

Karl Siebertz
David van Bebber
Thomas Hochkirchen

Statistische Versuchsplanung

Design of Experiments (DoE)

2. Auflage



Springer Vieweg

VDI-Buch

Weitere Bände in der Reihe <http://www.springer.com/series/3482>

Karl Siebertz · David van Bebber
Thomas Hochkirchen

Statistische Versuchsplanung

Design of Experiments (DoE)

2. Auflage



Springer Vieweg

Karl Siebertz
Aldenhoven, Deutschland

Thomas Hochkirchen
Aachen, Niederlande

David van Bebber
Aachen, Deutschland

VDI-Buch
ISBN 978-3-662-55742-6 ISBN 978-3-662-55743-3 (eBook)
<https://doi.org/10.1007/978-3-662-55743-3>

Die Deutsche Nationalbibliothek verzeichnetet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Springer Vieweg
© Springer-Verlag GmbH Deutschland 2010, 2017
Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsgesetz zugelassen ist, bedarf der vorherigen Zustimmung des Verlags.
Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.
Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.
Der Verlag, die Autoren und die Herausgeber gehen davon aus, dass die Angaben und Informationen in diesem Werk zum Zeitpunkt der Veröffentlichung vollständig und korrekt sind. Weder der Verlag noch die Autoren oder die Herausgeber übernehmen, ausdrücklich oder implizit, Gewähr für den Inhalt des Werkes, etwaige Fehler oder Äußerungen. Der Verlag bleibt im Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutionsadressen neutral.

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier

Springer Vieweg ist Teil von Springer Nature
Die eingetragene Gesellschaft ist Springer-Verlag GmbH Deutschland
Die Anschrift der Gesellschaft ist: Heidelberger Platz 3, 14197 Berlin, Germany

*Für unsere Familien,
die sehr viel Verständnis für unser Hobby
aufgebracht haben und bereit waren, auf viele
Stunden gemeinsamer Freizeit zu verzichten.
Ohne die hervorragende Unterstützung durch
unsere Familien hätten wir das Buch niemals
schreiben können.*

Geleitwort

Nur wenige Methoden haben eine so langfristige Bedeutung für das Arbeitsleben eines Ingenieurs wie die statistische Versuchsplanung. CAD-Programme ändern sich schnell, so dass gelerntes Wissen schnell veralten kann. Auch Berechnungsverfahren ändern sich im Laufe der Zeit, allein durch die Verfügbarkeit immer schnellerer Rechner. Prüfstände sehen heute anders aus als vor zehn Jahren. Oft ändert sich im Verlauf der beruflichen Tätigkeit auch das Arbeitsgebiet, was neues spezifisches Fachwissen verlangt. Nach insgesamt 25 Jahren praktischer Erfahrung mit der statistischen Versuchsplanung in der industriellen Anwendung fällt die Bilanz sehr positiv aus. Die Investition hat sich gelohnt. Es gab viele persönlich erlebte Beispiele für eine erfolgreiche Anwendung. Wissen wurde nicht wertlos, sondern hat sich vermehrt. Natürlich ändern sich auch hier die Grenzen. Dinge sind nun möglich, die noch vor zehn Jahren kaum vorstellbar waren. Mit den Möglichkeiten steigt der methodische Anspruch und auch das Risiko, die Methode falsch anzuwenden.

Wir haben uns entschlossen, die Erfahrungen aufzuschreiben, neues Wissen zu sammeln und die Expertise aus verschiedenen Bereichen zusammenzutragen. So entstand die Idee zu diesem Buch. Ein Buch aus der Praxis für die Praxis. Keiner der Autoren ist an einer Hochschule tätig. Alle mussten nach Feierabend und am Wochenende auf dieses Ziel hin arbeiten. Vielen unserer Leser wird es genauso gehen. Deshalb ist dieses Buch anders geschrieben als ein Hochschulskript. Es geht nicht um wilde Gleichungen und aufwendige Herleitungen, auch nicht um Klausuraufgaben. Es geht um einen fundierten Einblick in eine mathematische Methode, die zur Zeit noch nicht zur üblichen Ingenieursausbildung gehört, obwohl viele Ingenieure im Laufe ihres Arbeitslebens dieses Wissen brauchen.

Design of Experiments wird auch in anderen Disziplinen eingesetzt, zum Beispiel Sozialwissenschaften. In der Chemie ist die statistische Versuchsplanung mittlerweile etabliert. Die in diesem Buch vermittelten Grundlagen sind natürlich in allen Bereichen gültig. Allerdings wollten wir dem Buch ein klares Profil verleihen und haben uns auf die Anwendung im Bereich der klassischen Ingenieurwissenschaften konzentriert. Dort kennen wir uns aus und können dem Leser mitten aus der Praxis berichten.

Stoffauswahl und didaktische Aufarbeitung sind über die Jahre gereift, als Ergebnis vieler selbst durchgeführter Schulungsmaßnahmen und Beratungen für Ingenieure. Ein Teil des Erfolges hängt mit der cleveren Anwendung der Methode zusammen. Clever heißt in diesem Zusammenhang, den technischen Sachverhalt passend aufzuarbeiten, den besten unter mehreren Lösungswegen zu wählen und die richtigen Schlüsse aus den Ergebnissen zu ziehen. Wir haben uns bemüht, den Anwender hier möglichst viel zu unterstützen und nicht einfach die DoE-Methode mathematisch neutral auszurollen. Dieser Anwendungsbezug ist naturgemäß anwendungsspezifisch.

Was erwartet Sie?

Kapitel 1 liefert die Grundlagen für den Einstieg in die Methode der statistischen Versuchsplanung. Hierbei geht es um die Klärung der Fachbegriffe und den groben Überblick. **Kapitel 2** ist den Versuchsplänen gewidmet, mit Beschreibung der gängigen Feldkonstruktionen und ihren spezifischen Eigenschaften. Ziel ist die Unterstützung des Anwenders bei der Auswahl des für sein Problem am besten geeigneten Versuchsplans.

Statistik spielt natürlich bei diesem Verfahren eine zentrale Rolle, insbesondere bei experimentellen Anwendungen. Wie groß muss die Stichprobe sein? Welcher Effekt ist real? Wie genau ist die Vorhersage? Dies sind nur wenige der Fragen die sich immer wieder stellen. Die gängige Ingenieursausbildung kann nur begrenzt darauf eingehen. Es gibt viele Statistikbücher. Leider setzen diese Bücher oft so viel voraus, dass viele Ingenieure im Selbststudium nicht bis zum Ende des Buches durchhalten. Im Rahmen dieses Buches wird ein passendes Bündel geschnürt, quasi ein statistischer *survival kit*. Interessierte Leser seien aber dennoch dazu aufgefordert, weitere Quellen zu nutzen. Statistik lohnt sich.

Bei der Darstellung der Statistik haben wir uns entschlossen, zwei sich ergänzende Kapitel zu schreiben. **Kapitel 3** vermittelt einen Einstieg in die verfügbaren Kontrollverfahren und orientiert sich dabei am Ablauf der praktischen Anwendung. Die vertiefende Betrachtung in **Kapitel 4** geht im Detail auf Teststreuung ein und ist in der Reihenfolge dargestellt, die sich aus der statistischen Theorie ergibt.

Kapitel 5 stellt einige Varianten und Erweiterungen der statistischen Versuchsplanung vor. Hierzu gehören Parameterdesign und Toleranzdesign, aber auch der Umgang mit mehreren Qualitätsmerkmalen.

Ergänzend zur klassischen DoE geht das Buch in der zweiten Hälfte sehr ausführlich auf Computersimulationen ein. Dies erscheint uns notwendig, denn wir sehen in der Praxis eindeutig die Entwicklung zum Metamodelling. Computermodelle werden nicht länger als einfacher virtueller Test verwendet. In der Zukunft geht es um strukturierte Parametervariationen, die Bündelungen zahlreicher Rechnungen zu kompaktem Wissen und die vom Computermodell losgelöste Beschreibung der Ergebnisgrößen als Funktion der Eingangsgrößen. Dies ist ein Modell des Modells, das Metamodell. Hier tritt die DoE in Konkurrenz zu anderen Verfahren, kann sich in vielen Fällen behaupten, muss aber in anderen Fällen Platz machen für völlig andere methodische Ansätze. Der zuständige Berechnungsingenieur muss wissen, wo die Grenzen des Verfahrens liegen und welche Alternativen möglich sind.

Nach dem einführenden **Kapitel 6** liefert **Kapitel 7** die Versuchspläne für komplexe Zusammenhänge. **Kapitel 8** stellt Metamodelle vor und geht dabei auf unterschiedliche Verfahren ein, damit der Anwender die Alternativen zur DoE mit ihren Stärken und Schwächen kennenlernen. Die **Kapitel 9** und **Kapitel 10** behandeln Optimierung und Sensitivitätsanalyse mit Schwerpunkt auf nichtlinearen Methoden.

Kapitel 11 schließt den Hauptteil des Buches mit einer Betrachtung der Strategie ab. Hierbei geht es um praktische Tipps und eine Hilfestellung bei der konkreten Umsetzung. Im Anhang befindet sich ein mathematisches Modell mit vollständiger Dokumentation, als erstes eigenes Anwendungsbeispiel für engagierte Leser. Außerdem enthält der Anhang noch zahlreiche Berechnungsergebnisse, zum Vergleich der klassischen DoE mit anderen multivariaten Analyseverfahren.

Eine einzelne Person hätte dieses Buch nicht schreiben können. Wir haben uns die Kapitel aufgeteilt und jeder konnte sich in seinem Spezialgebiet einbringen. Die Basiskapitel bilden in erweiterter Form einen Trainingskurs ab, den Karl Siebertz seit vielen Jahren hält. Thomas Hochkirchen hat sich als erfahrener Statistiker der statistischen Modellbildung gewidmet. Der Forschungsingenieur David van Bebber hat zahllose Literaturquellen durchforstet und das noch junge Gebiet des Metamodeling von Computersimulationen aufgearbeitet. Dem Leser fällt vermutlich auf, dass jeder Autor seinen eigenen Stil hat. Die Inhalte sind natürlich aufeinander abgestimmt, aber wir hielten es für sinnvoll, das Buch nicht auf einen gemeinsamen Stil weizuspülen.

DoE ist universell einsetzbar und eine sehr robuste Methode. Bereits mit wenigen Grundkenntnissen kann man das Verfahren erfolgreich anwenden. Auf der anderen Seite steckt eine enorme Komplexität hinter dem Gesamtgebilde DoE + Metamodeling. Dieser Spagat zwischen einer leicht verständlichen Einführung und einer exakten Darstellung des gesamten Leistungsspektrums ist nur mit einer progressiv ansteigenden Stoffdichte zu bewerkstelligen. Der Schwierigkeitsgrad steigt daher zu den hinteren Kapiteln stark an. Dies ist leider unvermeidbar.

In der Praxis zeigt sich eine anwachsende Kluft zwischen den Wissensständen von Berechnungsingenieuren und Projektkoordinatoren. Testingenieure befinden sich im Mittelfeld, aber die Grenzen zwischen Test und Simulation beginnen zu verschwimmen. Klassisches Ingenieurwissen wird bei Projektkoordinationen häufig nicht mehr abgefragt und klingt ab. Computersimulationen wagen sich hingegen in immer schwierigeres Terrain. In der Praxis treten Kommunikationsprobleme auf. Statistische Versuchsplanung und Metamodeling können einen wertvollen Beitrag zur Kommunikation und zur Speicherung des Wissens innerhalb eines Unternehmens leisten, weil Ablauf und Ergebnisdarstellungen strukturiert sind.

Viele Leser werden die Methode über lange Jahre benutzen können und ihr Wissen kontinuierlich erweitern. Daher ist es vermutlich kein ernstes Problem, wenn man nicht beim ersten Durchblättern alle Kapitel verständlich findet. Schön ist mitunter ja auch ein Ausblick für fortgeschrittene Anwendungen. Wichtig sind der gefahrlose Einstieg, Erfolgsergebnisse und die Motivation zur weiteren Vertiefung. Dieses Buch soll dafür ein geeigneter Begleiter sein.

Lang ist die Liste derer, die uns unterstützt haben. In der Tat wurde dieses Buch zu einem internationalen Projekt. Softwarehersteller haben ihre Programme zur Verfügung gestellt, um die Analysen anzufertigen und um den aktuellen Leistungsstand der verfügbaren Programme auszuloten. An dieser Stelle möchten wir uns bedanken bei: Caroline Chopek und Cathy Higgins (Statgraphics), Heidi Hansel Wolfe und Tryg Helseth (Design Expert), Dr. Ian Cox und Win LeDinh (JMP). Ohne diese hervorragenden Programme hätte sich das Buch in die Richtung einer (der vielen) trockenen und theoretischen Abhandlungen bewegt, die dem Leser viel Vorstellungskraft abverlangen. Die Erstellung des Manuskripts erfolgte parallel auf drei Rechnern und zwei völlig verschiedenen Betriebssystemen mit \TeX . Eine ungezählte Schar ehrenamtlicher Programmierer hat dies ermöglicht. Dank gebührt auch Christel Siebertz, Anne van Bebber, Susanne Blömer, Christian Gudrian und Bernd Tück für die sorgfältige Durchsicht des Manuskripts. Dr. Alois Mauthofer hat viele wertvolle Anregungen beigesteuert, die in die Konzeption des Buches eingeflossen sind. Unser besonderer Dank gilt dem Springer-Verlag, für das entgegengebrachte Vertrauen und die professionelle Umsetzung unseres Manuskripts in ein schönes Buch.

Aachen, Januar 2010

Dr. Karl Siebertz,

Dr. David van Bebber,

Dr. Thomas Hochkirchen

Geleitwort zur 2. Auflage

Die stetige Nachfrage hat uns dazu ermuntert, das Buch zu aktualisieren und inhaltlich auszubauen. Statistische Versuchsplanung und artverwandte Beschreibungsverfahren sind eben zeitlos schön. Die zweite Auflage wurde stark erweitert, um der stürmischen Entwicklung in den Bereichen Metamodelle und Optimierung Rechnung zu tragen. Außerdem kamen DoE-Beispiele in einem eigenen Kapitel hinzu und die Komponentenanalyse bekam ein eigenes Kapitel. Den aufmerksamen Leser(inne)n möchten wir für ihre Anregungen und Fehlermeldungen danken. Glücklicherweise wurden nur sehr wenige Korrekturen nötig.

Der Trend zu komplexeren Anwendungen hat sich in den vergangenen Jahren verstärkt. Dies betrifft nicht nur CAE Anwendungen, sondern auch Anwendungen im Testbereich, über automatisierte Prüfstände. Immer häufiger kommen sehr große Versuchspläne zum Einsatz. Die gleichzeitige Optimierung mehrerer Qualitätsmerkmale wird zum Regelfall. Das Kapitel der DoE-Beispiele greift diese Thematik bereits auf, bleibt aber noch im Rahmen der statistischen Versuchsplanung, allerdings ergänzt um die Principal-Component-Analysis. Wir entschieden uns für zwei Beispiele aus der Praxis, die vollständig dokumentiert sind, um den typischen Ablauf einer gesamten Anwendung zu verdeutlichen. An dieser Stelle möchten wir uns bedanken bei: Dr. Detlef Neuenhaus, für ein Beispiel aus der Welt des Stahlbaus und bei Dr. Rainer Lach, für ein Beispiel aus der Motorenentwicklung. In beiden Fällen erfolgte die DoE Beratung aus dem Kreise der Autoren, die CAE Berechnungen der konkreten Beispiele jedoch von Dr. Neuenhaus bzw. Dr. Lach.

Bereits die erste Auflage führte in komplexere Versuchspläne, Metamodelle und Optimierung ein. In der zweiten Auflage des Buches wurde dieser Teil sehr stark erweitert. Der mathematische Anspruch steigt dadurch unweigerlich, aber dies entspricht auch den Anforderungen in der Praxis. Mit der gestiegenen Flexibilität in Bezug auf die Größe der Versuchspläne steigt auch die Auswahl der einsetzbaren Methoden. Praktiker sehen sich zunehmend mit der Notwendigkeit konfrontiert, bereits zu Beginn einer Studie zwischen methodischen Ansätzen entscheiden zu müssen, die untereinander nur begrenzt kompatibel sind. Hier soll das Buch eine Entscheidungshilfe bieten, weswegen uns eine möglichst vollständige und konsistente Darstellung wichtig erscheint. Mit der klassischen statistischen Versuchsplanung haben derartige Metamodelle nicht unmittelbar zu tun, sie erfüllen aber letztlich die gleiche Aufgabe: quantitative Beschreibung des Systemverhaltens. Insofern macht es durchaus wieder Sinn, alle Verfahren in einem einzigen Buch darzustellen. Das erweiterte Strategiekapitel am Ende des Buches bildet eine Klammer und soll bei der Entscheidung für das passende Verfahren helfen.

Farbige Abbildungen haben sich bewährt, um komplizierte Sachverhalte einfacher und prägnanter darzustellen, als dies mit monochromen Abbildungen möglich wäre. Nur im Druck erzeugen farbige Abbildungen zusätzliche Kosten. Bereits jetzt haben elektronische Kopien des Buches zahlenmäßig eine weitaus größere Verbreitung gefunden, als gedruckte Kopien. Dieser Trend wird sich vermutlich fortsetzen. In der zweiten Auflage wurde deswegen konsequent auf farbige Abbildungen Wert gelegt.

Auch 2017 gehört die statistische Versuchsplanung noch immer nicht zur Standardausbildung junger Ingenieurinnen und Ingenieure. Im Rahmen eines Wahlfachs für Maschinenbau im Masterstudiengang einer recht bekannten ortsansässigen Hochschule unterrichten wir einmal im Jahr die Grundlagen der statistischen Versuchsplanung. Hinzu kommen firmeninterne Kurse für neue Kollegen, in erster Linie junge Hochschulabsolventen. Es zeigt sich, dass nach wie vor der Bedarf nach einer kompakten Einführung besteht. Das Buch bleibt also der ursprünglichen Linie treu und liefert das gesamte Spektrum, von der einfachen Darstellung der Grundlagen, bis zur fundierten Beschreibung der komplexeren Verfahren.

Dank gebührt Birgit Siebertz und Anne van Bebber für die aufgebrachte Geduld und Nachsicht mit ihren Feierabendautoren. Für umfangreichen Korrekturen der überarbeiteten und erweiterten Kapitel möchten wir besonders folgenden Personen danken: Dr. Heiko Baum, Klaus Peter Heinig, Dr. Claudia Herudek, Dr. Bert Hobein, Dr. Anselm Hopf, Dr. Christian Hoymann, Florian Huth, Dr. Bernd Jungbluth, Dr. Helmut Kindl, Hans Günter Quix, Andreas Schmitt und Anne van Bebber. Dr. Anselm Hopf danken wir zusätzlich für die sorgsame Durchsicht des Literaturverzeichnisses. Der Ford Werke GmbH danken wir für die Freigabe der Bilder und Ergebnisse für das zweite DoE-Beispiel. Unser besonderer Dank gilt auch diesmal dem Springer-Verlag, für das entgegengebrachte Vertrauen und die professionelle Umsetzung unseres Manuskripts in ein schönes Buch.

Aachen, Juli 2017

Dr. Karl Siebertz,

Dr. David van Bebber;

Dr. Thomas Hochkirchen

Inhaltsverzeichnis

1	Grundlagen	1
1.1	Einleitung	1
1.2	Grundbegriffe	2
1.2.1	Systemgrenzen	3
1.2.2	Qualitätsmerkmal	4
1.2.3	Parameter und Faktoren	5
1.2.4	Stufen	6
1.2.5	Vergleich zu traditionellen Verfahren	6
1.3	Auswertung	9
1.3.1	Fallstudie	9
1.3.2	Effekt	12
1.3.3	Wechselwirkung	15
1.3.4	Lineares Beschreibungsmodell	21
	Literaturverzeichnis	25
2	Versuchspläne	27
2.1	Einleitung	27
2.2	Screening Versuchspläne	28
2.2.1	Konzept	28
2.2.2	Reguläre Felder nach dem Yates-Standard	32
2.2.3	Irreguläre Felder nach Plackett-Burman	33
2.2.4	Fallstudie	35
2.3	Versuchspläne für ein quadratisches Beschreibungsmodell	39
2.3.1	Central-Composite-Design	40
2.3.2	Box-Behnken-Design	42
2.3.3	Monte-Carlo-Verfahren	44
2.3.4	Fallstudie	45
2.4	Grenzen des Beschreibungsmodells	48
2.5	Mischungspläne	52
2.5.1	Simplex-Lattice-Design	52

2.5.2 Simplex-Centroid-Design	53
2.6 Individuell erstellte Versuchspläne	54
2.6.1 Auswahlkriterien	55
2.6.2 Einschränkungen des Faktorraums	57
2.7 Die Mutter aller Versuchspläne	57
Literaturverzeichnis	59
3 Kontrollverfahren	61
3.1 Einleitung	61
3.2 Versuchsplan	62
3.2.1 Fallbeispiel	63
3.2.2 Korrelationsmatrix	64
3.2.3 Varianz-Inflations-Faktor (VIF)	65
3.2.4 Fraction of Design Space (FDS)	65
3.2.5 Hebelwerte	66
3.3 Beschreibungsmodell	68
3.3.1 Half-Normal-Plot	68
3.3.2 Varianzanalyse	73
3.4 Genauigkeit der Vorhersage	77
3.4.1 Fallbeispiel	77
3.4.2 Residual-Plots	78
3.4.3 Löschdiagnosen	81
3.4.4 Box-Cox Transformation	83
3.4.5 Bestätigungsläufe	84
Literaturverzeichnis	85
4 Statistische Modellbildung	87
4.1 Einleitung	87
4.2 Warum Statistik?	88
4.3 Randomisierung, Wiederholung, Blockbildung — Fishers Brücke in die Statistik	93
4.3.1 Randomisierung	93
4.3.2 Wiederholung	96
4.3.3 Blockbildung	99
4.4 Wieso “Null”hypothese? Der Grundgedanke aller statistischen Tests	101
4.4.1 Ein Beispiel	101
4.4.2 α - und β -Risiko	102
4.4.3 Versuchsumfang	106
4.5 “Der” Test für DoE: Fishers Varianzanalyse	111
4.5.1 Varianzzerlegung	111
4.5.2 Die Anova-Tabelle	114
4.5.3 Von der Testgröße zur Wahrscheinlichkeit	117
4.5.4 Auswertung bei Blockbildung	121
4.5.5 Faktorelimination	122
4.5.6 Versuchszahl	128

4.6	Modellvalidierung	133
4.7	Zusammenfassung: Von den Daten zum Modell in 7 Schritten	136
	Literaturverzeichnis	137
5	Varianten der statistischen Versuchsplanung	139
5.1	Einleitung	139
5.2	Umgang mit mehreren Qualitätsmerkmalen	140
5.2.1	Multiple-Response-Optimisation	140
5.2.2	Sequentielle Methode und Ersatzgrößen	145
5.2.3	Principal Component Analysis	146
5.3	Robustes Design	147
5.3.1	Parameterdesign	147
5.3.2	Toleranzdesign	153
5.4	Umgang mit kategorialen Faktoren	156
	Literaturverzeichnis	158
6	DoE Beispiele	159
6.1	Einleitung	159
6.2	Die Schutzplanke	160
6.2.1	Systembeschreibung und Versuchsplan	160
6.2.2	Auswertung der Qualitätsmerkmale	163
6.3	Der Ventiltrieb	165
6.3.1	Systembeschreibung	165
6.3.2	Versuchsplan und Stufenfestlegung	167
6.3.3	Auswertung der Qualitätsmerkmale	169
6.3.4	Optimierung	172
	Literaturverzeichnis	177
7	Computer-Experiment	179
7.1	Einleitung	179
7.2	Aufbau und Analyse von Computer-Experimenten	179
7.2.1	Vergleich von Computer- und physikalischem Experiment	181
7.2.2	Testfelder für Computer-Experimente	182
7.2.3	Metamodelle	186
7.2.4	Analyse und Optimierung	186
	Literaturverzeichnis	188
8	Versuchspläne für komplexe Zusammenhänge	189
8.1	Einleitung	189
8.2	Gütekriterien für Testfelder	190
8.2.1	MiniMax und MaxiMin	191
8.2.2	Entropie	193
8.2.3	Gleichverteilung (<i>Uniformity</i>)	194
8.2.4	Vergleich verschiedener Gütekriterien	197
8.3	Konstruktionsmethoden gleichverteilter Testfelder	198
8.3.1	(Quasi) Monte-Carlo	199

8.3.2 Orthogonale Testfelder	204
8.3.3 Latin Hypercube	205
8.3.4 Gleichverteilte Testfelder (<i>Uniform Designs</i>)	209
8.4 Optimierung von Testfeldern	214
8.5 Ungleichverteilte Testfelder	219
8.6 Faktorbereiche entfernen	220
8.7 Versuchsplanerweiterungen und Voranalyse von Messdaten	222
8.8 Wie viele Messungen soll ich machen?	224
8.9 Zusammenfassung	226
Literaturverzeichnis	227
9 Metamodelle	231
9.1 Einleitung	231
9.2 Lineare Regression	232
9.3 Polynome	234
9.3.1 Faktorwahl	234
9.4 Robuste Regression	239
9.5 Adaptive Basis-Funktions Konstruktion	241
9.6 Kernel- und Lokale Polynom-Regression	243
9.7 Regressionsbaum	248
9.8 Splines	254
9.9 Support Vector Machines zur Klassifikation	258
9.9.1 Klassifikation linear separierbarer Daten	258
9.9.2 Klassifikation nicht komplett linear separierbarer Daten	262
9.9.3 Nichtlineare Support Vector Machines	264
9.10 Support Vector Regression	266
9.10.1 Regression mit linearer ϵ -unempfindlicher Straffunktion	266
9.10.2 v -SVR-Verfahren zur automatischen ϵ -Bestimmung	269
9.10.3 Regression mit quadratischer ϵ -unempfindlicher Straffunktion	270
9.10.4 Least Square Support Vector Regression	272
9.11 Ridge Regression	273
9.12 Kleinster und gewichteter Abstand	275
9.13 Kriging	276
9.13.1 Kriging-Regression verrauschter Daten	281
9.13.2 Universal Kriging	283
9.14 Radial Basis Funktion	284
9.14.1 RBF-Regression verrauschter Daten	285
9.14.2 Polynom Erweiterung	288
9.14.3 Reduzierung der Zentren	289
9.15 Gauß Prozess Modelle	290
9.15.1 Bedingte Verteilung und Randverteilung	291
9.15.2 Vorhersagen mittels bedingter Verteilung	293
9.15.3 Betrachtung im Funktionsraum	294
9.15.4 Kovarianzfunktionen	295

9.15.5 Anpassung der Hyperparameter	300
9.16 Künstliche Neuronale Netzwerke	302
9.17 Kombinerte Modelle	313
9.18 Qualität von Metamodellen	314
9.19 Zusammenfassung	316
Literaturverzeichnis	319
10 Optimierung	325
10.1 Einleitung	325
10.2 Dominanz	326
10.2.1 Priorität und Grenzwert	328
10.3 Randbedingungen	332
10.4 Reduktion auf eine Zielgröße	333
10.5 Naturanaloge Optimierungsverfahren	336
10.5.1 Partikelschwarmoptimierung	336
10.5.2 Glühwürmchen	340
10.5.3 Fledermaus	342
10.5.4 Blütenbestäubung	344
10.5.5 Symbiotische Organismus Suche	346
10.5.6 Erweiterung auf mehrere Zielgrößen	348
10.6 Genetische Evolutionsverfahren für mehrerer Zielgrößen	353
10.6.1 Kreuzung	355
10.6.2 Mutation	357
10.6.3 Exemplarische Verfahren (NSGA-II und ϵ -MOEA)	358
10.7 Qualität multidimensionaler Pareto-Grenzen	367
10.7.1 R-Indikatoren	368
10.7.2 Hypervolumen	374
10.8 Zusammenfassung	376
Literaturverzeichnis	377
11 Korrelationsanalyse	381
11.1 Pearson Korrelation	381
11.2 Scheinkorrelation und verdeckte Korrelation	383
11.3 Signifikanz einer Korrelation	383
11.3.1 Permutationstest	384
11.4 Rangkorrelation	385
11.4.1 Spearman	387
11.4.2 Kendalls τ	388
11.5 Nichtlineare Korrelation	389
Literaturverzeichnis	394
12 Komponentenanalyse	395
12.1 Einleitung	395
12.2 Hauptkomponentenanalyse (PCA)	396
12.3 Kernel-Hauptkomponentenanalyse (kPCA)	400

12.4	Unabhängige Komponentenanalyse (ICA)	403
12.4.1	Unabhängigkeit	405
12.4.2	Nicht-Normalverteilt	406
12.4.3	Datenvorbereitung	409
12.4.4	FastICA	410
	Literaturverzeichnis	413
13	Sensitivitätsanalyse	415
13.1	Einleitung	415
13.2	Sensitivitätsanalyse bei Linearen Modellen	416
13.2.1	Normierte Regressionskoeffizienten	416
13.2.2	Partialsumme der Quadrate	417
13.2.3	Partieller Determinationskoeffizient	418
13.2.4	Predictive Error Sum of Squares	418
13.2.5	Partielle Korrelationsfaktoren	418
13.3	Sensitivitätsanalyse bei nichtlinearen Modellansätzen	419
13.3.1	Korrelationsverhältnis	419
13.3.2	Sobol's Kennzahl	422
13.3.3	FAST (Fourier Amplitude Sensitivity Test)	425
13.4	Zusammenfassung	429
	Literaturverzeichnis	430
14	Strategie	431
14.1	Einleitung	431
14.2	Qualitative Systembeschreibung	432
14.3	Versuchsdurchführung und Auswertung	434
14.4	CAE	436
14.5	Software	437
15	Strategie für komplexe Systeme	441
15.1	Vorbereitung und Planung	441
15.2	Versuchsplan erstellen	451
15.3	Experiment (Messung und Rechnung)	453
15.4	Kontrolle der Daten	453
15.5	Erzeugung von Metamodellen	454
15.6	Kontrolle der Metamodelle	455
15.7	Analyse der Daten und Metamodelle	456
15.8	Optimierung	457
15.9	Prüfung der Analyse und Optimierungsergebnisse	457
15.10	Dokumentation	458
15.11	Schlusswort	458

A	Berechnungsmodell zum Fallbeispiel Rasensprenger	461
A.1	Nomenklatur	462
A.2	Berechnung	462
A.3	Erweiterungen	466
A.4	Quellcode	468
	Literaturverzeichnis	473
B	Computer-Experiment	475
B.1	Rasensprenger mit erweitertem Faktorraum	475
B.2	Testfelder und Metamodelle	479
B.3	Sensitivitätsanalyse	491
B.4	Optimierung	492
	Nomenklatur	495
	Abkürzungen und Markennamen	499
	Sachverzeichnis	501

Kapitel 1

Grundlagen

1.1 Einleitung

Die statistische Versuchsplanung – als Methode zur effizienten Planung und Auswertung von Versuchsreihen – wurde bereits in den 20er Jahren des vergangenen Jahrhunderts entwickelt, ist also älter als vielfach angenommen. 1935 schrieb R. A. FISHER das erste Fachbuch darüber [3]. Frühe deutschsprachige Literatur zu diesem Thema erschien in den 70er Jahren [1]. Erst in den 80er Jahren hat sich die Methode weltweit durchgesetzt, ist aber immer noch nicht standardisierter Bestandteil einer Ingenieursausbildung.

Dieses Buch möchte die Lücke schließen und dem in der Praxis tätigen Ingenieur eine fundierte, aber gleichzeitig leicht verständliche Einführung an die Hand geben. Die mathematischen Herleitungen sind auf das erforderliche Minimum reduziert, damit genug Raum bleibt, um auch komplexere Fälle behandeln zu können, die in der Praxis auch durchaus auftreten.

Kaum eine andere Methode ist derart universell in allen Ingenieursdisziplinen einsetzbar. Allerdings erwachsen daraus auch Unterschiede in Bezug auf die mathematischen Anforderungen. Reale Versuche sind häufig sehr aufwändig und die Zahl der kontrolliert einstellbaren Parameter hat enge Grenzen. Die verwendeten Versuchspläne sind in der Regel einfach, aber dafür erschwert die Teststreuung die Interpretation der Versuchsergebnisse. Typische Fragestellungen betreffen hier die notwendige Stichprobengröße oder die Unterscheidung von realen und scheinbaren Effekten. Anwendungen in der Verfahrenstechnik erfordern oft maßgeschneiderte Versuchspläne, um kritischen Einstellungen aus dem Weg zu gehen. Im Gegensatz dazu ermöglichen Computersimulationen mittlerweile sehr große Versuchspläne mit vielen Faktoren. Dies erfordert leistungsfähige Beschreibungsmodelle, um auch die nichtlinearen Zusammenhänge zu erfassen. Alle oben genannten Fälle sind praxisrelevant und die statistische Versuchsplanung bietet das nötige Rüstzeug, um die spezifischen Probleme zu lösen.

Auf CAE-Anwendungen geht dieses Buch bewusst sehr detailliert ein, da diese Anwendungen in Zukunft aller Voraussicht nach an Bedeutung gewinnen wer-

den. Dies liegt an der stetigen Leistungssteigerung der Computer, dem allgemeinen Trend zur Einsparung von Prototypen und an der Notwendigkeit zur Kommunikation [5] des Modellverhaltens. Letzteres wird von Berechnungsingenieuren oft unterschätzt. Bei komplexen Anwendungsfällen tritt die statistische Versuchsplanung in Konkurrenz zu anderen Verfahren der multivariaten Datenanalyse¹, kann sich aber nach wie vor gut behaupten. Einfache Handhabung, Effizienz, Stabilität und klare Darstellung der Ergebnisse sind die Gründe dafür. Berechnungsingenieure sind in der Regel sehr versiert in Mathematik, daher setzen die entsprechenden Kapitel des Buches im Vergleich zu den übrigen Kapiteln etwas mehr voraus.

Wegen der standardisierten Vorgehensweise wird die statistische Versuchsplanung mittlerweile von zahlreichen Auswerteprogrammen unterstützt. Dies trägt natürlich zur weiteren Verbreitung des Verfahrens bei. Der Anwender hat daher im Regelfall nichts mehr mit der Konstruktion von Versuchsplänen oder der Lösung von Gleichungssystemen zu tun. Allerdings ist es von Vorteil, wenn man die spezifischen Eigenschaften der zur Auswahl stehenden Versuchspläne kennt und die automatisch erstellten Diagramme korrekt interpretieren kann. In diesem Buch wird ganz bewusst eine neutrale Darstellung verfolgt, die den Leser nicht an eine bestimmte Software bindet.

Wer sich einmal mit der statistischen Versuchsplanung vertraut gemacht hat, wird diese über viele Jahre mit Erfolg einsetzen können und mit der Zeit auch komplexe Anwendungen beherrschen. Ähnlich wie bei etablierten Ingenieurwissenschaften, zum Beispiel Mechanik oder Thermodynamik, bleibt das Grundlagenwissen zeitlos aktuell. Daher lohnt sich die Mühe der Einarbeitung in die statistische Versuchsplanung und Fachwissen lässt sich in Ruhe ansammeln. Die Kapitel 1-6 bieten das nötige Rüstzeug, um mit der Methode zu beginnen. Die folgenden Kapitel decken auch schwierigere Fälle ab, damit dieses Buch auch langfristig als Nachschlagewerk dienen kann.

1.2 Grundbegriffe

Ein Grund für den durchschlagenden Erfolg der statistischen Versuchsplanung liegt in der weltweit standardisierten [8, 10, 4] Vorgehensweise, insbesondere der Darstellung der Ergebnisse. Dieser Normierungseffekt begünstigt eine effiziente Kommunikation innerhalb und zwischen den beteiligten Unternehmen. Die Grundbegriffe, wie zum Beispiel *Faktor*, *Effekt* oder *Wechselwirkung*, bilden sozusagen das Vokabular der Methode. Gleichzeitig verdeutlicht dieses Kapitel einige Unterschiede zur traditionellen Vorgehensweise bei der Versuchsplanung und geht auf die Vorbereitung einer Versuchsreihe ein.

¹ zum Beispiel: neuronale Netze, Kriging oder multivariate adaptive regression splines

1.2.1 Systemgrenzen

Das System ist das zu untersuchende Gebilde. Es muss klar definierte Grenzen haben, die Systemgrenzen. Eingangsgrößen gehören entweder zum System oder liegen außerhalb des Systems. Dies ist ein entscheidender Unterschied, denn nur für die zum System gehörigen Größen kann im Rahmen der Untersuchung eine optimale Einstellung gefunden werden. Bei allen übrigen Eingangsgrößen muss man mit Variationen rechnen, kann also in der Praxis keine feste Einstellung voraussetzen. Das System muss sozusagen mit einer beliebigen Kombination der übrigen Einstellgrößen zureckkommen. Bei der Abgrenzung des Systems kann ein Blockschaubild sehr hilfreich sein. Das Blockschaubild zeigt an, mit welchen benachbarten Systemen das untersuchte System in Verbindung steht und wie die Verbindung im konkreten Fall aussieht.

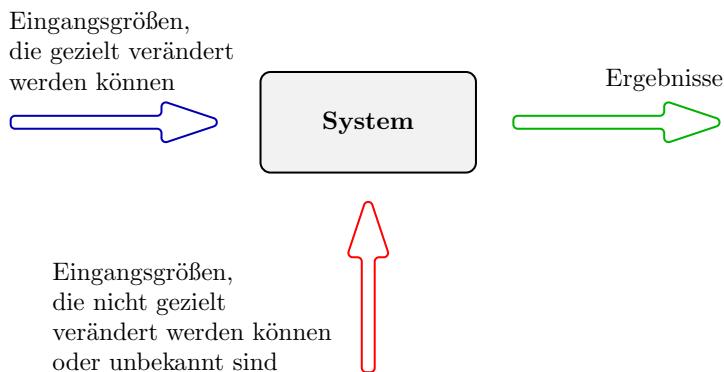


Abb. 1.1 Schematische Darstellung des untersuchten Systems. Eine Fülle von Eingangsgrößen wirkt darauf ein. Jedes System hat Grenzen und erzeugt Ergebnisse.

In der Praxis taucht an dieser Stelle auch ein nicht zu unterschätzendes psychologisches Moment auf. Nur selten bleibt die Bearbeitung eines Versuchsplanes innerhalb einer Abteilung, mitunter sind auch mehrere Unternehmen beteiligt. Die Abgrenzung des Systems entscheidet darüber, wer unmittelbar betroffen ist. Wird das System zu weit abgesteckt, fällt der Wirkungsgrad der Untersuchung, wegen der großen Anzahl beteiligter Parteien. Ist das System zu eng abgegrenzt, bleiben viele Einflussfaktoren ungenutzt. Möglicherweise verlieren dann potentiell wichtige Teammitglieder das Interesse an der Untersuchung. All dies hat zunächst nichts mit Statistik zu tun, aber die statistische Versuchsplanung erzwingt eine frühzeitige Festlegung, was sich in der Praxis oft als das eigentliche Erfolgsgeheimnis entpuppt.

1.2.2 Qualitätsmerkmal

Jedes System hat eine oder mehrere Funktionen. Die Erfüllung der Funktionen äußert sich in messbaren Ergebnissen, mit Hilfe derer sich gute von schlechten Systemen eindeutig unterscheiden lassen. Daher werden diese (positiven) Ergebnisse auch Qualitätsmerkmale genannt. Im weiteren Verlauf wird das System nur noch hinsichtlich des Qualitätsmerkmals untersucht. Selten genügt in der Praxis ein einziges Qualitätsmerkmal, um alle Anforderungen auszudrücken. Mehrere Qualitätsmerkmale sind unkritisch, denn sie lassen sich unabhängig voneinander erfassen und auswerten. Sogar eine gemeinsame Optimierung (Multiple Response Optimisation) ist mit geringem Aufwand möglich.

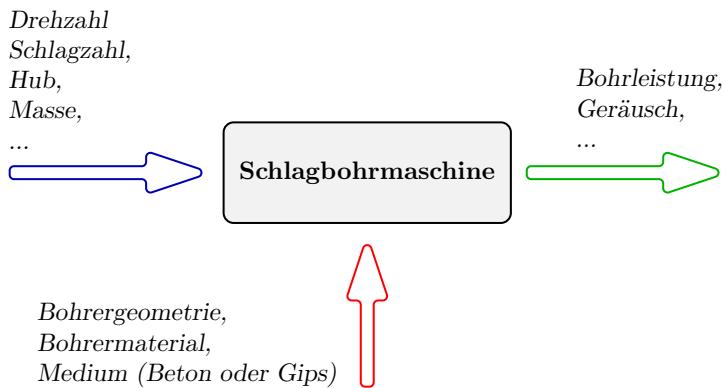


Abb. 1.2 Schematische Darstellung am Beispiel einer Bohrmaschine. Der Hersteller kann die konstruktiven Parameter der Bohrmaschine festlegen, jedoch wählt der Kunde die Bohrer aus. Hohe Bohrleistung und möglichst moderate Lärmentwicklung gehören zu den Qualitätsmerkmalen.

Qualitätsmerkmale müssen kontinuierliche Größen sein, sonst gelingt keine Effektberechnung. Notfalls muss man Hilfsgrößen einführen, um ein digitales Systemverhalten beschreiben zu können. (Zum Beispiel die Zahl der im Prüfzeitraum ausgefallenen Prototypenteile.)

Auch wenn es zunächst erstaunen mag, die Definition der Systemgrenzen und der Qualitätsmerkmale entscheidet über den späteren Erfolg der gesamten Untersuchung. Gelingt dieser Schritt, dann bringen auch einfache Versuchspläne einen deutlichen Erkenntnisgewinn. Ist im Gegensatz dazu das System falsch abgegrenzt oder wird die Systemleistung durch unklug gewählte Qualitätsmerkmale nicht ausreichend erfasst, scheitert jeder Versuchsplan.

1.2.3 Parameter und Faktoren

Die Menge aller Eingangsgrößen nennt man Parameter. Selbst wenn zunächst nur die Möglichkeit besteht, wenige Einflussgrößen zu untersuchen, ist es sinnvoll, von Anfang an eine möglichst vollständige Parameterliste zusammenzustellen, um die Prioritäten festlegen zu können und eventuelle spätere Versuchsreihen zu unterstützen. Die Parameterliste hat auch einen gruppodynamischen Aspekt, denn es fördert die konstruktive Zusammenarbeit, wenn alle Experten des Teams ihre favorisierten Parameter beisteuern können. Die nachfolgenden Untersuchungen führen immer zu einer Selektion, also bleiben Einflussgrößen unerkannt, wenn sie nicht von Anfang an in Betracht gezogen werden.

Die im Versuchsplan enthaltenen Parameter heißen Faktoren, stellen also eine sorgfältig ausgewählte Teilmenge dar. Natürlich bieten sich zunächst die Faktoren an, die nach den verfügbaren Informationen einen großen Einfluss auf das System haben. Im Zweifelsfall sollte man aber immer eine höhere Zahl von Faktoren testen. Der statistischen Versuchsplanung genügt eine grobe Einteilung der Parameter in zwei Gruppen: Faktoren, die untersucht werden und übrige Parameter, die beobachtet und möglichst konstant gehalten werden. Eine weitere Priorisierung muss nicht erfolgen. Dies ist ein Vorteil, denn es erspart dem Team lange Diskussionen

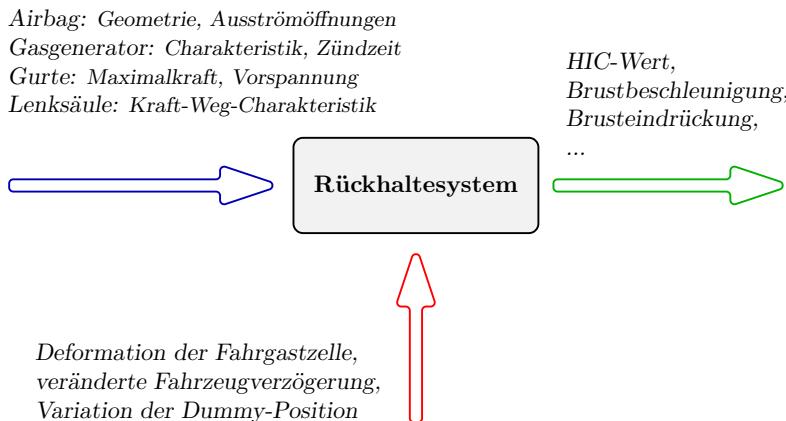


Abb. 1.3 Schematische Darstellung am Beispiel eines Pkw-Rückhaltesystems. Aufgabe des Systems ist es, die Insassenbelastungen im Falle eines Aufpralls zu minimieren. Üblicherweise zählt die Frontstruktur nicht zum Rückhaltesystem, liegt also außerhalb der Systemgrenzen. Die Fülle der Einflussgrößen erfordert hier häufig eine sorgfältige Auswahl der untersuchten Faktoren, zumindest solange kein passendes Berechnungsmodell vorliegt.

Faktoren müssen gezielt und reproduzierbar eingestellt werden. Der Versuchsplan sieht viele verschiedene Einstellungskombinationen vor und es ist wichtig, dass sich die Kombinationen nicht im konkreten Fall gegenseitig ausschließen. Im Bedarfsfall lassen sich spezielle Versuchspläne erzeugen, um einzelne Kombinationen zu vermeiden. Diese haben in der Regel jedoch Nachteile. Mit der Zahl der Faktoren

steigt natürlich auch der Aufwand. Außerdem verlangt die statistische Versuchsplanung eine lückenlose Abarbeitung des Versuchsplans. Erfolgsentscheidend ist nun die geschickte Auswahl der Faktoren, wobei sich in der Praxis oft eine Gruppierung in mehrere voneinander unabhängige Versuchspläne anbietet, um Kompatibilitätsprobleme zu umgehen.

1.2.4 Stufen

Die Einstellungen der Faktoren nennt man *Stufen* oder *Level*. Jeder Faktor wird auf mindestens zwei unterschiedlichen Stufen getestet. Die Stufen sind festgelegte Zustände oder Sachverhalte für die jeweilige Faktoreneinstellung. Der Effekt eines Faktors hängt natürlich von der Einstellungsvariation ab, also dem Stufenabstand. Die Stufenabstände der Faktoren eines Versuchsplans müssen demnach aufeinander abgestimmt sein. Sofern feste Sachverhalte die Stufenwahl vorgeben, existiert kein Handlungsspielraum. In vielen Fällen gehört es jedoch zu den vorbereitenden Arbeiten, die möglichen Stufenabstände auszuloten. Geringe Stufenabstände gehen mit kleinen Effekten einher. Das ist nicht mathematisch, sondern physikalisch begründet. Kleine Änderungen haben in der Regel kleine Auswirkungen. Ist die Änderung zu klein, wird die Messung der Wirkung schwierig sein. Extrapolationen sind grundsätzlich riskant, weil sich außerhalb des untersuchten Parameterbereichs das Systemverhalten mitunter sprunghaft ändert.

Gerade in frühen Phasen der Untersuchung sind große Stufenabstände ratsam. Begrenzt werden die Stufenabstände durch die Forderung nach einem funktions-tüchtigem System. Zunächst muss jeder einzelne Faktor in einem realistischen Einstellbereich bleiben. Die gleichzeitige Variation vieler Faktoren kann darüber hinaus für das System eine drastische Änderung darstellen. Im Zweifelsfall sind Vorversuche sehr nützlich, um zu überprüfen, ob das untersuchte System auch wirklich bei allen geplanten Einstellungskombinationen funktioniert.

Unter Kodierung versteht man eine einheitliche Schreibweise mit dem Ziel, die Faktorenstufen zu kennzeichnen. Es gibt verschiedene Konventionen, zum Beispiel $-/+$, $-1/1$ oder $1/2$. Ein Versuchsplan mit vielen Faktoren lässt sich kodiert sehr kompakt darstellen. Standardisierte Versuchspläne sind immer kodiert angegeben, auch in kommerziellen Auswerteprogrammen.

1.2.5 Vergleich zu traditionellen Verfahren

Bei einem vollfaktoriellen Versuchsplan (Vollfaktorplan) werden alle Kombinationen getestet. Der Versuchsaufwand n_r ergibt sich aus der Zahl der Faktoren n_f und der Zahl der Stufen n_l [7].

$$n_r = n_l^{n_f} \quad (1.1)$$

<i>A</i>	<i>B</i>	<i>C</i>	<i>y</i>
—	—	—	y_1
+	—	—	y_2
—	+	—	y_3
+	+	—	y_4
—	—	+	y_5
+	—	+	y_6
—	+	+	y_7
+	+	+	y_8

Tabelle 1.1 Einfacher Versuchsplan mit drei Faktoren und acht Versuchsläufen. Jeder Faktor wird in zwei unterschiedlichen Einstellungen getestet. Die beiden Stufen sind mit – und + kodiert. Acht Versuche erlauben es, alle Kombinationen zu testen.

Bei 7 Faktoren auf 2 Stufen ergeben sich also 128 Kombinationen. Die statistische Versuchsplanung bietet effektive Möglichkeiten, um den Versuchsaufwand zu verringern, damit viele Faktoren oder nichtlineare Zusammenhänge mit vertretbarem Aufwand untersucht werden können.

Der Umgang mit vielen Variablen erfordert immer große Sorgfalt bei der Versuchsplanung und Versuchsdurchführung. In der Schulphysik gilt der Grundsatz, dass immer nur eine Einflussgröße verändert werden darf, während die anderen Einflussgrößen konstant bleiben müssen. Es geht um die eindeutige Zuordnung der Effekte zu den jeweiligen Faktoren. Diese Motivation ist absolut richtig, allerdings bietet die statistische Versuchsplanung eine alternative Strategie an, bei der trotz gleichzeitiger Variation mehrerer Faktoren ebenfalls eine eindeutige Zuordnung möglich ist. Die “ein Faktor nach dem anderen Methode” der Schulphysik hat einen gravierenden Nachteil, der in der Praxis oft zu Fehlinterpretationen führt.

Grundsätzlich wird genau ein Ausgangspunkt im Einstellbereich der Faktoren gewählt, typischerweise eine Ecke im Faktorraum². Alle Variationen beziehen sich auf diesen Ausgangspunkt. Unklar bleibt, wie das System reagiert, wenn man einen anderen Ausgangspunkt wählt. Letztlich wird bei der Interpretation der Ergebnisse vorausgesetzt, dass die Wirkung eines Faktors unabhängig von der Einstellung der anderen Faktoren ist. Reale Systeme verhalten sich jedoch oft anders. Hier liegt eine wesentliche Stärke der statistischen Versuchsplanung, denn sie untersucht gleichmäßig den gesamten Faktorraum. Jeder Faktor durchläuft mehrere Umstellvorgänge, ausgehend von unterschiedlichen Randbedingungen.

Wie stellt nun die statistische Versuchsplanung sicher, dass die Wirkungen der jeweiligen Faktoren getrennt voneinander untersucht werden können? Die Antwort steckt im der Konstruktion der Versuchspläne. *Orthogonal* ist ein Versuchsplan dann, wenn keine Kombination aus jeweils zwei Spalten miteinander korreliert. Anders ausgedrückt, die Einstellungsmuster aller Faktoren sind voneinander unabhängig. *Ausgewogen* ist ein Versuchsplan dann, wenn für die Faktorstufen jedes beliebigen Faktors die Einstellungen der anderen Faktoren gleichmäßig aufgeteilt sind. Sortiert man zum Beispiel alle Einstellungen des Versuchsplans nach *A*– und *A*+, dann taucht *B*– auf beiden Seiten gleich oft auf, *B*+ ebenfalls usw. . Daher ist

² Unter Faktorraum ist in diesem Zusammenhang ein mehrdimensionales Gebilde zu verstehen, das den Einstellbereich aller untersuchten Faktoren abbildet.

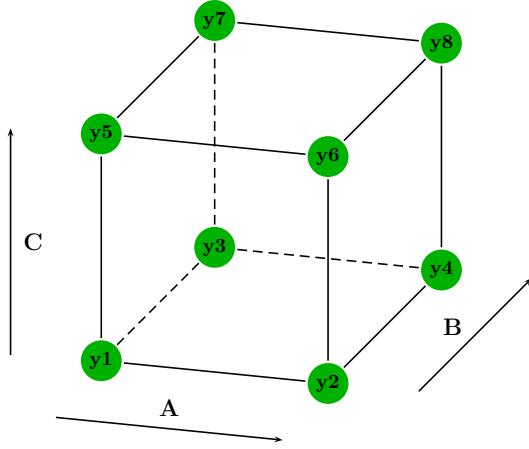


Abb. 1.4 Grafische Darstellung eines vollfaktoriellen Versuchsplans für drei Faktoren auf jeweils zwei Stufen.

es aus mathematischer Sicht kein Problem, eine kleine Wirkung eines Faktors zu erkennen, auch wenn die übrigen Faktoren eine wesentlich größere Wirkung haben. Standardisierte Versuchspläne erfüllen die beiden oben genannten Bedingungen³. Problematischer sind Spezialkonstruktionen oder Fälle, bei denen in der konkreten Durchführung nicht exakt der Versuchspran eingehalten wurde. Das Kapitel Kontrollverfahren wird darauf noch näher eingehen.

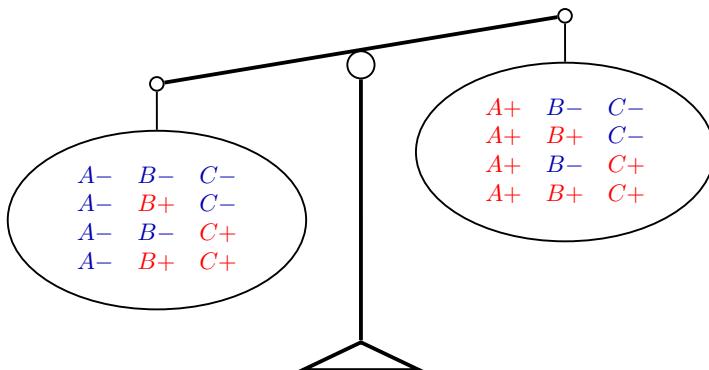


Abb. 1.5 Symbolische Darstellung der Ausgewogenheit eines Versuchsplans. Auf beiden Seiten tauchen alle Einstellungen der Faktoren B und C gleich häufig auf, daher kann auch ein kleiner Effekt von A sicher erkannt werden.

³ Versuchspläne für nichtlineare Zusammenhänge gewichten die Neutralstellung oft höher als die Randwerte, sind aber in jedem Fall orthogonal konstruiert.

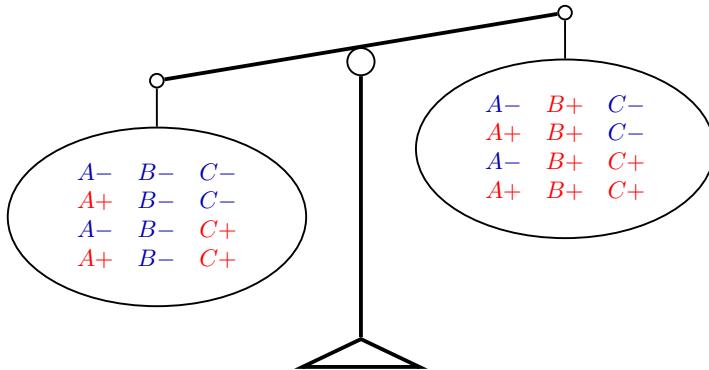


Abb. 1.6 Der gleiche Versuchsplan, sortiert nach den Einstellungen des Faktors B

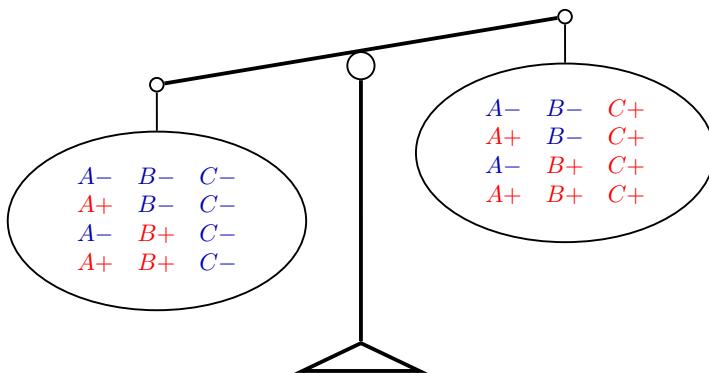


Abb. 1.7 Der gleiche Versuchsplan, sortiert nach den Einstellungen des Faktors C

1.3 Auswertung

In diesem Kapitel werden grundlegende Arbeitsschritte zur Auswertung von Versuchsplänen dargestellt. Effektendiagramme und Wechselwirkungsdiagramme sind standardisierte graphische Darstellungen der aus den Versuchsergebnissen abgeleiteten Beschreibungsfunktion. Die Vorgehensweise bei der Erstellung dieser Diagramme ist für alle Versuchspläne im Prinzip gleich und wird anhand einer Fallstudie erörtert.

1.3.1 Fallstudie

Um die Auswertung der Versuchsreihe zu verdeutlichen, dient im weiteren Verlauf eine Fallstudie. Dieses ist so gewählt, dass es auch aufwändige Untersuchun-



Abb. 1.8 Fallstudie Rasensprenger. Dieses technische System eignet sich hervorragend zur Anwendung der statistischen Versuchsplanung. Hohe Drehzahl und grosse Reichweite sind wichtige Kundenanforderungen aus der Sicht des Kindes. Niedriger Wasserverbrauch ist eine zusätzliche Kundenanforderung aus der Sicht der Eltern.

gen gestattet. Das Fallbeispiel *Rasensprenger* [9] wird in diesem Buch durchgängig eingesetzt, um dem Leser die Einarbeitung in viele, voneinander unabhängige Beispiele zu ersparen. Ein einheitliches Beispiel ermöglicht darüber hinaus den Direktvergleich verschiedener Versuchspläne und Optimierungsansätze, was sonst nicht möglich wäre.

Typischerweise hat ein Rasensprenger die Funktion, eine Fläche gleichmäßig zu bewässern. An warmen Sommertagen gesellt sich zu dieser Hauptfunktion noch die Nebenfunktion der Kinderbelustigung. Aus der Sicht der Kinder ist dies die Hauptfunktion und bei der angestrebten langen Nutzungsdauer gelangt mehr Wasser auf den Rasen, als eigentlich nötig wäre. Insgesamt lassen sich drei unabhängige Qualitätsmerkmale identifizieren: große Reichweite, hohe Drehzahl und geringer Wasserverbrauch. Betrachtet wird das System *Rasensprenger* ab Zuleitung hinter dem Absperrhahn. Die konstruktiven Parameter sind: vertikaler Düsenwinkel α , tangentialer Düsenwinkel β , Düsenquerschnitt A_q , Durchmesser, Reibung (trocken und flüssig) sowie der Wasserdruck. Bis auf die Faktorenauswahl und Stufenfestlegung ist nun alles Nötige definiert. Die Fallstudie kann mit elementaren Gleichungen der Strömungsmechanik und einigen vereinfachenden Annahmen numerisch gelöst werden. Die dargestellten Ergebnisse resultieren aus diesem nichtlinearen Simulationsmodell, was jedoch den Arbeitsablauf bei der Anwendung der statistischen Versuchsplanung in keiner Weise im Vergleich zu realen Experimenten verändert.

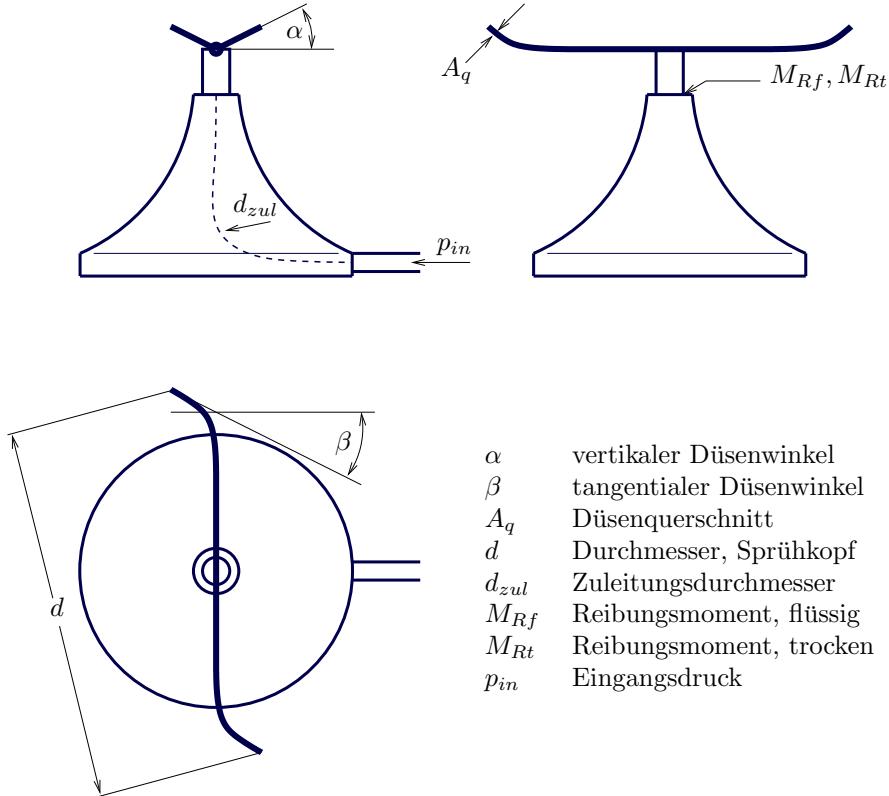


Abb. 1.9 Schematische Darstellung eines Rasensprengers. Bis zu acht Parameter sind frei einstellbar. Die verwendeten Gleichungen finden sich im Anhang. Ein eigens erstelltes Matlab/Octave Programm löst die Gleichungen numerisch und liefert Ergebnisse für jede Parameterkombination.

Faktor / Parameter	Symbol	Einstellung		Einheit
		-	+	
vertikaler Düsenwinkel	α	A	15	45
tangentialer Düsenwinkel	β	B	0	30
Düsenquerschnittsfläche	A_q	C	2	4 mm^2
Durchmesser, Sprühkopf	d	-	150	mm
Reibungsmom., trocken	M_{Rt}	-	0,015	Nm
Reibungsmom., flüssig	M_{Rf}	-	0,015	$\frac{\text{Nm}}{\text{s}}$
Eingangsdruck	p_{in}	-	1,5	bar
Zuleitungsdurchmesser	d_{zul}	-	7,5	mm

Tabelle 1.2 Einstellungstabelle für die erste Untersuchung. Als Faktoren variiert werden die Düsenwinkel und die Düsenquerschnittsfläche. Alle weiteren Parameter bleiben in einer mittleren Einstellung.

In der ersten Untersuchung beschränkt sich die Faktorenauswahl auf die Düsenwinkel und die Düsenquerschnittsfläche, wobei die übrigen Parameter konstant bleiben. Der einfache Vollfaktorplan besteht aus acht Kombinationen der drei Faktoren mit jeweils zwei Stufen. Jedes Qualitätsmerkmal ist eine kontinuierliche, skalare

Kenngröße und die drei Qualitätsmerkmale werden unabhängig voneinander ausgewertet.

A (α)	B (β)	C (A_q)	Drehzahl [1/s]	Reichweite [m]	Verbrauch [l/min]
—	—	—	4,1286	4,4088	4,1690
+	—	—	2,7671	5,0178	4,1535
—	+	—	3,4447	4,5387	4,1604
+	+	—	2,2742	5,0691	4,1495
—	—	+	8,0984	4,8512	8,3900
+	—	+	5,5804	6,4937	8,2846
—	+	+	6,8151	5,2425	8,3321
+	+	+	4,6980	6,6427	8,2565

Tabelle 1.3 Ergebnistabelle der ersten Untersuchung. Die drei Qualitätsmerkmale wurden gleichzeitig ermittelt, müssen aber unabhängig voneinander ausgewertet werden. Hierzu wird nacheinander jede der drei Ergebnisspalten als y_1, \dots, y_8 interpretiert.

1.3.2 Effekt

Die Wirkung eines Faktors auf das System wird durch den sogenannten *Effekt* gekennzeichnet. Als Effekt gilt die Differenz zweier Mittelwerte, dem Mittelwert bei der Einstellung + und dem Mittelwert bei der Einstellung – [6]. Der Effekt quantifiziert also die mittlere registrierte Veränderung des Qualitätsmerkmals, beim Wechsel der Faktoreneinstellung von – nach +. Dieses Verfahren heißt Kontrastmethode. Bereits bei diesem einfachen Versuchsplan erfolgen vier unabhängige Umschaltvorgänge für jeden Faktor, also hat der Effekt eine gewisse Stabilität in Bezug auf eventuelle Versuchsstreuungen und repräsentiert gleichzeitig mehrere Startbedingungen.

Der Effekt des Faktors A berechnet sich aus:

$$E_A = \frac{y_2 + y_4 + y_6 + y_8}{4} - \frac{y_1 + y_3 + y_5 + y_7}{4} \quad (1.2)$$

Für die Effektberechnung des Faktors B werden die gleichen Versuchswerte herangezogen, allerdings in einer anderen Gruppierung:

$$E_B = \frac{y_3 + y_4 + y_7 + y_8}{4} - \frac{y_1 + y_2 + y_5 + y_6}{4} \quad (1.3)$$

Analog ergibt sich der Effekt des Faktors C:

$$E_C = \frac{y_5 + y_6 + y_7 + y_8}{4} - \frac{y_1 + y_2 + y_3 + y_4}{4} \quad (1.4)$$

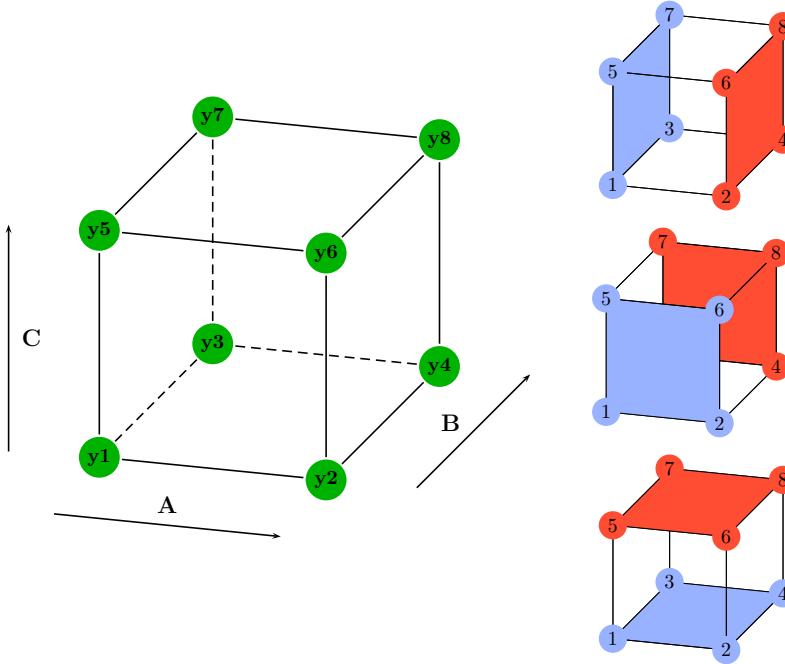


Abb. 1.10 Grafische Darstellung der Effektberechnung. Die Haupteffekte jedes Faktors berechnen sich aus der Differenz der jeweiligen Stufenmittelwerte. Die gleichen Versuchsdaten werden für jeden Faktor anders gruppiert, entsprechend der Faktoreinstellung.

A (α)	B (β)	C (A_q)	Drehzahl [1/s]
—	—	—	4,1286
+	—	—	2,7671
—	+	—	3,4447
+	+	—	2,2742
—	—	+	8,0984
+	—	+	5,5804
—	+	+	6,8151
+	+	+	4,6980
MW +	3,83	4,31	6,30
MW -	5,62	5,14	3,15
Effekt	-1,79	-0,84	3,14

Tabelle 1.4 Effektberechnung für das Qualitätsmerkmal Drehzahl. MW steht für *Mittelwert bei der jeweiligen Einstellung*

Das Effekt-Diagramm ist eine standardisierte Darstellung der Effekte. Alle Auswerteprogramme bieten diese Darstellung an und im Laufe der Jahre hat sich die Darstellung nur kosmetisch verändert. Auf der horizontalen Achse werden die Faktoren der Reihe nach aufgeführt, jeweils mit den untersuchten Stufen. Die Einheit

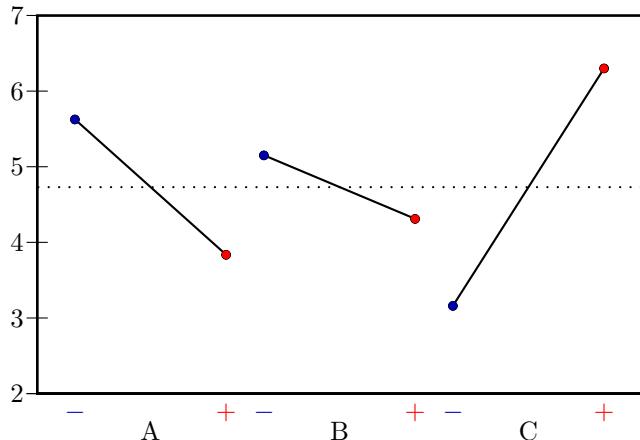


Abb. 1.11 Effekt-Diagramm am Beispiel des Qualitätsmerkmals Drehzahl.

ist durch die Stufenkodierung normiert (dimensionslos), denn die jeweiligen Faktoren können völlig unterschiedliche physikalische Einheiten besitzen und nur eine normierte Darstellung erlaubt den übersichtlichen Direktvergleich der Effekte.

Die vertikale Achse zeigt den Wert eines Qualitätsmerkmals in der jeweiligen Einheit. Bei mehreren Qualitätsmerkmalen ist es erforderlich, voneinander unabhängige Effekt-Diagramme zu erstellen. Nach passender Aufteilung der Achsen werden die Stufenmittelwerte eingetragen und für jeden Faktor getrennt mit einer direkten Linie verbunden. Die Steigung dieser Linie kennzeichnet den Effekt. Zur Kontrolle dient eine gestrichelte Linie auf der Höhe des Gesamtmittelwertes. Der Mittelwert der jeweiligen Stufenmittelwerte muss für jeden Faktor dem Betrag des Gesamtmittelwertes entsprechen, denn in beiden Berechnungen sind alle Versuchswerte enthalten. Mit anderen Worten: Alle Effektlinien müssen die gestrichelte Linie des Gesamtmittelwertes schneiden und zwar genau in der Mitte der jeweiligen Effektlinie. Auswerteprogramme arbeiten hierbei in der Regel fehlerfrei, deshalb kann die gestrichelte Hilfslinie bei automatisierter Auswertung wegfallen.

Die Fallstudie zeigt deutlich, wie unterschiedlich die Effekte der gleichen Faktoren auf verschiedene Qualitätsmerkmale sein können. Der steilere vertikale Düsenwinkel reduziert die Drehzahl, erhöht die Reichweite und hat nur einen geringen Einfluss auf den Wasserverbrauch. Bezuglich des optimalen Düsenquerschnitts zeichnet sich ein Zielkonflikt ab, denn ein großer Querschnitt bewirkt gleichzeitig eine hohe Drehzahl und einen hohen Wasserverbrauch⁴.

⁴ Der Wasserverbrauch soll jedoch möglichst niedrig sein.

	A (α)	B (β)	C (A_q)	Reichweite [m]
	—	—	—	4, 4088
	+	—	—	5, 0178
	—	+	—	4, 5387
	+	+	—	5, 0691
	—	—	+	4, 8512
	+	—	+	6, 4937
	—	+	+	5, 2425
	+	+	+	6, 6427
MW +	5, 81	5, 37	5, 81	
MW —	4, 76	5, 19	4, 76	
Effekt	1, 05	0, 18	1, 05	

Tabelle 1.5 Effektberechnung für das Qualitätsmerkmal Reichweite.

	A (α)	B (β)	C (A_q)	Verbrauch [l/min]
	—	—	—	4, 1690
	+	—	—	4, 1535
	—	+	—	4, 1604
	+	+	—	4, 1495
	—	—	+	8, 3900
	+	—	+	8, 2846
	—	+	+	8, 3321
	+	+	+	8, 2565
MW +	6, 21	6, 22	8, 32	
MW —	6, 26	6, 25	4, 16	
Effekt	-0, 05	-0, 02	4, 16	

Tabelle 1.6 Effektberechnung für das Qualitätsmerkmal Wasserverbrauch.

1.3.3 Wechselwirkung

Der Effekt kennzeichnet die mittlere Veränderung des Qualitätsmerkmals in Folge einer Stufenvariation. Dies schließt aber nicht aus, dass eine Abhängigkeit vom Ausgangszustand ⁵ bestehen kann. In der Tat ist dies in der Praxis oft zu beobachten. Bereits im einfachen Beispiel des Rasensprengers ergibt sich eine derartige Abhängigkeit. Hier hängt der Einfluss des Düsenquerschnitts auf die Drehzahl von den Düsenwinkeln ab. Wenn der Effekt eines Faktors von der Einstellung eines anderen Faktors abhängt, nennt man dies eine Wechselwirkung oder einen Wechselwirkungseffekt. Um den Effekt eines Faktors begrifflich davon abzugrenzen, wird dieser auch Haupteffekt genannt.

Ohne die Grenzen der Physik zu überschreiten, lässt sich ein drastisches Beispiel für starke Wechselwirkungen konstruieren. In diesem Fall besteht das System aus einem umbauten Raum, mit dem Ziel einer möglichst guten Schallabsorption exter-

⁵ also der Einstellung der übrigen Faktoren

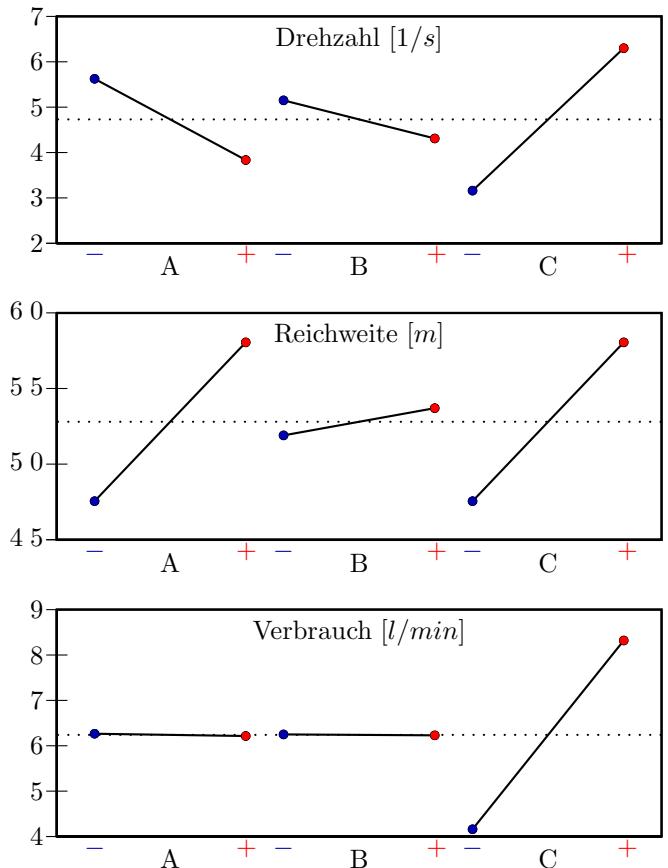


Abb. 1.12 Effekt-Diagramm für alle Qualitätsmerkmale im Direktvergleich. Die Effekte der jeweiligen Faktoren auf die Qualitätsmerkmale können völlig unterschiedlich sein.

ner Geräusche. Nehmen wir an, dieser Raum habe zwei Fenster, dann lassen sich daraus zwei Faktoren ableiten: Fenster 1 und Fenster 2, jeweils mit den Einstellungen *offen* und *geschlossen*. Die subjektive Beurteilung der Schalldämmung auf einer Skala von 1 bis 10 könnte bei guten Fenstern ein klares Ergebnis liefern.

A	B	QM
-	-	1
+	-	2
-	+	2
+	+	10

Tabelle 1.7 Ergebnistabelle einer fiktiven Fallstudie. Das Qualitätsmerkmal *Schallabsorption* hängt von den beiden Faktoren *Fenster 1* und *Fenster 2* ab. Nur wenn beide Fenster geschlossen sind, bleibt der Lärm draußen.

In dieser Konstellation hängt die Wirkung der Faktoren stark von der Einstellung des jeweiligen anderen Faktors ab. Ist das andere Fenster offen, bleibt die erreichte

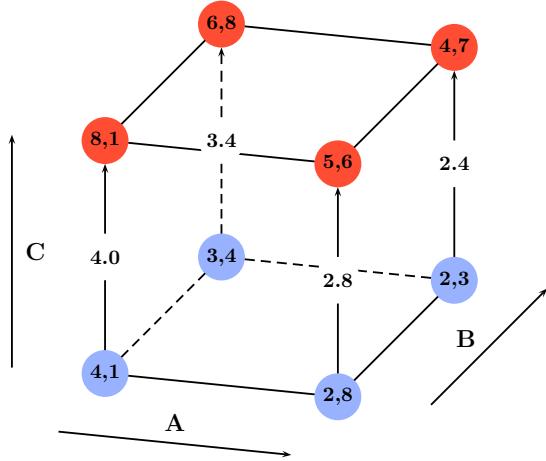


Abb. 1.13 Einfluss des Düsenquerschnitts auf die Drehzahl des Rasensprengers. In Abhängigkeit von den Randbedingungen ergibt sich ein unterschiedlicher Wert. Es liegen sogenannte Wechselwirkungen vor.



Abb. 1.14 Gedankenexperiment zur Verdeutlichung einer starken Wechselwirkung. Ausgehend vom Ziel einer guten Schallabsorption, gibt es nur eine Lösung: beide Fenster müssen geschlossen bleiben. (The Sims 3 ©2009 Electronic Arts Inc. The Sims is a trademark or registered trademark of Electronic Arts Inc. in the U.S. and/or other countries. All Rights Reserved. Used with permission.)

Schalldämmung nahezu gleich. Bei geschlossenem zweiten Fenster hingegen ändert sich das Qualitätsmerkmal Schalldämmung signifikant. Die Berechnungsmethode ist unkompliziert. Zunächst erzeugt man die *Wechselwirkungsspalte*, also eine neue Spalte zur Berechnung des Wechselwirkungseffekts. Die Effekte der Faktoren werden nun zur besseren Abgrenzung *Haupteffekte* genannt. Die Wechselwirkungsspalte zeigt an, ob die Faktoren auf gleicher (+) oder ungleicher (-) Stufe stehen. Der Wechselwirkungseffekt berechnet sich analog zum Haupteffekt aus der Differenz der Stufenmittelwerte. Ein großer Wechselwirkungseffekt lässt auf eine starke

Wechselwirkung schließen. Man kann den Wechselwirkungseffekt auch als Einfluss der Randbedingungen interpretieren.

A	B	AB	QM
—	—	+	1
+	—	—	2
—	+	—	2
+	+	+	10
MW +	6,0	6,0	5,5
MW —	1,5	1,5	2,0
Effekt	4,5	4,5	3,5

Tabelle 1.8 Berechnung der Wechselwirkungsspalte $A \times B$. Die aus der Effektberechnung bekannte Kontrastmethode lässt sich analog anwenden. Der Wechselwirkungseffekt gibt an, wie stark die für eine Randbedingung aufgeteilte Effektiline im Wechselwirkungsschaubild vom Haupteffekt abweicht.

Im konkreten Beispiel beträgt der Haupteffekt von A 4,5. Dies ist der gemittelte Wert für beide Randbedingungen (Fenster B offen und Fenster B geschlossen). Bezogen auf die einzelnen Randbedingungen berechnet sich der Effekt von A aus dem mittleren Effekt und der Wechselwirkung mit dem jeweiligen Vorzeichen.

$$E_{A(B-)} = E_A - E_{AB} = 4,5 - 3,5 = 1 \quad (1.5)$$

$$E_{A(B+)} = E_A + E_{AB} = 4,5 + 3,5 = 8 \quad (1.6)$$

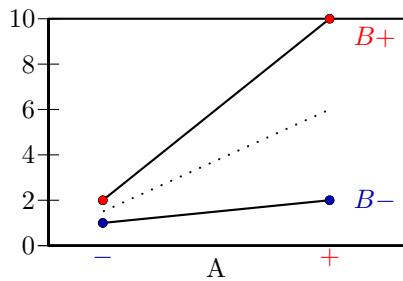


Abb. 1.15 Schematische Darstellung eines Wechselwirkungsdiagramms. Der Effekt eines Faktors (hier A) wird für zwei unterschiedliche Randbedingungen dargestellt, in diesem Fall B− und B+. Die punktierte Hilfslinie zeigt den Haupteffekt von A und dient hier nur zur Orientierung. Die Wechselwirkung entspricht der Steigungsänderung. Sind die eingezeichneten Linien parallel, gibt es keine Wechselwirkung. Bei starken Wechselwirkungen können sich die Linien auch kreuzen.

In der konkreten Anwendung ist es oft hilfreich, die Stufen in Richtung der zu erwartenden Veränderung zu kodieren, zum Beispiel – für die derzeitige Standardeinstellung und + für die nach Experteneinschätzung bestmögliche neue Einstellung des Faktors. Dies hat den Vorteil, dass die Vorzeichen der Haupteffekte sofort interpretierbar sind. Auch die Vorzeichen der Wechselwirkungseffekte kann man in

diesem Fall klar zuordnen. Hat der Wechselwirkungseffekt das gleiche Vorzeichen wie die Haupteffekte, wirkt die Wechselwirkung verstärkend, ansonsten abschwächend. Letzteres ist in der Praxis häufiger zu beobachten und bedeutet, dass sich bei gleichzeitiger Umstellung mehrerer Faktoren insgesamt ein geringerer Effekt einstellt, als es die Summe der Haupteffekte vermuten lässt.

Angenommen, es gäbe zwei konstruktive Maßnahmen zur Verringerung der Insassenbelastung beim Frontalaufprall und jede einzelne Maßnahme reduziert den Belastungskennwert HIC um 100, dann ist es sehr wahrscheinlich, dass beide Maßnahmen bei gleichzeitigem Einsatz einen etwas geringeren Effekt bringen, zum Beispiel 160. Dies steht auch im Einklang mit physikalischen Überlegungen, denn ein gutes Rückhaltesystem bietet im Vergleich zum schlechten Rückhaltesystem ein geringeres Verbesserungspotential. In anderen Bereichen der Technik gibt es analoge Beobachtungen, zum Beispiel in der Motorenentwicklung.

Bei zweistufigen Versuchsplänen ist die Konstruktion der "Wechselwirkungsspalten" einfach und ohne Auswerteprogramm möglich. Hierzu interpretiert man eine Stufeneinstellung als -1 , die andere als $+1$. Nun müssen lediglich die Werte der ausgewählten Spalten zeilenweise miteinander multipliziert werden. Sind die Einstellungen gleichsinnig (also $-/-$ oder $+/+$) erscheint ein $+$, ansonsten (also $-/+$ oder $+/-$) ein $-$.

Bei drei Faktoren entsteht möglicherweise eine Dreifachwechselwirkung. Die zugehörige Kontrastspalte ergibt sich analog zur Zweifachwechselwirkung durch Multiplikation der betroffenen Spalten. Für alle Spalten ist die Kontrastmethode zur Auswertung anwendbar. Das aus den Spalten gebildete Feld ist bei Vollfaktorplänen immer orthogonal, da alle neu gebildeten Spalten weder untereinander noch mit den bereits existierenden Spalten korrelieren.

A	B	C	AB	AC	BC	ABC	y
$-$	$-$	$-$	$+$	$+$	$+$	$-$	y_1
$+$	$-$	$-$	$-$	$-$	$+$	$+$	y_2
$-$	$+$	$-$	$-$	$+$	$-$	$+$	y_3
$+$	$+$	$-$	$+$	$-$	$-$	$-$	y_4
$-$	$-$	$+$	$+$	$-$	$-$	$+$	y_5
$+$	$-$	$+$	$-$	$+$	$-$	$-$	y_6
$-$	$+$	$+$	$-$	$-$	$+$	$-$	y_7
$+$	$+$	$+$	$+$	$+$	$+$	$+$	y_8

Tabelle 1.9 Erzeugung der Wechselwirkungsspalten bei einem Versuchsplan mit drei Faktoren und acht Versuchsläufen. Es entstehen vier neue Spalten, drei Spalten für die Zweifachwechselwirkungen und eine für die Dreifachwechselwirkung.

Tabelle 1.10 zeigt das Ergebnis des Qualitätsmerkmals *Reichweite*. Nur die Wechselwirkung $A \times C$ erreicht einen deutlichen Betrag, ist aber kleiner als die Haupteffekte der beteiligten Faktoren. Zur Erstellung eines Wechselwirkungsdiagramms muss zunächst der vorhandene Versuchsplan auf einen Vollfaktorplan der beiden untersuchten Faktoren zurückgeführt werden. Am Beispiel $A \times B$ wird deut-

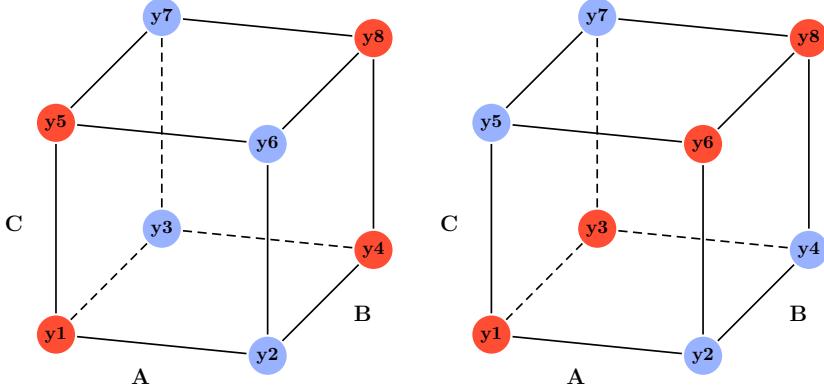


Abb. 1.16 Graphische Darstellung der Wechselwirkungen $A \times B$ und $A \times C$. Die Punkte gleichen Vorzeichens liegen auf einer Diagonalen in der Ebene AB (links) bzw. AC (rechts).

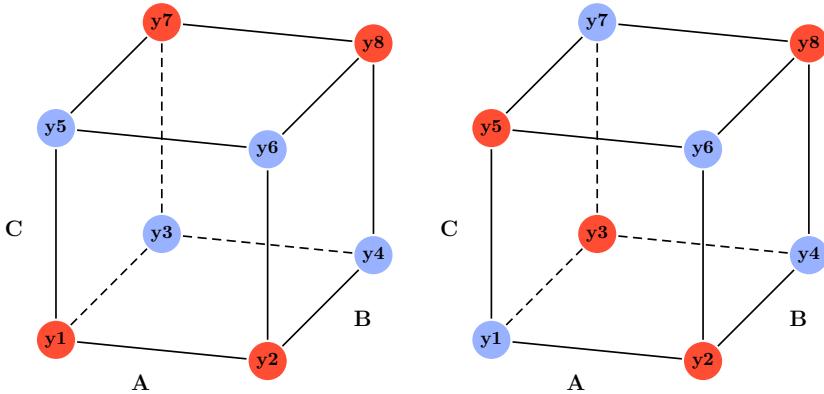


Abb. 1.17 Graphische Darstellung der Wechselwirkungen $B \times C$ und $A \times B \times C$. Die Punkte gleichen Vorzeichens liegen auf einer Diagonalen in der Ebene BC (links) bzw. bilden einen Tetraeder (rechts).

lich, dass alle Kombinationen von A und B im Versuchsplan mit acht Versuchsläufen doppelt auftauchen, jeweils für $C-$ und $C+$. Für die Auswertung interessieren nun die Mittelwerte aus den beiden Einstellungen. Das Wechselwirkungsdiagramm bildet genau diese vier Punkte ab, also $A-/A+$ bei $B-$ und $A-/A+$ bei $B+$. Auswerteprogramme erstellen diese Schaubilder automatisch.

Jede Zweifachwechselwirkung lässt sich auf zwei verschiedene Weisen darstellen. Entweder ist der Effekt des ersten Faktors bei zwei verschiedenen Randbedingungen des zweiten Faktors aufgetragen oder der Effekt des zweiten Faktors bei verschiedenen Randbedingungen des ersten Faktors. Welche Darstellung gewählt wird, hängt vom Anwendungsfall ab. Bei automatisierter Auswertung ist der Aufwand so gering, dass man im Zweifelsfall für die signifikanten Wechselwirkungen immer beide Darstellungen produziert, um den Sachverhalt zu veranschaulichen.

Versuch	<i>A</i>	<i>B</i>	<i>C</i>	<i>AB</i>	<i>AC</i>	<i>BC</i>	<i>ABC</i>	<i>Reichweite</i>
1	—	—	—	+	+	+	—	4,409
2	+	—	—	—	—	+	+	5,018
3	—	+	—	—	+	—	+	4,539
4	+	+	—	+	—	—	—	5,069
5	—	—	+	+	—	—	+	4,851
6	+	—	+	—	+	—	—	6,494
7	—	+	+	—	—	+	—	5,243
8	+	+	+	+	+	+	+	6,643
MW +	5,81	5,37	5,81	5,24	5,52	5,33	5,26	
MW —	4,76	5,19	4,76	5,32	5,05	5,24	5,30	
Effekt	1,05	0,18	1,05	-0,08	0,48	0,09	-0,04	

Tabelle 1.10 Auswertung des Qualitätsmerkmals *Reichweite*.

1.3.4 Lineares Beschreibungsmodell

Haupteffekte und Wechselwirkungen quantifizieren die Wirkung der Faktoren auf das Qualitätsmerkmal. Letztlich entsteht eine beschreibende Gleichung, die das zu Grunde liegende physikalische Phänomen zwar nicht erklärt, aber die Zusammenhänge quantifiziert. In vielen Fällen reicht dies in der Praxis völlig aus. Experten haben oft bereits eine sehr gute qualitative Vorstellung der Zusammenhänge "ihres" Systems. Häufig sind die Systeme viel zu komplex für analytische Verfahren oder man hat nur eine geringe Aussicht darauf, alle erforderlichen Eingangsdaten für die exakte analytische Modellierung zu bekommen.

Ein lineares Beschreibungsmodell mit drei Faktoren sieht folgendermaßen aus:

$$y = c_0 + c_1x_1 + c_2x_2 + c_3x_3 + c_{12}x_1x_2 + c_{13}x_1x_3 + c_{23}x_2x_3 + \varepsilon \quad (1.7)$$

Es liefert für jede Kombination der Eingangsgrößen x_1, x_2, x_3 einen Näherungswert des Qualitätsmerkmals y . $c_0 \dots c_{23}$ sind Modellkonstanten, in der Literatur oft mit β bezeichnet, jedoch ist c in der Ingenieurswelt gebräuchlicher. Die Abweichung ε ist bei einem passenden Modell klein, im Vergleich zur Variation des Qualitätsmerkmals. Für eine beliebige Zahl von Faktoren ergibt sich die Gleichung für das Beschreibungsmodell analog. Bei der Summation ist die Symmetrie der Koeffizientenmatrix entlang der Hauptdiagonale zu beachten.⁶ Ferner dürfen die Wechselwirkungen nicht doppelt aufsummiert werden.

$$y = c_0 + \sum_{i=1}^{n_f} c_i x_i + \sum_{i=1}^{n_f-1} \sum_{j=i+1}^{n_f} c_{ij} x_i x_j + \varepsilon \quad (1.8)$$

⁶ $c_{ij} = c_{ji}$

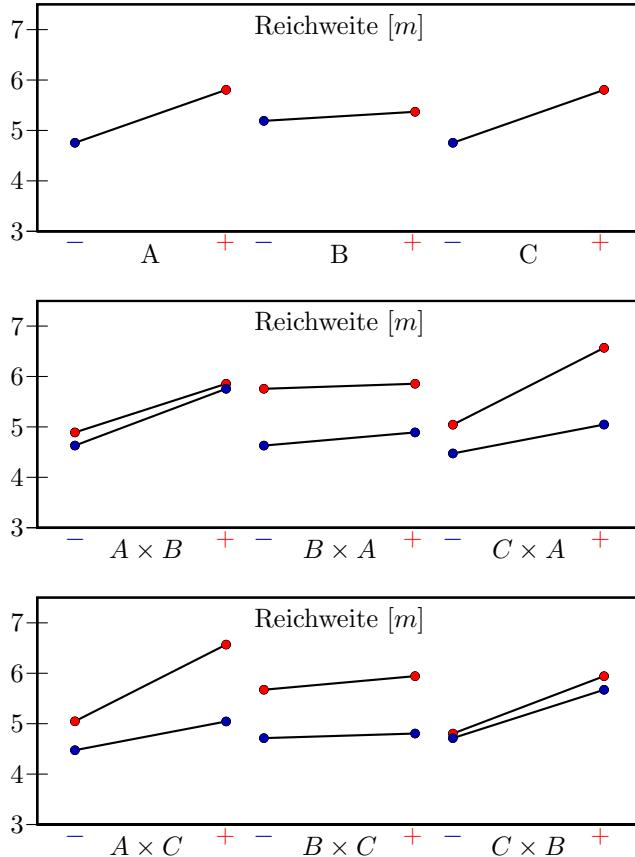


Abb. 1.18 Haupteffekte und Zweifachwechselwirkungen der Fallstudie für das Qualitätsmerkmal *Reichweite*. Die Wechselwirkungen zeigen den Effekt des erstgenannten Faktors bei den verschiedenen Einstellungen des zweitgenannten Faktors, also unter verschiedenen Randbedingungen.

Die Zahl der Modellkonstanten erhöht sich mit jedem Faktor um die aktuelle Zahl der Faktoren. Die Summation über alle Faktoren ergibt dann die Gesamtzahl der Modellkonstanten

$$n_m = 1 + \sum_{i=1}^{n_f} i \quad \text{mit} \quad n_{m_i} = n_{m_{i-1}} + n_{f_i} \quad (1.9)$$

Im linearen Beschreibungsmodell sind alle Haupteffekte und Wechselwirkungseffekte enthalten. Darüber hinaus taucht als Konstante der Gesamtmittelwert des Qualitätsmerkmals über die Versuchsreihe auf. Bei normierten Wertebereichen der Faktoren von -1 bis 1 durchläuft jeder Faktor in der Beschreibungsgleichung die Stufenbreite 2, also müssen alle Effekte durch 2 dividiert werden, um von den Effekten zu den Modellkonstanten zu kommen. Alle Konstanten haben dann die phy-

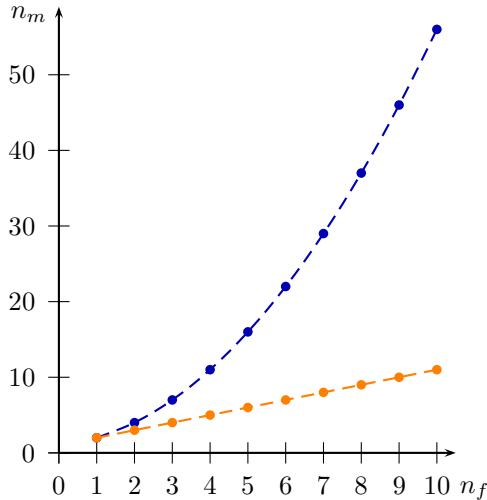


Abb. 1.19 Anstieg der Zahl der Modellkonstanten mit der Anzahl der Faktoren. Unter Berücksichtigung der Wechselwirkungen ist der Anstieg progressiv (dunkle Kurve), ansonsten linear. In jedem Fall bleibt die Zahl unter der Anzahl der Versuche eines Vollfaktorplans. Bei 10 Faktoren auf jeweils zwei Stufen verlangt dieser 1024 Versuche.

sikalische Einheit des Qualitätsmerkmals. \hat{y} bezeichnet die Vorhersage des Qualitätsmerkmals für die gewählte Einstellung, \bar{y} den Gesamtmittelwert.

$$\hat{y} = \bar{y} + \frac{E_A}{2}x_1 + \frac{E_B}{2}x_2 + \frac{E_C}{2}x_3 + \frac{E_{AB}}{2}x_1x_2 + \frac{E_{AC}}{2}x_1x_3 + \frac{E_{BC}}{2}x_2x_3 \quad (1.10)$$

Auch eine nicht normierte Beschreibung ist möglich, verliert aber an Übersichtlichkeit, weil die Konstanten unterschiedliche physikalische Einheiten bekommen können und sich die Werte der Konstanten wegen der individuellen Stufenbreiten mitunter schlecht vergleichen lassen. Auswerteprogramme bieten beide Varianten an. Man muss sich bei der Versuchsplanerstellung im Programm für eine der beiden Varianten entscheiden.

Trotz des einfachen Aufbaus ist das lineare Beschreibungsmodell im praktischen Einsatz erstaunlich leistungsfähig. Dies hat verschiedene Gründe:

1. Die statistische Versuchsplanung setzt auf die Strategie, die Wirkung vieler Faktoren möglichst einfach zu beschreiben. Im Gegensatz dazu konzentrieren sich andere Verfahren⁷ auf wenige Faktoren und beschreiben deren Einfluss komplex. In der Praxis wird die zulässige Zahl der Modellkonstanten schnell durch die maximal mögliche Zahl der Versuchsläufe beschränkt, also wird man selten viele Faktoren komplex beschreiben können.
2. Die Nichtlinearität der Zusammenhänge wird oft überschätzt. Wenn die Stufenabstände nicht allzu groß sind, liefert die Linearisierung oft überraschend gute Ergebnisse.
3. Die Wechselwirkungseffekte werden oft unterschätzt. In fast allen technischen Systemen spielen Wechselwirkungen eine beträchtliche Rolle und sind häufig viel dominanter als Nichtlinearitäten der Effekte einzelner Faktoren.

⁷ Zum Beispiel die Methode der neuronalen Netze. [2]

4. Das lineare Modell ist anschaulich und kommunizierbar. Ein hochkomplexes Beschreibungsmodell ersetzt letztlich eine “black box” durch eine andere “black box”, während sich das vergleichsweise simple lineare Modell hervorragend eignet, um das Systemverhalten jedermann verständlich zu machen, insbesondere den Entscheidungsträgern. Dies ist ein wertvoller Beitrag, um wissenschaftliche Methoden auch in Zukunft fest in der Praxis zu verankern [5].

Das lineare Beschreibungsmodell erfordert einen zweistufigen Versuchsplan, kann aber auch über Regressionsverfahren aus den Ergebnissen von Versuchsplänen mit mehr als zwei Stufen gebildet werden. In jedem Fall ist die im Modell vorgenommene Interpolation nur bei kontinuierlich einstellbaren Faktoren sinnvoll. Extrapolationen sind grundsätzlich unzulässig, weil sich außerhalb der getesteten Einstellbereiche neue physikalische Wirkungen ergeben können oder sonstige Unstetigkeiten das Systemverhalten möglicherweise drastisch beeinflussen.

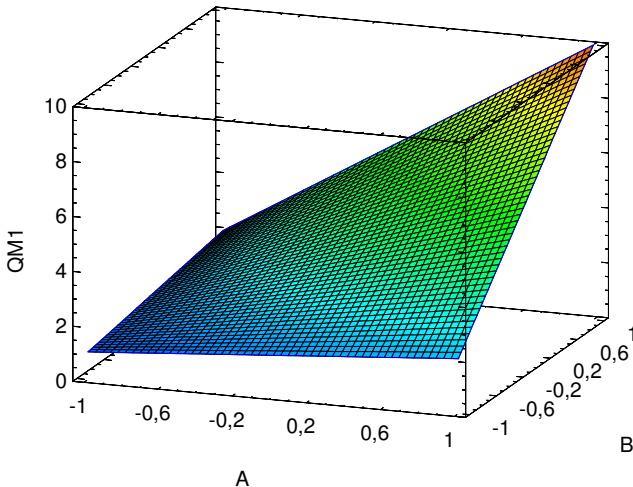


Abb. 1.20 Darstellung des linearen Beschreibungsmodells am Beispiel: Abhängigkeit des Qualitätsmerkmals *Schalldämpfung* von den Faktoren A und B.

Bei nichtlinearen Zusammenhängen arbeitet das Modell ungenau. Allerdings kann ein zweistufiger Versuchsplan nachträglich erweitert werden, um die nötigen Informationen für ein quadratisches Modell zu liefern. Schnelle Abhilfe bringt immer ein reduzierter Stufenabstand, sofern die Kontraste dann noch ausreichend genau messbar sind und die Stufenabstände der Faktoren noch immer zueinander passen. Als Behelfslösung kann eine Achsenttransformation sehr wirksam sein, insbesondere dann, wenn für einzelne Faktoren der physikalische Zusammenhang zum Qualitätsmerkmal analytisch ableitbar oder anderweitig bekannt ist. In diesem Fall bildet man die Faktoreinstellung x auf eine Zwischengröße \tilde{x} ab und berechnet das lineare Modell mit \tilde{x} .

Literaturverzeichnis

1. Bandemer, H., Bellmann, A., Jung, W., Richter, K.: *Optimale Versuchsplanung*. Verlag Harri Deutsch, Zürich, Frankfurt am Main, Thun (1976) 1
2. Bothe, H.H.: *Neuro-Fuzzy Methoden*. Springer Verlag, Berlin Heidelberg New York (1998) 23, 48
3. Fisher, R.A.: *The Design of Experiments*. Oliver and Boyd, Edinburgh and London (1935) 1, 91, 104
4. Fowlkes, W., Creveling, C.: *Engineering Methods for Robust Product Design*. Addison-Wesley, Reading, MA (1995) 2, 34, 58, 139
5. Gauch, H.H.: *Scientific Method in Practice*. Cambridge University Press, Cambridge New York (2003) 2, 24
6. Grove, D., Davis, T.: *Engineering, Quality and Experimental Design*. Longman Scientific and Technical, Harlow (1992) 12
7. Kleppmann, W.: *Taschenbuch Versuchsplanung*. Carl Hanser Verlag, München (2008) 6
8. Montgomery, D.C.: *Design and Analysis of Experiments*. John Wiley, New York (2001) 2, 32
9. Rutten, K., Siebertz, K., van Bebber, D., Hochkirchen, T., Ketelaere, B.D.: *The Garden Sprinkler: An Interactive Web-Based Application For Teaching Response Surface Methodology*. In: European Network for Business and Industrial Statistics, ENBIS-13. Ankara (2013) 10, 159
10. Toutenburg, H., Gössl, R., Kunert, J.: *Quality Engineering, Eine Einführung in Taguchi-Methoden*. Prentice Hall, München (1997) 2

Kapitel 2

Versuchspläne

2.1 Einleitung

Oft wird die statistische Versuchsplanung fast ausschließlich mit der Konstruktion von Versuchsplänen in Verbindung gebracht. In der Tat ist dies ein sehr wichtiger und eigenständiger Teil der Methode. Im Gegensatz zu den Anfängen der statistischen Versuchsplanung, bieten die verfügbaren Auswerteprogramme eine hervorragende Unterstützung mit vorkonfektionierten Feldern und beherrschen vielfach auch die Erstellung maßgeschneiderter Versuchspläne für den speziellen Anwendungsfall. Wichtig ist nach wie vor die Vermittlung der Strategien hinter den jeweiligen Feldkonstruktionen, damit der Anwender weiß, welche Auswahlmöglichkeit besteht.

Nach Anwendungsfall gruppiert, behandelt dieses Kapitel alle gängigen Feldkonstruktionen. Zunächst stehen Screening-Versuchspläne auf dem Programm, mit denen eine hohe Zahl von Faktoren untersucht werden kann. Detailuntersuchungen werden oft mit einem quadratischen Beschreibungsmodell durchgeführt, um den vorhandenen Nichtlinearitäten Rechnung zu tragen. Auch diese Modelle haben Grenzen, was in einem eigenen Abschnitt diskutiert wird. Mischungspläne verwendet man oft in der Verfahrenstechnik, denn sie berücksichtigen die Randbedingung, dass bei Mischungen die Summe aller Anteile der beteiligten Komponenten 100% ergibt. In Sonderfällen sind maßgeschneiderte Versuchspläne erforderlich. Für die automatische Erstellung dieser Versuchspläne gibt es mehrere Optimierungskriterien, die in einem eigenen Abschnitt vorgestellt werden. Als kleiner Exkurs in die Geschichte der Versuchsplanung bilden die umstrittenen Latin Squares den Abschluss dieses Kapitels.

2.2 Screening Versuchspläne

Zu den wesentlichen Stärken der statistischen Versuchsplanung gehört Effizienz, also die Möglichkeit, mit minimalem Versuchsaufwand viele Faktoren zu untersuchen. Hierzu gibt es speziell konstruierte Versuchspläne, die nahezu alle in der Praxis auftretenden Anforderungen abdecken und eine sichere Analyse gewährleisten. Nur in Ausnahmefällen ist eine Sonderkonstruktion nötig. In diesem Abschnitt wird zunächst die grundsätzliche Strategie dieser Versuchspläne erläutert. Anschließend erfolgt eine Vorstellung der gebräuchlichen Feldkonstruktionen mit Direktvergleich der Ergebnisse anhand eines Fallbeispiels.

2.2.1 Konzept

Bei einer hohen Zahl von Faktoren ist der Vollfaktorplan nicht mehr durchführbar. Screening Versuchspläne haben die Aufgabe, bei minimalem Informationsverlust mit möglichst wenigen Versuchen auszukommen. In der Literatur finden sich dafür verschiedene Bezeichnungen, unter anderem: screening designs, fractional factorial designs, Screening Versuchspläne, teilstatistische Versuchspläne, Teilstatistische Pläne oder fraktionelle faktorielle Versuchspläne.

Grundsätzlich stellt der Versuchsplan ein lineares Gleichungssystem dar. Jeder Versuch liefert eine Gleichung. Daher ist es möglich, Beschreibungsmodelle anzupassen, deren Parameterzahl der Zahl der Versuchsläufe entspricht. Günstiger ist jedoch ein Überschuss an Gleichungen. Dies hat den Vorteil, dass eine Kontrolle des Beschreibungsmodells möglich ist. Einzelheiten dazu finden sich im Kapitel *Kontrollverfahren*.

Ausgehend von einem Vollfaktorplan für vier Faktoren auf jeweils zwei Stufen, lässt sich ein Beschreibungsmodell mit 16 Konstanten aufstellen. Eine Konstante ist der Gesamtmittelwert, vier Konstanten entfallen auf die Haupteffekte, sechs Konstanten auf die Zweifachwechselwirkungen, vier auf die Dreifachwechselwirkungen und eine auf die Vierfachwechselwirkung. Unter der Annahme, dass die Terme höherer Ordnung keine signifikanten Werte erreichen, sind letztlich nur zehn Modellkonstanten relevant. Die Feldkonstruktion liefert jedoch 15¹ orthogonale Spalten.

Nur vier dieser 15 Spalten werden als Einstellungsmuster für die Faktoren genutzt, und zwar die Spalten mit den Haupteffekten. Die verbleibenden Spalten dienen zunächst nur dazu, die Modellkonstanten höherer Ordnung zu berechnen. Hier setzt die Strategie der Teilstatistischen Pläne an und deklariert eine für das Beschreibungsmodell unbedeutende Spalte zur Einstellungsvorschrift für den nächsten Faktor. Dieser Strategie folgend, besteht ebenso die Möglichkeit, von einem kleineren Feld auszugehen und es mit einem zusätzlichen Faktor zu belegen.

¹ Zur Berechnung des Gesamtmittelwertes könnte man eine 16. Spalte bilden, die vollständig mit + kodiert ist.

A	B	AB	C	AC	BC	ABC	D	AD	BD	ABD	CD	ACD	BCD	$ABCD$	y
-	-	+	-	+	+	-	-	+	+	-	+	-	-	-	y_1
+	-	-	-	-	+	+	-	-	+	+	+	+	-	-	y_2
-	+	-	-	+	-	+	-	+	-	+	+	-	+	-	y_3
+	+	+	-	-	-	-	-	-	-	-	+	+	+	-	y_4
-	-	+	+	-	-	-	+	+	+	-	+	+	+	-	y_5
+	-	-	+	+	-	-	-	-	+	+	-	-	+	+	y_6
-	+	-	+	-	+	-	-	+	-	+	-	+	-	+	y_7
+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	y_8
-	-	+	+	+	-	+	-	-	-	+	-	+	+	-	y_9
+	-	-	-	+	+	+	-	-	-	-	-	-	+	+	y_{10}
-	+	-	-	+	-	+	-	-	-	-	+	-	-	+	y_{11}
+	+	+	-	-	-	-	+	+	+	+	-	-	-	-	y_{12}
-	-	+	+	-	-	+	-	-	+	+	-	-	-	+	y_{13}
+	-	-	+	-	-	+	-	-	-	+	+	-	-	-	y_{14}
-	+	-	+	-	+	-	+	-	-	+	-	+	-	-	y_{15}
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	y_{16}

Tabelle 2.1 Vollfaktorieller Versuchsplan mit vier Faktoren auf zwei Stufen und 16 Versuchen. Jede Zeile liefert eine Gleichung und in jeder Spalte steht eine Unbekannte.

A BCD	B ACD	AB CD	C ABD	AC BD	BC AD	ABC D	y
-	-	+	-	+	+	-	y_1
+	-	-	-	-	+	+	y_2
-	+	-	-	+	-	+	y_3
+	+	+	-	-	-	-	y_4
-	-	+	+	-	-	+	y_5
+	-	-	+	+	-	-	y_6
-	+	-	+	-	+	-	y_7
+	+	+	+	+	+	+	y_8

Tabelle 2.2 Teilstudieller Versuchsplan mit vier Faktoren auf zwei Stufen und 8 Versuchen. Jede Zeile liefert eine Gleichung und in jeder Spalte steht eine Summe aus zwei Unbekannten.

Der Vollfaktorplan für drei Faktoren auf jeweils zwei Stufen besteht aus acht Versuchsläufen und liefert sieben orthogonale Spalten. Drei dieser Spalten sind durch Haupteffekte besetzt, drei durch Zweifachwechselwirkungen. In der siebten Spalte befindet sich die als vernachlässigbar klein eingestufte Dreifachwechselwirkung. Nutzt man diese Spalte als Einstellungsvorschrift für den vierten Faktor, dann erhöht sich natürlich auch die Zahl der Unbekannten. Durch Multiplikation der entsprechenden Spalten lässt sich leicht nachweisen, dass Haupteffekte und Dreifachwechselwirkungen in den gleichen Spalten stehen. In den übrigen Spalten befinden sich jeweils zwei Zweifachwechselwirkungen. Letztlich entsteht ein Gleichungssystem mit 16 Unbekannten und 8 Gleichungen, bei dem jeweils zwei Unbekannte auf der linken Seite stehen. Eine Trennung der jeweiligen Zweiergruppen ist unmöglich, die Zweiergruppen sind miteinander vermischt. In der Praxis ist dies jedoch weniger schwerwiegend, als es zunächst erscheint, da die Terme höherer Ordnung in guter Näherung zu Null gesetzt werden können. Gesamtmittelwert und Haupteffekte las-

sen sich also sicher bestimmen. Bei den Zweifachwechselwirkungen bleibt jedoch die Unsicherheit der Zuordnung. Das ist der Preis für die gesteigerte Effizienz.

Trotz der Einschränkung ist das gezeigte Feld mit acht Versuchsläufen bei vier Faktoren in der Praxis extrem erfolgreich. Immer dann, wenn es auf eine schnelle Durchführung der Versuchsreihe ankommt, kann dieser Versuchsplan punkten. Acht Versuchsläufe bilden die untere Grenze für eine statistische Auswertung, vier Faktoren finden sich praktisch immer und die Einstellungsmuster sind relativ simpel. Es gibt erfolgreiche Anwendungsberater der statistischen Versuchsplanung, die nie ein anderes Feld eingesetzt haben.

<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>y</i>
—	—	+	—	+	+	—	<i>y</i> ₁
+	—	—	—	—	+	+	<i>y</i> ₂
—	+	—	—	+	—	+	<i>y</i> ₃
+	+	—	—	—	—	—	<i>y</i> ₄
—	—	+	+	—	—	+	<i>y</i> ₅
+	—	—	+	+	—	—	<i>y</i> ₆
—	+	—	+	—	+	—	<i>y</i> ₇
+	+	+	+	+	+	+	<i>y</i> ₈

Tabelle 2.3 Teilstudieller Versuchsplan mit sieben Faktoren auf zwei Stufen und 8 Versuchen. Die Wechselwirkungen sind nun sogar mit den Haupteffekten vermengt. Trotzdem ist dieser Versuchsplan sinnvoll, um eine hohe Zahl von Faktoren mit geringem Aufwand zu sichten.

Im Extremfall lässt sich dieses Feld mit sieben Faktoren belegen. Dann sind alle Spalten als Einstellungsvorschrift genutzt und nur 6,25% aller möglichen Kombinationen werden getestet. Eine Auswertung der Wechselwirkungen ist in diesem Fall völlig unmöglich und alle Haupteffekte sind mit mehreren Zweifachwechselwirkungen vermengt. Das Feld ist nun *gesättigt*. Trotz dieser Einschränkungen ist der Versuchsplan leistungsfähiger als die traditionelle “ein Faktor nach dem anderen Methode”, die bei sieben Faktoren exakt gleich viele Versuchsläufe benötigt. Jeder Faktor wird bei diesem Versuchsplan viermal verstellt, also reduziert sich bei der Effektberechnung durch Mittelwertbildung die Versuchsstreuung. Außerdem erfolgt die Verstellung jeweils aus einer unterschiedlichen Ausgangsposition. Der traditionelle Ansatz kann diese Vorteile nicht bieten, ohne dass sich der Aufwand vervielfacht.

Unter Vernachlässigung der Terme höherer Ordnung benötigt das lineare Beschreibungsmodell bei fünf Faktoren genau 16 Konstanten, also sollte es möglich sein, mit 16 Versuchen alle erforderlichen Informationen zu bekommen. Der gezeigte Versuchsplan erfüllt genau diese Aufgabe. Im Vergleich zum Vollfaktorplan ergibt sich immerhin eine Ersparnis von 50%, da nur jede zweite Kombination vorkommt. In den Zeilen 1,2,3,5 und 9 befinden sich übrigens die Einstellungen der “ein Faktor nach dem anderen Methode”, insofern kann dieser Versuchsplan hilfreich sein, wenn man sich erst spät für die statistische Versuchsplanung entscheidet oder auf jeden Fall die Variation der einzelnen Faktoren testen möchte ². Dieser

² Bei Computermodellen tritt dieser Fall mitunter auf, wenn die Variation einzelner Faktoren mit einer gravierenden Modelländerung einhergeht.

A	B	AB	C	AC	BC	DE	D	AD	BD	CE	CD	BE	AE	E	y
—	—	+	—	+	+	—	—	+	+	—	+	—	—	+	y_1
+	—	—	—	—	+	+	—	—	+	+	+	+	—	—	y_2
—	+	—	—	+	—	+	—	+	—	+	+	—	+	—	y_3
+	+	+	—	—	—	—	—	—	—	—	+	+	+	+	y_4
—	—	+	+	—	—	+	—	+	+	—	—	+	+	—	y_5
+	—	—	+	+	—	—	—	—	+	+	—	—	+	+	y_6
—	+	—	+	—	+	—	—	+	—	+	—	+	—	+	y_7
+	+	+	+	+	+	—	—	—	—	—	—	—	—	—	y_8
—	—	+	—	+	+	—	+	—	—	+	—	+	+	—	y_9
+	—	—	—	—	+	+	+	+	—	—	—	—	+	+	y_{10}
—	+	—	—	+	—	+	+	—	+	—	—	+	—	+	y_{11}
+	+	+	—	—	—	—	+	+	+	+	—	—	—	—	y_{12}
—	—	+	+	—	—	+	+	—	—	+	+	—	—	+	y_{13}
+	—	—	+	+	—	—	+	—	—	—	+	—	—	—	y_{14}
—	+	—	+	—	+	—	+	—	+	—	+	—	+	—	y_{15}
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	y_{16}

Tabelle 2.4 Teilfaktorieller Versuchsplan mit fünf Faktoren auf zwei Stufen und 16 Versuchen. Alle Haupteffekte und Zweifachwechselwirkungen sind sicher voneinander trennbar.

Versuchsplan ist sehr gut auszuwerten [10]. Alle Haupteffekte und Wechselwirkungen liegen frei und das Feld ist groß genug für eine sichere statistische Analyse. Mit fünf Faktoren besteht meistens ausreichender Spielraum für die Untersuchung der wichtigsten Parameter und der Gesamtaufwand bleibt im praktikablen Rahmen. Jeder Faktor wird acht Mal auf + und acht Mal auf — getestet. Die Effektberechnung ist daher so stabil, dass in der Regel auf eine Versuchswiederholung verzichtet werden kann.

Der Begriff Auflösung bewertet die Vermengungsstruktur. Man unterscheidet grob in vier Grundtypen. Die Auflösungsstufen sind international genormt. Auswerteprogramme geben die Auflösungsstufen der zur Auswahl stehenden Felder in der Regel an.

Auflösung	Eigenschaften
III	Haupteffekte sind mit Zweifachwechselwirkungen vermischt. Das Feld ist dicht besetzt und nur zum Screening geeignet.
IV	Haupteffekte sind mit Dreifachwechselwirkungen vermischt und Zweifachwechselwirkungen untereinander. Das Feld ist geeignet, um Haupteffekte sicher zu bestimmen, Zweifachwechselwirkungen lassen sich jedoch nicht eindeutig zuordnen.
V	Haupteffekte sind mit Vierfachwechselwirkungen vermischt und Dreifachwechselwirkungen mit Zweifachwechselwirkungen. Ein derartiges Feld kann ohne Schwierigkeiten das lineare Beschreibungsmodell versorgen.
V+	Haupteffekte und Zweifachwechselwirkungen sind praktisch unvermischt. Im Vergleich zur Auflösungsstufe V ist jedoch kein großer Genaugkeitsgewinn zu erwarten.

Tabelle 2.5 Auflösung von Versuchsplänen.

Die Auflösungsstufen (engl. resolution) sind international genormt und alle Auswerteprogramme kennzeichnen die zur Auswahl stehenden Felder entsprechend. Der Versuchsplan nach dem YATES-Standard mit acht Versuchen und vier Faktoren hat die Auflösungsstufe IV. Bei einer Belegung mit sieben Faktoren reduziert sich die Auflösung auf Stufe III. Das Feld mit 16 Versuchen hat bei einer Belegung mit fünf Faktoren die Auflösungsstufe V.

Einen Hinweis auf die Auflösungsstufe liefert der sogenannte Generator des Feldes. Der Generator ist die Kombination der Faktorspalten, die bei einer Multiplikation eine Spalte erzeugt, die nur + enthält [7]. Beim Feld mit 8 Versuchsläufen und vier Faktoren ist dies die Kombination $ABCD$. Multipliziert man eine Spalte mit sich selbst, entsteht ebenfalls diese “Identitätsspalte” I. Daraus lässt sich eine Rechenregel ableiten, um die vermengten Effekte zu finden: Multiplikation mit dem Generator und Kürzung der doppelt auftretenden Spalten. In unserem Beispiel ist A mit BCD vermischt, denn $AABCD$ entspricht BCD . Dieser Generator liefert die Auflösungsstufe IV. Je nach Belegung gibt es mehrere Generatoren. Der kürzeste Generator bestimmt die Auflösungsstufe.

In der Literatur finden sich verschiedene Bezeichnungen für die gleichen Versuchspläne. Üblich ist die Bezeichnung nach Zahl der möglichen Kombinationen mit Angabe der Reduktionsstufe. 2^{7-4} beispielsweise kennzeichnet einen Versuchsplan für sieben Faktoren auf zwei Stufen, mit der Reduktionsstufe 4, also werden $2^{7-4} = 2^3 = 8$ Versuchsläufe nötig sein. 2^{5-1} kennzeichnet einen Versuchsplan für fünf Faktoren, mit 16 Versuchsläufen und der Reduktionsstufe 1. Hierbei werden 50% aller Kombinationen getestet.

Oft werden die Felder auch einfach nach ihrer Größe benannt, mit dem Zusatz L als Symbol für die vorhandene Orthogonalität. L8 beispielsweise bezeichnet den Versuchsplan nach YATES mit acht Zeilen und sieben Spalten, unabhängig von der Belegung. Die nächst größere Konstruktion heißt L16, gefolgt von L32, L64 usw. .

2.2.2 Reguläre Felder nach dem Yates-Standard

Die bislang vorgestellten Felder sind nach dem YATES-Standard aufgebaut. Frank Yates hat in den Dreißigerjahren des vorigen Jahrhunderts eine leicht erlernbare Systematik entwickelt, um beliebig große orthogonale Felder zu konstruieren. Das Prinzip erinnert ein wenig an die ineinander verschachtelten russischen Holzpüppchen, denn in jedem größeren Feld sind kleinere Felder enthalten. Die Verdopplung erfolgt durch Addition einer Zusatzspalte und vierfacher Kopie des Ausgangselementes, wobei eine Kopie invertiert wird. Die Zusatzspalte ist in der oberen Hälfte mit – besetzt und in der unteren Hälfte mit + .

Auch dieses Feld lässt sich schnell verdoppeln, wenn man es als neues Ausgangselement auffasst und die gleichen Verdopplungsregeln anwendet. Je nach Bedarf entstehen auf diese Weise beliebig große Felder, die alle orthogonal und ausgewogen sind. Versuchspläne nach diesem Standard haben lediglich den Nachteil, dass die Größenabstufung recht grob ist (4, 8, 16, 32, 64, ...).

$\begin{array}{c} \text{Das Ausgangsfeld:} \\ \begin{array}{c} \text{--- +} \\ \text{+ ---} \\ \text{--- +} \\ \text{+ ++} \end{array} \end{array}$	$\text{Das vergrößerte Feld:} \quad \left \begin{array}{c} \text{--- +} \\ \text{+ ---} \\ \text{--- +} \\ \text{+ ++} \\ \hline \text{--- +} \\ \text{+ ---} \\ \text{--- +} \\ \text{+ ++} \end{array} \right $
---	--

Abb. 2.1 Verdopplung der Feldgröße von vier auf acht Versuche durch Kopie des Basisfeldes. Der YATES-Standard stellt sicher, dass auch das vergrößerte Feld orthogonal ist.

$\begin{array}{c} \text{--- +} \\ \text{+ ---} \\ \text{--- +} \\ \text{+ ++} \\ \hline \text{--- +} \\ \text{+ ---} \\ \text{--- +} \\ \text{+ ++} \end{array}$	\rightarrow	$\left \begin{array}{c} \text{--- +} \quad \text{+ + -} \\ \text{+ ---} \quad \text{+ + +} \\ \text{--- +} \quad \text{+ + -} \\ \text{+ ++} \quad \text{+ + +} \\ \hline \text{--- +} \quad \text{+ + +} \\ \text{+ ---} \quad \text{+ + +} \\ \text{--- +} \quad \text{+ + +} \\ \text{+ ++} \quad \text{+ + +} \end{array} \right $
--	---------------	---

Abb. 2.2 Verdopplung der Feldgröße von acht auf 16 Versuche durch Kopie des vergrößerten Basisfeldes.

2.2.3 Irreguläre Felder nach Plackett-Burman

PLACKETT und BURMAN haben 1946 eine neue Feldkonstruktion veröffentlicht, die bis heute äußerst erfolgreich ist. Ernsthaftige Konkurrenz dazu ist im Bereich der zweistufigen Felder nicht in Sicht. Die PLACKETT-BURMAN-Konstruktion erzeugt sogenannte irreguläre Felder der Auflösungsstufe III. Irregulär heißt, dass das Produkt zweier Spalten nur zu diesen Spalten orthogonal ist, aber nicht zu den anderen Spalten des Feldes. Die Haupteffekte sind daher mit Anteilen der Zweifachwechselwirkungen vermengt. Wechselwirkungen schlagen also nicht zu 100% in die Haupteffektberechnung durch, verfälschen aber alle Haupteffekte der nicht an der jeweiligen Zweifachwechselwirkung beteiligten Faktoren. $A \times B$ beispielsweise findet sich abgeschwächt in allen Haupteffekten, außer in A und B .

PLACKETT-BURMAN-Felder entstehen durch zyklische Vertauschung der Kodierungskette von Spalte zu Spalte. Im abgebildeten Beispiel fällt das dadurch entstehende diagonale Streifenmuster auf. Die letzte Zeile wird jedoch für alle Spalten auf $-$ gesetzt. Die Länge der Kodierungskette bestimmt die Feldgröße. PLACKETT-BURMAN-Felder gibt es in sehr feinen Abstufungen von 8 - 96 Versuchsläufen mit

A	B	C	D	E	F	G	H	I	J	K	y
+	-	+	-	-	-	+	+	+	-	+	y_1
+	+	-	+	-	-	-	+	+	+	-	y_2
-	+	+	-	+	-	-	-	+	+	+	y_3
+	-	+	+	-	+	-	-	-	+	+	y_4
+	+	-	+	+	-	+	-	-	-	+	y_5
+	+	+	-	+	+	-	+	-	-	-	y_6
-	+	+	+	-	+	+	-	+	-	-	y_7
-	-	+	+	+	-	+	+	-	+	-	y_8
-	-	-	+	+	-	+	+	+	-	+	y_9
+	-	-	-	+	+	+	-	+	+	-	y_{10}
-	+	-	-	-	+	+	+	-	+	+	y_{11}
-	-	-	-	-	-	-	-	-	-	-	y_{12}

Tabelle 2.6 Versuchsplan nach PLACKETT-BURMAN mit 12 Versuchen und bis zu 11 Faktoren. Alle Spalten sind orthogonal, also bleiben die Haupteffekte auch bei voller Belegung eindeutig trennbar. Die Wechselwirkungen sind jedoch abgeschwächt mit den Haupteffekten vermengt.

dem Inkrement 4. Sehr gängig sind die Felder für 12, 20 und 24 Versuche, offenbar weil diese Größe in der Praxis am häufigsten benötigt wird. Die grundsätzlichen Eigenschaften der Felder sind jedoch von der Größe unabhängig.

TAGUCHI hat sich oft der PLACKETT-BURMAN-Felder bedient, diese aber umsortiert, um die Versuchsdurchführung zu vereinfachen[5]. Orthogonale Felder bleiben orthogonal, auch wenn man Zeilen und Spalten vertauscht. Diese Regel lässt sich ausnutzen, um die Zahl der Stufenwechsel zu minimieren. Oft sind einzelne Faktoren nur mit hohem Aufwand zu verstehen, andere hingegen mit wenig Mühe. Die vorsortierten Versuchspläne tragen dem Rechnung und bieten eine von links nach rechts ansteigende Zahl von Stufenwechseln an. Aufwändige Faktoren kommen in die erste Spalte und müssen nur noch ein einziges Mal verstellt werden.

Ein PLACKETT-BURMAN-Feld mit 12 Versuchen wird auch $L12$ genannt und lässt sich mit bis zu 11 Faktoren belegen. In diesem Fall testet man weniger als 0,6% aller möglichen Kombinationen, kann aber in der Regel die signifikanten Faktoren entdecken. Empfehlenswert ist, 1-3 Spalten unbelegt zu lassen, um die auftretende Verfälschung durch kummulierte Wechselwirkungsanteile abschätzen zu können. Die leeren Spalten enthalten nur Überlagerungen der Zweifachwechselwirkungen, die belegten Spalten die Summe aus Haupteffekt und Überlagerung.

Bei größeren Feldern wird die Effizienzsteigerung noch dramatischer. Das mit 19 Faktoren belegte $L20$ -Feld nach PLACKETT-BURMAN testet nur noch 0,0038% aller möglichen Kombinationen, bietet aber dennoch eine sehr stabile Effektberechnung, weil für jeden Faktor bei allen Stufen Daten mit völlig unterschiedlichen Ausgangspositionen vorliegen.

In manchen Fällen reicht die Auflösungsstufe III nicht aus, weil die auftretenden Wechselwirkungen zu stark sind. Dann ist die Auflösungsstufe IV erforderlich. Es gibt eine einfache Methode, um *jedes beliebige Feld* der Auflösungsstufe III in

<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>H</i>	<i>I</i>	<i>J</i>	<i>K</i>	<i>L</i>	<i>M</i>	<i>N</i>	<i>O</i>	<i>P</i>	<i>Q</i>	<i>R</i>	<i>S</i>	<i>y</i>
+	-	+	+	-	-	-	-	+	-	+	-	+	+	+	+	-	-	+	<i>y</i> ₁
+	+	-	+	+	-	-	-	+	-	+	-	+	+	+	+	-	-	-	<i>y</i> ₂
-	+	+	-	+	+	-	-	-	+	-	+	-	+	+	+	+	-	-	<i>y</i> ₃
-	-	+	+	-	+	+	-	-	-	+	-	+	-	+	+	+	+	+	<i>y</i> ₄
+	-	-	+	+	-	+	+	-	-	-	+	-	+	-	+	+	+	+	<i>y</i> ₅
+	+	-	-	+	+	-	+	+	-	-	-	+	-	+	-	+	+	+	<i>y</i> ₆
+	+	+	-	+	+	-	+	+	-	-	-	+	-	+	-	+	-	+	<i>y</i> ₇
+	+	+	+	-	+	+	-	+	+	-	-	-	+	-	+	-	+	-	<i>y</i> ₈
-	+	+	+	+	-	+	+	-	+	+	-	-	-	+	-	+	-	+	<i>y</i> ₉
+	-	+	+	+	-	+	+	-	+	+	-	-	-	-	+	-	-	-	<i>y</i> ₁₀
-	+	-	+	+	+	-	+	+	-	+	+	-	-	-	-	-	+	-	<i>y</i> ₁₁
+	-	+	-	+	+	-	+	+	-	+	+	-	-	-	-	-	-	-	<i>y</i> ₁₂
-	+	-	+	-	+	+	+	-	+	+	-	+	-	-	-	-	-	-	<i>y</i> ₁₃
-	-	+	-	+	-	+	+	+	-	+	-	+	-	+	-	+	-	-	<i>y</i> ₁₄
-	-	-	+	-	+	-	+	+	-	+	-	+	-	+	-	+	-	-	<i>y</i> ₁₅
-	-	-	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	-	<i>y</i> ₁₆
+	-	-	-	-	+	-	+	-	+	+	-	+	-	+	-	+	-	-	<i>y</i> ₁₇
+	+	-	-	-	-	+	-	+	-	+	+	+	-	-	+	+	-	-	<i>y</i> ₁₈
-	+	+	-	-	-	-	+	-	+	-	+	+	+	-	-	+	+	+	<i>y</i> ₁₉
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	<i>y</i> ₂₀

Tabelle 2.7 Versuchsplan nach PLACKETT-BURMAN mit 20 Versuchen und bis zu 19 Faktoren. Das Feld enthält weniger als 0,004% aller Kombinationen, ist also sehr effizient.

die Auflösungsstufe IV zu überführen. Die sogenannte Faltung (fold over) besteht aus einer schlichten Kopie des ursprünglichen Versuchsplans mit gleichzeitiger Invertierung.³ Durch diesen Trick lassen sich alle Vermengungen von Haupteffekten und Zweifachwechselwirkungen auflösen. Für die Praxis bedeutet dies, dass man getrost mit einem Versuchsplan der Auflösungsstufe III die Untersuchung beginnen kann und erst im Bedarfsfall mit einer zweiten Versuchsreihe gleichen Umfangs die Haupteffekte freilegen muss. Hierbei geht kein Ergebnis verloren, denn der erweiterte Versuchsplan beinhaltet auch die vorangegangene Versuchsreihe.

2.2.4 Fallstudie

Vor der Entscheidung für einen Versuchsplan stellt sich natürlich die Frage, inwie weit das Ergebnis eines Teilstukturplans vom Ergebnis des Vollfaktorplans abweicht. Welcher Versuchsplan ist der beste? Kann man durch die Wahl des “falschen” Versuchsplans auf die falsche Fährte gelangen? Einsteiger in die Methode der statistischen Versuchsplanung sind hier oft sehr verunsichert, insbesondere dann, wenn sie unmittelbar vor der Entscheidung einen Einblick in die Vielfalt der Versuchsplä-

³ Dies bedeutet: Vorzeichenwechsel bei der Kodierung.

<i>A</i>	<i>B</i>	<i>E</i>	<i>C</i>	<i>F</i>	<i>G</i>	<i>D</i>	<i>BE</i>	<i>AE</i>	<i>AB</i>	<i>AF</i>	<i>AC</i>	<i>AD</i>	<i>AG</i>	<i>y</i>
							<i>CF</i>	<i>CG</i>	<i>CD</i>	<i>BG</i>	<i>BD</i>	<i>BC</i>	<i>BF</i>	
							<i>DG</i>	<i>DF</i>	<i>FG</i>	<i>DE</i>	<i>EG</i>	<i>EF</i>	<i>CE</i>	
-	-	-	-	+	+	+	-	-	-	-	+	+	+	<i>y</i> ₁
+	-	-	-	-	+	+	+	-	-	-	-	+	+	<i>y</i> ₂
-	+	-	-	+	-	+	-	+	-	-	+	-	+	<i>y</i> ₃
+	+	-	+	-	-	+	+	-	+	-	-	-	+	<i>y</i> ₄
-	-	+	+	-	-	+	-	-	+	+	-	-	+	<i>y</i> ₅
+	-	+	-	+	-	+	-	+	-	+	-	-	+	<i>y</i> ₆
-	+	+	-	-	+	-	-	+	+	-	-	+	-	<i>y</i> ₇
+	+	+	+	+	+	+	+	+	+	+	+	+	+	<i>y</i> ₈
+	+	+	-	-	-	+	-	-	-	+	+	+	-	<i>y</i> ₉
-	+	+	+	+	-	-	+	-	-	-	-	+	+	<i>y</i> ₁₀
+	-	+	+	-	+	-	-	-	+	-	+	-	+	<i>y</i> ₁₁
-	-	+	-	+	+	+	-	-	+	-	-	-	-	<i>y</i> ₁₂
+	+	-	-	+	+	-	-	-	+	-	-	-	+	<i>y</i> ₁₃
-	+	-	+	-	+	+	-	-	+	-	+	-	-	<i>y</i> ₁₄
+	-	-	+	+	-	+	-	-	+	-	-	+	-	<i>y</i> ₁₅
-	-	-	-	-	-	-	+	+	+	+	+	+	-	<i>y</i> ₁₆

Tabelle 2.8 Gefalteter Versuchsplan mit nun 16 Versuchen und 7 Faktoren. Das ursprüngliche Feld mit acht Versuchen wurde erweitert. Dadurch lassen sich die Wechselwirkungen von den Haupteffekten trennen.

ne erhalten haben⁴. Jedes Problem ist anders gelagert und es gibt kein Kochrezept für die Auswahl des optimalen Versuchsplans. Der nachfolgende Direktvergleich mehrerer Versuchspläne an einem konkreten Beispiel zeigt jedoch, dass man auf verschiedene Weise zum gleichen Ziel gelangen kann. Die Methode der statistischen Versuchsplanung ist verblüffend robust und oft auch dann erfolgreich, wenn der Anwender sie nicht vollständig verstanden hat.

Symbol	Parameter	Einheit	Einstellung		
			-	0	+
A	α	o	15	45	
B	β	o	0	30	
C	A_q	mm ²	2	4	
D	d	mm	100	200	
E	M_{Ri}	Nm	0,01	0,02	
F	M_{Rf}	Nm	0,01	0,02	
G	p_{in}	bar	1	2	
	d_{zul}	mm		7,5	

Tabelle 2.9 Einstellungstabelle. Sieben der acht Parameter wurden in der folgenden Studie variiert. Der Zuleitungsdurchmesser blieb konstant.

⁴ Manchmal ist es besser, wenn die Auswahl nicht zu groß ist. Deshalb wird in diesem Buch auch der Ansatz verfolgt, nur “gute” Versuchspläne vorzustellen.

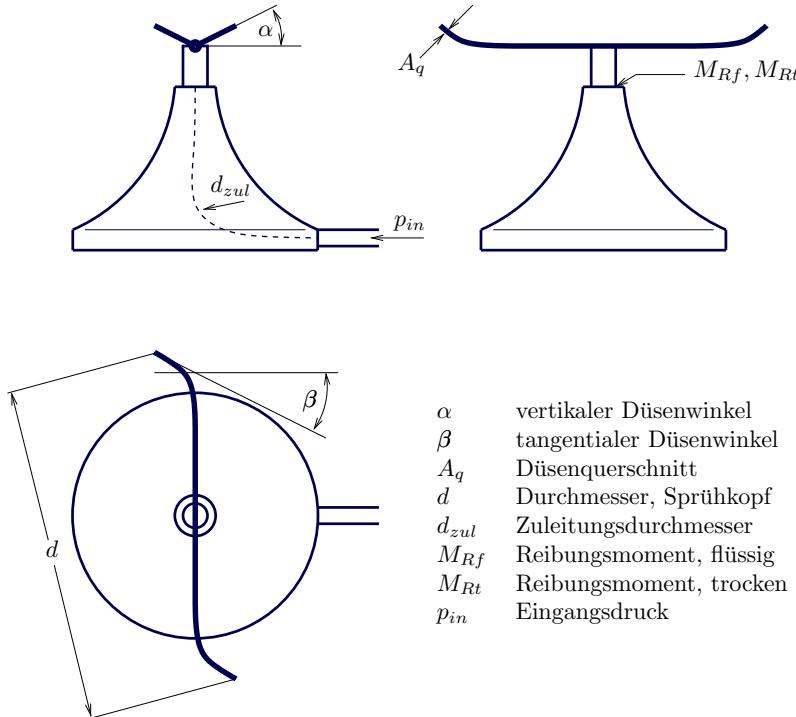


Abb. 2.3 Schematische Darstellung eines Rasensprengers.

Das bereits vorgestellte Rasensprengerbeispiel bietet die Möglichkeit, sieben Faktoren zu variieren. Der zugehörige Versuchsplan besteht aus 128 Versuchsläufen. Das vollbesetzte L_8 Feld nach YATES und das L_{12} -Feld nach PLACKETT-BURMAN nutzen jeweils nur einen Bruchteil dieser Kombinationen. Im Folgenden wird aufgezeigt, wie sich die Haupteffekte im Direktvergleich darstellen, wobei der Vollfaktorplan quasi die “Musterlösung” liefert.

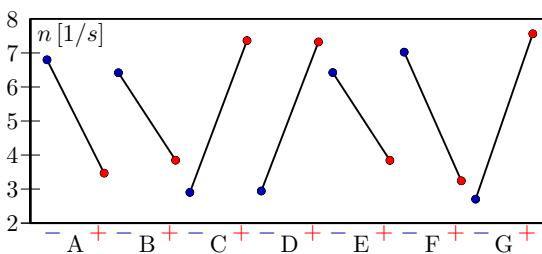


Abb. 2.4 Ergebnis des teilst faktoriellen Versuchsplans mit 8 Einstellungen (L_8). Auswertung des Qualitätsmerkmals *Drehzahl*.

Die Abweichungen liegen in Anbetracht des drastisch reduzierten Versuchsaufwandes bei beiden Teilst faktorplänen sehr nah an den Ergebnissen des Vollfaktor-

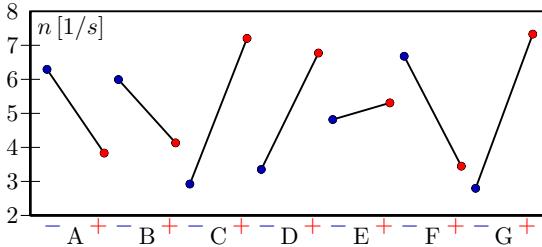


Abb. 2.5 Ergebnis des teilfaktoriellen Versuchsplans mit 12 Einstellungen (L12). Auswertung des Qualitätsmerkmals *Drehzahl*.

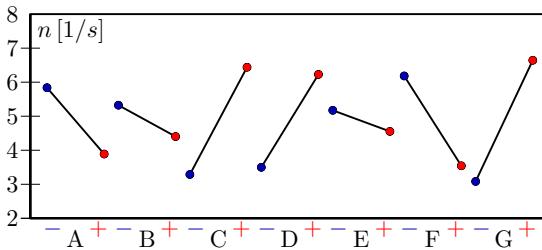


Abb. 2.6 Ergebnis des teilfaktoriellen Versuchsplans mit 128 Einstellungen (L128). Auswertung des Qualitätsmerkmals *Drehzahl*.

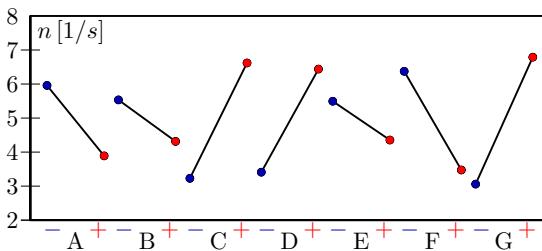


Abb. 2.7 Ergebnis des teilfaktoriellen Versuchsplans mit 16 Einstellungen (L8 + fold over). Auswertung des Qualitätsmerkmals *Drehzahl*.

