j = 1, \ldots, N_i, \tag{5.4}$$

zur Analyse des Verhaltens der Zielfunktionen bezüglich deren Minimierung dargestellt. Die Ergebnisse sind in Abbildung 5.21 dargestellt. Zusätzlich sind auch die niedrigdimensionalen Darstellungen der Ergebnisse von MCR-BANDS, also die jeweiligen lokalen Extrempunkte, eingezeichnet. Diese sind gut als lokale Maxima beziehungsweise lokale Minima auf den Flächen erkennbar. Die lokalen Extremstellen sind dieselben wie in Abbildung 5.18 (siehe auch die mittels ∘ gekennzeichneten zulässigen Lösungen in Abbildung 5.22).

**Bemerkung 5.2.**

1. *In den Grafiken von Abbildung 5.21 sind die Flächen zu den Punkten aus (5.3) und (5.4) dargestellt. Für die Berechnungen der jeweiligen z-Koordinaten werden die Zielfunktionen $g_i(T)$ unter verschiedenen Nebenbedingungen maximiert beziehungsweise minimiert. Zur Erzeugung der Grafiken wurde mit der* MATLAB*-Routine* fmincon *dieselbe Methode wie in*

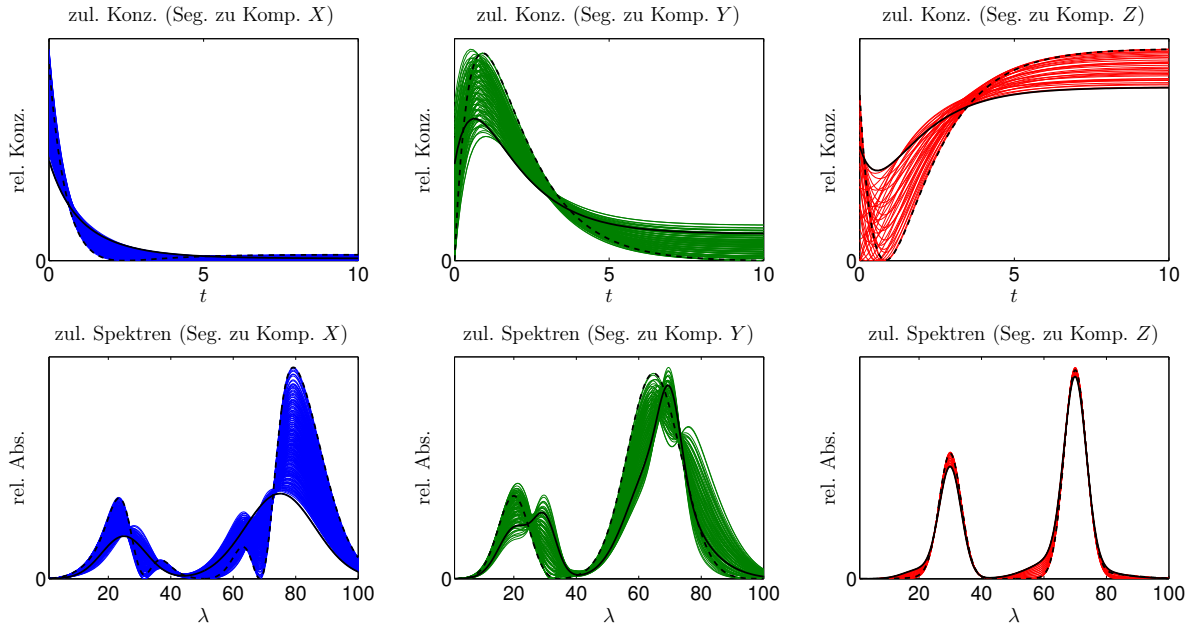


**Abbildung 5.19:** Zulässige Profile für $C$ und $A$ (farbig) und die begrenzenden Lösungen (schwarz) für Datensatz 2 mit der Skalierung wie in (5.1) und (5.2). Nur für die Grafiken oben links und unten rechts sind die speziell bestimmten Lösungen begrenzend. Die gestrichelten schwarzen Linien gehören zu den in Abbildung 5.16 mittels × gekennzeichneten zulässigen Lösungen und die durchgezogenen schwarzen Linien zu den mit ∘ gekennzeichneten. Bezüglich der aufgestellten Optimierungsprobleme repräsentieren die durchgezogenen Linien die lokalen Maximalstellen und die gestichelten Linien die lokalen Minimalstellen.

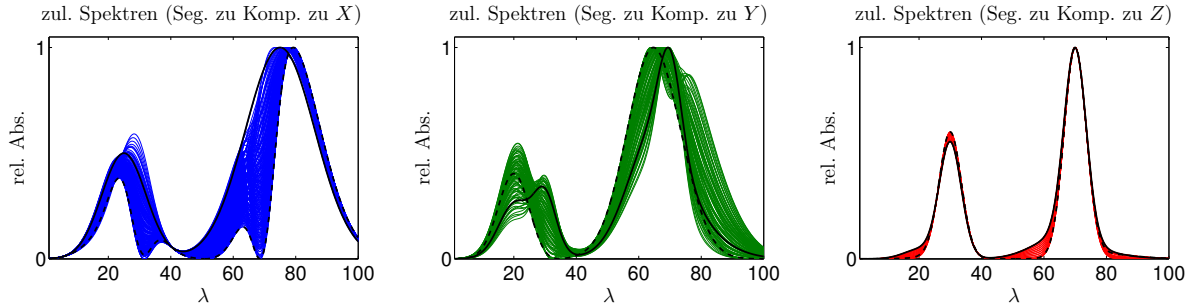


**Abbildung 5.20:** Darstellung mit normierten Profilen $a/\max|a|$. Die Darstellungsart ist auch nicht besser geeignet als die in Abbildung 5.19 mit der direkten Skalierung $a = (1, x^{\mathbf{T}})V^{\mathbf{T}}$.

*der* MCR-BANDS*-Implementierung genutzt. Anhand der vielen leichten Ausreißer in den Oberflächen ist zu erahnen, dass einige der Optimierungen nicht auf (zumindest lokale) Extrema führten. Starke Ausreißer, die durch fehlgeschlagene Optimierungen entstanden, wurden eliminiert und sind nicht mit dargestellt. Die Leistungsstärke von* fmincon *scheint zur Lösung der jeweils aufgestellten Optimierungsprobleme unter Umständen nicht ausreichend zu sein, was umgekehrt eine leichte Abhängig der Resultate bei* MCR-BANDS *von der initialen nichtnegativen Matrixfaktorisierung bedeutet.*

2. *Eine Vermutung zu der Verbindung zwischen* MCR-BANDS *und den Mengen zulässiger Lösungen ist, dass die niedrigdimensionalen Darstellungen der begrenzenden Lösungen stets auf dem Rand von $\mathcal{M}_A$ beziehungsweise $\mathcal{M}_C$ liegen. Unter der Annahme, dass das gewählte Gitter zu den Grafiken in Abbildung 5.21 ausreichend fein ist und die Optimierungen zu den Randpunkten erfolgreich verliefen, kann diese Vermutung mittels der Abbildungen 5.18 und 5.21 für $s = 3$, widerlegt werden. Beim grünen Segment von $\mathcal{M}_A$ (siehe Abbildung 5.18) liegt eine extremale Lösung nicht auf dem Rand. Dies kann durch die Auswertung mittels der $P_1^{(i,j)}$ für dieses Segment (siehe Abbildung 5.21, mittlere Grafik) als korrekt bestätigt*

