CARL VON OSSIETZKY universität OLDENBURG

Forschungszentrum Jülich
in der Helmholtz-Gemeinschaft

Carl von Ossietzky Universität Oldenburg
Diplomstudiengang Mathematik
mit dem Schwerpunkt Biowissenschaften

DIPLOMARBEIT

Robuste Parameterschätzung, Modelldiskriminierung und optimale Versuchsplanung am Beispiel von In-vitro-Datensätzen zur Benzaldehydlyase

vorgelegt von: Anna Zofia Siudak
vorgelegt am: 17.12.2007

Betreuende Gutachterin: Prof. Dr. Christine Müller
Zweiter Gutachter: Dr. Peter Harmand
Betreuer im Forschungszentrum Jülich: Dr.-Ing. Eric von Lieres

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

Abkürzung	**Bedeutung**
Abb.	Abbildung
abs. Häufigkeit	absolute Häufigkeit
AIC	Akaike-Informations-Kriterium
AIC_C	Akaike-Informations-Kriterium für kleine Datensätze
BA	Benzaldehyd
BAL	Benzaldehydlyase
BZ	Benzoin
C-C-Bindung	Bindung zwischen zwei Kohlenstoffatomen
DALD	Dimethoxyaldehyd
DGL-System	Differentialgleichungssystem
DGL-Löser	numerischer Differentialgleichungslöser
DHPP	(R)-3,3-Dimethoxy-1-phenyl-2-hydroxypropanon
DMSO	Dimethylsulfoxid
ee	Enantiomerenüberschuss (engl. *enantiomeric excess*)
EC - Nummer	Klassifikationsnummer für Enzyme (engl. *Enzyme Classification Number*)
gKQSS	gewichtete Kleinste-Quadrat-Summen-Schätzung
HPLC	Hochleistungsflüssigchromatographie (engl. *high performance liquid chromatography*)
i.i.d.	Bezeichnung für stochastisch unabhängige Zufallsvariablen mit gleicher Verteilung (engl. *independent and identically distributed*)
KLI	Kullback-Leibler-Information
MCData	Simulierte Datensätze (parametrischer Bootstrap)
ML-Schätzer	Maximum-Likelihood-Schätzer
mM	Millimolar
pH-Wert	Maß, das die Stärke einer Säure oder Base angibt (lat. *potentia Hydrogenii* bzw. *pondus Hydrogenii*)
SIMUL	simulierte Datensätze, bei denen der Startparameter verrauscht wurde
ThDP	Thiamindiphosphat
TOL	Toleranzkriterien für die Optimierung
U	Einheit der Enzymaktivität

Symbolverzeichnis

Symbol	**Bedeutung**
a_0	Anfangsbedingungen eines Differentialgleichungssystems
Corr(Y)	Korrelationsmatrix zum Zufallsvektor Y
Cov(Y)	Kovarianzmatrix zum Zufallsvektor Y
E(Y)	Erwartungswert des Zufallsvektors Y
E(h)	Abbruchfehler der numerischen Differentiation
$F(\theta)$	Fisher-Informationsmatrix zum Parametersatz θ
$F(\zeta)$	Informationsmatrix auf Basis des Versuchsplans ζ
$g_1(t), \ldots, g_n(t)$	Lösungen eines Differentialgleichungssystems
$H(G; \theta; a_0)$	vektorielle Schreibweise des Differentialgleichungssystems mit Parametern θ und Anfangsbedingungen a_0
J	Systemmatrix eines Regressionsmodells
$L(\theta; M)$	Maximum-Likelihoodfunktion
$l(\theta; M)$	Log-Likelihoodfunktion
L_D	$L(G; \theta; M_{1\bullet}; T_D)$, Matrix der simulierten Messwerte
l_{Dij}	Einträge der Matrix L_D
M	Matrix der Messdaten
$M_{i\bullet}$	i-te Zeile der Matrix M
$M_{\bullet j}$	j-te Spalte der Matrix M
m_{ij}	Einträge der Matrix M
n	Dimension eines Differentialgleichungssystems
N	Stichprobengröße
$\mathcal{N}_p(\mu, \Sigma)$	p-dimensionale Normalverteilung mit Erwartungswert μ und Kovarianzmatrix Σ
$O(h^n)$	Maß für die Konvergenzgeschwindigkeit
p	Anzahl der Elemente des Parametervektors
P	Punktgitter
s	Anzahl der Elemete des Vektors T
s_D	Anzahl der Elemente des Vektors T_D
$S(\theta; M)$	Score-Funktion
$Sens_M$	Output-Sensitivität
$Sens_{\hat{\theta}}$	Parameter-Sensitivität
T_D	Vektor der experimentellen Messzeitpunkte
tr(M)	Spur der Matrix M ($\sum_i m_{ii}$)
$u_G(t)$	Gitterfunktion
V	Raum aller möglichen Tupel von Eingangsvariablen
V^+	Menge aller Versuchspläne $\zeta \in V$, für die det $F(\zeta) \neq 0$ gilt
Var(Y)	Varianz des Zufallsvektors Y

$x_1, \ldots, x_m$	Eingangsvariablen des Systems
$\delta^{(q)}$	q-ter Abstiegsschritt eines Optimierungsverfahrens
ε	Matrix der Messwertfehler im Regressionsmodell
ε_G	Matrix, die das Grundrauschen der Messwerte enthält
$\epsilon(M; L_D(G; (\theta))$	Vektor aller Residuen der gKQSS
Γ	Matrix der Eigenvektoren
Λ	Matrix der Eigenwerte
Θ	Parameterraum
θ	unbekannter Parametervektor
$\theta^{(1)}$	Startparametersatz für eine Optimierungsroutine über $\theta \in \Theta$
$\hat{\theta}$	Schätzung für den Parametersatz θ
$\hat{\theta}_{KQSS}$	Schätzung für θ mit der Methode der kleinsten Quadrate
$\hat{\theta}_{LTS}$	Least-Trimmed-Squares-Schätzung für θ
$\hat{\theta}_{ML}$	Maximum-Likelihood-Schätzung für θ
ζ	Versuchsplan

Danksagung

Die vorliegende Diplomarbeit wurde am Institut für Biotechnologie 2 im Forschungszentrum Jülich angefertigt.

Ganz herzlich möchte ich mich bei Herrn Dr.-Ing. Eric von Lieres bedanken, der mir das Thema gestellt und mich während der Diplomarbeit hervorragend betreut und beraten hat, ohne mich und meine Neugier einzuengen.

Bei Frau Prof. Dr. Christine Müller und Herrn Dr. Peter Harmand bedanke ich mich ebenfalls für die Betreuung der Diplomarbeit als erste Gutachterin und zweiter Gutachter, sowie für weiterführende und motivierende Anmerkungen in der Arbeitsphase.

Weiterhin möchte ich mich bei Frau Dr. Katharina Nöh bedanken, die mir geduldig und unermüdlich mit Ratschlägen zum mathematischen Teil dieser Arbeit zur Seite gestanden hat.

Ich bedanke mich bei Herrn Dipl.-Ing. Stephan Noack, Frau Maria Brune und Herrn Joel Andersson für die herzliche Atmosphäre der Bürogemeinschaft, die fachliche Unterstützung und die angeregten interdisziplinären Diskussionen.

Bedanken möchte ich mich auch bei Herrn Dr. Jürgen Hubbuch und der Aufarbeitungsgruppe des IBT-2 für ihre Unterstützung und die herzliche Aufnahme in die Gruppe.

Ich bedanke mich bei Herrn Dr. Stephan Lütz, Herrn Dipl.-Ing. Sven Kühl und Herrn Sebastian Grefen für ihre fachliche Unterstützung und Mitarbeit.

Weiterhin möchte ich mich bei Herrn Prof. Dr. C. Wandrey sowie der Forschungszentrum Jülich GmbH für die Bereitstellung der Mittel bedanken und für die Möglichkeit, meine Diplomarbeit im Forschungszentrum anzufertigen.

Mein besonderer Dank gilt meinen Eltern, die mich nicht nur finanziell unterstützt, sondern mich die Jahre meines Studiums unermüdlich mit Rat, Tat und ihrer Liebe begleitet haben.

HERR, DEINE LIEBE IST WIE GRAS UND UFER,
WIE WIND UND WEITE UND WIE EIN ZUHAUS...

ANDERS FROSTENSON/ERNST HANSEN

1. Einleitung

1.1. Motivation

Mit fortschreitender Methodenentwicklung in der Biotechnologie und anderen Wissenschaften, erzeugen neue Geräte, Experimentier- und Messmethoden immer größere Datenmengen. Diese können nicht nur computergestützt verwaltet und gegebenenfalls visualisiert, sondern auch ausgewertet werden. Die Auswertung der Daten soll Ergebnisse liefern, die als Argumente im wissenschaftlichen Diskurs oder als Grundlage weiterer Experimente dienen können. Deshalb ist es nicht nur wichtig, dass die Auswertung korrekt und mathematisch fundiert vorgenommen wird, sondern auch, dass die Grenzen ihrer Aussagekraft benannt werden können.

Größere Speicherkapazitäten und Rechnergeschwindigkeiten eröffnen in der Versuchsauswertung und -planung zunehmend die Möglichkeit, das betrachtete System als mathematisches Modell darzustellen. Aus den zur Verfügung stehenden Daten müssen zunächst mit Hilfe mathematischer Methoden unbekannte Parameter des Modells geschätzt werden. Auf Grundlage solcher Modelle und Parameterschätzungen können dann verschiedene Systemzustände untersucht und Voraussagen für die Ergebnisse zukünftiger Experimente getroffen werden.

Werden diese Experimente realisiert, so kann anhand der erhaltenen Messdaten überprüft werden, ob die bisher angenommenen Modelleigenschaften den Ergebnissen entsprechen, was in der Regel zu einer Verwerfung oder Verbesserung des Ausgangsmodells führt. Auf diese Weise sind Modell und Experiment in einem Optimierungs-Kreislauf miteinander verknüpft, wie in Abbildung 1.1 verdeutlicht wird. Dieser Kreislauf führt in den meisten Fällen schließlich zu einem Modell, das Voraussagen in gewünschter Qualität liefert.

In dieser Arbeit wurden Daten ausgewertet, die aus Experimenten zum Enzym Benzaldehydlyase (BAL) stammen. Die genaue Funktionsweise des Enzyms in der betrachteten Reaktion ist noch unbekannt. Für verschiedene Hypothesen zu seiner Funktion, beziehungsweise zum Verlauf der beobachteten Reaktion mitsamt ihrer Teilreaktionen sollten mathematische Modelle aufgestellt werden. Diese Modelle wurden gemäß gebräuchlicher Modellvorstellungen für Enzym-Kinetiken formuliert. Sie enthielten unbekannte kinetische Parameter, die auf Grundlage der in dieser Arbeit betrachteten Messdaten bestimmt werden sollten.

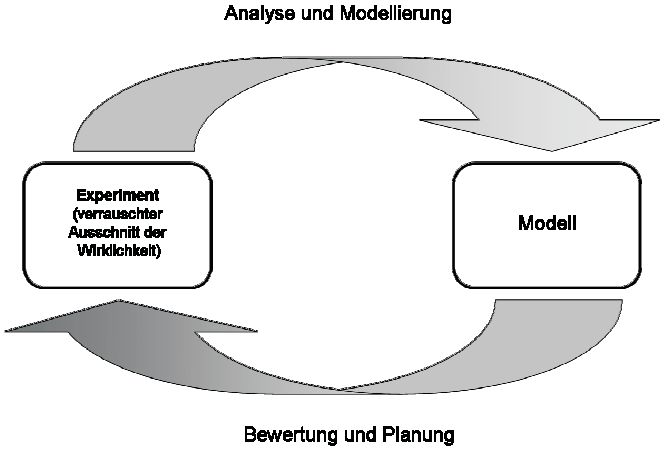

Abbildung 1.1.: Wechselwirkung zwischen Experiment und Modell

Durch mathematische Modelldiskriminierung sollte anschließend aus allen Modellformulierungen ein den anderen überlegenes Modell für die BAL-katalysierte Reaktion ausgewählt werden. Um dieses Modell an Stellen zu verbessern, die sich bis zu diesem Arbeitsschritt als unzureichend oder fehlerhaft herausgestellt haben würden, sollten mit Hilfe der optimalen Versuchsplanung auf seiner Grundlage neue Experimente ausgearbeitet werden. Nach der Durchführung der geplanten Experimente sollte außerdem der Erfolg der Planung selber bewertet werden.

1.2. Aufbau der Arbeit

Wie in Abbildung 1.1 dargestellt, handelt es sich bei einer Prozessoptimierung um einen Kreislauf zwischen Experiment und Modell, der vorzugsweise dann endet, wenn letzteres einer vorher definierten Qualität gerecht wird. In dieser Diplomarbeit wird auf Basis dreier Experimente zur BAL eine „erste Umrundung“ vollzogen, von der Modellformulierung bis zu der Planung eines neuen Experiments.

Der Einleitung folgt ein Kapitel, das gemäß des aktuellen Wissensstands Auskunft über die BAL und ihre Funktionsweise gibt. Dem folgen die für die Datenauswertung benötigten mathematischen Grundlagen. Im vierten Kapitel wird der Verlauf der drei Experimente geschildert, auf deren Messdaten die Modellierung in dieser Arbeit beruht. Ergänzend folgt eine Diskussion der Datenlage, sowie des einzigen mathematischen Modellvorschlags für die betrachtete Reaktion, der in der aktuellen Forschung zu finden ist. Auf Basis der Datenlage, des bisherigen Wissens über die BAL sowie zweier verschiedener Hypothesen zum Reaktionsverlauf werden daraufhin die grundlegenden Differentialgleichungsmodelle (Grundmodelle) dieser Arbeit hergeleitet.

Die Ergebnisse der mit verschiedenen statistischen Methoden berechneten Schätzungen der unbekannten kinetischen Parameter der Grundmodelle und erweiterter

Modelle werden in Kapitel 5 dargestellt. Anschließend werden die Ergebnisse der Modelldiskriminierung und die sich daraus ergebenden Konsequenzen für die optimale Versuchsplanung diskutiert. Es folgen Szenarien der optimalen Versuchsplanung für unterschiedliche Zielsetzungen und Empfehlungen zur Durchführung eines neuen Experiments. Im letzten Kapitel wird eine Zusammenfassung der Ergebnisse dieser Arbeit gegeben, der ein Ausblick auf weitere erstrebenswerte Experimente und Modellverbesserungen folgt. Den Schluss bildet ein persönlicher Kommentar zu dieser Arbeit.

2. Einführung in das Thema

2.1. Enzyme

2.1.1. Ein Enzym – Was ist das überhaupt?

In dieser Arbeit werden Messdaten betrachtet, die aus einer enzymatischen Umsetzung stammen. Dieses Kapitel stellt die Grundlagen der Funktionsweise von Enzymen dar. Erst diese Grundlagen ermöglichen ein Verständnis des Experiments und die Formulierung von mathematischen Modellen für die Reaktion.

Enzyme katalysieren chemische Reaktionen, das bedeutet, dass sie die für eine Umsetzung nötige Aktivierungsenergie herabsetzen und die katalysierte Reaktion sehr viel schneller als eine Reaktion ohne Katalysator verläuft (vergleiche Abb. 2.1). Der Unterschied in der Reaktionsgeschwindigkeit zu einer nicht katalysierten Reaktion kann bei einem Faktor von bis zu 10^7 liegen. Enzyme sind lebenswichtig für alle Organismen der Erde, da diese für ihr Wachstum Energie aus chemischen Reaktionen nutzen müssen.

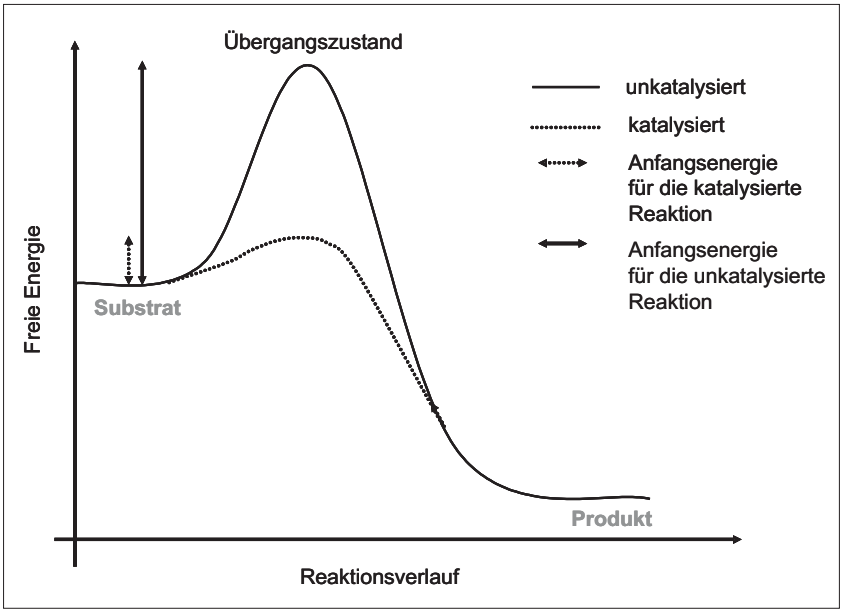

Abbildung 2.1.: Energiediagramm einer enzymatischen Reaktion

Werden beispielsweise Kohlenstoff-Kohlenstoff-Bindungen (C-C-Bindungen) von Zuckermolekülen gespalten, so kann die dabei frei werdende Bindungsenergie durch

den Organismus genutzt werden. Solche Reaktionen würden aber beispielsweise in wässriger, neutraler Umgebung bei 37°C (so sind die Bedingungen im menschlichen Organismus) viel zu langsam ablaufen. Erst Enzyme sorgen für einen schnellen und damit nutzbaren Umsatz des Zuckers. Die katalytischen Fähigkeiten von Enzymen liegen in ihrer Molekülstruktur begründet.

Enzyme bestehen aus aneinandergereihten Aminosäuren. Wechselwirkungen zwischen diesen Säuren sorgen für eine dreidimensionale Struktur des Moleküls, die seine Funktionalität ausmacht. Manche Unterstrukturen der Enzyme wechselwirken spezifisch mit einem Substrat, vergleichbar mit einer Tasche in der Enzymkonformation, in die nur ein bestimmtes Molekül passt. Deshalb werden solche Strukturen auch als *Bindungstaschen* des Enzyms bezeichnet. Durch die Bindung eines Substrats oder mehrerer Substrate an das Enzym entsteht ein *Enzym-Substrat-Komplex*. Die Substratspezifität ist nicht bei jedem Enzym gleich stark und oft katalysieren Enzyme die Umsetzung einiger ähnlicher Substrate.

Nun erfolgt die eigentliche Umsetzungsreaktion, die durch verschiedene katalytische Mechanismen gefördert wird. Es treten beispielsweise reaktionsbegünstigende Orientierungs- und Nachbarschaftseffekte nach Wechselwirkung zweier Reaktionspartner mit dem Enzym ein. Veränderungen in der Enzymkonformation, das heißt in der dreidimensionalen Struktur des Enzyms, können Substrate spalten oder zu Bindungen zwischen Substraten führen. Oft tragen auch an das Enzym gebundene Moleküle wie Vitamine oder Metallionen, sogenannte *Kofaktoren*, zur katalytischen Umsetzung bei.

Nach der Umsetzung des Substrats werden alle Produkte vollständig vom Enzym abgespalten. Das bedeutet, dass Enzyme bei den von ihnen katalysierten Reaktionen nicht verbraucht werden. In manchen Fällen muss das Enzym allerdings nach der Katalyse erst zu seiner ursprünglichen Konformation regeneriert werden, bevor es wieder funktionsfähig ist.

Die Anzahl der verschiedenen Enzyme im Organismus wird auf genetischer Ebene bestimmt und reguliert. Die *Enzymaktivität* hingegen wird von mehreren Faktoren direkt während der Reaktion beeinflusst. Sie ist definiert als die Substratmenge, die pro Zeiteinheit in Anwesenheit einer bestimmten Menge des Enzyms umgesetzt wird und wird in der Einheit U = μmol $\cdot$ min^{-1} angegeben, sofern die Bestimmung unter standardisierten Bedingungen erfolgt [Mic99]. Die Enzymaktivität kann zum einen durch den pH-Wert oder die Temperatur der Reaktionsumgebung gesteigert (*Aktivation*) oder gehemmt (*Inhibition*) werden, zum anderen aber auch durch die Einwirkung von Metaboliten in seiner Umgebung (Substrate, Produkte, Aktivatoren, Inhibitoren, Kofaktoren, etc.).

Bei der Inhibition wird zwischen kompetetiver und allosterischer Hemmung unterschieden (vergl. Abb. 2.2). Kompetetive Inhibitoren reagieren reversibel mit einer

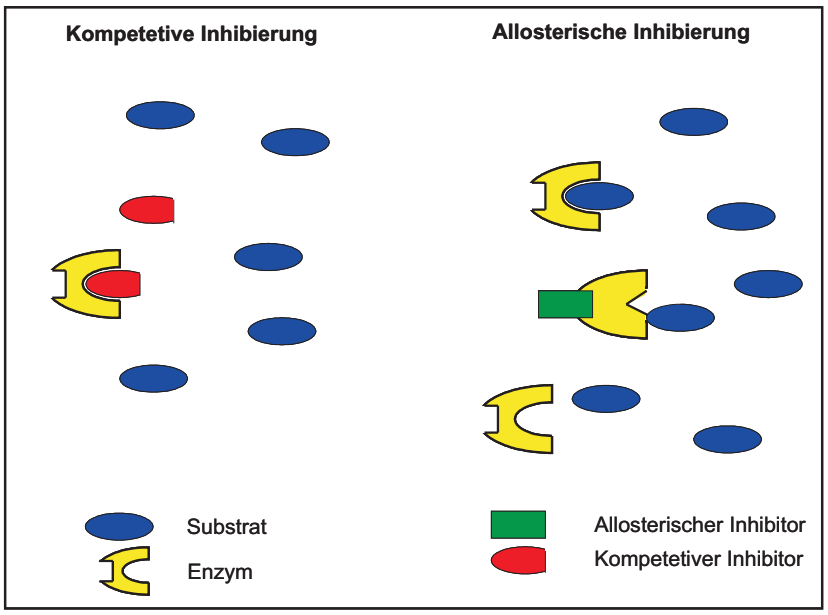

Abbildung 2.2.: Kompetetive und allosterische Enzyminhibierung

Substrat-Bindungsstelle des Enzyms und bilden Enzym-Inhibitor-Komplexe, die die Bindung des Substrats und dessen Umsetzung zu Reaktionsprodukten verhindern. Diese Art der Hemmung kann durch Erhöhung der Substratkonzentration oder Verringerung der Inhibitor-Konzentration vermindert werden [Mic99]. Bei der allosterischen Hemmung bindet ein Inhibitor an eine andere Stelle des Enzyms und verändert dadurch die Struktur der Bindungstasche, so dass das Substrat nicht mehr mit dem Enzym wechselwirken kann. Ein besonders häufiger Fall tritt ein, wenn das Reaktionsprodukt selbst der Inhibitor ist und die weitere Umsetzung des Substrats zum Abklingen bringt.

2.1.2. Die Nutzung enzymatisch katalysierter Reaktionen

Chemische Reaktionen, die durch Enzyme katalysiert werden, sind der Menschheit schon sehr lange bekannt. Die Herstellung von Bier und Wein sowie die Produktion von Joghurt und Käse sind auf enzymatische Reaktionen zurückzuführen. Verfahren zur Nutzung der Fermentation für die Herstellung von Genussmitteln oder zum Zwecke der einfacheren Verdaulichkeit von Milchprodukten sind also seit der Antike bekannt.

Das Nutzungsspektrum für Enzyme ist heutzutage erheblich breiter geworden und trotzdem kommen ständig neue Möglichkeiten hinzu. Das Fachgebiet, in dem die bekannten katalytischen Prozesse und die Erforschung neuer Verfahren zusammengefasst sind, heißt *Biokatalyse*. Die Palette der Methoden für die Produktion ist breit gefächert und reicht von der Verwendung isolierten reinen Enzyms (*In-Vitro-Methoden*) bis zu dem Einsatz von Organismen, die das Produkt *in vivo* herstellen. Hierbei wird es nach Abschluss der Reaktion durch Zellaufschluss gewonnen. Enzy-

me sind beispielsweise in Waschmitteln enthalten und viele Nahrungsmittelzusätze (Aminosäuren, Vitamine) werden in großem Maßstab mit Hilfe von Enzymen produziert.

Während bei der chemischen Synthese von Molekülen oft hohe Temperaturen und ein gewisser Aufwand zur Herstellung einer geeigneten Reaktionsumgebung nötig sind, katalysieren Enzyme die Umsetzung zu den gewünschten Endprodukten unter vergleichsweise milden Bedingungen, was einen geringeren Verbrauch von Energieressourcen bedeutet. Außerdem ist die hohe Spezifität der Enzyme gegenüber dem Ausgangs- und dem Endprodukt ein Vorteil, vor allem, wenn beispielsweise ein kostengünstiges Substratgemisch eingesetzt, aber eine hohe Reinheit des Endproduktes erreicht werden soll. Eine besondere Rolle spielt hier die Enantioselektivität von Enzymen.

Enantiomere sind chemische Verbindungen, die zwar die gleiche Summenformel besitzen, aber eine unterschiedliche Struktur im Raum haben. Man unterscheidet dabei zwischen der R-Konfiguration und der S-Konfiguration eines Moleküls, angelehnt an das lateinische *rectus* und *sinister*, was „rechts“ und „links“ bedeutet. Chemisch gesehen spricht man beim Vorkommen beider Konfigurationen von der Moleküleigenschaft der *Chiralität* (gr. *chiros* = Hand).

Die Konfigurationen sind räumlich so ausgerichtet, dass die eine ein Spiegelbild der anderen ist, sie also nur durch eine Spiegelachse zur Deckung gebracht werden können, genau wie die rechte und die linke Hand [NC01]. Da von der Struktur eines Moleküls seine physikalischen Eigenschaften abhängen, ist es möglich, chirale Moleküle zu unterscheiden. Enantiomere können zum Beispiel durch ihre Wechselwirkung mit polarisiertem Licht unterschieden werden, wobei die R-Konfiguration rechts-, die L-Konfiguration linksdrehend ist. Außerdem ist eine Unterscheidung manchmal auch mit menschlichen Sinnen möglich; beispielsweise riecht (R)-Carvon nach Minze, (S)-Carvon aber nach Kümmel. Ein Gemisch aus Enantiomeren heißt *Racemat* beziehungsweise *racemisches Gemisch*.

Durch die spezielle Konformation der Bindungstasche eines Enzyms kann es bei chiralen Substraten in der Regel nur mit einer der beiden Konfigurationen wechselwirken. Diese Eigenschaft heißt *Stereospezifität*. Dadurch entsteht bei der Umsetzung auch nur eine der möglichen Konfigurationen des Produkts. Diese durch die Stereospezifität entstehende Enantiomerenreinheit bei Produkten von enzymatisch katalysierten Reaktionen beträgt oft mehr als 99% der gewünschten Konfiguration im Produkt und wird in *ee* angegeben. Enantiomerenreinheit ist bei chiralen Wirkstoffen in der Medizin besonders notwendig, bei denen nur ein Enantiomer die gewünschte therapeutische Wirkung hat, das andere dagegen keine Effekte zeigt oder im schlimmsten Falle schädlich ist [Pet00].

Die Stereospezifität der Enzyme spielt eine große Rolle in der Biokatalyse, da bei

einer künstlichen Synthese im Labor immer ein Racemat entsteht [NC01]. Für die klassische organische Chemie ist es allerdings sehr aufwändig, Racemate aufzutrennen. Es müssen komplizierte Verfahren angewendet werden, welche die Auftrennung im Grunde ähnlich wie Enzyme bewerkstelligen. Auch hier geschieht die Trennung durch spezielle Unterstrukturen, an die nur eines der Enantiomere binden kann. Dieser Aufwand entfällt bei einer Synthese, die von einem stereospezifischen Enzym katalysiert wird.

2.1.3. Die Benzaldehydlyase und ihre Eigenschaften

In dieser Arbeit wird das Enzym Benzaldehydlyase (BAL, EC-Nummer 4.1.2.38) betrachtet. Die BAL besitzt vier Untereinheiten, die bei einer Reaktion als Bindungstaschen für die Substrate dienen können und verfügt sowohl über Ligase- als auch über Lyasefähigkeiten. Das bedeutet, dass die BAL sowohl die Spaltung als auch die Bildung von C-C-Bindungen katalysieren kann.

Diese Eigenschaft ist besonders wichtig, denn gerade Verbindungen zwischen Kohlenstoffatomen sind das Grundgerüst jedes Metabolismus. Aus einfachen Molekülen entstehen auf diese Weise größere Komplexe. Viele grundlegende Moleküle des Lebens (Fette, Zucker, etc.) sind Verkettungen von Kohlenstoffatomen. Aus der Spaltung dieser Ketten kann der Organismus wiederum Energie gewinnen. Die katalytische Eigenschaft der BAL wird durch den Kofaktor Thiamindiphosphat unterstützt, ein Molekül, das sich von Thiamin (Vitamin B_1) ableitet (siehe Abb. 2.4).

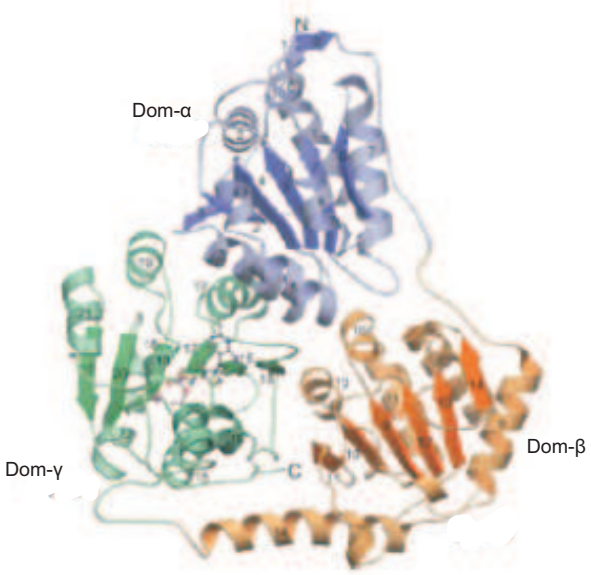

Abbildung 2.3.: Dreidimensionale Darstellung der Benzaldehydlyase [Küh07]

Die BAL katalysiert unter anderem die Bildung von Benzoinen aus aromatischen Aldehyden und die Verbindung aromatischer mit aliphatischen Aldehyden, was (R)-

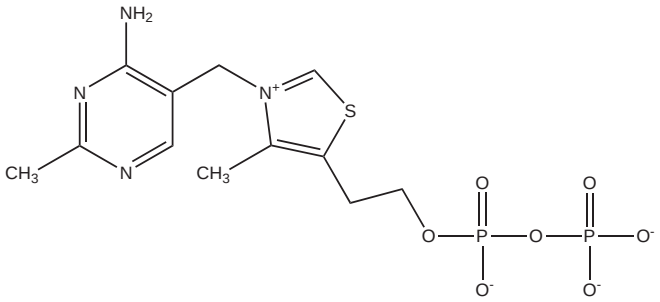

Abbildung 2.4.: Thiamindiphosphat

Hydroxyphenylpropanone ergibt. Einige der Endprodukte von der BAL katalysierter Reaktionen sind Basismoleküle für Medikamente, beispielsweise für Antiallergika [KMK+99] und Antidepressiva [FHG+00]. Diese Wirkstoffe werden zur Zeit noch mittels chemischer Synthese hergestellt. Die besonders hohe Stereoselektivität der BAL bei der Produktion von Hydroxyphenylpropanonen macht sie zu einer interessanten Alternative zur klassischen chemischen Synthese. Mehr als 99% des Umsatzergebnisses sind rechtsdrehende Moleküle.