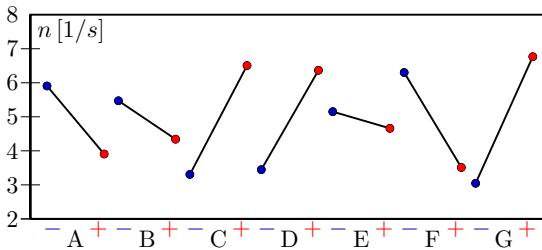


Abb. 2.8 Ergebnis des teilfaktoriellen Versuchsplans mit 24 Einstellungen (L12 + fold over). Auswertung des Qualitätsmerkmals *Drehzahl*.

plans. Die stärksten Effekte werden in der richtigen Reihenfolge erkannt und haben das richtige Vorzeichen. Unsicher ist lediglich die Berechnung des Effektes von Faktor E. Das L8-Feld überschätzt den Effekt, während das L12-Feld den Effekt etwa in der richtigen Stärke, aber mit falschem Vorzeichen angibt. Dies liegt an der Vermengung der Haupteffekte mit Wechselwirkungen. In beiden Fällen wird durch Faltung die Genauigkeit stark verbessert, ohne den Versuchsaufwand allzu stark in die Höhe zu treiben. Die Angst, durch die falsche Wahl des Versuchsplans alle Chancen auf den Erfolg zu verspielen, ist also völlig unbegründet.

2.3 Versuchspläne für ein quadratisches Beschreibungsmodell

Bei nichtlinearen Zusammenhängen zwischen Faktor und Qualitätsmerkmal stößt das lineare Beschreibungsmodell an seine Grenzen. Mit dem quadratischen Beschreibungsmodell bietet sich eine leistungsfähige Erweiterung an. Hierzu wird das lineare Modell um die quadratischen Terme der Haupteffekte ergänzt.

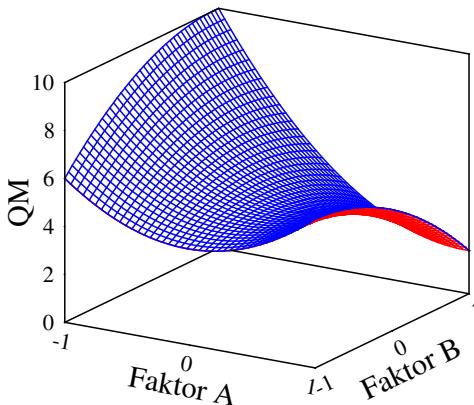


Abb. 2.9 Beispiel einer quadratischen Funktion zur Beschreibung des Qualitätsmerkmals in Abhängigkeit von zwei Faktoren A und B.

Zweistufige Versuchspläne beinhalten keine mittlere Einstellung, reichen also für ein quadratisches Beschreibungsmodell nicht aus. Mit der Zahl der Einstellungsstufen steigt die Zahl der Kombinationen erheblich. Bei fünf Faktoren beispielsweise ergeben sich 243 Kombinationen bei drei Stufen, im Vergleich zu 32 Kombinationen bei zwei Stufen. Umso wichtiger ist nun die passende Strategie für einen Teilstufenplan, weil Vollfaktorpläne in der Praxis nicht mehr handhabbar sind. Der folgende Abschnitt stellt einige Strategien zur Konstruktion von Teilstufenplänen für nichtlineare Zusammenhänge vor und beleuchtet deren Vor- und Nachteile.

Dieser Abschnitt beschränkt sich auf die Betrachtung eines quadratischen Beschreibungsmodells. Bei Computersimulationen besteht oft die Möglichkeit, sehr viele Berechnungsläufe durchzuführen und weitaus komplexere Beschreibungsmodelle zu verwenden. Das Kapitel *Versuchspläne für nichtlineare Zusammenhänge* spannt den Bogen weiter und beschreibt eine multivariate Analyse ohne vorherige Festlegung auf ein Beschreibungsmodell.

Die Zahl der Unbekannten steigt beim quadratischen Modell im Vergleich zum linearen Modell⁵ nur geringfügig an. Für jeden Faktor benötigt man einen zusätzlichen Term. Bei acht Faktoren stehen also 45 Modellkonstanten den 6561 Kombinationen eines dreistufigen Vollfaktorplans gegenüber. Somit ist eine Reduzierung des Aufwandes ohne Informationsverlust möglich. Die Zahl der Modellkonstanten steigt progressiv mit der Zahl der Faktoren an (siehe Abb. 2.10).

⁵ Mit Wechselwirkungen.

$$n_m = \sum_{i=1}^{n_f+1} i \quad (2.1)$$

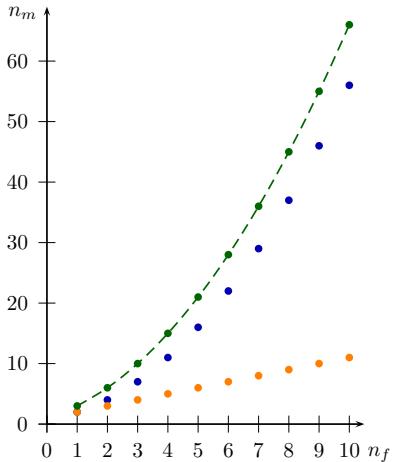


Abb. 2.10 Anstieg der Zahl der Modellkonstanten mit der Anzahl der Faktoren. Unter Berücksichtigung der quadratischen Effekte ist der Anstieg progressiv (gestrichelte Kurve), steigt aber im Vergleich zum linearen Modell mit Wechselwirkungen für zweistufige Versuchspläne (mittlere Kurve) nur moderat an. Lineare Modelle, die nur die Haupteffekte berücksichtigen, sind weniger anspruchsvoll (untere Kurve).

2.3.1 Central-Composite-Design

Das Central-Composite-Design (CCD) baut auf einem zweistufigen Versuchsplan auf. Dadurch entsteht die Möglichkeit, zunächst das System mit einem zweistufigen Versuchsplan zu untersuchen und erst bei Bedarf die fehlenden Versuchsläufe zu ergänzen. Das Central-Composite-Design besteht immer aus einem „Würfel“ und einem „Stern“. Der „Würfel“ ist ein zweistufiger Versuchsplan, in der Regel ein Teilstufenplan der Auflösungsstufe IV oder V. Ein „Stern“ entsteht durch Variation der einzelnen Faktoren, ausgehend von der Mittelstellung, dem sogenannten *center point*. Der Stufenabstand dieser Variation übersteigt den Stufenabstand des Würfels, also wird letztlich jeder Faktor auf fünf Stufen getestet.

Der über den Würfel hinausragende Stern stellt in der Praxis oft ein Problem dar, weil die vom Versuchsplan verlangten Einstellungen mitunter nicht durchführbar sind. Wenn im konkreten Fall daher die Stufenbreite nicht über die Stufenbreite des Würfels hinaus vergrößert werden kann, greift man auf das *face centered* CCD zurück und bleibt damit bei einer auf die Würfeldimensionen reduzierten Abmessung des Sterns. Allerdings sind die Eigenschaften dieser Konstruktion im Vergleich zum regulären CCD schlechter, weil die quadratischen Effekte untereinander korrelieren. Ein Blick auf die Korrelationsmatrix empfiehlt sich also, im Rahmen der Versuchsauswertung. Der Quotient der Stufenbreiten heißt Verlängerungsfaktor. Orthogonalität und Drehbarkeit (rotatability) verlangen vorgegebene Verlängerungsfaktoren in Abhängigkeit von der Zahl der Faktoren. Drehbar ist ein Design dann, wenn die

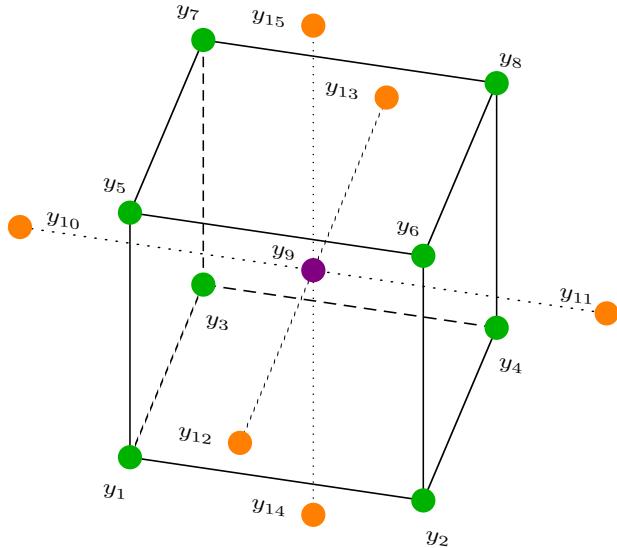


Abb. 2.11 Central-Composite-Design. Aufbauend auf einen zweistufigen Versuchsplan (Würfel) bieten zusätzliche Versuche (Stern), die Möglichkeit, auch nichtlineare Zusammenhänge zu untersuchen.

A	B	C	y
—	—	—	y_1
+	—	—	y_2
—	+	—	y_3
+	+	—	y_4
—	—	+	y_5
+	—	+	y_6
—	+	+	y_7
+	+	+	y_8
0	0	0	y_9
—	0	0	y_{10}
++	0	0	y_{11}
0	—	0	y_{12}
0	++	0	y_{13}
0	0	—	y_{14}
0	0	++	y_{15}

Tabelle 2.10 Central-Composite-Design. Aufbauend auf einen zweistufigen Versuchsplan (Würfel) wird zunächst der Zentralpunkt getestet. Dann folgen Variationen für jeweils einen Faktor, die über die Würfelfgrenzen hinaus gehen können. Drei Faktoren lassen sich auf diese Weise mit nur 15 Versuchen auf jeweils fünf Stufen untersuchen.

Varianz der Vorhersage nur noch vom Abstand zum Zentralpunkt abhängt und nicht von der Richtung. Dies ist ein Anspruch, der über die Orthogonalität hinaus geht. In der Praxis sind die dafür erforderlichen Verlängerungsfaktoren selten realisierbar.

Das DRAPER-LIN CCD arbeitet mit dichter besetzten zweistufigen Teilstufenplänen als Würfel. Dies reduziert die Zahl der Versuchsläufe bis knapp an das theoretische Minimum zur Versorgung des quadratischen Modells, geht aber mit einer deutlich schlechteren Korrelationsmatrix einher. Hier korrelieren sogar Haupteffekte mit Wechselwirkungen, also ist das Feld nur unter Vorbehalt einsetzbar. An dieser

Faktoren	α_{orth}	α_{rot}
3	1,29	1,68
4	1,48	2,00
5	1,61	2,00
6	1,78	2,38
7	1,94	2,83
8	2,05	2,83

Tabelle 2.11 Verlängerungsfaktor α beim Central-Composite-Design in Abhängigkeit von der Faktorenzahl. Das drehbare Design stellt noch höhere Ansprüche als das orthogonale Design.

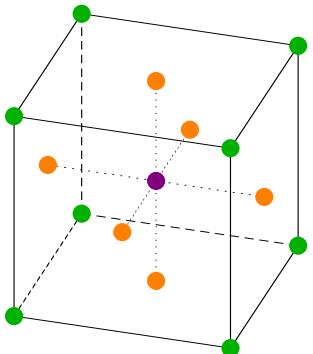


Abb. 2.12 Face-Centered-Central-Composite-Design. Der zusätzliche Stern ragt nicht über den Würfel hinaus. Dies vermeidet Probleme bei der Versuchsdurchführung, liefert aber eine schlechtere Trennung der quadratischen Effekte untereinander.

Stelle sei der Rat gestattet, im Zweifelsfall lieber ein sauberes zweistufiges Feld zu verwenden, als ein mit starken Kompromissen behaftetes mehrstufiges Feld. Oft werden die Nichtlinearitäten überschätzt, oder es genügt nach sorgfältiger Voruntersuchung eine Detailuntersuchung mit einer geringeren Faktorenzahl.

Das Central-Composite-Design wird häufig eingesetzt [9], weil die Zahl der benötigten Versuchsläufe nur moderat mit der Faktorenzahl ansteigt und die Eigenschaften der Felder insgesamt sehr gut sind [2].

Faktoren	Unbekannte	CCD	Draper-Lin
3	10	15	
4	15	25	17
5	21	27	23
6	28	45	29
7	36	79	39
8	45	81	53

Tabelle 2.12 Zahl der benötigten Einstellungen beim Central-Composite-Design in Abhängigkeit von der Faktorenzahl. DRAPER-LIN-CCD verwenden als Würfel einen dichter besetzten Teilstellfaktorplan.

2.3.2 Box-Behnken-Design

Das Box-Behnken-Design geht auf BOX und BEHNKEN zurück. Immer dann, wenn die Ecken des Faktorraums kritisch sind, bietet sich das Box-Behnken-Design an, denn es lässt genau diese aus. Natürlich muss dem Anwender dann klar sein, dass die Beschreibungsfunktion nicht mehr in den Ecken gilt, weil diese außerhalb des unter-

suchten Bereichs liegen und Extrapolationen grundsätzlich unzulässig sind. Bei bekannt nichtlinearen Zusammenhängen kann jedoch bereits zum Zeitpunkt der Versuchsplanung das Optimum im mittleren Bereich des Faktorraums vermutet werden, weshalb die oben genannte Einschränkung in der Praxis vielfach nicht ins Gewicht fällt. So wird zum Beispiel dieses Feld sehr erfolgreich bei der Motorenentwicklung eingesetzt. Die Zusammenhänge sind bei dieser Anwendung immer nichtlinear und die Eckpunkte in vielen Fällen nicht ansteuerbar, weil die Motoren in den extremen Betriebspunkten nicht mehr starten.

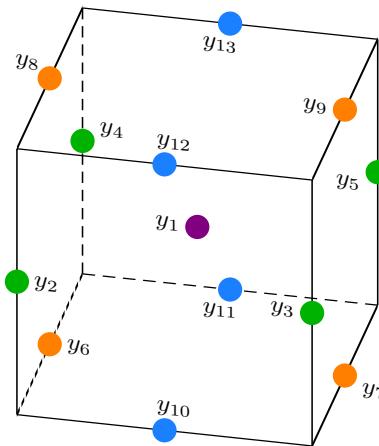


Abb. 2.13 Box-Behnken-Design. Die Ecken werden nicht besetzt, sondern die Mittelpunkte der Würfelkanten und der Zentralpunkt. Alle Kombinationen von jeweils zwei Faktoren auf den Stufen -1 und 1 bilden einen Ring. Hierbei bleiben die übrigen Faktoren bei einer mittleren Einstellung.

A	B	C	y
0	0	0	y_1
—	—	0	y_2
+	—	0	y_3
—	+	0	y_4
+	+	0	y_5
—	0	—	y_6
+	0	—	y_7
—	0	+	y_8
+	0	+	y_9
0	—	—	y_{10}
0	+	—	y_{11}
0	—	+	y_{12}
0	+	+	y_{13}

Tabelle 2.13 Box-Behnken-Design. Dieser Versuchspunkt ist völlig eigenständig und baut nicht auf einem zweistufigen Versuchspunkt auf. Um den Zentralpunkt werden ringförmig jeweils zwei Faktoren in allen Kombinationen getestet. 13 Versuche genügen für 3 Faktoren. Bei realen Versuchen (kein CAE) fügt man Wiederholungen des Zentralpunktes hinzu, um die Teststreuung abschätzen zu können.

Das Box-Behnken-Design setzt sich aus “Ringen” zusammen. Jeweils zwei Faktoren werden in allen Kombinationen auf zwei Stufen variiert, die übrigen Faktoren bleiben auf einer mittleren Einstellung. Dadurch entsteht ein sehr gut konditioniertes Feld mit sauberer Auflösung der Wechselwirkungen und der quadratischen Effekte. Im Gegensatz zum face centered Central-Composite-Design korrelieren die quadratischen Effekte nur schwach, mit Werten zwischen 0,07 und 0,2. Nachteilig ist bei großer Faktorenzahl der Überhang der mittleren Einstellung im Vergleich zu den Randeinstellungen⁶. Daher wird diese Konstruktion in der Literatur nur für 3-5 Faktoren ohne Einschränkung empfohlen, lässt sich jedoch im Bedarfsfall auch für eine höhere Zahl von Faktoren verwenden. Das Box-Behnken-Design ist ebenfalls sehr effizient und absolut praxistauglich.

Faktoren	Unbekannte	BBB
3	10	13
4	15	25
5	21	41
6	28	49
7	36	57
8	45	81

Tabelle 2.14 Zahl der benötigten Einstellungen beim Box-Behnken-Design in Abhängigkeit von der Faktorenzahl.

2.3.3 Monte-Carlo-Verfahren

Der Name dieses Verfahrens erinnert nicht ohne Grund an ein Spielkasino. Man nutzt den Zufallsgenerator, um die Faktoreneinstellungen quasi “auszuwürfeln”. Da alle Spalten unabhängig voneinander ausgewürfelt werden, ergeben sich bei genügend großen Feldern nur schwache Korrelationen. Das auf zunächst sonderbar erscheinende Weise erzeugte Feld ist sozusagen von Natur aus weitgehend orthogonal. Die größte Stärke dieser Konstruktion liegt in der Tatsache, dass sehr viele Einstellungen für jeden Faktor gefahren werden. Ein “over-fit” ist damit ausgeschlossen.

Nachteilig ist die hohe Zahl der erforderlichen Versuchsläufe. Die zufällige Festlegung der Faktoreneinstellung arbeitet natürlich weitaus weniger effizient als eine speziell ausgeklügelte Feldkonstruktion. Eine gleichmäßige Abdeckung des mehrdimensionalen Faktorraumes erfordert also sehr viele Versuchsläufe. Latin Hypercubes reduzieren den Aufwand ohne Verlust der Vorteile um etwa 50%. Bei realen Versuchen wird man selten die gewürfelten Einstellungen umsetzen können, weshalb dieses Verfahren üblicherweise den CAE-Studien vorbehalten bleibt. Wenn die Zahl der Versuchsläufe keine große Rolle spielt, zum Beispiel bei schnellen CAE-Modellen mit automatisierter Ablaufsteuerung, ist das Monte-Carlo-Verfahren eine gute Wahl.

⁶ Dadurch ist die Dämpfung der Teststreuung an den Rändern schlechter als bei zweistufigen Feldern gleicher Größe. Dies kann im Einzelfall die Auswertung behindern.

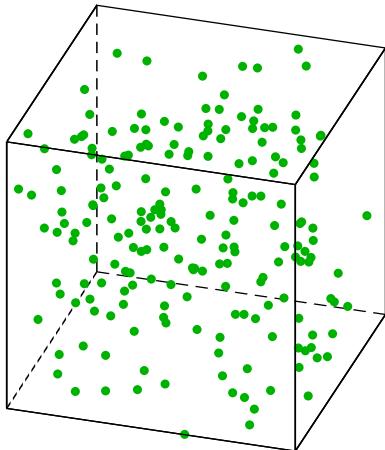


Abb. 2.14 Monte-Carlo-Design. Per Zufallsgenerator werden die Faktoreinstellungen bestimmt. Der Versuchsplan ist weitgehend orthogonal und nicht auf ein bestimmtes Beschreibungsmodell festgelegt. Allerdings benötigt man viele Versuchsläufe. In diesem Fall zeigen sich selbst bei 200 Einstellungen noch relativ große „Löcher“.

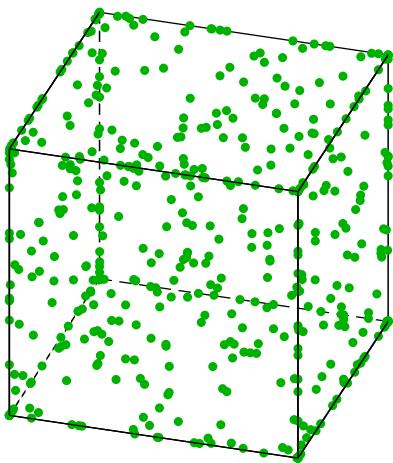


Abb. 2.15 Space-Filling-Design. Gezeigt werden die ersten drei Dimensionen eines Space-Filling-Designs mit 500 Versuchen. Das Feld wurde von JMP für acht Dimensionen gerechnet.

Eine weitere Variante zur Reduktion der erforderlichen Versuchszahl sind die sogenannte Space-Filling-Designs. Hier geht man zunächst von zufälligen Kombinationen aus, sorgt aber parallel dazu für eine möglichst gleichmäßige Verteilung im Faktorraum, um „Löcher“ zu vermeiden. Das Kapitel *Versuchspläne für komplexe Zusammenhänge* geht darauf im Detail ein.

2.3.4 Fallstudie

Vor der Entscheidung für einen Versuchspräzisierung stellt sich natürlich die Frage, inwie weit die Feldkonstruktion das Endergebnis beeinflusst. Welcher Versuchspräzisierung ist der beste? Kann man durch die Wahl des „falschen“ Versuchspräzisierungen auf die falsche

Fährte gelangen? Dies sind die gleichen Fragen wie im Abschnitt *Screening Versuchspläne*. Es gibt keine allgemein gültige Antwort darauf, ansonsten hätte ja nur eine Feldkonstruktion überlebt.

Symbol	Parameter	Einheit	Einstellung	
			min	max
A	α	o	15	45
B	β	o	0	30
C	A_q	mm^2	2	4
D	d	mm	100	200
E	M_{Rt}	Nm	0,01	0,02
F	M_{Rf}	$\frac{Nm}{s}$	0,01	0,02
G	p_{in}	bar	1	2
H	d_{zul}	mm	5	10

Tabelle 2.15 Einstellungstabelle, Fallstudie Rasensprenger. Alle acht Parameter wurden in der folgenden Studie variiert. Es gab jeweils mehr als zwei Stufen, daher die Angabe von *min* und *max*, anstelle von – und +.

Das bereits vorgestellte Rasensprengerbeispiel (siehe Abb. 2.3) bietet die Möglichkeit, acht Faktoren zu variieren. Erprobt wurden vier verschiedene Felder. Als Musterlösung dient ein Space-Filling-Design mit 500 Versuchen. Hier kann ein over-fit mit Sicherheit ausgeschlossen werden. Als zweites Feld geht ein Space-Filling-Design mit 82 Versuchsläufen in's Rennen. Der dritte Kandidat ist ein Latin-Hypercube-Design mit 129 Versuchen. Zum Abschluss wurde noch ein klassisches flächenzentriertes (face centered) Central-Composite-Design verwendet.

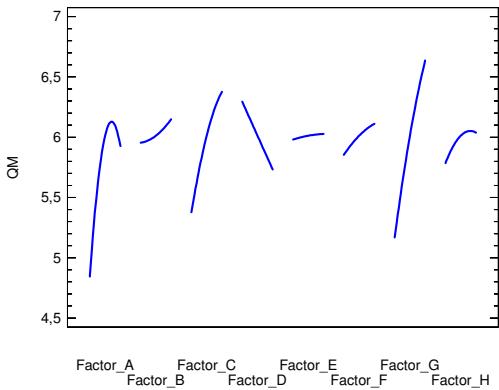


Abb. 2.16 Ergebnis der Simulation mit 500 Versuchen, Space-Filling-Design. Effekt-Diagramm für das Qualitätsmerkmal: Reichweite

Die Unterschiede sind erstaunlich gering, in Anbetracht der völlig unterschiedlichen Feldkonstruktionen, bei denen de facto keine einzige Versuchseinstellung in mehr als einem Feld vorkam. Letztlich kommt man auch in diesem Fall auf mehreren Wegen zum gleichen Ziel. Dies ist ein Verdienst des vergleichsweise robusten quadratischen Modells. Der unerfahrene Anwender braucht also keine Angst vor Misserfolgen zu haben und dem erfahrenen Anwender steht jederzeit eine reiche Auswahl an Versuchsplänen zur Verfügung.

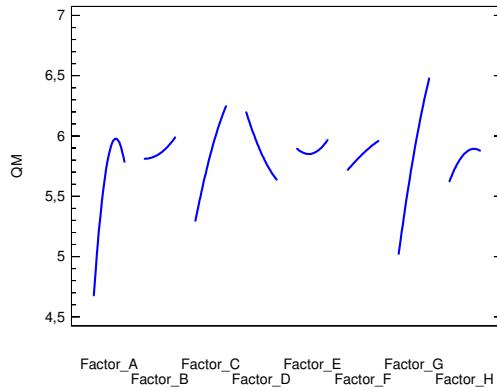


Abb. 2.17 Ergebnis der Simulation mit 82 Versuchen, Space-Filling-Design. Effekt-Diagramm für das Qualitätsmerkmal: Reichweite

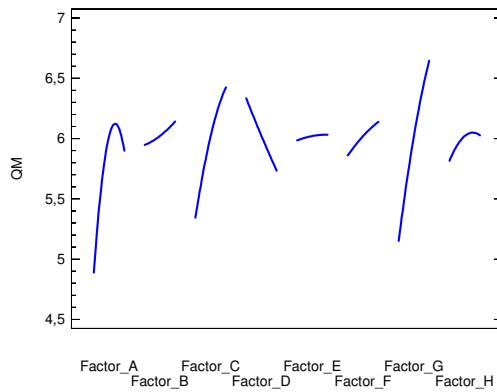


Abb. 2.18 Ergebnis der Simulation mit 129 Versuchen, Latin-Hypercube-Design. Effekt-Diagramm für das Qualitätsmerkmal: Reichweite

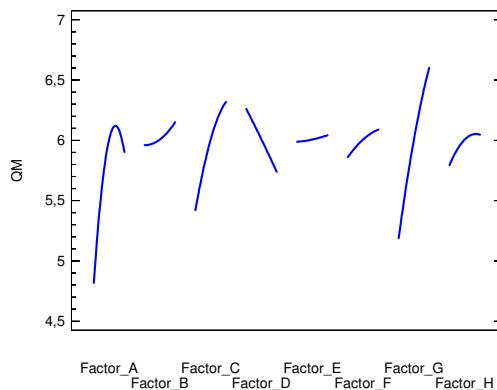


Abb. 2.19 Ergebnis der Simulation mit 82 Versuchen, Central-Composite-Design (face-centered). Effekt-Diagramm für das Qualitätsmerkmal: Reichweite

Das quadratische Modell konnte diese Aufgabe übrigens souverän meistern. Selbst beim 500er Feld lässt sich über 99 % der Gesamtvarianz mit dem Beschreibungsmodell abdecken. Die kleineren Felder liegen naturgemäß bei diesem Indi-

kator über den großen Feldern, weil weniger Freiheitsgrade im System vorhanden sind und sich das Modell dann besser an die vorhandenen Versuchsdaten anpassen kann. Immerhin wurde im acht-dimensionalen Raum und bis an den Rand der numerischen Stabilitätsgrenzen des Rasensprengermodells variiert. Es gibt ausgeprägte Wechselwirkungen und Nichtlinearitäten. Dies bezieht sich auf die Basisvariante des Rasensprengermodells. Für die aufwendigeren Untersuchungen in den CAE-Kapiteln wurde das Modell erweitert, um die einfachen Beschreibungsmodelle über ihre Leistungsgrenzen zu bringen. Natürlich hat das quadratische Modell seine Grenzen. Es ist grundsätzlich stetig und differenzierbar. Daher kann es weder “Sprünge” noch “Knicke” abbilden.

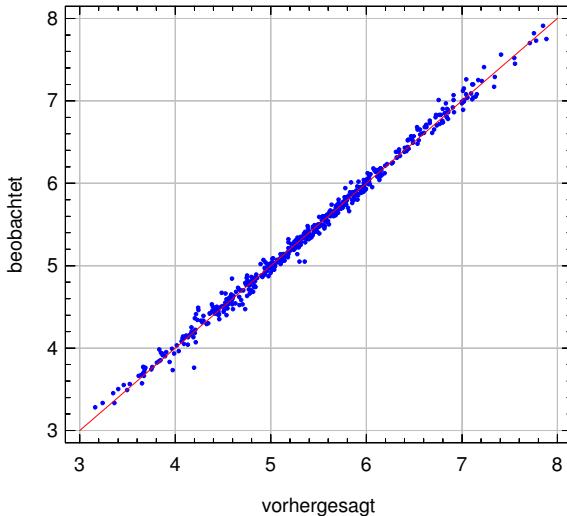


Abb. 2.20 Residual-Plot, Simulation mit 500 Versuchen, Space-Filling-Design. Bis auf vereinzelte Ausreißer kann das Beschreibungsmodell die Versuchsreihe gut abbilden.

2.4 Grenzen des Beschreibungsmodells

Wenn das quadratische Beschreibungsmodell nicht mehr ausreicht, kann ein kubisches Beschreibungsmodell in vielen Fällen das Problem lösen, ohne einen Methodenwechsel in Richtung neuronaler Netze [3] oder Kriging zu erzwingen. An den Versuchplan werden dann allerdings höhere Anforderungen gestellt und auch die Zahl der Unbekannten steigt drastisch an. Die Gefahr eines over-fits ist hier gegeben, daher empfehlen sich Versuchspläne mit sehr vielen Zwischenstufen, zum Beispiel Space-Filling-Designs oder Latin-Hypercubes.

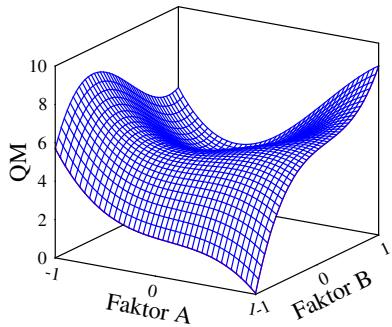


Abb. 2.21 Beispiel einer kubischen Funktion zur Beschreibung des Qualitätsmerkmals in Abhängigkeit von zwei Faktoren A und B.

Die Handhabung kubischer Beschreibungsmodelle ist aufwendig. Hilfreich ist dann eine Software⁷, die automatisch alle unbedeutenden Terme kürzen kann, um die Beschreibungsfunktion einigermaßen kompakt zu halten.

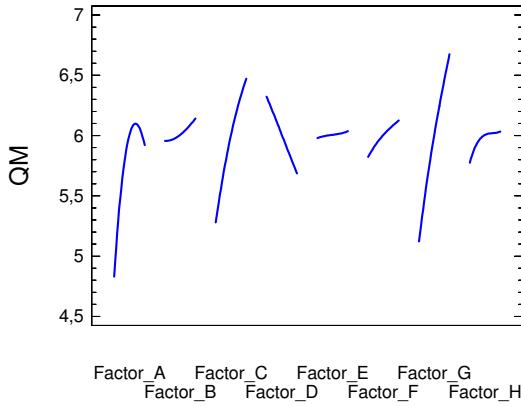


Abb. 2.22 Ergebnis der Simulation mit 500 Versuchen, Space-Filling-Design, kubisches Beschreibungsmodell. Dargestellt ist das Effekt-Diagramm für das Qualitätsmerkmal: Reichweite. Die Unterschiede zum quadratischen Modell sind in diesem Fall nur gering.

$$n_m = \sum_{i=1}^{n_f+1} i + n_f^2 + (n_f - 2)^2 \quad \text{für } n_f > 1 \quad (2.2)$$

Bei drei Faktoren ergeben sich bereits 20 Modellkonstanten, bei vier Faktoren 35, bei fünf Faktoren 55, usw.. Eine Beschreibungsgleichung für zwei Faktoren x_1 und x_2 sieht dann folgendermaßen aus:

$$y = c_0 + c_1 x_1 + c_2 x_2 + c_{12} x_1 x_2 + c_{11} x_1^2 + c_{22} x_2^2 + c_{111} x_1^3 + c_{112} x_1^2 x_2 + c_{122} x_1 x_2^2 + c_{222} x_2^3 \quad (2.3)$$

⁷ Design Expert® ist beispielsweise dazu in der Lage.

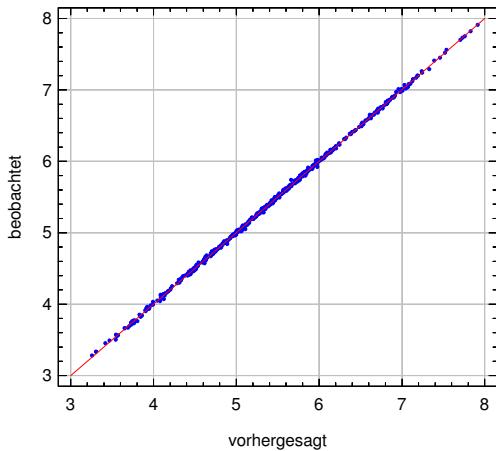


Abb. 2.23 Residual-Plot des kubischen Beschreibungsmodells am Beispiel Rasensprenger. Simulation mit 500 Versuchen, Space-Filling-Design. Das Modell dritter Ordnung erreicht einen $R^2_{adjusted}$ Wert von 99,95 %, wobei noch über 300 Freiheitsgrade verbleiben, um eine stabile Statistik aufzubauen. In diesem Fall lagen tatsächlich signifikante Dreifachwechselwirkungen und kubische Effekte vor.

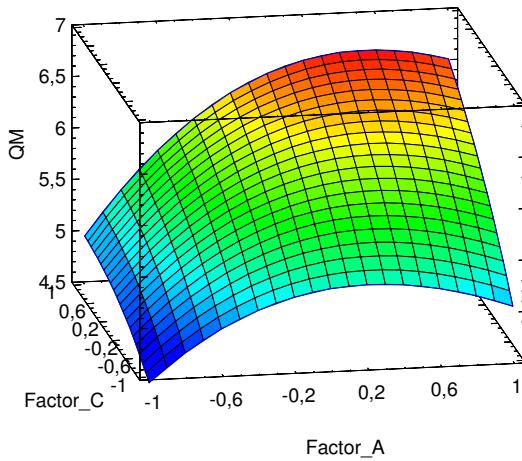


Abb. 2.24 Reichweite des Rasensprengers als Funktion vom vertikalem Düsenwinkel α und Düsenquerschnitt. Space-Filling-Design, kubisches Beschreibungsmodell. Die übrigen Faktoren stehen in der mittleren Einstellung.

Auch ein Beschreibungsmodell vierter Ordnung ist möglich (quartic). Die Zahl der Modellkonstanten steigt dann rasant an, also erfordert dies große Felder und eine automatisierte Elimination der nicht signifikanten Terme. Bei einer geringen Zahl von Faktoren kann das Modell vierter Ordnung jedoch sehr hilfreich sein und erweitert den Anwendungsbereich der DoE. Wo sind die Grenzen? Eine allgemeingültige Antwort gibt es nicht, allerdings einige grundsätzliche Überlegungen. Hierzu hilft die Betrachtung eines eindimensionalen Problems. Nur wenn der grundsätzliche Verlauf der Messdaten durch diese Funktionsklasse abbildungbar ist, kann die Regression erfolgreich sein.

Im Einzelfall kann eine logarithmische Transformation des Qualitätsmerkmals (also der Ergebnisgröße) die Grenzen noch etwas weiter treiben, ist also immer einen Versuch wert. Allerdings wird die Regression nicht grundsätzlich besser. Daraus ist in diesen Fällen eine sorgfältige Prüfung des Beschreibungsmodells notwen-

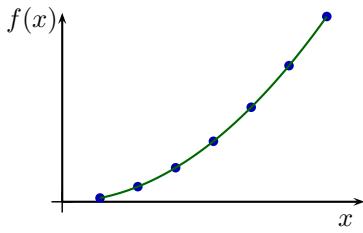


Abb. 2.25 Progressiver Verlauf. Dies ist kein Problem und erfordert in der Regel nur ein Modell zweiter Ordnung, auch wenn das Extremum nicht am Rand liegt. Die Terme dritter und vierter Ordnung bieten weitere Möglichkeiten, also ist diese Kategorie unkritisch.

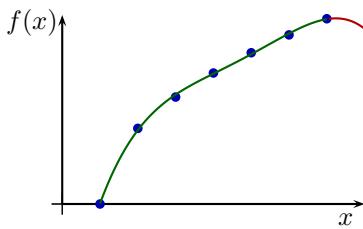


Abb. 2.26 Degressiver Verlauf. Dies ist ebenfalls möglich, jedoch nur mit einem Modell dritter oder vierter Ordnung. An den Rändern läuft die Regression aus dem Ruder, also wird jede Extrapolation mit Sicherheit scheitern. Insgesamt tun sich Polynome mit degressiven Verläufen ein wenig schwer.

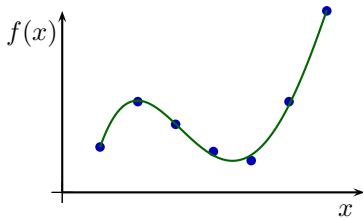


Abb. 2.27 Wendepunkt mit lokalen Extrema. Ein Modell dritter Ordnung kommt damit zurecht. An den Rändern ist auch hier Vorsicht geboten. Die Lage der Extrema wird möglicherweise nicht exakt vorhergesagt. Bei Optimierungen sind Bestätigungsläufe angebracht.

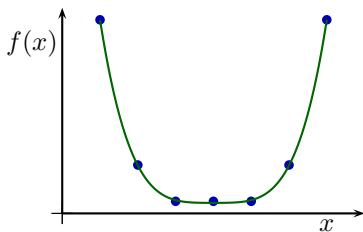


Abb. 2.28 Wannenförmiger Verlauf. Ein Modell vierten Ordnung kann dies überraschend gut abbilden, weil sich die Terme höherer Ordnung gegenseitig kontrollieren.

dig. Sollten die Polynome als Funktionsklasse scheitern, kann auch eine allgemeine Regression in Betracht gezogen werden. Dies ist kein DoE Standard, aber mit

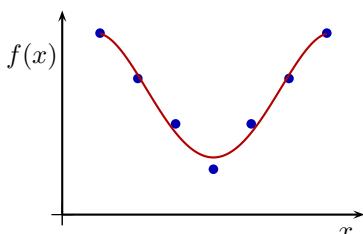


Abb. 2.29 Nicht differenzierbarer Verlauf. Hier ist Schluss. Auch ein Modell vierten Ordnung kommt nicht in die Ecke hinein und wird den vermutlich interessantesten Teil der Daten nicht gut abbilden.

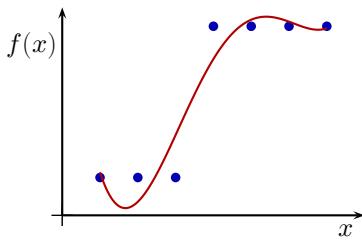


Abb. 2.30 Sprungfunktion. Auch an diesem Fall heißt sich selbst das Modell vierter Ordnung die Zähne aus. Eine grobe Abbildung ist machbar, jedoch kommt es zu Überschwingen und unrealistischen Verläufen an den Rändern.

der gleichen Software und den gleichen Versuchsdaten machbar. Im Wesentlichen erweitert sich dadurch die Auswahl an Funktionstermen und in Folge dessen die Flexibilität bei der Anpassung an die Testdaten. Wer in der Praxis mit derart nicht-linearen Zusammenhängen zu tun hat, kennt in der Regel sein System schon aus Vorversuchen und hat eine qualitative Idee vom erwarteten Verlauf. Normalerweise sind die realisierbaren Stufenabstände jedoch begrenzt, insbesondere dann, wenn viele Faktoren gleichzeitig variiert werden. Daher arbeiten auch einfache Modelle zweiter Ordnung in erstaunlich vielen Fällen absolut zuverlässig.

2.5 Mischungspläne

Anwendungen in der Chemie und der Verfahrenstechnik beziehen sich oft auf Mischungen. Im Gegensatz zur allgemeinen Anwendung liegt hier eine zusätzliche Randbedingung vor, die der Faktorraum eingrenzt: Die Summe aller Mischungsanteile ist 100%. Diese Randbedingung reduziert den Faktorraum um eine Dimension. Zum Beispiel steht bei drei Faktoren letztlich nur noch ein zweidimensionaler Bereich zur Verfügung, der die Randbedingung erfüllt. Sehr häufig ist der Faktorraum noch durch weitere Randbedingungen eingeschränkt, weil nicht jedes mögliche Mischungsverhältnis in der Realität darstellbar ist bzw. einen Sinn ergibt. Diese Einschränkungen sind jedoch von Fall zu Fall unterschiedlich. Daher ist dieser Abschnitt nur kurz, denn die allgemeinen Mischungspläne sind nur dann anwendbar, wenn der Faktorraum keine weiteren Einschränkungen enthält, ansonsten kommen *maßgeschneiderte* Versuchspläne zum Einsatz.

$$\sum_{i=1}^{n_f} x_i = 1 \quad (2.4)$$

2.5.1 Simplex-Lattice-Design

Das Simplex-Lattice-Design (Simplexgitterplan) testet zunächst die Ecken des verbleibenden Faktorraums, also die Mischungen mit jeweils vollem Anteil einer Komponente. Die verbleibenden Komponenten haben dann den Anteil 0. In Abhängig-

			x_1	x_2	x_3	
linear	quadratisch	kubisch	1	0	0	y_1
			0,5	0,5	0	y_2
			0,5	0	0,5	y_3
			0	1	0	y_4
0	0,5	0,5	y_5			
0	0	1	y_6			
			1	0	0	y_1
			0,5	0,5	0	y_2
			0,5	0	0,5	y_3
			0	1	0	y_4
			0	0,5	0,5	y_5
			0	0	1	y_6
			0,5	0,5	0,5	y_7
			0,5	0,5	0,5	y_8
			0,5	0,5	0,5	y_9
			0	0	1	y_{10}

Tabelle 2.16 Simplex-Lattice-Design für drei Mischungskomponenten und verschiedene Beschreibungsmodelle. Die Bildung der Beschreibungsmodelle erfolgt analog zur Bildung bei konventionellen Versuchsplänen. Allerdings ist der Faktorraum eingeschränkt, weil in jedem Fall die Randbedingung für die Mischung eingehalten werden muss.

keit vom gewünschten Beschreibungsmodell kommen weitere Punkte hinzu, wobei die Mischungsanteile in jeweils gleichen Stufenabständen variieren. Für das quadratische Modell kommt der Anteil 0,5 hinzu, beim kubischen Modell werden die Anteile 0,3 und 0,6 getestet.

2.5.2 Simplex-Centroid-Design

x_1	x_2	x_3	
1	0	0	y_1
0	1	0	y_2
0	0	1	y_3
0,5	0,5	0	y_4
0,5	0	0,5	y_5
0	0,5	0,5	y_6
0,5	0,5	0,5	y_7

Tabelle 2.17 Simplex-Centroid-Design für drei Mischungskomponenten. Dieser Versuchsplan ist geeignet für ein lineares, ein quadratisches und ein reduziertes kubisches Beschreibungsmodell, ohne die Wechselwirkungsterme dritter Ordnung.

Das Simplex-Centroid-Design sieht grundsätzlich die Verwendung des Zentralpunktes vor. Die Bildungsvorschrift orientiert sich an einer gleich gewichteten Aufteilung mit steigender Zahl von Komponenten in jeweils allen Permutationen. Zunächst kommen alle Permutationen für eine Komponente auf 100%, dann alle Permutationen mit jeweils zwei Komponenten auf 50%, dann alle Permutationen mit drei Komponenten auf 33% usw..

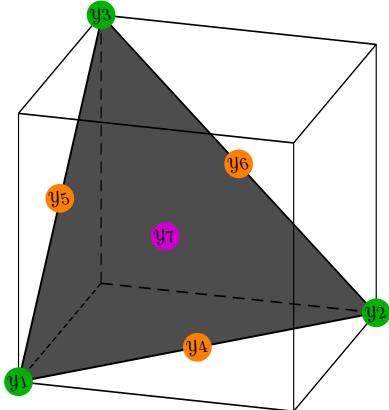


Abb. 2.31 Simplex-Centroid-Design für drei Mischungskomponenten. Aus dem dreidimensionalen Faktorraum wird ein zweidimensionaler Bereich, der die Randbedingung für die Mischung erfüllt. In diesem Bereich sind die sieben Versuchspunkte verteilt. Die Eckpunkte entstehen, wenn nur eine Komponente eingesetzt wird. Auf den Kanten des Dreiecks liegen Mischungen von jeweils zwei Komponenten. Der Zentralpunkt entsteht durch die Mischung aller drei Komponenten.

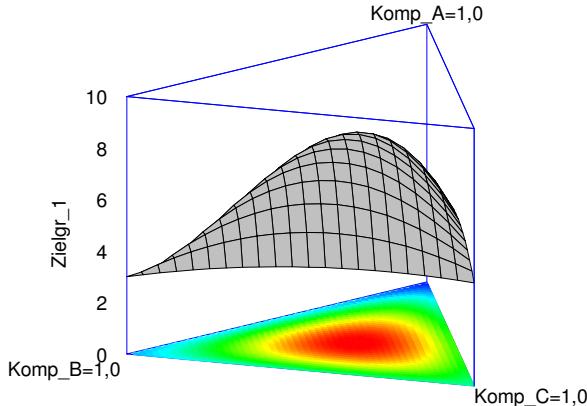


Abb. 2.32 Typische Ergebnisdarstellung eines Simplex-Centroid-Designs für drei Mischungskomponenten.

2.6 Individuell erstellte Versuchspläne

In der Praxis kann es gute Gründe geben, von den vorkonfektionierten Versuchsplänen abzuweichen und einen Versuchsplan zu erstellen, der quasi für das aktuelle Experiment maßgeschneidert wird. In Chemie und Verfahrenstechnik kommt dies allein deshalb oft vor, weil Mischungen nicht in beliebiger Zusammensetzung zu brauchbaren Ergebnissen führen, sondern nur in eingegrenzten Gebieten. Gemischtstufige Felder sind ebenfalls nicht trivial und werden dann erforderlich, wenn die Faktoren in unterschiedlich vielen Stufen zu testen sind. Mitunter ist die Zahl der durchführbaren Versuche auch so begrenzt, dass man bestrebt ist, exakt das erforderliche Minimum an Versuchen durchzuführen.

In diesen Fällen kommen sogenannte *optimale* Versuchspläne zum Einsatz, die nach bestimmten Kriterien aus einem Vollfaktorplan die wichtigsten Einstellungen herauspicken. An die Stelle eines Vollfaktorplans kann auch eine Kombination aus zweistufigem Vollfaktorplan und weiteren *Kandidaten* treten. Als Kandidat gilt hier eine mögliche Einstellung der Faktoren, die nicht im zweistufigen Vollfaktorplan enthalten ist, zum Beispiel Zentralpunkt oder Kantenmitten (vgl. Box-Behnken-Plan). Natürlich muss auch hier gewährleistet sein, dass die Effekte voneinander zu trennen sind.

An dieser Stelle taucht eine große Hürde auf, da der Anwender vor der Versuchsreihe sein Beschreibungmodell festlegen muss. Der Auswahlalgorithmus berücksichtigt nur die Effekte des vorher ausgewählten Beschreibungsmodells bei der Selektion der optimalen Kombinationen. Kennt man sein System gut, ergibt sich dadurch im Vergleich zu vorkonfektionierten Plänen ein gewisses Einsparpotential. Kennt man sein System nicht so gut, wird der optimale Versuchspläne die vorkonfektionierten Pläne kaum schlagen können, bringt aber möglicherweise zusätzliche Komplikationen mit sich. Der Anwender muss neben dem Beschreibungsmodell auch die Zahl der verfügbaren Versuchsläufe und das Auswahlkriterium festlegen. Aus diesen Angaben errechnet der Computer dann den bestmöglichen Kompromiss in Bezug auf das Auswahlkriterium. In jedem Fall ist es ratsam, sich das resultierende Feld genau anzusehen und die Eigenschaften zu prüfen. Oft wird leider die Zahl der für eine saubere Untersuchung erforderlichen Versuche unterschätzt oder der Anwender kennt den Unterschied der Auswahlkriterien nicht. Blindes Vertrauen in den *optimalen* Versuchspläne führt dann zur Enttäuschung.

2.6.1 Auswahlkriterien

Gängig ist das sogenannte D-optimale Design. Hierzu wird die Koeffizientenmatrix \mathbf{X} analysiert, die in Abhängigkeit vom Beschreibungsmodell und dem Versuchsplan entsteht. Um zu verstehen, was es damit auf sich hat, ist ein kleiner Exkurs in die Regressionsanalyse [8, 4, 11] notwendig. Letztlich laufen alle alle bislang gezeigten Beschreibungsmodelle auf eine lineare Regression hinaus. Auch die Wechselwirkungen und quadratischen Effekte sind formal wie lineare Effekte berechenbar, wenn man sogenannte *transformierte Eingangsgrößen* einführt. Eine transformierte Eingangsgröße entsteht aus einer oder mehreren Eingangsgrößen durch eine feste mathematische Verknüpfung, zum Beispiel Multiplikation. Im Gleichungssystem erfordert jede transformierte Eingangsgröße eine zusätzliche Spalte. De facto muss also ein Gleichungssystem mit n_c Spalten und n_r Zeilen gelöst werden.

$$n_c = n_f + n_t + 1 \quad (2.5)$$

Zur Faktorenzahl n_f kommt noch die Zahl der zusätzlichen transformierten Eingangsgrößen n_t hinzu. Außerdem erfordert der Gesamtmittelwert eine Konstante. n_r ist die Zahl der Versuchsläufe. Die Ergebnisse der Versuchsläufe bilden einen Vek-

tor \mathbf{y} . Dieser wird durch das Beschreibungsmodell angenähert, wobei ein Restfehler ε bleibt. Das Beschreibungsmodell seinerseits ist eine Linearkombination der Eingangsgrößen (inclusive der transformierten Eingangsgrößen) und der Konstanten. Die Linearkombination entsteht einfach durch Multiplikation der Matrix \mathbf{X} mit dem Vektor \mathbf{c} .

$$\mathbf{y} = \mathbf{X}\mathbf{c} + \varepsilon \quad (2.6)$$

Um an die Modellkonstanten \mathbf{c} zu kommen, muss die folgende Gleichung gelöst werden:

$$\mathbf{c} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \quad (2.7)$$

An dieser Stelle setzen die Bewertungsverfahren an. Der D-optimale Versuchsplan minimiert die Determinante des Terms $(\mathbf{X}'\mathbf{X})^{-1}$. Dies entspricht einer Maximierung der Determinante von $(\mathbf{X}'\mathbf{X})$. $(\mathbf{X}'\mathbf{X})$ wird auch als *Informationsmatrix* bezeichnet. Lapidar ausgedrückt bringt die maximale Determinante auch die maximale Information. So wird eine möglichst stabile Berechnung der Modellkonstanten erreicht. Der Suchalgorithmus stellt bei vorgegebenem Beschreibungsmodell jeweils die Kandidatenliste zusammen und rechnet eine neue Matrix \mathbf{X} aus. Die Optimierung nach dem oben genannten Kriterium liefert dann die bestmögliche Auswahl der Faktoreinstellungen. D-optimale Versuchspläne minimieren das Volumen des gemeinsamen Vertrauensbereiches des Vektors \mathbf{c} .

Alternativ dazu kann auch die mittlere Varianz der Regressionskoeffizienten optimiert werden. Dies geschieht im A-optimalen Design. Hierzu wird die Summe der Hauptdiagonalelemente von $(\mathbf{X}'\mathbf{X})^{-1}$ minimiert, die Spur (trace) dieser Matrix.

Das G-optimale Design minimiert das größte Element der Hauptdiagonale in der sogenannten Hutmatrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Dies beeinflusst die maximal auftretende Varianz der Vorhersagewerte im gesamten Faktorraum. Es gilt nämlich folgende Beziehung zwischen den vorhergesagten Ergebnissen $\hat{\mathbf{y}}$ und den tatsächlichen Ergebnissen \mathbf{y} :

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (2.8)$$

I-optimale und V-optimale Versuchspläne richten hingegen ihr Augenmerk auf die mittlere Vorhersagegüte im Faktorraum.

Eine genaue mathematische Beschreibung der verschiedenen Kriterien, mit Herleitungen und Vergleich in Bezug auf ihre Auswirkung auf den letztendlich resultierenden Versuchsplan sprengt den Rahmen dieses Buches. Zur Regressionsanalyse gibt es eigene Bücher [4], die im Detail alle Rechenschritte durchgehen.

2.6.2 Einschränkungen des Faktorraums

Es gibt zwei mögliche Varianten: Einschränkungen, die jeweils nur einen Faktor betreffen, oder Einschränkungen die sich auf eine Kombination aus Faktoreinstellungen beziehen. Im Endeffekt führen beide Restriktionstypen dazu, dass ein Teil der möglichen Kombinationen von vornherein ausscheidet, also bei der Suche nach dem optimalen Versuchsplan nicht weiter betrachtet wird.

Die erste Variante lässt sich leicht durch die Wahl der Einstellgrenzen klären.

$$x_{i\min} \leq x_i \leq x_{i\max} \quad (2.9)$$

Die zweite Variante erfordert zusätzliche Restriktionen, die in der Regel implizit ausgedrückt werden, zum Beispiel:

$$x_1 + x_2 \leq 10 \quad (2.10)$$

Nicht alle Versuchsplanungsprogramme gestatten es, die Restriktionen derart detailliert einzugeben. Hier sollte der Anwender vor einer Kaufentscheidung mit dem Softwareanbieter Rücksprache halten, sofern diese Funktionalität von Bedeutung ist. Alternativ dazu kann man im Einzelfall mit Pseudofaktoren durchaus zum Ziel kommen. Pseudofaktoren sind mathematische Konstrukte, die reale Faktoreinstellungen miteinander verknüpfen. Im Versuchsplan und der nachfolgenden Auswertung werden sie wie reale Faktoren behandelt. Für den tatsächlichen Test muss man natürlich aus der verlangten Einstellung der Pseudofaktoren wieder die Einstellungen der realen Faktoren ausrechnen, was bei einfachen Verknüpfungen aber kein Hexenwerk ist.

2.7 Die Mutter aller Versuchspläne

Obwohl diese Kategorie von Versuchsplänen keine besonders gute Vermengungsstruktur aufweist⁸, gelten die griechisch-lateinischen Quadrate von Leonhard Euler (1707 bis 1783) als die ältesten brauchbaren Versuchspläne. Arabische und indische Schmuckstücke mit Mustern vergleichbarer Anordnungen datieren zwar zurück bis ca. 1200, wurden jedoch nicht mathematisch analysiert.

Shall	we	all	dye ?
We	shall	dye	all!
All	dye	shall	we?
Dye	all!	-we	shall.

Tabelle 2.18 Schon vor langer Zeit wurden düstere Gedanken mit Mathematik verknüpft. Diese Inschrift findet sich in Cornwall auf dem Grab von Hannibal Basset (1686-1708). 291 Jahre später prägten die Brüder Wachowski den Ausspruch: "The matrix is everywhere."

⁸ 1989 sah sich Stuart HUNTER [6] dazu veranlasst, vor diesen Feldern in einer Publikation zu warnen: Let's All Beware the Latin Square.

Was hat das alles mit einem Versuchsplan zu tun? Nun, kommen wir zurück zum Ursprung der DoE und die Arbeiten von R. Fisher, ca. um 1920. Fisher hatte zweidimensionale Gebilde als Testfelder zur Verfügung. Dies im wahrsten Sinne des Wortes, denn er betrieb landwirtschaftliche Züchtungsforschung. Auf diesen zweidimensionalen Gebilden wollte er natürlich mehr als zwei Faktoren untersuchen, idealerweise auch auf mehr als zwei Stufen. Die Euler'schen Quadrate waren als mathematische "Spielerei" damals bekannt. Ihr Einsatz zur Festlegung der Beplantung auf einem Acker war hingegen innovativ. Eine wunderbare Zusammenstellung der historischen Entwicklung findet sich übrigens bei ANDERSON [1].

a α	b δ	c β	d γ
d β	c γ	b α	a δ
b γ	a β	d δ	c α
c δ	d α	a γ	b β

Tabelle 2.19 Griechisch-lateinisches Quadrat nach Leonhard Euler (1707-1783) In jeder Zeile und in jeder Spalte kommen alle Buchstaben genau einmal vor. Dieses Feld ist ausgewogen.

Vereinfachend betrachten wir nun ein Feld mit neun Segmenten. Trivial ist die Belegung mit zwei Faktoren auf jeweils drei Stufen. Diese werden einfach zeilenweise bzw. spaltenweise angeordnet. Mit der Idee von Euler ergibt sich nun die Möglichkeit, zwei weitere Faktoren einzubringen, ohne die Zahl der Segmente zu erhöhen.

B	+	C+	D+	C-	D0	C0	D-	
	0	C0	D0	C+	D-	C-	D+	
	-	C-	D-	C0	D+	C+	D0	
		-		0			+	
					A			

Tabelle 2.20 L9 Versuchsplan mit neun Einstellungen und vier Faktoren auf jeweils drei Stufen. Für A- kommt jede Stufe von C und D genau einmal vor, bei B+ ebenso. Bei C0 oder D- findet man jede Stufe von A und auch von B. Jede beliebige Kombination funktioniert. Das Feld ist ausgewogen.

Felder dieser Art sind sehr dicht besetzt, daher klein, aber auch schwer auszuwerten. Taguchi hat diese Felder sehr gerne eingesetzt [5], war damit jedoch bei Statistikern umstritten. Trotz aller Kritik war Taguchi erfolgreich, was letztlich auch ein wenig auf die Robustheit der Methode zurückzuführen ist. Nur die Stufenmittelwerte der Faktoreinstellungen lassen sich aus den Ergebnissen sicher ablesen, keine quadratischen Effekte und keine Wechselwirkungen.

Literaturverzeichnis

1. Anderson, L.D.: *The history of latin squares*. Aalborg University, Dept. of Mathematical Sciences (2007) 58
2. Anderson, M., Whitcomb, P.: *RSM Simplified*. Productivity Press, New York (2005) 42, 65, 82, 83
3. Bothe, H.H.: *Neuro-Fuzzy Methoden*. Springer Verlag, Berlin Heidelberg New York (1998) 23, 48
4. Fahrmeir, L., Kneib, T., Lang, S.: *Regression*. Springer Verlag, Berlin Heidelberg (2009) 55, 56, 61, 65, 213
5. Fowlkes, W., Creveling, C.: *Engineering Methods for Robust Product Design*. Addison-Wesley, Reading, MA (1995) 2, 34, 58, 139
6. Hunter, S.J.: *Let's all Beware the Latin Square*. Quality Engineering **1** (4), pp. 453 – 465 (1989) 57
7. Montgomery, D.C.: *Design and Analysis of Experiments*. John Wiley, New York (2001) 2, 32
8. Pokropp, F.: *Lineare Regression und Varianzanalyse*. Oldenbourg Verlag, München Wien (1999) 55
9. Schulte, H., Platzbäcker, W., Siebertz, K., Lach, R.: *Design of Experiments (DoE) in der Motorenentwicklung*, chap. Hydrodynamic Bearing Calculation as a Potential DoE Application within the Engine Development Process, pp. 1–19. Expert Verlag, Renningen (2003) 42
10. Siebertz, K.: *Front Impact Occupant Models with Finite Element Structures to Investigate Lower Leg Loads*. In: European MADYMO User's Conference, Stuttgart (1999) 31
11. Toutenburg, H., Fieger, A.: *Deskriptive Statistik*. Prentice Hall, München London (1998) 55

Kapitel 3

Kontrollverfahren

3.1 Einleitung

Die Durchführung einer Versuchsreihe ist in der Regel kostspielig. Oft besteht aus organisatorischen Gründen nicht die Möglichkeit einer Wiederholung. Umso wichtiger ist es, von der Planung über die Auswertung bis zur Festlegung einer Systemverbesserung die richtigen Kontrollverfahren einzusetzen. In diesem Kapitel sind alle wesentlichen Kontrollverfahren zusammengefasst. Dort wo es nötig ist, werden einige statistische Grundlagen nachgeliefert. Der korrekte Einsatz der Kontrollverfahren ist auch für Ingenieure leicht erlernbar und setzt kein Mathematikstudium voraus.

Grundsätzlich ist es bei Versuchsreihen immer notwendig, alle Arbeitsschritte zu kontrollieren. Insofern sind die hier erwähnten Kontrollverfahren keine Schwäche der statistischen Versuchsplanung, sondern eine Stärke, da der strukturierte Versuchsplan im Vergleich zu einzelnen "ad hoc Versuchen" neue Möglichkeiten eröffnet, um die Interpretation der Ergebnisse abzusichern. In diesem Zusammenhang könnte man auch von einer Diagnose sprechen. Es geht letztlich darum, effizient und zielstrebig potentielle Fehler zu finden.

Die bei der statistischen Versuchsplanung verwendeten globalen Beschreibungsmodelle bauen auf Regressionsverfahren auf. Regressionsverfahren haben sich stetig im Laufe der letzten beiden Jahrhunderte entwickelt [2]. Sie sind sehr gut analysiert und bieten eine Fülle von Kontrollverfahren, um ihre korrekte Anwendung sicherzustellen, denn bei jedem Arbeitsschritt können Probleme auftauchen.

Nur sehr wenige Auswerteprogramme werden in einer deutschen Version angeboten. Auch die weiterführende Fachliteratur ist fast ausschließlich in englischer Sprache verfügbar. Daher erschien es an dieser Stelle sinnvoll, die Originalbezeichnungen in englischer Sprache beizubehalten und keine Übersetzung zu bringen, die in der praktischen Anwendung nur Verwirrung stiftet.

Versuchsplan Zunächst gilt es, einen Versuchsplan aufzustellen, der die Erwartungen an die Versuchsreihe erfüllt. Beschreibungsmodell und Versuchsplan sind miteinander verknüpft, also muss man sich bereits bei der Planung darüber Ge-

danken machen. Manchmal verläuft die Versuchsdurchführung anders als geplant und dann stellt sich die Frage, ob die verfügbaren Ergebnisse eine sichere Auswertung zulassen.

Beschreibungsmodell Das Beschreibungsmodell ist das zentrale Ergebnis der Untersuchung. Welcher Faktor spielt eine Rolle? Sind die Effekte reproduzierbar? Gibt es Wechselwirkungen oder Nichtlinearitäten? In diesem Kapitel werden einige der Prüfverfahren kurz vorgestellt. Das Kapitel *statistische Modellbildung* vertieft die Thematik und liefert mit anschaulichen Beispielen die erforderlichen statistischen Grundlagen.

Genauigkeit der Vorhersage Bereits die nicht reproduzierbare Versuchsstreuung führt zu Abweichungen zwischen Vorhersage und Testergebnis. Hinzu kommen Ausreißer, die Vereinfachungen des Modells und fehlende Einflussgrößen, deren Einstellung während der Versuchsdurchführung eventuell nicht konstant war. Wie gut ist das Beschreibungsmodell? Gab es Pannen bei der Versuchsdurchführung? Hierzu gibt es passende Kontrollverfahren.

Die lückenlose Absicherung gehört zu den unbestrittenen Stärken der statistischen Versuchsplanung. Man wird nicht für jede Versuchsreihe alle Prüfverfahren brauchen, aber es ist gut, wenn man weiß, dass diese im Zweifelsfall verfügbar sind.

3.2 Versuchsplan

Nur ein korrekter Versuchsplan liefert sinnvolle Ergebnisse. Schlecht konditionierte Versuchspläne entstehen durch:

- einen unpassenden Versuchsplan zu Beginn der Untersuchung,
- notwendige Abweichungen vom Versuchsplan, weil einzelne Kombinationen nicht durchführbar sind,
- Fehler bei der Versuchsdurchführung. Dadurch entsteht natürlich in der Realität ein neuer Versuchsplan, der in der Regel nicht mehr orthogonal ist.

Versuchsplan und gesuchtes Beschreibungsmodell müssen zusammenpassen. Zum Beispiel kann ein zweistufiger Versuchsplan keine quadratischen Effekte liefern und Screening-Versuchspläne lassen in der Regel nicht die Analyse von Wechselwirkungen zu. Eine sorgfältige Bedarfsanalyse im Vorfeld führt schnell zu einem vernünftigen Kompromiss zwischen Aufwand und Genauigkeit.

Vorversuche mit extremen Kombinationen reduzieren das Risiko unerfreulicher Überraschungen, aber oft stellt sich erst während der Versuchsdurchführung heraus, dass eine Kombination Schwierigkeiten bereitet. Zum Beispiel gibt es bei der Kalibrierung von Verbrennungsmotoren mitunter Einstellungen, bei denen das Gemisch nicht mehr zündet. Obwohl die Faktoren einzeln betrachtet moderat verstellt wurden, erweist sich die Kombination als kritisch. Dann muss ein Ersatzpunkt angefahren werden, der so nahe wie möglich an der ursprünglich geplanten Einstellung liegt. Der nun entstandene Versuchsplan ist nicht mehr vollständig orthogonal, aber in der Regel noch immer sehr gut verwendbar.

Ein weiterer Klassiker ist der sogenannte “Dreher”, also eine vertauschte Reihenfolge der Stufeneinstellungen einzelner Faktoren. Besonders gefährdet sind Messreihen, bei denen das Material verbraucht wird, zum Beispiel Crash-Tests. Ein einziger Fehler erzeugt mitunter schon ein logistisches Problem, wenn Ersatzmaterial fehlt. Oft wird bei solchen Situationen aus Verlegenheit mit dem verbleibenden Material eine weitere falsche Kombination getestet, nur um die Versuchsreihe formal abzuarbeiten.

Wichtig ist also die Erfassung des in der Realität getesteten Versuchsplans. Es ist bequem, einen Versuchsplan aufzustellen und sich erst dann wieder mit der Angelegenheit zu beschäftigen, wenn die Ergebnisse vorliegen. In der Praxis scheitert diese Vorgehensweise häufig, weil die eigentliche Durchführung der Versuchsreihe oft in anderen Fachabteilungen oder beauftragten Firmen stattfindet, ohne dass die Sachbearbeiter ausreichend informiert wurden. Im Zweifelsfall ist es immer besser, in eine gute Versuchsvorbereitung zu investieren, als im Nachhinein mit statistischen Tricks eine schlecht ausgeführte Messreihe retten zu wollen. Die tatsächlich gefahrenen Einstellungen müssen dokumentiert werden und bilden den realen Versuchsplan. Für die Auswertung ist nur der reale Versuchsplan relevant.

3.2.1 Fallbeispiel

In der Praxis kann es mitunter vorkommen, dass ein Versuchslauf nicht durchgeführt wurde oder sich die Durchführung als unmöglich herausstellt. Einen solchen Fall greifen wir nun heraus. Angenommen, der gewünschte Versuchsplan besteht aus acht Versuchen mit vier Faktoren auf jeweils zwei Stufen. Wenn nun der letzte Versuchslauf fehlt, entsteht ein realer Versuchsplan, der weder ausgewogen noch orthogonal ist. Die brennende Frage ist, ob die existierenden Ergebnisse bereits eine Analyse zulassen oder ob der fehlende Versuchslauf unbedingt nachgeholt werden muss.

A	B	AB	C	AC	BC	ABC	y
BCD	ACD	CD	ABD	BD	AD	D	
—	—	+	—	+	+	—	y_1
+	—	—	—	—	+	+	y_2
—	+	—	—	+	—	+	y_3
+	+	+	—	—	—	—	y_4
—	—	+	+	—	—	+	y_5
+	—	—	+	+	—	—	y_6
—	+	—	+	—	+	—	y_7
+	+	+	+	+	+	+	y_8

Tabelle 3.1 Teilstudieller Versuchsplan mit vier Faktoren auf zwei Stufen und 8 Versuchen. So war die Versuchsreihe ursprünglich geplant.

<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>y</i>
—	—	—	—	y_1
+	—	—	+	y_2
—	+	—	+	y_3
+	+	—	—	y_4
—	—	+	+	y_5
+	—	+	—	y_6
—	+	+	—	y_7

Tabelle 3.2 Teilstudieller Versuchsplan mit vier Faktoren auf zwei Stufen und 7 Versuchen. Der letzte Versuchslauf fehlt, der reale Versuchsplan bedarf einer kritischen Prüfung.

3.2.2 Korrelationsmatrix

Auswerteprogramme bieten die Korrelationsmatrix als Diagnoseinstrument an. Die Korrelationsmatrix zeigt auf einen Blick die Korrelationskoeffizienten der Effekte. Im Idealfall korrelieren die Effekte nur mit sich selbst, also zeigt die perfekte Matrix den Wert 1 in der Hauptdiagonalen und alle anderen Werte sind 0.

Bei Teilstudioplänen der Auflösungsstufe III sind die Haupteffekte mit Zweifachwechselwirkungen vermenkt, also tauchen in der Korrelationsmatrix in den betroffenen Zellen Werte auf, deren Betrag größer als Null ist. Teilstudiopläne der Auflösungsstufe IV liefern Korrelationen der Wechselwirkungen untereinander. Bei regulären Versuchsplänen kommen in der Korrelationsmatrix nur die Werte -1, 0 oder 1 vor. Bei irregulären Versuchsplänen gibt es auch Zwischenwerte. Abweichungen vom ursprünglichen Versuchsplan bewirken immer eine deutliche Änderung der Korrelationsmatrix. Oft sind fast alle Werte ungleich Null. Entscheidend ist der Betrag des Korrelationsfaktors in Verbindung mit der Effektstärke. Ein starker Haupteffekt kann einen schwachen Effekt bis zur Unkenntlichkeit überlagern, mitunter schon bei Korrelationskoeffizienten von 0,2 oder 0,3. Schwache Zweifachwechselwirkungen hingegen, werden auch bei einem Korrelationsfaktor von 0,5 einen starken Haupteffekt nur geringfügig verfälschen.

	MW	A	B	C	D
MW	1	0,25	0,25	0,25	0,25
A	0,25	1	0,25	0,25	0,25
B	0,25	0,25	1	0,25	0,25
C	0,25	0,25	0,25	1	0,25
D	0,25	0,25	0,25	0,25	1

Tabelle 3.3 Korrelationsmatrix der Effekte (correlation matrix of effects / coefficients). Die Effekte bzw. die Regressionskoeffizienten des Fallbeispiels korrelieren untereinander. Dies erschwert die Auswertung. Mit MW ist der Mittelwert gekennzeichnet, dieser stellt eine eigenständige Modellkonstante dar.

Rechnet man die Korrelationskoeffizienten der Faktoreinstellungen aus, kommt man zur Korrelationsmatrix der Faktoren (correlation matrix of factors). Einzelne Elemente der Matrix lassen sich durch eine lineare Regression der beteiligten Faktoren (nach PEARSON) leicht überprüfen. Auch hier ist im Fallbeispiel eine Korrelation feststellbar. Diese Matrix betrachtet die Faktoreinstellungen und nicht die Modellkonstanten, also fehlen Zeile und Spalte für dem Mittelwert.

	A	B	C	D
A	1	-0,16	-0,16	-0,16
B	-0,16	1	-0,16	-0,16
C	-0,16	-0,16	1	-0,16
D	-0,16	-0,16	-0,16	1

Tabelle 3.4 Korrelationsmatrix der Faktoren (correlation matrix of factors). Die Einstellungen der Faktoren korrelieren untereinander. Die Werte der Koeffizienten entsprechen den jeweiligen Korrelationskoeffizienten nach Pearson.

3.2.3 Varianz-Inflations-Faktor (VIF)

Bei einer größeren Zahl von Faktoren wird die Korrelationsmatrix unübersichtlich. Dann bietet sich die Betrachtung der Varianz-Inflations-Faktoren (variance inflation factor, VIF) an [1, 2]. Der Varianz-Inflations-Faktor gibt an, mit welcher Varianzverstärkung zu rechnen ist. Ein Wert von 1 entspricht dem Idealzustand. Ab einem Wert von 5 ist Vorsicht geboten. Ab 10 ist der betroffene Effekt praktisch nicht mehr auswertbar. Kollinearitäten im Versuchsplan führen zu einem schlecht konditionierten Gleichungssystem. Dadurch wird die Trennung der Effekte beeinträchtigt. Der VIF wertet die Korrelation der erklärenden Variablen untereinander aus.

$$VIF_j = \frac{1}{1 - R_j^2} \quad (3.1)$$

R_j^2 ist das Bestimmtheitsmaß einer erklärenden Variablen j in Bezug zu allen übrigen erklärenden Variablen. Eine erklärende Variable ist in diesem Zusammenhang nicht nur ein Faktor, sondern beinhaltet auch alle separat im Beschreibungsmodell berechneten Effekte höherer Ordnung, zum Beispiel Wechselwirkungen oder quadratische Effekte. $R_j^2 = 0$ ist der ideale Wert, dann ist die Variable j völlig unabhängig von allen anderen Variablen. $R_j^2 = 1$ belegt eine Kollinearität. Dann ist der Effekt der Variablen j nicht bestimmbar. Jede Modellkonstante hat einen Vertrauensbereich, der von der Teststreuung abhängt. Der VIF vergrößert die Unsicherheit, verbreitert also den Vertrauensbereich. Anders ausgedrückt, die Varianz der betroffenen Modellkonstanten wird größer und zwar um Faktor VIF im Vergleich zu einem idealen Versuchsplan.

Faktor	VIF
A	1,14
B	1,14
C	1,14
D	1,14

Tabelle 3.5 Varianz-Inflations-Faktoren (VIF) für das Fallbeispiel.

3.2.4 Fraction of Design Space (FDS)

Hobbyfotografen wissen, dass zu den Rändern hin die Abbildungsleistung der Objektive nachlässt. Praktisch alle Objektive liefern im Zentrum des Bildfeldes ihr Opti-

timum. Gute Objektive unterscheiden sich von schlechten Objektiven durch einen entsprechenden Ausgleich, um den Leistungsabfall zu begrenzen. Bei Versuchsplänen ist dies ähnlich. Die Genauigkeit der Vorhersage hängt neben der allgemeinen Teststreuung auch davon ab, wo sich der gesuchte Punkt befindet. Im mittleren Bereich des Faktorraums kann man in der Regel von einer genaueren Vorhersage ausgehen. Zu den Rändern hin steigt die Unsicherheit. Der FDS-Plot [6] zeigt an, wie stark dieser Anstieg für den betroffenen Versuchsplan ist. Hierzu werden der kumulierte Anteil des Faktorraums (design space) auf der Abszisse und die Standardabweichung auf der Ordinate aufgetragen.

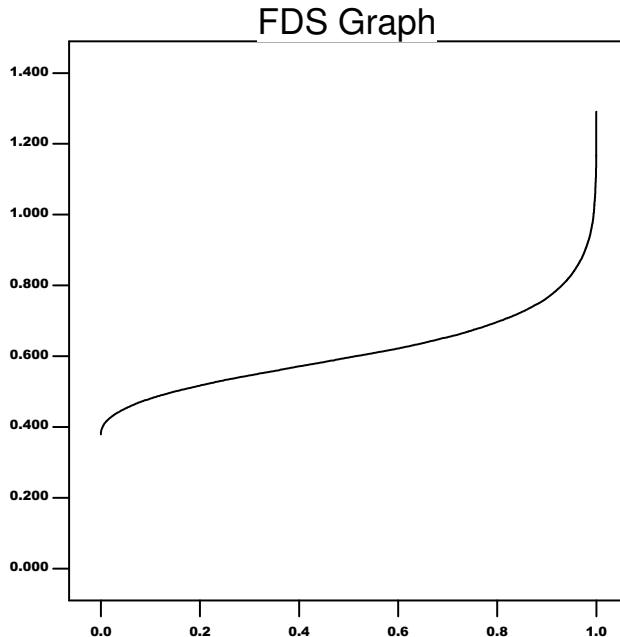


Abb. 3.1 FDS-Plot für das Fallbeispiel. Aufgetragen wird die erwartete Standardabweichung des Mittelwertes über den kumulierten Anteil des Faktorraums (Fraction of Design Space). Im Zentrum des Faktorraums ist die Genauigkeit höher als am Rand. Eine Extrapolation über die Grenzen des Faktorraums hinaus ist nicht zulässig.

3.2.5 Hebelwerte

Es kann durchaus Fälle geben, in denen nicht alle Versuchsläufe die gleiche Bedeutung für das Gleichungssystem haben. So ist zum Beispiel der Zentralpunkt beim Box-Behnken-Design ein sehr wichtiger Punkt. Ausreißer wirken sich hier dramatischer aus, als bei anderen Punkten. Von der einfachen eindimensionalen Regression

kennt man diesen Effekt. Ist ein Punkt weit vom Zentrum der Punktwolke entfernt, hat er einen großen Einfluss auf die Regressionskoeffizienten. Er arbeitet sozusagen mit einem langen Hebel. Aus dieser Betrachtung stammt der Begriff *Hebelwert* (leverage). Bei der Planung der Versuche kann man dem leicht Rechnung tragen, indem Punkte mit großem Hebelwert mehrfach getestet werden. Die Versuchswiederholungen an den neuralgischen Punkten dämpfen die Teststreuung gezielt und vermeiden böse Überraschungen.

Die Abbildung der getesteten Werte y auf die vorhergesagten Werte \hat{y} erfolgt über die sogenannte Hutmatrixt \mathbf{H} .

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \quad (3.2)$$

Diese ist definiert als

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad (3.3)$$

\mathbf{X} bezeichnet die Koeffizientenmatrix des linearen Gleichungssystems, mit dem die Modellkonstanten \mathbf{c} berechnet werden.

$$\mathbf{y} = \mathbf{X}\mathbf{c} + \boldsymbol{\varepsilon} \quad (3.4)$$

Die Diagonalelemente h_{ii} der Hutmatrixt sind die Hebelwerte für die jeweiligen Versuchswerte y_i .

$$\frac{1}{n_r} \leq h_{ii} \leq 1 \quad (3.5)$$

Ein Hebelwert nahe bei Eins hat einen sehr starken Einfluss auf das gesamte Gleichungssystem. Es gibt eine Faustformel zur Berechnung der kritischen Hebelwerte, in Abhängigkeit von der Zahl der Modellkonstanten n_m .

$$h_{ii_{krit}} = 2 \frac{n_m}{n_r} \quad (3.6)$$

Versuchslauf	Hebelwert
1	1
2	0,6
3	0,6
4	0,6
5	0,6
6	0,6
7	0,6

Tabelle 3.6 Hebelwerte (leverage) für das Fallbeispiel. Versuchslauf 1 hat einen hohen Hebelwert, weil sein “Gegenspieler” fehlt. In Versuchslauf 8 hätten alle Faktoren die entgegengesetzte Einstellung gehabt.

Das untersuchte Fallbeispiel ist absichtlich einfach gehalten. Hier hätte man nicht alle Tests gebraucht, um zur Entscheidung zu kommen, den fehlenden Versuchslauf nachzuholen. In der Praxis ist es leider nicht immer so offensichtlich. Die gezeigten Verfahren arbeiten jedoch umso zuverlässiger, je größer die Versuchspläne sind.

Insofern besteht kein Grund zur Sorge. Auch aufwendige Versuchspläne lassen sich schnell kontrollieren.

3.3 Beschreibungsmodell

Zufällige Streuung ist der ständige Begeiter einer jeden Messreihe. Selbst bei CAE Studien können durch Rundungsfehler und numerische Artefakte unkontrollierte Variationen der Ergebnisse entstehen, die einer Versuchsstreuung ähneln. Sobald eine Steuung vorliegt, liefern zwei voneinander unabhängige Versuchsgruppen nicht mehr den gleichen Mittelwert, auch wenn das System keinerlei systematische Änderung erfahren hat. Bei der Versuchsauswertung entsteht nun die Notwendigkeit, wahre Effekte von scheinbaren Effekten zu unterscheiden. Ein wahrer Effekt ist reproduzierbar und beschreibt die Auswirkung einer Systemveränderung auf die Systemleistung. Ein scheinbarer Effekt ist das zufällige Produkt der Versuchsstreuung und daher nicht reproduzierbar. Scheinbare Effekte verschlechtern die Genauigkeit des Beschreibungsmodells. Bei einer Wiederholung der Versuchsreihe nehmen sie andere Werte an und können sogar ihr Vorzeichen wechseln. Es geht also darum, “die Spreu vom Weizen zu trennen”.

In diesem Kapitel wird zunächst das Konzept der Prüfverfahren vorgestellt. Die Trennung von wahren und scheinbaren Effekten ist allerdings von elementarer Bedeutung für die gesamte Methode. Das folgende Kapitel *Statistische Modellbildung* liefert die nötigen statistischen Grundlagen und geht auf diese zentrale Problematik daher im Detail ein.

3.3.1 Half-Normal-Plot

Hinter allen Prüfverfahren zur Effektkategorisierung steckt ein Hypothesentest. Details dazu finden sich in empfehlenswerten Standardwerken der angewandten Statistik [5, 4], brauchen also nicht im Rahmen dieses Buches hergeleitet zu werden. Allerdings sollte jeder Anwender wissen, welche Grundidee dahinter steckt. Ausgehend von der Annahme, dass alle Effekte scheinbar sind, werden die Effekte gesucht, die zu stark sind, um als zufällig zu gelten. Je stärker ein Effekt im Vergleich zu den anderen Effekten ist, umso unwahrscheinlicher ist seine zufällige Entstehung. Für die Beurteilung braucht man also einige Effekte zum Vergleich und ein wenig Statistik zur Abschätzung der Wahrscheinlichkeit.

Scheinbare Effekte haben unterschiedliche Beträge und Vorzeichen, sind aber bei genügend großen Versuchsplänen immer normalverteilt, unabhängig vom untersuchten System. Dies liegt an der Mittelwertbildung bei der Effektberechnung, wodurch der zentrale Grenzwertsatz zum Tragen kommt. Die Häufigkeitsverteilung der Mittelwerte statistisch unabhängiger Zufallsvariablen nähert sich immer mehr einer Gaußverteilung an, je größer die Zahl der Einzelwerte ist. In den Abbildungen

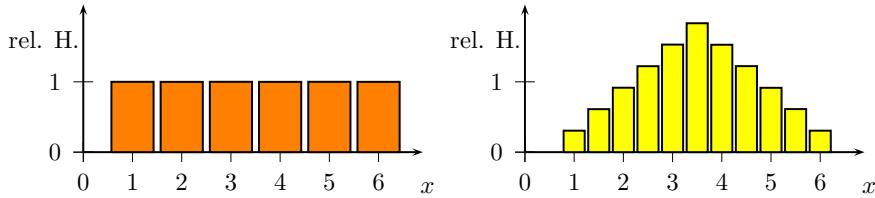


Abb. 3.2 Würfeexperiment mit 48 Millionen Würfen. Auftretenshäufigkeit von Einzelwürfen im Vergleich zu Mittelwerten aus jeweils zwei Würfen. Der Mittelwert 3,5 kann aus mehreren Kombinationen entstehen, ist also im Vergleich zum Mittelwert 1 häufiger zu beobachten. Die relative Häufigkeit bezieht sich auf den arithmetischen Mittelwert der Beobachtungen pro Kategorie.

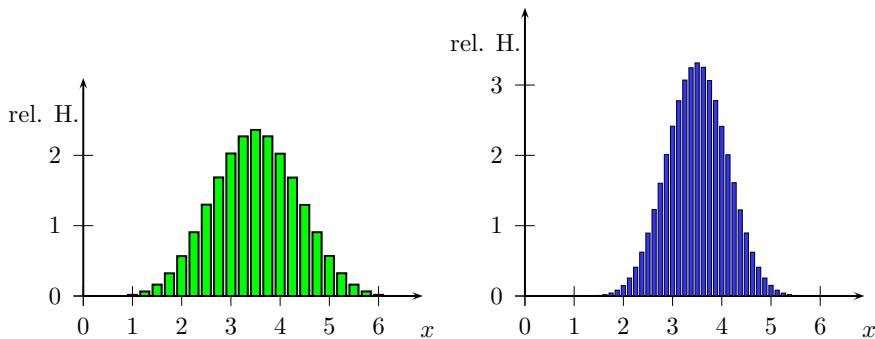


Abb. 3.3 Würfeexperiment mit 48 Millionen Würfen. Auftretenshäufigkeit von Mittelwerten aus jeweils vier und jeweils acht Würfen. Obwohl die Ausgangsverteilung einer nahezu exakten Gleichverteilung entsprach, ergibt sich bereits bei Gruppen von jeweils acht Würfen eine fast perfekte Gaußverteilung der Mittelwerte. Die relative Häufigkeit bezieht sich auf den arithmetischen Mittelwert der Beobachtungen pro Kategorie.

3.2 und 3.3 ist dies am Beispiel einer Gleichverteilung gezeigt. Ein einziger Wurf eines Würfels liefert für jede Augenzahl die gleiche Auftretenshäufigkeit. Der Mittelwert von acht Würfen hingegen folgt bereits weitgehend einer Gaußverteilung. Mittlere Werte zwischen 3 und 4 sind in diesem Beispiel sehr viel wahrscheinlicher als kleine oder große Werte.

Der von Cuthbert DANIEL 1959 entwickelte Half-Normal-Plot nutzt die Gültigkeit des zentralen Grenzwertsatzes und hat bis heute nicht an Bedeutung verloren. Ausgehend von einer Normalverteilung der Effekte, erfolgt zunächst die Betragsbildung, dann der Eintrag in eine spezielle Darstellung, das sogenannte Wahrscheinlichkeitsnetz. Die besondere Konstruktion der Abszisse bewirkt, dass die kumulierte Häufigkeitsfunktion einer Normalverteilung als Gerade erscheint. Hierzu werden die Quantilschritte aufgetragen, also die Flächenschwerpunkte gleichgroßer Teilflächen unter der Normalverteilungskurve. Durch die Betragsbildung genügt die Betrachtung einer Seite der Kurve und zwar der Seite rechts des Mittelwertes. Die flächengleichen Stücke kennzeichnen Bereiche gleicher Wahrscheinlichkeit, also kann man bei einer Normalverteilung davon ausgehen, dass sich diese gleichmäßig auf die abgesteckten Bereiche verteilen. Anders ausgedrückt, wenn alle Effekte

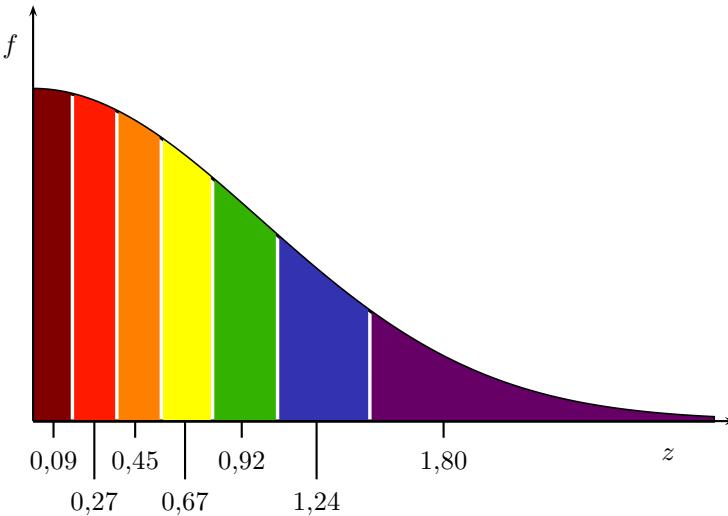


Abb. 3.4 Aufteilung der Normalverteilungskurve in flächengleiche Teilstücke. Die Kurve ist symmetrisch, daher genügt die Betrachtung einer Seite. Die Teilstücke kennzeichnen Bereiche gleicher Auftretenswahrscheinlichkeit. Die Quantilschritte kennzeichnen die jeweiligen Flächenschwerpunkte der Quantile. Bei einem rein zufälligen Prozess sind die Effektbeträge proportional zu den Quantilschritten.

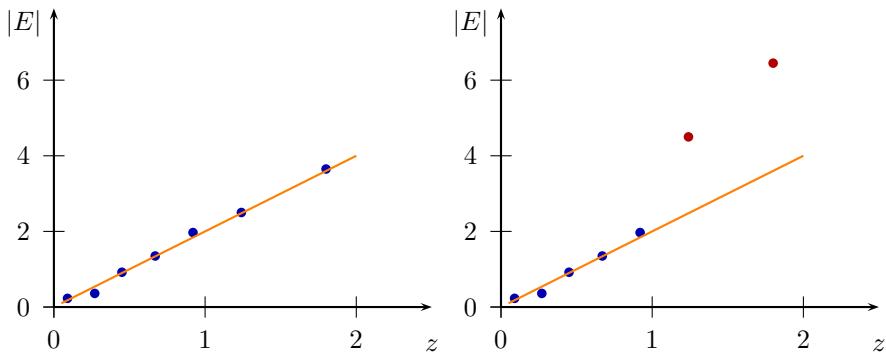


Abb. 3.5 Daniel-Plots für einen Versuchsplan mit acht Versuchsläufen. Scheinbare Effekte sind reine Zufallsprodukte und daher proportional zu den Quantilschritten. Wahre Effekte lassen sich nicht durch Zufall erklären, sie sind stärker, liegen also oberhalb der Ausgleichsgeraden. Auf der linken Seite ist kein wahrer Effekt zu erkennen, auf der rechten Seite hingegen zwei wahre Effekte.

nur scheinbarer Natur sind, ordnen sich die Effektbeträge proportional zu den Quantilschritten an. Hierzu werden die Effektbeträge in aufsteigender Reihenfolge sortiert und mit den Quantilschritten zu Wertepaaren gruppiert. Die Ordinate des Half-Normal-Plots hat die Einheit des Qualitätsmerkmals. Die Werte der Quantilschritte sind in vielen Statistikbüchern tabelliert. Bei computergestützter Auswertung übernimmt das Auswerteprogramm die Erstellung ohnehin automatisch.

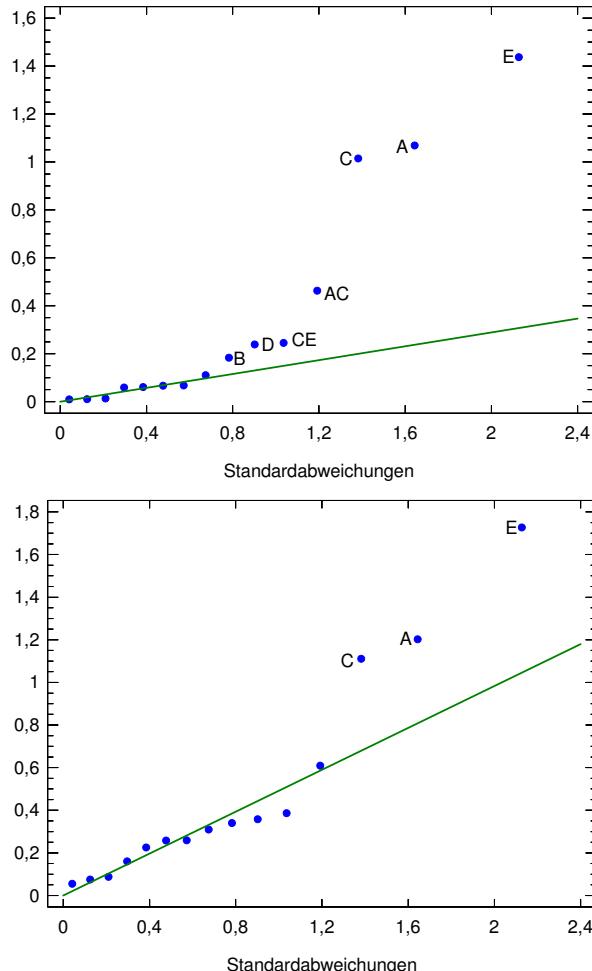


Abb. 3.6 Daniel-Plots für einen Versuchsplan mit fünf Faktoren und 16 Versuchsläufen. Untersucht wurde das Qualitätsmerkmal *Reichweite* beim Rasensprenger (siehe Anhang), mit den Faktoren: Düsenwinkel α (A), β (B), Düsenquerschnitt (C), flüssige Reibung (D) und Druck (E). Den Ergebnissen wurde ein normalverteiltes Rauschen überlagert, um Teststreuung zu simulieren. Oben ist der Rauschanteil niedrig ($\sigma = 0,02m$), unten ist der Rauschanteil sehr hoch ($\sigma = 0,5m$, also in der Größenordnung der starken Effekte). Auch bei starker Teststreuung dämpft der Versuchsplan die Auswirkung der Störung wirkungsvoll. Die starken Effekte bleiben erkennbar.

Die scheinbaren Effekte ordnen sich entlang einer Geraden an, da Quantilschritte und Effektbeträge in einem proportionalen Zusammenhang stehen. Wahre Effekte liegen oberhalb der Geraden, weil sie stärker sind, als es die Zufallsstreuung erwarten lässt. Wegen der Sortierung in aufsteigender Reihenfolge befinden sich typischerweise die scheinbaren Effekte in der Nähe des Koordinatenursprungs und die wahren Effekte in der rechten oberen Ecke. Insbesondere bei kleineren Ver-

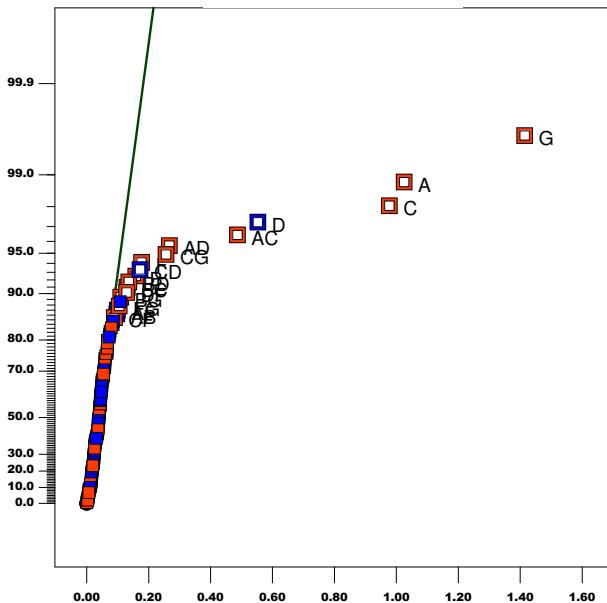


Abb. 3.7 Daniel-Plot der Effekte. Der Half-Normal-Plot kann auch achsenvertauscht aufgetragen sein. Viele Programme liefern dies als Voreinstellung. Die Effektbeträge (als unfreie Größe) werden dann auf der Abszisse und die kumulierte Häufigkeit auf der Ordinate dargestellt. Die Achsentransformation ist analog zur Transformation durch die Quantilschritte, nur die Skalen sind anders. Daher kann man auch hier die wahren Effekte daran erkennen, dass sie nicht zur Ausgleichsgeraden passen.

suchsplänen streuen die scheinbaren Effekte um die Gerade, weil die Mittelwerte nur aus wenigen Messergebnissen gebildet werden. Als wahre Effekte gelten nicht alle Punkte oberhalb der Geraden, sondern nur diejenigen, die grundsätzlich nicht zum Verlauf der Geraden passen. In der Praxis wird man drei Gruppen vorfinden: 1. Effekte, die sich sicher den scheinbaren Effekten zuordnen lassen und um eine Ausgleichsgerade gruppieren. 2. Starke Effekte, die deutlich über der Geraden liegen und mit hoher Wahrscheinlichkeit als wahre Effekte reproduzierbar sind. 3. Die "Grauzone" aus Effekten, die nicht eindeutig zuzuordnen sind.

Kleine Versuchspläne liefern weniger Punkte zur Konstruktion des Half-Normal-Plots. Bei Versuchsplänen der Auflösung III und IV sind die Effekte vermengt, was dazu führen kann, dass kein Effektbetrag klein ist. Dann liegt die Ausgleichsgerade etwas höherversetzt und schneidet nicht mehr den Koordinatenursprung. In den meissten Fällen reicht jedoch der Half-Normal-Plot völlig aus, um eine sichere Entscheidung zu treffen. Die Erstellung dieser Diagramme übernehmen alle guten Auswerteprogramme. Für den Anwender genügt es also, die Grundidee dieses Diagramms zu verstehen und ein Gefühl für die Zuordnung der Effekte in die jeweiligen Kategorien zu entwickeln. In der "Grauzone" löst der Half-Normal-Plot im Vergleich zur Varianzanalyse besser auf und stellt deswegen in jedem Fall eine sinnvolle Ergänzung dar.

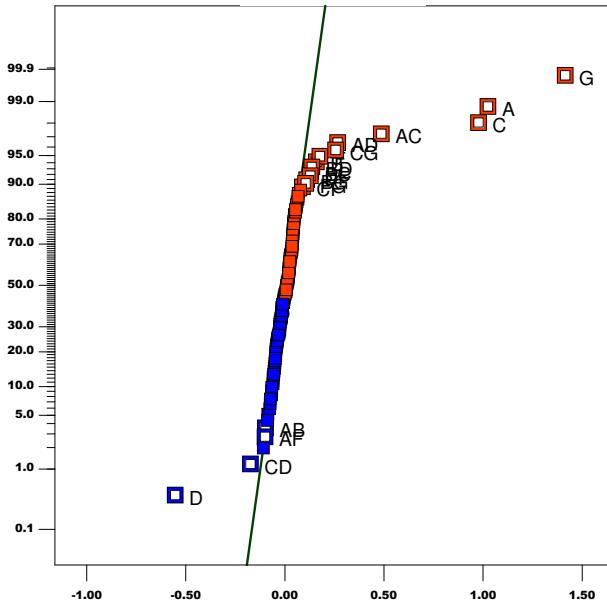


Abb. 3.8 Full-Normal-Plot der Effekte. Beim Full-Normal-Plot entfällt die Betragsbildung der Effekte. Auch hier zeichnen sich wahre Effekte dadurch aus, dass sie nicht zur Ausgleichsgeraden passen. Sie tauchen aber in zwei Regionen auf, je nach ihrem Vorzeichen.

3.3.2 Varianzanalyse

Die Varianzanalyse (ANalysis Of VAriance, ANOVA) bildet das rechnerische Ge- genstück zum Half-Normal-Plot. Die Aufgabe ist gleich, die Grundidee des Hypothesentests ebenfalls. Im Unterschied zur graphischen Lösung berechnet die ANOVA die Wahrscheinlichkeitswerte, um wahre Effekte von scheinbaren Effekten zu unterscheiden. Die Varianzanalyse liefert darüber hinaus auch zusätzliche Informationen zur Güte des Beschreibungsmodells. Auch die Varianzanalyse gehört zum normalen Leistungsumfang eines Auswerteprogramms, also genügt an dieser Stelle die Erläuterung des Grundprinzips.

Ausgangspunkt für die ANOVA ist die in der Versuchsreihe aufgetretene Gesamtvarianz. Diese berechnet sich aus den Werten aller Versuche und dem Gesamtmittelwert.

$$V_{ges} = \frac{1}{n_r} \sum_{i=1}^{n_r} (y_i - \bar{y})^2 \quad (3.7)$$

Im zweiten Schritt erfolgt die Berechnung der Teilvarianzen, die sich den einzelnen Faktoren zuordnen lassen. Abbildung x zeigt, wie man sich diese Aufteilung vorstellen kann. Der Effekt des jeweiligen Faktors bewirkt im Mittel eine Verschiebung aller Ergebnisse um den halben Effektbetrag. Die verbleibende Varianz steht

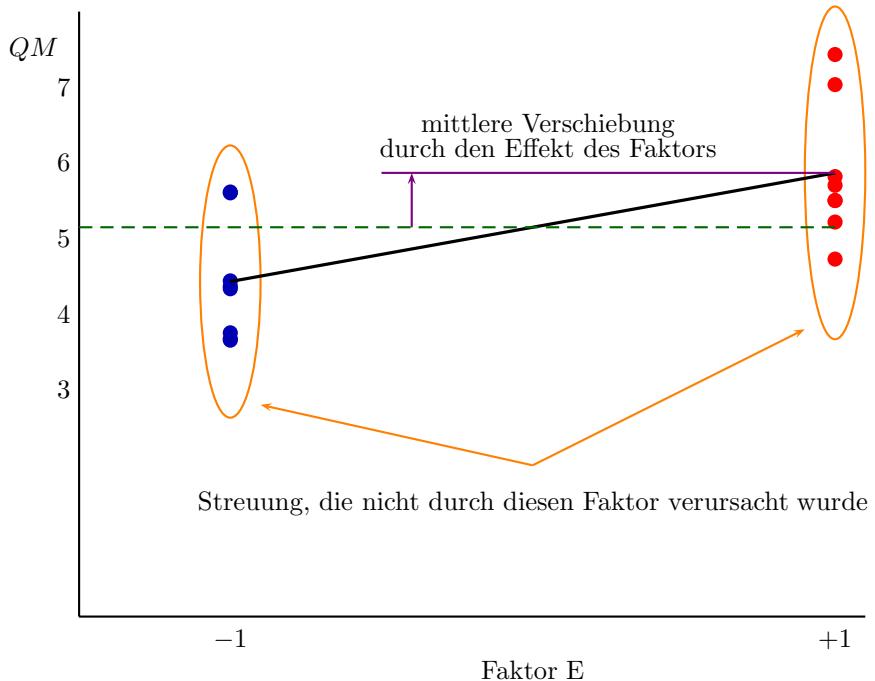


Abb. 3.9 Aufteilung der auftretenden Streuung. Der untersuchte Faktor (hier Faktor E, also *Druck* aus dem Fallbeispiel *Rasensprenger*) verschiebt im Mittel jeden Datenpunkt um den halben Effektbetrug. Die übrige Streuung wird von anderen Faktoren verursacht oder fällt unter die allgemeine Teststreuung.

nicht mit dem untersuchten Faktor in Verbindung, denn sie tritt bei konstanter Einstellung dieses Faktors auf. Der Effekt des Faktors erzeugt somit eine Teilvarianz.

$$V_j = \left(\frac{E_j}{2} \right)^2 \quad (3.8)$$

Unabhängige Teilvarianzen überlagern sich additiv, also entsteht die Gesamtvarianz aus ihrer Summe. Ein System mit n_r Gleichungen und n_m Modellkonstanten hat $(n_r - n_m)$ Freiheitsgrade, denn jede Modellkonstante (auch der Gesamtmittelwert) bindet einen Freiheitsgrad. V_E bezeichnet den Erwartungswert der Teilvarianz pro Freiheitsgrad.

$$V_E = \frac{V_{ges} - \sum_{j=1}^{n_m-1} V_j}{n_r - n_m} \quad (3.9)$$

Von der Gesamtvarianz werden also zunächst alle Teilvarianzen der Effekte abgezogen. Der nicht durch die Effekte erklärbare Restbetrag verteilt sich auf die übrigen Freiheitsgrade. Im nächsten Arbeitsgang erfolgt ein Vergleich der Teilvarianzen mit diesem Erwartungswert. Das liefert eine dimensionslose Kennzahl, das soge-

nannte F-Verhältnis (F-ratio). Liegt die Teilvarianz deutlich über dem Erwartungswert ist das F-Verhältnis deutlich größer als 1 und die Wahrscheinlichkeit ist groß, dass es sich hierbei um einen wahren Effekt handelt. Bei großen Feldern stellt 4 die Signifikanzgrenze dar, unter der Annahme einer fünfprozentigen Irrtumswahrscheinlichkeit. Bei kleinen Feldern oder voll besetzten Feldern¹ ist der berechnete Erwartungswert weniger stabil, denn die Zahl der verbleibenden Freiheitsgrade ist gering. Bei gleicher Irrtumswahrscheinlichkeit muss daher der Grenzwert für das F-Verhältnis ansteigen. Die Statistik benutzt quasi einen Sicherheitszuschlag. Statt der Varianz verwenden Auswerteprogramme meist die “Quadratsumme” oder (sum of squares), weil n_r , also die Zahl der Versuchsläufe, bei der Quotientenbildung (F-Verhältnis) ohnehin entfällt. Im Kapitel *Statistische Modellbildung* wird die ANOVA im Detail erklärt.

Ursache	Quadratsumme	FG	F-Quotient	p-Wert
A	4,566110	1	491,12	0,0000
B	0,134044	1	14,42	0,0067
C	4,115280	1	442,63	0,0000
D	0,227843	1	24,51	0,0017
E	8,258640	1	888,28	0,0000
AC	0,853954	1	91,85	0,0000
AD	0,048472	1	5,21	0,0564
CE	0,240382	1	25,85	0,0014
Rest	0,065081	7		

Tabelle 3.7 ANOVA für das Fallbeispiel Rasensprenger mit 16 Versuchsläufen und fünf Faktoren auf jeweils zwei Stufen. Die Teststreuung ist in diesem Fall so stark, dass kleinere Effekte als Scheineffekte eingestuft werden und nicht den Grenzwert F_{krit} erreichen. Diese wurden in der Tabelle bereits aussortiert. Zur Berechnung von F wird zunächst der nicht durch die Effekte erklärbare Anteil (Rest) durch die Zahl der verfügbaren Freiheitsgrade geteilt (hier: 16-9, also 7). Diese “erwartete Quadratsumme pro Freiheitsgrad” gilt als Referenz (hier: 0,00930). F ist der Quotient aus Quadratsumme und Referenzwert. Der p-Wert bezeichnet die statistische Irrtumswahrscheinlichkeit für die Entscheidung, den Effekt als *wahr* einzustufen.

Auswerteprogramme berechnen aus dem F-Verhältnis unter Berücksichtigung der verfügbaren Freiheitsgrade automatisch die Irrtumswahrscheinlichkeit für die Annahme, dass der jeweilige Effekt wahr ist. Für diesen p-Wert gilt 0,05 üblicherweise als Grenze, entsprechend einer fünfprozentigen Irrtumswahrscheinlichkeit. Unter 0,05 gilt der Effekt als wahr, über 0,05 als scheinbarer Effekt. Natürlich steckt in der Festlegung des Grenzwertes eine gewisse Willkür und bei kleinen Feldern besteht die Gefahr, dass wahre Effekte ausgemustert werden. Der Hypothesentest sichert leider nur eine Richtung, schützt also davor, einen scheinbaren Effekt als wahren Effekt anzusehen. Unklar bleibt folglich, mit welcher Wahrscheinlichkeit ein wahrer Effekt irrtümlicherweise als scheinbarer Effekt eingestuft wird und dem Beschreibungsmodell verloren geht.

¹ Voll besetzte Felder nennt man auch: “gesättigt”.

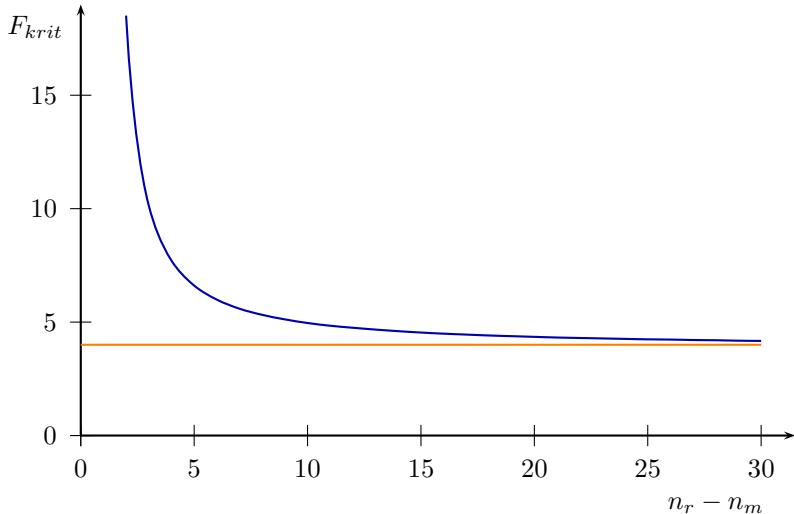


Abb. 3.10 Grenzkurve für F. Bei einer hohen Zahl von verfügbaren Freiheitsgraden konvergiert der Grenzwert gegen 4. Wenn nur wenige Freiheitsgrade zur Verfügung stehen, muss F einen wesentlich höheren Wert erreichen, um mit gleicher statistischer Sicherheit die wahren von den scheinbaren Effekten zu trennen.

Versuchswiederholungen liefern ebenfalls Freiheitsgrade. Bei Anwendung der Varianzanalyse ist es immer ratsam, keine Mittelwertbildung außerhalb des Auswerteprogramms durchzuführen, sondern die Versuche tatsächlich einzeln zu betrachten. Hierzu kopiert man den Versuchsplan entsprechend der Zahl der Wiederholungen untereinander² und trägt die einzelnen Versuchsergebnisse ein. Beispiel: Ein Versuchsplan mit 8 Kombinationen und 2 Wiederholungen liefert insgesamt 24 Versuchsergebnisse. Ist er mit 7 Faktoren besetzt, verbleiben 16 Freiheitsgrade für die Varianzanalyse.

Auf diese Weise bietet die ANOVA im Gegensatz zum Half-Normal-Plot auch die Möglichkeit, die Varianz der Versuchswiederholungen zu analysieren. Der Vergleich der Gesamtvarianz mit der Summe der durch die Effekte erzeugten Teilvarianzen gibt Aufschluss über die Güte des Beschreibungsmodells, allerdings nur in Bezug auf die getesteten Kombinationen. Ist die Differenz klein, erklären die Effekte die Systemantwort fast vollständig. Hierbei ist jedoch Vorsicht geboten. Bleiben scheinbare Effekte im Beschreibungsmodell, so wird dieses fälschlicherweise zu genau an die jeweiligen Testdaten angepasst (over-fit). Die Varianzanalyse gaukelt dann eine hohe Modellgüte vor, weil die Restvarianz klein ist. Eine Versuchswiederholung oder Tests mit neuen Kombinationen führen in diesem Fall zu enttäuschenden Ergebnissen. Ohne die scheinbaren Effekte steigt zwar die Restvarianz, aber

² Viele Auswerteprogramme bieten diese Option automatisch an. In anderen Fällen kann man dies leicht durch ein benutzerdefiniertes Feld bewerkstelligen.

das Beschreibungsmodell ist trotzdem besser, weil es nicht “hinter den verrauschten Testdaten herläuft”.

Im Verbund geben die Varianzanalyse und der Half-Normal-Plot ein hohes Maß an Sicherheit bei der Beurteilung des gewählten Beschreibungsmodells. Die statistische Analyse ersetzt jedoch kein Fachwissen. Aus diesem Grund ist es immer ratsam, nach einer physikalischen Erklärung für die Effekte zu suchen. Oft bringt dieser Dialog zwischen dem Statistiker und dem Anwender aus der entsprechenden Fachabteilung das eigentliche Systemverständnis.

3.4 Genauigkeit der Vorhersage

Eine wesentliche Aufgabe der Methode besteht in der Vorhersage neuer Einstellungen innerhalb des Faktorraums. Natürlich muss dies mit hinreichender Genauigkeit geschehen, sonst sind die Vorhersagen unbrauchbar und führen im schlimmsten Fall zu falschen Schlussfolgerungen. Wie bei allen Modellen, muss man grundsätzlich mit Abweichungen zwischen Vorhersage und Realität rechnen. Es stellen sich nun Fragen nach dem Ausmaß und den Ursachen der Abweichungen. In der Praxis wird man immer eine Kombination mehrerer Ursachen vorfinden und es gilt diese voneinander abzugrenzen, damit im Bedarfsfall Defizite gezielt beseitigt werden können. Auch hier liefert die klassische Regressionsanalyse ein ganzes Arsenal an Kontrollverfahren, um sicher zum Ziel zu gelangen.

Nach Anwendung der Kontrollverfahren muss die Ursachenanalyse natürlich auf die gesamte Versuchsdurchführung ausgedehnt werden. Oft gibt es simple Übertragungsfehler oder Fehler bei der Umrechnung der Rohdaten in die untersuchten Qualitätsmerkmale. In vielen Fällen führen gerade die unerwarteten Ergebnisse zu wertvollen Erkenntnissen, sei es indem sie auf unberücksichtigte Faktoren hinweisen oder auf physikalische Effekte, mit denen nicht gerechnet wurde. Dieser Teil der Analyse hat nicht mehr mit Statistik zu tun, es geht vielmehr um eine fachlich fundierte Interpretation der Ergebnisse.

3.4.1 Fallbeispiel

Rasensprenger, siehe Abbildung 3.11 und Anhang. Untersucht wurde das Qualitätsmerkmal *Reichweite* mit den Faktoren: Düsenwinkel α (A), β (B), Düsenquerschnitt (C), Durchmesser (D), trockene Reibung (E), flüssige Reibung (F) und Druck (G), entsprechend den Einstellungen aus Tabelle 2.9. Als Versuchsplan kam ein Vollfaktorplan mit 128 Versuchen zum Einsatz. Überlagertes normalverteiltes Rauschen mit $\sigma = 0,1\text{m}$ dient als Ersatz für die Teststreuung. Einen absichtlich eingebauten Ausreißer in Versuchslauf 13 gilt es zu entdecken.

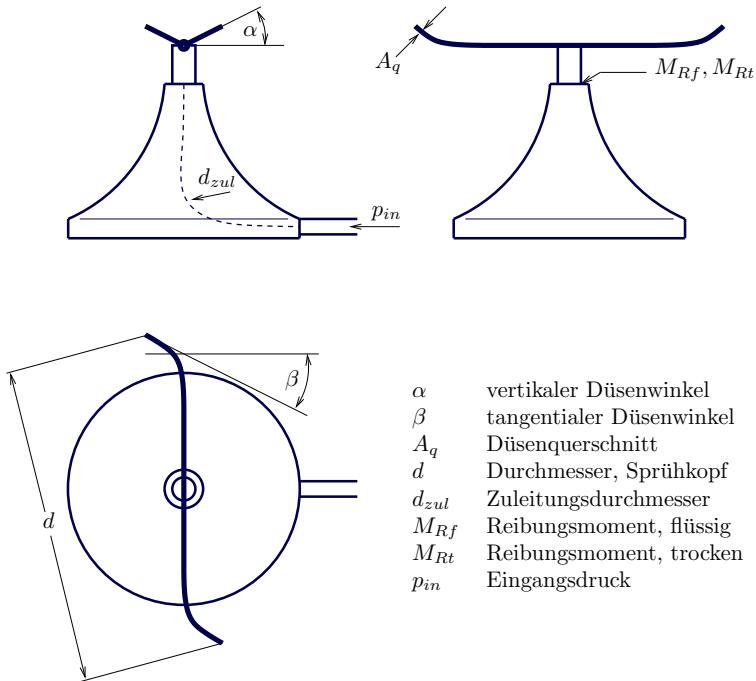


Abb. 3.11 Schematische Darstellung eines Rasensprengers.

3.4.2 Residual-Plots

Nutzt man das Beschreibungsmodell, um Vorhersagewerte für die bereits getesteten Kombinationen zu berechnen, eröffnet sich ohne zusätzlichen Versuchsaufwand eine weitere Kontrollmöglichkeit. Der Residualplot zeigt die Abweichung zwischen Vorhersage und Testergebnis. Große Abweichungen können verschiedene Ursachen haben, zum Beispiel einzelne Ausreißer, ein zu grobes Beschreibungsmodell, eine hohe Teststreuung oder einen dejustierten Versuchsaufbau. Unterschiedliche Darstellungsweisen erleichtern die Ursachenanalyse.

Die Darstellung *predicted vs. actual* zeigt die Vorhersage im Vergleich zu den gemessenen Werten. Diese Darstellung eignet sich hervorragend für einen schnellen Überblick, weil sie die Abweichungen sofort in Relation zu den auftretenden Werten setzt. Dies vereinfacht die Entscheidung darüber, ob die erreichte Modellgüte bereits den Erwartungen entspricht oder weitere Analysen notwendig werden. Streng genommen ist dies noch kein Residual-Plot, findet sich aber bei Auswerteprogrammen typischerweise in dieser Kategorie.

Die Darstellung *residual vs. run order* zeigt die Residuen in der Reihenfolge der Versuche und ist eines der wichtigsten Diagnosewerkzeuge, weil es unmittelbar aufzeigt, in welchen Versuchen die Abweichungen zwischen Vorhersage und Testergebnis besonders gross sind. Viele potentielle Fehler lassen sich hier able-

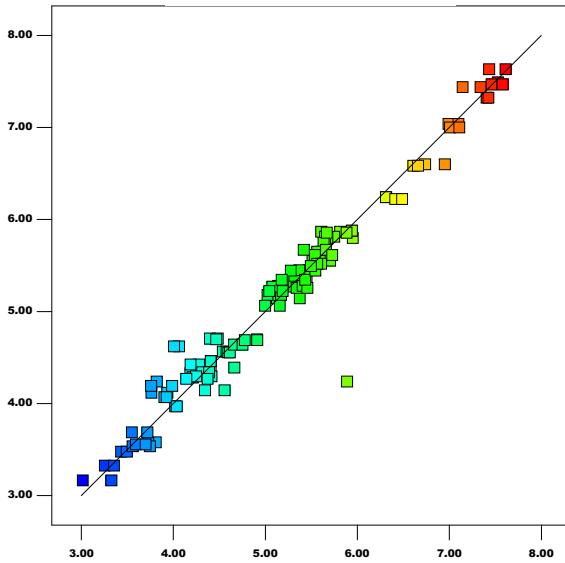


Abb. 3.12 Predicted vs. actual. Hier wird die Vorhersage über den Testergebnissen aufgetragen. Die Teststreuung macht sich bemerkbar. Auch der Ausreißer fällt auf.

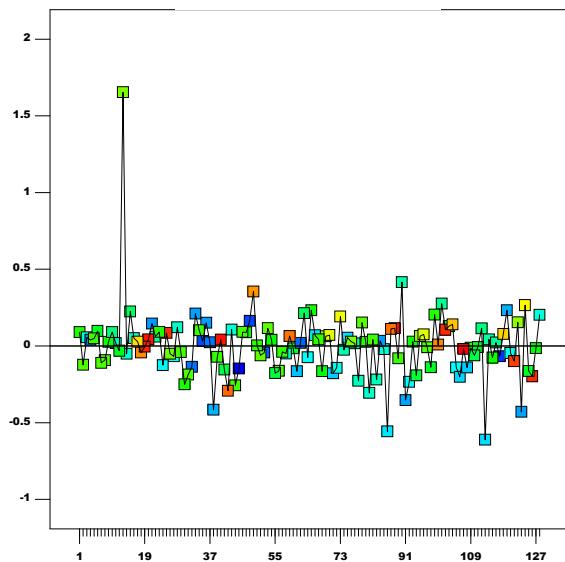


Abb. 3.13 Residual vs. run order. Hier wird die Abweichung zwischen Vorhersage und Testergebnis in der Reihenfolge der Versuche aufgetragen. Der Ausreißer fällt auf und kann eindeutig einem Versuchslauf zugeordnet werden.

sen. Ausreißer fallen auf, aber zum Beispiel auch ein mitten in der Versuchsreihe verstellter Versuchsaufbau.

In vielen Fällen ist die Vorhersage im mittleren Bereich der Versuchsergebnisse wesentlich besser als bei den niedrigsten oder höchsten Ergebnissen. Diese und

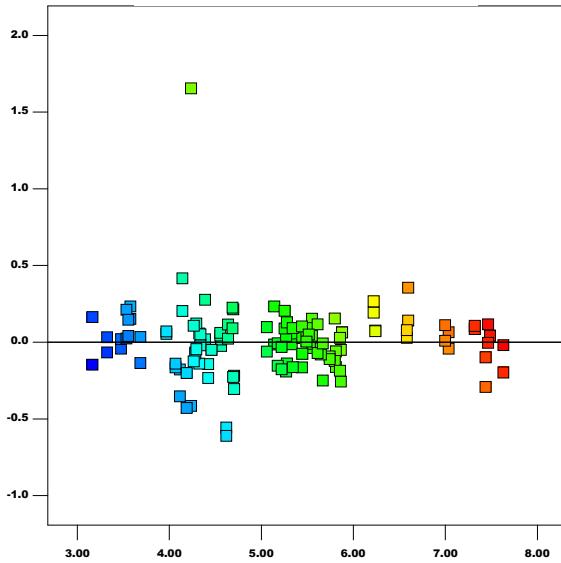


Abb. 3.14 Residual vs. predicted. Hier werden die Residuen über der Vorhersage aufgetragen. Ein systematischer Trend würde dabei auffallen, ist hier jedoch nicht erkennbar.

andere systematische Ungenauigkeiten zeigt die Darstellung *residual vs. predicted*. Die Kombination aus geringer Teststreuung und ungenauem Beschreibungsmodell fällt hier sehr gut auf, sofern vorhanden.

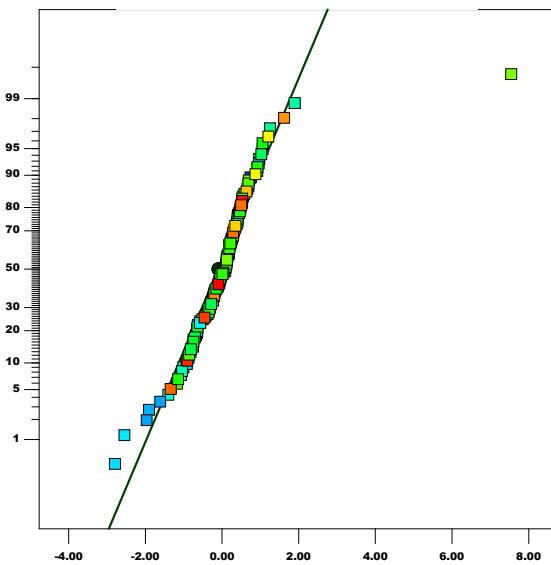


Abb. 3.15 Full-Normal Plot der Residuen. Auch hier fällt der Ausreißer sofort auf, ansonsten lässt die Verteilung der Residuen auf eine zufällige Streuung schließen.

Rein zufällige Schwankungen folgen der Normalverteilung. Daher liegt es nahe, den Full-Normal-Plot nicht nur für die Effekte, sondern auch für die Residuen ein-

zusetzen. Jede systematische Abweichung zwischen Vorhersage und Testergebnis fällt dadurch auf, dass sie nicht mit der Normalverteilung erklärbar ist. Man bildet also lediglich eine Ausgleichsgerade und findet schnell die potentiellen Ausreißer, verschiedene Gruppen von Testläufen oder systematische Schwächen des Modells.

Die Diagnose mit Hilfe der Residuen ist in der Regel sehr sicher und erfordert keine zusätzlichen Versuche. Große Versuchspläne bieten von vornherein mehr Freiheitsgrade als kleine Versuchspläne an. Aufwendige Modelle hingegen zehren viele Freiheitsgrade auf. Das Wechselspiel zwischen Versuchsplan und Beschreibungsmodell ist letztlich entscheidend. Es geht um eine genaue und reproduzierbare Vorhersage des Systemverhaltens, ohne unrealistische Anpassung an streuungsbehaftete Versuchswerte.

3.4.3 Löschdiagnosen

3.4.3.1 DFFITS

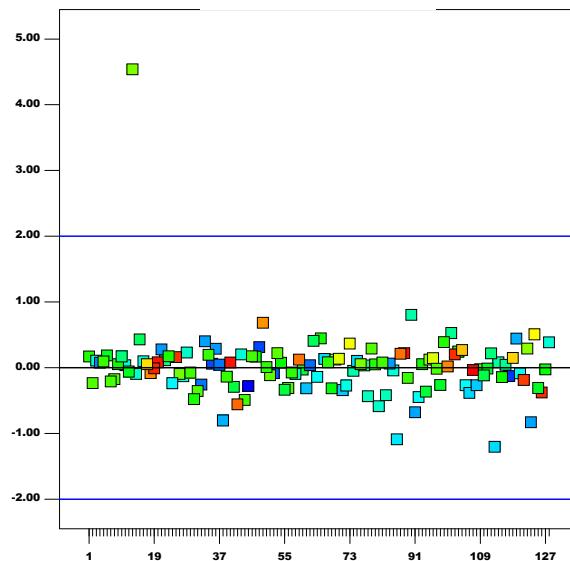


Abb. 3.16 DFFITS-Test. Hier wird die Veränderung der Vorhersage für den Entfall des jeweiligen Testlaufes eingetragen. Hierzu rechnet die Auswertung eine neue Vorhersage ohne den zu prüfenden Lauf und vergleicht diese mit der ursprünglichen Vorhersage. Große Abweichungen sind ein sicheres Zeichen für Ausreißer. Je nach Versuchsplan sind nicht alle Läufe gleich wichtig für das Gleichungssystem, was bei dieser Analyse berücksichtigt wird.

Fällt im Residual-Plot ein möglicher Ausreißer auf, so stellt sich sofort die Frage, wie stark dieser Versuch das gesamte Ergebnis beeinflusst. Die Kenngröße DFFITS

übernimmt genau diese Prüfung [3, 1] und gehört damit zur Kategorie der *Löschdiagnosen*. Hierzu klappert das Auswerteprogramm alle einzelnen Testwerte ab und berechnet neue Vorhersagen für die Testwerte, ohne den jeweiligen Testwert bei der Modellbildung zu berücksichtigen. Die Differenz zwischen der neuen Vorhersage und der ursprünglichen Vorhersage wird auf die geschätzte Teststreuung (ohne Berücksichtigung des Testwertes i) bezogen und als DFFITS deklariert, *difference in fits*. Hierbei gehen auch die jeweiligen Hebelwerte h_{ii} ein. $\hat{\sigma}_{-i}$ ist ein Schätzwert für die Streuung der gesamten Messreihe, ohne Berücksichtigung der Einzelmessung i .

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(-i)}}{\hat{\sigma}_{(-i)} \sqrt{h_{ii}}} \quad (3.10)$$

3.4.3.2 Cook-Distanz

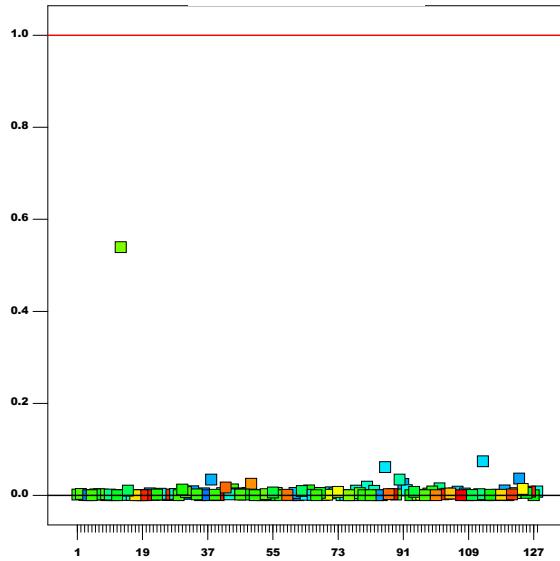


Abb. 3.17 COOK-Test. Die Cook-Distanz gibt Aufschluss darüber, wie stark ein einzelner Wert die Vorhersage für die gesamte Messreihe beeinflusst.

Auch die Cook-Distanz gehört zu den Löschdiagnosen, allerdings wird hier der Einfluss auf die Vorhersage *aller* Ergebnisse ausgewertet. Die Differenz zwischen den Vorhersagen mit dem Punkt y_i und ohne den Punkt y_i wird quadriert, aufsummiert und anders normiert als bei der Berechnung des DFFITS. Ausreißer zeichnen sich dadurch stärker ab. Ab einem Wert von 1 gilt der Ausreißer als kritisch und verfälscht die gesamte Vorhersage. n_m bezeichnet die Zahl der Modellkonstanten und \hat{V} die geschätzte Varianz.

$$D_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(-i)})^2}{n_m \hat{V}} \quad (3.11)$$

3.4.4 Box-Cox Transformation

Wenn sich eine starke Abhängigkeit der Residuen von den vorhergesagten Werten zeigt (residuals vs. predicted) kann die Box-Cox Transformation möglicherweise Abhilfe schaffen. Eine mathematische Transformation des Qualitätsmerkmals ist ohne weiteres zulässig, da die Definition des Qualitätsmerkmals selber einer gewissen Willkür unterliegt. Beispielsweise kann man den Kraftstoffverbrauch eines Fahrzeuges in Litern pro 100 Kilometern angeben oder auch in Miles per Gallon. Die Nachgiebigkeit von Schraubenfedern lässt sich über einen vorgegebenen Weg oder eine vorgegebene Kraft messen. Lebensdauerangaben finden über die mittlere Lebensdauer oder eine Ausfallrate statt, usw. . Die Transformation der Ergebnisgröße (Qualitätsmerkmal) kann die Abhängigkeit von den Faktoren der Untersuchung mathematisch günstiger gestalten, was sich über ein genaueres Beschreibungsmo dell auszahlt [3, 1].

Glücklicherweise gibt es auch hier ein passendes Instrument, um dem Anwender zeitraubende Routinearbeit zu ersparen. Der Box-Cox Plot zeigt an, ob sich eine Transformation lohnt und welche Transformation im speziellen Fall zu den geringssten Residuen führt. Hierzu wird eine allgemein formulierte Funktionsklasse betrachtet, die sogenannte Power-Law-Family. Damit ist kein erfolgreicher Familienbetrieb von Rechtsanwälten gemeint, sondern eine clever formulierte mathematische Transformation, die extrem flexibel ist, aber trotzdem nur von einem Parameter abhängt.

$$z = \begin{cases} y^\lambda & , \lambda \neq 0 \\ \ln(y) & , \lambda = 0 \end{cases} \quad (3.12)$$

Mit dieser Transformation kann man sehr viele Verläufe realisieren. Bei $\lambda = 1$ wird keine Transformation durchgeführt (identische Abbildung), $\lambda = -1$ erzeugt den Kehrwert, $\lambda = 0,5$ die Wurzelfunktion und so weiter.

Mit einer passenden Erweiterung konnten Box und Cox die Transformation so umformen, dass die Ergebniswerte unabhängig von λ in den gleichen Einheiten erscheinen.

$$z = \begin{cases} \frac{y^\lambda - 1}{\lambda g^{\lambda-1}} & , \lambda \neq 0 \\ \ln(y)g & , \lambda = 0 \end{cases} \quad (3.13)$$

$$g = (y_1 y_2 \dots y_{n_r})^{\frac{1}{n_r}} \quad (3.14)$$

Im Box-Cox Plot wird die Summe der Fehlerquadrate (residual sum of squares) als Funktion von λ aufgetragen, woraus sich unmittelbar der optimale Wert für λ ablesen lässt.

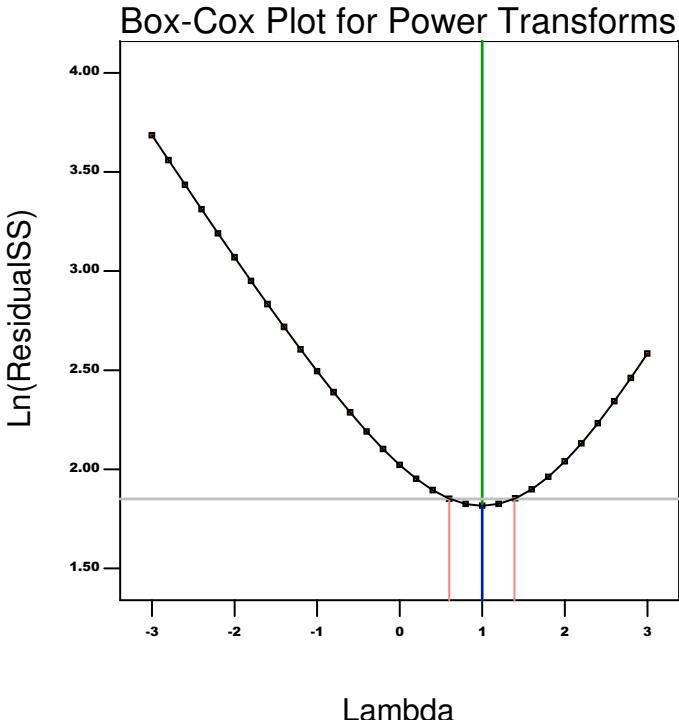


Abb. 3.18 Box-Cox Plot. Für das Fallbeispiel ist offenbar keine Transformation sinnvoll, denn bei $\lambda = 1$ liegt das Minimum der Kurve.

3.4.5 Bestätigungsläufe

Auch wenn die bereits gezeigten Kontrollverfahren das Risiko einer Fehlinterpretation bereits stark reduzieren, bleibt letztlich immer nur der Bestätigungslauf übrig, um absolute Gewissheit zu geben. Insbesondere bei Teilstudiengängen mit dichter Belegung ist die Chance sehr gering, dass die vorgeschlagene Einstellung bereits getestet wurde. Es lohnt sich in diesen Fällen immer, logistisch einen Nachversuch mit der optimierten Einstellung des Systems vorzusehen. Wenn die Möglichkeit besteht, sind auch zwei weitere Einstellungen empfehlenswert: Die Einstellung aller Faktoren auf einen Mittelwert (der sogenannte "center point") gibt Aufschluss über auftretende Nichtlinearitäten. Die Wiederholung der ersten Einstellung kann als zusätzlicher Datenpunkt hilfreich sein, um zeitliche Veränderungen aufzudecken, wie sie zum Beispiel durch Verschleiss oder eine sukzessive Verstellung entstehen.

Literaturverzeichnis

1. Anderson, M., Whitcomb, P.: *RSM Simplified*. Productivity Press, New York (2005) 42, 65, 82, 83
2. Fahrmeir, L., Kneib, T., Lang, S.: *Regression*. Springer Verlag, Berlin Heidelberg (2009) 55, 56, 61, 65, 213
3. Mason, R.L., Gunst, R.F., Hess, J.L.: *Statistical Design and Analysis of Experiments*. Wiley Interscience (2003) 82, 83
4. Papula, L.: *Mathematik für Ingenieure und Naturwissenschaftler, Band 3*. Vieweg, Braunschweig (1997) 68
5. Sachs, L., Hedderich, J.: *Angewandte Statistik, Methodensammlung mit R*. Springer-Verlag Berlin Heidelberg (2009) 68, 234
6. Whitcomb, P.: *FDS - A Power Tool for Designers of Optimization Experiments*. Stat-Teaser pp. 1–3 (2008) 66

Kapitel 4

Statistische Modellbildung

4.1 Einleitung

Nachdem wir in den vorigen Kapiteln einen Weg in die Anwendung der statistischen Versuchsplanung gefunden haben, wird es in der Folge darum gehen, Kapitel 3 aufzugreifen und die dort zu findende Darstellung der wesentlichen Kontrollverfahren zu vertiefen beziehungsweise, wo sinnvoll, zu ergänzen. Es wird dabei um Verfahren gehen, mit denen sichergestellt wird, dass man

- eine *sinnvolle* statistische Auswertung durchführt, die es erlaubt, “echte” Effekte von Effekten zu trennen, die durch das Messrauschen nur vorgetäuscht werden,
- eine dafür sinnvolle Anzahl von Versuchswiederholungen benutzt
- von einem *validen* Vorhersagmodell ausgeht
- mit der schließlich vorgeschlagenen Systemverbesserung richtig liegt.

Nach einigen Gedanken zur Frage, warum wir uns überhaupt mit Statistik befassen müssen (keine Panik — es geht...!) werden wir zunächst einige über 80 Jahre alte Prinzipien bei der statistischen Versuchsplanung kennen lernen: Randomisierung, Versuchswiederholung und Blockbildung — drei Maßnahmen, die uns helfen werden, das “Signal” vom “Rauschen” zu trennen (Kap. 4.3). Diese Trennung führt uns zum Thema des statistischen Testens: Was heißt eigentlich “statistisch signifikant”? Da dieses Thema zwar von zentraler Bedeutung ist, in der Regel aber weder gut erklärt noch von Anwendern gut verstanden wird, werden wir den Grundgedanken des Testens einigen Raum gönnen (Kap. 4.4), bevor wir auf den Kern-Test für geplante Experimente, die sogenannte Varianzanalyse (Analysis of Variance, ANOVA, vgl. 4.5) zu sprechen kommen. Leider sind alle diese Vorbereitungen nötig, um eine zentrale Frage beantworten (und die Antwort verstehen und anwenden) zu können: Wie viele Versuche muss man eigentlich durchführen, um zu gesicherten, verlässlichen Ergebnissen zu gelangen (4.5.6)? Last, but not least, müssen wir noch einmal auf die Residuenanalyse und verwandte Themen zurückkommen: Nur wenn die Voraussetzungen der ANOVA erfüllt sind, gelangt man nämlich zu sinnvollen Ergebnissen. Gilt dies nicht, wird alle “Statistik” zur Makulatur... (vgl. 4.6).

Was aber hat die Auswertung von Versuchen überhaupt mit Statistik zu tun?

4.2 Warum Statistik?

To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of.

— Sir Ronald Aylmer Fisher, ca. 1938¹

The truth is that we all live in a non-stationary world;
a world in which external factors never stay still

— George Box, 1989²

Liegen die Ergebnisse der zuvor geplanten Messreihen vor, kann man sich durchaus Situationen vorstellen, wie sie Christer Hellstrand, Statistiker beim Kugellagerhersteller SKF, 1989 in einem technischen Bericht beschrieben hat: Dort ging es darum, die Wirkung der Modifikationen dreier Designparameter A, B und C auf die Lebensdauer bestimmter Kugellager zu ermitteln. Hellstrand ([9], später auch George Box, vgl. [1]) beschreibt, wie durch einen simplen Vollfaktorplan mit 2^3 Versuchen eine maßgebliche Interaktion entdeckt wurde, die durch “normale” *one factor at a time*-Versuche niemals hätte entdeckt werden können: Während die einzelnen Haupteffekte, wie in Abbildung 4.1 sichtbar wird, eher klein sind, ist der Einfluss der Modifikationen dramatisch, wenn alle drei Modifikationen *zugleich* durchgeführt werden. Die bisherige Lebensdauer von 17 Stunden im Falle dreier nicht modifizierter Parameter konnte durch gleichzeitige Nutzung aller Modifikationen auf 128 Stunden erhöht werden! Dies ist natürlich ein beachtliches — und später zu Recht von Box als Beispiel für den Wert der Betrachtung von Interaktionen hervorgehobenes — Resultat. An dieser Stelle sei aber auch auf einen weiteren Aspekt des Ergebnisses hingewiesen: Im Falle derartig klarer und eindeutiger Versuchsergebnisse erübriggt sich im Grunde die Anwendung statistischer Methoden, denn die Resultate sprechen für sich. Wer würde seinen Chefs keine klare Empfehlung auf der Basis von Abbildung 4.1 geben?

Leider sind die Ergebnisse der meisten Experimente in der Regel nicht so klar. Dies liegt natürlich vor allem daran, dass zufällige Streuung der Begleiter einer jeden Messreihe ist. Sobald jedoch eine Streuung vorliegt, liefern zwei voneinander unabhängige Versuchsgruppen nicht mehr den gleichen Mittelwert, auch wenn das System keinerlei systematische Änderung erfahren hat. Bei der Versuchsauswertung entsteht damit die Notwendigkeit, wahre Effekte von scheinbaren Effekten zu unterscheiden. Während ein wahrer Effekt reproduzierbar ist und die Auswirkung einer Systemveränderung auf die Systemleistung beschreibt, ist ein “scheinbarer”

¹ Den Statistiker hinzuzuziehen, nachdem das Experiment durchgeführt wurde, könnte nicht mehr bedeuten als ihn um eine Autopsie zu bitten: Er könnte sagen, woran der Versuch gestorben ist.

² In Wahrheit leben wir alle in einer nicht-stationären Welt; einer Welt, in der externe Faktoren niemals stillstehen.

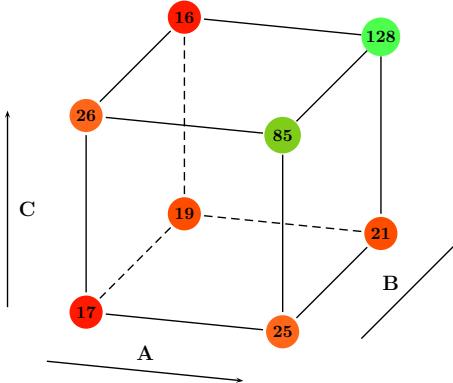


Abb. 4.1 Schematische Darstellung des Einflusses der Designparameter A, B und C: Während die Veränderung von Parameter A die Lebensdauer von 17 auf 25 Stunden, B auf 19 Stunden und C auf 26 Stunden erhöhte, wurde festgestellt, dass die gemeinsame Veränderung aller Parameter zu einer Erhöhung der Lebensdauer des Kugellagers auf 128 Stunden führte!

Effekt das *zufällige Produkt der Versuchsstreuung* — und daher *nicht reproduzierbar*. Scheinbare Effekte verschlechtern somit die Genauigkeit des Beschreibungsmodells, denn bei einer Wiederholung der Versuchsreihe nehmen sie andere Werte an und können sogar ihr Vorzeichen wechseln.

Es geht also darum, “die Spreu vom Weizen zu trennen” — die Quellen unkontrollierbarer Variabilität zu enttarnen und deren Effekte möglichst gut zu verstehen. So stellte sich für den Pionier der Theorie der statistischen Versuchsplanung, R.A. Fisher, die Frage, ob der bessere Ernteertrag eines mit einem neuen Dünger gedüngten Feldes aufgrund der Wirkung des neuen Düngers zu erklären ist — oder beispielsweise aufgrund einer besseren Bodenbeschaffenheit oder sonnigeren Lage des ertragreicherchen Feldes. Ähnliche Fragen stellen sich natürlich auch in industriellen Anwendungen. Leben wir in einer “stationären” Welt, in der wir die der DoE zugrunde liegenden Faktoren kontrollieren, alle anderen Parameter aber konstant halten können?

Mit den Worten von George Box: “*To see if you believe in .. stationarity in your particular kind of work, think of the size of chance differences you expect in measurements taken n steps apart (in time or in space)*” — um festzustellen, ob man an Stationarität in seinem speziellen Arbeitsgebiet glauben kann, denke man an die Größe von zufälligen Unterschieden, die man n Schritte (zeitlich oder räumlich) entfernt erwartet... ([2, S.2]).

Die Antwort liegt natürlich auf der Hand: Wir leben in einer nichtstationären Welt. So basiert beispielsweise jeder Produktionsprozess auf der Einbeziehung von Menschen und Maschinen. Menschen aber ändern ihr Verhalten mit der Zeit, und verschiedene *Operator* werden meist verschiedene Gewohnheiten haben. Maschinen verschleißt, müssen neu eingestellt werden usw. Eingesetzte Materialien können sich von Charge zu Charge ändern (müssen und sollten aber nicht)...

Alle diese Faktoren verursachen Variabilität — neben der schon eingangs erwähnten Messungsgenauigkeit —, und an dieser Stelle kommen Wahrscheinlichkeitsrechnung und Statistik zum Einsatz, die mit den Konzepten von “zufälligen Veränderlichen” und “Wahrscheinlichkeitsverteilungen” eine universelle Sprache zur Beschreibung und Analyse von Variabilität anbieten. Wie groß muss ein Faktoreffekt sein, um als “signifikant” zu gelten? Wie genau müssen wir messen können, um sicher zu sein, dass ein kleiner Effekt “real” ist? Wann sehen wir nur Messrauschen?