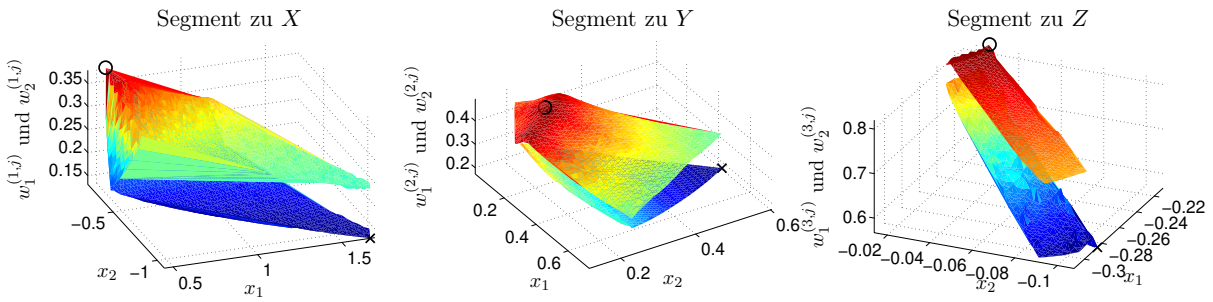


**Abbildung 5.21:** Darstellungen der Flächen zu den Punkten aus (5.3) und der Flächen zu den Punkten aus (5.4) für die Segmente von $\mathcal{M}_A$ zu Datensatz 2. Die oberen Flächen gehören zu den Punkten aus (5.3), die unteren zu den Flächen aus (5.4). Die mittels MCR-BANDS bestimmten lokalen Extrempunkte der einzelnen Flächen sind mittels ∘ und × gekennzeichnet.

*werden. Es handelt sich nicht um ein falsch bestimmtes lokales Extremum.*

### 5.5.5 Abhängigkeit von der initialen Zerlegung

Die Optimierung der Zielfunktion (4.39) wird in MCR-BANDS zu einer vorgegebenen initialen nichtnegativen Matrixfaktorisierungen vorgenommen. In diesem Abschnitt wird die Methode auf Datensatz 2 mit zwei verschiedenen initialen Zerlegungen angewendet und die Lösungen werden verglichen.

Die mittels fmincon bestimmten maximalen Werte $w_1^{(\max)}, \ldots, w_3^{(\max)}$ und minimalen Werte $w_1^{(\min)}, \ldots, w_3^{(\min)}$ der Zielfunktionen $g_i(T)$, $i = 1, 2, 3$, sind für die eine initiale Zerlegung

$$w^{(\max)} = (3.7859 \cdot 10^{-1},\ 4.9398 \cdot 10^{-1},\ 8.2013 \cdot 10^{-1}),$$
$$w^{(\min)} = (1.3108 \cdot 10^{-1},\ 1.7327 \cdot 10^{-1},\ 5.6726 \cdot 10^{-1})$$

sowie zu einer anderen initialen Zerlegung

$$w^{(\max)} = (3.7861 \cdot 10^{-1},\ 4.8673 \cdot 10^{-1},\ 8.2018 \cdot 10^{-1}),$$
$$w^{(\min)} = (1.3106 \cdot 10^{-1},\ 1.7328 \cdot 10^{-1},\ 5.6726 \cdot 10^{-1}).$$

In Abbildung 5.22 sind die niedrigdimensionalen Darstellungen der einzelnen Lösungen zusammen mit den Mengen $\mathcal{M}_A$ und $\mathcal{M}_C$ dargestellt. Für jeweils ein Segment stimmen je zwei zugehörige Lösungen nicht überein.

## 5.6 Kritische Zusammenfassung

In diesem Kapitel wurden verschiedene Methoden zur Berechnung von $\mathcal{M}_A$ und verschiedene Ansätze zur Klassifizierung von $x \in \mathbb{R}^{s-1}$ anhand zweier Datensätze verglichen. Bei der Anwendung auf ungestörte Daten stellte sich die geometrische Konstruktion als ultimative Methode heraus. Einzig in Bezug auf die Steuerbarkeit durch Parameter offenbarten sich leichte Vorteile für die numerischen Methoden. In Bezug auf Genauigkeit und Rechenaufwand zeigten sich bei den numerischen Methoden starke Unterschiede. Für die Methode der Dreieckseinschließung ergab sich durch die fehlende Adaptivität ein Nachteil, dafür ist die Steuerbarkeit hier am besten. Die Adaptivität ist der große Vorteil der Polygon inflation Methoden, was sich im Verhältnis aus Genauigkeit und Rechenaufwand widerspiegelt. Die Flexibilität, auch im Hinblick auf $s \geq 4$, ist der große Vorteil der Strahlenmethode. Die nicht adaptiv gesteuerte Variante hat in Bezug auf das Verhältnis zwischen Genauigkeit und Rechenaufwand sowie in Bezug auf die Steuerbarkeit

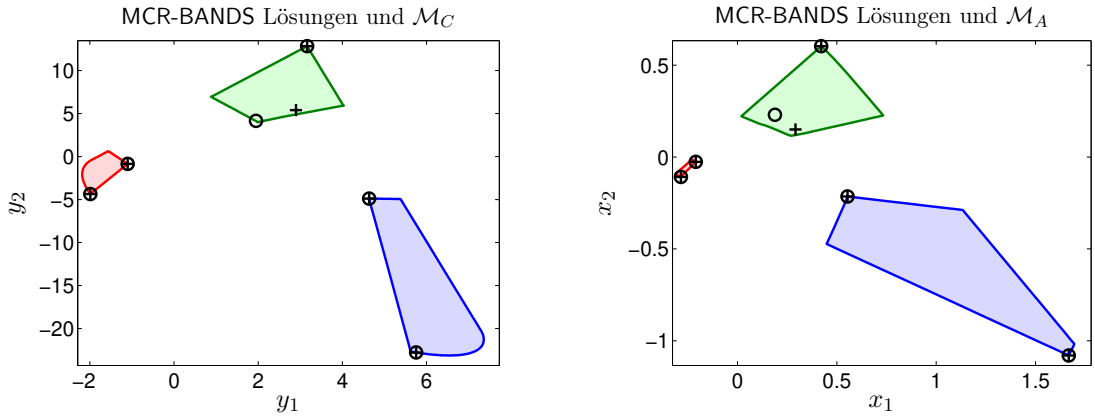


**Abbildung 5.22:** Analyse der Abhängigkeit der niedrigdimensionalen Darstellungen der Resultate von MCR-BANDS von den Startwerten (initiale Zerlegungen) am Beispiel des Datensatzes 2. Für die eine initiale Zerlegung sind die Ergebnisse mit ∘ gekennzeichnet, für die andere mit +. Offensichtlich stimmen jeweils nur fünf und nicht alle sechs Lösungen überein.

klare Nachteile. Das Potential der adaptiv gesteuerten Variante des Strahlenalgorithmus konnte aufgezeigt werden.

Weiter wurden drei Möglichkeiten zur Berücksichtigung von Störungen bei den Berechnungen von $\mathcal{M}_A$ und $\mathcal{M}_C$ am konkreten Beispiel des Datensatzes 3 analysiert. Der Vorteil der konstruktiven Bestimmung innerer Randpunkte erwies sich bei der verallgemeinerten geometrischen Konstruktion als Handicap.[4] Die beiden numerischen Ansätze, basierend auf den Funktionen $f(x, S)$ aus (4.2) und $\mathsf{ssq}(x, S)$ aus (4.6), erzeugen sehr ähnliche und überzeugende Ergebnisse. Die Resultate der verallgemeinerten geometrischen Konstruktion weichen davon stark ab. Zwar kann es bei der Bewertung für störungsbehaftete Daten kein *richtig* oder *falsch* geben. Es ist jedoch zumindest fragwürdig, dass die, mittels kinetischer Modellierung bestimmten, Lösungen teils nicht in den geometrisch konstruierten Mengen zulässiger Lösungen enthalten sind.