Erstmals entdeckt wurde die BAL 1986 in einem Bakterium namens *Pseudomonas fluorescens* Biovar I, welches in der Lage war, Benzoin oder Anisoin als alleinige Kohlenstoffquelle zu verwerten [GV89]. Bei dem in der vorliegenden Arbeit untersuchten Enzym handelt es sich um ein sehr ähnliches Protein, das mit Hilfe einer gentechnisch veränderten Variante des Bakteriums *Escherichia coli* hergestellt wurde [Hil05].

Die Enzymaktivität ist von verschiedenen äußeren Faktoren wie Kofaktorkonzentration oder pH-Wert abhängig (siehe Abschnitt 2.1.1) und wird in der Einheit U angegeben. Dabei ist 1 U BAL als die Enzymmenge definiert, die nötig ist, um ausgehend von einer Benzaldehyd-Anfangskonzentration von 30 mM unter Standardbedingungen (siehe [Küh07]) in einer Minute 1 μmol Benzoin zu synthetisieren.

2.2. Modelle

2.2.1. Kinetische Modelle – Historischer Anriss

Der zeitliche Ablauf chemischer Reaktionen, beziehungsweise die zeitliche Änderung der Konzentrationen der Reaktionspartner und -produkte, wird als *Kinetik* bezeichnet. Eine Reaktionskinetik wird hauptsächlich durch die *Reaktionsgeschwindigkeit* und die Anzahl und Anfangskonzentrationen der Reaktionspartner bestimmt. Bei katalysierten Reaktionen hat zudem die Katalysatorkonzentration einen wichtigen Einfluss auf die Reaktionskinetik [Mic99].

Betrachtet man Ein-Substrat-Enzymkinetiken, so erkennt man eine Sättigungskurve, die zu Beginn fast linear ist, dann in eine gekrümmte Übergangsphase eintritt und

sich schließlich einem Maximalwert annähert (vergl. Abb. 2.5). Die auf makroskopischer Ebene stattfindende Umsetzung von Substrat (S) und Enzym (E) zu Produkt (P) (vergleiche Abschnitt 2.1.1) kann für dieses System folgendermaßen dargestellt werden:

$$S + E \underset{K_{-1}}{\overset{K_1}{\rightleftharpoons}} [SE] \overset{K_2}{\rightarrow} E + P \tag{2.1}$$

wobei mit [SE] der *Substrat-Enzym-Komplex* und mit K_1, K_2 und K_{-1} die Reaktionsgeschwindigkeiten der Hin- und Rückreaktionen bezeichnet werden. Zusätzlich wird die Annahme getroffen, dass beim zweiten Schritt keine Rückreaktion des Produkts mit dem Enzym stattfindet.

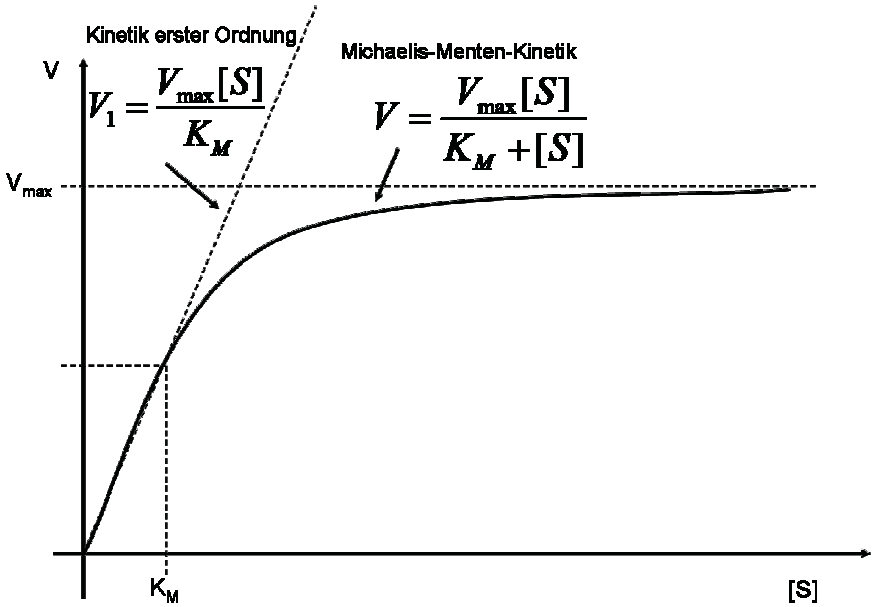

Abbildung 2.5.: Reaktionsordnungen in Abhängigkeit von der Substratkonzentration

Auf dieser Grundlage konnte 1925 von Briggs und Haldane die Reaktionsgeschwindigkeit einer Ein-Substrat-Reaktion beschrieben werden [BH25]. Benennt man die zu Anfang eingesetzte Konzentration des Enzyms mit c_{E0}, die von der Zeit t abhängigen Konzentrationen der anderen an der Reaktion beteiligten Stoffe analog, so gilt

$$\frac{d\,c_{[SE]}}{dt} \;=\; K_1 \cdot (c_{E0} - c_{[SE]}) \cdot c_S - K_{-1} \cdot c_{[SE]} - K_2 \cdot c_{[SE]}$$

mit

$$c_{E0} = c_E + c_{[SE]} \;\Leftrightarrow\; c_E = c_{E0} - c_{[SE]}.$$

Nimmt man an, dass die Reaktion im Gleichgewicht ist, so gilt:

$$\frac{d\,c_{[SE]}}{dt} \;=\; 0.$$

Dann ergibt sich für die Konzentration von [SE]:

$$c_{[SE]} = \frac{K_1 \cdot c_{E0} \cdot c_S}{K_{-1} + K_2 + K_1 \cdot c_S}.$$

Unter der Annahme, dass in (2.1) $[SE] \stackrel{K_2}{\rightarrow} E + P$ der geschwindigkeitsbestimmende Schritt der Katalyse ist, also für die Reaktionsgesamtgeschwindigkeit v

$$v \;=\; K_2 \;\cdot\; c_{[SE]}$$

gilt, erhält man:

$$v \;=\; \frac{K_1 \cdot K_2 \cdot c_{E0} \cdot c_S}{K_{-1} + K_2 + K_1 \cdot c_S} = \frac{K_2 \cdot c_{E0} \cdot c_S}{\frac{K_{-1}+K_2}{K_1} + c_S}.$$

Definiert man nun

$$\begin{aligned} V_{max} &:= K_2 \cdot c_{E0} \\ K_M &:= \frac{K_{-1} + K_2}{K_1} \end{aligned}$$

so erhält man für die Reaktionsgeschwindigkeit die Gleichung der *Michaelis-Menten-Kinetik*

$$v \;=\; \frac{V_{max} \cdot c_S}{K_M + c_S}. \tag{2.2}$$

Dabei beschreibt V_{max}, wie in Abbildung 2.5 zu sehen ist, die maximale Reaktionsgeschwindigkeit, die bei Substratsättigung des Enzyms erreicht wird. K_M, also die Substratkonzentration, bei der die Umsatzgeschwindigkeit $v = \frac{V_{max}}{2}$ beträgt, wird auch als *Michaelis-Menten-Konstante* oder *Halbsättigungskonstante* bezeichnet.

Dieses Modell ist nur für Ein-Substrat-Kinetiken gültig, kann aber durch Modifikationen anderen Situationen leicht angepasst werden. Treten bei der Reaktion mehrere Substrate oder auch verschiedene Arten von Inhibitoren auf, kann ein ergänzter Ansatz bei (2.1) verwendet werden, der dann zu analogen Formulierungen für die gesuchten Reaktionsgeschwindigkeiten führt (siehe [Bis02], [CB95] und [Seg95]).

2.2.2. Graphische Bestimmung kinetischer Parameter

Wesentliche Parameter einer Kinetik sind die Anzahl der Reaktionspartner zu Beginn der Reaktion, die Konzentrationen, in denen sie zu dem Zeitpunkt vorliegen und die Reaktionsgeschwindigkeit. Diese hängt im Michaelis-Menten-Modell wiederum von der maximalen Geschwindigkeit V_{max} und der Halbsättigungskonstante K_M ab. Bei einer katalysierten Reaktion ist zudem noch die Konzentration des Katalysators

relevant.

Während Anzahl und Anfangskonzentrationen der Reaktionspartner im Allgemeinen bekannt sind, ist die experimentelle Bestimmung von V_{max} und K_M der Michaelis-Menten-Kinetik etwas aufwändiger. Eine gängige Methode ist es, diese Parameter durch Messungen von *Anfangskinetiken* zu bestimmen. Dabei werden gleich zu Beginn der Reaktion in möglichst kurzen Abständen Proben genommen. Diese Anfangsmessungen werden dann durch Regressionsgeraden angenähert und so kann über die Kinetik erster Ordnung (vergl. Abb. 2.5) sowohl V_{max} als auch K_M bestimmt werden. Zudem gibt es etliche Möglichkeiten, eine Kinetik als lineare Funktion aufzutragen und aus den Schnittpunkten des Graphen mit den Koordinatenachsen Näherungswerte für die gesuchten Parameter abzulesen (z. Bsp. Eadie-Hofstee-Diagramm, Lineweaver-Burk-Diagramm, etc.) [Bis02].

Die kinetischen Parameter der BAL-katalysierten Umsetzung von Substraten wurden bisher, soweit dies möglich war, über Anfangskinetiken bestimmt und in einigen Fällen mit Ergebnissen aus der Anpassung der Datenpunkte an mathematische Modelle mit Hilfe des Programms Scientist® verglichen [Küh07].

2.2.3. Mathematische Modelle – Wahrheitsanspruch und Nutzen

> All models are wrong, but some are useful.
> George Box, 1976

In dieser Arbeit werden mathematische Differentialgleichungsmodelle verwendet, um die enzymatische Umsetzung zu beschreiben, die von der BAL katalysiert wird. Bevor in den nächsten Kapiteln konkret auf die mathematischen Modelle zur Beschreibung der Reaktionskinetik eingegangen wird, soll der Begriff des „Modells“ sowie einige Vor- und Nachteile der Benutzung von mathematischen Modellen diskutiert werden.

Das Wort „Model“ stammt von dem lateinischen Wort *modulus* ab, das „Muster, Form“ bedeutet. Diese Bedeutung wurde im Laufe der Zeit insofern ausgeweitet, dass mit „Modell“ heute allgemein ein vereinfachendes Abbild der Wirklichkeit bezeichnet wird. Vereinfachungen können zum Beispiel darin bestehen, dass große Bauwerke in kleinem Maßstab nachgebildet werden, sei es zur Veranschaulichung oder für Untersuchungen, die an dem großen Objekt nicht oder nur sehr schwer durchführbar wären (zum Beispiel Experimente im Windkanal). Es kann sich bei einem Modell auch um eine größengetreue Nachahmung des Originals unter Fortlassen von Details handeln, um das für einen bestimmten Zweck Wesentliche hervorzuheben oder Kosten und Aufwand einer detailgetreueren Kopie zu sparen. Ein

Modell zeichnet sich also durch eine bewusste Vernachlässigung von Merkmalen des Originals aus, wobei oft die Ungenauigkeit in einer unnachahmlichen Komplexität des Originals begründet ist.

Modelle haben entweder einen ästhetischen Wert oder werden für einen bestimmten Zweck aufgestellt und müssen dann im Hinblick auf diesen interpretiert und bewertet werden. Dieser Pragmatismus betrifft vor allem Modelle in den Geistes- und Naturwissenschaften. Hier dienen Modelle meist dazu, komplexe Zusammenhänge zu veranschaulichen, zu vereinfachen oder sie mit dem Computer darstellen und untersuchen zu können. Beispiele dafür sind Wirtschaftsmodelle, Kommunikationsmodelle und Räuber-Beute-Modelle in der Ökologie.

Modelle in den Wissenschaften haben vieles gemeinsam:

- Es müssen Annahmen getroffen werden, um eine Ausgangssituation und die Struktur des Modells konkret zu benennen. Diese können Erfahrung, Experimente oder Aussagen eines anderen Modells zur Grundlage haben.
- Modelle beschreiben die Beziehungen oder Wechselwirkungen zwischen verschiedenen Einheiten/Zuständen des Modells.
- Das Modell kann überprüft, beziehungsweise seine (Vor-)Aussagekreft mit Hilfe neuer Experimente bewertet werden

Kann ein Modell nicht überprüft werden, ist es für die Wissenschaft nutzlos. In den meisten Fällen zieht eine Überprüfung die Verwerfung oder eine Verbesserung des Modells nach sich, bis es die gewünschte Voraussagegenauigkeit erreicht. Diese kann so lange erhöht werden, wie sich eine Kosten-Nutzen-Abschätzung dafür aussprechen kann. Wie in Abb. 2.6 gezeigt, gibt es ein Optimum dieser Abschätzung. Die Fundamentalität α ist dabei ein theoretisches Konstrukt für die Genauigkeit des Modells im Verhältnis zur Realität.

Mathematische Modelle versuchen, die Beziehungen zwischen verschiedenen Modelleinheiten mit mathematischen Ausdrücken und Gleichungen darzustellen. Das hat den Nachteil, dass das Modell in einer analytisch oder numerisch untersuchbaren Form darstellbar sein muss, also Zielwerte eines Systems sich als mathematische Funktion des Anfangszustands ausdrücken lassen müssen. Dieser Grad der Abstrahierung ist beispielsweise für Kommunikationsmodelle nicht möglich. Zudem muss eine Entscheidung darüber getroffen werden, ob es genügt, wenn das Modell eine gewisse Voraussagekraft erreicht, oder ob es zusätzlich dazu geeignet sein soll, die Vorgänge innerhalb des modellierten Systems näher zu untersuchen. Im letzteren Fall ist die Formulierung der Modellgleichungen an im System auftretende physikalische, biologische, wirtschaftliche oder andere Gesetzmäßigkeiten gebunden.

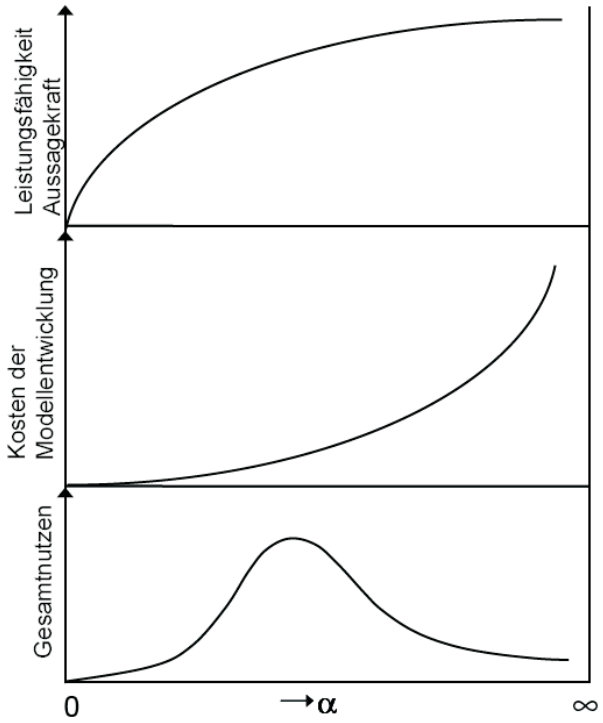

Abbildung 2.6.: Kosten-Nutzen-Verhältnis in der Modellierung mit zunehmender Fundamentalität α [Wan02]

Ein Vorteil mathematischer Modelle besteht darin, dass sie mit allen Mitteln der Mathematik analysiert und ausgewertet werden können. Dadurch wird es möglich, verschiedene Systemzustände genau zu untersuchen, beziehungsweise Auswirkungen einer Veränderung der Anfangsparameter auf das ganze System nachvollziehen und berechnen zu können. Ein weiterer Vorteil ist es, dass mit mathematischen Kriterien bewertet werden kann, wie gut verschiedene Modellierungsfunktionen die Eigenschaften des repräsentierten Systems darstellen und damit ein direkter Modellvergleich angestellt werden kann. Einige mathematische Grundlagen und Methoden zu diesen Themen, die vor allem aus den Teilbereichen der Statistik und der Stochastik stammen, werden im nächsten Kapitel eingeführt.

3. Mathematische Grundlagen und Theorie

3.1. Einleitung

In diesem Kapitel werden die benötigten mathematischen Grundlagen zum Verständnis der in dieser Arbeit verwendeten Methoden ausgearbeitet. Da die mathematischen Modelle, die verwendet werden, um den Reaktionsverlauf der enzymatischen Katalyse durch die Benzaldehydlyase zu beschreiben, Differentialgleichungsmodelle (DGL-Modelle) sind, wird eine detaillierte Darstellung eines solchen Modells zu Beginn dieses Kapitels gegeben. Anschließend werden verschiedene Lösungswege für DGL-Systeme dargelegt.

Zur Schätzung der unbekannten Parameter der DGL-Systeme (inverses Problem) wird in dieser Arbeit ein statistischer Ansatz verwendet. Dafür ist die Aufstellung eines Regressionsmodells auf Basis des DGL-Systems nötig. Im Rahmen der Parameterschätzung werden zudem die Implementierung der statistischen Methoden sowie die Verwendung numerischer Optimierungsroutinen und sinnvoller Nebenbedingungen diskutiert.

Zwei verschiedene Methoden, um die Genauigkeit der Parameterschätzungen und ihre Sensitivität bezüglich der zu ihrer Schätzung verwendeten Messwerte analysieren zu können, werden in den Abschnitten 3.6.5 und 3.7 vorgestellt. Eine der Methoden bedient sich dabei der Linearisierung des nichtlinearen Regressionsmodells, also der Approximation der Modellfunktion durch ein lineares Modell. Dieses Konzept mit seinen Eigenschaften wird in Abschnitt 3.7.1 näher betrachtet. Die Etablierung einer solchen „Asymptotisierung“ schafft die Möglichkeit, nichtlineare Modelle theoretisch zu behandeln. Ist die Konsistenz der Parameterschätzung gegeben, so kann das Modell für große Stichprobengrößen nämlich nahezu als lineares Modell betrachtet werden, dessen Eigenschaften approximativ mit wachsender Datenmenge auch wirklich angenommen werden [BB89]. Anschließend wird der Zusammenhang zwischen der Kovarianzmatrix einer Parameterschätzung und der Fisher-Informationsmatrix skizziert, die für die optimale Versuchsplanung benötigt wird.

Unterschiedliche Annahmen zu dem Reaktionsverlauf der enzymatischen Reaktion führen zur Formulierung mehrerer alternativer Differentialgleichungsmodelle. Aus

Vergleichskriterien, wie zum Beispiel der Anzahl der Modellparameter oder der Summe der Abweichungen der simulierten Werte von den Messwerten, kann ein Kriterium für die generelle „Überlegenheit“ eines Modells über ein anderes definiert werden. Das in dieser Arbeit verwendete Akaike-Informations-Kriterium (AIC) zur Modelldiskriminierung bezieht beide Vergleichskriterien ein und wird im Rahmen von Abschnitt 3.9.1 hergeleitet.

Schließlich werden einige Kriterien zur optimalen Versuchsplanung erläutert, die in dieser Arbeit verwendet wurden, um auf Grundlage der Parameterschätzung und der Modelldiskriminierung einen neuen Versuch zu planen. Diese Kriterien verwenden das Prinzip der „Informationsmaximierung“, auf welches in Abschnitt 3.10 genauer eingegangen wird.

3.2. Notation

3.2.1. Das Differentialgleichungssystem

Bei dem in dieser Arbeit betrachteten Modell handelt es sich um ein autonomes n-dimensionales Differentialgleichungssystem (DGL-System) 1. Ordnung [SW95] in der allgemeinen nichtlinearen Form

$$\begin{aligned}
\frac{d}{dt}g_1(t) &= h_1(g_1(t),\ldots,g_n(t);\theta;a_0)\\
\frac{d}{dt}g_2(t) &= h_2(g_1(t),\ldots,g_n(t);\theta;a_0)\\
&\vdots\\
\frac{d}{dt}g_n(t) &= h_n(g_1(t),\ldots,g_n(t);\theta;a_0)
\end{aligned}$$

wobei $g_1(t),\ldots,g_n(t)$ die Lösungen des DGL-Systems und $h_1,\ldots,h_n$ hinreichend oft stetig differenzierbare reellwertige Funktionen sind. Vektoriell lässt sich das System schreiben als

$$\frac{d}{dt}G(t) = H(G(t);\theta;a_0) \tag{3.1}$$

Überdies gelten die Anfangsbedingungen

$$G(t_0) = a_0 \quad \text{mit} \quad t_0 = 0.$$

Die Dimension n des DGL-Systems gibt die Anzahl der an der enzymatischen Reaktion beteiligten Partner an, θ ist der zu dem System gehörige Parametervektor mit p Elementen.

Es folgen einige Notationen zur numerischen Lösung eines DGL-Systems. Die Methode selber und ihre Anwendung in dieser Arbeit werden in Abschnitt 3.3.3 ausgearbeitet. Die numerische Lösung des DGL-Systems 3.1 stellt eine diskrete Approximation auf einem Punktgitter an die exakte Lösung $G(t)$ dar.

Definition 3.1 *Auf dem äquidistanten* Punktgitter $P = \{t_0, t_1, \ldots, t_m\}$ *mit* $t_0 < t_1 < \cdots < t_m$ *(äquidistant bedeutet, dass für die Schrittweite* $t_{j+1} - t_j = h_{SW}$ *für alle* $j = 0, \ldots, m-1$ *gilt) wird nun die* Gitterfunktion

$$u_G(t) : P \to \mathbb{R}^n \quad mit \quad t \in P$$

definiert.

Nun kann der Differentialgleichungslöser (DGL-Löser) als Verfahrensvorschrift verstanden werden, die für alle $t \in P$ eine Gitterfunktion $u_G(t) \in \mathbb{R}^n$ berechnet. Sei nun $T = (t_0, t_1, \ldots, t_{s-1})$ mit $t_0 < t_1 < \cdots < t_{s-1} \in P$ der Vektor der Zeitpunkte, zu denen der DGL-Löser das DGL-System betrachten soll, so erhält man nach Anwendung der Verfahrensvorschrift des Lösers eine $s \times n$ - dimensionale Matrix, die mit $L(G; \theta; a_0; T)$ bezeichnet werden soll.
Also gilt:

$$L(G; \theta; a_0; T) = \begin{bmatrix} u_G(t_0)^T \\ u_G(t_1)^T \\ \vdots \\ u_G(t_{s-1})^T \end{bmatrix}$$

Zusätzlich zu dem Funktionenvektor G, den Parametern θ, den Anfangsbedingungen $G(t_0) = u_G(t_0) = a_0$ und dem Zeitvektor T hängt $L(G; \theta; a_0; T)$ vom verwendeten DGL-Löser ab.

3.2.2. Weitere Notationen

Die $s_D \times n$ - dimensionale Datenmatrix M sei diejenige, in der die Messungen der Konzentrationen der Reaktionspartner spaltenweise verzeichnet sind, wobei $T_D = (t_{D0}, t_{D1}, \ldots)$ den Vektor der Messzeitpunkte des Experiments darstelle und s_D die Anzahl seiner Elemente sei. Die Elemente der Matrix M werden dann entsprechend mit m_{ij} angesprochen; dieselbe Regelung gilt auch für alle anderen Matrizen und Vektoren analog.

Eine besondere Rolle spielt die $s_D \times n$ - dimensionale Matrix $L_D := L(G; \theta; M_{1\bullet}; T_D)$ mit $M_{i\bullet}$ als der i-ten Zeile der Matrix M ($M_{\bullet j}$ bezeichne die j-te Spalte der Matrix M), da diese die Lösungen des Systems zu den Messzeiten des Experiments und

mit seinen Anfangskonzentrationen zum Zeitpunkt t_{D0} darstellt. Diese simulierten Werte werden mit M verglichen.

Im Folgenden werden vor allem bei Beweisen, aber auch an anderen Stellen, an denen die Komponente a_0 und die zeitliche Abhängigkeit der im Vektor $G(t)$ aufgeführten Funktionen nicht wesentlich sind, diese nicht explizit aufgeführt, also

$$H(G;\theta) := H(G(t);\theta;a_0)$$

Des weiteren werden mit $\theta \in \Theta$ ein unbekannter p-dimensionaler Parameter und mit $\hat{\theta}$ eine Schätzung desselben bezeichnet, wobei letztere je nach Art der Schätzung noch weiter spezifiziert werden kann.

3.3. Lösen von Differentialgleichungssystemen

3.3.1. Analytisches Lösen von Differentialgleichungssystemen

Es gibt verschiedene Möglichkeiten, Systeme mit gewöhnlichen Differentialgleichungen und Anfangswertproblemen analytisch zu lösen. Allerdings können explizite Lösungen nur in wenigen Spezialfällen angegeben werden, da die Integrale bei der Picard-Iteration oft nicht geschlossen auswertbar sind. Es wurde versucht, die in der Arbeit auftretenden Differentialgleichungssysteme analytisch per Hand sowie durch Computeralgebraprogramme (Maple™, Mathematica®) zu lösen. Da auf diesem Weg keine explizite Lösungsfunktionen gefunden werden konnten, war es nötig, einen alternativen Lösungsweg zu verwenden.

3.3.2. Verfahren zur numerischen Differentiation

Um die Verfahren zur numerischen Lösung von DGL-Systemen besser erläutern zu können, sollen an dieser Stelle zwei Verfahren zur numerischen Differentiation eingeführt werden. Sie beruhen auf dem Satz von Taylor, der im Anhang in Abschnitt C aufgeführt ist.

Definition 3.2 *Der* Vorwärtsdifferenzenquotient *für eine mindestens einmal stetig differenzierbare Funktion f(x) ist definiert durch*

$$f'(x) \approx \frac{f(x+h)-f(x)}{h} \tag{3.2}$$

Der zentrale Differenzenquotient *für eine mindestens zweimal stetig differenzierbare*

Funktion f(x) ist definiert durch

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h} \tag{3.3}$$

Satz 3.1 *Der durch die Approximation entstehende Abbruchfehler beim Vorwärtsdifferenzenquotienten ist größer als beim zentralen Differenzenquotienten. Genauer gesagt, beträgt der Abbruchfehler E(h) beim Vorwärtsdifferenzenquotienten $O(h)$ und beim zentralen Differenzenquotienten $O(h^2)$. Dabei bedeutet $(E(h) = O(h^n), h \to 0)$, dass der Fehler E(h) mit derselben Geschwindigkeit wie h^n für $h \to 0$ gegen Null geht.*

Beweis:
Der Beweis findet sich im Anhang (Teil C).

□

3.3.3. Numerisches Lösen von Differentialgleichungssystemen

Zur numerischen Lösung von DGL-Systemen wurde in dieser Arbeit ein Einschrittverfahren verwendet. Grundlage der Einschrittverfahren ist das *Euler-Verfahren*, das sich aus den Methoden zur numerischen Differentiation herleiten lässt.

Herleitung des Euler-Verfahrens:

Die Auswertung einer Differentialgleichung, hier mit

$$\frac{d\,y(t)}{dt} = f(y(t))$$

gegeben, an der Stelle $t = t_j$ lautet:

$$\frac{d\,y(t_j)}{dt} = f(y(t_j)). \tag{3.4}$$

Wird nun die linke Seite der Gleichung 3.4 durch den Vorwärtsdifferenzenquotienten (3.2) angenähert, so erhält man

$$\frac{d\,y(t_j)}{dt} = \frac{y(t_{j+1}) - y(t_j)}{h_j} + E_j$$

mit $t_{j+1} = t_j + h_j$ und $E_j = E(h_j)$ wie in Satz 3.1 definiert. Setzt man dies in (3.4) ein, so ergibt sich

$$\begin{aligned} \frac{y(t_{j+1}) - y(t_j)}{h_j} + E_j &= f(y(t_j)) \\ \Leftrightarrow \qquad y(t_{j+1}) &= y(t_j) + h_j \cdot f(y(t_j)) - h_j \cdot E_j \end{aligned}$$

Nach Satz 3.1 gilt, dass E_j gegen 0 konvergiert, falls das für h_j gilt. So kann aus der Auswertung der Differentialgleichung an der Stelle t_j approximativ die Auswertung an der Stelle t_{j+1} berechnet werden.

Es genügt also bei der Verwendung von Einschrittverfahren zur Berechnung der Gitterfunktion u_G, einen Startwert zum Zeitpunkt t_0 vorzugeben, in dieser Arbeit beispielsweise die Konzentrationen der Substrate und Produkte des Experimentes zum Zeitpunkt $t_0 = 0$, damit die Gitterfunktion iterativ an den folgenden Zeitpunkten $t_1, \ldots, t_{s-1}$ berechnet werden kann. In Matlab® sind verschiedene Einschritt-Verfahren implementiert, die sich beispielsweise in der Größe des Abbruchfehlers $h_j \cdot E_j$ unterscheiden. Die in dieser Arbeit verwendete Routine basiert auf dem Einschrittverfahren von Dormand-Prince, einem expliziten Runge-Kutta-Verfahren vierter Ordnung. Weitere Informationen zu diesem DGL-Löser finden sich in [DP58], [SW95].

3.4. Allgemeine statistische Grundlagen

An dieser Stelle werden nach [FH95] einige Definitionen aus der Statistik aufgeführt, die zum weiteren Verständnis des Textes nötig sind.

Definition 3.3 *Sei $Y = (Y_1, \ldots, Y_N)^T$ ein Zufallsvektor. Dann heißt*

$$E_Y(Y) = (E(Y_1), \ldots, E(Y_N))^T$$

Erwartungswertvektor von Y. Ist der Bezug auf die Zufallsvariable ersichtlich, so wird das Y im Index fortgelassen.