Wir benötigen Methoden die uns helfen, zufällige Unterschiede (Pseudo-Effekte) von systematischen Unterschieden zu trennen.

Aufgrund der Bedeutung dieser Aussage und der vielen in der Praxis auftretenden Missverständnisse widmen wir den Grundgedanken statistischen Testens in der Folge viel Raum. Da wir wie bisher davon ausgehen, dass die meisten Anwender der statistischen Versuchsplanung Standard-Software zur Durchführung ihrer Berechnungen einsetzen, wird der Fokus dabei allerdings nicht, wie sonst leider allzu oft, auf der Darstellung der manuellen Rechenschritte liegen, sondern auf einer sorgfältigen Darstellung der Grundprinzipien. Anders gesagt: Es soll *nicht* darum gehen, *wie* die Software rechnet, sondern *was* sie rechnet und *wie* der jeweilige *Output* zu interpretieren ist.

Zuvor sollen jedoch die eng damit zusammenhängenden und seit den Kindertagen der DoE in den zwanziger Jahren des vorigen Jahrhunderts klar definierten drei Grundprinzipien R.A. Fishers als “Brücke in die Statistik” diskutiert werden — Randomisierung, Replikation (Wiederholung) und Blockbildung (Kapitel 4.3). Damit ist der Rahmen abgesteckt, zunächst den zentralen Grundgedanken aller statistischen Tests zu erläutern (Kapitel 4.4), um anschließend diesen Gedanken auf die Varianzanalyse, eine ebenfalls bereits durch Fisher eingeführte Prozedur zur Auswertung randomisierter statistischer Versuchspläne zu übertragen (Kapitel 4.5). Erst mit diesem Instrumentarium wird der Rahmen geschaffen, endgültige, abgesicherte Aussagen aus den Versuchsergebnissen abzuleiten:

“The statistical approach to experimental design is necessary if we want to draw meaningful conclusions from data” — der statistische Zugang zur Versuchsplanung ist nötig, wenn wir sinnvolle Schlussfolgerungen aus den Daten ziehen wollen ([10, S. 11]).

Wenn in diesem Kapitel wiederholt zwei Herren zu Wort kommen, so liegt dies an deren fundamentaler Bedeutung für die Entwicklung sowohl der statistischen Versuchsplanung als auch ihrer “Vermarktung” im industriellen Umfeld (diese Auswahl ist, zugegebenermaßen, etwas subjektiv).

Bei **Sir Ronald Aylmer Fisher** (1890-1962) handelt es sich letztlich um den Erfinder der DoE, deren Wurzeln mindestens bis ins Jahr 1926 zurückgehen. Fisher hatte als junger Statistiker an der landwirtschaftlichen Versuchsanstalt Rothamsted in England einen äußerst folgenreichen Aufsatz über *The Arrangement of Field Experiments* ([7]) — das Arrangieren von Feldexperimenten im wahrsten Sinne des Wortes — veröffentlicht, in dem alle bis heute gültigen Grundprinzipien der Versuchsplanung im wesentlichen entwickelt wurden. Fisher, der ab 1933 verschiedene Professuren in England innehatte und sich später in Australien niederließ, gilt als Mitbegründer der gesamten modernen Statistik. Neben mehr als 300 Aufsätzen

veröffentlichte er bis zu seinem Tode mehrere einflussreiche Bücher, darunter *Statistical Methods for Research Workers*, das seit seiner ersten Auflage im Jahre 1925 insgesamt 14 Neuauflagen und Übersetzungen erlebte ([6]), sowie 8 Auflagen eines Lehrbuchs mit dem Titel *The Design of Experiments*, die zwischen 1935 und 1966 erschienen ([8]). Obwohl — oder gerade: weil — er stets die Anwendungen der Statistik im Auge hatte, ist sein Name eng mit statistischen Themen wie Varianz- und Kovarianzanalyse, Maximum Likelihood-Schätzung oder Diskriminanzanalyse verbunden. Auf Fisher wird in der Folge mehrfach zurückzukommen sein.

Bevor wir nun näher auf den zweiten der erwähnten Herren, **George Edward Pelham Box** (geboren 1919) eingehen, bietet sich noch eine Fußnote zu einem frühen Vorgänger Box' in der industriell orientierten Statistik an, zu dem englischen Wissenschaftler **William Sealey Gosset** (1876-1937). Gosset, vielleicht in der Statistik sogar besser bekannt unter seinem Pseudonym "student", das sich auch in der "student t-Verteilung" und im "student t-test" wiederfindet, kann nämlich in einem gewissen Sinne als Urahm der *Six Sigma*-Bewegung angesehen werden, als früher Anwender statistischer Verfahren zur Verbesserung industrieller Prozesse. Gosset-/Student, der sowohl Chemie als auch Mathematik studiert hatte, entwickelte ab 1899 für die Guinness-Brauerei in Dublin statistische Methoden, die er bei der Planung und Auswertung chemischer Experimente anwandte. Guinness war eine Firma, in der auf wissenschaftliche Methoden gesetzt wurde; bereits 1900 wurden die Guinness Research Laboratories gegründet, in denen Forschungen zu Themen wie Kosten oder Qualität von Gerste und Hopfen betrieben wurden. Gosssets Arbeit machte ihn mit dem Problem kleiner Stichprobenumfänge — geringer Versuchszahlen — bekannt, und aus seiner Beschäftigung mit den Konsequenzen daraus resultierte seine 1908 veröffentlichte Arbeit "The probable error of a mean" ([11]), die er — aus Angst seines Arbeitgebers vor der Veröffentlichung von Betriebsgeheimnissen — nicht unter seinem eigenen Namen publizieren konnte (klingt dies nicht noch immer modern?).

In diesem Sinne hatte **George E.P. Box** also sowohl einen frühen "Großvater" (Gosset) als auch einen berühmten Schwiegervater (er heiratete Fishers zweite Tochter Joan Fisher Box, die später eine vielzitierte Biographie ihres Vaters schrieb). Box, zunächst wie Gosset als Chemiker ausgebildet, wurde ebenfalls durch praktische Probleme motiviert, sich mit statistischen Themen auseinander zu setzen, da er während des zweiten Weltkrieges an Giftgasexperimenten für die britische Armee beteiligt war. Da er keine adäquate statistische Beratung finden konnte, eignete er sich die Grundlagen autodidaktisch an, um nach dem Krieg in London Statistik zu studieren, wo er 1953 über Abweichungen von den der Varianzanalyse zugrunde liegenden Annahmen (vgl. Kapitel 4.5) promovierte. 1960, Box war mittlerweile in die USA übergesiedelt, wurde er Professor für Statistik an der *University of Wisconsin*, wo er bis zu seiner Pensionierung im Jahre 1992 tätig war. Dort war er der erste *chairman* des neu gegründeten *Departments of Statistics* und gründete 1985, gemeinsam mit seinem früheren Doktoranden **William S. Hunter** (1937-1986) das *Center for Quality and Productivity Improvement*.

Wie Fisher veröffentlichte auch Box zahlreiche Artikel und Bücher, unter anderem Klassiker wie den Band *Statistics for Experimenters* ([5]), aber auch unbekanntere wie den (äußerst lesenswerten) Sammelband *Improving Almost Anything* ([4]). Sein Name ist durch Box-Behnken- Designs, Box-Jenkins Modelle der Zeitreihenanalyse und die Box-Cox-Transformation (vgl. Kap. 3.4.4) fest in der Statistik verankert; auch die *Response Surface Methode* wurde von ihm maßgeblich mit entwickelt.

Die Geschichte der industriellen Statistik ist noch weitgehend ungeschrieben. Eine ausführlichere Darstellung käme nicht umhin, beispielsweise die Beiträge von **Walter A. Shewhart** (1891-1967) für die statistische Qualitätskontrolle oder von **W. Edwards Deming** (1900-1993) zu würdigen. Zu Deming sei lediglich angemerkt, dass er nach erfolgreicher Tätigkeit in Japan seit Beginn der achtziger Jahre in den USA Aufmerksamkeit fand — die amerikanische Wirkung seiner Arbeiten fällt somit in die selbe Zeit, in der Box und Hunter das *Center for Quality and Productivity Improvement* gründeten — und in der Forscher von Motorola die Grundzüge von *Six Sigma* entwickelten. Die Zeit für einen forcierten Einsatz statistischer Methoden zur Effizienz- und Qualitätssteigerung war Mitte der achtziger Jahre des letzten Jahrhunderts offenbar reif.

Six Sigma, Mitte der Achtziger bei Motorola entstanden, wurde zu einer erfolgreichen, Statistik-basierten Qualitätsmethode und Managementstrategie, nachdem es zunächst 1995 erfolgreich von General Electric adaptiert wurde. Während es mittlerweile relativ ruhig um *Six Sigma* geworden zu sein scheint, kann man zumindest feststellen, dass es zu einer Demokratisierung der Anwendung statistischer Methoden in vielen Großunternehmen geführt hat: Durch die Ausbildung zahlreicher Mitarbeiter, die unter anderem einen Schwerpunkt auf die statistische Versuchsplanung legt, wurde das Thema in vielen Köpfen — auch auf Management-Ebene — präsent.

Alle in der Folge dargestellten statistischen Analysemethoden basieren auf einer Grundannahme — der Annahme, dass der Versuchsplan korrekt “abgefahren” wurde und dass die vorliegenden Daten für sich genommen stimmig und plausibel sind. Dies ist auch der Startpunkt der Betrachtung vieler gängiger Lehrbücher. Leider zeigt die Erfahrung jedoch, dass man in der Praxis nicht ohne weiteres von dieser Annahme ausgehen kann. Es ist daher zwingend notwendig, sich vor jeder weitergehenden Analyse zunächst einige Gedanken über die Qualität der gelieferten Daten zu machen. Auf diese Notwendigkeit und die zur Kontrolle der generierten Daten empfohlenen Verfahren wurde bereits ausführlich eingegangen (vgl. Kapitel 3.2). Zusätzlich zu den dort bereits erwähnten Aspekten der Datenkontrolle sei an dieser Stelle noch darauf hingewiesen, dass auch die Reihenfolge, in der die Versuche tatsächlich durchgeführt wurden, von der geplanten Reihenfolge abweichen kann. Wir werden in der Folge sehen, warum eine randomisierte Reihenfolge optimal ist, und warum in der Auswertung nicht berücksichtigte Abweichungen von dieser Reihenfolge zu gravierenden Fehlaussagen führen können.

4.3 Randomisierung, Wiederholung, Blockbildung — Fishers Brücke in die Statistik

In all cases, however, we recognize randomization as a postulate necessary to the validity of our conclusions, and the modern experimenter is careful to make sure that this postulate is justified

— R.A. Fisher, 1947³

Statistische Aussagen über das Verhalten von Mittelwerten oder Varianzen mehrerer Versuchsergebnisse gehen in der Regel von *unabhängig* voneinander durchgeführten Versuchen aus — es darf, grob gesprochen, keinen Zusammenhang zwischen den einzelnen Versuchen geben.

4.3.1 Randomisierung

Es gehört zu den großen Verdiensten des nun schon oft erwähnten Pioniers der statistischen Versuchsplanung, R.A. Fisher, über die Konsequenzen dieser Aussage bei seinen landwirtschaftlichen Experimenten gestolpert zu sein: die Erträge benachbarter Versuchsflächen (plots) waren natürlich ähnlicher als die Erträge weiter voneinander entfernter plots. Von der idealen Situation “zufälliger” Störungen durch unkontrollierbare Störgrößen, wie in Abbildung 4.2-a symbolisiert, waren seine Feldversuche weit entfernt — die von ihm gefundenen Störungen glichen weit eher den systematischen, in Abbildung 4.2-b visualisierten Größen (man beachte, dass diese Graphiken nicht nur als räumliche Abfolge, sondern auch als zeitliche Abfolge interpretiert werden können, etwa wenn eine steigende Umgebungstemperatur einen unkontrollierbaren Einfluss auf die Zielgrößen hat oder wenn die Abnutzung eines Werkzeugs im Produktionsprozess eine Rolle spielt).

Stellt man sich vor, dass man in einer derartigen Situation beispielsweise drei Faktoren mit jeweils 2 Stufen testen will, so wird sofort einsichtig, dass der Einfluss der (unkontrollierbaren) Störung im ersten Fall zwar die Variabilität der Messungen erhöht, aber aufgrund der zufälligen und symmetrischen Verteilung keine systematische Verzerrung verursacht, während diese Unabhängigkeit der Messungen im zweiten Fall gestört wird — je höher eine Messung, desto größer die Chance auf eine weitere hohe Messung.

Statistische Aussagen, die zu Unrecht auf einer Unabhängigkeitsannahme beruhen, werden aber falsch: Es kann sein, dass man “signifikante” Faktoreinflüsse erkennt, die in Wahrheit auf den Einfluss der Störgrößen zurückzuführen sind.

³ Auf jeden Fall erkennen wir die Randomisierung als notwendige Bedingung für die Gültigkeit unserer Schlussfolgerungen, und der moderne Versuchsleiter stellt sorgfältig sicher, dass dieses Postulat erfüllt ist.

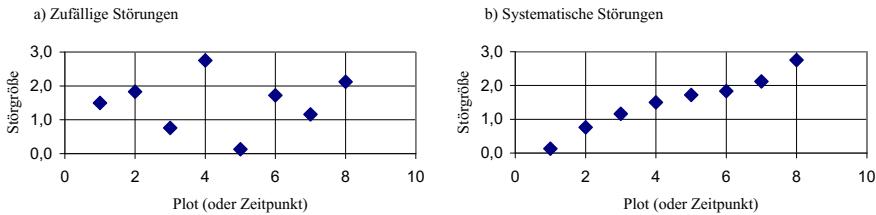


Abb. 4.2 Zwei Varianten derselben Störungen (Messfehler): Während die Störungen in Teil a) zufällig wirken, unterliegen sie in Teil b), rechte Seite, einer systematischen Ordnung. Es liegt auf der Hand, wieso Versuche durch systematische Störungen verfälscht werden, dass Fisher also zu Recht beunruhigt war.

Dies hängt natürlich vor allem vom Verhältnis der Störgrößen zu den zu messenden “echten” Effekten ab — das wir in der einen oder anderen Form im Laufe der gesamten statistischen Thematik betrachten werden.

Fishers großer Wurf bestand nun darin, die Versuchsreihenfolge von jeder Systematik zu lösen und stattdessen eine “zufällig ausgewürfelte” Reihenfolge zu nutzen. Um Missverständnissen vorzubeugen: Dies bedeutet *nicht*, dass es der Willkür des Durchführenden überlassen bleibt, wann er welchen Versuch durchführt — es bedeutet, dass die Versuchsreihenfolge (bzw. Zuordnung von plots zu Düngemitteln, Probanden zu Behandlungen etc.) *bei der Planung* der Experimente, etwa durch die Nutzung eines Zufallszahlengenerators, festgelegt wird.

Der Effekt der Randomisierung kann an obigem Beispiel sehr schön verdeutlicht werden, wenn man die den Graphiken zugrunde liegenden Daten nutzt.

Plot (oder Zeitpunkt)	Labor Zufällige Störung	Feld Systematische Störung
1	1.50	0.13
2	1.83	0.76
3	0.76	1.16
4	2.75	1.50
5	0.13	1.72
6	1.72	1.83
7	1.16	2.12
8	2.12	2.75

Tabelle 4.1 Beispielstörungen in Labor- und Feldversuchen

Es lässt sich leicht nachrechnen, wie der Beitrag dieser Störungen zu den ermittelten Faktoreffekten im Falle der systematischen Störung von der Reihenfolge der Versuchsdurchführung abhängt. Dazu betrachten wir sowohl einen nicht randomisierten als auch einen zufällig “ausgewürfelten”, das heißt vollständig randomisierten Vollfaktorplan, wie in Tabelle 4.2 dargestellt.

Vertauscht man die Versuchsreihenfolge wie in Tabelle 4.2 dargestellt, kann man errechnen, dass sich der Beitrag der systematischen Störung zu den Haupteffekten

Plot/Zeitpunkt	A	B	C	A	B	C	Plot/Zeitpunkt
1	—	—	—	+	+	—	7
2	—	—	+	—	—	+	2
3	—	+	—	—	+	—	3
4	—	+	+	+	—	—	5
5	+	—	—	+	—	+	6
6	+	—	+	+	+	+	8
7	+	+	—	—	+	+	4
8	+	+	+	—	—	—	1

Tabelle 4.2 Ein nicht randomisierter Versuchsplan neben einem vollständig randomisierten Plan

deutlich verkleinert (insbesondere für Faktor A, was angesichts der Tatsache, dass A in den ersten Versuchen stets auf —, in den letzten Versuchen stets auf + stand, mehr als verständlich ist).

“Effekt”-Beitrag	Nicht randomisiert	Randomisiert
A	1.22	-0.40
B	0.77	-0.37
C	0.43	0.22

Tabelle 4.3 Randomisierung verkleinert in der Regel den Beitrag des systematischen Fehlers zu den Faktoreffekten

Legt man stattdessen die “Laborsituation” einer echt zufälligen, stationären Störung zugrunde, stellt sich dieser Effekt natürlich deutlich weniger ausgeprägt dar.

Das bisher Gesagte kann kaum treffender zusammengefasst werden als mit den Worten von George Box anlässlich einer R. A. Fisher Memorial Lecture von 1988 ([3, S. 617]):

“... Fisher’s invention of statistical experimental design .. did much to move science out of laboratory into the real world. This was a major step in human progress.

The theory of experimental design that he developed, ..., solved the problem of how to conduct valid experiments in a world which is naturally non-stationary and non-homogeneous; a world, moreover, in which unseen ‘lurking variables’ are linked in unknown ways with the variables under study, thus inviting misinformation and confusion.”

George Box über R.A. Fisher: “Fishers Entwicklung des statistischen Designs von Versuchen .. war ein wesentlicher Schritt zum menschlichen Fortschritt, da die Wissenschaft aus dem Labor in die reale Welt geholt wurde. Die von ihm entwickelte Theorie des experimentellen Designs löste das Problem, wie man valide Experimente in einer natürlicherweise nicht stationären und homogenen Welt ausführen sollte — einer Welt, in der unbekannte, im ‘Hinterhalt’ liegende Variablen in unbekannter Weise mit den studierten Variablen zusammenspielen und damit Fehlinformation und Verwirrung einladen...”

Besser kann man die vorhergehenden Betrachtungen kaum zusammenfassen, die gezeigt haben, dass und warum die Existenz von Störgrößen in Experimenten

verstanden und berücksichtigt werden muss. Wir haben bereits gesehen, dass die Randomisierung einen wichtigen Beitrag dazu leistet.⁴ In der Folge werden wir sehen, dass auch die beiden anderen, von Fisher zu Recht so deutlich herausgehobenen Prinzipien der Wiederholung und Blockbildung dazu dienen, das “Signal” vom “Rauschen” zu unterscheiden und das “Rauschen” zu minimieren.

4.3.2 Wiederholung

Es ist natürlich insbesondere das Stichwort der Wiederholung von Versuchen, das beim Gedanken an die Verminderung des Versuchsrauschens in den Sinn kommt: Wir alle wissen, dass der Mittelwert einer erhöhten Anzahl von Messungen eine genauere Schätzung der gemessenen Größe liefert. “Mehr messen und dann mitteilen” reduziert die Versuchsstreuung, genauer gesagt die potenzielle Schwankung des ermittelten Mittelwertes.

Mathematisch formuliert: Ist σ^2 die Varianz der Originalgröße, so reduziert sich die Varianz des Mittelwertes aus n Messungen zu $\frac{\sigma^2}{n}$. Da Mittelwerte mehrerer, unabhängig voneinander durchgeführter Messungen dem Zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung folgen (wenn n genügend groß ist), kann man für genügend große Stichprobenumfänge feststellen, dass die zu erwartenden Mittel aus n Messungen einer Größe mit Mittelwert μ und Varianz σ^2 — wie auch immer diese Größe selbst verteilt sein mag — einer Normalverteilung mit Mittelwert μ und Varianz $\frac{\sigma^2}{n}$ folgen. Das bedeutet, dass die real dann zu findenden Mittelwerte aus n Messungen immer dichter am eigentlichen Wert μ liegen, da ca. 95% aller Mittelwerte im Bereich $\mu \pm 2\sqrt{\frac{\sigma^2}{n}}$ liegen werden, der mit zunehmendem n immer schmäler wird.

Ein Beispiel: Ein fairer Würfel zeigt alle Augenzahlen mit der gleichen Wahrscheinlichkeit $p_1 = p_2 = \dots = p_6 = \frac{1}{6}$. Der erwartete “mittlere gewürfelte Wert” μ lässt sich bestimmen als

$$\mu = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \dots + \frac{1}{6} \cdot 6 = 3.5,$$

die Varianz σ^2 , die “mittlere quadratische Abweichung vom Mittelwert”, damit als

$$\sigma^2 = \frac{1}{6} \cdot (1 - 3.5)^2 + \frac{1}{6} \cdot (2 - 3.5)^2 + \dots + \frac{1}{6} \cdot (6 - 3.5)^2 = 2.92.$$

⁴ In der industriellen Praxis führt die vollständige Randomisierung oft zu Problemen, da nicht alle Faktoren einfach zu verstehen sind — man denke etwa an Ofentemperaturen, die man aus Effizienzgründen lieber einfach von Versuch zu Versuch erhöhen möchte, statt zwischendurch immer wieder auf Abkühlphasen zu warten. Für diese Fälle gibt es spezielle Versuchspläne, sogenannte *split plot designs*, die dies berücksichtigen — sowohl in der Planung als auch der späteren Prozedur zur Auswertung dieser Versuche.

Würfelt man nun n mal und errechnet die mittlere Augenzahl aus diesen n Würfen, so ergibt sich eine Chance von 95%, Mittelwerte zwischen

$$\mu - 2\sqrt{\frac{\sigma^2}{n}} = 3.5 - 2 \cdot \sqrt{\frac{2.92}{n}}$$

und

$$\mu + 2\sqrt{\frac{\sigma^2}{n}} = 3.5 + 2 \cdot \sqrt{\frac{2.92}{n}}$$

zu finden, wie in Tabelle 4.4 gezeigt.

n	$\mu - 2\sqrt{\frac{\sigma^2}{n}}$	$\mu + 2\sqrt{\frac{\sigma^2}{n}}$
50	3.02	3.98
100	3.16	3.84
500	3.35	3.65
1000	3.39	3.61
10000	3.47	3.53

Tabelle 4.4 Grenzen, zwischen denen 95% der Mittelwerte von n gewürfelten Augenzahlen liegen (der “wahre” Wert ist 3.5)

Die zunehmende Zahl von Versuchen ermöglicht eine zunehmend genauere Bestimmung des mittleren Wertes, hier der mittleren Augenzahl von 3.5. Man sieht die Reduktion der Versuchsstreuung durch die Erhöhung des Stichprobenumfangs.

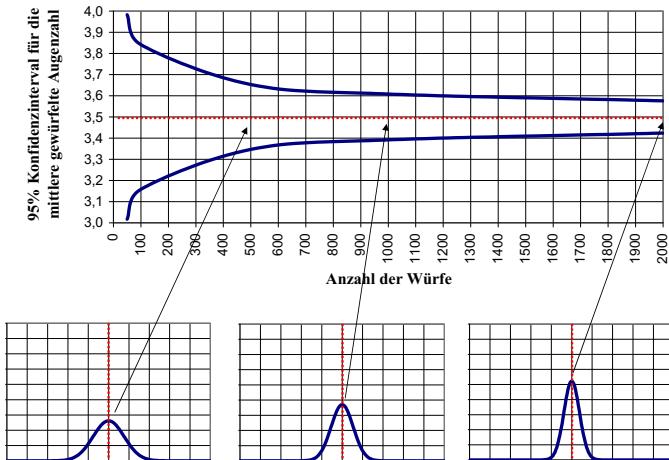


Abb. 4.3 Veranschaulichung von Tabelle 4.4. Es zeigt sich deutlich, wie sich das Konfidenzintervall mit zunehmender Versuchszahl verkleinert — allerdings zeigt sich auch, dass der Effekt bei kleineren Stichprobenumfängen bzw. Versuchszahlen am drastischsten ist.

Ein weiterer, mit der Wiederholung von Versuchen verbundener Aspekt der Versuchsstreuung liegt vermutlich weniger nahe — obwohl er im Grunde vielleicht sogar noch bedeutender ist: *Erst die Wiederholung von Versuchen ermöglicht, die Versuchsstreuung abzuschätzen!*

Warum dies von so fundamentaler Bedeutung für die Auswertbarkeit von Versuchen ist, zeigt die folgende beispielhafte Frage: Angenommen, wir wollten die Wirkung eines Faktors durch jeweils 5 Versuche auf zwei verschiedenen Stufen bestimmen und hätten bei der ersten Stufe ein mittleres Ergebnis von 17, bei der zweiten Stufe von 20 gefunden. Ist der Unterschied “signifikant”, oder könnte es sich um einen zufälligen, durch die Versuchsstreuung bedingten Effekt handeln? Die Antwort auf diese Frage kann nicht ohne zusätzliches Wissen gegeben werden. Sie hängt nämlich von unserer Einschätzung ab, ob der gemessene Unterschied von 3 wirklich messbar oder ein Produkt des Messrauschen ist.

Fall (1). Die Einzelmessungen bei Faktorstufe eins seien 16.8, 16.9, 17.0, 17.1 und 17.2 bzw. 19.7, 19.8, 20.0, 20.2 und 20.3 bei der zweiten Stufe.

In diesem Fall sind wir sicherlich auch ohne weitere statistische Analyse davon überzeugt, einen realen Effekt gefunden zu haben, da die Versuchsergebnisse trotz der vorhandenen Streuung weit genug “auseinander” liegen. Dies sieht im zweiten Fall anders aus:

Fall (2). Die Einzelmessungen bei der ersten Stufe seien 15, 16, 17, 18 und 19, diejenigen bei der zweiten Stufe seien 18, 19, 20, 21 und 22.

Während im zweiten Fall ein Effekt gleicher Größe gemessen wurde, überlappen sich die Messungen zu einem gewissen Grad. Einzelmessungen von 18 und 19 sind bei beiden Faktorstufen aufgetreten. Sollen wir dem Effekt (Unterschied zwischen den jeweils mittleren Messwerten) glauben?

Wir werden Fragen dieser Art in Kapitel 4.5 beantworten, in dem wir eine Methode darstellen, die gemessenen Unterschiede in intelligenter Art und Weise in Relation zur Versuchsstreuung zu setzen. Bevor wir aber zunächst den Abschnitt über die Wiederholung von Versuchen beenden, sei eine wichtige Warnung formuliert:

Die “Wiederholung von Versuchen” muss, um Aussagekraft zu gewinnen, eine “echte” Wiederholung des Versuchs — nicht lediglich der Messung — sein. Die Faktorstufen müssen jeweils neu eingestellt bzw. justiert werden; dies ist mehr als eine bloße Wiederholung der Messungen bei unveränderten Faktoreinstellungen.

Warum ist das so?

Die Antwort auf diese Frage liegt nach dem oben Gesagten (hoffentlich) auf der Hand: Während eine bloße Wiederholung der Messung lediglich Aufschluss über die Streuung im Meßsystem gibt, ermöglicht eine “echte” Versuchswiederholung auch die Abschätzung desjenigen Teils der Versuchsstreuung, die sich aus der Einstellbarkeit der Faktoren (Reproduzierbarkeit) ergibt.

Im Rahmen der “echten”, über die bloße Wiederholung von Messungen hinausgehenden Wiederholung von Versuchen gibt es also drei zentrale Gedanken:

1. Um die gemessenen Unterschiede, etwa zwischen zwei Faktoreinstellungen, zu bewerten, sollten sie in Relation zur Streuung der Versuche gesetzt werden.
2. Die Versuchsstreuung sollte reduziert werden.
3. Die Versuche sollten im Ganzen wiederholt werden, nicht nur als Messung.

Während der erste dieser Punkte in Kapitel 4.5 wieder aufgegriffen wird, erlaubt uns der zweite Punkt eine perfekte Überleitung zum Thema der Blockbildung, die einen weiteren wichtigen Baustein zur Reduktion unerwünschter Streuung bildet.

4.3.3 Blockbildung

Auch die Blockbildung unterstützt die Reduktion systematischer Fehler — in diesem Fall von Fehlern, die durch Faktoren bedingt werden, die zwar bekannt, aber nicht kontrollierbar sind, da sie nicht zum betrachteten System gehören. Die Grundidee der Blockbildung ist, den Einfluss dieser Fehler quantifizierbar zu machen und somit von der Berechnung der Faktoreffekte zu trennen.

Was ist damit gemeint?

Man kann sich leicht Szenarien vorstellen, bei denen systematische Verzerrungen durch außerhalb des Systems liegende Faktoren auftreten. So kann es sein, dass

- eine Charge eines Materials nicht ausreicht, um alle Prototypen zu bauen, und obwohl das Experiment nicht auf den Einfluss des Materials abzielt, Verzerrungen durch Unterschiede zwischen verschiedenen Chargen durchaus denkbar sind
- nicht alle Messungen innerhalb einer Woche oder einer Produktionsstätte durchgeführt werden können, obwohl mit unterschiedlichen Ergebnissen für verschiedene Zeiträume oder Orte zu rechnen ist.

Störfaktoren dieser Art, an denen kein echtes Interesse besteht, deren Effekte aber zur Vermeidung systematischer Fehler von den “echten” (interessierenden) Effekten getrennt werden sollen, können sogenannte “Blöcke” definieren — Gruppierungen des Versuchsplans in Bereiche mit relativ homogenen Randbedingungen. Zur Illustration sei erneut ein Beispiel aus der Landwirtschaft angeführt:

Möchte eine Bäuerin den Ertrag von Weizen auf ihren Feldern optimieren, so könnte sie beispielsweise einem Vollfaktorplan mit zwei Sorten von Weizen und zwei verschiedenen Düngemitteln folgen, um aus den vier möglichen Kombinationen diejenige mit dem höchsten Ertrag zu ermitteln. Sie könnte dazu weiterhin ihr Versuchsfeld in vier gleichgroße Teile einteilen und — da sie das Prinzip der Randomisierung verstanden hat — zufällig entscheiden, welche Kombination von Weizen und Dünger sie auf welcher Teilparzelle anpflanzen und düngen möchte. Mit dieser Strategie würde sie die (potenziellen) Effekte von Störfaktoren wie Bodenbeschaffenheit oder Sonneneinstrahlung “zufällig verteilen”, die Effekte aber zugleich zum Preis eines erhöhten Messrauschen verstecken — während eigentlich eine andere

Vorgehensweise auf der Hand liegt: die Unterteilung ihres großen Versuchsfeldes in eine Anzahl kleinerer, bezüglich Bodenbeschaffenheit und Sonneneinstrahlung möglichst homogener Teile — Blöcke — und die Anwendung der Randomisierung innerhalb dieser Blöcke. Damit kann sie noch immer erreichen, dass keine Weizensorte und kein Dünger bevorzugt wird, hat aber zugleich die Möglichkeit, den potenziellen (!) Einfluss der Störgrößen (Blöcke) zu analysieren und so zusätzliche Informationen zu gewinnen. Ein möglicher, innerhalb dreier Blöcke randomisierter Versuchsplan könnte dann aussehen wie in Tabelle 4.5 dargestellt: Alle vier Faktorkombinationen erscheinen in jedem Block; ihre Anordnung innerhalb der Blöcke ist rein zufällig.

BLOCK	Dünger	Weizen
1	+	+
1	-	-
1	+	-
1	-	+
2	+	+
2	+	-
2	-	+
2	-	-
3	+	+
3	-	+
3	-	-
3	+	-

Tabelle 4.5 Ein vollständig randomisierter Versuchsplan mit Blöcken erlaubt, unkontrollierbare Einflüsse berechenbar zu machen und damit in der Auswertung zu berücksichtigen.

Der dargestellte Versuchsplan beinhaltet drei Wiederholungen für jede Faktorkombination, woraus sich die insgesamt 12 Versuche ergeben. In dieser Situation gibt es keine Vermengung von Effekten; sowohl die Haupteffekte von Weizensorte und Düngemittel als auch deren Wechselwirkungen sind eindeutig identifizierbar und zusätzlich vom Effekt des Faktors “Block” zu trennen.

Auch wenn die Anzahl durchführbarer Versuche eine Rolle spielt — so wie in den meisten industriellen Anwendungen — kann man noch immer mit Blöcken operieren. Nichts ist allerdings umsonst — der Preis ist gegebenenfalls eine Vermengung (confounding) der Blockeffekte mit den Wechselwirkungen. Sollte sich allerdings im Rahmen der Varianzanalyse (siehe Kapitel 4.5) herausstellen, dass der Faktor Block in Wahrheit keinen nachweisbaren Einfluss hat, kann er ohne Verlust einfach aus der weiteren Analyse ausgeschlossen werden. Damit wird deutlich, wie die Gestaltung des Versuchsplans dazu beitragen kann, Fehler durch unkontrollierbare Störgrößen zu erkennen und zu minimieren. Wir werden in Kapitel 4.5 auf die Auswertung — die Quantifizierung der Einflüsse der Faktoren nach Berücksichtigung der Blockeffekte — zurückkommen. Bevor wir jedoch die Varianzanalyse (Analysis of Variance, ANOVA) als Mittel zur Identifikation “signifikanter” Effekte darstellen, wollen wir zunächst zum besseren Verständnis die Grundgedanken aller statistischen Signifikanztests erörtern.

4.4 Wieso “Null”hypothese? Der Grundgedanke aller statistischen Tests

In relation to any experiment we may speak of this hypothesis as the “null hypothesis,” and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.

— R.A. Fisher, 1935⁵

4.4.1 Ein Beispiel

Am einfachsten nähert man sich den Grundgedanken des statistischen Testens mit einem griffigen Beispiel: Wir sitzen mit einem Kollegen in einer Gaststätte und werfen mit Münzen aus, wer das nächste Bier bezahlen muss. Wer in einer abwechselnden Folge von Münzwürfen mit einer aus seiner Börse gegriffenen Münze als nächstes einen *Kopf* wirft, ist an der Reihe, das Bier zu bezahlen.

Während wir uns sicher sind, dass unsere Münze fair ist (wir haben sie zufällig aus unserer Geldbörse gefischt), ist uns unklar, ob dies auch für die Münze des Kollegen gilt. Beide Münzen sollten fair sein, also beispielsweise mit der gleichen Chance das Merkmal *Kopf* zeigen — in der Sprache der statistischen Versuchsplanung: der Faktor *Münze* sollte keinen Einfluss auf das Ergebnis (die Anzahl gewürfelter Köpfe) haben...

Etwas formaler kann man diese Frage wie folgt formulieren: Es sei $p_1 = 0.5$ die Wahrscheinlichkeit, mit der unserer Münze Kopf zeigt und p_2 die uns unbekannte Wahrscheinlichkeit für die andere Münze. Wir hoffen nun, dass der Faktor Münze keinen Einfluss hat ($p_2 = 0.5$), während sich ein für uns relevanter Einfluss in der Beziehung $p_2 > 0.5$ äußert. Wann aber werfen wir dem Kollegen vor, dass er schummelt?

1. Wenn er in 10 Würfen 7 Köpfe erzielt hat?
2. Wenn er in 50 Würfen 35 Köpfe erzielt hat?
3. Wenn er in 100 Würfen 70 Köpfe erzielt hat?

Während es qualitativ auf der Hand liegt, dass die konstante Erfolgsrate von 70% bei steigender Versuchszahl zunehmend nervös machen sollte (wir können mit zunehmender Sicherheit davon ausgehen, dass der Effekt des Faktors Münze signifikant ist), fehlt uns eine konkrete, begründ- und nachvollziehbare Entscheidungsregel.

⁵ In Bezug auf alle Experimente können wir bei dieser Hypothese von “Nullhypothese” sprechen, und es sollte angemerkt werden, dass die Nullhypothese niemals bewiesen wird, sondern möglicherweise im Rahmen eines Experiments widerlegt. Man könnte sagen, dass Experimente nur existieren, um den Fakten eine Chance zu geben, die Nullhypothese zu widerlegen.

Diese wird durch das Hilfsmittel der statistischen Tests geliefert.

Alle statistischen Tests funktionieren nach einem identischen Strickmuster: Man unterstellt die Gültigkeit einer Hypothese (also beispielsweise: $p_2 = 0.5$ bzw.: der Faktor Münze hat keinen Effekt) und bestimmt die Abweichung von der auf der Gültigkeit der Hypothese basierenden Erwartung in einer geeigneten Form. Ist die Münze fair ($p_2 = 0.5$), so erwarten wir

1. bei 10 Würfen 5 Köpfe
 - 7 geworfene Köpfe bedeuten eine Abweichung von 2 Köpfen
2. bei 50 Würfen 25 Köpfe
 - 35 geworfene Köpfe bedeuten eine Abweichung von 10 Köpfen
3. bei 100 Würfen 50 Köpfe
 - 70 geworfene Köpfe bedeuten eine Abweichung von 20 Köpfen

Ist die Abweichung zu “groß”, gehen wir davon aus dass die Hypothese widerlegt ist (während 6 Köpfe in 10 Würfen wohl mit der Hypothese vereinbar sind, trifft dies für 10 Köpfe in 10 Würfen wohl eher nicht zu). Dabei ist zu beachten, dass bei der Wahl der Akzeptanzgrenze zwei entgegengesetzte Ziele verfolgt werden:

- Hält man die Hypothese (der fairen Münze) bereits bei kleinen Abweichungen für widerlegt, bezahlt man auf keinen Fall zu viel Bier — der Mitspieler könnte jedoch zu Recht verärgert sein, wenn man ihn zu Unrecht des Schummelns bezichtigt
- Hält man dagegen die Hypothese erst bei großen Abweichungen für widerlegt, verkleinert man zwar das Risiko für Streit — zahlt aber im Zweifelsfall im wahrssten Wortsinn die Zeche...

4.4.2 α - und β -Risiko

Die selben Zusammenhänge kann man auch in einem seriöseren Anwendungsfall darstellen: Testet man den Effekt eines neuen Düngers auf den Ernteertrag (Hypothese: Der Dünger hat *keinen* Effekt), so besteht der trade-off darin zu entscheiden, ab wann man die Hypothese verwirft, also das neue *Produkt* lieber...

- ... zu Unrecht akzeptiert (sogenannter Fehler 1. Art), da man die Nullhypothese zu Unrecht zurückweist
- ... zu Unrecht zurückweist (sogenannter Fehler 2. Art), da man die Nullhypothese zu Unrecht akzeptiert

Wie man Tabelle 4.6 sieht, werden die beiden möglichen Fehlentscheidungen — und das Risiko, sie zu treffen — unterschiedlich bezeichnet und behandelt. In langen Jahren der Anwendung statistischer Tests hat sich die Vorgehensweise bewährt, das α -Risiko zu kontrollieren (damit wird beispielsweise sichergestellt, dass die Wahrscheinlichkeit, einem nicht signifikanten Effekt aufzusitzen, maximal 5% oder maximal 10% beträgt) und über eine Betrachtung des β -Risikos die benötigte Anzahl von Versuchen zu bestimmen.

	Hypothese akzeptiert (Entscheidung: $p_2 = 0.5$)	Hypothese verworfen (Entscheidung: $p_2 > 0.5$)
Nullhypothese: Effekt nicht vorhanden ($p_2 = 0.5$)	OK	<ul style="list-style-type: none"> • Kollege verärgert • falschen Dünger gekauft • Fehler 1. Art • Hypothese zu Unrecht verworfen • nicht signifikanter Effekt als wichtig angesehen • α- Risiko
Alternativhypothese: Effekt vorhanden ($p_2 > 0.5$)	<ul style="list-style-type: none"> • Autor zahlt Zeche • guten Dünger zurückgewiesen • Fehler 2. Art • Hypothese zu Unrecht akzeptiert • Signifikanter Effekt nicht erkannt • β-Risiko 	OK

Tabelle 4.6 Die vier möglichen Situationen: Richtige und falsche Entscheidungen, α - und β -Risiko. Wenn manche Autoren den Fehler 1. Art auch als "Produzentenrisiko" und den Fehler 2. Art als "Konsumentenrisiko" bezeichnen, so liegt dies daran, dass hier gedanklich von der *Annahmekontrolle einer WarenSendung* ausgegangen wird. Im Gegensatz zur Standardsituation in der statistischen Versuchsplanung wird hier also nicht auf **Effekte**, sondern auf **Defekte** fokussiert. In diesem Kontext wird das ungerechtfertigte Zurückweisen der Sendung (Fehler 1. Art) zum Risiko des Produzenten, ein Übersehen von Fehlern (Fehler 2. Art) zum Risiko des Konsumenten.

Im Rahmen des bisher Gesagten ist ein Punkt von zentraler Bedeutung:

Die dargestellte Konstruktion besagt, dass es "nur" die Möglichkeit gibt, eine Hypothese zu verwerfen (wenn die Beobachtung zu stark von den möglichen Konsequenzen der Hypothese abweicht), aber niemals die Möglichkeit, die Gültigkeit der Hypothese zu beweisen. Aus diesem Grund hat es sich eingebürgert, von einer bereits von Fisher so genannten "Nullhypothese" auszugehen (es gibt keinen Effekt) und das "zu zeigende" Ergebnis als Alternative zur "Null" anzunehmen. Damit wir die Evidenz, mit der man die Nullhypothese verwirft, zur Evidenz, mit der man sein Ergebnis gezeigt hat!

Eine Beobachtung. Bei Wahl eines vorgegebenen α s lässt sich zu *jeder* Versuchszahl eine Akzeptanzschwelle für die Abweichung angeben, ab der die Nullhypothese verworfen wird. So weiß man aus der Wahrscheinlichkeitsrechnung, dass – ist die Münze fair, d.h. ist die Erfolgswahrscheinlichkeit $p = 0.5$ – die Anzahl gewürfelter *Köpfe* einer sogenannten *Binomialverteilung* folgt.

Man kann deshalb für jeden Versuchsumfang n feststellen, wie groß die Wahrscheinlichkeit ist, in 70% (oder mehr) aller Würfe zu gewinnen, wenn man von der Gültigkeit der Hypothese $p = 0.5$ ausgeht.

Die Betrachtung der Histogramme in Abbildung 4.4 bestätigt unsere “qualitative” Vermutung, dass der erste Versuch ($n=10$) noch keinen Anlass bietet, mit dem Kollegen zu streiten (da er mit einer fairen Münze immerhin eine Chance von 17.2% hatte, 7 oder mehr Köpfe in 10 Versuchen zu werfen), während das Ergebnis des dritten Versuchs ($n=100$) ein unzweifelhaftes Indiz dafür liefert, dass man betrogen wurde, da die Chance für 70 oder mehr Köpfe bei einer fairen Münze fast Null ist.

ACHTUNG — TRUGSCHLUSSGEFAHR!

An dieser Stelle sei die Warnung vor einem typischen Trugschluss noch einmal wiederholt, der lautet, dass die Hypothese (einer fairen Münze) im ersten Fall bewiesen, im zweiten widerlegt wurde. *Dies ist falsch — nur der zweite Halbsatz gilt!* Da es nur eine verschwindend geringe Wahrscheinlichkeit von 0.003% gibt, mit einer fairen Münze in 100 Versuchen 70 oder mehr Köpfe zu werfen, kann hier der logisch korrekte Schluss gezogen werden, dass die Münze unfair ist — das Risiko, hierbei falsch zu liegen, lässt sich mit eben diesen $p = 0.003\%$ genau quantifizieren, dem Wert des α -Risikos. Hingegen ist im ersten Fall die Hypothese nicht *widerlegbar* — daraus folgt jedoch keinesfalls ihr Beweis. Man kann nicht logisch zwingend folgern, dass die Münze fair ist, nur weil sie ein mit Fairness vereinbares Ergebnis gezeigt hat!

“Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis”... — man könnte sagen, dass Experimente nur existieren, um den Fakten eine Chance zu geben, die Nullhypothese zu verwerfen. ([8, S. 18])

Die oben angestellten Betrachtungen zeigen im übrigen auch, dass man bereits im Vorfeld, also beispielsweise vor der Durchführung der 10, 50 oder 100 Münzwurfversuche, Akzeptanzschwellen für die Ablehnung der Hypothese berechnen kann, wenn man das α -Risiko vorgibt. So ergibt sich für $\alpha = 10\%$ ein Ablehnungsbereich von

- 8 oder mehr Köpfen im Falle von 10 Versuchen (da diese bei einer fairen Münze nur mit einer Wahrscheinlichkeit von ca. 5.5% eintreten)
- 31 oder mehr Köpfen bei 50 Versuchen (Eintrittswahrscheinlichkeit 5.9%)
- 57 oder mehr Köpfen bei 100 Versuchen (Eintrittswahrscheinlichkeit 9.7%)

Reduziert man das α -Risiko, etwa auf $\alpha = 1\%$ (da einem der Kollege lieb und teuer ist und man keinen unnötigen Streit riskieren möchte), so ergibt sich für

- 10 Versuche, dass man die Hypothese nicht schon ab 8 Köpfen ablehnen kann, sondern erst ab 9 geworfenen Köpfen,
- 50 Versuche, dass man die Hypothese nicht schon ab 31 Köpfen ablehnen kann, sondern erst ab 34 geworfenen Köpfen, und bei
- 100 Versuche, dass man die Hypothese nicht schon ab 57 Köpfen ablehnen kann, sondern erst ab 62 geworfenen Köpfen.

Die mangelnde Konfliktbereitschaft hat ihren Preis — das Risiko, die Zeche selbst zu bezahlen steigt...

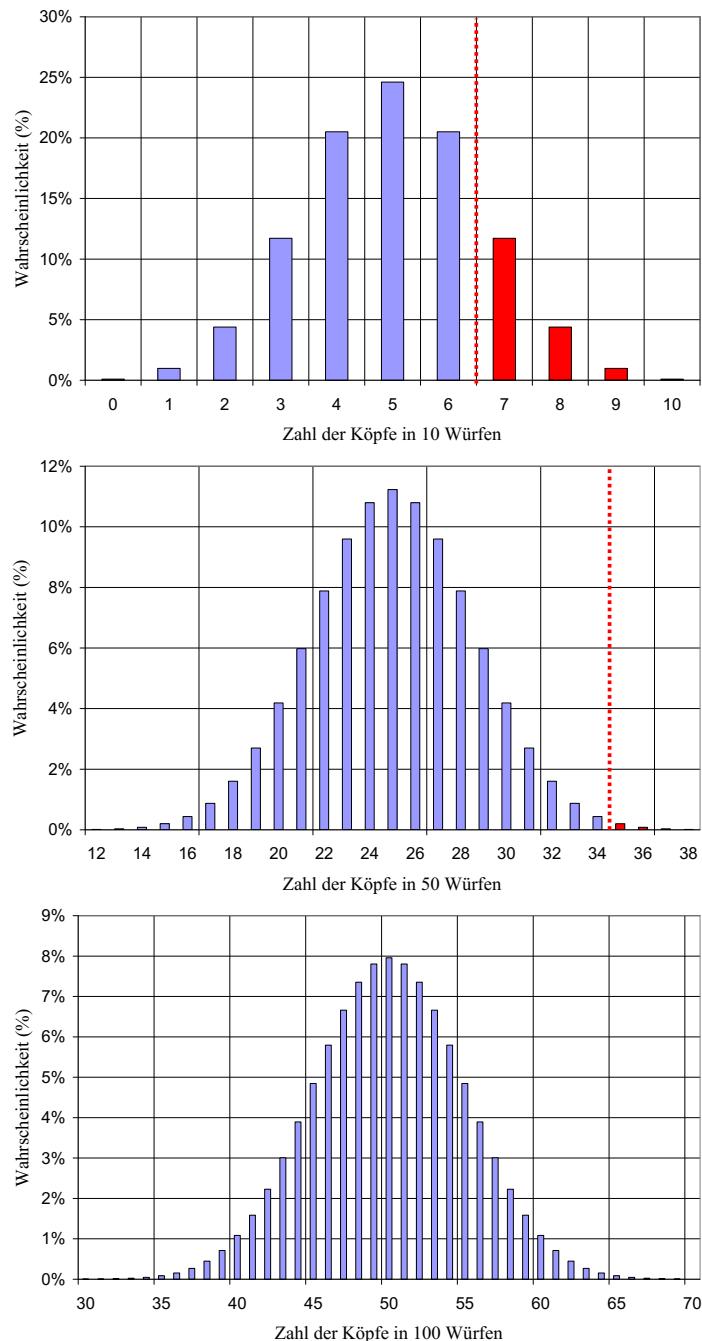


Abb. 4.4 Die Chance, 7 oder mehr Köpfe in 10 Versuchen zu erzielen, beträgt immerhin 17.2%. Die Chance, 35 oder mehr Köpfe in 50 Versuchen zu erzielen, beträgt nur noch 0.3%, und die Chance auf 70 oder mehr Köpfe in 100 Versuchen ist praktisch Null.

4.4.3 Versuchsumfang

Wie viele Münzwürfe (bzw. Versuche) sollte man also durchführen, um beiden Risiken — α — und β —Risiko — gerecht zu werden?

Es liegt nach dem oben Gesagten auf der Hand, dass die Betrachtung des α —Risikos allein nicht zielführend ist, da man zu einem gegebenen Wert und einer gegebenen Anzahl von Versuchen stets eine “Akzeptanzschwelle” finden kann, die die Kontrolle des α —Risikos ermöglicht. Wir müssen auch das β —Risiko betrachten, die Wahrscheinlichkeit, die Hypothese zu akzeptieren (obwohl sie unter Umständen auch falsch ist). Aus praktischen Gründen betrachtet man in diesem Zusammenhang allerdings meist nicht β , sondern die Größe $1 - \beta$: die Wahrscheinlichkeit, die Hypothese zu verwerfen (bzw. die Alternative anzunehmen — eine unfaire Münze, einen besseren Dünger oder einen signifikanten Faktoreffekt). $1 - \beta$, die sogenannte *Power* eines Tests, beschreibt mit anderen Worten die Fähigkeit, signifikante Unterschiede (zwischen Münzen, Düngemitteln oder Faktoreinstellungen) als solche zu erkennen.

Die Power eines Tests. Wie nun bereits mehrfach wiederholt, ist das β —Risiko durch die Wahrscheinlichkeit beschrieben, für einen gegebenen Wert $p_2 > 0.5$ die Hypothese *nicht* abzulehnen, obwohl sie in der Tat falsch ist — das heißt (für obiges Beispiel mit $\alpha = 10\%$), durch die Wahrscheinlichkeit, in maximal 7 von 10, 30 von 50 oder 56 von 100 Münzwürfen mit einem *Kopf* zu gewinnen. Wie bereits in den vorhergehenden Abschnitten lassen sich diese Wahrscheinlichkeiten auch hier mit Hilfe der Binomialverteilung bestimmen.

So ergibt sich für verschiedene Werte von p_2 und Versuchszahlen n die in Tabelle 4.7 dargestellten Werte für das β —Risiko (die Wahrscheinlichkeit, eine unfaire Münze nicht als solche zu erkennen). Anders herum formuliert, zeigt eine Aufstellung der Werte für $1 - \beta$ die vom wahren Wert p_2 abhängige Power der Tests, vergleiche Tabelle 4.8.

p_2	$n = 10$	$n = 50$	$n = 100$	$n = 500$
0.55	0.90	0.80	0.62	0.17
0.60	0.83	0.55	0.24	0.00
0.75	0.47	0.01	0.00	0.00

Tabelle 4.7 Verschiedene, für $\alpha = 10\%$ ermittelte Werte für das β —Risiko (die Gefahr, Abweichungen von der Hypothese “ $p = 0.5$ ” nicht zu erkennen)...

Man erkennt unschwer, dass kleinere Abweichungen (beispielsweise $p_2 = 0.55$ statt 0.5) eher unerkannt bleiben als größere, und dass die Versuchszahl bei der Gefahr, Abweichungen zu übersehen, eine große Rolle spielt.

Ebenso deutlich wird nun, wie das Zusammenspiel zwischen Versuchszahl, Genauigkeit (Auflösung) und Risikobereitschaft ist.

p_2	$n = 10$	$n = 50$	$n = 100$	$n = 500$
0.55	0.10	0.20	0.38	0.83
0.60	0.17	0.45	0.76	1.00
0.75	0.53	0.99	1.00	1.00

Tabelle 4.8 ...und die entsprechende *Power* des durch die Forderung nach $\alpha \leq 10\%$ definierten Tests. Liegt die wahre Erfolgsschance bei $p_2 = 55\%$, so wird diese Abweichung von der Hypothese einer fairen Münze beim Test mit 10 Würfen lediglich mit einer Chance von 10% erkannt, beim Test mit 500 Würfen aber mit einer Wahrscheinlichkeit von 83%

Zu einem kontrollierbaren α -Risiko (eine wahre Hypothese zu Unrecht zu verwerfen, bzw. einen nicht signifikanten Effekt irrtümlich für real zu halten) lässt sich ein Test definieren, beispielsweise

1. "Werfe die Münze 100 mal und verwerfe die Hypothese einer fairen Münze, wenn mindestens 57 mal Kopf geworfen wurde" ($\alpha = 10\%$)
2. "Werfe die Münze 100 mal und verwerfe die Hypothese einer fairen Münze, wenn mindestens 62 mal Kopf geworfen wurde" ($\alpha = 1\%$)

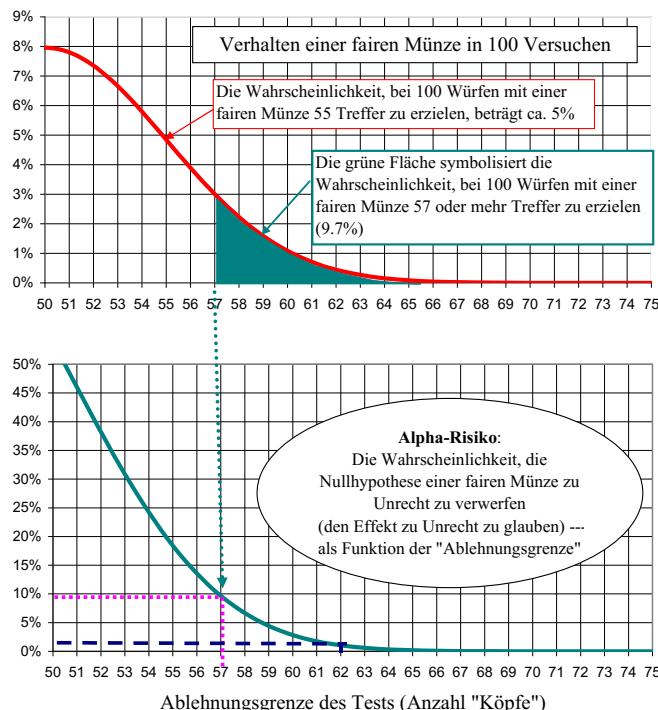


Abb. 4.5 Die Wahrscheinlichkeit, eine faire Münze zu Unrecht als unfair zu betrachten als Funktion der "Ablehnungsgrenze".

In Abhängigkeit vom wahren (aber bei der Versuchsplanung natürlich unbekannten) Wert p_2 (der Wahrscheinlichkeit, mit der Münze des Kollegen Kopf zu werfen) lässt sich das β -Risiko für den gegebenen Stichprobenumfang (z.B. $n = 100$) ermitteln — die Wahrscheinlichkeit, die falsche Hypothese *nicht* als solche zu erkennen.

1. Im ersten Fall handelt es sich um die aus p_2 resultierende Wahrscheinlichkeit, maximal 56 Köpfe zu werfen,
2. ... im zweiten Fall um die Wahrscheinlichkeit, maximal 61 Köpfe zu erzielen.

Diese Wahrscheinlichkeiten finden sich (für $n = 100$) in Tabelle 4.9

p_2	$\alpha = 10\%$	$\alpha = 1\%$
0.50	0.90	0.99
0.55	0.62	0.90
0.60	0.24	0.62
0.65	0.04	0.23
0.70	0.00	0.03
0.75	0.00	0.00

Tabelle 4.9 β -Risiken für verschiedene α s. Die Chance, eine Abweichung von der Hypothese einer fairen Münze nicht zu erkennen, liegt bei 24 %, wenn die wahre Erfolgschance der Münze bei $p_2 = 0.6$ liegt und der Test ein α -Risiko von 10% in Kauf nimmt. Testet man “vorsichtiger” und akzeptiert lediglich ein α -Risiko von 1%, erhöht sich diese Chance auf 62%.

Aus diesen Werten für β lässt sich die Power der Tests, $1 - \beta$, wieder einfach ermitteln (siehe Tabelle 4.10). Die Daten in Tabelle 4.10 zeigen erneut das Zusammenspiel zwischen α - und β -Risiko (bzw. Power): Bei identischer Versuchszahl lassen sich Tests mit geringerem α -Risiko finden (“spätere” Ablehnung der Nullhypothese) — zum Preis eines erhöhten β -Risikos (einer geringeren Power).

p_2	$\alpha = 10\%$	$\alpha = 1\%$
0.50	0.10	0.01
0.55	0.38	0.10
0.60	0.76	0.38
0.65	0.96	0.77
0.70	1.00	0.97
0.75	1.00	1.00

Tabelle 4.10 Power zweier Tests mit je 100 Versuchen für verschiedene Werte von α . Man beachte, dass die genannten Werte die Wahrscheinlichkeit für die Ablehnung der Hypothese einer fairen Münze darstellen — es ist also kein Zufall, dass die Werte in der ersten Zeile, $p_2 = 0.5$, gerade das α -Risiko wiedergeben, denn mit dieser Wahrscheinlichkeit wird die Hypothese abgelehnt, wenn sie eigentlich korrekt ist...

Das Geheimnis zur Bestimmung des Versuchsumfangs besteht also darin, ein der Problemstellung angemessenes α -Risiko und eine passende “Schmerzgrenze”, ab der Unterschiede erkannt werden sollen, samt zugehöriger Power (Wahrscheinlichkeit, mit der der Unterschied an der Schmerzgrenze erkannt wird) zu bestimmen.

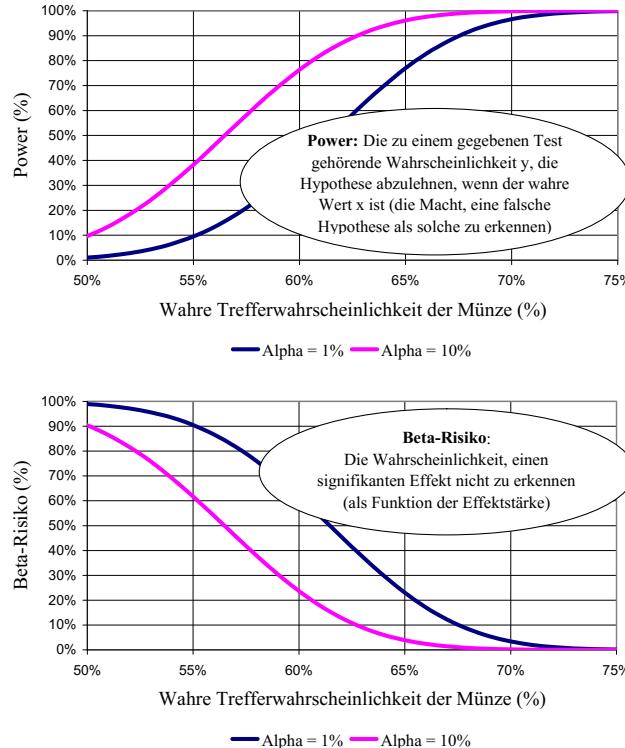


Abb. 4.6 Zusammenspiel zwischen *Power*, α - und β -Risiko des Tests für das Münzwurfbeispiel: Reduziert man das α -Risiko von 10% auf 1%, verringert sich die Power deutlich, vor allem, wenn die Münze des Kollegen eine Erfolgschance zwischen 55% und 65% hat. Bei stärkeren Effekten (z.B. bei Münzen mit einer Erfolgschance von mehr als 70%) ist der Verlust an *Power* minimal.

An dieser Stelle müssen wir leider darauf hinweisen, dass sich die nun naheliegende und in der Praxis immer wieder gestellte Frage nach "richtigen" Werten für α und β nicht allgemeingültig beantworten lässt. Es wurde bereits mehrfach erwähnt, dass das gewählte Risiko vom Sicherheitsbedürfnis abhängt. Drei Faktoren, die die Antwort beeinflussen, sollen jedoch nicht unerwähnt bleiben:

- 1. Relevanz der Antwort.** Soll gezeigt werden, dass ein neues Design der Bremsanlage eines Sportwagens die Bremsleistung im Vergleich zum alten Design verbessert, setzt man sicherlich eine hohe Sicherheits-Schwelle, bevor man das neue Design übernimmt, da man die Gefahr eines Irrtums (das neue System verbessert die Bremsleistung *nicht*) sehr streng kontrollieren sollte (z.B. $\alpha = 0.001\%$). Handelt es sich dagegen um die Frage, ob Kunden dieses Fahrzeugtyps lieber runde oder rechteckige Schalter für die Bedienung der Klimaanlage hätten, kann man wohl eher nachlässiger sein (z.B. $\alpha = 10\%$).

2. **Stärke des Effekts.** Man denke an Abbildung 4.6: Was “bekommt” man für die “eingesetzten” 9% α -Risiko, wenn man von 1% auf 10% akzeptables α -Risiko erhöht? Der Zugewinn an *Power* pro Prozentpunkt α -Risiko zeigt, dass die 9% in diesem Beispiel “effizient” eingesetzt sind, solange die wahre Erfolgswahrscheinlichkeit zwischen 51% und 66% liegt; danach ist der Effekt so stark, dass er auch mit dem weniger aggressiven Test ($\alpha = 1\%$) erkannt wird.
3. **Versuchskosten.** Leider ist es in der Praxis meist so, dass die Anzahl durchzuführender Versuche von vornherein beschränkt ist — während der Statistiker aus α und β den benötigten Versuchsumfang bestimmen möchte, ist in der Realität fast immer die Frage zu beantworten, *ob* mit dem bezahlbaren Versuchsumfang eine angemessene Power erreicht werden kann.

Fassen wir den bisherigen Stand der Dinge zusammen:

Grundprinzipien des statistischen Testens

- Um Unterschiede, zum Beispiel zwischen Faktoreinstellungen, zu bewerten, sollten sie in Relation zur Streuung der jeweiligen Versuche gesetzt werden.
- Ein Effekt ist “signifikant”, wenn er *nicht* mit der “Nullhypothese” einer rein zufälligen Abweichung vereinbar ist.
- Um dies zu testen, wird eine passende Kenngröße definiert, die den Grad der Abweichung von der Nullhypothese quantifiziert (z.B. die Anzahl gewürfelter Köpfe in 10 Versuchen).
- Die Wahrscheinlichkeit dafür, dass diese Kenngröße — wenn die Hypothese gilt — bestimmte Werte annimmt, muss bestimmbar sein (z.B. als Binomialverteilung, die uns erlaubt, die Chance für 7 oder mehr Köpfe in 10 Würfen zu bestimmen, wenn die Münze fair ist).
- Damit lässt sich zu jedem gefundenen Versuchsergebnis (= Wert der definierten Kenngröße) ein Wahrscheinlichkeitswert (“*p value*”) p angeben, der die Wahrscheinlichkeit dafür beschreibt, die gemessene oder eine größere Abweichung zu finden, wenn man die Gültigkeit der Hypothese unterstellt. Der *p value* beschreibt also das Risiko, bei einer Ablehnung der Nullhypothese (Annahme eines signifikanten Effekts) falsch zu liegen.
- Typische Schranken sind 10%, 5% oder 1% — ist der *p value* kleiner als eine solche Schranke, weist man die Nullhypothese in der Regel zurück und geht von einem signifikanten Effekt aus.
- Der benötigte Versuchsumfang lässt sich über die verlangte *power* der Tests bestimmen, nämlich über den benötigten Grad der Auflösung (z.B.: Man will eine gefälschte Münze erkennen, wenn die Chance für Kopf größer als 60% ist.) und die Chance, Abweichungen von der Nullhypothese bei dieser Auflösung als solche zu erkennen.

Wir werden dies nun mit mehr Leben füllen und zeigen, wie diese Grundgedanken im Rahmen der Varianzanalyse zum Tragen kommen.

4.5 “Der” Test für DoE: Fishers Varianzanalyse

The analysis of variance is not a mathematical theorem, but rather a convenient method of arranging the arithmetic.

— R.A. Fisher, 1934⁶

Wie nun schon mehrfach erwähnt, ist es das Ziel der Varianzanalyse, Mittelwerte in verschiedenen Gruppen (zum Beispiel durch verschiedene Faktoreinstellungen definiert) auf signifikante Unterschiede zu vergleichen. Dazu wird die gesamte, in den Daten vorhandene Variation in zwei Teile zerlegt — die Variation *zwischen den Gruppen* (die Faktoreffekte) sowie die Variation *innerhalb der Gruppen* (das durch die Faktoren nicht erklärte “Rauschen”). Durch den Vergleich dieser beiden Teile ist man in der Lage, die Nullhypothese, dass alle Gruppen den gleichen Mittelwert haben (das der Faktor keinen Effekt hat) gegen die Alternativhypothese zu testen, dass mindestens ein Gruppenmittel von den anderen verschieden ist (die Daten erlauben es, auf einen signifikanten Effekt des Faktors zu schließen).

4.5.1 Varianzzerlegung

Startpunkt der Betrachtungen ist die sogenannte *Total Sum of Squares* (TSS), die bis auf einen konstanten Faktor der Varianz der insgesamt vorhandenen Daten entspricht. So ergibt sich für insgesamt n_r Messungen y_1, \dots, y_{n_r} des Qualitätsmerkmals y mit Mittelwert \bar{y} die Beziehung

$$TSS = \sum_{i=1}^{n_r} (y_i - \bar{y})^2,$$

man betrachtet also die quadratische Differenz aller Messungen vom Gesamtmittelwert. Je mehr dieser Variabilität durch die Faktorunterschiede erklärt wird, desto sicherer hat man die das Qualitätsmerkmal treibenden Kräfte verstanden.

Man kann nun die *Sum of Squares Between Groups* (SSB) so durch die Abweichungen der Gruppenmittelwerte vom Gesamtmittelwert \bar{y} definieren, dass diese Größe den durch die Gruppenunterschiede (Faktoreffekte) abgebildeten Anteil an der TSS beschreibt; analog kann man eine *Sum of Squares Within Groups* (SSW) bestimmen, die den Rest der Variabilität aufnimmt, der nicht durch die Faktorunterschiede erklärt wird. Die Definitionen von SSB und SSW sind so wählbar, dass die Beziehung

$$TSS = SSB + SSW$$

gilt.

⁶ Die Varianzanalyse ist kein mathematisches Theorem, sondern eine bequeme Methode, die Arithmetik zu arrangieren.

Statt mit einer Formel wollen wir diese Definitionen an dieser Stelle lieber mit dem Beispiel des Rasensprengers mit Leben füllen, der in Kapitel 1.3.1 schon ausführlich dargestellt wurde. Dort wurde ein einfacher Vollfaktorplan vorgestellt, mit dem Drehzahl, Reichweite und Verbrauch eines Rasensprengers ermittelt wurden, und für die Reichweite wurden Ergebnisse gemäß Tabelle 4.11 dokumentiert.

A (α)	B (β)	C (A_q)	Reichweite [m]
—	—	—	4.4088
+	—	—	5.0178
—	+	—	4.5387
+	+	—	5.0691
—	—	+	4.8512
+	—	+	6.4937
—	+	+	5.2425
+	+	+	6.6427

Tabelle 4.11 Auszug aus Tabelle 1.3 - Reichweite des Rasensprengers

Wie kann man die Gesamtvariabilität der gemessenen Reichweiten bestimmen?

Als durchschnittliche Reichweite in allen 8 Versuchen wurde $5.2831m$ gemessen, und nach dem oben Gesagten ergibt sich

$$TSS = (4.4088m - 5.2831m)^2 + \dots + (6.6427m - 5.2831m)^2 = 4.94m^2$$

Dies entspricht $n_r = 8$ mal der Varianz der acht Messungen und damit der in den Daten enthaltenen Gesamtvariabilität.

Der Gruppenmittelwert (mittlere Reichweite bei Faktor A auf level “—”) der Gruppe A— lässt sich aus Tabelle 4.11 als $4.7603m$ errechnen, derjenige der Gruppe A+ als $5.8058m$. Damit ergibt sich, wenn man alle Messungen berücksichtigt, für die SSW, die Abweichungen *innerhalb* der Gruppen, die Größe

$$\begin{aligned} SSW &= \underbrace{(4.4088m - 4.7603m)^2 + \dots + (5.2425m - 4.7603m)^2}_{A-} \\ &\quad + \underbrace{(5.0178m - 5.8058m)^2 + \dots + (6.6427m - 5.8058m)^2}_{A+} \\ &= 2.75m^2, \end{aligned}$$

während sich für die Abweichungen *zwischen* den Gruppen — erneut unter Berücksichtigung aller Messungen — die Größe

$$\begin{aligned} SSB &= 4 \cdot (4.7603m - 5.2831m)^2 + 4 \cdot (5.8058m - 5.2831m)^2 \\ &= 2.19m^2 \end{aligned}$$

ergibt, da sowohl in der Gruppe A— mit Gruppenmittelwert $4.7603m$ als auch in der Gruppe A+ mit Mittelwert $5.8058m$ vier mal gemessen wurde.

Die gesamte in den gemessenen Reichweiten des Rasensprengers vorhandene Variabilität von $4.94m^2$ wurde also in einen durch die unterschiedlichen Einstellungen von Faktor A erklärten Anteil von $2.19m^2$ und in einen "Fehleranteil" von $2.75m^2$ zerlegt. Anders gesagt:

$$\frac{2.19}{4.94} \cdot 100\% = 44.6\%$$

der Variabilität der Reichweiten wird durch die unterschiedlichen Einstellungen des Faktors A erklärt.

Wie sieht dies für Faktor B aus?

Da die Gesamtvariabilität der Reichweiten TSS unverändert bleibt, reicht es aus, die *Sum of Squares Between Groups* B- und B+ auszurechnen. Hierfür ergibt sich

$$\begin{aligned} SSB &= 4 \cdot (5.3733m - 5.2831m)^2 + 4 \cdot (5.1929m - 5.2831m)^2 \\ &= 0.0651m^2. \end{aligned}$$

Für den durch die Veränderung von B nicht erklärten Teil ergibt sich damit automatisch

$$SSW = TSS - SSB = 4.94m^2 - 0.07m^2 = 4.87m^2.$$

Veränderungen des Faktors B erklären damit lediglich 1.4% der Veränderungen der Reichweite.

Man beachte, dass diese Ergebnisse völlig mit der Gestalt des in Abbildung 1.12 reproduzierten Effektdiagrams in Einklang sind, in dem der unterschiedliche Einfluss der Faktoren A und B ebenfalls zu sehen ist.

In der letzten Gleichung wurde sichtbar, dass der durch Faktoreffekte nicht erklärte Anteil der Variabilität des Qualitätsmerkmals automatisch dem Fehler zugeschlagen wurde. Dies ist ein durchgängiges Prinzip. So kann man ebenfalls die beiden Haupteffekte gemeinsam betrachten und die Gesamtvariabilität von 4.94 in die Teile $SSB = 2.19 + 0.07 = 2.26$ und $SSW = 4.94 - 2.26 = 2.68$ zerlegen, und dieses Spiel lässt sich natürlich beliebig weit treiben — bis dem Fehler kein Raum mehr gelassen wird.

Wenn wir aus den 8 Messungen 8 Koeffizienten ermitteln — die Konstante und die Koeffizienten für die Haupteffekte A, B und C, die Wechselwirkungen AB, AC und BC sowie die Dreifachwechselwirkung ABC — lösen wir letztlich ein Gleichungssystem mit 8 Gleichungen und 8 Unbekannten, welches die Daten dann perfekt beschreibt, ohne dem Zufall (Rauschen) dann noch Raum zu lassen.

Mit den in Tabelle 1.10 angegebenen jeweiligen Gruppenmitteln und dem Gesamtmittel von $5.28m$ kann man problemlos für jede Modellkomponente die SSB berechnen wie in Tabelle 4.12 angegeben (wobei sich Abweichungen zu den oben genannten Werten durch Rundung erklären).

Damit kann man auf jeden Fall zunächst einmal die einzelnen potenziellen Modelleingangsgrößen Größen ihrer Bedeutung für die Erklärung der gemessenen Unterschiede in den Reichweiten nach sortieren: A/C, AC, B, AB/AC, ABC. Wo aber soll man den "Schnitt" machen? Während A sicherlich ein signifikanter Faktor ist

	A	B	C	AB	AC	BC	ABC
+	5.81	5.37	5.81	5.24	5.52	5.33	5.26
-	4.76	5.19	4.76	5.32	5.05	5.24	5.30
SSB(Faktor)	2.21	0.06	2.21	0.01	0.44	0.02	0.00
% von total	44.6%	1.3%	44.6%	0.3%	8.9%	0.3%	0.1%

Tabelle 4.12 Zerlegung der Gesamtvariabilität der gemessenen Reichweiten in ihre Anteile

und die Dreifachwechselwirkung ABC sicher nicht, liegt Faktor B in einer Grauzone. Ist der Einfluss von Faktor B auf die Reichweite des Rasensprengers signifikant, oder könnte es sich hier um ein zufälliges, durch Variabilität in den Messungen verursachtes Ergebnis handeln?

4.5.2 Die Anova-Tabelle

Um diese Frage zu beantworten, wollen wir in der Folge schrittweise eine Tabelle aufbauen und erläutern, die zum Standard-Output eines jeden Auswerteprogramms gehört. Dabei wird zu sehen sein, dass die Logik dieser Ausgabe sehr leicht verständlich gemacht werden kann, wenn man schrittweise vorgeht.

Zur Vorbereitung der nun folgenden Überlegungen müssen wir nur den Begriff der Freiheitsgrade (*degrees of freedom, DF*) veranschaulichen. Dies geht am einfachsten, wenn man an die Verteilung von Gegenständen in n Boxen denkt, die synonym mit $n - 1$ Freiheitsgraden ist. Nachdem man die ersten $n - 1$ Boxen mit einer beliebigen Anzahl von Gegenständen gefüllt hat, liegt die Anzahl der Gegenstände für die letzte Box nämlich fest — man hat also lediglich $n - 1$ “Entscheidungen” zu treffen. Es hat sich gezeigt, dass man mit Hilfe des Konzepts der Freiheitsgrade einen sinnvollen Signifikanztest auf der Basis der oben dargestellten Zerlegungen der Gesamtvariabilität definieren kann.

Dazu definiert man zu jedem Faktor bzw. potenziellen Modelleingang (also auch zu allen betrachteten Wechselwirkungen) die Zahl der Freiheitsgrade, wobei

1. ein Modellparameter mit n_l Stufen über $n_l - 1$ Freiheitsgrade verfügt,
2. ein Versuchsplan mit insgesamt n_r Messungen (runs) über $n_r - 1$ Freiheitsgrade,
3. die Differenz zwischen den gesamten und den durch die Parameter im Modell berücksichtigten Freiheitsgrade dem Fehler zugeschlagen wird (vergleiche Tabelle 4.13: Die *nicht* durch die Parameter “verbrauchten” Freiheitsgrade werden der Error-Zeile zugeschlagen).

Unser Beispielversuchsplan zum Rasensprenger war in allen Faktoren zweistufig — sowohl die Haupteffekte als auch die Wechselwirkungen können also durch Messungen in zwei Gruppen beschrieben werden, so dass jeweils ein Freiheitsgrad zur Verfügung steht.

Wir werden nun schrittweise die angekündigte Tabelle aufbauen, die vier verschiedene Unterbeispiele enthält: Die ersten Spalten (Spalten 1-4) zeigen die Zerle-

gung der Variabilität, die sich aus dem “vollen” Modell ergibt, das alle Haupteffekte und Wechselwirkungen ohne Berücksichtigung ihrer Signifikanz enthält; im zweiten Teil (Spalten 5-8) beginnt die Verschiebung der Restvariabilität in den Bereich des Fehlers (durch Weglassen der Wechselwirkungen AB , BC und ABC), schließlich zeigen wir noch einen Schritt mit drei Eingangsgrößen (Faktoren A , B und AC , Spalten 9-12) sowie der Vollständigkeit halber eine Zerlegung mit Faktor A (die wir oben bereits vorgerechnet haben, Spalten 13-16). Somit ergibt sich der Anfang unserer schrittweise aufzubauenden Tabelle wie in Tabelle 4.13 gezeigt:

Faktor	DF	SSB	MS	F	DF	SSB	MS	F	DF	SSB	MS	F	DF	SSB	MS	F
Spalte	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	1	2.21			1	2.21			1	2.21			1	2.21		
B	1	0.06			1	0.06										
C	1	2.21			1	2.21			1	2.21						
AB	1	0.01														
AC	1	0.44			1	0.44			1	0.44						
BC	1	0.02														
ABC	1	0.00														
ERROR	0	0.00			3	0.03			4	0.09			6	2.74		
TOTAL	7	4.95			7	4.95			7	4.95			7	4.95		
Error (% Total)		0%				0.6%				1.8%				55.4%		

Tabelle 4.13 *Sum of Squares Between Groups* für verschiedene Anzahlen von Modellparametern und die resultierenden, unterschiedlichen *Sum of Squares Within Groups*, hier gezeigt als “Error”. Es ist *SSB* die Summe der “*SSBs*”, die zu den jeweiligen Faktoren gehören und *SSW* der entsprechende Wert unter “*ERROR*”.

Man sieht: Im ersten Fall sind alle 7 Freiheitsgrade durch die Modellparameter “belegt” — für den Zufall (=Variabilität/Fehler) bleibt kein Raum. Jeder aus dem Modell entfernte Parameter erhöht allerdings den “Raum für den Zufall”, der — von allen gängigen Programmen als *Error* bezeichnet — mit sinkender Zahl von Modellparametern mehr Raum enthält. Man sieht jedoch auch, dass die ersten Reduktionen des Modells den Fehler deutlich weniger vergrößert haben als die letzte!

Für die endgültige Bewertung des Einflusses einzelner Faktoren und des Fehlers erscheint es sinnvoll, den jeweiligen “Erklärungsbeitrag” *SSB* in Relation zu den zur Lieferung dieses Beitrags benötigten Freiheitsgraden zu setzen — womit man die sogenannten *Mean Squares* (MS) enthält. Man sieht am Beispiel von Tabelle 4.13, dass damit der jeweilige Fehler unterschiedlich bewertet wird, kann sich aber auch unschwer vorstellen, dass diese Unterscheidung sinnvoll ist, wenn zwei Faktoren zwar den gleichen Beitrag *SSB* zur Erklärung der Gesamtvariabilität liefern, ein Faktor dafür allerdings nur zwei Stufen ($DF = 1$), der andere aber 6 Stufen ($DF = 5$) benötigt.

Aus diesem Grunde erweitert man die Zerlegungstabellen um eine weitere Spalte, welche die *Mean Squares* beinhaltet, die man als durchschnittlichen Beitrag pro Freiheitsgrad auffassen kann (siehe Tabelle 4.14).

Faktor	DF	SSB	MS	F	DF	SSB	MS	F	DF	SSB	MS	F	DF	SSB	MS	F
Spalte	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	1	2.21	2.21	*	1	2.21	2.21	221	1	2.21	2.21	98.2	1	2.21	2.21	4.8
B	1	0.06	0.06	*	1	0.06	0.06	6								
C	1	2.21	2.21	*	1	2.21	2.21	221	1	2.21	2.21	98.2				
AB	1	0.01	0.01	*												
AC	1	0.44	0.44	*	1	0.44	0.44	44	1	0.44	0.44	19.6				
BC	1	0.02	0.02	*												
ABC	1	0.00	0.00	*												
ERROR	0	0.00	0.00		3	0.03	0.01		4	0.09	0.02		6	2.74	0.46	
TOTAL	7	4.95			7	4.95			7	4.95			7	4.95		
Error (% Total)		0%				0.6%				1.8%				55.4%		

Tabelle 4.14 Mean Squares. In diesem Beispiel unterscheiden sich SSB und MS lediglich in der Zeile, die den Fehler enthält, da nur dort mehr als ein Freiheitsgrad vorhanden ist.

Damit stehen wir nun schon vor dem vorletzten Schritt, um zu einer endgültigen Einschätzung der Signifikanz der einzelnen Beiträge zu gelangen. Erinnern wir uns an die Ausführungen aus Kapitel 4.4. Dort wurde bereits darauf hingewiesen, dass die gemessenen Unterschiede zwischen Faktorstufen in geeigneter Weise in eine Relation zur Streuung der Versuche (den Unterschieden *innerhalb* der jeweiligen Stufen) gesetzt werden sollte — und es hat sich gezeigt, dass man dies in sinnvoller Weise durch den Vergleich der *Mean Squares* tun kann.

Dazu betrachtet man pro Faktor die Größe

$$F = \frac{MS(Faktor)}{MS(Error)},$$

um die unsere Zerlegungstabelle nun ergänzt wird (siehe Tabelle 4.15).

Faktor	DF	SSB	MS	F	DF	SSB	MS	F	DF	SSB	MS	F	DF	SSB	MS	F
Spalte	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	1	2.21	2.21	*	1	2.21	2.21	221	1	2.21	2.21	98.2	1	2.21	2.21	4.8
B	1	0.06	0.06	*	1	0.06	0.06	6								
C	1	2.21	2.21	*	1	2.21	2.21	221	1	2.21	2.21	98.2				
AB	1	0.01	0.01	*												
AC	1	0.44	0.44	*	1	0.44	0.44	44	1	0.44	0.44	19.6				
BC	1	0.02	0.02	*												
ABC	1	0.00	0.00	*												
ERROR	0	0.00	0.00		3	0.03	0.01		4	0.09	0.02		6	2.74	0.46	
TOTAL	7	4.95			7	4.95			7	4.95			7	4.95		
Error (% Total)		0%				0.6%				1.8%				55.4%		

Tabelle 4.15 Einführung der Testgröße F , die die *Mean Squares* der Faktoren zur SSW in Beziehung setzt.

Bei Betrachtung von Tabelle 4.15 sollte man sich drei Fragen stellen:

1. An wen erinnert der Buchstabe F in dieser allgemein akzeptierten Darstellung?
2. Warum stehen in Spalte 4 keine Werte, sondern nur Sternchen?
3. Wie stehen die sogenannten F -Ratios in Relation zur Signifikanz der Ergebnisse?
Glauben wir eher bei großem oder eher bei kleinem F an einen signifikanten Faktor?

Wer diesem Kapitel von Anfang an gefolgt ist, wird an dieser Stelle natürlich überhaupt nicht mehr überrascht sein, dass der Buchstabe F für die aus den Versuchsergebnissen hergeleiteten Kenngrößen an R.A. Fisher erinnert — Fisher war derjenige, der die Varianzanalyse Hand in Hand mit den Grundideen der statistischen Versuchsplanung entwickelt hat.

Auch die Einführung der Sternchen sollte nicht überraschen — da dem Fehler im “vollen” Modell kein Raum gelassen wurde, müsste man zur Bestimmung der Größe F durch Null dividieren ... Um in solchen Situationen zu einer Entscheidung zu gelangen, welche Größen man als erste entfernen sollte, bietet sich übrigens ein Blick auf den *half normal plot* (vergleiche Kapitel 3.3.1) an, der uns graphisch hilft, die aussichtsreichsten Kandidaten zu entlarven.

Vergegenwärtigt man sich schließlich an dieser Stelle noch einmal den bisherigen Gedankengang, so stellt man fest, dass F in einem sehr klar definierten Sinn die Stärke des “Signals” (Faktoreffekts) ins Verhältnis zur Stärke des “Rauschens” (error = nicht erklärter Teil der Variabilität) setzt. Je größer F , desto weniger sollten wir geneigt sein, an zufällige Effekte zu glauben — womit wir an die Zusammenfassung von Kapitel 4.4 anknüpfen können, in der es hieß:

Ein Effekt ist “signifikant”, wenn er *nicht* mit der “Nullhypothese” einer rein zufälligen Abweichung vereinbar ist. Um dies zu testen, wird eine passende Kenngröße definiert, die den Grad der Abweichung von der Nullhypothese quantifiziert (...).

Dies haben wir mit der Definition von F nun erledigt. Obwohl die Konstruktion von F vielleicht nicht unmittelbar auf der Hand liegt, hat sie einen Vorteil: F wurde so konstruiert, dass auch der nächste Punkt aus der Zusammenfassung erfüllbar ist:

Die Wahrscheinlichkeit dafür, dass diese Kenngröße — wenn die Hypothese gilt — bestimmte Werte annimmt, muss bestimmbar sein (...).

Hier zeigt sich das Zusammenspiel von Fishers Gedanken zur Versuchsplanung und Varianzanalyse: Die sinnvollsten Ergebnisse kann man erreichen, wenn man bei der Planung der Experimente ihre spätere Auswertbarkeit direkt mit berücksichtigt.

4.5.3 Von der Testgröße zur Wahrscheinlichkeit

Aus didaktischen Gründen beschränken wir uns in der Folge auf die Darstellung für ein Experiment mit nur einem Faktor. Diesem liegt eine Modellannahme der Art

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

zugrunde, die besagt, dass man das j -te Messergebnis auf Faktorstufe i zerlegen kann in drei Teilkomponenten

μ = Mittelwert aller Messungen

τ_i = Effekt der i -ten Stufe

ε_{ij} = durch den Modell nicht erfasster Fehler

(bei n_s Stufen und n Messungen je Stufe gilt $1 \leq i \leq n_s, 1 \leq j \leq n$).

Dabei, so die Modellannahme, sollen die Fehler ε_{ij} unabhängig voneinander sein und einer Normalverteilung mit Mittelwert Null und konstanter Varianz σ^2 folgen, dem Standardmodell für zufälliges Rauschen. Es soll, anders gesagt, keine systematischen Differenzen zwischen Messung und Vorhersage " $\mu + \tau_i$ " geben als die durch eine Gaußsche Glockenkurve beschriebenen — sonst hat man einen (oder mehrere) wichtige Faktoren übersehen.

Sind diese Annahmen und die Nullhypothese erfüllt, so kann man die Wahrscheinlichkeitsverteilung der Größe F ermitteln.

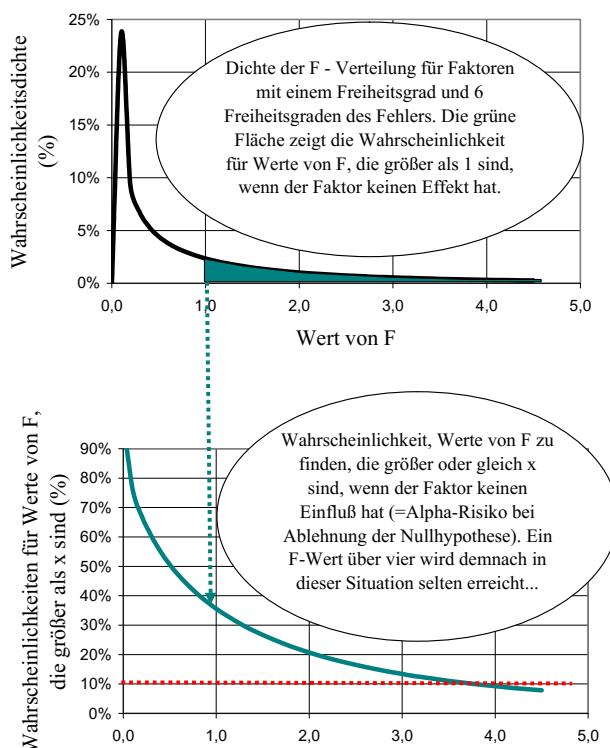


Abb. 4.7 Die Graphik zeigt exemplarisch die Dichte der $F_{1,6}$ -Verteilung, die benutzt wird, wenn in 8 Versuchen ein Freiheitsgrad für einen einzelnen, zweistufigen Faktor und 6 Freiheitsgrade für den Fehler vorliegen.

Es gilt:

In allen durch die jeweiligen Faktoreinstellungen definierten Gruppen seien die Messungen unabhängig voneinander, normalverteilt mit Mittelwert $\mu + \tau_i$ und pro Gruppe identischer Varianz σ^2 .

Sind diese Bedingungen erfüllt, so folgt die Größe F unter Gültigkeit der Nullhypothese, dass ein Faktor keinen Effekt hat (dass also $H_0 : \tau_1 = \dots = \tau_{n_s} = 0$ gilt), einer sogenannten F-Verteilung.

Da diese Verteilung in allen Auswerteprogrammen (ebenso wie in Tabellenkalkulationsprogrammen) hinterlegt ist, braucht ihre mathematisch etwas kompliziertere Formel an dieser Stelle nicht zu interessieren; wichtig ist lediglich, dass sie durch zwei Parameter, die sogenannten Freiheitsgrade f_1 und f_2 , beschrieben wird (man findet in vielen Statistikbüchern auch Tabellen). Man muss sich lediglich den Ablauf des Hypothesentests merken:

Grundprinzip der Varianzanalyse

1. Man unterstellt, dass der zu untersuchende Faktor *keinen* Effekt hat. Dies ist die Nullhypothese.
2. Wenn die Nullhypothese gilt, so folgt aus den oben formulierten Annahmen, dass die Testgröße F einer F_{f_1, f_2} -Verteilung folgt mit

$$f_1 = DF(\text{Faktor})$$

und

$$f_2 = DF(\text{Error}).$$

3. Damit lässt sich die Wahrscheinlichkeit dafür bestimmen, in einem Experiment einen F-Wert zu finden, der mindestens so groß ist wie der vorgefundene — dies ist der sogenannte *p*-Wert oder *p-value*.
4. Ist diese Wahrscheinlichkeit klein, so verwirft man die Nullhypothese, da die Abweichung von der Erwartung zu groß ist, und geht von einem signifikanten Effekt des betreffenden Faktors aus.
5. Hat man durch schrittweises Entfernen nicht signifikanter Faktoren schließlich ein endgültiges, reduziertes Modell gefunden, so muss man abschließend die Gültigkeit der oben genannten Modellannahmen überprüfen (vgl. Kapitel 4.6).

Der letzte Punkt ist von fundamentaler Bedeutung, denn die ermittelten *p-values*, auf deren Basis wir über Signifikanz entscheiden, haben nur Gültigkeit, wenn die folgenden Bedingungen erfüllt sind:

1. Die Messungen auf jeder Faktorstufe sind normalverteilt...
2. ...mit jeweils gleicher Varianz...
3. ...und die ermittelten Residuen (Modellfehler) sind unabhängig und ebenfalls normalverteilt um den Gesamtmittelwert.

Es muss also, will man zu endgültig abgesicherten Aussagen gelangen, in jedem Einzelfall überprüft werden, ob man von diesem Annahmen ausgehen kann. In der Praxis geht man dabei so vor, dass man zunächst sein endgültiges Modell ermittelt (also alle “nicht signifikanten” Faktoren ausklammert) und dann die Modellfehler (Residuen) des endgültigen Modells untersucht. Wie dies funktioniert — und warum es mehr ist als nur “statistische Nabelschau” — wird in Abschnitt 4.6 erläutert.

Kommen wir zunächst auf das Beispiel der Reichweite des Rasensprengers zurück. Wir können nun nämlich die zuvor bereits schrittweise aufgebauten Zerlegungstabellen unter Berücksichtigung der $F_{DF(Faktor),DF(Error)}$ -Verteilung ergänzen. Dazu gehen wir schrittweise vor, beginnend mit einem ersten Ansatz, der alle Faktoren und zwei Zweifachwechselwirkungen enthält.

Faktor	DF	SSB	MS	F	p
Spalte	1	2	3	4	5
A	1	2.21	2.21	442	0.002
B	1	0.06	0.06	12	0.074
C	1	2.21	2.21	442	0.002
AC	1	0.44	0.44	88	0.011
BC	1	0.02	0.02	4	0.184
ERROR	2	0.01	0.005		
TOTAL	7	4.95			

Tabelle 4.16 Ein Zwischenstand, wie er sich nach Entfernung von ABC und AB ergibt.

Da in diesem Beispiel alle Modellparameter über die gleiche Anzahl von Freiheitsgraden verfügen, können wir hier für alle Parameter von einer $F_{1,2}$ -Verteilung ausgehen, um die entsprechenden p -Werte zu ermitteln und dabei feststellen:

1. Die Chance, ein Verhältnis von 442 von Faktoreffekt zum Rauschen zu haben, wenn Faktor A in Wahrheit keinen Effekt hat, ist 0.2%.
2. Für Faktor B ist diese Chance 7.4%, für die Wechselwirkung BC 18.4%.

Zur Erinnerung: Für die Wechselwirkung BC bedeutet dies, dass die Wahrscheinlichkeit für einen Fehler erster Art, das α -Risiko, 18.4% ist. Man würde die Hypothese *nicht* verwerfen, dass dieser Modellparameter *keinen* Einfluss hat — man hat keinen Beweis für einen signifikanten Einfluss gefunden und würde ihn im nächsten Schritt aus dem Modell entfernen. Dabei ist zu bedenken, dass der F -Wert sich “innerhalb des Modells” bestimmt; so sehen wir in Tabelle 4.15, je nach Anzahl vorhandener Modellparameter, unterschiedliche F -Werte für den Faktor A. Es ist deshalb ratsam, die Entscheidung über die Signifikanz von Parametern schrittweise zu fällen — und nicht alle im Rahmen des ersten Modells als “nicht signifikant” erscheinenden Parameter zugleich zu entfernen. So ergibt sich aus Tabelle 4.16 nach Eliminierung der Wechselwirkung BC die Situation aus Tabelle 4.17.

An dieser Stelle bietet sich nun ein Hinweis zur Auswertung bei Blockbildung (vergleiche Kapitel 4.3) an — ein Hinweis, der bei einer ersten Lektüre auch überlesen werden mag...

Faktor	DF	SSB	MS	F	p
Spalte	1	2	3	4	5
A	1	2.21	2.21	221	0.001
B	1	0.06	0.06	6	0.092
C	1	2.21	2.21	221	0.002
AC	1	0.44	0.44	44	0.001
ERROR	3	0.03	0.01		
TOTAL	7	4.95			

Tabelle 4.17 Veränderung des Zwischenstands nach Entfernung von BC.

4.5.4 Auswertung bei Blockbildung

Wie berücksichtigt man Blöcke bei der statistischen Versuchsauswertung?

Der Sinn der Blockbildung wurde in Abschnitt 4.3 erläutert, das Vorgehen bei der Planung der Versuche anhand eines Beispiels erklärt. Dabei ging es um eine Bäuerin, die den Weizenertrag auf ihren Feldern optimieren wollte und dazu einen einfachen Vollfaktorplan mit 2 Faktoren (Weizensorte und Düngemittel) plante. Die Blöcke sollten dazu dienen, etwaige Einflüsse von Bodenbeschaffenheit, Sonneneinstrahlung etc. berücksichtigen zu können. Der entsprechende Versuchsplan wurde in Tabelle 4.5 vorgestellt. Tabelle 4.18 zeigt nun die (hypothetischen) Resultate ihrer Messungen. Wertet man dieses Ergebnis ohne Berücksichtigung der Blöcke aus, so

Block	Dünger A	Weizen B	Ertrag [dt/ha]
1	+	+	84.25
1	-	-	73.18
1	+	-	73.69
1	-	+	62.64
2	+	+	74.78
2	+	-	77.48
2	-	+	58.59
2	-	-	66.64
3	+	+	92.79
3	-	+	70.11
3	-	-	71.32
3	+	-	82.04

Tabelle 4.18 Hypothetische Ernteerträge, in Dezitonnen pro Hektar.

ergibt sich analog zu den obigen Beispielen, das in Tabelle 4.19 gezeigte Bild.

Das auf allen Faktoren beruhende Modell lässt $\frac{288.3}{979.2} = 29\%$ der Gesamtvariabilität unerklärt. Die Signifikanz der Wechselwirkung wäre fraglich. Berücksichtigt man dagegen die Blöcke wie einen weiteren Faktor, so ergibt sich ein anderes Bild (vgl. Tabelle 4.20): es verbleibt ein unerklärter Rest an Variabilität in der Größenordnung von $\frac{98.8}{979.2} = 10\%$.

Faktor	DF	SSB	MS	F	p
A	1	567.9	567.9	15.757	0.004
B	1	0.1	0.1	0.003	0.959
AB	1	122.9	122.9	3.410	0.102
ERROR	8	288.3	36.04		
TOTAL	11	979.2			

Tabelle 4.19 Varianzanalyse der Ernteerträge, ohne Berücksichtigung der Blöcke. Faktor Dünger ist signifikant, Faktor Weizen nicht. Bei der Wechselwirkung handelt es sich um einen Zweifelsfall.

Faktor	DF	SSB	MS	F	p
BLOCK	2	189.5	94.8	5.75	0.04
A	1	567.9	567.9	34.49	0.001
B	1	0.1	0.1	0.01	0.935
AB	1	122.9	122.9	7.47	0.034
ERROR	6	98.8	16.5		
TOTAL	11	979.2			

Tabelle 4.20 Varianzanalyse der Ernteerträge, diesmal mit Berücksichtigung der Blöcke. Die Blöcke sind signifikant, der Faktor Dünger bleibt signifikant, und die Wechselwirkung wird deutlich sichtbar, nachdem man den Beitrag der Blöcke vom Error abgezogen hat.

Die Bäuerin hat also ohne viel Aufwand die Erklärungskraft des Modells deutlich steigern können, auch wenn letztlich die Faktoren “hinter” den Blöcken nicht kontrollierbar sind. Der zusätzliche Nutzen ist aber größer: Die Interaktion von Weizen und Düngemittel wurde vom Effekt der Blöcke versteckt — sie wurde erst nach Berücksichtigung des Beitrags der Blöcke zur Gesamtvariabilität sichtbar! Wäre der Faktor “Block” *nicht* signifikant gewesen, hätte er ohne Verlust einfach aus der weiteren Betrachtung ausgeschlossen werden können. Die Bildung von Blöcken erlaubt es also, ohne viel Aufwand Faktoren zu berücksichtigen, die außerhalb des Systems liegen, aber als Störgröße durchaus Einfluss haben. Somit können Einflüsse innerhalb des Systems sichtbar gemacht werden, die ansonsten — wie die Wechselwirkung von Weizensorte und Düngemittel — verborgen geblieben wären.

4.5.5 Faktorelimination

Nach diesem Einschub, der zeigt, wie auch Blockbildung mit der Varianzanalyse auswertbar ist und wieso die Zuordnung von Versuchen zu Blöcken ein sinnvoller Weg sein kann, dass “Messrauschen” zu minimieren, sei nun auf die Prinzipien der Faktorelimination eingegangen.

Grundprinzip der Faktorelimination

Es hat sich bewährt, Parameter in einer bestimmten Reihenfolge zu eliminieren:

1. Wurde geblockt, so untersuche man zunächst, ob der Faktor "Block" signifikant ist. Wenn ja, sollte er im Modell verbleiben, wenn nein entfernt werden.
2. Man beginnt dann mit der Analyse der Wechselwirkungen der höchsten Stufe. Diese sollten als erstes entfernt werden, da sie in der Regel den kleinsten Beitrag liefern (Beispiel: die Dreifachwechselwirkung *ABC* auf die Reichweite des Rasensprengers).
3. Dann arbeitet man sich schrittweise durch die Wechselwirkungen der niedrigeren Stufen (im Falle des Rasensprengers also die Zweifachwechselwirkungen).
4. Schließlich eliminiert man nicht signifikante Haupteffekte.
5. Dabei ist zu beachten, dass nicht signifikante Parameter trotzdem erhalten bleiben müssen, wenn sie in Wechselwirkungen höherer Stufe auftauchen. Selbst wenn der Parameter *C* keinen signifikanten Haupteffekt hätte, aber in eine signifikante Wechselwirkung *AC* verstrickt wäre, müsste *C* im Modell bleiben. Man nennt dies auch "Bewahren der hierarchischen Integrität" des Modells.

Im Rahmen unseres Rasensprenger-Beispiels würde es nun zunächst einmal vom akzeptierten α -Risiko abhängen, ob man Faktor *B* im Modell belässt oder nicht. Ausgehend von 10%, würde man *B* für signifikant halten und im Modell belassen, ausgehend von 5% würde man *B* entfernen (vgl. Tabelle 4.17).

Man sollte hier allerdings Augenmaß und gesunden Menschenverstand walten lassen. Gibt es inhaltlich sinnvolle Gründe, warum *B* einen Einfluss haben sollte, ist es natürlich ratsam, *B* auch im Modell zu belassen (gegebenenfalls kann man versuchen, durch zusätzliche Experimente ein klareres Bild zu erhalten). Auf der anderen Seite sollte jedes Modell so einfach wie möglich sein — auf überflüssige Erklärungsparameter sollte man schon deswegen verzichten, weil sie zusätzlichen Aufwand (zum Beispiel für künftige Messungen und Kontrollen) beinhalten.

Damit stellt sich eine zentrale Frage: Verändert sich die Erklärungskraft des Modells, wenn man *B* weglässt?

Ohne *B* bekommt die Zerlegungstabelle die in Tabelle 4.21 gezeigte Form, in der nur noch zweifelsfrei signifikante Parameter verblieben sind.

Zum Vergleich der Güte der Modelle mit verschiedener Anzahl von Parametern bietet sich neben der jeweiligen Residuenanalyse (Kap. 4.6) an, die jeweilige Erklärungskraft zu vergleichen, die wir oben schon als Anteil erklärter Variabilität

$$\frac{SSB}{SST} = 1 - \frac{SSW}{SST}$$

kennen gelernt haben und nun auch formal mit ihrem "offiziellen Namen" R^2 bezeichnen wollen.

Faktor	DF	SSB	MS	F	p
Spalte	1	2	3	4	5
A	1	2.21	2.21	98.2	0.001
C	1	2.21	2.21	98.2	0.001
AC	1	0.44	0.44	19.6	0.011
ERROR	4	0.09	0.02		
TOTAL	7	4.95			

Tabelle 4.21 Ein endgültiges Modell.

Mit R^2 bezeichnet man den durch ein mathematisches Modell erklärten Anteil an der Gesamtvariabilität der Daten; im Falle der Varianzanalyse gilt $R^2 = \frac{SSB}{SST}$.

Bereits in der Konstruktion der Zerlegungstabellen wurde allerdings klar, dass man mit mehr Parametern stets eine kleinere Restvarianz SSW und damit eine größere erklärte Varianz SSB erreicht. Darum hat es sich zur Vermeidung von falschen Schlüssen eingebürgert, zusätzlich zu R^2 eine adjustierte Größe $R^2_{adjusted}$, kurz R^2_{adj} , zu betrachten, die die "Effizienz" des Modells im Hinblick auf die Anzahl n_m der geschätzten Modellkonstanten berücksichtigt, die man aus den insgesamt verfügbaren n_r Messungen schätzen will:

$$R^2_{adj} = 1 - \frac{\frac{SSW}{n_r - n_m}}{\frac{SST}{n_r - 1}} = 1 - \frac{n_r - 1}{n_r - n_m} \cdot (1 - R^2)$$

Man erkennt unschwer, dass der Grad $\frac{R^2_{adj}}{R^2} \cdot 100\%$ der Korrektur sowohl von n_r und n_m als auch vom erreichten R^2 abhängt. Was dies konkret bedeutet, zeigt Tabelle 4.22, die diese Größen für den Rasensprenger mit $n_r = 8$ exemplarisch zueinander in Beziehung setzt.

Konstanten n_m	$R^2 = 70\%$	$R^2 = 80\%$	$R^2 = 90\%$	$R^2 = 95\%$	$R^2 = 99\%$
1	70%	80%	90%	95%	99%
2	65%	77%	88%	94%	99%
3	58%	72%	86%	93%	99%
4	48%	65%	83%	91%	98%
5	30%	53%	77%	88%	98%
6	0%	30%	65%	83%	97%
7		0%	30%	65%	93%

Tabelle 4.22 Die Tabelle zeigt R^2_{adj} ; hier für $n_r = 8$. Das Verhältnis von R^2 zu R^2_{adj} hängt von der Anzahl der Modellkonstanten n_m und R^2 ab.

Bei überragend guten Modellen, also solchen mit hoher Erklärungskraft R^2 , spielt die Adjustierung kaum eine Rolle. Je schwächer jedoch die Erklärungskraft des Modells ist, desto größer der Effekt der Adjustierung.

Für unser Beispiel der Reichweite des Rasensprengers bedeutet dies, dass wir angesichts des hohen erzielten R^2 von 98%-99% von R^2_{adj} keine weiteren Aufschlüsse erhoffen dürfen. Um den Effekt dennoch an diesem Beispiel verdeutlichen zu können, haben wir die in der Simulation “gemessenen” Reichweiten durch künstliches Rauschen gestört, indem wir normalverteilte Zufallszahlen mit Mittelwert 0 und einer kleinen Standardabweichung addiert haben. Tabelle 4.23 zeigt die sich für diese künstlich verrauschten Werte ergebende Reduktion von Parametern mit den zugehörigen Werten von R^2 und R^2_{adj} .

Anzahl Parameter	Faktoren	R^2	R^2_{adj}
7	A, B, C, AB, AC, BC	97.1%	79.6%
6	A, B, C, AC, BC	96.9%	89.1%
5	A, B, C, AC	96.6%	92.2%
4	A, C, AC	96.4%	93.7%
3	A, C	91.8%	88.6%
2	A	40.6%	30.7%

Tabelle 4.23 Während R^2 mit der Anzahl am Modell beteiligter Parameter wächst, zeigt R^2_{adj} einen Wendepunkt. Das “effizienteste” Modell verfügt über vier Parameter, nämlich A, C, AC und eine Konstante.

Die gezeigten Verläufe von R^2 und R^2_{adj} sind typisch. Während R^2 mit der Anzahl der am Modell beteiligten Parameter steigt, zeigt R^2_{adj} typischerweise ein Maximum an anderer Stelle. Die Erklärungskraft des Modells steigt zunächst mit weiteren Parametern, aber irgendwann kommt es zum Wendepunkt, ab dem die Integration neuer Modellparameter nicht mehr hilft. So würde im hier konstruierten Fall die Modelleffizienz durch Hinzufügen des Parameters B zurückgehen.

Die oben gemachten Berechnungen wurden, um sie händisch nachvollziehbar zu halten, mit den angegebenen zwei Nachkommastellen gemacht. Rechnet man mit den vollen, in Tabelle 4.11 angegebenen vier Stellen, so ergibt sich, dass das endgültige Modell auf den Parametern A, C und AC beruhen sollte. Dieses Modell würde sich wie folgt darstellen:

$$\text{Reichweite} = 5.2831m + 0.5228m \cdot A + 0.5245m \cdot C + 0.2379m \cdot AC,$$

wobei sich die jeweiligen Koeffizienten aus den halbierten Faktoreffekten ergeben und auf die durch -1 und $+1$ normierten Faktoreinstellungen anzuwenden sind. Damit ergibt sich der in Tabelle 4.24 gezeigte Vergleich von Versuchsergebnis und Modellvorhersage.

Erinnert man sich daran, dass die “Messungen” zu diesem Beispiel aus einer Computersimulation stammen, so verwundert es natürlich nicht, dass das hier er-

A	B	C	Messung [m]	Vorhersage [m]	Abweichung (Residuum) [m]
-	-	-	4.41	4.47	-0.06
+	-	-	5.02	5.04	-0.03
-	+	-	4.54	4.47	+0.06
+	+	-	5.07	5.04	+0.03
-	-	+	4.85	5.05	-0.20
+	-	+	6.49	6.57	-0.07
-	+	+	5.24	5.05	+0.20
+	+	+	6.64	6.57	+0.07

Tabelle 4.24 Der Vergleich von Messung und Vorhersage des reduzierten Modells zeigt, dass die zentralen Parameter in der Lage sind, die Reichweite des Rasensprengers sehr genau vorherzusagen.

mittelte Modell so gut funktioniert — viel Raum für Messfehler und Variabilität ist in dieser “Laborsituation” nicht geblieben. Man sollte sich jedoch auch daran erinnern, dass das ursprüngliche Simulationsmodell auf acht Parametern beruhte (vgl. Anhang A), aus denen die bedeutenden Parameter (und deren Wechselwirkung) schnell und letztlich unkompliziert herausgefischt werden konnten — dabei natürlich immer stillschweigend vorausgesetzt, dass die Annahmen der ANOVA erfüllt sind, was wir noch prüfen müssen.

Wir werden auf diese Prüfung im nächsten Abschnitt zurückkommen. Zuvor wollen wir jedoch noch einige Aspekte der Varianzanalyse behandeln, die im Laufe dieses Kapitels an verschiedenen Stellen bereits angesprochen wurden — beginnend mit einer Erinnerung an die in Kapitel 4.4 vorgestellten “Grundprinzipien des statistischen Testens”, die wir nun noch einmal mit den Grundgedanken der Varianzanalyse verbinden wollen.

- Um Unterschiede, zum Beispiel zwischen Faktoreinstellungen, zu bewerten, sollten sie in Relation zur Streuung der jeweiligen Versuche gesetzt werden.

Genau dies geschieht bei der Varianzanalyse. Die Größen SSB und SSW wurden genau so konstruiert, dass dieser zentrale Gedanke abgebildet wird — dies wurde oben ausführlich erläutert.

- Ein Effekt ist “signifikant”, wenn er *nicht* mit der “Nullhypothese” einer rein zufälligen Abweichung vereinbar ist.
- Um dies zu testen, wird eine passende Kenngröße definiert, die den Grad der Abweichung von der Nullhypothese quantifiziert (z.B. die Anzahl gewürfelter Köpfe in 10 Versuchen).

Im Fall der Varianzanalyse handelt es sich um die Größe

$$F = \frac{MS(Faktor)}{MS(Error)},$$

die Abweichung von der Annahme eines *nicht* vorhandenen Einflusses — je stärker der Einfluss des Faktors, desto größer wird F . Gibt es umgekehrt keinen Einfluss des Faktors, d.h. sind die Mittelwerte in beiden Gruppen gleich, so wird SSB und damit auch $MS(Faktor)$ und somit F zu Null.

- Die Wahrscheinlichkeit dafür, dass diese Kenngröße — wenn die Hypothese gilt — bestimmte Werte annimmt, muss bestimbar sein (z.B. als Binomialverteilung, die uns erlaubt, die Chance für 7 oder mehr Köpfe in 10 Würfen zu bestimmen, wenn die Münze fair ist).

Es wurde dargestellt, dass die Größe F einer sogenannten F -Verteilung mit $f_1 = DF(Faktor)$ und $f_2 = DF(Error)$ Freiheitsgraden folgt, wenn die Nullhypothese stimmt, dass der Faktor keinen Einfluss hat und die Bedingungen für die Varianzanalyse erfüllt sind. Damit ergibt sich automatisch der Bezug zu den beiden nächsten Punkten in unseren Grundprinzipien des statistischen Testens:

- Damit lässt sich zu jedem gefundenen Versuchsergebnis (= Wert der definierten Kenngröße) ein Wahrscheinlichkeitswert (“ p value”) p angeben, der die Wahrscheinlichkeit dafür beschreibt, die gemessene oder eine größere Abweichung zu finden, wenn man die Gültigkeit der Hypothese unterstellt. Der p value beschreibt also das Risiko, bei einer Ablehnung der Nullhypothese (Annahme eines signifikanten Effekts) falsch zu liegen.
- Typische Schranken sind 10%, 5% oder 1% — ist der p value kleiner als eine solche Schranke, weist man die Nullhypothese in der Regel zurück und geht von einem signifikanten Effekt aus.

Es wurde bereits verdeutlicht, wie dies beispielsweise für die Bestimmung der signifikanten Effekte für den Rasensprenger aussieht. Lediglich zum letzten der erarbeiteten Punkte fehlen noch einige (wenige) Anmerkungen:

- Der benötigte Versuchsumfang lässt sich über die verlangte *power* der Tests bestimmen, nämlich über den benötigten Grad der Auflösung (z.B.: Man will eine gefälschte Münze erkennen, wenn die Chance für Kopf größer als 60% ist.) und die Chance, Abweichungen von der Nullhypothese bei dieser Auflösung als solche zu erkennen.

Die dabei zugrunde liegende Idee ist natürlich, genügend Daten zu sammeln, um Effekte “praktisch brauchbarer” Größen ermitteln zu können, ohne übertrieben viele Versuche zu machen — jeder Versuch kostet schließlich Geld...

4.5.6 Versuchszahl

Wie also gelangt man zu einer Abschätzung der benötigten Anzahl von Versuchen?

Nach den bisherigen Ausführungen ist klar, dass die benötigte Versuchszahl von verschiedenen Parametern abhängt:

- Wie schon in den einführenden Beispielen (Münzwürfe in der Gaststätte) gilt es zunächst, das α -Risiko festzulegen — das Risiko, einen Faktor zu Unrecht als signifikant anzusehen, obwohl er in Wahrheit keinen Einfluss hat.
- In einem zweiten Schritt muss man überlegen, welche “Auflösung” man erwartet — also welche Kraft, signifikante Effekte als solche zu erkennen. Diese Kraft, in Abschnitt 4.4 als *Power* eingeführt, hängt natürlich von der wahren Stärke des Effektes (des Unterschiedes zwischen den Gruppenmittelwerten) ab — je größer der Effekt, desto leichter ist er zu erkennen. Wir möchten an dieser Stelle von “praktischer Signifikanz” (im Unterschied zur statistischen Signifikanz) reden, der inhaltlichen Relevanz des Effektes einer gewissen Größenordnung. Während uns sehr kleine Effekte meist egal sein können, stellt sich damit die Frage, ab welcher Größenordnung die Erkenntnis eines Effektes relevant wird — ab wann wir diese Erkenntnis nutzen können, etwa um das betrachtete System zu optimieren. Neben α benötigen wir also noch ein inhaltlich begründetes Maß für die Stärke Δ des Unterschiedes zwischen den Faktorstufen, sowie ein Maß für die Power $1 - \beta(\Delta)$, mit der wir diesen Unterschied konstatieren können wollen.
- Ferner benötigen wir zur endgültigen Abschätzung der benötigten Versuchszahl die Anzahl der Faktorstufen. Gibt es Faktoren mit unterschiedlich vielen Stufen, ist die maximale Stufenzahl relevant.

Bis hierher war die Darstellung einfach, denn es handelte sich bisher nur um eine Auflistung von Parametern, die wir selber frei wählen können (natürlich geleitet von inhaltlichen Überlegungen):

α	Akzeptables Risiko, einem nicht signifikanten Effekt “aufzusitzen”
$1 - \beta$	“Power”: Gewünschte Wahrscheinlichkeit, einen Effekt zu erkennen...
Δ	...wenn er mindestens diese Größe Δ hat (“praktische Signifikanz”!) und ...
Stufen	... maximal diese Anzahl von Faktorstufen eingeplant wird

Tabelle 4.25 Frei wählbare Eingangsparameter zur Bestimmung der benötigten Versuchszahl

Leider benötigen wir aber darüber hinaus allerdings auch noch einen weiteren Parameter zur Bestimmung der benötigten Versuchszahl: eine Abschätzung der dem System bzw. Modell zugrunde gelegten Standardabweichung σ . Dies begründet sich aus der nun schon mehrfach diskutierten Tatsache, dass die Merkmalsunterschiede zum “Rauschen” ins Verhältnis gesetzt werden müssen und passt zur Grundannahme der Varianzanalyse, dass die Standardabweichung in allen verglichenen Gruppen identisch sein muss (zur Erinnerung: diese Annahme bleibt noch zu prüfen!):

σ	Standardabweichung der Messwerte (Schwankung um die Faktormittelwerte)
----------	--

Tabelle 4.26 Keine Wahl: Die Standardabweichung als weiterer Eingangsparameter

Je größer die Standardabweichung (das Rauschen) ist, desto mehr Versuche werden wir benötigen, um die statistische Signifikanz von praktisch signifikanten Unterschieden nachweisen zu können. Wie aber gelangen wir zu einer Abschätzung von σ , der einzigen Größe, die wir nicht "frei" festlegen können?

Die einfachste Möglichkeit ist natürlich, dass schon Erfahrungswerte zur Messstreuung vorliegen, von denen man ausgehen kann. Ist dies nicht der Fall, kann man im Rahmen einer Vorstudie eine Abschätzung für σ gewinnen oder versuchen, Expertenschätzungen einzuholen. Bleibt man unsicher, sollte man verschiedene Schätzwerte für σ ausprobieren und die jeweils benötigten Versuchszahlen ansehen — möglicherweise zeigt sich, dass man ohnehin keine genauere Abschätzung braucht, da man in einer Situation ist, in der kleinere Veränderungen von σ ohnehin keine Auswirkungen (mehr) haben.

Die mit diesen Parametern durchzuführenden Berechnungen sind relativ komplex (genaueres findet man z.B. bei [10, S. 101ff.]), aber glücklicherweise bieten manche Statistikprogramme Verfahren an, die uns das Leben erleichtern.

Ein Beispiel: Eine Fahrzeugtechnikerin möchte drei verschiedene neue Technologien, die helfen sollen, den Normverbrauch eines Fahrzeugs zu reduzieren, mit einem nicht modifizierten Basisfahrzeug vergleichen. Aus der Erfahrung ist bekannt, dass die Standardabweichung des dazu benutzten Prüfstandes in der Größenordnung von $0.1l/100km$ liegt.

1. Da man nicht ohne Grund in neue, teure Technologien investieren möchte, soll das α -Risiko auf 5% beschränkt sein — die Wahrscheinlichkeit, einen Effekt einer Technologie festzustellen, der nicht existiert, soll maximal 5% betragen.
2. Existiert ein Effekt, so will man ihn mit 80% Sicherheit erkennen, wenn er eine Verbesserung von...
3. ...mindestens $0.25l/100km$ bedeutet.
4. Sie misst auf vier verschiedenen Stufen und sucht einen Unterschied zwischen zwei beliebigen Faktorstufen.
5. Die der Anzahl benötigter Versuche zugrunde liegende Standardabweichung wird mit $0.1l/100km$ angenommen.

Mit diesen Parametern kann man mit Programmhilfe ermitteln, dass 5 Versuche pro Stufe benötigt werden, denn damit erreicht man eine Power von 84.7%, während mit 4 Wiederholungen lediglich eine von 71% erreicht würde. Hält man auch kleinere Verbesserungen für praktisch signifikant, erhöht sich die Anzahl benötigter Versuche schnell, wenn man die Power von 80% beibehalten möchte, wie aus Tabelle 4.27 ersichtlich wird.

Tabelle 4.27 enthält im übrigen zwei gute Nachrichten:

1. Die gezeigten Versuchsumfänge gelten *je Faktorstufe*. Hat man aber mehrere Faktoren, so wird jede Stufe ohnehin schon mehrfach gemessen!

Effekt Δ (l/100km)	benötigte Anzahl von Versuchen $\sigma = 0.1l/100km$	$\sigma = 0.2l/100km$
0.10	23	89
0.15	11	40
0.20	7	23
0.25	5	15
0.30	4	11

Tabelle 4.27 Die benötigte Zahl von Messungen hängt sowohl von der detektierenswerten Effektstärke Δ als auch vom Messrauschen σ ab.

2. Die Tabelle zeigt gleiche Werte, wenn das Verhältnis des zu messenden Effektes Δ zur Standardabweichung σ identisch ist. Um einen Effekt der Größe der Standardabweichung mit 80% Wahrscheinlichkeit zu erkennen, werden 23 Versuchswiederholungen benötigt, für das 1.5-fache von σ noch 11.

Diese Normierbarkeit des Effektes (Effektstärke gemessen als Vielfaches der angenommenen Standardabweichung) können wir nutzbar machen, um eine einfache Tabelle zu generieren, die für verschiedene (gängige) Werte von α , Power und Stufenzahl die benötigte Anzahl von Messungen pro Stufe zeigt.

Hätte dies nicht Tabelle 4.28, die Technikerin nur eine einzelne neue Technologie gegen das Basisfahrzeug verglichen (2 Faktorstufen), so hätte sie, um einen Effekt von $0.25l/100km$ bei einer Standardabweichung von $0.1l/100km$ mit 80% Sicherheit erkennen zu können, 4 Versuche pro Stufe benötigt ($\alpha = 5\%$). Hätte sie zusätzlich noch einen zweiten zweistufigen Faktor berücksichtigt, so hätte sie jede Faktorstufe des ersten Faktors automatisch zweimal gemessen — eine einzige Durchführung des gesamten Versuchsplans hätte also gereicht.

Ein weiteres Beispiel soll verdeutlichen, wie wir dies ausnutzen können.

Ein Kollege der Fahrzeugtechnikerin möchte herausfinden, ob und welche der folgenden Faktoren

- A = Anzahl der Speichen (2 Stufen)
- B = Durchmesser (3 Stufen)
- C = Dicke (5 Stufen)

Einfluss auf die wahrgenommene Griffigkeit eines Lenkrades haben. Zu diesem Zweck sollen Probanden die 30 möglichen Faktorkombinationen anhand von 30 Prototypen mit Schulnoten von “1” (sehr gut) bis “6” (ungenügend) bewerten.

Als “praktisch signifikant” wird erachtet, wenn es zu einem Faktor Stufen gibt, die mindestens eine Notenstufe auseinander liegen. Bei einem α -Risiko von 10% soll ein solcher Einfluss mit einer Power von mindestens 80 % erkannt werden.

Ein Blick in Tabelle 4.28 zeigt, dass eine Versuchszahl, die für den fünfstufigen Faktor ausreicht, erst recht für die anderen Faktoren genügt. Nun wird aber bereits bei einer einfachen Durchführung des gesamten Vollfaktorplans jede dieser fünf Stufen bereits sechs mal gemessen, womit sich, je nach Standardabweichung der Benotungen, das in Tabelle 4.29 gezeigte Bild ergibt.

Δ/σ	$\alpha=10\%$				$\alpha=5\%$				$\alpha=1\%$				
	60%	70%	80%	90%	60%	70%	80%	90%	60%	70%	80%	90%	
2 Stufen	0.50	30	39	51	70	41	51	64	86	66	79	96	121
	0.75	14	18	23	32	19	23	29	39	31	36	44	55
	1.00	8	11	14	18	11	14	17	23	18	21	26	32
	1.50	5	6	7	9	6	7	9	11	9	11	13	15
	2.00	3	4	4	6	4	5	6	7	6	7	8	10
	2.50	3	3	3	4	3	4	4	5	5	5	6	7
	3.00	2	3	3	3	3	3	4	4	4	5	5	6
	3.50	2	2	3	3	3	3	3	4	4	4	4	5
	4.00	2	2	2	3	3	3	3	3	3	4	4	4
3 Stufen	0.50	39	49	63	85	51	63	79	103	80	95	113	141
	0.75	18	23	29	38	24	29	36	47	37	43	51	64
	1.00	11	13	17	22	14	17	21	27	22	25	30	37
	1.50	6	7	8	11	7	8	10	13	11	12	14	18
	2.00	4	4	5	7	5	5	6	8	7	8	9	11
	2.50	3	3	4	5	4	4	5	6	5	6	7	8
	3.00	3	3	3	4	3	3	4	5	4	5	5	6
	3.50	2	3	3	3	3	3	3	4	4	4	4	5
	4.00	2	2	3	3	3	3	3	3	3	4	4	4
4 Stufen	0.50	45	56	72	96	59	72	89	115	90	106	126	156
	0.75	21	26	33	43	27	33	40	52	41	48	57	70
	1.00	12	15	19	25	16	19	23	30	24	28	33	40
	1.50	6	7	9	12	8	9	11	14	12	13	16	19
	2.00	4	5	6	7	5	6	7	9	7	8	10	12
	2.50	3	4	4	5	4	4	5	6	6	6	7	8
	3.00	3	3	3	4	3	4	4	5	4	5	5	6
	3.50	2	3	3	3	3	3	3	4	4	4	5	5
	4.00	2	2	3	3	3	3	3	3	3	4	4	4
5 Stufen	0.50	50	62	79	104	65	79	97	125	98	115	136	168
	0.75	23	28	36	47	30	36	44	56	45	52	61	76
	1.00	13	17	21	27	17	21	25	32	26	30	35	43
	2.00	4	5	6	8	5	6	7	9	8	9	10	12
	2.50	3	4	4	5	4	5	5	6	6	6	7	9
	3.00	3	3	4	4	3	4	4	5	5	5	6	7
	3.50	2	3	3	4	3	3	4	4	4	4	5	5
	4.00	2	2	3	3	3	3	3	4	3	4	4	5

Tabelle 4.28 Benötigte Anzahl von Messungen pro Stufe, wenn man bei den gegebenen Parametern Effekte einer gegebenen Stärke mit der entsprechenden Power messen will. **Beispiel:** Haben die Faktoren maximal 2 Stufen und möchte man bei einem α -Risiko von 5% Abweichungen in der Größe der doppelten Standardabweichung ($\frac{\Delta}{\sigma} = 2$) mit einer Power von 80% erkennen, so werden mindestens 6 Messungen je Stufe benötigt.

σ (Notenstufen)	Δ (=1 Notenstufe)	Wiederholungen (je Stufe)	Wiederholungen (Vollfaktorplan)
0.25	$4 \cdot \sigma$	3	1
0.5	$2 \cdot \sigma$	6	1
1	$1 \cdot \sigma$	21	4

Tabelle 4.29 Anzahl benötigter Versuchswiederholungen je Stufe bei der Lenkradstudie, in Abhängigkeit von der unbekannten Standardabweichung der Bewertungen, wenn $\alpha = 10\%$ und $1 - \beta = 80\%$ angenommen wird. Da im Rahmen des Vollfaktorplans jede Stufe ohnehin mindestens sechs mal gemessen wird, ergeben sich die in der letzten Spalte dargestellten Werte.

Fassen wir unter Bezugnahme auf das letzte Beispiel zusammen:

Prinzipien zur Bestimmung des benötigten Versuchsumfangs

1. Man bestimme das akzeptable α -Risiko, also die maximal akzeptable Chance, einen Effekt fälschlicherweise für signifikant zu halten. Gängige Werte sind $\alpha = 10\%$, $\alpha = 5\%$ oder manchmal auch $\alpha = 1\%$.

Für die Lenkradstudie wählen wir exemplarisch $\alpha = 10\%$.

2. Man bestimme, ab wann ein Effekt "praktisch signifikant" wird, ab wann man also Effekte gerne detektieren möchte. Die Größe Δ dieses Effektes bestimmt letztlich den erforderlichen Versuchsumfang.

Im Rahmen der Lenkradstudie wählen wir $\Delta = 1$ (Notenstufe).

3. Basierend auf der (gegebenenfalls im Rahmen einer Pilotstudie) zu schätzenden Standardabweichung σ der Messapparatur ermittle man den normierten "praktisch signifikanten" Effekt $\frac{\Delta}{\sigma}$ und bestimme ferner die gewünschte *Power* des Versuchs, also die Chance, einen Einfluss dieser Stärke zu finden, wenn er existiert.

Unterstellen wir, dass die Probanden alle Lenkräder mit einer Standardabweichung von $\sigma = 1$ (Notenstufe) bewerten, so können wir von $\frac{\Delta}{\sigma} = 1$ ausgehen, einem Effekt, den wir mit 80% Wahrscheinlichkeit erkennen wollen.

4. Man bestimme die maximale Anzahl S von Faktorstufen, die für einen einzelnen Faktor im Rahmen des Versuchsplans vorkommt.

Es gilt natürlich $S = 5$.

5. Zu dieser Anzahl S von Stufen, der Annahmen für α und *Power* $1 - \beta$ und der Größe $\frac{\Delta}{\sigma}$ ermittle man die benötigte Anzahl A von Versuchswiederholungen aus Tabelle 4.28.

Für die erwartete Power von 80% ergibt sich $A = 21$; bei 70% wäre $A = 17$.

6. Man ermittle die Mindestzahl von Messungen, die ohnehin für jede Faktorstufe durchgeführt werden; teilt man die benötigte Anzahl A von Versuchen aus dem letzten Schritt durch diesen Wert, erhält man, gegebenenfalls nach Aufrunden, die für den Gesamtversuchsplan benötigte Anzahl von Wiederholungen.

Jedes Lenkrad wird mindestens $\frac{30}{5} = 6$ mal bewertet. Mit $\frac{21}{6} = 3.5$ ergibt sich, dass man sein Ziel mit 4 Wiederholungen des Gesamtversuchsplans (also $4 \cdot 30 = 120$ Einzelbewertungen) erreicht. Kann man auch mit einer Mindest-Power von 70% leben, kommt man mit 3 Wiederholungen (90 Einzelversuchen) aus, da $\frac{17}{6} = 2.83$ ist.

7. Die Wahl dieser Abschätzungen erfolgte so, dass man für alle Faktoren auf der "sicheren Seite" ist — für manche der Faktoren ist die Power damit unter Umständen sogar deutlich höher als gefordert.

So werden die Stufen des Faktors A ohnehin je 15 mal getestet, so dass man gemäß Tabelle 4.28 für diesen Faktor mit dem einfach durchgeföhrten Versuchsplan auskäme.

4.6 Modellvalidierung

Essentially, all models are wrong, but some are useful.

— George Box, 1987⁷

Nach Abschluss der im Laufe dieses Kapitels dargestellten Arbeiten sind wir in der Lage, ein endgültiges mathematisches Modell des Verhaltens des betrachteten Systems zu formulieren und dieses zur Optimierung der Qualitätsmerkmale zu nutzen. Dazu nutzt man die mit der Varianzanalyse bestätigten Haupteffekte und Wechselwirkungen aus, um die Zielgröße in die gewünschte Richtung zu treiben. Wir können also die ursprüngliche Aufgabe der Systemoptimierung abschließen und diejenigen Einstellungen der betrachteten Parameter ermitteln, die für das gewünschte Systemverhalten sorgen. Zwei Punkte sind allerdings noch offen: Es wurde bereits erwähnt, dass die Voraussetzungen für die Varianzanalyse zu prüfen sind, und letztlich bleibt es auch bei erfolgreicher Validierung dieser Voraussetzungen stets wünschenswert, das mittels eines *mathematischen* Modells ermittelte Wissen über Einflüsse auch *mit dem gesunden Menschenverstand* zu validieren. Der vorliegende Abschnitt schließt diese offenen Punkte.

Residuenanalyse, mehr als statistische Nabelschau.

Seit der Einführung der Grundprinzipien der Varianzanalyse wurde bereits mehrfach darauf hingewiesen, dass die Annahmen, die der Bestimmung der p -Werte zugrunde liegen, auch überprüft werden müssen. Das gesamte Modell basiert nämlich auf der Idee, dass die Fehler auf jeder (Kombination von) Faktorstufe(n) unabhängig voneinander und normalverteilt sind und die gleiche Varianz haben (eine weitere Forderung wurde bisher “unterschlagen”: auch die Eingangswerte sollen unabhängig voneinander sein. Dies ist aber bei orthogonalen Designs stets der Fall — aus diesem Grund wurden ja gerade orthogonale Designs gewählt!).

Inhaltlich mag man dies als Forderung danach interpretieren, dass die Streuung der Messapparatur von den Faktoreinstellungen unabhängig ist — das “Rauschen” soll auf allen Stufen die gleiche Stärke haben, und es soll kein System in den Vorher-sagefehlern geben. Man sollte dies nicht als Problem, sondern als Chance begreifen: wie bereits an anderer Stelle ausführlich dargestellt (vgl. 3.4.2), bietet die Analyse der Modellfehler die Möglichkeit, Ausreißer in den Messdaten zu erkennen und gegebenenfalls verstehen zu lernen, so dass hier durchaus die Chance besteht, mehr über das betrachtete System und sein Verhalten zu lernen.

Mathematisch gesehen handelt es sich um eine wesentliche Voraussetzung, ohne die man die Verteilung der F -Werte nicht bestimmen kann. Kurz gesagt: Die stets auf der Basis der F -Verteilung berechenbaren p -Werte, mit denen über die Signifikanz der Effekte entschieden wird, sind bedeutungslos, wenn man nicht davon ausgehen kann, dass die F -Werte auch wirklich der F -Verteilung folgen.

⁷ Im wesentlichen sind alle Modelle falsch, aber manche sind nützlich.

Dies ist jedoch nur der Fall, wenn die bewusst nun schon so oft zitierten Annahmen auch wirklich gültig sind. Was also bleibt zu prüfen?

Zunächst einmal ist zu beachten, dass die folgenden Betrachtungen auf die *Residuen*, und nicht auf die Messwerte oder die Vorhersagen anzuwenden sind. Dies sind die Modellfehler — die Abweichungen zwischen Modellvorhersagen und “echten” Messwerten. Wie bereits an anderer Stelle dargestellt (vgl. Kap. 3.4.2), gibt es eine Reihe vor allem graphischer Hilfsmittel, die uns nun helfen.

1. Unabhängigkeit der Modellfehler

Um die Unabhängigkeit zu prüfen, hilft ein Diagramm, in dem die Fehler (“residuals”) gegen die Reihenfolge ihrer Entstehung (“run order”) aufgetragen werden, wie exemplarisch in Abbildung 3.13 gezeigt. Dieses Diagramm sollte keinen erkennbaren Trend aufweisen — die Kenntnis eines Wertes darf nicht ermöglichen, das Ergebnis der folgenden Versuche abzuschätzen (wie es bei der Existenz eines Trends der Fall wäre).

2. Normalverteilung der Residuen

Auch in diesem Fall hilft ein Verweis auf Kapitel 3, denn dort wurde bereits ein einfaches Verfahren vorgestellt, mit dem sich die Normalität der Fehler kontrollieren lässt: der sogenannte *Normal Probability Plot*, manchmal auch als *full normal plot* bezeichnet (vgl. Abb. 3.15). Folgen die Residuen einer Normalverteilung, so sollte ungefähr eine gerade Linie entstehen. Dieser graphische Test lässt sich durch formale statistische Tests ergänzen, die erneut auf dem in diesem Kapitel vorgestellten Strickmuster folgen: Mit verschiedenen Tests lässt sich die Abweichung von der Normalverteilungs-Annahme quantifizieren, so dass ein p -Wert ermittelt werden kann, der — falls klein — signifikante Abweichungen signalisiert. Es mag lohnen, in diesem Zusammenhang darauf hinzuweisen, dass wir an dieser Stelle *keine* signifikante Abweichung sehen wollen. War ein kleiner p -Wert in der Elimination von Faktoren Trumpf, so ist es hier genau umgekehrt...

Erwähnenswerte Tests, die in vielen Programmen auch implementiert sind, sind

- der Anderson-Darling-Test
- der Kolmogoroff-Smirnoff-Test
- unter Umständen auch der χ^2 (sprich: Schi-Quadrat) Anpassungstest

Darüber hinaus mag natürlich auch ein Histogramm der Residuen helfen, Abweichungen von der Glockenkurve zu erkennen.

3. Gleiche Varianzen

Trägt man die Fehler gegen die vorhergesagten Werte ab, so entsteht ein “residual versus predicted-plot”, wie in Abbildung 3.14 exemplarisch dargestellt. Wenn man bedenkt, dass sich zu Messungen aus einer Gruppe eine senkrechte Linie durch den Gruppenmittelwert x ergibt, wird klar, woran Abweichungen von der Varianzhomogenität erkennbar werden: Während man ein homogenes waagerechtes Band rund um die x -Achse erwarten würde, wären bei Gruppen mit größerer Varianz deutlich mehr Ausreißer nach oben oder unten zu erwarten. Somit gibt es auch hier eine einfache Methode, Verletzungen der Grundannahmen visuell zu erkennen.

Auch die Varianzhomogenität (Gleichheit der Varianzen in verschiedenen Gruppen) lässt sich im übrigen formal testen: hier sind der sogenannte *Levene-Test* und der *Bartlett-Test* zu nennen, die auch in statistischen Programmen implementiert sind. Wie im Falle der Normalverteilungsannahme würden auch hier kleine p -Werte auf signifikante Abweichungen von der Annahme gleicher Varianzen hinweisen. In diesem Fall bieten sich Transformationen der gemessenen Daten an, wie sie in Kapitel 3.4.4 besprochen wurden — oftmals gelingt es, die Annahmen doch noch zu retten und so das Instrumentarium der Varianzanalyse voll nutzen zu können.

Insgesamt erscheint es empfehlenswert, die benötigten Tests durch die graphische Analyse von vier Diagrammen vorzunehmen:

1. Ein Histogramm der Residuen.

Man erwartet eine um Null zentrierte Glockenkurve.

2. Einen Normal Probability Plot der Residuen.

Es sollte eine gerade Linie entstehen. Bei Zweifeln kann man einen formalen Test „nachlegen“.

3. Ein Diagramm der Residuen gegen die *run order*.

Es sollte kein Trend erkennbar sein.

4. Ein Diagramm der Residuen gegen die vorhergesagten Werte.

Es sollte ein gleichmäßiges, um die x -Achse zentriertes Band ohne Muster zeigen.

Damit ist — sollten alle Tests funktioniert haben — der Mathematiker zufrieden; es sollte aber auch klar geworden sein, dass inhaltlich gewonnen wurde, wenn wir uns vergewissert haben, dass kein systematischer Effekt übersehen wurde, dass also der *nicht* erklärte Teil der Messdatenvariabilität lediglich “Rauschen” ist.

Einen weiteren Test sollte unser Modell jedoch auch noch bestehen: den des gesunden Menschenverstandes...

Der gesunde Menschenverstand legt nahe, dass wir uns abschließend fragen, ob uns die gefundenen Erkenntnisse inhaltlich plausibel vorkommen und ob sie unseren Erwartungen entsprechen:

- Können wir uns erklären, warum manche Faktoren — die ja anfangs ausgewählt wurden, da man an ihren Einfluss glaubte — am Ende vielleicht doch nicht signifikant sind? Ist der Einfluss der als signifikant erkannten Faktoren plausibel? Was ist mit der Plausibilität der Wechselwirkungen?
- Erscheint die Stärke der Haupteffekte sinnvoll?
- Wie ist die Plausibilität der Vorzeichen? Gibt es hier Überraschungen?
- Hilft uns das Ergebnis, das ursprüngliche (Optimierungs-)Problem zu lösen?

Dabei lasse man sich durch etwaige Zweifel nicht entmutigen: es gibt stets Neues zu lernen — und auch das Durchführen mittels DoE geplanter Experimente bleibt ein iteratives Geschäft.

4.7 Zusammenfassung: Von den Daten zum Modell in 7 Schritten

The statistician who supposes that his main contribution to the planning of an experiment will involve statistical theory, finds repeatedly that he makes his most valuable contribution simply by persuading the investigator to explain why he wishes to do the experiment...

— Gertrude Cox, 1951⁸

- Datencheck.** Wie sieht der real abgefahrene Versuchsplan aus? Gibt es offensichtliche Fehler? Lücken? War der Versuchsplan wirklich orthogonal?
- Modellbildung** durch schrittweise Elimination von Faktoren und Wechselwirkungen. Ausgehend von einem mehr oder weniger “vollen” Modell eliminiere man schrittweise diejenigen Parameter, deren Einfluss auf die Zielgröße nicht nachgewiesen werden kann; man entferne also, beginnend mit den Mehrfachwechselwirkungen, alle Größen mit großem p -Wert, bis man bei einem Modell angelangt ist, das lediglich signifikante Eingangsgrößen enthält und eine brauchbare Erklärungskraft R^2 hat. Gelingt es nicht, zu “brauchbaren” Werten von R^2 zu gelangen, so hat man offenbar signifikante Einflussfaktoren übersehen.
- Residuenanalyse.** Ist man bei einem endgültigen Modell angelangt, muss man die Annahmen prüfen: Sind die Residuen unabhängig und normalverteilt? Ist die Annahme homogener Varianzen gerechtfertigt? Neben der mathematischen Notwendigkeit, dies zu prüfen (denn sonst sind die p -Werte, mit denen die Signifikanz der Effekte nachgewiesen werden soll, Makulatur), gibt es auch einen durchaus erwähnenswerten inhaltlichen Nutzen, wenn sichergestellt wurde, dass man keinen systematischen Zusammenhang übersehen hat.
- Liegt das Modell nach seiner Validierung fest, so kann man es in Form einer Gleichung darstellen, die die signifikanten Effekte aufnimmt. Basierend auf dieser Gleichung kann man das System im Hinblick auf die Zielgröße **optimieren**.
- Testläufe.** Auch die Durchführung bestätigender Testläufe mit den optimalen Einstellungen dient letztlich der Modellvalidierung. Man hat nur gewonnen, wenn die ermittelte optimale Einstellung zum Erfolg führt ... !
- Schließlich gehört auch die **Dokumentation** der Ergebnisse zum professionellen Arbeiten. Nur so wird sichergestellt, dass unsere Erkenntnisse nachvollziehbar bleiben und ins *corporate knowledge* eingehen können, und nur so kann man auch in Zukunft nachweisen, dass handwerklich sauber gearbeitet wurde (z.B. durch den Nachweis, dass die Modelle validiert wurden).
- Team Event.** War ein ganzes Team an Planung, Durchführung und Auswertung der Experimente beteiligt, stellt ein abschließendes Team Event sicher, dass man auch in Zukunft mit Unterstützung rechnen kann, wenn man sie benötigt. Würfeln in der Gaststätte bietet sich an ...

⁸ Der Statistiker, der annimmt, dass sein Hauptbeitrag zur Planung von Experimenten mit statistischer Theorie zu tun hat, findet oft, dass sein wertvollster Beitrag darin besteht, den Forscher dazu zu bringen zu erklären, warum er das Experiment ausführen will... .