Die natürlichen Anforderungen an einen Algorithmus zur Approximation der Mengen zulässiger Lösungen sind, dass er schnell und stabil arbeitet und auch für gestörte Daten sinnvolle Resultate generiert. Diesbezüglich erweisen sich die Polygon inflation Methoden (direkter und inverser Typ) als gut geeignet. Einzig die adaptiv gesteuerte Strahlenmethode kann grundsätzlich mithalten. Der Algorithmus der Dreieckseinschließung arbeitet stabil, hat jedoch bei dem Nachteil des Mehraufwandes keinen entscheidenden Vorteil gegenüber den Polygon inflation Methoden. Die (verallgemeinerte) geometrische Konstruktion ist zwar sehr schnell, hat aber in Bezug auf die Anwendung für gestörte Daten klare Defizite.

[4] Der Ansatz der verallgemeinerten geometrischen Konstruktion [86–88] ist in dieser Schrift nicht näher erläutert.

# 6 Zusammenfassung und Perspektiven

Die vorliegende Schrift befasst sich mit der Mehrdeutigkeit der nichtnegativen Matrixfaktorisierung. Da es nicht sinnvoll ist, die Menge möglicher Faktorisierungen direkt anzugeben, wird der Zugang verfolgt, die Mengen möglicher Spalten des ersten und möglicher Zeilen des zweiten Faktors zu bestimmen, die zu nichtnegativen Matrixfaktorisierungen erweitert werden können. Bedingt durch die Zusatzforderung, dass die Faktoren vollen Ranges sein sollen, gelingt die Konstruktion möglicher Faktorisierungen mittels einer abgeschnittenen Singulärwertzerlegung. Unter der Festlegung einer speziellen Skalierung der Spalten des ersten beziehungsweise der Zeilen des zweiten Faktors führt dies auf die in dieser Schrift untersuchten Mengen zulässiger Lösungen. Dass die Mengen zulässiger Lösungen für ein $D$ überhaupt definiert werden können, ist, ebenso wie es viele ihrer Eigenschaften sind, formal nur an eine schwache Voraussetzung gebunden: Das Faktorisierungsproblem darf sich nicht in Subfaktorisierungsprobleme zerlegen lassen. Andernfalls ließe sich eine Aufspaltung in nicht weiter reduzierbare Teilsysteme vornehmen und die Mengen zulässiger Lösungen für diese betrachten. Im Hinblick auf die Entwicklung von Algorithmen zur geometrischen Konstruktion und zur numerischen Approximation der Mengen zulässiger Lösungen sind wichtige Eigenschaften detailliert untersucht worden. Bezogen auf $\mathcal{M}_A$ sind die Beschränktheit entscheidend sowie, dass der Nullpunkt nicht zu $\mathcal{M}_A$ aber zur Obermenge $\mathcal{F}_A$ gehört und dass der Schnitt mit einem, vom Ursprung ausgehenden, Strahl unterbrechungsfrei ist. Weiterhin ist die geometrische Interpretation einer zulässigen Lösung über Konvexkombinationen von zentraler Bedeutung.

Ein Schwerpunkt dieser Schrift liegt, neben der Analyse der Mengen zulässiger Lösungen, auf der Entwicklung und Analyse effizienter Methoden zu deren Berechnung beziehungsweise Approximation. Die Algorithmen gliedern sich in geometrisch konstruktive und numerisch approximative. Die geometrisch konstruktiven Ansätze sind sehr schnell, jedoch nur für ungestörte Daten problemlos anwendbar und bislang auf $s \in \{2, 3\}$ limitiert. Die numerisch approximativen Methoden arbeiten für ungestörte wie für gestörte Daten gleichermaßen stabil, sind aber rechenaufwendiger und liefern nur Näherungen. Für die numerischen Methoden liegen grundsätzlich keine Limitierungen bezüglich $s$ vor. Jedoch steigt für $s \geq 4$ der Rechenaufwand deutlich. Des Weiteren ist etwa die Strahlenmethode nicht auf eine spezielle Topologie der Mengen zulässiger Lösungen beschränkt.

In Anbindung an die Mengen zulässiger Lösungen ergeben sich unter anderem folgende Fragestellungen und Forschungsschwerpunkte für die Zukunft:

- Tiefgründige Analyse der Menge zulässiger Lösungen für Probleme mit Rangdefizit: In dem weiterführenden Abschnitt 3.7 wird der Fall betrachtet, dass die Ausgangsmatrix keine nichtnegative Faktorisierung mit Faktoren vollen Ranges besitzt und es wird eine verallgemeinerte Menge zulässiger Lösungen eingeführt und untersucht. Die bisherigen Untersuchungen sind zu erweitern und noch nicht berücksichtigte Fälle sind zu untersuchen. Weiter sind geometrisch konstruktive Algorithmen zur Berechnung der verallgemeinerten Mengen zulässiger Lösungen zu entwickeln. Diese werden grundlegend anders aufgebaut sein; für $\operatorname{rank}_+(D) > \operatorname{rank}(D) = 3$ wird die geometrische Konstruktion nicht mehr auf Dreiecken basieren, sondern stattdessen auf Polygonen mit $m = \operatorname{rank}_+(D)$ Ecken.
- Erweiterung für Tensorfaktorisierungen: Die nichtnegative Matrixfaktorisierung gehört zur Klasse der nichtnegativen Tensorfaktorisierungen. Allgemein wird bei der nichtnegativen

Tensorfaktorisierung zu einem $m$-dimensionalen Tensor $T \in \mathbb{R}_+^{n_1 \times \cdots \times n_m}$ eine Zerlegung in $X^{(j)} \in \mathbb{R}_+^{n_j \times s}$, $j = 1, \ldots, m$, zu minimalem $s$ gesucht, sodass

$$T_{i_1,\ldots,i_m} = \sum_{\ell=1}^{s} \prod_{j=1}^{m} X_{i_j \ell}^{(j)}, \qquad i_j = 1, \ldots, n_j,\ j = 1, \ldots, m.$$

Für $m = 3$ ist die Frage nach der Mehrdeutigkeit der Faktorisierung in [98, 99, 165] untersucht. Konkret werden Bedingungen aufgestellt, wie anhand einer Faktorisierung festgestellt werden kann, ob diese bis auf triviale Mehrdeutigkeiten eindeutig ist. Dabei stellt sich heraus, dass eine Zerlegung für $m = 3$ häufig eindeutig ist. Es wird zu untersuchen sein, inwiefern die Resultate und Ansätze der vorliegenden Schrift auf nichtnegative Tensorfaktorisierungen für $m \geq 3$ erweitert werden können und wie sich Teile von Tensorfaktorisierungen in Form der Mengen zulässiger Lösungen niedrigdimensional darstellen lassen.

- Analyse der Topologie der Mengen zulässiger Lösungen für $s \geq 4$: Die Topologien der Mengen zulässiger Lösungen für $s = 2$ (zwei disjunkte und zusammenhängende Teilintervalle, siehe Abschnitt 4.2) und für $s = 3$ (eine oder $3m$, $m \in \mathbb{N} \setminus \{0\}$, Zusammenhangskomponenten [86]) sind geklärt. Zu untersuchen ist, welche Anzahlen an Zusammenhangskomponenten für $s \geq 4$ möglich sind.