Definition 3.4 *Sei Y wie in Definition 3.3, $\sigma_{ij} = Cov(Y_i, Y_j), i \neq j$ die Kovarianz zwischen Y_i und Y_j, $\sigma_{ii} = Cov(Y_i, Y_i) = Var(Y_i)$ die Varianz von Y_i. Dann heißt*

$$Cov(Y) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1N} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2N} \\ \vdots & \vdots & & \vdots \\ \sigma_{N1} & \sigma_{N2} & \ldots & \sigma_{NN} \end{bmatrix}$$

Kovarianzmatrix *von Y.*

Bemerkung:
Es gilt

$$\mathrm{Cov}(Y) = E(YY^T) - E(Y)E(Y)^T.$$

Die Kovarianzmatrix eines Zufallsvektors ist stets symmetrisch und positiv semidefinit [FH95].

Definition 3.5 *Sei Y ein Zufallsvektor wie in Definition 3.3. Mit den* Korrelationen

$$\rho(Y_i, Y_j) = \rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

heißt

$$Corr(Y) = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1N} \\ \rho_{21} & 1 & \dots & \rho_{2N} \\ \vdots & \vdots & & \vdots \\ \rho_{N1} & \rho_{N2} & \dots & 1 \end{bmatrix}$$

Korrelationsmatrix *von Y [FH95].*

Definition 3.6 *Der Zufallsvektor Y heißt* p-dimensional normalverteilt *mit Erwartungswert $E(Y) = \mu$ und $Cov(Y) = \Sigma$ (Notation: $Y \sim \mathcal{N}_p(\mu, \Sigma)$), falls die Dichte von Y durch*

$$f_Y(y) = \frac{1}{(\sqrt{2\pi})^p \sqrt{det(\Sigma)}}\, e^{[-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)]}; \qquad y \in \mathbb{R}^p$$

gegeben ist, wobei Σ positiv definit ist.

Konfidenz-Ellipsoide

Für einen Zufallsvektor $Y \sim \mathcal{N}_p(\mu, \Sigma)$ und eine feste Konstante $c \in \mathbb{R}_{>0}$ gilt:

$$(Y - \mu)^T \cdot \Sigma^{-1} \cdot (Y - \mu) \;\; = \;\; c^2.$$

Das bedeutet, dass Punkte gleicher gemeinsamer Dichte auf dem Rand von Ellipsoiden liegen, die den Erwartungswert zum Mittelpunkt haben. Zerlegt man Σ mit Hilfe der Spektralzerlegung zu $\Sigma = \Gamma\Lambda\Gamma^T$ mit $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$ als der Diagonalmatrix der Eigenwerte von Σ und Γ der orthogonalen Matrix, die die korrespondierenden Eigenvektoren von Σ enthält, so weisen die Hauptachsen der Konfidenz-Ellipse in die Richtungen der Eigenvektoren und besitzen den Wurzeln der zugehörigen Eigenwerte proportionale Längen. Genauer gilt:

$$\sum_{i=1}^{p} \frac{Z_i^2}{\lambda_i} = c^2 \tag{3.5}$$

mit $Z = \Gamma^T(Y - \mu)$ [MKB79].

3.5. Nichtlineare Regressionsmodelle

Betrachtet man die Messdaten als durch Messungenauigkeiten verrauschtes Ergebnis einer Funktion, die von verschiedenen Variablen (beobachtbaren Eingangsvariablen und unbekannten Parametern $\theta_1, \ldots, \theta_p$) abhängt, so spricht man von einem Regressionsmodell. Dieses wird als nichtlineares Regressionsmodell bezeichnet, wenn die so genannte *Regressionsfunktion* nichtlinear in mindestens einem der Parameter ist.

Das in dieser Arbeit betrachtete nichtlineare Regressionsmodell lautet:

$$M \quad = \quad H(G;\theta) + \varepsilon_G + \varepsilon = L_D + \varepsilon_G + \varepsilon \tag{3.6}$$

mit $H(G;\theta)$ aus (3.1) und $L_D = L(G;\theta;a_0;T_D)$ wie in Abschnitt 3.2.1 definiert. ε_G ist eine $s_D \times n$ - dimensionale Matrix, bei der jedes Element ein für die Messwerte vorher definiertes Grundrauschen ist und ε ist die ebenfalls $s_D \times n$ - dimensionale Matrix der Messwertfehler.

Die Aufstellung des Regressionsmodells (3.6) basiert auf folgenden Annahmen:

A1 Die Fehler der Messwerte als Elemente ε_{ij} von ε unterliegen einer Normalverteilung mit Mittelwert 0 und bekannter Varianz σ_{ij}. Die Varianz ist proportional zum Messwert m_{ij} und wurde von den Experimentatoren angegeben.

A2 Das Grundrauschen ε_{Gij} entspricht der Sensibilitätsgrenze der Messapparatur und ist bekannt. In dieser Arbeit wurde beispielsweise $\varepsilon_{Gij} = 0.1$ mM gesetzt (vergl. Abschnitt 5.2)

Beide Annahmen A1 und A2 sind kritisch zu betrachten, da erstens die Fehler experimenteller Messdaten nicht immer normalverteilt sein müssen und zweitens die Angaben der Varianzen und der Sensibilitätsgrenze nur auf der Erfahrung des Experimentators beruhen, also keinesfalls exakt sind.

3.6. Parameterschätzung

In Abschnitt 3.3.3 wurde betrachtet, wie bei Vorgabe fester Parameter das DGL-System gelöst werden kann (*Vorwärtsproblem*). In einem zweiten Schritt wurde das Regressionsmodell (3.6) mit Hilfe eines DGL-Systems aufgestellt, welches die Voraussagewerte, beziehungsweise simulierten Werte bereitstellt. Nun geht es darum, aus realen Messergebnissen einer enzymatischen Reaktion die unbekannten Parameter des nichtlinearen Regressionsmodells 3.6 zu schätzen (*inverses Problem*).

Die in dieser Arbeit verwendete Schätzmethode leitet sich aus dem statistischen Zugang zu diesem Problem her. Statistische Zugänge sind etwa die Methode der

kleinsten Quadrate (siehe Abschnitt 3.6.1) oder das Maximum-Likelihood-Prinzip (siehe Abschnitt 3.8). Ein anderer möglicher Zugang, der hier nicht verwendet wird, betrachtet Regularisierungsmethoden aus der Theorie inverser Probleme, deren Anliegen die Analyse und Lösung inkorrekt gestellter oder instabiler Probleme ist [KS04], [ABT04].

3.6.1. Methode der kleinsten Quadrate

Die im Folgenden vorgestellte Methode der kleinsten Quadrate setzt als Schätzung den Parametersatz an, der für einen möglichst geringen Abstand zwischen der Regressionsfunktion und den Messwerten sorgt.

Definition 3.7 *Geht man von dem in Abschnitt 3.5 angeführten nichtlinearen Regressionsmodell 3.6 aus, so heißt* $\hat{\theta}_{KQSS}$ gewichtete Kleinste-Quadrat-Summen-Schätzung *(gKQSS) falls*

$$\hat{\theta}_{KQSS} \quad = \quad arg\min_{\theta\in\Theta} \sum_{i=1}^{s_D} \sum_{j=1}^{n} w_{ij}(m_{ij} - l_{Dij})^2 \tag{3.7}$$

mit Gewichten $w_{ij} \in \mathbb{R}$ *gilt.* m_{ij} *sind dabei die Einträge der Messwertematrix* M *und* $l_{Dij} \in L(G;\theta;M_{1\bullet};T_D)$ *die damit zu vergleichenden Voraussagewerte, die sich als Lösung des Differentialgleichungssystems (siehe Abschnitte 3.2.2 und 3.5) ergeben.*

Mit den Residuen

$$\epsilon_{ij} := m_{ij} - l_{Dij}$$

ist die zu minimierende Zielfunktion *mit*

$$\sum_{i=1}^{s_D} \sum_{j=1}^{n} w_{ij}\epsilon_{ij}^2 \tag{3.8}$$

gegeben.

3.6.2. Wahl der Gewichte in der gKQSS

Die Gewichte w_{ij} vor dem Abstandstermen ϵ_{ij} werden dafür verwendet, Messwerten mit kleinerer Varianz mehr Einfluss auf die Schätzung zu gestatten als Messwerten mit größerer Messunsicherheit. Eine Möglichkeit für die Wahl der Gewichte ist es deshalb, direkt die inversen Varianzen der Messwerte für w_{ij} in Gleichung 3.7 einzusetzen.

Da allerdings das im Modell angenommene Grundrauschen in die Gewichte aufgenommen werden sollte, wurden diese folgendermaßen definiert:

$$w_{ij}^{-1} \quad = \quad a \cdot m_{ij} + \varepsilon_{Gij}$$

mit

$$a := \frac{Var(m_{max}) - \varepsilon_{Gij}}{m_{max}} \tag{3.9}$$

wobei m_{max} der größte Eintrag der Matrix M sei. Diese Art der Gewichtung stellt sicher, dass m_{max} nur mit dem Inversen seiner Varianz gewichtet wird, denn das Grundrauschen ist hier unerheblich. Ein Messwert nahe bei 0, der aufgrund der Annahme A1 zum Regressionsmodell aus Abschnitt 3.5 eine sehr kleine Varianz hat, wird dagegen hauptsächlich mit dem Grundrauschen ε_{Gij} gewichtet. Die Situation ist in Abb. 3.1 dargestellt und für die unbekannte Steigung a der gestrichelten Verbindungsgeraden ergibt sich gerade (3.9).

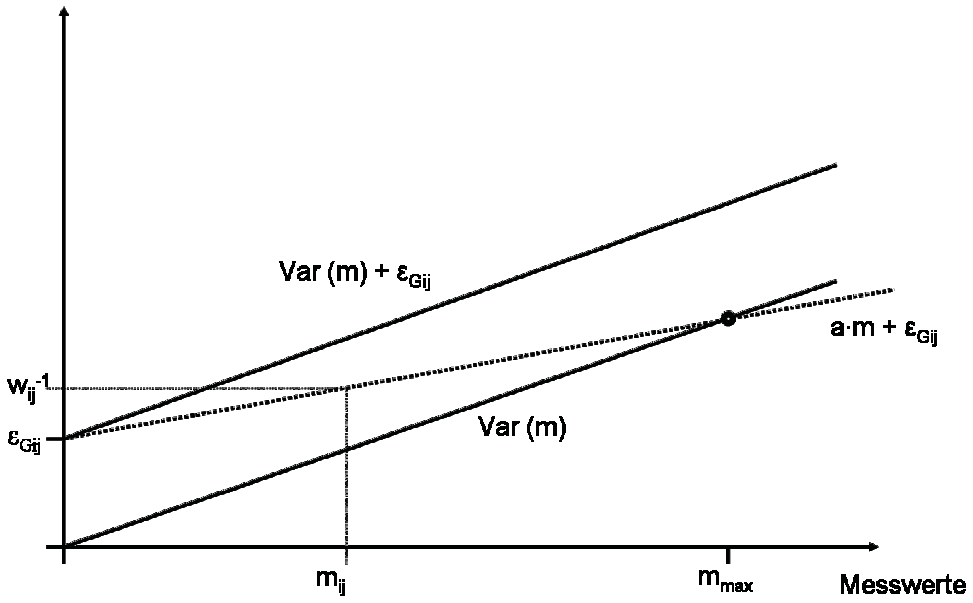

Abbildung 3.1.: Bestimmung der Gewichte für die gKQSS

3.6.3. Least-Trimmed-Squares-Schätzung

Ein bekanntes Problem der gKQSS ist, dass Ausreißer unter den Messwerten einen sehr großen Einfluss auf die Schätzung haben, beziehungsweise sie stark verfälschen können. Um zusätzlich eine Methode zu verwenden, die robust, also nicht so empfindlich gegenüber Ausreißern ist, wurden die Parametersätze zum Vergleich auch mit der *Least-Trimmed-Squares-Methode* (LTS-Methode) geschätzt, bei der eine vorher festgelegte Anzahl der größten Residuen nicht in die Minimierung miteinbezogen werden.

Definition 3.8 *Seien das Modell und der Datensatz wie in Definition 3.7 definiert. Die Regressionsschätzung $\hat{\theta}_{LTS}$ mit der LTS-Methode bezüglich k ist definiert durch:*

$$\hat{\theta}_{LTS} \quad = \quad arg\min_{\theta\in\Theta} \sum_{i=1}^{s_D\cdot n-k} \epsilon_{(i)}(M, L_D(G;\theta))$$

mit $\epsilon_i(M, L_D(G;\theta)) = |m_{i1} - l_{Di1}|$ für $i = 1, \ldots, s_D$, $\epsilon_i(M, L_D(G;\theta)) = |m_{i2} - l_{Di2}|$ für $i = s_D + 1, \ldots, 2 \cdot s_D$, etc. und den geordneten Residuen $\epsilon_{(1)}(M, L_D(G;\theta)) \leq \epsilon_{(2)}(M, L_D(G;\theta)) \leq \cdots \leq \epsilon_{(s_D\cdot n)}(M, L_D(G;\theta))$ [Mül05].

In dieser Methode nehmen also die k größten Residuen keinen Einfluss auf die Schätzung.

3.6.4. Numerisches Verfahren zur Optimierung

Als numerisches Verfahren zur Optimierung einer Zielfunktion, das beispielsweise für die Methode der kleinsten Fehlerquadrate (siehe Abschnitt 3.6.1) notwendig ist, wurde die Matlab®-Routine *lsqnonlin* verwendet, die auf dem Levenberg-Marquardt-Algorithmus basiert [SW03]. Sie ist eine Erweiterung des Gauss-Newton-Algorithmus speziell für die Optimierung der Zielfunktion der Methode der kleinsten Quadrate (3.8).

Definition 3.9 *Habe das Minimierungsproblem die Form*

$$arg\min_{\theta\in\Theta} \sum_{i=1}^{s_D\cdot n} \epsilon_i(M, L_D(G;\theta))^2 \tag{3.10}$$

mit $\epsilon_i(M, L_D(G;\theta))$ wie in Definition 3.8. Sei

$$\epsilon(M, L_D(G;\theta)) = (\epsilon_1(M, L_D(G;\theta)), \ldots, \epsilon_{s_D\cdot n}(M, L_D(G;\theta)))$$

der Residuenvektor *und $\theta^{(1)}$ ein zu Beginn des Algorithmus geeignet gewählter Startparameter. Der q-te Schritt $(q = 1, 2, 3, \ldots)$ des* Gauss-Newton-Algorithmus *lautet*

$$\theta^{(q+1)} = \theta^{(q)} + \delta^{(q)} \tag{3.11}$$

mit

$$\delta^{(q)} = -(K^{(q)T}K^{(q)})^{-1}K^{(q)T}\epsilon^{(q)}$$

wobei gilt:

$$K^{(q)} \quad = \quad \frac{\partial\,\epsilon^{(q)}}{\partial\theta} \quad \text{und} \quad \epsilon^{(q)} = \epsilon(M, L_D(G;\theta^{(q)}))$$

Das Prinzip des Levenberg-Marquardt-Algorithmus beruht nun auf möglichst großen Verbesserungsschritten im Optimierungsvorgang durch die Einführung eines Faktors η in die Schritte des Gauss-Newton-Algorithmus. Der q-te Schritt des verwendeten Levenberg-Marquardt-Algorithmus nutzt ebenfalls (3.11) allerdings mit

$$\delta^{(q)} = -(K^{(q)T}K^{(q)} + \eta^{(q)})^{-1}K^{(q)T}\epsilon^{(q)}. \tag{3.12}$$

Bei jedem Schritt wird der Faktors $\eta^{(q)}$ bestimmt, indem der Abstiegsschritt $\delta^{(q)}$ maximiert wird:

$$\eta^{(q)} = arg \max_{\eta \in \mathbb{R}^{s_D \cdot n}} \delta^{(q)}$$

Es findet also bei jedem Schritt eine Optimierung zwischen der Gauss-Newton-Methode ($\eta^{(q)} \longrightarrow 0$) und der Methode des steilsten Abstiegs ($\eta^{(q)} \longrightarrow \infty$) statt.

Optimierungsvorgang

Der Optimierungsvorgang zur Parameterbestimmung mit Hilfe der gKQSS oder der LTS-Schätzung ist in Abbildung 3.2 dargestellt. Nach der Wahl eines Startwertes $\theta^{(1)}$

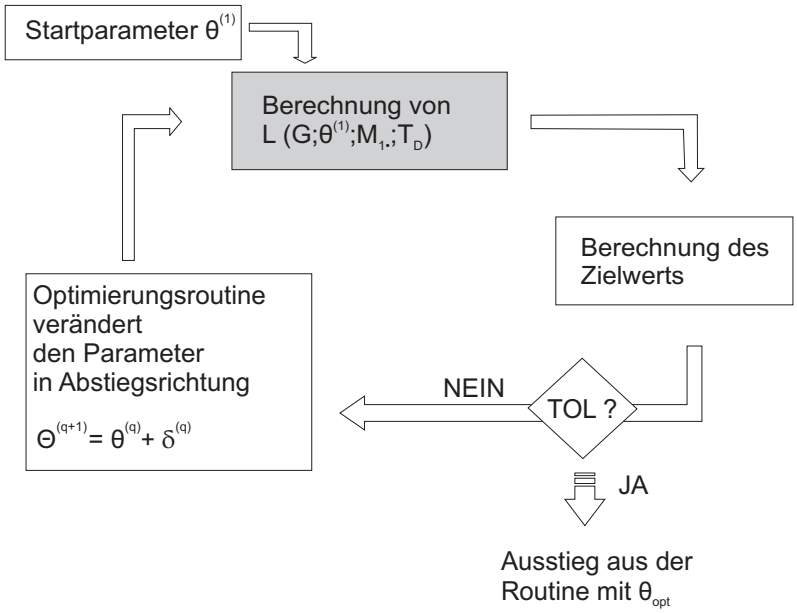

Abbildung 3.2.: Optimierungszyklus zur Parameterbestimmung

beginnt die Optimierung mit der Berechnung von $L(G; \theta^{(1)}; M_{1\bullet}; T_D)$, der Matrix der simulierten Messwerte bezüglich des Startparametersatzes. Im nächsten Schritt wird dann der Zielwert der zur minimierenden Funktion berechnet. Dann wird durch Überprüfung vorher festgelegter Toleranzkriterien (TOL) entschieden, ob die Optimierung fortzusetzen oder an dieser Stelle abzubrechen ist. Toleranzkriterien können erfüllt sein, wenn eine Unterschreitung der minimalen Veränderung des

Zielwerts vorliegt oder die Anzahl der bisher durchgeführten Iterationen einen Maximalwert übersteigt. Wird die Optimierung fortgeführt, so verändert das gewählte Verfahren den Parametersatz in Abstiegsrichtung. Mit dem dadurch erhaltenen neuen Parametersatz $\theta^{(2)}$ wird die Routine erneut aufgenommen.

In dem in Abbildung 3.2 grau hinterlegten Schritt muss das DGL-System bei jeder Iteration mit einem neuen Parametersatz gelöst werden. Dieser Vorgang bestimmt die Rechenzeit der Optimierung, die bei den in dieser Arbeit betrachteten Modellen bei etwa vier bis zehn Sekunden lag. Ein Beispiel für die Bildschirmausgabe einer in dieser Arbeit durchgeführten Optimierung zur Parameterschätzung mit der Methode der kleinsten Quadrate findet sich im Anhang Teil D.

Wahl der Startwerte und Nebenbedingungen für die Optimierung

Die Startwerte für die Optimierung wurden durch einen Zufallszahlengenerator gewählt, um eine möglichst breite Streuung der Startparametersätze im Parameterraum Θ zu gewährleisten. Jede Optimierung wurde mehrfach mit verschiedenen Startwerten gestartet (*Multi-Start-Strategie*) um die Reproduzierbarkeit des Ergebnisses zu überprüfen. Trotzdem gibt es (anders als bei analytischen Lösungen eines Minimierungsproblems) keine absolute Sicherheit bezüglich der Globalität des gefundenen Minimums.

Um nur Parametersätze innerhalb sinnvoller biologischer Grenzen schätzen zu lassen, wurde in dieser Arbeit die Optimierung des Zielwerts (3.8) unter Nebenbedingungen verwendet. Die in dieser Arbeit verwendeten Nebenbedingungen waren:

$$0 \leq \theta_i \leq 1000 \quad \text{für} \quad 1 \leq i \leq p. \tag{3.13}$$

Auf diese Art und Weise wurde nur im Raum $[0, \ldots, 1000]^p$ nach Parameterschätzungen gesucht.

3.6.5. Parameterschätzung mit parametrischem Bootstrap

Eine erste Methode, die es erlaubt, mehr über die Verteilung der Parameterschätzungen, beziehungsweise den Einfluss des Messrauschens auf die Schätzung zu erfahren, ist die Simulation weiterer künstlicher Datensätze unter der Annahme der Normalverteilung der Messwerte. Für diese werden erneut die Modellparameter geschätzt, so dass eine Verteilung von Parameterschätzungen generiert werden kann. Zur Simulation von Datensätzen wurden in dieser Arbeit zwei unterschiedliche Methoden verwendet, die nun näher erläutert werden sollen.

- **Methode MCData** (*parametrischer Bootstrap* [ET98])
  Gegeben sei ein Parametersatz θ_{MC}. Nach Anwendung des DGL-Lösers auf

das betrachtete DGL-System erhält man die $s_D \times n$ - dimensionale Matrix $L(G; \theta_{MC}; a_0; T_D)$. Diese wird durch das Addieren einer Matrix mit $\mathcal{N}(0,1)$-verteilten Elementen verrauscht und stellt danach einen neuen künstlichen Datensatz dar.

- **Methode SIMUL**
  Gegeben sei ein Parametersatz θ_{SIM}. Dieser Parametersatz wird nun durch Addition normalverteilter Fehlerwerte beliebig oft verrauscht. Mit jedem dieser verrauschten Parameter wird nun nach Methode MCData vorgegangen. Dieser Ansatz ist realistischer als Methode MCData, da er nicht nur ein Messrauschen für den Datensatz simuliert, sondern auch davon ausgeht, dass der Startparametersatz verrauscht ist.

Simulationsmethoden sind keineswegs ideale Instrumente, um statistische Schätzer sehr präzise zu bestimmen, sondern sind eher geeignet, um mit einfachen Mitteln einen ersten Eindruck zu Schätzwerten zu erhalten. Über die Fehlerrechnung, beziehungsweise das Gauß'sche Gesetz der Fehlerfortpflanzung, kann nachgewiesen werden, dass der auf Basis künstlicher Datensätze erhaltene Schätzer für große Replikationszahlen mit einer Geschwindigkeit von $\sqrt{N}$, mit N als der Anzahl der verwendeten Datenpunkte, gegen den Erwartungswert konvergiert. Das bedeutet beispielsweise, dass für doppelte Genauigkeit der Methode viermal mehr Datenpunkten als ursprünglich benötigt werden. Die genaue Parameterschätzung mit Simulationsmethoden ist deshalb sehr aufwändig und zeitraubend, verglichen beispielsweise mit der gKQSS.

Vorteilhaft ist hingegen, dass man neben der Bestimmung von Parameterschätzungen auch eine empirische Verteilung für den Schätzer erhält und seine Sensitivität im Bezug auf die Messwerte untersuchen kann. Simulationsmethoden haben den Vorteil gegenüber anderen Methoden zur Sensitivitätsanalyse nichtlinearer Modelle, wie beispielsweise der Linearisierung des Regressionsmodells, die in Abschnitt 3.7 angesprochen wird, dass sie direkt auf das zu betrachtende Modell angewendet werden können, ohne dass es modifiziert werden muss. Damit sind sie besonders gut geeignet, um die Verteilungen, Varianzen und Sensivitäten von Parameterschätzungen nichtlinearer und/oder sehr komplexer Modelle zu analysieren.

3.7. Kovarianzmatrix und Informationsmatrix

3.7.1. Linearisierung eines nichtlinearen Regressionsmodells

Eine Methode, die Varianzen der geschätzten Parameter zu bestimmen, ist die Berechnung der Kovarianzmatrix für die Schätzung des Parametersatzes. In der Diagonalen der Kovarianzmatrix sind nach Definition 3.4 die Varianzen der geschätzten

Parameterwerte zu finden. Die Linearisierung der nichtlinearen Regressionsfunktion (3.6) um $\hat{\theta}$ ist die in dieser Arbeit verwendete Methode, um für einen geschätzten Parametersatz $\hat{\theta}$ die Kovarianzmatrix zu bestimmen. Durch die Linearisierung der Regressionsfunktion können Ergebnisse aus der Theorie der linearen Modelle verwendet werden. Die Linearisierung und ihre Eigenschaften werden nach [Páz93], [SW03] dargestellt, wobei $\tilde{M}$ der Vektor mit den Beobachtungen sei, also

$$\tilde{M} = \begin{bmatrix} M_{\bullet 1} \\ M_{\bullet 2} \\ \vdots \\ M_{\bullet n} \end{bmatrix}$$

und $\tilde{M} \in \mathbb{R}^{s_D \cdot n}$.

Definition 3.10 *Ein* multivariates lineares Regressionsmodell *habe die allgemeine Form*

$$\tilde{M} = J \cdot \theta + \varepsilon$$

mit $J \in \mathbb{R}^{s_D \cdot n \times p}$ *als* Systemmatrix *und* $\varepsilon \sim \mathcal{N}(0, \Sigma)$ *mit* $\Sigma \in \mathbb{R}^{s_D \cdot n \times s_D \cdot n}$*, wobei* J *und* Σ *bekannt seien.*

Um vom allgemeinen Regressionsmodell zu seiner linearisierten Form zu gelangen, muss nun die Systemmatrix der Linearisierung bestimmt werden.

Satz 3.2 *Die Systemmatrix* J *der linearisierten Form des allgemeinen nichtlinearen Regressionsmodells 3.6 ist gegeben durch*

$$J = \left. \frac{\partial H(G;\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}} \tag{3.14}$$

mit einem gegebenen erwartungstreuen Schätzer $\hat{\theta}$*.*

Beweis:
Der Linearisierung liegt eine Taylor-Entwicklung der Funktion $H(G;\theta)$ um den Punkt $\hat{\theta}$ zugrunde, die nach Vernachlässigung höherer Terme für $\hat{\theta}$ nahe θ zu

$$H(G;\theta) \approx H(G;\hat{\theta}) + \left. \frac{\partial H(G;\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}} \cdot (\theta - \hat{\theta})$$

führt. Setzt man diese Näherung in das nichtlineare Regressionsmodell (3.6) ein, so erhält man

$$\tilde{\tilde{M}} = J\theta + \varepsilon_G + \varepsilon$$

mit

$$\tilde{\tilde{M}} = \tilde{M} - H(G;\hat{\theta}) + J\hat{\theta} \quad \text{und} \quad J = \left.\frac{\partial H(G;\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}}.$$

□

Satz 3.3 *(Eigenschaften des linearisierten Regressionsmodells)*
Sei $\tilde{M} = J\theta + \varepsilon$ mit $\varepsilon \sim \mathcal{N}(0,\Sigma)$ ein linearisiertes Regressionsmodell mit stochastisch unabhängigen Fehlern ε und einer nichtsingulären Matrix $J^T \cdot \Sigma^{-1} \cdot J \in \mathbb{R}^{p\times p}$. Sei $\hat{\theta}$ eine KQSS-Parameterschätzung. Dann gilt

1. *$\hat{\theta}$ ist ein erwartungstreuer Schätzer für θ_0:*

$$E(\hat{\theta}) \;=\; \theta_0$$

2. *Die* Output-Sensitivität *eines linearisierten Regressionsmodells bei $\hat{\theta}$ ist gegeben durch:*

$$Sens_M := \left.\frac{\partial H(G;\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} = J \tag{3.15}$$

3. *Die* Parameter-Sensitivität *von $\hat{\theta}$ berechnet sich aus*

$$Sens_{\hat{\theta}} = (J^T \cdot \Sigma^{-1} \cdot J)^{-1} \cdot J^T \cdot \Sigma^{-1} \tag{3.16}$$

4. *$\hat{\theta}$ ist der beste lineare erwartungstreue Schätzer (BLUE) für θ_0, das bedeutet, dass $\hat{\theta}$ in der Klasse aller linearen unverzerrten Schätzer die kleinste Kovarianzmatrix besitzt[1]. Zudem gilt $\hat{\theta} \sim \mathcal{N}(\theta_0, Cov(\hat{\theta}))$ und die Kovarianzmatrix hat die Form*

$$\begin{aligned} Cov(\hat{\theta}) &= (Sens_M^T \cdot \Sigma^{-1} \cdot Sens_M)^{-1} && (3.17)\\ &= Sens_{\hat{\theta}} \cdot \Sigma \cdot Sens_{\hat{\theta}}^T && (3.18) \end{aligned}$$

Beweis: [Páz93], [SW03]

□

Damit liefern (3.17) und (3.18) die gesuchten Berechnungsvorschriften für die Kovarianzmatrix eines Schätzers $\hat{\theta}$. Diese Vorschriften enthalten die Output-Sensitivität $Sens_M$ beziehungsweise Parameter-Sensitivität $Sens_{\hat{\theta}}$ und die Matrix Σ der Messfehlervarianzen. Da die numerische Berechnung der Sensitivitäten recht aufwändig ist, wird sie stellvertretend für beide Sensitivitäten im folgenden Abschnitt für die Output-Sensitivität dargestellt.

[1] A,B $\in \mathbb{R}^{p\times p}$, $A < B \Leftrightarrow B - A$ ist positiv definit

Nach Gleichung (3.15) ist es für die Berechnung der Output-Sensitivität nötig, die Regressionsfunktion (3.6) nach den Parametern abzuleiten. Die Implementierung dieser Vorschrift ist etwas aufwändig, da das Modell durch ein DGL-System gegeben ist und deshalb numerische Differentiation verwendet wird. Es gilt unter Verwendung des zentralen Differenzenquotienten (3.3)

$$(Sens_M)_{\bullet j} = \frac{\partial H(G;\theta)}{\partial \theta_j} \approx \frac{H(G;\theta + h_j) - H(G;\theta - h_j)}{2h \cdot e_j} \tag{3.19}$$

wobei $h = (h_1, \ldots, h_p)$ wie in Abschnitt 3.3.3 definiert ist und e_j der j-te Einheitsvektor sei. Das Berechnungsschema für die Implementierung der Berechnung der Output-Sensitivität wird in Abb. 3.3 dargestellt. Der geschwindigkeitsbestimmende Schritt der Berechnung ist grau hinterlegt.