Literaturverzeichnis

1. Box, G.: *Do Interactions Matter?* Tech. rep., Center for Quality and Productivity Improvement, University of Wisconsin - Madison (1989) 88
2. Box, G.: *Must We Randomize Our Experiment?* Tech. rep., Center for Quality and Productivity Improvement, University of Wisconsin - Madison (1989) 89
3. Box, G.: *Quality Improvement: An Expanding Domain for the Application of Scientific Method.* Phil. Transact. Royal Soc. London, Series A **327**, pp. 617–630 (1989) 95
4. Box, G., Friends: *Improving Almost Anything: Ideas and Essays.* John Wiley and Sons, NJ (2006) 92
5. Box, G., Hunter, W.G., Hunter, J.S.: *Statistics for Experimenters. An Introduction to Design, Data Analysis, and Model Building.* John Wiley and Sons, NJ (1978) 92
6. Fisher, R.A.: *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh (1925). URL <http://psychclassics.yorku.ca/Fisher/Methods/index.htm>. (abgerufen 11/2016) 91
7. Fisher, R.A.: *The Arrangement of Field Experiments.* Journal of the Ministry of Agriculture of Great Britain **33**, pp. 503–513 (1926) 90
8. Fisher, R.A.: *The Design of Experiments.* Oliver and Boyd, Edinburgh and London (1935) 1, 91, 104
9. Hellstrand, C.: *The Necessity of Modern Quality Improvement and Some Experience With its Implementation in the Manufacture of Rolling Bearings.* Tech. rep., Center for Quality and Productivity Improvement, University of Wisconsin - Madison (1989) 88
10. Montgomery, D.C.: *Design and Analysis of Experiments.* John Wiley and Sons, Hoboken, NJ (2001/2009) 90, 129, 234
11. Student: *The Probable Error of a Mean.* Biometrika **VI**(1), pp. 1–25 (1908) 91

Kapitel 5

Varianten der statistischen Versuchsplanung

5.1 Einleitung

Nur in wenigen Fällen besteht die Aufgabe darin, ein einziges Qualitätsmerkmal zu optimieren. Allein durch den Kostendruck kommt im Regelfall neben der Systemleistung ein weiterer Aspekt hinzu. Vielfach tauchen mehrere Leistungskriterien auf, die nur in Ausnahmefällen so miteinander korrelieren, dass die jeweiligen optimalen Einstellungen identisch sind. Die gleichzeitige Optimierung mehrerer Qualitätsmerkmale ist nicht Bestandteil der klassischen statistischen Versuchsplanung. Allerdings liefert die statistische Versuchsplanung eine hervorragende Ausgangsposition und lässt sich problemlos mit Optimierungsverfahren koppeln. Auswerteprogramme tragen dem Rechnung und haben die “Multiple-Response-Optimisation” in ihr Standardrepertoire aufgenommen. Dieses Kapitel gibt einen kompakten Überblick zu diesem Thema, ohne auf die Hintergründe der numerischen Mathematik weiter einzugehen. Fortgeschrittene Leser finden im Kapitel *Optimierung* weitere Informationen.

Technische Systeme werden im Alltagsgebrauch anders beansprucht als unter idealen Bedingungen im Labor. Für den Endkunden ist die optimale Leistungsentfaltung unter Laborbedingungen irrelevant, nur das Systemverhalten im Alltag zählt. Hochgezüchtete Systeme verhalten sich immer etwas “nervös” in Bezug auf veränderte Randbedingungen. Auf der anderen Seite zwingt der Kostendruck den Konstrukteur dazu, unnötige Sicherheitsreserven abzubauen und das System genau den Anforderungen anzupassen. Es besteht somit die Notwendigkeit, den Alltagsgebrauch durch den Kunden im Labor des Konstrukteurs nachzubilden[5].

Die statistische Versuchsplanung ist geradezu prädestiniert, um diese Aufgabe zu übernehmen, da es letztlich darum geht, das Systemverhalten in Abhängigkeit vieler Variablen zu untersuchen. Es bedarf nur einer kleinen Erweiterung der bereits vorgestellten Methodik, um einen generischen Ansatz für die Untersuchung der “Robustness” herzuleiten. Parameterdesign und Toleranzdesign sind geradezu klassische Erweiterungen der statistischen Versuchsplanung. Beide sind sehr eng mit GENICHI TAGUCHI verbunden [2]. TAGUCHI hat bereits in den 50er Jahren des

vorigen Jahrhunderts diese Verfahren erfolgreich beim Wiederaufbau der japanischen Industrie eingesetzt und dadurch nachhaltig zur Verbreitung der statistischen Versuchsplanung beigetragen.

5.2 Umgang mit mehreren Qualitätsmerkmalen

5.2.1 Multiple-Response-Optimisation

Für die Optimierungsrechnung ist eine skalare Bewertungsfunktion erforderlich, die gleichzeitig alle Qualitätsmerkmale berücksichtigt. Diese Aufgabe erscheint schwieriger als sie ist. Den ersten Arbeitsgang erledigt bereits die traditionelle statistische Versuchsplanung, denn für jedes Qualitätsmerkmal liegt eine Beschreibungsfunktion vor. Im zweiten Arbeitsgang gilt es, die Beschreibungsfunktionen miteinander zu koppeln, um aus mehreren Ergebniswerten einen Wert der alles entscheidenden Bewertungsfunktion zu berechnen. Die Kopplung erfolgt über Rampenfunktionen. Durch diese Transformation werden die Qualitätsmerkmale dimensionslos, ihre Gewichtung lässt sich einstellen und Ausreißer können das Gesamtergebnis nicht mehr verzerrn. Jede Rampe kann nur Werte zwischen 0 und 1 annehmen. Der Wert 0 entspricht einem sehr schlechten Ergebnis, 1 ist für ein sehr gutes Ergebnis vorgesehen. Die Rampenfunktion selbst ist normalerweise eine lineare Abbildung des Qualitätsmerkmals, im Bedarfsfall ist aber auch eine nichtlineare Abbildung möglich. Grundsätzlich gibt es drei Basisvarianten: Minimierungsaufgabe, Maximierungsaufgabe, Zielwert treffen. Durch passende Definition der Eckpunkte bei den Rampenfunktionen werden alle Varianten auf eine Aufgabe zurückgeführt: Maximierung des Rampenwertes.

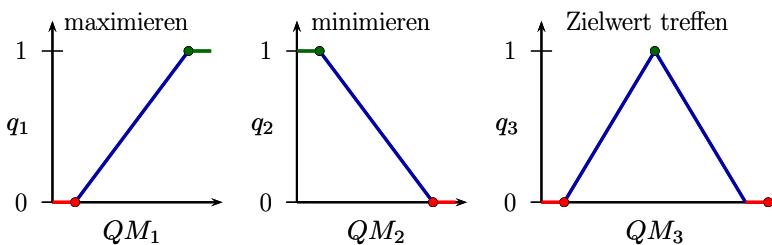


Abb. 5.1 Rampenfunktionen zur Multiple-Response-Optimisation. Jedes Qualitätsmerkmal QM wird auf einen dimensionslosen Kennwert q abgebildet, der im Wertebereich zwischen 0 und 1 liegt. Es gibt drei Grundaufgaben: maximieren, minimieren und Zielwert treffen. Die Eckpunkte muss der Anwender definieren. Sie kennzeichnen, ab wann ein völlig akzeptables, bzw. völlig unakzeptables Ergebnis vorliegt.

Der erste Eckpunkt beschreibt sozusagen die Schmerzgrenze, also ein Ergebnis des Qualitätsmerkmals, welches völlig inakzeptabel ist. Noch schlechtere Er-

gebnisse machen keinen Unterschied mehr. Dieser Eckpunkt könnte zum Beispiel einen gesetzlich zulässigen Grenzwert darstellen, liegt man darüber, ist das Produkt unbrauchbar. Die Rampenfunktion liefert dafür den Wert 0. Bei Zielwertaufgaben gibt es zwei Eckpunkte, auf beiden Seiten des Optimums. Der verbleibende Eckpunkt definiert das erhoffte Ergebnis. Hier ist Bescheidenheit ratsam, denn völlig überzogene Erwartungen an jedes Qualitätsmerkmal führen zu einer schlecht konditionierten Bewertungsfunktion. An dieser Stelle lässt sich auch eine Gewichtung der Qualitätsmerkmale unterbringen. Bei moderaten Anforderungen für die weniger wichtigen Qualitätsmerkmale hat die Optimierungsrechnung automatisch mehr Spielraum, um in Richtung der wichtigen Qualitätsmerkmale zu arbeiten.

Sei y der Wert des Qualitätsmerkmals, mit y_1 , bzw y_2 zur Definition der Eckpunkte, dann ergibt sich die Rampenfunktion durch eine Fallunterscheidung.

$$q = \begin{cases} 0 & \forall y \leq y_1 \\ \frac{y-y_1}{y_2-y_1} & \forall y_1 < y < y_2 \\ 1 & \forall y \geq y_2 \end{cases} \quad (5.1)$$

Die Rampendefinitionen für die Minimierungsaufgabe und das Erreichen eines Zielwertes ergeben sich analog dazu.

Die Multiple-Response-Optimisation findet im Rahmen der Auswertung, also nach der Versuchsreihe statt. Veränderungen der Eckpunkte zur Definition einer passenden Rampenfunktion kosten nur sehr wenig Zeit und erfordern keine neuen Versuchsdaten. Es macht also durchaus Sinn, mehrere Eckpunkte auszuprobieren und dabei das Systemverhalten kennenzulernen. Welches Optimierungspotential ein System besitzt, hängt nicht zuletzt auch davon ab, wie stark sich die einzelnen Qualitätsmerkmale widersprechen.

Die Kopplung zu einer Wunschfunktion besteht in der einfachen Multiplikation der Funktionswerte aller Rampen. Die Wunschfunktion (desirability function) nimmt also ebenfalls Werte zwischen 0 und 1 an. Bereits ein unakzeptables Ergebnis eines einzelnen Qualitätsmerkmals drückt die Gesamtbewertung auf 0. Der Wert 1 hingegen wird nur dann erreicht, wenn gleichzeitig bei allen Qualitätsmerkmalen sehr gute Ergebnisse zu erwarten sind. Die Optimierung der Wunschfunktion kann mit beliebigen Solvieren erfolgen. Ein einfacher Simplexalgorithmus leistet bereits gute Dienste, denn die Zielfunktion ist kompakt und in der Regel gut konditioniert. Es empfiehlt sich allerdings, mehrere Startwerte vorzusehen, um wirklich das globale Optimum zu finden. n_q bezeichnet die Zahl der zu optimierenden Qualitätsmerkmale.

$$D = (q_1 q_2 \dots q_{n_r})^{\frac{1}{n_q}} = \left(\prod_{i=1}^{n_q} q_i \right)^{\frac{1}{n_q}} \quad (5.2)$$

Neben der Wahl der Eckpunkte gibt es zwei weitere Möglichkeiten, die Gewichtung der Qualitätsmerkmale untereinander zu beeinflussen. Die erste Variante sieht eine nichtlineare Transformation bei Bildung der Rampenfunktion vor. Der Zielbereich bleibt gleich, aber die Krümmung ändert sich. Dies erfolgt durch einfaches

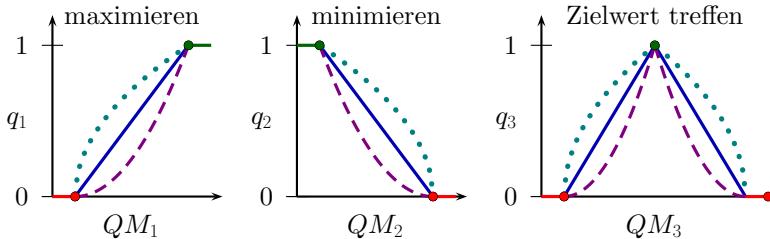


Abb. 5.2 Nichtlineare Gewichtung der Rampenfunktionen. Die Kennwerte können mit einem beliebigen Exponenten auf nichtlineare Kennwerte transformiert werden. Exponenten unter 1 (gepunktet) liefern einen degressiven Verlauf, tendenziell mit höheren Werten. Exponenten über 1 liefern einen progressiven Verlauf und die Bewertung wird insgesamt strenger.

Exponieren. Exponenten unter 1 liefern einen sanfteren Übergang in den angestrebten Bereich und schwächen das Kriterium ab. Exponenten über 1 wirken progressiv und erzeugen bereits knapp neben dem zweiten Eckpunkt niedrige Werte. Das Kriterium geht somit stärker in die Gesamtberechnung ein. Für jedes Qualitätsmerkmal kann ein eigener Exponent gewählt werden. Die folgende Gleichung zeigt dies am Beispiel einer Maximierungsaufgabe.

$$q = \begin{cases} 0 & \forall y \leq y_1 \\ \left(\frac{y-y_1}{y_2-y_1}\right)^s & \forall y_1 < y < y_2 \\ 1 & \forall y \geq y_2 \end{cases} \quad (5.3)$$

Die zweite Variante geht von individuellen Exponenten für jeden Rampenwert bei der Bildung der Wunschfunktion aus, also nachdem die Abbildung auf q erfolgt ist. Beide Verfahren sind kombinierbar. In der Praxis erzeugt dies schnell ein heilloses Chaos, weil sich der Anwender nicht ohne Weiteres die Wirkung derartiger nichtlinearer Abbildungen vorstellen kann. An dieser Stelle sei empfohlen, zunächst über die Wahl der Eckpunkte eine Gewichtung der Qualitätsmerkmale vorzunehmen, also bei linearen Abbildungen zu bleiben. Reicht dies nicht aus, ist die Variante 1 vorzuziehen, weil die Rampenwerte für jedes Qualitätsmerkmal im Zweifelsfall besser nachzuvollziehen sind.

Als einfache Fallstudie dient der Rasensprenger mit den drei Faktoren: horizontaler Düsenwinkel $15^\circ < \alpha < 45^\circ$, tangentialer Düsenwinkel $0^\circ < \beta < 30^\circ$ und Düsenquerschnitt $2mm^2 < A_q < 4mm^2$. Der Vollfaktorplan mit acht Versuchsläufen versorgt nur ein lineares Modell (inclusive Wechselwirkungen), trotzdem ist die Wunschfunktion nichtlinear. Das gemeinsame Optimum für die drei Qualitätsmerkmale liegt daher nicht in einer der Ecken, sondern bei $\alpha = 25,3^\circ$, $\beta = 0^\circ$, $A_q = 2,75mm^2$.

Natürlich lässt sich die Methode auch auf komplexere Fälle anwenden, zum Beispiel den Rasensprenger mit 8 Faktoren, einem 500er Versuchsplan mit Space-Filling Design und verschiedenen Beschreibungsmodellen. Das riesige Feld ermöglicht problemlos die elegante Umschiffung der ersten Klippe: unterschiedliche Startwerte. Allzu leicht landet der Suchalgorithmus in einem lokalen Minimum. Je mehr

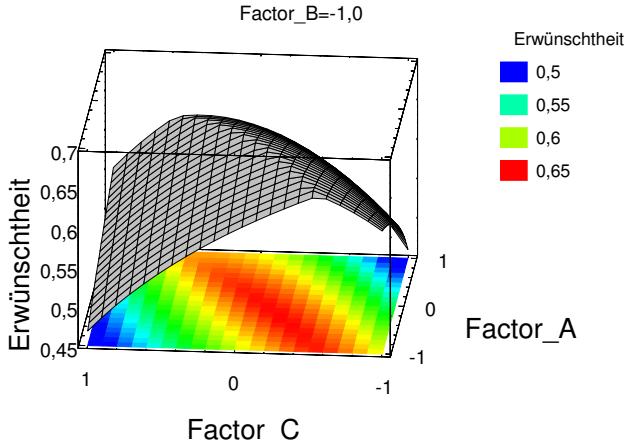


Abb. 5.3 Wunschfunktion für die gleichzeitige Optimierung aller Qualitätsmerkmale in Abhängigkeit von α und A_q . β steht auf -1 , alle anderen Parameter auf 0.

QM	Einh.	Grenzen	Typ	Optimum
Drehzahl	$\left[\frac{1}{s}\right]$	2 5	max	5,285
Reichweite	$[m]$	2 7	max	5,615
Verbrauch	$\left[\frac{l}{min}\right]$	1 10	min	5,733

Tabelle 5.1 Definition der Rampenfunktionen und resultierendes Optimum. Der Bestätigungslauf der postulierten Einstellung liefert hier bessere Werte als die Vorhersage.

Dimensionen der Faktorraum hat, desto wahrscheinlicher tritt dieser Effekt auf. Abhilfe schafft eine systematische Suche, die viele Startpunkte verwendet und sich dann für die beste Kombination entscheidet, sozusagen das beste lokale Optimum. Als Standardoption bietet es sich an, jeden Punkt des Versuchsplans gleichzeitig als Startwert zu verwenden. Dies läuft automatisch ab und bei mittelgroßen Feldern in Sekundenschnelle, auch mit handelsüblichen Rechnern. Damit erreicht man die nachfolgenden (Best-) Werte:

Qualitätsmerkmal Bez.	Einh. u. G. o. G.	Rampendef.		Beschreibungsmodell		
		Typ	kubisch	quadratisch	linear	
Drehzahl	$\left[\frac{1}{s}\right]$	2 8	max	8,012	7,820	7,890
Reichweite	$[m]$	2 7	max	5,537	5,545	5,406
Verbrauch	$\left[\frac{l}{min}\right]$	1 10	min	4,998	4,891	4,942

Tabelle 5.2 Definition der Rampenfunktionen für das komplexe Rasensprengerbeispiel und resultierende Optima für verschiedene Beschreibungsmodelle. Das kubische Modell und das quadratische Modell treffen vergleichbar gute Optima. Das lineare Modell mit Wechselwirkungen fällt etwas ab, schneidet aber immer noch überraschend gut ab. Alle tabellierten Ergebnisse wurden aus Bestätigungsläufen gewonnen.

Der Aufwand für die jeweiligen Beschreibungsmodelle unterscheidet sich natürlich erheblich. Während das kubische Modell für acht Faktoren nach ca. 500 Läufen verlangt, kann das quadratische Modell bereits mit einem 129er Latin Hypercube Design bestens versorgt werden. Das genügsame lineare Modell mit Wechselwirkungen braucht nur ein zweistufiges 64er Feld mit Auflösungsstufe V. Bei schnellen Computermodellen spielt die Feldgröße eine untergeordnete Rolle. Hier wird man im Zweifelsfall nicht kleckern, sondern klotzen. Bei aufwendigen CAE-Modellen oder gar bei physikalischen Tests sind die Möglichkeiten begrenzt. Daher ist es gut zu wissen, dass in vielen Fällen auch ein einfaches BeschreibungsmodeLL hervorragend arbeiten kann.

Symbol	Größe Einheit	BeschreibungsmodeLL			verbrauchsoptimiert
		kubisch	quadratisch	linear	
α	°	30,3	30,3	39	32,2
β	°	18,8	0,6	6,5	28,2
A_g	mm ²	2,03	2	2	2
d	mm	186	161	200	188
M_{Rl}	Nm	0,013	0,011	0,01	0,012
M_{Rf}	Nm s	0,01	0,01	0,01	0,01
p_{in}	bar	2	2	2	1
d_{zul}	mm	10	9,66	9,52	8,81
Drehzahl	[$\frac{1}{s}$]	8,012	7,820	7,890	3,421
Reichweite	[m]	5,537	5,545	5,406	4,417
Verbrauch	[$\frac{l}{min}$]	4,998	4,891	4,942	3,424

Tabelle 5.3 Optimierte Rasensprengereinstellung für die drei unterschiedlichen Beschreibungsmodelle. Das Ausgangsfeld ist in allen Fällen gleich (Space Filling Design, 500 Einstellungen). Die Zahl der Modellkonstanten wurde entsprechend reduziert. Das Lineare Modell enthält auch Zweifach-Wechselwirkungen. Beim kubischen Modell fällt auf, dass für M_{Rl} nicht der niedrigste Wert favorisiert wird. Dies liegt daran, dass der optimale Wert für die Drehzahl bereits erreicht wurde und der Minimalwert für M_{Rl} gleichzeitig den Verbrauch erhöhen würde, also insgesamt ungünstiger wäre. Die verbrauchsoptimierte Einstellung resultiert aus einer anderen Rampendefinition, mit strengerem Anforderungen an den Verbrauch und gelockerten Anforderungen für die übrigen Qualitätsmerkmale.

Nun stellt sich die Frage, ob die berechneten Optima auch bei vergleichbaren Faktoreinstellung erreicht wurden. Hier zeigen sich gleichermaßen Chancen und Risiken der Methode. Zum einen stellt man schnell fest, dass es ähnlich gute Ergebnisse an völlig unterschiedlichen Positionen im n_f -dimensionalen Faktorraum gibt. Einige Programme bieten sogar Listen der besten Kombinationen an, also nicht nur einen Optimalwert. Das Risiko besteht darin, sich auf einen Punkt festzulegen, der nur aufwendig zu realisieren ist, obwohl möglicherweise ein fast ebenso guter Punkt existiert, der viel kostengünstiger wäre. Daraus erwächst die Chance, durch kleine Planspiele mit dem BeschreibungsmodeLL, den Faktorraum nach einer preiswerten Lösung abzugrasen. Diese Übung geht sehr schnell und verlangt keine neuen Versuche, weil alles mit dem BeschreibungsmodeLL durchgeführt werden kann. Idealerweise setzen sich Systemfachmann und Statistiker zusammen, denn der Systemfachmann kann schnell beurteilen, welche Variante besonders günstig ist. Oft reicht

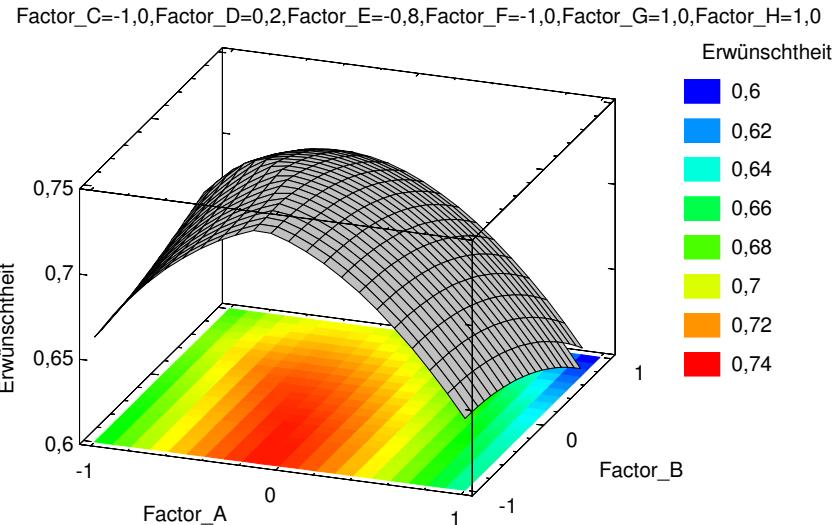


Abb. 5.4 Wunschkfunktion für die gleichzeitige Optimierung aller Qualitätsmerkmale. Gezeigt ist die Abhängigkeit von den Düsenwinkeln, bei optimaler Einstellung der übrigen Faktoren. Alle Einheiten sind kodiert, also im standardisierten Wertebereich von -1 bis 1.

es schon, den Suchraum etwas einzuzgrenzen, um die teuren Varianten auszuschließen. Mitunter wird es bei stärkeren Einschränkungen nötig, die Rampenfunktionen leicht zu modifizieren, um die Aufgabe lösbar zu halten. Überhaupt ist es immer ratsam, verschiedene Gewichtungen der Qualitätsmerkmale zu eruieren. Durch die Rampendefinitionen lässt sich dies leicht bewerkstelligen. Sehr schnell gelangt man dadurch zu einem guten Systemverständnis und lotet risikofrei die Grenzen aus. In der Praxis ist dieser Lernprozess sehr spannend und belohnt alle Beteiligten für die Mühen bei der Datengewinnung.

5.2.2 Sequentielle Methode und Ersatzgrößen

Ohne Auswerteprogramme bleiben zwei alternative Methoden zur gleichzeitigen Behandlung mehrerer Qualitätsmerkmale. Die sequentielle Methode geht von einer klaren Reihenfolge der Qualitätsmerkmale in Bezug auf ihre Bedeutung aus. Das wichtigste Merkmal wird zuerst optimiert und legt für einige der Faktoren bereits die Einstellung fest. Mit Hilfe der verbleibenden Faktoren erfolgt die Optimierung des nächsten Qualitätsmerkmals. Mehr als zwei Qualitätsmerkmale lassen sich auf diese Weise in der Praxis nicht verarbeiten.

Eine gleichzeitige Optimierung mehrerer Qualitätsmerkmale ergibt sich automatisch, wenn vor der Modellbildung aus den Qualitätsmerkmalen eine zusammenfassende Ersatzgröße gebildet wird [4]. Dies kann ein real existierendes Be-

wertungskriterium sein, wie zum Beispiel die aus dem HIC-Wert und der Brustbeschleunigung des Dummies gebildete US-NCAP Bewertung von Rückhaltesystemen bei Fahrzeugen. Für das Systemverständnis kann diese Methode sehr hilfreich sein, denn alle Ergebnisdarstellungen (Effektdiagramm, Wechselwirkungsdiagramm, ...) beziehen sich unmittelbar auf die Gesamtbewertung. Nachteilig ist jedoch die schnell eintretende Nichtlinearität der Gesamtbewertung, bedingt durch die Verknüpfung der Qualitätsmerkmale. In der Praxis ist die Multiple-Response-Optimisation leistungsfähiger, weil die Verknüpfung der Qualitätsmerkmale erst nach der Bildung der Beschreibungsmodelle erfolgt.

5.2.3 Principal Component Analysis

Die Principal Component Analysis (PCA) ist ein geeignetes Verfahren, um die Abhängigkeit der Qualitätsmerkmale untereinander zu analysieren. Oft korrelieren mehrere Qualitätsmerkmale miteinander und der Ergebnisraum hat weniger unabhängige Freiheitsgrade, als zunächst angenommen. Die PCA stammt aus der Strukturanalyse und wird sehr erfolgreich eingesetzt, um die Kopplung von Knotenbewegungen aufzudecken. Auch die Psychologie nutzt die PCA, um Kopplungen von Assoziationen bei Kundenbefragungen herzustellen. Mit der statistischen Versuchsplanung hat die PCA zunächst nichts zu tun, die beiden Methoden lassen sich aber hervorragend kombinieren. Die statistische Versuchsplanung liefert bei minimalem Versuchsaufwand die bestmögliche Abdeckung des Faktorraums (der Eingangsgrößen) und damit gut konditionierte Eingangsdaten für die PCA. Die anschließende PCA bringt die Qualitätsmerkmale miteinander in Verbindung, was die statistische Versuchsplanung ihrerseits nicht leisten kann, da sie nur die Abbildung der Eingangsgrößenvariation auf jeweils ein Qualitätsmerkmal erzeugt. Die Korrelation der Qualitätsmerkmale untereinander reduziert die Zahl der unabhängigen Freiheitsgrade im Ergebnisraum. Dadurch wird das Optimierungsproblem einfacher. Bereits vor der Multiple-Response-Optimisation zeigt sich aufgrund der PCA, ob die Qualitätsmerkmale gegeneinander arbeiten oder unter einen Nenner zu bringen sind.

Die PCA ist eine Hauptachsentransformation im Ergebnisraum der Qualitätsmerkmale. Jeder Versuch liefert einen Punkt in diesem mehrdimensionalen Raum. Die PCA berechnet passende Ellipsoide, um die Punktwolke bestmöglich einzuschließen. Die Ausdehnung der Achsen entspricht den Eigenwerten, die Richtung der Ellipsoidachsen zeigen die Eigenvektoren. Zusätzlich gibt die PCA Aufschluss darüber, welcher Anteil der Gesamtvariation durch die entsprechende Zahl der Ellipsoid-Dimensionen erklärt werden kann. Der Scree-Plot stellt die Eigenwerte in fallender Reihenfolge dar. Große Eigenwerte (über 1) kennzeichnen wichtige Dimensionen, kleine Eigenwerte deuten darauf hin, dass in der betroffenen Dimension keine große Varianz mehr stattfindet. Der Bi-Plot stellt die Richtung der Qualitätsmerkmale im Koordinatensystem der stärksten Hauptachsen dar. Hier zeigt sich sehr schnell, wie die Qualitätsmerkmale zusammenhängen. Oft erwächst aus diesen Darstellungen ein sehr gutes Systemverständnis. Eine Multiple-Response-Optimisation

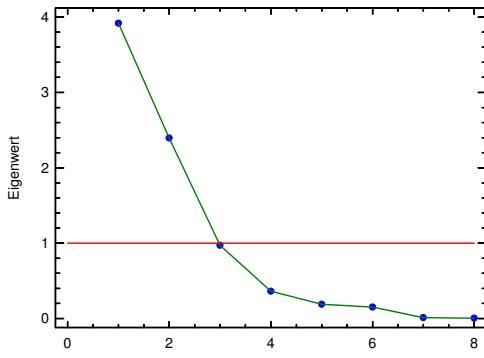


Abb. 5.5 Fallbeispiel Rasensprenger. Scree-Plot für acht Ergebnisgrößen. Zusätzlich zu den drei Qualitätsmerkmalen wurden noch aufgezeichnet: Antriebsmoment, absolute Tropfengeschwindigkeit, deren vertikale Komponente, relative Tropfengeschwindigkeit (jeweils beim Düsenaustritt) und resultierende Flugzeit der Wassertropfen. Zwei der acht Dimensionen erklären bereits 80 % der Ergebnisvariation

der in Hauptachsen transformierten Qualitätsmerkmale ist möglich, wird aber in der Praxis selten anschaulich sein, wenn man den Hauptachsen keine physikalische Bedeutung geben kann. Gelingt jedoch eine derartige Zuordnung, dann ist die PCA sehr wertvoll, um die für das System wirklich entscheidenden Qualitätsmerkmale zu finden.

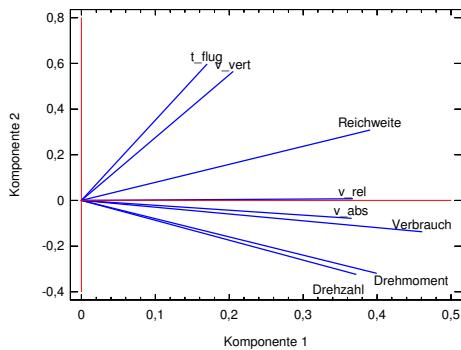


Abb. 5.6 Fallbeispiel Rasensprenger. Bi-Plot für acht Ergebnisgrößen. In gewisser Weise kann man eine Analogie zum Tauziehen herstellen. Alle Qualitätsmerkmale ziehen in fast die gleiche Richtung. Da jedoch der Verbrauch minimiert und Reichweite + Drehzahl maximiert werden sollen, arbeiten die Merkmale gegeneinander. Die zweite Komponente zeigt den Konflikt zwischen vertikaler Tropfengeschwindigkeit und horizontaler Tropfengeschwindigkeit.

5.3 Robustes Design

5.3.1 Parameterdesign

5.3.1.1 Parameterdiagramm

Das Parameterdiagramm (oder auch: P-Diagramm) ist der Dreh- und Angelpunkt des Parameterdesigns. Alle weiteren Schritte bauen darauf auf. Selbst wenn keine

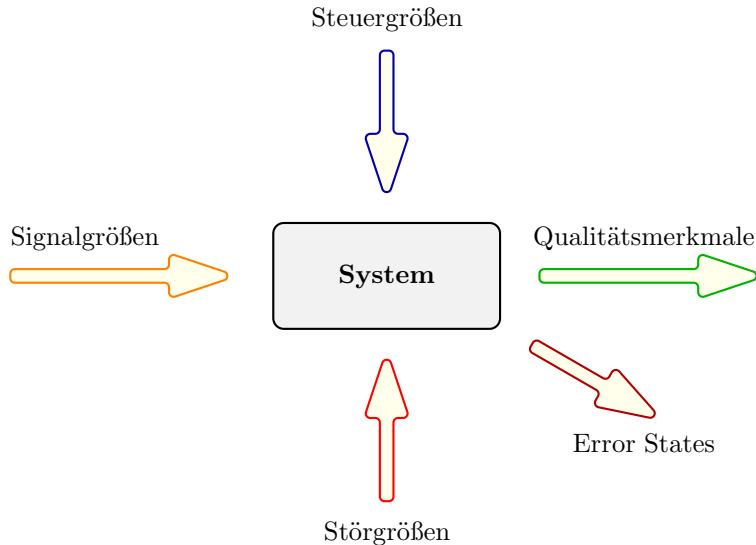


Abb. 5.7 Parameterdiagramm. Steuer-, Signal- und Störgrößen nehmen Einfluss auf das System. Ein Teil der Ergebnisse ist erwünscht, die Qualitätsmerkmale. Error States kennzeichnen die unerwünschten Ergebnisse.

konkrete Versuchsreihe geplant ist, kann das Parameterdiagramm einen wertvollen Beitrag zum Systemverständnis leisten. Die bereits bei den Grundbegriffen eingeführte schematische Sichtweise wird nun erweitert. Beinflussbare Konstruktionsparameter heißen Steuergrößen. Nicht, oder nur mit unverhältnismäßig großem Aufwand beinflussbare Größen teilen sich in zwei Gruppen auf. Signalgrößen stecken den Betriebsbereich des Systems ab, während Störgrößen die in der Praxis unkontrollierbaren Einwirkungen auf das System beschreiben. Außer den gewünschten Ergebnissen werden nun auch die unerwünschten Ergebnisse betrachtet, die sogenannten "error states". Geblieben ist die abstrakte Beschreibung des Systems als "graue Kiste" mit genau definierten Systemgrenzen, einer klaren Vorstellung von der gewünschten Systemleistung und einer im Zweifelsfall eher überdimensionierten Liste aller Einflussgrößen.

Die in der Literatur gängige Darstellung des Parameterdiagramms hat eine andere Anordnung als die in diesem Buch gezeigte Darstellung. Üblicherweise erscheinen die Störgrößen oben und die Steuergrößen unten. In der konkreten Anwendung hat dies jedoch Nachteile bei der Platzaufteilung, weil typischerweise die Zahl der Steuergrößen die Zahl der untersuchten Störgrößen übersteigt. Außerdem verbessert es die Lesbarkeit, wenn zusammenhängende Aspekte auch visuell gruppiert sind. Unerwünschte Ergebnisse stehen im kausalen Zusammenhang mit den Störgrößen und befinden sich im unteren Bereich des Parameterdiagramms. Steuergrößen werden hingegen genutzt, um ein möglichst positives Ergebnis zu erhalten, also Optimierung der Systemleistung bei gleichzeitiger Vermeidung der unerwünschten Ergebnisse.

Die Strategie des Parameterdesigns besteht darin, die Variation von Störgrößen und Signalgrößen zuzulassen, also weder dagegen anzukämpfen, noch den Betriebsbereich des Systems einzuschränken. Natürlich setzt dies eine genaue Analyse des Systemverhaltens voraus, denn nur wenn die Steuergrößen richtig eingestellt sind, richten die Störgrößen keinen großen Schaden an.

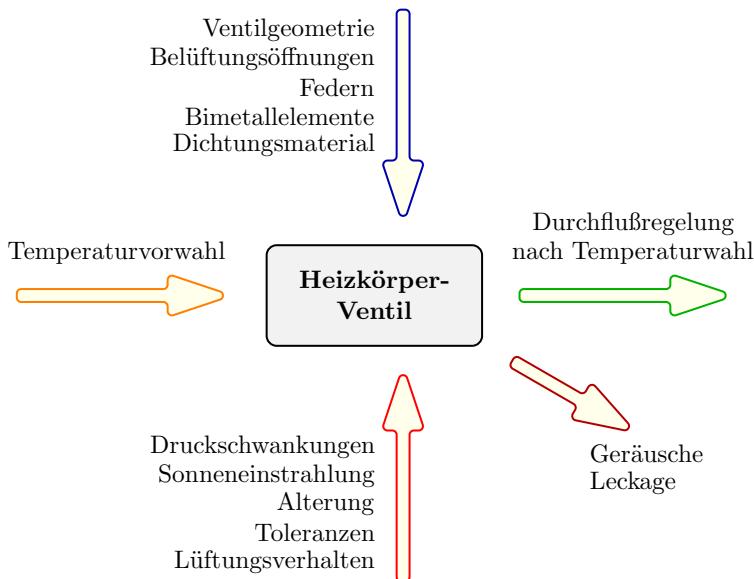


Abb. 5.8 Parameterdiagramm am Beispiel eines Thermostatventils für Heizkörper. Geräusche und Leckage können hier als unerwünschte Ergebnisse auch bei guter Erfüllung des Qualitätsmerkmals auftreten.

Die Abgrenzung der Einflussgrößen voneinander erfordert immer ein wenig Erfahrung und eine klare Definition der jeweiligen Kategorien. Steuergrößen gehören immer zum System, liegen also innerhalb der Systemgrenzen. Sie sind vom Konstrukteur beeinflussbar und es besteht im konkreten Fall auch die Möglichkeit, die optimale Einstellung im Rahmen eines Variationsspielraums vorzugeben. Störgrößen liegen außerhalb des Systems und unterliegen im Alltagsgebrauch einer nicht kontrollierbaren Schwankung. Dies kann verschiedene Ursachen haben, zum Beispiel Alterung, äußere Umwelteinflüsse, unterschiedlicher Kundengebrauch, Bauteiltoleranzen¹ oder die Störung durch benachbarte Systeme. Signalgrößen kennzeichnen den Betriebsbereich des Systems, variieren also auch beim bestimmungsgemäßen Gebrauch unter günstigen Bedingungen. Die Bremsanlage eines Fahr-

¹ Bauteiltoleranzen werden explizit im Toleranzdesign untersucht. Theoretisch kann man die Toleranzen aber auch als Störgrößen auffassen, da die Produktionsstreuung innerhalb der vorgegebenen Toleranz nicht mehr kontrollierbar ist. Die Toleranz als solche wird dadurch zum eigenständigen Parameter und gehört nicht mehr zu den Steuergrößen.

zeugs zum Beispiel muss bei dichtem Verkehr wohl dosiert arbeiten, aber im Bedarfsfall auch brachial verzögern können. Sitze müssen für unterschiedliche Personen gleichermaßen komfortabel sein, HIFI-Anlagen müssen bei jeder Lautstärke gut klingen, Fotoapparate bei unterschiedlichen Lichtverhältnissen gute Bilder machen, etc.. Bei der Anfertigung eines Parameterdiagramms gibt es oft längere Diskussionen über die Zuordnung zu Signalgröße oder Störgröße. Hier lohnt der Aufwand nicht, denn beide Kategorien werden im späteren Versuchsplan zusammengefasst, landen quasi in einem Topf.

Bei der Definition der “error states” weicht das Parameterdiagramm von der in einer FMEA üblichen Definition ab. Dies ist vielfach nicht bekannt und führt deshalb oft zu Missverständnissen. Die FMEA rechnet auch die unzureichende Systemleistung zu den “error states”, im Parameterdiagramm wird dies bereits im Qualitätsmerkmal erfasst. Zur Vorbereitung für eine statistische Versuchsplanung ist es notwendig, eine schwache Systemleistung im Qualitätsmerkmal registrieren zu können, denn nur so lassen sich später gute Systeme eindeutig von schlechten Systemen unterscheiden. Darüberhinaus sollten die “error states” bei guten Systemen natürlich nicht vorkommen.

5.3.1.2 Versuchsplan

				N_C	+	-	-	+
				N_B	-	-	+	+
				N_A	-	+	-	+
A	B	C	D					
-	-	-	-		y_{11}	y_{12}	y_{13}	y_{14}
+	-	-	+		y_{21}	y_{22}	y_{23}	y_{24}
-	+	-	+		y_{31}	y_{32}	y_{33}	y_{34}
+	+	-	-		y_{41}	y_{42}	y_{43}	y_{44}
-	-	+	+		y_{51}	y_{52}	y_{53}	y_{54}
+	-	+	-		y_{61}	y_{62}	y_{63}	y_{64}
-	+	+	-		y_{71}	y_{72}	y_{73}	y_{74}
+	+	+	+		y_{81}	y_{82}	y_{83}	y_{84}

Tabelle 5.4 Feldkonstruktion mit innerem und äußerem Feld. Das äußere Feld beinhaltet Kombinationen der Signal- und Störgrößen, um die Systemleistung bei den verschiedensten Randbedingungen zu evaluieren.

Alle Kontrollgrößen kommen in das *innere Feld*, einen traditionellen Versuchsplan, typischerweise mit der Auflösungsstufe IV oder V. Das *äußere Feld* stellt eine Erweiterung der Methode gegenüber der klassischen Versuchsplanung dar, es ist ein um 90° gedrehtes Feld, typischerweise mit geringerer Auflösung, zum Beispiel III.

Im äußeren Feld befinden sich Störgrößen und Signalgrößen. Die Kombinationen des äußeren Feldes bilden eine Art Testprogramm für die Versuchsläufe des inneren Feldes. Jede Steuergrößenkombination wird diesem Testprogramm ausgesetzt. Daraus erwächst die Möglichkeit, das Systemverhalten unter realistischen Bedingungen zu testen, also nicht nur die optimale Systemleistung zu erfassen, sondern auch die Systemleistung bei Störungen. Eine gute Steuergrößenkombination liefert gleichzeitig eine hohe mittlere Systemleistung und geringe Schwankungen der Systemleistung in Folge von Störungen.

Der Versuchsaufwand berechnet sich multiplikativ aus den beiden Feldgrößen. Somit besteht sofort der Bedarf nach kompakten Feldern, sonst sprengt der Aufwand schnell den Rahmen der Studie. Störgrößen und Signalgrößen sind als Faktoren weniger interessant im Vergleich zu den Steuergrößen. Hier zählen nur die extremen Einstellungen mit großer Auswirkung auf das System. Kenntnisse über die Wechselwirkungen der Störgrößen untereinander sind nahezu wertlos, solange man sich sicher sein kann, dass die “schlimmste Kombination”² im Versuchsplan enthalten ist. Die Auflösungsstufe III reicht für das äußere Feld völlig aus. Da nur die Wirkung interessiert, kann man auch mehrere Parameter zu einem Faktor zusammenfassen, auch wenn diese völlig unterschiedliche physikalische Einheiten haben. Der Konstrukteur eines Startermotors könnte beispielsweise eine geringe Batteriespannung und eine hohe Motorreibung als kombinierte Störgröße ansehen. Hohe Spannung in Verbindung mit niedriger Reibung ergibt den günstigsten Fall, niedrige Spannung mit hoher Reibung die “schlimmste Kombination”.

Grundsätzlich kann man jedes innere mit jedem äußeren Feld kombinieren. Sogar eine Mischung von regulären (innen) mit irregulären (außen) Feldern ist unkritisch. Die Ergebnismatrix ermöglicht immer eine saubere Analyse der Wechselwirkungen zwischen Störgrößen und Steuergrößen. Manche Auswerteprogramme bieten eine Analyse mit innerem und äußerem Feld an, beschränken sich aber auf wenige Konfigurationen der Felder. Die Beschränkung ist unnötig, daher empfiehlt sich in diesen Fällen die eigene Erstellung eines Versuchsplans aus der für den Anwendungsfall optimalen Kombination von innerem und äußerem Feld mit nachfolgender Auswertung über die Multiple-Response-Optimisation³.

5.3.1.3 Auswertung

Die Wechselwirkungen zwischen Störgrößen und Steuergrößen sind letztlich dafür verantwortlich, dass es überhaupt eine robuste Einstellung geben kann. Gesucht wird die Einstellung der Steuergrößen, die gleichzeitig den Effekt der Störgrößen

² Als “schlimmste Kombination” gilt hier die Störgrößenkombination, die das System am meissten stört, also die Systemleistung signifikant beeinträchtigt. Im Zweifelsfall sind Vorversuche ratsam, bei denen für eine Steuergrößenkombination verschiedene Störgrößenkombinationen getestet werden.

³ Dieses Verfahren wird im Buch explizit erläutert und ist generell einsetzbar, wenn es darum geht, mehrere Qualitätsmerkmale gleichzeitig zu optimieren.

reduziert und die Leistungsanforderungen an das System erfüllt. Ohne Auswerteprogramm muss man in mehreren Arbeitsgängen vorgehen. Zunächst werden die Steuergrößen in drei Gruppen eingeteilt: Steuergrößen, die eine starke Wechselwirkung mit der Störgröße eingehen, Steuergrößen mit großem Einfluss auf das Qualitätsmerkmal, Steuergrößen ohne signifikanten Einfluss. Die erste Gruppe sorgt für die robuste Einstellung und wird mit höchster Priorität festgelegt. Mit der zweiten Gruppe kann man Einfluss auf das Qualitätsmerkmal nehmen und eventuelle Verschlechterungen ausgleichen, sofern die robuste Einstellung die Systemleistung beeinflusst hat.

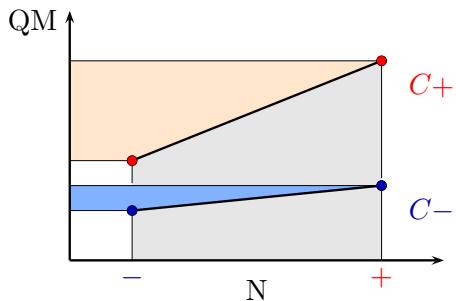


Abb. 5.9 Wechselwirkungsdiagramm. Die Störgröße N lässt sich nicht an ihrer Variation hindern. Eine geschickte Wahl der Steuergrößeneinstellung reduziert jedoch die Auswirkung der Störgröße auf das Qualitätsmerkmal. In diesem Fall ist $C-$ die richtige Wahl.

Mit Hilfe eines Auswerteprogramms besteht die Möglichkeit einer Optimierungsrechnung. Hierbei liefert das äußere Feld die notwendigen Variationen für einen eigenen Kennwert, als Qualitätsmerkmal für die Streuung. Im weiteren Verlauf wird dann nur das innere Feld untersucht, allerdings nun mit mehreren Qualitätsmerkmalen, denn eine hohe mittlere Systemleistung bleibt neben der geringen Streuung natürlich das Ziel. Zur Beschreibung der mittleren Systemleistung lässt sich ebenfalls aus den Variationen des äußeren Feldes für jede Steuergrößenkombination ein Kennwert berechnen. Dies kann zum Beispiel der arithmetische Mittelwert sein. Nach Rückführung der Ergebnismatrix auf zwei Ergebnisspalten mit unabhängigen Qualitätsmerkmalen kommt die Multiple-Response-Optimisation zum Zuge, die zum Standardrepertoire der Auswerteprogramme gehört.

In älteren Fachbüchern nimmt die Darstellung des Signal-Rausch-Verhältnisses (Signal to Noise Ratio, oder auch S/N ratio) einen größeren Raum ein. Mitunter ist es auch hilfreich, nicht den Absolutbetrag der Streuung zu betrachten, sondern einen auf die Signalstärke bezogenen Wert. Für den Anwender verwirrend ist jedoch die Vielzahl der unterschiedlichen Definitionen des Signal-Rausch-Verhältnisses mit der zugehörigen Fallunterscheidung. In vielen Anwendungsfällen bringt diese Transformation keinen erkennbaren Vorteil, sondern verkompliziert nur die Auswertung. Das Signal-Rausch-Verhältnis wird logarithmisch aufgetragen, was eine Betragsbildung des Quotienten notwendig macht. Daraus erwächst ein weiterer Nachteil. Durch die Betragsbildung verschwinden die Vorzeichenwechsel bei den Null-durchgängen. Aus numerischer Sicht ist ein Vorzeichenwechsel sehr hilfreich, um

N_C	+	-	-	+						
N_B	-	-	+	+						
N_A	-	+	-	+						
A	B	C	D		Leistung	Streuung				
-	-	-	-	y_{11}	y_{12}	y_{13}	y_{14}	}	L_1	S_1
+	-	-	+	y_{21}	y_{22}	y_{23}	y_{24}		L_2	S_2
-	+	-	+	y_{31}	y_{32}	y_{33}	y_{34}		L_3	S_3
+	+	-	-	y_{41}	y_{42}	y_{43}	y_{44}		L_4	S_4
-	-	+	+	y_{51}	y_{52}	y_{53}	y_{54}		L_5	S_5
+	-	+	-	y_{61}	y_{62}	y_{63}	y_{64}		L_6	S_6
-	+	+	-	y_{71}	y_{72}	y_{73}	y_{74}		L_7	S_7
+	+	+	+	y_{81}	y_{82}	y_{83}	y_{84}		L_8	S_8

Tabelle 5.5 Transformation des Ergebnisfeldes zu zwei unabhängigen Qualitätsmerkmalen. Ein Merkmal beschreibt die Systemleistung, zum Beispiel als Mittelwert der Ergebnisse für eine Steuergrößenkombination. Das zweite Merkmal beschreibt die Streuung, verursacht durch die verschiedenen Einstellungen des äußeren Feldes. Welche Transformation besonders günstig ist, hängt vom Einzelfall ab.

Nulldurchgänge zu finden. Bei einigen Anwendungsfällen⁴ beschreibt der Nulldurchgang den perfekten Auslegungspunkt.

5.3.2 Toleranzdesign

Strikte Kundenorientierung zeichnet viele erfolgreiche Qualitätsmethoden aus. Toleranzdesign macht dabei keine Ausnahme und ersetzt die traditionelle bauteilorientierte Sichtweise durch eine klare Fokussierung auf die Systemleistung. Letztlich dient jede Einengung der Bauteiltoleranzen nur dazu, die Streuung der Systemleistung zu reduzieren. Wirkt sich eine Bauteiltoleranz nicht auf die Streuung der Systemleistung aus, macht die Toleranzeneinengung keinen Sinn, denn der Kunde merkt nichts davon. Wenn sich hingegen eine reduzierte Streuung der Systemleistung einstellt, steigt der Marktwert des Produkts, denn in fast allen Fällen erhöht sich dadurch auch der Gebrauchswert, was der Kunde unmittelbar feststellt.

Der Mehrwert eines Produkts schafft finanziellen Spielraum für die Einengung der Bauteiltoleranzen. Die Entscheidung wird also immer auf einer sorgfältigen Kosten-Nutzen-Analyse beruhen. Erst durch die Quantifizierung des Nutzens in Form einer Wertsteigerung ist diese Analyse möglich. Außerdem erzeugt die Ab-

⁴ Im einfachsten Fall besteht das äußere Feld aus zwei Einstellungen: ideal, mit Störung. Das die Streuung beschreibende Qualitätsmerkmal ist dann lediglich die Differenz aus den beiden Ergebnissen bei gleicher Steuergrößenkombination. Diese Differenz soll bei optimaler Auslegung verschwinden (Nulldurchgang) und kann durchaus ihr Vorzeichen wechseln.

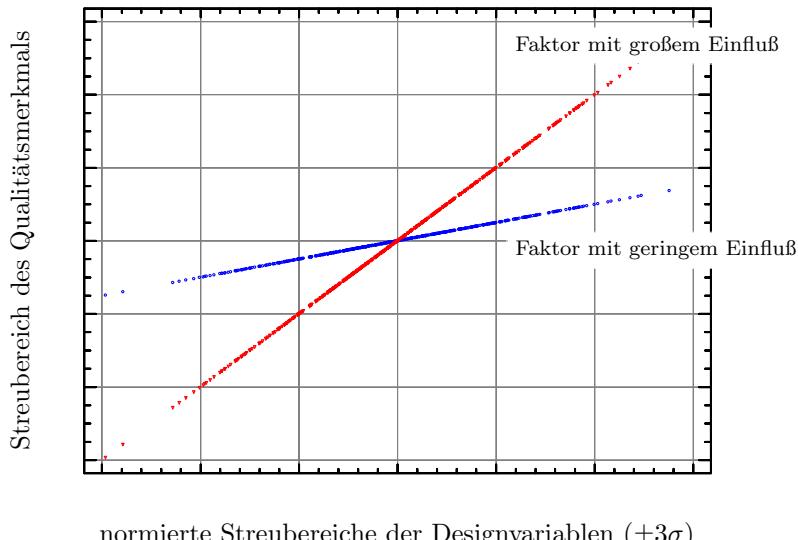


Abb. 5.10 Fiktives Beispiel für die Auswirkung mehrerer Toleranzen auf die Streuung des Qualitätsmerkmals. Die Abszisse ist normiert aufgetragen, denn die untersuchten Streuungen der jeweiligen Faktoren können durchaus völlig unterschiedliche physikalische Einheiten haben.

bildung auf die Streuung der Systemleistung einen einheitlichen Maßstab für alle untersuchten Faktoren. Bei bauteilorientierten Toleranzfestlegungen gelingt dies nicht.

Die konkrete Anwendung erfolgt völlig analog zur herkömmlichen Versuchsplanung, allerdings mit sehr kleinen Stufenabständen, entsprechend den zunächst festgelegten Bauteiltoleranzen. Wegen der geringen Stufenabstände sind nur geringe Kontraste zu erwarten. Das Qualitätsmerkmal verändert sich also in einem sehr schlecht messbaren, geringem Umfang. Dies macht bei realen Versuchen immer eine Messmittelfähigkeitsanalyse erforderlich. Berechnungsverfahren sind hier im Vorteil, weil in der Regel die Rundungsfehler weit unter den Kontrasten liegen.

Die Aufteilung der Gesamtvarianz V_{ges} ergibt sich aus der additiven Überlagerung der Teilvarianzen. Zunächst produziert jeder untersuchte Faktor einen Varianzanteil V_i . Die Auswertung der Versuchsreihe wird eine Restvarianz V_R aufzeigen, die durch eventuelle Versuchsstreuung, unkontrolliert schwankende Parameter und eine begrenzte Genauigkeit des Beschreibungsmodells zu erklären ist. Möglicherweise wurden signifikante Parameter nicht im Rahmen der Versuchsreihe variiert. Diese schwanken dann entweder unkontrolliert oder es ist gelungen, sie im Rahmen der Versuchsreihe konstant zu halten. In jedem Fall treiben sie in der Realität die Gesamtvarianz des Qualitätsmerkmals nach oben, also kommt zu den genannten Termen noch ein unbekannter Varianzanteil V_u hinzu.

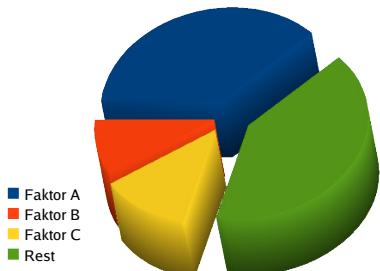


Abb. 5.11 Beispiel für die Aufteilung der Gesamtvarianz. Die Varianzanalyse zeigt auf, welche Anteile den jeweiligen Faktoren zuzuordnen sind. Der verbleibende Rest wurde nicht durch die untersuchten Faktoren erzeugt. Eine Toleranzeneingang kann nur den jeweiligen Varianzanteil günstig beeinflussen. Daraus ergibt sich das Verbesserungspotential in Bezug auf die Systemleistung.

$$V_{ges} = \sum_{i=1}^{n_f} V_i + V_R + V_u \quad (5.4)$$

Die individuellen Varianzanteile V_i der Faktoren sind proportional zum Quadrat des Effektbetrags. Daraus lässt sich schnell ausrechnen, welchen Varianzanteil eine reduzierte Bauteiltoleranz produziert. Innerhalb der getesteten Stufenabstände ist zunächst eine lineare Interpolation zulässig, so liefert die halbierte Bauteiltoleranz den halben Effektbetrag und ein Viertel der Teilvarianz. Eine Extrapolation muss immer nachträglich überprüft werden, denn eine Erweiterung der Toleranz kann zu neuen physikalischen Effekten führen, die im Rahmen der ursprünglichen Versuchsreihe nicht aufgetreten sind.

$$V_{i_{neu}} = V_{i_{orig}} \cdot \frac{\Delta^2 x_{i_{neu}}}{\Delta^2 x_{i_{orig}}} \quad (5.5)$$

Die Quantifizierung des Nutzens erfordert eine Abbildung der Streungsreduktion auf einen monetären Gegenwert. Natürlich ist diese Abbildung sehr produkt spezifisch und kann sich mit den Kundenerwartungen auch im Laufe der Zeit ändern. Offenbar trifft die Parabel als Abbildungsfunktion in vielen Fällen recht gut und wird üblicherweise angesetzt. Bei allen anderen Funktionen ergibt sich außerdem ein höherer Berechnungsaufwand. Ausgangspunkt ist eine Negativbetrachtung, also monetärer Verlust durch Streuung. Ohne Streuung ist der Verlust Null, mit Streuung wächst der Verlust quadratisch an. Die Proportionalitätskonstante erhält man durch Marktstudien oder Vergleichsmessungen an Konkurrenzprodukten. Wenn die Bewertung der Toleranz asymmetrisch ist, wird jedem Ast der Parabel eine eigene Konstante zugewiesen.

Abgesehen von der Kosten-Nutzen-Analyse bringt das Toleranzdesign weitere Vorteile. Zunächst schließt die Methode mit der Unart ab, pauschal alle Bauteiltoleranzen zu verringern, was in jedem Fall unnötige Produktionskosten erzeugt. Der Direktvergleich der positiven Auswirkung ermöglicht eine Rangfolge der notwendigen Veränderungen, auch wenn es sich um völlig unterschiedliche Bauteiltoleranzen mit jeweils eigenen physikalischen Einheiten handelt.

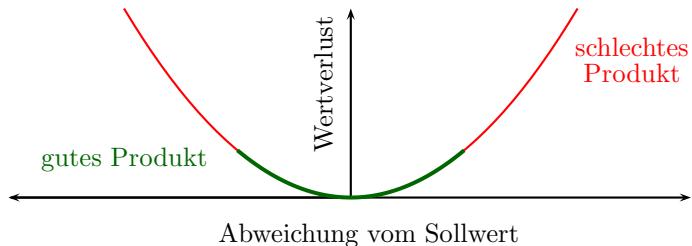


Abb. 5.12 Die grundsätzliche Idee der Kosten-Nutzen Rechnung. Jede Abweichung vom Sollwert des Qualitätsmerkmals erzeugt einen Wertverlust des Produkts. Gute Produkte liefern hier eine geringere Streuung. Geht man von einer quadratischen Bewertungskurve aus, ist der Wertverlust proportional zur Gesamtvarianz des Qualitätsmerkmals. Liegen die Kosten der Abstellmaßnahme unter dem verursachten Wertverlust, lohnt sich die Maßnahme.

Die Eingangsdaten für Computermodelle unterliegen oft einer erheblichen Unsicherheit. Daraus können Variationen der Ergebnisgrößen erwachsen, die weit über den eigenlichen Modellfehlern liegen. Kurzum, eine Verfeinerung des Modells würde keinen Sinn machen, bevor die Eingangsdaten in höherer Genauigkeit verfügbar sind. DOLTSINIS [1] und MARCZYK [3] bringen eindrucksvolle Beispiele dazu. Toleranzdesign und die im Kapitel *Sensitivitätsanalyse* vorgestellte Methodik sind artverwandt. Bei der Sensitivitätsanalyse wird der Bogen weiter gespannt. An die Stelle von lediglich zwei Einstellungen je Faktor kommen dann sehr große Stichproben mit einer repräsentativen Verteilung. Bei nichtlinearen Zusammenhängen zwischen Faktoreinstellung und Qualitätsmerkmal kommt es in der Regel zu einer starken Verformung der Verteilungsfunktion. Die stochastische Auswertung liefert auch in diesen Fällen ein sehr klares Bild von der zu erwartenden Steigung des Qualitätsmerkmals.

5.4 Umgang mit kategorialen Faktoren

In der Praxis sind die meisten Faktoren numerischer Natur, also kontinuierlich einstellbar. Das lineare Modell ist auf diesen Faktortyp ausgerichtet, denn es erlaubt Vorhersagen für Zwischenstufen. Es gibt jedoch auch kategoriale (categorical) Faktoren. Diese Faktoren lassen sich nicht zählen, wiegen oder messen. Ihre Einstellung erfolgt immer in festen Stufen.

Faktor	Stufe 1	Stufe 2	Stufe 3
Lieferant	Müller	Meier	Schulze
Verbindungsart	schrauben	nieten	kleben
Verunreinigung	Ei	Blut	Kakao
Musikrichtung	Jazz	Pop	Klassik
Geometrie	rund	eckig	
Verstärkungsblech	ohne	mit	

Tabelle 5.6 Beispiele für kategoriale Faktoren.

Wie wertet man diese Faktoren aus? Wie geht man mit Kombinationen aus numerischen und kategorialen Faktoren um? Es gibt mehrere Lösungswege dafür.

Ein Weg besteht darin, den kategorialen Charakter zunächst zu ignorieren und den Faktor wie einen numerischen Faktor zu behandeln. Insbesondere in der Phase der Faktorsichtung ist dies völlig unkritisch, denn dann geht es nicht um die Einstellung von mittleren Werten sondern nur um die Identifikation der signifikanten Faktoren. Ein Mittelwert erscheint zunächst sinnlos, bei näherer Betrachtung lassen sich aber in vielen Fällen die Eigenschaften kategorialer Faktoren auf einer numerischen Skala beschreiben. Wenn der als numerischer Faktor getarnte kategoriale Faktor eine große Bedeutung hat, wird die Abbildung auf die entsprechende numerische Eigenschaft erforderlich.

Wenn zum Beispiel die Verbindungsart *schrauben* besser ist als die Verbindungsart *kleben*, lohnt sich eine genaue Analyse der Unterschiede. Möglicherweise ist das charakteristische Merkmal die maximale Scherkraft, also letztlich eine numerische Variable. Geometrien lassen sich in vielen Fällen durch Morphing stufenlos ineinander überführen. Clevere Formulierungen mit Shape Variablen machen dann aus kategorialen Faktoren wieder numerische Faktoren. Ist ein Bauteil mal vorhanden und mal nicht, kann häufig eine numerische Variable den Sachverhalt elegant beschreiben. Die Variable (E-Modul, Blechdicke, etc.) wird dann so definiert, dass sie den Wert Null annimmt, wenn das Bauteil nicht vorhanden ist. Die Produkte verschiedener Lieferanten haben eigene charakteristische Eigenschaften, die sich normalerweise ebenfalls in Zahlen ausdrücken lassen. Sollte der Faktor *Lieferant* eine große Bedeutung haben, ist es ohnehin höchste Zeit, der Sache auf den Grund zu gehen und die Unterschiede zu analysieren. Wenn der Faktor *Musikrichtung* zum Beispiel bei der optimalen Einstellung des MP3-Players eine Rolle spielt, steckt vermutlich das Zusammenspiel zwischen Frequenzgang und den jeweils wichtigsten Frequenzändern dahinter. Auch dies sind letztlich numerische Variablen.

Warum dieser Trick? Versuchsplan und Auswertung vereinfachen sich, wenn man nur mit numerischen Faktoren arbeitet. Dann sind alle Versuchspläne und Beschreibungsmodelle einsetzbar. Auch die Optimierung funktioniert dann reibungslos im gesamten Faktorraum. Vereinzelte zweistufige kategoriale Faktoren kann man bei der Optimierung nacheinander auf die beiden Stufen festsetzen und die Optima vergleichen. Bei dreistufigen kategorialen Faktoren ist diese Taktik nicht mehr zulässig, weil die Wechselwirkungsterme zwischen den numerischen und den kategorialen Faktoren des Beschreibungsmodells unsinnig sind.

Sollte die Rückführung auf einen numerischen Faktor nicht möglich sein, sind einschränkende Randbedingungen zu beachten. Zwischenwerte machen in diesem Fall keinen Sinn. Man vergleicht lediglich die Ergebnisse der jeweiligen Stufen untereinander. Die Varianzanalyse gibt dann Aufschluss über Effekte und Wechselwirkungen. Als Darstellungen kommen hierbei auch Streudiagramme zum Einsatz, aufgeteilt nach den Faktorstufen. Die Mischung von numerischen Faktoren und kategorialen Faktoren ist aufwendig, da man natürlich für die numerischen Faktoren nach wie vor gerne ein Beschreibungsmodell hätte. Hierzu lassen sich ein beliebiger Versuchsplan für die numerischen Faktoren mit einem Vollfaktorplan für die

kategorialen Faktoren multiplikativ verbinden. Für jede Einstellung der kategorialen Faktoren folgt dann ein eigenes Beschreibungsmodell der numerischen Faktoren.

Die sauberste Lösung ist in jedem Fall der Vollfaktorplan, also ein Versuchsplan mit allen Kombinationen. Dies ist in der Praxis jedoch nur bei einer geringen Zahl von Faktoren durchführbar. DoE-Auswerteprogramme bieten allerdings dafür eine eigene Analyse an, die speziell auf kategoriale Faktoren zugeschnitten ist.

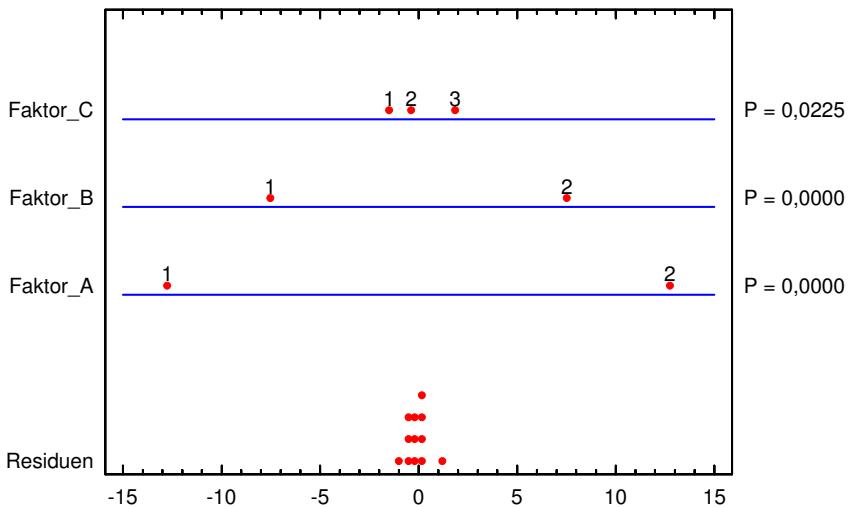


Abb. 5.13 Grafische ANOVA einer Versuchsreihe mit drei kategorialen Faktoren. Die Abbildung zeigt die Abweichung der jeweiligen Stufenmittelwerte vom Gesamtmittelwert im Vergleich zu den Residuen. In diesem Beispiel wurden die Faktoren A und B auf zwei Stufen und der Faktor C auf drei Stufen variiert. Bei allen Faktoren zeigt sich eine signifikante Differenz der jeweiligen Stufenmittelwerte im Vergleich zur Varianz der Residuen. Folglich handelt es sich um wahre Effekte.

Literaturverzeichnis

1. Doltsinis, I.: *Stochastic Analysis of Multivariate Systems in Computational Mechanics and Engineering*. Int. Center for Numerical Methods in Engineering, Barcelona (1999) 156
2. Fowlkes, W., Creveling, C.: *Engineering Methods for Robust Product Design*. Addison-Wesley, Reading, MA (1995) 2, 34, 58, 139
3. Marczyk, J.: *Principles of Simulation-Based Computer Aided Engineering*. FIM Publications, Madrid (1999) 156
4. Siebertz, K., Midoun, D.: *CAE Driven Parameter Studies Using the DoE Method and Tailored Objective Functions*. In: MADYMO User's Conference, Windsor Canada (1998) 145
5. Toutenburg, H.: *Versuchsplanung in der Industrie*. Prentice Hall, München (1996) 139

Kapitel 6

DoE Beispiele

6.1 Einleitung

Die nachfolgenden Beispiele kommen aus der industriellen Praxis und sind nicht eigens konstruiert worden, um ein Lehrbuch zu unterstützen. Vorausgesetzt wird die in den Kapiteln 1-5 gebrachte Theorie. An dieser Stelle sei angemerkt, dass das Rasensprengerbeispiel des Buches mittlerweile Gegenstand einer eigenen Publikationen geworden ist [4]. Eine interaktive Version ist im Internet verfügbar. Diese wurde von Dr. Koen Rutten erstellt, während seiner Promotion an der KU Leuven.

Das erste Beispiel verdeutlicht die Vorgehensweise zum Screening von Parametern. Außerdem zeigt es, wie die Principal Component Analysis genutzt werden kann, um den Zusammenhang zwischen mehreren Qualitätsmerkmalen zu untersuchen. Das zweite Beispiel ist etwas aufwändiger. Zunächst erfolgt eine qualitative Systembeschreibung über ein Parameterdiagramm. Dann werden nichtlineare Zusammenhänge und Wechselwirkungen zwischen den Faktoren und den unterschiedlichen Qualitätsmerkmalen untersucht. Hier kommt eine Multiple Response Optimization zum Einsatz.

Die Darstellung in diesem Kapitel konzentriert sich auf die Anwendung der DoE Methode. Beide Beispiele sind in wissenschaftlichen Publikationen veröffentlicht worden [2] [3] [1]. Dort findet der geneigte Leser weitere Informationen zu den technischen Hintergründen der Beispiele oder die Deutung der Ergebnisse.

Zu Dank verpflichtet sind die Autoren den Experten, die an den gezeigten Beispielen gearbeitet haben und die einer Verwendung für dieses Buch zustimmten. Dr.-Ing. Detlef Neuenhaus hat alle Berechnungen zum Beispiel Schutzplanke mit dem von ihm selbst entwickelten Mehrkörpersimulationsprogramm durchgeführt. Dr.-Ing. Rainer Lach bearbeitete das Beispiel Ventiltrieb, inklusive der Berechnungen und Skizzen. Die Veröffentlichung des zweiten Beispiels erfolgt mit freundlicher Genehmigung der Ford Werke GmbH.

6.2 Die Schutzplanke

Schutzplanken leisten einen wesentlichen Beitrag zur Sicherheit im Straßenverkehr. Moderne Schutzplanken sind nachgiebig und können einen Teil der Anprallenergie absorbieren. Es entsteht ein Zielkonflikt zwischen Bauraum und Deformationsweg, weshalb ein derartiges Schutzsystem sorgfältig optimiert sein sollte. Im nachfolgenden Beispiel geht es um eine Schutzplanke aus Stahl und die Sichtung der relevanten Parameter, weshalb ein Screening-Array verwendet wurde. Die Bundesanstalt für Straßenwesen verwendet in ihrer Einsatzempfehlung für Fahrzeug-Rückhaltesysteme den Begriff *Fahrzeug-Rückhaltesystem* statt *Schutzplanke*. Im Verwechslungen mit dem Insassen-Rückhaltesystem zu vermeiden (welches im Grundlagenkapitel als Beispiel verwendet wird), sei in diesem Zusammenhang die Verwendung des Begriffes Schutzplanke gestattet.

6.2.1 Systembeschreibung und Versuchsplan



Abb. 6.1 Schutzplanke in typischer Ausführung. Fotografiert am Aachener Kreuz.

Für die Konstruktion einer derartigen Schutzplanke gibt es eine Vielzahl konstruktiver Parameter. Die Blechdicken sind variierbar, ebenso die Geometrie der Profile, die Abstände der Pfosten und natürlich auch die verwendete Stahlsorte. Dies sind Steuergrößen im Sinne der DoE. Die Belastung im Falle eines Anpralls hängt von der Fahrzeugkonstruktion ab. Um vergleichbare Bedingungen zu schaffen, wurden in der EN 1317 Fahrzeugtypen und die zugehörigen Anprallarten normiert. Es gibt elf verschiedene Typen und Anprallarten. Im Rahmen des Beispiels wurden nur zwei davon exemplarisch untersucht. TB11 repräsentiert ein leichtes Fahrzeug (900kg), welches mit $100 \frac{km}{h}$ unter einem Winkel von 20° kollidiert. TB32 repräsentiert ein mittelschweres Fahrzeug (1500kg), welches mit einer Anprallgeschwindigkeit von $110 \frac{km}{h}$ unter einem Winkel von 20° kollidiert. Die Schutzplanke soll viele

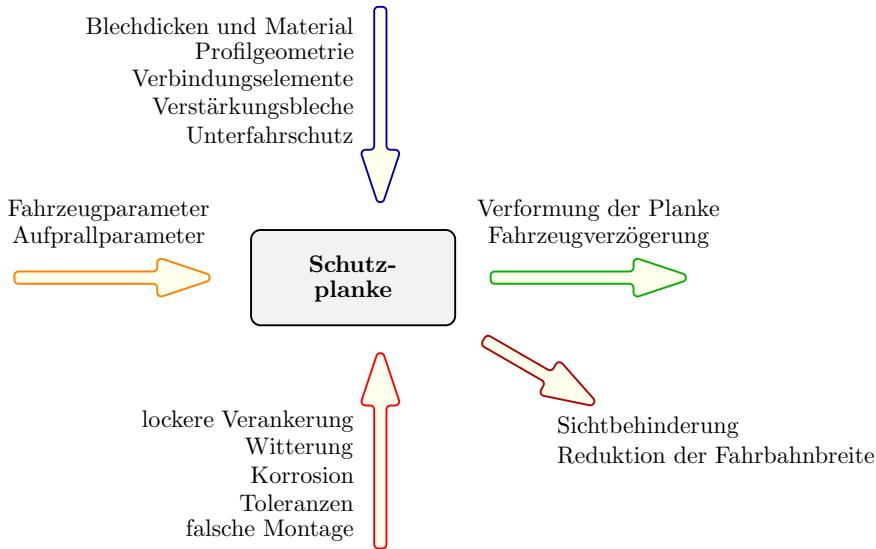


Abb. 6.2 Qualitative Systembeschreibung einer Schutzplanke mit Hilfe des P-Diagramms.

unterschiedliche Anprallsituationen sicher meistern, dies sind somit die Signalgrößen im Sinne der DoE. Als Störgrößen tauchen Einbaubedingungen, Umwelteinflüsse und Toleranzen auf. Sichtbehinderungen oder eine reduzierte Fahrbahnbreite können als Error States vorkommen.

Kodierung	Faktor	Stufe -	Stufe +
A	Dicke des Verstärkungsprofils	3mm	5mm
B	Abstand der Pfosten	0,666 m	1,000 m
C	Verschraubung, Pfosten - Leitplanke	M10	M12
D	Verstärkungsblech	mit	ohne
E	Plankenprofil	Typ B	Typ A
F	Unterfahrschutz	ohne	mit
G	Stahlsorte	S235	S355
H	aufprallendes Fahrzeug	TB11	TB 32

Tabelle 6.1 Faktoren des Versuchsplans und Stufenfestlegung

Als Qualitätsmerkmale für diese Untersuchung galten vier unterschiedliche Größen: Der Wirkungsbereich bezeichnet die Summe aus Bautiefe und dynamischer Verformung. Synonym zum Begriff *Wirkungsbereich* ist auch der Begriff *Arbeitsweite* üblich, denn dies ist der benötigte Platzbedarf des Schutzsystems. Ein zu großer Wirkungsbereich reduziert die effektive Fahrbahnbreite. Die dynamische Verformung bezeichnet die maximale laterale Verschiebung des Schutzsystems in Folge des Anpralls.

Der Kennwert, Acceleration Severity Index, *ASI* ist ein Maß für die Stärke der Fahrzeugverzögerung. Dies ist in der Norm EN 1317 definiert. Natürlich ist eine moderate Verzögerung wünschenswert. Berechnet wird der *ASI* aus den auf ein Zeitfenster von 50ms gemittelten Spitzenverzögerungen und den für die jeweiligen Dimensionen definierten Grenzwerten.

$$ASI = \sqrt{\frac{\hat{a}_x^2}{\hat{a}_x^2} + \frac{\hat{a}_y^2}{\hat{a}_y^2} + \frac{\hat{a}_z^2}{\hat{a}_z^2}} \quad (6.1)$$

Die entsprechenden Grenzwerte liegen bei $\hat{a}_x = 12g$, $\hat{a}_y = 9g$ und $\hat{a}_z = 10g$, für die longitudinale (x), laterale (y) und vertikale (z) Richtung. Je nach erreichtem *ASI* wird das Schutzsystem in Kategorien eingeteilt. Über 1,4 gehört es zu Kategorie C, zwischen 1,4 und 1,0 zu Kategorie B und unter 1,0 zu Kategorie A. Gleichzeitig wird allerdings für Kategorie A und Kategorie B die Einhaltung eines Belastungsgrenzwertes gefordert, dem *THIV* (Theoretical Head Impact Velocity), der die theoretische Aufprallgeschwindigkeit des Kopfes im Fahrzeug beschreibt. Bei $THIV \geq 33 \frac{km}{h}$ fällt das Schutzsystem in Kategorie C.

<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>H</i>	Arbeitsweite [mm]	Verformung [mm]	<i>ASI</i> [-]	<i>THIV</i> [$\frac{km}{h}$]
+	-	+	-	-	-	+	+	536	233	1,350	17,89
+	+	-	+	-	-	-	+	731	455	0,971	16,65
-	+	+	-	+	-	-	-	427	194	1,194	14,46
+	-	+	+	-	+	-	-	397	176	1,255	14,35
+	+	-	+	+	-	+	-	391	172	1,294	14,32
+	+	+	-	+	+	-	+	616	364	1,270	17,85
-	+	+	+	-	+	+	-	400	173	1,275	14,35
-	-	+	+	+	-	+	+	643	276	1,334	18,25
-	-	-	+	+	+	+	-	564	308	1,301	16,25
+	-	-	-	+	+	+	-	350	125	1,372	14,68
-	+	-	-	-	+	+	+	577	271	1,266	16,97
-	-	-	-	-	-	-	-	430	202	1,237	14,57

Tabelle 6.2 Ergebnistabelle der Untersuchung. Die vier Qualitätsmerkmale wurden gleichzeitig ermittelt, müssen aber unabhängig voneinander ausgewertet werden. Hierzu wird nacheinander jede der Ergebnisspalten interpretiert.

Bei der Wahl des Versuchsplans stand die Effizienz im Vordergrund. Daher bot sich das Screening-Array nach PLACKETT BURMAN an. Mit nur zwölf Einstellungen ließen sich auf diese Weise die acht als wichtig eingeschätzten Faktoren untersuchen. Diese Konfiguration gestattet es allerdings nicht, die Wechselwirkungen zu

analysieren. Bei Berechnungen treten keine Zufallsstreuungen auf¹, also macht ein Half-Normal-Plot bei diesem Beispiel keinen Sinn.

6.2.2 Auswertung der Qualitätsmerkmale

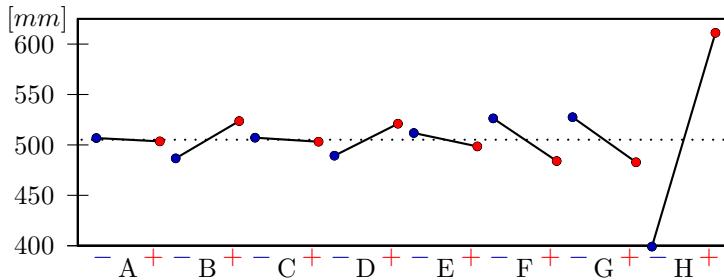


Abb. 6.3 Grafische Darstellung der Effektberechnung für das Qualitätsmerkmal: Wirkungsbereich (Arbeitsweite).

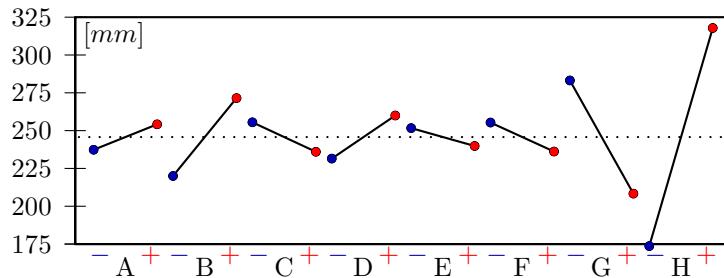


Abb. 6.4 Grafische Darstellung der Effektberechnung für das Qualitätsmerkmal: dynamische Verformung.

Wirkungsbereich (Arbeitsweite): Faktor H hat den stärksten Effekt. Dies leuchtet ein, denn mit Fahrzeugmasse und Anprallgeschwindigkeit steigt auch die zu absorbierende Anprallenergie. Dynamische Verformung: Auch hier dominiert Faktor H, allerdings weniger stark als im vorigen Qualitätsmerkmal. ASI: Hier zeigt sich der Zielkonflikt zwischen Nachgiebigkeit und moderater Verzögerung. ASI und THIV hängen nicht unmittelbar miteinander zusammen. Beim THIV spielt das *timing* des Anprallvorgangs eine große Rolle, was eine intuitive Vorhersage der Ergebnisse na-

¹ Bei aufwändigen Crash Berechnungen mit expliziten FE-Solvoren kann es zu numerischen Artefakten kommen, die einer Zufallsstreuung ähneln, aber das ist die Ausnahme.

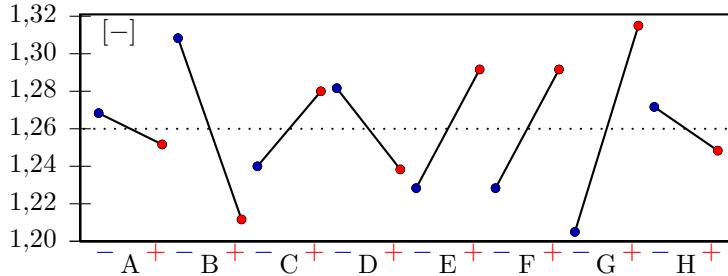


Abb. 6.5 Grafische Darstellung der Effektberechnung für das Qualitätsmerkmal: ASI.

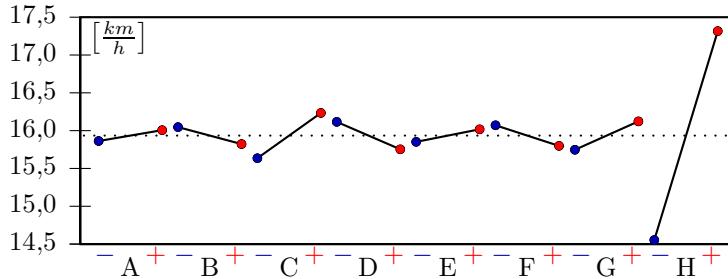


Abb. 6.6 Grafische Darstellung der Effektberechnung für das Qualitätsmerkmal: THIV.

hezu unmöglich macht. Die Ergebnisse zeigen deutlich, dass die vier Qualitätsmerkmale irgendwie zusammenhängen. Spannend ist nun die Frage, wieviele Dimensionen der *Ergebnisraum* wirklich hat. An dieser Stelle erweist sich die Principal Component Analysis (PCA) als nützlich.

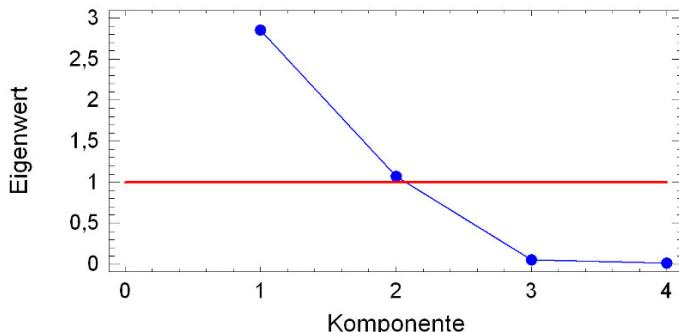


Abb. 6.7 Scree-Plot der Principal Component Analysis. Etwa 98 Prozent der Variation lässt sich durch zwei Komponenten erklären. Das Problem ist also im Grunde nicht vierdimensional, sondern zweidimensional.

Insgesamt steckten sehr viele Erkenntnisse in nur zwölf Berechnungsläufen. Die DoE Methode konnte in Kombination mit der PCA den Einfluss von acht Faktoren

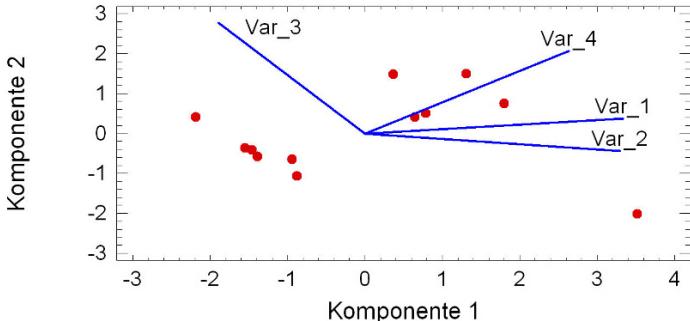


Abb. 6.8 Bi-Plot der Principal Component Analysis. Der Zielkonflikt zwischen Nachgiebigkeit (Variable 1 und Variable 2) und dem Belastungskennwert ASI (Variable 3) ist offenkundig. Es wird unmöglich sein, alle drei Qualitätsmerkmale gleichzeitig zu minimieren, also bleibt es bei einem Kompromiss. Variable 4 (THIV) zeigt kein eindeutiges Verhalten. Für eine Interpretation reicht die Zahl der Datenpunkte hier nicht aus.

auf vier unterschiedliche Qualitätsmerkmale quantifizieren und darüberhinaus auch eine Verbindung zwischen den Qualitätsmerkmalen aufzeigen.

6.3 Der Ventiltrieb

Die Ventilsteuerung ist eine zentrale Baugruppe des Viertakt-Verbrennungsmotors. Moderne Motoren erreichen ihre hohe Leistungsdichte und Wirtschaftlichkeit nur bei einem optimalen Gasaustausch zwischen den Arbeitstakten. Das gezeigte Beispiel entstand bei der Entwicklung eines 6-Zylinder Dieselmotors.

6.3.1 Systembeschreibung

Auch hier lohnt es sich, mit der qualitativen Systembeschreibung zu beginnen. Dies schafft einen Überblick und leitet das weitere Vorgehen ein. Die Liste der Kontrollgrößen ist lang. In gewissen Grenzen sind die geometrischen Parameter der beteiligten Komponenten variierbar. Die Motordrehzahl ist eine Signalgröße, denn sie kennzeichnet den Betriebsbereich des Ventiltriebs und zählt weder zu den Kontrollgrößen, noch zu den Störgrößen. Relevante Störgrößen sind in diesem Zusammenhang die Massen- und Federtoleranzen. Der Kontakt mit dem Nocken muss zu jeder Zeit gewährleistet sein, möglichst ohne hohe Reibungsverluste in Folge zu starker Feder Spannungen. Die beiden genannten Störgrößen lassen sich im Versuchplan zu einer kombinierten Störgröße zusammenfassen. Ihre Wirkung auf die Systemleistung ist gleich, allerdings mit umgekehrten Vorzeichen. Somit muss die leichte Masse mit der hohen Federvorspannung kombiniert werden und umgekehrt. Schwere Masse in



Abb. 6.9 Darstellung des Ventiltriebs für einen der sechs Zylinder. Jeder Zylinder wird durch vier Ventile versorgt. Die Zylinder sind in zwei Bänken angeordnet. Quelle: Ford

Kombination mit geringer Vorspannung der Feder ist hierbei der ungünstige Fall. Error States wurden im Rahmen der Studie nicht explizit untersucht, sollen aber dennoch genannt werden. Geräuschenwicklung und Reibungsverluste des Ventiltriebs können bei modernen Motoren durchaus eine Rolle spielen und sind durch sorgfältige Auslegung in Grenzen zu halten.

Es gibt viele Qualitätsmerkmale, die einen guten Ventiltrieb von einem schlechten Ventiltrieb unterscheiden. Zunächst sei die Hertz'sche Pressung genannt. Dies ist die Kontaktspannung. Eine hohe Kontaktspannung zwischen bewegten Teilen führt unweigerlich zu schwierigen Bedingungen für die Schmierung, also droht Verschleiß. Auch eine hohe Gleitgeschwindigkeit kann die Schmierung erschweren. Hohe Kräfte am Spielausgleich verlangen nach einer stärkeren Auslegung. Die Nockenkonkavität sollte sein, um den permanenten Kontakt zwischen Nocken und Rolle zu gewährleisten. Kontaktverlust kann fatale Folgen für die Ventilkinematik haben und Verschleiß erzeugen, deshalb erzwingt eine hohe Konkavität eine hohe Vorspannung, die ihrerseits Reibung erzeugt. Die Fertigungskosten der Nockenwelle steigen erheblich, wenn Konkavität verlangt wird.

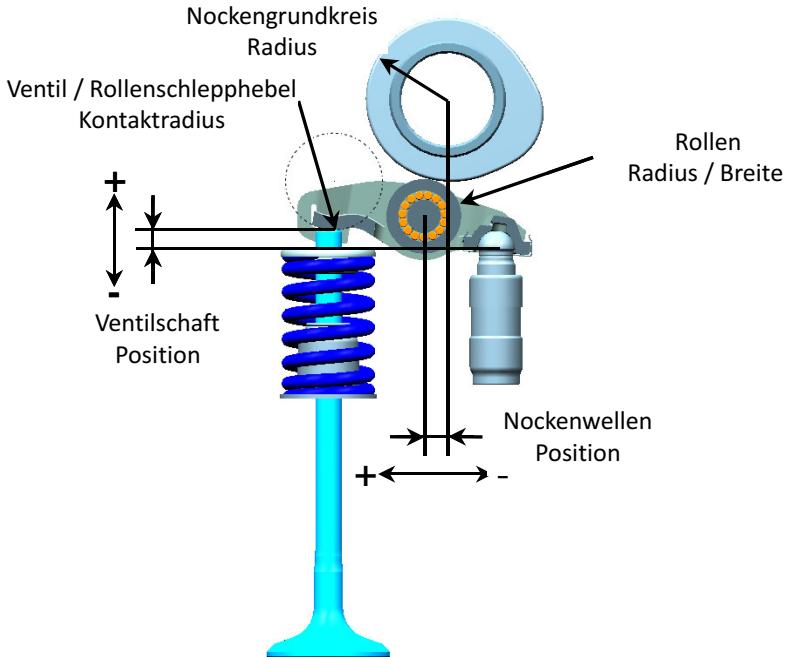


Abb. 6.10 Skizze des untersuchten Systems, mit Kennzeichnung relevanter Parameter. Quelle: Ford

6.3.2 Versuchsplan und Stufenfestlegung

Damit ist der Rahmen der Untersuchung grob abgesteckt. Nun kommt es zur Auswahl der Faktoren, des Versuchsplans und der Definition der Stufen. Bereits vor Beginn der Studie war bekannt, dass die Wechselwirkungen nicht vernachlässigbar sein werden und darüber hinaus auch nichtlineare Zusammenhänge zwischen einzelnen Faktoren und Qualitätsmerkmalen vorliegen. Ausgewählt wurden die in der Tabelle vorgestellten Faktoren, auf jeweils drei Stufen. Die kombinierte Störgröße reduziert den Aufwand, im Vergleich zu zwei separaten Störgrößen. Aufwand ist insgesamt ein gutes Stichwort, denn die Zahl aller Kombinationen bei acht Faktoren auf drei Stufen liegt bei 6561. Natürlich ist der Vollfaktorplan keine gute Option für diese Anwendung. Es handelt sich hier um eine CAE Studie, also wäre die Monte-Carlo Methode durchaus möglich gewesen, allerdings nur mit circa 500 Kombinationen oder mehr. Ein Latin-Hypercube könnte die nötige Zahl auf rund 300 drücken, was aber immer noch relativ hoch ist. Die klassische Variante, mit einem inneren Feld für die Steuergrößen und einem äußeren Feld für die Signal- und Störgrößen bietet den Vorteil einer sauberen Ermittlung aller Wechselwirkungen, welche die Systemstreuung beeinflussen. Nachteilig ist jedoch der hohe Aufwand bei dreistufigen Versuchsplänen. In diesem Fall hätte man ein inneres Feld der

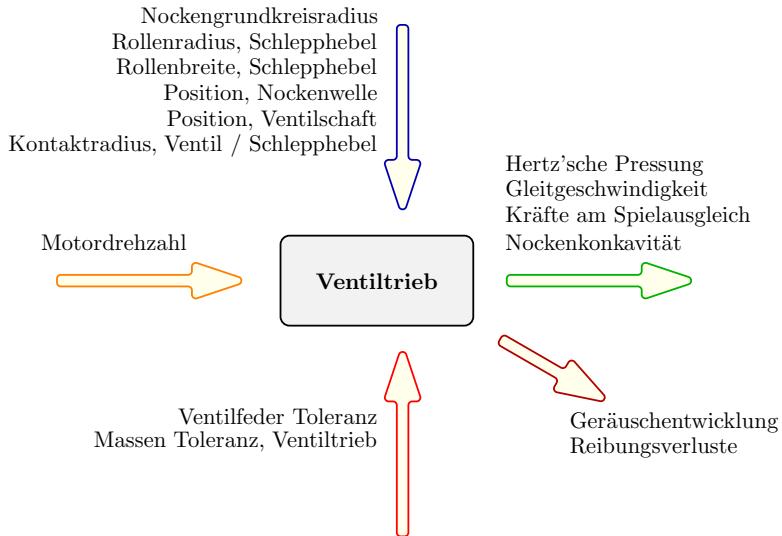


Abb. 6.11 Qualitative Systembeschreibung des Ventiltriebs mit Hilfe des P-Diagramms.

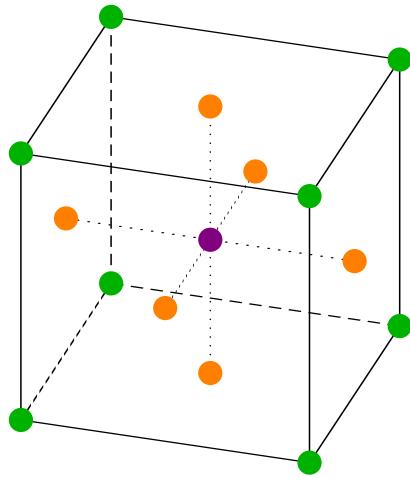


Abb. 6.12 Versuchsplan: Face Centered Central Composite Design.

Auflösung IV für sechs Faktoren auf drei Stufen mit einem äußeren Feld für zwei Faktoren kombinieren müssen, wovon mindestens einer der beiden Faktoren auf drei Stufen darzustellen ist. Das wäre auf 270 Kombinationen hinausgelaufen. Alternativ dazu bietet ein einziges Feld für acht Faktoren auf drei Stufen fast die gleiche Informationsfülle, bei weniger als einem Drittel der Kombinationen. In Frage kommen die Box-Behnken Konstruktion und das Central-Composite Design. Box-Behnken Felder betonen die mittlere Einstellung, sobald die Zahl der Faktoren über vier ansteigt. Für acht Faktoren empfiehlt sich diese Konstruktion also nicht.

Kode	Faktor	Einheit	-1	0	+1
A	Nockengrundkreisradius	mm	16	17	18
B	Rollenradius	mm	8	8,5	9
C	effektive Rollenbreite	mm	5,5	7	8,5
D	horizontale Position, Nockenwelle	mm	-6,6	0	6,6
E	vertikale Position, Nockenwelle	mm	3	4	5
F	Kontaktradius, Ventil/Schlepph.	mm	8	11	14
G	Motordrehzahl	U/min	750	2500	5000
H	Ventiltriebstoleranzen, Masse und Vorspannung (unten)	%	105 95	100 100	95 105

Tabelle 6.3 Faktoren des Versuchsplans und Stufenfestlegung

Es bleibt also das Central-Composite Design. Hierbei wird ein ursprünglich zweistufiger Versuchsplan (der Würfel) um eine Variation einzelner Faktoren um die mittlere Einstellung aller Faktoren (den Stern) erweitert. Es gibt zwei Varianten, die sich in Verlängerungsfaktor des Sterns unterscheiden. Perfekt orthogonale Central-Composite Designs für acht Faktoren haben eine Achsenlänge von 2,05, bezogen auf den Abstand zwischen mittlerer Einstellung und Stufe – bzw. Stufe+. Diese hohen Kontraste waren im vorliegenden Beispiel nicht mehr umsetzbar. Als Kompromiss erwies sich das Face-Centered-Central-Composite Design, mit einem Verlängerungsfaktor von 1, also ragt der Stern nicht über den Würfel hinaus. Erkauft wird diese praktische Vereinfachung mit einer leichten Korrelation der quadratischen Effekte untereinander. Der Betrag der Korrelation liegt unter 0,14, also ist dies zu verkraften. In diesem Fall bleibt es bei 81 Kombinationen, 64 für den Würfel und 17 für den eingepassten Stern. Unter Annahme eines Beschreibungsmodells zweiter Ordnung liegen 45 Modellkonstanten vor (Mittelwert, lineare Effekte quadratische Effekte, zweifache Wechselwirkungen). Der Versuchsplan liefert also trotz seines kompakten Aufbaus noch einen hinreichenden Informationsüberschuss, um einen over-fit zu vermeiden. Die Draper-Lin Variante, mit nur 53 Kombinationen wäre hier kritisch.

6.3.3 Auswertung der Qualitätsmerkmale

Schreiten wir zur Tat. Die Qualitätsmerkmale werden nach der Ermittlung der erreichten Werte zunächst einzeln ausgewertet. Die Zusammenfassung erfolgt erst später. Die Effekt-Diagramme quantifizieren die schon qualitativ erwarteten nichtlinearen Zusammenhänge. Darüber hinaus zeigt sich, wie unterschiedlich die jeweiligen Qualitätsmerkmale auf die Variation derselben Faktoren reagieren. Für den Fachmann ist bereits dieser Teil der Analyse sehr aufschlussreich.

Bei den Wechselwirkungen ergibt sich ein relativ aufgeräumtes Bild. Dargestellt sind nur die Wechselwirkungen von signifikanter Größe. In der Summe sind dies weniger Wechselwirkungen, als man von einem derart komplexen System hätte er-

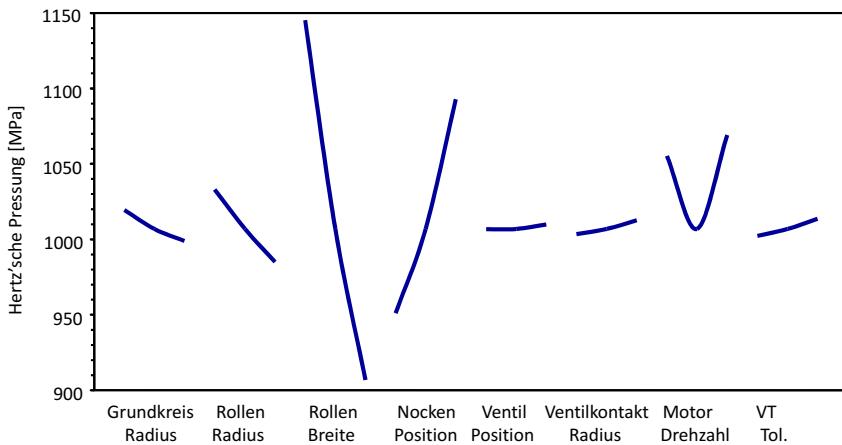


Abb. 6.13 Grafische Darstellung der Effektberechnung für das Qualitätsmerkmal: Hertz'sche Pressung zwischen Nocken und Rolle.

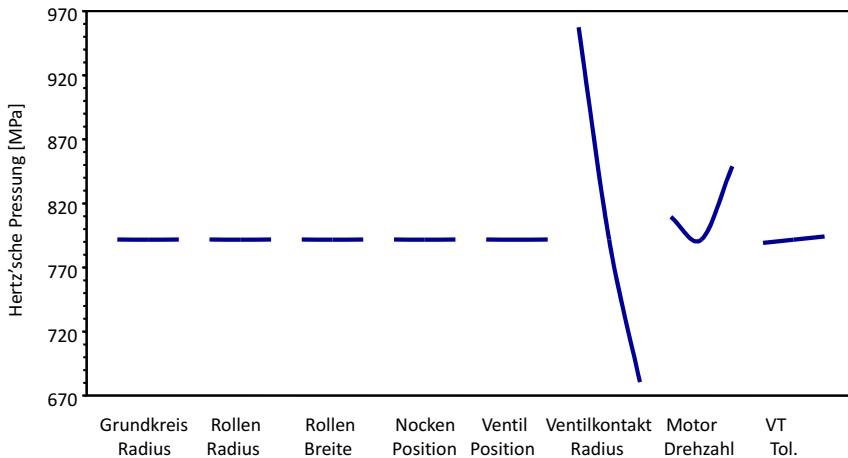


Abb. 6.14 Effekte für das Qualitätsmerkmal: Hertz'sche Pressung zwischen Schlepphebel und Ventil.

warten können. Interessant sind vor allem die Wechselwirkungen zwischen Steuergrößen und Störgrößen bzw. Signalgrößen. In diesem Fall sind dies die Faktoren G (Motordrehzahl) und H (kombinierte Störgröße). Hier zeigt sich, welchen Einfluss die Signal- und Störgrößen besitzen, in Abhängigkeit von der Einstellung der Steuergrößen.

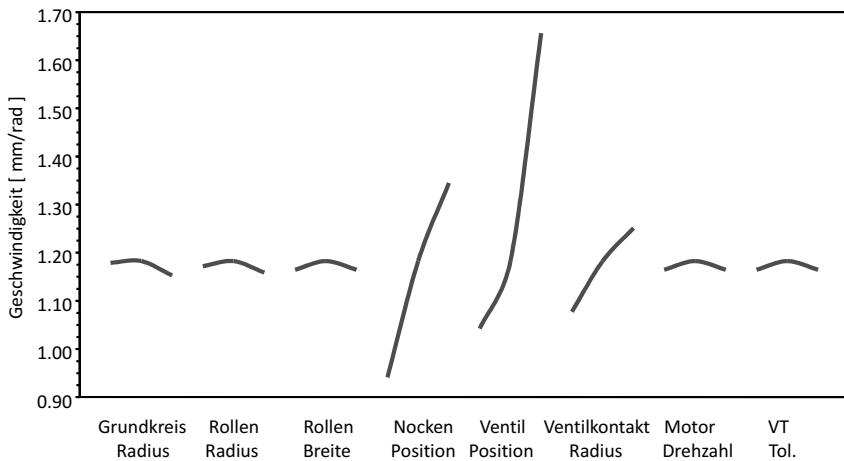


Abb. 6.15 Effekte für das Qualitätsmerkmal: Gleitgeschwindigkeit zwischen Schlepphebel und Ventil.

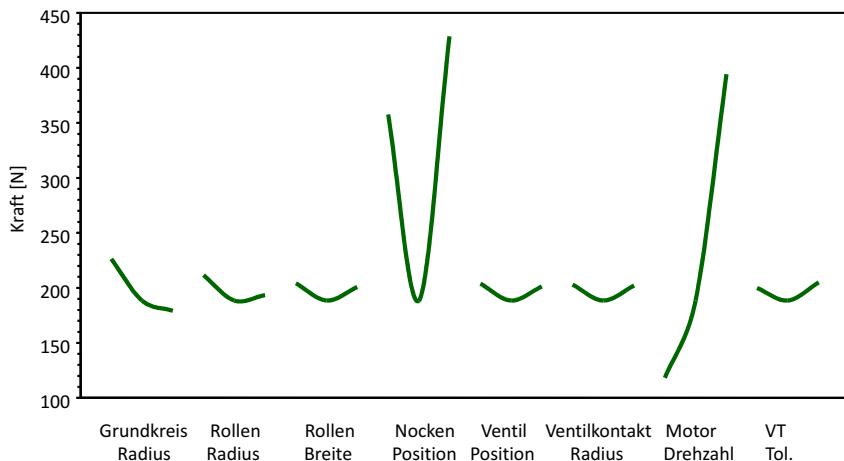


Abb. 6.16 Effekte für das Qualitätsmerkmal: Seitenkraft am Spielausgleich.

So informativ die Einzelanalysen der Qualitätsmerkmale auch sein mögen, letztendlich wird immer die gemeinsame Optimierung aller Qualitätsmerkmale für eine ausgereifte Konstruktion nötig sein. Blindes Vertrauen in den Optimierungsalgorithmus kann allerdings zu falschen Schlussfolgerungen führen. Zunächst empfiehlt sich eine Analyse der Zusammenhänge zwischen den Qualitätsmerkmalen. Hierbei hilft die Principal-Component-Analysis (PCA). Der Scree-Plot zeigt, dass mindestens drei Komponenten eine Rolle spielen, eher vier. Für sich alleine genommen, wäre dies ein Hinweis auf ein recht komplexes System. Allerdings soll das fünfte

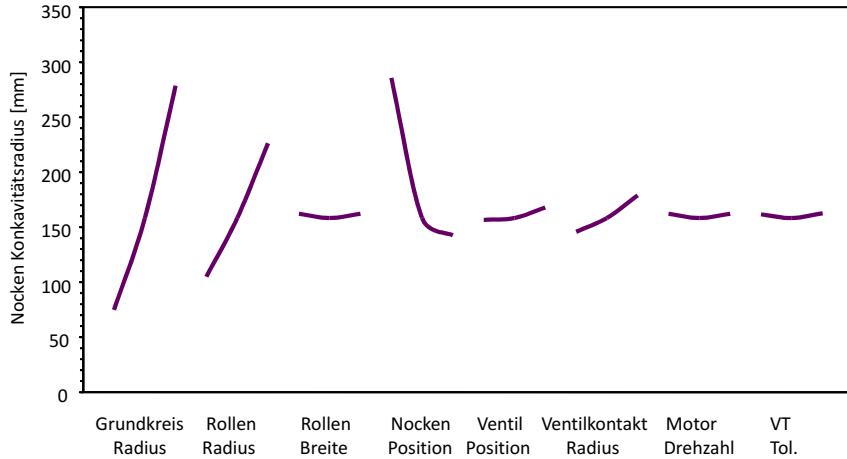


Abb. 6.17 Effekte für das Qualitätsmerkmal: Konkavität des Nockens.

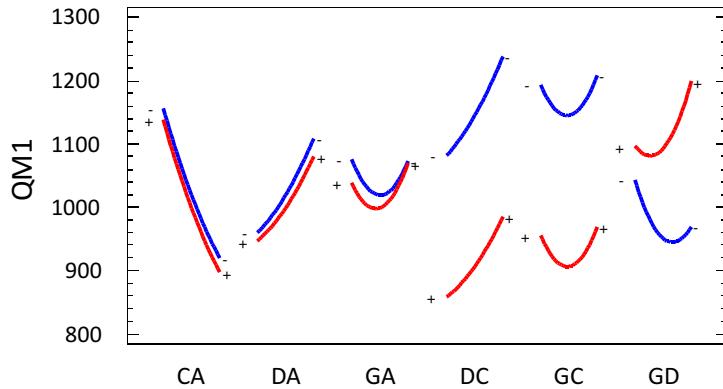


Abb. 6.18 Wechselwirkungen für das Qualitätsmerkmal: Hertz'sche Pressung zwischen Nocken und Rolle.

Qualitätsmerkmal (Konkavität) im Gegensatz zu allen anderen Qualitätsmerkmalen maximiert werden. Daher kann man im Geiste die Richtung des Zeigers für dieses Qualitätsmerkmal im Biplot umkehren. Dann ziehen grob gesehen alle Qualitätsmerkmale in eine Richtung. Das lässt auf ein gutes Optimierungspotential hoffen.

6.3.4 Optimierung

Bei der Kombination der Qualitätsmerkmale geht es darum, die mehrdimensionale Systembewertung auf eine einzige Dimension zu reduzieren. Erst dann kann der

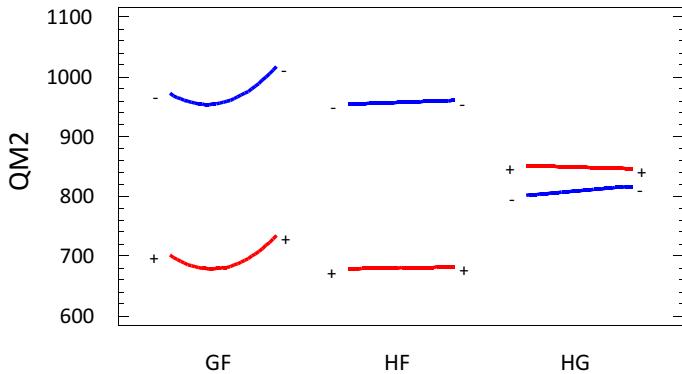


Abb. 6.19 Wechselwirkungen für das Qualitätsmerkmal: Hertz'sche Pressung zwischen Schlepphebel und Ventil.

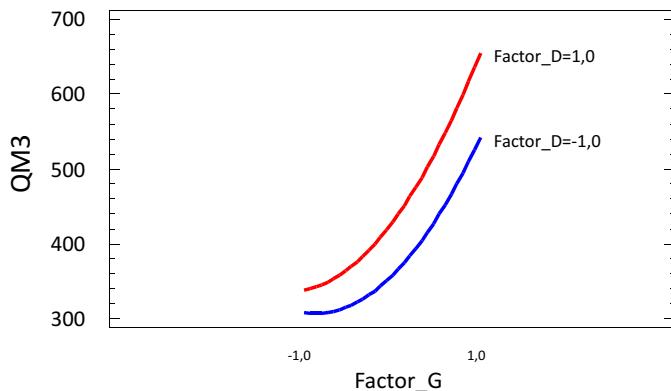


Abb. 6.20 Wechselwirkungen für das Qualitätsmerkmal: Seitenkraft am Spielausgleich.

Optimierungsalgorithmus automatisiert im gegebenen Suchraum zum Optimum finden. Hierzu werden die Qualitätsmerkmale über Rampenfunktionen dimensionslos gemacht. Die Rampenfunktionen liefern Funktionswerte zwischen 0 und 1, wobei 0 ein sehr schlechtes Ergebnis darstellt und 1 ein sehr gutes Ergebnis, im Sinne der Systembewertung. Hier ist Augenmaß gefragt, denn eine zu strenge Definition der Rampen stellt den Optimierungsalgorithmus vor eine unlösbare Aufgabe. Sofern man keine konkrete Anforderung bezüglich einzelner Qualitätsmerkmale hat, ist die einfachste Vorgehensweise die, von den verfügbaren Kombinationen auszugehen und die jeweiligen Extremwerte als Referenzwerte für die Rampen festzulegen. Die dimensionslosen Kennwerte der einzelnen Qualitätsmerkmale werden miteinander multipliziert. So entsteht die eindimensionale Bewertungsfunktion, Desirability Function. Der Optimierungsalgorithmus wird dann versuchen, eine einzige Einstellung zu finden, die gleichzeitig für jedes Qualitätsmerkmal den bekannten

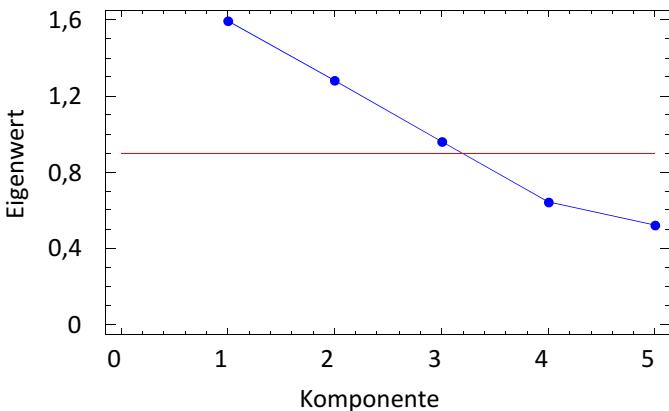


Abb. 6.21 Scree-Plot der Principal Component Analysis.

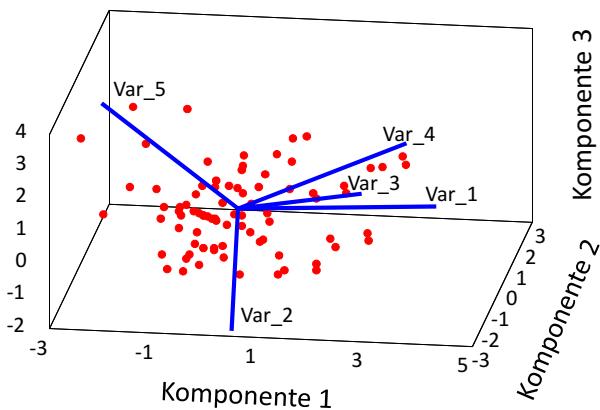


Abb. 6.22 Komponentendarstellung im Biplot.

Bestwert schlägt. Da die bekannten Bestwerte für die jeweiligen Qualitätsmerkmale typischerweise bei verschiedenen Einstellungen auftreten, ist diese Aufgabe anspruchsvoller, als sie zunächst erscheinen mag.

Wie geht man nun mit den Signal- und Störgrößen um? Es würde wenig Sinn ergeben, eine optimale Einstellung dieser Größen zu ermitteln, denn in der Realität werden genau diese Größen variieren. Im gezeigten Beispiel wurde deshalb zunächst von einer mittleren Einstellung ausgegangen, um ein Optimum für die Steuergrößen zu erhalten. Dann folgte eine Sensitivitätsanalyse, um die Streuung der Systemleistung in Folge der Variation der Signal- und Störgrößen abzuschätzen. Im gefundenen Optimum besteht immer eine Sensitivität in Bezug auf die Einstellungen der Faktoren. Hier sind die Faktoren G und H gesondert zu betrachten, weil deren Werte im praktischen Gebrauch des untersuchten Systems schwanken werden.

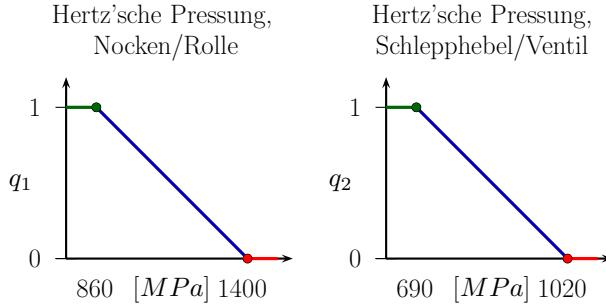


Abb. 6.23 Rampenfunktionen zur Multiple-Response-Optimisation, q_1 und q_2 .

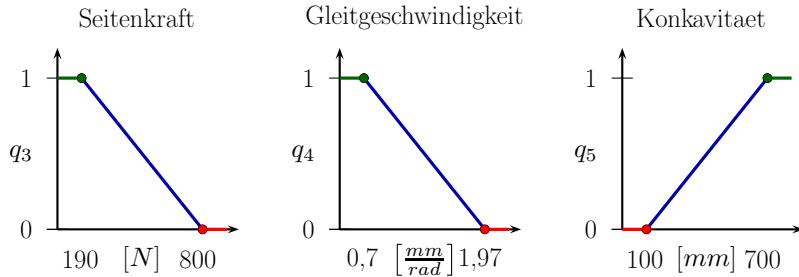


Abb. 6.24 Rampenfunktionen zur Multiple-Response-Optimisation, q_3 , q_4 und q_5 .

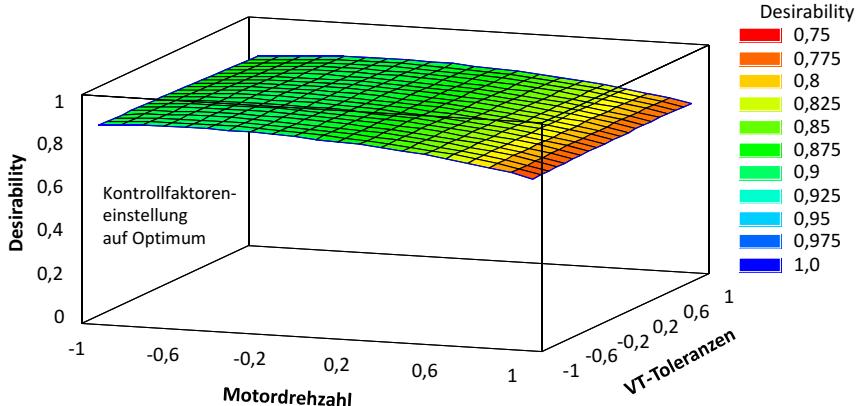


Abb. 6.25 Sensitivität im Optimum der Steuergrößen.

Die gefundene Sensitivität ist jedoch moderat, also ist das vorgeschlagene Optimum verwendbar.

Bei der Multiple-Response-Optimisation macht es Sinn, verschiedene Szenarien zu betrachten. Es gibt immer lokale Optima. Im Falle eines hohen Optimierungspotentials können die lokalen Optima sehr nahe an das globale Optimum kommen.

Die Einstellungen der Faktoren können dabei jedoch stark voneinander abweichen. Mit anderen Worten: Es gibt mitunter völlig verschiedene Einstellungen, die vergleichbar gute Ergebnisse im Sinne der Systembewertung liefern. Szenarienanalysen bestehen darin, den Spielraum einzelner Faktoren gezielt einzugehen, um den Optimierungsalgorithmus gezielt in ein lokales Optimum zu steuern. Hierbei wird man kostengünstige Einstellungen der Faktoren bevorzugen. Dieser Vorgang ist iterativ und erfordert die Zusammenarbeit der Fachleute bei der Festlegung der Szenarien. Allerdings ist die Berechnung sehr schnell und kann während der Besprechung erfolgen.

Abschließend seien noch die erzielten Ergebnisse gezeigt, hierbei bewusst der „worst case“, also das jeweils schlechteste Ergebnis in Abhängigkeit der Variation von Faktor G und Faktor H, bei konstanter Einstellung der Steuergrößen im vorgeschlagenen Optimum. Im Vergleich zur Spanne der Ergebnisse aus dem ursprünglichen Versuchsplan zeigt sich für alle Qualitätsmerkmale ein sehr gutes Ergebnis. Beim Qualitätsmerkmal Seitenkraft ergab sich eine moderate Verbesserung. Dies ist jedoch vertretbar, weil der erreichte Wert konstruktiv gut beherrschbar ist und nur bei sehr hohen Motordrehzahlen vorkommt. Abschließend wurden mit der vorgeschlagenen Einstellung Bestätigungsläufe gerechnet, die eine gute Übereinstimmung mit den prognostizierten Werten zeigten. Die Optimierung ist somit gelungen.

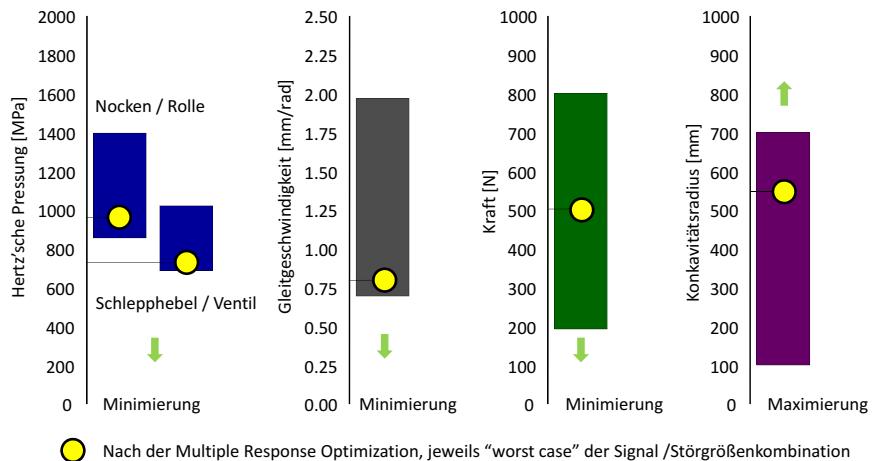


Abb. 6.26 Ergebnis der Multiple-Response-Optimisation.

Literaturverzeichnis

1. Lach, R., Kien, R., Hans, C., Siebertz, K., Platzbäcker, W.: *Design of Experiments (DoE) in der Motorenentwicklung*, chap. DoE Application within Analytical Valve Train Development, pp. 14–27. Expert Verlag, Renningen (2005) 159
2. Neuenhaus, D., Siebertz, K.: *Multibody System Simulations to Analyze Limit States in Structural Engineering using Experimental Design*. In: 1st Joint International Conference on Multibody System Dynamics, May 25-27, 2010, Lappeenranta, Finland. (2010) 159
3. Neuenhaus, D., Siebertz, K.: *Multibody Dynamics and Multiple Response Optimization - A Strong Team for Robust Design in Structural Engineering*. In: Multibody Dynamics 2011, ECCOMAS Thematic Conference, J.C. Samin, P. Fisette (eds.), Brussels, Belgium, 4-7 July 2011. (2011) 159
4. Rutten, K., Siebertz, K., van Bebber, D., Hochkirchen, T., Ketelaere, B.D.: *The Garden Sprinkler: An Interactive Web-Based Application For Teaching Response Surface Methodology*. In: European Network for Business and Industrial Statistics, ENBIS-13. Ankara (2013) 10, 159

Kapitel 7

Computer-Experiment

7.1 Einleitung

Seit der Entwicklung des Z3 von KONRAD ZUSE im Jahr 1941, welcher von vielen Wissenschaftlern als erster funktionstüchtiger Computer angesehen wird, steigen die Rechenleistung und Speicherkapazität von Computern kontinuierlich an. In den letzten Jahrzehnten ist für die Entwicklung technischer Systeme der Einsatz von Computern unverzichtbar geworden. Die Anwendungsmöglichkeiten erstrecken sich dabei von Prüfstandssteuerung, Datenverwaltung und -analyse bis hin zur Simulation komplexer Systeme. Ziel des Einsatzes von Computersystemen während des Entwicklungsprozesses ist unter anderem die Verbesserung der Produktqualität wie zum Beispiel die Verringerung der Ausfallwahrscheinlichkeit oder die Erhöhung der Effizienz. Parallel dazu werden Entwicklungszeiten und -kosten durch den gezielten Einsatz von Computersystemen verringert. Vergleichbar mit Versuchsplungen für physikalische Experimente existieren verschiedenste statistische Methoden, die einen effizienten Einsatz von Computer-Ressourcen im Entwicklungsprozess unterstützen. Eine grundlegende Einführung zu dem Thema *Computer-Experiment* bieten folgende Literaturstellen [4, 5, 3, 1].

7.2 Aufbau und Analyse von Computer-Experimenten

In vielen Arbeitsgebieten weisen die zu analysierenden und zu optimierenden Systeme eine steigende Komplexität auf, wobei sich diese einerseits in der wachsenden Faktoranzahl und andererseits in den umfangreicher abzubildenden Zusammenhängen zwischen Faktoren und Systemantworten widerspiegelt. Simulationsmodelle (im weiteren auch *Modelle* genannt) dieser Systeme sind daher in der Regel ebenfalls komplex und weisen lange Simulationszeiten auf. Die Rechenleistung der zur Verfügung stehenden Computersysteme steigt zwar kontinuierlich an, jedoch entwickeln sich die eingesetzten Modelle ebenfalls weiter, was einerseits zur

Erhöhung der Simulationsgüte führt, jedoch andererseits die durchschnittliche Rechenzeit ansteigen lässt. Aus diesem Grund ist die effiziente Ausnutzung der vorhandenen Computer-Ressourcen auch heute und in naher Zukunft noch immer eine entscheidende Grundvoraussetzung für den sinnvollen Einsatz von Computer-Experimenten. Die Komplexität der eingesetzten Modelle macht eine direkte analytische Lösung der betrachteten Aufgabenstellungen (zum Beispiel Leistungsoptimierung eines technischen Systems) meist unmöglich. Weiterhin ist eine direkte Optimierung der betrachteten Systeme mittels klassischen Optimierungsverfahren durch lange Rechenzeiten inakzeptabel. Für die meisten Entwicklungsprozesse werden daher nicht die Originalmodelle sondern Ersatzmodelle (Approximationsmodelle, Transfermodelle, Metamodelle, Transferfunktionen) verwendet, die das technische System beziehungsweise das komplexe Simulationsmodell ausreichend genau beschreiben und gleichzeitig nur minimale Computer-Ressourcen erfordern. Vergleichbar zu physikalischen Experimenten wird dabei das Approximationsmodell aus Daten erzeugt, die mittels eines Versuchsplans aus dem zu untersuchenden Simulationsmodell ermittelt wurden. Dadurch wird ein vereinfachtes Ersatzmodell des komplexen Simulationsmodells erstellt, welches lediglich die signifikanten Effekte beinhaltet. Diese Modelle werden in der Literatur meist als Metamodelle oder Transferfunktionen bezeichnet [2].

Abbildung 7.1 zeigt schematisch ein typisches Vorgehen bei Computer-Experimenten (CE) zur Verbesserung oder Neuentwicklung technischer Systeme.

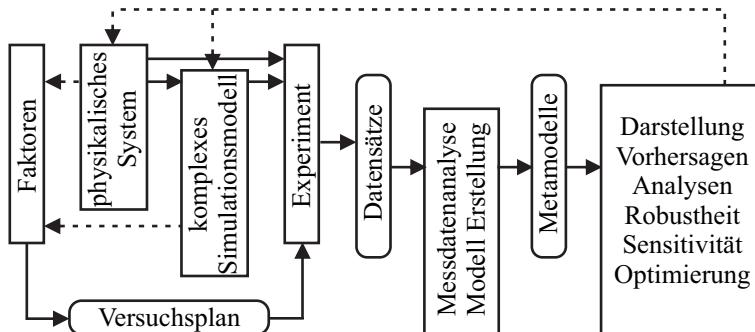


Abb. 7.1 Computer-Experiment

Im ersten Schritt wird basierend auf physikalischen Zusammenhängen und Erfahrungen technischer Experten ein komplexes Simulationsmodell aufgebaut. Abhängig vom aktuellen Entwicklungsziel und auf Basis bereits vorliegenden Fachwissens über das zu untersuchende System werden die Modellfaktoren in ihren Variationsgrenzen definiert. Mit Hilfe eines speziell für Computer-Experimente ausgelegten Versuchsplans (*Design of Experiment*) werden im Anschluss mittels des komplexen Simulationsmodells Datensätze erzeugt, die es ermöglichen ausreichend genaue Metamodelle für komplexe und nichtlineare Zusammenhänge zwischen Faktoren des Modells und den zu analysierenden Systemantworten zu ermitteln. Bevor

die benötigten Metamodelle erzeugt werden können ist es jedoch notwendig, die Datensätze auf ihre Qualität zu überprüfen (Kapitel 8.7). Mit den Metamodellen können anschließend Analysen des Systemverhaltens viel effizienter als mit dem Originalmodell oder dem physikalischen System selbst durchgeführt werden. Dabei können neben einfachen grafischen Darstellungen gesuchter Zusammenhänge zwischen Faktoren und Systemantworten auch Sensitivitätsanalysen bis hin zu Optimierungen mit Berücksichtigung der Robustheit gegenüber Störgrößen oder anderen Randbedingungen durchgeführt werden. Die Ergebnisse und Schlussfolgerungen dieser Analysen fließen anschließend in die Verbesserung des technischen Systems und wenn notwendig in das dazugehörige komplexe Simulationsmodell ein. Hierbei handelt es sich meist um einen iterativen Prozess, wobei die Anzahl der notwendigen Iterationen durch die gezielte Wahl von Versuchsplänen, Algorithmen zur Erstellung von Metamodellen, Analysen und Optimierungen minimiert wird.

7.2.1 Vergleich von Computer- und physikalischem Experiment

Versuchspläne, welche in den letzten Jahrzehnten speziell für physikalische Experimente entwickelt wurden, wie zum Beispiel Teilstufentests oder Central Composite Designs, sind zwar grundsätzlich auch für Computer-Experimente verwendbar, können jedoch nicht von den speziellen Eigenschaften der Computer-Simulationen profitieren. Der direkte Vergleich zwischen Computer- und physikalischen Experimenten zeigt einige unterschiedliche Eigenschaften welche in Tabelle 7.1 aufgelistet sind.

Computer	Physikalisch
deterministisch	stochastisch
keine Messfehler	Messfehler
hohe Faktoranzahl	geringe Faktoranzahl
flexible Stufenwahl	eingeschränkte Stufenwahl
einfache Änderung von Faktoren	Änderung von Faktoren oft schwierig oder zeitaufwendig

Tabelle 7.1 Vergleich von Computer- und physikalischem Experiment

Ein System beziehungsweise Simulationsmodell wird als *deterministisch* bezeichnet, wenn eine eindeutige Beziehung zwischen Eingangsdaten (Faktoren) und Ausgangsdaten (Gütekriterien, Systemantworten) vorliegt. Gleiche Eingangsdaten liefern immer identische Ausgangsdaten. Liefert ein System hingegen für gleiche Eingangsdaten unterschiedliche Ausgangsdaten (zum Beispiel durch Messstreuung), wird das System als *stochastisch* bezeichnet. Simulationsmodelle sind deterministisch, so dass Wiederholungsmessungen beziehungsweise -simulationen mit identischen Faktoren, wie sie in physikalischen Experimenten zur Dämpfung der Versuchsstreuung notwendig sind, in Computer-Experimenten unnötig sind.

Tendenziell ist festzustellen, dass in Computer-Experimenten mehr Faktoren in den Untersuchungen berücksichtigt werden als in physikalischen Experimenten. Dieses basiert unter anderem darauf, dass Faktoren in Simulationsmodellen leichter zu verändern oder kontrollieren sind als in physikalischen Versuchsaufbauten. Die Umgebungstemperatur oder die Eigenschaften einer eingesetzten Flüssigkeit großer technischer Systeme können im physikalischen Experiment nur schwer verändert werden, was in einem Simulationsmodell jedoch meist keine Schwierigkeit darstellt. Zusätzlich sind physikalische Experimente auf vorhandene oder kurzfristig lieferbare Bauteile beschränkt. Durchmesser von Rohrleitungen können beispielsweise kurzfristig nur in Standardmaßen geliefert werden, was gerade bei engen Zeitfestsätzen in der Entwicklung neuer Produkte die Spielräume in den Versuchsplänen stark einschränkt. Diese starken Einschränkungen treten bei Verwendung von Simulationsmodellen nicht auf.

Ein weiterer wichtiger Punkt im Vergleich zwischen Computer- und physikalischem Experiment ist die Veränderung von Faktoreinstellungen. Die Änderung eines Faktors in einem physikalischen Experiment erfordert oft einen Umbau der Prüfeinrichtung. Als Beispiel sei hier die Änderung eines Rohrdurchmessers genannt, welcher den Wechsel der zu verändernden Rohrleitung bedingt. Zur Minimierung des Versuchsaufwands in physikalischen Experimenten kommt daher der genauen und geschickten Versuchsplanung (zum Beispiel Versuchsreihenfolge) eine hohe Bedeutung zu. Bei der Versuchsreihenfolge im physikalischen Experiment sind darüber hinaus zusätzlich zeitliche Veränderungen des Systems (Drift) zu berücksichtigen, so dass Faktorvariationen randomisiert erfolgen sollten, was einer Sortierung der Versuchsreihenfolge zur Minimierung der Versuchsumbaus widerspricht. Beim Einsatz von Simulationsmodellen sind Überlegungen zur Versuchsreihenfolge von nur geringer Relevanz, da Faktoren leicht zu ändern sind und die zeitliche Reihenfolge der Simulation keinen Einfluss auf das Simulationsergebnis aufweist.

7.2.2 Testfelder für Computer-Experimente

Die im Kapitel 7.2.1 aufgeführten Eigenschaften von Computer-Experimenten ermöglichen es spezielle Testfelder zu entwickeln, welche Simulationsmodelle ideal ausnutzen und eine maximale Informationsmenge aus einer vorgegebenen Anzahl an Versuchen liefern. Abbildung 7.2a zeigt einen einfachen klassischen Voll faktorplan mit zwei Faktoren und 9 Versuchsläufen. Da nur die Stufen $-1, 0, 1$ besetzt sind, ergeben sich relativ große Bereiche im Inneren des Faktorraums, die keine Testpunkte enthalten, wodurch keine Informationen für diese Bereiche vorliegen werden. Weiterhin können *Pseudo*-Wiederholungen durch insignifikante Faktoren auftreten, die nicht zur Erhöhung der Informationsmenge beitragen. Betrachtet wird dazu der Fall, dass Faktor 2 keinen oder einen vernachlässigbaren Einfluss auf die Ausgangsvariable des betrachteten System aufweist (*z.B.* $y(x) = x_1^3$). Jeweils drei Testpunkte mit gleicher Faktorstufe für Faktor x_1 liefern das gleiche Ergebnis für y , da die Variation von x_2 kein Einfluss auf y aufweist (Abbildung 7.3). Die drei

Testpunkte ergeben jeweils für den Haupteffekt von x_1 auf y genauso viele Informationen wie ein einzelner Testpunkt. Es wurden somit zwei Pseudo-Wiederholungen durchgeführt, die trotz erhöhtem Testaufwand keine zusätzlichen Informationen für die Analyse des Zusammenhangs zwischen x_1 und y liefern. Zu beachten ist, dass keine Versuchsstreung durch Mittelwertbildung gedämpft werden muss.

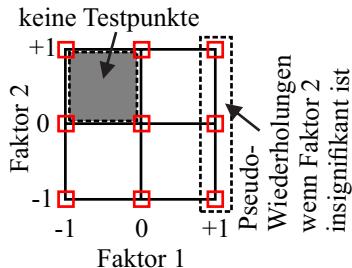


Abb. 7.2a Vollfaktorplan

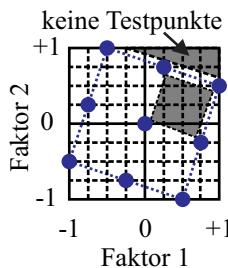


Abb. 7.2b LatinHypercube (LHC)

Abbildung 7.2b zeigt im Vergleich einen *Latin Hypercube* Testplan (Kapitel 8.3.3), welcher bei 2 Faktoren und 9 Testläufen einem rotierten Vollfaktorplan ähnelt. Im Gegensatz zum Vollfaktorplan wird hier jeder Faktor auf 9 äquidistante Stufen (gleich der Anzahl an Testpunkten) getestet, die jeweils nur einmal verwendet werden. Die Orthogonalität der Faktoren (Skalarprodukt der Spalten ist Null) ist in diesem Beispiel weiterhin gegeben, so dass die Haupteffekte sauber getrennt werden können. Die Testpunkte sind so verteilt, dass der Raum gleichmäßig ausgefüllt wird. In diesem Zusammenhang wird von einem *Spacefilling Design* oder gleichverteiltem Testfeld gesprochen. Im Vergleich zum Vollfaktorplan fällt der innere Bereich ohne Testpunkte deutlich geringer aus, so dass hier genauere Vorhersagen für die Ausgangsvariable y in Abhängigkeit der Faktoren getroffen werden können. Zu beachten ist jedoch, dass im dargestellten Beispiel keine Testpunkte in den Ecken des Faktorraums vorhanden sind, so dass bei Vorhersagen in diesen Bereichen auf meist ungenaue Extrapolationen zurückgegriffen werden muss. Pseudo Wiederholungen können im dargestellten Latin Hypercube nicht auftreten, da jede Faktorstufe jedes Faktors nur ein einziges Mal verwendet wird.

Abbildung 7.3 zeigt die unterschiedliche Datenmenge, die zur Bildung eines Metamodells durch die 9 Testpunkte bei Verwendung der beiden Testfelder zur Verfügung steht. Durch die Berücksichtigung des insignifikanten Faktors x_2 und die Wahl des Vollfaktorplans kann nur ein einfaches Modell für den Zusammenhang zwischen x_1 und y ermittelt werden. Durch die Verwendung des LHCs mit gleicher Anzahl an Versuchsläufen werden deutlich mehr Informationen über den Zusammenhang zwischen x_1 und y gewonnen. Gerade bei steigender Faktoranzahl werden häufig auch insignifikante Faktoren mitberücksichtigt, so dass durch die Verwendung speziell für Computer-Experimente optimierter Testfelder der Informationsgewinn im Vergleich zu klassischen Testfeldern deutlich gesteigert werden kann. Komplexe Systeme werden somit durch gleichen Versuchsaufwand deutlich genauer dargestellt.

Mittels des Vollfaktorplans wird in diesem Beispiel korrekter Weise kein Effekt von x_2 ermittelt. Bei Verwendung des LHCs kann je nach verwendetem Auswertealgorithmus ein geringer Einfluss von x_2 auf y ermittelt werden. Dieser wird jedoch im Verhältnis zum Effekt von x_1 klein bleiben. Im Vergleich zum Informationsgewinn für signifikante Faktoren (hier x_1) ist dies akzeptabel.

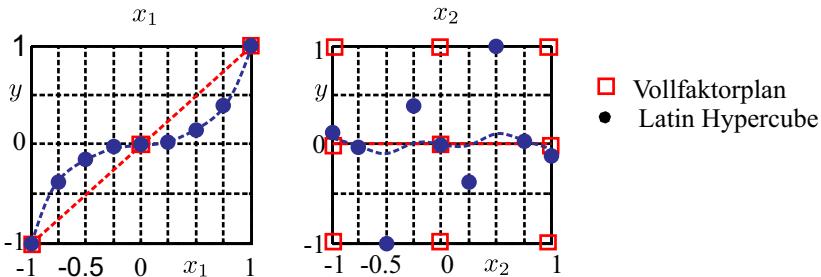


Abb. 7.3 Fallbeispiel zur Verwendung eines Vollfaktorplans oder LHCs

Insignifikante Faktoren

Insignifikante Faktoren treten in der Praxis deutlich häufiger auf als im allgemeinen vermutet. Aufgrund der Möglichkeit viele Faktoren mit einzubeziehen und einer kontinuierlich steigenden Komplexität der untersuchten Systeme, werden häufig einige zusätzliche Faktoren, bei denen nicht sicher ist ob sie notwendig sind, berücksichtigt. Weiterhin werden meistens mehrere Systemantworten mit einem Versuchsplan und einer Versuchsdurchführung analysiert und optimiert. Von den gewonnenen Daten werden Metamodelle für alle Systemantworten abgeleitet. Dabei weist nicht jeder Faktor einen Einfluss auf jede Systemantwort auf. Diese sogenannten insignifikanten Faktoren treten bei separater Betrachtung der einzelnen Systemantworten auf. Zur Veranschaulichung wird hier die Konstruktion eines fiktiven Tisches betrachtet, welcher als Designelement ein Loch des Durchmessers D und die Exzentrizität E aufweist (Abbildung 7.4). Insgesamt werden sechs Faktoren bei der Analyse berücksichtigt.

- A Anzahl N der Füße
- B Abstand R der Füße zum Mittelpunkt
- C Durchmesser D der Bohrung
- D Exzentrizität E der Bohrung zum Mittelpunkt
- E Oberflächenbeschaffenheit durch unterschiedliche Fertigungsverfahren
- F Dicke der Lackierung

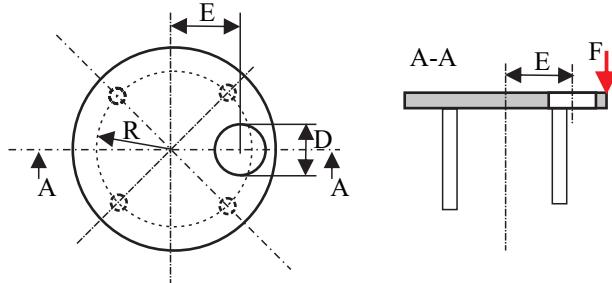


Abb. 7.4 Fallbeispiel zur Verwendung eines Vollfaktorplans oder LHCs

Verschiedene Systemkriterien sind im weiteren zu berücksichtigen:

- 1 Kippstabilität bei Kraft F am äußeren Rand
- 2 Gesamtgewicht
- 3 Rutschwiderstand der Tischoberfläche
- 4 Fertigungskosten
- 5 Materialkosten

Tabelle 7.2 zeigt eine erste Abschätzung, welche Faktoren einen signifikanten Einfluss auf die einzelnen Systemantworten aufweisen. Die Exzentrizität hat zum Beispiel auf kein Kriterium einen Einfluss und könnte somit ganz entfallen. Sollen eventuell im Anschluss noch weitere (bislang nicht betrachtete) Systemantworten berücksichtigt werden auf die Faktor D einen Einfluss haben könnte, wird dieser trotzdem weiter berücksichtigt. Die restlichen Faktoren weisen lediglich auf einen Teil der Systemantworten einen deutlichen Einfluss auf. Die Lackierungsdicke (F) hat zum Beispiel einen Einfluss auf den Rutschwiderstand, da bei größerer Dicke die Oberflächenstruktur des Fertigungsverfahrens (E) stärker geglättet wird. Der Einfluss auf das Gesamtgewicht sowie die Materialkosten ist hier im Verhältnis zu anderen Faktoren gering.

Tabelle 7.2 Faktoren und Systemkriterien

	A	B	C	D	E	F	signifikant	insignifikant
1	X	X					2	4
2	X		X		(X)		2-3	3-4
3				X	X		2	4
4	X			X			2	4
5	X	X			(X)		3	3

Bei Betrachtung einer einzelnen Systemantwort existiert keine, die von mehr als drei Faktoren abhängig ist. Auf die Kippstabilität wirkt sich zum Beispiel nur die Anzahl N der Füße sowie der Abstand R aus. Bei geschickter Planung des Versuchsplans kann im Endeffekt von drei anstatt sechs Faktoren ausgegangen werden (Kapitel 8).

7.2.3 Metamodelle

Basierend auf den gewonnenen Daten des Computer-Experiments wird ein Metamodell erzeugt, welches mit ausreichender Genauigkeit und minimaler Rechenzeit das komplexe Simulationsmodell beziehungsweise das zu analysierende System abbildet. Durch die speziell auf Computer-Experimente zugeschnittenen Testfelder ist ein Satz optimaler Informationen zur Erzeugung des Metamodells gegeben und sollte im Weiteren vollständig ausgenutzt werden. Im Vergleich zu physikalischen Experimenten werden durch Computer-Experimente meist komplexe Zusammenhänge erfasst, die vor der Analyse nicht oder nur teilweise bekannt sind. Aus diesem Grund werden Verfahren zur Metamodellerstellung benötigt, die sich flexibel und selbstständig (ohne Vorgabe von funktionellen Abhängigkeiten zwischen Faktoren und Systemantworten) an komplexe Zusammenhänge anpassen können. Verfahren welche einen vordefinierten Zusammenhang zwischen Faktoren und Ausgangsvariablen voraussetzen (wie beispielsweise die lineare Regression), können nur gute Vorhersagemodelle erzeugen, wenn das zu analysierende System den Annahmen entspricht. Da bei komplexen Simulationsmodellen der Zusammenhang zwischen Faktoren und Systemantworten (Qualitätsmerkmalen) jedoch typischerweise nicht bekannt ist, führt eine falsche Wahl der Grundfunktion schnell zu ungenauen Metamodellen. Um dieser Problematik aus dem Wege zu gehen, werden Verfahren wie beispielsweise Neuronale Netze oder Kriging eingesetzt, die sich eigenständig an gegebene Datenpunkte und Zusammenhänge anpassen (Kapitel 9.16 und 9.13).