- Mehrskalenansätze zur Beschleunigung der Approximationsmethoden: Den größte Anteil des Rechenaufwands zur numerischen Approximation der Mengen zulässiger Lösungen machen die Klassifizierungen einzelner $x$ aus. Für jede Klassifizierung wird ein nichtlineares Ausgleichsproblem gelöst und so eine einzelne nichtnegative Faktorisierung berechnet. Für festes $s$ hängt der Aufwand der einzelnen Optimierungen von $f(x, S)$ aus (4.2) hauptsächlich von $k$ und $n$ ab. Aus [15, 143, 150] sind Mehrskalenansätze zur Beschleunigung der Berechnung einzelner Faktorisierungen bekannt. Ähnliche Ansätze sollen für die Berechnung der Mengen zulässiger Lösungen entwickelt und analysiert werden. Dazu werden die Mengen zulässiger Lösungen in einer Grobgitterhierarchie für unterschiedlich stark ausgedünnte Matrizen berechnet. Die Resultate für die Ränder von $\mathcal{M}_C$ und $\mathcal{M}_A$ einer Grobgitterstufe werden zur nächstfeineren prolongiert. Für diese Stufe werden so deutlich weniger Klassifizierungsaufrufe zur Berechnung von randnahen Punkten benötigt. Der Ansatz ist erfolgreich, wenn die eingesparte Rechenzeit durch die guten Anfangsnäherungen den zeitlichen Mehraufwand für die Berechnungen der vorherigen Mengen zulässiger Lösungen übersteigt. Die unterschiedlichen Skalierungen, bedingt durch die Normierungen der einzelnen Singulärvektoren der jeweils unterschiedlich dimensionalen Daten, sind zu beachten. Erste Testrechnungen wurden bereits durchgeführt und die Resultate sind vielversprechend.

- Entwicklung stabiler und adaptiv gesteuerter Methoden für $s \geq 4$: Die in Abschnitt 4.7 vorgestellte Strahlenmethode ist in ihrer einfachen Form ein nicht adaptiver Algorithmus zur Approximation der Mengen zulässiger Lösungen mittels eines Gitters im $\mathbb{R}^{s-2}$. In Abschnitt 4.7.9 ist eine adaptive Steuerung für $s = 3$ erläutert, deren Effizienz für einige Beispiele nachgewiesen wurde. Für $s = 4$ kann die adaptive Steuerung in Kombination mit dem Polyhedron inflation Algorithmus erfolgen. Stabile Ansätze zur adaptiven Steuerung etwa der Strahlenmethode für $s \geq 5$ sind zu entwickeln und zu untersuchen.

- Menge zulässiger kinetischer Parameter bei der Anpassung kinetischer Modelle: Für die Zerlegung spektroskopischer Daten kann die Berechnung einer Faktorisierung oft mit der Anpassung eines kinetischen Modells verknüpft werden. Dabei werden nichtnegative Faktorisierungen gesucht, die mit einem kinetischen Modell konsistent sind [30, 63, 112, 144]. Dies kann auf eine eindeutige Zerlegung und eindeutige kinetische Parameter führen. Unter Umständen gibt es jedoch eine Menge zulässiger kinetischer Parameter. Ein klassisches Beispiel ist die *slow-fast ambiguity* [5, 77], wobei sich für eine konsekutive Reaktion erster Ordnung mit $s - 1$ Teilreaktionen bis zu $(s - 1)!$ verschiedene nichtnegative Zerlegungen ergeben

können, für die $C$ dem kinetischen Modell genügt. In [160, 162] ist die Zerlegungsuneindeutigkeit bei simultaner kinetischer Modellierung für Reaktionssysteme erster Ordnung mit teilweise reversiblen Reaktionen untersucht. In [162] wird der Begriff *D-consistent reaction rate constants* eingeführt: ein Vektor kinetischer Parameter wird als *D-konsistent* bezeichnet, falls er für das kinetische Modell und die zugrunde liegende Diskretisierung auf eine Matrix von Konzentrationsprofilen führt, die durch die linksseitigen Singulärvektoren von $D$ dargestellt werden kann. Die Menge aller dieser Vektoren kinetischer Konstanten sei mit $\mathcal{K}$ bezeichnet. Unter Hinzunahme der Nichtnegativitätsrestriktionen für $A$ wird von $\mathcal{K}$ eine Teilmenge $\mathcal{K}^+$ abgespalten. Mitunter kommen zur Approximation von $\mathcal{K}^+$ Methoden zum Einsatz, die ähnliche Funktionsweisen haben, wie die in dieser Schrift vorgestellten Algorithmen zur Approximation der Mengen zulässiger Lösungen. Es sollen die Mengen zulässiger kinetischer Parameter unter dem Einfluss von Störungen untersucht und weitere stabile Methoden zu deren Approximation entwickelt werden.

# Literaturverzeichnis


[1] H. Abdollahi, M. Maeder and R. Tauler. Calculation and meaning of feasible band boundaries in multivariate curve resolution of a two-component system. *Anal. Chem.*, 81(6):2115–2122, 2009.

[2] H. Abdollahi and R. Tauler. Uniqueness and rotation ambiguities in multivariate curve resolution methods. *Chemom. Intell. Lab. Syst.*, 108(2):100–111, 2011.

[3] A. Aggarwal, H. Booth, J. O'Rourke, S. Suri and C. K. Yap. Finding minimal convex nested polygons. *Inform. Comput.*, 83(1):98–110, 1989.

[4] M. Akbari and H. Abdollahi. Investigation and visualization of resolution theorems in self modeling curve resolution (SMCR) methods. *J. Chemom.*, 27(10):278–286, 2013.

[5] N. W. Alcock, D. J. Benton and P. Moore. Kinetics of series first-order reactions. Analysis of spectrophotometric data by the method of least squares and an ambiguity. *Trans. Faraday Soc.*, 66:2210–2213, 1970.

[6] M. Alinaghi, R. Rajkó and H. Abdollahi. A systematic study on the effects of multi-set data analysis on the range of feasible solutions. *Chemom. Intell. Lab. Syst.*, 153:22–32, 2016.

[7] T. Azzouz and R. Tauler. Application of multivariate curve resolution alternating least squares (MCR-ALS) to the quantitative analysis of pharmaceutical and agricultural samples. *Talanta*, 74(5):1201–1210, 2008.

[8] R. B. Bapat and T. E. S. Raghavan. *Nonnegative matrices and applications*, volume 64 of *Encyclopedia Math. Appl.* Cambridge university press, New York, 1997.

[9] A. Barvinok. *A course in convexity*, volume 54 of *Grad. Stud. Math.* American Mathematical Society, Providence, 2002.

[10] A. Berman and R. J. Plemmons. *Nonnegative matrices in the mathematical sciences*, volume 9 of *Classics Appl. Math.* SIAM, Philadelphia, 1994.

[11] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca and R. J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Stat. Data Anal.*, 52(1):155–173, 2007.

[12] S. Beyramysoltan, H. Abdollahi and R. Rajkó. Newer developments on self-modeling curve resolution implementing equality and unimodality constraints. *Anal. Chim. Acta*, 827:1–14, 2014.

[13] S. Beyramysoltan, R. Rajkó and H. Abdollahi. Investigation of the equality constraint effect on the reduction of the rotational ambiguity in three-component system using a novel grid search method. *Anal. Chim. Acta*, 791:25–35, 2013.

[14] Å. Björck. *Numerical methods for least squares problems*, volume 51 of *Other Titles in Applied Mathematics*. SIAM, Philadelphia, 1996.

[15] M. Bojahr. Multilevelstrategien zur Spektrenfaktorisierung. Diploma thesis, Universität Rostock, 2009.

[16] O. S. Borgen, N. Davidsen, Z. Mingyang and Ø. Øyen. The multivariate N-component resolution problem with minimum assumptions. *Microchim. Acta*, 89:63–73, 1986.

[17] O. S. Borgen and B. R. Kowalski. An extension of the multivariate component-resolution method to three components. *Anal. Chim. Acta*, 174:1–26, 1985.

[18] R. G. Brereton. *Chemometrics: Data analysis for the laboratory and chemical plant.* John Wiley & Sons, Chichester, 2003.

[19] R. G. Brereton. *Applied chemometrics for scientists.* John Wiley & Sons, Chichester, 2007.

[20] R. Bro and N. D. Sidiropoulos. Least squares algorithms under unimodality and non-negativity constraints. *J. Chemom.*, 12(4):223–247, 1998.

[21] R. Bro and A. K. Smilde. Principal component analysis. *Anal. Methods*, 6(9):2812–2831, 2014.