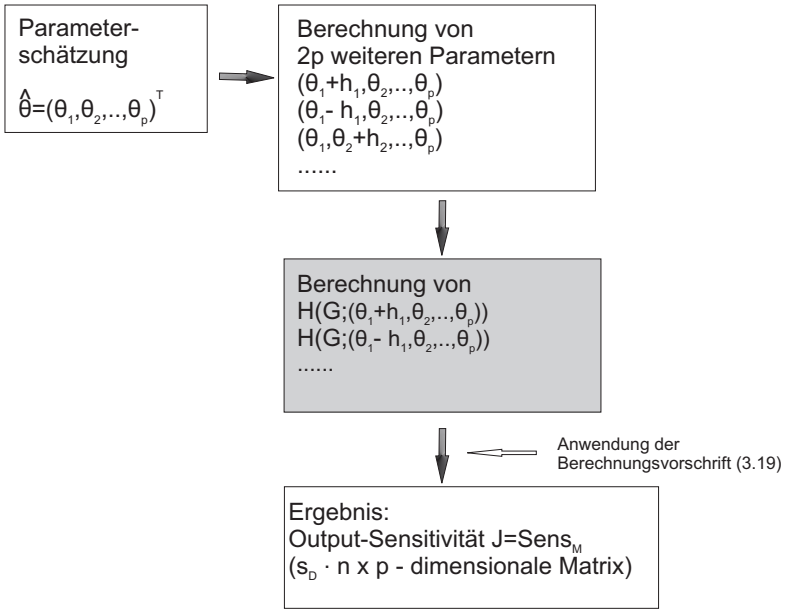

Abbildung 3.3.: Berechnung der Sensitivitätsmatrizen

3.8. Maximum-Likelihood-Prinzip und Cramér-Rao-Schranke

Um die in Abschnitt 3.7 definierten Sensitivitätsmatrizen mit der für die optimale Versuchsplanung benötigten Fisher-Informationsmatrix in Beziehung zu setzen, werden in diesem Kapitel einige Ergebnisse aus der Theorie des Maximum-Likelihood-Prinzips nach [FH95] dargestellt. Das Maximum-Likelihood-Prinzip besteht darin, bei Vorliegen der Beobachtung M einen Parameterschätzwert $\hat{\theta}_{ML} \in \Theta$ so zu wählen,

dass M gerade für $\hat{\theta}_{ML}$ eine maximale Wahrscheinlichkeitsdichte zukommt (siehe Definition 3.12).

Definition 3.11 *Sei $f(M_{\bullet 1}, \ldots, M_{\bullet n}; \theta)$ die gemeinsame Dichte von $M_{\bullet 1}, \ldots, M_{\bullet n}$, wobei θ ein unbekannter zu schätzender Parametervektor sei. Die* Likelihoodfunktion *der Stichprobe ist*

$$L(\theta; M_{\bullet 1}, \ldots, M_{\bullet n}) \quad = \quad f(M_{\bullet 1}, \ldots, M_{\bullet n}; \theta).$$

Die Log-Likelihoodfunktion *werde bezeichnet durch*

$$l(\theta; M_{\bullet 1}, \ldots, M_{\bullet n}) \quad = \quad \ln f(M_{\bullet 1}, \ldots, M_{\bullet n}; \theta).$$

Für den Fall unabhängiger $M_{\bullet 1}, \ldots, M_{\bullet n}$ mit den Dichten $f_i(M_{\bullet i}; \theta)$ erhält man

$$L(\theta; M_{\bullet 1}, \ldots, M_{\bullet n}) = \prod_{i=1}^{n} f_i(M_{\bullet i}; \theta)$$

bzw.

$$l(\theta; M_{\bullet 1}, \ldots, M_{\bullet n}) = \sum_{i=1}^{n} \ln f_i(M_{\bullet i}; \theta)$$

Diese Größen sollen kurz als $L(\theta; M)$ beziehungsweise $l(\theta; M)$ bezeichnet werden. Die Likelihoodfunktion wird für eine feste Stichprobe $M_{\bullet 1}, \ldots, M_{\bullet n}$ als Funktion von $\theta \in \Theta$ aufgefasst [FH95].

Definition 3.12 *$\hat{\theta}_{ML}$ heißt* Maximum-Likelihood-Schätzer für θ, *wenn*

$$L(\hat{\theta}_{ML}; M) \quad \geq \quad L(\theta; M) \quad \textit{für alle} \quad \theta \in \Theta$$

beziehungsweise

$$l(\hat{\theta}_{ML}; M) \quad \geq \quad l(\theta; M) \quad \textit{für alle} \quad \theta \in \Theta$$

gilt.

Definition 3.13 *Der Vektor der ersten Ableitungen der Log-Likelihoodfunktion wird auch als* Score-Funktion *bezeichnet:*

$$S(\theta; M) \quad = \quad \frac{\partial}{\partial \theta} l(\theta; M) = \frac{1}{L(\theta; M)} \frac{\partial}{\partial \theta} L(\theta; M).$$

Für $M_{\bullet 1}, \ldots, M_{\bullet n}$ stochastisch unabhängig und gleich verteilt (i.i.d.) heißt

$$F(\theta) = E\left(-\frac{\partial^2 l(\theta; M)}{\partial\theta\partial\theta^T}\right) \tag{3.20}$$

die Fisher-Informationsmatrix

Es gilt also für die in Definition 3.12 definierte ML-Schätzung $\hat{\theta}_{ML}$ als Maximum der Likelihood-Funktion:

$$S(\hat{\theta}_{ML}; M) \;=\; \frac{\partial}{\partial\theta} l(\hat{\theta}; M) = 0 \tag{3.21}$$

Satz 3.4 *Sind Integration und Differentiation vertauschbar (diese Bedingung wird im folgenden als Regularitätsbedingung $\mathcal{R}$ bezeichnet, siehe Anhang C) und sind $M_{\bullet 1}, \ldots, M_{\bullet n}$ i.i.d. so gilt:*
(a) $E(S(\theta; M)) = 0$
(b) $Cov(S(\theta; M)) = E(S(\theta; M)S(\theta; M)^T) = F(\theta)$

Beweis zu (a):
Es gilt:

$$\int L(\theta; M)\mathrm{d}M = 1$$

Differenziert man nach θ, so erhält man:

$$0 \overset{\mathcal{R}}{=} \int \frac{\partial L(\theta; M)}{\partial\theta}\mathrm{d}M = \int \frac{\partial l(\theta; M)}{\partial\theta} L(\theta; M)\mathrm{d}M = E(S(\theta; M)).$$

Beweis zu (b):
Aus Teil (a) folgt:

$$\mathrm{Cov}(S(\theta; M)) = E(S(\theta; M)S(\theta; M)^T).$$

Weiterhin gilt:

$$\begin{aligned}
F(\theta) &= E\left(-\frac{\partial^2 l(\theta;M)}{\partial\theta\partial\theta^T}\right)\\
&= E\left(-\frac{\partial}{\partial\theta}\frac{\frac{\partial}{\partial\theta^T}L(\theta;M)}{L(\theta;M)}\right)\\
&= E\left(-\frac{\left(\frac{\partial^2}{\partial\theta\partial\theta^T}L(\theta;M)\right)\cdot L(\theta;M)-\frac{\partial}{\partial\theta}L(\theta;M)\cdot\frac{\partial}{\partial\theta^T}L(\theta;M)}{L(\theta;M)\cdot L(\theta;M)}\right)\\
&= -E\left(\frac{\frac{\partial^2}{\partial\theta\partial\theta^T}L(\theta;M)}{L(\theta;M)}\right)+E\left(\frac{\frac{\partial}{\partial\theta}L(\theta;M)\frac{\partial}{\partial\theta^T}L(\theta;M)}{L(\theta;M)\cdot L(\theta;M)}\right)\\
&= -\int\frac{\frac{\partial^2}{\partial\theta\partial\theta^T}L(\theta;M)}{L(\theta;M)}\cdot L(\theta;M)\,\mathrm{d}M+\int\frac{\frac{\partial}{\partial\theta}L(\theta;M)\frac{\partial}{\partial\theta^T}L(\theta;M)}{L(\theta;M)\cdot L(\theta;M)}\cdot L(\theta;M)\,\mathrm{d}M\\
&= -\int\frac{\partial^2}{\partial\theta\partial\theta^T}L(\theta;M)\,\mathrm{d}M+\int\frac{\partial}{\partial\theta}l(\theta;M)\frac{\partial}{\partial\theta^T}l(\theta;M)L(\theta;M)\,\mathrm{d}M\\
&\overset{\mathcal{R}}{=} 0+E(S(\theta;M)S(\theta;M)^T).
\end{aligned}$$

□

Ein erster Zusammenhang zwischen der Informationsmatrix $F(\theta)$ und der Kovarinazmatrix $\mathrm{Cov}\,(\hat{\theta})$ ergibt sich durch folgenden Satz [MKB79]:

Satz 3.5 *(Die Cramér-Rao-Schranke)*
Ist $\hat{\theta}_{ML}$ ein unverzerrter Maximum-Likelihood-Schätzer für θ, so gilt

$$Cov(\hat{\theta}_{ML}) \geq F(\theta)^{-1} \tag{3.22}$$

mit $F(\theta)$ gegeben durch (3.20).

Das bedeutet, dass die Inverse der Informationsmatrix die bestmögliche Kovarianzmatrix für die Schätzung eines Parameters ist.

Zum Beweis dieses Satzes werden folgende drei Lemmata benötigt:

Lemma 3.6 *Sei $S(\theta;M)$ die Score-Funktion wie in Definition 3.13 und $t(\theta;M)$ eine beliebige Funktion, die von M und θ abhängig sei und einen Vektor der gleichen Elementanzahl wie $S(\theta;M)$ zum Funktionswert hat, dann gilt mit der Regularitätsbedingung $\mathcal{R}$:*

$$E(S(\theta;M)t(\theta;M)^T) = \frac{\partial}{\partial\theta}E(t(\theta;M)) - E\left(\frac{\partial t(\theta;M)^T}{\partial\theta}\right) \tag{3.23}$$

Beweis:
Es gilt:

$$E\left(t(\theta;M)^T\right) = \int t(\theta;M)^T L(\theta;M)\,\mathrm{d}M$$

Differenziert man beide Seiten im Hinblick auf θ, so erhält man

$$\frac{\partial}{\partial\theta}E(t(\theta;M)^T) \overset{\mathcal{R}}{=} \int \frac{\partial l(\theta;M)}{\partial\theta} t(\theta;M)^T L(\theta;M)\,\mathrm{d}M + \int \frac{\partial t(\theta;M)^T}{\partial\theta} L(\theta;M)\,\mathrm{d}M$$

und damit auch (3.23).

□

Lemma 3.7 *Sei $S(\theta;M)$ wie in Definition 3.13 und $\hat{\theta}$ ein erwartungstreuer Schätzer für θ, dann gilt:*

$$E(S(\theta;M)\hat{\theta}^T) = I \tag{3.24}$$

mit I als der Einheitsmatrix passender Dimension.

Beweis:
Der Beweis dieses Satzes ergibt sich unmittelbar aus Gleichung (3.23), da

$$\frac{\partial\hat{\theta}}{\partial\theta} = 0$$

□

Lemma 3.8 *Seien A und B symmetrische Matrizen mit $B>0$, $A, B \in \mathbb{R}^{n\times n}$.*

1. *Das Maximum über alle $x \in \mathbb{R}^n$ für x^TAx unter der Bedingung, dass $x^TBx = 1$ gilt, ist gegeben durch den Eigenvektor von $B^{-1}A$, der zum größten Eigenwert von $B^{-1}A$ korrespondiert.*

2. *Das Maximum von a^Tx mit $a, x \in \mathbb{R}^n$ unter der Bedingung, dass $x^TBx = 1$ gilt, ist gegeben durch $(a^TB^{-1}a)^{\frac{1}{2}}$. Außerdem gilt:*

$$\max_{x\in\mathbb{R}^n} \frac{(a^Tx)^2}{x^TBx} = a^TB^{-1}a \tag{3.25}$$

für

$$x_{max} = \frac{B^{-1}a}{a^TB^{-1}a}. \tag{3.26}$$

Beweis zu 1.:
Sei $B^{\frac{1}{2}}$ die Matrix, für die gilt

$$B = B^{\frac{1}{2}} \cdot B^{\frac{1}{2}}$$

und gelte weiterhin $y = B^{\frac{1}{2}} \cdot x$. Dann kann die Maximierung von $x^T Ax$ unter der oben angegebenen Bedingung formuliert werden als:

$$\max_{y \in \mathbb{R}^n} y^T B^{-\frac{1}{2}} A B^{-\frac{1}{2}} y \qquad \text{mit} \qquad y^T y = 1 \tag{3.27}$$

Sei $\Gamma \Lambda \Gamma^T$ die Spektralzerlegung der symmetrischen Matrix $B^{-\frac{1}{2}} A B^{-\frac{1}{2}}$ (vergleiche Abschnitt 3.4) und $z = \Gamma^T y$. Dann gilt

$$z^T z = y^T \Gamma \Gamma^T y = y^T y$$

und man kann (3.27) schreiben als

$$\max_{z \in \mathbb{R}^n} z^T \Lambda z = \max_{z \in \mathbb{R}^n} \sum_i \lambda_i z_i^2 \qquad \text{mit} \qquad z^T z = 1.$$

λ_i seien also die Eigenwerte von $B^{-\frac{1}{2}} A B^{-\frac{1}{2}}$ und λ_1 sei der größte von ihnen, so gilt

$$\max_{z \in \mathbb{R}^n} \sum_i \lambda_i z_i^2 \leq \lambda_1 \max_{z \in \mathbb{R}^n} \sum_i z_i^2 = \lambda_1$$

Da für zwei Matrizen $C \in \mathbb{R}^{n \times p}$ und $D \in \mathbb{R}^{p \times n}$ gilt, dass die nichttrivialen Eigenwerte von CD und DC gleich sind, so gilt hier, dass jeder Eigenwert von $B^{-\frac{1}{2}} A B^{-\frac{1}{2}}$ auch ein Eigenwert von $B^{-1}A$ ist. Es folgt zudem, dass $x = B^{-1}\Gamma_{\bullet 1}$ der zum größten Eigenwert λ_1 korrespondierende Eigenvektor ist.

Beweis zu 2.:
Die Behauptung folgt aus *1.* bei Betrachtung von $x^T Ax = (a^T x)^2 = x^T (aa^T) x$.

□

Beweis zu Satz 3.5 (Cramér-Rao-Schranke):
Betrachte $\text{Corr}(\alpha, \gamma)$ mit $\alpha = a^T \hat{\theta}_{ML}$ und $\gamma = c^T s$ mit $s := S(\theta; M)$, c, a passender Dimension. Nach Satz 3.4, Lemma 3.7 und den Rechenregeln für die Kovarianzmatrix gilt

$$\text{Cov}\,(\alpha, \gamma) = a^T \text{Cov}\,(\hat{\theta}_{ML}, s) c = a^T c \tag{3.28}$$

$$\text{Cov}\,(\gamma) = c^T \text{Cov}\,(s) c = c^T F(\theta) c \tag{3.29}$$

Also gilt

$$\begin{aligned} \mathrm{Corr}^2(\alpha,\gamma) &= \frac{\mathrm{Cov}^2(\alpha,\gamma)}{\mathrm{Cov}(\alpha)\mathrm{Cov}(\gamma)} && (3.30)\\ &= \frac{(a^T c)^2}{a^T Cov(\hat{\theta}_{ML}) a c^T F(\theta) c} \leq 1. && (3.31) \end{aligned}$$

Maximiert man die linke Seite der Gleichung (3.30) in Bezug auf c, so ergibt sich nach Lemma 3.8 für alle a

$$\begin{aligned} \frac{a^T F(\theta)^{-1} a}{a^T \mathrm{Cov}(\hat{\theta}_{ML}) a} &\leq 1 \\ \Leftrightarrow a^T(\mathrm{Cov}(\hat{\theta}_{ML}) - F(\theta)^{-1})a &\geq 0 \end{aligned}$$

und das wiederum ist äquivalent zu 3.22.

□

Unter den Bedingungen von Satz 3.3 zu den Eigenschaften eines linearisierten Regressionsmodells lässt sich für den Zusammenhang zwischen $Cov(\hat{\theta}_{ML})$ und $F(\theta)$ die viel stärkere Aussage formulieren [FH95]:

$$F(\theta) = \nu \left[Cov(\hat{\theta}_{ML}) \right]^{-1} = \nu \left(Sens_M^T \cdot \Sigma^{-1} Sens_M \right) \tag{3.32}$$

mit einem geeigneten Faktor ν.

Bei Betrachtung der Fisher-Informationsmatrix anstatt der Kovarianzmatrix einer Parameterschätzung erübrigt sich eine Matrix-Inversion. Das ist vorteilhaft, vor allem, wenn die Matrix $F(\theta)$ numerisch schlecht konditioniert ist. Beide Matrizen, sowohl $F(\theta)$ als auch $\mathrm{Cov}(\hat{\theta}_{ML})$ enthalten eine gewisse „Informationsmenge“ über den unbekannten Parametersatz und die Güte seiner Bestimmbarkeit. Auf dieses Konzept wird in Abschnitt 3.10 näher eingegangen.

Da im linearen Regressionsmodell die Schätzung für θ mit der Methode der kleinsten Quadrate gleich der Schätzung ist, die man mit Hilfe des Maximum-Likelihood-Prinzips erhält, können alle in diesem Kapitel formulierten Erkenntnisse auch approximativ unter den Bedingungen von Satz 3.3 auf den in dieser Arbeit definierten Schätzer $\hat{\theta}_{KQSS}$ (siehe Definition 3.7) übertragen werden.

3.9. Modelldiskriminierung

In einem Regressionsmodell definiert die Regressionsfunktion die Abhängigkeit zwischen den Eingangsvariablen und den Messwerten. Allerdings gibt es oft mehrere

Annahmen, welcher Art diese Abhängigkeit sein kann. Das bedeutet, dass mehrere mögliche Regressionsfunktionen zur Auswahl stehen und deshalb Kriterien gefunden werden müssen, um sie anhand der Messdaten vergleichen und nach ihrer Güte sortieren zu können. Zwei Eigenschaften der betrachteten Modelle können intuitiv als Bewertungskriterium dienen. Zum einen, wie gut die Daten durch das Modell angepasst werden, zum anderen, wieviele unbekannte Parameter das Modell enthält.

Die Güte der Anpassung wiederum lässt sich unterschiedlich definieren; eine sehr einfache Möglichkeit ist es, die Summe der Residuenquadrate zu betrachten, die den Abstand der Messpunkte von der Regressionsfunktion ausdrücken (siehe 3.10). Bei gleicher Güte zweier Modelle kann dann das Prinzip der Sparsamkeit (engl. *principle of parsimony*) angewendet werden, welches dasjenige bevorzugt, bei dem weniger Parameter vorkommen [BJ76]. Das in dieser Arbeit verwendete Akaike-Kriterium zur Modelldiskriminierung bewertet beide Modelleigenschaften und wird nun näher erläutert.

3.9.1. Akaike-Kriterium

Definition 3.14 *Erfülle das Regressionsmodell die Annahme normalverteilter Fehler ($\varepsilon \sim \mathcal{N}(0, \Sigma)$), so lautet das* Akaike-Informationskriterium (AIC)*:*

$$AIC \quad = \quad N \ln\left(\frac{\sum_{ij} \epsilon_{ij}^2}{N}\right) + 2K$$

wobei K die Anzahl der im Regressionsmodell zu schätzenden Parameter ist, ϵ_{ij} wie in Definition 3.7 und $N = s_D \cdot n$ die Anzahl der Messwerte [BA02].

Für kleine Stichprobengrößen aus denen mehrere Parameter geschätzt werden sollen, ist das *Small Sample AIC* nach [HT89] definiert:

$$AIC_C \quad = \quad AIC + \frac{2K(K+1)}{N-K-1} \tag{3.33}$$

Anzumerken ist, dass aus dem Wert des AIC für ein einzelnes Modell nichts abgeleitet werden kann. Informativ ist das AIC erst im Vergleich der berechneten AIC-Werte mehrerer Modelle [FH95]. Zudem müssen bei der Berechnung von K in (3.33) alle geschätzten Regressionsparameter beachtet werden, also neben allen Einträgen des Parametervektors θ auch gegebenenfalls unbekannte Messwertvarianzen. Da diese in der vorliegenden Arbeit als bekannt angenommen wurden, gilt für die Berechnungen des AIC sowie seine Herleitung $K = p$. Es folgt eine univariate Herleitung des AIC nach [BA02].

Sei $x = (x_1, \ldots, x_N)$ ein Datenvektor und $h(x)$ die „wahre“ Dichte, die ihm zugrunde liegt. $[g(x;\theta)]_{\theta\in\Theta}$ sei die Klasse aller betrachteten Modelle und $g(x;\theta_0)$ das beste Modell dieser Klasse. Dann lautet die *Kullback-Leibler-Information (KLI)* für dieses Modell

$$I(h, g(\cdot;\theta_0)) \;=\; \int h(x) \ln\left(\frac{h(x)}{g(x;\theta_0)}\right) \mathrm{d}x. \tag{3.34}$$

$I(h, g(\cdot;\theta_0))$ ist also abhängig von dem unbekannten Wert für θ_0 und der unbekannten wahren Dichte $h(\cdot)$, aber nicht vom Datensatz x, da über x integriert wird. Sei nun $y = y_1, \ldots, y_N$ ein weiterer, von x stochastisch unabhängiger Datensatz, der aus der (unbekannten) Funktion $h(\cdot)$ hervorgeht, so kann für θ_0 auf Basis dieses Datensatzes ein ML-Schätzer $\hat{\theta}_{ML} = \hat{\theta}(y)$ errechnet werden. Entsprechend kann dann eine Schätzung für (3.34) angegeben werden mit:

$$I(h, g(\cdot;\hat{\theta}(y))) \;=\; \int h(x) \ln\left(\frac{h(x)}{g(x;\hat{\theta}(y))}\right) \mathrm{d}x$$

Wäre nun $g(x;\theta_0)$ das „perfekte“ Modell, so würde gelten $I(h, g(\cdot;\theta_0)) = 0$. Allerdings bleibt zu bemerken, dass selbst wenn man das perfekte Modell gefunden hätte, also $g(x;\theta_0) = h(x)$ gälte, trotzdem der aus den verrauschten Messdaten geschätzte Parameter $\hat{\theta}(y)$ nie genau gleich θ_0 sein würde. Für jeden Wert $\hat{\theta}(y)$, der ungleich θ_0 ist, gilt aber: $I(h, g(\cdot;\hat{\theta}(y))) > I(h, g(\cdot;\theta_0))$, was impliziert, dass die KLI immer ein Wert größer als 0 ist. Da es sich um Zufallsexperimente handelt, gilt somit

$$E_Y[I(h, g(\cdot;\hat{\theta}(y)))] \;>\; I(h, g(\cdot;\theta_0)).$$

Das Modell $g(x;\theta) \in [g(x;\theta)]_{\theta\in\Theta}$ ist so zu wählen, dass $E_Y[I(h, g(\cdot;\hat{\theta}(y)))]$ minimiert wird. Es gilt

$$\begin{aligned} E_Y[I(h, g(\cdot;\hat{\theta}(y)))] &= \int h(x)\ln(h(x))\mathrm{d}x - E_Y\left[\int h(x)\ln(g(x;\hat{\theta}(y)))\mathrm{d}x\right] \\ &= C - E_Y E_X\left[\ln(g(X;\hat{\theta}(y)))\right] \end{aligned}$$

mit einem konstanten Wert C. Es kann nun ein Minimum durch Maximierung des Ausdrucks

$$T := E_Y E_X[\ln(g(X;\hat{\theta}(y)))] \tag{3.35}$$

gefunden werden, beziehungsweise durch Maximierung des Ausdrucks

$$\int h(y) \left[\int h(x) \ln(g(x; \hat{\theta}(y))) \mathrm{d}x \right] \mathrm{d}y.$$

Durch Betrachtung der Taylor-Entwicklung des Ausdrucks $\ln g(x; \hat{\theta}(y))$ um den Punkt θ_0, erhält man

$$\begin{aligned} \ln g(x; \hat{\theta}(y)) &\approx \ln g(x; \theta_0(y)) + \left[\frac{\partial \ln g(x; \theta_0(y))}{\partial \theta} \right]^T [\hat{\theta}(y) - \theta_0(y)] \\ &+ \frac{1}{2} [\hat{\theta}(y) - \theta_0(y)]^T \left[\frac{\partial^2 \ln g(x; \theta_0(y))}{\partial \theta^2} \right] [\hat{\theta}(y) - \theta_0(y)] \end{aligned} \quad (3.36)$$

wobei $\theta_0(y) = \theta_0$ gilt, da θ_0 nicht von y abhängt. Betrachtet man nun den Erwartungswert bezüglich der Zufallsvariablen X, so gilt:

$$\begin{aligned} E_X \left[\ln g(X; \hat{\theta}(y)) \right] &\approx E_X \left[\ln g(X; \theta_0) \right] + E_X \left[\left[\frac{\partial \ln g(X; \theta_0)}{\partial \theta} \right]^T \right] [\hat{\theta}(y) - \theta_0] \\ &+ \frac{1}{2} [\hat{\theta}(y) - \theta_0]^T \left[E_X \left[\frac{\partial^2 \ln g(X; \theta_0)}{\partial \theta^2} \right] \right] [\hat{\theta}(y) - \theta_0] \end{aligned} \quad (3.37)$$

Da außerdem

$$E_X \left[\frac{\partial \ln g(x; \theta_0)}{\partial \theta} \right] = 0$$

gilt (vergl. Satz 3.4), entfällt im Ausdruck (3.37) der lineare Term:

$$E_X \left[\ln g(X; \hat{\theta}(y)) \right] \approx E_X \left[\ln g(X; \theta_0) \right] + \frac{1}{2} [\hat{\theta}(y) - \theta_0]^T \left[E_X \left[\frac{\partial^2 \ln g(X; \theta_0)}{\partial \theta^2} \right] \right] [\hat{\theta}(y) - \theta_0].$$

Mit der Definition der Fisher-Informationsmatrix (Definition 3.13) gilt nun also

$$\begin{aligned} E_X \left[\ln g(X; \hat{\theta}(y)) \right] &\approx E_X \left[\ln g(X; \theta_0) \right] \\ &\quad - \frac{1}{2} [\hat{\theta}(y) - \theta_0]^T \cdot F(\theta_0) \cdot [\hat{\theta}(y) - \theta_0] \end{aligned} \quad (3.38)$$

und nun gilt mit Hilfe des Satzes über die Erwartungswerte quadratischer Formen (siehe Anhang Teil C)

$$E_Y E_X \left[\ln g(X; \hat{\theta}(y)) \right] \approx E_X \left[\ln g(X; \theta_0) \right] - \frac{1}{2} \mathrm{tr} \left[F(\theta_0) E_Y \left([\hat{\theta}(y) - \theta_0][\hat{\theta}(y) - \theta_0]^T \right) \right]$$

wobei

$$E_Y \left([\hat{\theta}(y) - \theta_0][\hat{\theta}(y) - \theta_0]^T \right) = \mathrm{Cov} \left(\hat{\theta}(y) \right) =: \Sigma_{\hat{\theta}}$$

gilt, da $\hat{\theta}$ ein erwartungstreuer Schätzer ist. Also ist in (3.35)

$$T \approx E_X[\ln g(X;\theta_0)] - \frac{1}{2}\mathrm{tr}\,[F(\theta_0)\Sigma_{\hat{\theta}}]. \tag{3.39}$$

Durch Taylor-Entwicklung von $\ln g(x;\theta_0)$ um $\hat{\theta}(x)$ erhält man dagegen

$$\begin{aligned}\ln g(x;\theta_0) &\approx \ln g(x;\hat{\theta}(x)) + \left[\frac{\partial \ln g(x;\hat{\theta}(x))}{\partial\theta}\right]^T [\theta_0 - \hat{\theta}(x)] \\ &+ \frac{1}{2}[\theta_0 - \hat{\theta}(x)]^T \left[\frac{\partial^2 \ln g(x;\hat{\theta}(x))}{\partial\theta^2}\right] [\theta_0 - \hat{\theta}(x)].\end{aligned}$$

Da $\hat{\theta}(x)$ als ML-Schätzer allerdings die Lösung der Score-Gleichungen ist (vergl. (3.21)), gilt:

$$\frac{\partial \ln g(x;\hat{\theta}(x))}{\partial\theta} = 0.$$

Also kann auch hier der lineare Term vernachlässigt werden. Im Zusammenhang mit dem Erwartungswert wird analog zu (3.38) formuliert:

$$E_X\left[\ln g(X;\theta_0)\right] \approx E_X\left[\ln g(X;\hat{\theta}(x))\right] - \frac{1}{2}\mathrm{tr}\,[F(\hat{\theta}(x))\Sigma_{\hat{\theta}}]$$

Da $F(\hat{\theta}(x)) \to F(\theta_0)$ für $N \to \infty$ gilt (siehe [BA02], [FH95]), gilt nun mit Hilfe der Approximation (3.39)

$$\begin{aligned}E_X\left[\ln g(X;\theta_0)\right] &\approx E_X\left[\ln g(X;\hat{\theta}(x))\right] - \frac{1}{2}\mathrm{tr}\,[F(\hat{\theta}(x))\Sigma_{\hat{\theta}}] \\ \overset{(3.39)}{\Rightarrow} T &\approx E_X\left[\ln g(X;\hat{\theta}(x))\right] - \mathrm{tr}\,[F(\theta_0)\Sigma_{\hat{\theta}}].\end{aligned}$$

Das Kriterium entspricht also von seiner Struktur her folgendem Ausdruck, wenn die Spur $\mathrm{tr}[F(\theta_0)\Sigma_{\hat{\theta}}]$ geschätzt wird:

$$\hat{T} \approx \ln g(x;\hat{\theta}(x)) - \hat{\mathrm{tr}}\,[F(\theta_0)\Sigma_{\hat{\theta}}]$$

beziehungsweise unter der Annahme $\Sigma_{\hat{\theta}} = \mathrm{Cov}\,(\hat{\theta}(y)) \approx F(\theta_0)$ (siehe Gleichung 3.32):

$$\hat{T} \approx \ln g(x;\hat{\theta}(x)) - \hat{\mathrm{tr}}\,\left[F(\theta_0)F(\theta_0)^{-1}\right].$$

Aus konventionellen Gründen hat sich eine Multiplikation des zu minimierenden Kriteriums mit 2 durchgesetzt [BA02]:

$$\hat{T} \approx -2\ln(g(x;\hat{\theta}(x))) + \underbrace{2\,\hat{\mathrm{tr}}\,[I]}_{=2\cdot K}$$

mit I als der K-dimensionalen Einheitsmatrix, was zum allgemeinen AIC führt

$$AIC \quad = \quad -2\ln(g(x;\hat{\theta}(x))) + 2K.$$

Da unter Normalverteilungssannahme für die Fehler der Messwerte die Log-Likelihoodfunktion mit eingesetzten ML-Schätzern $\hat{\theta}(x) = \hat{\theta}_{ML}, \hat{\sigma}^2_{ML}$

$$l(\hat{\theta}_{ML};x) \quad = \quad \ln\left((2\pi\hat{\sigma}^2_{ML})^{-\frac{N}{2}}\exp\left(\frac{\sum_{i=1}^{N}(x_i - v_i(x;\hat{\theta}_{ML}))^2}{2\hat{\sigma}^2_{ML}}\right)\right)$$

lautet, wobei $v_i(x;\hat{\theta}_{ML})$ den Voraussagewert für den Messpunkt x_i darstellen soll, kann das AIC auch notiert werden als

$$AIC \quad = \quad N\ln(2\pi\hat{\sigma}_{ML}) + 2K + C$$

wobei C eine positive Konstante ist. Da der Logarithmus eine streng monoton steigende Funktion ist, kann der Faktor 2π vernachlässigt werden. So ergibt sich mit ϵ_{ij} aus Definition 3.7

$$AIC \quad = \quad N\ln\left(\frac{\sum_{ij}\epsilon^2_{ij}}{N}\right) + 2K.$$

□

3.10. Optimale Versuchsplanung

Erstmalig eingeführt bei [Fis35], basiert optimales Versuchsdesign darauf, die manipulierbaren Faktoren, beziehungsweise Umweltbedingungen des Experiments, für a priori bestimmte Zwecke (beispielsweise Parameterschätzung, Modelldiskriminierung) so einzustellen, das möglichst wenige Messungen mit jeweils möglichst hohem Informationsgehalt bezüglich des Zwecks gemacht werden können [SCS00]. Manipulierbare Eingangsvariablen bei Experimenten zu enzymatischen Reaktionen sind die Anfangskonzentrationen der eingesetzten Substrate, die Konzentration des Enzyms zu Beginn der Reaktion und die Anordnung der Messzeitpunkte.