7.2.4 Analyse und Optimierung

Sind ausreichend genaue Metamodelle mit geringen Rechenzeiten vorhanden, können umfangreiche Analysen in kurzer Zeit durchgeführt werden. Diese Analysen beginnen meist mit grafischen Darstellungen der Zusammenhänge zwischen Faktoren und Systemantworten (Abbildung 7.5a). Mittels der Metamodelle werden Systemantworten für geänderte Faktoreinstellungen in Millisekunden approximiert, so dass der Faktoreinfluss direkt (online) beobachtet und analysiert werden kann. Dadurch ist ein effizientes Hilfsmittel gegeben, welches einen schnellen Einblick und ein gutes Verständnis für Zusammenhänge zwischen Faktoren und Ausgangsvariablen liefert.

Neben allgemeinen Analysen zum Verständnis der Systemzusammenhänge werden ebenfalls Optimierungen durchgeführt (Kapitel 10). Hierbei sind typischerweise mehrere Zielgrößen zu berücksichtigen, die nicht unabhängig voneinander sind und sich widersprechende Ziele verfolgen. Neben klassischen Verfahren, die mehrere Zielgrößen beispielsweise mit Gewichtungsfaktoren zu einer globalen Zielgröße zusammenfassen, kommen gerade bei Computer-Experimenten Verfahren zur Bestimmung der *Pareto-Grenze* zum Einsatz. Abbildung 7.5b zeigt beispielhaft ein System mit zwei Qualitätsgrößen, die jeweils minimiert werden sollen. Jeder dargestellte Kreis ist genau ein mögliches Ergebnis, welches durch eine Kombination der

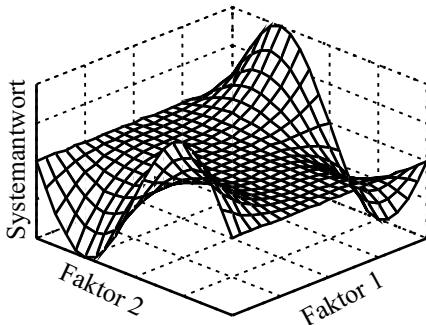


Abb. 7.5a Grafische Darstellung

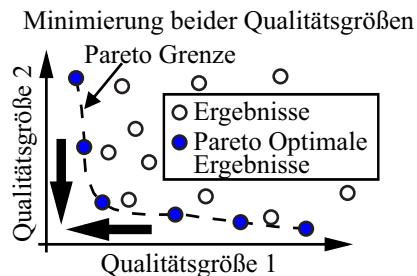


Abb. 7.5b Pareto-Grenze

gegebenen Faktoren erzielt wird. Bei Verwendung eines klassischen Optimierungsverfahrens, bei dem beide Zielgrößen (Z_1, Z_2) zu einer globalen Zielgröße Z_{global} zusammengefasst werden (zum Beispiel: $Z_{global} = Z_1 w_1 + Z_2 [1 - w_1]$ mit $0 \leq w_1 \leq 1$), ist das Ergebnis der Optimierung genau *ein* Punkt auf der Pareto-Grenze. Welcher Punkt auf der Pareto-Grenze gefunden wird, hängt dabei entscheidend von der Kombination der einzelnen Zielgrößen zur globalen Zielgröße ab. In unserem Beispiel wäre somit die Wahl des Gewichtungsfaktors w_1 entscheidend für das erzielte Ergebnis. In der Praxis wird dann mit dem ermittelten Ergebnis und den dazugehörigen Faktoreinstellungen weiter gearbeitet. Die Art der Zielgrößenkombination und die Wahl der Gewichtungen ist jedoch mehr oder weniger subjektiv und willkürlich, wodurch das erzielte Optimum und die dazugehörigen Faktoreinstellungen ebenfalls variabel sind. Falls die gefundenen Faktoreinstellungen aus einem Grund nicht akzeptabel sind (z.B. hoher Fertigungsaufwand oder hohe Kosten der benötigten Bauteile) kann ein weiteres Optimum auf der Pareto-Grenze nur gefunden werden, wenn die Art der Zielgrößenkombination oder die Gewichtungsfaktoren verändert werden und eine neue Optimierung durchgeführt wird.

Für jeden Punkt der Pareto-Grenze gilt, dass keine Zielgröße verbessert werden kann ohne eine andere zu verschlechtern. Diese Eigenschaft wird auch als *Pareto-optimal* bezeichnet. Ziel einer sinnvolleren Optimierung für mehrere Zielgrößen ist es eine Gruppe von Pareto-optimalen Punkten auf der Pareto-Grenze zu finden, die gleichmäßig auf der gesamten Pareto-Grenze verteilt sind. In der Praxis ergibt sich dadurch die Möglichkeit aus den gefundenen Pareto-optimalen Lösungen, welche nicht durch vordefinierte Gewichtungen eingeschränkt wurden, einen sinnvollen Kompromiss nicht nur zwischen den Zielgrößen zu wählen, sondern auch gleichzeitig die dazugehörigen Faktoreinstellungen zu betrachten. In der Praxis zeigt sich, dass sich durch kleine Verschiebungen auf der Pareto-Grenze deutlich bessere Faktoreinstellungen ergeben können, die zum Beispiel eine einfachere oder kostengünstigere Fertigung ermöglichen.

Diese und andere Analysen fließen zurück in das zu entwickelnde Produkt oder in die Verbesserung des Simulationsmodells. Auch wenn heutzutage physikalische Experimente und Prototypentests meistens nicht verzichtbar sind, wird der Entwick-

lungsprozess durch einen iterativen Einsatz von Computer-Experimenten verkürzt und das Ergebnis spürbar verbessert. Es zeigt sich gerade bei anschließenden physikalischen Experimenten, dass der Bereich für das physikalische Experiment durch vorgeschaltete Computer-Experimente auf enge und sinnvolle Bereiche eingegrenzt werden kann.

Literaturverzeichnis

1. Bates, R.A., Buck, R.J., Riccomagno, E., Wynn, H.P.: *Experimental Design and Observation for Large Systems*. Journal of the Royal Statistical Society, Series B (58), pp. 77–94 (1996) 179
2. Kleijnen, J.P.C.: *Statistical tools for simulation practitioners*. Marcel Dekker, New York (1986) 180
3. Koehler, J.R., Owen, A.B.: *Computer experiments*. In: Handbook of Statistics. Elsevier Science (1996) 179
4. Sacks, J., Schiller, S.B., Welch, W.J.: *Designs for computer experiments*. Tech. rep., Technometrics (1989) 179
5. Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P.: *Design and Analysis of Computer Experiments*. Statistical Sience 4, pp. 409–423 (1989) 179, 193, 205

Kapitel 8

Versuchspläne für komplexe Zusammenhänge

8.1 Einleitung

Die Komplexität von Systemen, die mit Hilfe statistischer Methoden analysiert und optimiert werden, steigt kontinuierlich an. Dies äußert sich einerseits in der zu berücksichtigen Faktanzahl und andererseits in der Komplexität abzubildender Zusammenhänge zwischen Faktoren und Systemantworten. Gerade der Einsatz von Simulationsmodellen in Kombination mit Computerexperimenten (Kapitel 7) eröffnet deutlich umfangreichere Aufgabenstellungen als es aus dem Bereich physikalischer Experimente bekannt war. Die Grundlage zur Bearbeitung der Analyse- und Optimierungsaufgaben muss eine Informationsmenge sein, die alle signifikanten Eigenschaften des betrachteten Systems enthält. Klassische Testfelder für physikalische Experimente mit wenigen Faktorstufen sind zur Ermittlung der benötigten Daten meist nicht ausreichend. Zur Erläuterung wird dazu eine unbekannte Funktion $y = f(x)$ betrachtet, von der drei Stützstellen bekannt sind (Abbildung 8.1, 'Kreise'). Aus mathematischer Sicht ist durch drei Punkte eine quadratische Funktion

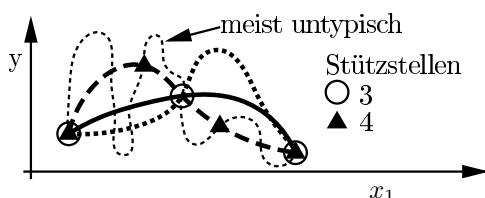


Abb. 8.1 Stützstellen zur Bestimmung einer unbekannten nichtlinearen Funktion

exakt definiert. In den meisten Anwendungsfällen kann jedoch von einem rein quadratischem Zusammenhang nicht ausgegangen werden. Neben der quadratischen Lösung existieren unendlich viele weitere Funktionszusammenhänge (Bild 8.1), welche durch die drei gegebenen Stützstellen verlaufen. In der Praxis zeigt sich, dass in den häufigsten Fällen zwar von komplexeren Zusammenhängen als linear

oder quadratisch ausgegangen werden muss, jedoch hoch komplexe Zusammenhänge nicht zu erwarten und somit zu vernachlässigen sind. Eine sinnvolle Vorhersage der zu beschreibenden Zusammenhänge wird durch eine leichte Erhöhung der Stützstellenanzahl (Faktorstufen) ermöglicht. Im dargestellten Beispiel ist bereits durch eine vierte Stützstelle (Abbildung 8.1, 'Dreiecke') der Zusammenhang zwischen x und y deutlich besser beschrieben. Da komplexe nichtlineare Modelle hauptsächlich bei Verwendung von Computer-Experimenten benötigt werden, können die speziellen deterministischen Eigenschaften, wie keine Versuchsstreung (siehe Kapitel 7), zur Konstruktion von Testfeldern Berücksichtigung finden. Bei der Auslegung eines Versuchsfelds (Testfelds) für Computer-Experimente wird grundsätzlich eine Gruppe von *ideal* verteilten Testpunkten $D_{n_r} = \{x_1, \dots, x_{n_r}\}$ aus einem Faktorraum der Dimension n_f (= *Faktanzahl*) gesucht. Die ideale Verteilung kann dabei je nach Anwendungsfall oder Vorkenntnissen unterschiedlich ausfallen. Sind keine Vorabinformationen vorhanden ist das erste übergeordnete Ziel der Testpunktverteilung eine exakte und robuste Vorhersage des globalen Mittelwerts der Funktion $y = f(\mathbf{x})$ zu erreichen, bei Berücksichtigung des gesamten Faktorraums C^{n_f} . Zur Vereinfachung der im Weiteren dargestellten Gleichungen und Herleitungen wird als Faktorraum der Einheitsraum $C^{n_f} = [0, 1]^{n_f}$ verwendet.

$$\frac{1}{n_r} \sum_{i=1}^{n_r} f(\mathbf{x}_i) = \int_{C^{n_f}} f(\mathbf{x}) \quad (8.1)$$

8.2 Gütekriterien für Testfelder

Vergleichbar zum Entwicklungsprozess technischer Systeme wird auch zur Optimierung von Testfeldern ein Qualitätskriterium benötigt. Neben grundsätzlichen Kriterien wie Orthogonalität oder Symmetrie wurden in unterschiedlichen Literaturquellen verschiedene Beurteilungskriterien für Testfelder vorgestellt, welche in den folgenden Kapiteln näher betrachtet werden.

Liegen grundlegende Informationen über das zu untersuchende System vor (z.B. Messdaten oder grundlegende Zusammenhänge zwischen Faktoren und Ausgangsvariablen), sollten diese in die Konstruktion der Testfelder integriert werden, wodurch speziell an die jeweiligen Vorabinformationen angepasste Prüfpläne entstehen. Eine Verallgemeinerung ist in diesen Fällen durch die Vielzahl an unterschiedlichen Vorabinformationen nicht möglich, so dass sich die folgenden Erläuterungen auf Anwendungen beziehen, bei denen nur geringe oder keine Informationen vorhanden sind. In diesen Fällen liegt das Augenmerk auf einem gleichmäßigen und maximalen Informationsgewinn im gesamten Faktorraum. In der Praxis zeigt sich, dass der Einsatz dieser Felder in vielen Anwendungen mit Vorabinformationen ebenfalls sinnvolle Ergebnisse liefert.

8.2.1 MiniMax und MaxiMin

JOHNSON et al. [31] schlägt in seiner Arbeit vor, den minimalen beziehungsweise maximalen Abstand zwischen allen Testpunkten als Qualitätskriterium zu verwenden. Dabei muss eine Abstandsdefinition $d(\mathbf{x}_i, \mathbf{x}_j)$ zweier Testpunkte $\mathbf{x}_{i,j}$ folgende Eigenschaften erfüllen:

$$\begin{aligned} d(\mathbf{x}_1, \mathbf{x}_2) &\geq 0 \\ d(\mathbf{x}_1, \mathbf{x}_2) &= d(\mathbf{x}_2, \mathbf{x}_1) \\ d(\mathbf{x}_1, \mathbf{x}_2) &\leq d(\mathbf{x}_1, \mathbf{x}_3) + d(\mathbf{x}_3, \mathbf{x}_2) \\ \text{mit } \mathbf{x}_{1,2,3} &\in C^{n_f} \end{aligned} \quad (8.2)$$

Diese Eigenschaften werden von der *euklidischen Norm* d_e erfüllt.

$$d_e(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{n_f} (x_{1l} - x_{2l})^2} \quad (8.3)$$

Aus Effizienzgründen wird häufig die Betrachtung des Quadrats d_e^2 (Absolutbetrag) vorgezogen, obwohl diese Formulierung die dritte Eigenschaft aus Gleichung 8.2 nicht garantiert. Ein eindimensionales Beispiel mit den Punkten $x_1 = 1, x_2 = 5$ und $x_3 = 3$ veranschaulicht die Verletzung der Bedingung.

$$\begin{aligned} d_e^2(x_1, x_2) &= 4^2 = 16 \\ d_e^2(x_1, x_3) &= 2^2 = 4 \\ d_e^2(x_3, x_2) &= 2^2 = 4 \\ d_e^2(x_1, x_2) &= 16 > 8 = d_e^2(x_1, x_3) + d_e^2(x_1, x_3) \\ d_e(x_1, x_2) &= 4 \leq 4 = d_e(x_1, x_3) + d_e(x_1, x_3) \end{aligned} \quad (8.4)$$

Im zweidimensionalen Fall ist die Forderung nur erfüllt, wenn der Winkel zwischen $\overrightarrow{\mathbf{x}_1\mathbf{x}_3}$ und $\overrightarrow{\mathbf{x}_3\mathbf{x}_2}$ kleiner oder gleich 90° ist (Abbildungen 8.2). Auf das klassische *MaxiMin*-Kriterium hat dies jedoch keinen Einfluss da zur Beurteilung der Testfelder lediglich der kleinste oder größte Abstand zweier Felder verglichen wird [28]. Sei nun D die Menge aller Abstände zwischen den Testpunkten $(\mathbf{x}_1 \dots, \mathbf{x}_{n_r})$ eines Testfelds T .

$$D = \{d(\mathbf{x}_i, \mathbf{x}_k)\} \quad 1 \leq i < k \leq n_r \quad (8.5)$$

Die Verteilung der Testpunkte im Faktorraum C^{n_f} wird als gleichmäßig angenommen (*space filling*), wenn der kleinste Wert von D maximiert wird. Das Ziel des sogenannten *MaxiMin*-Kriteriums ist es folglich, den Wert $\min(D)$ durch geschickte Wahl der Testpunkte zu maximieren. Entsprechend wird beim *MiniMax*-Kriterium der Maximalwert von D minimiert.

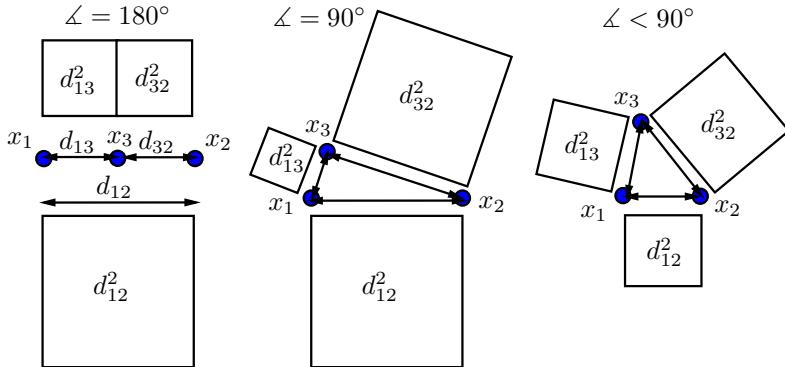


Abb. 8.2 Grafische Darstellung der Forderung $d(\mathbf{x}_1, \mathbf{x}_2) \leq d(\mathbf{x}_1, \mathbf{x}_3) + d(\mathbf{x}_3, \mathbf{x}_2)$

MORRIS et al. [49] hat einige Jahre nach der Veröffentlichung von Johnson das *MaxiMin*-Kriterium erweitert. Die Abstände D werden zuerst aufsteigend sortiert. Auf Basis der sortierten Abstände werden folgende Kennwerte ermittelt:

$$\begin{aligned}
 d_1 &\quad \text{kleinster Abstand in } D \\
 J_1 &\quad \text{Häufigkeit des Abstands } d_1 \text{ in } D \\
 d_2 &\quad \text{zweitkleinster Abstand in } D \\
 J_2 &\quad \text{Häufigkeit des Abstands } d_2 \text{ in } D \\
 &\dots
 \end{aligned} \tag{8.6}$$

Zur Erzeugung des Testfelds werden anschließen folgende Zielgrößen definiert und optimiert.

$$\begin{aligned}
 1a) &\text{ maximiere } d_1 \\
 1b) &\text{ minimiere } J_1 \\
 2a) &\text{ maximiere } d_2 \\
 2b) &\text{ minimiere } J_2 \\
 &\dots
 \end{aligned} \tag{8.7}$$

Die Verwendung eines einfachen skalaren Kriteriums ist in der Praxis einfacher als die in Liste 8.7 dargestellten Kriterien, so dass typischerweise das verallgemeinerte Kriterium aus Gleichung 8.8 minimiert wird [32, 18].

$$\text{MaxiMin}_p = \left[\sum_{k=1}^m \frac{J_k}{d_k^p} \right]^{1/p} \tag{8.8}$$

Dabei sind m und p ganzzahlige positive Werte, die vom Anwender zu wählen sind. Eine Erhöhung von p verringert den Einfluss von größeren Abständen auf das Gütekriterium. Sollen alle Abstände eines Testfelds betrachtet werden, so vereinfacht sich das *Minimax* _{p} -Kriterium zu Gleichung 8.9, was eine Sortierung der Abstände

in D und die Bestimmung von J_k überflüssig macht [32, 18].

$$\text{MaxiMin}_p = \left[\sum_{1 \leq i < k \leq n_r} d(\mathbf{x}_i, \mathbf{x}_k)^{-p} \right]^{1/p} \quad (8.9)$$

Im Gegensatz zum klassischen *MaxiMin*-Kriterium hat die Verwendung des quadrierten Abstands d_e^2 im Vergleich zum nichtquadrierten Abstands d_e einen Einfluss auf das Kriterium MaxiMin_p . Abbildung 8.3 zeigt die Korrelation zwischen beiden Varianten für $p = 2$ und 8 bei zufälligen orthogonalen Latin Hypercube Designs (LHD: Kapitel 8.3.3) mit $n_r = 65$ Datenpunkten und $n_f = 4$ Faktoren [28]. Bei kleinen Werten der Gütekriterien, was einem gesuchtem Testfeld mit guter Gleichverteilung entspricht, zeigt sich eine gute Korrelation. Erst im uninteressanten Bereich mit größeren Werten des Gütekriteriums steigt die Abweichung im dargestellten Beispiel an, was darauf schließen lässt, dass auch im erweiterten Kriterium beide Varianten grundsätzlich einsetzbar sind.

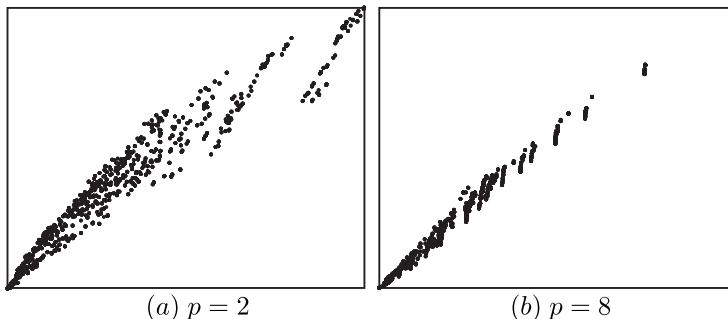


Abb. 8.3 Korrelation des MaxiMin_p -Kriteriums für quadrierte und nichtquadrierte Distanz bei unterschiedlichen Werten des Parameters p .

8.2.2 Entropie

Die Entropie ist nach SHANNON ein Maß für die Informationsmenge in einem Datensatz [63]. Die Maximierung der Entropie zur Erstellung eines optimalen Testplans wurde von SHEWRY und WYNN bereits 1987 vorgeschlagen [65]. In der Praxis wird zur Bestimmung eines optimalen Testfelds die Determinante der Korrelationsmatrix $\det(R)$ maximiert [61, 36]. Wird von einer Normalverteilung der gemessenen Daten ausgegangen, wird normalerweise die Korrelationsfunktion in Gleichung 8.10 verwendet.

$$r_{ij} = e^{-\sum_{k=1}^{n_f} \Phi_k (x_{ik} - x_{jk})^2} \quad (8.10)$$

Das Qualitätskriterium ist abhängig von der gewählten Korrelation (siehe auch Kapitel 9.13) und sollte zu den gemessenen Daten passen. Leider ist die Korrelation der Daten in den meisten Fällen erst nach der Vermessung bekannt. Soll im Anschluss ein Kriging Modell an die Daten angepasst werden, so ist es sinnvoll, die gleiche Korrelationsfunktion zu verwenden.

8.2.3 Gleichverteilung (Uniformity)

FANG et al. schlägt zur Beurteilung von Testfeldern verschiedene Kriterien vor, welche die Gleichverteilung (Uniformity) von Testpunkten im Faktorraum beschreiben [11, 74, 13, 16, 42, 12, 15, 18]. Eine Grundvoraussetzung für ein allgemeingültiges Gütekriterium ist dabei, dass es unabhängig von Vertauschungen von Faktorspalten oder der Versuchsreihenfolge ist. Zwei Testfelder X_1 und X_2 , die durch Vertauschungen von Spalten oder Zeilen ineinander überführt werden können, müssen daher als gleichwertig beurteilt werden [18].

Diskrepanz

Sei T ein Testfeld mit n_f Faktoren und n_r Testläufen sowie \mathbf{z} ein beliebiger Punkt aus dem Faktorraum C^{n_f} . Weiterhin gibt $N(T, [0, \mathbf{z}])$ die Anzahl aller Testpunkte \mathbf{x}_i des Testfeldes T an, die in den Raum $[0, \mathbf{z}]$ fallen (Bild 8.4a). Das Verhältnis der Anzahl N zur Gesamtanzahl der Testpunkte im Vergleich zum eingeschlossenen Volumen des n_f -dimensionalen (normierten) Raums ergibt die Diskrepanz (discrepancy) des Punkts \mathbf{z} .

$$\text{Diskrepanz}(\mathbf{z}) = \left| \frac{N(T, [0, \mathbf{z}])}{n_r} - \text{Vol}([0, \mathbf{z}]) \right| \quad (8.11)$$

Abbildung 8.4a zeigt beispielhaft drei Diskrepanzen eines Monte-Carlo-Testfelds mit zwei Faktoren (x_1, x_2) und 30 Testpunkten.

L_p -Diskrepanz

Die Mittlere L_p Norm der Diskrepanz im gesamten Faktorraum C^{n_f} wird als Stern (*star*)-Diskrepanz bezeichnet und dient als Maß für die Gleichverteilung der Testpunkte im Faktorraum C^{n_f} .

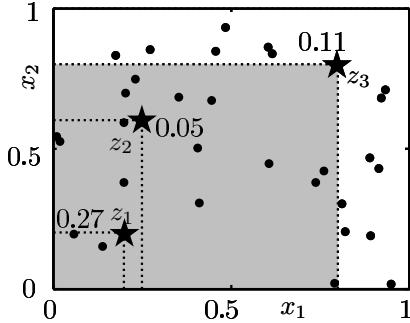


Abb. 8.4a Diskrepanz eines zweidimensionalen Monte-Carlo-Testfeldes

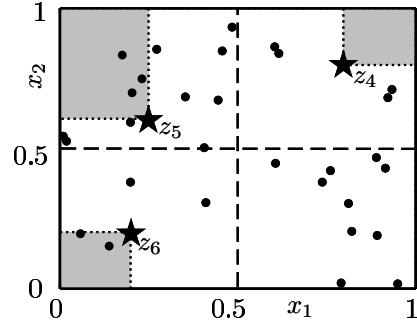


Abb. 8.4b Zentrierte L_2 -Diskrepanz

$$D_p(T) = \left\{ \int_{C^{n_f}} \text{Diskrepanz}(z)^p \right\}^{\frac{1}{p}} \quad (8.12)$$

Die Stern-Diskrepanz mit $p \rightarrow \infty$ spielt bei Quasi-Monte-Carlo-Methoden und in der Statistik eine große Rolle, ist jedoch schwierig zu berechnen. Algorithmen zur Berechnung oder Abschätzung der Diskrepanz mit $p \rightarrow \infty$ wurden von BUND-SCHUH et al. [6] sowie WINKLER et al. [78] vorgestellt. Es gilt vereinfacht:

$$D_\infty(T) = \max_{\mathbf{z} \in C^{n_f}} (\text{Diskrepanz}(\mathbf{z})) \quad (8.13)$$

Ist hingegen $p = 2$, so kann nach WARNOCK die L_2 -Diskrepanz für ein Testfeld T analytisch durch folgende Gleichung abgeschätzt werden [75]:

$$[D_2(T)] = 3^{n_f} \frac{2^{1-n_f}}{n_r} \sum_{i=1}^{n_r} \prod_{j=1}^{n_f} (1 - x_{ij}^2) + \frac{1}{n_r^2} \sum_{k=1}^{n_r} \sum_{i=1}^{n_r} \prod_{j=1}^{n_f} [1 - \max(x_{kj}, x_{ij})] \quad (8.14)$$

Die L_2 -Diskrepanz vernachlässigt Eigenschaften von Unterräumen des Faktorraums C^{n_f} und ist somit für Computer-Experimente nur eingeschränkt geeignet [18]. Weiterhin ist die allgemeine L_p -Diskrepanz nicht unabhängig von einer Rotation des Testfelds, da der Ursprung '0' eine besondere Bedeutung in der Berechnungsvorschrift einnimmt. Dieses erschwert einen aussagekräftigen Vergleich verschiedener Testfelder mit dem L_p -Kriterium.

Zentrierte und Einhüllende Diskrepanz

Zur Ermittlung verbesserter Diskrepanz-Kriterien wurden verschiedene Forderungen aufgestellt, die eine sinnvolle Beurteilung von Testfeldern sicher stellen [18]:

1. Invariant gegenüber Permutation von Spalten und/oder Zeilen
2. Invariant gegenüber der Rotation des Testfelds T
3. Berücksichtigung der Gleichverteilung über C^{n_f} und über jeden nichtleeren Unterraum C^u von C^{n_f}
4. Sinnvolle geometrische Bedeutung des Kriteriums
5. Leicht zu berechnen
6. Erfüllung der Koksma-Hlawka-Ungleichung (s.u.)
7. Das Kriterium ist widerspruchsfrei zu anderen Kriterien aus dem Bereich DoE

Koksma-Hlawka-Ungleichung: [50]

Ist f eine gleichmäßig verteilte Funktion mit endlicher Varianz $V(f)$ (im Sinne von Hardy und Krause), die in C^{n_f} definiert ist, und T eine Verteilung von Punkten in C^{n_f} , so gilt:

$$\left| \frac{1}{n_r} \sum_{i=1}^{n_r} f(\mathbf{x}_i) - E(f) \right| \leq V(f) D_\infty(T) \quad (8.15)$$

Basierend auf den aufgelisteten Bedingungen schlägt HICKERNELL [25, 26] verschiedene erweiterte Diskrepanz-Kriterien vor.

Sei R_z ein Rechteck, welches eindeutig durch die Wahl eines Punktes \mathbf{z} aus C^{n_f} definiert wird und die Bedingung (2) nicht verletzt. Sei weiterhin R_{z_u} die Projektion von R_z in den Unterraum C^{n_u} . Die modifizierte Diskrepanz, welche alle nicht leeren Unterräume berücksichtigt, ist dadurch allgemein mittels Gleichung 8.16 definiert:

$$[D_m(T)]^2 = \sum_{u \neq \emptyset} \int_{C^{n_u}} \left| \frac{N(T, R_{z_u})}{n_r} - \text{Vol}(R_{z_u}) \right|^2 du \quad (8.16)$$

Bei der **zentrierten L_2 -Diskrepanz** (ZD) wird ein Rechteck R_z zwischen dem Punkt \mathbf{z} und dem Eckpunkt des Faktorraums, der dem Punkt \mathbf{z} am nächsten liegt, aufgespannt (Abbildung 8.4b). Ist ein Testfeld T in C^{n_f} gegeben, lässt sich das ZD-Kriterium wie folgt berechnen [25]:

$$\begin{aligned} [ZD(T)]^2 &= \left(\frac{13}{12} \right)^{n_f} - \frac{2}{n_r} \sum_{i=1}^{n_r} \prod_{j=1}^{n_f} \left[1 + \frac{1}{2} |x_{ij} - 0.5| - \frac{1}{2} |x_{ij} - 0.5|^2 \right] \\ &\quad + \frac{1}{n_r^2} \sum_{i=1}^{n_r} \sum_{k=1}^{n_r} \prod_{j=1}^{n_f} \left[1 + \frac{1}{2} |x_{ij} - 0.5| + \frac{1}{2} |x_{kj} - 0.5| - \frac{1}{2} |x_{ij} - x_{kj}| \right] \end{aligned} \quad (8.17)$$

Anstelle des Rechtecks R_z zwischen einem Punkt \mathbf{z} und einer Ecke des Faktorraums kann ebenfalls ein Rechteck durch zwei beliebige Punkte \mathbf{z}_1 und \mathbf{z}_2 im Faktorraum aufgespannt werden (R_{z_1, z_2}) (Abbildung 8.5).

$$R_{z_1, z_2} = \begin{cases} [\mathbf{z}_1, \mathbf{z}_2) & \mathbf{z}_1 \leq \mathbf{z}_2 \\ [0, \mathbf{z}_2) \cup [\mathbf{z}_1, 1) & \mathbf{z}_1 > \mathbf{z}_2 \end{cases} \quad (8.18)$$

Durch die Verwendung des Rechtecks R_{z_1, z_2} wird die **Einhüllende Diskrepanz**

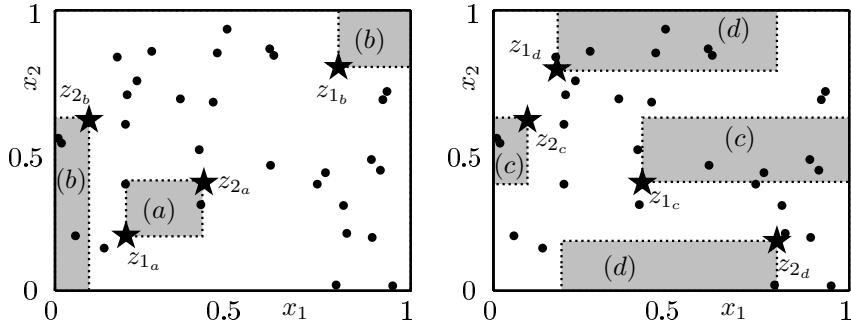


Abb. 8.5 Einhüllende Diskrepanz

(Wrap Around Discrepancy) definiert und kann für ein Testfeld T wie folgt berechnet werden [72, 26, 18]:

$$[ED(T)]^2 = -\left(\frac{4}{3}\right)^{n_f} + \frac{1}{n_r^2} \sum_{i=1}^{n_r} \sum_{k=1}^{n_r} \prod_{j=1}^{n_f} \left[\frac{3}{2} - |x_{ij} - x_{kj}| (1 - |x_{ij} - x_{kj}|) \right] \quad (8.19)$$

Die zentrierte (ZD) und die einhüllende (ED) Diskrepanz erfüllen beide die ersten sechs Bedingungen für verbesserte Diskrepanz-Kriterien [25, 26]. Weitere Diskrepanz-Kriterien finden sich in Arbeiten von HICKERNELL [26, 27].

8.2.4 Vergleich verschiedener Gütekriterien

Zum Vergleich der dargestellten Gütekriterien werden 1000 verschiedene LHDs (Kapitel 8.3.3) mit $n_r = 65$ Testpunkten und $n_f = 4$ Faktoren erzeugt und die dazugehörigen Gütekriterien grafisch gegenübergestellt (Abbildung 8.6). Zwischen den Kriterien *einhüllende Diskrepanz*, *zentrierte Diskrepanz*, *MiniMax_p* und der negativen *Entropie* ist eine mehr oder weniger deutliche Korrelation vorhanden. Wird ein Testfeld mit einem dieser Kriterien optimiert, ergeben sich zwar nicht zwangsläufig die besten Werte für die anderen Gütekriterien, jedoch kann davon ausgegangen werden, dass auch die anderen Kriterien dieses Feld positiv bewerten. Das einfache *MiniMax(MAX)*-Kriterium zeigt nur eine geringe Korrelation mit den restlichen Kriterien. Das *MiniMax* und *MaxiMin* Kriterium sollte nur mit Bedacht oder besser direkt das *MiniMax_p*-Kriterium verwendet werden. Besteht in der Praxis die Möglichkeit, mehrere Kriterien gleichzeitig zu verwenden, so ist das Testfeld zu wählen, welches für alle Kriterien gute Werte ergibt. Die Bandbreite der Gütekriterien kann durch eine Analyse der Kriterien-Streuung bei zufälligen Testfeldern abgeschätzt werden. Für die einhüllende Diskrepanz *ED* stellt FANG et al. in [22, 18] eine allgemeine Berechnungsvorschrift zur Bestimmung der unteren Grenze eines Testfeldes $U(n_r, n_s^{n_f})$ dar, die hier nicht weiter aufgeführt wird.

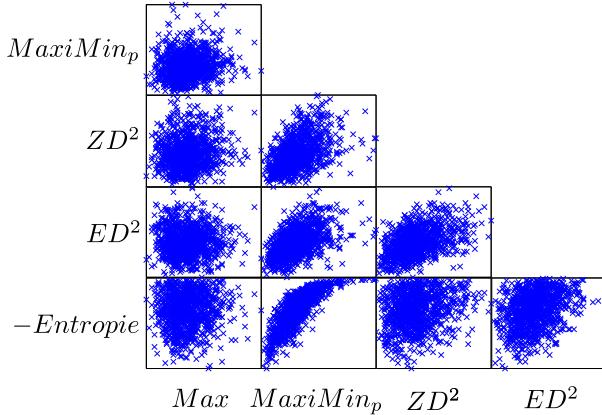


Abb. 8.6 Korrelation verschiedener Gütekriterien

8.3 Konstruktionsmethoden gleichverteilter Testfelder

Zur Erzeugung eines gleichverteilten Testfeldes T mit n_r Versuchsläufen und n_f Faktoren sind über die letzten Jahrzehnte unterschiedliche Algorithmen entwickelt worden. Verfahren zur Erzeugung von Testfeldern für Computer-Experimente gehen meist davon aus, dass keine oder nur wenig Vorinformationen über das zu untersuchende System bzw. Simulationsmodell vorhanden sind. Die Verteilung der Testpunkte im Faktorraum $C^{n_f} = [0, 1]^{n_f}$ werden daher so gewählt, dass in allen Bereichen des Faktorraums so viele Informationen wie möglich, bei gegebener Anzahl von Versuchsläufen, gewonnen werden. Zur Beurteilung der Qualität eines Testfelds werden dabei verschiedene Kriterien verwendet, welche in Kapitel 8.2 dargestellt sind.

Algorithmen zur Erzeugung von Testfeldern für Computer-Experimente profitieren von den speziellen Eigenschaften der Simulationsmodelle, wie keine Versuchsstreuung oder einfache Variation von Faktoreinstellungen (Kapitel 7). Das grundlegende Optimierungsziel eines Testfelds ist dabei die Minimierung der Varianz einer unverzerrten Vorhersage des globalen Mittelwerts der zu analysierenden Ausgangsvariablen y .

$$\bar{y}(T) = \frac{1}{n_r} \sum_{i=1}^{n_r} y(x_i) \quad (8.20)$$

$$E[\hat{y}(x)] = E[y(x)] \quad (8.21)$$

8.3.1 (Quasi) Monte-Carlo

Basierend auf dem *Gesetz der großen Zahlen*, welches in seiner einfachsten Form aussagt, dass sich die relative Häufigkeit eines Zufallsergebnisses immer weiter an dessen Wahrscheinlichkeit annähert je öfter das Zufallsexperiment durchgeführt wird, werden *Monte-Carlo*-Testfelder erzeugt. Für jeden Versuchslauf i und jeden Faktor j wird dabei ein zufälliger Wert im Definitionsbereich des betrachteten Faktors gewählt. Bei Verwendung des Einheitsraums C^{nf} gilt somit:

$$x_{ij} = \text{rand}(0, 1) \quad \text{rand : Zufallszahl (Gleichverteilung)} \quad (8.22)$$

Die Erzeugung von *echten* Zufallszahlen ist mittels eines Computers, der keine zufällige Variation kennt, nur bedingt möglich. Daher wurden verschiedene Algorithmen entwickelt, mit denen es möglich ist, sogenannte *Pseudo*-Zufallszahlen zu erzeugen (linearer und multiplikativer Kongruenzgenerator, Mersenne-Twister, usw.) [35, 57, 45, 55, 46, 59, 44]. Die erzeugten Zufallszahlen hängen dabei vom gewählten Algorithmus und einem gewählten Startwert (*seed*) ab. Bei gleichem Startwert und Algorithmus werden identische *Zufallszahlen* erzeugt, so dass häufig der Startwert in Abhängigkeit der aktuellen Zeit gewählt wird, um variierende Zufallszahlen zu erzeugen.

Bei Verwendung der Monte-Carlo-Methode wird eine große Anzahl von Versuchsläufen benötigt, da der absolute Fehler des vorhergesagten Erwartungswertes lediglich mit $\frac{1}{\sqrt{n_r}}$ gegen 0 konvergiert. Eine Halbierung des Fehlers wird somit durch eine Vervierfachung der Testpunkteanzahl erreicht. Wird im Vergleich dazu ein kartesisches Gitter in den Faktorraum gelegt und an jedem Gitterschnittpunkt ein Testpunkt plaziert, so kann der Fehler bereits mit $\frac{1}{n_r}$ gegen 0 konvergieren. Dieses deterministische Verfahren hat jedoch den Nachteil, dass vor dem Computer-Experiment die Feinheit des Gitters (Gitterabstand) festgelegt werden muss und typischerweise alle geplanten Testpunkte zur anschließenden Analyse benötigt werden. Gewünscht ist hingegen ein Monte-Carlo-Verfahren, welches den Faktorraum gleichmäßig und *zufällig* ausfüllt, wobei eine bessere Konvergenz als $\frac{1}{\sqrt{n_r}}$ erreicht wird. Der Faktorraum soll dabei am Anfang grob und bei steigender Anzahl der Testläufe kontinuierlich feiner abgetastet werden. Abbildung 8.7 zeigt im linken Teil eine Sequenz von 64 Testpunkten, welche gleichmäßig über den 2-dimensionalen Faktorraum verteilt sind. Die folgenden 64 Punkte der gleichen Sequenz füllen die Zwischenräume und verfeinern somit die vorhandene Struktur, wie im rechten Diagramm zu sehen ist. Obwohl es sich hier nicht um zufällige sondern deterministisch ermittelte Testpunkte handelt, werden die Verfahren im allgemeinen Sprachgebrauch als Quasi-Monte-Carlo-Methoden (oder auch *Low Discrepancy Procedures*) bezeichnet.

Halton- und Hammersley-Sequenz

Einen einfach zu berechnenden Algorithmus liefert die *Halton-Sequenz*, welche auf der *Van-der-Corput-Sequenz* basiert [59]. Jedem Faktor wird dabei eine unterschied-

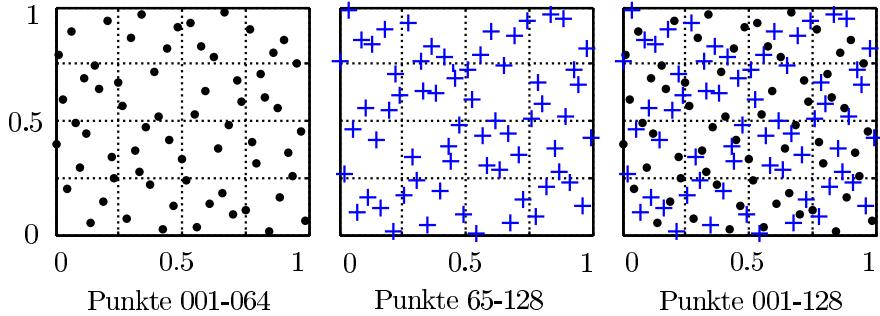


Abb. 8.7 Quasi-Monte-Carlo (Sequenz)

liche Primzahl ($2, 3, 5, 7, 11, 13, 17, 19, 23, 29, \dots$) [8] zugewiesen, welche im Folgenden die Zahlenbasis b_j für den jeweiligen Faktor j ist. Um den Wert eines Faktors j für den Testpunkt \mathbf{x}_i zu berechnen, wird der Wert i im Zahlensystem zur Basis b_j dargestellt.

$$i = \sum_{k=0}^{b_j-1} a_{ijk} b_j^k \quad a_{ijk} \in \{0, 1, \dots, b_j - 1\} \quad (8.23)$$

Der Zahlenwert für x_{ij} lässt sich anschließend durch Gleichung 8.24 ermitteln.

$$x_{ij} = \sum_{k=0}^{b_j-1} \frac{a_{ijk}}{b_j^{k+1}} \quad (8.24)$$

Ein Beispiel für zwei Faktoren und die Primzahlen $b_j = \{2, 3\}$ findet sich in Tabelle 8.1 und Abbildung 8.8.

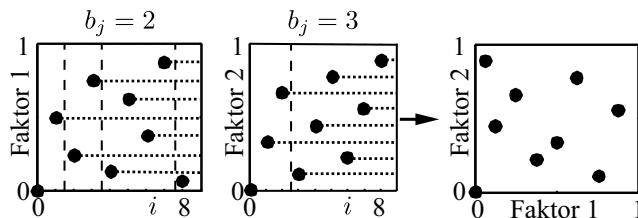


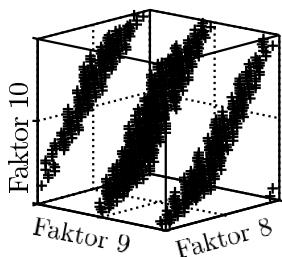
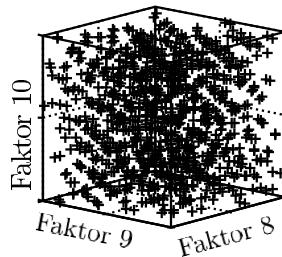
Abb. 8.8 Halton-Sequenz für $b_j = \{2, 3\}$

Die gewählte Primzahl b_j bestimmt den Wertebstand des Faktors j während des ersten *groben* Durchlaufs durch den Faktorraum $[0,1]$. Eine größere Primzahl führt dabei zu kleineren Abständen und zu einer größeren benötigten Anzahl an Testläufen n_r , um die dazugehörige Dimension zu durchlaufen. Bei gegebenen Primzahlen wird daher empfohlen, $n_r > \prod b_j$ zu wählen, wodurch die Anzahl von Testläufen bei

i	2^0	2^1	2^2	$1/2^1$	$1/2^2$	$1/2^3$	x_{ij}
$b_j = 2 \rightarrow$	0	0	0	0	0	0	0
	1	1	0	1/2	0	0	1/2
	2	0	1	0	1/4	0	1/4
	3	1	1	1/2	1/4	0	3/4
	4	0	0	1	0	1/8	1/8
	5	1	0	1	1/2	0	6/8
	6	0	1	1	0	1/4	3/8
	7	1	1	1	1/2	1/4	7/8
	:	:	:	:	:	:	:
i	3^0	3^1	3^2	$1/3^1$	$1/3^2$	$1/3^3$	x_{ik}
$b_k = 3 \rightarrow$	0	0	0	0	0	0	0
	1	1	0	1/3	0	0	1/3
	2	2	0	2/3	0	0	2/3
	3	0	1	0	1/9	0	1/9
	4	1	1	0	1/3	1/9	0
	5	2	1	0	2/3	1/9	0
	6	0	2	0	0	2/9	0
	7	1	2	0	1/3	2/9	0
	8	2	2	0	2/3	2/9	0
				:	:	:	:

Tabelle 8.1 Halton-Sequenz für 2 Faktoren

Erhöhung der Faktanzahl schnell ansteigt. Bei zu geringer Anzahl von Testpunkten ergeben sich Regionen im Faktorraum, welche durch die Sequenz noch nicht vollständig ausgefüllt sind, so dass wichtige Informationen aus diesen Bereichen fehlen. In Abbildung 8.9a ist zur Veranschaulichung für ein Testfeld mit 10 Faktoren und 1024 Testpunkten die Verteilung der Testpunkte für die Faktoren 8 bis 10 bei Verwendung einer Halton-Sequenz dargestellt.

Abb. 8.9a Halton-Sequenz ($n_f = 10$)Abb. 8.9b Sobol-Sequenz ($n_f = 10$)

Die *Hammersley-Sequenz* ersetzt die Werte eines Faktors einer Halton-Sequenz durch eine aufsteigende Zahlenfolge mit konstanten Abständen. Dadurch kann die

benötigte Anzahl an Primzahlen um eins reduziert werden, so dass bereits früher als in der Halton-Sequenz eine Gleichverteilung im Faktorraum erzielt wird.

Kronecker-Sequenz

Die *Kronecker-Sequenz* bezeichnet eine Zahlenreihe, welche durch die Nachkommastellen reeller Zahlen definiert ist. Der Ausdruck $\{\alpha\}$ steht dabei für die Nachkommastellen einer reellen Zahl α (z.B. $\{3.2341\} = 0.2341$). Bei gegebenem α ist die Kronecker-Sequenz definiert durch: $(\{1\alpha\}, \{2\alpha\}, \{3\alpha\}, \dots)$. Wird für jeden Faktor j eine unterschiedliche Primzahl b_j und $\alpha_j = \sqrt[b_j]{b_j}$ gewählt, so entspricht die Sequenz dem sogenannten *Torus-Algorithmus*. Bei Verwendung der Primzahl $b_j = 5 \rightarrow \alpha_j = \sqrt[5]{5} = 2.2361$ für Faktor j ergibt sich folglich die Sequenz zu: $0.2361, 0.4721, 0.7082, 0.9443, 0.1803, \dots$

Faure-Sequenz

Die *Faure-Sequenz* baut auf der Halton-Sequenz auf, verwendet jedoch nur eine Primzahl. Typischerweise wird die kleinste Primzahl b gewählt, welche größer oder gleich der Faktorenanzahl n_f ist. Die Werte des ersten Faktors x_{i1} entsprechen der Halton-Sequenz zur gewählten Basis b . Alle weiteren Faktoren j werden in Abhängigkeit des jeweiligen vorhergehenden Faktors $j - 1$ oder direkt in Abhängigkeit zum ersten Faktor bestimmt.

$$x_{ij} = \sum_{k=0}^{b_j-1} \frac{a_{ijk}}{b_j^{k+1}} \quad \text{mit } a_{ijk} = \left[\sum_{m \geq k}^{b_j-1} (j-1)^{m-k} \binom{m}{k} a_{i1m} \right] \bmod b_j \quad (8.25)$$

und $\binom{m}{k} = \frac{m!}{k!(m-k)!}$

$$\text{oder } a_{ijk} = \left[\sum_{m \geq k}^{b_j-1} \binom{m}{k} a_{i(j-1)m} \right] \bmod b_j \quad (8.26)$$

Bei drei Faktoren ($n_f = 3$) wird beispielsweise die Basis $b = 3$ gewählt. Die Werte für den ersten Faktor x_{i1} sind, wie in Tabelle 8.2 dargestellt, durch die Halton-Sequenz festgelegt.

Der Testpunkt $x_{9,3}$ mit $i = 9$ und $j = 3$ wird wie in Gleichung 8.27 dargestellt berechnet. Bei steigender Faktoranzahl zeigt sich am Anfang der Faure-Sequenz eine Häufung von Testpunkten in der Nähe des Ursprungs O . Aus diesem Grund werden häufig die ersten $(b^4 - 1)$ Punkte übersprungen [23], wodurch jedoch die Gefahr besteht, Bereiche des Faktorraums nicht gleichmäßig zu füllen.

TEZUKA führt eine *Verallgemeinerte Faure-Sequenz* (Generalized Faure Sequence)

$k \downarrow$	$i \rightarrow$	0	1	2	3	4	5	6	7	8	9	...
0	$a_{i1_0} \rightarrow$	0	1	2	0	1	2	0	1	2	0	...
1	$a_{i1_1} \rightarrow$	0	0	0	1	1	1	2	2	2	0	...
2	$a_{i1_2} \rightarrow$	0	0	0	0	0	0	0	0	0	1	...

$$x_{i1} \rightarrow | 0 \ 1/3 \ 2/3 \ 1/9 \ 4/9 \ 7/9 \ 2/9 \ 5/9 \ 8/9 \ 1/27 \ \dots$$

$$x_{i2} \rightarrow | 0 \ 1/3 \ 2/3 \ 4/9 \ 7/9 \ 1/9 \ 8/9 \ 2/9 \ 5/9 \ 16/27 \ \dots$$

$$x_{i3} \rightarrow | 0 \ 1/3 \ 2/3 \ 7/9 \ 1/9 \ 4/9 \ 5/9 \ 8/9 \ 2/9 \ 13/27 \ \dots$$

Tabelle 8.2 Faure-Sequenz

ein, welche ebenfalls auf der Van-der-Corput-Sequenz aufbaut, allerdings Polynome zur Anordnung der Faktorwerte verwendet [73, 58].

$$\begin{aligned} k = 0 : m = 0 : a_{9,3_0,m=0} &= 2^0 \begin{pmatrix} 0 \\ 0 \end{pmatrix} a_{9,1_0} = 1 \cdot 1 \cdot 0 = 0 \\ m = 1 : a_{9,3_0,m=1} &= 2^1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} a_{9,1_1} = 2 \cdot 1 \cdot 0 = 0 \quad \left. \right\} a_{9,3_0} = [0 + 0 + 4] \bmod 3 = 1 \\ m = 2 : a_{9,3_0,m=2} &= 2^2 \begin{pmatrix} 2 \\ 0 \end{pmatrix} a_{9,1_2} = 4 \cdot 1 \cdot 1 = 4 \\ k = 1 : m = 1 : a_{9,3_1,m=1} &= 2^0 \begin{pmatrix} 1 \\ 1 \end{pmatrix} a_{9,1_1} = 1 \cdot 1 \cdot 0 = 0 \\ m = 2 : a_{9,3_1,m=1} &= 2^1 \begin{pmatrix} 2 \\ 1 \end{pmatrix} a_{9,1_2} = 2 \cdot 2 \cdot 1 = 4 \quad \left. \right\} a_{9,3_1} = [0 + 4] \bmod 3 = 1 \\ k = 2 : m = 2 : a_{9,3_2,m=2} &= 2^0 \begin{pmatrix} 2 \\ 2 \end{pmatrix} a_{9,1_2} = 1 \cdot 1 \cdot 1 = 1 \quad \left. \right\} a_{9,3_2} = [1] \bmod 3 = 1 \\ \Rightarrow x_{9,3} &= \frac{a_{9,3_0}}{3} + \frac{a_{9,3_1}}{9} + \frac{a_{9,3_2}}{27} = \frac{1}{3} + \frac{1}{9} + \frac{1}{27} = \frac{13}{27} \end{aligned} \tag{8.27}$$

Sobol-Sequenz

Eine *Sobol-Sequenz* verwendet für alle Dimensionen (Faktoren) die Primzahlbasis $b_j = 2$ [69]. Der erste Faktor wird mittels der *Van-der-Corput-Sequenz* bestimmt. Alle anderen Faktoren werden durch Permutation des ersten Faktors erzeugt. Das Ergebnis der Sobol-Sequenz wird dabei von der Wahl verschiedener *primitiver* Polynome für den Permutationsalgorithmus beeinflusst. Es zeigt sich, dass die Sobol-Sequenz neben der verallgemeinerten Faure-Sequenz eine sehr gute Gleichverteilung aufweist (Abbildung 8.9b). In der Literatur und im Internet finden sich verschiedene Algorithmen, mit denen bereits Sobol-Sequenzen bis 2500 Dimensionen erzeugt wurden [59, 69, 2, 23, 66, 7].

Alle dargestellten Sequenzen sind Spezialfälle der sogenannten (t,m,s) -Netze und (t,s) -Sequenzen, welche 1992 durch NIEDERREITER beschrieben wurden [50, 62].

Hybrid-Quasi-Monte-Carlo

Unter den Begriff *Hybrid-Quasi-Monte-Carlo* Methoden wird eine Mischung aus Quasi-Monte-Carlo-Methoden und *zufälligen* Permutationen zusammengefasst. So können beispielsweise die Einstellungen für den ersten Faktor mit der Van der Corput-Sequenz bestimmt werden, und alle weiteren Faktoren sind lediglich zufällige Permutationen des ersten Faktors. Da bei diesem Verfahren nur die Basis $b = 2$ verwendet wird, weisen die erzeugten Testfelder meist auch in höheren Dimensionen eine gute Gleichverteilung auf.

8.3.2 Orthogonale Testfelder

Orthogonale Testfelder (*Orthogonal Arrays*) $OA(n_r, n_s, s, t)$ bezeichnen Felder mit n_r Versuchsläufen und n_s Faktoren, wobei jeder Faktor s Stufen aufweist (Gleichung 8.28). Die Stärke t gibt an, dass in beliebigen t Spalten (Faktoren) der $n_r \times n_s$ Matrix jedes mögliche t -Tupel der s Stufen vorkommt und gleichhäufig auftritt [52, 24, 3]. Im Internet finden sich umfangreiche Listen verschiedener orthogonaler Felder [67, 68, 38].

$$OA(9, 4, 3, 2) \rightarrow \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 \\ 0 & 2 & 2 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 2 & 0 \\ 1 & 2 & 0 & 2 \\ 2 & 0 & 2 & 2 \\ 2 & 1 & 0 & 1 \\ 2 & 2 & 1 & 0 \end{pmatrix} \quad \begin{array}{l} \text{neun zweier-Tupel mit 3 Stufen} \\ 00, 01, 02, 10, \dots \end{array} \quad (8.28)$$

Testfelder mit unterschiedlichen Stufen für verschiedene Faktoren werden als gemischte orthogonale Testfelder $OA(n_r, n_f, s_1^{k_1} \times s_2^{k_2} \times \dots \times s_l^{k_l}, t)$ (*mixed orthogonal arrays*) bezeichnet. Dabei geben die Parameter $k_1 \dots k_l$ die Häufigkeit der Stufen $s_1 \dots s_l$ an, so dass gilt: $n_f = \sum_i k_i$. Für jede verwendete Stufe s_j muss dabei gelten, dass n_r ein ganzzahliges Vielfaches von s_j ist ($n_r = s_j a$, $a \in \mathbb{Z}$). Für einige Kombinationen von Faktoren, Stufen und Versuchsläufen, wie beispielsweise $OA(36, 15, 3^6 4^8, 2)$ existieren keine OA. Für diese Felder können sogenannte fast orthogonale Felder (*nearly orthogonal arrays*: NOA) eingesetzt werden. XU stellt dazu in [79, 80] ein Verfahren vor, welches ein NOA spaltenweise aufbaut und dabei die D_2 -Diskrepanz (siehe Kapitel 8.2.3) minimiert.

Orthogonale Design-Tabellen

Unter den Begriff orthogonale Design-Tabellen L_n (*orthogonal design tables*) werden spezielle orthogonale Testfelder der Stärke $t = 2$ zusammengefasst, die folgende Eigenschaften aufweisen:

1. Jede Stufe eines Faktors (Spalte) kommt gleich häufig vor.
2. Alle Stufenkombinationen zweier Faktoren kommen gleich häufig vor.

Somit gilt beispielsweise $L_n(s^{n_f}) = OA(n_r, n_f, s, 2)$ oder

$$L_n(s_1^{k_1} s_2^{k_2} \cdots s_l^{k_l}) = OA(n_r, n_f, s_1^{k_1} s_2^{k_2} \cdots s_l^{k_l}, 2) \text{ mit } n_f = \sum_j k_j.$$

8.3.3 Latin Hypercube

Ein *Latin Hypercube Design LHD* (n_r, n_f) ist eine $n_r \times n_f$ Matrix X^{LHD} , bei der jede Spalte aus einer zufälligen Permutation der Zahlen $\{1, 2, 3, \dots, n_r\}$ besteht. Ein *Latin Hypercube Sampling LHS* wird aus einem *LHD* erzeugt, indem von jedem Wert des *LHD* eine Zufallszahl aus dem Bereich $[0, 1]$ abgezogen wird und anschließend jeder Wert durch n_r geteilt wird, wodurch ein Testfeld im Einheitsraum C^{n_f} entsteht (Abbildung 8.10) [47, 61].

$$x_{ij} = \frac{x_{ij}^{LHD} - \text{rand}[0, 1]}{n_r} \quad x_{ij}^{LHD} \in \{1, 2, \dots, n_r\} \quad (8.29)$$

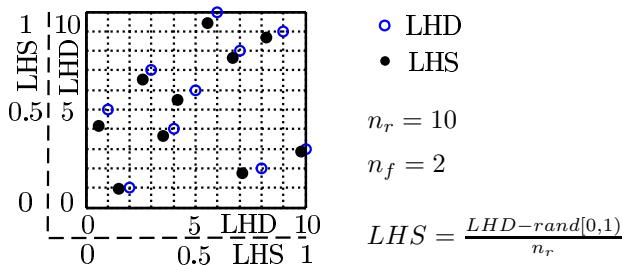


Abb. 8.10 Latin Hypercube Design und Latin Hypercube Sampling

Da das *LHS* den Faktorraum in Schichten (*strata*) aufteilt, wird es auch als Erweiterung des *Stratified Sampling* betrachtet [61]. Werden die Zufallszahlen durch den konstanten Wert 0.5 ersetzt, so wird ein zentriertes *LHS* (*MLHS*) (*Midpoint- oder Centered LHS*) erzeugt.

Die grundlegende Konstruktionsmethode eines *LHS* bzw. *LHD* garantiert kein gleichverteiltes und korrelationsfreies Testfeld, wie in Abbildung 8.11 beispielhaft

gezeigt wird. Das *schlechtere* Testfeld [o] erfüllt zwar die Konstruktionskriterien für ein LHS mit 20 Testpunkten und zwei Faktoren, die Faktoren weisen jedoch die größtmögliche Korrelation und nur eine schlechte Gleichverteilung im Faktorraum auf. Eine sinnvolle Auswertung der daraus ermittelten Testergebnisse ist nicht möglich. Das zweite Feld [•] weist hingegen deutlich bessere Eigenschaften auf und ist vorzuziehen.

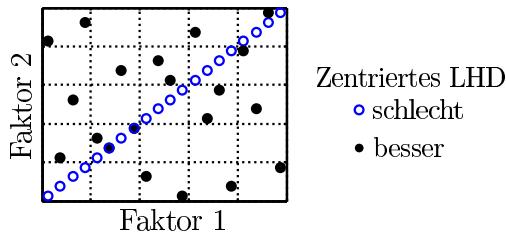


Abb. 8.11 Zwei theoretische LHS mit 20 Testpunkten

Es zeigt sich, dass bei gutem Aufbau eines LHS die Varianz des globalen Mittelwerts \hat{y} deutlich geringer ist als bei Verwendung eines zufälligen Monte-Carlo-Felds mit gleicher Testpunkteanzahl [47, 70, 51].

Zur weiteren Verringerung der Varianz eines LHS kann dieses auf Basis von orthogonalen Testfeldern (orthogonal arrays) erzeugt werden (Kapitel 8.3.2) [52, 54, 71, 40]. Diese Testfelder werden in der Literatur entweder *orthogonal array-based Latin Hypercubes* oder *randomized orthogonal arrays* (OA-based LHD) genannt. Dabei werden jeder *Stufe* des orthogonalen Testfelds Zufallszahlen (Stufen) aus der Menge $\{1 \dots n_r\}$ zugewiesen. Beispielsweise wird dazu das orthogonale Testfeld $L_n(3^4)$ betrachtet (Tabelle 8.3, siehe auch Kapitel 2.7). Jede Stufe tritt in diesem Feld für jeden Faktor genau drei mal auf. Jeder Stufe werden im ersten Schritt drei beliebige unterschiedliche ganze Zahlen aus dem Bereich 1 bis n_r zugewiesen. Im dargestellten Beispiel sind dies:

$$1 \rightarrow \{1, 2, 3\} \quad 2 \rightarrow \{4, 5, 6\} \quad 3 \rightarrow \{7, 8, 9\}$$

In jeder Spalte des orthogonalen Feldes werden im zweiten Schritt die Stufen des orthogonalen Feldes durch eine zufällige Permutation der zugewiesenen Gruppe ersetzt (Tabelle 8.3 und Abbildung 8.12).

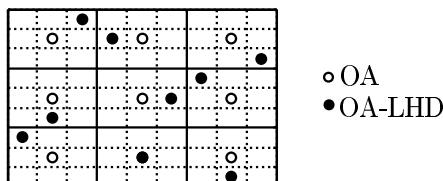


Abb. 8.12 OALHD Struktur im Vergleich zum grundlegendem OA für 2 Faktoren

$L_9(3^4)$				$OA - LHD(9, 9^4)$			
1	1	1	1		1	3	1
1	2	2	2		2	4	6
1	3	3	3		3	9	7
2	1	2	3		5	2	5
2	2	3	1	→	6	5	9
2	3	1	2		4	8	3
3	1	3	2		8	1	8
3	2	1	3		7	6	2
3	3	2	1		9	7	4

Tabelle 8.3 Latin Hypercube Testfeld auf Basis eines orthogonalen Testfelds

Da LHDs grundsätzlich durch Permutationen der Zahlenfolge $\{1, \dots, n_r\}$ erzeugt werden, existieren insgesamt $n_r^{n_f}$ mögliche verschiedene Testfelder. Die Suche nach einem ausgeglichenen Feld wird dadurch schnell rechenintensiv. Bei $n_r = 20$ und $n_f = 5$ existieren bereits $3.5035 \cdot 10^{73}$ mögliche Testfelder, was einen Vergleich aller Felder für die Praxis zu zeitaufwändig macht. In der Praxis zeigt sich, dass symmetrische LHDs vorteilhafte Eigenschaften für Computer-Experimente aufweisen [56, 49]. Ein symmetrisches LHD ist symmetrisch zum Zentrum, welches für jede Dimension bei $(n_r + 1)/2$ liegt. Somit gibt es zu jedem Testpunkt $x_j = \{x_{j_1}, \dots, x_{j_{n_f}}\}$ einen gespiegelten Punkt $x_j^* = \{n_r + 1 - x_{j_1}, \dots, n_r + 1 - x_{j_{n_f}}\}$, wobei gilt $x_j + x_j^* = \{n_r + 1, n_r + 1, \dots\}$ (Tabelle 8.4).

Punkte: x_j	1	3	8	5		1	8	6
	2	8	3	2		2	3	2
	3	7	5	3		3	6	3
	4	5	2	1		4	9	1
	8	6	1	4		5	5	5
gespiegelte Punkte: x_j^*	7	1	6	7		9	2	4
	6	2	4	6		8	7	8
	5	4	7	8		7	4	7
						6	1	9

Tabelle 8.4 Symmetrisches LHD mit gerader und ungerader Stufenanzahl

Weiterhin zeigt sich, dass Testfelder mit geringer Korrelation zwischen den einzelnen Faktoren (Dimensionen) vorteilhaft sind, um im Anschluss Metamodelle mit guter Vorhersagequalität zu erzeugen [29, 53]. YE [81] stellt aus diesem Grund ein Verfahren vor, mit dem paarweise (jeweils zwei Faktoren) orthogonale LHDs erzeugt werden (OLHD). Die paarweise Orthogonalität garantiert dabei die unabhängige Bestimmung aller Haupteffekte. Die Konstruktionsmethode nach Ye ist begrenzt auf LHDs mit folgenden Größen:

$$n_r = \begin{cases} 2^{1+\frac{i+1}{2}} + 1 & , i \text{ ungerade} \\ 2^{1+\frac{i}{2}} + 1 & , i \text{ gerade} \end{cases} \quad \text{mit } i \geq n_f \quad (8.30)$$

Mit der Hilfsgröße $m = \log_2(n_r - 1)$ wird ein OLHD in wenigen Schritten erzeugt. Die grundlegende Konstruktionsmethode besteht aus 6 Schritten (a-f) und wird hier an einem Beispiel mit $n_f = 6$ und $n_r = 17 \rightarrow m = 4$ näher erläutert (Tabelle 8.5).

- Erzeuge eine Matrix M mit $s = 2(m - 1)$ Spalten und $z = (n_r - 1)/2$ Zeilen. Die erste Spalte ist eine zufällige Permutation der Zahlen $1, 2, \dots, s$ und wird als Erzeugungsvektor e bezeichnet, da aus ihm alle weiteren Spalten abgeleitet werden. Für die Spalten $j = 2, \dots, m - 1$ wird e in aufeinanderfolgenden Gruppen von 2^{j-1} Zahlen aufgeteilt und in umgekehrter Reihenfolge aufgetragen. Die Spalte m ist dann der komplett umgekehrte Erzeugungsvektor e . Die folgenden Spalten $j > m$ werden durch Aufteilen der Spalte m in Gruppen mit je 2^{j-m} Elementen und der Umkehrung erzeugt (Tabelle 8.5).
- Erzeuge eine sogenannte *PlusMinus*-Matrix mit der gleichen Dimension wie M , wobei die erste Spalte mit +1 gefüllt ist. Die Spalten $j = 2, \dots, m$ werden mit abwechselnden Gruppen von -1 und +1 der Größe 2^{j-1} belegt. Die weiteren Spalten werden durch die Multiplikation (Interaktion) zwischen den Spalten 2 und $j - m + 2$ bestimmt.
- Multipliziere die in Schritt a und b erzeugten Felder und füge eine Zeile mit Nullen hinzu.
- Kopiere die Zeilen $1, \dots, (n_r - 1)/2$ in die Zeilen $(n_r - 1)/2 + 2, \dots, n_r$ und multipliziere diese mit -1.
- Im letzte Schritt wird zur Erzeugung des Felds der Wert $(n_r + 1)/2$ addiert, so dass für jeden Faktor die Stufen $1, \dots, n_r$ auftreten.

a:	b:	c,d:	e:																																																																																																																																																																																																																																																												
<table border="1"> <tr><td>1</td><td>2</td><td>4</td><td>8</td><td>7</td><td>5</td></tr> <tr><td>2</td><td>1</td><td>3</td><td>7</td><td>8</td><td>6</td></tr> <tr><td>3</td><td>4</td><td>2</td><td>6</td><td>5</td><td>7</td></tr> <tr><td>4</td><td>3</td><td>1</td><td>5</td><td>6</td><td>8</td></tr> <tr><td>5</td><td>6</td><td>8</td><td>4</td><td>3</td><td>1</td></tr> <tr><td>6</td><td>5</td><td>7</td><td>3</td><td>4</td><td>2</td></tr> <tr><td>7</td><td>8</td><td>6</td><td>2</td><td>1</td><td>3</td></tr> <tr><td>8</td><td>7</td><td>5</td><td>1</td><td>2</td><td>4</td></tr> </table>	1	2	4	8	7	5	2	1	3	7	8	6	3	4	2	6	5	7	4	3	1	5	6	8	5	6	8	4	3	1	6	5	7	3	4	2	7	8	6	2	1	3	8	7	5	1	2	4	<table border="1"> <tr><td>+1</td><td>-1</td><td>-1</td><td>-1</td><td>+1</td><td>+1</td></tr> <tr><td>+1</td><td>+1</td><td>-1</td><td>-1</td><td>-1</td><td>-1</td></tr> <tr><td>+1</td><td>-1</td><td>+1</td><td>-1</td><td>-1</td><td>+1</td></tr> <tr><td>+1</td><td>+1</td><td>+1</td><td>-1</td><td>+1</td><td>-1</td></tr> <tr><td>+1</td><td>-1</td><td>-1</td><td>+1</td><td>+1</td><td>-1</td></tr> <tr><td>+1</td><td>+1</td><td>-1</td><td>+1</td><td>-1</td><td>+1</td></tr> <tr><td>+1</td><td>-1</td><td>+1</td><td>+1</td><td>-1</td><td>-1</td></tr> <tr><td>+1</td><td>+1</td><td>+1</td><td>+1</td><td>+1</td><td>+1</td></tr> </table>	+1	-1	-1	-1	+1	+1	+1	+1	-1	-1	-1	-1	+1	-1	+1	-1	-1	+1	+1	+1	+1	-1	+1	-1	+1	-1	-1	+1	+1	-1	+1	+1	-1	+1	-1	+1	+1	-1	+1	+1	-1	-1	+1	+1	+1	+1	+1	+1	<table border="1"> <tr><td>+1</td><td>-2</td><td>-4</td><td>-8</td><td>+7</td><td>+5</td></tr> <tr><td>+2</td><td>+1</td><td>-3</td><td>-7</td><td>-8</td><td>-6</td></tr> <tr><td>+3</td><td>-4</td><td>+2</td><td>-6</td><td>-5</td><td>+7</td></tr> <tr><td>+4</td><td>+3</td><td>+1</td><td>-5</td><td>+6</td><td>-8</td></tr> <tr><td>+5</td><td>-6</td><td>-8</td><td>+4</td><td>+3</td><td>-1</td></tr> <tr><td>+6</td><td>+5</td><td>-7</td><td>+3</td><td>-4</td><td>+2</td></tr> <tr><td>+7</td><td>-8</td><td>+6</td><td>+2</td><td>-1</td><td>-3</td></tr> <tr><td>+8</td><td>+7</td><td>+5</td><td>+1</td><td>+2</td><td>+4</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> </table>	+1	-2	-4	-8	+7	+5	+2	+1	-3	-7	-8	-6	+3	-4	+2	-6	-5	+7	+4	+3	+1	-5	+6	-8	+5	-6	-8	+4	+3	-1	+6	+5	-7	+3	-4	+2	+7	-8	+6	+2	-1	-3	+8	+7	+5	+1	+2	+4	0	0	0	0	0	0	<table border="1"> <tr><td>+1</td><td>-2</td><td>-4</td><td>-8</td><td>+7</td><td>+5</td></tr> <tr><td>+2</td><td>+1</td><td>-3</td><td>-7</td><td>-8</td><td>-6</td></tr> <tr><td>+3</td><td>-4</td><td>+2</td><td>-6</td><td>-5</td><td>+7</td></tr> <tr><td>+4</td><td>+3</td><td>+1</td><td>-5</td><td>+6</td><td>-8</td></tr> <tr><td>+5</td><td>-6</td><td>-8</td><td>+4</td><td>+3</td><td>-1</td></tr> <tr><td>+6</td><td>+5</td><td>-7</td><td>+3</td><td>-4</td><td>+2</td></tr> <tr><td>+7</td><td>-8</td><td>+6</td><td>+2</td><td>-1</td><td>-3</td></tr> <tr><td>+8</td><td>+7</td><td>+5</td><td>+1</td><td>+2</td><td>+4</td></tr> <tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>-1</td><td>+2</td><td>+4</td><td>+8</td><td>-7</td><td>-5</td></tr> <tr><td>-2</td><td>-1</td><td>+3</td><td>+7</td><td>+8</td><td>+6</td></tr> <tr><td>-3</td><td>+4</td><td>-2</td><td>+6</td><td>+5</td><td>-7</td></tr> <tr><td>-4</td><td>-3</td><td>-1</td><td>+5</td><td>-6</td><td>+8</td></tr> <tr><td>-5</td><td>+6</td><td>+8</td><td>-4</td><td>-3</td><td>+1</td></tr> <tr><td>-6</td><td>-5</td><td>+7</td><td>-3</td><td>+4</td><td>-2</td></tr> <tr><td>-7</td><td>+8</td><td>-6</td><td>-2</td><td>+1</td><td>+3</td></tr> <tr><td>-8</td><td>-7</td><td>-5</td><td>-1</td><td>-2</td><td>-4</td></tr> </table>	+1	-2	-4	-8	+7	+5	+2	+1	-3	-7	-8	-6	+3	-4	+2	-6	-5	+7	+4	+3	+1	-5	+6	-8	+5	-6	-8	+4	+3	-1	+6	+5	-7	+3	-4	+2	+7	-8	+6	+2	-1	-3	+8	+7	+5	+1	+2	+4	0	0	0	0	0	0	-1	+2	+4	+8	-7	-5	-2	-1	+3	+7	+8	+6	-3	+4	-2	+6	+5	-7	-4	-3	-1	+5	-6	+8	-5	+6	+8	-4	-3	+1	-6	-5	+7	-3	+4	-2	-7	+8	-6	-2	+1	+3	-8	-7	-5	-1	-2	-4
1	2	4	8	7	5																																																																																																																																																																																																																																																										
2	1	3	7	8	6																																																																																																																																																																																																																																																										
3	4	2	6	5	7																																																																																																																																																																																																																																																										
4	3	1	5	6	8																																																																																																																																																																																																																																																										
5	6	8	4	3	1																																																																																																																																																																																																																																																										
6	5	7	3	4	2																																																																																																																																																																																																																																																										
7	8	6	2	1	3																																																																																																																																																																																																																																																										
8	7	5	1	2	4																																																																																																																																																																																																																																																										
+1	-1	-1	-1	+1	+1																																																																																																																																																																																																																																																										
+1	+1	-1	-1	-1	-1																																																																																																																																																																																																																																																										
+1	-1	+1	-1	-1	+1																																																																																																																																																																																																																																																										
+1	+1	+1	-1	+1	-1																																																																																																																																																																																																																																																										
+1	-1	-1	+1	+1	-1																																																																																																																																																																																																																																																										
+1	+1	-1	+1	-1	+1																																																																																																																																																																																																																																																										
+1	-1	+1	+1	-1	-1																																																																																																																																																																																																																																																										
+1	+1	+1	+1	+1	+1																																																																																																																																																																																																																																																										
+1	-2	-4	-8	+7	+5																																																																																																																																																																																																																																																										
+2	+1	-3	-7	-8	-6																																																																																																																																																																																																																																																										
+3	-4	+2	-6	-5	+7																																																																																																																																																																																																																																																										
+4	+3	+1	-5	+6	-8																																																																																																																																																																																																																																																										
+5	-6	-8	+4	+3	-1																																																																																																																																																																																																																																																										
+6	+5	-7	+3	-4	+2																																																																																																																																																																																																																																																										
+7	-8	+6	+2	-1	-3																																																																																																																																																																																																																																																										
+8	+7	+5	+1	+2	+4																																																																																																																																																																																																																																																										
0	0	0	0	0	0																																																																																																																																																																																																																																																										
+1	-2	-4	-8	+7	+5																																																																																																																																																																																																																																																										
+2	+1	-3	-7	-8	-6																																																																																																																																																																																																																																																										
+3	-4	+2	-6	-5	+7																																																																																																																																																																																																																																																										
+4	+3	+1	-5	+6	-8																																																																																																																																																																																																																																																										
+5	-6	-8	+4	+3	-1																																																																																																																																																																																																																																																										
+6	+5	-7	+3	-4	+2																																																																																																																																																																																																																																																										
+7	-8	+6	+2	-1	-3																																																																																																																																																																																																																																																										
+8	+7	+5	+1	+2	+4																																																																																																																																																																																																																																																										
0	0	0	0	0	0																																																																																																																																																																																																																																																										
-1	+2	+4	+8	-7	-5																																																																																																																																																																																																																																																										
-2	-1	+3	+7	+8	+6																																																																																																																																																																																																																																																										
-3	+4	-2	+6	+5	-7																																																																																																																																																																																																																																																										
-4	-3	-1	+5	-6	+8																																																																																																																																																																																																																																																										
-5	+6	+8	-4	-3	+1																																																																																																																																																																																																																																																										
-6	-5	+7	-3	+4	-2																																																																																																																																																																																																																																																										
-7	+8	-6	-2	+1	+3																																																																																																																																																																																																																																																										
-8	-7	-5	-1	-2	-4																																																																																																																																																																																																																																																										
		↓																																																																																																																																																																																																																																																													
		-1	→																																																																																																																																																																																																																																																												

Tabelle 8.5 Erzeugung eines OLHD (Beispiel)

Die Konstruktion eines OLHD durch Permutation seines Erzeugungsvektors führt zu $[(n_r - 1)/2]!$ möglichen OLHDs. Bei lediglich $n_r = 33$ Testpunkten ergeben sich dadurch bereits $2.09 \cdot 10^{13}$ verschiedene Felder. Zur Auswahl eines bevorzugten Feldes werden die in Kapitel 8.2 dargestellten Qualitätskriterien eingesetzt.

8.3.4 Gleichverteilte Testfelder (Uniform Designs)

Ein Testfeld $U(n_r, s_1 \times s_2 \times \dots \times s_{n_f})$, mit n_r Testpunkten und n_f Faktoren, bei dem jeder Faktor j genau s_j Stufen aufweist, die jeweils gleich häufig auftreten, wird *U-type, ballanced* oder *lattice Design* genannt [18, 43, 41]. Treten bei verschiedenen Faktoren die gleichen Stufenzahlen auf, so werden diese Felder vereinfacht durch $U(n_r, s_1^{r_1} \times \dots \times s_k^{r_k})$ dargestellt, wobei gilt: $n_f = \sum_j r_j$. Weisen alle Faktoren die gleiche Stufenanzahl s auf, so wird von einem symmetrischen U-type Design $U(n_r, s^{n_f})$ gesprochen. Typischerweise sind den einzelnen Stufen ganzzahlige Werte aus $u_{ij} \in \{1, \dots, s_j\}$ zugewiesen. Die Transformation in den Einheitsraum $C^{n_f} = [0, 1]^{n_f}$ wird durch folgende Rechenvorschrift ermöglicht:

$$x_{ij} = \frac{2u_{ij} - 1}{2s_j} \quad , \quad i = 1, \dots, n_r \quad , \quad j = 1, \dots, n_f \quad (8.31)$$

Das dadurch im Einheitsraum erzeugte Testfeld wird als induziertes (*induced*) Design D_U von U bezeichnet [18]. Bei Verwendung einer identischen Stufenanzahl s für alle Faktoren n_f besteht das induzierte Design D_U lediglich aus den folgenden Elementen $\{\frac{1}{2s}, \frac{3}{2s}, \dots, \frac{2s-1}{2s}\}$.

Sei nun D ein Maß für die Gleichverteilung eines Feldes D_U wie sie in Kapitel 8.2.3 dargestellt werden. Das U-type Design $U(n_r, s^{n_f})$, welches das Kriterium D optimiert, wird als *Uniform Design* bezüglich D bezeichnet $[U_D(n_r)]$ [43]. Bei der Betrachtung eines einzelnen Faktors ist jede Punktereihenfolge mit äquidistanten Abständen ein Uniform Design (UD) [21]. Im Faktorbereich $[0, 1]$ ist somit das UD durch die Punkte $\{\frac{1}{2n_r}, \frac{3}{2n_r}, \dots, \frac{2n_r-1}{2n_r}\}$ definiert. Bei Verwendung mehrerer Faktoren $n_f > 1$ wird die Suche nach einem UD schnell rechenaufwändig, so dass verschiedene Methoden eingesetzt werden, um die in Frage kommenden Testfelder zu begrenzen.

Ein häufig eingesetztes Verfahren ist dabei die Gute-Gitterpunkt-Methode (GGM) (*good lattice point method (GLP)*) [37, 18]. Im ersten Schritt werden alle Zahlen $h < n_r$ gesucht, für die der größte gemeinsame Teiler (GGT) von h und n_r eins ist: $GGT(h, n_r) = 1$. Für $n_r = 21$ sind das beispielsweise die folgenden 12 Zahlen: $h \in \{1, 2, 4, 5, 8, 10, 11, 13, 16, 17, 19, 20\}$. Nachdem anschließend für jeden Faktor j eine unterschiedliche Zahl h_j gewählt wurde, werden die Elemente der Matrix U wie folgt berechnet:

$$u_{ij} = \begin{cases} (h_j i) \bmod n_r & \text{für } h_j i \bmod n_r \neq 0 \\ n_r & \text{für } h_j i \bmod n_r = 0 \end{cases} \quad (8.32)$$

Der Vektor aller gewählten h_j wird dabei als Erzeugungsvektor bezeichnet. Das Design mit dem besten Wert für ein gewähltes Gleichverteilungs-Kriterium (Kapitel 8.2) wird als (nahezu) Gleichverteiltes Design (*nearly uniform design*) bezeichnet.

Bei zwei Faktoren und beispielsweise $n_r = 21$ Testpunkten ergeben sich daraus $\binom{12}{2} = \frac{12!}{2!(12-2)!} = 66$ Kombinationen der Kandidaten aus h und folglich genauso viele mögliche GGM-Testfelder. Der Erzeugungsvektor $(1, 13)$ ergibt im dargestellten Beispiel ein Testfeld mit guter Gleichverteilung nach der zentrierten Diskrepanz (ZD) [18].

Faktor 1 = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Faktor 2 = 13 5 18 10 2 15 7 20 12 4 17 9 1 14 6 19 11 3 16 8 21

Da auch bei der GGM-Methode die Anzahl der möglichen Testfelder bei steigender Testpunkte- n_r und Faktanzahl n_f schnell zunimmt, wird ein Erzeugungsvektor (*power generator*) der folgenden Form vorgeschlagen, wobei die endgültigen Elemente h_j unterschiedlich sein müssen [18]:

$$(h_1, \dots, h_{n_f}) = (1, a, a^2, \dots, a^{n_f-1}) \bmod n_r, \quad 0 < a < n_r \quad (8.33)$$

Auch hier wird wie in Gleichung 8.32 die *mod* Funktion soweit verändert, dass sie Werte zwischen 1 und n_r liefert (0 durch n_r ersetzen). Werden Testfelder mit geringer Testpunkteanzahl gesucht, existieren in einigen Fällen nur wenige Kandidaten h ($n_r = 6 \rightarrow h \in \{1, 5\}$). Die daraus kombinierbaren Erzeugungsvektoren führen meist zu Feldern mit schlechter Gleichverteilung. In diesen Fällen kann ein Testfeld für $n_r + 1$ Testpunkte erzeugt und der letzte Testpunkt gestrichen werden. Besonders wenn $n_r + 1$ eine Primzahl ist, stehen deutlich mehr Kandidaten für h zur Verfügung ($n_r = 6 + 1 = 7 \rightarrow h \in \{1, 2, 3, 4, 5, 6\}$) und das erzeugte Testfeld weist meist eine bessere Gleichverteilung als die erzeugten Originalfelder auf [18, 74, 14]. Verschiedene Uniform Designs wurden bereits von FANG, MA und WINKLER im Internet zur Verfügung gestellt [20].

Da Testfelder, die mit der *good lattice point*-Methode erzeugt werden und bei denen n_r keine Primzahl ist, schlechte Gleichverteilungen aufweisen können, schlagen MA und FANG [43] zur Verbesserung die sogenannte Schnitt (*cutting*) Methode vor. Dabei wird aus einem gleichverteilten Testfeld $U_p^{n_f} = U(p, p^{n_f})$ mit $p > n_r$ oder $p \gg n_r$ ein Testfeld mit n_r Testpunkten herausgeschnitten. p oder $p + 1$ ist eine Primzahl, so dass das verwendete Testfeld eine gute Gleichverteilung aufweist. Um ein Testfeld $U(n_r, n_r^{n_f})$ aus $U(p, p^{n_f})$ zu erzeugen, werden die Testpunkte (Zeilen) des Basisfelds U_p so sortiert, dass in einer gewählten Spalte j die Stufenwerte kontinuierlich ansteigen ($x_{i,j} < x_{i+1,j}$). Startend von einem beliebig gewählten

Testpunkt i werden anschließend n_r aufeinander folgende Punkte ausgewählt (Abbildung 8.13). Sollte das letzte Element von U_p erreicht werden, wird beim ersten Testpunkt von U_p fortgefahren. Anschließend werden die n_r gewählten Testpunkte gleichmäßig in jeder Dimension (Spalte) verteilt. Dazu werden die n_r Werte jeder Spalte entsprechend ihrer Größe durch die Zahlen $k = 1, 2, \dots, n_r$ ersetzt. Als Beispiel wird aus einem U_{30}^2 Testfeld ein $U(10, 10^2)$ Testfeld erzeugt, wobei nach der ersten Dimension sortiert wird und ab Position 25 zehn Testpunkte herausgeschnitten werden (Abbildung 8.13).

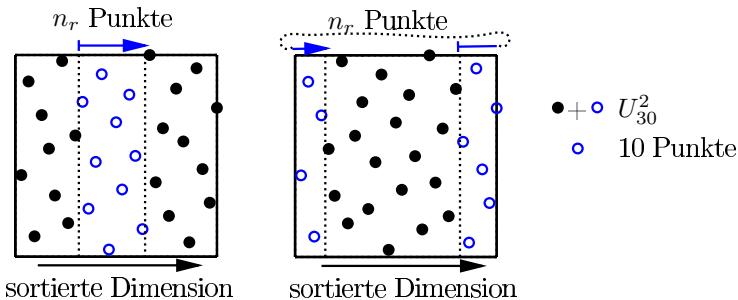


Abb. 8.13 Wahl von 10 Testpunkten aus einem gleichverteilten Testfeld

24	23	1	9	5	11	19	6	21	3	12	15	20	18	17	26	7	4	28	27	25	13	14	29	22	8	2	16	30	10
25	6	12	18	16	7	4	9	11	3	14	20	30	15	24	2	29	21	13	28	17	27	1	8	19	5	26	10	22	23
↓																													
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
12	26	3	21	16	9	29	5	18	23	7	14	27	1	20	10	24	15	4	30	11	19	6	25	17	2	28	13	8	22
↓																													
25	26	27	28	29	30	1	2	3	4																				
17	2	28	13	8	22	12	26	3	21																				
5	6	7	8	9	10	1	2	3	4																				
6	1	10	5	3	8	4	9	2	7																				

Abb. 8.14 Erzeugung eines $U(10, 10^2)$ aus einem $U(30, 30^2)$

Ein Basisfeld $U_p^{n_f}$ kann grundsätzlich in jeder Dimension sortiert werden, wodurch insgesamt n_f verschiedene sortierte Basisfelder entstehen. Aus jedem sortierten Basisfeld können p unterschiedliche Startpunkte zur Auswahl der n_r Testpunkte gewählt werden, wodurch insgesamt $n_f p$ verschiedene $U(n_r, n_r^{n_f})$ Testfelder erzeugt werden. Das Testfeld mit der besten Gleichverteilung (z.B. geringste ZD, Kapitel 8.2.3) wird zur weiteren Verwendung gewählt. Die Cutting-Methode erzeugt in vielen Fällen Testfelder mit besserer Gleichverteilung als die einfache *good lattice point*-Methode. Weiterhin werden durch ein Basisfeld der Form $U(p, p^{n_f})$ mit ge-

ringem Rechenaufwand gute Felder für n_f Faktoren und $n_r \leq p$ Testpunkte erzeugt [43, 18].

Eine besondere Gruppe der U-type Designs bilden *Latin Squares* (LS) der Form $U(n_r, n_r^{n_r})$. Basierend auf einem Erzeugungsvektor e (Permutation der Zahlen $1 \cdots n_r$) wird dabei durch eine Verschiebung der Elemente ein Testfeld konstruiert. Bei Verwendung einer einfachen *Linksverschiebung* um genau eine Stelle

$$L(x_{i1}, x_{i2}, \dots, x_{in_r}) = (x_{i2}, x_{i3}, \dots, x_{in_r}, x_{i1}) \quad (8.34)$$

wird ein *links zyklischer Latin Square* (LCLS: left cycle latin square) wie folgt bestimmt:

$$x_{i+1} = L(x_i) \quad , i = 1, \dots, n_r - 1 \quad (8.35)$$

Durch die eindeutige Definition des LS durch einen Erzeugungsvektor existieren bei gegebenem n_r , genau $n_r!$ unterschiedliche LCLS. Ein gleichverteiltes Testfeld mit n_r Testpunkten und $n_f \leq n_r$ Faktoren wird in zwei Schritten mittels eines LS der Größe n_r erzeugt [17, 18].

1. Erzeuge ein LS mit n_r Testpunkten und bestmöglicher Gleichverteilung für ein gewähltes Gütekriterium aus Kapitel 8.2.3.
2. Wähle n_f Spalten des LS, so dass das gewählte Gütekriterium für das entstandene Testfeld $U(n_r, n_r^{n_f})$ optimiert wird.

Ein Vorteil dieses Verfahrens ist die Möglichkeit, mit geringem Aufwand verschiedene Testfelder mit $n_f < n_r$ aus einem LS zu erzeugen. Der Optimierungsaufwand zur Erzeugung des LS ist somit nur einmal für eine Testpunkteanzahl n_r durchzuführen.

Asymmetrische Testfelder der Form $U(n_r, s_1^{r_1} \times \dots \times s_k^{r_k})$ können aus einem symmetrischen Testfeld $U(n_r, n_r^{n_f})$ mit wenig Aufwand ermittelt werden, wenn n_r ein ganzzahliges Vielfaches jeder Stufenanzahl s_j ist. Zur Konstruktion des gesuchten Testfelds werden jeder Stufe eines Faktors j genau n_r/s_j zufällige Stufen des symmetrischen Testfelds zugewiesen. So kann beispielsweise aus einem $U(12, 12^2)$ ein $U(12, 6^1 \times 4^1)$ wie folgt erstellt werden:

$$U(12, 12^2) = \begin{pmatrix} 4 & 8 \\ 10 & 11 \\ 9 & 5 \\ 8 & 9 \\ 5 & 4 \\ 3 & 2 \\ 6 & 12 \\ 7 & 1 \\ 1 & 6 \\ 11 & 3 \\ 2 & 10 \\ 12 & 7 \end{pmatrix} \quad \text{mit} \quad \begin{pmatrix} \{1,2\} \rightarrow 1 & \{1,2,11\} \rightarrow 1 \\ \{3,4\} \rightarrow 2 & \{3,4,12\} \rightarrow 2 \\ \{5,6\} \rightarrow 3 & \{5,6,10\} \rightarrow 3 \\ \{7,8\} \rightarrow 4 & \{7,8,9\} \rightarrow 4 \\ \{9,10\} \rightarrow 5 & \\ \{11,12\} \rightarrow 6 & \end{pmatrix}$$

$$\rightarrow U(12, 6^1 \times 4^1) = \begin{pmatrix} 2 & 4 \\ 5 & 1 \\ 5 & 3 \\ 4 & 4 \\ 3 & 2 \\ 2 & 1 \\ 3 & 2 \\ 4 & 1 \\ 1 & 3 \\ 6 & 2 \\ 1 & 3 \\ 6 & 4 \end{pmatrix}$$

Durch unterschiedliche Zuweisungen der s_j Stufen zu den gegebenen n_r Stufen des Basisfelds entstehen Felder mit unterschiedlich guten Gleichverteilungen, wobei das Feld mit der besten Gleichverteilung nach dem gewählten Gütekriterium (Kapitel 8.2) im Weiteren verwendet wird. Das dargestellte Ersetzungsverfahren kann ebenfalls zur Erzeugung eines $U(n_r, n_s^{n_f})$ mit $n_f < n_r$ verwendet werden.

Ein alternatives Verfahren zur Erzeugung asymmetrischer Testfelder der Form $U(n_r, s_1^{r_1} \times \dots \times s_k^{r_k})$ ist die *Collapsing*-Methode [19, 18]. Dabei werden zwei gleichverteilte Testfelder (Uniform Designs) $U_v(n_{rv}, s_1 \times s_2 \times \dots \times s_v)$ und $U_w(n_{rw}, n_{rw}^w)$ zu einem kombinierten Testfeld $U_{v,w}(n_{rv}n_{rw}, s_1 \times s_2 \times \dots \times s_v \times n_{rw}^w)$ zusammengefasst. Die Konstruktion basiert auf einem Kronecker-Produkt \otimes [5, 10] der Testfelder U_v und U_w mit jeweils einem ein-dimensionalen Vektor aus Einsen der Länge n_{rv} bzw. n_{rw} .

$$U_{v,w} = \left(1_{n_{rv}} \otimes U_v : U_w \otimes 1_{n_{rw}} \right) \quad (8.36)$$

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix} \quad \text{mit } A = (a_{ij})$$

Betrachten wir zur Erläuterung das folgende Beispiel [18]:

$$U_v = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 2 \\ 3 & 2 & 2 \\ 4 & 2 & 1 \end{pmatrix}, U_w = \begin{pmatrix} 1 & 3 \\ 2 & 2 \\ 3 & 1 \end{pmatrix} \rightarrow U_{v,w} = \left(\begin{array}{c|c} \begin{matrix} 1 & 1 & 1 \\ 2 & 1 & 2 \\ 3 & 2 & 2 \\ 4 & 2 & 1 \end{matrix} & \begin{matrix} 1 & 3 \\ 1 & 3 \\ 1 & 3 \\ 1 & 3 \end{matrix} \\ \hline \begin{matrix} 1 & 1 & 1 \\ 2 & 1 & 2 \\ 3 & 2 & 2 \\ 4 & 2 & 1 \end{matrix} & \begin{matrix} 2 & 2 \\ 2 & 2 \\ 2 & 2 \\ 2 & 2 \end{matrix} \\ \hline \begin{matrix} 1 & 1 & 1 \\ 2 & 1 & 2 \\ 3 & 2 & 2 \\ 4 & 2 & 1 \end{matrix} & \begin{matrix} 3 & 1 \\ 3 & 1 \\ 3 & 1 \\ 3 & 1 \end{matrix} \end{array} \right) \quad (8.37)$$

Bei Verwendung von gleichverteilten Testfeldern für U_v und U_w ergeben sich kombinierte Testfelder mit ebenfalls guter Gleichverteilung [18].

8.4 Optimierung von Testfeldern

Zur Beurteilung der Güte eines Testfeldes T werden verschiedene Qualitätskriterien eingesetzt (Kapitel 8.2). Jedes Kriterium ist dabei eine direkte Funktion des Testfelds $q(T)$ und kann so definiert werden, dass die Minimierung von Δq zu einem optimalen Testfeld bezüglich des gewählten Qualitätskriteriums führt. Die Optimierung eines Testfeldes wird dabei grundsätzlich in fünf Schritten durchgeführt:

1. Wahl eines Basis-Testfeldes T_0 .
2. Erzeugung eines neuen Testfelds T_i , welches auf einem vorhergehenden Testfeld (meistens T_{i-1}) beruht.
3. Berechnung des Qualitätskriteriums $q(T_i)$ für das neue Testfeld T_i bzw. die Änderung des Qualitätskriteriums zum vorhergehenden Testfeld $\Delta q = q(T_i) - q(T_{i-1})$.
4. In Abhängigkeit der Veränderung des Qualitätskriteriums Δq wird das Testfeld T_i als neues Basis-Testfeld akzeptiert oder verworfen.
5. Basierend auf einem Stop-Kriterium wird die Optimierung beendet oder mit Schritt (2) fortgesetzt.

Erzeugung eines neuen Testfelds

Zur Erzeugung eines neuen Testfeldes T_i wird in vielen Fällen eine einfache Vertauschung von Faktorstufen innerhalb eines Faktors (Spalte) favorisiert. Dabei werden im ersten Schritt zufällig ein Faktor und zwei Testpunkte ausgewählt [41]. Anschließend werden die Faktorstufen der beiden Testpunkte für den gewählten Faktor

vertauscht (Gleichung 8.38). Durch die Vertauschung innerhalb eines Faktors wird sichergestellt, dass die korrekte Stufenanzahl und -häufigkeit für jeden Faktor auch nach der Vertauschung erhalten bleibt. Eine gleichzeitige Vertauschung mehrerer Faktorstufen oder Faktoren ist jedoch ebenfalls möglich. Werden Testfeldkonstruktionen verwendet, die auf einem Erzeugungsvektor basieren (z.B. orthogonale LHD, Kapitel 8.3.3), so wird die Vertauschung lediglich im Erzeugungsvektor durchgeführt, da sich das restliche Feld aus diesem Vektor vollständig ergibt. Neben der *zufälligen* Vertauschung von Testpunkten ist je nach Wahl des Qualitätskriteriums ebenfalls eine *gezielte* Vertauschung von Elementen zur Verbesserung des Gütekriteriums möglich.

Im Fall von symmetrischen Testfeldern muss die Symmetriebedingung aufrecht erhalten werden, so dass eventuell nach der Vertauschung von zwei Elementen eines Faktors zwei weitere zur Wiederherstellung der Symmetrie vertauscht werden müssen.

$$T_{i-1} = \begin{pmatrix} 3 & 4 & 1 \\ 2 & 1 & 4 \\ 1 & 3 & 3 \\ 4 & 2 & 2 \end{pmatrix} \xrightarrow{\substack{\text{Faktor :3} \\ \text{Testpunkte :1,3}}} T_i = \begin{pmatrix} 3 & 4 & 3 \\ 2 & 1 & 4 \\ 1 & 3 & 1 \\ 4 & 2 & 2 \end{pmatrix} \quad (8.38)$$

Ersetzungsregeln

Ob ein neues Testfeld T_i das vorherige Feld T_{i-1} ersetzt, wird durch eine Ersetzungsregel entschieden. Die einfachste Regel ist der lokale Suchalgorithmus (LS: local search), welcher in Abhängigkeit der Qualitätsänderung $\Delta q = q(T_i) - q(T_{i-1})$ definiert wird:

$$LS \left\{ \begin{array}{ll} \text{ersetzen} & , \Delta q \leq 0 \\ \text{nicht ersetzen} & , \Delta q > 0 \end{array} \right. \quad (8.39)$$

Wie der Name des Verfahrens bereits impliziert, besteht die Gefahr, dass die lokale Suche (LS) in lokalen Minima verharrt. Sollte die Qualitätsfunktion q verschiedene lokale Minima aufweisen, ist es somit notwendig, die Optimierung von verschiedenen Startpunkten (Basisfeldern) zu beginnen, um das globale Optimum oder ein annähernd gutes Testfeld zu ermitteln. Zur Verbesserung der Sucheigenschaften wurden erweiterte Regeln entwickelt, welche einen meist variablen Schwellwert S_k (*threshold*) anstelle der festen Grenze $\Delta q = 0$ einsetzen.

$$TA : \left\{ \begin{array}{ll} \text{ersetzen} & , \Delta q \leq S_k \\ \text{nicht ersetzen} & , \Delta q > S_k \end{array} \right. \quad (8.40)$$

Bei der *Threshold Accepting* (TA) Methode wird der Schwellwert während des Optimierungsprozesses kontinuierlich, startend von einem vorgegebenen Anfangswert, verringert [9, 77], was beispielsweise in der folgenden Form durchgeführt wird.

$$S_k = \frac{S_0}{n_{TA}} [n_{TA} - k]_+ \quad \text{mit } k \leq 0$$

Dabei ist n_{TA} eine ganze Zahl größer Null, welche die Anzahl der Stufen angibt, in denen der Startwert S_0 bis auf Null reduziert wird. Dabei kann der Schwellwert nach jedem oder nach mehreren Optimierungsschritten reduziert werden. Umso größer der Schwellwert, desto wahrscheinlicher ist es, dass auch ein schlechteres Testfeld als das vorhergehende akzeptiert wird. Dadurch besteht die Möglichkeit, sich aus einem lokalem Minimum zu entfernen. Wenn die maximale Bandbreite des Qualitätskriteriums $\Delta q_{max} = q_{max} - q_{min}$ bekannt ist, so wird empfohlen den Startwert mit $S_0 = 0.1\Delta q_{max}$ festzulegen.

Eine weitere Variante der Ersetzungsregel ist das *Simulated Annealing* (SA: Simulierte Abkühlung), welches sich an dem Abkühlverhalten einer Masse orientiert [34]. Der *Temperatur*-Faktor T_k nimmt dabei vergleichbar mit dem Schwellwert S_k beim TA über der Optimierungszeit kontinuierlich ab $T_k = \alpha T_{k-1}$ (α : Abkühlungsgeschwindigkeit). Im Gegensatz zu TA wird als Schwelle jedoch nicht direkt die *Temperatur* als Schwellwert verwendet, sondern eine Zufallszahl zwischen 0 und 1 (Gleichung 8.41). Es zeigt sich, dass die Qualität des SA deutlich von der Wahl des Abkühlungsfaktors α und der Starttemperatur T_0 abhängt.

$$SA : \begin{cases} \text{ersetzen} & , e^{-\frac{\Delta q}{T_k}} \geq r \\ \text{nicht ersetzen} & , e^{-\frac{\Delta q}{T_k}} < r \end{cases} \quad \text{mit } r = \text{rand}(0, 1) \quad (8.41)$$

Eine Kombination und Erweiterung der Verfahren LS und SA bietet das stochastische Evolutionsverfahren (SE), welches zusätzlich zur Ersetzung bei $\Delta q \leq 0$ eine zufallsgesteuerte Ersetzung mit variabler maximaler Schwelle S_k ermöglicht [60]. Der Faktor r wird dabei in jedem Optimierungsschritt neu bestimmt.

$$SE : \begin{cases} \text{ersetzen} & , \Delta q \leq S_k r \\ \text{nicht ersetzen} & , \Delta q > S_k r \end{cases} \quad \text{mit } r = \text{rand}(0, 1) \quad (8.42)$$

Das SE startet typischerweise mit einem kleinen Schwellwert S_0 , was die Gefahr einschließt, in einem lokalen Minimum (Testfeld: T_{lm}) zu verharren. Scheint dieses zu geschehen, wird der Schwellwert kontinuierlich erhöht, bis ein besseres Ergebnis (Testfeld) gefunden wird als das letzte lokale Minimum T_{lm} . Danach wird erneut auf den kleineren Schwellwert S_0 umgeschaltet und weiter im lokalen Umfeld optimiert. Das Verfahren zeigt zwar gute Konvergenzeigenschaften, jedoch ist die Entscheidung, wann der Schwellwert erhöht werden muss, schwierig, so dass ein verbessertes SE (*enhanced stochastic evolutionary algorithmus*) entwickelt wurde [30]. Der Algorithmus basiert dabei auf einer inneren Schleife, welche neue Testfelder generiert und gleichzeitig bewertet. Eine überlagerte äußere Schleife kontrolliert in Abhängigkeit vom Verhältnis zwischen Anzahl der verbesserten Testfelder und insgesamt berechneter Felder eine Anpassung des Schwellwerts S_k . Dabei wird zwischen den zwei grundsätzlichen Zuständen a) lokale Optimierung (S_{klein}) und b) globale Erforschung ($S_{groß}$) unterschieden [30].

Reduktion des Rechenaufwands

Der Rechenaufwand der Testfeldoptimierung hängt wesentlich vom Aufwand zur Berechnung des Gütekriteriums q bzw. der Veränderung Δq ab. Wenn zur Erzeugung eines neuen Testfelds nur geringe Veränderungen durchgeführt werden, ist es in vielen Fällen nicht notwendig, das Gütekriterium durch eine vollständige Neuberechnung zu ermitteln. Bei gegebenem Gütekriterium müssen dann lediglich die Teile des Berechnungsalgorithmus erneut analysiert werden, die sich durch die Vertauschungen im Testfeld verändert haben. Betrachten wir dazu eine einfache Vertauschung von zwei Faktorstufen i_1 und i_2 eines Faktors k , wie es bereits in Gleichung 8.38 dargestellt wurde. Wird zur Beurteilung des Testfelds das einfache MaxiMin- oder MiniMax-Kriterium (Kapitel 8.2.1) verwendet, so werden lediglich die neuen Abstände zwischen den vertauschten Testpunkten und allen anderen benötigt. Eine komplette Neuberechnung würde bei insgesamt 4 Testpunkten somit $\sum_{i=1}^{n_r-1} i = 6$ Abstände berechnen. Es verändern sich lediglich die Abstände mit den zwei vertauschten Testpunkten und somit $2(n_r - 2) = 4$ Abstände. Bei Erhöhung der Testpunkteanzahl zum Beispiel auf $n_r = 100$ ergibt sich dadurch bereits eine deutliche Verringerung des Rechenaufwands von 4950 auf 196 Abstandsberechnungen (Reduktion: 96%).

Für die Gütekriterien **MaxiMin_p**, **ZD₂** und **Entropie** (Kapitel 8.2) geben JIN, CHEN und SUDJANTO [30] einfache Berechnungsvorschriften für die Vertauschung zweier Elemente eines Faktors k ($x_{i_1,k}$ und $x_{i_2,k}$) an.

$$\text{MiniMax}_{p,\text{NEU}} = \left[\begin{array}{l} \left[\text{MiniMax}_{p,\text{ALT}} \right]^p \\ + \sum_{1 \leq j \leq n_r, j \neq i_1, i_2} \left[d_{\text{NEU}}^{-p}(i_1, j) - d_{\text{ALT}}^{-p}(i_1, j) \right] \\ + \sum_{1 \leq j \leq n_r, j \neq i_1, i_2} \left[d_{\text{NEU}}^{-p}(i_2, j) - d_{\text{ALT}}^{-p}(i_2, j) \right] \end{array} \right]^{\frac{1}{p}} \quad (8.43)$$

Zur Berechnung der **L₂-Diskrepanz** eines zentrierten Testfelds Z wird die folgende symmetrische Matrix $C = (c_{ij})$ benötigt.

$$c_{ij} = \begin{cases} \frac{1}{n_r^2} \prod_{m=1}^{n_f} \frac{1}{2} (2 + |z_{im}| + |z_{jm}| - |z_{im} - z_{jm}|) & , \text{ für } i \neq j \\ \frac{1}{n_r^2} \prod_{m=1}^{n_f} (1 + |z_{im}|) - \frac{2}{n_r} \prod_{m=1}^{n_f} \left(1 + \frac{1}{2} |z_{im}| - \frac{1}{2} z_{im}^2 \right) & , \text{ für } i = j \end{cases} \quad \text{mit } z_{im} = x_{im} - 0.5 \quad (8.44)$$

Die zentrierte Diskrepanz wird dann durch Gleichung 8.17 berechnet.

$$[\text{ZD}]^2 = \left(\frac{13}{12} \right)^{n_f} + \sum_{i,j=1}^{n_r} c_{ij} \quad (8.45)$$

Zur vereinfachten Berechnung der Qualitätsänderung werden folgende Hilfsfunktionen definiert.

$$\begin{aligned} g_i &= \prod_{m=1}^{n_f} (1 + |z_{im}|) \quad h_i = \prod_{m=1}^{n_f} \left(1 + \frac{1}{2} |z_{im}| - \frac{1}{2} z_{im}^2\right) \\ \alpha(i_1, i_2, m) &= \frac{1+|z_{i_2 m}|}{1+|z_{i_1 m}|} \quad \beta(i_1, i_2, m) = \frac{2-|z_{i_2 m}|}{2-|z_{i_1 m}|} \\ \gamma(i_1, i_2, m, j) &= \frac{2+|z_{i_2 m}|+|z_{jm}|-|z_{i_2 m}-z_{jm}|}{2+|z_{i_1 m}|+|z_{jm}|-|z_{i_1 m}-z_{jm}|} \end{aligned} \quad (8.46)$$

Für die Vertauschung von $x_{i_1 k}$ und $x_{i_2 k}$ werden weiterhin folgende Gleichungen definiert.

$$\begin{aligned} c_{i_1 j, \text{NEU}} &= c_{j i_1, \text{NEU}} = \gamma(i_1, i_2, k, j) c_{i_1 j, \text{ALT}} \\ c_{i_2 j, \text{NEU}} &= c_{j i_2, \text{NEU}} = \frac{c_{i_2 j, \text{ALT}}}{\gamma(i_1, i_2, k, j)} \\ c_{i_1 i_1, \text{NEU}} &= \frac{\alpha(i_1, i_2, k) g_{i_1}}{n_r^2} - \frac{2\alpha(i_1, i_2, k) \beta(i_1, i_2, k) h_{i_1}}{n_r} \\ c_{i_2 i_2, \text{NEU}} &= \frac{g_{i_2}}{n_r^2 \alpha(i_1, i_2, k)} - \frac{2h_{i_2}}{n_r \alpha(i_1, i_2, k) \beta(i_1, i_2, k)} \end{aligned} \quad (8.47)$$

Bei den dargestellten Gleichungen ist zu beachten, dass sie sich alle auf die x - beziehungsweise z -Werte des vorherigen Testfelds beziehen. Mit den Definitionen aus Gleichungen 8.45 bis 8.47 wird dann die zentrierte L_2 -Diskrepanz für das neue Testfeld durch Gleichung 8.48 berechnet [30].

$$\begin{aligned} [ZD]_{\text{NEU}}^2 &= [ZD]_{\text{ALT}}^2 + c_{i_1 i_1, \text{NEU}} - c_{i_1 i_1, \text{ALT}} + c_{i_2 i_2, \text{NEU}} - c_{i_2 i_2, \text{ALT}} \\ &\quad + 2 \sum_{j=1, j \neq i_1, i_2}^{n_r} (c_{i_1 j, \text{NEU}} - c_{i_1 j, \text{ALT}} + c_{i_2 j, \text{NEU}} - c_{i_2 j, \text{ALT}}) \end{aligned} \quad (8.48)$$

Zur Berechnung des **Entropie**-Kriteriums wird typischerweise eine Korrelationsmatrix $R = [r_{ij}]$ verwendet (Kapitel 8.2.2). Da R eine positiv definite Matrix ist, existiert eine Cholesky-Zerlegung der Form $R = U'U$, wobei U eine Dreiecks-Matrix mit $u_{ij} = 0, i < j$ ist. Die Determinante der Korrelationsmatrix ist dadurch mittels Gleichung 8.49 berechenbar [30].

$$|R| = \prod_{i=1}^{n_r} u_{ii}^2 \quad (8.49)$$

Die Determinante des neuen Testfelds $|R_{\text{NEU}}|$ kann nicht direkt aus der Determinante des vorhergehenden Testfelds $|R_{\text{ALT}}|$ berechnet werden. Mit $n_1 = \min(i_1, i_2)$, wobei $i_{1,2}$ die vertauschten Testpunkte darstellen, kann die Korrelationsmatrix R folgendermaßen aufgeteilt werden.

$$R = \begin{bmatrix} (R_1)_{n_1 \times n_1} & (R_2)_{n_1 \times (n_r - n_1)} \\ R'_2 & (R_3)_{(n_r - n_1) \times (n_r - n_1)} \end{bmatrix} \quad (8.50)$$

Die Cholesky-Zerlegung U wird damit durch die Zerlegungen der Matritzen R_1, R_2 und R_3 bestimmt, wobei U_3 eine obere Dreiecksmatrix ist [30].

$$U = \begin{bmatrix} (U_1)_{n_1 \times n_1} & (U_2)_{n_1 \times (n_r - n_1)} \\ 0 & (U_3)_{(n_r - n_1) \times (n_r - n_1)} \end{bmatrix} \quad (8.51)$$

Die Terme der neuen Matrix U_{NEU} werden nach der Aufteilung durch folgenden Algorithmus berechnet [18].

$$\begin{aligned} a) 1 \leq i, j \leq n_1 &\rightarrow u_{ij,\text{NEU}} = u_{ij} \\ b) i \leq n_1, n_1 \leq j \leq n_r &\rightarrow u_{ij,\text{NEU}} = \frac{r_{ij,\text{NEU}} - \sum_{k=1}^{i-1} u_{ki,\text{NEU}} u_{kj,\text{NEU}}}{u_{ii,\text{NEU}}} \\ c) n_1 \leq i \leq n_r &\rightarrow \begin{cases} 1) & u_{ii,\text{NEU}} = \sqrt{1 - \sum_{k=1}^{i-1} u_{ki,\text{NEU}}^2} \\ 2) i+1 \leq j \leq n_r &\rightarrow u_{ij} = \frac{r_{ij,\text{NEU}} - \sum_{k=1}^{i-1} u_{ki,\text{NEU}} u_{kj,\text{NEU}}}{u_{ii,\text{NEU}}} \end{cases} \end{aligned} \quad (8.52)$$

Die Reduktion des Rechenaufwands hängt bei diesem Verfahren direkt von $n_1 = \min(i_1, i_2)$ ab. Umso größer n_1 ist, desto größer ist die Reduzierung des Rechenaufwands.

8.5 Ungleicheverteilte Testfelder

In Kapitel 8.3 wurden verschiedenste Verfahren zur Erzeugung gleichverteilter Testfelder vorgestellt. Sind andere Verteilungen notwendig, so können diese mit Hilfe der Verteilungsfunktion (*Cumulative Normal Distribution*) aus gleichverteilten Feldern erzeugt werden [39]. Betrachten wir als Beispiel die Normalverteilung (Gauß-Verteilung, Glockenkurve, *Standard Normal Distribution*), welche wohl neben der Gleichverteilung die am häufigsten benötigte Verteilungsform darstellt. Sei nun x eine Zufallsvariable mit der folgenden Wahrscheinlichkeitsdichte f (Mittelwert = 0 und Varianz = 1) und Verteilungsfunktion F (Abbildung 8.15):

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{s^2}{2}} ds \quad (8.53)$$

Die y -Achse der Wahrscheinlichkeitsdichte weist den gleichen Definitionsbereich wie die normierten gleichverteilten Testfelder auf ($0 \leq F(x) \leq 1$). Da die Funktion F monoton steigend ist, kann zu jedem y -Wert genau ein dazugehöriger x -Wert ermittelt werden (Abbildung 8.15). Durch die inverse Verteilungsfunktion $F^{-1}(x)$ werden somit gleichverteilte Testpunkte in eine Normalverteilung oder jede beliebige durch $F(x)$ bestimmte Verteilung umgewandelt. Die Berechnung der inversen Normalverteilung erfolgt mittels einer numerischen Approximation, wie sie von MORO vorgeschlagen wird [48, 33].

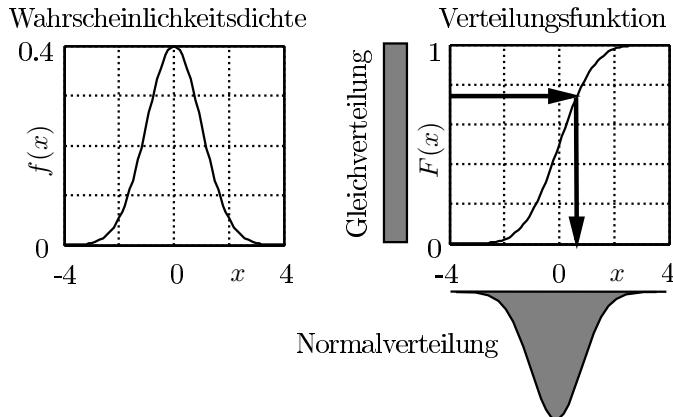


Abb. 8.15 Normalverteilung: Verteilungsfunktion und Wahrscheinlichkeitsdichte

$$\begin{aligned}
 x &= y \frac{[(a_4y^2+a_3)y^2+a_2]y^2+a_1}{\{([(b_4y^2+b_3)y^2+b_2]y^2+b_1)\}y^2+1} , \text{ wenn } |y| < 0.42 \\
 x &= (c_1 + d \{c_2 + d [c_3 + de]\}) \operatorname{sign}(y) , \text{ wenn } |y| \geq 0.42 \\
 e &= c_4 + d \{c_5 + d \{c_6 + d [c_7 + d (c_8 + dc_9)]\}\}
 \end{aligned} \tag{8.54}$$

mit $y = F(x) - 0.5$

$$\begin{aligned}
 \text{und } d &= \log(-\log(1 - F(x))) , \text{ wenn } y > 0 \\
 d &= \log(-\log(F(x))) , \text{ wenn } y \leq 0
 \end{aligned}$$

$$\begin{aligned}
 a_{1-4} &= [2.50662823884, -18.61500062529, 41.39119773534, -25.44106049637] \\
 b_{1-4} &= [-8.4735109309, 23.08336743743, -21.06224101826, 3.13082909833] \\
 c_{1-3} &= [0.337475482272615, 0.976169019091719, 0.160797971491821] \\
 c_{4-6} &= [2.76438810333863E - 2, 3.8405729373609E - 3, 3.951896511919E - 4] \\
 c_{7-9} &= [3.21767881768E - 5, 2.888167364E - 7, 3.960315187E - 7]
 \end{aligned}$$

Der zentrale Bereich der Verteilung ($|y| < 0.42$) wird dabei durch ein Verfahren von BEASLEY und SPRINGER [4] berechnet. Die schwieriger zu berechnenden Randbereiche ($|y| \geq 0.42$) werden zur Verbesserung der Genauigkeit durch eine Chebyschev-Reihe ermittelt [59]. Weitere Verfahren oder Verbesserung finden sich in den folgenden Literaturstellen [33, 76, 1, 64].

8.6 Faktorbereiche entfernen

In der Praxis können oder sollen nicht immer aller Faktorkombinationen eines Testfelds gemessen oder berechnet werden. Dieses kann zum Beispiel daran liegen, dass

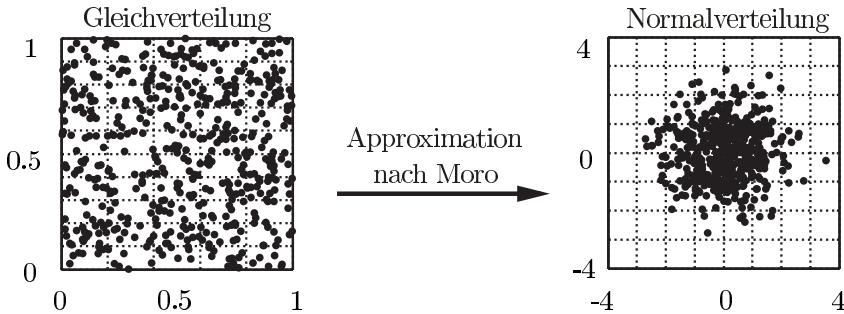


Abb. 8.16 Erzeugung einer Normalverteilung aus einer Gleichverteilung (Monte Carlo)

Kombinationen physikalisch nicht einstellbar sind oder bereits im Vorfeld als unsinnig eingestuft wurden. Sollen trotzdem als Basis gleichverteilte Testfelder mit rechtwinkligem Faktorraum verwendet werden, müssen aus den Testfeldern Unterbereiche entfernt werden. Nehmen wir als Beispiel an, dass in einem zweidimensionalen Faktorraum alle Punkte im Bereich $x_2 \geq 0.75 - 0.5x_1$ nicht geprüft werden sollen. Im einfachsten Fall werden alle Punkte, die in diesen Bereich fallen gelöscht (Abbildung 8.17, links). Alternativ ist es möglich die Punkte auf den erlaubten Rand des Faktorbereichs zu verschieben. Hierbei kann der Punkt entweder entlang einer Linie zum Zentrum des Faktorbereichs verschoben werden oder im normierten Raum auf den nächsten Randpunkt (Mitte). Ist die wahre Begrenzung des Faktorraums nicht bekannt und soll diese während der Versuchsdurchführung ermittelt werden, so wird im ersten Schritt die Begrenzung in einen problemlos messbaren Bereich gelegt. Im zweiten Schritt wird der gewählte Randpunkt \mathbf{x}_{i_1} aus dem vorherigen Ansatz geprüft. Anschließend werden entlang der Verbindungsgerade zwischen \mathbf{x}_{i_1} und der eigentlich zu prüfenden Faktorkombination \mathbf{x}_i solange Faktorkombinationen geprüft, bis die wahre messtechnische Grenze bei \mathbf{x}_{i_g} erreicht wurde. Die wahre Grenze des Faktorraums kann dann durch alle gefundenen Grenzpunkte bestimmt werden. Die Entfernung von kleinen Bereichen aus dem Testfeld ist zwar nicht optimal, aber meist akzeptabel, wenn die Randbedingungen ebenfalls in der anschließenden Analyse und Optimierung berücksichtigt werden.

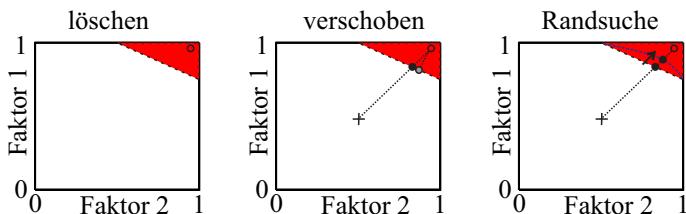


Abb. 8.17 Faktorbereich entfernen

8.7 Versuchsplanerweiterungen und Voranalyse von Messdaten

Alle folgenden Analysen, Metamodelle und Optimierungen basieren auf Messdaten, welche mittels eines optimierten Versuchsplans gewonnenen wurden. Messdaten mit hoher Qualität sind der Schlüssel zu einer erfolgreichen Projektdurchführung. Nachdem die Faktoren inklusive ihrer Variationsbereiche sowie der normierten Versuchspläne aufgestellt wurden, muss daher eine robuste Versuchsdurchführung sicher gestellt werden. Der beste Schritt in diese Richtung ist in den meisten Fällen eine Automatisierung der Versuche. Viele Fehlerursachen, wie zum Beispiel falsches Einstellen von Faktoren oder fehlerhafte Dokumentation von Systemantworten, können dadurch weitestgehend ausgeschlossen werden. In Computerexperimenten ist eine Automatisierung meist einfach, wohingegen im Bereich der physikalischen Experimente die Automatisierungsmöglichkeiten und der dazu benötigte Aufwand stark variieren. Auch wenn der Aufwand nicht vernachlässigbar ist, sollte die grundsätzliche Automatisierung in keinem Fall vernachlässigt werden, da häufig nur dadurch eine zeiteffiziente und erfolgreiche Projektdurchführung gewährleistet wird. Der am Ende eines Projekts verursachte Mehraufwand von vereinzelten Fehlmessungen überschreiten meist deutlich den zeitlichen Aufwand für eine robuste Prüfstandautomatisierung. Ein Grund dafür besteht darin, dass Fehlmessungen häufig erst nach einer kompletten Analyse, Modellbildung und Optimierung durch kuriose Ergebnisse entdeckt werden und dadurch ein Großteil des Entwicklungsprozesses wiederholt werden muss.

Neben klassischen Fehlmessungen sind zeitliche Veränderungen des zu prüfenden Systems während der Messung oder zwischen den Messungen eine Fehlerquelle, die vermieden oder wenigstens dokumentiert werden muss. Dazu wird mindestens ein Referenzpunkt mit konstanten Faktoren gleichmäßig über den Versuchsplan verteilt. Auch wenn ein Referenzpunkt nur eine feste Faktorkombination prüft kann bereits in vielen Fällen durch Veränderung des Messergebnisses am Referenzpunkt eine Veränderung des Systems erkannt werden.

Wenn auf Basis der Messdaten Metamodelle erzeugt werden, die sich flexible an komplexe Zusammenhänge anpassen können, ist es zur Abschätzung der Modellvorhersagegenauigkeit ebenfalls sinnvoll, zusätzliche Datenpunkte mit zufälligen Faktorkombinationen über den Versuchsplan zu verteilen. Diese Punkte werden später nicht zur Erzeugung der Metamodelle verwendet, sondern lediglich zur Ermittlung der Modellvorhersagegenauigkeit. Ein typischer erweiterter Versuchsplan ist wie folgt aufgebaut.

Kompletter Versuchsplan → $|r|t|v|v|v|v|\dots|v|r|v|v|v|t|v|v|v|r|v|\dots|v|t|r|$

$v \rightarrow$ eigentlicher Versuchsplan

$t \rightarrow$ Testpunkt

$r \rightarrow$ Referenzpunkt

Neben den bereits erwähnten manuellen Einstellungsfehlern von Faktoren, welche

durch eine Automatisierung minimiert werden, treten weitere Abweichungen vom geplanten Versuchsplan auf. Dies können Hysteresen in Stellantrieben, einfache Fehlfunktionen des Prüfstands oder Systembegrenzungen sein, welche die Einstellung einer speziellen Faktorkombination verhindern. Besonders problematisch sind Abweichungen, wenn in der weiteren Modellbildung und Analyse davon ausgegangen wird, dass die gewünschten Faktoreinstellungen verwendet wurden. In diesen Fällen kommt es zu Fehlinterpretationen und ungenauen Modellvorhersagen. Werden jedoch die realen Einstellungen ebenfalls gemessen, können die fehlerhaften Einstellungen schnell erkannt, sowie die Ursachen analysiert und behoben werden. Metamodelle sollten immer auf Basis der realen Faktoreinstellungen erstellt werden. Dabei muss jedoch vorher der real gefahrene Versuchsplan auf seine Eignung für weitere Analysen geprüft werden. Hierbei ist beispielsweise zu prüfen, ob die Faktoren noch unabhängig voneinander sind oder wie sich die gemessenen Faktorbereiche verändert haben, wobei meist nicht mehr von einem rechtwinkligen Faktorraum ausgegangen werden kann. Ist der real gefahrene Versuchsplan verwendbar, können in vielen Fällen die realen Faktoreinstellungen ohne großen Qualitätsverlust zur Analyse und Modellbildung eingesetzt werden. Abweichungen der Faktoren ($x_{set} - x_{real}$) werden zur besseren Interpretation relativ zu ihrem eigentlich geplanten Variationsbereich $[x_{min}, x_{max}]$ angegeben.

$$\text{Fehler}_{rel} = \frac{x_{set} - x_{real}}{x_{max} - x_{min}} \quad (8.55)$$

Soll ein Faktor x_i zum Beispiel im Bereich $x_i \in [2, 10]$ variiert werden und wird ein einzustellender Wert von $x_{i, set} = 5$ im Versuch mit $x_{i, real} = 4.8$ eingestellt, so liegt ein relativer Fehler von $\frac{0.2}{8} = 2.5\%$ vor.

Entsprechend der klassischen Versuchsplanung sollten die Messpunkte in einer zufälligen Reihenfolge geprüft werden, wodurch im Anschluss der Versuchsdurchführung verschiedene Qualitätsmerkmale anhand eines einfachen Punktediagramms kontrolliert werden können. Zur Erläuterung zeigt Abbildung 8.18 verschiedene fiktive Messergebnisse. Im Normalfall ähnelt das Punktediagramm jeder Qualitätsgröße dem Diagramm A. Der Referenzpunkt (rot) weist für alle Messungen nahezu den gleichen Messwert auf, was für ein stabiles Systemverhalten spricht. Die Messpunkte des eigentlichen Versuchsplans (schwarz) zeigen über dem gesamten Versuchsablauf eine nahezu gleichbleibende Varianz, welche durch die zufällige Reihenfolge der Messpunkte bedingt ist. Die Verteilungsdichte der Systemantworten muss dabei nicht konstant sein, da diese von den Systemeigenschaften abhängt. So können beispielsweise in einem Bereich der Systemantwort deutlich mehr Messpunkte liegen als in anderen Bereichen. Wichtig ist nur, dass diese Verteilung über den Versuchsverlauf (Versuchsnummern) annähernd gleich bleibt. Die in grün dargestellten Validierungspunkte variieren über den gesamten Qualitätsmerkmalsbereich, was vermuten lässt, dass auch diese Punkte wichtige Bereiche des Systems abbilden. Diagramm B weist im Gegensatz zu A eine Verschiebung der mittleren Systemantworten während der Versuchsdurchführung auf. Einerseits verschiebt sich der Referenzpunkt und andererseits der Mittelwert der Daten des eigentlichen Versuchsplans (schwarz), wobei die Varianz gleich bleibt. Dieses tritt gelegentlich auf,

wenn ein Versuch unterbrochen und zu einem späteren Zeitpunkt weitergeführt wird oder ein Offset (Fehler, Rekalibrierung) im Messsystem auftritt. Im Vergleich dazu zeigt Diagramm D eine Veränderung der Varianz bei konstantem Mittelwert der Versuchspunkte (schwarz), was ebenfalls ein Zeichen für eine Systemveränderung während der Versuchsdurchführung ist. Dieses Verhalten wird beispielsweise durch Temperaturänderungen oder Einlaufprozesse hervorgerufen. Diagramm C zeigt hingegen einen Ausreißer (orange gekennzeichnet), der eventuell durch eine fehlerhafte Einstellung eines Faktors oder eine Fehlmessung der Systemantwort auftritt. Auf der anderen Seite kann dieses aber auch durch ein deutlich anderes Systemverhalten genau für diese Faktorkombination hervorgerufen werden. In diesem Fall weist der Ausreißer darauf hin, dass in diesem Faktorbereich eventuell eine höhere Messpunktedichte notwendig gewesen wäre, um das komplexe Systemverhalten zu beschreiben. Egal, ob es sich um eine Veränderung des Referenzpunktes, der Varianz oder einen Ausreißer handelt, ist es unabdingbar den Grund für das Verhalten zu analysiert und zu verstehen, bevor weitere Schritte unternommen werden. Ein häufiger Fehler ist das unüberlegte Löschen von „angeblichen“ Ausreißern, bei denen sich später herausstellt, dass genau dort ein besonderes Systemverhalten vorliegt.

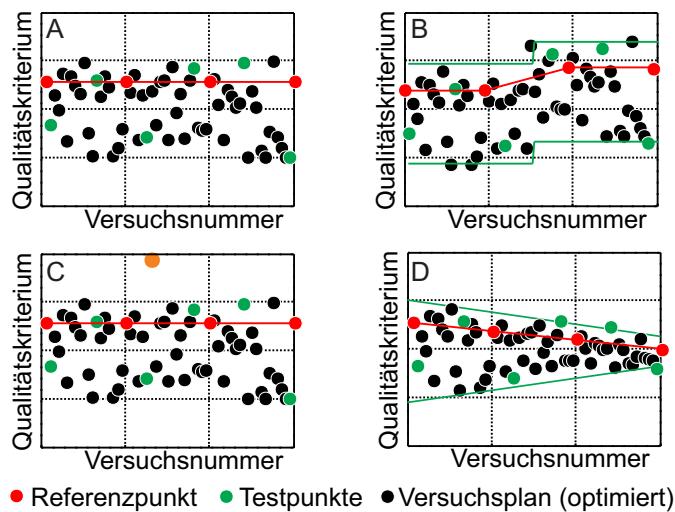


Abb. 8.18 Punktendiagramme von Messdaten

8.8 Wie viele Messungen soll ich machen?

Die wohl am häufigsten gestellte Frage ist und bleibt „Wie viele Messungen soll ich machen?“. Im Gegensatz zur klassischen Versuchsplanung, bei der auf Basis von Faktoranzahl und des gewünschten Modells (linear, quadratisch) schnell eine sinn-

volle Antwort gegeben werden kann, ist hier die Antwort „Kommt ganz darauf an!“ oder „Mehr wäre wahrscheinlich besser!“. Ein erster Anhaltspunkt zur benötigten Anzahl bietet jedoch auch hier die klassische Versuchsplanung. Im ersten Schritt könnte von einem quadratischen Zusammenhang ausgegangen werden, wodurch in Kombination mit der Faktoranzahl und einem klassischen Versuchsfeld, wie Central Composite, eine erste Versuchsanzahl bestimmt wird. Da aber die Versuchspläne für komplexe Zusammenhänge wahrscheinlich nicht zur Abbildung von einfachen Zusammenhängen gewählt wurden, sollte dieser Wert eher als untere Grenze verstanden werden. Eine höhere Systemkomplexität führt zu mehr benötigten Testpunkten. Auf der anderen Seite unterstützt der raumfüllende Aufbau und die Verwendung von mehreren Faktorstufen die Abbildung von komplexeren Zusammenhängen, so dass eine moderate Erhöhung meist ausreichend ist. Das Thema „Pseudo Wiederholungen“ (siehe Kapitel 7.2.2) kann die benötigte Anzahl der Testpunkte reduzieren, wenn beispielsweise auf eine Systemantwort niemals alle Faktoren einen signifikanten Einfluss aufweisen. Werden zum Beispiel 8 Faktoren geprüft und jede Systemantwort ist von maximal vier Faktoren abhängig, so kann bei der Planung von Punkteanzahlen für raumfüllende Versuchspläne von vier anstatt acht Faktoren ausgegangen werden.

Die reale Mess- oder Simulationszeit für jeden Datenpunkt ist in einigen Fällen der begrenzende Parameter und muss bei der Planung stets berücksichtigt werden. Sinnvoller Weise ist ein Versuchsplan so aufzustellen, dass er in einem Durchlauf komplett abgearbeitet werden kann, so dass unerwünschte Effekte durch Rekalibrierungen oder Ähnliches weitestgehend auszuschließen sind. Der verständliche Wunsch, so wenig Messpunkte wie möglich zu verwenden, birgt die Gefahr entscheidende Systemeigenschaften zu übersehen. Ein anschauliches Beispiel zur Ermittlung der Punktedichte in einem interessanten Bereich des Faktorraums ist häufig hilfreich bei der Diskussion über die benötigte Testpunkteanzahl. Nehmen wir ein Versuchsfeld mit 8 Faktoren und gehen davon aus, dass am Ende der Analyse der interessante oder optimale Bereich jeweils innerhalb von 20 Prozent der ursprünglichen Variationsbreite jedes Faktors liegt. Weiterhin sollen 300 Messpunkte aufgenommen werden. Sind am Ende lediglich 20 Prozent vom Variationsbereich des ersten Faktors $x_1 \in [0.2, 0.4]$ von Interesse, so liegen in diesem Ausschnitt des Versuchsplans $300 \times 0.2 = 60$ Datenpunkte (Abbildung 8.19). Eine Reduzierung des zweiten Faktors x_2 auf 20% ($x_2 \in [0.5, 0.7]$) resultiert in $60 \times 0.2 = 12$ Datenpunkten im betrachteten Ausschnitt. Werden alle Faktorbereiche auf einen Anteil von p_i reduziert, so liegen im verbleibenden Ausschnitt $n_r^* = n_r \prod_{i=1}^{n_f} p_i$ Datenpunkte. In dem hier dargestellten Beispiel mit einem relativ großen Anteil von 20 Prozent sind dieses dann nur noch $300 \times 0.2^8 = 0.000768$ Datenpunkte. Auch wenn dies eine sehr vereinfachte Betrachtungsweise ist, vermittelt sie dem Anwender ein grundlegendes Gefühl für das Zusammenspiel von Messpunkte- und Faktoranzahl. Bei komplexen Systemzusammenhängen in begrenzten Faktorbereichen ist es möglich, dass diese nicht optimal abgebildet werden. Jedoch weisen die meisten Analysen erstaunlich gute Ergebnisse mit relativ wenigen Datenpunkten auf, da die grundlegenden Zusammenhänge eines Systems bereits abgebildet werden. Sollten Unstimmigkeiten

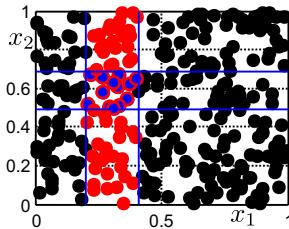


Abb. 8.19 Interessanter Ausschnitt eines Versuchsplans

oder Modellfehler während der Analyse auftreten, ist es häufig ausreichend, lediglich in einzelnen begrenzten Faktorbereichen nachträglich die Messdatendichte zu erhöhen.

8.9 Zusammenfassung

Die Anzahl von Analysen technischer Systeme mit nichtlinearen Zusammenhängen steigt in der Praxis stetig an. Besonders bei Computer-Experimenten ist durch die hohe Flexibilität der verwendeten Simulationsmodelle die Analyse von nichtlinearen Zusammenhängen für eine erfolgreiche und zielstrebige Entwicklung unumgänglich geworden. Zur Ermittlung der erforderlichen Datenbasis werden Testfelder benötigt, die mit geringem Aufwand ein Maximum an Informationen über die zu analysierenden Zusammenhänge (Faktoren und Qualitätskriterien des Systems) liefern. Sind über die zu untersuchenden Systeme (Zusammenhänge) bereits gesicherte Daten oder Informationen (Messdaten, grundlegende Funktionszusammenhänge) vorhanden, können diese zur Entwicklung spezieller Testfelder eingesetzt werden. In den überwiegenden Anwendungen sind jedoch keine oder nur wenige gesicherte Informationen verfügbar, so dass Testfelder benötigt werden, welche Informationen im gesamten Faktorbereich gleichmäßig ermitteln. In der Praxis werden diese allgemeinen Testfelder auch für Analysen mit Vorwissen eingesetzt, da sie schnell zu erstellen und robust anzuwenden sind. Die Auslegung von allgemeinen Testfeldern basiert grundsätzlich auf einer genauen und stabilen Vorhersage des globalen Mittelwerts einer zu analysierenden Systemantwort durch gute Gleichverteilung der Testpunkte im gesamten Faktorraum.

Zur Beurteilung der Güte eines Testfelds existieren verschiedene Kriterien, wie *Diskrepanz* oder Minimax_p , welche ein Maß für die Gleichverteilung der Testpunkte im Faktorraum darstellen. Grundsätzlich zeigt sich, dass zwischen den Gütekriterien Korrelationen vorhanden sind und ein optimiertes Testfeld auch durch andere Gütekriterien positiv beurteilt wird.

Zur Erzeugung von Testfeldern wurden in den letzten Jahrzehnten verschiedene Ansätze und Algorithmen entwickelt. Diese beginnen bei Monte-Carlo- oder Quasi-Monte-Carlo-Methoden, die das Testfeld bei steigender Anzahl von Testpunkten

immer feiner ausfüllen und gehen bis zu Latin Hypercubes oder gleichverteilten Testfeldern (*Uniform Designs*), welche bei gegebener Testpunkt-, Faktor- und Stufenanzahl gezielt Testfelder aufbauen und optimieren.

Die Optimierung von Testfeldern basiert dabei auf der Verbesserung eines gewählten Gütekriteriums durch zufällige oder gezielte Vertauschung von Elementen eines Testfelds. Zur Reduktion des Rechenaufwands werden vorteilhafter Weise nur die Bereiche erneut analysiert, welche zum vorhergehenden Testfeld verändert wurden. Werden für eine Analyse Testfelder benötigt, bei denen die Faktoren nicht gleichverteilt sind, so können diese mittels der gewünschten Verteilungsfunktion direkt aus einem gleichverteilten Testfeld erstellt werden. Dadurch ist es nicht notwendig, für jede Verteilungsfunktion spezielle Testfelder und Optimierungsalgorithmen zu entwickeln.

Testfelder können an simulatorische oder experimentelle Grenzen durch Entfernung oder Verschiebung von Teilbereichen angepasst werden. Wichtig bei der Anpassung ist, dass die Grenzen und Randbedingungen in anschließenden Analysen und Optimierungen ebenfalls Berücksichtigung finden.