[22] S. L. Campbell and G. D. Poole. Computing nonnegative rank factorizations. *Linear Algebra Appl.*, 35:175–182, 1981.

[23] J.-C. Chen. The nonnegative rank factorizations of nonnegative matrices. *Linear Algebra Appl.*, 62:207–217, 1984.

[24] J.-H. Chen and L.-P. Hwang. Reconstruction of mass spectra of components of unknown mixtures based on factor analysis. *Anal. Chim. Acta*, 133(3):271–281, 1981.

[25] W. Chew, E. Widjaja and M. Garland. Band-target entropy minimization (BTEM): An advanced method for recovering unknown pure component spectra. Application to the FT-IR spectra of unstable organometallic mixtures. *Organometallics*, 21(9):1982–1990, 2002.

[26] A. Cichocki, R. Zdunek, A. H. Phan and S.-I. Amari. *Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation.* John Wiley & Sons, Chichester, 2009.

[27] J. E. Cohen and U. G. Rothblum. Nonnegative ranks, decompositions, and factorizations of nonnegative matrices. *Linear Algebra Appl.*, 190:149–168, 1993.

[28] P. Comon. Independent component analysis, A new concept? *Signal Processing*, 36(3):287–314, 1994.

[29] S. Dahlke, W. Dahmen, M. Griebel, W. Hackbusch, K. Ritter, R. Schneider, C. Schwab and H. Yserentant, editors. *Extraction of quantifiable information from complex systems*, volume 102 of *Lect. Notes Comput. Sci. Eng.* Springer, 2014.

[30] A. de Juan, M. Maeder, M. Martínez and R. Tauler. Combining hard and soft-modelling to solve kinetic problems. *Chemom. Intell. Lab. Syst.*, 54(2):123–141, 2000.

[31] A. de Juan and R. Tauler. Multivariate curve resolution-alternating least squares for spectroscopic data. In C. Ruckebusch, editor, *Resolving Spectral Mixtures*, volume 30 of *Data Handling Sci. Technol.*, pages 5–51. Elsevier, Cambridge, 2016.

[32] A. de Juan, Y. Vander Heyden, R. Tauler and D. L . Massart. Assessment of new constraints applied to the alternating least squares method. *Anal. Chim. Acta*, 346(3):307–318, 1997.

[33] W. Den and E. R. Malinowski. Investigation of copper (II)-Ethylenediaminetetraacetate complexation by window factor analysis of ultraviolet spectra. *J. Chemom.*, 7(2):89–98, 1993.

[34] J. Dennis, D. Gay and R. Welsch. Algorithm 573: An adaptive nonlinear least-squares algorithm. *ACM Trans. Math. Softw.*, 7(3):369–383, 1981.

[35] J. Dennis, D. Gay and R. Welsch. An adaptive nonlinear least-squares algorithm. *ACM*

*Trans. Math. Softw.*, 7(3):348–368, 1981.

[36] B. Dong, M. M. Lin and M.T. Chu. Nonnegative rank factorization-a heuristic approach via rank reduction. *Numer. Algorithms*, 65(2):251–274, 2014.

[37] D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processes*, NIPS, pages 1141–1148. 2003.

[38] C. Eckard and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[39] L. Eldén. *Matrix methods in data mining and pattern recognition*, volume 4 of *Fundam. Algorithms*. SIAM, Philadelphia, 2007.

[40] H. W. Engl, M. Hanke and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Math. Appl.* Kluwer Academic Publishers, Dordrecht, 2000.

[41] C. Fischer, S. Schulz, H.-J. Drexler, C. Selle, M. Lotz, M. Sawall, K. Neymeyr and D. Heller. The influence of substituents in diphosphine ligands on the hydrogenation activity and selectivity of the corresponding rhodium complexes as exemplified by ButiPhane. *ChemCatChem*, 4(1):81–88, 2012.

[42] R. Franke, D. Selent and A. Börner. Applied hydroformylation. *Chem. Rev.*, 112(11):5675–5732, 2012.

[43] G. F. Frobenius. Über Matrizen aus nicht negativen Elementen. *S.-B. Preuss. Akad. Wiss.*, 26:456–477, 1912.

[44] H. Gampp, M. Maeder, C. J. Meyer and A. D. Zuberbühler. Calculation of equilibrium constants from multiwavelength spectroscopic data III: Model-free analysis of spectrophotometric and ESR titrations. *Talanta*, 32(12):1133–1139, 1985.

[45] H. Gampp, M. Maeder, C. J. Meyer and A. D. Zuberbühler. Calculation of equilibrium constants from multiwavelength spectroscopic data IV: Model-free least-squares refinement by use of evolving factor analysis. *Talanta*, 33(12):943–951, 1986.

[46] R. Gargallo, R. Tauler and A. Izquierdo-Ridorsa. Application of a multivariate curve resolution procedure to the analysis of second-order melting data of synthetic and natural polynucleotides. *Anal. Chem.*, 69(9):1785–1792, 1997.

[47] M. Garrido, I. Lázaro, M. S. Larrechi and F. X. Rius. Multivariate resolution of rank-deficient near-infrared spectroscopy data from the reaction of curing epoxy resins using the rank augmentation strategy and multivariate curve resolution alternating least squares approach. *Anal. Chim. Acta*, 515(1):65–73, 2004.

[48] P. J. Gemperline. Target transformation factor analysis with linear inequality constraints applied to spectroscopic-chromatographic data. *Anal. Chem.*, 58(13):2656–2663, 1986.

[49] P. J. Gemperline. Mixture Analysis using factor analysis I: Calibration and quantitation. *J. Chemom.*, 3(4):549–568, 1989.

[50] P. J. Gemperline. Computation of the range of feasible solutions in self-modeling curve resolution algorithms. *Anal. Chem.*, 71(23):5398–5404, 1999.

[51] P. J. Gemperline and E. Cash. Advantages of soft versus hard constraints in self-modeling curve resolution problems. Alternating least squares with penalty functions. *Anal. Chem.*, 75(16):4236–4243, 2003.

[52] P. J. Gemperline and J. C. Hamilton. Evolving factor analysis applied to flow injection analysis data. *J. Chemom.*, 3(3):455–461, 1989.

[53] S. Ghaheri, S. Masoum and A. Gholami. Resolving of challenging gas chromatography-mass spectrometry peak clusters in fragrance samples using multicomponent factorization approaches based on polygon inflation algorithm. *J. Chromatogr. A*, 1429:317–328, 2016.

[54] N. Gillis and F. Glineur. On the geometric interpretation of the nonnegative rank. *Linear Algebra Appl.*, 437(11):2685–2712, 2012.

[55] A. Golshan, H. Abdollahi, S. Beyramysoltan, M. Maeder, K. Neymeyr, R. Rajkó, M. Sawall and R. Tauler. A review of recent methods for the determination of ranges of feasible solutions resulting from soft modelling analyses of multivariate data. *Anal. Chim. Acta*, 911:1–13, 2016.

[56] A. Golshan, H. Abdollahi and M. Maeder. Resolution of rotational ambiguity for three-component systems. *Anal. Chem.*, 83(3):836–841, 2011.

[57] A. Golshan, H. Abdollahi and M. Maeder. The reduction of rotational ambiguity in soft-modeling by introducing hard models. *Anal. Chim. Acta*, 709:32–40, 2012.

[58] A. Golshan, M. Maeder and H. Abdollahi. Determination and visualization of rotational ambiguity in four-component systems. *Anal. Chim. Acta*, 796:20–26, 2013.

[59] G. H. Golub, A. Hoffman and G. W. Stewart. A generalization of the Eckart-Young-Mirsky matrix approximation theorem. *Linear Algebra Appl.*, 88-89:317–327, 1987.

[60] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3 of *Johns Hopkins Stud. Math. Sci.* Johns Hopkins University Press, Baltimore, 2012.