In dieser Arbeit wurden Versuchspläne sequentiell erstellt, das bedeutet ein Vorgehen mit folgenden Schritten:

1. Schätzung des gesuchten Parameters, wobei bereits bekanntes Wissen miteinbezogen wird (siehe Kapitel 3.6). Es ist auch möglich, einen Versuch nur auf der Grundlage einer Schätzung für die Verteilung eines Parameters zu planen [AD92]. Diese Methode wurde in der vorliegenden Arbeit nicht verwendet.

2. Linearisierung des Regressionsmodells (siehe Kapitel 3.7)
3. Bestimmung der optimalen Versuchspläne für das lineare Modell
4. Durchführung der Versuchspläne und Verbesserung der Parameterschätzung auf Grundlage der Ergebnisse

Hauptziel der optimalen Versuchsplanung ist es meist, Varianzen zu verringern. Dabei kann es sich um Varianzen zu schätzender Parameter handeln oder aber auch um Varianzen für Messwertvorraussagen [AD92].

In dieser Arbeit wurden drei Ziele mit Hilfe der optimalen Versuchsplanung formuliert und verfolgt:

- Verminderung der Varianzen der geschätzten Modellparameter
- Verminderung der Korrelationen zwischen den geschätzten Modellparametern
- Ermöglichung einer deutlicheren Modelldiskriminierung

Es stellt sich nun die Frage nach geeigneten Kriterien, um Versuchsplanung im Hinblick auf die oben angeführten drei Ziele durchführen zu können. Da Varianzen in der Kovarianzmatrix verzeichnet sind, soll zunächst ein Blick auf diese „Informationsquelle“ geworfen werden. Visualisiert man beispielsweise die Kovarianzmatrix zweier Parameter θ_1 und θ_2 mit Hilfe der Konfidenz-Ellipse nach einer Hauptachsentransformation wie in Abschnitt 3.4 beschrieben, so kann sie bei gleicher Achsenskalierung unterschiedliche Formen annehmen, wie in Abbildung 3.4 gezeigt. Die Varianzen der Parameter werden durch Projektion der Konfidenz-Ellipse auf die jeweiligen Achsen in Rot angezeigt.

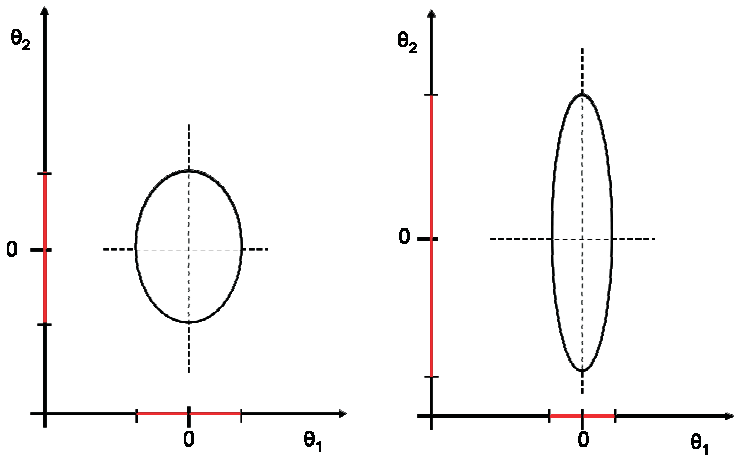

Abbildung 3.4.: Beispiele für Konfidenz-Ellipsen

Welche der Konfidenz-Ellipsen als „besser“ angesehen werden kann, kommt ganz auf die Betrachtungsweise an. Soll Parameter θ_1 sehr genau bestimmt werden, so

ist die rechte Konfidenzellipse in Abb. 3.4, beziehungsweise das Experiment, das diese Konfidenzellipse ergibt, geeigneter. Sollen hingegen möglichst alle Parameter gleichzeitig gut bestimmt werden, so ist der Versuch, auf dessen Grundlage die Kovarianzmatrix für die linke Konfidenzellipse in Abb. 3.4 geschätzt wurde, dem anderen überlegen.

Deshalb ist es nötig, Kriterien zu formulieren, die vom Ziel der Versuchsplanung abhängen. Da wie in Abschnitt 3.4 beschrieben, die Form des Konfidenz-Ellipsoids von den Eigenwerten und Eigenvektoren der Kovarianzmatrix abhängt, liegt es nahe, Funktionale zu formulieren, die über diese Größen auf die Form Einfluss nehmen können. Solche Funktionale werden im folgenden für die verschiedenen Ziele der Modelldiskriminierung definiert.

Bemerkung:
Wegen Gleichung (3.32) lassen sich die Kriterien der optimalen Versuchsplanung als Funktional der Kovarianzmatrix oder der Fisher-Informationsmatrix formulieren. Die Minimierung eines Funktionals der Kovarianzmatrix ist gleichbedeutend zu der Maximierung desselben Funktionals der Informationsmatrix (Prinzip der *Informationsmaximierung*). Die Maximierung eines Funktionals der Informationsmatrix lässt sich auch als Maximierung der zweiten Ableitung der Log-Likelihoodfunktion deuten, also ihrer Krümmung. Ist die Krümmung der Maximum-Likelihoodfunktion größer, können Parameter mit kleineren Varianzen geschätzt werden.

3.10.1. Verminderung der Varianz einer Parameterschätzung

Zur Verminderung der Varianzen der Modellparameterschätzung wurde die *D-Optimalität* verwendet, die im folgenden als Minimierungsproblem definiert wird.

Definition 3.15 *Sei m die Anzahl der Eingangsvariablen eines Experiments und $V \in \mathbb{R}^m$ der Raum aller möglichen Tupel von Eingangsvariablen. Der Versuchsplan ζ sei definiert als*

$$\zeta := (x_1, \ldots, x_m) \quad mit \quad x_1, \ldots, x_m \in \mathbb{R} \quad \zeta \in V$$

Zusätzlich gelte:

$$V^+ := \{\zeta \in \mathbb{R}^m | \det F(\zeta) \neq 0\}$$

wobei $F(\zeta)$ abkürzend für die Informationsmatrix stehen soll, die man bei gegebenem Regressionsmodell und gegebenen Versuchsplan ζ für den auf dieser Basis geschätzten Parametersatz nach Formel 3.32 erhält.

Definition 3.16 *[BB94] Ein konkreter Versuchsplan ζ^* heißt* D-optimal*, wenn gilt*

$$\det\, F^{-1}(\zeta^*) = \min_{\zeta \in V^+} \det\, F^{-1}(\zeta).$$

Die Minimierung der Determinante der Inversen der Informationsmatrix entspricht nach Gleichung 3.32 einer Minimierung der Determinante der Kovarianzmatrix für den geschätzten Parametersatz $\hat{\theta}$. Für das Konfidenz-Ellipsoid bedeutet dies, dass das zu der Determinante von $F^{-1}(\zeta)$ proportionale Volumen des Ellipsoids minimiert wird.

3.10.2. Verminderung der Korrelationen zwischen Parameterschätzungen

Zur Verminderung der Korrelationen zwischen den Modellparametern wurde die *E-Optimalität* verwendet, die unter Voraussetzungen von Definition 3.15 folgendermaßen definiert wird:

Definition 3.17 *[BB94]*
Ein konkreter Versuchsplan ζ^* heißt E-optimal*, wenn gilt*

$$\lambda_{\max}(F^{-1}(\zeta^*)) = \min_{\zeta \in V^+} \lambda_{\max}(F^{-1}(\zeta)).$$

$\lambda_{\max}(F^{-1}(\zeta))$ *stelle den größten Eigenvektor der inversen Fisher-Informationsmatrix dar.*

Die Verwendung dieses Kriteriums zur Verminderung der Korrelationen wird dadurch begründet, dass durch die Minimierung des größten Eigenwertes der Kovarianzmatrix zugleich die Ausdehnung des Konfidenz-Ellipsoids in die angezeigte Richtung minimiert wird. Dadurch werden gleichzeitig auch die Korrelationen der anderen Parameter mit dem zu diesem Eigenwert gehörigen Parameter verringert.

3.10.3. Optimale Versuchsplanung zur Modelldiskriminierung

Ziel der optimalen Versuchsplanung zur Modelldiskriminierung ist es, die Eingangsparameter eines Versuchs so zu wählen, dass zwei oder mehrere Modelle sich gut voneinander unterscheiden lassen [AD92]. Der Abstand zwischen zwei Modellen kann zum einen durch das in Kapitel 3.9 angesprochene Akaike-Kriterium definiert werden, aber auch direkt über den tatsächlichen Abstand voneinander berechnet werden. Lag bei den zu vergleichenden Modellen die gleiche Parameteranzahl vor, so wurde für die optimale Versuchsplanung letztere Methode verwendet.

Als Voraussetzung für die optimale Versuchsplanung muss eines der Modelle als „korrekt“ angenommen werden. Anschließend wird der Abstand der anderen Modelle zu diesem Ausgangsmodell berechnet. Für dieses Ausgangsmodell muss also bereits eine Parameterschätzung vorliegen. Dies ist insofern problematisch, als durch die Güte der Parameterschätzung auch die Güte der Modelldiskriminierung bestimmt wird. Im Folgenden wird das als korrekt angenommene *Ausgangsmodell* mit $H_1(G;\hat{\theta}_\zeta) = H_1(G)$ bezeichnet, wobei $\hat{\theta}_\zeta$ die Schätzung für θ unter den Bedingungen des Versuchsplans ζ sei.

Definition 3.18 *Der* Abstand *des Modells $H_2(G;\hat{\theta}_\zeta)$ zum Ausgangsmodell sei definiert durch*

$$\Delta(\zeta) \;:=\; \sum_{i=1}^{s_D}\sum_{j=1}^{n} |l_{1Dij} - l_{2Dij}|$$

mit $l_{kDij} \in L_k(G;\hat{\theta}_\zeta;a_0;T_D)$ für k=1,2.

Es ist also nun derjenige Versuchsplan ζ^* optimal für die Modelldiskriminierung, für den gilt

$$\Delta(\zeta^*) \;=\; \max_{\zeta\in V^+}\Delta(\zeta)$$

Eine Beschränkung bei der Anzahl der zu untersuchenden Modelle ist ratsam, da sonst eine „kombinatorische Explosion“ auftreten kann.

4. Ausgangslage

4.1. Modellrelevantes Wissen zur Benzaldehydlyase

Eines der Ziele dieser Arbeit war, auf Basis von Datensätzen eines Experiments zur Benzaldehydlyase (BAL) ein mathematisches Modell für die BAL-katalysierte Reaktion von BA zu DHPP aufzustellen. In diesem Kapitel werden zunächst alle bisher in der Literatur vorhandenen Informationen zur BAL erfasst, um Hinweise für die Modellformulierung zu erhalten. Weiterhin werden modellrelevante Ergebnisse bisheriger Experimente zur BAL, sowie das in der Literatur bisher einzige mathematische Modell zur Beschreibung der betrachteten Reaktion dargestellt.

Im Jahr 1989 veröffentlichten B. Gonzales und R. Vicuña die Entdeckung, dass das Bakterium *Pseudomonas fluorescens* Biovar I autonom auf einer Benzoinquelle wachsen, es also Benzoin als einzige Kohlenstoff- und Energiequelle nutzen kann. Sie fanden das dafür verantwortliche Enzym, die BAL, isolierten es aus dem Bakterium, bestimmten seinen Kofaktor Thiamindiphosphat (ThDP, siehe Abb. 2.4) und sein Molekulargewicht. Zudem schlugen sie einen Reaktionsmechanismus für die BAL-katalysierte Benzoinspaltung in zwei Benzaldehydmoleküle vor, der auf den Erkenntnissen über andere ThDP-abhängige Enzyme basiert [GV89].

Demir et al. publizierten 2001 ein Papier über Experimente, bei denen sie mit Hilfe der BAL von Benzoin abgeleitete Stoffe synthetisierten sowie Hydroxypropiophenone aus aromatischen Aldehyden und Acetaldehyden. Sie konnten die Synthese in wässrigem Medium unter Zugabe des Lösungsmittels Dimethylsulfoxid (DMSO) und einer Pufferlösung nachweisen. Bei diesen Untersuchungen legten sie einen Schwerpunkt auf die Selektivität der BAL im Bezug auf chirale Substrate und Produkte (siehe Kapitel 2) und erforschten Umsetzungen über ein weites Substratspektrum [DEH$^+$02].

Von E. Janzen wurde 2002 die nach gentechnischer Veränderung (*Biotransformation*) stattfindende Überexpression der BAL in *E. coli*, also einem sehr gut erforschten Organismus, als Grundlage für weitere Experimente in grösserem Maßstab und als Vorbereitung für die industrielle Nutzung des Enzyms beschrieben. Der Einfluss der Umgebungsbedingungen wie pH-Wert und Einsatz verschiedener organischer Lö-

sungsmittel auf die Enzymaktivität nach der Biotransformation wurde beispielsweise von [Jan02] und [JMKJ+06] untersucht.

Ein Konzept für die Konformationsänderung der BAL und die auftretenden Wechselwirkungen bei der Umsetzung von Substraten wurde 2004 aufgestellt [Dün04]. Es orientierte sich an den Erkenntnissen über andere ThDP-abhängige Enzyme. In [Hil05], [Sti04], [Küh07] und [HKP+07] wurden die Kinetiken spezifischer Synthesereaktionen (BA zu BZ und auch von Folgereaktionen zu Hydroxypropiophenonen) näher untersucht, sowohl durch Analysen der Anfangskinetiken als auch durch Anpassung der Daten an Differentialgleichungsmodelle mit Hilfe des Programms Scientist®.

Die BAL wird heute noch nicht in größerem industriellen Maßstab genutzt. Bisher wurden vor allem ihre katalytischen Eigenschaften untersucht, das heißt die Auswirkung äußerer Einflüsse auf die Reaktionsgeschwindigkeit und die Stabilität der BAL. Das geschah vor allem, um ihren industriellen Nutzen abzuschätzen, sowie Vorteile der enzymatischen Umwandlung gegenüber anderen Verfahren aufzeigen zu können. Aufgrund von Ergebnissen bisheriger Experimente gelang es auch, ein vorteilhaftes Reaktorkonzept für die Nutzung der BAL aufzustellen, nämlich den *Enzym-Membran-Reaktor* [Küh07]. Die Bestimmung der kinetischen Parameter der Reaktion ist ein weiterer erforderlicher Schritt, um das Potential des Enzyms für die industrielle Nutzung zur Produktion von DHPP zu ermessen.

4.2. Datenlage dieser Arbeit

4.2.1. Versuchsbeschreibung

Die Experimente zur BAL, deren Messdaten Grundlage dieser Arbeit sind (siehe Anhang Teil A), wurden in einem *Batch-Reaktor* (auch *Satz-Reaktor*) durchgeführt. Die Substrate und das Enzym werden dabei mitsamt einem Kaliumphosphatpuffer und einem Lösungsmittel (*Kosolvent*) in ein Becherglas gegeben. Die Reaktion läuft anschließend unter stetigem Rühren ab, bis ihr Gleichgewichtszustand erreicht ist. Im Gegensatz zu einem *Fed-Batch-Reaktor*, bei dem kontinuierlich Substrat zudosiert wird, erfährt das System im Batch-Reaktor nach Einstellung eines Anfangszustandes keine weiteren Einwirkungen von außen. In einem idealen Batch-Reaktor liegt eine vollständige Durchmischung ohne Temperatur- oder Konzentrationsgradienten vor.

Es können während des Reaktionsverlaufs zu jedem Zeitpunkt Proben mit einer Pipette entnommen werden. Mit Hilfe der Hochleistungsflüssigchromatographie (HPLC) werden in den Proben die Konzentrationen der an der Reaktion beteiligten Stoffe gemessen, wobei die Konzentrationen chiraler Stoffe (siehe Kapitel 2) anteilig gemessen werden können. Um die HPLC benutzen zu können, muss zu der Probe

ein Lösungsmittel, in diesem Fall Acetonitril, gegeben werden. Erst dann kann die chromatographische Aufspaltung erfolgen, die für die Bestimmung der in der Probe enthaltenen Stoffkonzentrationen notwendig ist. Das Kosolvent Dimethylsulfoxid (DMSO) wird dieser Reaktion zugegeben, da eines der Produkte, nämlich Benzoin, in Wasser schwer löslich ist, aber nur in Lösung durch die BAL umgesetzt werden kann. Leider wirkt DMSO sich negativ auf die Enzymstabilität aus [KBB+04].

Die Wahl eines geeigneten pH-Werts für die Reaktionsumgebung, der durch eine Pufferlösung eingestellt wird, ist ein Abwägen zwischen hoher Enzymaktivität (optimale Aktivität ungefähr bei pH 8.75) und hoher Enzymstabilität (Optimum ungefähr bei pH 7). Bei den hier betrachteten Experimenten lag der pH-Wert der Reaktionsumgebung bei pH=8. Die Zusammensetzung des Puffers, die Konzentration des Kosolvents sowie die Einstellungen der HPLC sind [Küh07] zu entnehmen.

4.2.2. Der Datensatz und seine Tücken

Bei den für die Modellierung verfügbaren Daten handelt es sich um die Messungen aus drei Batchversuchen, wobei zu Beginn eine variierende Menge BA, eine durch ungenaue Einwaage unabsichtlich abweichende Menge des Enzyms BAL und Dimethoxyaldehyd (DALD) im Überschuss (100 mM) in das Becherglas gegeben wurden (siehe Tabelle 4.1). Der Überschuss an DALD war nötig, um das thermodynamische Gleichgewicht der Reaktion auf die Produktion von DHPP zu legen [Küh07].

	Anfangskonz. BA [mM]	Anfangskonz. BAL [mg/ml]
Batch A	56.0	0.22
Batch B	31.5	0.29
Batch C	18.7	0.20

Tabelle 4.1.: Anfangskonzentrationen der Batch-Experimente

Die Konzentrationen jeweils von BA, BZ, DHPP sowie DALD wurden kontinuierlich zu verschiedenen Zeitpunkten gemessen (nach 0, 16, 21, 30, 45, 60, 90, 120, 150, 180, 210, 240, 300 Minuten). Eine zusätzliche Konzentrationsmessung fand nach 1325 Minuten statt, um zu überprüfen, ob der Gleichgewichtszustand der Reaktion nach Abschluss der vorherigen Messungen wirklich schon erreicht war. Die Ergebnisse dieser Zusatzmessung wurden aber im Folgenden nicht miteinbezogen, da die numerische Lösung der DGL-Systeme der mathematischen Modelle zu der enzymatischen Reaktion sonst erheblich zeitraubender gewesen wäre.

Ebenfalls nicht miteinbezogen wurde der Konzentrationsverlauf des DALD, da die Messungen starken Schwankungen unterlagen und von den Experimentatoren selbst als sehr unsicher eingestuft wurden. Es ist erst möglich, die Konzentration des DALD in der Probe zu messen, wenn es zu einem anderen Molekül weiterreagiert. Für

diese Reaktion wird die Lösung über eine Zeitdauer von 20 Minuten auf 100 °C erhitzt und auf diese aufwändige Prozedur sind die Messschwankungen vermutlich zurückzuführen. Das bedeutet, dass der statistischen Analyse insgesamt $3 \cdot 3 \cdot 13 = 117$ Datenpunkte zur Verfügung standen.

Die Messung der Benzoinkonzentration war trotz Zugabe des Lösungsmittels DMSO durch ausfallende Benzoinflocken (*Agglomerationen*) erschwert. Diese Agglomerationen waren während der Reaktion chemisch inaktiv, wurden aber durch Zugabe von Acetonitril, das zur chromatographischen Messung nötig ist, wieder gelöst und daher in die Konzentrationsmessung miteinbezogen. Die wahre Benzoinkonzentration im Reaktionsverlauf ist also in einem unbekannten Ausmaß kleiner als die gemessene. HPLC-Messungen niedrigerer Konzentrationen unterliegen zudem einer Sensibilitätsgrenze des Geräts. Laut Aussagen der Experimentatoren können Konzentrationen von etwa 0.05 mM noch detektiert werden. Natürliche Unregelmäßigkeiten bei Pipettenvolumina stellen einen weiteren Faktor für Ungenauigkeiten in Messungen dar.

Jeder Batchversuch wurde nur einmal angesetzt und der Konzentrationsverlauf der Reaktanden über die Zeit gemessen. Diese Experimentalstrategie macht es schwer, die Daten statistisch zu untersuchen, da es für keinen Datenpunkt Referenzpunkte gibt. Schon eine einzelne Wiederholung des Versuchs hätte Erkenntnisse über Messgenauigkeit beziehungsweise Varianzen der Messwerte geben können. In dieser Arbeit wurden von den Experimentatoren geschätzte Werte für die Messwertvarianzen in die modellgestützte Datenauswertung implementiert.

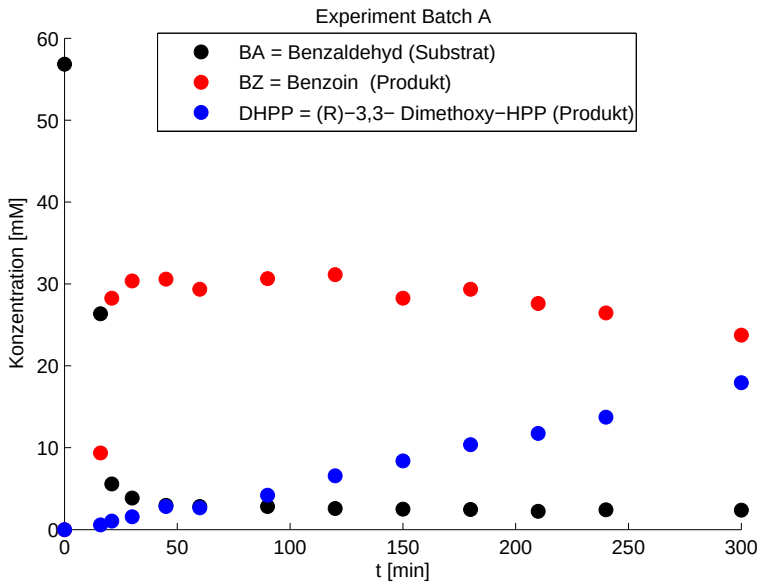

Abbildung 4.1.: Experimentalverlauf Batch A

Abbildung 4.1 zeigt ein Beispiel für die graphische Aufbereitung der Messdaten. In ihr wird der der zeitliche Verlauf der Reaktion in Batch A dargestellt. Es ist deutlich

zu erkennen, dass BA in der katalytischen Umsetzung verbraucht wird und BZ und DHPP als Produkte entstehen. Um ein Modell für die Gesamtreaktion erstellen zu können, wurden nun Hypothesen beziehungsweise Reaktionsschemata für alle Teilreaktionen der beteiligten Stoffe und deren zeitliche Abfolge formuliert.

4.3. Hypothesen zum Reaktionsverlauf

4.3.1. Reaktionsschemata

Um die Zusammenhänge einer kinetischen Reaktion mathematisch zu erfassen, kann man aus den einzelnen Schritten der Reaktion und der Annahme, dass die Abspaltung des Produktes vom Enzym der geschwindigkeitsbestimmende Schritt ist, ein DGL-System aufstellen, wie in Abschnitt 2.2.1 beschrieben wurde. Das bedeutet, dass bekannt sein muss, welche Reaktionsteilnehmer mit dem Enzym wechselwirken, welche Produkte dabei entstehen und überdies, in welcher Reihenfolge die einzelnen Schritte ablaufen. Sollten Inhibitoren oder Aktivatoren an dem System beteiligt sein, so ist es nötig, ihre Konzentration sowie die Art der Inhibierung oder Aktivierung zu kennen, um die Formulierung des DGL-Systems entsprechend anzupassen.

Bei der in dieser Arbeit betrachteten enzymatischen Umsetzung des Benzaldehyds (BA) zum Endprodukt (R)-3,3-Dimethoxy-1-phenyl-2-hydroxypropanon (DHPP) sind der genaue Mechanismus und die Abfolge der Reaktion unbekannt. Es gibt bisher nur einige Hypothesen zum Reaktionsverlauf, die sich auf die Funktionsweise anderer Enzyme stützen, die ThDP als Kofaktor haben und mit denen die BAL deshalb verglichen werden kann [DEH+02].

In den Abbildungen 4.2 [Reaktionsschema A] und 4.3 [Reaktionsschema B] sind zwei aus diesen Hypothesen hergeleitete Reaktionsschemata dargestellt. Sie unterscheiden sich im Reaktionsprodukt der Umsetzung des Benzoins (BZ) durch die BAL. Während in Reaktionsschema A ein BZ-Molekül mit zwei DALD-Molekülen zu zwei DHPP-Molekülen reagiert, entsteht in Reaktionsschema B aus der Reaktion eines BZ-Moleküls mit einem DALD-Molekül jeweil ein BA- und ein DHPP-Molekül. Das BA-Molekül kann hier wieder als Substrat für eine weitere Umsetzung zu Benzoin in die Reaktion eingehen (gestrichelter Pfeil).

Zwar stellen die Pfeile in den Reaktionsschemata alle vorstellbaren Reaktionen dar, aber es ist unbekannt, welche von ihnen tatsächlich stattfinden. Die BAL kann theoretisch alle aufgezeichneten Reaktionen katalysieren. Außerdem ist, bis auf Hinweise aus einigen Messungen außerhalb des Rahmens dieser Arbeit, die auf eine Inhibierung der BAL durch das DALD hinweisen [Küh07], unbekannt, ob und von welchen Reaktionsteilnehmern die BAL aktiviert oder inhibiert wird.

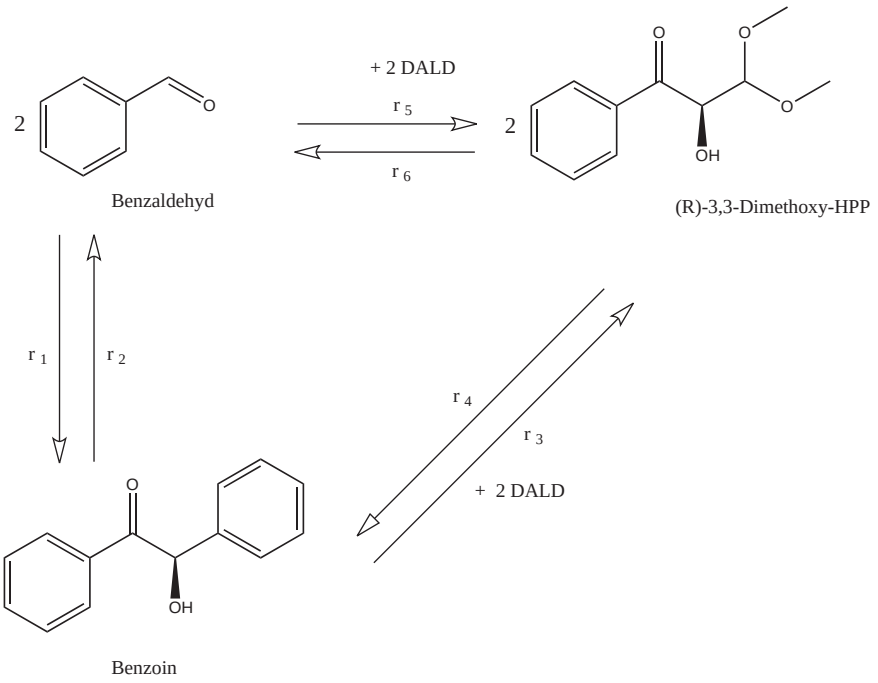

Abbildung 4.2.: Erste Hypothese für die Reaktion von Benzaldehyd zu DHPP [Reaktionsschema A]

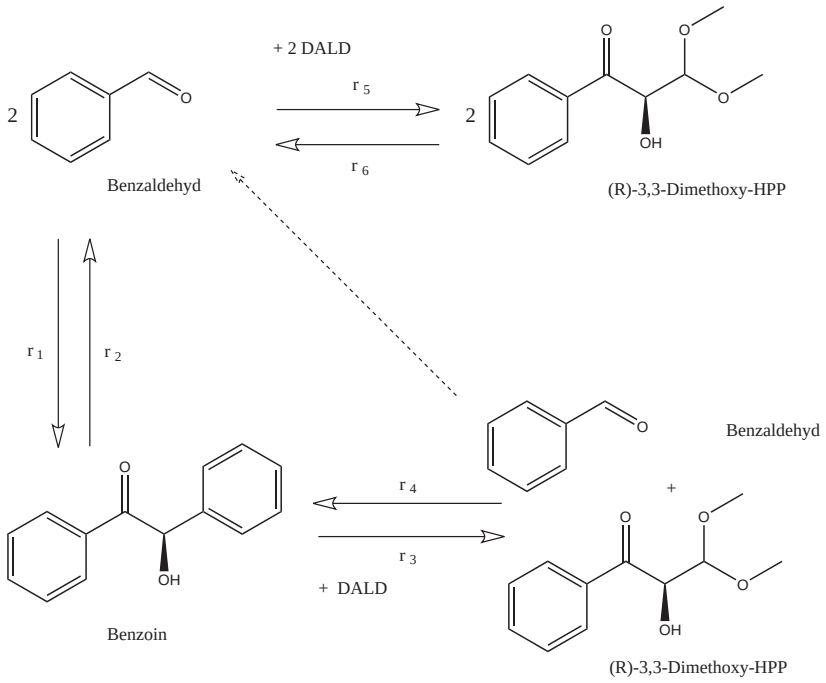

Abbildung 4.3.: Zweite Hypothese für die Reaktion von Benzaldehyd zu DHPP [Reaktionsschema B]

Für die Formulierung eines mathematischen Modells, das die Reaktion beschreibt, ist es nötig, sich für einige Reaktionsschritte zu entscheiden, um die Anzahl der Modellparameter zu begrenzen und das Modell insgesamt nur so komplex wie nötig zu wählen. Für jeden in dem Modell beachteten Teilschritt der Reaktion kommen unter Annahme einer Michaelis-Menten-Kinetik mindestens zwei unbekannte Parameter zu dem Modell hinzu.