Literaturverzeichnis

1. Acklam, P.J.: *Inverse normal cumulative distribution* (2009) 220
2. Antonov, I., Saleev, V.: *An Economic Method of Computing LP Tau-Sequences*. USSR Computational Mathematics and Mathematical Physics **19**, pp. 252–256 (1980) 203
3. Barton, R.: *Metamodeling: A State of the Art Review*. In: Proceedings of the 1994 Winter Simulation Conference (1994) 204
4. Beasley, J., Springer, S.: *Algorithm AS 111: The percentage points of the Normal Distribution*. Applied Statistics **26**, pp. 118–121 (1977) 220
5. Bronstein, I.N., Semendjajew, K.A., Musiol, G.: *Taschenbuch der Mathematik*. Harri Deutsch (2008) 213
6. Bundschuh, P., Zhu, Y.: *A method for exact calculation of the discrepancy of low-dimensional finite point sets (I)*. Abhandlungen aus Math. Seminar (Univ. Hamburg) **63**, pp. 115–133 (1993) 195
7. Burkardt, J.: *Homepage* (2009). URL <http://people.scs.fsu.edu/~burkardt/>. (abgerufen 11/2016) 203
8. Caldwell, C.: *The Prime Pages* (2009). URL <http://primes.utm.edu>. (abgerufen 11/2016) 200
9. Dueck, G., Scheuer, T.: *Threshold accepting: A general purpose algorithm appearing superior to simulated annealing*. J. of Comp. Physics **90**, pp. 161–175 (1990) 215
10. Fahrmeir, L., Kneib, T., Lang, S.: *Regression*. Springer Verlag, Berlin Heidelberg (2009) 55, 56, 61, 65, 213
11. Fang, K.: *The uniform design: application of number-theoretic methods in experimental design*. Acta Math. Appl. Sinica **3**, pp. 363–372 (1980) 194
12. Fang, K.: *Theory, method and applications of the uniform design*. Inter. J. Reliability, Quality and Safety Engineering **9**, pp. 305–315 (2002) 194
13. Fang, K., Hickernell, F.: *The uniform design and its application*. Bulletin of The International Statistical Institute pp. 339–349 (1995) 194
14. Fang, K., Li, J.: *Some new results on uniform design*. Chinese Science Bulletin **40**, pp. 268–272 (1994) 210
15. Fang, K., Lin, D.: *Uniform designs and their application in industry*. In: Handbook on Statistics in Industry, pp. 131–170. Elsevier, North-Holland, Amsterdam (2003) 194

16. Fang, K., Lin, D., Winker, P., Zhang, Y.: *Uniform design: Theory and applications*. Technometrics **42**, pp. 237–248 (2000) 194
17. Fang, K.T., Li, R.: *Bayesian statistical inference on elliptical matrix distributions*. J. Multivar. Anal. **70**(1), pp. 66–85 (1999) 212
18. Fang, K.T., Li, R., Sudjianto, A.: *Design and Modeling for Computer Experiments (Computer Science & Data Analysis)*. Chapman & Hall/CRC (2005) 192, 193, 194, 195, 197, 209, 210, 212, 213, 214, 219, 238, 254, 255, 417, 418, 419
19. Fang, K.T., Lu, X., Winker, P.: *Lower bounds for centered and wrap-around L2-discrepancies and construction of uniform designs by threshold accepting*. J. Complex. **19**(5), pp. 692–711 (2003) 213
20. Fang, K.T., Ma, C.X., Maringer, D., Tang, Y., Winker, P.: *Uniform Designs* (2005). URL <http://www.math.hkbu.edu.hk/UniformDesign/main.html>. (abgerufen 11/2016) 210
21. Fang, K.T., Ma, C.X., Winker, P.: *Centered L2-discrepancy of random sampling and Latin hypercube design and construction of uniform designs*. Mathematics of Computation **71**(237), pp. 275–296 (2000). URL <http://www.ams.org/mcom/2002-71-237/S0025-5718-00-01281-3/S0025-5718-00-01281-3.pdf>. (abgerufen 11/2016) 209
22. Fang, K.T., Tang, Y., Yin, J.: *Lower bounds for wrap-around L2-discrepancy and constructions of symmetrical uniform designs*. J. Complex. **21**(5), pp. 757–771 (2005) 197
23. Galanti, S., Jung, A.R.: *Low-Discrepancy Sequences: Monte Carlo Simulation of Option Prices*. Journal of Derivatives pp. 63–68 (1997) 202, 203
24. Hedayat, A., Sloane, N., Stufken, J.: *Orthogonal arrays: Theory and Applications*. Springer-Verlag, New York (1999) 204
25. Hickernell, F.: *A generalized discrepancy and quadrature error bound*. Math. Comp. **67**, pp. 299–322 (1998) 196, 197
26. Hickernell, F.: *Random and Quasi Random Point Sets*, chap. Lattice rules: How well do they measure up?, pp. 106–166. Springer Verlag, Berlin, New York (1998) 196, 197
27. Hickernell, F.: *Monte Carlo and Quasi-Monte Carlo Methods 1998*, chap. What affects the accuracy of quasi-Monte Carlo quadrature?, pp. 16–55. Springer Verlag, Berlin (2000) 197
28. Huth, F.: *Simulation eines Hochdruck-Benzineinspritzsystems*. Master's thesis, RWTH Aachen (2011) 191, 193, 287
29. Iman, R.L., Conover, W.J.: *A distribution-free approach to inducing rank correlation among input variables*. Communications in Statistics - Simulation and Computation **11**, pp. 311 – 334 (1982) 207
30. Jin, R., Chen, W., Sudjianto, A.: *An efficient algorithm for constructing optimal design of computer experiments*. Journal of Statistical Planning and Inference **134**(1), pp. 268–287 (2005) 216, 217, 218
31. Johnson, M.E., Moore, L.M., Ylvisaker, D.: *Minimax and maximin distance designs*. Journal of Statistical Planning and Inference **26**(26), pp. 131–148 (1990) 191
32. Joseph, V.R., Ying, H.: *Orthogonal-Maximin Latin Hypercube Designs*. Statistica Sinica **18**, pp. 171–186 (2008) 192, 193
33. Joy, C., Boyle, P., Tan, K.: *Quasi-Monte Carlo Methods in Numerical Finance*. Management Science **42**, pp. 926–938 (1996) 219, 220
34. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: *Optimization by simulated annealing*. Science **220**, pp. 671–680 (1983) 216
35. Knuth, D.: *The Art of Computer Programming: seminumerical algorithms*, vol. 2. Addison-Wesley (2002) 199
36. Ko, C.W., Lee, J., Queyranne, M.: *An Exact Algorithm for Maximum Entropy Sampling*. Operations Research **43**(4), pp. 684–691 (1995) 193
37. Korobov, M.: *The approximate computation of multiple integrals*. Dokl. Akad. Nauk SSSR **124**, pp. 1207–1210 (1959) 209
38. Kuhfeld, W.F.: *Orthogonal Arrays*. Advanced Analytics Division, SAS (2016). URL <http://support.sas.com/techsup/technote/ts723.html>. (abgerufen 11/2016) 204

39. Law, A.M., Kelton, W.D.: *Simulation modeling and analysis*. McGraw-Hill Science/Engineering/Math (2000) 219
40. Leary, S., Bhaskar, A., Keane, A.: *Optimal orthogonal-array-based latin hypercubes*. Journal of Applied Statistics **30**, pp. 585–598 (2003) 206
41. Li, W.W., Wu, C.F.J.: *Columnwise-Pairwise Algorithms With Applications to the Construction of Supersaturated Designs*. Technometrics (39), pp. 171–179 (1997) 209, 214
42. Liang, Y., Fang, K., Xu, Q.: *Design and its applications in chemistry and chemical engineering*. Chemom. Intell. Lab. Systems **58**, pp. 43–57 (2001) 194
43. Ma, C., Fang, K.T.: *A new approach to construction of nearly uniform designs*. International Journal of Materials and Product Technology **20**, pp. 115–126 (2004). URL <http://www.math.hkbu.edu.hk/UniformDesign/ud2003.file/Theory/UD-03-Ma-30.pdf>. (abgerufen 11/2016) 209, 210, 212
44. Matsumoto, M.: *SIMD-oriented Fast Mersenne Twister (SFMT): twice faster than Mersenne Twister* (2009). URL <http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/SFMT/index.html>. (abgerufen 11/2016) 199
45. Matsumoto, M., Nishimura, T.: *Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator*. ACM Trans. on Modelling and Computer Simulation **8**, pp. 117–121 (1998) 199
46. Matsumoto, M., Saito, M.: *SIMD-oriented Fast Mersenne Twister: A 128-bit Pseudorandom Number Generator*. In: Monte Carlo and Quasi-Monte Carlo Methods 2006, pp. 607–622. Springer Verlag (2008) 199
47. McKay, M., Beckman, R., Conover, W.: *A comparison of three methods for selecting values of input variables in the analysis of output of a computer code*. Technometrics **21**, pp. 239–245 (1979) 205, 206
48. Moro, B.: *The Full Monte*. Risk **8** (1995) 219
49. Morris, M., Mitchell, T.: *Exploratory design for computational experiments*. Journal of Statistical Planning and Inference **34**(26), pp. 381–402 (1995) 192, 207
50. Niederreiter, H. (ed.): *Random number generation and quasi-Monte Carlo methods, CBMS-NSF regional conference series in applied mathematics*, vol. 63. Society for Industrial and Applied Mathematics (1992) 196, 204
51. Owen, A.B.: *A central limit theorem for Latin Hypercube sampling*. Journal of the Royal Statistical Society **54**, pp. 541–551 (1992) 206
52. Owen, A.B.: *Orthogonal Arrays for Computer Experiments, Integration and Visualisation*. Statistica Sinica (2), pp. 439–452 (1992) 204, 206
53. Owen, A.B.: *Controlling Correlations in Latin Hypercube Samples*. Journal of the American Statistical Association **89**, pp. 1517–1522 (1994) 207
54. Owen, A.B.: *Lattice Sampling Revisited: Monte Carlo Variance of Means Over Randomized Orthogonal Arrays*. Annals of Statistic **22**, pp. 930–945 (1994). URL http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&handle=euclid-aos/1176325504. (abgerufen 11/2016) 206
55. Panneton, F., L'Ecuyer, P., Matsumoto, M.: *Improved long-period generators based on linear recurrences modulo 2*. ACM Trans. on Mathematical Software **32**, pp. 1–16 (2006) 199
56. Park, J.: *Optimal Latin-Hypercube Designs for Computer Experiments*. Journal of Statistical Planning and Inference (39), pp. 95–111 (1994) 207
57. Park, S., Miller, K.: *Random number generators: good ones are hard to find*. Association for Computing Machinery **31**, pp. 1192–2001 (1988) 199
58. Paskov, S., Traub, J.: *Faster valuation of financial derivatives*. Journal of Portfolio Management **22**, pp. 113–120 (1995) 203
59. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press (2007) 199, 203, 220, 254, 265, 276, 277, 281, 285, 307, 334, 381, 398, 400
60. Saab, Y., Rao, V.: *Combinatorial optimization by stochastic evolution*. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **10**, pp. 525–535 (1991) 216

61. Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P.: *Design and Analysis of Computer Experiments*. Statistical Sience **4**, pp. 409–423 (1989) 179, 193, 205
62. Schmid, W.C., Schürer, R.: *MinT, the online database for optimal parameters of (t, m, s) -nets, (t, s) -sequences, orthogonal arrays, linear codes, and OOAs!* (2009). URL <http://mint.sbg.ac.at/index.php?i=0>. (abgerufen 11/2016) 204
63. Shannon, C.E.: *A Mathematical Theory of Communication*. Bell System Technical Journal **27**, pp. 379,423, 623,656 (1948) 193
64. Shaw, W.: *Refinement of the Normal Quantile: A benchmark Normal quantile based on recursion, and an appraisal of the Beasley-Springer-Moro, Acklam, and Wichura (AS241) methods*. Tech. rep., Financial Mathematics Group, King's College London (2007) 220
65. Shewry, M., Wynn, H.: *Maximum entropy design*. Journal of Applied Statistics (14), pp. 165–170 (1987) 193
66. da Silva, M.E., Barbe, T.: *Quasi Monte Carlo in Finance: Extending for High Dimensional Problems*. Economia Aplicada **9** (2005) 203
67. Sloane, N.: *Orthogonale Felder* (2009). URL <http://neilsloane.com/oadir/>. (abgerufen 11/2016) 204
68. Sloane, N.: *NOA: Program for Constructing Orthogonal Arrays, Near-Orthogonal Arrays and Supersaturated Designs* (2016). URL <http://designcomputing.net/gendex/noa/>. (abgerufen 11/2016) 204
69. Sobol, I.M.: *On the distribution of points in a cube and the approximate evaluation of integrals*. U.S.S.R. Computational Math. and Math. Phys **7**, pp. 86–112 (1967) 203
70. Stein, M.: *Large sample properties of simulations using latin hypercube sampling*. Technometrics **29**, pp. 143–151 (1987) 206
71. Tang, B.: *Orthogonal array-based latin hypercubes*. Journal of the American Statistical Association **88**, pp. 1392–1397 (1993) 206
72. Tang, Y., Fang, K.T.: *Constructions for Uniform Designs under Wrap-around Discrepancy*. In: Symposium on the Uniform Experimental Design (2003) 197
73. Tezuka, S.: *Random and Quasi-Random Point Sets, Lecture Notes in Statistics*, chap. Financial applications of Monte Carlo and quasi-Monte Carlo methods, pp. 303–332. Springer Verlag, Berlin (1998) 203
74. Wang, Y., Fang, K.: *A note on uniform distribution and experimental design*. KeXue TongBao **26**, pp. 485–489 (1981) 194, 210
75. Warnock, T.: *Computational investigations of low discrepancy point sets*. In: Applications of Number Theory to Numerical Analysis, pp. 319–343. Zaremba, S.K. (1972) 195
76. Wichura, M.: *Algorithm AS 241: The Percentage Points of the Normal Distribution*. Applied Statistics **37**, pp. 477–484 (1988) 220
77. Winker, P.: *Optimization Heuristics in Econometrics : Applications of Threshold Accepting*. Wiley, Chichester (2000) 215
78. Winker, P., Fang, K.T.: *Application of Threshold-Accepting to the Evaluation of the Discrepancy of a Set of Points*. SIAM J. Numer. Anal. **34**(5), pp. 2028–2042 (1997) 195
79. Xu, H.: *An Algorithm for Constructing Orthogonal and Nearly Orthogonal Arrays with Mixed Levels and Small Runs*. Tech. rep., Department of Statistics, University of California, Los Angeles (2000). URL <http://repositories.cdlib.org/cgi/viewcontent.cgi?article=1210&context=uclastat>. (abgerufen 11/2016) 204
80. Xu, H.: *An Algorithm for Constructing Orthogonal and Nearly Orthogonal Arrays with Mixed Levels and Small Runs*. Technometrics **44**, pp. 356–368 (2002) 204
81. Ye, K.: *Orthogonal Column Latin Hypercubes and their application in Computer Experiments*. Journal of the American Statistical Association **93**, pp. 1430–1439 (1998) 207

Kapitel 9

Metamodelle

9.1 Einleitung

Der direkte Einsatz komplexer Simulationsmodelle ist durch lange Rechenzeiten in vielen Fällen nur eingeschränkt zur Analyse technischer Systeme sinnvoll. Aus diesem Grund werden alternativ sogenannte Metamodelle (auch Transferfunktionen, Surrogate-, Approximations- oder Ersatzmodelle genannt) verwendet, welche mit deutlich geringeren Rechenzeiten und ausreichend genauen Vorhersagen das komplexe Simulationsmodell abbilden. Die Rechenzeiten von Metamodellen liegen dabei im Bereich von Millisekunden, während die ursprünglichen Modelle teilweise Stunden, Tage oder Wochen für die Berechnung eines einzelnen Ergebnisses benötigen.

Die Metamodelle werden dabei vorteilhafter Weise aus Daten gewonnen, die auf Basis eines raumfüllenden Testfelds (siehe Kapitel 8) ermittelt wurden. Diese Testfelder sind dabei so ausgelegt, dass die erzeugten Daten über den gesamten zu analysierenden Faktorraum eine maximale Informationsmenge bei minimaler Versuchsanzahl ergeben. Sie liefern Informationen um nicht nur lineare, sondern auch komplexe Zusammenhänge zwischen Eingangsvariablen (Faktoren) und den zu analysierenden Ausgangsvariablen ausreichend genau abzubilden. Zur Erzeugung der benötigten Metamodelle werden Algorithmen eingesetzt, welche vorhandene, beziehungsweise speziell erzeugte Trainingsdaten optimal zur Modellierung des zu untersuchenden Systems nutzen.

Vor der Analyse eines technischen Systems oder eines Simulationsmodells sind in vielen praktischen Fällen nur geringe Kenntnisse über den genauen Zusammenhang zwischen Eingangsvariablen (Faktoren) und den zu betrachtenden Ausgangsvariablen bekannt. Gerade in diesen Fällen sind Algorithmen zur Erstellung der Metamodelle notwendig, die keine expliziten Vorgaben über die zu erwartenden Zusammenhänge benötigen und sich selbstständig und flexibel an die jeweilige Komplexität des betrachteten Systems anpassen.

Werden aus Unkenntnis Verfahren verwendet, welche eine Vorgabe von Zusammenhängen benötigen, können bei falscher Wahl dieser Abhängigkeiten nur unzurei-

chend genaue Metamodelle erstellt werden. Liegen jedoch Informationen über die Zusammenhänge der Faktoren und der zu analysierenden Ausgangsvariablen vor, so zeigt sich, dass auch mit klassischen Verfahren, wie zum Beispiel linearer Regression, ausreichend genaue Metamodelle erzeugt werden können.

Unabhängig von der gewählten Methode zur Metamodellerzeugung gilt, dass die Modelle für eine weitere Analyse erst dann eingesetzt werden können, wenn sichergestellt ist, dass sie eine ausreichende Genauigkeit bei der Vorhersage von neuen Kombinationen der Faktoren aufweisen. Bei Verwendung von klassischen Regressionsmodellen ist dieses durch eine Analyse der Residuen¹ an den gegebenen Stützstellen (Trainingsdaten) möglich (Gleichverteilung, Absolutwerte, ...). Werden hingegen selbst anpassende Algorithmen ohne Vorgabe fester Zusammenhänge verwendet, so ist die Betrachtung der Residuen meist nicht ausreichend, da diese an den Stützstellen durch die Flexibilität der Modellanpassung schnell kleine Absolutwerte annehmen. In diesen Fällen sind weitere Prüfungen der Vorhersagequalität der Metamodelle notwendig. Dazu werden typischerweise die Abweichungen der Vorhersage an Datenpunkten herangezogen, die nicht zur Erzeugung des Metamodells verwendet wurden.

In der Praxis zeigt sich, dass die hier dargestellten Modellverfahren nicht nur für Computerexperimente, sondern ebenfalls äußerst gut im Bereich von physikalischen Experimenten einsetzbar sind.

9.2 Lineare Regression

Die lineare Regression wird bei vielen Analysen und Entwicklungsaufgaben zur Erzeugung von Metamodellen eingesetzt. Durch die umfangreichen Erfahrungen, welche über den Einsatz von linearen Regressionen vorliegen, ist dieses Verfahren auch bei komplexeren Systemen mit gutem Erfolg einsetzbar. Gerade wenn *alle* grundlegenden Zusammenhänge zwischen Ein- und Ausgangsvariablen bekannt sind, liefert die lineare Regression bei richtiger Anwendung ausreichend genaue Metamodelle in Kombination mit einem geringen Rechenaufwand. Da eine Vielzahl von Büchern zur linearen Regression existiert, werden hier nur die wichtigsten Grundideen dargestellt, die für das Verständnis folgender Verfahren hilfreich sind. Das allgemeine lineare Regressionsmodell weist die in Gleichung 9.1 dargestellte Form auf.

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_{n_f} x_{in_f} + \varepsilon_i \quad (9.1)$$

y_i ist dabei die zu analysierende Ausgangsvariable und x_{i1}, \dots, x_{in_f} sind unabhängige Faktoren welche y_i beeinflussen. b_0, \dots, b_{n_f} sind unbekannte Konstanten, welche zur Anpassung des linearen Regressionsmodells mittels gegebener Messdaten (Trainingsdaten) bestimmt werden. Der Term ε steht für einen normalverteilten Fehler, der den Teil des zu beschreibenden Systems darstellt, der nicht durch das lineare Regressionsmodell erklärt werden kann. ε weist einen Erwartungswert von Null und

¹ Abweichung zwischen realem und approximiertem Wert

eine Varianz σ^2 auf $[\varepsilon \sim \mathcal{N}(0, \sigma^2)]$. Allgemein wird das lineare Regressionsmodell wie folgt dargestellt:

$$y_i = \sum_{j=0}^{n_f} b_j x_{ij} + \varepsilon_i \quad \text{mit } x_{i0} = 1 \quad (9.2)$$

Dieses entspricht in Matrix-Schreibweise:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_r} \end{pmatrix}, \quad X = \begin{pmatrix} x_{10} & \cdots & x_{1n_f} \\ \vdots & \ddots & \vdots \\ x_{n_r 0} & \cdots & x_{n_r n_f} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n_f} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{n_r} \end{pmatrix} \quad (9.3)$$

Zur Ermittlung der gesuchten Konstanten b_0 bis b_{n_f} wurde bereits Anfang des 19^{ten} Jahrhunderts die *Methode der kleinsten Fehlerquadrate* (method of least squares) eingeführt. Dabei wird die Summe $S(\mathbf{b})$ der quadratischen Abweichungen zwischen den wahren Datenpunkten y_i und den Vorhersagen des linearen Regressionsmodells \hat{y}_i minimiert.

$$S(\mathbf{b}) = \sum_{i=1}^{n_r} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n_r} \left(y_i - \sum_{j=0}^{n_f} x_{ij} b_j \right)^2$$

$$\Rightarrow S(\mathbf{b}) = (\mathbf{y} - X\mathbf{b})'(\mathbf{y} - X\mathbf{b}) \quad (9.4)$$

Das Ableiten nach \mathbf{b} und zu Null setzen von $S(\mathbf{b})$ führt zu der Gleichung:

$$X' \mathbf{y} = X' X \mathbf{b} \quad (9.5)$$

Ist $X' X$ invertierbar, können die gesuchten Koeffizienten \mathbf{b} mit Gleichung 9.6 abgeschätzt werden:

$$\hat{\mathbf{b}} = (X' X)^{-1} X' \mathbf{y} \quad (9.6)$$

Die Varianz σ^2 des Fehlers ε kann mittels Gleichung 9.7 und den gegebenen Trainingsdaten abgeschätzt werden.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n_r} (y_i - \hat{y}_i)^2}{n_r - n_f - 1} = \frac{\sum_{i=1}^{n_r} \left(y_i - \sum_{j=0}^{n_f} x_{ij} \hat{b}_j \right)^2}{n_r - n_f - 1} = \frac{\sum_{i=1}^{n_r} (y_i - \hat{y}_i)^2}{n_r - n_p} \quad (9.7)$$

n_p steht dabei für die Anzahl der zu bestimmenden Parameter. Bei einem linearen Regressionsmodell entspricht dies der Faktanzahl plus eins, für den konstanten Term ($n_p = n_f + 1$). Eine Überprüfung der Gültigkeit und der Vorhersagegenauigkeit eines linearen Regressionsmodells ist gerade bei komplexeren Systemen unabdingbar. Hierzu werden typischerweise die Abweichungen zwischen gemessenen Datenpunkten und den Vorhersagen des linearen Regressionsmodells analysiert (Residuen): $r_i = y_i - \hat{y}_i$

Die Residuen müssen dabei über alle Trainingsdaten normalverteilt sein und für die

durchzuführende Analyse ausreichend geringe Absolutwerte aufweisen. Eine weitere Darstellung der linearen Regression findet sich in Kapitel 1.3.4 und in einer Vielzahl von mathematischen Fachbüchern [82, 64].

9.3 Polynome

Polynome werden in unterschiedlichsten Bereichen zur Bildung von Metamodellen eingesetzt, wobei typische polynomiale Metamodelle die folgende Form aufweisen:

$$\hat{y}_i(\mathbf{x}_i) = b_0 + b_1x_{i1} + b_2x_{i1}^2 + b_3x_{i2} + b_4x_{i2}^2 + \cdots + b_kx_{i1}x_{i2} + \cdots \quad (9.8)$$

Durch einfache Substitution der nichtlinearen Terme mit Ersatzvariablen wird ein Polynom in ein lineares Gleichungssystem umgewandelt, so dass auch hier die Regressionskoeffizienten mittels der *Methode der kleinsten Fehlerquadrate* berechenbar sind. Im dargestellten Beispiel würden folgende Ersetzungen durchgeführt werden:

$$\begin{aligned} x_{in_f+1} &= x_{i1}^2, & x_{in_f+2} &= x_{i2}^2, & \cdots, & x_{in_f+3} &= x_{i1}x_{i2}, & \cdots \\ \Rightarrow \hat{y}_i(\mathbf{x}_i) &= b_0 + b_1x_{i1} + b_2x_{in_f+1} + b_3x_{i2} + b_4x_{in_f+2} + \cdots + b_kx_{in_f+3} + \cdots \end{aligned} \quad (9.9)$$

Verallgemeinert kann jede Hauptvariable und jede Substitution auch durch eine Basisfunktion $g_j(\mathbf{x}_i)$ dargestellt werden, wobei $g_j(\mathbf{x}_i)$ für jede beliebige Kombination der Hauptfaktoren x_{i1}, \dots, x_{in_f} steht. In dem hier dargestellten Beispiel wäre dieses $g_0(\mathbf{x}_i) = 1$, $g_1(\mathbf{x}_i) = x_{i1}$ oder $g_{n_f+3}(\mathbf{x}_i) = x_{i1}x_{i2}$.

Die Anzahl der Polynom-Terme steigt mit Erhöhung der Faktoranzahl oder des Grads des Polynoms stark an, wodurch ein hoher Rechenaufwand zur Bestimmung der Konstanten b_i und eine hohe Anzahl von Datenpunkten notwendig wird. Aus diesem Grund werden Polynome meist auf maximal quadratische oder kubische Terme beschränkt, wobei in modernen Softwaretools auch höhere Ordnungen zur Verfügung stehen.

Bei komplexeren Polynomen ist weiterhin zu berücksichtigen, dass die Terme meist nicht orthogonal sind, was die eindeutige Zuweisung von Effekten zu einzelnen Termen erschwert. Sollte dieses zu Problemen während einer Analyse führen, kann ein Verfahren von AN und OWEN Abhilfe schaffen, welches basierend auf univariablen Funktionen orthogonale Polynome erzeugt [2].

9.3.1 Faktorwahl

In der Praxis treten eine Vielzahl von Faktoren (Parameter) auf, die während der Analyse eines technischen Systems berücksichtigt werden können. Die schnell steigende Anzahl n_p der möglichen Basisfunktion $g_j(\mathbf{x}_i)$ $[1, x_{i1}, x_{i2}, x_{i1}x_{i2}, x_{i1}^2, x_{i1}^3, \dots]$ eines allgemeinen linearen (polynom) Modells (Gleichung 9.10) mit n_f Faktoren und ei-

ner maximalen Potenz von p lässt sich durch Gleichung 9.11 bestimmen [47].

$$\hat{y}_i(\mathbf{x}_i) = \sum_{j=0}^{n_p-1} b_j g_j(\mathbf{x}_i) \quad (9.10)$$

$$n_p = \prod_{k=1}^p \left(1 + \frac{n_f}{k}\right) \quad (9.11)$$

Zum Beispiel führen zwei Faktoren ($x_1, x_2 \rightarrow n_f = 2$) und eine maximale Potenz von $p = 3$ bereits zu zehn Basisfunktionen.

$$\begin{aligned} \hat{y}(x_1, x_2) &= b_0 + b_1 x_1 + b_2 x_1^2 + b_3 x_1^3 + b_4 x_2 + b_5 x_2^2 + b_6 x_2^3 \\ &\quad + b_7 x_1 x_2 + b_8 x_1^2 x_2 + b_9 x_1 x_2^2 \\ n_b &= \left[1 + \frac{2}{1}\right] \cdot \left[1 + \frac{2}{2}\right] \cdot \left[1 + \frac{2}{3}\right] = 3 \cdot 2 \cdot \frac{5}{3} = 10 \end{aligned} \quad (9.12)$$

Komplexe Metamodelle mit vielen Basisfunktionen neigen zu Overfitting und somit schlechten Vorhersagen für unbekannte Faktorkombinationen (siehe Kapitel 9.16, Abbildung 9.47). Weiterhin steigt die benötigte Anzahl an Trainingsdaten mit der Anzahl der Basisfunktionen, da minimal so viele Trainingsdaten wie Basisfunktionen vorhanden sein müssen um alle Parameter b_j zu bestimmen. Im Gegensatz dazu bilden zu einfache Modelle einen komplexen Zusammenhang nicht ausreichend genau ab, was ebenfalls zu einer schlechten Approximationsgüte führt. Somit ist neben der Bestimmung der optimalen Parameter \mathbf{b} eines Regressionsmodells ebenfalls die Auswahl der richtigen Basisfunktionen erforderlich [47, 37, 77] (siehe auch Kapitel 9.5).

Existieren n_p mögliche Basisfunktionen, können diese in $n_k = 2^{n_p}$ Varianten kombiniert werden. Eine Berechnung und Beurteilung jedes der n_k Submodelle ist meist zu rechen- und zeitaufwendig, so dass es sinnvoller ist, während der Erstellung des Metamodells sequentiell eine Auswahl der möglichen Basisfunktionen durchzuführen.

Zur Bestimmung signifikanter Faktoren existieren verschiedene Verfahren, die ein *gutes*, aber nicht unbedingt das für die Aufgabe *beste* Metamodell erzeugen [47, 37, 77].

Sequentielle Vorwärts Selektion

Bei der *Sequentiellen Vorwärts Selektion* (sequential forward selection, SFS) wird mit einem einfachen Basismodell M_0 gestartet (siehe auch Kapitel 9.5). Dieses Modell enthält meistens nur einen konstanten Term ($\hat{y} = b_0$). Sind bereits Faktoren bekannt, die in jedem Fall für das Metamodell signifikant sind, werden diese ebenfalls dem Basismodell hinzugefügt. Anschließend wird das Modell in einem iterativen Prozess mit neuen Basisfunktionen erweitert. In jedem Schritt wird dazu dem aktuellen Metamodell jede noch nicht verwendete Basisfunktion einzeln hinzuge-

fügt und die Verbesserung der Modellgüte bestimmt. Die Basisfunktion g^* , welche die Modellgüte am meisten verbessert, wird schlussendlich im Modell verwendet. Im folgenden Iterationsschritt wird das so erweiterte Modell als Basismodell $M_i = M_{i-1} + g^*$ verwendet. Die Erweiterung des Metamodells endet, wenn eine vorgegebene Modellgüte erreicht oder keine signifikante Verbesserung mehr erzielt wird [61].

Sequentielle Rückwärts Selektion

Die *Sequentielle Rückwärts Selektion* (sequential backward selection, SBS) geht im Gegensatz zur Vorwärts-Selektion von einem Metamodell aus, welches alle zur Verfügung stehenden Basisfunktionen enthält. In jeder Iteration wird die Basisfunktion gelöscht, deren Entfernung zur höchstmöglichen Verbesserung des Gütekriteriums führt [61].

Sequentielle Gleitende Vorwärts Selektion

Der Begriff *Sequentielle Gleitende Vorwärts Selektion* (Sequential Floating Forward Selection, SFFS) bezeichnet eine Kombination der Vorwärts- und Rückwärts-Selektion, wobei nach jedem zugefügten Faktor geprüft wird, ob ein oder iterativ mehrere beliebige Basisfaktoren aus dem Metamodell entfernt werden können [71, 23]. Der Prozess wird beendet, wenn in der Vorwärtsphase keine Basisfunktion mehr hinzugefügt wird (siehe SFS).

Gütekriterien

Zur Beurteilung der Modellgüte werden in der Praxis verschiedene Kriterien verwendet, die im einfachsten Fall lediglich von den Residuen² abhängen (R^2). Zur Vermeidung von Overfitting an verrauschte oder fehlerhaft ermittelte Daten ist es sinnvoll einen weiteren Term einzuführen, der zu viele Parameter n_p im Verhältnis zur Anzahl der Trainingsdaten n_r bestraft und das Gütekriterium verschlechtert [1, 87, 57, 61].

In der Tabelle 9.1 bezieht sich SSE_p auf das aktuell zu untersuchende Modell und $\hat{\sigma}_a$ auf das Modell mit allen Parametern (siehe Gleichung 9.7). Der Einsatz des *BIC*-Kriteriums führt im Vergleich zum *AIC*-Kriterium typischerweise zu weniger Faktoren im optimierten Modell. *AICc* ist eine Erweiterung des *AIC* Kriteriums zum Einsatz bei wenigen Trainingsdaten. Das C_p -Kriterium muss mit der Randbedingung $C_p \approx p$ minimiert werden. Neben diesen häufig eingesetzten Gütekriterien existieren erweiterte Kriterien wie *CIC*, *DIC*, *EIC*, *MAIC3*, *CAIC*, *ABIC*, auf die hier nicht weiter eingegangen wird [79, 91, 45, 7, 78]. Es ist anzumerken, dass klei-

² Abweichung zwischen Trainingsdaten und Modellvorhersage

ne Änderungen in den Trainingsdaten teilweise zu deutlich anderen Metamodellen führen.

Name	Gleichung
Bestimmtheitsmaß	$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^{n_r} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_r} (y_i - \bar{y})^2}$
angepasstes R^2	$R_{adj}^2 = 1 - \left(\frac{n_r - 1}{n_r - n_p} \right) (1 - R^2) = 1 - \frac{SSE}{SST} \frac{n_r - 1}{n_r - n_p}$
Likelihood für Regression	$-2\ln L = n_r \ln \left(\frac{SSE}{n_r} \right)$
Aikaike's Informationskriterium	$AIC = -2\ln L + 2(n_p + 1)$
AIC_c , kleine Datenmenge	$AIC_c = AIC + \frac{2(n_p + 1)(n_p + 2)}{n_r - (n_p + 1) - 1}$
Bayessches Informationskriterium	$BIC = -2\ln L + \ln(n_r)(n_p + 1)$
Mallows C_p	$C_p = \frac{SSE_p}{\hat{\sigma}_u} - n_r + 2n_p$
PRESS, Leave-One-Out	$PRESS = \sum_{i=1}^{n_r} (y_i - \hat{y}_i^{\setminus i})^2 = \sum_{i=1}^{n_r} \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}} \right)^2$
PRESS R^2	$\text{PRESS } R^2 = 1 - \frac{PRESS}{SST}$

Tabelle 9.1 Gütekriterien

mit

$$\begin{aligned}
 SSE &= \sum_{i=1}^{n_r} (y_i - \hat{y}_i)^2 \\
 SSR &= \sum_{i=1}^{n_r} (\hat{y}_i - \bar{y})^2 \\
 SST &= \sum_{i=1}^{n_r} (y_i - \bar{y})^2 = SSE + SSR \\
 MSE &= \frac{1}{n_r} \sum_{i=1}^{n_r} (y_i - \hat{y}_i)^2 \\
 H &= X (X'X)^{-1} X' \\
 n_p &\text{: Anzahl der Parameter (Basisfunktionen) inklusive Konstante} \\
 \hat{y}_i^{\setminus i} &\text{: Vorhersage von } y_i \text{ mit einem Modell welches ohne } \mathbf{x}_i \\
 &\text{trainiert wurde (siehe auch Kapitel 9.18).}
 \end{aligned} \tag{9.13}$$

Straffunktionen

Alternativ zu den dargestellten Gütekriterien, welche neben den Vorhersageabweichungen an den Trainingsdaten, die Anzahl der Basisfunktionen (Parameter) be-

grenzen, kann alternativ die Summe der Fehlerquadrate plus eine zusätzliche Straffunktion für die Regressionskoeffizienten zur Bestimmung des Polynoms minimiert werden [40, 63, 18].

$$\min_b \left[\sum_{i=1}^{n_r} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=0}^{n_p} b_j^2 \right] = \min_b \left[SSE + \lambda \sum_{j=0}^{n_p} b_j^2 \right] \quad (9.14)$$

Bei steigendem λ streben die Absolutwerte der Regressionskoeffizienten b_j insignifikanter Faktoren schnell gegen Null, wodurch eine automatische Variablenelektion stattfindet. Die erzeugten Metamodelle werden bei steigendem λ einfacher, wodurch die Varianz des Metamodells sinkt, jedoch der Approximationsfehler für die einzelnen Datenpunkte steigt. Die erzeugten Metamodelle werden typischerweise durch Kreuzvalidierung beurteilt, wodurch das geeignete λ und somit Metamodell gewählt werden kann (siehe Kapitel 9.18). Damit die Straffunktion unabhängig von den Dimensionen der einzelnen Faktoren ist, werden die Daten auf eine einheitliche Varianz von eins normiert. Dieses Verfahren ist eine einfache Version der *Ridge Regression* (siehe Kapitel 9.11). Die gesuchten Regressionskoeffizienten können in Abhängigkeit von λ direkt nach Gleichung 9.15 bestimmt werden, wobei I die quadratische Einheitsmatrix ist [109]:

$$\begin{aligned} \hat{b}_\lambda &= (X'X + \lambda I)^{-1} X'Y \\ \hat{Y} &= X\hat{b}_\lambda \end{aligned} \quad (9.15)$$

Im sogenannten **LASSO** Verfahren wird anstelle der quadratischen Straffunktion die Summe der Absolutwerte verwendet $\lambda \sum_{j=0}^{n_p} |b_j|$ [97, 98]. Diese Straffunktion drückt die Regressionskoeffizienten b_j insignifikanter Faktoren deutlich stärker gegen Null, wobei die Koeffizienten im Gegensatz zur Ridge-Regression wirklich Null werden, so dass eine eindeutige Faktorselektion ermöglicht wird. Neben der (einfachen) Ridge Regression und LASSO existieren in der Literatur weitere Ansätze für Straffunktionen und weiterentwickelte Verfahren, auf die hier nicht weiter eingegangen wird [28, 22, 39]. Da einige Metamodellverfahren in eine allgemeine lineare Form (Gleichung 9.16) umgewandelt werden können, sind die dargestellten Ansätze auch bei diesen Modelltypen grundsätzlich einsetzbar [25]. Dabei sind g_j beliebige Basisfunktionen, die jeweils von mehreren Faktoren abhängen können.

$$\hat{y}_i = \sum_{j=0}^{n_p-1} b_j g_j(\mathbf{x}_i) \quad (9.16)$$

Alle bisher dargestellten Verfahren zur Faktorwahl basieren auf einer festen vorgegebenen Anzahl möglicher Basisfunktionen, aus denen die beste Kombination ausgewählt wird. Die Gruppe der möglichen Basisfunktionen wird dabei in den meisten Fällen durch die Anzahl der Faktoren, der maximal zu betrachtende Potenz der Faktoren und den Grad der zu betrachtenden Interaktionen bestimmt. Bei steigender

Anzahl der Faktoren oder der Komplexität der betrachteten Funktionen (z.B. Grad des Polynoms) wird das Verzeichnis der möglichen Basisfunktionen schnell so umfangreich, dass eine effektive Faktorauswahl nur schwer möglich ist. Hier setzt das Verfahren *Adaptive Basis-Funktions Konstruktion* (ABFC, Adaptive Basis Function Construction) an, bei der in jedem Schritt lediglich ein Teil möglicher neuer Basisfunktionen betrachtet wird (siehe Kapitel 9.5). Neben der Reduktion möglicher Basisfunktionen in jedem Selektionsschritt muss zusätzlich keine maximale Potenz für Faktoren beziehungsweise Komplexität des Polynoms vor der Erstellung angegeben werden.

9.4 Robuste Regression

Fehlmessungen beziehungsweise Ausreißer in den Trainingsdaten führen zu fehlerhaften Regressionsmodellen (Abbildung 9.1, blau), die nicht zur sinnvollen Analyse und Optimierung von Systemen einsetzbar sind. Zur Analyse dieser Trainingsdaten wurden daher verschiedene *robuste* Regressionsmethoden entwickelt, welche Ausreißer automatisch detektieren und in der weiteren Analyse sinnvoller berücksichtigen beziehungsweise vernachlässigen können.

In einer einfachen Methode wird im ersten Schritt eine klassischen Regression durchgeführt. Ist der gewählte Modellansatz (z.B. linear oder quadratisch) korrekt und das Rauschen auf den Trainingsdaten normalverteilt, müssten alle Residuen $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}}$, d.h. die Abweichungen zwischen Trainingsdaten und Modellvorhersage, ebenfalls normalverteilt sein, so dass die Standardabweichung der Residuen σ

bestimmt werden kann $\sigma^2 = \frac{\sum_{i=1}^{n_r} r_i}{n_r - n_f - 1}$. Weist das vom Betrag größte Residuum einen Wert $|r| > 3\sigma$ auf, so ist diese Abweichung unwahrscheinlich und der dazugehörige Datenpunkt wird als Ausreißer betrachtet und aus den Trainingsdaten gelöscht. Danach wird der Prozess solange wiederholt bis alle Residuen-Beträge $|r| \leq 3\sigma$ sind. Eine weitaus geschicktere Methode basiert auf einer automatischen Gewichtung jedes einzelnen Trainingspunkts. Ausreißern oder Datenpunkten, die nicht zu dem gewählten Modellansatz passen, wird ein geringeres Gewicht zugeordnet als den restlichen Punkten [20, 41, 43, 93, 61]. Im ersten Schritt wird auch hier mit Hilfe der kleinsten Fehlerquadrate eine erste Abschätzung der Parameter $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ durchgeführt. Anschließend werden die Residuen wie oben beschrieben berechnet und zusätzlich mit einem Korrekturfaktor k_i beaufschlagt [20].

$$\begin{aligned} k_i &= \frac{1}{\sqrt{1 - l_i}} \text{ mit} \\ l_i &= \sum_{j=1}^{n_r} E_{i,j}^2 \text{ und } E = \left[R(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \right]' \end{aligned} \tag{9.17}$$

E ist dabei eine $n_r \times n_r$ Matrix und R die Dreiecksmatrix der QR-Zerlegung (QR-Faktorisierung), so dass gilt $\mathbf{X} = \mathbf{Q}\mathbf{R}$ und $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$. Im zweiten Schritt werden der

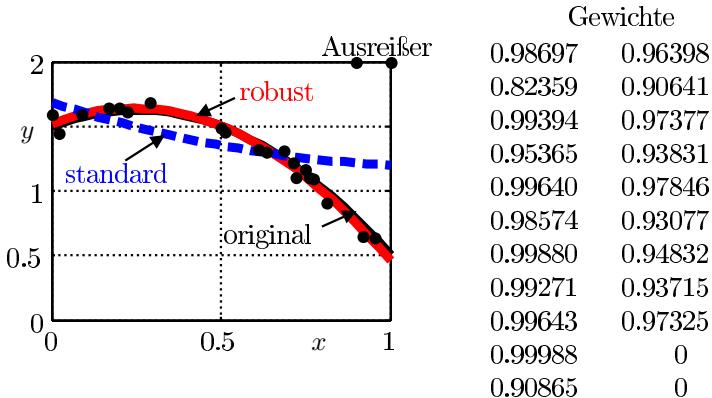


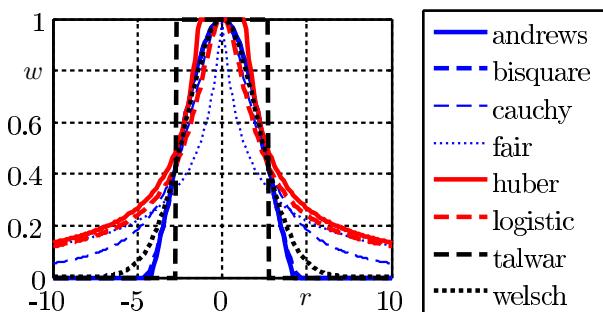
Abb. 9.1 Beispiel zur robusten Regression

Parametervektor $\hat{\mathbf{b}}$ und ein Daten-Gewichtungsvektor $\hat{\mathbf{w}}$ solange iterativ korrigiert, bis diese konvergieren. Dabei wird folgender Algorithmus 1 verwendet.

- 1 $\hat{\mathbf{b}} = (X'X)^{-1} X'y$
- 2 QR-Zerlegung von X: $X = QR$ und $Q'Q = I$
- 3 $k_i = \frac{1}{\sqrt{1-l_i}}$ mit $l_i = \sum_{j=1}^{n_r} E_{i,j}^2$ und $E = [R(X'X)^{-1} X']'$
- 4 **solange die Parameter $\hat{\mathbf{b}}$ nicht konvergiert sind tue**
 - 5 Berechne aktuelle Residuen: $\mathbf{r} = \mathbf{y} - X\hat{\mathbf{b}}$
 - 6 Korrigiere Residuen: $r_{k_i} = r_i k_i$
 - 7 Sortiere Absolutwerte der Residuen: $\mathbf{r}_s = \text{sort}\{|\mathbf{r}_k|\}$
 - 8 Berechne die absolute Median Abweichung ohne die ersten $n_f - 1$ Residuen:
 $s = \frac{1}{0.6745} \text{median}(r_{s_j})$ mit $j = n_f \dots n_r$
(Der Wert 0.6745 macht die Abschätzung unverzerrt zur Normalverteilung)
 - 9 Berechne Gewichte $\hat{\mathbf{w}}$ mit Hilfe einer Gewichtungsfunktion f_w aus Tabelle 9.2
(Abbildung 9.2) sowie dem dazugehörigen Tuningfaktor t [61].
 $w_i = f_w\left(\frac{r_{k_i}}{s \cdot t}\right)$
 - 10 Berechne gewichtete Trainingsdaten: $X_{w_{i,j}} = X_{i,j} \sqrt{w_i}$ und $y_{w_i} = y_i \sqrt{w_i}$
 - 11 Berechne neue Parameter $\hat{\mathbf{b}}$ mit den gewichteten Trainingsdaten $\hat{\mathbf{b}} = (X'_w X_w)^{-1} X'_w \mathbf{y}_w$
- 12 **Ende**

Algorithmus 1 : robuste Regression

Name	Gleichung	Faktor t
andrews	$f_w = \begin{cases} \frac{\sin r}{r} & \text{wenn } r < \pi \\ 0 & \text{wenn } r \geq \pi \end{cases}$	1.339
bisquare	$f_w = \begin{cases} (1 - r^2)^2 & \text{wenn } r < 1 \\ 0 & \text{wenn } r \geq 1 \end{cases}$	4.685
cauchy	$f_w = \frac{1}{1+r^2}$	2.385
fair	$f_w = \frac{1}{1+ r }$	1.4
huber	$f_w = \frac{1}{\max(1, r)}$	1.345
logistic	$f_w = \frac{\tanh(r)}{r}$	1.205
talwar	$f_w = \begin{cases} 1 & \text{wenn } r < 1 \\ 0 & \text{wenn } r \geq 1 \end{cases}$	2.795
welsch	$f_w = e^{-r^2}$	2.985

Tabelle 9.2 Gewichtungsfunktionen für robuste Regression [61]**Abb. 9.2** Gewichtungsfunktionen für robuste Regression

9.5 Adaptive Basis-Funktions Konstruktion

In den letzten Jahren zeigt sich gerade im Bereich von Computerexperimenten ein kontinuierlicher Anstieg der zu berücksichtigen Faktoren und Interaktionen sowie eine Erhöhung der abzubildenden Komplexität zwischen Faktoren X und Systemantworten y . Sollen die Zusammenhänge mittels einer polynomialen Regression abgebildet werden, so steigt die Anzahl der möglichen Basisfunktionen n_p ($x_1, x_1^2, x_1x_2, \dots$) dramatisch mit der Anzahl an Faktoren und dem Polynomgrad an (siehe Kapitel 9.3.1 und 9.3).

$$\hat{y}_i = \sum_{j=0}^{n_p-1} b_j x_{ij} \quad (9.18)$$

Die klassische Faktorauswahl mit sequentieller Vorwärts/Rückwärts Selektion gerät dabei durch die riesige Anzahl zu berücksichtigender und analysierender Basisfunktion-Kombinationen an ihre rechnerischen/zeitlichen Grenzen. Eine Reduktion der zu betrachtenden Faktoren oder der Komplexität des Modells durch Verkleinerung des maximalen Polynomgrads ist meistens nicht erwünscht, da dadurch eventuell ein zu einfaches (underfitted) Modell erzeugt wird, welches die Systemzusammenhän-

ge nicht mehr abbilden kann. Die Adaptive Basis-Funktions Konstruktion (*Adaptive Basis Function Construction, ABFC*) umgeht dieses Problem, indem es nicht auf ein großes im Vorfeld definiertes und festes Verzeichnis von Basisfunktionen und deren Kombinationen zurückgreift, sondern in jedem Iterationsschritt auf Basis des momentanen Modells ein kleines neues Verzeichnis mit möglichen Basisfunktionen generiert [46, 47]. Ein positiver Effekt ist, dass die im Vorfeld häufig schwierige Definition des maximalen Polynomgrads oder der zu berücksichtigenden Interaktionen entfällt, da diese automatisch ermittelt werden. Aufbauend auf ein einfaches Modell, dass typischerweise wie beim SFS (Kapitel 9.3.1) nur einen konstanten Term enthält, werden durch (a) Hinzufügen von einfachen linearen Basisfunktionen, (b) Löschen von Basisfunktionen und (c) Manipulation von vorhandenen Basisfunktionen, neue mögliche Kombinationen von Basisfunktionen (Modellstrukturen) erzeugt. Beim ABFC werden vier Basisoperationen verwendet [46, 47]:

1. Hinzufügen einer linearen Basisfunktion mit nur einem linearen Faktor x_i
2. Kopie einer vorhandenen Basisfunktion und Erhöhung des Exponenten eines Faktors um eins
3. Reduzierung eines Exponenten um eins bei einer bereits existierenden Basisfunktion
4. Löschen einer Basisfunktion

Abbildung 9.3 zeigt an einem einfachen Beispiel mit drei Faktoren die Änderungen des Modells, die sich durch die vier Operationen ergeben können [47]. Als aktuell gegebene Modellstruktur wird $\hat{y} = b_0 + b_1x_1 + b_2x_2x_3^3$ angenommen, welche durch die folgende Matrix dargestellt wird.

$$\hat{y} = b_0 + b_1x_1 + b_2x_2x_3^3 \rightarrow \begin{vmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 3 \end{vmatrix} \quad (9.19)$$

Wenn die dargestellten Operationen im klassischen SFS Algorithmus anstelle der Auswahl aus einem vorgegebenen Verzeichnis zur Faktorauswahl (Kapitel 9.3.1) eingesetzt werden, so wird von *Floating Adaptive Basis Function Construction (F-ABFC)* gesprochen. Zur Beurteilung der erzeugten Modellstrukturen wird zum Beispiel das Gütekriterium AICc (siehe Kapitel 9.3.1) verwendet [46, 47]. Besonders bei wenigen Faktoren ($n_f < 4$) wächst die Anzahl der möglichen Basisfunktionen nur sehr langsam. Bei Konstruktion eines komplexen Modells kann der Algorithmus dadurch in einem lokalen Minimum stecken bleiben und ein zu einfaches Modell erzeugen. Um dieses zu verhindern ist es möglich jede ausgewählte Operation rekursiv mehrmals einzusetzen, wobei jedoch nur die gerade verwendete Operation wiederholt wird und nicht mit anderen Operationen kombiniert wird. Im Fall, dass der zweite Operator mit dem Term [010] verwendet wird, würden durch die ersten zwei Iterationen die in Abbildung 9.4 dargestellten Basisfunktionen erzeugt werden. Doppelte Einträge werden dabei nur einmal berücksichtigt [47].

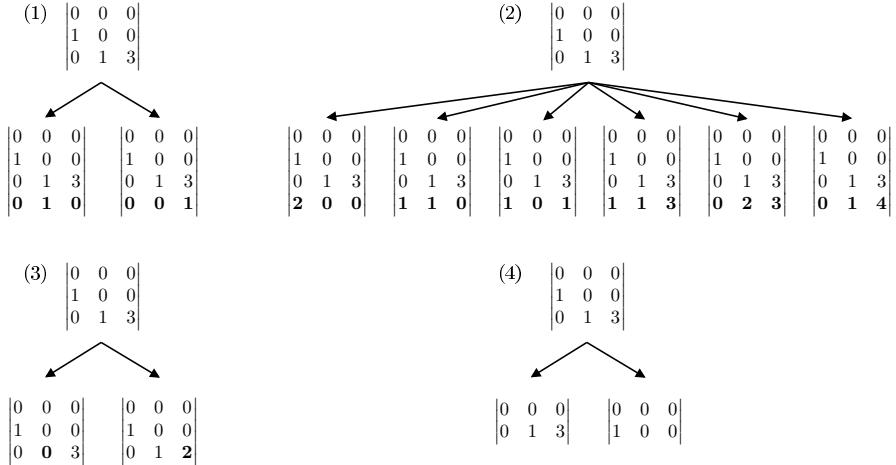


Abb. 9.3 Operationen für die Adaptive Basis-Funktion Konstruktion (Beispiel)

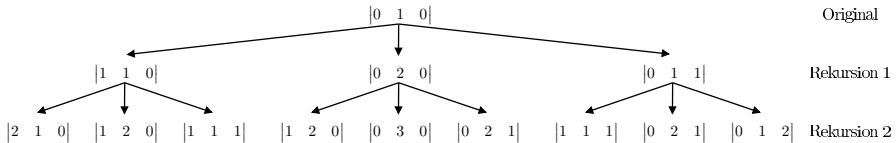


Abb. 9.4 Rekursive Erstellung von zusätzlichen Basisfunktionen (Beispiel)

9.6 Kernel- und Lokale Polynom-Regression

Ziel einer Kernel-Regression ist die Abschätzung eines Funktionswerts y_0 an einer Faktorkombination \mathbf{x}_0 auf Basis gewichteter Messdaten. Die häufigsten Einsatzgebiete dieser Methode weisen eine geringe Faktoranzahl $n_f \leq 2$ auf. Die grundlegende Form der Approximationsgleichung lautet dabei:

$$\hat{y}_0 = \sum_{i=1}^{n_r} b_i(\mathbf{x}_0, \mathbf{x}_i) y_i \quad (9.20)$$

Jedem der n_r Datenpunkte wird ein Gewicht b zugeordnet, welches nicht nur von den gegebenen Datenpunkten selbst, sondern zusätzlich von der Faktoreinstellung des gesuchten Punkts \mathbf{x}_0 abhängt. NARDARAYA [65, 66] und WATSON [106] schlagen zur Bestimmung der Gewichte \mathbf{b} im eindimensionalen Fall folgende Berechnungsvorschrift vor:

$$b_i(\mathbf{x}_0, \mathbf{x}_i) = \frac{K\left(\frac{x_i - x_0}{h}\right)}{\sum_{k=1}^{n_r} K\left(\frac{x_k - x_0}{h}\right)} \quad (9.21)$$

Die Funktion $K(u) \geq 0$ ist dabei eine symmetrische Funktion, die mit steigendem Absolutwert von u abnimmt und als *Kernel-Funktion* bezeichnet wird. Die Band-

breite h ist dabei ein Faktor, welcher die Glättung der Approximationsfunktion beeinflusst. Als Kernel $K(u)$ werden verschiedene Funktionen eingesetzt, wobei die meisten nur im Bereich $|u| \leq 1$ definiert sind. Im übrigen Bereich ist der Funktionswert Null (Abbildung 9.5).

$$\begin{aligned}
 \text{Konstant} \quad K(u) &= 0.5 \quad |u| \leq 1 \\
 \text{Dreieck} \quad K(u) &= \max(0, 1 - |u|) \quad |u| \leq 1 \\
 \text{Epanechnikov} \quad K(u) &= \max\left(0, \frac{3}{4}(1 - u^2)\right) \quad |u| \leq 1 \\
 \text{Quadratisch} \quad K(u) &= \max\left(0, \frac{15}{16}(1 - u^2)^2\right) \quad |u| \leq 1 \\
 \text{Kubisch} \quad K(u) &= \max\left(0, \frac{35}{32}(1 - u^2)^3\right) \quad |u| \leq 1 \\
 \text{Kosinus} \quad K(u) &= \max\left(0, \frac{\pi}{4} \cos\left(\frac{\pi}{2}u\right)\right) \quad |u| \leq 1 \\
 \text{Gauß} \quad K(u) &= \frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}}
 \end{aligned} \tag{9.22}$$

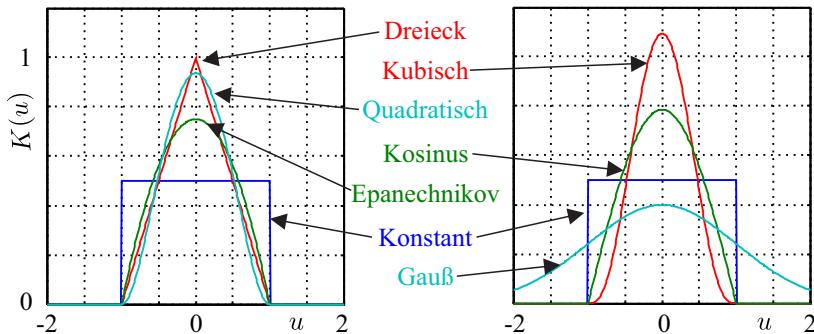


Abb. 9.5 Kernel-Funktionen

Für alle Kernel-Funktionen gilt:

$$\int_{-\infty}^{\infty} K(u) du = 1 \text{ und } K(-u) = K(u) \tag{9.23}$$

Die Wahl der Kernel-Funktion K hat im Vergleich zur Wahl der Bandbreite h nur einen geringen Einfluss auf die Qualität des erzeugten Modells [58, 35]. Wird h sehr klein gewählt, so erhalten lediglich die Messdaten in der lokalen Umgebung des gesuchten Punkts \mathbf{x}_0 ein deutliches Gewicht für die Approximation des gesuchten Funktionswerts. Das erzeugte Metamodell ist in diesen Fällen meist zu stark an die gegebenen Messdaten angepasst (*overfitting*). Die Approximation der bekannten Messdaten ist dann zwar sehr genau jedoch können hohe Abweichungen bei Faktorkombinationen auftreten, die nicht zum Training des Metamodells verwendet wurden. Große Werte für h führen auf der anderen Seite zu einer starken Glättung des Funktionszusammenhangs, so dass die Vorhersage des Funktionswerts immer

weiter auf den globalen Mittelwert \bar{y} der Trainingsdaten zuläuft.

Ein grundsätzliches Problem bei der Approximation tritt an den Rändern des Faktorraums auf, da dort die Datenpunkte nicht mehr symmetrisch um den gesuchten Punkt x_0 verteilt sind. Wird beispielsweise eine eindimensionale Funktion f betrachtet, die am Rand des Faktorraums kontinuierlich fallend ist, so stehen zur Approximation lediglich Datenpunkte mit Funktionswerten größer als y_0 zur Verfügung. Die Vorhersage \hat{y}_0 wird dadurch größer sein als der wahre Funktionswert y_0 . Typischerweise steigt daher der Fehler zum Rand hin an (Abbildung 9.6).

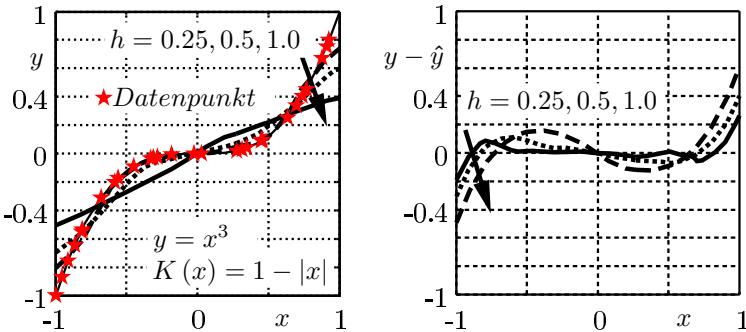


Abb. 9.6 Kernel-Regression, Beispiel: $y = x^3$

Die Kernel-Regression wird immer dann sinnvoll eingesetzt, wenn eine hohe Anzahl von Trainingsdaten vorliegt, welche mit einer Messstreuung überlagert sind. Eine Anhäufung von Datenpunkten in bestimmten Regionen des Faktorraums (Clustering) führt automatisch zu einer stärkeren Gewichtung dieser Regionen, so dass eine gute Gleichverteilung der Datenpunkte vorteilhaft ist.

Zur Erhöhung der Approximationsgenauigkeit kann an der Stelle x_0 nicht nur ein gewichteter Mittelwert, sondern ebenfalls ein Polynom mit dem Grad $d \neq 0$ bestimmt werden (*Lokale Polynom-Regression*) [14, 36], wodurch die Approximationsfehler an den Rändern des Faktorraums sinken. Im eindimensionalen Fall (1 Faktor) wird dadurch eine Funktion der folgenden Form gesucht:

$$\begin{aligned}\hat{y}(x_0) &= b_0(x_0) + b_1(x_0)x_0 + b_2(x_0)x_0^2 + \cdots + b_d(x_0)x_0^d & (9.24) \\ &= \sum_{k=0}^d b_k(x_0)x_0^k = \mathbf{x}_0 \mathbf{b} \\ \text{mit } \mathbf{x}_0 &= [1, x_0, x_0^2, \dots, x_0^d] \text{ und } \mathbf{b} = [b_0, b_1, \dots, b_d]'\end{aligned}$$

Mit der Definition der Trainingsmatrix X und Gewichtungsmatrix W aus Gleichungen 9.25 und 9.26 werden die Regressionskoeffizienten \mathbf{b} entsprechend Gleichung 9.27 bestimmt (siehe auch Gleichung 9.6). Auf die Normierung der Gewichtungsmatrix W mittels der summierten Kernelfunktionen kann grundsätzlich auch

verzichtet werden.

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^d \\ 1 & x_2 & x_2^2 & \cdots & x_2^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n_r} & x_{n_r}^2 & \cdots & x_{n_r}^d \end{pmatrix} \quad (9.25)$$

$$W = \frac{\operatorname{diag} \left[K \left(\frac{x_0 - x_1}{h} \right), K \left(\frac{x_0 - x_2}{h} \right), \dots, K \left(\frac{x_0 - x_{n_r}}{h} \right) \right]}{\sum_{i=1}^{n_r} K \left(\frac{x_0 - x_i}{h} \right)} \quad (9.26)$$

$$\hat{\mathbf{b}} = (X'WX)^{-1} X'W\mathbf{y} \quad (9.27)$$

Abbildung 9.7 zeigt deutlich die erreichte Verbesserung des Metamodells bei Verwendung einer lokalen Polynom-Regression mit dem Grad $d = 2$. In der Praxis zeigt sich, dass Polynome des Grads $d \leq 2$ meistens für eine gute Approximation des Randbereichs ausreichen.

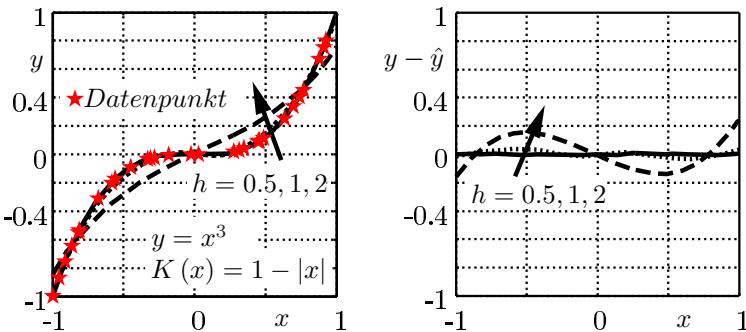


Abb. 9.7 Lokale Polynom-Regression (Polynom Grad 2)

Bei mehreren Faktoren ($n_f > 1$) wird überwiegend ein linearer Ansatz mit $d = 1$ eingesetzt. Für jeden Faktor kann dabei ebenfalls eine unterschiedliche Bandbreite h_j gewählt werden, was den Aufwand zur optimalen Wahl der Bandbreiten erschwert. Die mehrdimensionale Kernel-Funktion muss weiterhin die folgende Grundvoraussetzung erfüllen.

$$\int_{-\infty}^{\infty} K(u) du = 1 \quad (9.28)$$

Die einfachste Möglichkeit zur Erzeugung einer mehrdimensionalen Kernel-Funktion ist die Multiplikation eindimensionaler Kernel-Funktionen, die für jede Dimension (Faktor) separat berechnet werden.

$$K\left(\frac{x_{i1} - x_{01}}{h_1}, \frac{x_{i2} - x_{02}}{h_2}, \dots, \frac{x_{in_f} - x_{0n_f}}{h_{n_f}}\right) = K\left(\frac{x_{i1} - x_{01}}{h_1}\right) K\left(\frac{x_{i2} - x_{02}}{h_2}\right) \dots K\left(\frac{x_{in_f} - x_{0n_f}}{h_{n_f}}\right) \quad (9.29)$$

Zur Approximation von y_0 an der Stelle \mathbf{x}_0 wird entsprechend dem eindimensionalen Fall folgendes Gleichungssystem definiert:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n_f} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n_f} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n_r1} & x_{n_r2} & \cdots & x_{n_r n_f} \end{pmatrix} \quad (9.30)$$

$$W = \text{diag} \left[K\left(\frac{x_{01} - x_{11}}{h_1}\right) K\left(\frac{x_{02} - x_{12}}{h_2}\right) \dots K\left(\frac{x_{0n_f} - x_{1n_f}}{h_{n_f}}\right), \right. \quad (9.31)$$

$$K\left(\frac{x_{01} - x_{21}}{h_1}\right) K\left(\frac{x_{02} - x_{22}}{h_2}\right) \dots K\left(\frac{x_{0n_f} - x_{2n_f}}{h_{n_f}}\right),$$

...

$$\left. K\left(\frac{x_{01} - x_{n_r1}}{h_1}\right) K\left(\frac{x_{02} - x_{n_r2}}{h_2}\right) \dots K\left(\frac{x_{0n_f} - x_{n_r n_f}}{h_{n_f}}\right) \right] \quad (9.32)$$

$$\hat{\mathbf{b}} = (X'WX)^{-1}X'W\mathbf{y} \quad (9.33)$$

Abbildung 9.8 zeigt die Approximation des zwei-dimensionalen Beispiels aus Gleichung 9.156 bei Verwendung unterschiedlicher Bandbreiten h (gleiche Bandbreite für alle Dimensionen). Durch eine steigende Bandbreite nähert sich das Metamodell immer weiter der einfachen linearen Regression, was zu deutlich höheren Approximationsfehlern führt. Kleinere Bandbreiten liefern genauere Vorhersagen im inneren Bereich, weisen aber bei Extrapolation höhere Fehler auf. Soll zur weiteren Verbesserung der Approximationsgenauigkeit ein komplexeres Polynom verwendet werden, so fließt dieses in die Bestimmung von X und \mathbf{x}_0 ein [$\mathbf{x}_0 = (1, x_{01}, x_{02}, x_{01}^2, x_{02}^2)$]. In verschiedenen Literaturstellen wird die Wahl der Bandbreite diskutiert [6, 49, 83, 105, 21, 51].

Eine interessante Anwendung der lokalen Polynom-Regression bietet die Kombination mit einem Kriging Verfahren (Kapitel 9.13). Dabei wird der globale Trend des zu approximierenden Zusammenhangs durch die lokale Polynom-Regression modelliert und durch ein Kriging Modell die verbleibende Abweichung zwischen dem globalen Trend und den bekannten Datenpunkten [42].

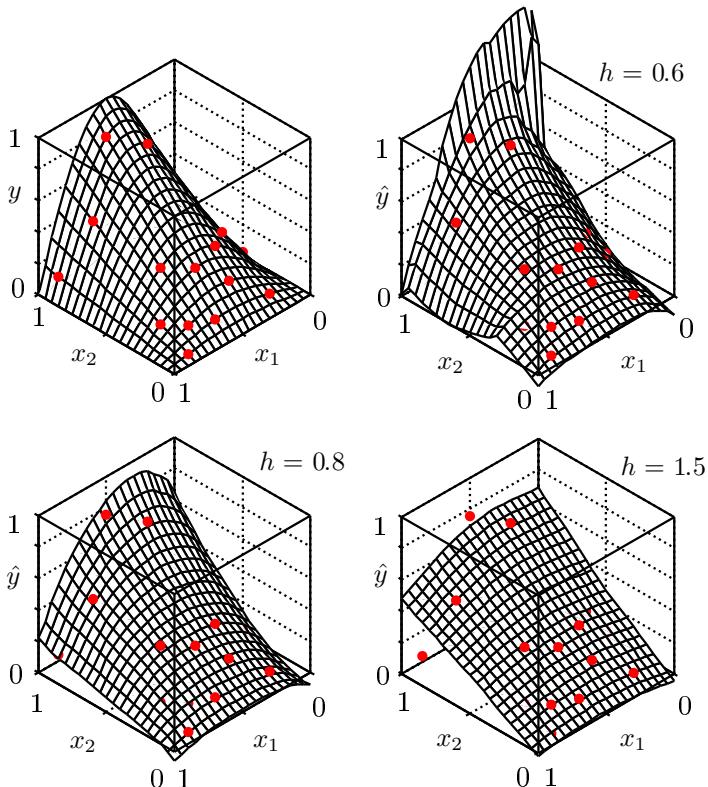


Abb. 9.8 Lokale Polynom-Regression: Beispiel

9.7 Regressionsbaum

Ein Regressionsbaum ist eine Erweiterung des Klassifikationsbaums, wobei in diesem Abschnitt nicht näher auf Klassifikationen eingegangen wird [9, 59]. Grundsätzlich zerlegen Regressionsbäume gegebenen Trainingsdaten (X, y) sukzessiv in Untergruppen und approximieren den Funktionswert y einer Untergruppe mit dem Mittelwert der enthaltenen Daten. Die Vorhersage des in Abbildung 9.9 dargestell-

ten Regressionsbaums wird durch Algorithmus 2 beschrieben.

```

1 wenn  $x_2 < a$  dann
2   |   wenn  $x_1 < b$  dann
3   |   |   y = 1.9
4   |   sonst
5   |   |   y = 0.1
6   |   Ende
7 sonst
8   |   y = 1.0
9 Ende

```

Algorithmus 2 : Beispiel eines einfachen Regressionsbaums

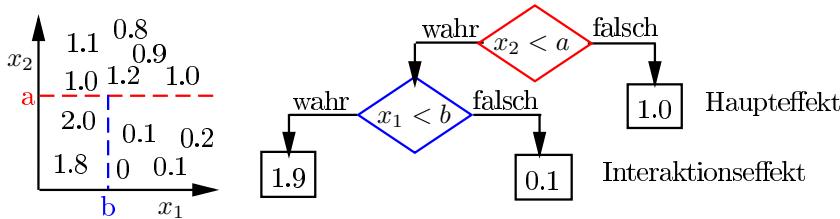


Abb. 9.9 Beispiel eines Regressionsbaums

Der Approximationsfehler $\text{Err}(B)$ eines Regressionsbaums B wird durch den Mittelwert der Abweichungen aller Trainingsdaten zum Mittelwert des zugehörigen Blatts berechnet (Gleichung 9.34). Ein Blatt ist jeweils eine Datengruppe, welche nicht weiter unterteilt wurde und sich somit am unterem Ende der Aststruktur befindet [59, 9].

$$\text{Err}(B) = \frac{1}{n_r} \sum_{b=1}^{n_b} \sum_{i=1}^{n_{rb}} (y_{bi} - \bar{y}_b)^2 \quad (9.34)$$

Ein Blatt b das n_b Blätter aufweist, enthält dabei n_{rb} Datenelemente y_{bi} und weist einen Mittelwert von \bar{y}_b auf. Der Vorhersagefehler einer Baumstruktur wird durch Aufteilen eines Blatts b in zwei neue Unterblätter (l : links und r : rechts) genau dann minimiert, wenn die Summe der Quadratfehler beider Subblätter minimal wird.

$$\min \left\{ \sum_{i=1}^{n_{rl}} (y_{li} - \bar{y}_l)^2 + \sum_{i=1}^{n_{rr}} (y_{ri} - \bar{y}_r)^2 \right\} \quad (9.35)$$

Soll bei der Suche nach der besten Schnittposition eines Blatts nicht jedes mal der Mittelwert jeder möglichen neuen Untergruppe berechnet werden, ist es ebenfalls möglich Gleichung 9.36 zu maximieren, die lediglich auf dem unveränderten Mittelwert des zu teilenden Gesamtknotens \bar{y} basiert.

$$\max \left\{ \frac{\left(\sum_{i=1}^{n_r} y_{l,i} - \bar{y} \right)^2}{n_{r_l}} + \frac{\left(\sum_{i=1}^{n_r} y_{r,i} - \bar{y} \right)^2}{n_{r_r}} \right\} \quad (9.36)$$

Der Suchalgorithmus zur Bestimmung der besten Schnittposition eines Blatts b , welche aus der besten Schnittvariable j und dem besten Schnittwert x^* besteht, ist in Algorithmus 3 dargestellt.

- 1 für alle Dimensionen j der aktuellen Blatt-Datenmenge X tue**
 - 2 Sortiere Daten aufsteigend nach den Faktorwerten \mathbf{x}_j der aktuellen Dimension j**
 - 3 Finde mögliche Schnittpositionen i in der aktuell sortierten Dimension j für die gilt**
 - $x_{i,j} < x_{i+1,j}$
 - Der Schnittpunkt $x_{i,j}^*$ für die Position i in Dimension j liegt genau zwischen den Datenpunkten: $x_{i,j}^* = \frac{x_{i,j} + x_{i+1,j}}{2}$
 - 4 für alle gefundenen Schnittpunkte x^* der Dimension j tue**
 - Berechne Mittelwerte der durch x^* aufgeteilten Untermengen

$$\bar{y}_l = \frac{1}{n_l} \sum_{x_{i,j} < x_{i,j}^*} y_i$$

$$\bar{y}_r = \frac{1}{n_r} \sum_{x_{i,j} \geq x_{i,j}^*} y_i$$
 - Berechne das Qualitätskriterium Q für den aktuellen Schnittpunkt $x_{i,j}^*$ in der Dimension j

$$Q_{x_{i,j}^*} = \sum_{x_{i,j} < x_{i,j}^*} (y_i - \bar{y}_l)^2 + \sum_{x_{i,j} \geq x_{i,j}^*} (y_i - \bar{y}_r)^2 \text{ (siehe auch Gleichung 9.35)}$$
 - 7 Ende**
 - 8 Ende**
 - Wähle Schnittpunkt $x_{i,j}^*$, der zum besten Qualitätskriterium Q führt
 - Erzeuge Untermengen X_L und X_R mit gewähltem Schnittpunkt und führe den Algorithmus jeweils mit den Untermengen durch, solange kein Stopp-Kriterium erfüllt ist
- Algorithmus 3 :** Teilen einer Blatt-Datenmenge

Die Konstruktion eines Regressionsbaums mit der Blattgröße eins für alle Blätter besitzt zwar den geringsten $Err(B) = 0$ Wert, jedoch zeigen diese Bäume die Tendenz zum Overfitting (siehe Kapitel 9.16). In vielen Fällen ist es sinnvoller einen kleineren Baum zu erzeugen, wobei entweder die Erstellung des Baums frühzeitig gestoppt wird oder ein umfangreicher Baum nachträglich beschnitten (*pruning*) wird. Zwei übliche Stoppkriterien bei der Erzeugung eines Regressionsbaums sind in diesem Zusammenhang verbreitet:

1. Die Mindestanzahl γ von Daten in einem Astknoten oder Blatt, also die Größe einer Gruppe, die noch geteilt werden darf beziehungsweise die minimale Datenanzahl eines Blatts.
2. Maximale Verunreinigung β eines Knotens (*Impurity*)

$$I = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 < \beta$$

Kleine Werte für γ und β führen zu großen Bäumen, die zum Overfitting neigen und große Werte bergen das Risiko signifikante Effekte nicht zu berücksichtigen. Typische Werte sind $\gamma = 5 \dots 10$ und $\beta = 1e^{-6}I_0$, wobei I_0 die Verunreinigung (*Impurity*) der ungeteilten gesamten Datenmenge beschreibt.

Abbildung 9.10 zeigt zwei Approximationen einer Datenmenge X bei Verwendung unterschiedlicher Grenzen für die teilbare Blattgröße γ . Größere γ -Werte führen dabei zu einer stärkeren Glättung.

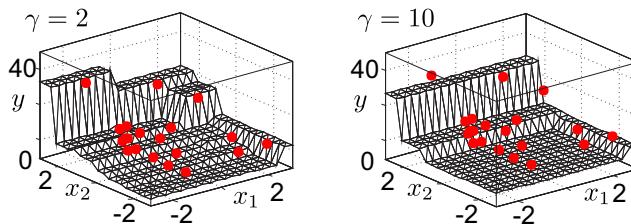


Abb. 9.10 Approximation zweier Regressionsbäume mit unterschiedlicher minimaler teilbaren Blattgröße

Die optimalen Größen für γ und β sind im Vorfeld nicht immer bekannt, so dass im ersten Schritt meist ein großer Baum B erzeugt (kleine Werte für γ und β) wird, welcher im zweiten Schritt an einem zu bestimmenden Knoten k beschnitten (*pruned*) wird. Durch die Beschneidung wird ein gesamter Ast B_k vom Baum B entfernt und ein verkleinerter (optimierter) Baum B ohne B_k entsteht ($B \setminus B_k$).

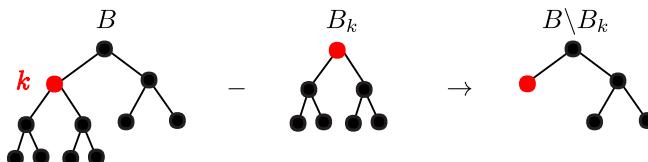


Abb. 9.11 Beschnittener Regressionsbaum

Zur Bestimmung der momentan besten Schnittposition können je nach Anzahl der vorhandenen Daten zwei unterschiedliche Verfahren eingesetzt werden. Bei einer großen Datenanzahl ist es möglich einen kleinen zufällig ausgewählten Anteil der Daten (n_v Elemente) zur Validierung einzusetzen. Die Validierungsdaten werden dabei nicht zum Training des Baums verwendet, wodurch die Vorhersagegenauigkeit dieser Punkte als Gütekriterium verwendet werden kann. Die Güte jedes beschnittenen Baums wird durch die Quadratsumme der Abweichungen der Validierungspunkte (SSE: Error Sum of Squares) bestimmt. Der Teilbaum mit dem kleinsten SSE-Wert wird als bester Teilbaum ausgewählt.

$$SSE_v = \sum_{i=1}^{n_v} (\widehat{y}_{v_i} - y_{v_i})^2 \quad (9.37)$$

Sind nur wenige Daten vorhanden, ist es sinnvoller alle Daten zum Training des Baums einzusetzen, so dass keine separaten Validierungsdaten vorhanden sind. In diesem Fall wird ein Kompromiss zwischen der Komplexität des Baums und der Vorhersagegenauigkeit der Trainingsdaten gesucht. Die Komplexität eines Baums wird hierzu durch die Anzahl der Blätter des Baums beschrieben, so dass die folgende Kenngröße E minimiert werden muss [9]:

$$E(B \setminus B_k) = Err(B \setminus B_k) + \alpha |\text{Blätter}(B \setminus B_k)| \quad (9.38)$$

Der Parameter $\alpha > 0$ ist dabei ein Maß für die gewünschte Komplexität des Baums und wird mit der Anzahl der Blätter des geschnittenen Baums multipliziert. Eine Vergrößerung von α führt zu einem kleineren Baum mit weniger Blättern [59, 9]. Bevor ein Baum geschnitten wird, werden zuerst alle Blätter (l, r) entfernt, welche die Verunreinigung (I : *impurity*) im Vergleich zum ungeteilten Elternknoten (e) nicht verändern.

$$I_e = I_l + I_r \quad (9.39)$$

Danach wird der Baum in einem rekursiven Prozess verkleinert, wobei in jedem Schritt der schwächste Astknoten gesucht und entfernt wird. Die Stärke eines Astes $g(B_k)$ wird durch die Verunreinigung des Schnittknotens k und des abgeschnittenen Astes, welche durch die Summe der Verunreinigungen der zugehörigen Blätter berechnet wird, sowie der Anzahl der Blätter des abgeschnittenen Astes bestimmt [59].

$$g(B_k) = \frac{I(k) - I(B_k)}{|\text{Blätter}(B_k)| - 1} \quad (9.40)$$

Sollten verschiedene Schnittknoten die kleinste und gleiche Stärke g aufweisen, wird der Ast mit den wenigsten Knoten entfernt. Durch den iterativen Prozess werden verschiedene Teilbäume mit unterschiedlicher Vorhersagequalität erzeugt. Der finale Teilbaum wird im Anschluss durch eine Kreuzvalidierung (CV: *cross validation*) ausgewählt.

Die Struktur eines Regressionsbaums ist nicht robust, wodurch kleine Änderungen in den Trainingsdaten bereits starke Änderungen in der Baumstruktur hervorrufen können. Trotz alledem ist die eigentliche Vorhersagegenauigkeit eines Baums robust, da die Vorhersage lediglich durch eine andere Struktur abgebildet wird. Aus diesem Grund kann durch eine Kreuzvalidierung nicht die Struktur des Regressionsbaums optimiert werden, sondern lediglich die Komplexität des Baums. Die Auswahl der optimalen Komplexität wird durch Algorithmus 4 ermöglicht, welcher aus folgenden grundlegenden Teilschritten besteht [59].

- Kompletten Baum B aus allen Trainingsdaten erzeugen und Äste entfernen, welche die Impurity nicht verändern.

- Rekursiv alle möglichen geschnittenen Bäume B_p erzeugen und die Stärke des jeweils abgeschnittenen Astes g_p merken, wobei g_p durch weiteres Beschneiden des Baums kontinuierlich ansteigt.
- Jedem geschnittenen Baum B_p eine mittlere Aststärke g_p^m zwischen g_p und g_{p+1} zuordnen.
- Zur Kreuzvalidierung die Trainingsdaten in n_{CV} Teilgruppen aufteilen.
 - Für jede Teilgruppe i einen CV-Baum mit allen zugehörigen geschnittenen Bäumen auf Basis der Trainingsdaten ohne die Teilgruppen i erzeugen.
 - Für jeden g_p^m -Wert, den geschnittenen CV-Baum suchen der gerade noch einen Wert $g_p \leq g_p^m$ aufweist.
 - Mit diesen CV-Bäumen jeweils die Vorhersage der momentanen Teilgruppe i durchführen, so dass für jede Komplexität g_p^m eine Vorhersage der Teilgruppe i mit einem CV-Baum vorliegt.
- Im Anschluss sind alle Trainingsdaten einmal für jede Komplexität g_p^m durch CV-Bäume vorhergesagt und es kann durch die mittlere Abweichung der Vorhersagen, die optimale Komplexität g_p^m gewählt werden.

- 1 Berechne kompletten Baum B aus allen Daten und vorgegebenen Randbedingungen für minimale Blattgröße und Verunreinigung
 - 2 Entferne alle Blätter, welche die Verunreinigung (*Impurity*) nicht verändern
 $I_{Knoten} = I_{links} + I_{rechts}$
 - 3 Rekursive Erzeugung von n_p geschnittenen (*pruned*) Bäume B_p inklusive der zugehörigen Astfehler g_p , welche durch die kleiner werdenden geschnittenen Bäume kontinuierlich größer werden ($g_i \leq g_{i+1}$)
 - 4 Erzeuge mittlere Astfehler mit $g_k^m = \sqrt{g_k g_{k+1}}$ für $k = 1 \dots n_p - 1$. Der letzte geschnittene Baum B_{n_p} erhält den Wert $g_{n_p}^m = \inf$
 - 5 Teile die Trainingsdaten X, y in n_{CV} Teilgruppen CV_i
 - 6 **für alle** Teilgruppen CV_i mit $i = 1 \dots n_{CV}$ **tue**
 - 7 Teile Datenmenge in aktuelle Trainings- und Validierungsdaten, wobei die Validierungsdaten jeweils der aktuellen Teilgruppe CV_i entsprechen und alle anderen Daten aus X, y zum Training verwendet werden (X_t, y_t, X_v, y_v)
 - 8 Erzeuge Regressionsbaum B_{CV} , der nur auf den aktuellen Trainingsdaten beruht
 - 9 Erzeuge geschnittene Bäume B_{CV_q} aus B_{CV} und zugehörige Astfehler g_{CV_i} **für alle** g_k^m mit $k = 1 \dots n_p$ **tue**
 - 10 Finde ein g_{CV_i} , so dass gilt: $g_{CV_i} \leq g_k^m < g_{CV_{i+1}}$
 - 11 Vorhersage der aktuellen Validierungsdaten \hat{y}_v und Speichern der Werte als Teil-Ergebnis $V_{v_i, k}$ für die mittlere Stufe g_k^m , (v_i bezeichnet die Positionen der aktuellen Teilgruppe i in den Trainingsdaten X und y)
 - 12 **Ende**
 - 13 **Ende**
 - 14 Alle Validierungspunkte sind nun für alle g_k^m Werte mit passenden geschnittenen CV-Bäumen B_{CV_q} vorhergesagt worden, so dass eine Vorhersage-Matrix der Größe $V_{n_r \times n_p}$ existiert.
 - 15 **für alle** g_k^m mit $k = 1 \dots n_p$ **tue**
 - 16 Berechne den quadratischen Fehler jedes Datenpunkts für aktuellen g_k^m -Wert:
 $e_{i,k} = (V_{i,k} - y_i)^2$
 - 17 Berechne den mittlerer quadratischen Fehler: $\bar{e}_k = \frac{\sum_{i=1}^{n_r} e_{i,k}}{n_r}$
 - 18 Berechne die Varianz des quadratischen Fehlers: $var_k = \frac{\sum_{i=1}^{n_r} (e_{i,k} - \bar{e}_k)^2}{n_r}$
 - 19 Abschätzung des Standardfehlers: $\sigma_k = \sqrt{\frac{var_k}{n_r}}$
 - 20 **Ende**
 - 21 Wähle minimalen mittleren quadratische Fehler $\min_k \{\bar{e}_k\}$ an Position k^* und berechne Grenzwert: $e_g = \bar{e}_{k^*} + \sigma_{k^*}$
 - 22 Finde g_k^m -Wert an Position k^{opt} für den der mittlere quadratische Fehler noch gerade kleiner ist als der Grenzwert e_g : $\max_k \{\bar{e}_k\}$ mit $\bar{e}_{k^{opt}} < e_g$
 - 23 Wähle den geschnittenen Originalbaum $B_{k^{opt}}$ mit $g_{k^{opt}}^m$ als optimalen Regressionsbaum
- Algorithmus 4 :** Wahl eines Regressionsbaums durch Kreuzvalidierung

9.8 Splines

Zur Approximation von Zusammenhängen zwischen Ein- und Ausgangsvariablen werden verschiedene Arten von Splines eingesetzt. Bekannte Verfahren sind zum Beispiel B-Splines oder kubische Splines [5, 70, 104, 24, 25, 92]. Grundsätzlich wird bei diesem Verfahren der Faktorraum durch sogenannte Knoten (Stützstellen)

in Unterbereiche aufgeteilt und die Zusammenhänge durch unterschiedliche Polynome in jedem Teilbereich approximiert. Zur Veranschaulichung wird die Konstruktion eines Splines für *eine* Eingangsvariable (Faktor) basierend auf der sogenannten *Power Basis* betrachtet. Im ersten Schritt werden l verschiedene Knotenpunkte K (Stützpunkte) im ein-dimensionalen Faktorraum festgelegt $\{K_1, K_2, \dots, K_l\}$. Dieser wird hier durch eine äquidistante Aufteilung des Faktorbereichs definiert. Ein Spline der Ordnung $p > 1$, wobei p eine ganze Zahl ist, wird dann allgemein mittels der *Power Basis* aus Gleichung 9.41 beschrieben [25]:

$$s(x) = b_0 + b_1 x + b_2 x^2 + \dots + b_p x^p + \sum_{k=1}^l b_{p+k} (x - K_k)_+^p \quad (9.41)$$

mit $(a)_+ = \max(0, a)$

$s(x)$ ist ein Polynom p^{ten} Grades zwischen jedem Knotenpunkt und hat im gesamten Definitionsbereich $p - 1$ stetige Ableitungen [92, 25]. Auf eine Darstellung weiterer in der Praxis gebräuchlicher Verfahren wird hier verzichtet.

Werden mehrere Faktoren betrachtet, so kann die benötigte mehrdimensionale Spline-Basis durch ein Tensorprodukt erzeugt werden. Zur Vereinfachung werden die Eingangsvariablen x_1, \dots, x_{n_f} dabei so normiert, dass sie den gleichen Definitionsbereich (zum Beispiel $0 \dots 1$ oder $-1 \dots 1$) aufweisen. Weiterhin können zur weiteren Vereinfachung gleiche Knotenpositionen für alle Dimensionen festgelegt werden $\{K_1, K_2, \dots, K_l\}$, was zu den folgenden Grundfunktionen für jeweils einen Faktor x_j führt:

$$\begin{aligned} s_{0,j}(x_j) &= 1, \quad s_{1,j}(x_j) = x_j, \quad s_{2,j}(x_j) = x_j^2, \dots, \quad s_{p,j}(x_j) = x_j^p \\ s_{p+1,j}(x_j) &= (x_j - K_1)_+^p, \dots, \quad s_{p+l,j}(x_j) = (x_j - K_l)_+^p \end{aligned} \quad (9.42)$$

Die mehrdimensionalen Basisfunktionen lassen sich nun durch Multiplikation jeweils eines Terms $s_{0,j}, \dots, s_{p+l,j}$ für jede Faktordimension j erzeugen:

$$B_{r_1, \dots, r_{n_f}}(\mathbf{x}) = \prod_{j=1}^{n_f} s_{r_j,j}(x_j) \quad (9.43)$$

Zwei einfache Beispiele für eine Basisfunktionen mit $n_f = 3$, $l = 5$ und $p = 2$ sind in Gleichung 9.44 gegeben:

$$\begin{aligned} B_{0,0,0} &= s_{0,1}(x_1) \cdot s_{0,2}(x_2) \cdot s_{0,3}(x_3) = 1 \cdot 1 \cdot 1 = 1 \\ B_{2,0,5} &= s_{2,1}(x_1) \cdot s_{0,2}(x_2) \cdot s_{6,3}(x_3) = x_1^2 \cdot 1 \cdot (x_3 - K_4)^2 \end{aligned} \quad (9.44)$$

Werden mittels des Tensorprodukts n_b Basisfunktionen B_k erzeugt, so ist das Regressionsmodell durch die Gleichung 9.45 definiert.

$$\hat{y}(\mathbf{x}) = \sum_{k=1}^{n_b} b_k B_k(\mathbf{x}) \quad (9.45)$$

Dabei sind b_k mit Hilfe von Trainingsdaten zu bestimmende Parameter, welche beispielsweise wie bei der linearen Regression (Kapitel 9.2) durch die Methode der kleinsten Fehlerquadrate berechnet werden.

Die Übertragung vom ein-dimensionalen Ansatz in mehrere Dimensionen wird durch das Tensor-Produkt auf einfache Weise ermöglicht. Problematisch ist jedoch der exponentielle Anstieg der möglichen Basisfunktionen [$n_{b_{max}} = (p + 1 + l)^{n_f}$] bei Erhöhung der Faktanzahl. Zur Bestimmung der Konstanten b_k werden $n_r \geq n_b$ Datensätze benötigt. Im oben dargestellten Beispiel mit drei Faktoren, fünf Stützstellen und einem Polynomgrad von zwei ergeben sich bereits $(2 + 1 + 5)^3 = 512$ benötigte Datensätze. Zur Reduktion der erforderlichen Datenpunkte ist eine Auswahl aus allen möglichen Basisfunktionen notwendig, was sich in der Praxis als schwierig erweist.

Multivariate Adaptive Regression Splines

Eine praxisorientierte Variante der Splinemethodik bietet das von FRIEDMAN 1991 eingeführte Verfahren *Multivariate Adaptive Regression Splines* [29, 30, 96]. In diesem Algorithmus werden neben der Basisfunktion $B_{0,\dots,0} = 1$ nur Basisfunktionen aus Kombinationen der Terme $[x_j - K]_+$ und $[K - x_j]_+$ verwendet. Betrachten wir zur Erläuterung die Funktion $y(x) = e^{2x}$ im Bereich $x \in [0, 1]$, welche in Abbildung 9.12 als gepunktete Linie dargestellt ist. Auf Basis von 21 Testpunkten, die mit einer normalverteilten Störung $\mathcal{N}(0, 0.5)$ [$\mu = 0$ und $s = 0.5$] überlagert sind, werden zwei einfache Metamodelle ermittelt. Die lineare Regression führt zu dem Modell in Gleichung 9.46, welches zwar die steigende Tendenz des Zusammenhangs richtig abbildet, jedoch in großen Teilbereichen deutliche Abweichungen aufweist (Abbildung 9.12, rechts).

$$\hat{y}_{IR} = 0.301 + 6.1192x \quad (9.46)$$

Ein Regression Spline Modell mit nur einem Knotenpunkt bei $K = 0.6083$ liefert bereits eine deutlich bessere Approximation des wahren Zusammenhangs.

$$\hat{y} = 3.606 + 9.684[x - 0.6083]_+ - 4.2141[0.6083 - x]_+ \quad (9.47)$$

Die Konstante 3.606 entspricht dabei dem Mittelwert der 21 Testpunkte. Die Approximation der Ausgangsvariable y wird durch die Addition von drei Termen erzielt, die in Abbildung 9.13a separat skizziert sind. Durch Erhöhung der Knotenzahl kann der Zusammenhang in weitere Bereiche (Terme) aufgeteilt und die Vorhersagequalität des Modells weiter erhöht werden. Komplexe Zusammenhänge zwischen Ein- und Ausgangsvariablen werden somit bei Betrachtung lediglich eines einzelnen Faktors durch eine Addition von linearen Funktionen approximiert.

Mehrdimensionale Zusammenhänge ($n_f > 1$) enthalten typischerweise Interaktionen zwischen Faktoren, welche durch Multiplikation einzelner Terme ($[\bullet]_+$) dargestellt werden, so dass auch im mehrdimensionalen Raum komplexe Funktionszusammenhänge durch einfache Grundgleichungen dargestellt werden können (Abbil-

dung 9.13b).

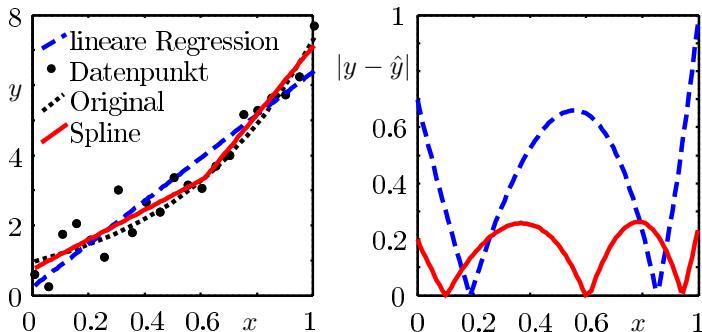


Abb. 9.12 Einfacher ein-dimensionaler Regression Spline

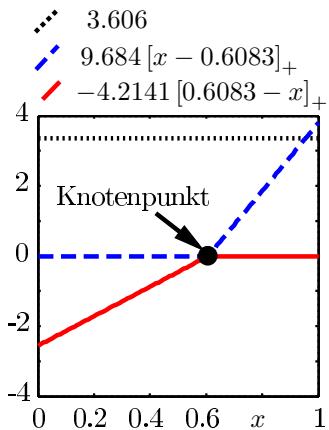


Abb. 9.13a Regression Spline: Terme

$$\begin{aligned} y &= 0.5 + 0.4 [x_1 - 0.8]_+ \\ &\quad - 0.3 [x_2 - 0.2]_+ \\ &\quad - 0.8 [0.6 - x_2]_+ [0.5 - x_1]_+ \end{aligned}$$

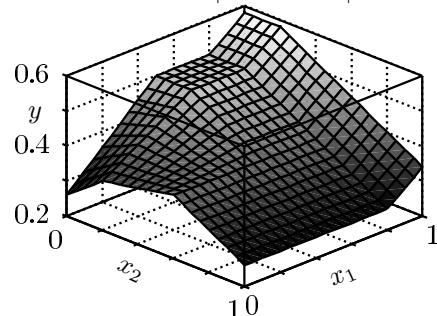


Abb. 9.13b mehrdimensionaler Regression Spline

Die eigentliche Herausforderung ist auch in diesem Verfahren die Auswahl der Knotenpunkte (Anzahl und Position) sowie die Ermittlung der dazugehörigen Konstanten. Zur Lösung dieser Aufgabe führt FRIEDMAN einen zweistufigen Algorithmus ein [29, 30].

Der erste Teil des Algorithmus (*forward pass*) startet mit einem Metamodell, welches nur aus einer konstanten Basisfunktion $B_0(\mathbf{x}) = 1$ besteht. Anschließend werden Schritt für Schritt neue Basisfunktionen erzeugt und dem Modell hinzugefügt. Dabei wird in jedem Schritt eine neue Basisfunktion mit einer bereits im Modell vorhandenen Basisfunktion multipliziert (siehe auch Gleichung 9.43). Es wird dabei immer die Kombination gewählt, welche den Modellfehler am stärksten reduziert. Grundsätzlich werden dem Modell in jedem Schritt zwei Basisfunktionen hinzuge-

fügt, die sich nur darin unterscheiden, dass sie jeweils eine Seite des Knotenpunkts ($x \geq K, x \leq K$) beschreiben ($b_1 [x - K]_+, b_2 [K - x]_+$). Zur Auswahl der besten Terme für eine Basisfunktion wird beispielsweise die *Methode der kleinsten Fehlerquadrate* eingesetzt. Die Erweiterung des Metamodells wird solange fortgeführt bis ein vorgegebener Fehler unterschritten oder eine maximale Anzahl an Basisfunktionen erreicht wurde. Normalerweise wird durch den *forward pass* ein überangepasstes (*overfitted*) Metamodell erzeugt, das zwar die gegebenen Testpunkte gut approximiert, bei dem die Vorhersagegenauigkeit für neue (unbekannte) Faktorkombinationen jedoch ungenügend ist [29].

Aus diesem Grund wird im zweiten Teil des Algorithmus (*backward pass*) das Modell wieder gestutzt (*pruned*), so dass eine bessere Allgemeingültigkeit erreicht wird. Dabei wird in jedem Schritt die Basisfunktion des Metamodells entfernt, welche den geringsten Effekt auf die Vorhersagequalität aufweist. Im Vergleich zum ersten Teil des Algorithmus, wo immer nur das nächste Basisfunktions-Paar gesucht wird, kann hier jede beliebige einzelne Basisfunktion gelöscht werden [29]. Zur Reduktion des Rechenaufwands im *forward pass* schlägt FRIEDMAN verschiedene Erweiterungen seines Verfahrens vor, die hier nicht weiter erläutert werden [30].

9.9 Support Vector Machines zur Klassifikation

Support Vector Machines oder auch Support Vector Networks wurden zur Klassifizierung und Mustererkennung eingeführt und sind in diesem Bereich weit verbreitet [100, 101]. Erweiterungen des grundlegenden Verfahrens ermöglichen ebenfalls den Einsatz zur Regression, wobei dann von Support Vector Regression gesprochen wird [19, 89, 80].

9.9.1 Klassifikation linear separierbarer Daten

Bei der Klassifizierung von Daten werden n_r Trainingspunkte betrachtet, von denen jeder Eingangsvektor \mathbf{x}_i, n_f verschiedene Attribute, Dimensionen oder Faktoren aufweist. Zusätzlich gehört jeder Datenpunkt zu einer von zwei Klassen $y_i \in \{-1, +1\}$. Die Trainingsdaten liegen somit in folgender Form vor:

$$\{\mathbf{x}_i, y_i\} \text{ wobei } i = 1 \dots n_r; y_i \in \{-1, 1\}, \mathbf{x}_i \in R^{n_f} \quad (9.48)$$

Im ersten Schritt wird angenommen, dass die Daten vollständig linear trennbar sind, was bedeutet, dass die Daten komplett durch eine $[n_f - 1]$ -dimensionale Hyperebene in zwei unterschiedliche Gruppen aufteilbar sind. Eine Hyperebene lässt sich in Normalenform durch $\mathbf{w} \cdot \mathbf{x} + b = 0$ beschreiben, wobei \mathbf{w} der Normalenvektor der Hyperebene ist und $\frac{b}{\|\mathbf{w}\|}$ den senkrechten Abstand der Hyperebene zum

Ursprung darstellt.

Die Datenpunkte, die der trennenden Hyperebene am nächsten liegen und die Lage der Hyperebene festlegen, werden *Support Vectors* genannt. Das Ziel des Support Vector Machine (SVM) Algorithmus ist es, die Hyperebene so im n_f -dimensionalen Raum auszurichten, dass sie so weit wie möglich von den nächstliegenden Datenpunkten, also den Support Vectors, entfernt liegt (Abbildung 9.14) [27, 80]. Die

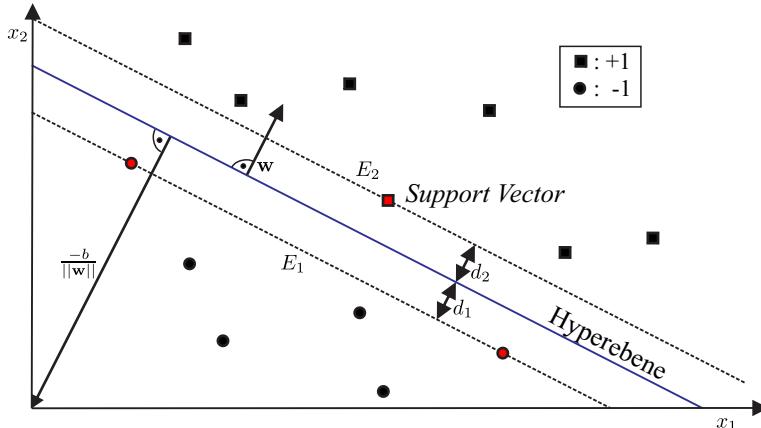


Abb. 9.14 $n_f - 1$ dimensionale Hyperebene zwischen zwei linear trennbaren Datensätzen im zweidimensionalen Faktorraum

Implementierung des Verfahrens wählt die Variablen \mathbf{w} und b genau so, dass für alle Trainingsspunkte gilt:

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1 \text{ für } y_i = +1 \quad (9.49)$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 \text{ für } y_i = -1 \quad (9.50)$$

Die Gleichungen können durch Gleichung 9.51 zusammengefasst werden:

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0 \quad \forall i = 1 \dots n_r \quad (9.51)$$

Zur Trennung der Daten müssen die Punkte (Support Vectors) auf der Hülle der Hyperebene gefunden werden. Die beiden einhüllenden n_f -dimensionalen Ebenen E_1 und E_2 werden über die folgenden Gleichungen definiert:

$$\mathbf{x}_i \cdot \mathbf{w} + b = +1 \text{ für } E_1 \quad (9.52)$$

$$\mathbf{x}_i \cdot \mathbf{w} + b = -1 \text{ für } E_2 \quad (9.53)$$

Nach Abbildung 9.14 wird d_1 als der Abstand zwischen der Hyperebene und E_1 und analog d_2 als Abstand zu E_2 definiert. Die Hyperebene liegt genau in der Mitte der beiden Ebenen, so dass $d_1 = d_2$ ist. Der Abstand zwischen E_1 und E_2 wird als *Margin* der SVM bezeichnet. Um die Hyperebene so weit als möglich von den Sup-

port Vectors entfernt auszurichten, muss folglich der Rand ($d = d_1 + d_2$) maximiert werden. Die geometrische Breite d des Randes ergibt sich bei einer Betrachtung unter Zuhilfenahme von zwei Punkten \mathbf{x}^+ und \mathbf{x}^- , die auf dem Rand liegen sowie Gleichungen 9.49 und 9.50, zu [16]:

$$d = \frac{1}{2} \left(\frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot \mathbf{x}^+ - \frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot \mathbf{x}^- \right) \quad (9.54)$$

$$d = \frac{1}{2\|\mathbf{w}\|} (\mathbf{w} \cdot \mathbf{x}^+ - \mathbf{w} \cdot \mathbf{x}^-) = \frac{1}{2\|\mathbf{w}\|} (1 - b - [-1 - b]) \quad (9.55)$$

$$d = \frac{1}{\|\mathbf{w}\|} \quad (9.56)$$

Um die Breite der Hülle zu maximieren muss somit $\frac{1}{\|\mathbf{w}\|}$ maximiert oder $\|\mathbf{w}\|$ minimiert werden [27, 80].

$$\min \|\mathbf{w}\| \text{ mit } y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0 \quad \forall i = 1 \dots n_r \quad (9.57)$$

Anstatt $\|\mathbf{w}\|$ zu optimieren, lässt sich äquivalent $\frac{1}{2}\|\mathbf{w}\|^2$ verwenden, wodurch sich dass Optimierungsproblem später mittels Quadratischer Programmierung (QP) lösen lässt [27, 80].

$$\min \frac{1}{2}\|\mathbf{w}\|^2 \text{ mit } y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0 \quad \forall i = 1 \dots n_r \quad (9.58)$$

$$\|\mathbf{w}\|^2 = \sum w_i^2 = \langle \mathbf{w} \cdot \mathbf{w} \rangle \quad (9.59)$$

Daraus ergibt sich mittels der Methode nach Lagrange das primale Problem:

$$L_P \equiv \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n_r} \alpha_i [y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1] \quad (9.60)$$

$$\Rightarrow L_P \equiv \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n_r} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{n_r} \alpha_i \quad (9.61)$$

$$\text{mit } \alpha_i \geq 0 \quad \forall i = 1 \dots n_r \quad (9.62)$$

Ziel ist es nun, Werte für \mathbf{w} und b zu finden, welche die L_P minimieren und Werte für α , welche L_P unter Beibehaltung der Bedingung $\alpha_i \geq 0$ maximieren [27, 80]. Die Ableitungen von L_P nach \mathbf{w} und b , welche zu Null gesetzt werden, ergeben sich zu:

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{n_r} \alpha_i y_i \mathbf{x}_i \quad (9.63)$$

$$\frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n_r} \alpha_i y_i = 0 \quad (9.64)$$

Einsetzen der Ableitungen in Gleichung 9.61 ergibt die duale Formulierung L_D , welche lediglich von α abhängt und maximiert werden muss [27, 80]:

$$L_D \equiv \frac{1}{2} \sum_{i,j=1}^{n_r} y_i y_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j - \underbrace{\sum_{i,j=1}^{n_r} y_i y_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j}_{=0} - b \sum_{i=1}^{n_r} \alpha_i y_i + \sum_{i=1}^{n_r} \alpha_i \quad (9.65)$$

$$L_D \equiv \sum_{i=1}^{n_r} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n_r} y_i y_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j \quad (9.66)$$

Vereinfacht lässt sich der Ausdruck wie folgt darstellen:

$$L_D \equiv \sum_{i=1}^{n_r} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n_r} \alpha_i H_{ij} \alpha_j \quad \text{wobei} \quad H_{ij} \equiv y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \quad (9.67)$$

$$L_D \equiv \sum_{i=1}^{n_r} \alpha_i - \frac{1}{2} \alpha' \mathbf{H} \alpha \quad \text{mit} \quad \alpha_i \geq 0 \ \forall i \quad \text{und} \quad \sum_{i=1}^{n_r} \alpha_i y_i = 0 \quad (9.68)$$

Die duale Formulierung L_D benötigt zur Berechnung nur das Skalarprodukt $\mathbf{x}_i \mathbf{x}_j$. Dies ist vorteilhaft, weil damit im späteren Verlauf ebenfalls die Vorgehensweise über Kernel-Funktionen ermöglicht wird (Kapitel 9.9.3). Aus der Aufgabe L_P zu minimieren hat sich durch die Methode von Lagrange ein Maximierungsproblem von L_D ergeben:

$$\max_{\alpha} \left[\sum_{i=1}^{n_r} \alpha_i - \frac{1}{2} \alpha' \mathbf{H} \alpha \right] \quad \text{mit} \quad \alpha_i \geq 0 \quad \forall i = 1 \dots n_r \quad \text{und} \quad \sum_{i=1}^{n_r} \alpha_i y_i = 0 \quad (9.69)$$

Hierbei handelt es sich um ein konvexes quadratisches Optimierungsproblem, welches durch einen QP-Solver gelöst wird. Zur Information sei hier darauf hingewiesen, dass die allgemeine Beschreibung eines quadratischen Optimierungsproblems wie folgt definiert ist und auf die weitere Lösung nicht weiter eingegangen wird [102].

$$\min_x \left[\frac{1}{2} \mathbf{x}' \mathbf{H} \mathbf{x} + \mathbf{f}' \mathbf{x} \right] \quad (9.70)$$

unter den Nebenbedingungen:

$$\mathbf{A} \mathbf{x} \leq \mathbf{b} \quad (9.71)$$

$$\mathbf{A}_{eq} \mathbf{x} = \mathbf{b}_{eq} \quad (9.72)$$

$$\mathbf{L} \mathbf{B} \leq \mathbf{x} \leq \mathbf{U} \mathbf{B} \quad (9.73)$$

Mit dem ermittelten α Vektor wird im Anschluss durch die Gleichung 9.63 der Vektor \mathbf{w} bestimmt. Die gesuchte Menge S der Support Vektoren wird durch die Trainingsdaten definiert deren α_i -Werte größer gleich Null sind ($\alpha_i > 0$). Zur Klassifizierung bekannter und unbekannter Faktorkombinationen fehlt abschließend lediglich der Wert für b , welcher mit Hilfe des Zusammenhangs bestimmt wird, dass jeder Support Vector folgende Gleichung erfüllt [27, 80]:

$$y_s(\mathbf{x}_s \cdot \mathbf{w} + b) = 1 \quad (9.74)$$

Nach Einsetzen von Gleichung 9.63 ergibt sich daraus:

$$y_s(\mathbf{x}_s \cdot \sum_{m \in S} \alpha_m y_m \mathbf{x}_m + b) = 1 \quad (9.75)$$

Wird der Ausdruck 9.75 mit y_s multipliziert und gleichzeitig die Bedingung $y_s^2 = 1$ verwendet (siehe Gleichungen 9.52 und 9.53), so ergibt sich zur Bestimmung von b [27]:

$$y_s^2 (\mathbf{x}_s \cdot \sum_{m \in S} \alpha_m y_m \mathbf{x}_m + b) = y_s \quad (9.76)$$

$$b = y_s - \sum_{m \in S} \alpha_m y_m \mathbf{x}_m \cdot \mathbf{x}_s \quad (9.77)$$

Anstatt einen zufällig ausgewählten Support Vector für die Bestimmung von b zu verwenden, ist es vorteilhaft, den Durchschnitt über alle Support Vectors aus S zu bestimmen [27]:

$$b = \frac{1}{n_s} \sum_{s \in S} \left(y_s - \sum_{m \in S} \alpha_m y_m \mathbf{x}_m \cdot \mathbf{x}_s \right) \quad (9.78)$$

Damit sind die Variablen \mathbf{w} und b ermittelt, welche die optimale Orientierung der Hyperebene und ebenfalls das *Support Vector Machine* Modell festlegen. Die Klassifikation neuer Datenpunkte \mathbf{x}_0 lässt sich mit den im vorigen errechneten Größen nach Gleichung 9.79 durchführen [27]:

$$y_0 = \text{sgn}(\mathbf{w} \cdot \mathbf{x}_0 + b) \quad (9.79)$$

9.9.2 Klassifikation nicht komplett linear separierbarer Daten

Um auch Daten verarbeiten zu können, die nicht komplett linear trennbar sind, werden Schlupfvariablen ξ_i (*slack variables*) eingeführt, welche die Forderungen aus den Gleichungen 9.49 und 9.50 aufweichen (Abbildung 9.15). Durch diese Erweiterung werden nun auch fehl-klassifizierte Punkte zugelassen, wobei das Verfahren als *soft margin* SVM bekannt ist [16]. Jeder Datenpunkt erhält dabei eine eigene Schlupfvariable $\xi_i, i = 1, \dots, n_r$:

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1 - \xi_i \quad \text{für } y_i = +1 \quad (9.80)$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 + \xi_i \quad \text{für } y_i = -1 \quad (9.81)$$

$$\xi_i \geq 0 \quad \forall i \quad (9.82)$$

Beide Gleichungen lassen sich auch in diesem Fall in eine Gleichung zusammenfassen.

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i \geq 0 \quad \text{wobei} \quad \xi_i \geq 0 \quad \forall i = 1 \dots n_r \quad (9.83)$$

Datenpunkten auf der falschen Seite der Hyperebene wird dadurch ein Strafwert ξ_i zugewiesen, der umso größer ist, je weiter der Punkt von der Hyperebene entfernt liegt. Um die Anzahl der Fehlklassifikationen möglichst gering zu halten wird die Minimierungsfunktion aus Gleichung 9.59 folgendermaßen erweitert [27]:

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n_r} \xi_i \quad \text{mit } y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i \geq 0 \quad \forall i = 1 \dots n_r \quad (9.84)$$

Hierbei kontrolliert der Parameter C die Abwägung zwischen der Summe der zugewiesenen Strafwerte und der Breite der Hyperebenenhülle. Die erweiterte Optimierungsaufgabe lässt sich mit Lagrange-Multiplikatoren vereinfachen, so dass aus dem Minimierungsproblem bezüglich \mathbf{w} , b und ξ_i ein Maximierungsproblem bezüglich α und μ wird. Dabei muss gelten: $\alpha_i \geq 0, \mu_i \geq 0 \quad \forall i = 1 \dots n_r$.

$$L_P \equiv \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n_r} \xi_i - \sum_{i=1}^{n_r} \alpha_i [y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i] - \sum_{i=1}^{n_r} \mu_i \xi_i \quad (9.85)$$

$$\Rightarrow L_P \equiv \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n_r} \xi_i - \sum_{i=1}^{n_r} \alpha_i [y_i(\mathbf{x}_i \cdot \mathbf{w} + b)] + \sum_{i=1}^{n_r} \alpha_i - \sum_{i=1}^{n_r} \alpha_i \xi_i - \sum_{i=1}^{n_r} \mu_i \xi_i \quad (9.86)$$

Die Differenzierung nach \mathbf{w} , b und ξ_i und zu Null setzen ergibt:

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{n_r} \alpha_i y_i \mathbf{x}_i \quad (9.87)$$

$$\frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n_r} \alpha_i y_i = 0 \quad (9.88)$$

$$\frac{\partial L_P}{\partial \xi_i} = 0 \Rightarrow C = \alpha_i + \mu_i \quad (9.89)$$

Eingesetzt in Gleichung 9.86 ergibt dies das duale Problem L_D :

$$L_D \equiv \frac{1}{2} \sum_{i,j=1}^{n_r} y_i y_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j + C \sum_{i=1}^{n_r} \xi_i - \sum_{i,j=1}^{n_r} y_i y_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j - b \underbrace{\sum_{i=1}^{n_r} \alpha_i y_i}_{=0} + \sum_{i=1}^{n_r} \alpha_i - \sum_{i=1}^{n_r} \alpha_i \xi_i - \sum_{i=1}^{n_r} \mu_i \xi_i \quad (9.90)$$

$$\Rightarrow L_D \equiv \sum_{i=1}^{n_r} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n_r} y_i y_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j + C \sum_{i=1}^{n_r} \xi_i - \sum_{i=1}^{n_r} \alpha_i \xi_i - \sum_{i=1}^{n_r} \mu_i \xi_i \quad (9.91)$$

$$\Rightarrow L_D \equiv \sum_{i=1}^{n_r} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n_r} y_i y_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j \quad (9.92)$$

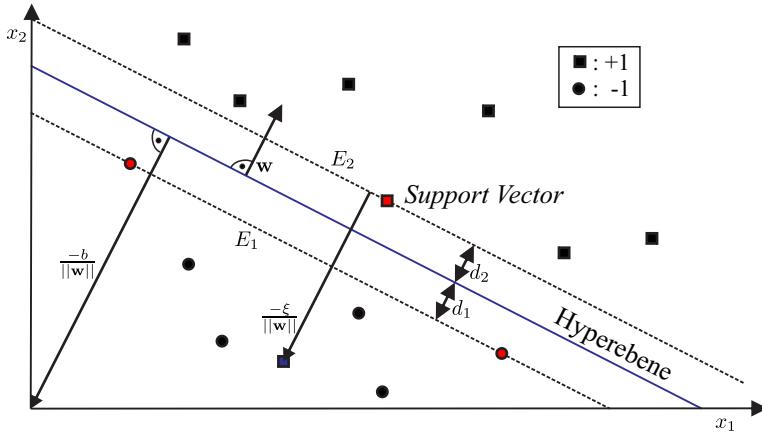


Abb. 9.15 Hyperebene zwischen zwei nicht vollständig linear trennbarer Datensätzen

Die Gleichung 9.92 ist identisch mit Gleichung 9.66. Durch die Forderung $\mu_i \geq 0 \forall i$ in Kombination mit Gleichung 9.89 folgt jedoch, dass $\alpha_i \leq C$ sein muss. Das quadratische Optimierungsproblem ergibt sich damit zu:

$$\max_{\alpha} \left[\sum_{i=1}^{n_r} \alpha_i - \frac{1}{2} \alpha' \mathbf{H} \alpha \right] \quad \text{mit} \quad 0 \leq \alpha_i \leq C \quad \forall i \quad \text{und} \quad \sum_{i=1}^{n_r} \alpha_i y_i = 0 \quad (9.93)$$

Der Wert für b (siehe Gleichung 9.80) lässt sich nun entsprechend Gleichung 9.78 bestimmen. Hierbei muss allerdings beachtet werden, dass als Support Vektoren nur diejenigen Vektoren benutzt werden, deren α_i -Wert zwischen 0 und C liegen: $0 < \alpha_i \leq C$ [27].

9.9.3 Nichtlineare Support Vector Machines

Bisher wurden nur lineare Fälle betrachtet, bei denen die Matrix \mathbf{H} mittels eines Vektorproduktes bestimmt wird.

$$H_{ij} = \mathbf{x}_i' \mathbf{x}_j = k(\mathbf{x}_i, \mathbf{x}_j) \quad (9.94)$$

Das Vektorprodukt ist ein Spezialfall der Funktions-Familie *Kernel-Funktionen* $[k(\mathbf{x}_i, \mathbf{x}_j)]$. Allen Varianten der Kernel-Funktionen ist gemein, dass sie zwei Vektoren eine Zahl (Skalar) zuordnen. Mit der passenden Gestaltung der Kernelfunktion ist es möglich, Daten in höherdimensionale Räume (Featureraum) zu transformieren, in dem die Daten einfacher (z.B. linear) als im Original-Faktorraum zu trennen

sind. Zur Transformation wird dazu eine Abbildungsfunktion der folgenden Form verwendet [27]:

$$\mathbf{x} \mapsto \phi(\mathbf{x}) \quad (9.95)$$

Ein Vorteil der verwendeten *Kernel*-Methode ist es, dass die benötigten Skalarprodukte im Featureraum auch ohne die explizite Berechnung der Transformation ϕ bestimmt werden können. Ein einfaches und eindimensionales Beispiel für nicht linear trennbare Daten ist in Abbildung 9.16 (a) gezeigt. Bereits durch die Transformation mittels einer quadratischen Funktion $\mathbf{x} \mapsto \phi(\mathbf{x})$ in den zweidimensionalen Featureraum ist eine einfache lineare Trennung der vorher nicht linear trennbaren Daten möglich (Abbildung 9.16 (b)) [62]. Es existieren verschiedene Kernel-

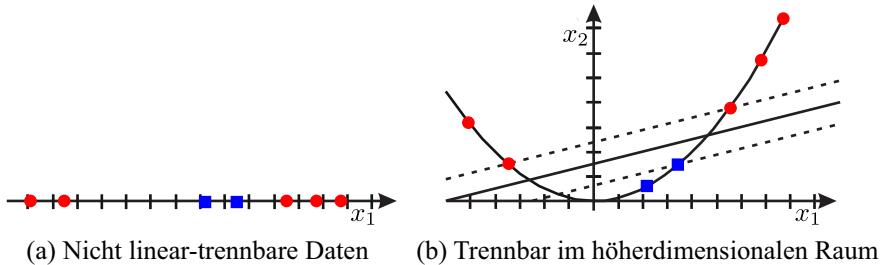


Abb. 9.16 Transformation von nicht linear trennbaren Daten in einen höherdimensionalen Featureraum

Funktionen, wobei einige für spezielle Aufgaben zugeschnitten sind. Die Konstruktion solcher speziellen Kernel-Funktionen ist teilweise aufwendig und verlangt vertiefte Kenntnisse über das zu lösende Problem [16]. Es existieren aber generische Kernel-Funktionen, die gute Ergebnisse bei vielen realen Problemen liefern [70].

linear	$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$
power	$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^d$
polynom	$K(\mathbf{x}_i, \mathbf{x}_j) = (a\mathbf{x}_i \cdot \mathbf{x}_j + b)^d$
sigmoid	$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(a\mathbf{x}_i \cdot \mathbf{x}_j + b)$
Gauss RBF	$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{1}{2} \mathbf{x}_i - \mathbf{x}_j ^2/\sigma^2)$

Die Gauß-RBF Funktion erfüllt die Anforderungen an Kernel-Funktionen formal nicht. Allerdings wurde gezeigt, dass sie sich trotzdem zum Einsatz als Kernel für SVM eignet [89]. In den folgenden Kapiteln wird ausschließlich mit Kernelfunktionen gearbeitet, wobei die folgende Notation gilt:

$$\mathbf{x}_i \cdot \mathbf{x}_j \mapsto K(\mathbf{x}_i, \mathbf{x}_j) \quad (9.96)$$

9.10 Support Vector Regression

Anstatt einer Klassifizierung unbekannter Variablen \mathbf{x}_0 in zwei Kategorien $y_0 = \pm 1$ lässt sich mittels Support Vector Machines auch eine Regression durchführen, wobei der Ausgangsvariable y Zahlenwerte aus \mathbb{R} zugewiesen werden.

Die Trainingsdaten \mathbf{x}_i und y_i liegen hierbei in folgender Form vor:

$$\{\mathbf{x}_i, y_i\} \text{ mit } i = 1, \dots, n_r, \quad y_i \in \mathbb{R}, \quad \mathbf{x} \in \mathbb{R}^{n_f} \quad (9.97)$$

$$y_i = \mathbf{w} \cdot \mathbf{x}_i + b \quad (9.98)$$

9.10.1 Regression mit linearer ε -unempfindlicher Straffunktion

Im Vergleich zur Klassifikation mittels SVM, bei der ein einfacher Strafwert (Schlupfvariable) zur Verarbeitung nicht vollständig linear trennbarer Daten eingesetzt wird, muss bei der Regression eine komplexere Straffunktion (*penalty function*) verwendet werden. Die in Abbildung 9.17 dargestellte lineare ε -intensive Straffunktion weist bei kleinen Abweichungen $|y_i - \hat{y}_i| < \varepsilon$ keinen Strafwert auf ($\xi = 0$) [27, 16]. Allen Punkten außerhalb dieser Grenze wird ein von der Entfernung abhängiger Strafwert (Schlupfvariable) ξ^+ oder ξ^- zugewiesen (Abbildungen 9.18, [16]). Hierbei wird unterschieden, auf welcher Seite der Hülle sich der Punkt befindet (+,-) [27, 108, 80].

$$\begin{aligned} y_i - \hat{y}_i &\leq \varepsilon + \xi_i^+ \\ y_i - \hat{y}_i &\geq -\varepsilon - \xi_i^- \end{aligned} \quad (9.99)$$

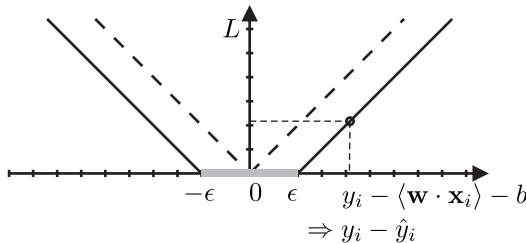


Abb. 9.17 Lineare ε -unempfindlicher Straffunktion (loss function)

Daraus folgt die zu minimierende Fehlerfunktion (Gleichung 9.100) für ein Regressionsmodell [16].

$$C \sum_{i=1}^{n_r} (\xi_i^+ + \xi_i^-) + \frac{1}{2} \|\mathbf{w}\|^2, \quad \text{mit } \xi_i^{+,-} \geq 0 \quad (9.100)$$

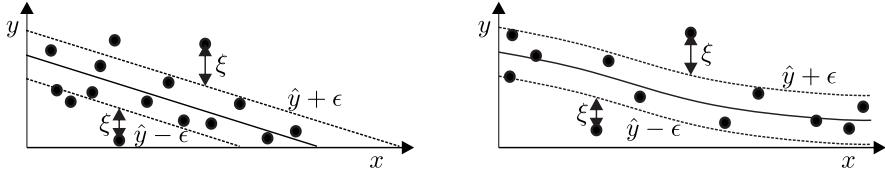


Abb. 9.18 Lineare und nichtlineare Regression mit ϵ -intensiver Hülle

Die Konstante $C > 0$ beschreibt dabei den Kompromiss zwischen der Flachheit der Funktion (Regression) und der Toleranz gegenüber Abweichungen außerhalb der ϵ -Hülle. Zur Lösung der Minimierungsaufgabe werden vier Lagrange Multiplikatoren eingeführt

$$\alpha_i^+ \geq 0, \quad \alpha_i^- \geq 0, \quad \mu_i^+ \geq 0, \quad \mu_i^- \geq 0 \quad \forall i = 1 \dots n_r \quad (9.101)$$

wodurch sich folgendes primale Problem ergibt [27]:

$$\begin{aligned} L_P = & C \sum_{i=1}^{n_r} (\xi_i^+ + \xi_i^-) + \frac{1}{2} \|\mathbf{w}\|^2 \\ = & \sum_{i=1}^{n_r} \alpha_i^+ (\epsilon + \xi_i^+ + y_i - \hat{y}_i) - \sum_{i=1}^{n_r} \alpha_i^- (\epsilon + \xi_i^- - y_i + \hat{y}_i) \\ & - \sum_{i=1}^{n_r} (\mu_i^+ \xi_i^+ + \mu_i^- \xi_i^-) \end{aligned} \quad (9.102)$$

Nach Einsetzen von Gleichung 9.98 ($\hat{y}_i = \mathbf{w} \cdot \mathbf{x}_i + b$) und Ableitung nach b, \mathbf{w}, ξ_i^+ sowie ξ_i^- ergibt sich [108, 27]:

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{n_r} (\alpha_i^- - \alpha_i^+) \mathbf{x}_i \quad (9.103)$$

$$\frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n_r} (\alpha_i^- - \alpha_i^+) = 0 \quad (9.104)$$

$$\frac{\partial L_P}{\partial \xi_i^+} = 0 \Rightarrow C = \alpha_i^+ + \mu_i^+ \quad (9.105)$$

$$\frac{\partial L_P}{\partial \xi_i^-} = 0 \Rightarrow C = \alpha_i^- + \mu_i^- \quad (9.106)$$

Substitution der Gleichungen aus 9.103 bis 9.106 in 9.102 ergibt das duale Problem (Gleichung 9.107)[108]:

$$\begin{aligned}
L_D = & C \underbrace{\sum_{i=1}^{n_r} (\xi_i^+ + \xi_i^-)}_{+A} + \frac{1}{2} \underbrace{\sum_{i,j=1}^{n_r} (\alpha_i^- - \alpha_i^+) (\alpha_j^- - \alpha_j^+) \langle \mathbf{x}_i, \mathbf{x}_j \rangle}_{\frac{1}{2}B} - \varepsilon \sum_{i=1}^{n_r} (\alpha_i^+ + \alpha_i^-) \\
& - \underbrace{\sum_{i=1}^{n_r} (\alpha_i^- \xi_i^- + \alpha_i^+ \xi_i^+)}_{-D} + \sum_{i=1}^{n_r} y_i (\alpha_i^- - \alpha_i^+) \\
& - \underbrace{\sum_{i,j=1}^{n_r} (\alpha_i^- - \alpha_i^+) (\alpha_j^- - \alpha_j^+) \langle \mathbf{x}_i, \mathbf{x}_j \rangle}_{-B} \\
& - b \underbrace{\sum_{i=1}^{n_r} (\alpha_i^- - \alpha_i^+)}_{=0} - C \underbrace{\sum_{i=1}^{n_r} (\xi_i^+ + \xi_i^-)}_{-A} + \sum_{i=1}^{n_r} (\alpha_i^- \xi_i^- + \alpha_i^+ \xi_i^+) \\
= & - \frac{1}{2} \sum_{i,j=1}^{n_r} (\alpha_i^- - \alpha_i^+) (\alpha_j^- - \alpha_j^+) \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \varepsilon \sum_{i=1}^{n_r} (\alpha_i^+ + \alpha_i^-) + \sum_{i=1}^{n_r} y_i (\alpha_i^- - \alpha_i^+) \tag{9.107}
\end{aligned}$$

Zur Verallgemeinerung kann der Term $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \mathbf{x}_i' \mathbf{x}_j$ durch eine Kernelfunktion (Gleichung 9.96) ersetzt werden. Mit den Voraussetzungen $\mu_i^+ \geq 0$ und $\mu_i^- \geq 0$ sowie den partiellen Ableitungen $\frac{\partial L_P}{\partial \xi_i^+} = 0$ und $\frac{\partial L_P}{\partial \xi_i^-} = 0$ aus Gleichung 9.106 folgt, dass $\alpha_i^+ \leq C$ und $\alpha_i^- \leq C$ ist. Das vollständige Optimierungsproblem ergibt sich damit zu:

$$\max_{\alpha^+, \alpha^-} \left[-\frac{1}{2} \sum_{i,j=1}^{n_r} (\alpha_i^- - \alpha_i^+) (\alpha_j^- - \alpha_j^+) \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \varepsilon \sum_{i=1}^{n_r} (\alpha_i^+ + \alpha_i^-) + \sum_{i=1}^{n_r} y_i (\alpha_i^- - \alpha_i^+) \right] \tag{9.108}$$

Die dazugehörigen Nebenbedingungen lauten:

$$0 \leq \alpha_i^{+, -} \leq C \text{ und } \sum_{i=1}^{n_r} (\alpha_i^+ - \alpha_i^-) = 0 \quad \forall i = 1 \dots n_r \tag{9.109}$$

Dieses Problem lässt sich ebenfalls mittels eines QP-Solver (*Quadratic Problem*) lösen, wozu es in die folgende Form gebracht wird.

$$\max_{\alpha} \left[\frac{1}{2} \alpha' \mathbf{H} \alpha + \mathbf{f}' \alpha \right] \tag{9.110}$$

Hierbei gilt angelehnt an RUNNERSSON und GUNN [80, 33]:

$$\alpha = \begin{bmatrix} \alpha^- \\ \alpha^+ \end{bmatrix}, \quad \mathbf{f} = \varepsilon + \begin{bmatrix} -\mathbf{y} \\ \mathbf{y} \end{bmatrix} \text{ und } \mathbf{H} = \begin{bmatrix} \mathbf{G} & -\mathbf{G} \\ -\mathbf{G} & \mathbf{G} \end{bmatrix} \text{ mit } \mathbf{G} = K(\mathbf{X}, \mathbf{X}') \tag{9.111}$$

Die Lösung dieses Problems liefert die Werte für α , wobei die gesuchten Support Vektoren durch α Werte zwischen Null und der Konstanten C ($0 < |\alpha^+ - \alpha^-| < C$) definiert sind. Ähnlich zur Klassifikation aus Kapitel 9.9.1 wird der Bias-Wert b mittels der gefundenen Support Vektoren ermittelt [48, 27].

$$b = y_s - \varepsilon - \sum_{i=1}^{n_r} (\alpha_i^- - \alpha_i^+) K(\mathbf{x}_i, \mathbf{x}_s) \quad (9.112)$$

oder

$$b = \frac{1}{N_s} \sum_{s \in S} \left[y_s - \varepsilon - \sum_{i=1}^{n_r} (\alpha_i^- - \alpha_i^+) K(\mathbf{x}_i, \mathbf{x}_s) \right] \quad (9.113)$$

Approximationen von y_0 für beliebige Faktorkombinationen x_0 lassen sich dann durch die Gleichung 9.114 berechnen.

$$y_0 = \sum_{i=1}^{n_r} (\alpha_i^- - \alpha_i^+) K(\mathbf{x}_i, \mathbf{x}_0) + b \quad (9.114)$$

9.10.2 v -SVR-Verfahren zur automatischen ε -Bestimmung

Da in den meisten Fällen der optimale Wert für ε nicht bekannt ist, wird dieser im v -SVR-Verfahren automatisch ermittelt. Die Größe von ε wird dabei durch den Ersatz-Parameter v bestimmt, welcher eine Abwägung zwischen der Modellkomplexität und der Wahrscheinlichkeit, viele Werte außerhalb der ε -Hülle zu erhalten, ist.

Die Fehlerfunktion der v -Regression lautet [86, 80]

$$C \left(v\varepsilon + \frac{1}{n_r} \sum_{i=1}^{n_r} (\xi_i^+ + \xi_i^-) \right) + \frac{1}{2} \|\mathbf{w}\|^2 \quad (9.115)$$

und wird mit den bekannten Nebenbedingungen minimiert.

$$y_i - \hat{y}_i \leq \varepsilon + \xi_i^+, \quad y_i - \hat{y}_i \geq -\varepsilon - \xi_i^- \quad \text{und} \quad \xi_i^{+,-} \geq 0 \quad (9.116)$$

Nach der Methode von Lagrange ergibt sich das duale Problem in Gleichung 9.117.

$$\max_{\alpha^-, \alpha^+} \left[\sum_{i=1}^{n_r} (\alpha_i^- - \alpha_i^+) y_i - \frac{1}{2} \sum_{i,j=1}^{n_r} (\alpha_i^- - \alpha_i^+) (\alpha_j^- - \alpha_j^+) K(\mathbf{x}_i, \mathbf{x}_j) \right] \quad (9.117)$$

mit den folgenden Randbedingungen [86, 81]

$$\sum_{i=1}^{n_r} (\alpha_i^+ - \alpha_i^-) = 0, \quad 0 \leq \alpha_i^{+,-} \leq \frac{C}{n_r} \quad \text{und} \quad \sum_{i=1}^{n_r} (\alpha_i^+ + \alpha_i^-) \leq Cv \quad (9.118)$$

Dieses Problem lässt sich ebenfalls mit dem QP-Löser berechnen, wobei die oben genannten Randbedingungen berücksichtigt werden müssen (Gleichung 9.110).

$$\alpha = \begin{bmatrix} \alpha^- \\ \alpha^+ \end{bmatrix}, \mathbf{f} = \begin{bmatrix} -\mathbf{y} \\ \mathbf{y} \end{bmatrix} \text{ und } \mathbf{H} = \begin{bmatrix} \mathbf{G} & -\mathbf{G} \\ -\mathbf{G} & \mathbf{G} \end{bmatrix} \text{ mit } \mathbf{G} = K(\mathbf{X}, \mathbf{X}') \quad (9.119)$$

Die Support Vektoren werden nach ihren Vorzeichen gruppiert und folgendermaßen bestimmt.

$$S^+ : 0 < \alpha^- - \alpha^+ < \frac{C}{n_r} \text{ und } S^- : -\frac{C}{n_r} < \alpha^- - \alpha^+ < 0 \quad (9.120)$$

Die Werte für b und ε werden dann durch folgendes Gleichungssystem bestimmt [81]:

$$\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \cdot \begin{bmatrix} b \\ \varepsilon \end{bmatrix} = \begin{bmatrix} y^+ - (\alpha^- - \alpha^+) G(\mathbf{x}_+, \mathbf{x}_{1 \dots n_r}) \\ (\alpha^- - \alpha^+) G(\mathbf{x}_-, \mathbf{x}_{1 \dots n_r}) - y^- \end{bmatrix} \quad (9.121)$$

In diesem Gleichungssystem stehen $y^{+,-}$, $\mathbf{x}_{+,-}$ und $\alpha^{+,-}$ für jeweils einen ausgewählten Support Vector aus den Gruppen $S^{+,-}$. Vorhersagen werden im Anschluss durch die folgende Gleichung ermittelt:

$$y_0 = \sum_{i=1}^{n_r} (\alpha_i^- - \alpha_i^+) K(\mathbf{x}_i, \mathbf{x}_0) + b \quad (9.122)$$

9.10.3 Regression mit quadratischer ε -unempfindlicher Straffunktion

Bei Verwendung der quadratischen ε -unempfindlichen Straffunktion steigt der Strafwert außerhalb der ε -Hülle quadratisch anstatt linear an (Abbildung 9.19) [16]. Die zu minimierende Fehlerfunktion ändert sich dadurch nur geringfügig und die Nebenbedingungen bleiben gleich.

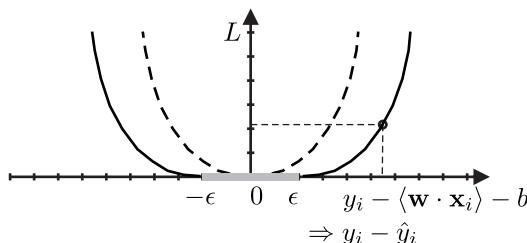


Abb. 9.19 Quadratische ε -unempfindliche Straffunktion (loss function)

$$C \sum_{i=1}^{n_r} (\xi_i^{+2} + \xi_i^{-2}) + \|\mathbf{w}\|^2 \quad (9.123)$$

$$y_i - \hat{y}_i \leq \varepsilon + \xi_i^-, \hat{y}_i - y_i \leq \varepsilon + \xi_i^+ \text{ und } \xi_i^{-,+} \geq 0 \quad (9.124)$$

Mit der Methode nach Lagrange lässt sich auch hier das duale Problem herleiten [16]:

$$\max_{\alpha^+, \alpha^-} \left[\sum_{i=1}^{n_r} (\alpha_i^- - \alpha_i^+) y_i - \varepsilon \sum_{i=1}^{n_r} (\alpha_i^- + \alpha_i^+) \cdots \right. \\ \left. - \frac{1}{2} \sum_{i,j=1}^{n_r} (\alpha_i^- - \alpha_i^+) (\alpha_j^- - \alpha_j^+) K(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{C} \delta_{ij} \right] \quad (9.125)$$

Das sogenannte Kronecker-Delta δ_{ij} ist in Gleichung 9.126 definiert und in der Matrixnotation gleichbedeutend mit der Einheitsmatrix \mathbf{I} .

$$\delta_{ij} = \begin{cases} 1 & \text{falls } i = j \\ 0 & \text{falls } i \neq j \end{cases} \quad (9.126)$$

Die zu dem dualen Problem gehörenden Nebenbedingungen sind:

$$\alpha_i^{-,+} \geq 0 \text{ und } \sum_{i=1}^{n_r} (\alpha_i^- - \alpha_i^+) = 0 \quad (9.127)$$

Identisch zur linearen Strafffunktion lässt sich diese Problem durch einen QP-Solver lösen [80].

$$\alpha = \begin{bmatrix} \alpha^- \\ \alpha^+ \end{bmatrix}, \mathbf{f} = \varepsilon + \begin{bmatrix} \mathbf{y} \\ +\mathbf{y} \end{bmatrix} \quad (9.128)$$

$$\mathbf{H} = \begin{bmatrix} \mathbf{G} + \frac{1}{C} \mathbf{I} & -\mathbf{G} - \frac{1}{C} \mathbf{I} \\ -\mathbf{G} - \frac{1}{C} \mathbf{I} & \mathbf{G} + \frac{1}{C} \mathbf{I} \end{bmatrix} \text{ mit } \mathbf{G} = K(\mathbf{x}, \mathbf{x}') \quad (9.129)$$

Die gesuchten Support Vektoren S sind durch $0 < \alpha < C$ definiert und b wird mittels der Elemente von S bestimmt.

$$b = y_s - \varepsilon - \sum_{m=1}^{n_r} ((\alpha_m^- - \alpha_m^+) K(\mathbf{x}_m, \mathbf{x}_s)) - \frac{\alpha_m^- - \alpha_m^+}{C} \quad (9.130)$$

oder als Mittelwert aller Support Vektoren

$$b = \frac{1}{n_s} \sum_{s \in S} \left[y_s - \varepsilon - \sum_{m=1}^{n_r} ((\alpha_m^- - \alpha_m^+) K(\mathbf{x}_m, \mathbf{x}_s)) - \frac{\alpha_m^- - \alpha_m^+}{C} \right] \quad (9.131)$$

Vorhersagen an einem unbekannten Datenpunkt \mathbf{x}_0 lassen sich mit Gleichung 9.132 ermitteln.

$$y_0 = \sum_{i=1}^{n_r} (\alpha_i^i - \alpha_i^+) K(\mathbf{x}_i, \mathbf{x}_0) + b \quad (9.132)$$

9.10.4 Least Square Support Vector Regression

SUYKENS schlägt eine Least Square Support Vector Methode vor, die anstelle eines konvex quadratischen Problems einen Satz von linearen Gleichungen löst [95]. Das Minimierungsproblem wird dabei von

$$\min \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n_r} \xi_i \right\} \text{ mit } y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i \geq 0 \text{ und } \xi_i \geq 0 \quad \forall i = 1 \dots n_r \quad (9.133)$$

folgendermaßen umformuliert.

$$\min \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + \gamma \sum_{i=1}^{n_r} e_i^2 \right\} \text{ mit } y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + e_i = 0 \quad \forall i = 1 \dots n_r \quad (9.134)$$

Das primäre Problem ändert sich dadurch von

$$L_P = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n_r} \xi_i - \sum_{i=1}^{n_r} \alpha_i [y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i] - \sum_{i=1}^{n_r} \mu_i \xi_i \quad (9.135)$$

$$\text{zu } L_P = \frac{1}{2} \|\mathbf{w}\|^2 + \gamma \sum_{i=1}^{n_r} e_i^2 - \sum_{i=1}^{n_r} \alpha_i [y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + e_i] \quad (9.136)$$

wobei α_i nun positiv oder negativ werden kann [95]. Die Nebenbedingungen folgen aus den Kuhn-Tucker Bedingungen [26].