[61] D. A. Gregory and N. J. Pullman. Semiring rank: Boolean rank and nonnegative rank factorizations. *J. Combin. Inform. System Sci*, 8(3):223–233, 1983.

[62] H. Günzler and H. M. Heise. *IR-Spektroskopie*. VCH, Weinheim, 1996.

[63] H. Haario and V. M. Taavitsainen. Combining soft and hard modelling in chemical kinetic models. *Chemom. Intell. Lab. Syst.*, 44(1-2):77–98, 1998.

[64] J. C. Hamilton and P. J. Gemperline. Mixture analysis using factor analysis. II: Self-modeling curve resolution. *J. Chemom.*, 4(1):1–13, 1990.

[65] P. C. Hansen. *Discrete inverse problems: insight and algorithms*, volume 7 of *Fundam. Algorithms*. SIAM, Philadelphia, 2010.

[66] W. Härdle and L. Simar. *Applied multivariate statistical analysis*. Springer, Berlin, 2007.

[67] Z. He, S. Xie, R. Zdunek, G. Zhou and A. Cichocki. Symmetric nonnegative matrix factorization: Algorithms and applications to probabilistic clustering. *IEEE Trans. Neural Netw.*, 22(12):2117–2131, 2011.

[68] B. Hemmateenejad, Z. Shojaeifard, M. Shamsipur, K. Neymeyr, M. Sawall and A. Mohajeri. Solute-induced perturbation of methanol-water association. *RSC Adv.*, 5(87):71102–71108, 2015.

[69] R. C. Henry. Duality in multivariate receptor models. *Chemom. Intell. Lab. Syst.*, 77(1-2):59–63, 2005.

[70] R. C. Henry and B. M. Kim. Extension of self-modeling curve resolution to mixtures of more than three components: Part 1. Finding the basic feasible region. *Chemom. Intell. Lab. Syst.*, 8(2):205–216, 1990.

[71] N.-D. Ho. *Nonnegative matrix factorization algorithms and applications*. PhD thesis, Université catholique de Louvain, 2008.

[72] B. Hofmann. *Mathematik inverser Probleme*. Math. Ingen. Naturwiss. Teubner, Stuttgart,

1999.

[73] M. Hopp. Effektive numerische Berechnung der Oberflächendurchschnitte dreidimensionaler Objekte. Master's thesis, Universität Rostock, 2017.

[74] K. Huang, N. D. Sidiropoulos and A. Swami. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Trans. Signal Process.*, 62(1):211–224, 2014.

[75] A. Hyvärinen, J. Karhunen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Netw.*, 13:411–430, 2000.

[76] A. Hyvärinen, J. Karhunen and E. Oja. *Independent component analysis*, volume 46 of *Adapt. Learn. Syst. Signal Process. Commun. Control*. John Wiley & Sons, New York, 2001.

[77] W. G. Jackson, J. M. Harrowfield and P. D. Vowles. Consecutive, irreversible first-order reactions. Ambiguities and practical aspects of kinetic analyses. *Int. J. Chem. Kinet.*, 9(4):535–548, 1977.

[78] M. Jalali-Heravi and H. Parastar. Recent trends in application of multivariate curve resolution approaches for improving gas chromatography-mass spectrometry analysis of essential oils. *Talanta*, 85(2):835–849, 2011.

[79] J. Jaumot, A. de Juan and R. Tauler. MCR-ALS GUI 2.0: New features and applications. *Chemom. Intell. Lab. Syst.*, 140:1–12, 2015.

[80] J. Jaumot, N. Escaja, R. Gargallo, C. González, E. Pedroso and R. Tauler. Multivariate curve resolution: A powerful tool for the analysis of conformational transitions in nucleic acids. *Nucl. Acids Res.*, 30(17):e92, 2002.

[81] J. Jaumot, R. Gargallo, A. de Juan and R. Tauler. A graphical user-friendly interface for MCR-ALS: A new tool for multivariate curve resolution in MATLAB. *Chemom. Intell. Lab. Syst.*, 76(1):101–110, 2005.

[82] J. Jaumot, P. J. Gemperline and A. Stang. Non-negativity constraints for elimination of multiple solutions in fitting of multivariate kinetic models to spectroscopic data. *J. Chemom.*, 19(2):97–106, 2005.

[83] J. Jaumot and R. Tauler. MCR-BANDS: A user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution. *Chemom. Intell. Lab. Syst.*, 103(2):96–107, 2010.

[84] R. A. Johnson and D. W. Wichern. *Applied multivariate statistical analysis*. Pearson Education International, New Jersey, 2002.

[85] A. Jürß. Über Lösungsmengen nichtnegativer Matrixfaktorisierungen. Master's thesis, Universität Rostock, 2013.

[86] A. Jürß. *Über nichtnegative Matrixfaktorisierungen und geometrische Algorithmen zur Approximation ihrer Lösungsmengen*. PhD thesis, Universität Rostock, 2017.

[87] A. Jürß, M. Sawall and K. Neymeyr. On generalized Borgen plots. I: From convex to affine combinations and applications to spectral data. *J. Chemom.*, 29(7):420–433, 2015.

[88] A. Jürß, M. Sawall and K. Neymeyr. On generalized Borgen plots II: The line-moving algorithm and its numerical implementation. *J. Chemom.*, 30(11):636–650, 2016.

[89] H. R. Keller and D. L. Massart. Evolving factor analysis. *Chemom. Intell. Lab. Syst.*, 12(3):209–224, 1991.

[90] R. Kellner, J.-M. Mermet, M. Otto, M. Valcárcel and H. M. Widmer, editors. *Analytical chemistry*. Wiley-VCH, Weinheim, 2004.

[91] B. M. Kim and R. C. Henry. Extension of self-modeling curve resolution to mixtures of more than three components: Part 2. Finding the complete solution. *Chemom. Intell. Lab. Syst.*, 49(1):67–77, 1999.

[92] B. M. Kim and R. C. Henry. Extension of self-modeling curve resolution to mixtures of more than three components: Part 3. Atmospheric aerosol data simulation studies. *Chemom. Intell. Lab. Syst.*, 52(2):145–154, 2000.

[93] H. Kim and H. Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM J. Matrix Anal. Appl.*, 30(2):713–730, 2008.

[94] H. Kim, H. Park and L. Eldén. Non-negative tensor factorization based on alternating large-scale non-negativity-constrained least squares. *Proceedings of the 7th IEEE international conference on bioinformatics & bioengineering*, 2:1147–1151, 2007.

[95] J. Kim, Y. He and H. Park. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *J. Global Optim.*, 58(2):285–319, 2014.

[96] J. Kim and H. Park. Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM J. Sci. Comput.*, 33(6):3261–3281, 2011.

[97] R. Kramer. *Chemometric techniques for quantitative analysis*. CRC Press, Boca Raton, 1998.

[98] J. B. Kruskal. Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl.*, 18(2):95–138, 1977.

[99] J. B. Kruskal. Rank, decomposition, and uniqueness for 3-way and N-way arrays. In R. Coppi and S. Bolasco, editors, *Multiway data analysis*, pages 7–18. Elsevier (North-Holland), Amsterdam, 1989.

[100] C. Kubis, W. Baumann, E. Barsch, D. Selent, M. Sawall, R. Ludwig, K. Neymeyr, D. Hess, R. Franke and A. Börner. Investigation into the equilibrium of iridium catalysts for the hydroformylation of olefins by combining in situ high-pressure FTIR- and NMR-spectroscopy. *ACS Catal.*, 4(7):2097–2108, 2014.

[101] C. Kubis, M. Sawall, A. Block, K. Neymeyr, R. Ludwig, A. Börner and D. Selent. An operando FTIR spectroscopic and kinetic study of carbon monoxide pressure influence on rhodium-catalyzed olefin hydroformylation. *Chem.-Eur. J.*, 20(37):11921–11931, 2014.