4.3.2. Diskussion eines Modellvorschlags aus der Literatur

An dieser Stelle soll das bisher einzige in der Literatur beschriebene DGL-Modell für die von der BAL katalysierte Reaktion von BA zu DHPP dargestellt werden. Es enthält einige Schritte der Gesamtreaktion und wurde in einer einfachen, sowie in einer um einige Gleichungen und Parameter erweiterten Form mit Hilfe des Programms Scientist® an die Messdaten aus den drei Batchversuchen (siehe Abschnitt 4.2.2) angepasst [Küh07].

Reaktionsschema B (vergl. Abb. 4.3) wurde als Grundlage für dieses Modell ausgewählt. Begründet wurde dies damit, dass alle an einer Umsetzung beteiligten Moleküle mit dem Enzym zusammentreffen müssen, bevor eine Reaktion stattfinden kann. Deshalb schien es den Experimentatoren intuitiv wahrscheinlicher, dass die in den Abbildungen 4.2 und 4.3 mit r_3 bezeichnete Umsetzung mit nur zwei Molekülen (BZ und DALD) stattfindet, anstatt mit dreien (BZ und zwei Moleküle DALD).

Der Modellvorschlag lautet [Küh07]:

$$\begin{aligned}
\frac{d\ c_{BA}}{dt} &= -2r_1 + r_3 \\
\frac{d\ c_{BZ}}{dt} &= r_1 - r_3[-r_7] \\
\frac{d\ c_{DHPP}}{dt} &= r_3 \\
\frac{d\ c_{DALD}}{dt} &= -r_3
\end{aligned}$$

mit der potentiellen Ergänzung folgender Gleichungen:

$$\begin{aligned}
\frac{d\ c_{BZ_c}}{dt} &= r_7 \\
\frac{d\ c_{E_{act}}}{dt} &= -r_8 \\
\frac{d\ c_{E_{deact}}}{dt} &= r_8
\end{aligned}$$

und

$$\begin{aligned}
r_1 &= E \cdot V_{max,BABZ} \cdot \left(\frac{c_{BA}}{K_{m,BABZ}\,(1 + \frac{c_{DALD}}{K_{i,DALD}}) + c_{BA}} \right)^2 \\
r_3 &= E \cdot V_{max,BZDHPP} \cdot \frac{c_{BZ}}{K_{m,BZDHPP}\,(1 + \frac{c_{DALD}}{K_{i,DALD}} + c_{BZ})} \\
r_7 &= k_c \cdot c_{BZ} - k_s \cdot c_{BZ_c} \\
r_8 &= k_{act} \cdot E_{act} - k_{deact} \cdot E_{deact}.
\end{aligned}$$

In diesem Modellvorschlag bezeichnet c_X die zeitabhängige Konzentration des Stoffes X,

E die Anfangskonzentration der BAL (gesetzt als feste Konstante E = 0.25 mg/ml) und $K_{i,DALD}$ die Inhibierungskonstante für eine Inhibition der BAL durch das DALD. Die kinetischen Parameter V_{max} und K_m sind wie in Abschnitt 2.2.1 erklärt.

Die Erweiterung des Modells um Schritt r_7 soll zusätzlich die Kristallisation des Benzoins berücksichtigen. Dabei stellen die unbekannten Parameter k_s und k_c eine Solubilisierungskonstante und eine Kristallisationskonstante dar. Für die Konzentration des kristallinen Benzoins c_{BZ_c} waren allerdings keine Messungen vorhanden. Der Konzentrationsverlauf wurde durch die Anpassung des Modells an die experimentellen Daten mitbestimmt.

Eine weitere Ergänzung des Modells (r_8) unterscheidet zwischen dem Enzym im aktiven Zustand (E_{act}) und im inaktiven Zustand (E_{deact}). Zu diesen beiden Zuständen wurden die unbekannten Parameter k_{act} und k_{deact} mit der Einheit min^{-1} in das System eingeführt. Es gibt ebenfalls keine Messreihen zu den Konzentrationen der Enzymzustände E_{deact} und E_{act}, deshalb wurde der Konzentrationsverlauf wieder durch Anpassung des Modells an die Daten bestimmt. Die für die Parameterschätzung verwendete Datenreihe zum Konzentrationsverlauf des DALD stammte ebenfalls nicht aus experimentellen Messungen, sondern wurde vorher mit Hilfe der angenommenen Michaelis-Menten-Kinetik simuliert und dann zum Anpassen verwendet.

Dieses DGL-Modell hat in der einfachen Formulierung fünf beziehungsweise in der erweiterten Formulierung neun Parameter. Die sich aus einer Anpassung mit Hilfe des Programms Scientist® ergebenden Werte für die Parameter, sowohl für das einfache als noch viel mehr für das erweiterte Modell, sind aufgrund der vielen Annahmen, die zu ihrer Berechnung getroffen wurden, problematisch. Vor allem die Annahmen bezüglich der Konzentrationsverläufe von DALD, des kristallinen Benzoins und der beiden Enzymzustände, die nicht durch Daten gestützt werden und das System nur mit weiteren Unbekannten belasten, machen die Parameterschätzung fragwürdig.

Aus diesen Gründen war es eines der vorrangigen Ziele dieser Arbeit, auf Basis der Reaktionsschemata DGL-Modelle aufzustellen, die das System gut beschreiben aber nicht komplexer als notwendig sind. Die Schätzung der Parameter der Modelle sollte nachvollziehbar sein und keinen übermäßig einschränkenden Annahmen unterliegen, sowie nur auf vorhandenen Datenreihen beruhen. Die Formulierung dieser Modelle wird im folgenden Abschnitt dargestellt.

4.3.3. Abgeleitete Grundmodelle dieser Arbeit

> If a particular model (parametrization) does not make biological sense, this is reason to exclude it from the set of candidate models, particularly in case where causation is of interest [BA02].

Eines der Ziele dieser Arbeit war, mathematische Modelle für die BAL-katalysierte Reaktion des BA zum DHPP aufzustellen. Um den Modellen biologischen Sinn zu verleihen, sollte sich ihre Formulierung an bestehenden Modellen für Reaktionskinetiken orientieren. Aus der begrenzten Anzahl von 117 Messpunkten (vergl. Abschnitt 4.2.2) und dem Kriterium der Parametersparsamkeit (principle of parsimony [BJ76]) heraus wurde angestrebt, die betrachteten Teilreaktionen der Katalyse auf ein nötiges Mindestmaß zu beschränken.

Aus den beiden Reaktionsschemata A und B in den Abbildungen 4.2 und 4.3 ergibt sich, dass jeweils die Schritte r_1 und r_3 ausreichen, um eine Reaktion zu beschreiben, bei der BA zu DHPP umgesetzt wird. BZ ensteht bei dieser Reaktion als Zwischenprodukt. Nach einer Reduzierung der Reaktionsschemata A und B auf diese beiden Schritte konnten die beiden sich im Produkt von Reaktionsschritt r_3 unterscheidenden DGL-Modelle Modell 1.1 und Modell 2.1 für die katalysierte Reaktion von BA zu DHPP hergeleitet werden. Auf die Einfügung von Inhibierungen oder Kristallisationsparametern in die Reaktionsmodelle wurde vorerst verzichtet, ebenso wie auf die Verwendung komplexerer Kinetiken als die der Michaelis-Menten-Kinetik.

Die Formulierung der beiden auf zwei Schritte reduzierten Reaktionsschemata als Differentialgleichungsmodelle mit Michaelis-Menten-Kinetik unter Beachtung der Reaktion von BA zu BZ und der Folgereaktion zu DHPP lautet:

Modell 1.1 [Reaktionsschema A]

$$\frac{d\, c_{BA}}{dt} = -2r_1 \tag{4.1}$$

$$\frac{d\, c_{BZ}}{dt} = r_1 - r_3 \tag{4.2}$$

$$\frac{d\, c_{DHPP}}{dt} = 2r_3 \tag{4.3}$$

Modell 2.1 [Reaktionsschema B]

$$\frac{d\ c_{BA}}{dt} = -2r_1 + r_3 \tag{4.4}$$

$$\frac{d\ c_{BZ}}{dt} = r_1 - r_3 \tag{4.5}$$

$$\frac{d\ c_{DHPP}}{dt} = r_3 \tag{4.6}$$

mit

$$r_1 = V_{max,BABZ} * \left(\frac{c_{BA}}{K_{m,BABZ} + c_{BA}}\right)^2 \tag{4.7}$$

$$r_3 = V_{max,BZDHPP} * \frac{c_{BZ}}{K_{m,BZDHPP} + c_{BZ}} \tag{4.8}$$

Dabei ist in den unbekannten Parametern $V_{max,...}$ die bekannte Enzymkonzentration als Faktor mit inbegriffen. Es gilt also für den Wert eines um die Enzymkonzentration c_{E0} bereinigten $V_{max,ber}$:

$$V_{max,ber} = \frac{V_{max,...}}{c_{E0}}$$

$V_{max,ber}$ wird in der Einheit der Enzymaktivität pro Gewicht (U mg^{-1}) angegeben und die K_M-Werte mit der Einheit M (Molarität). Die Quadrierung des zweiten Terms in Gleichung 4.7 beruht auf der Tatsache, dass zwei Moleküle BA zu einem Molekül Benzoin umgesetzt werden.

Diese Formulierung dieser beiden Grundmodelle erlaubt es, auf einfache Weise weitere Teilschritte der Reaktion oder Inhibitionen zu ergänzen und zu überprüfen, inwiefern diese zur Verbesserung der Modellgüte beitragen. Die Ergänzung von weiteren Reaktionsschritten war insofern von Interesse, weil unter anderem aufgrund der Datenbasis geklärt werden sollte, ob Schritt r_5 von Bedeutung für die Reaktion ist, beziehungsweise ob diese Teilreaktion überhaupt stattfindet.

Eine Inhibition der BAL durch DALD konnte leider nicht ergänzt werden, da zu der Schätzung des unbekannten Inhibitionsparameters Messwerte für die DALD-Konzentration im Reaktionsverlauf nötig gewesen wären. Diese wurden allerdings von vorneherein aufgrund ihrer starken Schwankungen vernachlässigt (vergl. Abschnitt 4.2.2). Deshalb war es nur möglich, Substrat- und Produktinhibierungen zu ergänzen.

Für den Einbau einer solchen Inhibition wird im Nenner der Michaelis-Menten Kinetik der Faktor

$$\left(1 + \frac{c_{Inh}}{K_{I,Inh}}\right)$$

mit dem K_M-Wert multipliziert, wobei c_{Inh} die Konzentration des Inhibitors und $K_{I,Inh}$ die unbekannte Inhibitionskonstante für diese Reaktion bezeichnen.

4.3.4. Zusammenfassung

Der Ansatz, nur wirklich nötige Teilschritte der Reaktion in das Modell aufzunehmen und von stark auf das Wesentliche reduzierten DGL-Modellen auszugehen, unterscheidet sich wesentlich von dem in Abschnitt 4.3.2 vorgestellten Modellansatz, der in der dort beschriebenen Weise an dem Vorhaben scheitert, von vorneherein möglichst viele Parameter, Beobachtungen und Annahmen einzubeziehen.

Die Formulierung der Differentialgleichungsmodelle Modell 1.1 und Modell 2.1 (siehe Seite 55) war die Basis aller folgenden Berechnungen. Auf diese beiden Modelle sowie ihre Ergänzungen durch weitere Teilreaktionen oder Inhibitionen konnten die in Kapitel 3 beschriebenen Methoden zur Parameterschätzung und Modelldiskriminierung angewendet werden. Durch die Modelldiskriminierung sollte es dann auch ermöglicht werden, von den beiden Reaktionsschemata A und B (vergl. Abb. 4.2 und 4.3) das wahrscheinlichere zu bestimmen.

5. Ergebnisse

5.1. Implementierung

Die Implementierung der verschiedenen mathematischen Methoden zur Parameterschätzung, Modelldiskriminierung und optimalen Versuchsplanung erfolgte unter MatLab®. Dabei wurden dort vorhandene Optimierungsroutinen und Routinen zur numerischen Lösung von Differentialgleichungssystemen (DGL-Löser) verwendet.

Die Programmierung und die Rechenzeiten wurden wesentlich dadurch bestimmt, dass keine explizite Gleichung für das nichtlineare Modell gegeben war, sondern alle Ergebnisse nur durch wiederholtes Anwenden des DGL-Lösers erhalten werden konnten. Ein Optimierungsvorgang, bei dem Parameter eines Modells geschätzt wurden, dauerte etwa vier bis zehn Sekunden. Dabei wurde das DGL-System mehrfach gelöst; in der Optimierung, die im Anhang in Abschnitt D dargestellt ist, beispielsweise 160-mal. Die mathematischen Methoden, bei denen für tausend künstliche erzeugte Datensätze jeweils Parameterschätzungen und Kovarianzmatrizen berechnet wurden (MCData und SIMUL, siehe Abschnitt 3.6.5), benötigten mehrere Stunden Rechenzeit.

Eine korrekte Implementierung ist ausschlaggebend für vertrauenswürdige Ergebnisse, mit denen man wissenschaftlich argumentieren kann. Deshalb wurde viel Wert auf stetes Überprüfen des Programmcodes gelegt. Es wurden keine Programme aus anderen Quellen im Rahmen dieser Arbeit verwendet.

5.2. Parameterschätzung

Alle in dieser Arbeit untersuchten Differentialgleichungsmodelle für die von der BAL katalysierte Reaktion von BA zu DHPP enthielten mindestens vier unbekannte kinetische Parameter. Diese unbekannten Parameter, zusammengefasst in einem Vektor $\theta = (V_{max,BABZ}; K_{M,BABZ}; \dots)$, wurden mit den vier in Kapitel 3 allgemein beschriebenen Methoden auf Basis der aus drei Batchversuchen erhaltenen Datensätze (siehe Abschnitt 4.2.2) geschätzt. Die Schätzmethoden haben verschiedene mathematische Parameter, die vor der konkreten Anwendung spezifiziert werden

müssen. Die Methoden und die Konfiguration dieser Parameter werden im Folgenden aufgeführt.

- **Methode der gewichteten kleinsten Quadrate (gKQSS)**
  Das Grundrauschen der Messdaten in dem für diese Methode betrachteten Regressionsmodell (siehe Abschnitt 3.5) wurde auf $\varepsilon_{Gij} = 0.1$ mM gesetzt. Das ist etwas höher als die von den Experimentatoren prognostizierte Sensibilitätsgrenze des verwendeten Messgeräts von etwa 0.5 mM (siehe Abschnitt 4.2.2). Der Wert für ε_{Gij} wurde auf 0.1 mM gesetzt, um sicherzugehen, dass die wahre Sensibilitätsgrenze nicht höher ist, als die für das Modell angenommene. Die Gewichte der gKQSS wurden entsprechend der in Abschnitt 3.6.2 beschriebenen Methode berechnet.

- **Least-Trimmed-Squares-Methode (LTS)**
  Diese Methode wurde für k=2 verwendet, das bedeutet, dass die zwei größten Residuen (Abstände zwischen den Messwerten und den durch das Modell simulierten Werten) nicht in die Schätzung des Parametersatzes miteinbezogen wurden.

- **Methode des parametrischen Bootstrap (MCData)**
  Als Ausgangsparameter, mit dem tausend künstliche Datensätze für jedes Modell erzeugt wurden, wurde immer $\hat{\theta}_{KQSS}$ verwendet, also die jeweilige Schätzung des Parametersatzes mit der gKQSS für das betrachtete Modell. Für diese künstlichen Datensätze wurden wiederum Parameter geschätzt. Das Ergebnis der Schätzung mit dieser Methode ist der Mittelwert aus den tausend Parameterschätzungen.

  Die Anzahl der simulierten Datensätze wurde auf tausend festgelegt, da gezeigt werden konnte, dass die Methode innerhalb dieses Rahmens konvergiert (vergl. Abschnitt 5.2.2).

- **Methode des parametrischen Bootstrap mit verrauschtem Anfangsparameter (SIMUL)**
  Der Ausgangsparameter wurde wie bei Methode MCData gewählt und vor dem Simulieren neuer Datensätze verrauscht, wie in Abschnitt 3.6.5 beschrieben. Danach wurde analog zu Methode MCData vorgegangen.

Die Ergebnisse der Parameterschätzungen mit diesen vier Methoden für die beiden Grundmodelle (Modell 1.1 und Modell 2.1, siehe Seite 56) finden sich im Anhang (Teil B) in den Tabellen B.1 und B.2 und sind in den Abbildungen 5.1 und 5.2 dargestellt. Es wird darauf hingewiesen, dass die in diesem Kapitel aufgeführten Schätzungen für die V_{max}-Werte alle noch die in den Batch-Experimenten eingesetzte

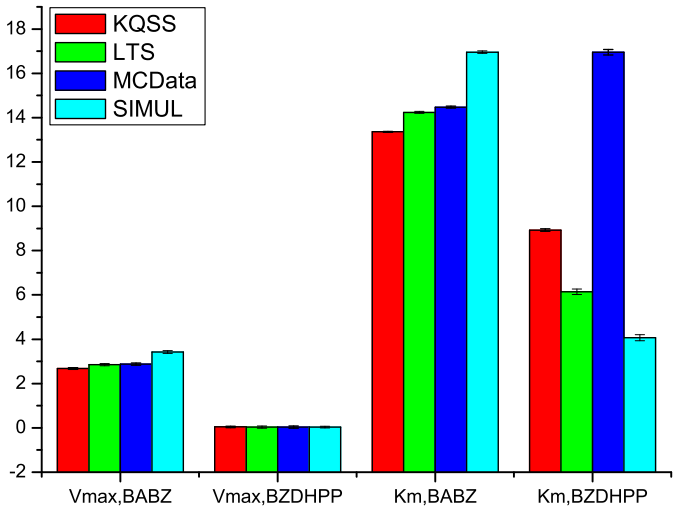

Abbildung 5.1.: Parameterschätzungen für Modell 1.1 im Vergleich

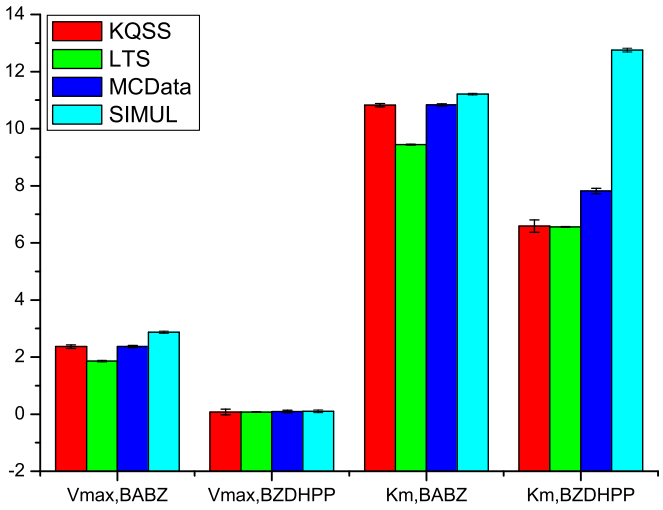

Abbildung 5.2.: Parameterschätzungen für Modell 2.1 im Vergleich

Enzymkonzentration als Faktor beinhalten, die bei etwa 0.2 mg ml^{-1} lag (siehe Tabelle 4.1 und Abschnitt 4.3.3).

Man erkennt in den Abbildungen 5.1 und 5.2, dass die Werte der Parameterschätzungen zu den jeweiligen Modellen sich sehr ähnlich sind, unabhängig davon, mit welcher Methode sie geschätzt wurden. Das bedeutet, dass keine Schätzung von den anderen stark abweichende Werte ergibt. Die Parameter haben kleine Standardabweichungen, sind also sehr gut bestimmt.

Abbildung 5.3 zeigt das Ergebnis der Anpassung von Modell 2.1 an die Daten von Batchversuch A mit dem Parametersatz ${}_{21}\hat{\theta}_{KQSS}$ (Die Zahl links von θ gibt das Modell, die Abkürzung rechts die Schätzmethode an). Die Anpassungen an die

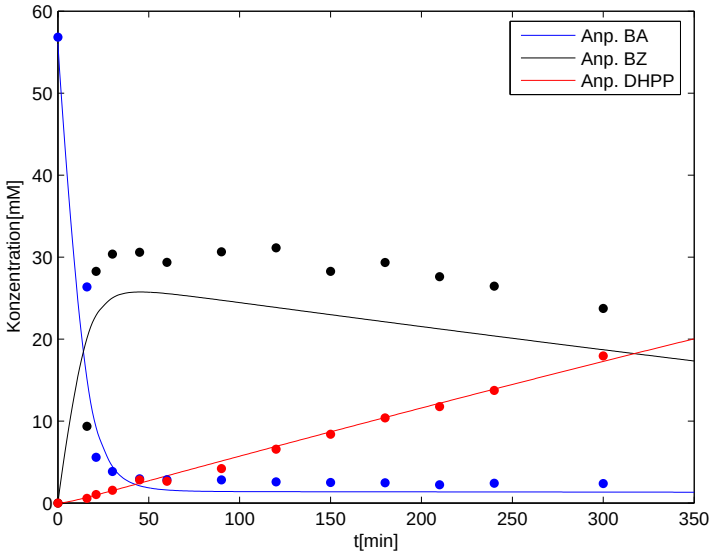

Abbildung 5.3.: Modellanpassung an die Daten aus Batch A

anderen beiden Batch-Datensätze sehen ähnlich aus.

5.2.1. Vergleich der auf zwei Arten geschätzten Kovarianzmatrizen

Die Standardabweichungen der Parameterschätzungen in den Tabellen B.1 und B.2 sind Kovarianzmatrizen entnommen, die auf Grundlage von Satz 3.3 in Kapitel 3 berechnet wurden. Da in Satz 3.3 zwei Möglichkeiten der Berechnung angeführt werden, nämlich die Formeln (3.17) und (3.18), wurden beide im Vorhinein verglichen. Der Vergleich fand für die Kovarianzmatrizen zu Parametersatz ${}_{21}\hat{\theta}_{KQSS}$ statt. Die Elemente des Störvektors $h \in \mathbb{R}^p$ der zur Berechnung der Kovarianzmatrizen angewendeten numerischen Differentiation (siehe Abschnitt 3.7.1) wurden auf h_j=0.3 für $j = 1, \ldots, p$ gesetzt, da sich die Berechnungen der Kovarianzmatrix für ${}_{21}\hat{\theta}_{KQSS}$ um h_j=0.3 als numerisch stabil erwiesen.

Es wurden zwei Beobachtungen gemacht:

- Zur Berechnung der in Formel (3.17) benötigten Output-Sensitivität musste der Optimierungsalgorithmus nur achtmal neu gestartet werden, im Vergleich zu 234-mal für die Berechnung der in Formel (3.18) verwendeten Parameter-Sensitivität (vergl. Abb. 3.3). Das ist eine bedeutende Zeitersparnis, da jede Optimierung etwa vier bis zehn Sekunden dauert.

- Die Berechnung der Kovarianzmatrix auf Basis von Formel (3.17) mit dem oben definierten Vektor h war numerisch stabiler. Wählte man um 0.1 kleinere oder größere Werte für h_j, so erhielt man hier die gleichen Einträge in der

Kovarianzmatrix, während sie bei der anderen Methode bereits Schwankungen in der ersten Nachkommastelle unterlagen.

Der Vergleich der beiden Berechnungsarten führte zu der Entscheidung, für die Berechnung der übrigen Kovarianzmatrizen Formel (3.17) zu verwenden.

5.2.2. Verteilung der Parameterschätzung

Mit den Methoden MCData und SIMUL sollten nicht nur die unbekannten Parameter geschätzt, sondern auch die empirischen Verteilungen der Schätzungen bestimmt werden. Zuerst wurde die Konvergenz der Methoden MCData und SIMUL überprüft. Dies ist in Abbildung 5.4 beispielhaft dargestellt. Sie zeigt die Konvergenz für die Schätzung des Parameters $V_{max,BZDHPP}$, der hier mit der Methode SIMUL für Modell 2.1 geschätzt wurde. Die x-Achse zeigt die Anzahl der in die Schätzung eingehenden Datensätze an, die Schätzung selber ist mit einem Plus gekennzeichnet und die Standardabweichungen in beide Richtungen mit Kreisen.

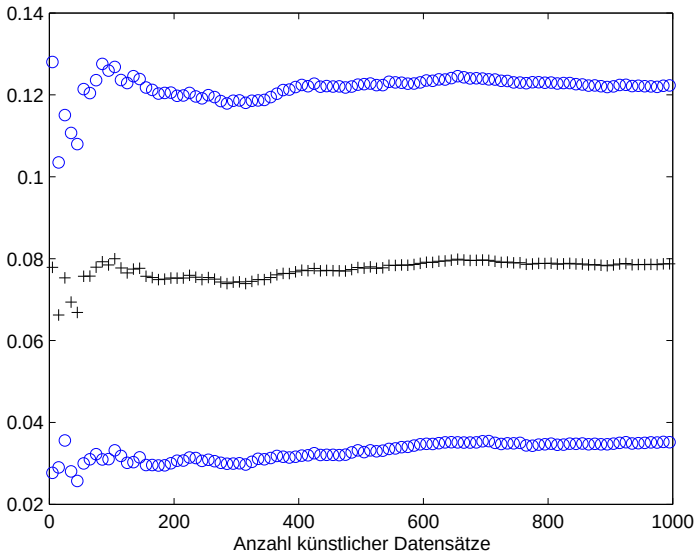

Abbildung 5.4.: Beispiel für die Konvergenz eines Parameters (hier $V_{max,BZDHPP}$ aus Modell 2.1) bei Zunahme der zur Schätzung verwendeten Datensätze

Die Konvergenz stellte sich für alle Parameter spätestens nach etwa 600 bis 800 Simulationsschritten ein. Das bedeutet, dass nach der Erzeugung dieser Anzahl von Datensätzen der Mittelwert der dazu berechneten Parameterschätzungen gegen einen Zielwert konvergierte. Alle folgenden Ergebnisse, die mit Hilfe der Methoden MCData und SIMUL erhalten wurden, basieren auf jeweils tausend künstlich erzeugten Datensätzen, da diese für ein aussagekräftiges Ergebnis ausreichend waren, wie hier gezeigt wurde.

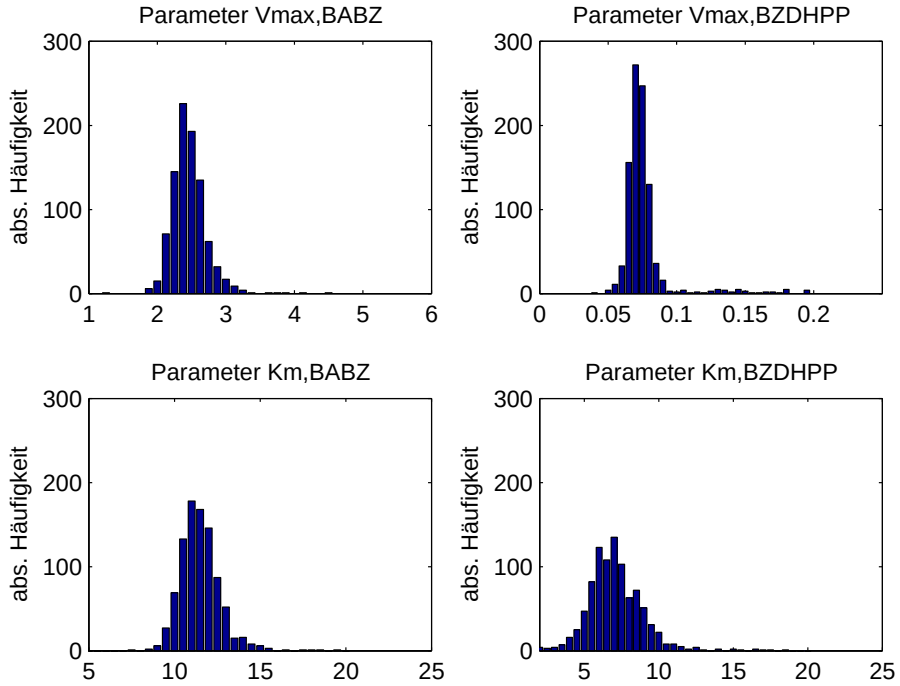

Abbildung 5.5.: Histogramme für die Parameterhäufigkeiten bei der Parameterschätzung mit MCData für Modell 2.1

Aus den jeweils tausend mit den Methoden MCData und SIMUL ermittelten Parametersätzen für die verschiedenen Modelle wurden anschließend empirische Verteilungen aufgezeichnet. Ein Beispiel für die Verteilung der tausend Parameterschätzungen für Modell 2.1 (Methode MCData) zeigt Abbildung 5.5. Die verstreuten Parameterschätzungen an den Rändern der Verteilung sind zu einem großen Teil dadurch zu erklären, dass es sich dabei um Schätzungen aus Optimierungsvorgängen handelt, die in einem Nebenminimum beendet wurden oder aufgrund des Erreichens einer maximalen Anzahl von Iterationen abgebrochen wurden.

Man erkennt in Abb. 5.5 sehr deutlich die glockenförmigen empirischen Verteilungen der einzelnen Parameterschätzungen, die auf Normalverteilung schließen lassen. Die Parameterschätzung für $V_{max,BZDHPP}$ unterliegt der geringsten Streuung, lässt sich also genauer schätzen als die anderen Parameter. Im Gegensatz dazu unterliegt der Parameter $K_{m,BZDHPP}$ der breitesten Streuung.

Diese Ergebnisse sollten mit der anderen Methode zur Bestimmung der Standardabweichungen der Parameter, der in Abschnitt 5.2.1 beschriebenen Berechnung der Kovarianzmatrix vergleichbar sein. Die Ergebnisse dieser Methode, die in den Tabellen B.1 und B.2 als Standardabweichungen hinter den Schätzwerten verzeichnet sind, zeigen tatsächlich das gleiche Streuungsverhalten der Parameter wie Abb. 5.5.