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{n_r} \alpha_i y_i \mathbf{x}_i \quad (9.137)$$

$$\frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n_r} \alpha_i y_i = 0 \quad (9.138)$$

$$\frac{\partial L_P}{\partial e_i} = 0 \Rightarrow \alpha_i = \gamma e_i \quad (9.139)$$

$$\frac{\partial L_P}{\partial \alpha_i} = 0 \Rightarrow y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + e_i = 0 \quad (9.140)$$

b und α können durch das folgende Gleichungssystem bestimmt werden [26, 95, 94].

$$\begin{bmatrix} 0 & -Y' \\ Y & \Omega + \gamma^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad (9.141)$$

mit

$$\begin{aligned} Y' &= [y_1, y_2, \dots, y_{n_r}] : \text{bekannte Systemantworten} \\ I &: \text{Einheitsmatrix der Größe } n_r \times n_r \\ \Omega &: \text{Kernel Matritze mit } \Omega_{i,j} = K(x_i, x_j) \end{aligned} \quad (9.142)$$

Die Vorhersage für einen Datenpunkt \mathbf{x}_0 erfolgt durch Gleichung 9.143.

$$y_0 = \sum_{i=1}^{n_r} \alpha_i K(\mathbf{x}_i, \mathbf{x}_0) + b \quad (9.143)$$

mit $b = \frac{\sum (H^{-1} \mathbf{y})}{\sum H^{-1}}$, $\alpha_i = \sum_{j=1}^{n_r} y_j H_{i,j}^{-1} - b \sum_{j=1}^{n_r} H_{i,j}^{-1}$ und $H = \Omega + \gamma^{-1} I$

9.11 Ridge Regression

Im Vergleich zur Support Vector Regression verwendet die Ridge Regression typischerweise einfachere Kernel-Funktionen und basiert auf der Minimierung der quadratischen Fehlerfunktion der klassischen linearen Regression.

$$\min \frac{1}{2} \sum_{i=1}^{n_r} (y_i - \mathbf{w} \cdot \mathbf{x}_i - b)^2 = \min \frac{1}{2} \sum_{i=1}^{n_r} (y_i - \hat{y}_i)^2 \quad (9.144)$$

Gerade bei der Analyse von höher dimensionalen Problemen tendiert dieser Ansatz zu Overfitting. Dem entgegenzuwirken wird die Norm von \mathbf{w} in Verbindung mit einem Regulierungsfaktor λ als Straffunktion eingeführt $\lambda \|\mathbf{w}\|^2$. Dieses Vorgehen wird als weight-decay-Methode bezeichnet [107]. Die Minimierungsfunktion ergibt sich dadurch zu

$$\min \left[\lambda \|\mathbf{w}\|^2 + \sum_{i=1}^{n_r} \xi_i^2 \right] \text{ mit } \xi_i = y_i - \mathbf{w} \cdot \mathbf{x}_i - b \quad (9.145)$$

Mit Hilfe eines Lagrange Multiplikators α lässt sich das Minimierungsproblem in ein primäres Problem umwandeln.

$$L_P = \lambda \|\mathbf{w}\|^2 + \sum_{i=1}^{n_r} \xi_i^2 + \sum_{i=1}^{n_r} \alpha_i (y_i - \mathbf{w} \cdot \mathbf{x}_i - b - \xi_i) \quad (9.146)$$

Die partielle Differenzierung nach den Variablen \mathbf{w} , b und ξ ergibt:

$$\begin{aligned} \frac{\partial L_P}{\partial \mathbf{w}} &= 2\lambda \mathbf{w} - \sum_{i=1}^{n_r} \alpha_i \mathbf{x}_i = 0 \Rightarrow \mathbf{w} = \frac{1}{2\lambda} \sum_{i=1}^{n_r} \alpha_i \mathbf{x}_i \\ \frac{\partial L_P}{\partial b} &= - \sum_{i=1}^{n_r} \alpha_i = 0 \Rightarrow \sum_{i=1}^{n_r} \alpha_i = 0 \\ \frac{\partial L_P}{\partial \xi} &= 2 \sum_{i=1}^{n_r} \xi_i - \sum_{i=1}^{n_r} \alpha_i = 0 \Rightarrow \sum_{i=1}^{n_r} \xi_i = \frac{1}{2} \sum_{i=1}^{n_r} \alpha_i \end{aligned} \quad (9.147)$$

Die Kombination aller Gleichungen resultiert in das duale Problem mit zugehöriger Maximierungsaufgabe [107, 85], wobei bereits eine Verallgemeinerung mittels Kernelfunktion K eingeführt wurde (siehe Kapitel 9.9.3).

$$\begin{aligned} L_D &= \lambda \frac{1}{4\lambda^2} \sum_{i=1}^{n_r} \sum_{j=1}^{n_r} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{4} \sum_{i=1}^{n_r} \alpha_i^2 + \sum_{i=1}^{n_r} \alpha_i y_i \\ &\quad - \frac{1}{2\lambda} \sum_{i=1}^{n_r} \sum_{j=1}^{n_r} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) - b \sum_{i=1}^{n_r} \alpha_i - \frac{1}{2} \sum_{i=1}^{n_r} \alpha_i^2 \\ L_D &= \sum_{i=1}^{n_r} \alpha_i y_i - \frac{1}{4\lambda} \sum_{i=1}^{n_r} \sum_{j=1}^{n_r} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{4} \sum_{i=1}^{n_r} \alpha_i^2 \end{aligned} \quad (9.148)$$

$$\max_{\alpha} \left(\mathbf{y}' \alpha - \frac{1}{4\lambda} \alpha' \mathbf{K} \alpha - \frac{1}{4} \alpha' \alpha \right) \quad \text{mit } \sum_{i=1}^{n_r} \alpha_i = 0 \quad (9.149)$$

Bei der klassischen Ridge Regression wird vom unbiased Fall $b = 0$ ausgegangen. Dadurch ist der Term $b \sum_{i=1}^{n_r} \alpha_i$ aus Gleichung 9.148 bereits Null und die Nebenbedingung $\sum_{i=1}^{n_r} \alpha_i = 0$ kann vernachlässigt werden. In diesem Fall wird α folgendermaßen bestimmt [80]

$$\frac{dL_D}{d\alpha} = \mathbf{y} - \frac{1}{2\lambda} \mathbf{K} \alpha - \frac{1}{2} \alpha = 0 \Rightarrow \alpha = (\mathbf{K} + \lambda \mathbf{I})^{-1} 2\lambda \mathbf{y} \quad (9.150)$$

Vorhersagen für einen Datenpunkt \mathbf{x}_0 erfolgen durch:

$$y_0 = \frac{1}{2\lambda} \sum_{i=1}^{n_r} K(\mathbf{x}_i, \mathbf{x}_0) \alpha_i = \sum_{i=1}^{n_r} K(\mathbf{x}_i, \mathbf{x}_0) (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{y} \quad (9.151)$$

Im allgemeinen Fall ($b \neq 0$) wird α mittels eines QP Solvers bestimmt [80] (siehe auch Kapitel 9.9).

$$\max_{\alpha} \left[\frac{1}{2} \alpha' \mathbf{H} \alpha + \mathbf{f}' \alpha \right] \quad (9.152)$$

$$H = \frac{1}{2} \left[\frac{1}{\lambda} \mathbf{K} + \mathbf{I} \right] \quad \text{und } f = -\mathbf{y} \quad (9.153)$$

Der Parameter b lässt sich dann mit einem beliebig bekannten Punkt k berechnen:

$$b = y_k - \mathbf{w} \mathbf{x}_k - \xi_k = y_k - \frac{1}{2\lambda} \sum_{i=1}^{n_r} [K(\mathbf{x}_i, \mathbf{x}_k) \alpha_i] - \frac{1}{2} \alpha_k \quad (9.154)$$

Mit bekanntem b und α können dann Vorhersagen y_0 für die Faktorkombination \mathbf{x}_0 berechnet werden [80].

$$y_0 = \frac{1}{2\lambda} \sum_{i=1}^{n_r} K(\mathbf{x}_i, \mathbf{x}_0) \alpha_i + b \quad (9.155)$$

9.12 Kleinster und gewichteter Abstand

Das Verfahren des kleinsten Abstands, welches auch unter Polygon Verfahren oder *Nearest Point* bekannt ist, approximiert Funktionswert y_0 an der Faktorkombination \mathbf{x}_0 durch den Funktionswert des räumlich am nächsten liegenden und bekannten Datenpunkts. Ein zweidimensionaler Faktorraum wird dadurch in Polygone mit jeweils konstantem Funktionswert aufgeteilt. An den Grenzen der Polygone treten Unstetigkeiten auf, so dass keine kontinuierlichen Metamodelle erzeugt werden. Häufig weist dieses Modellverfahren gerade an den Polygongrenzen hohe Approximationsfehler auf. Abbildung 9.20 zeigt die Approximation der Testfunktion aus Gleichung 9.156 bei 10 bekannten Trainingspunkten:

$$y = \sin(\pi x_1) x_2 \quad (9.156)$$

Soll nicht nur der nächste sondern mehrere Datenpunkte in der Umgebung ei-

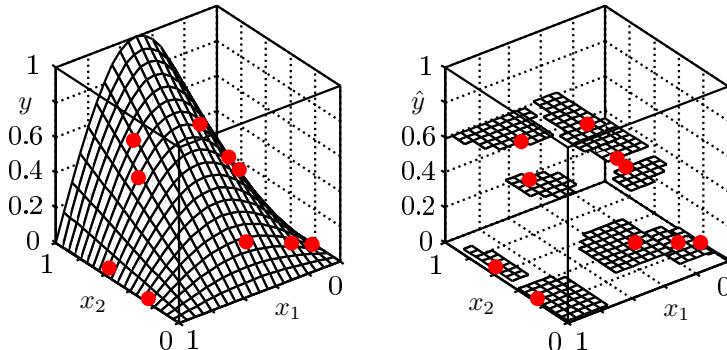


Abb. 9.20 Metamodell mittels kleinstem Abstand

nes Punkts \mathbf{x}_0 für die Approximation herangezogen werden, wird im einfachsten Fall der Einfluss einzelner Datenpunkte durch ihren Abstand $d_i = \|\mathbf{x}_i - \mathbf{x}_0\|_2$ zum Punkt \mathbf{x}_0 ermittelt. Die Inverser-Abstand-Gewichtungs Methode (*Inverse-Distance-Weighting Method*) approximiert den gesuchten Funktionswert $y(\mathbf{x}_0)$ wie folgt.

$$\hat{y}(\mathbf{x}_0) = \frac{\sum_{i=1}^{n_r} \frac{y(\mathbf{x}_i)}{d_i}}{\sum_{i=1}^{n_r} \frac{1}{d_i}} \quad \text{mit } d_i = \|\mathbf{x}_i - \mathbf{x}_0\|_2 \quad (9.157)$$

Hierbei ist es dem Anwender überlassen ob alle Datenpunkte, nur Datenpunkte bis zu einer maximalen Entfernung oder eine feste Anzahl von Datenpunkten (die dem gesuchten Punkt am nächsten sind) zur Berechnung verwendet werden. Da das Ergebnis direkt von der Abstandsbestimmung abhängt, ist es notwendig die Eingangsvariablen (Faktoren) zu normieren. Ein grundlegendes Problem des Verfahrens ist,

dass die Verteilung der Trainingsdaten im Faktorraum nicht berücksichtigt wird. Liegen in den Daten Faktorbereiche mit höherer Datendichte vor, so nimmt die Bedeutung dieser Bereiche für die Approximation zu, da alle Datenpunkte gleichberechtigt verwendet werden. Die Vorhersage eines guten Metamodells sollte jedoch nicht davon abhängen, ob in einem lokalen Faktorbereich mehr oder weniger Messpunkte vorhanden sind.

Die Kriging Methode (Kapitel 9.13) berücksichtigt daher neben den Abständen der einzelnen Testpunkte zur gesuchten Faktorkombination \mathbf{x}_0 ebenfalls die Verteilung der gegebenen Trainingsdaten. Tritt eine höhere Dichte von Messdaten an einem Faktorbereich auf, so wird das Gewicht jedes einzelnen Datenpunkts aus dieser Gruppe automatisch reduziert.

9.13 Kriging

Kriging ist ein Modellverfahren, das nach dem Südafrikanischen Bergbauingenieurs Danie G. Krige benannt ist. Bei einer Kriging-Interpolation wird ein unbekannter Funktionswert $y(\mathbf{x}_0)$ an einer Faktorkombination \mathbf{x}_0 durch ein gewichtetes Mittel von bekannten Nachbarpunkten geschätzt. Grundlage für die Schätzung ist dabei ein Variogramm, in dem räumliche Zusammenhänge bekannter Datenpunkte festgehalten werden. Zur Approximation eines Funktionswerts $y(\mathbf{x}_0)$ ist es notwendig den Erwartungswert der quadratischen Abweichung zweier Punkte in Abhängigkeit von deren Abstands \mathbf{r} zu ermitteln. Ein solches Konstrukt wird als Variogramm $2\gamma(\mathbf{r})$ bezeichnet [67]:

$$2\gamma(r) = 2\gamma(\|\mathbf{x}_i - \mathbf{x}_j\|_2) = E(y(\mathbf{x}_i) - y(\mathbf{x}_j))^2 = \text{Var}(y(\mathbf{x}_i) - y(\mathbf{x}_j)) \quad (9.158)$$

Das Variogramm lässt sich aus empirischen Trainingsdaten nach Gleichung 9.159 ermitteln [67, 70]:

$$2\gamma(r) = \frac{1}{n_{r^*}} \sum_{i=1}^{n_{r^*}} (y(\mathbf{x}_i) - y(\mathbf{x}_i + \mathbf{r}))^2 \quad (9.159)$$

Hierbei steht n_{r^*} für die Anzahl von Datenpunkt-kombinationen mit dem Abstand r . In der Praxis werden die Abstände manchmal zur Vereinfachung in Klassen eingeteilt. Eine alternative und meist verwendete Darstellung des Variogramms $[2\gamma(r)]$ ist das Semi-Variogramm $[\gamma(r)]$ [67, 84]:

$$\gamma(r) = \frac{1}{2} \text{Var}(y(\mathbf{x}_1) - y(\mathbf{x}_2)) \quad (9.160)$$

Die Eigenschaften eines Variogramms um den Ursprung sind von besonderer Bedeutung, wobei für den Sonderfall $r = 0$ immer $\gamma(0) = 0$ gilt. Empirische Semi-Variogramme werden in der Praxis für kleine Abstände einen Wert ungleich Null annehmen. Diese Abweichung wird mit *Nugget-Effekt* c_n bezeichnet (Abbildung 9.21). Der Nugget Effekt entsteht durch Fehlmessungen, Rauschen, Toleranzen des Messmittels sowie örtliche Variationen unterhalb eines minimalen Messabstands. Der mi-

nimale Messabstand kann dabei zum Beispiel durch wirtschaftliche Randbedingungen oder wie im Bergbau durch eine kleinste Körnung verursacht werden. Der Funktionswert eines Semi-Variogramms steigt grundsätzlich mit dem Abstand r an und nähert sich in den meisten Fällen ab einer Entfernung α_0 (*Range*) einer Schwelle c_0 (*Sill*) (Abbildung 9.21). Die Schwelle entspricht dabei ungefähr der Varianz σ^2 des Gesamtsystems ($\gamma(\infty) = c_0 \approx \sigma^2$). Falls die Trainingsdaten gleichmäßig im Faktorraum verteilt sind, kann die Schwelle (*Sill*) in vielen Fällen mit Gleichung 9.161 abgeschätzt werden [50, 17, 4].

$$\widehat{c}_0 = \widehat{c}_n + \widehat{\sigma}_0^2 = \frac{1}{n_r} \sum_{i=1}^{n_r} (y_i - \bar{y})^2 \quad (9.161)$$

Bei der Bestimmung eines Variogramms aus einer begrenzten Menge empirischer

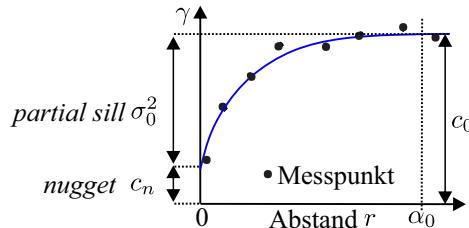


Abb. 9.21 Semi-Variogramm

Trainingsdaten ergeben sich einzelne diskrete Funktionswerte. Um einen funktionalen Zusammenhang für beliebige Abstände r zu erhalten, wurden verschiedene theoretische Semivariogramme eingeführt, welche an die Trainingsdaten angepasst werden [84, 67]. Im Folgenden sind einige bekannte theoretische Variogramme dargestellt, wobei sie teilweise zu den Originalquellen leicht modifiziert wurden [84, 67, 70, 15, 103].

Linear

$$\gamma(r) = \begin{cases} 0 & |r| = 0 \\ c_n + \sigma_0^2 \frac{|r|}{\alpha_0} & |r| \neq 0 \end{cases}$$

Potenz (*power*)

$$\gamma(r) = \begin{cases} 0 & |r| = 0 \\ c_n + \sigma_0^2 \left(\frac{|r|}{\alpha_0} \right)^\beta & |r| \neq 0 \text{ mit } 0 \leq \beta < 2, \text{ e.g. } \beta = 0.5 \end{cases}$$

Exponentiell

$$\gamma(r) = \begin{cases} 0 & |r| = 0 \\ c_n + \sigma_0^2 \left(1 - e^{-\beta \frac{|r|}{\alpha_0}} \right) & |r| \neq 0 \text{ mit } \beta > 0 \text{ e.g. } \beta = 4 \end{cases}$$

Spherical

$$\gamma(\mathbf{r}) = \begin{cases} 0 & |r| = 0 \\ c_n + \sigma_0^2 \left(\frac{3}{2} \frac{|r|}{\alpha_0} - \frac{1}{2} \left(\frac{|r|}{\alpha_0} \right)^3 \right) & 0 \leq |r| \leq \alpha_0 \\ c_n + \sigma_0^2 & |r| > \alpha_0 \end{cases}$$

Rational Quadratisch

$$\gamma(\mathbf{r}) = \begin{cases} 0 & |r| = 0 \\ c_n + \sigma_0^2 \frac{\left(\frac{|r|}{\alpha_0} \right)^2}{\beta + \left(\frac{|r|}{\alpha_0} \right)^2} & |r| \neq 0 \text{ mit } \beta > 0, \text{ e.g. } \beta = 0.01 \end{cases}$$

Gauß

$$\gamma(r) = \begin{cases} 0 & |r| = 0 \\ c_n + \sigma_0^2 \left(1 - e^{-\beta \left(\frac{|r|}{\alpha_0} \right)^2} \right) & |r| \neq 0 \text{ mit } \beta > 0, \text{ e.g. } \beta = 4 \end{cases}$$

Kubisch

$$\gamma(r) = \begin{cases} 0 & |r| = 0 \\ c_n + \sigma_0^2 \left[7 \left(\frac{|r|}{\alpha_0} \right)^2 - \frac{35}{4} \left(\frac{|r|}{\alpha_0} \right)^3 + \frac{7}{2} \left(\frac{|r|}{\alpha_0} \right)^5 - \frac{3}{4} \left(\frac{|r|}{\alpha_0} \right)^7 \right] & 0 < |r| \leq \alpha_0 \\ c_n + \sigma_0^2 & |r| > \alpha_0 \end{cases}$$

Pentaspherical

$$\gamma(r) = \begin{cases} 0 & |r| = 0 \\ c_n + \sigma_0^2 \left[\frac{15}{8} \frac{|r|}{\alpha_0} - \frac{5}{4} \left(\frac{|r|}{\alpha_0} \right)^3 + \frac{3}{8} \left(\frac{|r|}{\alpha_0} \right)^5 \right] & 0 < |r| \leq \alpha_0 \\ c_n + \sigma_0^2 & |r| > \alpha_0 \end{cases}$$

Sine Hole Effect (Welle)

$$\gamma(\mathbf{r}) = \begin{cases} 0 & |r| = 0 \\ c_n + \sigma_0^2 \left[1 - \frac{\sin \left(\frac{|r|}{\beta} \right)}{\frac{|r|}{\beta}} \right] & |r| \neq 0 \text{ mit } \beta > 0, \text{ e.g. } \beta = \frac{\alpha_0}{4\pi} \end{cases}$$

Abbildung 9.22 zeigt zur Veranschaulichung verschiedene Semi-Variogramme mit folgenden Parametern: $\alpha_0 = 0.5$, $c_n = 0.1$, $c_0 = 2.2$, $\beta = \text{siehe Gleichungen}$
 Die vorgestellten Semi-Variogramme weisen alle die in Gleichung 9.162 gezeigte Grundform auf, wodurch eine Abschätzung der partiellen Schwelle σ_0^2 mittels gegebener Trainingsdaten in einer allgemeinen Form möglich ist.

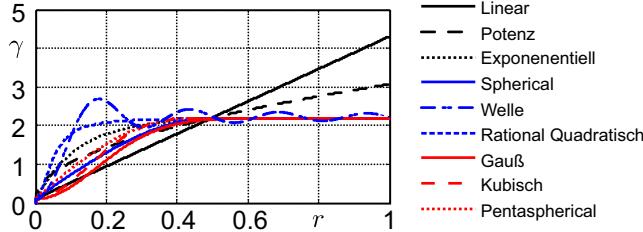


Abb. 9.22 Theoretische Semi-Variogramme

$$\gamma(r_{ij}) = c_n + \sigma_0^2 f(r_{ij}) \quad (9.162)$$

$$\frac{(y(\mathbf{x}_i) - y(\mathbf{x}_j))^2}{2} = \gamma(r_{ij}) + \epsilon = c_n + \sigma_0^2 f(r_{ij}) + \epsilon \quad \forall i, j = 1 \dots n_r \quad (9.163)$$

$$\Rightarrow \frac{(y(\mathbf{x}_i) - y(\mathbf{x}_j))^2}{2} - c_n = \sigma_0^2 f(r_{ij}) \quad (9.164)$$

Mit der Substitution $a_{ij} = f(r_{ij})$ und $b_{ij} = \frac{(y(\mathbf{x}_i) - y(\mathbf{x}_j))^2}{2} - c_n$ wird dann σ_0^2 nach Gleichung 9.165 abgeschätzt (siehe auch Kapitel 9.2).

$$\widehat{\sigma}_0^2 = (\mathbf{a}' \mathbf{a})^{-1} \mathbf{a}' \mathbf{b} = \frac{\sum \mathbf{a} \mathbf{b}}{\sum \mathbf{a}^2} \quad (9.165)$$

Seien nun für eine Menge X im n_f -dimensionalen Faktorraum die Systemantworten \mathbf{y} bekannt. Soll nun die Approximation \hat{y}_0 an einer beliebigen Stelle \mathbf{x}_0 im Faktorraum ein *Best Linear Unbiased Estimator* (BLUE) sein, müssen folgende Eigenschaften erfüllt sein [90]:

- (Best) Minimaler *Mittlere Quadratischer Fehler*: $\min(\sigma_E^2 = M_{eanSquaredError})$
- (Linear) Linearkombination der bekannten Daten: $\hat{y}(\mathbf{x}_0) = \sum_{i=1}^{n_r} \lambda_i y(\mathbf{x}_i)$
- (Unbiased) Approximation ist erwartungstreue: $\sum_{i=1}^{n_r} \lambda_i = 1$

Der mittlere quadratische Fehler MSE ist definiert durch [90]:

$$MSE(\lambda) = \sigma_E^2 = E[(y_0 - \hat{y}_0)^2] = E \left[\left(y(\mathbf{x}_0) - \sum_{i=1}^{n_r} \lambda_i y(\mathbf{x}_i) \right)^2 \right] \quad (9.166)$$

Nach SOENDERGERATH [90] kann dann $MSE(\lambda)$ nach Gleichung 9.167 berechnet werden.

$$MSE(\lambda) = 2 \sum_{i=1}^{n_r} \lambda_i \gamma(|\mathbf{x}_0 - \mathbf{x}_i|) - \sum_{i=1}^{n_r} \sum_{j=1}^{n_r} \lambda_i \lambda_j \gamma(|\mathbf{x}_i - \mathbf{x}_j|) \quad (9.167)$$

Zur Minimierung des MSE unter Berücksichtigung der Nebenbedingung $\sum \lambda = 1$ wird die Gleichung mit einem Lagrange Multiplikator τ erweitert.

$$\Phi(\lambda, \tau) = 2 \sum_{i=1}^{n_r} \lambda_i \gamma(|\mathbf{x}_o - \mathbf{x}_i|) - \sum_{i=1}^{n_r} \sum_{j=1}^{n_r} \lambda_i \lambda_j \gamma(|\mathbf{x}_i - \mathbf{x}_j|) + 2\tau \left(\sum_{i=1}^{n_r} \lambda_i - 1 \right) \quad (9.168)$$

Die Ableitungen von $\Phi(\lambda, \tau)$ nach λ und τ sowie das anschließende zu Null setzen führt zu den gesuchten Kriging Gleichungen, wobei γ_{ij} eine Abkürzung für $\gamma(|\mathbf{x}_i - \mathbf{x}_j|)$ darstellt.

$$\frac{\partial \Phi}{\partial \lambda_i} = 2\gamma_{0i} - 2 \sum_{j=1}^{n_r} \lambda_j \gamma_{ij} + 2\tau = 0 \Rightarrow \gamma_{0i} = \sum_{j=1}^{n_r} \lambda_j \gamma_{ij} - \tau \quad (9.169)$$

$$\frac{\partial \Phi}{\partial \tau} = 2 \left(\sum_{i=1}^{n_r} \lambda_i - 1 \right) = 0 \Rightarrow \sum_{i=1}^{n_r} \lambda_i - 1 = 0 \quad (9.170)$$

In vereinfachter Matrixschreibweise entsprechen diese Gleichungen $\boldsymbol{\gamma} = \mathbf{K} \cdot \boldsymbol{\lambda}$.

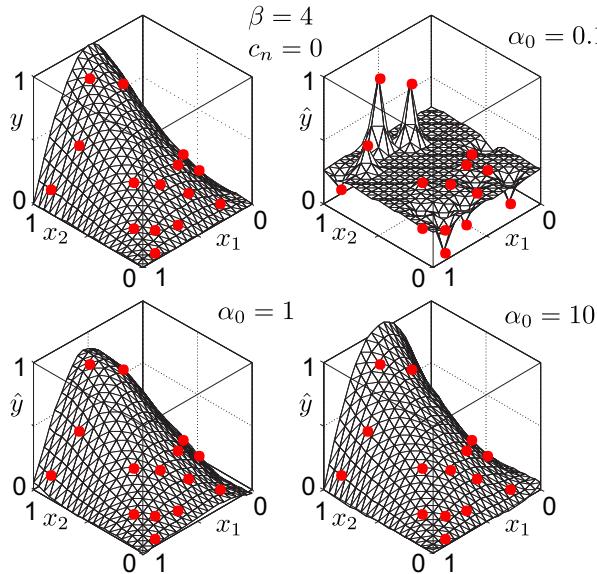
$$\begin{pmatrix} \gamma_{01} \\ \gamma_{02} \\ \vdots \\ \gamma_{0n_r} \\ 1 \end{pmatrix} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n_r} & 1 \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n_r} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{n_r 1} & \gamma_{n_r 2} & \cdots & \gamma_{n_r n_r} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{n_r} \\ -\tau \end{pmatrix} \quad (9.171)$$

Durch die Lösung $\boldsymbol{\lambda} = \mathbf{K}^{-1} \cdot \boldsymbol{\gamma}$ ergeben sich die Gewichte für die *Kriging* Interpolation, so dass sich Vorhersagen an einer Faktorkombination \mathbf{x}_0 nach folgender Linearkombination berechnen lassen:

$$\hat{y}(\mathbf{x}_0) = \sum_{i=1}^{n_r} \lambda_i y(\mathbf{x}_i) \quad (9.172)$$

Gerade der Abstand α_0 hat ähnlich wie β einen deutlichen Einfluss auf die Interpolation. Abbildung 9.23 zeigt zur Veranschaulichung den Einfluss einer α_0 Variation bei Verwendung von $c_n = 0$, $\beta = 4$ und dem theoretischen Variogramm *Gauß* auf die Vorhersage der Funktion 9.156.

Zur Reduktion der Rechenzeit ist es sinnvoll die Gleichung 9.172 aufzuteilen und einen Hilfsvektor \mathbf{h} (Gleichung 9.173) aus gegebenen Trainingsdaten X, \mathbf{y} zu erstellen. Dieser wird später zur Approximation neuer Faktorkombinationen \mathbf{x}_0 verwendet (Gleichung 9.174). Dieses Vorgehen ist ebenfalls bei den folgenden Variationen des Kriging-Verfahrens möglich, wobei \mathbf{y} mit der passenden Anzahl von Nullen erweitert wird.

Abb. 9.23 Variation von α_0 (Gauß)

$$\mathbf{h} = \mathbf{K}^{-1} \cdot \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_r} \\ 0 \end{pmatrix} \quad (9.173)$$

$$\hat{y}(\mathbf{x}_0) = \mathbf{h} \cdot \boldsymbol{\gamma} \quad (9.174)$$

9.13.1 Kriging-Regression verrauschter Daten

Bei dem bisher dargestellten Kriging Algorithmus handelt es sich um ein reines Interpolationsverfahren, wobei das Metamodell immer exakt durch die gegebenen Trainingsdaten verläuft. Sind diese jedoch mit einem Rauschsignal überlagert, ist dieses nicht sinnvoll, da lediglich das grundlegende Systemverhalten aber nicht das Rauschen abgebildet werden soll. Die Abweichung zu den verrauschten Trainingsdaten kann durch eine Diagonalmatrix Σ von Varianzen (σ_i^2) in der Matrix \mathbf{K} berücksichtigt werden $\mathbf{K}^* = \mathbf{K} - \Sigma$ [70]. Da sich die Anpassung lediglich auf die Trainingsdaten bezieht und die Nebenbedingung $\sum_{i=1}^{n_f} \lambda_i = 1$ unverändert bestehen bleibt, wird die Korrekturmatrixt Σ^* durch Nullen erweitert.

$$\Sigma^* = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \sigma_{n_r}^2 & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix} \quad (9.175)$$

Nun lassen sich die gesuchten Gewichte durch die korrigierte Gleichung 9.176 bestimmen und zur Approximation mittels Gleichung 9.172 einsetzen. Bei Verwendung normierter Faktorgrößen, wird häufig zur Vereinfachung die gleiche Varianz für alle Faktoren eingesetzt $\sigma_i = \sigma_{const}$. Mittels Kreuzvalidierung (Kapitel 9.18) kann σ_{const} robust optimiert werden.

$$\lambda = (\mathbf{K} - \Sigma^*)^{-1} \cdot \gamma \quad (9.176)$$

Zur Veranschaulichung des σ^2 Einflusses auf die Modellvorhersage wurde die Testfunktion (Gleichung 9.156) mit einem zufälligen Rauschen im Bereich $e = [-0.15, 0.15]$ überlagert und der σ^2 Wert variiert (Abbildung 9.24). Im Fall einer Interpolation $\sigma^2 = 0$ bildet das Modell das Rauschen ab, obwohl lediglich der grundlegende Zusammenhang gesucht ist (*Overfitting*, siehe Kapitel 9.16). $\sigma^2 = 0.5$ zeigt hingegen eine zu große Abweichung und Dämpfung des wirklichen zugrunde liegenden Zusammenhangs. Ein akzeptabler Kompromiss in den dargestellten Beispielen ist $\sigma^2 = 0.1$.

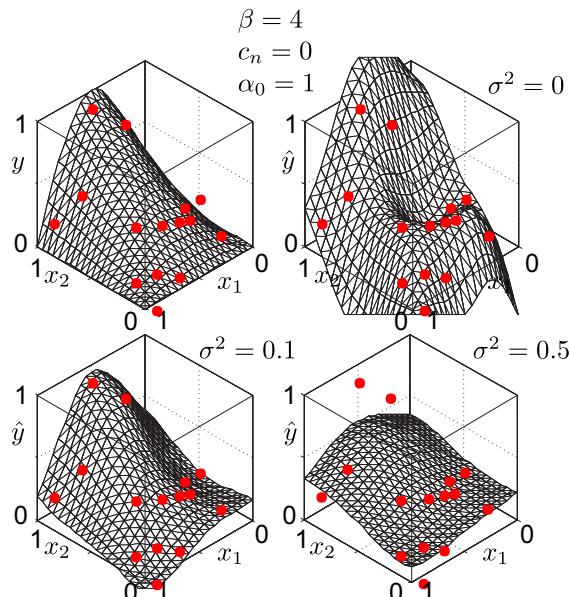


Abb. 9.24 Variation der erlaubten Abweichung σ^2

9.13.2 Universal Kriging

Während beim einfachen (*ordinary*) Kriging von einem konstanten Mittelwert aus gegangen wird, verwendet das universelle (*universal*) Kriging eine Mittelfunktionsfunktion $\mu(\mathbf{x})$ (Gleichung 9.177 und Abbildung 9.25) [67, 38, 3, 75, 55, 32, 31].

$$\mu(\mathbf{x}) = \sum_{k=0}^{n_b} \beta_k f_k(\mathbf{x}) \text{ mit } f_0 = 1 \quad (9.177)$$

β_k sind dabei Parameter die parallel zu λ_i mittels der Trainingsdaten angepasst werden und $f_k(\mathbf{x})$ sind im Voraus definierte Basisfunktionen (siehe auch Kapitel 9.3). Das Polynom $\mu(\mathbf{x})$ soll grundlegende und einfache Zusammenhänge zwischen den Faktoren und der Qualitätsgröße abbilden, wobei die verbleibenden Abweichungen der Trainingsdaten zum Polynom $\mu(\mathbf{x})$ durch Kriging erfasst werden. Das Polynom

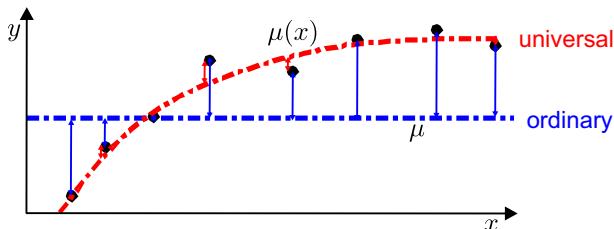


Abb. 9.25 Ordinary und Universal Kriging

$\mu(\mathbf{x})$ wird durch Erweiterung der \mathbf{K} -Matrix berücksichtigt [67, 38, 3, 75, 55, 32, 31].

$$\begin{pmatrix} \gamma_{01} \\ \gamma_{02} \\ \vdots \\ \gamma_{0n_r} \\ f_0(\mathbf{x}_0) \\ f_1(\mathbf{x}_0) \\ f_2(\mathbf{x}_0) \\ \vdots \\ f_{n_b}(\mathbf{x}_0) \end{pmatrix} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n_r} & f_0(\mathbf{x}_1) & f_1(\mathbf{x}_1) & f_2(\mathbf{x}_1) & \cdots & f_{n_b}(\mathbf{x}_1) \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n_r} & f_0(\mathbf{x}_2) & f_1(\mathbf{x}_2) & f_2(\mathbf{x}_2) & \cdots & f_{n_b}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{n_r 1} & \gamma_{n_r 2} & \cdots & \gamma_{n_r n_r} & f_0(\mathbf{x}_{n_r}) & f_1(\mathbf{x}_{n_r}) & f_2(\mathbf{x}_{n_r}) & \cdots & f_{n_b}(\mathbf{x}_{n_r}) \\ f_0(\mathbf{x}_1) & f_0(\mathbf{x}_2) & \cdots & f_0(\mathbf{x}_{n_r}) & 0 & 0 & 0 & \cdots & 0 \\ f_1(\mathbf{x}_1) & f_1(\mathbf{x}_2) & \cdots & f_1(\mathbf{x}_{n_r}) & 0 & 0 & 0 & \cdots & 0 \\ f_2(\mathbf{x}_1) & f_2(\mathbf{x}_2) & \cdots & f_2(\mathbf{x}_{n_r}) & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{n_b}(\mathbf{x}_1) & f_{n_b}(\mathbf{x}_2) & \cdots & f_{n_b}(\mathbf{x}_{n_r}) & 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{n_r} \\ \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{n_b} \end{pmatrix} \quad (9.178)$$

Zur Veranschaulichung des universal Krigings zeigt Abbildung 9.26 ein zweidimensionales Beispiel mit quadratischer Mittelfunktion $\mu(\mathbf{x})$.

Die Vorhersage einer Systemantwort y_0 für die Faktorkombination \mathbf{x}_0 lässt sich dann durch die Gleichung 9.179 bestimmen.

$$\hat{y}_0 = \sum_{k=0}^{n_b} f_k(\mathbf{x}_0) + \sum_{i=1}^{n_r} \lambda_i y_i \quad (9.179)$$

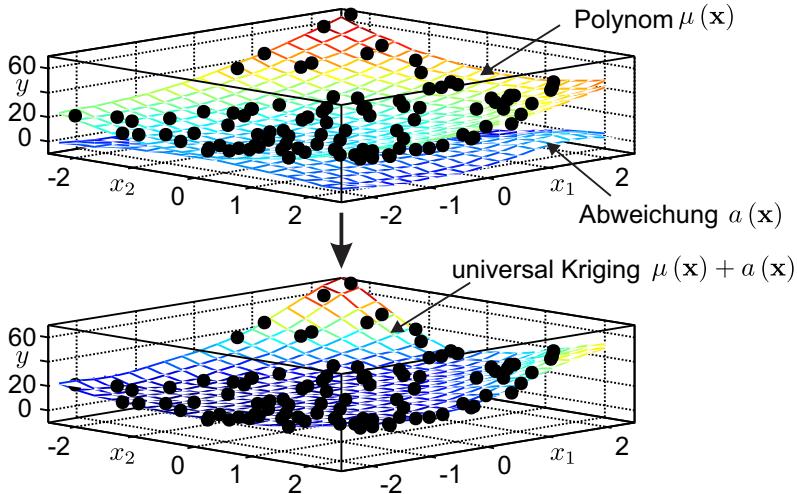


Abb. 9.26 Universal Kriging: Beispiel

9.14 Radial Basis Funktion

Die Methode Radial Basis Funktion (RBF) geht grundsätzlich davon aus, dass ein Funktionswert y_0 an der Faktorkombination \mathbf{x}_0 überwiegend von Trainingsdaten in der unmittelbaren Umgebung von \mathbf{x}_0 bestimmt wird und der Einfluss eines Datenpunkts \mathbf{x}_i mit dem radialen Abstand $r = |\mathbf{x}_0 - \mathbf{x}_i|$ zu \mathbf{x}_0 sinkt. Der Einfluss wird im Algorithmus durch eine im Vorfeld gewählte Radial-Basis-Funktion $\phi(r)$ definiert, die lediglich vom Abstand r zweier Datenpunkte abhängt. Die Approximation (Interpolation) des Funktionswerts y_0 berechnet sich durch eine gewichtete Linear-kombination der Basisfunktion $\Phi(r_i)$ für jeden Traingspunkt \mathbf{x}_i (Gleichung 9.180) [99].

$$\hat{y}(\mathbf{x}_0) = \sum_{i=1}^{n_r} w_i \phi(|\mathbf{x}_0 - \mathbf{x}_i|) \quad (9.180)$$

Abbildung 9.27 zeigt beispielhaft eine RBF Interpolation \hat{x} eines eindimensionalen Zusammenhangs auf Basis von drei bekannten Stützstellen [54]. Die benötigten Gewichte \mathbf{w} werden mittels bekannter Trainingsdaten und Lösen der Gleichung $\phi \mathbf{w} = \mathbf{y}$ bestimmt, wobei ϕ eine $n_r \times n_r$ -Matrix ist.

$$\begin{pmatrix} \phi(|\mathbf{x}_1 - \mathbf{x}_1|) & \phi(|\mathbf{x}_1 - \mathbf{x}_2|) & \cdots & \phi(|\mathbf{x}_1 - \mathbf{x}_{n_r}|) \\ \phi(|\mathbf{x}_2 - \mathbf{x}_1|) & \phi(|\mathbf{x}_2 - \mathbf{x}_2|) & \cdots & \phi(|\mathbf{x}_2 - \mathbf{x}_{n_r}|) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(|\mathbf{x}_{n_r} - \mathbf{x}_1|) & \phi(|\mathbf{x}_{n_r} - \mathbf{x}_2|) & \cdots & \phi(|\mathbf{x}_{n_r} - \mathbf{x}_{n_r}|) \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_{n_r} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_r} \end{pmatrix} \quad (9.181)$$

$$\mathbf{w} = \phi^{-1} \cdot \mathbf{y} \quad (9.182)$$

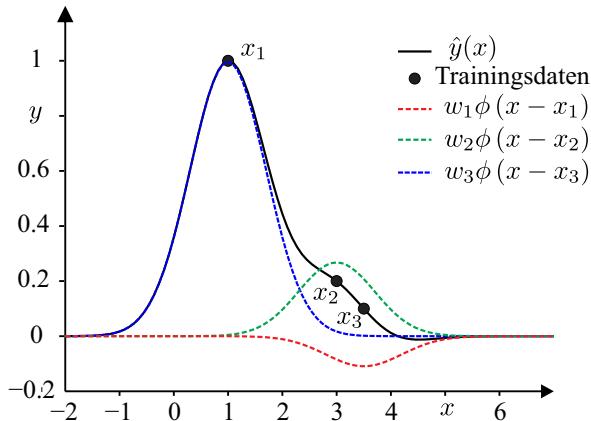


Abb. 9.27 Interpolation mit Radial Basis Funktionen: Beispiel

In der Literatur finden sich verschiedene Radial Basis Funktionen (siehe Tabelle 9.3), deren grundlegende Verläufe in Abbildung 9.28 dargestellt sind [99, 70]. In den Radial Basis Funktionen ist r_0 ein Skalierungsfaktor, der größer sein sollte als der durchschnittliche Abstand r aller Trainingspunkte aber auch so klein, dass die Form des abzubildenden Zusammenhangs aufgelöst werden kann (*Feature Size*). Der Einfluss des Skalierungsfaktors auf die Interpolation der Testfunktion 9.156 ist beispielhaft in Abbildung 9.29 dargestellt.

multiquadratisch	$\phi(r) = \sqrt{r^2 + r_0^2}$
invers multiquadratisch	$\phi(r) = \frac{1}{\sqrt{r^2 + r_0^2}}$
invers quadratisch	$\phi(r) = \frac{1}{r^2 + r_0^2}$
thin-plate spline	$\phi(r) = r^{2k} \ln \frac{r}{r_0}, k \in \mathbb{N}$
polyharmonischer spline	$\phi(r) = r^{2k-1}, k \in \mathbb{N}$
Gauß	$\phi(r) = e^{-\frac{1}{2} \frac{r^2}{r_0^2}}$

Tabelle 9.3 Radial Basis Funktionen

9.14.1 RBF-Regression verrauschter Daten

Das dargestellte RBF-Interpolationsverfahren verläuft immer genau durch die gegebenen Trainingsdaten. Liegen verrauschte Daten vor, so ist es sinnvoll eine Abweichung an den gegebenen Trainingsdaten zu erlauben. Dazu wird ein Parameter ρ eingeführt, der eine Abwägung zwischen Glättung des Funktionszusammenhangs

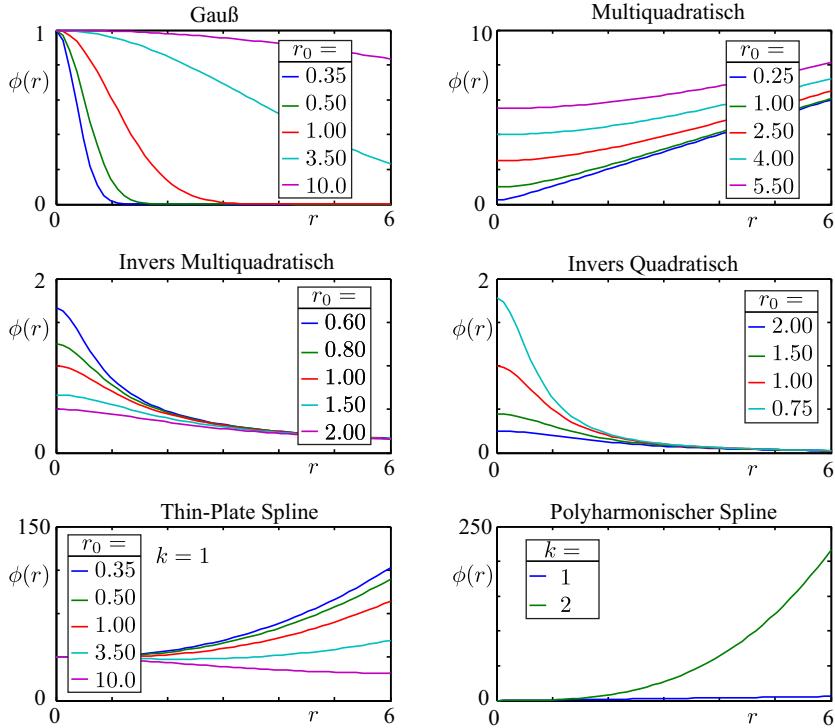


Abb. 9.28 Radial Basis Funktionen

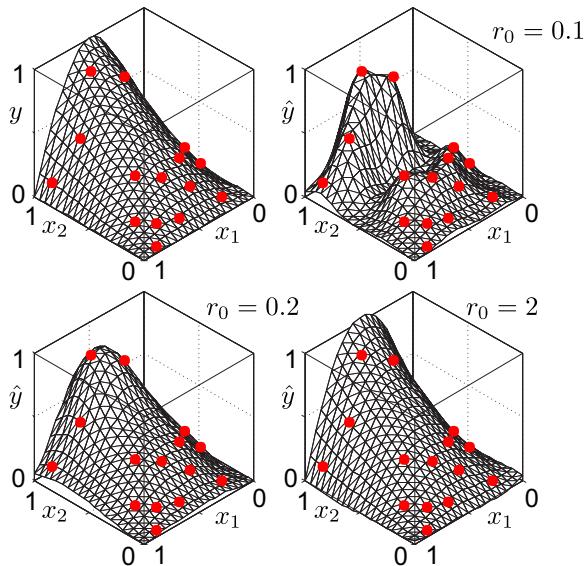


Abb. 9.29 Variation des Skalierungsfaktors r_0 (Gauß)

gegenüber Treue zu den Trainingsdaten darstellt. Für jeden Traingspunkt muss nun die allgemeine Gleichung 9.183 gelten.

$$y(\mathbf{x}_j) = \sum_{i=1}^{n_r} w_i \phi(|\mathbf{x}_j - \mathbf{x}_i|) - \rho w_j \quad (9.183)$$

Daraus ergibt sich das vollständige Gleichungssystem 9.184 ($\phi^* \cdot \mathbf{w} = \mathbf{y}$), welches zur Bestimmung der Gewichte \mathbf{w} und Vorhersage von Funktionswerten $\hat{y}(\mathbf{x}_0)$ entsprechend Gleichung 9.182 und 9.180 verwendet wird [54, 44].

$$\begin{pmatrix} \phi(|\mathbf{x}_1 - \mathbf{x}_1|) - \rho & \phi(|\mathbf{x}_1 - \mathbf{x}_2|) & \cdots & \phi(|\mathbf{x}_1 - \mathbf{x}_{n_r}|) \\ \phi(|\mathbf{x}_2 - \mathbf{x}_1|) & \phi(|\mathbf{x}_2 - \mathbf{x}_2|) - \rho & \cdots & \phi(|\mathbf{x}_2 - \mathbf{x}_{n_r}|) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(|\mathbf{x}_{n_r} - \mathbf{x}_1|) & \phi(|\mathbf{x}_{n_r} - \mathbf{x}_2|) & \cdots & \phi(|\mathbf{x}_{n_r} - \mathbf{x}_{n_r}|) - \rho \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_{n_r} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_r} \end{pmatrix} \quad (9.184)$$

Die Vorhersage des Funktionswerts y_0 an der Faktorkombination \mathbf{x}_0 wird durch Gleichung 9.180 durchgeführt. Zur Veranschaulichung des Einflusses von ρ auf die Glättung von Modellen, zeigen die Abbildungen 9.30 und 9.31 unterschiedliche Modellvorhersagen der verrauschten Testfunktionen 9.185 und 9.156.

$$y(x) = 2 \sin(x) \sin\left(\frac{x}{5}\right) \quad (9.185)$$

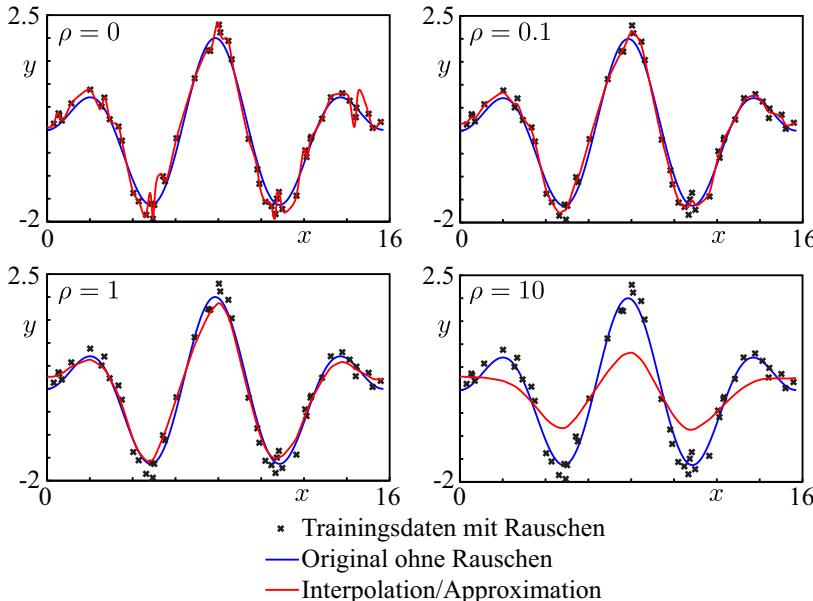


Abb. 9.30 RBF: Variation von ρ (multiquadratic, $r_0 = 0.5$)

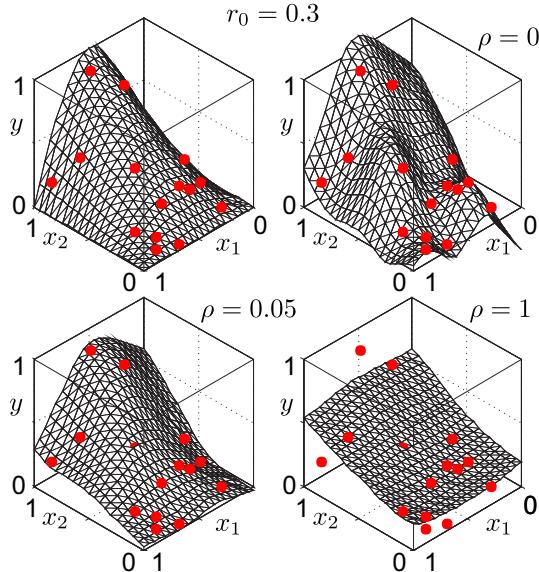


Abb. 9.31 Variation von ρ (multiquadratic, $r_0 = 0.3$)

9.14.2 Polynom Erweiterung

Einfache lineare Zusammenhänge können mit Radial Basis Funktionen (durch ihr nichtlineares Verhalten) oft schlechter abgebildet werden als durch einfache Polynome. Daher ist es vorteilhaft, kombinierte Modelle aus einem Polynom und einem RBF aufzubauen, so dass die Stärken beider Modelltypen ausgenutzt werden. Dazu wird die grundlegende RBF Gleichung 9.180 mit einem Polynom bestehend aus n_b Termen f_k erweitert.

$$\hat{y}(\mathbf{x}_0) = \sum_{k=1}^{n_b} \beta_k f_k(\mathbf{x}_0) + \sum_{i=1}^{n_r} w_i \phi(|\mathbf{x}_0 - \mathbf{x}_i|) \quad (9.186)$$

$$\text{mit } \sum_{i=1}^{n_r} w_i f_k(\mathbf{x}_i) = 0, \forall k = 1 \dots n_b \quad (9.187)$$

Durch das zusätzliche Polynom inklusive zugehöriger Nebenbedingung ergibt sich das erweiterte lineare Gleichungssystem 9.188.

$$\begin{bmatrix} \phi & F \\ F^T & 0 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{w} \\ \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} \quad (9.188)$$

ϕ ist dabei unverändert zum Fall ohne Erweiterung und F ist definiert durch:

$$F_{ij} = f_j(x_i) \text{ mit } i = 1, \dots, n_r \text{ und } j = 1, \dots, n_b \quad (9.189)$$

$$\begin{pmatrix} \phi(|\mathbf{x}_1 - \mathbf{x}_1|) & \phi(|\mathbf{x}_1 - \mathbf{x}_2|) & \cdots & \phi(|\mathbf{x}_1 - \mathbf{x}_{n_r}|) & f_1(\mathbf{x}_1) & f_2(\mathbf{x}_1) & \cdots & f_{n_b}(\mathbf{x}_1) \\ \phi(|\mathbf{x}_2 - \mathbf{x}_1|) & \phi(|\mathbf{x}_2 - \mathbf{x}_2|) & \cdots & \phi(|\mathbf{x}_2 - \mathbf{x}_{n_r}|) & f_1(\mathbf{x}_2) & f_2(\mathbf{x}_2) & \cdots & f_{n_b}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi(|\mathbf{x}_{n_r} - \mathbf{x}_1|) & \phi(|\mathbf{x}_{n_r} - \mathbf{x}_2|) & \cdots & \phi(|\mathbf{x}_{n_r} - \mathbf{x}_{n_r}|) & f_1(\mathbf{x}_{n_r}) & f_2(\mathbf{x}_{n_r}) & \cdots & f_{n_b}(\mathbf{x}_{n_r}) \\ f_1(\mathbf{x}_1) & f_1(\mathbf{x}_2) & \cdots & f_1(\mathbf{x}_{n_r}) & 0 & 0 & \cdots & 0 \\ f_2(\mathbf{x}_1) & f_2(\mathbf{x}_2) & \cdots & f_2(\mathbf{x}_{n_r}) & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{n_b}(\mathbf{x}_1) & f_{n_b}(\mathbf{x}_2) & \cdots & f_{n_b}(\mathbf{x}_{n_r}) & 0 & 0 & \cdots & 0 \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_{n_r} \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{n_b} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_r} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad (9.190)$$

Mit $\phi^* = \begin{bmatrix} \phi & F \\ F^T & 0 \end{bmatrix}$, $\mathbf{w}^* = \begin{bmatrix} \mathbf{w} \\ \boldsymbol{\beta} \end{bmatrix}$ und $\mathbf{y}^* = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix}$ können alle benötigten Parameter w_i und β_i entsprechend Gleichung 9.182 ermittelt werden. Zur Veranschaulichung der Approximation mittels RBF plus Polynom zeigt Abbildung 9.32 ein einfaches zweidimensionales Beispiel.

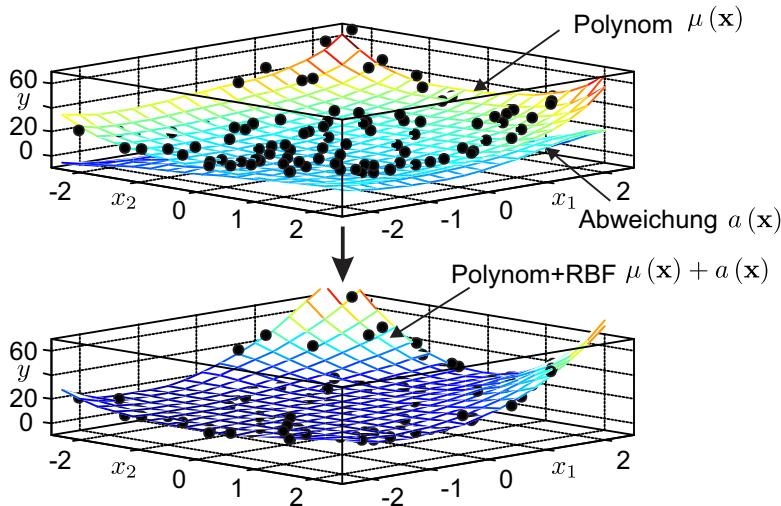


Abb. 9.32 RBF plus Polynom: Beispiel

9.14.3 Reduzierung der Zentren

Die normale RBF-Interpolation nutzt alle Trainingspunkte zur Vorhersage, wodurch sich ein komplexes Modell ergibt. Gerade bei einer großen Anzahl von Trainingsdaten steigt die Rechenzeit dadurch deutlich an. Existieren jedoch Trainingsdaten, die nur einen geringen Einfluss auf die Güte des Modells aufweisen, können diese oh-

ne signifikante Reduzierung der Modellgüte entfernt werden (Abbildung 9.33) [11]. Carr schlägt dazu einen Greedy Algorithmus 5 vor, welcher iterativ einzelne Punkte dem Modell zufügt bis eine maximale Abweichung δ an allen Trainingsdaten unterschritten wird [11].

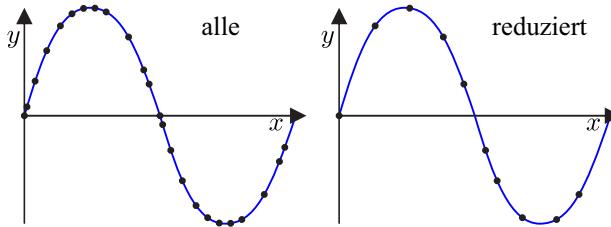


Abb. 9.33 Reduzierung der Zentren

- 1 $\varepsilon_{max} = \infty$
- 2 Wähle eine kleine Teilmenge X_0 aus den gegebenen Trainingsdaten X
- 3 Berechne RBF für die Teilmenge X_0
- 4 Approximiere \hat{y} aller Trainingsdaten X mit erzeugtem RBF-Modell
- 5 Berechne Residuen der Trainingsdaten: $\varepsilon = y - \hat{y}$
- 6 $\varepsilon_{max} = \max(|\varepsilon|)$
- 7 solange $\varepsilon_{max} > \delta$ tue
 - 8 Füge Trainingspunkt i mit größtem $|\varepsilon_i|$ zu Teilmenge hinzu: $X_j = X_{j-1} \cup x_i$
 - 9 Berechne RBF für die Teilmenge X_j
- 10 Approximiere \hat{y} aller Trainingsdaten X mit erzeugtem RBF-Modell
- 11 Berechne Residuen der Trainingsdaten: $\varepsilon = y - \hat{y}$
- 12 $\varepsilon_{max} = \max(|\varepsilon|)$
- 13 Ende

Algorithmus 5 : Greedy zur Reduzierung der RBF Zentren

9.15 Gauß Prozess Modelle

Der Gauß Prozess ist eine Verallgemeinerung der ein- und mehrdimensionalen Gaußverteilung (Abbildung 9.34) für unendlich viele Variablen. Eine eindimensionale Gaußverteilung der Datenmenge X wird durch zwei Parameter (Mittelwert μ und Standardabweichung σ) vollständig definiert.

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \text{mit} \quad p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (9.191)$$

$p(x)$ beschreibt dabei die Wahrscheinlichkeit (Wahrscheinlichkeitsdichte) für das Auftreten eines Punkts x in der Datenmenge X . Im n_f dimensionalen Fall wird die

Gaußverteilung durch einen Mittelwertsvektor μ der Länge n_f und einer Kovarianzmatrix Σ der Größe $n_f \times n_f$ definiert, wobei $|\Sigma|$ der Determinante von Σ entspricht.

$$\mathbf{X} \sim \mathcal{N}_{n_f}(\mu, \Sigma) \quad \text{mit} \quad p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^{n_f} |\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)' \Sigma^{-1} (\mathbf{x}-\mu)} \quad (9.192)$$

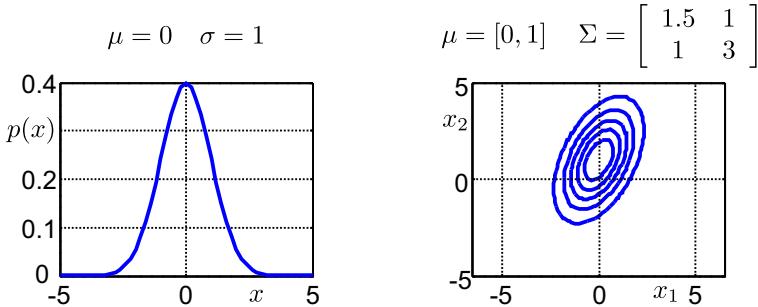


Abb. 9.34 Wahrscheinlichkeitsdichte der ein- und zweidimensionalen Gaußverteilung

Eine Erweiterung für unendlich viele Variablen wird durch die Substitution des endlichen Mittelvektors und der Kovarianzmatrix durch die Funktionen $m(\mathbf{x})$ und $k(\mathbf{x}, \mathbf{x}^*)$ erreicht. $m(\mathbf{x})$ beschreibt dabei den Erwartungswert einer beliebigen Funktion $f(\mathbf{x})$ und $k(\mathbf{x}, \mathbf{x}^*)$ die Kovarianz zweier Punkte im Faktorraum.

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \quad \text{und} \quad k(\mathbf{x}, \mathbf{x}^*) = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}^*) - m(\mathbf{x}^*))] \quad (9.193)$$

Der Gauß Prozess wird somit durch Gleichung 9.194 vollständig beschrieben.

$$f(\mathbf{x}) = \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}^*)) \quad (9.194)$$

Im Gegensatz zu unendlich großen Vektoren oder Matrizen können die Funktionen $m(\mathbf{x})$ und $k(\mathbf{x}, \mathbf{x}^*)$ (unabhängig von der genauen Form) explizit angegeben werden.

9.15.1 Bedingte Verteilung und Randverteilung

In realen Anwendungen ist meist eine endliche Informationsmenge vorhanden, so dass eine sinnvolle Analyse nur in einer endlichen Anzahl von Dimensionen und Datenpunkten möglich ist. Um dies zu erreichen wird auf die Theorie der bedingten Verteilung (*conditional distribution*) sowie der Randverteilung (*marginal distribution*) zurückgegriffen, die hier zur Veranschaulichung für zwei normalverteilte und unabhängige Variablen x_1, x_2 im zweidimensionalen Raum (*joint gaussian*) betrachtet werden [74] (Abbildung 9.35 und 9.36).

$$X \sim \mathcal{N} \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma'_{12} & \Sigma_{22} \end{bmatrix} \right) \quad (9.195)$$

Die *bedingte Verteilung* beschreibt die Verteilung der Variable x_1 für den Fall, dass Variable x_2 einen vorgegebenen Wert ($x_2 = a$) aufweist ($x_1|x_2 = a$) [74] (Abbildung 9.35).

$$x_1|x_2 = a \sim \mathcal{N} (\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12}) \quad (9.196)$$

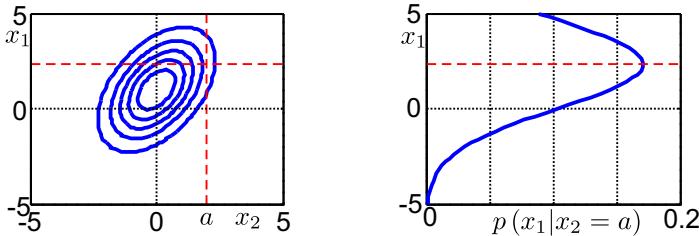


Abb. 9.35 Bedingte Verteilung

Im Gegensatz dazu beschreibt die *Randverteilung* oder Marginalverteilung die Wahrscheinlichkeitsverteilung einer Variablenuntergruppe beziehungsweise im zweidimensionalen Fall einer einzelnen Variable x_1 oder x_2 (Abbildung 9.36).

$$x_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}) \quad (9.197)$$

Die *bedingte Verteilung* sowie die *Randverteilung* einer multivariaten Gaußverteilung sind nach Gleichung 9.196 und 9.197 ebenfalls wieder Gaußverteilungen, was für die spätere Modellbildung eine entscheidende Rolle spielt. Die Eigenschaften der Randverteilung ist lediglich abhängig von der jeweils betrachteten Untergruppe. Nun können die unendlich vielen Einträge eines Gauß Prozesses in zwei unabhängige Gruppen **a** und **b** aufgeteilt werden. In der ersten Gruppe **a** werden alle Dimensionen, die von Interesse sind³ und in der zweiten Gruppe **b** werden alle restlichen (unendlich viele) Dimensionen zusammengefasst [74].

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma'_{ab} & \Sigma_{bb} \end{bmatrix} \right) \quad (9.198)$$

Im Folgenden kann dann lediglich die Randverteilung **a** betrachtet werden, so dass der Gauß Prozess mit unendlich vielen Variablen auf einfache Weise in eine handhabbare endliche Form gebracht wird.

³ was das auch immer im Einzelfall bedeutet

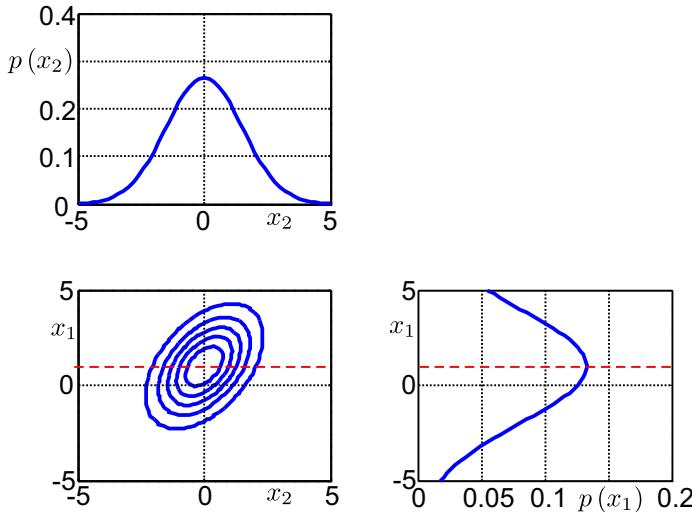


Abb. 9.36 Randverteilung

9.15.2 Vorhersagen mittels bedingter Verteilung

Seien nun Funktionswerte \mathbf{y} an den Punkten X bekannt und die Funktionswerte \mathbf{y}^* an den Stellen X^* gesucht. Für die gemeinsame Verteilung gilt:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}^* \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K(X, X) & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{bmatrix} \right) \quad (9.199)$$

Die Approximation von \mathbf{y}^* bei den gegebenen Randbedingungen X^*, X, \mathbf{y} ist ebenfalls als Normalverteilung darstellbar [74].

$$\mathbf{y}^* | X^*, X, \mathbf{y} \sim \mathcal{N} \left(K(X^*, X) K(X, X)^{-1} \mathbf{y}, K(X^*, X^*) - K(X^*, X) K(X, X)^{-1} K(X, X^*) \right) \quad (9.200)$$

Sind die bekannten Daten \mathbf{y} mit einem Rauschen $\mathcal{N}(0, \sigma_n)$ überlagert, so kann dieses beim Kovarianzterm $K(X, X)$ berücksichtigt werden, wodurch sich nur eine leicht veränderte Form der Verteilung ergibt [74].

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}^* \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K(X, X) + \sigma_n^2 I & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{bmatrix} \right) \quad (9.201)$$

$$\begin{aligned} \mathbf{y}^* | X^*, X, \mathbf{y} \sim & \mathcal{N} \left(K(X^*, X) [K(X, X) + \sigma_n^2 I]^{-1} \mathbf{y}, \right. \\ & \left. K(X^*, X^*) - K(X^*, X) [K(X, X) + \sigma_n^2 I]^{-1} K(X, X^*) \right) \quad (9.202) \end{aligned}$$

Die Vorhersage unbekannter Datenpunkte wird grundsätzlich wie bei anderen Modellansätzen als Linearkombination verschiedener Basisfunktionen $k(\mathbf{x}_i, \mathbf{x}^*)$ be-

trachtet. $k(\mathbf{x}_i, \mathbf{x}^*)$ ist dabei eine Funktion, die vom gesuchten Punkt \mathbf{x}^* und jeweils einem bekannten zweiten Punkt \mathbf{x}_i abhängt [74].

$$\hat{\mathbf{y}}^* = \sum_{i=1}^{n_r} \beta_i k(\mathbf{x}_i, \mathbf{x}^*) \quad \text{mit } \boldsymbol{\beta} = [K(X, X) + \sigma_n^2 I]^{-1} \mathbf{y} \quad (9.203)$$

Die Vorhersagegenauigkeit oder Varianz unbekannter Datenpunkte setzt sich aus einer vorher (Prior) gewählten Grund-Kovarianz $K(\mathbf{x}^*, \mathbf{x}^*)$ abzüglich eines positiven Terms, der lediglich von der Verteilung bekannter Datenpunkte im Faktorraum abhängt, zusammen [74]. Die Systemantworten \mathbf{y} der bekannten Datenpunkte besitzen dabei keinen Einfluss auf die Varianz. Bei der Voraussetzung einer Normalverteilung wird die logarithmische marginale Wahrscheinlichkeit, welche ein Maß für die Güte des Modells darstellt⁴, wie in Gleichung 9.204 berechnet [74].

$$\begin{aligned} \log p(\mathbf{y}|X) = & -\frac{1}{2} \mathbf{y}' [K(X, X) + \sigma_n^2 I]^{-1} \mathbf{y} \\ & -\frac{1}{2} \log (\det(K(X, X) + \sigma_n^2 I)) - \frac{n_r}{2} \log(2\pi) \end{aligned} \quad (9.204)$$

$\frac{n_r}{2} \log(2\pi)$ ist dabei lediglich ein konstanter Normierungsfaktor.

9.15.3 Betrachtung im Funktionsraum

Ist für einen Gauß Prozess $f(\mathbf{x}) = \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}^*))$ die Mittelfunktion $m(\mathbf{x})$ und Kovarianzfunktion $k(\mathbf{x}, \mathbf{x}^*)$ gegeben (Prior), kann die Wahrscheinlichkeitsverteilung von $f(\mathbf{x})$ im Funktionsraum angegeben werden. Wird beispielsweise $m(x) = 0$, $k(x, x^*) = e^{0.5|x-x^*|^2}$ und $\sigma_n = 0$ im eindimensionalen Faktorraum gewählt, so liegt die gesuchte Funktion $f(x)$, solange kein weiterer Datenpunkt bekannt ist, mit $\sim 95\%$ Wahrscheinlichkeit im Bereich $[m(x) \pm 2k(x, x)] = [-2, 2]$, wie es in Abbildung 9.37 (oben links) dargestellt ist. Der genaue Verlauf von $f(x)$ liegt mit zugehöriger Wahrscheinlichkeit irgendwo in diesem Bereich (mit 5% Wahrscheinlichkeit außerhalb). Bei den vorgegebenen Prior-Informationen ist der wahrscheinlichste Verlauf $f(x) = m(x) = 0$. Wenn nun ein erster Datenpunkt bekannt ist (Abbildung 9.37, oben rechts), wird die wahrscheinliche Lage der Funktion $f(x)$ in der Nähe dieses Punktes beeinflusst.

⁴ größer ist besser

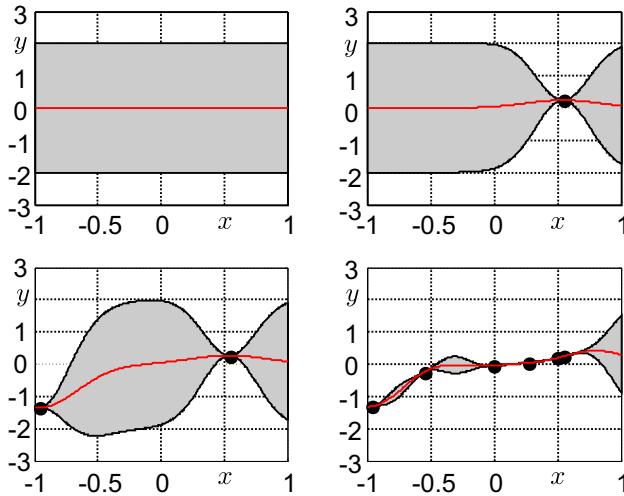


Abb. 9.37 Beispiel eines eindimensionalen Gauß Prozess Modells

Die Stärke der Beeinflussung ist durch die gegebene Kovarianzfunktion definiert, wobei der Einfluss bei steigender Entfernung abnimmt. Weiter Datenpunkte erhöhen die Informationsdichte und Vorhersagegenauigkeit für die gesuchte Funktion $f(x)$. RASMUSSEN zeigt eine einfache Berechnungsvorschrift auf Basis der Cholesky Zerlegung zur Bestimmung der Approximation \hat{y}^* an einer Stelle \mathbf{x}^* inklusive Vorhersage der Varianz $\mathbb{V}[y^*]$ und der zugehörigen logarithmischen marginalen Wahrscheinlichkeit $\log p(\mathbf{y}|X)$ (Algorithmus 6) [74].

- 1 $L = \text{cholesky}(k(X, X) + \sigma_n^2)$
- 2 $\alpha = L' \backslash (L \backslash \mathbf{y})$
- 3 $\mathbf{k}_* = k(X, \mathbf{x}^*)$
- 4 $\hat{y}^* = \mathbf{k}'_* \alpha$
- 5 $\mathbf{v} = L \backslash \mathbf{k}_*$
- 6 $\mathbb{V}[y^*] = k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{v}' \mathbf{v}$
- 7 $\log p(\mathbf{y}|X) = -\frac{1}{2} \mathbf{y}' \alpha - \sum_{i=1}^{n_r} \log(L_{ii}) - \frac{n_r}{2} \log 2\pi$

Algorithmus 6 : Gauß Prozess Approximation

9.15.4 Kovarianzfunktionen

Für die Erstellung eines Gauß Prozess Modells müssen grundsätzlich unendlich viele Funktionen mittels Mittelvektoren und Kovarianzmatrizen dargestellt werden. Beispielhaft wird hier die Abbildung von unendlich vielen linearen sowie Hügel-

funktionen betrachtet.

Eine Schar von linearen Funktionen im eindimensionalem Faktorraum lässt sich durch die Gleichung $f(x) = ax + b$ darstellen. Die Parameter a und b weisen dabei eine voneinander unabhängige Normalverteilung um Null auf [72].

$$f(x) = ax + b \text{ mit } a \sim \mathcal{N}(0, \alpha) \text{ und } b \sim \mathcal{N}(0, \beta) \quad (9.205)$$

Der erwartete Mittelwert sowie die Kovarianz der Funktionsschar lässt sich mittels Integration über a und b bestimmen [72]:

$$\mathbb{E}[f(x)] = \mu(x) = \iint f(x)p(a)p(b)dadb = \int axp(a)da + \int bp(b)db = 0 \quad (9.206)$$

$$\begin{aligned} \mathbb{E}[(f(x) - \mu(x))(f(x^*) - \mu(x^*))] &= \mathbb{E}[(f(x) - 0)(f(x^*) - 0)] = k(x, x^*) \quad (9.207) \\ &= \iint (ax + b)(ax^* + b)p(a)p(b)dadb \\ &= \int a^2 xx^* p(a)da + \int b^2 p(b)db + (x + x^*) \int abp(a)p(b)dadb \\ &= \alpha xx^* + \beta \end{aligned}$$

Im Gegensatz zu stationären Kovarianzfunktionen, die lediglich vom Abstand zweier Punkte abhängen und nicht von der Lage im Faktorraum, ist die Kovarianzmatrix der linearen Funktionen nur von der Lage abhängig und spiegelt dadurch den Einfluss der Entfernung linearer Funktionen auf den Funktionswert wieder.

Betrachten wir im Gegensatz dazu eine Funktion, die aus einer Linarkombination unendlich vieler Hügelfunktionen $f_p(x) = e^{-(x-x_p)^2}$ besteht. x_p beschreibt dabei die Position des Extremums der Funktion. Abbildung 9.38 zeigt ein Beispiel mit drei im Bereich $[0, 1]$ verteilten Hügeln und die daraus resultierende Linierkombination (rot). Im allgemeinen Fall kann bei der Linearkombination von $n_p + 1$ gleichverteilten Stützstellen im Bereich $[0, 1]$ ausgegangen werden.

$$\begin{aligned} f(x) &= \lim_{n_p \rightarrow \infty} \frac{1}{n_p + 1} \sum_{i=0}^{n_p} \alpha_i e^{-\left(x - \frac{i}{n_p}\right)^2} \quad \text{mit } \alpha_i \sim \mathcal{N}(0, 1), \forall i = 0 \dots n_p \quad (9.208) \\ &= \int_{-\infty}^{\infty} \alpha(u) e^{-(x-u)^2} du \quad \text{mit } \alpha(u) \sim \mathcal{N}(0, 1) \end{aligned}$$

Mittelwert- als auch Kovarianzfunktion ergeben sich nach RASMUSSEN zu [72]:

$$\mathbb{E}[f(x)] = \mu(x) = \int e^{-(x-u)^2} \int \alpha p(\alpha) d\alpha du = 0 \quad (9.209)$$

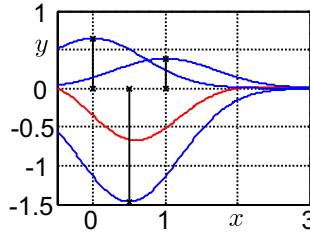


Abb. 9.38 Linearkombination dreier Hügelfunktionen

$$\begin{aligned} \mathbb{E}[(f(x) - \mu(x))(f(x^*) - \mu(x^*))] &= \mathbb{E}[(f(x) - 0)(f(x^*) - 0)] = k(x, x^*) \quad (9.210) \\ &= \int e^{-(x-u)^2 - (x^*-u)^2} du \\ &= \int e^{-2\left(u-\frac{x+x^*}{2}\right)^2 + \frac{(x+x^*)^2}{2} - x^2 - x^{*2}} du \propto e^{-\frac{(x-x^*)^2}{2}} \end{aligned}$$

Die quadratisch-exponentielle Kovarianzfunktion repräsentiert somit eine unendliche Anzahl von gaußähnlichen Funktionen (Hügelfunktionen), die im gesamten normierten Definitionsbereich $[0, 1]$ existieren. Weitere beispielhafte Kovarianzfunktionen sind in Tabelle 9.4 dargestellt, wobei diese zur Vereinfachung zu den Originalquellen [73, 74, 111] leicht verändert wurden. Simplere Kovarianzfunktionen wie zum Beispiel die lineare Kovarianzfunktion $k(\mathbf{x}, \mathbf{x}^*) = \mathbf{x}' \mathbf{x}^*$ können durch geschickte Wahl der Hyperparameter erzeugt werden (hier Polynomfunktion mit $c = 0$ und $p = 1$). Δx^2 steht in den Gleichungen für den quadratischen euklidischen Abstand zweier betrachteter Punkte \mathbf{x} und \mathbf{x}^* .

$$\Delta x^2 = (\mathbf{x} - \mathbf{x}^*)'(\mathbf{x} - \mathbf{x}^*) \quad (9.211)$$

Alle anderen Parameter dienen zur Anpassung der Funktionen an die Modellierungsaufgaben und müssen in einem späteren Schritt bestimmt werden. Jede der gezeigten Kovarianzfunktionen kann grundsätzlich mit einem konstanten Faktor multipliziert (b) oder beaufschlagt (a) werden, was zu zwei zusätzlichen Parametern führt.

$$k_{neu}(\mathbf{x}, \mathbf{x}^*) = a + b k(\mathbf{x}, \mathbf{x}^*) \quad (9.212)$$

Komplexere Kovarianzfunktionen werden durch Multiplikation oder Addition mehrere Kovarianzfunktionen erzeugt, wodurch dem Anwender eine Vielzahl von Möglichkeiten zur Verfügung steht.

$$k_{neu}(\mathbf{x}, \mathbf{x}^*) = \sum_i k_i(\mathbf{x}, \mathbf{x}^*) \text{ oder } k_{neu}(\mathbf{x}, \mathbf{x}^*) = \prod_i k_i(\mathbf{x}, \mathbf{x}^*) \quad (9.213)$$

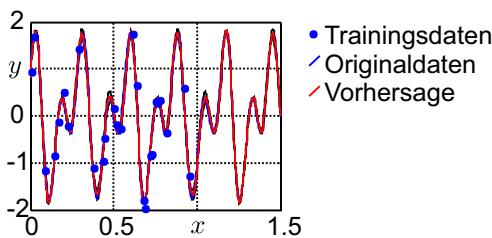
Zusätzliche Kovarianzfunktionen sowie Erweiterungen, zum Beispiel für periodische Systemcharakteristiken, finden sich bei RASMUSSEN [73].

Polynom	$(\mathbf{x}' \mathbf{x}^* + c)^p$
Quadratisch Exponentiell	$e^{-\frac{1}{2} \frac{\Delta x^2}{l^2}}$
Quadratisch Rational	$\left(1 + \frac{1}{2p} \frac{\Delta x^2}{l^2}\right)^{-p}$
Matern 1,2,3	$c = 1 : e^{-d_l}$ $c = 3 : (1 + d_l) e^{-d_l}$ $c = 5 : \left(1 + d_l + \frac{d_l^2}{3}\right) e^{-d_l}$ mit $d_l = \sqrt{\frac{c}{l^2} \Delta x^2}$
Neuronales Netzwerk	$\sin^{-1} \left(\frac{1 + \mathbf{x}' \mathbf{x}^*}{\sqrt{l + 1 + \mathbf{x}' \mathbf{x}} \sqrt{l + 1 + \mathbf{x}^* \mathbf{x}'}} \right)$
Periodisch	$e^{-\frac{2}{l^2} \sin^2 \left(\frac{\pi}{p} \ \mathbf{x} - \mathbf{x}^*\ \right)}$
Stückweise Polynom	$\max \left(1 - \frac{\Delta x}{l}, 0\right)^{j+v} f_v(r)$ mit $v = 0 : f_v(r) = 1$ $v = 1 : f_v(r) = 1 + (j+1)r$ $v = 2 : f_v(r) = 1 + (j+2)r + \frac{j^2 + 4j + 3}{3} r^2$ $v = 3 : f_v(r) = 1 + (j+3)r + \frac{6j^2 + 36j + 45}{15} r^2 + \frac{j^3 + 9j^2 + 23j + 15}{15} r^3$ $j = \left\lfloor \frac{n_f}{2} \right\rfloor + v + 1$ $r = \frac{\ \mathbf{x} - \mathbf{x}^*\ }{l}$

Tabelle 9.4 Kovarianzfunktionen

Beispiele

Periodische Zusammenhänge im eindimensionalen Faktorraum können mit der Kovarianzfunktion $k(x, x^*) = e^{-\frac{2}{l^2} \sin^2 \left(\frac{\pi}{p} \Delta x \right)}$ abgebildet werden. Abbildung 9.39 zeigt dazu die Vorhersage und Extrapolation der Funktion $y = \cos(22x) + \sin(44x)$, wobei 26 leicht verrauschte Trainingspunkte verwendet wurden. Die Extrapolation eines periodischen Signals ist mit diesem Verfahren, wie im Beispiel zu sehen, möglich.

**Abb. 9.39** Periodische Funktion mit verrauschten Stützstellen und Extrapolation

Eine Kombination verschiedener Kovarianzfunktionen ist sinnvoll, wenn unterschiedliche charakteristische Zusammenhänge in den Daten abgebildet werden müssen.

sen. RASMUSSEN [72] zeigt dazu ein Beispiel auf Basis des atmosphärischen CO_2 -Anstiegs zwischen 1958 und 2008 (Abbildung 9.40, blau) [12]. Die erste Kovarianzfunktion K_1 dient zur Abbildung eines gleichmäßigen Anstiegs des CO_2 -Anteils über die Monate. K_2 bildet die jährlichen (periodischen) Schwankungen ab, wobei der exponentielle Zusatzterm eine abklingende Abweichung vom exakten periodischen Verhalten ermöglicht. Zur Darstellung von Ungleichmäßigkeiten über die Jahre wird ein quadratischer Term K_3 verwendet. Die letzten zwei Terme (K_4 und K_5) dienen zur Darstellung von Messrauschen. Die vollständige Kovarianzfunktion wird durch die Summe aller Terme K_i erzeugt.

$$K_1 = p_1^2 e^{-\frac{d^2}{2p_2^2}} \quad (9.214)$$

$$K_2 = p_3^2 e^{-\frac{d^2}{2p_4^2}} e^{-\frac{2\sin^2(\pi \frac{d}{p_{12}})}{p_5^2}} \quad (9.215)$$

$$K_3 = p_6^2 \left(1 + \frac{d^2}{2p_8 p_7^2}\right)^{-p_8} \quad (9.216)$$

$$K_4 = p_9^2 e^{-\frac{d^2}{2p_{10}^2}} \quad (9.217)$$

$$K_5 = p_{11}^2 \delta \quad (9.218)$$

$$K = K_1 + K_2 + K_3 + K_4 + K_5 = \sum_{i=1}^5 K_i \quad (9.219)$$

Abbildung 9.40 zeigt die Vorhersagegüte des angepassten Gauß Prozess Models. Im Bereich der bekannten Messdaten liegt eine gute Übereinstimmung mit geringer Varianz vor. Die Extrapolation zeigt eine sinnvolle Weiterführung des ansteigenden Trends mit überlagerter periodischer Schwingung, wobei die vorhergesagte Toleranz kontinuierlich mit Entfernung zu den Messdaten ansteigt.

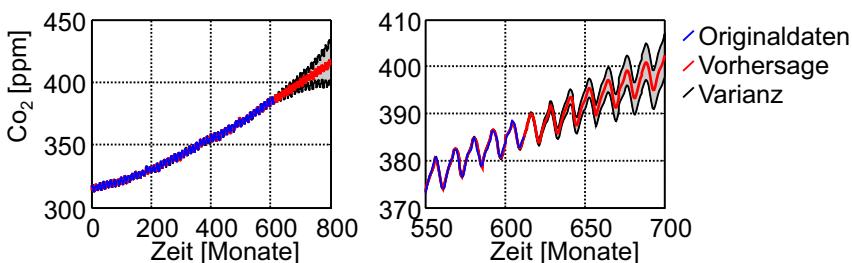


Abb. 9.40 Vorhersage eines periodischen CO_2 Anstiegs

Normierung und Skalierung

Durch Einführung eines Skalierungsfaktors $\frac{1}{\lambda_i}, i = 1 \dots n_f$ für jede Dimension kann der Einfluss der unterschiedlichen Dimensionen/Faktoren auf das Modell beeinflusst werden. Zur Skalierung wird dabei die Matrix X mit der Diagonalmatrix $diag(\lambda^{-1})$ multipliziert ($X_{neu} = X diag(\lambda^{-1})$). Die Skalierungsfaktoren werden entsprechend aller anderen Parameter an das Regressionsproblem angepasst. In der Literatur wird in diesem Zusammenhang auch von „automatic relevance determination“ (ARD) gesprochen [68]. Eine Vergrößerung von λ_i ist dabei gleichbedeutend mit der Reduzierung des Einflusses beziehungsweise der Bedeutung eines Faktors, wodurch nach WILLIAMS eine (automatische) Faktorauswahl ermöglicht wird [110]. Bei der Betrachtung vieler Faktoren ist es manchmal sinnvoll eine Faktorenanalyse durchzuführen, wobei die Faktoren mittels einer Linearkombination auf neue Hauptachsen im Faktorraum projiziert werden (siehe Kapitel 12). Werden nun nur die wichtigsten Hauptachsen zur Analyse verwendet ist eine deutliche Reduktion der Dimensionenzahl möglich. Dazu wird die Matrix X zusätzlich mit einer Lambda Matrix $\Lambda_{n_R \times k}$ mit ($k < n_R$) multipliziert [74].

9.15.5 Anpassung der Hyperparameter

Zur Anpassung eines Gauß Prozess Modells an Trainingsdaten sind neben der Auswahl geeigneter Kovarianzfunktionen K_i ebenfalls alle freien Parameter der Kovarianzfunktionen sowie weitere Adoptionsparameter (zum Beispiel λ) zu bestimmen. Alle zu ermittelnden Parameter werden unter der Bezeichnung Hyperparameter Θ zusammengefasst. RASMUSSEN und WILLIAMS [74] zeigen, dass ein guter Kompromiss zwischen Modellanpassung an die Trainingsdaten und Komplexität des Modells erzielt wird, wenn zur Bestimmung der Hyperparameter die aus Gleichung 9.204 bekannte Randverteilung zur Beurteilung eingesetzt wird.

$$\log p(\mathbf{y}|X, \Theta) = -\frac{1}{2}\mathbf{y}'K_y^{-1}\mathbf{y} - \frac{1}{2}\log(\det(K_y)) - \frac{n_r}{2}\log(2\pi) \quad (9.220)$$

$$\text{mit } K_y = K(X, X, \Theta) + \sigma_n^2 I \quad (9.221)$$

K_y bezieht sich dabei auf die eventuell verrauschten \mathbf{y} Werte und $K(X, X, \Theta)$ nur auf die gesuchte rauschfreie Grundfunktion. Lediglich der erste Term der Randverteilung $-\frac{1}{2}\mathbf{y}'K_y^{-1}\mathbf{y}$ ist von den Systemvariablen \mathbf{y} abhängig und beschreibt die Anpassung des Models an die Trainingsdaten. Der zweite Term ist ein Maß für die Komplexität des Modells und ist lediglich von der Lage der Trainingsdaten X im Faktorraum und Θ abhängig. Der letzte Term ist ein konstanter Normierungsterm der von der Messdatenanzahl abhängt. Die Hyperparameter Θ , welche die Randverteilung maximieren, liefern den „besten“ Kompromiss zwischen der Modellanpassung an die Trainingsdaten und Komplexität auf Basis der bekannten Trainingsdaten. Das Maximum der Randverteilung wird zum Beispiel mittels eines Gradienten-

verfahrens ermittelt, wobei die Möglichkeit besteht, dass ein lokales Optimum der Hyperparameter gefunden wird. Aus diesem Grund ist es ratsam mehrere Optimierungen von unterschiedlichen Startwerten durchzuführen und im Anschluss die Parameter mit dem besten Qualitätsergebnis zu verwenden. Zur stabilen und schnellen Optimierung mittels Gradientenverfahren ist der Einsatz von analytischen partiellen Ableitungen der Randverteilung nach den einzelnen Hyperparametern hilfreich, welche nach RASMUSSEN wie folgt berechnet werden [74, 72].

$$\begin{aligned}\frac{\partial}{\partial \Theta_i} \log p(\mathbf{y}|X, \Theta) &= \frac{1}{2} \mathbf{y}' K^{-1} \frac{\partial K}{\partial \Theta_i} K^{-1} \mathbf{y} - \frac{1}{2} \text{tr} \left(K^{-1} \frac{\partial K}{\partial \Theta_i} \right) \\ &= \frac{1}{2} \text{tr} \left((\alpha \alpha' - K^{-1}) \frac{\partial K}{\partial \Theta_i} \right)\end{aligned}\quad (9.222)$$

mit $\alpha = K^{-1} \mathbf{y}$

tr steht dabei für die Spur (*trace*) einer Matrix. Ableitungen verschiedener Kovarianzfunktionen finden sich zum Beispiel im Quelltext zur MatlabToolbox GPML (Gaussian Process for Machine Learning) von RASMUSSEN [73]. Zur einfacheren Ableitung der hier exemplarisch dargestellten Kovarianzfunktion $K = \sigma_n^2 e^{-\frac{1}{2} \frac{d^2}{l^2}}$ wird diese leicht umformuliert (d : Abstand zweier Datenpunkte).

$$K(d, \Theta_1, \Theta_2) = e^{2\Theta_2} e^{-\frac{1}{2} \frac{d^2}{e^{2\Theta_1}}} \quad (9.223)$$

Die partiellen Ableitungen lassen sich dadurch wie in Gleichungen 9.224 und 9.225 berechnen und sind lediglich von Komponenten abhängig, die bereits zur Erstellung von $K(d, \Theta_1, \Theta_2)$ bestimmt wurden.

$$\frac{\partial K(d, \Theta_1, \Theta_2)}{\partial \Theta_1} = \frac{d^2}{e^{2\Theta_1}} e^{2\Theta_2} e^{-\frac{1}{2} \frac{d^2}{e^{2\Theta_1}}} = \frac{d^2}{e^{2\Theta_1}} K(d, \Theta_1, \Theta_2) \quad (9.224)$$

$$\frac{\partial K(d, \Theta_1, \Theta_2)}{\partial \Theta_2} = 2e^{2\Theta_2} e^{-\frac{1}{2} \frac{d^2}{e^{2\Theta_1}}} = 2K(d, \Theta_1, \Theta_2) \quad (9.225)$$

In vielen Modellverfahren werden benötigte Parameter mittels Kreuzvalidierung oder Leave-One-Out Kreuzvalidierung (LOO) robust ermittelt (siehe Kapitel 9.18). Ist ein Trainingspunkt \mathbf{x}_i aus den Trainingsdaten X entfernt worden, so wird die logarithmische Vorhersagewahrscheinlichkeit wie in Gleichung 9.226 berechnet [74].

$$\log p(y_i|X, \mathbf{y}_{-i}, \Theta) = -\frac{1}{2} \log \sigma_i^2 - \frac{(y_i - \mu_i)^2}{2\sigma_i^2} - \frac{1}{2} \log(2\pi) \quad (9.226)$$

μ_i und σ_i stehen für den Mittelwert sowie die Varianz des Datensatzes ohne den i -ten Datenpunkt und können ohne großen Aufwand aus dem Komplettdatensatz (K^{-1} und \mathbf{y}) ermittelt werden [74].

$$\mu_i = y_i - \frac{[K^{-1}\mathbf{y}]_i}{[K^{-1}]_{ii}} \text{ und } \sigma_i^2 = \frac{1}{[K^{-1}]_{ii}} \quad (9.227)$$

Die resultierende logarithmische Vorhersagewahrscheinlichkeit L_{LOO} wird dann durch die Summe aller Einzelwahrscheinlichkeiten bestimmt (Pseudo-Likelihood).

$$L_{LOO}(X, \mathbf{y}, \Theta) = \sum_{i=1}^{n_r} \log p(y_i | X, \mathbf{y}_{-i}, \Theta) \quad (9.228)$$

Die Ableitung von L_{LOO} nach einem Parameter Θ_j wird nach RASMUSSEN mittels Gleichung 9.229 bestimmt [74].

$$\frac{\partial L_{LOO}}{\partial \Theta_j} = \sum_{i=1}^{n_r} \frac{\alpha_i [Z_j \alpha]_i - \frac{1}{2} \left(1 + \frac{\alpha_i^2}{[K^{-1}]_{ii}} \right) [Z_j K^{-1}]_{ii}}{[K^{-1}]_{ii}} \quad (9.229)$$

α und Z_j können auch in diesem Fall aus dem gesamten Trainingsdatensatz X bestimmt werden [74].

$$\alpha = K^{-1}\mathbf{y} \text{ und } Z_j = K^{-1} \frac{\partial K}{\partial \Theta_j} \quad (9.230)$$

Im Vergleich zur logarithmischen Likelihood der Randverteilung $\log p(\mathbf{y}|X, \Theta)$, welche die Wahrscheinlichkeit der Trainingsdaten bei gegebener Annahmen des Modells (Kovarianzfunktionen) ermittelt, kann L_{LOO} ebenfalls eine Abschätzung liefern, ob die Modellannahmen für die Trainingsdaten erfüllt sind. Die Anwendung von L_{LOO} ist daher bei ungefähr gleichem Rechenaufwand in vielen Fällen robuster [104, 74]. Zu den hier dargestellten grundlegenden Verfahren zur Regression existieren ebenfalls Adaptionen zur Klassifizierung, welche umfangreich von RASMUSSEN dargestellt werden [74]. Weiterhin bestehen verschiedenste Abwandlungen und Erweiterungen des Verfahrens, die zum Beispiel andere Likelihood-Funktionen verwenden, besonders geeignet für große Datensätze sind oder Anstelle des Mittelwerts $\mu \equiv 0$ des Gauß Prozesses einen konstanten, linearen oder polynomialen Zusammenhang annehmen. Ein guter Einstieg in weiterführende Themen bieten dabei die Arbeiten von RASMUSSEN et.al. [74, 72, 73].

9.16 Künstliche Neuronale Netzwerke

Künstliche Neuronale Netzwerke (KNN) sind Metamodelle die durch biologische Nervensysteme inspiriert wurden. Die Hauptelemente eines KNNs bilden *Neuronen*, welche durch Informationsaustausch untereinander verschiedene Aufgaben (zum Beispiel Approximation eines Funktionswerts y_0 an einer Faktorkombination \mathbf{x}_0) erfüllen können. KNNs werden dazu mit bekannte Versuchsdaten trainiert. Trainieren bedeutet in diesem Zusammenhang, entsprechend den biologischen Vorgängen im Gehirn, die Anpassung der Verbindungen und somit des Informationsaustauschs

zwischen den einzelnen Neuronen.

Die ersten Künstlichen Neuronalen Netzwerke wurden bereits 1943 von dem Neurophysiologen WARREN MCCULLOCH und dem Logiker WALTER PITTS vorgestellt. Bedeutsame Weiterentwicklungen und Einsatzmöglichkeiten sind jedoch erst zwischen 1970 und 1980 erzielt worden. In Kombination mit modernen Computersystemen bieten KNNs bei richtiger Anwendung eine gegen Störgrößen und einzelne Datenfehler robuste Methode zur Metamodellerzeugung [60].

Neben spezialisierten Netzwerktypen wird häufig das *einfache Feedforward Netzwerk* zur Erzeugung von Metamodellen eingesetzt. Abbildung 9.41 zeigt eine schematische Darstellung dieses Grundtyps, bei dem die Neuronen (Punkte) in unterschiedliche Ebenen aufgeteilt werden. Die erste Ebene (Eingangs Ebene) erhält die verschiedenen Faktoren (x_1, \dots, x_{n_f}) des untersuchten Systems als Eingangssignale. Auf der gegenüberliegenden Seite befindet sich die Ausgangsebene, in der sich die gesuchte Approximation \hat{y} wiederfindet. Künstliche Neuronale Netzwerk können dabei eine oder direkt mehrere Ausgangsvariablen approximieren. In den meisten Fällen ist es jedoch sinnvoll für jede zu untersuchende Ausgangsvariable y ein eigenes KNN zu verwenden, da dieses nur dadurch speziell an die zu untersuchende Ausgangsvariable angepasst wird. Zwischen Ein- und Ausgangsebene befinden sich eine oder mehrere *versteckte* Ebenen, wobei in den meisten praktischen Fällen nicht mehr als zwei versteckte Ebenen benötigt werden.

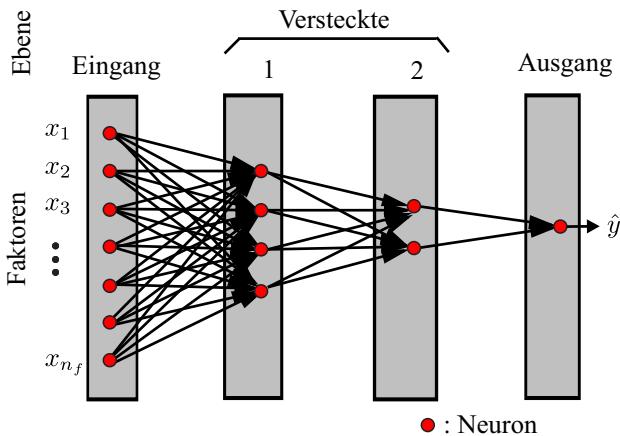


Abb. 9.41 Feedforward Netzwerk

Aktivierung eines Neurons

Die Hauptinformation jedes einzelnen Neurons ist seine *Aktivierung*, welche durch eine reelle Zahl $a = [0, 1]$ repräsentiert wird. Die Aktivierung in einem *einfachen*

Feedforward Netzwerk hängt dabei nur von der Aktivierung der Neuronen aus der vorhergehenden Ebene ab. Abbildung 9.42 zeigt schematisch die Berechnungsmethode zur Ermittlung der Aktivierung *eines* Neurons j aus Ebene $k+1$ in Abhängigkeit *aller* Neuronen aus Ebene k und einer Konstanten $b_{k,0,j}$. Im ersten Schritt wird dazu jedem Neuron aus Ebene k ein Gewicht⁵ $b_{k,i,j}$ zugeordnet und die Summe aller gewichteten Aktivierungen und der Konstante $b_{k,0,j}$ gebildet.

$$S_{k,j} = b_{k,0,j} + a_{k,1}b_{k,1,j} + a_{k,2}b_{k,2,j} + \dots + a_{k,n_k}b_{k,n_k,j} = \sum_{i=0}^{n_k} a_{k,i}b_{k,i,j} \text{ mit } a_{k,0}=1 \quad (9.231)$$

Die Aktivierung des Neurons j aus Ebene $k+1$ wird im zweiten Schritt durch eine im Voraus definierte Aktivierungsfunktion $a(S)$ bestimmt.

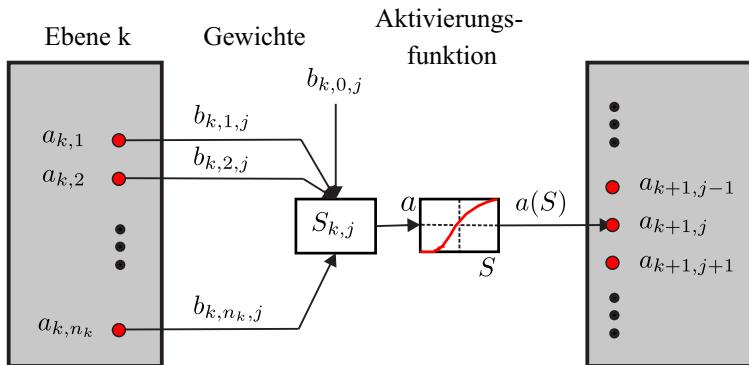


Abb. 9.42 Aktivierung eines Neurons

Der größte Teil aller Anwendungen verwendet dazu sigmoidale (s-förmige) Aktivierungsfunktionen, wie sie in Abbildung 9.43 und Gleichung 9.232 dargestellt sind [60].

$$\begin{aligned} a(S) &= \frac{1}{1+e^{-S}} \text{ mit } a'(S) = a(S)[1-a(S)] \\ a(S) &= \tanh(S) = \frac{e^S - e^{-S}}{e^S + e^{-S}} \text{ mit } a'(S) = 1 - \tanh^2(S) = \frac{1}{\cosh(S)} = \operatorname{sech}(S) \\ a(S) &= \frac{2}{\pi} \arctan\left(\frac{\pi}{2}S\right) \text{ mit } a'(S) = -\frac{1}{\sqrt{1-\left(\frac{\pi}{2}S\right)^2}} \end{aligned} \quad (9.232)$$

Allen Aktivierungsfunktionen ist gemein, dass ihre theoretischen Extremwerte nie- mals erreicht werden. Die genaue Form der Aktivierungsfunktion hat einen gerin- gen Einfluss auf die Qualität des erzeugten Künstlichen Neuralen Netzwerks. Eine Auswirkung weist die Wahl jedoch auf die Geschwindigkeit des Trainingsprozesses auf. Durch die einfache Berechnung der Ableitung der ersten dargestellten Aktivie-

⁵ $b_{k,i,j}$: Gewicht i der Ebene k und Zielneuron j der folgenden Ebene $k+1$

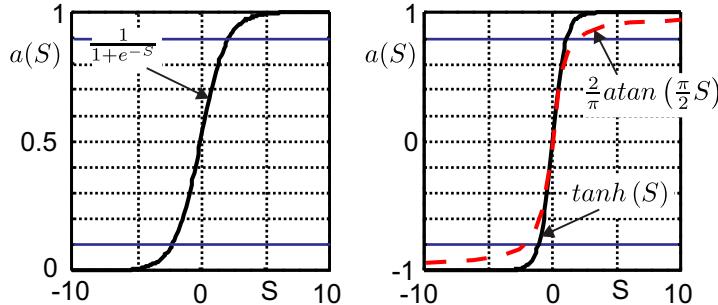


Abb. 9.43 Aktivierungsfunktionen

rungsfunktion (logistische Funktion, $a(S) = \frac{1}{1+e^{-S}}$) wird diese in der Praxis häufig den anderen Funktionen vorgezogen. Aktivierungen von $a \geq 0.9$ werden dabei als komplett aktiviert und $a \leq 0.1$ als nicht aktiviert interpretiert, so dass lediglich der Funktionsbereich $[0.1, 0.9]$ verwendet wird. Durch die flache Form der sigmoidalen Funktionen bei großen Absolutwerten wird der Einfluss von Extremwerten (zum Beispiel Ausreißern) in den Daten deutlich gedämpft.

Training durch Fehlerrückführung

Zur Bestimmung der verschiedenen Gewichte \mathbf{b} wird das Künstliche Neurale Netzwerk mit bekannten Daten trainiert. Dazu wird beispielsweise die Methode der Fehlerrückführung (*Backpropagation*) eingesetzt, welche mittels eines Gradientenverfahrens den quadratischen Fehler der Approximation $(y - \hat{y})^2$ minimiert. Im allgemeinen Fall kann ein neuronales Netzwerk $n_a \geq 1$ Ausgangsvariablen (y_1, \dots, y_{n_a}) aufweisen. Der quadratische Fehler E_m des Netzwerks für einen Datenpunkt (Faktorkombination) m ist die Summe aller einzelnen Fehlerterme der Ausgangsvariablen.

$$E_m = \frac{1}{n_a} \sum_{l=1}^{n_a} [y_{m,l} - \hat{y}_{m,l}]^2 = \frac{1}{n_a} \sum_{l=1}^{n_a} \left[y_{m,l} - a(S_{*,l}^{(m)}) \right]^2 \quad (9.233)$$

$S_{*,l}^{(m)}$ steht in diesem Zusammenhang für die gewichtete Summe der letzten versteckten Ebene $*$, welche in die Aktivierungsfunktion $a(S)$ zur Berechnung der l^{ten} Ausgangsvariable für den Datensatz m eingesetzt wird. Werden n_r bekannte Datensätze zum Training des Netzwerks verwendet, so berechnet sich der gesamte mittlere Fehler durch Gleichung 9.234.

$$E = \frac{1}{n_r} \sum_{m=1}^{n_r} E_m \quad (9.234)$$

Bei gegebenen aktuellen Werten der Gewichte \mathbf{b} wird im ersten Schritt des Trainingsverfahrens der Gradient des Fehlers E in Abhängigkeit der Gewichte $\frac{\partial E}{\partial \mathbf{b}}$ be-

stimmt. Dazu wird der Einfluss der Gewichte auf den Fehler ausgehend von der letzten Ebene (Ausgangsebene) *rückwärts* bis zur ersten Ebene (Eingangsebene) berechnet, wodurch auch der Name des Verfahrens (Fehlerrückführung) zu erklären ist.

Die partielle Ableitung des Fehlers E nach Gewicht $b_{k,i,j}$ zwischen einem Ausgangsneuron j und einem Neuron i aus der vorherigen Ebene k mit der momentanen Aktivierung $a_{k,i}$ lässt sich berechnen durch Gleichung 9.235 (Abbildung 9.44). Die Ableitung wird in dieser Gleichung für einen einzigen Trainingspunkt l bestimmt.

$$\frac{\partial E}{\partial b_{k,i,j}} = -a_{k,i} a' (S_{k,j}) [y_j - \hat{y}_j] = -a_{k,i} \delta_{k,j} \quad (9.235)$$

Im Gegensatz zur Ausgangsebene, in der die Zielwerte y_j bekannt sind, existieren für alle anderen Ebenen keine direkten Zielwerte, wodurch die partiellen Ableitungen in Abhängigkeit der $\delta_{k+1,m}$ und $b_{k+1,j,m}$ Werte der folgenden Ebene bestimmt werden müssen (Abbildung 9.44).

$$\delta_{k-1,j} = a' (S_{k-1,j}) \sum_{m=1}^{n_{k+1}} \delta_{k,m} b_{k,j,m} \quad (9.236)$$

$$\frac{\partial E}{\partial b_{k-1,i,j}} = -a_{k-1,i} \delta_{k-1,j} \quad (9.237)$$

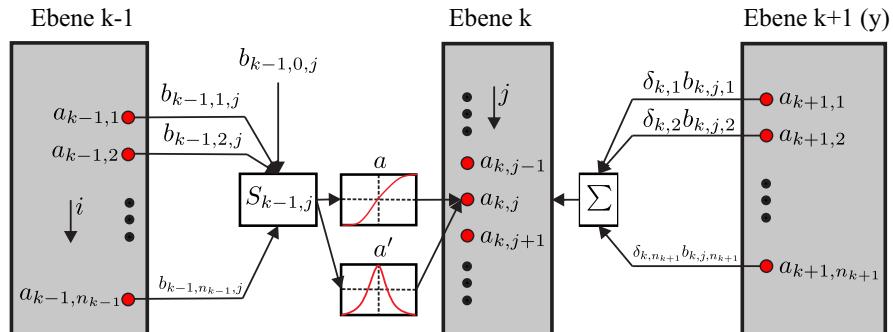


Abb. 9.44 Berechnung der partiellen Ableitung

Durch die Gleichungen 9.235 und 9.237 werden somit die Gradienten des Fehlers E in Abhängigkeit jedes Gewichts $b_{k,i,j}$ bei Betrachtung eines bekannten Datensatzes bestimmt. Bei der Berücksichtigung von n_r Datensätzen wird der Gradient jedes Gewichts $b_{k,i,j}$ durch die Summe der einzelnen Gradienten berechnet:

$$\left[\frac{\partial E}{\partial b_{k,i,j}} \right]_{\Sigma} = \sum_{l=1}^{n_r} \left[\frac{\partial E}{\partial b_{k,i,j}} \right]_l \quad (9.238)$$

Nachdem durch die Bestimmung der Gradienten ein Richtungsvektor B_g zur Minimierung von E bestimmt wurde, wird die Länge s entlang dieses Vektors gesucht, die ausgehend von den momentanen Gewichten B_0 den Fehler E minimiert.

$$\min_s E(B_0 + sB_g) \quad (9.239)$$

MASTERS schlägt zur Bestimmung einer guten Schrittweite s das Konjugierte-Gradienten-Verfahren vor, welches hier nur kurz vorgestellt wird [60]. Im ersten Schritt des Verfahrens werden drei Punkte entlang der Suchrichtung ermittelt, wobei der mittlere Punkt den geringsten Fehler aufweisen muss. Dadurch kann die Suche nach der optimalen Schrittweite auf den Bereich zwischen den äußeren Punkten beschränkt werden (siehe Abbildung 9.45). Im zweiten Schritt wird das Minimum mit Hilfe einer Parabel abgeschätzt. Sollte sich ein Minimum außerhalb der Grenzen ergeben, so wird ein Punkt innerhalb der Grenzen bei einem vorgegebenen Verhältnis gebildet. Der zweite Schritt wird solange wiederholt, bis ein stabiles Minimum gefunden wurde. In einem iterativen Prozess wird das gesamte Training so lange fortgesetzt, bis ein Abbruchkriterium erreicht wird (zum Beispiel Unterschreiten einer maximalen Fehlergrenze oder Überschreiten einer maximalen Anzahl an Iterationen). Aus mathematischer Sicht ist die Minimierung des Fehlers komplex, da meist viele lokale Minima und Bereiche mit sehr flachen Gradienten existieren [60, 70, 10].

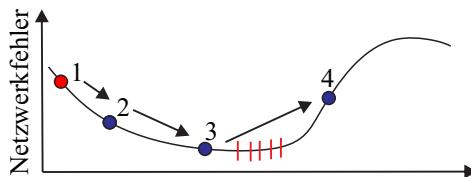


Abb. 9.45 Minimierung der Fehlerfunktion

Normierung

Normalerweise sind Neuronale Netzwerke so aufgebaut, dass die Neuronen einen Aktivierungsbereich von 0.1 bis 0.9 aufweisen. Daten, welche durch ein Neuronales Netzwerk approximiert werden sollen, müssen daher vor dem Training auf diesen Bereich normiert werden. Zur Vereinfachung wird in einigen Fällen eine Normierung auf den Bereich [0, 1] verwendet (n: normiert, r: real).

$$R(x) \rightarrow N(x) : n = \frac{x - x_{\min}}{x_{\max} - x_{\min}} d + m \quad (9.240)$$

$$N(x) \rightarrow R(x) : x = \frac{n - m}{d} (x_{\max} - x_{\min}) + x_{\min} \quad (9.241)$$

Dabei bezeichnet d die Spannbreite des normierten Bereichs ($d = 0.8$ oder $d = 1.0$), m die untere Grenze des Zielbereichs ($m = 0.1$ oder $m = 0.0$), und $x_{\min,\max}$ den kleinsten beziehungsweise größten Wert der Trainingsdaten. Daten aus physikalischen Experimenten sind hingegen mit Rauschen überlagert und werden sinnvoller durch die Verwendung des Mittelwerts μ , der Standardabweichung σ und einem gewählten Z-Wert mit folgenden Gleichungen normiert.

$$R(x) \rightarrow N(x) : n = \frac{r}{\sigma} x + \left(0.1 - r \left[\frac{\mu}{\sigma} + x_{\min}^* \right] \right) \quad (9.242)$$

$$N(x) \rightarrow R(x) : x = \frac{\sigma}{r} n + \left(\mu + \sigma \left[x_{\min}^* - \frac{0.1}{r} \right] \right) \quad (9.243)$$

mit

$$r = \frac{0.8}{x_{\max}^* - x_{\min}^*}$$

$$x_{\min,\max}^* = \pm Z\sigma + \mu$$

One-of-N Kodierung

Zur Erhöhung der Vorhersagequalität von Künstlichen Neuronalen Netzwerken ist es gerade bei komplexen Zusammenhängen sinnvoll den Faktorraum in überlappende Unterbereiche aufzuteilen. Dieses wird beispielsweise durch die *One-Of-N* Kodierung erreicht, welche aus dem Bereich der Fuzzy-Technologie bekannt ist (Abbildung 9.46). Eine Eingangsvariable (Faktor) x_j wird dazu durch N Variablen $x_{j_0}, \dots, x_{j_{N-1}}$ ersetzt, wodurch das Neuron der Eingangsebene für die Variable x_j durch N neue Neuronen (Eingangsvariablen) substituiert wird. Die einzustellenden Werte für die N Variablen $x_{j_0}, \dots, x_{j_{N-1}}$ werden wie in Gleichung 9.244 gezeigt in Abhängigkeit von der Originalvariablen x_j berechnet (Abbildung 9.46).

$$x_{jl} = \max \left(0, \frac{d_{x_j} - |\min(x_j) + ld_{x_j} - x_j|}{d_{x_j}} \right) \quad l = 0, \dots, N-1 \quad (9.244)$$

mit $d_{x_j} = \frac{\max(x_j) - \min(x_j)}{N-1}$

Durch die Aufteilung werden für unterschiedliche Bereiche des Faktors x_j unterschiedliche Neuronen und somit unterschiedliche Bereiche des Künstlichen Neuronalen Netzwerks aktiviert. Mittels der erhöhten Gewichtungsanzahl und den überlappenden Eingangsvariablen kann meist ein besseres Netzwerk trainiert werden,

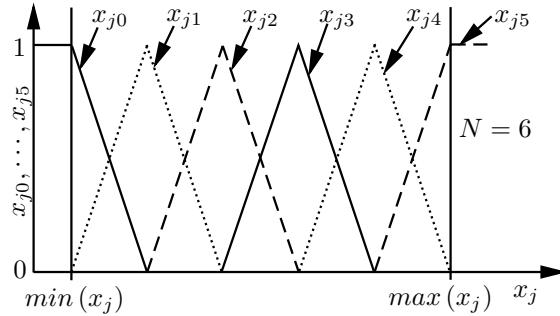


Abb. 9.46 One-of- N Kodierung mit $N=6$

als bei alleiniger Erhöhung der Anzahl durch mehr Neuronen in den versteckten Ebenen.

Anzahl von Schichten und Gewichten

In der Praxis tritt die Schwierigkeit auf, eine sinnvolle Wahl für die Anzahl von Schichten und Neuronen sowie die Iterationsanzahl des Trainingsprozesses zu finden. In vielen Fällen ist *eine* versteckte Ebene zur Abbildung auch komplexer Zusammenhänge ausreichend. Eine zweite versteckte Ebene ist bei sehr komplexen oder nicht-kontinuierlichen Zusammenhängen notwendig (zum Beispiel Sägezahn-Funktion). In allen anderen Anwendungen wird die Genauigkeit des Künstlichen Neuronalen Netzwerks durch eine zweite versteckte Ebene zwar geringfügig verbessert, jedoch nimmt die Trainingsgeschwindigkeit des Netzwerks deutlich ab.

Für die Wahl der Neuronenanzahl je Ebene existieren lediglich Faustformeln. Bei der Verwendung von einer versteckten Ebene wird als erster Anhaltspunkt die folgende Anzahl an Neuronen vorgeschlagen [60]:

$$n_{vE1} = \sqrt{n_a n_f} \quad \text{mit } n_a < n_f \quad (9.245)$$

Bei komplexen Zusammenhängen können mehr Neuronen und bei einfachen Zusammenhängen weniger Neuronen sinnvoll sein. Um ein Overfitting des Netzwerks zu vermeiden, sollte mit kleiner Neuronenanzahl gestartet werden. Für die Verwendung von zwei versteckten Ebenen existiert die folgende Faustformel zur Auslegung der Neuronenanzahl [60]:

$$n_{vE1} = n_a r^2 \quad n_{vE2} = n_a r \quad \text{mit } r = \sqrt[3]{\frac{n_f}{n_a}} \quad \text{und } n_a < n_f \quad (9.246)$$

Bei der Wahl der Neuronenanzahl ist zu beachten, dass die resultierende Anzahl n_b

der zu bestimmenden Gewichte \mathbf{b} schnell mit steigender Neuronenanzahl wächst und dadurch ebenfalls die Menge der benötigten Trainingsdaten $n_r \geq n_b$ (Gleichung 9.247).

$$\begin{array}{lll}
 \text{Versteckte Ebenen} & & \text{Anzahl Gewichte } n_b \\
 0 & & n_a(n_f + 1) \\
 1 & & n_{vE1}(n_f + 1) + n_a(n_{vE1} + 1) \\
 2 & & n_{vE1}(n_f + 1) + n_{vE2}(n_{vE1} + 1) + n_a(n_{vE2} + 1)
 \end{array} \quad (9.247)$$

n_f : Anzahl Neuronen der Eingangsebene [Faktoren]

$n_{vE1/2}$: Anzahl Neuronen der versteckten Ebene 1 bzw. 2

n_a : Anzahl Neuronen der Ausgangsebene [Ausgangsvariablen]

Overfitting und Testdaten

Künstliche Neuronale Netzwerke besitzen die gewünschte Fähigkeit sich an komplexe Funktionszusammenhänge anzupassen. Dieses ermöglicht jedoch ebenfalls eine ungewünschte Überanpassung an verrauchten Messdaten (zum Beispiel aus physikalische Experimenten). Abbildung 9.47 zeigt beispielhaft eine verrauchte Datenmenge eines linearen Zusammenhangs. Bei genügend Neuronen kann nicht nur der lineare Zusammenhang sondern auch das überlagerte Rauschen in den Trainingsdaten vollkommen abgebildet werden (gestrichelte Linie). Der Approximationsfehler der Trainingsdaten ist zwar gering, jedoch wird der eigentliche Zusammenhang nicht korrekt abgebildet. Zusätzliche Testdaten, die ebenfalls den grundlegenden Zusammenhang inklusive Messrauschen aufweisen, ermöglichen eine Überprüfung eines möglichen Overfittings. Solange sich das KNN an den grundlegenden Zusammenhang anpasst, verbessert sich der Fehler der Trainings und Testdaten. Sobald sich das Netzwerk an das Rauschen anpasst verringert sich zwar der Fehler der Trainingsdaten, jedoch steigt der Fehler der Testdaten. An diesem Punkt muss das Training beendet werden (Abbildung 9.47, rechts). Die gewählte Aufteilung zwischen Trainings- und Testdaten ist ein Kompromiss zwischen dem Wunsch so viele Daten wie möglich zum Training des Netzwerkes zu verwenden, um den grundlegenden Zusammenhang der Daten zu ermitteln, und genügend Testdaten um ein Overfitting detektieren zu können. Genauso wie die Trainingsdaten müssen die Testdaten den gesamten Faktorraum gleichmäßig abdecken. Sind deutlich mehr Trainingsdaten vorhanden, als zu bestimmende Gewichte $n_r \gg n_b$, ist das Risiko eines Overfittings geringer. Abbildung 9.48 zeigt schematisch die Aufteilung einer Datenmenge in Trainings- und Testdaten sowie das iterativen Training inklusive Qualitätsprüfung eines Netzwerks.

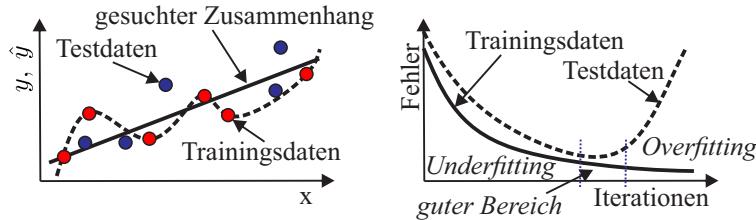


Abb. 9.47 Approximationsfehler und Overfitting

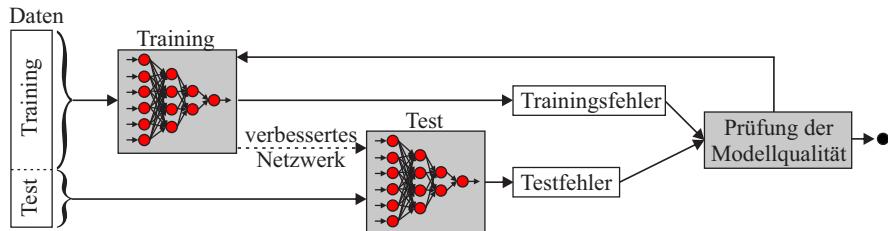


Abb. 9.48 Training und Prüfung eines Künstlichen Neuronalen Netzwerks

Beispiel

Abbildung 9.49 zeigt die Approximation der Beispieldfunktion aus Gleichung 9.156 bei Verwendung verschiedener Netzwerke mit *einer* verdeckten Ebene und unterschiedlicher Neuronenzahl. Bereits zwei Neuronen in der verdeckten Ebene genügen, um den gesuchten Zusammenhang ausreichend genau zu approximieren (Abbildung 9.49 [Mitte, oben] und 9.50). Durch die Verwendung von einem Neuron wird lediglich ein globaler Trend abgebildet und die Wahl von fünf Neuronen erhöht die Qualität der Approximation nur geringfügig. Die Gefahr des Overfittings steigt bei fünf Neuronen an, da bereits mehr Gewichte als Datenpunkte zum Training vorhanden sind. Die untere Zeile in Abbildung 9.49 zeigt die Approximation der jeweils gleichen Netzstrukturen, wenn zum Training andere Startwerte für die Gewichte als in der ersten Zeile verwendet werden. In dem dargestellten Beispiel wird dadurch ein deutlich anderer Zusammenhang zwischen Ein- und Ausgangsvariablen ermittelt. Die Summe der Approximationsfehler der gegebenen Trainingspunkte ist nahezu identisch zu den Modellen aus der ersten Zeile. Dieses zeigt, dass mit Künstlichen Neuronalen Netzwerken einerseits komplexe Zusammenhänge dargestellt werden können, auf der anderen Seite es aber ebenfalls notwendig ist, das ermittelte Metamodell auf seine Allgemeingültigkeit zu prüfen. Dieses wird in den meisten Fällen durch die Prüfung der Approximationsgenauigkeit an zusätzlichen bekannten Datenpunkten (Testpunkte), welche nicht zum Training des Netzes verwendet wurden, festgestellt. Grundsätzlich ist eine Prüfung des Metamodells nach seiner Erstellung nicht nur für Künstliche Neuronale Netzwerke sinnvoll, sondern für alle Metamodelle, die sich automatisch an komplexe Zusammenhänge anpassen können.

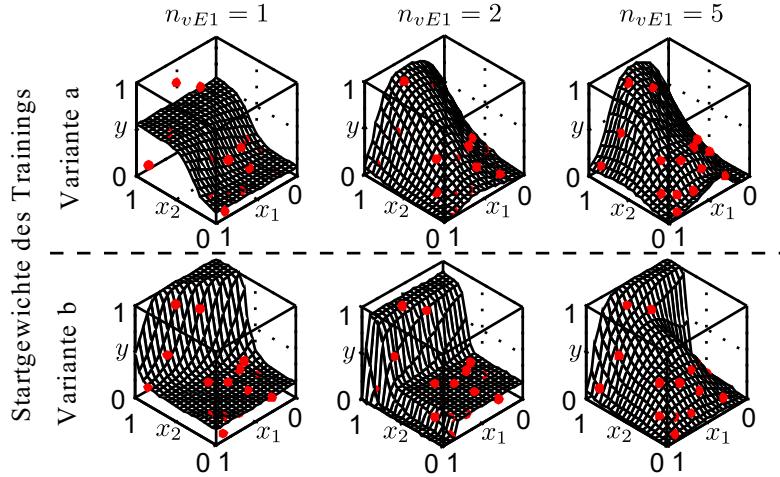


Abb. 9.49 Künstliche Neuronale Netzwerke: Beispiele

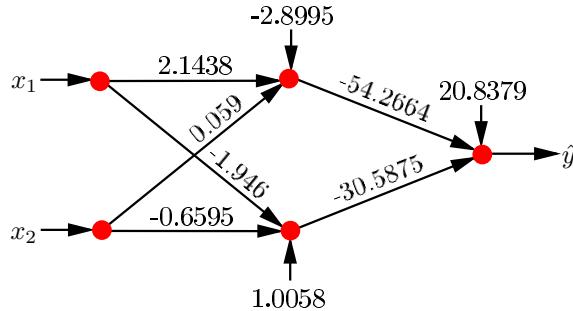


Abb. 9.50 KNN mit zwei Neuronen in einer versteckten Ebene

Schlussbemerkung

Neben den hier dargestellten Feedforward Netzwerken existieren deutlich komplexe Strukturen, in denen beispielsweise die Aktivierung eines Neurons nicht nur von den Neuronen der vorhergehenden Ebene, sondern auch anderen Ebenen abhängt. Grundsätzlich können mit Künstlichen Neuronalen Netzwerken komplexe Zusammenhänge gut abgebildet werden, wobei bereits einfache Feedforward Netzwerke mit einer oder zwei versteckten Ebenen in vielen Anwendungsfällen ausreichend sind. Die optimale Wahl der Netzwerkstruktur (z.B. Anzahl Neuronen pro Ebene) ist im Vorfeld häufig schwierig und führt in der Praxis dazu, dass verschiedene Netzwerke erzeugt werden und durch ein zu definierendes Beurteilungskriterium das vorteilhafteste für die aktuelle Aufgabe gewählt wird.

9.17 Kombinerte Modelle

Bei den meisten Verfahren zur Erstellung eines Regressionsmodells treten Probleme mit der Erwartungstreue (*bias*) und der Stabilität auf, was zu schlechten Modellvorschägen führt. Probleme mit der Erwartungstreue treten besonders dann auf, wenn die gleichen Daten zum Training und zur Validierung des Modells verwendet werden [47, 76, 77, 56, 46]. Das Risiko ist dabei unabhängig vom Gütekriterium, welches zur Beurteilung des Modells gewählt wird (siehe Kapitel 9.3.1). Selbst wenn zur Validierung ein komplett unabhängiger Datensatz verwendet wird, kann es zu Problemen bei dem Erwartungswert kommen [47, 52, 77]. Instabilität bezieht sich hingegen auf das Problem, dass kleine Änderungen in den Daten zu neuen (lokalen) Minima der verwendeten Gütefunktion führen und folglich zu unterschiedlichen Modellen [47, 8, 53, 34, 13]. Kleine Änderungen treten zum Beispiel durch Rauschen oder Löschen beziehungsweise Hinzufügen von Daten auf.

Die Kombination von verschiedenen Modellen kann diese Effekte reduzieren [47, 8, 69, 13, 46]. Typischerweise werden zwei Schritte zur Erzeugung eines Kombinationsmodells durchgeführt [13]. Im ersten Schritt werden verschiedene Modelle auf Basis des Datensatzes erzeugt. Hierbei werden verschiedene Modelltypen, Modellparameter oder (überlappende) Teildatensätze verwendet. Im zweiten Schritt werden die unterschiedlichen Modelle linear kombiniert. Abbildung 9.51 zeigt ein beispielhaftes Flussdiagramm zur Modellerstellung bei Verwendung unterschiedlicher Teilmengen zur Modellerstellung und Validierung. Nachdem die Teilmengen zufällig definiert wurden, werden drei Modelle mit jeweils zwei unterschiedlichen Teilmengen erzeugt. Die jeweils nicht verwendete dritte Menge wird zur Validierung verwendet. Sollte ein Modell mit fester Struktur erstellt werden, können die Parameter anschließend durch den Einsatz der gesamten Daten nochmals verfeinert werden. Das endgültige Modellergebnis wird im Anschluss aus dem Mittelwert oder Median aller drei (n_M) Sub-Modelle berechnet.

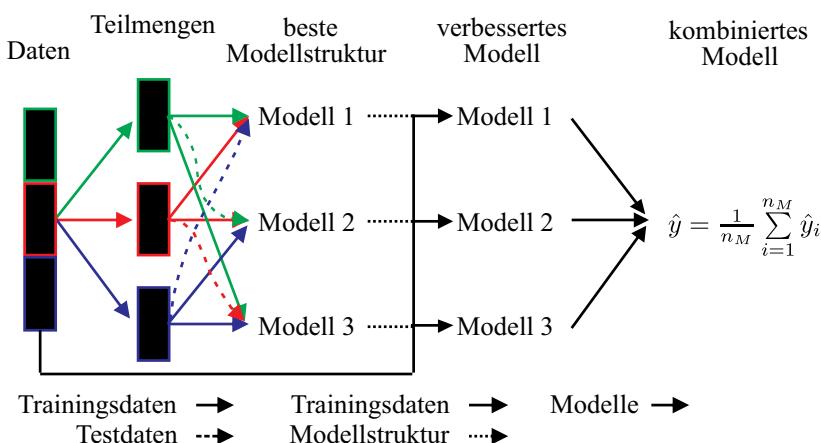


Abb. 9.51 Beispielprozess für kombinierte Modelle

9.18 Qualität von Metamodellen

Die hohe Flexibilität und Anpassungsfähigkeit der dargestellten Modellverfahren ermöglicht eine gute Approximation von unbekannten und komplexen Zusammenhängen zwischen Ein- und Ausgangsvariablen eines zu analysierenden Systems. Es besteht jedoch die Gefahr, Metamodelle zu erzeugen, welche sich zwar gut an die gegebenen Datenpunkte anpassen, deren Vorhersagegenauigkeit für nicht zum Training verwendete Faktorkombinationen jedoch gering sind. Das kann durch falsch gewählte Parameter (Abbildung 9.52, $\Theta = 100$), eine mathematische falsche Interpretation des Zusammenhangs⁶ (Abbildung 9.49, Zeile 2) oder ein Overfitting an die gegebenen Daten (Abbildung 9.47) auftreten. Daher ist die Überprüfung der Modellqualität für die Approximation neuer Faktorkombinationen eine niemals zu vernachlässigende Aufgabe.

Eine typische Beurteilungsgröße ist für die Qualität eines Metamodells der mittlere quadratische Approximationsfehler (MSE:*mean squared error*):

$$MSE = \frac{1}{n_r} \sum_{i=1}^{n_r} (y_i - \hat{y}_i)^2 \quad (9.248)$$

Alternativ zu MSE wird der Mittelwert der absoluten Differenz ($|y_i - \hat{y}_i|$) verwendet. Zur Stabilisierung des Ergebnisses ist auch der Einsatz des Medians anstelle des Mittelwerts möglich. Weitere Gütekriterien finden sich in Kapitel 9.3.1.

Eine Betrachtung des Approximationsfehlers ($y_i - \hat{y}_i$) an den Datenpunkten X_e , welcher zur Erzeugung des Metamodells verwendet wurde, wie es von linearen Regressionsmodellen bekannt ist, reicht bei komplexen Metamodellen nicht aus. So verläuft beispielsweise ein Metamodell auf Basis des einfachen Kriging-Ansatzes immer genau durch die zum Training verwendeten Datenpunkte, was eine sinnvolle Beurteilung des Metamodells durch Residuen der Trainingsdaten ausschließt. Entsprechendes gilt für Overfitting bei verrauschte Trainingsdaten.

Testdaten

Eine der einfachsten Möglichkeiten zur Beurteilung eines Metamodells ist die Ermittlung der Residuen oder des mittleren quadratischen Approximationsfehlers an zusätzlichen Datenpunkten X_t , welche nicht zur Erzeugung des Metamodells verwendet wurden. Sind die Residuen an diesen Datenpunkten in der gleichen Größenordnung wie die Residuen der Datenpunkte X_e und gleichzeitig in einer für die durchzuführenden Analysen akzeptablen Größe, so wird davon ausgegangen, dass das Metamodell den grundsätzlichen Zusammenhang zwischen Ein- und Ausgangsvariablen richtig abbildet und die Approximationsgenauigkeit ausreichend ist. Bei der Wahl der Testpunkte X_t ist darauf zu achten, dass der gesamte Faktorraum gleichmäßig und vollständig abgedeckt wird, da die Modellqualität für den gesamten Faktor-

⁶ basierend auf den gegebenen Daten

raum gewährleistet sein muss, um im Weiteren vertrauenswürdige Untersuchungen durchzuführen.

Der MSE von 10 zufällig gewählten Faktorkombinationen für die drei Kriging-Modelle aus Abbildung 9.52 zeigen bereits deutlich, dass im dargestellten Beispiel die Wahl von $\Theta = 0.01$ das sinnvollste Metamodell erzeugt.

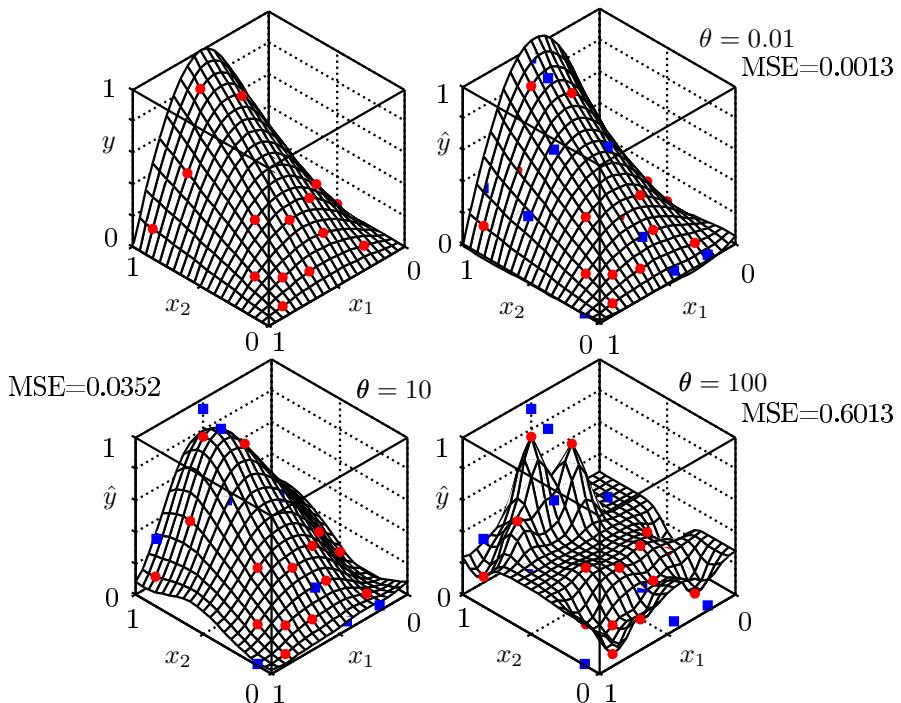


Abb. 9.52 Testdaten zur Überprüfung eines Metamodells

Grundsätzlich muss der Anwender bei einer begrenzten Anzahl an Datenpunkten entscheiden, wie viele Datenpunkte für die Erzeugung des Metamodells verwendet werden sollen und wie viele zur Kontrolle der Qualität.

Kreuzvalidierung

Werden zur Erzeugung von Trainingsdaten Simulationsmodelle mit langen Rechenzeiten oder langwierige physikalische Experimente verwendet, so ist es häufig notwendig, alle Datenpunkte zur Erzeugung des Metamodells zu verwenden, da ansonsten die vorhandene Komplexität der Zusammenhänge nicht abgebildet werden kann. In diesen Fällen wird die Kreuzvalidierung (*cross validation*) zur Prüfung der Modellqualität eingesetzt. Dabei werden drei Verfahren der Kreuzvalidierung unterschieden.

Die ***zufällige Untergruppen-Validierung*** unterteilt die vorhandenen Datenpunkte in eine Trainings- und eine Testgruppe, wobei die Testgruppe meist kleiner ist als die Gruppe der Trainingsdaten. Das Metamodell wird nur mit den Trainingsdaten erzeugt und die Qualität des Metamodells mit den Daten der Testgruppe beurteilt (zum Beispiel mit MSE). Dieses Validierung wird mehrmals mit unterschiedlichen (zufälligen) Aufteilungen der Daten wiederholt. Der Mittelwert aller MSE aus diesen Validierungen wird als Gütekriterium verwendet. Problematisch kann sich hierbei auswirken, dass die Aufteilung zufällig durchgeführt wird. So können einige Datenpunkte niemals und andere mehrmals zur Testgruppe gehören, was den Datenpunkten unterschiedliche Bedeutung in der Modellerzeugung beziehungsweise Kontrolle zukommen lässt.

Aus diesem Grund wird in den meisten Fällen die ***k-fold*** Kreuzvalidierung eingesetzt. Dabei werden die gegebenen Datenpunkte in k möglichst gleichgroße Untergruppen K_1, \dots, K_k unterteilt. Eine Untergruppe wird dann zum Test der Qualität und die restlichen $k - 1$ Untergruppen zur Erzeugung des Metamodells verwendet. Dieses wird für jede der k Gruppen durchgeführt und anschließend der Mittelwert aller k Qualitätswerte gebildet. Typischerweise wird $k = 10$ gewählt. Der Vorteil dieser Methode ist, dass jeder Punkt genau einmal als Testpunkt und $k - 1$ mal zur Erzeugung des Metamodells verwendet wird und somit jeder Datenpunkt gleichberüchtigt eingesetzt wird. Eine sinnvolle Erweiterung der k -fold Kreuzvalidierung ist die Erzeugung der Untergruppen in einer Weise, dass jede der k Gruppen annähernd den gleichen Mittelwert aufweist. Das kann zwar nicht garantieren, dass die k Untergruppen gleichmäßig verteilt sind, jedoch steigt die Wahrscheinlichkeit deutlich an.

Wird jeder einzelne Datenpunkt als Gruppe betrachtet ($k = n_r$), so wird von ***Leave-One-Out*** Kreuzvalidierung gesprochen. Hierbei handelt es sich um einen Sonderfall der k -fold Kreuzvalidierung, wobei das grundsätzliche Berechnungsverfahren gleich bleibt. In der Praxis wird die *Leave-One-Out* Kreuzvalidierung bei steigender Anzahl der gegebenen Datenpunkte rechenintensiv, so dass dieser Sonderfall hauptsächlich bei geringen Datenmengen angewendet wird.

Auch wenn die Kreuzvalidierung bei einem Vergleich verschiedener Metamodellsätze oder Parameter nicht garantieren kann, dass das Metamodell mit dem kleinsten mittleren MSE das für die Anwendung beste Modell darstellt, hat es sich in vielen Bereichen als sinnvolles Kriterium etabliert.

9.19 Zusammenfassung

Zur Analyse und Optimierung technischer Systeme werden immer häufiger komplexe Simulationsmodelle verwendet. Obwohl die zur Verfügung stehende Rechenleistung kontinuierlich ansteigt, ist die benötigte Rechenzeit weiterhin einer der kritischsten Punkte im Entwicklungs- und Analyseprozess. Parallel zur Rechenkapazität steigt nämlich die Komplexität der Simulationsmodelle zur Erhöhung der Simulationsqualität und der zu berücksichtigenden Faktoren ebenfalls stetig an. Eine

direkte Analyse mit einem komplexen Simulationsmodell ist dadurch in den meisten Fällen aus zeitlichem Grund nicht sinnvoll. Entsprechendes gilt natürlich auch für umfangreiche physikalische Experimente. Aus diesem Grund werden meistens Metamodelle gesucht, welche Zusammenhänge zwischen Eingangsvariablen (Faktoren) x und relevante Ausgangsvariablen y mit ausreichender Genauigkeit und minimaler Rechenzeit abbilden. Die erzeugten Modelle approximieren Systemantworten in wenigen Millisekunden, wohingegen komplexe Simulationsmodell oder Experimente Stunden, Tage oder Wochen benötigen.

Metamodelle werden auf Basis bekannter Trainingsdaten erzeugt, welche durch Messungen oder Simulationen ermittelt wurden. Dabei kann es sich um bereits vorhandene oder speziell für die Projektaufgabe ermittelte Daten handeln. Werden neue Daten erzeugt, ist es sinnvoll, speziell für die Metamodellerstellung ausgelegte Testfelder wie Latin Hypercubes (siehe Kapitel 8) zu verwenden. Sie liefern bei einer gegebenen Anzahl von bestimmbaren Trainingsdaten die maximale Informationsmenge über die gesuchten und unbekannten Zusammenhänge. In den meisten Fällen gilt dabei, dass eine Erhöhung der Datenpunkteanzahl und somit der Informationsmenge zu einem genaueren Metamodell führt.

Basierend auf den Trainingsdaten wird ein Metamodell erzeugt, welches signifikante Zusammenhänge abbildet und im Idealfall die vorhandenen Informationen optimal ausnutzt. Dabei ist eine automatische Anpassung des Metamodells an die gegebenen Trainingsdaten ohne Vorgabe eines grundlegenden Zusammenhangs zwischen Ein- und Ausgangsvariablen sinnvoll. Wird ein Modellverfahren mit Vorgabe eines grundlegenden Zusammenhangs verwendet, kann bei falscher Vorgabe des Zusammenhangs auch bei optimaler Trainingsdatenmenge kein angemessenes Metamodell erzeugt werden.

Metamodellverfahren sind in lokale und globale Methoden aufteilbar. Lokale Verfahren approximieren einen gesuchten Funktionswert y an einer Faktorkombination x_0 durch Datenpunkte in einer begrenzten Umgebung um x_0 . Globale Verfahren ermitteln hingegen eine Vorhersage von y_0 auf Basis aller Trainingsdaten.