[102] C. Kubis, D. Selent, M. Sawall, R. Ludwig, K. Neymeyr, W. Baumann, R. Franke and A. Börner. Exploring between the extremes: Conversion dependent kinetics of phosphite-modified hydroformylation catalysis. *Chem. Eur. J.*, 18(28):8780–8794, 2012.

[103] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen and S. H. Jensen. Theorems on positive data: On the uniqueness of NMF. *Comput. Intell. Neurosci.*, 2008:9 pages, 2008.

[104] C. L. Lawson and R. J. Hanson. *Solving least squares problems*, volume 15 of *Classics Appl. Math.* SIAM, Philadelphia, 1995.

[105] W. H. Lawton and E. A. Sylvestre. Self modelling curve resolution. *Technometrics*, 13(3):617–633, 1971.

[106] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factori-

zation. *Nature*, 401(6755):788–791, 1999.

[107] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Adv. Neural Inf. Process. Syst.*, 13:556–562, 2001.

[108] C. Li, E. Widjaja and M. Garland. Spectral reconstruction of in situ FTIR spectroscopic reaction data using band-target entropy minimization (BTEM): Application to the homogeneous rhodium catalyzed hydroformylation of 3,3-dimethylbut-1-ene using $Rh_4(CO)_{12}$. *J. Catal.*, 213(2):126–134, 2003.

[109] C.-J. Lin. On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *IEEE Trans. Neural Netw.*, 18(6):1589–1596, 2007.

[110] C.-J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Comput.*, 19(10):2756–2779, 2007.

[111] M. Maeder. Evolving factor analysis for the resolution of overlapping chromatographic peaks. *Anal. Chem.*, 59(3):527–530, 1987.

[112] M. Maeder and Y. M. Neuhold. *Practical data analysis in chemistry*, volume 26 of *Data Handling Sci. Technol.* Elsevier, Amsterdam, 2007.

[113] M. Maeder and A. Zilian. Evolving factor analysis, a new multivariate technique in chromatography. *Chemom. Intell. Lab. Syst.*, 3(3):205–213, 1988.

[114] M. Maeder and A. D. Zuberbuehler. The resolution of overlapping chromatographic peaks by evolving factor analysis. *Anal. Chim. Acta*, 181:287–291, 1986.

[115] A. Malik and R. Tauler. Ambiguities in multivariate curve resolution. In C. Ruckebusch, editor, *Resolving Spectral Mixtures*, volume 30 of *Data Handling Sci. Technol.*, pages 101–133. Elsevier, Cambridge, 2016.

[116] E. R. Malinowski. Window factor analysis: Theoretical derivation and application to flow injection analysis data. *J. Chemom.*, 6(1):29–40, 1992.

[117] E. R. Malinowski. *Factor analysis in chemistry.* Wiley, New York, 2002.

[118] R. Manne. On the resolution problem in hyphenated chromatography. *Chemom. Intell. Lab. Syst.*, 27(1):89–94, 1995.

[119] D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte and L. Kaufman. *Chemometrics: A textbook*, volume 2 of *Data Handling Sci. Technol.* Elsevier, Amsterdam, 1988.

[120] A. Meister. Estimation of component spectra by the principal components method. *Anal. Chim. Acta*, 161:149–161, 1984.

[121] H. Minc. *Nonnegative matrices.* John Wiley & Sons, New York, 1988.

[122] K. Neymeyr and M. Sawall. On an SVD-free approach to the complementarity and coupling theory: A note on the elimination of unknowns in sums of dyadic products. *J. Chemom.*, 30(1):30–36, 2016.

[123] K. Neymeyr and M. Sawall. On the set of solutions of the nonnegative matrix factorization problem. 39:1049–1069, 2018. SIAM J. Matrix Anal. Appl.

[124] K. Neymeyr, M. Sawall and D. Hess. Pure component spectral recovery and constrained matrix factorizations: Concepts and applications. *J. Chemom.*, 24(2):67–74, 2010.

[125] P. Paatero. A weighted non-negative least squares algorithm for three-way 'PARAFAC' factor analysis. *Chemom. Intell. Lab. Syst.*, 38(2):223 – 242, 1997.

[126] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with

optimal utilization of error estimates of data values. *Environmetrics*, 5:111–126, 1994.

[127] H. Parastar. Multivariate curve resolution methods for qualitative and quantitative analysis in analytical chemistry. In A. de la Peña, H. Goicoechea, G. Escandar and A. Olivieri, editors, *Fundamentals and and analytical applications of multiway calibration*, volume 29 of *Data Handling Sci. Technol.*, pages 293–345. Elsevier, 2015.

[128] V. P. Pauca, J. Piper and R. J. Plemmons. Nonnegative matrix factorization for spectral data analysis. *Linear Algebra Appl.*, 416(1):29–47, 2006.

[129] O. Perron. Zur Theorie der Matrices. *Math. Ann.*, 64(2):248–263, 1907.

[130] R. J. Plemmons. Regular nonnegative matrices. *Proc. Amer. Math. Soc.*, 39(1):26–32, 1973.

[131] M. Radolko. Adaptive Oberflächenapproximation für Lösungsmengen nichtnegativer Matrixfaktorisierungen. Master's thesis, Universität Rostock, 2014.

[132] N. Rahimdoust, M. Sawall, K. Neymeyr and H. Abdollahi. Investigating the effect of flexible constraints on the accuracy of self-modeling curve resolution methods in the presence of perturbations. *J. Chemom.*, 30(5):252–267, 2016.

[133] R. Rajkó. Natural duality in minimal constrained self modeling curve resolution. *J. Chemom.*, 20(3-4):164–169, 2006.

[134] R. Rajkó. Some surprising properties of multivariate curve resolution-alternating least squares (MCR-ALS) algorithms. *J. Chemom.*, 23(4):172–178, 2009.

[135] R. Rajkó. Studies on the adaptability of different Borgen norms applied in self-modeling curve resolution (SMCR) method. *J. Chemom.*, 23(6):265–274, 2009.

[136] R. Rajkó. Additional knowledge for determining and interpreting feasible band boundaries in self-modeling/multivariate curve resolution of two-component systems. *Anal. Chim. Acta*, 661(2):129–132, 2010.

[137] R. Rajkó, H. Abdollahi, S. Beyramysoltan and N. Omidikia. Definition and detection of data-based uniqueness in evaluating bilinear (two-way) chemical measurements. *Anal. Chim. Acta*, 855:21–33, 2015.

[138] R. Rajkó and K. István. Analytical solution for determining feasible regions of self-modeling curve resolution (SMCR) method based on computational geometry. *J. Chemom.*, 19(8):448–463, 2005.

[139] C. Ruckebusch, editor. *Resolving Spectral Mixtures*, volume 30 of *Data Handling Sci. Technol.* Elsevier, Cambridge, 2016.

[140] P. Sajda, S. Du and L. Parra. Recovery of constituent spectra using non-negative matrix factorization. In *Wavelets: Applications in Signal and Image Processing X*, volume 5207, pages 321–331. International Society for Optics and Photonics, 2003.

[141] R. Sandler and M. Lindenbaum. Nonnegative matrix factorization with earth mover's distance metric for image analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1590–1602, 2011.

[142] K. Sasaki, S. Kawata and S. Minami. Constrained nonlinear method for estimating component spectra from multicomponent mixtures. *Appl. Opt.*, 22(22):3599–3603, 1983.

[143] M. Sawall. *Regularisierte nichtnegative Matrixfaktorisierungen und ihre Anwendungen in der Spektroskopie*. PhD thesis, Universität Rostock, 2011.

[144] M. Sawall, A. Börner, C. Kubis, D. Selent, R. Ludwig and K. Neymeyr. Model-free multi-

variate curve resolution combined with model-based kinetics: algorithm and applications. *J. Chemom.*, 26(10):538–548, 2012.