5.2.3. Korrelationen zwischen den Parameterschätzungen

Korrelationen zwischen den Parameterschätzungen konnten beispielsweise für die Parameterschätzung ${}_{21}\hat{\theta}_{KQSS}$ gezeigt werden. Die für diesen Parameter berechnete Kovarianzmatrix ist in Tabelle 5.1, die daraus nach Definition 3.5 resultierende Korrelationsmatrix in Tabelle 5.2 dargestellt.

Cov $({}_{21}\hat{\theta}_{KQSS})$	$V_{max,BABZ}$	$V_{max,BZDHPP}$	$K_{M,BABZ}$	$K_{M,BZDHPP}$
$V_{max,BABZ}$	0.0036703	-0.00032335	0.0031365	-0.00024516
$V_{max,BZDHPP}$	-0.00032335	0.010301	-0.0004015	0.021916
$K_{M,BABZ}$	0.0031365	-0.0004015	0.0028417	-0.00042306
$K_{M,BZDHPP}$	-0.00024516	0.021916	-0.00042306	0.04786

Tabelle 5.1.: Kovarianzmatrix nach Formel (3.17) zu Parametersatz ${}_{21}\hat{\theta}_{KQSS}$

Corr $({}_{21}\hat{\theta}_{KQSS})$	$V_{max,BABZ}$	$V_{max,BZDHPP}$	$K_{M,BABZ}$	$K_{M,BZDHPP}$
$V_{max,BABZ}$	1	-0.05259	0.97119	-0.01850
$V_{max,BZDHPP}$	-0.05259	1	-0.07421	0.98704
$K_{M,BABZ}$	0.97119	-0.07421	1	-0.03628
$K_{M,BZDHPP}$	-0.01850	0.98704	-0.03628	1

Tabelle 5.2.: Korrelationsmatrix nach Formel (3.17) zu Parametersatz ${}_{21}\hat{\theta}_{KQSS}$

In der Korrelationsmatrix sind hohe Werte für die Korrelation der kinetischen Parameter $V_{max,BABZ}$ und $K_{M,BABZ}$ sowie von $V_{max,BZDHPP}$ und $K_{M,BZDHPP}$ angegeben. Für Korrelationsmatrizen auf Basis anderer Schätzmethoden war dies ähnlich. Korrelationen kinetischer Parameter sind nicht durch biologische Gesetzmäßigkeiten erklärlich. Deshalb sollte eines der Ziele der optimalen Versuchsplanung sein, diese Korrelationen zu verringern.

5.3. Modelldiskriminierung

5.3.1. Modellvarianten

Die Diskriminierung der Modelle 1.1 und 2.1 wurde mit Hilfe des Akaike-Kriteriums für kleine Datensätze (AIC_C, siehe Abschnitt 3.9.1) vorgenommen. Die Berechnung der AIC_C-Werte für die in Tabelle B.1 und Tabelle B.2 aufgeführten Parameter, die direkt aus den Datensätzen der Batch-Experimente geschätzt wurden ($\hat{\theta}_{KQSS}$, $\hat{\theta}_{LTS}$), wird in Tabelle 5.3 gezeigt.

Die Werte des AIC_C für Modell 2.1 sind kleiner als die für Modell 1.1. Das bedeutet, dass Modell 2.1 und damit auch das Reaktionsschema B als Grundlage dieses Modells dem anderen Modell/Reaktionsschema A überlegen ist. Allerdings könnte dieser Unterschied der AIC_C-Werte auch durch Messrauschen zustande gekommen sein.

	Modell 1.1	Modell 2.1
$\hat{\theta}_{KQSS}$	229.86	185.35
$\hat{\theta}_{LTS}$	229.02	163.31

Tabelle 5.3.: Vergleich der AIC-Werte der Grundmodelle

Um zu überprüfen, ob der Unterschied zwischen den Modellen vom Messrauschen abhängig ist, wurde mit der Methode MCData eine Häufigkeitsverteilung von tausend AIC_C-Differenzen (AIC_C-Wert Modell 1.1 minus AIC_C-Wert Modell 2.1) berechnet, wie in Abbildung 5.6 gezeigt wird.

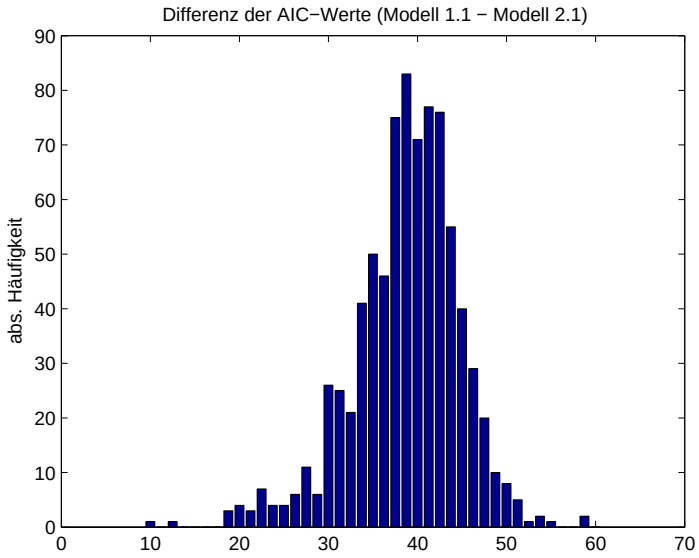

Abbildung 5.6.: Häufigkeitsverteilung der AIC_C-Differenzen (Modell 1.1 und Modell 2.1)

Aus der Verteilung der AIC_C-Differenzen ergibt sich, dass bezüglich des AIC_C und aller simulierter Datensätze Modell 2.1 Modell 1.1 überlegen ist. Messrauschen als Grund für die Unterschiede der Werte in Tabelle 5.3 kann also ausgeschlossen werden. Allerdings ist der Unterschied zwischen den Modellen nicht besonders groß, vor allem, wenn man die relativen Differenzen der AIC_C-Werte betrachtet. Ein weiteres Ziel der optimalen Versuchsplanung sollte deshalb sein, eine bessere Diskriminierung zwischen den beiden Modellen zu ermöglichen.

5.3.2. Modellergänzungen

Wurden zu den Grundmodellen weitere Teilreaktionen und Inhibitionen ergänzt, so konnten bezüglich des AIC_C keine wesentlichen Verbesserungen festgestellt werden. Ein Beispiel dafür ist das Ergebnis des Vergleichs von Modell 2.1 und Modell 2.2, bei dem als zusätzliche Teilreaktion die direkte Umsetzung des BA zu DHPP eingefügt wurde (Schritt r_5):

Modell 2.2:

$$\begin{aligned}
\frac{d\ c_{BA}}{dt} &= -2r_1 + r_3 - r_5 \\
\frac{d\ c_{BZ}}{dt} &= r_1 - r_3 \\
\frac{d\ c_{DHPP}}{dt} &= r_3 + r_5
\end{aligned}$$

mit den Geschwindigkeiten

$$\begin{aligned}
r_1 &= V_{max,BABZ} \cdot \left(\frac{c_{BA}}{K_{M,BABZ} + c_{BA}} \right)^2 \\
r_3 &= V_{max,BZDHPP} \cdot \frac{c_{BZ}}{K_{M,BZDHPP} + c_{BZ}} \\
r_5 &= V_{max,BADHPP} \cdot \frac{c_{BA}}{K_{M,BADHPP} + c_{BA}}
\end{aligned}$$

Das Ergebnis der Parameterschätzung mit den Methoden gKQSS und MCData ist in Tabelle 5.4 verzeichnet.

	${}_{22}\hat{\theta}_{KQSS}$	${}_{22}\hat{\theta}_{MCData}$
$V_{max,BABZ}$	2.4081 ± 0.0415	1.971 ± 0.0159
$V_{max,BZDHPP}$	0.067291 ± 0.0527	0.4132 ± 0.0243
$V_{max,BADHPP}$	0.049587 ± 0.5028	0.3463 ± 0.0666
$K_{M,BABZ}$	11.267 ± 0.0469	4.036 ± 0.0322
$K_{M,BZDHPP}$	6.2519 ± 0.1385	2.961 ± 0.0874
$K_{M,BADHPP}$	32.133 ± 0.9948	31.4331 ± 0.1477
$\sum_{ij} \epsilon_{ij}^2$	527.27	840.096

Tabelle 5.4.: Parameterschätzung zu Modell 2.2

Zum einen kann man hier erkennen, dass für die Schätzung mit der Methode der gewichteten kleinsten Quadrate die Fehlerquadratsumme $\sum_{ij} \epsilon_{ij}^2$ durch die Hinzunahme weiterer Parameter ($V_{max,BADHPP}$, $K_{M,BADHPP}$) im Vergleich zu der Schätzung für die Grundmodelle (Tabellen B.1 und B.2) vermindert wird. Zum anderen unterscheiden sich die Ergebnisse der verschiedenen Schätzmethoden deutlicher. Zur Modelldiskriminierung zwischen Modell 2.2 und 2.1 wurde wiederum das Akaike-Kriterium für kleine Datensätze (AIC_C) verwendet. Die Verteilung der AIC_C-Wert-Differenzen (AIC_C Modell 2.2 minus AIC_C Modell 2.1) wird in Abbildung 5.7 gezeigt.

Die Streuung der Differenzen ist breiter als bei der Diskriminierung der beiden Grundmodelle (vergl. Abb. 5.6). Obwohl für die meisten Datensätze Modell 2.2 Modell 2.1 geringfügig überlegen ist, gibt es auch einen nicht geringen Anteil von

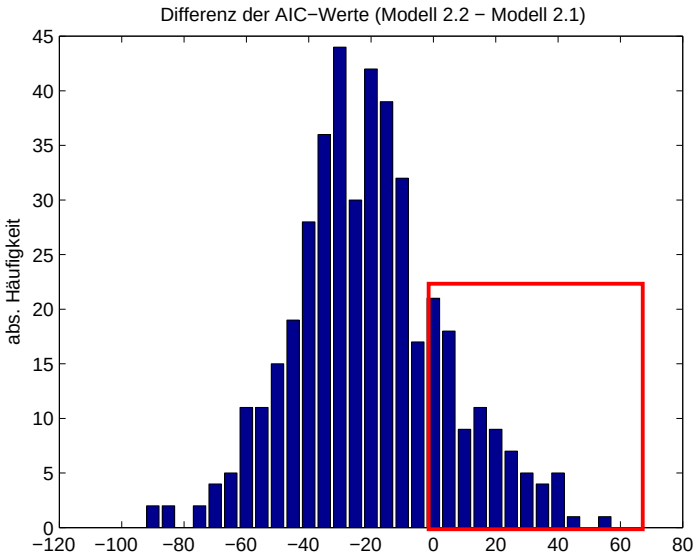

Abbildung 5.7.: Häufigkeitsverteilung der AIC_C-Differenzen (Modell 2.2 und Modell 2.1)

Datensätzen (innerhalb des roten Rechtecks in Abb. 5.7), bei denen es umgekehrt ist. Es kann also auf Basis dieser Datenlage keine eindeutige Verbesserung des Modells durch Hinzunahme des Reaktionsschritts r_5 festgestellt werden.

Dieses Ergebnis hat seine Ursache eventuell in der kleinen Anzahl der Messdaten. Zwar ist die Fehlerquadratsumme durch Hinzufügen weiterer Parameter für Modell 2.2 kleiner als für Modell 2.1, aber in der Struktur des zur Modelldiskriminierung verwendetem AIC_C wiegt der Strafterm, der für den Einsatz zusätzlicher Parameter zu dem Kriterium addiert wird, dies auf (siehe Definition des AIC_C 3.9.1). Die Messdatenanzahl geht beim AIC_C als Faktor der Fehlerquadratsumme ein. Stünden mehr Daten zur Verfügung, würde dieser Summand stärker gewichtet und die zusätzlichen Parameter würden die AIC_C-Werte nicht so stark erhöhen.

5.4. Optimale Versuchsplanung

In der optimalen Versuchsplanung sollen die Eingangsvariablen des Experiments derart beeinflusst werden, dass für ein bestimmtes Ziel ein Versuchsplan berechnet werden kann, dessen Information maximal ist (zum Prinzip der „Informationsmaximierung“ siehe Abschnitt 3.10). Die Eingangsvariablen, auf die bei dem Experiment zur BAL Einfluss genommen werden konnte, waren die Anfangssubstrat- und die Anfangsenzymkonzentration, sowie die Zeitpunkte der Messungen und die Dauer des Experiments. Da die beiden letzteren von den verwendeten Messinstrumenten und der Auswertungszeit für eine einzelne Messung abhängen, wurden zunächst nur Optimierungen über die beiden Anfangskonzentrationen betrachtet.

5.4.1. D-optimale Versuchsplanung

Wie in Abschnitt 3.10.1 beschrieben, dient ein D-optimaler Versuchsplan der Verkleinerung der Varianzen der Parameterschätzung. Er basiert auf einer Minimierung der Determinante der Kovarianzmatrix des geschätzten Parametersatzes. Die Abbildungen 5.8 und 5.9 zeigen jeweils die Auswirkung der Anfangssubstrat- beziehungsweise der Anfangsenzymkonzentration auf die Determinante der Kovarianzmatrix. Die Sternchen zeigen die Standardabweichung der Prognose an.

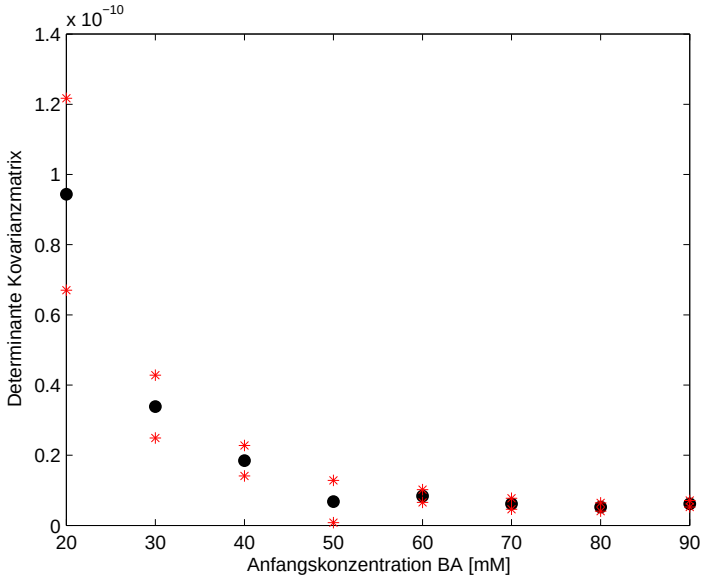

Abbildung 5.8.: Verhältnis zwischen der Anfangssubstratkonzentration und der Determinante der Kovarianzmatrix

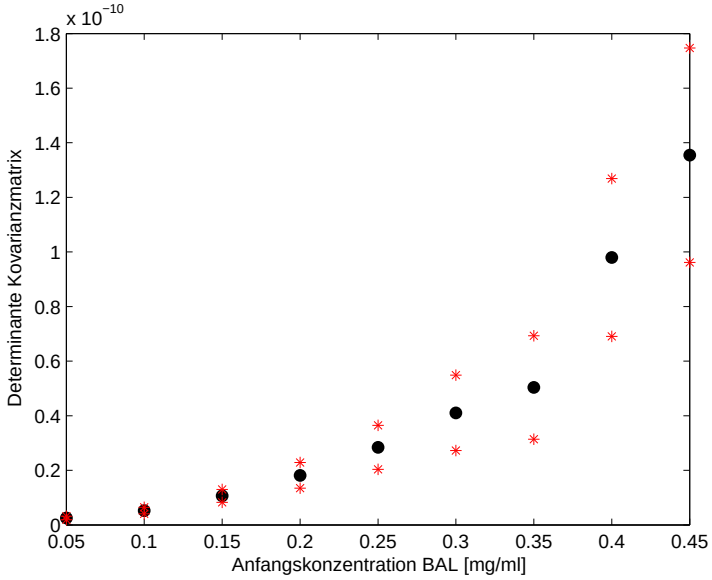

Abbildung 5.9.: Verhältnis zwischen der Anfangsenzymkonzentration und der Determinante der Kovarianzmatrix

Gäbe man also dem Ausgangssystem weniger Enzym zu, oder würde die Substratkonzentration erhöht, so würden sich die Parametervarianzen verringern.

Das Ergebnis lässt sich noch eingängiger mit Hilfe der Konfidenz-Ellipsoide (siehe Abschnitt 3.4) illustrieren. In Abbildung 5.10 wird das Konfidenz-Ellipsoid des Parametersatzes ${}_{21}\hat{\theta}_{KQSS}$ (durchgezogene Linie, Anfangskonzentration BAL 0.2 mg/ml) mit dem eines Parametersatzes verglichen, der aus einem Versuch mit einer Anfangsenzymkonzentration von 0.05 mg/ml, also einem Viertel der ursprünglichen Konzentration, berechnet würde (gestrichelte Linie).

Um das vierdimensionale Konfidenz-Ellipsoid anzeigen zu können, wurde es in verschiedene Ebenen projiziert; die betrachteten Parameter sind jeweils in der Abbildungsüberschrift verzeichnet. Zur besseren Übersichtlichkeit wurden die durch die Projektion entstehenden Ellipsen so transformiert, dass ihre Mittelpunkte im Ursprung liegen und die Projektion des Ellipsoids auf eine der Parameterachsen die relative Abweichung des Parameters in Prozent angibt. Man kann deutlich erken-

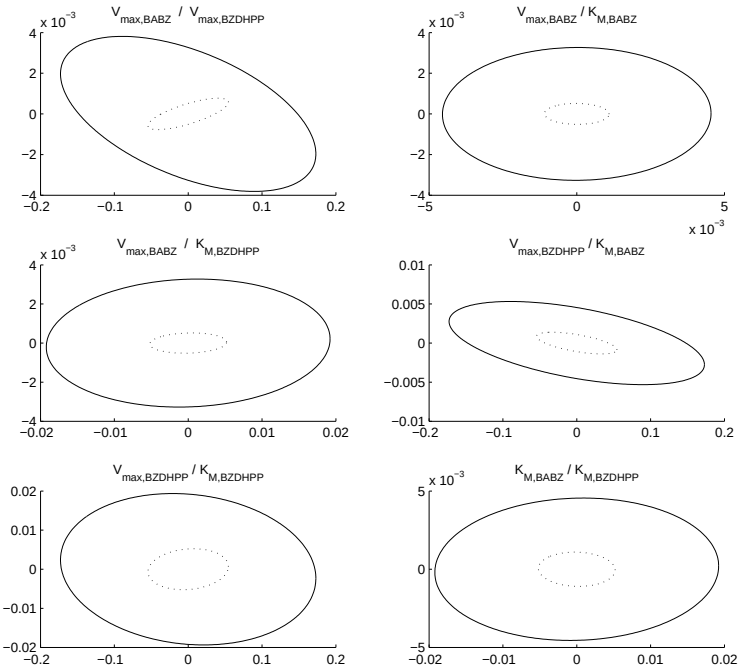

Abbildung 5.10.: Veränderung des Konfidenz-Ellipsoids durch Versuchsplanung

nen, dass ein Versuch, bei dem eine geringere Menge der BAL eingesetzt würde, das Konfidenz-Ellipsoid verkleinern würde, also die Parameterschätzungen kleinere Abweichungen hätten. In Abbildung 5.11 wurden zum Vergleich zusätzlich die Koordinatenachsen für jeden Parameter gleich skaliert. Dadurch erkennt man allerdings für einige Parameter die Verkleinerung der Konfidenz-Ellipsen nicht mehr.

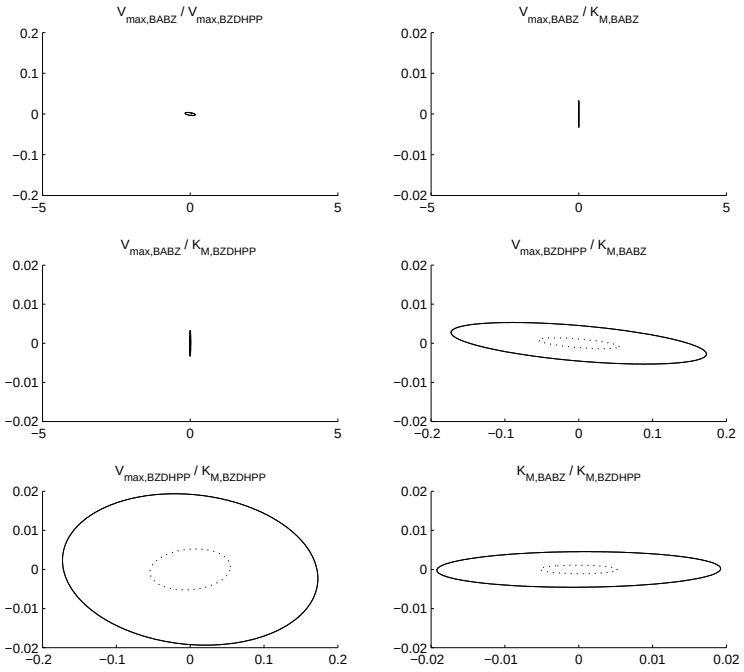

Abbildung 5.11.: Veränderung des Konfidenz-Ellipsoids durch Versuchsplanung (konstante Achsen)

5.4.2. E-optimale Versuchsplanung

Die E-optimale Versuchsplanung wurde zum Zweck der Verminderung der in dieser Arbeit auftretenden Korrelationen zwischen den geschätzten Parametern verwendet (siehe Abschnitt 3.10.2). Sie basiert auf der Minimierung des größten Eigenwerts der Kovarianzmatrix.

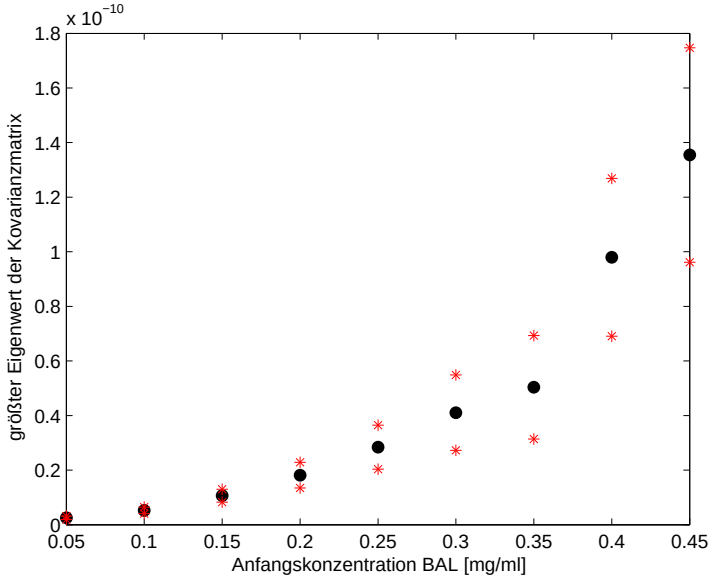

Abbildung 5.12.: Verhältnis zwischen der Anfangsenzymkonzentration und dem größten Eigenwert der Kovarianzmatrix

Die Abhängigkeit des größten Eigenwerts der Kovarianzmatrix von der Enzymkonzentration wird in Abbildung 5.12 dargestellt. Die ausgefüllten Kreise geben den prognostizierten größten Eigenwert für die in der x-Achse angezeigte Konzentration an und die Sternchen die Standardabweichung in beide Richtungen. Es ergibt sich, dass eine Verringerung der Anfangsenzymkonzentration auch die Korrelationen vermindern würde.

5.4.3. Versuchsplanung zur besseren Modelldiskriminierung

Aufgrund des relativ geringen Unterschieds der AIC_C-Werte für Modell 1.1 und Modell 2.1 (vergleiche Abbildung 5.6) sollte ein neues Experiment geplant werden, das es erlauben sollte, die beiden in Abschnitt 4.3.3 vorgeschlagenen Grundmodelle noch besser zu unterscheiden. Dazu wurde die Auswirkung der Anfangsenzym- und der Anfangssubstratkonzentration auf den Abstand zwischen den zu diskriminierenden Modellen untersucht, wobei der Begriff des Abstands in Definition 3.18 erklärt ist.

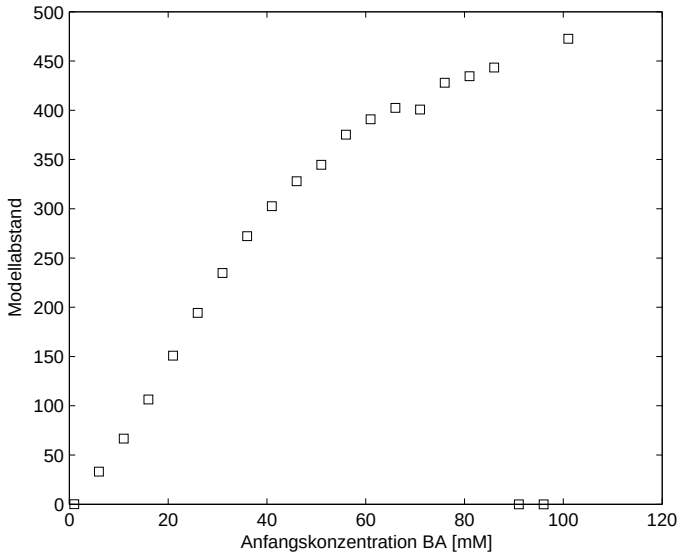

Abbildung 5.13.: Auswirkung der Anfangssubstratkonzentration auf den Abstand zwischen Modell 1.1 und 2.1

Je größer der Abstand ist, desto besser können die Modelle voneinander unterschieden werden. Für die Versuchsplanung wurden die mittels Methode MCData erzeugten Datensätze samt der dazu geschätzten Parametersätze ${}_{11}\hat{\theta}_{KQSS}$ und ${}_{21}\hat{\theta}_{KQSS}$ verwendet. Eines der Modelle musste zum Ausgangsmodell für die Abstandsberechnung bestimmt werden (siehe Definition 3.18). In dieser Arbeit wurde Modell 2.1 als Ausgangsmodell verwendet.

Abbildung 5.13 zeigt, dass der Abstand zwischen den Modellen mit Zunahme der Anfangssubstratkonzentration wüchse. Würde zu Beginn des Experiments weniger BAL eingesetzt, so wüchse der Abstand ebenfalls, wie in Abbildung 5.14 gezeigt ist.

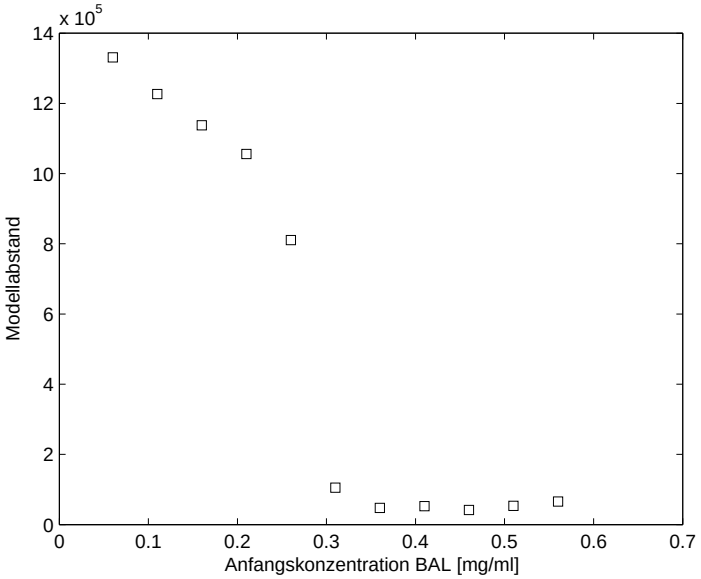

Abbildung 5.14.: Auswirkung der Anfangsenzymkonzentration auf den Abstand zwischen Modell 1.1 und 2.1

5.4.4. Ergebnisse der optimalen Versuchsplanung

Das Ergebnis der Versuchsplanung zu allen drei Zielen ist, dass eine Verringerung der Anfangsenzymkonzentration oder eine Erhöhung der Anfangssubstratkonzentration jeweils zum Ziel führt. Der durch die optimale Versuchsplanung erkannte antagonistische Zusammenhang zwischen diesen beiden Konzentrationen ist dadurch zu erklären, dass das Enzym-Substrat-Verhältnis in der katalysierten Reaktion beeinflusst werden muss. Ist es nicht möglich, mehr Substrat pro Enzym einzusetzen, so tritt der gewünschte Effekt auch ein, wenn weniger Enzym pro Substrat eingesetzt wird.

Es ist tatsächlich der Fall, dass die Anfangssubstratkonzentration des Versuchs nicht erhöht werden kann. Die Vergrößerung der BA-Konzentration zu Beginn des Experiments ist aus Gründen der geringen Löslichkeit des Stoffes nicht möglich. Die in Batch A eingesetzte Konzentration von 60 mM BA ist schon sehr nah an der Löslichkeitsgrenze von etwa 75 mM [HKP$^+$07] gelegen. Die Verringerung der Enzymkonzentration von 0.2 mg/ml auf bis zu 0.05 mg/ml stellt nach Aussage der Experimentatoren kein Problem dar und böte nach diesen Berechnungen eine gute Grundlage für ein neues Experiment.

5.5. Optimierung der Messzeitpunkte

Die Wahl der Messzeitpunkte ist eine weitere Eingangsvariable des Experiments. Sie verändert zwar nichts an seinem Verlauf, ist aber trotzdem entscheidend für die Auswertung der Messdaten. Bei dem in dieser Arbeit betrachteten Modell ist das

konkret daran zu erkennen, dass der Vektor der Messzeitpunkte in die numerische Lösung des DGL-Systems eingeht (siehe Abschnitt 3.2.2). Es kommt dabei sowohl auf die Anzahl als auch auf die Verteilung der Messzeitpunkte über die Experimentdauer an.

In dieser Arbeit wurden die Auswirkungen einiger Messanordnungen auf die drei Ziele der optimalen Versuchsplanung hin überprüft. Stellvertretend für die anderen Berechnungen sollen hier nur die Ergebnisse zu den Auswirkungen auf die Modelldiskriminierung, beziehungsweise den Abstand der Grundmodelle 1.1 und 2.1 angeführt werden. Zusätzlich wurde jeweils eine optimale Versuchsdauer bestimmt, die hier ebenfalls nur im Hinblick auf die Modelldiskriminierung dargestellt wird.