[145] M. Sawall, C. Fischer, D. Heller and K. Neymeyr. Reduction of the rotational ambiguity of curve resolution techniques under partial knowledge of the factors. Complementarity and coupling theorems. *J. Chemom.*, 26(10):526–537, 2012.

[146] M. Sawall, A. Jürß and K. Neymeyr. FACPACK: A software for the computation of multi-component factorizations and the area of feasible solutions, Revision 1.3. FACPACK homepage: http://www.math.uni-rostock.de/facpack/, 2015.

[147] M. Sawall, A. Jürß, H. Schröder and K. Neymeyr. On the analysis and computation of the area of feasible solutions for two-, three- and four-component systems. In C. Ruckebusch, editor, *Resolving Spectral Mixtures*, volume 30 of *Data Handling Sci. Technol.*, pages 135–184. Elsevier, Cambridge, 2016.

[148] M. Sawall, A. Jürß, H. Schröder and K. Neymeyr. Simultaneous construction of dual Borgen plots. I: The case of noise-free data. *J. Chemom.*, 31(12):e2954, 2017.

[149] M. Sawall, C. Kubis, E. Barsch, D. Selent, A. Börner and K. Neymeyr. Peak group analysis for the extraction of pure component spectra. *J. Iran. Chem. Soc.*, 13(2):191–205, 2016.

[150] M. Sawall, C. Kubis, A. Börner, D. Selent and K. Neymeyr. A multiresolution approach for the convergence acceleration of multivariate curve resolution methods. *Anal. Chim. Acta*, 891:101–112, 2015.

[151] M. Sawall, C. Kubis, R. Franke, D. Hess, D. Selent, A. Börner and K. Neymeyr. How to apply the complementarity and coupling theorems in MCR methods: Practical implementation and application to the Rhodium-catalyzed hydroformylation. *ACS Catal.*, 4(9):2836–2843, 2014.

[152] M. Sawall, C. Kubis, D. Selent, A. Börner and K. Neymeyr. A fast polygon inflation algorithm to compute the area of feasible solutions for three-component systems. I: Concepts and applications. *J. Chemom.*, 27(5):106–116, 2013.

[153] M. Sawall, A. Moog, C. Kubis, H. Schröder, D. Selent, R. Franke, A. Brächer, A. Börner and K. Neymeyr. Simultaneous construction of dual Borgen plots. II: Algorithmic enhancement for applications to noisy spectral data. *J. Chemom.*, 32(6):e3012, 2018.

[154] M. Sawall and K. Neymeyr. A fast polygon inflation algorithm to compute the area of feasible solutions for three-component systems. II: Theoretical foundation, inverse polygon inflation, and FACPACK implementation. *J. Chemom.*, 28(5):633–644, 2014.

[155] M. Sawall and K. Neymeyr. How to compute the area of feasible solutions. A practical study and users' guide to FACPACK. In M. Khanmohammadi, editor, *Current Applications of Chemometrics*, pages 97–134. Nova Science Publishers, New York, 2014.

[156] M. Sawall and K. Neymeyr. On the area of feasible solutions and its reduction by the complementarity theorem. *Anal. Chim. Acta*, 828:17–26, 2014.

[157] M. Sawall and K. Neymeyr. A ray casting method for the computation of the area of feasible solutions for multicomponent systems: Theory, applications and FACPACK-implementation. *Anal. Chim. Acta*, 960:40–52, 2017.

[158] M. Sawall, N. Rahimdoust, C. Kubis, H. Schröder, D. Selent, D. Hess, H. Abdollahi, R. Franke, Börner A. and K. Neymeyr. Soft constraints for reducing the intrinsic rotational ambiguity of the area of feasible solutions. *Chemom. Intell. Lab. Syst.*, 149:140–150, 2015.

[159] K. J. Schostack and E. R. Malinowski. Investigation of window factor analysis and matrix regression analysis in chromatography. *Chemom. Intell. Lab. Syst.*, 20(2):173–182, 1993.

[160] H. Schröder. Kinetische Modellierung für multivariate Faktormethoden. Master's thesis, Universität Rostock, 2013.

[161] H. Schröder, M. Sawall, C. Kubis, A. Jürß, D. Selent, A. Brächer, A. Börner, R. Franke and K. Neymeyr. Comparative multivariate curve resolution study in the area of feasible solutions. *Chemom. Intell. Lab. Syst.*, 163:55–63, 2017.

[162] H. Schröder, M. Sawall, C. Kubis, D. Selent, D. Hess, R. Franke, A. Börner and K. Neymeyr. On the ambiguity of the reaction rate constants in multivariate curve resolution for reversible first-order reaction systems. *Anal. Chim. Acta*, 927:21–34, 2016.

[163] E. Seneta. *Nonnegative matrices and Markov chains*. Springer Ser. Statist. Springer, New York, 2000.

[164] A. N. Skvortsov. Estimation of rotation ambiguity in multivariate curve resolution with charged particle swarm optimization. *J. Chemom.*, 28(10):727–739, 2014.

[165] A. Stegeman and N. D. Sidiropoulos. On Kruskal's uniqueness condition for the Candecomp/Parafac decomposition. *Linear Algebra Appl.*, 420(2-3):540–552, 2007.

[166] G. W. Stewart. On the early history of the singular value decomposition. *SIAM Rev.*, 35:551–556, 1993.

[167] R. Tauler. Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution. *J. Chemom.*, 15(8):627–646, 2001.

[168] R. Tauler. Application of non-linear optimization methods to the estimation of multivariate curve resolution solutions and of their feasible band boundaries in the investigation of two chemical and environmental simulated data sets. *Anal. Chim. Acta*, 595(1):289–298, 2007.

[169] R. Tauler, A. Smilde and B. Kowalski. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *J. Chemom.*, 9(1):31–58, 1995.

[170] L. B. Thomas. Rank factorization of nonnegative matrices (A. Berman). *SIAM Rev.*, 16(3):393–394, 1974.

[171] M. Tjahjono, X. Li, F. Tang, K. Sa-ei and M. Garland. Kinetic study of a complex triangular reaction system in alkaline aqueous-ethanol medium using on-line transmission FTIR spectroscopy and BTEM analysis. *Talanta*, 85(5):2534 – 2541, 2011.

[172] R. S. Varga. *Matrix iterative analysis*, volume 27 of *Springer Ser. Comput. Math.* Springer, Heidelberg, 2000.

[173] M. Vosough, C. Mason, R. Tauler, M. Jalali-Heravi and M. Maeder. On rotational ambiguity in model-free analyses of multivariate data. *J. Chemom.*, 20(6-7):302–310, 2006.

[174] Y.-X. Wang and Y.-J. Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Trans. Knowl. Data Eng.*, 25(6):1336–1353, 2013.

[175] E. Widjaja, C. Li, W. Chew and M. Garland. Band target entropy minimization. A robust algorithm for pure component spectral recovery. Application to complex randomized mixtures of six components. *Anal. Chem.*, 75(17):4499–4507, 2003.

[176] E. Widjaja, C. Li and M. Garland. Semi-batch homogeneous catalytic in-situ spectroscopic data. FTIR spectral reconstructions using Band-Target Entropy Minimization (BTEM) without spectral preconditioning. *Organometallics*, 21(9):1991–1997, 2002.

[177] Y.-L. Xie, P. K. Hopke and P. Paatero. Positive matrix factorization applied to a curve resolution problem. *J. Chemom.*, 12(6):357–364, 1999.

[178] X. Zhang and R. Tauler. Measuring and comparing the resolution performance and the extent of rotation ambiguities of some bilinear modeling methods. *Chemom. Intell. Lab. Syst.*, 147:47–57, 2015.

[179] G. M. Ziegler. *Lectures on polytopes*, volume 152 of *Grad. Texts in Math.* Springer, New York, 1994.