5.5.1. Vergleich von Messzeitanordnungen

Es wurden verschiedene Messzeitanordnungen formuliert, um sie vergleichen zu können. Beispielsweise sollte berechnet werden, ob die Halbierung der Messzeit einen Einfluss auf die Modelldiskriminierung hat. Außerdem war von Interesse, ob es sich lohnen würde, in der ersten Hälfte des Experiments doppelt so schnell zu messen, oder ob sogar eine exponentielle Verteilung der Messzeitpunkte über die Experimentdauer von Vorteil wäre. Folgende Messzeitanordnungen wurden miteinander verglichen:

- Ä_300: Äquidistante Messungen bis zur dreihundertsten Minute
- 2:1_300: Messungen bis zur dreihundertsten Minute, wobei zwei Drittel der Messungen in der ersten Hälfte der Zeit gemacht werden
- Ä_150: Äquidistante Messungen bis zur hundertfünfzigsten Minute
- 2:1_150: Messungen bis zur hundertfünfzigsten Minute, wobei zwei Drittel der Messungen in der ersten Hälfte der Zeit gemacht werden
- Expo: Messzeitpunkte deren Abstand exponentiell (2^x) bis zur dreihundertsten Minute zunimmt

In Abbildung 5.15 ist zu sehen, dass das Experiment nicht schon in der hundertfünfzigsten Minute abgebrochen werden, sondern mindestens bis zur dreihundertsten Minute fortlaufen sollte. Halbiert man die Messdauer, kann der Abstand zwischen den Modellen nicht mehr so gut berechnet werden. Dasselbe gilt für Messungen zu Messzeiten, deren Abstand exponentiell zunimmt. Ob es günstiger ist, in der ersten Hälfte des Experiments doppelt so schnell zu messen, kann auf Basis dieser Datenlage nicht entschieden werden. Der Einfluss der Messdauer auf die Auswertung des Experiments wurde anschließend noch einmal separat betrachtet.

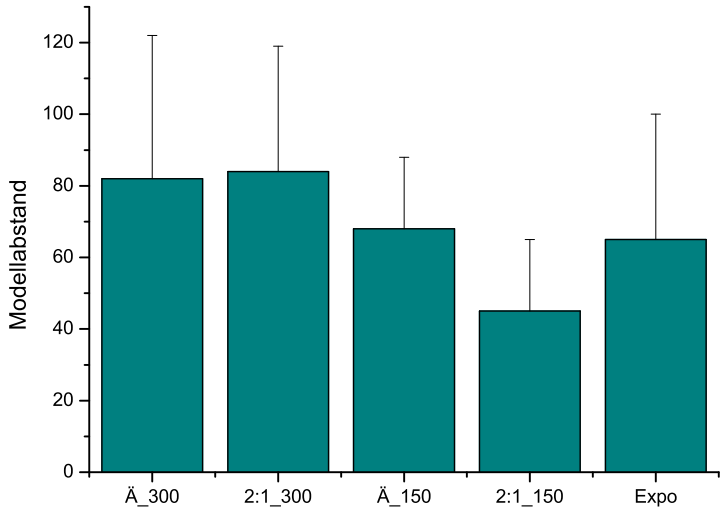

Abbildung 5.15.: Vergleich verschiedener Messzeitanordnungen

5.5.2. Optimierung der Messdauer

Die Messdauer für ein Experiment wird zur Zeit dadurch bestimmt, dass es abgebrochen wird, wenn sich bei einigen aufeinander folgenden Messungen kaum Veränderungen in den Werten zeigen. Die optimale Messdauer wurde im Rahmen dieser Arbeit auf der Grundlage der mathematischen Modelle betrachtet.

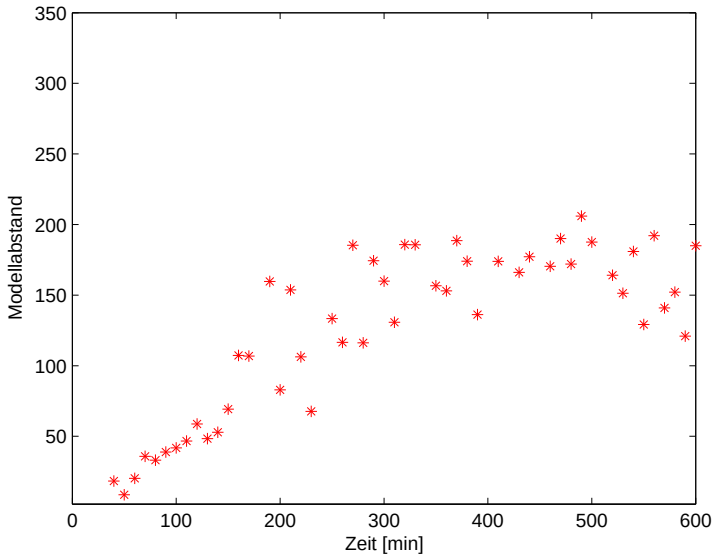

Abbildung 5.16.: Auswirkung der Messdauer auf den Informationszugewinn

In Abbildung 5.16 ist der nach Definition 3.18 berechnete Abstand zwischen Modell 1.1 und Modell 2.1 im Verhältnis zur Messdauer bei einer gleichbleibenden Anzahl von dreizehn Messpunkten dargestellt. Das Optimum liegt bei einer Messzeit

von etwa 400 Minuten. Es lohnt sich also nicht, das Experiment länger als 400 Minuten durchzuführen, wenn es zur Modelldiskriminierung verwendet werden soll. Diese Sättigungskurve bestätigt zugleich die Kurve der theoretischen Aufzeichnung des Modellierungsaufwands gegenüber der Modellfundamentalität (vergleiche Abbildung 2.6).

6. Zusammenfassung und Ausblick

Zusammenfassung

Die in der Einleitung formulierten mathematischen Ziele dieser Arbeit wurden vollständig erreicht. Zunächst sollte ein geeignetes mathematisches Modell gefunden werden, um die Reaktion von Benzaldehyd zu DHPP beschreiben zu können, die von der Benzaldehydlyase katalysiert wird. Dafür wurden zwei Hypothesen über den Verlauf der Reaktion aufgestellt und die aus ihnen hergeleiteten Differentialgleichungsmodelle mit Hilfe des Akaike-Kriteriums, einer mathematischen Methode zur Modelldiskriminierung, verglichen. Als Ergebnis der Modelldiskriminierung konnte einerseits von den beiden Hypothesen für den Reaktionsverlauf die wahrscheinlichere bestimmt werden und andererseits das Differentialgleichungsmodell 2.1 hergeleitet werden, das sich im Vergleich mit allen anderen Modellen als das beste erwies. Es hat nur vier unbekannte kinetische Parameter, die mit Hilfe der Parameterschätzung ziemlich genau bestimmt werden konnten.

Mit Methoden der Modelldiskriminierung konnte ebenfalls festgestellt werden, dass auf Basis der in dieser Arbeit verwendeten Datensätze keine weiteren Teilreaktionen des Versuchs wesentlich sind. Lediglich die Reaktion von Benzaldehyd zu Benzoin und die Folgereaktion zu DHPP sind notwendig für ein Modell, das hinreichend gute Voraussagen liefert. Durch die Modellierung konnten außerdem keinerlei Inhibitionen der BAL durch Substrate oder Produkte nachgewiesen werden.

Um auszuschließen, dass diese Ergebnisse von Messfehlern der in dieser Arbeit ausgewerteten Messdaten abhängen, wurden sie mit Hilfe von Bootstrap-Methoden überprüft. Dabei wurde festgestellt, dass die Ergebnisse der Diskriminierung der beiden Grundmodelle beständig sind, aber eine größere Signifikanz wünschenswert wäre. Bei der Diskriminierung zwischen Differentialgleichungsmodell 2.1 und ergänzten Modellen führten die Bootstrap-Methoden hingegen zu dem Ergebnis, dass für signifikante Aussagen eine breitere Datenbasis nötig ist. Die Erhöhung der Signifikanz der Modelldiskriminierung und die Verbesserung der Datenlage sollten im Anschluss mit den Mitteln der optimalen Versuchsplanung erreicht werden.

Die Planung eines neuen Experiments, auf den Ergebnissen der Parameterschät-

zung und der Modelldiskriminierung aufbauend, hatte zum Ergebnis, dass weitere Versuche mit möglichst hoher Benzaldehyd-Konzentration und geringer BAL-Konzentration durchgeführt werden sollten. Dann wäre es möglich, die Parameter genauer zu schätzen und verschiedene mathematische Modelle präziser voneinander unterscheiden und bewerten zu können. Die Mechanismen und kinetischen Parameter der Reaktion sind (mathematisch) deutlicher zu identifizieren und zu quantifizieren, wenn möglichst viel Substrat pro Enzym eingesetzt wird. Leider konnte das in dieser Arbeit geplante Experiment bisher trotz Interesse der Experimentatoren aus Zeitgründen nicht durchgeführt werden.

Ausblick

Die Durchführung eines neuen Experiments auf Basis der in dieser Arbeit vollzogenen Versuchsplanung, wird Auskunft über die Güte der Planung geben und zur Überprüfung der hier getroffenen Prognosen dienen. Zudem werden sich die kinetischen Parameter auf einer breiteren Datenbasis voraussichtlich wesentlich genauer bestimmen lassen. Des Weiteren wäre die Durchführung einer gezielten Versuchsplanung im Hinblick auf die Inhibition der BAL durch den Reaktionspartner Dimethoxyaldehyd wünschenswert. Diese Inhibition konnte in Einzelversuchen direkt nachgewiesen werden, führte bezüglich des Akaike-Kriteriums im mathematischen Modell aber nicht zu einer Verbesserung der Anpassung an die Daten. Bei weiteren Modellierungen sollten zudem die Messdaten zur Dimethoxyaldehyd-Konzentration einbezogen werden. Die Daten zu diesem weiteren Reaktionsteilnehmer wurden in dieser Arbeit wegen der hohen Messungenauigkeit und starken Schwankungen vernachlässigt.

Persönlicher Kommentar zu dieser Arbeit

Mein Anliegen für diese Arbeit war, im Studium erworbene Kenntnisse und Fähigkeiten an reellen Datensätze zielgerichtet anzuwenden. Die Herausforderung, über die Reproduktion alter Ergebnisse hinaus Neues zu erarbeiten, gefiel mir. Dass ich nicht nur mit realen, sondern auch wissenschaftlich relevanten, „frischen“ Daten arbeiten durfte, motivierte mich dabei zusätzlich.

Weniger bewusst waren mir im Vorhinein die Probleme und Hürden, die bei der Bearbeitung solcher Daten zu bewältigen sind. Die Schätzung der Varianzen für unreferenzierte Messungen war eines der Probleme. Auch die Messungenauigkeiten, die durch Agglomeration oder Flüchtigkeit der an der Reaktion beteiligten Stoffe entstanden, erforderten die Anwendung besonderer Methoden der Statistik, zum Beispiel einer speziellen Gewichtung in der Methode der kleinsten Quadrate. Die mathematische Umsetzung der verschiedenen Arten von durch den experimentellen

Aufbau verursachten Datenbeeinträchtigungen wird leider in Lehrbüchern kaum angesprochen.

Eine weitere Motivation für meine Arbeit, war der Wunsch, mit Naturwissenschaftlern anderer Disziplinen zusammen zu arbeiten und dabei ein weites Blickfeld (über aktuelle Forschungsthemen, Stimmung in Deutschland als Forschungslandschaft, etc.) zu erhalten.

Was meine persönlichen Erwartungen betraf, bin ich sehr zufrieden mit dem Verlauf meiner Arbeit am Forschungszentrum und dem Ergebnis. Das Thema war interessant und ich konnte viele meiner im Studium erlernten Fähigkeiten daran erproben. Andererseits erfordete seine Bearbeitung auch das Lernen neuer Methoden und das Erarbeiten der dahinterliegenden Theorie.

Die interdisziplinäre Arbeit hat mir in den meisten Fällen Vergnügen bereitet. Viele Arbeitsgruppenseminare und Doktorandenseminare, an denen ich teilnehmen konnte, zeigten mir das breite Spektrum an Themen, die am Institut für Biotechnologie 2 bearbeitet wurden und werden. Zusammenfassend gesehen, bot mir das Forschungszentrum in Jülich gute Arbeitsbedingungen, Betreuung und eine motivierende Atmosphäre, was ich genutzt habe und wofür ich sehr dankbar bin.

A. Datensatz

Zur Verfügung standen Datensätze aus drei Batchversuchen (siehe Kapitel 4). Diese Datensätze sind in den Tabellen A.1, A.2 und A.3 aufgeführt. Für eine beispielhafte graphische Darstellung eines der Datensätze siehe Abb. (4.1).

Messzeit [min]	c_{BA} [mM]	c_{BZ} [mM]	c_{DHPP} [mM]
0	56.827	0	0
16	26.362	9.3661	0.57463
21	5.5827	28.256	1.0422
30	3.8501	30.355	1.5634
45	2.9546	30.591	2.816
60	2.8327	29.351	2.6621
90	2.8131	30.645	4.195
120	2.5848	31.127	6.5699
150	2.5028	28.258	8.3837
180	2.4614	29.347	10.381
210	2.2304	27.614	11.755
240	2.4204	26.456	13.732
300	2.3755	23.739	17.937

Tabelle A.1.: Datensatz Batch A

Messzeit [min]	c_{BA} [mM]	c_{BZ} [mM]	c_{DHPP} [mM]
0	31.467	0	0
16	3.3338	12.162	1.8176
21	2.7474	12.053	2.5199
30	2.0867	13.289	3.2671
45	1.9264	12.693	4.2852
60	1.8771	12.5	4.3543
90	1.7697	12.03	5.3394
120	1.7797	12.721	6.9927
150	1.8259	11.955	8.6751
180	1.7131	10.887	9.4588
210	1.674	9.8131	11.537
240	1.5938	10.308	12.315
300	1.4489	8.3153	13.481

Tabelle A.2.: Datensatz Batch B

Messzeit [min]	c_{BA} [mM]	c_{BZ} [mM]	c_{DHPP} [mM]
0	18.622	0	0
16	3.2139	6.7029	0.91459
21	3.0077	7.8395	1.5313
30	2.869	7.7746	1.4764
45	2.0463	7.6137	1.7603
60	2.062	7.4807	2.6997
90	1.877	7.141	3.2706
120	1.7403	6.8793	4.1758
150	1.6467	6.0138	5.8694
180	1.6081	5.4998	6.4649
210	1.6759	5.4603	6.8229
240	1.5947	5.1337	7.5219
300	1.6486	4.2043	9.2429

Tabelle A.3.: Datensatz Batch C

B. Parameterschätzungen zu Modell 1.1 und Modell 2.1

Die Ergebnisse zur Parameterschätzung für die Modelle 1.1 und 2.1 (siehe Abschnitt 5.2) werden an dieser Stelle in Tabellenform und auf die vierte Nachkommastelle gerundet dargestellt.

	${}_{11}\hat{\theta}_{KQSS}$	${}_{11}\hat{\theta}_{LTS}$	${}_{11}\hat{\theta}_{MCData}$	${}_{11}\hat{\theta}_{SIMUL}$
$V_{max,BABZ}$	2.6795 ± 0.0332	2.8504 ± 0.048	2.8812 ± 0.0481	3.4212 ± 0.062
$V_{max,BZDHPP}$	0.0418 ± 0.027	0.0361 ± 0.046	0.0373 ± 0.0491	0.0328 ± 0.0405
$K_{M,BABZ}$	13.364 ± 0.0195	14.2401 ± 0.046	14.476 ± 0.0454	16.957 ± 0.0525
$K_{M,BZDHPP}$	8.9228 ± 0.0566	6.1421 ± 0.125	7.244 ± 0.1237	4.065 ± 0.137
$\sum_{ij} \epsilon_{ij}^2$	779.32	773.73	871.21	1819

Tabelle B.1.: Parameterschätzung zu Modell 1.1

	${}_{21}\hat{\theta}_{KQSS}$	${}_{21}\hat{\theta}_{LTS}$	${}_{21}\hat{\theta}_{MCData}$	${}_{21}\hat{\theta}_{SIMUL}$
$V_{max,BABZ}$	2.3651 ± 0.0606	1.8579 ± 0.0309	2.368 ± 0.0341	2.8722 ± 0.039
$V_{max,BZDHPP}$	0.0745 ± 0.1015	0.0746 ± 0.0067	0.0937 ± 0.0409	0.0990 ± 0.047
$K_{M,BABZ}$	10.823 ± 0.0533	9.4388 ± 0.0183	10.833 ± 0.035	11.214 ± 0.0281
$K_{M,BZDHPP}$	6.5895 ± 0.2188	6.5582 ± 0.0174	7.817 ± 0.0964	12.752 ± 0.0703
$\sum_{ij} \epsilon_{ij}^2$	532.72	441.24	628.5	1614.77

Tabelle B.2.: Parameterschätzung zu Modell 2.1

C. Ergänzungen zur mathematischen Theorie

Satz von Taylor

Zur Herleitung der Methoden zur numerischen Differentiation (vergl. Abschnitt 3.3.2) benötigt man folgenden Satz:

Satz C.1 ***Satz von Taylor*** *[Spä94]*
Ist eine beliebige skalare Funktion f n-mal stetig differenzierbar im geschlossenen Intervall $[a, b]$ und (n+1)-mal differenzierbar im offenen Intervall (a, b), dann gibt es für jedes $x_0 \in [a, b]$ und jedes $x \in [a, b]$ ein ζ mit $x_0 \leq \zeta \leq x$ bzw. $x \leq \zeta \leq x_0$ derart, dass gilt

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + \frac{f^{(n+1)}(\zeta)}{(n+1)!}(x - x_0)^{n+1} \tag{C.1}$$

Beweis zu Satz 3.1:
Wendet man nun Formel C.1 für n=1 an und entwickelt um den Punkt x+h, so erhält man:

$$f(x + h) = f(x) + hf'(x) + \frac{h^2}{2}f''(\zeta_1)$$

woraus sich dann der *Vorwärtsdifferenzenquotient* ergibt:

$$f'(x) \approx \frac{f(x + h) - f(x)}{h} \tag{C.2}$$

und der Abbruchfehler

$$E(h) := -\frac{h}{2}f''(\zeta)$$

Der Abbruchfehler wird verringert, wenn man Formel C.1 mit n=2 verwendet und die Gültigkeit der Voraussetzungen für f im Intervall $[x - h, x + h]$ annimmt. Zuerst

ergibt sich

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(\zeta_2) \quad \text{(C.3)}$$

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f'''(\zeta_3) \quad \text{(C.4)}$$

und duch Subtraktion ((C.3) - (C.4)) ergibt sich dann der zentrale Differenzenquotient

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h} \quad \text{(C.5)}$$

mit dem Abbruchfehler

$$E(h) = -\frac{h^2}{6}f'''(\tilde{\zeta})$$

□

Regularitätsbedingungen

Die Vertauschbarkeit von Grenzwert und Integral (vergl. Satz 3.4) ist durch den Satz von Beppo Levi für die monotone Konvergenz oder den Satz von Lebesgue für die majorisierte Konvergenz nachzuprüfen.

Satz C.2 ***(Satz von Beppo Levi)***
Es sei (f_k) eine monoton wachsende Folge von auf einer Menge B messbaren, nichtnegativen Funktionen und $f(x) = \lim f_k(x)$. Dann gilt

$$\int_B f(x)\,dx = \lim_{k\to\infty} \int_B f_k(x)\,dx.$$

Beweis:
[Wal95]

□

Satz C.3 ***(Satz von der majorisierten Konvergenz)***
Die Funktionen $f_k : B \longrightarrow \mathbb{R} \cup \{\pm\infty\}$ seinen messbar und es gelte $|f_k(x)| \leq |g(x)|$ in B für $k = 1, 2, \ldots$ mit g e ine Lebesgue-integrierbare Funktion. Der Limes $f(x) := \lim_{k\to\infty} f_k(x)$ existiere (punktweise) fast überall in B. Dann sind die Funktionen f_k und f über B integrierbar und es ist

$$\int_B f(x)\,dx = \lim_{k\to\infty} \int_B f_k(x)\,dx$$

Beweis:
[Wal95]

□

Erwartungswerte quadratischer Formen

Satz C.4 ***(Erwartungswerte quadratischer Formen)***
Sei Y ein N-dimensionaler Zufallsvektor und $A \in \mathbb{R}^{N \times N}$. Dann gilt:

$$E(Y^T AY) = E(Y)^T \cdot A \cdot E(Y) + tr(A \cdot Cov(Y))$$

Beweis:
[FH95]

□

D. Beispiele

Ein Beispiel für die Bildschirmausgabe bei Parameterschätzung unter Matlab® zeigt Abbildung D.1. Man kann erkennen, dass 31 Iterationen vollzogen wurden, bei denen insgesamt 160-mal das DGL-System gelöst wurde. Diese Berechnung dauerte etwa acht Sekunden.

```
                                        Norm of      First-order
 Iteration  Func-count     f(x)          step          optimality   CG-iterations
     0          5         38245.7                     2.13e+006
     1         10         18928.4       0.220735      7.95e+005           2
     2         15         7576.63       0.255576      1.77e+006           2
     3         20         1911.34       0.387164      5.34e+005           2
     4         25         993.192       0.165201      1.63e+005           1
     5         30         703.907       0.224948      2.37e+004           2
     6         35         636.247       0.385921      7.79e+005           2
     7         40         628.428      0.0101494       1.6e+004           2
     8         45         618.266      0.0592993      1.26e+004           2
     9         50         617.902       0.109353      1.87e+004           2
    10         55         614.313       0.435353      3.07e+005           2
    11         60          613.43     0.00915409            805           2
    12         65         537.444         3.3352           10.9           2
    13         70         534.522        0.59017      2.36e+004           2
    14         75         533.406     0.00575119            544           2
    15         80         533.308      0.0396214      1.06e+005           2
    16         85         533.125     0.00327455      7.72e+003           2
    17         90         533.125     0.00885428      7.72e+003           2
    18         95         533.056     0.00221357      6.36e+003           0
    19        100         533.056     0.00149387      6.36e+003           2
    20        105         533.041    0.000373468      1.57e+003           0
    21        110         533.009    0.000746937      3.13e+003           2
    22        115         533.009     0.00149387      3.13e+003           2
    23        120         533.009    0.000373468      3.13e+003           0
    24        125         533.005   9.33671e-005            705           0
    25        130         533.005    0.000186734            705           2
    26        135         533.002   4.66835e-005            721           0
    27        140         533.002   9.33671e-005            721           2
    28        145         533.002   2.33418e-005            721           0
    29        150         533.002   5.83544e-006            721           0
    30        155         533.002   1.45886e-006            721           0
    31        160         533.002   3.64715e-007            721           0
Optimization terminated: norm of the current step is less
 than OPTIONS.TolX.

erg =

    2.3939    0.0774   11.0003    7.2709  533.0023
```

Abbildung D.1.: Bildschirmausgabe für die Parameterschätzung

E. Verwendete Programme

DVI-Viewer Yap 2.4.1803
MatLab® 7.0.4.365 (R14)
MicroMath Scientist® 3.0
Maple™ 10
Mathematica® 6.0
CorelDraw Graphics Suite® 12
ChemDraw Ultra® 10.0
Adobe Reader® 7.0
Microsoft Office Excel® 2003
Microsoft Windows XP®

Literaturverzeichnis

[ABT04] Aster, R., B. Borchers und C. Thurber: *Parameter Estimation and Inverse Problems*. In: *International Geophysics*. Elsevier Academic Press, Amsterdam, 2004.

[AD92] Atkinson, A.C. und A.N. Donev: *Optimum Experimental Designs*. In: *Oxford Statistical Science Series*. Oxford University Press, Oxford, 1992.

[BA02] Burnham, K.P. und D.R. Anderson: *Model Selection and Multimodel Inference*. Springer Science + Business Media, Inc., New York, 2002.

[BB89] Bunke, H. und O. Bunke: *Nonlinear Regression, Functional Relations and Robust Methods*. In: *Wiley Series in Probability and Mathematical Statistics*. John Wiley & Sons, Inc., Berlin, 1989.

[BB94] Bandemer, H. und A. Bellmann: *Statistische Versuchsplanung*. B.G. Teubner Verlagsgesellschaft, Leipzig, 1994.

[BH25] Briggs, G.E. und J.B. Haldane: *A Note on the Kinetics of Enzyme Action*. Biochemical Journal, 19(2):338–339, 1925.

[Bis02] Bisswanger, H.: *Enzyme Kinetics*. Wiley-VCH Verlag GmbH, Weinheim, 2002.

[BJ76] Box, G.E.P. und G.M. Jenkins: *Time Series Analysis*. Holden-Day Inc., San Francisco, 1976.

[CB95] Cornish-Bowden, A.: *Fundamentals of Enzyme Kinetics*. Portland Press Ltd, London, 1995.

[DEH+02] Demir, A.S., E. Eren, B. Hosrik, Ö. Şeşenoglu, M. Pohl, E. Janzen, D. Kolter, R. Feldmann, P. Dünkelmann und M. Müller: *Enantioselective Synthesis of α-Hydroxy Ketones via Benzaldehyde Lyase-Catalyzed C-C Bond Formation Reaction*. Advanced Synthesis & Catalysis, 344(1):96 – 103, 2002.

[Dün04] Dünkelmann, P.: *Entwicklung eines Donor/Akzeptor-Konzeptes für die asymmetrische Synthese unsymmetrischer Benzoine mit Hilfe ThDP-abhängiger Enzyme.* Doktorarbeit, Universität Bonn, 2004.

[DP58] Dormand, J.R. und P.J. Prince: *A family of embedded Runge-Kutta formulae.* Journal of Computational Mathematics, 10:517–534, 1958.

[ET98] Efron, B. und R.J. Tibishirani: *An Introduction to the Bootstrap.* In: *Monographs on Statistics and Applied Probability 57.* Chapman & Hall/CRC, Boca Raton, 1998.

[FH95] Fahrmeir, L. und A. Hamerle: *Multivariate statistische Verfahren.* Walter de Gruyter & Co., Berlin, 1995.

[FHG+00] Fang, Q.K., Z. Han, P. Grover, D. Kessler, C.H. Senanayake und S.A. Wald: *Rapid access to enantiopure bupropion and its major metabolite by stereospecific nucleophilic substitution on an α-ketotriflate.* Tetrahedron:Asymmetry, 11(18):3659–3663(5), 2000.

[Fis35] Fisher, R.A.: *The Design of Experiments.* Oliver & Boyd, Edinburgh, 1935.

[GV89] Gonzàles, B. und R. Vicuña: *Benzaldehyde Lyase, a Novel Thiamine PPi-requiring Enzyme, from* Pseudomonas fluorescens *Biovar I.* Journal of Bacteriology, 171:2401 – 2405, 1989.

[Hil05] Hildebrand, F.: *Reaktionstechnische Untersuchungen zur enantioselektiven Synthese von Hydroxyphenylpropanonen durch Benzaldehydlyase.* Diplomarbeit, Rheinische Friedrich-Wilhelms-Universität Bonn, 2005.

[HKP+07] Hildebrand, F., S. Kühl, M. Pohl, D. Vasic-Racki, M. Müller, C. Wandrey und S. Lütz: *The Production of (R)-2-Hydroxy-1-phenyl-propan-1-one Derivatives by Benzaldehyde Lyase From* Pseudomonas fluorescens *in a Continuously Operated Membrane Reactor.* Biotechnology and Bioengineering, 96(5):835 – 843, 2007.

[HT89] Hurvich, C.M. und C-L. Tsai: *Regression and time series model selection in small samples.* Biometrika, 76:297–307, 1989.

[Jan02] Janzen, E.: *Die Benzaldehydlyase aus* Pseudomonas fluoreszens*: Biochemische Charakterisierung und die Untersuchung von Struktur-Funktionsbeziehungen.* Doktorarbeit, Heinrich-Heine Universität Düsseldorf, 2002.

[JMKJ+06] Janzen, E., M. Müller, D. Kolter-Jung, M.M. Kneen, M.J. McLeish und M. Pohl: *Characterization of Benzaldehyde lyase from* Pseudomonas fluorescens*: A versatile enzyme for asymmetric C-C bond formation.* Bioorganic Chemistry, 34:345 – 361, 2006.

[KBB+04] Krieger, N., T. Bhatnagar, J.C. Barrat, A.M. Baron, V.M. de Lima und D. Mitchell: *Non-Aqueous Biocatalysis in Heterogeneous Solvent Systems.* Food Technology and Biotechnology, 42(4):279–286, 2004.

[Küh07] Kühl, S.: *Enzymkatalysierte C-C Knüpfung: Reaktionstechnische Untersuchungen zur Synthese pharmazeutischer Intermediate.* Doktorarbeit, Universität Bonn, 2007.

[KMK+99] Kakeya, H., M. Morishita, H. Koshino, T. Morita, K. Kobayashi und H. Osada: *Cytoxazone: A novel cytokine modulator containing a 2-Oxazolidinone Ring produced by* Streptomyces sp. Journal of Organic Chemistry, 64:1052–1053, 1999.

[KS04] Kaipio, J. und E. Somersalo: *Statistical and Computational Inverse Problems.* In: *Applied Mathematical Sciences.* Springer, Berlin, 2004.

[Mic99] Michal, G. (Herausgeber): *Biochemical Pathways.* Spektrum Akademischer Verlag GmbH, Heidelberg Berlin, 1999.

[MKB79] Mardia, K.V., J.T. Kent und J.M. Bibby: *Multivariate Analysis.* Academic Press Inc. Ltd., London, 1979.

[Mül05] Müller, C.: *Skript zur Explorativen Datenanalyse, WS 2005/2006.* Universität Oldenburg, 2005.

[NC01] Nelson, D. und M. Cox: *Lehninger Biochemie.* Springer-Verlag, Berlin Heidelberg, 2001.

[Pet00] Petersen, K.-U.: *Händige Pharmaka: Chiral Switch - Auf der Suche nach optischen Isomeren.* Deutsches Ärzteblatt 97, 46:A–3089,B–2607,C–2314, 2000.

[Páz93] Pázman, A.: *Nonlinear Statistical Models.* In: *Mathematics and its Applications.* Kluwer, Dordrecht, 1993.

[SCS00] Saltelli, A., K. Chan und E.M. Scott: *Sensitivity Analysis.* In: *Wiley Series in Probability and Mathematical Statistics.* John Wiley & Sons, Inc., Chichester, 2000.

[Seg95] Segel, I.H.: *Enzyme Kinetics.* John Wiley & Sons, Inc., New York, 1995.

[Spä94] Späth, H.: *Numerik*. Friedr. Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig/Wiesbaden, 1994.

[Sti04] Stillger, T.: *Enantioselektive C-C Knüpfung mit Enzymen: Charakterisierung und reaktionstechnische Bearbeitung der Benzaldehydlyase aus* Pseudomonas fluorescens*Biovar I*. Doktorarbeit, Universität Bonn, 2004.

[SW95] Strehmel, K. und R. Weiner: *Numerik gewöhnlicher Differentialgleichungen*. B.G. Teubner, Stuttgart, 1995.

[SW03] Seber, G.A.F. und C.J. Wild: *Nonlinear Regression*. In: *Wiley Series in Probability and Mathematical Statistics*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2003.

[Wal95] Walter, W.: *Analysis 2*. Springer Verlag, Berlin Heidelberg, 1995.

[Wan02] Wandrey, C.: *Skript zur Vorlesung Reaktionstechnik*. Universität Bonn, 2002.

Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe.

Oldenburg, den 17.12.2007