Vor der Erzeugung eines Metamodells muss entschieden werden, ob die Vorhersage des Modells genau durch die bekannten Datenpunkte verlaufen soll oder ob die Approximation eines mittlerer Verlaufs mit Abweichungen zu den Trainingsdaten sinnvoller ist. Bei vorhandenem Rauschen in den Trainingsdaten ist der mittlere Verlauf des Zusammenhangs für eine anschließende Analyse vorzuziehen. Auch bei Verwendung von Computerexperimenten, bei denen keine direkte zufällige Streuung auftritt, kann durch kleine Änderungen der Faktoreinstellungen das Ergebnis deutlich schwanken, wenn zum Beispiel das Simulationsmodell am Ende der Simulationszeit nicht vollständig eingeschwungen ist, der Lösungsalgorithmus beziehungsweise die Simulationsschrittweite verändert wird oder andere zufällige Bestandteile im Modell enthalten sind. Auch in diesen Fällen ist die Vorhersage des mittleren Verlaufs sinnvoller.

Neben der Wahl von Faktoren inklusive dazugehöriger Variationsbreite aussagekräftigen Systemantworten sowie des Modelltyps ist die Überprüfung der Modellvorhersagegenauigkeit einer der wichtigsten Schritte in der Modellerstellung. Die Vernachlässigung der Qualitätsprüfung ist ein häufiger Grund für den missglückten

Einsatz von Metamodellen. Die Prüfung der Metamodelle kann durch die Beurteilung der Approximationsgenauigkeit von zusätzlichen Validierungsdaten erfolgen, die nicht zur Erzeugung des Metamodells verwendet wurden oder durch Verfahren wie Kreuzvalidierung, welche zur Erzeugung des Metamodells Datenpunkte auslassen, um diese anschließend zur Überprüfung zu verwenden.

Grundsätzlich sind bei allen Metamodellen Extrapolationen außerhalb der abgesicherten Trainingsdaten zu vermeiden oder nur mit Vorsicht und weiteren Kontrollen zu verwenden.

Metamodelle auf Basis von Polynomen wurden ursprünglich für Daten aus physikalischen Experimenten mit zufälligen Messfehlern eingesetzt. Durch die langjährigen Erfahrungen mit diesen Modellverfahren und der unkomplizierten Implementierung werden sie ebenfalls häufig bei Computerexperimente eingesetzt. Sie liefern bei der Modellierung von einfachen Zusammenhängen mit wenigen Faktoren und grundsätzlich bekannten Zusammenhängen ausreichend genaue Ergebnisse. Der Einsatz von Splines kann bei überschaubarer Faktoranzahl die Vorhersagequalität der Metamodelle weiter verbessern [88]. Tritt kein oder nur unwesentliches Rauschen in den Trainingsdaten auf, so wird bei stark nichtlinearen Zusammenhängen und einer Faktoranzahl $n_f < 50$ von SIMPSON et al. der Einsatz von Kriging empfohlen [88]. Künstliche Neuronale Netzwerke, welche schon lange in vielen Bereichen eingesetzt werden, liefern bei sorgfältigem Einsatz gerade bei komplexen Zusammenhängen und vielen Faktoren gute Ergebnisse. In den letzten Jahren haben sich parallel zu Künstlichen Neuronalen Netzwerken ebenfalls Algorithmen auf Basis von Support Vektor Regression und Gauß Prozess Modellen als robuste Verfahren für komplexe Zusammenhänge bewährt und etabliert. Das optimale Modellverfahren hängt von vielen Randbedingungen ab, so dass alle Empfehlungen lediglich als Hinweise gedeutet werden können. Zur Absicherung von Modellvorhersagen ist ein Vergleich oder eine Kombination verschiedener Modellverfahren sinnvoll.

Literaturverzeichnis

1. Akaike, H.: *A new look at the statistical identification model*. IEEE Transactions on Automatic Control **19**, pp. 716–723 (1974) 236
2. An, J., Owen, A.: *Quasi-regression*. J. Complexity **17**, pp. 588–607 (2001) 234
3. Bailey, T., Gatrell, A.: *Interactive spatial data analysis*. Longman Scientific & Technical (1995) 283
4. Barnes, R.J.: *The variogram sill and the sample variance*. Mathematical Geology **23**(4), pp. 673–678 (1991) 277
5. Bartels, R.H., Beatty, J.C., Barsky, B.A.: *An Introduction to Splines for Use in Computer Graphics and Geometric Modeling*. Morgan Kaufmann (1987) 254
6. Bowman, A.W., Azzalini, A.: *Applied smoothing techniques for data analysis: the Kernel approach with S-plus illustrations*. Oxford Statistical Science Series. Clarendon Press, Oxford (2004) 247
7. Bozdogan, H.: *Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions*. Psychometrika **52**, pp. 346–370 (1987) 236
8. Breiman, L.: *Heuristics of instability and stabilization in model selection*. The Annals of Statistics **24**(6), pp. 2350–2383 (1996) 313
9. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: *Classification and Regression Trees*. Chapman and Hall/CRC (1984) 248, 249, 252
10. Brent, R.P.: *Algorithms for Minimisation Without Derivatives*. Prentice Hall (1973) 307
11. Carr, J.C., Beatson, R.K., Cherrie, J., Mitchell, T.J., Fright, W.R., McCallum, B.C., Evans, T.R.: *Reconstruction and representation of 3D objects with radial basis functions*. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques, pp. 67–76. ACM, New York (2001) 290
12. CDIAC: *Atmospheric Concentrations of CO₂ from Mauna Loa, Hawaii* (2016). URL http://cdiac.ornl.gov/trends/co2/recent_mauna_loa_co2.html. (abgerufen 11/2016) 299
13. Cherkassky, V.S., Mulier, F.: *Learning from Data: Concepts, Theory and Methods*, 2nd edn. John Wiley & Sons, New York, USA (2007) 313
14. Cleveland, W.: *Robust locally weighted regression and smoothing scatter plots*. J. Amer. Stat. Assoc. **74**, pp. 829–836 (1979) 245
15. Cressie, N.: *Statistics for Spatial Data*. Wiley, New York (1993) 277
16. Cristianini, N., Shawe-Taylor, J.: *An introduction to Support Vector Machines and other kernel-based learning machines*. Cambridge University Press, Cambridge (2000) 260, 262, 265, 266, 270, 271
17. David, M., Blais, R.A.: *Geostatistical Ore Reserve Estimation*. Developments in geomathematics. Elsevier Scientific Pub. Co. (1977) 277
18. Draper, N.R., Smith, H.: *Applied Regression Analysis*, 3rd edn. Wiley (1998) 238
19. Drucker, H., Burges, C., Kaufman, L., Smola, A., Vapnik, V.: *Support Vector Regression Machines*. Advances in Neural Information Processing Systems **9**, p. 155–161 (1997) 258
20. Dumouchel, W., O'Brien, F.: *Computing and graphics in statistics*. In: A. Buja, P.A. Tukey (eds.) Computing and graphics in statistics, chap. Integrating a robust option into a multiple regression computing environment, pp. 41–48. Springer-Verlag New York (1991) 239
21. Duong, T., Hazelton, M.: *Plug-in Bandwidth Selectors for Bivariate Kernel Density Estimation*. Journal of Nonparametric Statistics **15**, pp. 17–30 (2003) 247
22. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: *Least Angle Regression*. Annals of Statistics **32**, pp. 407–499 (2002). URL http://www-stat.stanford.edu/~tibs/ftp/LeastAngle_2002.pdf. (abgerufen 11/2016) 238
23. Efroymson, M.: *Mathematical Methods for Digital Computers*, chap. Multiple regression analysis. John Wiley & Sons Inc (1960) 236
24. Eilers, P.H.C., Rijnmond, D.M., Marx, B.D.: *Flexible smoothing with B-splines and penalties*. Statistical Science **11**, pp. 89–121 (1996) 254

25. Fang, K.T., Li, R., Sudjianto, A.: *Design and Modeling for Computer Experiments (Computer Science & Data Analysis)*. Chapman & Hall/CRC (2005) 192, 193, 194, 195, 197, 209, 210, 212, 213, 214, 219, 238, 254, 255, 417, 418, 419
26. Fletcher, R.: *Practical Methods of Optimization*, 2nd edn. Wiley (2000) 272
27. Fletcher, T.: *Support Vector Machines Explained* (2009). URL <http://www.tristanfletcher.co.uk>. (abgerufen 11/2016) 259, 260, 261, 262, 263, 264, 265, 266, 267, 269
28. Frank, I.E., Friedman, J.H.: *A statistical view of some chemometrics regression tools*. *Tech-nometrics* **35**, pp. 109–148 (1993) 238
29. Friedman, J.: *Multivariate Adaptive Regression Splines (with discussion)*. *Annals of Statistics* **19**, pp. 1–141 (1991) 256, 257, 258
30. Friedman, J.: *Fast MARS*. Tech. Rep. 110, Stanford University Department of Statistics (1993) 256, 257, 258
31. Graham, J.: *Ordinary Kriging* (2015). URL <http://www.math.umt.edu/graham>. (abgerufen 11/2016) 283
32. Graham, J.: *Trend, Variation and Universal Kriging* (2015). URL <http://www.math.umt.edu/graham>. (abgerufen 11/2016) 283
33. Gunn, S.R.: *Support Vector Machines for Classification and Regression*. Tech. rep., University of Southampton (1998). URL <http://users.ecs.soton.ac.uk/srg/publications/pdf/SVM.pdf>. (abgerufen 11/2016) 268, 400
34. Guyon, I.: *An introduction to variable and feature selection*. *Journal of Machine Learning Research* **3**, pp. 1157–1182 (2003) 313
35. Härdle, W.: *Applied Nonparametric Regression*. Cambridge University Press (1990) 244
36. Hastie, T., Loader, C.: *Local regression: automatic kernel carpentry (with discussion)*. *Statistical Science* **8**, pp. 120–143 (1993) 245
37. Hastie, T., Tibshirani, R.: *Generalized Additive Models*. Chapman and Hall, London (1990) 235
38. Hengl, T.: *Universal Kriging - Algorithm* (2013). URL http://spatial-analyst.net/ILWIS/htm/ilwisapp/universal_kriging_algorithm.htm. (abgerufen 11/2016) 283
39. Hesterberg, T., Choi, N.H., Meier, L., Fraley, C.: *Least angle and L1 penalized regression: A review*. *Statistics Surveys* **2**, pp. 61–93 (2008) 238
40. Hoerl, A.E., Kennard, R.W.: *Ridge regression: Biased estimation for nonorthogonal problems*. *Technometrics* **12**, pp. 55–67 (1970) 238
41. Holland, P.W., Welsch, R.E.: *Robust Regression Using Iteratively Reweighted Least-Squares*. *Communications in Statistics: Theory and Methods* **A6**, pp. 813–827 (1977) 239
42. Host, G.: *Kriging by local polynomials*. *Computational Statistics and Data Analysis* **29**(3), pp. 295–312 (1999). URL <http://ideas.repec.org/a/eee/csdana/v29y1999i3p295-312.html>. (abgerufen 11/2016) 247
43. Huber, P., Ronchetti, E.M.: *Robust Statistics*. Wiley (2009) 239, 383
44. Huth, F.: *Simulation eines Hochdruck-Benzineinspritzsystems*. Master's thesis, RWTH Aachen (2011) 191, 193, 287
45. Ishiguro, M., Sakamoto, Y., Kitagawa, G.: *Bootstrapping log likelihood and EIC, an extension of AIC*. *Annals of the Institute of Statistical Mathematics* **49**, pp. 411–434 (1997) 236
46. Jekabsons, G.: *Ensembling adaptively constructed polynomial regression models*. *International Journal of Intelligent Systems and Technologies (IJIST)* **3**(2), pp. 56–61 (2008) 242, 313
47. Jekabsons, G.: *Machine Learning*, chap. Adaptive Basis Function Construction: an approach for adaptive building of sparse polynomial regression models, pp. 127–156. InTech (2010) 235, 242, 313
48. JiSheng Hao, L.M., Wang, W.: *IFIP International Federation for Information Processing*, vol. 288, chap. An New Algorithm for Modeling Regression Curve, pp. 86–91. Springer (2009) 269
49. Jones, M., Marron, J., Sheather, S.: *A Brief Survey of Bandwidth Selection for Density Estimation*. *Journal of the American Statistical Association* **91**, pp. 401–407 (1996) 247

50. Journel, A.G., Huijbregts, C.J.: *Mining Geostatistics*. Academic Press, London (1978) 277
51. King, M.L., Zhang, X., Hyndman, R.J.: *Bandwidth Selection for Multivariate Kernel Density Estimation Using MCMC*. Computational Statistics and Data Analysis **50**, pp. 3009–3031 (2004) 247
52. Kohavi, R., John, G.H.: *Wrappers for Feature Subset Selection*. Artificial Intelligence **97**, pp. 273–324 (1997) 313
53. Kotsiantis, S.B., Pintelas, P.E.: *Combining Bagging and Boosting*. International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering **1**(8), pp. 372–381 (2007) 313
54. Kürz, C.: *Radiale-Basis-Funktionen*. Seminararbeit, Fraunhofer-Institut für Algorithmen und Wissenschaftliches Rechnen (2008). (abgerufen 11/2016) 284, 287
55. Lichtenstern, A.: *Kriging methods in spatial statistics*. Master's thesis, Technische Universität München (2013) 283
56. Loughrey, J., Cunningham, P.: *Overfitting in Wrapper-Based Feature Subset Selection: The Harder You Try the Worse it Gets*. In: M. Bramer, F. Coenen, T. Allen (eds.) SGAI Conf., pp. 33–43. Springer (2004) 313
57. Mallows, C.: *Some Comments on Cp*. Technometrics **42**, pp. 87–94 (2000) 236
58. Marron, J.S., Nolan, D.: *Canonical kernels for density estimation*. Statistics & Probability Letters **7**(3), pp. 195–199 (1988). URL <http://ideas.repec.org/a/eee/stapro/v7y1988i3p195-199.html>. (abgerufen 11/2016) 244
59. Martinez, W.L., Martinez, A.R.: *Computational Statistics Handbook with Matlab*. Chapman & Hall/CRC (2002) 248, 249, 252
60. Masters, T.: *Practical Neural Network Recipes in C++*. Academic Press (1993) 303, 304, 307, 309
61. Mathworks: *Matlab Dokumentation* (2015). URL <https://de.mathworks.com/help/matlab/>. (abgerufen 03/2017) 236, 239, 240, 241, 384
62. Menzel, W., Zhang, J., Hendrich, N.: *Algorithmisches Lernen - Support Vector Machines*. Vorlesungsunterlagen (2009). URL http://tams-www.informatik.uni-hamburg.de/lectures/2009ss/vorlesung/Algorithmisches_Lernen/. (abgerufen 11/2016) 265
63. Miller, A.: *Subset Selection in Regression*. CRC Press Inc (2002) 238
64. Montgomery, D.C.: *Design and Analysis of Experiments*. John Wiley and Sons, Hoboken, NJ (2001/2009) 90, 129, 234
65. Nadaraya, E.: *On Estimating Regression*. Theory Probab. Appl. **9**, pp. 141–142 (1964) 243
66. Nadaraya, E.: *On Non-Parametric Estimates of Density Functions and Regression Curves*. Theory Probab. Appl. **10**, pp. 186–190 (1965) 243
67. Navratil, G.: *Ausgleichsrechnung II*. Vorlesungsunterlagen (2006). URL <ftp://geoinfo.tuwien.ac.at/navratil/Ausgleich2.pdf>. (abgerufen 11/2016) 276, 277, 283
68. Neal, R.M.: *Bayesian Learning for Neural Networks*. Springer-Verlag New York (1996) 300
69. Opitz, D., Maclin, R.: *Popular Ensemble Methods: An Empirical Study*. Journal of Artificial Intelligence Research **11**, pp. 169–198 (1999) 313
70. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press (2007) 199, 203, 220, 254, 265, 276, 277, 281, 285, 307, 334, 381, 398, 400
71. Pudil, P., Ferri, F., Novovicova, J., Kittler, J.: *Floating search methods for feature selection with nonmonotonic criterion functions*. In: Proceedings of International Conference on Pattern Recognition, vol. 2, pp. 279–283 (1994) 236
72. Rasmussen, C.E.: *Solving Challenging Non-linear Regression Problems by Manipulating a Gaussian Distribution*. Machine Learning Summer School, Cambridge (2009). URL http://mlg.eng.cam.ac.uk/mlss09/mlss_slides/Rasmussen_1.pdf. (abgerufen 11/2016) 296, 299, 301, 302
73. Rasmussen, C.E.: *GPML: Gaussian Process for Machine Learning (Matlab Toolbox)* (2013). URL <http://www.gaussianprocess.org/gpml/code/matlab/doc/>. (abgerufen 11/2016) 297, 301, 302

74. Rasmussen, C.E., Williams, C.K.I.: *Gaussian Processes for Machine Learning*. The MIT Press (2006). URL <http://www.gaussianprocess.org/gpml/>. (abgerufen 11/2016) 291, 292, 293, 294, 295, 297, 300, 301, 302
75. Rathbun, S.: *The universal kriging equations* (1994). URL <http://www.esapubs.org/archive/ecol/E079/001/kriging.htm>. (abgerufen 11/2016) 283
76. Reunanen, J.: *Overfitting in Making Comparisons Between Variable Selection Methods*. Journal of Machine Learning Research **3**, pp. 1371–1382 (2003). URL <http://www.jmlr.org/papers/volume3/reunanen03a/reunanen03a.pdf>. (abgerufen 11/2016) 313
77. Reunanen, J.: *Feature extraction: foundations and applications*, chap. Search strategies, pp. 119–137. Springer (2006) 235, 313
78. Rissanen, J.: *Modelling by Shortest Data Description*. Automatica **14**, pp. 465–471 (1978) 236
79. Rodriguez, C.C.: *The ABC of Model Selection: AIC, BIC and the New CIC*. In: K.H. Knuth, A.E. Abbas, R.D. Morris, J.P. Castle (eds.) Bayesian Inference and Maximum Entropy Methods in Science and Engineering. *American Institute of Physics Conference Series*, vol. 803, pp. 80–87 (2005) 236
80. Runarsson, T.P., Sigurdsson, S.: *Support Vector Machines*. Vorlesungsunterlagen (2003). URL <http://notendur.hi.is/~tpr/tutorials/svm/fyrirlestrar.html>. (abgerufen 11/2016) 258, 259, 260, 261, 266, 268, 269, 271, 274
81. Runarsson, T.P., Sigurdsson, S.: *Support Vector Machines - kernel-based learning methods, computational/statistical learning theory*. University of Iceland (2003). URL <https://notendur.hi.is/tpr/tutorials/svm/notes/>. (abgerufen 11/2016) 269, 270
82. Sachs, L., Hedderich, J.: *Angewandte Statistik, Methodensammlung mit R*. Springer-Verlag Berlin Heidelberg (2009) 68, 234
83. Sain, S., Baggerly, K., Scott, D.: *Cross-Validation of Multivariate Densities*. Journal of the American Statistical Association **89**, pp. 807–817 (1994) 247
84. SAS: *SAS/STAT 9.2 User's Guide - The Variogram Procedure* (2009). URL <http://support.sas.com/documentation/>. (abgerufen 11/2016) 276, 277
85. Saunders, C., Gammerman, A., Vovk, V.: *Ridge Regression Learning Algorithm in Dual Variables*. In: 15th International Conference on Machine Learning (1998) 273
86. Schölkopf, B., Bartlett, P., Smola, A., Williamson, R.: *Shrinking the Tube: A New Support Vector Regression Algorithm*. In: Advances in Neural Information Processing Systems, vol. 11, pp. 330–336. MIT Press (1999) 269
87. Schwarz, G.: *Estimating the Dimension of a Model*. The Annals of Statistics **6**, pp. 461–464 (1978) 236
88. Simpson, T., Lin, D., Chen, W.: *Sampling strategies for computer experiments: design and analysis*. International Journal of Reliability and Safety (IJRS) **2**(3), pp. 209–240 (2001) 318
89. Smola, A.J., Schölkopf, B.: *A Tutorial on Support Vector Regression*. Statistics and Computing pp. 199–222 (2004) 258, 265
90. Söndergerath, D.: *Multivariate Statistik*. TU-Braunschweig (2007). Vorlesungsunterlagen 279
91. Spiegelhalter, D., Best, N., Carlin, B., van der Linde, A.: *Bayesian Measures of Model Complexity and Fit (with Discussion)*. Journal of the Royal Statistical Society **64**, pp. 583–616 (2002) 236
92. Stone, C.J., Hansen, M., Kooperberg, C., Truong, Y.K.: *Polynomial splines and their tensor products in extended linear modeling*. Ann. Statist **25**, pp. 1371–1470 (1997) 254, 255
93. Street, J.O., Carroll, R.J., Ruppert, D.: *A Note on Computing Robust Regression Estimates Via Iteratively Reweighted Least Squares*. The American Statistician **42**(2) (1988) 239
94. Suykens, J., Brabanter, K.D., Karsmakers, P., Ojeda, F., Alzate, C., Brabanter, J.D., Pelckmans, K., Moor, B.D., Vandewalle, J.: *LS-SVMlab Toolbox User's Guide*. Technical report, Katholieke Universiteit Leuven (2011). URL <http://www.esat.kuleuven.be/sista/lssvmlab/>. (abgerufen 11/2016) 272
95. Suykens, J.A.K., Vandewalle, J.: *Least Squares Support Vector Machine Classifiers*. Neural Processing Letters **9**(3), pp. 293–300 (1999) 272

96. T. Hastie, R.T., Friedman, J.: *The Elements of Statistical Learning Data Mining, Inference and Prediction*. Springer (2001) 256
97. Tibshirani, R.: *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society **58**, pp. 267–288 (1996). URL <http://www-stat.stanford.edu/~tibs/lasso/lasso.pdf>. (abgerufen 11/2016) 238
98. Tibshirani, R.: *The LASSO method for variable selection in the COX model*. Statistics in medicine **16**, pp. 385–395 (1997) 238
99. du Toit, W.: *Radial Basis Function Interpolation*. Master's thesis, University of Stellenbosch (2008). URL <http://scholar.sun.ac.za/handle/10019.1/2002>. (abgerufen 11/2016) 284, 285
100. Vapnik, V., Cortes, C.: *Support Vector Networks*. Machine Learning **20**, pp. 273–297 (1995). URL <http://www.springerlink.com/content/k238jx04hm87j80g/>. (abgerufen 11/2016) 258
101. Vapnik, V.N.: *The Nature of statistical learning theory*. Springer, New York (2000) 258
102. Venkataraman, P.: *Applied Optimization with MATLAB Programming*, 2nd edn. Wiley Publishing, Hoboken (2009) 261
103. Wackernagel, H.: *Multivariate geostatistics - an introduction with applications*, 2nd edn. Springer, Berlin (1998) 277
104. Wahba, G.: *Splines Models for Observational Data*. Series in Applied Mathematics **59** (1990) 254, 302
105. Wand, M., M.C.Jones: *Multivariate Plug-in Bandwidth Selection*. Computational Statistics **9**, pp. 97–116 (1994) 247
106. Watson, G.: *Smooth regression analysis*. Sankhyā: The Indian Journal of Statistics A **26**, pp. 359–372 (1964) 243
107. Welling, M.: *Kernel Ridge Regression*. Paper (2005) 273
108. Wendt, H.: *Support Vector Machines for Regression Estimation and their Application to Chaotic Time Series Prediction*. Master's thesis, Technische Universität Wien (2005) 266, 267
109. Wermuth, N.: *Beobachtungen zur Ridge-Regression*. Jahrbücher für Nationalökonomie und Statistik **189**, pp. 300–307 (1975) 238
110. Williams, C.K.I., Rasmussen, C.E.: *Gaussian Processes for Regression*. In: Advances in Neural Information Processing Systems, vol. 8, pp. 514–520 (1996). URL <http://citeseeerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.25.8841>. (abgerufen 11/2016) 300
111. Wilson, A.G., Adams, R.P.: *Gaussian process kernels for pattern discovery and extrapolation* (2013) 297

Kapitel 10

Optimierung

10.1 Einleitung

Neben Analysen zur Bestimmung der Zusammenhänge zwischen kontrollierbaren Eingangsvariablen (Faktoren) und Ausgangsgrößen (Qualitätsmerkmalen) eines technischen Systems ist in fast allen Fällen eine Optimierung von einer oder mehreren Systemeigenschaften notwendig. Dabei sind Faktoreinstellungen gesucht, welche ausgewählte Zielgrößen (Qualitätsmerkmale) z_1, \dots, z_{n_z} für den angestrebten Systemeinsatz optimieren. Die Zielgrößen sollen dabei einen gewünschten Wert annehmen beziehungsweise minimiert oder maximiert werden. Zur Vereinfachung der eingesetzten Algorithmen können alle Optimierungsaufgaben in Minimierungsaufgaben umgewandelt werden. .

Optimierungsziel	Minimierung	
$\min [z(x)]$	$\Rightarrow \min [z(x)]$	
$\max [z(x)]$	$\Rightarrow \min [-z(x)]$	
$z(x) = c$	$\Rightarrow \min c - z(x) $	

(10.1)

Einige Optimierungsaufgaben lassen sich exakt lösen, wobei der Rechenaufwand schnell ansteigen kann und dadurch die exakte Lösung nicht sinnvoll beziehungsweise praktikabel ist. In vielen Anwendungen reicht es jedoch aus, Lösungen in der Nähe des globalen Optimums zu finden, wenn dieses mit akzeptablen Rechen- und Zeitaufwand geschieht. Hierzu können verschiedene heuristische Optimierungsverfahren eingesetzt werden. *Heuristik* stammt aus dem Griechischen (heuriskein: Finden) und fasst analytische Verfahren zusammen, die auf Basis von unvollständigem Wissen und Mutmaßungen in kurzer Zeit Schlussfolgerungen über ein zu optimierendes System treffen. Im Gegensatz zu einer exakten Lösung kann ein heuristisches Verfahren jedoch nicht garantieren, dass es sich wirklich um die optimalen Lösungen handelt [25]. Metaheuristik beschreibt, im Gegensatz zu problemspezifischen Heuristiken, abstrakte Algorithmen beziehungsweise Schritte, die theoretisch auf beliebige Problemstellungen anwendbar sind. Unter Metaheuristiken fallen be-

kannte Methoden wie Simuliertes Abkühlen (Kapitel 8.4) sowie genetische- oder naturinspirierte Optimierungsverfahren.

Bei der Betrachtung mehrerer Zielgrößen ist es nur in den wenigsten Fällen möglich eine Faktorkombination zu finden, die alle Zielgrößen gleichzeitig minimiert (optimiert), so dass nach einem optimalen Kompromiss zwischen allen Zielgrößen (und den dazugehörigen Faktoreinstellungen) gesucht wird. Abbildung 10.1 zeigt dazu ein fiktives Beispiel mit zwei Zielgrößen und zwei ausgewählten Faktoren: Gesucht ist ein Computersystem, welches eine Simulationsaufgabe in einer möglichst kurzen Zeit berechnet und gleichzeitig kostengünstig ist. Dabei werden beide Zielgrößen unter anderem durch die Anzahl der Prozessoren und die Rechengeschwindigkeit jedes einzelnen Prozessors beeinflusst. Zwischen den betrachteten Zielgrößen muss ein Kompromiss gefunden werden, wobei eventuell nicht nur die beiden Zielgrößen (Preis, Rechenzeit), sondern ebenfalls die dazugehörigen Faktoreinstellungen zu berücksichtigen sind. So kann bei ähnlichen Eigenschaften für die Zielgrößen (ähnlicher Preis und Rechengeschwindigkeit) zum Beispiel ein Einzelrechner einem Rechencluster vorgezogen werden.

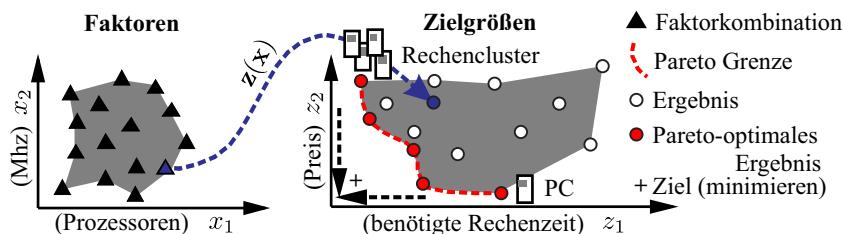


Abb. 10.1 Optimierung mehrerer Zielgrößen

Im ersten Schritt des Optimierungsprozesses mehrerer Zielgrößen sind Lösungen gesucht, die auf oder in der Nähe der *Pareto-Grenze* liegen, welche durch *Pareto-optimale* Ergebnisse definiert wird. Ein Pareto-optimales Ergebnis zeichnet sich dadurch aus, dass keine der betrachteten Zielgröße verbessert (verringert) werden kann ohne eine andere Zielgröße zu verschlechtern. Somit sind alle Ergebnisse auf der Pareto-Grenze als 'optimal' anzusehen. Sind genügend Punkte auf der Pareto-Grenze bestimmt, wird im zweiten Schritt ein akzeptabler Kompromiss zwischen den Zielgrößen unter Berücksichtigung der Faktoreinstellungen gewählt.

10.2 Dominanz

Zur Bestimmung, ob ein Datenpunkt x_0 mit den dazugehörigen Zielgrößen $z_i(x_0)$ zur gesuchten Pareto-Grenze gehört, wird überprüft, ob der Datenpunkt im Zielgrößenraum von anderen Punkten *dominiert* wird beziehungsweise, welche anderen Datenpunkte von ihm *dominiert* werden. Dabei dominiert ein Punkt a einen Punkt b

($a \prec b$: Pareto Dominanz), wenn er in mindestens einer Zielgröße besser und in allen anderen mindestens gleichwertig ist (Abbildung 10.2). Im Gegensatz dazu wird ein Punkt a von einem Punkt b dominiert wenn er in mindestens einer Zielgröße schlechter und in allen anderen Zielgrößen gleichwertig oder schlechter ist.

$$a \prec b : \begin{cases} z_i(a) \leq z_i(b) & \forall i = 1 \dots n_z \\ z_j(a) < z_j(b) & \exists j \in 1 \dots n_z \end{cases} \quad (10.2)$$

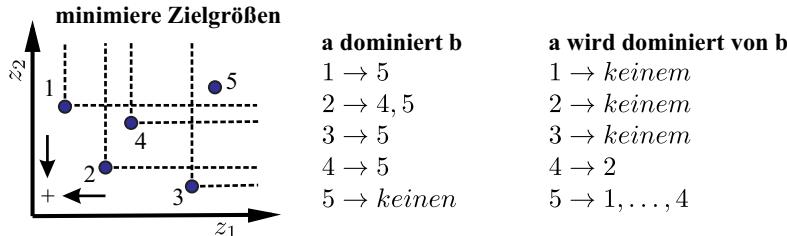


Abb. 10.2 Dominanz von Datenpunkten im Zielgrößenraum (minimieren, $n_z = 2$)

Wird nur ein Individuum (Datenpunkt) I_i betrachtet, existiert ein eindeutiger Bereich im Zielgrößenraum, welcher durch das Individuum I_i dominiert wird und ein zweiter welcher Individuen enthält, die das Individuum I_i dominieren (Abbildung 10.3). In allen anderen Bereichen kann keine eindeutige Aussage über die Dominanz des Individuums I_i getroffen werden, da einige Zielgrößen besser und andere schlechter sind.

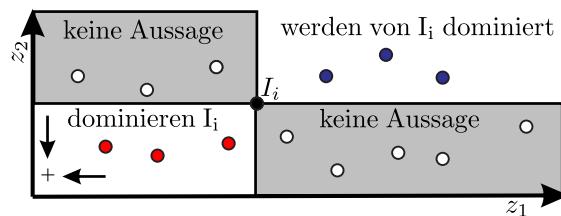


Abb. 10.3 Dominanz eines Individuums I_i (minimieren, $n_z = 2$)

Angelehnt an Veröffentlichungen von ZITZLER [62] können noch verschiedene Varianten der Dominanz definiert werden (Tabelle 10.1). Es ist jedoch zu beachten, dass unterschiedliche Veröffentlichungen und besonders andere Arbeitsgebiete, wie zum Beispiel die Spieltheorie, leicht andere Bezeichnungen und Definitionen der Dominanz verwenden.

strikte Dominanz	$z_1 \prec z_2$	$z_i(a) < z_i(b) \forall i = 1 \dots n_z$
Dominanz	$z_1 \prec z_2$	$z_i(a) \leq z_i(b) \forall i = 1 \dots n_z$ $z_j(a) < z_j(b) \exists j \in 1 \dots n_z$
schwache Dominanz	$z_1 \preceq z_2$	$z_i(a) \leq z_i(b) \forall i = 1 \dots n_z$
nicht vergleichbar	$z_1 \parallel z_2$	keine schwache Dominanz von z_1 und z_2

Tabelle 10.1 Varianten der Dominanz

10.2.1 Priorität und Grenzwert

In der klassischen Dominanzbetrachtung sind alle Zielgrößen gleichwertig, so dass keine Priorisierung der Zielgrößen während einer Optimierung durchgeführt werden kann. In der Praxis finden sich jedoch Anwendungen, in denen einige Zielgrößen wichtiger sind als andere, so dass eine Priorisierung sinnvoll erscheint. Weiterhin können Randbedingungen (siehe auch Kapitel 10.3) für Zielgrößen als unabhängige Optimierungsgrößen betrachtet werden, die jedoch lediglich bis zu einem vorgegebenen Grenzwert optimiert werden müssen und eine weitere Verbesserung zu gleichwertigen Ergebnissen führt. Wenn zum Beispiel eine Zielgröße z_1 kleiner gleich dem Wert $z_1 \leq 5$ sein muss, ist $z_1 = 5$ besser als $z_1 = 6$ aber $z_1 = 4$ kann als gleichwertig zu $z_1 = 5$ betrachtet werden. FONSECA führt zur Berücksichtigung von Prioritäten und Grenzwerten neben der standardmäßigen Zielgrößenmatrix $Z_{n_r \times n_z}$ einen Prioritäts- und Grenzwertvektor $p_{1 \times n_z}$ und $g_{1 \times n_z}$ ein [23, 24, 36]. Der Prioritätsvektor enthält ganzzahlige Werte $(1, 2, 3, \dots)$, wobei größere Werte höhere Prioritäten kennzeichnen. Typischerweise werden Randbedingungen große Prioritäten zugewiesen, da diese zuerst erfüllt werden müssen, bevor die eigentlichen Zielgrößen optimiert werden. Verschiedene Zielgrößen können gleiche Prioritäten aufweisen. So könnten im einfachsten Fall allen Randbedingungen die Priorität $p = 2$ und allen Optimierungsgrößen der Wert $p = 1$ zugewiesen werden. Der Grenzwertvektor enthält die Grenzwerte beziehungsweise Randbedingungen der Zielgrößen. Im Fall von Optimierungsgrößen ohne Grenzwert wird ein kleiner nicht erreichbarer Wert vorgegeben ($g \rightarrow -\infty$). Ob ein Individuum a einem Individuum b vorzuziehen ist $a \prec_g b$ wird mit folgenden Gleichungen bestimmt [23, 24, 36]:

$$a \prec_g b \left\{ \begin{array}{l} \left(a_p^a \prec b_p^a \right) \vee \\ \left(\left(a_p^a = b_p^a \right) \wedge \left[\left(b_p^a \not\leq g_p^a \right) \vee \left(a_p^a \prec b_p^a \right) \right] \right) \\ \left(a_p^a \prec b_p^a \right) \vee \\ \left(\left(a_p^a = b_p^a \right) \wedge \left[\left(b_p^a \not\leq g_p^a \right) \vee \left(a_{1\dots p-1}^a \prec_g b_{1\dots p-1}^a \right) \right] \right) \end{array} \right\} \begin{array}{l} \text{für } p = 1 \\ \\ \\ \text{für } p > 1 \end{array} \quad (10.3)$$

In Gleichung 10.3 bezeichnet der Term b_p die Zielgrößen des Individuum b mit der Priorität p . Der Zusatz $\overset{a}{\prec}$ grenzt die Zielgrößen auf diejenigen ein, bei denen das Individuum a die gegebenen Grenzwerte nicht erfüllt. Im Gegensatz dazu verweist $\overset{a}{\succ}$ auf die Zielgrößen bei denen Individuum a die Grenzwerte erfüllt. Die Kennung

b_p^a bezeichnet somit alle Zielgrößen von b , die eine Priorität von p aufweisen und bei denen das Individuum a die Grenzwerte der Zielgröße nicht erreicht. Durch \prec wird weiterhin ein klassischer Dominanztest gekennzeichnet und durch \prec_g der Dominanztest mit Priorität und Grenzwerte. Die zweite Zeile ($p > 1$) bedeutet somit: a ist b vorzuziehen, wenn

- 1) für alle Zielgrößen mit Priorität p und bei denen Individuum a die Grenzwerte nicht erreicht das Individuum a klassisch dominant zu b ist, **oder**
- 2) alle Zielgrößen mit Priorität p und bei denen Individuum a die Grenzwerte nicht erreicht das Individuum a identisch zu b ist **und gleichzeitig**
- 2.1) eine Zielgröße von Individuum b den Grenzwert nicht erreicht, wobei a den Grenzwert erreicht hat, oder
- 2.2) Individuum a dem Individuum b vorzuziehen ist, wenn die kleineren Prioritäten $1 \dots p-1$ berücksichtigt werden.

Durch die Einführung von Prioritäten und Grenzwerten können zwei unterschiedliche Individuen gleichwertig sein. Dieses tritt zum Beispiel auf, wenn zwei Individuen alle Randbedingungen (Optimierungsgroßen mit Grenzwert $\neq -\infty$) erfüllen und sich diese nur in den Zielgrößen mit höheren Prioritäten ($p > 1$, Randbedingungen) unterscheiden. Die Gleichwertigkeit von a und b ($a \equiv_g b$) wird durch Gleichung 10.4 bestimmt.

$$\left(a \overset{a}{\prec} b \right) \wedge \left(a_1 \overset{a}{\prec} b_1 \right) \wedge \left(b_{2 \dots p} \overset{a}{\leq} g_{2 \dots p} \right) \quad (10.4)$$

Abbildung 10.4 zeigt die Veränderung der Bevorzugung von Ergebnissen wenn die Zielgröße z_2 eine höhere Priorität aufweist als z_1 und beide Zielgrößen minimiert werden sollen ($g_i = -\infty$). Obwohl Punkt 3 lediglich Punkt 5 klassisch dominiert, ist er durch die Priorisierung ebenfalls den Punkten 1, 2 und 4 vorzuziehen. Untersuchen wir beispielhaft die Frage ob der Punkt drei (a) dem Punkt vier (b) vorzuziehen ist $a \prec_g b$. Im ersten Schritt ist $p = \max(p) = 2$ zu wählen, so dass in diesem Beispiel lediglich die Zielgröße z_2 betrachtet wird. a erfüllt die Zielgröße g_2 nicht, so dass z_2 in die Gruppe $\overset{a}{\prec}$ fällt. Da nur eine Zielgröße mit der Priorität $p = 2$ existiert, bleibt die Gruppe $\overset{a}{\prec}$ leer. In der Gruppe $\overset{a}{\leq}$, welche nur z_2 betrachtet, dominiert Punkt a den Punkt b , so dass die Bedingung $\left(a_p^a \prec b_p^a \right)$ bereits erfüllt ist und somit a , also Punkt 3, vorzuziehen ist.

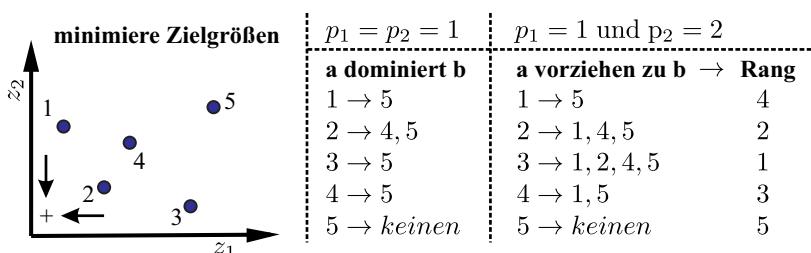


Abb. 10.4 Vergleich klassischer Dominanz und Dominanz mit Priorität

Der Rang eines Datenpunktes a , welcher zur Sortierung der Datenpunkte untereinander herangezogen wird, ist bei Verwendung der Dominanz mit Priorisierung gleich der Anzahl der Datenpunkte, die dem Datenpunkt vorzuziehen sind plus eins. Im Vergleich zur Bestimmung des Rangs in Kapitel 9 können in Spezialfällen bei gleicher Priorität für alle Zielgrößen und Grenzwerte von $p = -\infty$ leicht unterschiedliche Ergebnisse auftreten. Zur weiteren Veranschaulichung des Rangs zeigt Abbildung 10.5 die Veränderung des Rangs bei unterschiedlichen Grenzwerten für z_1 und z_2 . Im Fall $g_{1,2} = 2.7$ und $p_2 > p_1$ existieren beispielsweise zwei Einträge mit dem Rang eins. Des Weiteren ist der Punkt a an der Position $z = [2, 1]$ dem Punkt b mit $z = [2, 2]$ nicht vorzuziehen und nach Gleichung 10.4 gleichwertig. Die Zielgröße z_2 weist die höchste Priorität mit $p = 2$ auf und a erreicht in dieser Zielgröße den Grenzwert. Somit ist die Menge $a_2^{\frac{a}{2}}$ leer und die Dominanzbetrachtung $(a_p^{\frac{a}{p}} \prec b_p^{\frac{a}{p}})$ als falsch zu deklarieren. Im Gegensatz dazu können die leeren Mengen aber als gleich betrachtet werden, so dass die Gleichheitsbedingung $(a_p^{\frac{a}{p}} = b_p^{\frac{a}{p}})$ wahr ist. Alle Zielgrößen mit der Priorität $p = 2$ (also z_2) werden aber auch von b erfüllt, so dass die Bedingung $(b_p^{\frac{a}{p}} \not\leq g_p^{\frac{a}{p}})$ falsch ist. Im nächsten Schritt wird geprüft ob a bei der Betrachtung aller Prioritäten $p < 2$, also lediglich $p = 1$, dem Punkt b vorzuziehen ist. Da auch z_1 von beiden erfüllt ist sind die ersten drei Bedingungen aus Gleichung 10.3 für $p = 1$ [falsch, wahr, falsch], so dass der letzte Term $(a_p^{\frac{a}{p}} \prec b_p^{\frac{a}{p}})$ entscheidet, ob a vorzuziehen ist. Da a jedoch in der Zielgröße z_1 den Punkt b nicht dominiert, ist a nicht vorzuziehen. Die Gleichwertigkeit der beiden Punkte ist jedoch durch Gleichung 10.4 gegeben. Als weiteres soll der Punkt $z = [3, 2]$ betrachtet werden. Dieser ist allen Punkten mit $z_2 > 2.7$ (Grenzwert) oder $z_1 > 3$ vorzuziehen. Gleichwertig ist er zu allen Punkten mit $z_1 = 3$ und $z_2 < 2.7$. Da die Berechnung des Rangs für eine Datenmenge im einfachsten Fall die priorisierte Dominanz für jede Punktkombination berechnet und somit schnell einen hohen Rechenaufwand erzeugt, zeigt Lygoe [36, 23, 24] eine optimierte Rechenmethode mit der Bezeichnung *Progressive Preference Articulation Method of Fonseca and Fleming (PPAFF)*. Leicht abgewandelt ist die Berechnung in Algorithmus 7 dargestellt.

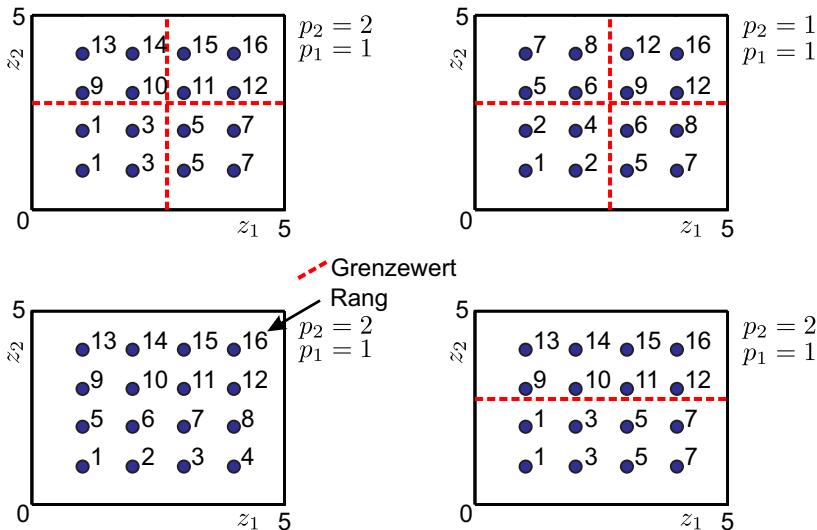


Abb. 10.5 Rang bei unterschiedlichen Prioritäten und Grenzwerten

- 1 Erzeuge einen Vektor R der Länge n_R , welcher für jeden Datenpunkt den aktuellen Rang=1 enthält.
- 2 **für alle Datenpunkte Z_i , ($i = 1 \rightarrow n_R$) tue**
 - 3 Erzeuge Testmenge T , die aus allen Datenpunkten bis auf Z_i besteht
 - 4 Erzeuge leere Menge D , die alle Datenpunkte enthalten soll, die Individuum Z_i dominiert
 - 5 **für alle Prioritäten p_j (von der höchsten bis zur kleinsten) tue**
 - 6 Bestimme alle Zielgrößen Z_p mit aktueller Priorität p_j
 - 7 Bestimme die Untermenge Z_p^\sim von Z_p bei der Z_i die Grenzwerte nicht erfüllt
 - 8 (1) Prüfe Dominanz von Z_i in der Untermenge der Ziele Z_p^\sim für jedes Element der aktuellen Testmenge T : $a \prec b$
 - 9 Füge Punkte aus T , die (1) erfüllen, der Menge D zu (2) Prüfe ob Z_i in der Untermenge Z_p^\sim identisch mit Elementen der aktuellen Testmenge T ist: $a = b$
 - 10 Bestimme die Untermenge Z_p^\sim von Z_p bei der Z_i die Grenzwerte erfüllt
 - 11 **wenn $Z_p^\sim \neq \emptyset$ dann**
 - 12 (3) berechne für jedes Element aus der Testmenge T ob mindestens eine Zielgröße aus der Menge Z_p^\sim nicht erreicht wird: $b \not\leq g$
 - 13 Füge Punkte aus T , die (2) und (3) erfüllen, der Menge D zu
 - 14 **Ende**
 - 15 Behalte lediglich Punkte in Testmenge T , die im letzten Test nicht dominiert wurden und Bedingung (2) erfüllen
 - 16 **Ende**
 - 17 **wenn $T \neq \emptyset \wedge Z_1^\sim \neq \emptyset$ dann**
 - 18 Prüfe Dominanz von Z_i zu den Elementen aus T bei Berücksichtigung von Z_1^\sim und füge dominante Elemente aus T der Menge D hinzu.
 - 19 **Ende**
 - 20 Der Rang R wird für alle Elemente aus D um eins erhöht.
 - 21 **Ende**

Algorithmus 7 : PPA_{FF}(modifiziert)

10.3 Randbedingungen

Neben der Berücksichtigung von Randbedingungen in der priorisierten Dominanzberechnung (siehe Kapitel 10.2.1) können diese auch separat betrachtet werden. Die Integration komplexer Randbedingungen für Faktoren und Zielgrößen ist in den meisten Optimierungsverfahren möglich. Zur Verallgemeinerung wird für jede Randbedingung $g(x, z)$, welche eine Funktion der Faktoren und Zielgrößen sein kann, eine Gleichung $c(x, z)$ so formuliert, dass bei Verletzung der Randbedingung ein negativer Funktionswert auftritt.

Randbedingung	Gleichung
$g(x, z) \leq K$	$c(x, z) = [K - g(x, z)] w$
$g(x, z) \geq K$	$c(x, z) = [g(x, z) - K] w$
$g(x, z) = K$	$c(x, z) = - g(x, z) - K w$
$K_1 \leq g(x, z) \leq K_2 \Rightarrow c(x, z) = \left[\frac{ K_1 - K_2 }{2} - g(x, z) - \frac{K_1 + K_2}{2} \right] w$	

Dabei sind K Konstanten und $w \geq 0$ optionale Gewichtungsfaktoren, die eine Normierung der Randbedingungen und einen direkten Vergleich untereinander ermöglichen.

Verletzt ein Testpunkt eine Randbedingung $c(x, z) < 0$, schließt dieses die Zugehörigkeit zur Pareto-Grenze aus. Werden zwei Individuen in einem Fitnesstest verglichen (Dominanz), so gewinnt das Individuum, welches keine Randbedingungen verletzt gegenüber einem Individuum, welches mindestens eine Randbedingung verletzt. Sollen oder können Randbedingungen im Fitnesstest nicht separat betrachtet werden, da dieses im verwendeten Algorithmus nicht vorgesehen ist, so wird jede Zielgröße bei Verletzung mindestens einer beliebigen Randbedingung durch einen Maximalwert für die Zielgröße plus den Absolutbetrag der Summe aller verletzten Randbedingungen ersetzt.

$$z_i^* = z_{i,max} + |\sum \min(0, c(x, z))| \quad \text{Verletzung} \rightarrow c(x, z) < 0 \quad (10.5)$$

Alternativ kann ein fester Strafwert z_{st} , welcher größer ist als der maximale Zielgrößenwert, dem eigentlichen Zielgrößenwert aufaddiert werden.

$$z_i^* = z_{i,st} + z_i \quad (10.6)$$

Dadurch wird sichergestellt, dass die Individuen mit Verletzung mindestens einer Randbedingung schlechtere Zielgrößenwerte aufweisen als die Individuen, die keine Verletzung aufweisen. Dieses führt zu einer automatischen Bevorzugung der Individuen ohne Verletzung. Somit können Optimierungsverfahren ohne spezielle interne Erweiterung für Randbedingungsverarbeitung verwendet werden, da die Berücksichtigung in den Bereich der Zielgrößenberechnung verschoben wird.

Grundsätzlich können die Gleichungen auch so definiert werden, dass bei Verletzung einer Randbedingung positive Werte auftreten. Dieses ist zum Beispiel in Kombination mit der priorisierten Dominanz sinnvoll wenn der zugehörige Grenzwert auf $g = 0$ eingestellt wird.

10.4 Reduktion auf eine Zielgröße

Eine auch heute noch weit verbreitete Methode verschiedene Zielgrößen gleichzeitig zu optimieren ist die Zusammenfassung in eine übergeordnete Zielgröße z^* .

$$z^* = g(z_1, \dots, z_{n_z}) \quad (10.7)$$

Die Schwierigkeit dabei ist eine sinnvolle Kombination $[g(z_1, \dots, z_{n_z})]$ der einzelnen Zielgrößen vor der Optimierung zu bestimmen. Dazu ist bereits vor der eigentlichen Optimierung ein umfangreiches Wissen über die Abhängigkeiten der einzelnen Zielgrößen und Faktoren notwendig.

Gewichtete Summe

Eine beliebte und einfache Kombination der Zielgrößen ist die *gewichtete Summe* bei der jede Zielgröße z_k mit einem vorher festgelegtem Gewicht w_k multipliziert wird.

$$z^* = \sum_{k=1}^{n_z} w_k z_k \quad \text{mit} \quad \sum_{k=1}^{n_z} w_k = 1 \quad (10.8)$$

Das optimierte Ergebnis hängt dabei direkt von dem Verhältnis der einzelnen Gewichte zueinander ab (Abbildung 10.6). Ist die Pareto-Grenze konvex, so kann grundsätzlich jeder Punkt der Grenze durch eine geeignete Kombination der Gewichte ermittelt werden. Bei konkaven Pareto-Grenzen sind jedoch eventuell wichtige Bereiche nicht ermittelbar.

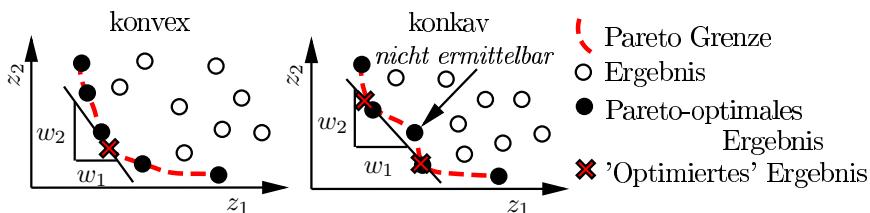


Abb. 10.6 Optimierung durch gewichtete Summe

Gewichtetes Produkt

Ein weiteres Kombinationsverfahren für Zielgrößen ist das *Produkt normierter und gewichteter Größen* (siehe Kapitel 5.2.1). Dazu werden die Zielgrößen z_1, \dots, z_{n_z} im

ersten Schritt auf den Bereich $[0, 1]$ normiert. Die Abbildung 10.7 zeigt eine typische Normierungen, wobei der normierte Funktionswert $z_{norm} = 1$ für ein ideales und $z_{norm} = 0$ für ein unzureichendes Ergebnis steht.

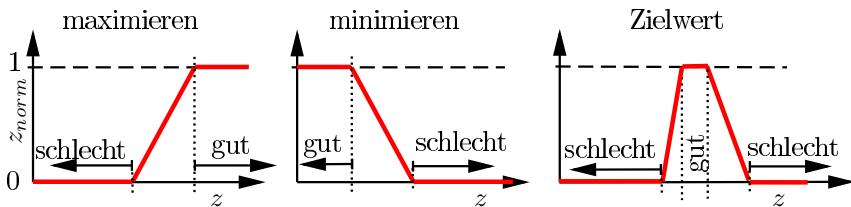


Abb. 10.7 Normierung von Zielgrößen

Neben den dargestellten linearen Normierungsfunktionen werden auch häufig sigmoidale (s-förmige) Funktionen verwendet, wie sie in Gleichung 9.232 (Kapitel 9.16) als Aktivierungsfunktion für künstliche Neuronale Netzwerke eingesetzt werden. Die übergeordnete Zielgröße z^* ist das (negative) Produkt der normierten Zielgrößen $z_{norm,k}$.

$$z^* = - \prod_{k=1}^{n_z} z_{norm,k} \quad (10.9)$$

Nimmt die übergeordnete Zielgröße den Wert $z^* = 0$ an, ist mindestens eine der Zielgrößen $z_{norm,k} = 0$ im nicht akzeptierten Bereich. Ein Wert von $z^* = -1$ bedeutet hingegen, dass alle normierten Zielgrößen im optimalen Bereich liegen.

Werden alle Zielgrößen in eine übergeordnete Zielgröße kombiniert, so kann diese mit klassischen Minimierungsverfahren optimiert werden. Dabei kommen meist Gradientenverfahren [42] aber auch Verfahren wie Simulated Annealing (Kapitel 8.4) oder genetische Optimierungen (Kapitel 10.6) zum Einsatz. Bei der Anwendung eines Gradientenverfahrens ist zu beachten, dass diese bei nicht definierten Bereichen im Faktorraum häufig Probleme aufweisen, so dass genetische Algorithmen hier Vorteile bieten können. Abbildung 10.8 zeigt schematisch den Optimierungsverlauf bei Verwendung von zwei verschiedenen Optimierungsstartpunkten ($Start_{1,2}$) eines Gradientenverfahrens beziehungsweise bei Kombination der Zielgrößen zu unterschiedlichen übergeordneten Zielgrößen $(z_{1,2}^*)$.

Je nach Startpunkt oder Kombination der Zielgrößen werden andere oder auch gleiche Lösungen auf der Pareto-Grenze gefunden. Bei Verwendung klassischer Verfahren wird lediglich ein Ergebnis ermittelt, welches zum großen Teil von dem vorher definierten Zusammenhang der Zielgrößen abhängt. Das ermittelte Ergebnis wird dann meistens in den folgenden Entwicklungsschritten als 'optimales' Ergebnis verwendet. Neben dem 'optimierten' Ergebnis befinden sich jedoch verschiedene Ergebnisse mit nahezu identischen oder ebenfalls akzeptablen Zielwertkombinationen. Diese alternativen Lösungen, die dem Anwender nicht zur Verfügung stehen, können jedoch entscheidende Vorteile gegenüber der gefundenen Lösung aufweisen.

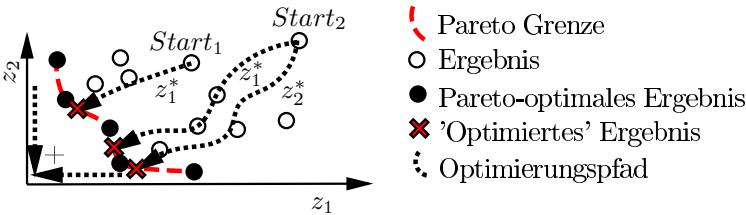


Abb. 10.8 Optimierung mittels einer übergeordneten Zielgröße

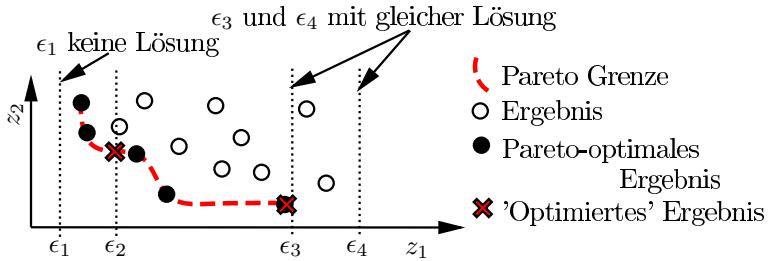
sen. Dieses können beispielsweise vorteilhafte Faktorkombinationen sein, die eine einfache Fertigung ermöglichen oder eine robustere Lösung darstellen. Weitere Ergebnisse auf der Pareto-Grenze sind nur durch eine erneute Optimierung mit neuen Zielgrößenkombinationen oder neuem Startpunkt ermittelbar. Der benötigte Rechenaufwand für die Optimierung wird dadurch vervielfacht und das grundsätzliche Problem der richtigen Gewichtungswahl und Kombinationsfunktion ist nicht gelöst, so dass das gefundene Ergebnis bei jeder Optimierung in gewissen Grenzen zufällig ist.

ε -Constraint

Eine Alternative zur Kombination der Zielgrößen zu einer übergeordneten Zielgröße z^* bietet die ε -Constraint-Methode [37], bei der lediglich eine der Zielgrößen $z_a \in \{z_1, \dots, z_{n_z}\}$ optimiert wird. Alle anderen Zielgrößen sind als Nebenbedingung mit festzulegenden Maximalwerten ε definiert (Abbildung 10.9):

$$z_b \leq \varepsilon_b \quad b = 1, \dots, n_z \quad b \neq a \quad (10.10)$$

Existiert eine Lösung für die Minimierungsaufgabe von z_a so ist diese Lösung eine Lösung der Pareto-Grenze. Durch die Optimierung der einzelnen Zielgrößen in Kombination mit unterschiedlichen Maximalwerten ε der übrigen Zielgrößen können grundsätzlich mit genügend Rechenaufwand alle Punkte der Pareto-Grenze ermittelt werden.

Abb. 10.9 ϵ -Constraint-Methode

10.5 Naturanaloge Optimierungsverfahren

Naturanaloge Optimierungsverfahren (*Nature Inspired Optimization*) sind Metaheuristiken, welche von biologischen beziehungsweise natürlichen Vorgängen inspiriert sind. Informatik und Biologie haben eine lange gemeinsame Geschichte, in der Informatiker Algorithmen zur Analyse biologischer Daten entwickelten und Biologen biologische Funktionsprinzipien untersuchten, welche neue Optimierungsverfahren inspirierten. Bekannte Verfahren sind zelluläre Automaten, neuronale Netze (Kapitel 9.16) oder evolutionäre Verfahren. Genetische Evolutionsverfahren werden ausführlich in Kapitel 10.6 behandelt.

10.5.1 Partikelschwarmoptimierung

Partikelschwarmoptimierung (*Particle Swarm Optimization*, PSO) ist ein metaheuristisches Optimierungsverfahren welches durch die Bewegung von Vögeln in einem Schwarm inspiriert ist. Die grundlegende Idee wurde 1995 von Kennedy und Eberhardt für Optimierungsaufgaben mit einer Zielgröße entwickelt und hat bereits eine weite Verbreitung erreicht [30], was durch eine einfache Implementierung gefördert wird. PSO verwendet einen Informationsaustausch zwischen verschiedenen Individuen und weist somit einige Ähnlichkeiten zu genetischen Verfahren (Kapitel 10.6) auf. Im Gegensatz zu genetischen Verfahren verwendet PSO immer reelle Faktorwerte und eine globale Kommunikation zwischen den Individuen anstelle von Kreuzung und Mutation. Die Bewegung eines Individuum (Vogels) resultiert dabei aus einer stochastischen und einer deterministischen Komponente. Abbildung 10.10 zeigt exemplarische einen Bewegungsschritt eines Vogels i zum Zeitpunkt t von Position x_i^t zur neuen Position im nächsten Zeit- beziehungsweise Generationsschritt x_i^{t+1} .

Das Individuum i befindet sich zum Zeitpunkt t an der Position x_i^t des Faktorraums und weist eine Fluggeschwindigkeit v_i^t auf. Es wird von der bisher besten bekannten Position g_i^t , welche vom gesamten Schwarm bislang entdeckt wurde, angezogen. Weiterhin ist jedem Individuum i die eigene bisher beste Position p_i^t bekannt, welche

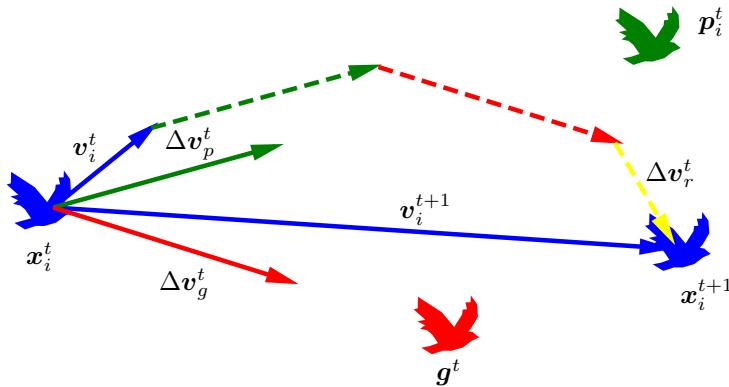


Abb. 10.10 Vereinfachter Bewegungsschritt eines Individuums während der Partikelschwarm-Optimierung

ebenfalls eine Anziehung auf das Individuum ausübt. Die Verwendung des globalen und persönlichen besten Ergebnisses erhöht deutlich die Diversität des Verfahrens. Während bei einer Optimierung mit einer Zielgröße immer nur ein globales bestes Ergebnis existiert, kann jedes Individuum sein individuelles bestes Ergebnis aufweisen. Zusätzlich wird die Flugbahn durch einen rein zufälligen Anteil Δv_r^t beeinflusst. Die Hauptaufgabe des zufälligen Anteils ist zu Verhindern, dass der Algorithmus in einem lokalen Minimum verharrt und ein großer Bereich des Faktorraums zufällig geprüft wird. Der Optimierungsalgorithmus bestimmt in jedem Schritt das globale beste Ergebnis und ist beendet, wenn sich dieses über eine vorgegebene Zeit nicht weiter verbessern lässt oder ein anderes Stopp-Kriterium, wie zum Beispiel eine voreingestellte Maximalzeit, erfüllt/erreicht ist. Die grundlegenden Algorithmusschritte einer Partikelschwarmoptimierung sind in Algorithmus 8 gegeben.

- 1 Initialisieren von zufälligen Startpunkten der Individuen x_i^0 im erlaubten Faktorbereich
- 2 Berechne die Qualitätskriterien der Individuen $f(x_i^0)$
- 3 Bestimme globales bestes Ergebnis g
- 4 Setze die persönlichen besten Zielgröße zu den aktuellen Zielgrößen $p_i = x_i$
- 5 **solange kein Stop-Kriterium erfüllt ist tue**
 - 6 **für alle Individuen i , ($i = 1 \rightarrow n_{ind}$) tue**
 - 7 Bestimme die neue Geschwindigkeit des Individuums i
 $v_i^{t+1} = f(x_i^t, v_i^t, g^t, p_i^t)$
 - 8 Bestimme die neue Position des Individuums i
 $x_i^{t+1} = f(x_i^t, v_i^{t+1})$
 - 9 Berechne die neue Zielgröße des Individuums $f(x_i^{t+1})$ im erlaubten Faktorraum
 - 10 Bestimme bislang persönliches bestes Ergebnis p_i^{t+1}
 - 11 **Ende**
 - 12 Bestimme bislang globales bestes Ergebnis g^{t+1}
 - 13 **Ende**

Algorithmus 8 : Partikelschwarmoptimierung

v_i^t beschreibt die aktuelle Geschwindigkeit und x_i^t die aktuelle Position des Individuums i zum Zeitpunkt t im Faktorraum und wird in jedem Schritt wie folgt aktualisiert.

$$x_i^{t+1} = x_i^t + v_i^{t+1} + \epsilon_{1,i} \cdot \lambda \odot x_{Diff}, \text{ mit} \quad (10.11)$$

$$v_i^{t+1} = \chi v_i^t + \alpha \epsilon_{2,i} \odot (g_i^t - x_i^t) + \beta \epsilon_{3,i} \odot (p_i^t - x_i^t) \quad (10.12)$$

In diesen Gleichungen sind die Variablen $\epsilon_1 \in [-1, 1]$, ϵ_2 und ϵ_3 beide $\in [0, 1]$ veränderliche Zufallszahlen, welche für jede Faktordimension gleiche oder unterschiedliche Werte annehmen können. α und β sind Parameter, welche in Abhängigkeit zweier Individuen die maximale Schrittweite beeinflussen. Ist dieser Wert größer als eins, so kann ein Individuum nicht nur bis, sondern über die globale oder persönliche beste Position beschleunigt werden. χ reduziert die aktuelle Geschwindigkeit jedes Individuums in jedem Schritt, so dass sich die Bewegungsgeschwindigkeit im Verlauf der Optimierung kontinuierlich verringert. Zusätzlich ist in den meisten Fällen eine Begrenzung der maximalen Fluggeschwindigkeit zum Beispiel auf 30 oder 80% Flugstrecke des Faktorbereichs pro Optimierungsschritt sinnvoll ($v_i \leq v_{max}$). Sollten die Faktorbereiche nicht normalisiert sein, so muss hier jede Faktordimension einzeln betrachtet werden. Mit Hilfe des Parameter λ wird die maximale zufällige Schrittweite eingestellt. Wenn durch x_{Diff} der Faktorbereich definiert ist, kann λ als normierter Faktor $\lambda \in [0, 1]$ vorgegeben werden. Abbildung 10.11 zeigt exemplarisch die Geschwindigkeit eines Individuums im Laufe eines Optimierungslaufs [38]. Nach anfänglich hohen Geschwindigkeiten fällt diese schnell ab und stabilisiert sich bei einem kleinen Wert nahe Null.

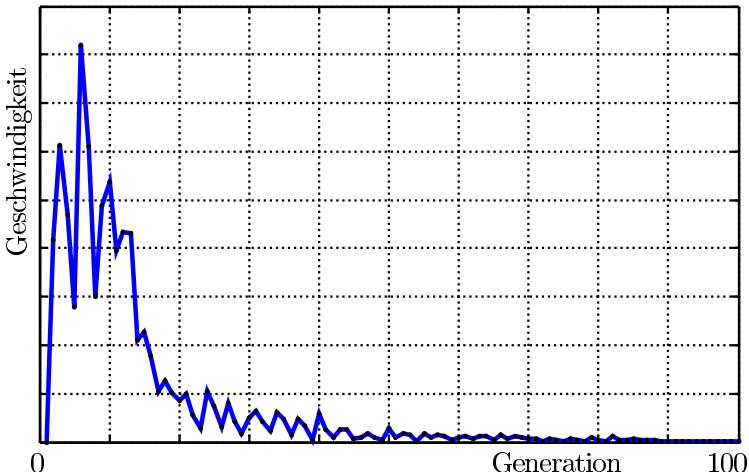


Abb. 10.11 Exemplarische Geschwindigkeit eines Individuums während der Partikelschwarm-Optimierung

Wird der erlaubte Faktorbereich während der Bewegung eines Individuums für einen Faktor k überschritten, so muss der betroffene Faktor auf die Grenze des erlaubten Bereichs zurückgesetzt werden. Die Geschwindigkeit in der Faktordimension k kann entweder unverändert bleiben, zu Null gesetzt oder umgekehrt werden. Alle dargestellten konstanten Parameter wie α oder β können ebenfalls variabel eingesetzt werden. In den meisten Fällen wird dann eine lineare Veränderung über den Optimierungsfortschritt eingeführt:

$$p = p_{start} + \frac{P_{end} - P_{start}}{generation_{max}} generation \quad (10.13)$$

Typische Parameterbereiche sind dabei:

$$\begin{aligned}\alpha &: 2.5 \rightarrow 0.5 \\ \chi &: 0.7 \rightarrow 0.4 \\ \beta &: 2.5 \rightarrow 0.5 \\ \lambda &: 0.1 \rightarrow 0.01\end{aligned}$$

Abbildung 10.12 zeigt exemplarisch eine Optimierung der Ackley Testfunktion für zwei Faktordimensionen [6, 49]. Die Startpositionen (+) sowie Endpositionen (·) von 20 Individuen zu Beginn und nach 20 Iterationen sowie die Flugbahn eines ausgewählten Individuums (gestrichelte Linie) sind zur Veranschaulichung dargestellt [38].

Ackley

$$f(x) = -a \exp \left(-b \sqrt{\frac{1}{n_f} \sum_{i=1}^{n_f} x_i^2} \right) - \exp \left(\frac{1}{n_f} \sum_{i=1}^{n_f} \cos(cx_i) \right) + a + \exp(1) \quad (10.14)$$

$$x_i \in [-5, 5]$$

Im zweidimensionalen Fall führt die Ackley Funktion zu folgender Form mit einem Minimum bei $f_{min} = f([0, 0]) = 0$.

$$f(x, y) = -20 \exp \left(-0.2 \sqrt{0.5(x^2 + y^2)} \right) - \exp(0.5(\cos(2\pi x) + \cos(2\pi y))) + 20 + \exp(1) \quad (10.15)$$

mit

$$n_f = 2 \quad a = 20 \quad b = 0.2 \quad c = 2\pi$$

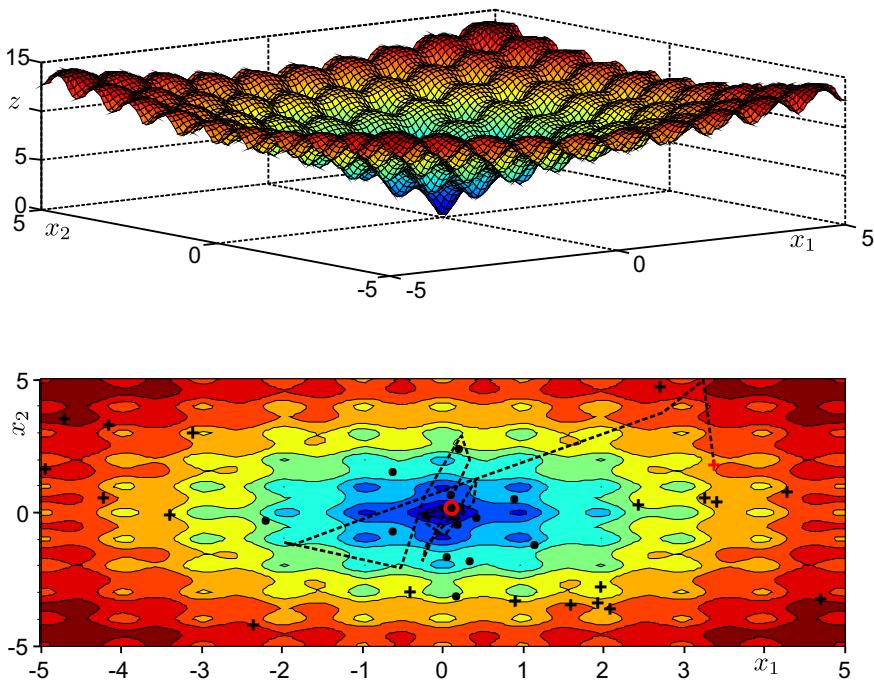


Abb. 10.12 Optimierung der Ackley Funktion mit PSO

10.5.2 Glühwürmchen

Der Glühwürmchen (*FireFly*) Algorithmus basiert auf der Helligkeit einzelner Glühwürmchen und deren Anziehung auf andere Artgenossen [51]. Zur Formulierung eines Optimierungsalgorithmus kann vereinfacht angenommen werden, dass die Anziehungskraft mit der Helligkeit steigt und mit dem Abstand zwischen Individuen sinkt. Yang beschreibt verschiedene grundlegende Prinzipien des Algorithmus:

1. Alle Glühwürmchen sind uni-sexuell und werden von jedem anderen Individuum gleichermaßen beeinflusst.
2. Die Anziehungskraft ist proportional zur Helligkeit des Glühwürmchens.
3. Das schwächer leuchtende Glühwürmchen wird immer zum stärker leuchtenden hingezogen.
4. Die Helligkeit nimmt über die Entfernung ab.
5. Wenn kein helleres Glühwürmchen vorhanden ist, kann eine zufällige Flugbahn gewählt werden.
6. Die Helligkeit eines Glühwürmchens kann mit der Zielgröße bzw. dem Gütekriterium gleichgesetzt werden.

Wenn die Leuchtintensität beziehungsweise Anziehungskraft eines Glühwürmchen mit β_0 definiert ist, so wird die reduzierte Leuchtkraft im Abstand r wie folgt angenommen.

$$\beta = \beta_0 e^{-\gamma r^m} \quad (10.16)$$

Dabei ist γ ein Absorptionskoeffizient, welcher die Abschwächung über r beschreibt und $m \geq 1$ ist ein zusätzlicher Tuningparameter. Der Abstand zweier Glühwürmchen i und j kann vereinfacht als euklidischer Abstand bestimmt werden, wobei eine Normierung aller Faktorbereiche auf $[0, 1]$ sinnvoll ist.

$$r = \sqrt{\sum_{k=1}^{n_F} (x_{ik} - x_{jk})^2} \quad (10.17)$$

Sollte das Glühwürmchen j eine höhere Helligkeit als i aufweisen, so wird die Bewegung von i folgendermaßen bestimmt.

$$x_i^{t+1} = x_i^t + \beta(x_j^t - x_i^t) + x_{Diff} \odot \epsilon_{1,i} \alpha_t \quad (10.18)$$

Der erste Term x_i^t beschreibt die aktuelle Position des Glühwürmchen i und der zweite Term $\beta(x_j^t - x_i^t)$ die Bewegung in Richtung des helleren Individuums j . Ein Wert von $\beta = 1$ würde bedeuten, dass sich das Glühwürmchen i nach diesem Schritt genau auf der gleichen Position wie j befindet. Ein Wert $0 < \beta < 1$ führt zu einer Position zwischen den Individuen und bei $\beta > 1$ fliegt das Glühwürmchen i über das Individuum j hinaus. Der dritte Term führt eine zufällige Komponente ein, welche sich über die Optimierungszeit verringert.

$$\alpha_t = \alpha_0 \theta^t, \quad \alpha_t = \alpha_0 \theta^{n_{gen}}, \quad 0 < \theta < 1 \quad (10.19)$$

α_0 ist eine normierte Konstante, welche die maximale zufällige Schrittweite am Anfang der Optimierung definiert. Die Normierung bezieht sich dabei auf den erlaubten Faktorbereich, der durch x_{Diff} beschrieben wird. Die Zufallszahl $\epsilon_{1,i} \in [-1, 1]$ kann in jedem Schritt, für jedes Individuum und für jeden Faktor unabhängig bestimmt werden.

Soll α einen vorgegebenen Endwert α_e bei konstantem θ und gegeben Startwert α_0 aufweisen, so wird der passende θ Wert folgendermaßen bestimmt.

$$\theta = \left(\frac{\alpha_e}{\alpha_0} \right)^{\frac{1}{t_{max}}} \quad (10.20)$$

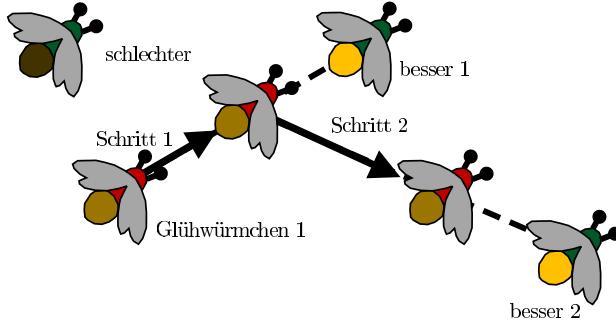
Nach Yang konvergiert der Algorithmus für eine beliebig große Anzahl von Glühwürmchen größer als die Anzahl der lokalen Optima $n_{Fliegen} \gg o_l$ [52, 55]. Der Algorithmus kann wie in Algorithmus 9 zusammengefasst werden.

Abbildung 10.13 zeigt exemplarisch einen Bewegungsschritt von Glühwürmchen i zum helleren Glühwürmchen eins und von dort zum Glühwürmchen zwei.

```

1 Initialisieren von zufälligen Startpunkten der Individuen  $x_i^0$  im erlaubten Faktorbereich
2 Berechne die Qualitätskriterien der Individuen  $f(x_i^0)$ 
3 solange kein Stop-Kriterium erfüllt ist tue
4   für alle Individuen  $i, (i = 1 \rightarrow n_{ind})$  tue
5     für alle Individuen  $j, (j = 1 \rightarrow n_{ind}), j \neq i$  tue
6       wenn  $j$  dominiert  $i$  dann
7         Bestimme euklidischen Abstand  $r_{i,j}$ 
8         Bewege Individuum  $i$  in Richtung  $j$ :  $x_i^{t+1} = f(x_{i,j}^t, \alpha, \beta, \varepsilon)$ 
9       Ende
10    Ende
11    wenn  $i$  nicht bewegt wurde dann
12      Bewege Individuum  $i$  zufällig im Faktorraum
13    Ende
14    Bestimme neue Zielgröße  $f(x_i^{t+1})$ 
15  Ende
16 Ende

```

Algorithmus 9 : Glühwürmchen Algorithmus**Abb. 10.13** Glühwürmchen Optimierungsschritt

10.5.3 Fledermaus

Der Fledermaus Algorithmus (*Bat Algorithm*) basiert auf der Orientierung mittels Echolot und wurde 2010 von YANG dargestellt [53]. Fledermäuse verwenden das Echo von ausgesendeten Impulsen nicht nur zur Orientierung und zum Umfliegen von Hindernissen, sondern ebenfalls um Nahrung aufzuspüren und ihren Schlafplatz zu finden. Basierend auf diesen Eigenschaften werden verschiedene Verallgemeinerungen zur Entwicklung eines Optimierungsalgorithmus aufgestellt [52, 53, 50, 55].

1. Fledermäuse verwenden Echolot zur Bestimmung von Entfernungen.
2. Fledermäuse fliegen mit einer Geschwindigkeit von v_i an der Position x_i und verwenden einen Echoton mit Frequenz f_i , Lautstärke L_i und Rate $r \in [0, 1]$ um Nahrung zu finden.
3. Frequenz und Rate des Impulses können je nach Entfernung zum Ziel variiert werden.

4. Die Lautstärke variiert zwischen vorgegebenen Grenzen L_{min} und L_{max} .
5. Die Frequenz variiert zufällig zwischen den vorgegebenen Grenzen $f_i \in [f_{min}, f_{max}]$.

Die Bewegung einzelner Fledermäuse wird vereinfacht durch folgende Vorschriften beschrieben.

$$x_i^{t+1} = x_i^t + v_i^{t+1} \quad (10.21)$$

$$v_i^{t+1} = v_i^t + (x_i^t - g^t) f_i \quad (10.22)$$

$$f_i = f_{min} + (f_{max} - f_{min}) \beta \quad (10.23)$$

$\beta \in [0, 1]$ ist dabei in jedem Bewegungsschritt ein zufälliger Wert (Vektor) zur Bestimmung einer neuen Frequenz für jede Faktordimension und jedem Individuum. g^t ist momentan das beste bekannte Ergebnis und leitet alle Individuen grundsätzlich in diese Region des Faktorraums. Abhängig von der momentanen Pulsrate eines Individuums ist es im vorgeschlagenen Algorithmus ebenfalls möglich, dass ein Individuum direkt an die Stelle des globalen besten Ergebnisses springt und mit einer zusätzlichen zufälligen Komponente nach einer neuen besseren Position in der lokalen Umgebung sucht.

$$x_i^{t+1} = g^t + \varepsilon_i \odot x_{Diff} \tilde{L}^t \quad (10.24)$$

$\varepsilon_i \in [0, 1]$ ist auch in diesem Fall ein Zufallsvektor, welcher zusätzlich mit der durchschnittlichen normierten Lautstärke aller Fledermäuse \tilde{L}^t multipliziert und mittels des Faktorbereichs x_{Diff} multipliziert wird. Alternativ zur mittleren Lautstärke ist es möglich einen festen oder variablen (abnehmenden) maximalen Sprungfaktor $[0, 1]$ zu verwenden. Sollte das Individuum nach dem letzten Bewegungsschritt eine bessere Position erreicht haben wird diese akzeptiert und ansonsten verworfen. Zur Verringerung der Wahrscheinlichkeit in einem lokalen Minimum stecken zu bleiben wird eine schlechtere Position zusätzlich mit einer Wahrscheinlichkeit L_i akzeptiert ($\varepsilon_{i,2} < L_i$). Die verwendete Pulsrate r und Lautstärke L (oder Amplitude A) kann konstant oder variabel über den Optimierungsverlauf angenommen werden und jedes Individuum kann unterschiedliche Werte aufweisen. In der Natur reduziert sich die Lautstärke und erhöht sich die Pulsrate bei Annäherung an Beute. Da während der Optimierung das wahre optimale Ergebnis nicht bekannt ist, werden Lautstärke und Pulsrate über den Optimierungsverlauf kontinuierlich erhöht beziehungsweise verringert, da von einer Annäherung an das Optimum während der Optimierung ausgegangen wird.

$$L_i^{t+1} = \alpha L_i^t, \alpha \in (0, 1), L^\infty = 0 \quad (10.25)$$

$$r_i^{t+1} = r_i^\infty [1 - e^{(-\gamma t)}], \gamma > 0 \quad (10.26)$$

Grundsätzlich wird das vorgeschlagene Verfahren wie in Algorithmus 10 dargestellt zusammengefasst.

```

1 Initialisieren von zufälligen Startpunkten der Individuen  $x_i^0$  im erlaubten Faktorbereich
2 Berechne die Qualitätskriterien der Individuen  $f(x_i^0)$ 
3 Bestimme die Geschwindigkeit  $v_i^0$ , die Pulsrate  $r_i^\infty$  und die Lautstärke  $L_i^0$  für jede Fledermaus
4 solang kein Stopp-Kriterium erfüllt ist tue
5   | Berechne durchschnittliche Lautstärke aller Fledermäuse  $\bar{L}_i^t$ 
6   | Bestimme globales bestes Ergebnis  $g^t$ 
7   | für alle Individuen  $i$ , ( $i = 1 \rightarrow n_{ind}$ ) tue
8     | Berechne zufällige Frequenz  $f_i^t \in [f_{min}, f_{max}]$ 
9     | Bestimme neue Geschwindigkeit und Position
10    |  $v_i^{neu} = f(v_i^t, x_i^t, g_i^t, f_i^t)$ 
11    |  $x_i^{neu} = f(x_i^t, v_i^{neu})$ 
12    | wenn  $rand > r_i^t$  dann
13      |   Überschreibe Position mit  $x_i^{neu} = f(g_i^t, \bar{L}_i^t)$ 
14    | Ende
15    | Berechnung des Qualitätskriteriums  $f(x_i^{neu})$ 
16    | wenn  $f(x_i^{neu}) \leq f(x_i^t)$  oder  $rand < L_i$  dann
17      |   Akzeptiere neue Position und Geschwindigkeit
18      |    $x_i^{t+1} = x_i^{neu}$  und  $v_i^{t+1} = v_i^{neu}$ 
19    | Ende
20    | Verringere  $L_i$  und erhöhe  $r_i$ 
21  | Ende
22 Ende

```

Algorithmus 10 : Fledermaus

10.5.4 Blütenbestäubung

Blütenbestäubung (*Flower Pollination*) ist ein ausgereifter evolutionärer Prozess, welcher den Fortbestand vieler Blumen, Sträucher und Bäume sichert und daher als Grundlage für einen Optimierungsalgorithmus verwendet werden kann. Über 90% aller blühenden Pflanzen verwenden biotische Bestäubung, was bedeutet, dass Ihre Pollen durch Bestäuber wie zum Beispiel Insekten oder andere Tiere transportiert werden. Die restlichen Pflanzen verwendet abiotische Bestäubung, wobei Wind und Diffusion im Wasser zum Transport verwendet wird [54]. Eine bekannte Gruppe der Bestäuber sind Bienen, die zusätzlich zur Bestäubungsfunktion eine Blütenstetigkeit aufweisen können [9, 55]. Blütenstetigkeit bedeutet, dass der Bestäuber bevorzugt oder ausschließlich eine Pflanzenart besucht. Blütenstetigkeit hat evolutionäre Vorteile, da sie den Pollentransport zu konspezifischen Pflanzen maximiert und somit die Reproduktion der Art. Für den Bestäuber ist auf der anderen Seite vorteilhaft, dass ausreichend Nahrungsquellen im limitierten Flugbereich vorhanden sind und diese mit wenig Aufwand erreicht und erhalten werden können [54]. Bestäubung wird weiterhin in Selbst- und Fremdbestäubung unterteilt. Fremdbestäubung bedeutet dabei, dass Pollen von anderen Pflanzen stammen, wobei Selbstbestäubung die Besamung einer einzelnen Pflanze beschreibt, wie es bei Pfirsichen vorkommt. Bei der Selbstbestäubung stammt der Pollen von der gleichen oder einer anderen Blüte der selben Pflanze und tritt meistens dann auf, wenn verlässliche Bestäuber nicht verfügbar sind. Selbstbestäubung findet in einem lokal begrenzten Bereich statt, wo-

bei biotische Fremdbestäubung eine deutlich größere Entfernung abdeckt und somit für den Optimierungsalgorithmus als globaler Anteil betrachtet wird. Die Bewegung der Bestäuber wird vereinfacht als zufällige Bewegung abgebildet, wobei von Yang vorgeschlagen wird eine Lévy Verteilung einzusetzen [54, 41]. Grundsätzlich werden von Yang folgende Idealisierungen angenommen:

1. Biotische und Fremdbestäubung werden als globaler Bestäubungsprozess betrachtet, wobei die Bestäuber Lévy Flüge ausführen.
2. Abiotische und Selbstbestäubung werden zur lokalen Bestäubung herangezogen.
3. Blütenstetigkeit kann grundsätzlich als Bestäubungswahrscheinlichkeit berücksichtigt werden und ist abhängig von der Ähnlichkeit beteiligter Pflanzen.
4. Das Verhältnis zwischen lokaler und globaler Bestäubung wird durch eine festgelegte Wahrscheinlichkeit $p \in [0, 1]$ kontrolliert.

Zur Vereinfachung wird angenommen, dass jede Pflanze nur eine Blüte und einen Pollen aufweist und lediglich einen lokalen sowie globalen Bewegungsalgorithmus aufweist. Im globalen Schritt wird das Samenkorn durch Bestäuber über eine lange Distanz transportiert, was den Fortbestand und die Reproduktion der stärksten Pflanzen sichert [54].

$$x_i^{t+1} = x_i^t + L(g^t - x_i^t) \quad (10.27)$$

x_i^t ist dabei die Position der Pflanze i im Faktorraum zum Zeitpunkt t und g^t die bisher beste gefundene Pflanzenposition. Der Parameter L ist nach Yang die „Bestäubungsstärke“ und ist äquivalent zu einer Schrittweite, die als Lévy Verteilung angesetzt werden sollte [54]. Zur Bestimmung der Lévy Verteilung wird dabei ein einfacher Algorithmus auf Basis der Gamma-Funktion Γ und der Normalverteilung \mathcal{N} vorgeschlagen.

$$L = L_f \frac{\mathcal{N}_1 \sigma}{|\mathcal{N}_2|^{\frac{1}{\lambda}}} \quad (10.28)$$

$$\sigma = \left[\frac{\Gamma(1 + \lambda) \sin(\pi \frac{\lambda}{2})}{\Gamma(\frac{1+\lambda}{2}) \lambda 2^{\frac{\lambda-1}{2}}} \right]^{\frac{1}{\lambda}} \quad (10.29)$$

Typische Parameterwerte sind in diesem Zusammenhang $L_f = 0.01$ und $\lambda = 1.5$ [54]. Jedes Individuum kann dabei in jedem Schritt und für jede Faktordimension eine eigene Schrittweite aufweisen. Die lokale Bestäubung und Blütenstetigkeit wird durch eine zufällige lokale Suche beschrieben, deren Suchrichtung durch zwei zufällig gewählte Pflanze beziehungsweise Pollen $x_{j,k}^t$ bestimmt wird.

$$x_i^{t+1} = x_i^t + \epsilon (x_j^t - x_k^t) \quad (10.30)$$

$$\epsilon \in [0, 1]$$

Das grundsätzliche Vorgehen ist in Algorithmus 11 dargestellt [54]:

```

1 Initialisieren von zufälligen Startpunkten der Individuen  $x_i^0$  im erlaubten Faktorbereich
2 Berechne die Qualitätskriterien der Individuen  $f(x_i^0)$ 
3 Definiere globale Wahrscheinlichkeit  $p \in [0, 1]$ 
4 solange kein Stop-Kriterium erfüllt ist tue
5   Bestimme globales bestes Ergebnis  $g^t$ 
6   für alle Individuen  $i, (i = 1 \rightarrow n_{ind})$  tue
7     wenn  $rand < p$  dann
8       Bestimme einen  $n_f$  dimensionalen Vektor aus der Lévy Distribution zur
         globalen Bestäubung
9        $x_i^{neu} = x_i^t + L(x_i^t - g^t)$ 
10      sonst
11        Bestimme  $n_f$  dimensionalen Zufallsvektor aus der Normalverteilung [0,1] zur
          lokalen Bestäubung
12        Wähle zufällige Individuen  $j, k \neq i$  und  $j \neq k$ 
13         $x_i^{neu} = x_i^t + \varepsilon (x_j^t - x_k^t)$ 
14      Ende
15      Bestimme Qualitätskriterium an neuer Position  $f(x_i^{t+1})$ 
16      wenn neue Position besser ist als alte Position dann
17        Akzeptiere neue Position  $x_i^{t+1} = x_i^{neu}$ 
18      Ende
19    Ende
20  Ende

```

Algorithmus 11 : Blütenbestäubung

10.5.5 Symbiotische Organismus Suche

Die Symbiotische Organismus Suche (*Symbiotic Organisms Search*) ahmt symbiotische Interaktionen verschiedener Organismen nach, welche ihnen die Verbreitung und das Überleben sicherstellen. Aus der Natur sind verschiedene symbiotische Beziehungen bekannt, die sich in zwei Arten einteilen lassen.

1. Obligate Parasiten: Organismen sind zwingend von anderen Organismen abhängig.
2. Fakultative Parasiten: Gelegentlich parasitierende Organismen, wobei ihre Entwicklung auch ohne parasitische Phase ablaufen kann.

Die drei bekanntesten symbiotischen Beziehungen sind dabei:

1. Mutualismus
2. Kommensalismus
3. Parasitismus

Mutualismus bezeichnet eine symbiotische Beziehung zweier unterschiedlicher Arten, bei denen beide ihren Nutzen aus der Beziehung ziehen. Bekannte Beispiele sind Ameise und Blattlaus oder Eiche und Eichelhäher. Bei **Kommensalismus** profitiert nur ein Partner der Beziehung, jedoch wird der zweite nicht negativ oder positiv beeinflusst. Dieses findet beispielsweise in den Beziehungen von Seepocken

und Buckelwalen oder Aasfressern aus Wüstenregionen, die größeren Jägern folgen statt. Im Gegensatz dazu wird beim **Parasitismus** der zweite Partner geschädigt [8]. Bekannt ist hier zum Beispiel der Brutparasitismus vom Kuckuck oder die Beziehung zwischen Menschen und blutsaugenden Insekten. Organismen entwickeln symbiotische Beziehungen um sich besser an ihre Umgebung anzupassen oder auf lange Sicht eine bessere Überlebenschance zu erhalten. Die *Symbiotische Organismus Suche* startet wie alle anderen Verfahren mit einer zufälligen Startpopulation im Faktorraum. Jedes Individuum repräsentiert eine mögliche Lösung des Optimierungsproblems und weist einen zugehörigen Gütewert auf. Neue mögliche Lösungen werden in drei Phasen ermittelt, welche die beschriebenen biologischen Interaktionen nachbilden. [8]. Alle Phasen kommen dabei ohne zusätzliche vom Benutzer zu definierende Parameter aus.

In der **Mutalismus** Phase gehen zwei Individuen x_i und x_j eine Beziehung ein um Vorteile in Ihrem Ökosystem zu erhalten. Neue Positionen im Faktorraum werden nach *Cheng* wie folgt bestimmt.

$$x_{i,j}^{neu} = x_{i,j} + \varepsilon_{i,j} (g^t - \bar{x}_{ij} b_{i,j}) \quad (10.31)$$

$$\bar{x}_{ij} = \frac{x_i + x_j}{2} \quad (10.32)$$

$$b_{i,j} \in \{1, 2\}$$

$$\varepsilon = \text{rand}(0, 1)$$

Der Benefit-Faktor $b_{i,j}$ nimmt dabei zufällig den Wert 1 oder 2 an und soll den Nutzen für jedes Individuum beschreiben, das es durch die Beziehung beziehungsweise Verbindung an der Position erreicht \bar{x}_{ij} . Der Wert g^t bezeichnet das momentan beste bekannte Ergebnis. Alternativ wird hier vom Autor vorgeschlagen den Schritt der Mutualismus Phase durch folgende Gleichung zu ersetzen, bei der $\varepsilon_{i,j}^* \in [0, 2]$ ein zufälliger Wert ist.

$$x_{i,j}^{neu} = x_{i,j} + \varepsilon_{i,j} b_{i,j} (g^t - \bar{x}_{ij}) \quad (10.33)$$

In der **Kommensalismus** Phase wird lediglich x_i beeinflusst und x_j bleibt unverändert.

$$x_i^{neu} = x_i \varepsilon_{i,j}^* (g^t - x_j) \quad (10.34)$$

$$\varepsilon \in [-1, 1]$$

In der **Parasitismus** Phase versucht ein Individuum, welches auf x_i basiert ein anderes Individuum x_j zu ersetzen. x_i bleibt selbst unverändert. Die neue Position wird bestimmt, indem einige zufällig ausgewählte Faktoren von x_i im Faktorraum variiert werden.

In allen Phasen gilt, dass ein Individuum nur an die neue Position verschoben wird, wenn sich die Qualitätsfunktion verbessert hat (Algorithmus 12).

```

1 Initialisieren von zufälligen Startpunkten der Individuen  $x_i^0$  im erlaubten Faktorbereich
2 Berechne die Qualitätskriterien der Individuen  $f(x_i^0)$ 
3 solange kein Stop-Kriterium erfüllt ist tue
4   Bestimme globales bestes Ergebnis  $g^t$ 
5   für alle Individuen  $i, (i = 1 \rightarrow n_{ind})$  tue
       Mutualismus:
6     wähle zufällig  $j \in [1, n_{ind}]$  mit  $j \neq i$ 
7     wähle zufällig  $b_{i,j} = 1$  oder 2
8      $x_{i,j}^{neu} = x_{i,j} + \varepsilon_{i,j} (g^t - \bar{x}_{i,j} b_{i,j})$ 
9      $\varepsilon_{i,j} = \text{rand}(0, 1)$ 
10    akzeptiere Ergebnis wenn  $x_{i,j}^{neu}$  besser ist als  $x_{i,j}$ 
       Kommensalismus:
11     wähle zufällig  $j \in [1, n_{ind}]$  mit  $j \neq i$ 
12      $x_i^{neu} = x_i \varepsilon (g^t - x_j)$ 
13      $\varepsilon = \text{rand}(0, 1)$ 
14     akzeptiere Ergebnis  $x_i^{neu}$  besser ist als  $x_i$ 
       Parasitismus:
15      $x_j^{neu} = x_i + \varepsilon_1 \odot x_{Diff} \odot \varepsilon_2$ 
16      $\varepsilon_1 = \text{rand}(-1, 1)$ 
17      $\varepsilon_2 = \text{rand}(0, 1) < 0.5$  (zufällige Auswahl der Faktoren)
18     akzeptiere Ergebnis wenn  $x_j^{neu}$  besser ist als  $x_j$ 
19   Ende
20 Ende

```

Algorithmus 12 : Symbiotische Organismus Suche Algorithmus

10.5.6 Erweiterung auf mehrere Zielgrößen

In den meisten realen Anwendungen müssen mehrere Zielgrößen gleichzeitig optimiert werden. Können oder sollen diese nicht wie in Kapitel 10.4 zu einer Zielgröße zusammengefasst werden, ist eine unabhängige Betrachtung der Qualitätsgrößen notwendig. Die bisher in diesem Kapitel gezeigten naturnahen Verfahren sind jedoch für eine Zielgröße entwickelt worden und werden nun auf die Berücksichtigung mehrerer Qualitätsgrößen erweitert. Dazu wurden bereits verschiedene Ansätze für die Partikelschwarmoptimierung vorgestellt und verglichen [1, 40]. Die zu lösende Hauptaufgabe ist dabei die Auswahl des globalen Ziels g^t für jedes Individuum, da im Gegensatz zur eindeutigen Lösung bei einer Qualitätsgröße im mehrdimensionalen Fall meist mehrere Pareto-optimale Lösungen existieren (siehe Kapitel 10.1). Das grundsätzliche Vorgehen für die Optimierung mehrerer Zielgrößen kann vereinfacht wie in Algorithmus 13 dargestellt zusammengefasst werden. Die einfachste Möglichkeit ein neues globales Ziel für ein Individuum i im mehrdimensionalen Fall zu wählen, ist die Rückführung der Zielgrößen auf eine globale Zielgröße, wie es in Kapitel 10.4 beschrieben wird. Für jedes Individuum werden dazu in jedem Optimierungsschritt zufällige Gewichte w_k für alle Qualitätsmerkmale o_k bestimmt und dadurch die gewichtete Summe der Kriterien berechnet.

- 1 Initialisieren von zufälligen Startpunkten der Individuen x_i^0 im erlaubten Faktorbereich
- 2 Berechne die Qualitätskriterien der Individuen $f(x_i^0)$
- 3 **solange kein Stopp-Kriterium erfüllt ist tue**
 - 4 **für alle Individuen i , ($i = 1 \rightarrow n_{ind}$) tue**
 - 5 Wähle ein globales Ziel für aktuelles Individuum g_i^t
 - 6 Bewege Individuum zu einer neuen Position x_i^{t+1}
 - 7 Akzeptiere oder verwirfe neue Positionen
 - 8 **Ende**
- 9 **Ende**

Algorithmus 13 : Optimierung mehrerer Zielgrößen

$$Z = \sum_{k=1}^{n_z} w_k z_k \quad (10.35)$$

$$\sum w_i = 1$$

$$w_k \in [0, 1]$$

Je nach Gewichtungswahl werden allen Individuen zufällige, aber je nach Gewichtung „gute“ Individuen zugewiesen. Die Wahl des Zielindividuums kann dabei immer nur aus der aktuellen Population gewählt werden. Ein vielversprechenderer Ansatz besteht in der Einführung eines zusätzlichen Archivs A , welches alle momentan bekannten Pareto-optimalen Lösungen beinhaltet [1]. Die Archivgröße kann unabhängig von der meist fixen Anzahl an Individuen variieren, so dass gefundene Pareto-optimale Lösungen gespeichert und gelöscht werden, wenn eine dominante Lösung (Kapitel 10.2) gefunden wird. Das globale Ziel eines Individuums wird in jedem Schritt aus dem externen Archiv A gewählt, wobei verschiedene Algorithmen zur Auswahl vorgeschlagen werden [1]. Das grundlegende Vorgehen (Algorithmus 13) zur Optimierung mehrerer Zielgrößen wird mit Funktionen zur Erstellung und Aktualisierung des externen Archivs erweitert (Algorithmus 14).

- 1 Initialisieren von zufälligen Startpunkten der Individuen x_i^0 im erlaubten Faktorbereich
- 2 Berechne die Qualitätskriterien der Individuen $f(x_i^0)$
- 3 **(NEU) Initialisiere das externe Archiv A**
- 4 **solange kein Stopp-Kriterium erfüllt ist tue**
 - 5 **für alle Individuen i , ($i = 1 \rightarrow n_{ind}$) tue**
 - 6 Wähle ein globales Ziel für aktuelles Individuum g_i^t **(NEU)** aus A
 - 7 Bewege Individuum zu einer neuen Position x_i^{t+1}
 - 8 Akzeptiere oder verwirfe neue Positionen
 - 9 **(NEU) wenn Individuum i von keinen Element aus A dominiert wird und nicht in A vorhanden ist dann**
 - 10 **(NEU) Füge Individuum i zu A**
 - 11 **(NEU) Lösche alle Elemente von A , die durch Individuum i dominiert werden**
 - 12 **Ende**
 - 13 **Ende**
 - 14 **Ende**

Algorithmus 14 : Optimierung mehrerer Zielgrößen mit externem Archiv

Im Laufe der Optimierung variiert die Größe des Archivs A . Es tendiert dazu dramatisch anzuwachsen, was zu einer deutlich langsameren Konvergenz der Algorithmen führt. Da die gewünschte Anzahl an Pareto-optimalen Lösungen in den meisten Fällen begrenzt ist, wird eine Limitierung der Archivgröße auf A_{max} sinnvoll. Im Fall, dass die Größe von A auf $|A| > A_{max}$ anwächst ist eine Reduzierung notwendig, die zum Beispiel auf Basis der „Crowding Distance“ (Kapitel 9) durchgeführt wird. Die Crowding Distance beurteilt dabei den Abstand von Pareto-optimalen Punkten zu den direkten Nachbarn und gibt ein Maß über die Dichte der Punkte innerhalb eines Bereichs der Pareto-Grenze [12, 15, 16, 13]. Die Elemente aus A mit den größten Crowding Distance Werten bleiben bei der Reduktion des Archivs erhalten, wobei dieses Vorgehen bei allen Verfahren zur Bestimmung des globalen Ziels (ROUNDS, RAND und PROB) Einsatz findet. Im PROB Algorithmus (Seite 351) können auch alternativ die dort berechneten Wahrscheinlichkeitswerte p_a verwendet werden. Zusätzlich zu den globalen Zielen werden beim Partikelschwarm Algorithmus ebenfalls persönlich beste Werte bzw. Ziele benötigt (Kapitel 10.5.1). Diese werden zum Beispiel durch das klassische Dominanzkriterium aktualisiert.

PURE RANDOM

Ist das externe Archiv A eingeführt wird im einfachsten Fall für jedes Individuum i ein zufälliges Element aus A als globales Ziel gewählt. Dieses entspricht grundsätzlich dem Vorgehen beim ϵ -MOEA Algorithmus aus Kapitel 9. Zur geschickteren Wahl der globalen Ziele werden die Verfahren *RANDOM*, *PROB* und *ROUNDS* vorgeschlagen [1].

RANDOM

Der RANDOM Algorithmus betrachtet jedes Individuum x_i einzeln und wählt ein zufälliges globales Ziel aus A , von dem es dominiert wird (Abbildung 10.14) [1]. Sollte kein dominantes Individuum in A vorhanden sein, so wird ein beliebiges Element aus A verwendet (Algorithmus 15).

```

1 für alle Individuen  $i$ , ( $i = 1 \rightarrow n_{ind}$ ) tue
2   |  $A_{dom} = \{a \in A \mid a \prec x_i\}$ 
3   | wenn  $A_{dom} \neq \emptyset$  dann
4   |   | Wähle  $g_i$  zufällig aus  $A_{dom}$ 
5   | sonst
6   |   | Wähle  $g_i$  zufällig aus  $A$ 
7   | Ende
8 Ende

```

Algorithmus 15 : RANDOM: Auswahl eines globalen Ziels

PROB

Um verstkt Elemente an den Rndern der Pareto-Grenze auszuwhlen werden Elemente aus A vorgezogen, welche wenige Individuen dominieren (Algorithmus 16).

```

1  fr alle Elemente  $a \in A$  tue
2     $X_a = \{x \in X | a \prec x\}$ , Anzahl von  $a$  dominierter Individuen
3     $p_a = \frac{1}{|X_a|+1}$ , Auswahlwahrscheinlichkeit basierend auf Anzahl dominierter Elemente
      (Modifikation zum Originalalgorithmus:  $p_a = \frac{1}{|X_a|}$ )
4  Ende
5  fr alle Individuen  $i$ , ( $i = 1 \rightarrow n_{ind}$ ) tue
6     $A_{dom} = \{a \in A | a \prec x_i\}$ 
7    wenn  $A_{dom} \neq \emptyset$  dann
8      Whle  $g_i$  zufllig aus  $A_{dom}$  mit der Wahrscheinlichkeit  $\propto p_a(A_{dom})$  [nur Elemente
      aus  $A_{dom}$  bercksichtigen]
9    sonst
10   Whle  $g_i$  zufllig aus  $A$  mit der Wahrscheinlichkeit  $\propto p_a(A)$ 
11  Ende
12 Ende

```

Algorithmus 16 : PROB: Auswahl eines globalen Ziels

ROUNDS

ROUNDS ist ein Algorithmus der dazu ausgelegt ist globale Ziele aus schwach besetzten Bereichen der Pareto-Grenze, also dem Archiv A , zu whlen. Dieses soll die Gleichverteilung der Lsungen auf der Pareto-Grenze verbessern, da neue Lsungen in schwach besetzte Bereiche geleitet werden. Grundstzlich versucht der Algorithmus fr jedes Individuum x_i ein globales Ziel g_i zu whlen, welches x_i dominiert. Wird das Individuum i von mehreren Elementen aus A dominiert, so werden Elemente bevorzugt, die wenige Individuen dominieren. Wenn ein Individuum als globales Ziel ausgewhlt wurde, steht es nicht fr weitere Individuen zur Verfgung. Dadurch wird sichergestellt, dass eine mglichst groe Anzahl verschiedener Elemente aus A verwendet werden. Im Fall, dass alle Individuen aus A bereits zugewiesen wurden oder die restlichen Individuen aus A das aktuelle Individuum nicht dominieren, werden alle bereits verwendeten Individuen aus A wieder reanimiert und stehen fr den Auswahlprozess wieder zur Verfgung. Sollte ein Individuum i von keinem Element aus A dominiert werden, zum Beispiel wenn es selber ein Element aus A ist, so kann zufllig ein Element aus A gewhlt werden. Im folgenden wird der Algorithmus grundstzlich dargestellt, wobei zu beachten ist, dass dieser an verschiedenen Stellen zum Original [1] leicht verndert wurde (Algorithmus 16).

```

1 Erzeuge Kopien der Individuen und des Archivs  $A$ 
 $X^* = X$ ,  $A^* = A$ 
2 solange  $X^* \neq \emptyset$  tue
3   wenn  $A^* = \emptyset$  dann
4     | Reinitialisiere  $A^* = A$ 
5   Ende
6   für alle  $a \in A^*$  tue
7     | Bestimme Individuen, die von  $a$  dominiert werden
8     |  $X_a^* = \{x \in X^* | a \prec x\}$ 
9   Ende
10   $a^* = \arg \min_{a \in A^* \wedge |X_a^*| > 0} (X_a^*)$  (1)
11   $a^*$  dominiert am wenigsten aber mindestens ein Individuum aus  $X^*$ 
12  wenn  $a^* = \emptyset$  und  $A^* \neq A$  dann
13    |  $A^* = A$ 
14    | gehe zu (1)
15  Ende
16  wenn  $a^* = \emptyset$  und  $A^* = A$  dann
17    | gehe zu (2)
18  Ende
19  Wähle ein beliebiges dominiertes Element von  $a^*$  aus  $X_{a^*}^*$ :  $x_n \in X_{a^*}^*$ 
20  Setze das globale Ziel von  $x_n$  zu  $a^*$ :  $g_{x_n}^* = a^*$ 
21  Lösche  $x_n$  aus  $X^*$ :  $X^* = X^* \setminus x_n$ 
22  Lösche  $a^*$  aus  $A^*$ :  $A^* = A^* \setminus a^*$ 
23 Ende
24 Wähle zufällige Elemente aus  $A$  für alle restlichen Individuen aus  $X^*$  (2)

```

Algorithmus 17 : ROUNDS: Auswahl eines globalen Ziels

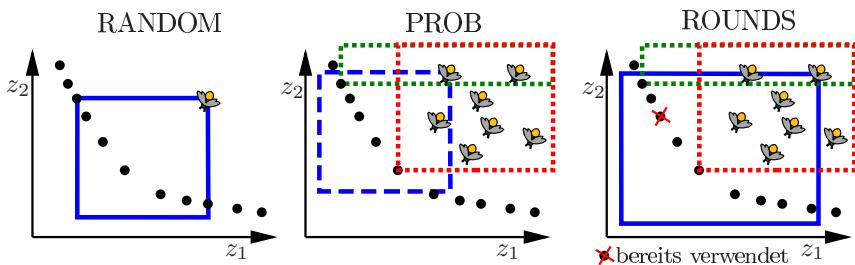


Abb. 10.14 Auswahlverfahren zur Bestimmung eines globalen Zielindividuums

Crowding Distance

Neben den dargestellten Verfahren ist ebenfalls die *Crowding Distance* aus Kapitel 10.6.3 zur Auswahl des globalen Ziels sinnvoll. Dabei werden Elemente aus A mit hohen Crowding Distance Werten bevorzugt.

10.6 Genetische Evolutionsverfahren für mehrerer Zielgrößen

Zur Bestimmung der Pareto-Grenze für mehrerer Zielgrößen (Abbildung 10.6) haben sich in vielen Bereichen genetische Optimierungsverfahren bewährt und durchgesetzt. Sie bieten die Möglichkeit alle Bereiche einer Pareto-Grenze robust zu bestimmen, auch wenn komplexe Zusammenhänge zwischen Faktoren und Zielgrößen oder unter Zielgrößen vorliegen. Weiterhin bereiten den Verfahren nicht definierte Bereiche im Faktorraum (Löcher) keine Schwierigkeiten und Randbedingungen für Faktoren und Zielgrößen lassen sich leicht in die Optimierungsalgorithmen integrieren. Bei steigender Anzahl der Zielgrößen steigt der Rechenaufwand zur Bestimmung der Pareto-Grenze schnell an, so dass typischerweise Metamodelle (Kapitel 9) mit kurzer Rechenzeit und hoher Genauigkeit anstelle von komplexen Simulationsmodellen eingesetzt werden.

In der Literatur finden sich verschiedenste genetische Optimierungsverfahren, die mehr oder weniger die biologische Evolution nachahmen. So werden Merkmale von einer zur nächsten Generation vererbt, mit anderen gekreuzt oder zufällig mutiert. In folgenden Generationen 'überleben' hauptsächlich Individuen (Faktoreinstellungen), die fitte (gute) Eigenschaften in den Zielgrößen aufweisen. Abbildung 10.15 zeigt den verallgemeinerten Ablauf einer genetischen Optimierung. Basierend auf einer vorhandenen Gruppe von Individuen mit unterschiedlichen Faktoreinstellungen (i^{te} Generation) werden durch ein zu bestimmendes Verfahren verschiedene „gute“ Individuen ausgewählt und als *Eltern* deklariert. Durch Kreuzung der verschiedenen Eltern entstehen sogenannte *Kinder*, wobei die Eigenschaften (Faktoreinstellungen) direkt von den Eltern und der Art der Kreuzung abhängen. Im Anschluss an die Kreuzung findet mit geringer Wahrscheinlichkeit eine Mutation der einzelnen Eigenschaften (Faktoreinstellungen) statt. Dazu werden die Faktoreinstellungen unabhängig von den Eltern verändert, wodurch gewährleistet wird, dass während des Optimierungsprozesses beliebige Faktoreinstellungen erzeugt werden können, die nicht aus der Elterngeneration durch Kreuzung erzeugbar sind. Dieses verhindert, dass das Optimierungsverfahren nicht in einem lokalen Minimum stecken bleibt und mit einer voreingestellten Wahrscheinlichkeit alle Bereiche des Faktorraums geprüft werden. Sind die Faktoreinstellungen der Kinder festgelegt, werden die dazugehörigen Zielgrößen bestimmt und im nächsten Schritt die neue Generation G_{i+1} ermittelt. Dazu werden durch ein zu bestimmendes Verfahren die fittesten (besten) Individuen aus den Kindern und der aktuellen i^{en} Generation gewählt. In einem iterativen Prozess werden diese vier Schritte solange wiederholt bis eine maximale Iterationsanzahl erreicht wurde oder keine signifikante Verbesserung der gefundenen Pareto-Grenze mehr erzielt wird.

Binäre Kodierung

Traditionell werden in Computer gestützten genetischen Optimierungsverfahren, die Faktoreinstellungen, binär kodiert. Seien x_u und x_o die untere und obere Grenze eines Faktors $x_u \leq x \leq x_o$ und n_b die Anzahl der verwendeten Bits (b_0, \dots, b_{n_b-1})

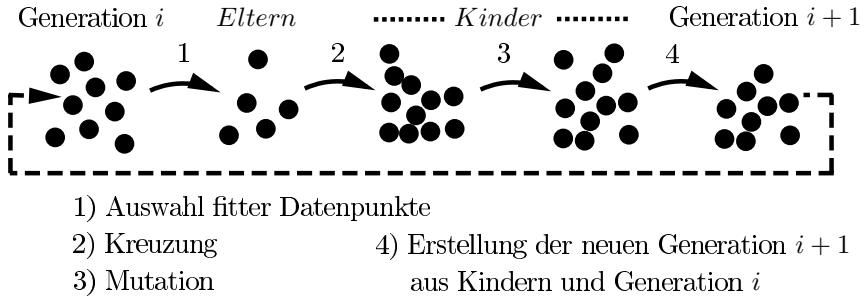


Abb. 10.15 Genetisches Evolutionsverfahren

zur Darstellung des Faktors x . Bei gegebener Binärdarstellung wird die Faktoreinstellung für x durch folgende Gleichung berechnet:

$$x = x_u + \sum_{k=0}^{n_b-1} 2^k b_k \frac{x_o - x_u}{2^{n_b} - 1} \quad (10.36)$$

Zur Veranschaulichung wird ein Rohrdurchmesser d mit zwei Bits kodiert ($2^2 = 4$ Stufen), wobei $x_u = 6$ und $x_o = 9$ sind (Tabelle 10.2). Der binäre Wert '10' ($b_1 = 1$ $b_0 = 0$) repräsentiert damit die Faktoreinstellung:

$$x = 6 + (2^0 0 + 2^1 1) \frac{9 - 6}{2^2 - 1} = 6 + 2 \frac{3}{3} = 8 \quad (10.37)$$

b_1	b_0	x
0	0	6
0	1	7
1	0	8
1	1	9

Tabelle 10.2 Binäre Kodierung

Wird eine minimale Auflösung Δ_{min} eines Faktors x für eine Optimierung benötigt, so muss die Anzahl n_b so gewählt werden, dass folgende Ungleichung erfüllt ist.

$$2^{n_b} \geq \frac{x_o - x_u}{\Delta_{min}} \quad (10.38)$$

Die binäre Kodierung der Faktoren begrenzt im Vergleich zur Verwendung des reellen Zahlenraums die möglichen Faktorstufen und somit auch die erzeugbaren Datenpunkte, was eine schnelle Konvergenz der Optimierung günstig beeinflussen kann. Eine Einschränkung der Lösungen durch die binäre Kodierung ist jedoch in einigen Fällen nicht erwünscht, so dass verschiedene Verfahren entwickelt wurden, die

neben der binären Kodierung ebenfalls einen sinnvollen Einsatz reeller Zahlen in genetischen Optimierungsverfahren ermöglicht.

10.6.1 Kreuzung

Die Kreuzung jedes Faktors x zweier Eltern a und b wird in der genetischen Optimierung mit einer vorgegebenen Wahrscheinlichkeit p_K durchgeführt, wobei typische Werte im Bereich $p_K \in [0.6, 1.0]$ liegen.

Binäre Kodierung

Die Kreuzung zweier *binär* kodierter Variablen (Faktoren) mit zwei Schnittpunkten ist beispielhaft in Abbildung 10.16 dargestellt. Im ersten Schritt des Kreuzungsalgorithmus werden zufällig zwei Schnittpunkte in der binären Darstellung ausgewählt. Anschließend werden die Kinder durch einfaches Vertauschen der Bits zwischen den zwei Schnittpunkten ermittelt. Durch die Kreuzung werden zwei Kinder erzeugt, die ähnliche Faktorwerte aufweisen wie deren Eltern. In der Literatur finden sich ebenfalls Verfahren mit nur einem Schnittpunkt, bei denen alle Bits auf einer Seite des Schnittpunkts vertauscht werden oder Verfahren bei denen für jedes Bit einzeln mit einer vorgegebenen Wahrscheinlichkeit entschieden wird, ob das Bit vertauscht wird.

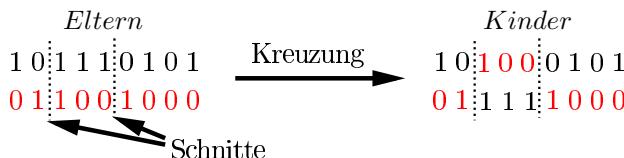


Abb. 10.16 Kreuzung (binär)

Führen zum Beispiel große Faktorwerte zu einer guten Zielgröße werden beide Eltern im Laufe der Optimierung mit hoher Wahrscheinlichkeit große Werte und somit Einsen in den oberen Bitbereichen aufweisen (Abbildung 10.17). Egal wo die Schnitte liegen, werden die Kinder ebenfalls hohe Faktorwerte aufweisen, so dass der Trend über die Generationen erhalten bleibt.

Reelle Faktoren

Zur Kreuzung *reeller* Zahlen sind verschiedene Algorithmen entwickelt worden (Fuzzy Recombination [48], BLX [20], SBX [14, 17, 4], vSBX [4, 5], PCX [16], XLM [46], PNX [5], PBX [35], UNDX [39]).

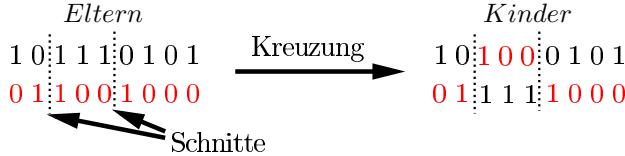


Abb. 10.17 Kreuzung (binär) mit hohen Faktorwerten

Das einfache BLX (*Blend Crossover*) erzeugt Kinder aus zwei Eltern a und b durch die zufällige Wahl einer reellen Zahl aus einem Bereich, der wie folgt definiert ist (Abbildung 10.18):

$$[\min(a, b) - \alpha |a - b|, \max(a, b) + \alpha |a - b|] \quad (10.39)$$

Bei steigendem α wird somit ein größerer Bereich durch die Kinder abgedeckt. Ein typischer Wert für α liegt bei $\alpha = 0.5$.

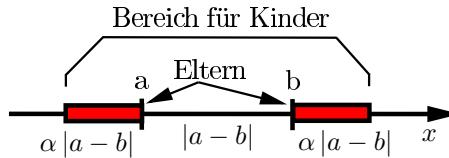


Abb. 10.18 Kreuzung durch BLX

Der vSBX Algorithmus ist eine Verbesserung des SBX (*Simulated Binary Crossover*) und verwendet zur Bestimmung der Faktoreinstellung k eines Kindes zwei unterschiedliche Gleichungen $k_{1,2}$, wobei vor der Kreuzung eine der beiden Gleichungen mit gleicher Wahrscheinlichkeit ausgewählt wird:

$$\begin{aligned} k_1 &= \begin{cases} 0.5[(1+\beta_1)a + (1-\beta_1)b] & 0 < u \leq 0.5 \\ 0.5[(3-\beta_2)a - (1-\beta_2)b] & 0.5 < u \leq 1 \end{cases} \\ k_2 &= \begin{cases} 0.5[(1-\beta_1)a + (1+\beta_1)b] & 0 < u \leq 0.5 \\ 0.5[-(1-\beta_2)a + (3-\beta_2)b] & 0.5 < u \leq 1 \end{cases} \end{aligned} \quad (10.40)$$

$$\beta_1 = \left(\frac{1}{2u}\right)^{\frac{1}{\eta_K+1}} \quad \beta_2 = \left(\frac{1}{2(1-u)}\right)^{\frac{1}{\eta_K+1}} \quad (10.41)$$

Die Zufallsvariable $u \in [0, 1]$ ist dabei für jede Variable (Faktor) und Kreuzung neu zu bestimmen. Der im Vorfeld festzulegende und meist konstante Parameter η_K bestimmt die Streuung der Kinder um die Eltern. Abbildung 10.19 zeigt beispielhaft den Einfluss von η_K auf die Kreuzung zweier Eltern, welche durch zwei Faktoren (x_1, x_2) definiert sind. Die Erhöhung von η_K führt zu einer Verringerung der Streuung. Typische Werte für η liegen zwischen 5 und 20.

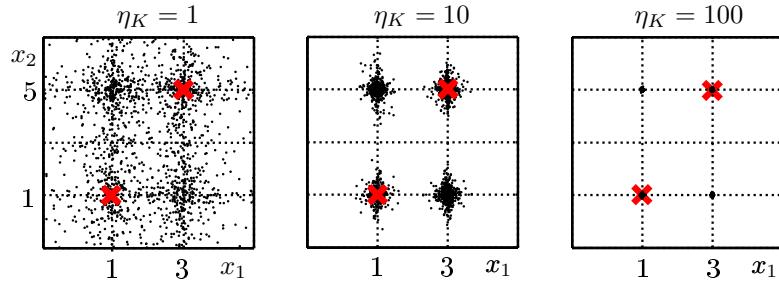


Abb. 10.19 Kreuzung reeller Zahlen (1,1) und (3,5) durch vSBX

10.6.2 Mutation

Die Mutation ändert zufällig Faktorstufen im Faktorraum, wodurch die Erzeugung von Datenpunkten (Kinder) ermöglicht wird, die nicht durch eine Kreuzung der ausgewählten Eltern erzielt werden können. Dadurch wird dem Optimierungsverfahren ermöglicht sich selbstständig aus lokalen Minima zu befreien, da in regelmäßigen Abständen zufällige Bereiche des Faktorraums und die dazugehörigen Zielgrößen geprüft werden. Die festzulegende Wahrscheinlichkeit der Mutation ist mit $p_M = 0 \dots 0.5$ deutlich geringer als die Wahrscheinlichkeit der Kreuzung (Kapitel 10.6.1).

Binäre Kodierung

Die Mutation eines *binär* kodierten Faktors wird durch eine zufällige Änderung einzelner Bits erzeugt (Abbildung 10.20), wobei jedes einzelne Bit mit der vorgegebenen Wahrscheinlichkeit p_M vertauscht wird.



Abb. 10.20 Mutation (binär)

Reelle Faktoren

Die Mutation eines *reellen* Faktors kann grundsätzlich durch die Wahl eines zufälligen Wertes aus dem Definitionsbereich erreicht werden, wobei sich jedoch zeigt, dass eine Streuung um den aktuellen Faktorwert sich positiv auf den Optimierungsprozess auswirkt. Typische Verfahren sind dabei *non-uniform-, Normverteilung* und

Polynom Mutation [43]. Bei gegebener unterer und oberer Grenze des zu mutierenden Faktors $x_u \leq x \leq x_o$ wird die *Polynom Mutation* durch folgende Gleichung berechnet.

$$x_M = x + \beta (x_o - x_u)$$

$$\text{mit } \beta = \begin{cases} \left(2u + [1 - 2u] \left[1 - \frac{x - x_u}{x_o - x_u}\right]^{\eta_{M+1}}\right)^{\frac{1}{\eta_{M+1}}} - 1 & u \leq 0.5 \\ 1 - \left(2[1 - u] + 2[u - 0.5] \left[1 - \frac{x_o - x}{x_o - x_u}\right]^{\eta_{M+1}}\right)^{\frac{1}{\eta_{M+1}}} & u > 0.5 \end{cases} \quad (10.42)$$

Dabei ist u eine Zufallszahl im Bereich $u \in [0, 1]$ und η_M eine Zahl größer Null, welche die Streubreite der Mutation beeinflusst. Abbildung 10.21 zeigt die Auswirkung von η_M auf die Mutation zweier Faktoren ($x_1 = 3, x_2 = 5$) mit $x_1, x_2 \in [1, 9]$. Typische Werte für η_M liegen im Bereich $\eta_M \in [5, 50]$.

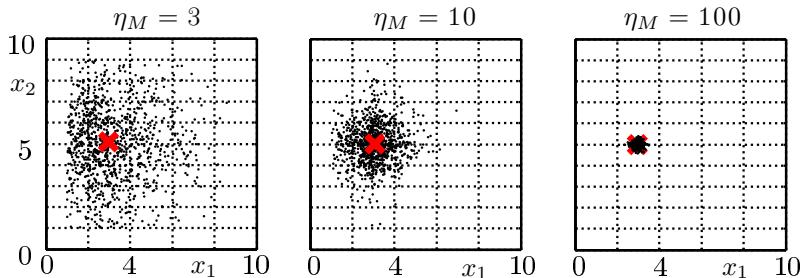


Abb. 10.21 Polynom Mutation reeller Zahlen ($x_1 = 3, x_2 = 5$)

10.6.3 Exemplarische Verfahren (NSGA-II und ε -MOEA)

Der erfolgreiche Einsatz genetischer Optimierungsmethoden in vielen Arbeitsbereichen hat zur Entwicklung verschiedener allgemeiner und auch spezifischer Algorithmen geführt. Bekannte Verfahren sind unter Anderen: VEGA [44], HLGA [34], MOGA [22, 23], NPGA [28], SPEA(2) [60, 61, 56, 59], NSGA-II [12, 19, 13], PAES [32], PESA [10] und ε -MOEA [18].

Im Rahmen dieses Buchs werden exemplarisch die Verfahren NSGA-II und ε -MOEA kurz dargestellt.

NSGA-II (Non-dominated Sorting Genetic Algorithm)

Das Verfahren NSGA-II findet sich in vielen kommerziellen und freien Optimierungstools [12, 19, 13]. Der prinzipielle Ablauf dieses Verfahrens ist in Algorithmus 18 dargestellt.

- 1 Wähle n_g zufällige Individuen, welche die Start-Generation G_0 bilden
- 2 Bestimme den *Rang* und die *Crowding Distance* (CD) für jedes der n_g Individuen
- 3 **solange kein Stop-Kriterium erfüllt ist tue**
- 4 Wähle geeignete Eltern E_i aus der aktuellen Generation G_i mittels Rang und CD
- 5 Berechne Kinder K_i durch Kreuzung der Eltern
- 6 Mutation der Kinder K_i
- 7 Bestimme *Rang* und *CD* für alle Individuen ($G_i \cup K_i$)
- 8 Erzeuge eine neue Generation G_{i+1} mittels Rang und CD aus $G_i \cup K_i$
- 9 **Ende**

Algorithmus 18 : NSGA II

Im Initialisierungsschritt des NSGA-II werden n_g zufällige Individuen ausgewählt, welche die erste Generation G_0 bilden. Zur Beurteilung und zum Vergleich der einzelnen Individuen wird der *Rang* und die *Crowding Distance* (CD) jedes einzelnen Individuums bestimmt.

Rang

Zur Bestimmung des Rangs eines Individuums werden alle Individuen in Grenzen (Fronten) eingeteilt (Abbildung 10.22). Der Rang *eins* enthält dabei alle Individuen der aktuellen Pareto-Grenze, also alle Individuen, die nicht von einem anderen Individuum dominiert werden. In Rang zwei befinden sich alle Individuen, die auf der Pareto-Grenze liegen, wenn alle Individuen des Rangs *eins* entfernt und somit nicht berücksichtigt werden. Alle weiteren Ränge enthalten entsprechend die Individuen, welche auf der Pareto-Grenze liegen, wenn alle Individuen der vorherigen Ränge entfernt werden. Da dieses grundsätzliche Bestimmungsverfahren viel Rechenzeit beansprucht, wurde von DEB et al. ein Sortieralgorithmus entwickelt, welcher die benötigte Rechenzeit deutlich reduziert [19].

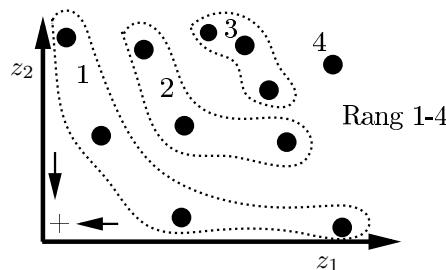


Abb. 10.22 Rang Zuordnung

Crowding Distance

Die Crowding Distance (CD) beurteilt die Individuen eines Rangs untereinander. Dabei ist die CD ein Maß für den Abstand eines Individuums zu seinen benachbarten Individuen. Zur Bestimmung der CD werden die Individuen eines Rangs im ersten Schritt separat für jede Zielgröße aufsteigend sortiert. Jedem Individuum i , welches aktuell für eine beliebige Zielgröße das kleinste oder größte (erste oder letzte) Element darstellt, wird der Wert $CD_i = \infty$ zugewiesen. Alle anderen Individuen besitzen bei separater Betrachtung der einen Zielgröße jeweils ein Individuum mit kleinerem oder größerem Qualitätswert. Die CD eines Individuums ist nun die Summe der Abstände zu den benachbarten Individuen für jede Zielgröße (Abbildung 10.23). Damit die Zielgrößen untereinander vergleichbar sind, werden diese durch die aktuelle Spannbreite der jeweiligen Zielgröße, welche durch die zwei aktuellen Randpunkte jeder Zielgröße definiert sind, normiert.

$$CD_i = \sum_{k=1}^{n_z} \frac{z_{k,i+1} - z_{k,i-1}}{z_{k,max} - z_{k,min}} \quad (10.43)$$

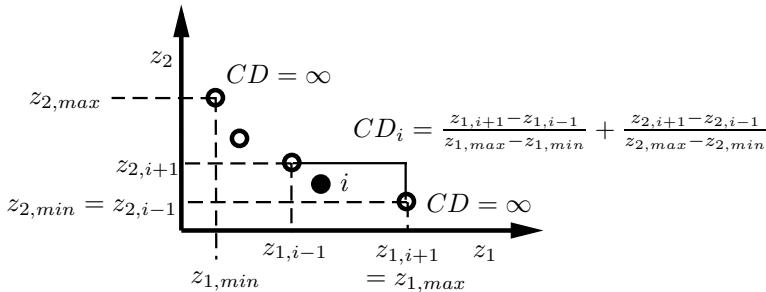


Abb. 10.23 Crowding Distance

Zur Auswahl eines Elternteils werden jeweils zwei zufällige Individuen aus der aktuellen Generation ausgewählt und deren Fitness verglichen, wobei das fittere Individuum ausgewählt wird. Die Fitness wird dabei durch den Rang und die Crowding Distance beschrieben. Hat ein Individuum einen geringeren Rang, so ist es aus einer Grenze, die näher an der gesuchten Pareto-Grenze liegt und somit zu bevorzugen. Weisen beide Individuen den gleichen Rang auf, so ist das Individuum zu wählen, welches in einem Bereich mit weniger Individuen liegt, also ein größere CD aufweist. Für eine gleichmäßige Verteilung von Individuen auf der betrachteten Grenze wird im nächsten Optimierungsschritt dadurch in dünn besetzten Gebieten nach neuen Individuen gesucht. Randpunkte werden durch die Zuweisung $CD = \infty$ automatisch inneren Individuen vorgezogen. Weisen zwei Individuen den gleichen Rang und CD auf, so wird ein zufälliges der beiden Individuen als Elternteil gewählt. Ein Individuum i ist einem Individuum j somit vorzuziehen, wenn:

$$Rang_i < Rang_j \vee [Rang_i = Rang_j \wedge CD_i > CD_j] \quad (10.44)$$

Aus zwei Eltern werden im nächsten Schritt durch Kreuzung (Kapitel 10.6.1) und Mutation (Kapitel 10.6.2) zwei neue Individuen (Kinder) berechnet. Die Auswahl von Eltern und Erzeugung der Kinder wird solange fortgesetzt bis eine vorgegebene Anzahl von neuen Individuen ermittelt wurde (zum Beispiel n_g neue Individuen). Zur Erzeugung der neuen Generation G_{i+1} werden die Individuen der vorhergehenden Generation G_i und alle neu erzeugten Individuen (Kinder) zusammen betrachtet und nach Rang und CD sortiert. Die besten n_g Individuen nach Gleichung 10.44 bilden anschließend die neue Generation G_{i+1} (Abbildung 10.24).

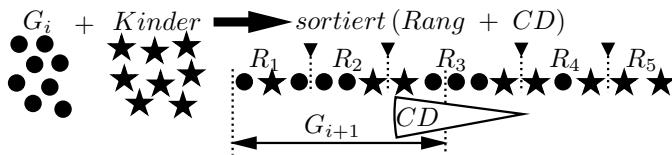


Abb. 10.24 Erzeugung einer neuen Generation (NSGA-II)

Beispiel

Zur Veranschaulichung des dargestellten Verfahrens wird ein Testproblem nach *Kursawe* mit zwei Zielgrößen und drei Faktoren betrachtet [33, 12].

$$\begin{aligned} z_1(x) &= \sum_{i=1}^2 -10e^{-0.2\sqrt{x_i^2+x_{i+1}^2}} \\ z_2(x) &= \sum_{i=1}^3 |x_i|^{0.8} + 5 \sin(x_i^3) \\ -5 \leq x_i &\leq 5 \end{aligned} \quad (10.45)$$

Eine Monte-Carlo-Simulation (Kapitel 8.3.1) mit 1000 Individuen zeigt eine deutliche Streuung im Zielgrößenraum und lässt die ungefähre Position der gesuchten Pareto-Grenze erkennen (Abbildung 10.25). Nur wenige Individuen liegen dabei auf oder in der Nähe der gefundenen Pareto-Grenze. Weiterhin zeigt sich, dass in der durchgeführten Monte-Carlo-Simulation die gefundenen Lösungen für kleine z_2 deutlich von der idealen Pareto-Grenze abweichen. Eine reine Monte-Carlo-Simulation ist somit zur Bestimmung der Pareto-Grenze nicht zielführend.

Abbildung 10.26 zeigt zum Vergleich den Verlauf der NSGA-II Optimierung bei einer gewählten Populationsgröße von $n_g = 32$. Ausgehend von der im gesamten Zielraum verteilten Startpopulation G_0 entwickelt sich die Population schnell in Richtung der gesuchten Pareto-Grenze. Bereits in der achten Generation G_8 und somit nach 288 Auswertungen der Testfunktion wird eine ähnliche Güte der Pareto-Grenze erreicht wie nach 1000 Auswertungen für die Monte-Carlo-Simulation. Nach 800 Auswertungen G_{24} ist die gefundene Pareto-Grenze bereits deutlich näher

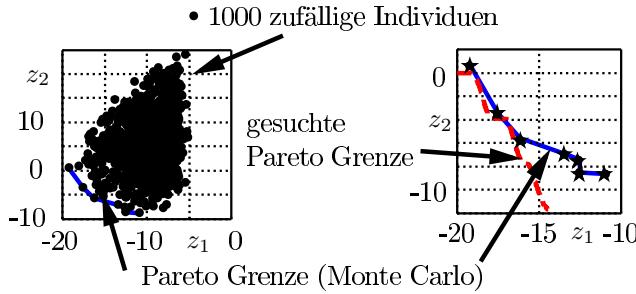


Abb. 10.25 Testproblem nach Kursawe, Monte-Carlo

an der idealen und gesuchten Grenze als nach der Monte-Carlo-Simulation (1000 Auswertungen). Nach 3000 Auswertungen und 92 Generationen liegen alle Individuen gleichmäßig verteilt an der gesuchten Pareto-Grenze (Abbildung 10.27). Die Konvergenzgeschwindigkeit der Optimierung hängt neben der Komplexität des betrachteten Systems ebenfalls von der Wahl der Optimierungsparameter (Populationsgröße, Kreuzungswahrscheinlichkeit, ...) ab. Unabhängig von der erreichten Konvergenzgeschwindigkeit zeigt der NSGA-II Algorithmus für die meisten Optimierungsprobleme ein robustes Lösungsverhalten.

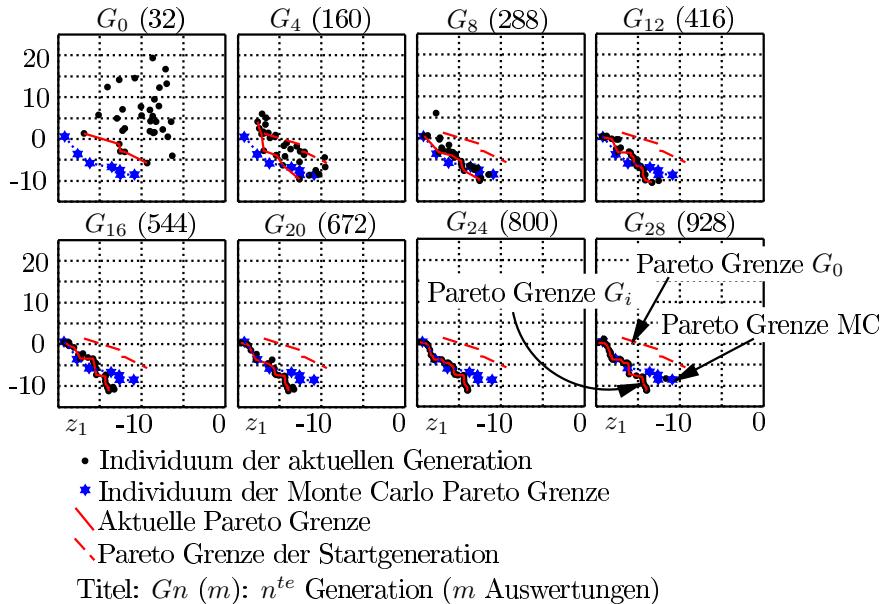


Abb. 10.26 Optimierungsverlauf (NSGA-II)

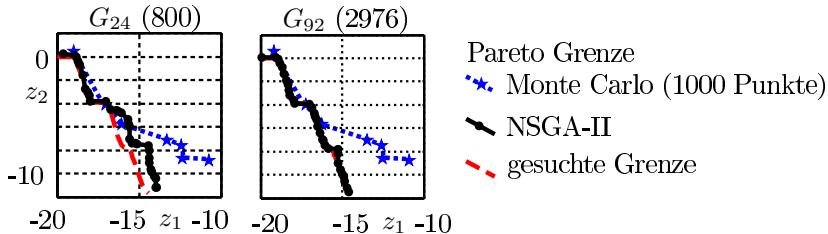
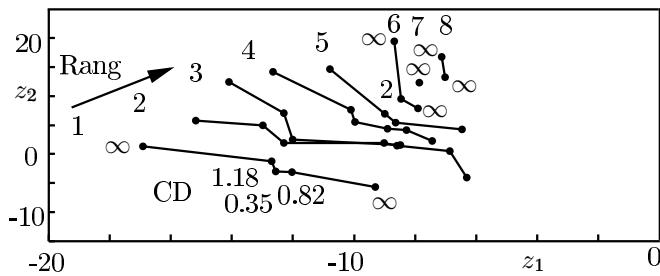


Abb. 10.27 Pareto-Grenze (NSGA-II)

Zur Veranschaulichung der Crowding Distance (CD) ist in Abbildung 10.28 die erste Generation der durchgeführten NSGA-II Optimierung mit Rang und einigen CDs dargestellt. Alle äußerer Punkte eines Rangs sind große Zahlen beziehungsweise ∞ zugewiesen. Im ersten Rang (aktuelle Pareto-Grenze) weist der mittlere Datenpunkt die geringste CD auf, so dass alle anderen Datenpunkte des Rangs im direkten Vergleich dem mittleren Datenpunkt vorzuziehen sind.

Abb. 10.28 Rang und CD der ersten Generation G₀ des Testproblems nach Kursawe

Werden zusätzlich folgende Randbedingungen der Optimierungsaufgabe hinzugefügt, so wird die in Abbildung 10.29 dargestellten Pareto-Grenze durch das NSGA-II Verfahren problemlos ermittelt.

$$\left. \begin{array}{l} z_1 \leq -17.5 \\ z_1 \geq -16.5 \\ z_2 \leq -11.0 \\ z_2 \geq -9.0 \end{array} \right\} \rightarrow g(z_1) = -0.5 + |-17.0 - z_1| \quad (10.46)$$

$$\left. \begin{array}{l} z_1 \leq -17.5 \\ z_1 \geq -16.5 \\ z_2 \leq -11.0 \\ z_2 \geq -9.0 \end{array} \right\} \rightarrow g(z_2) = -1.0 + |-10.0 - z_2|$$

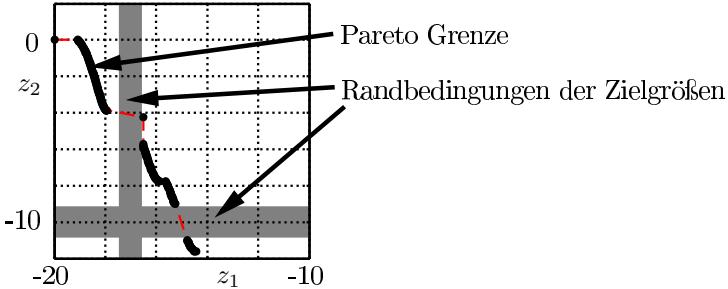
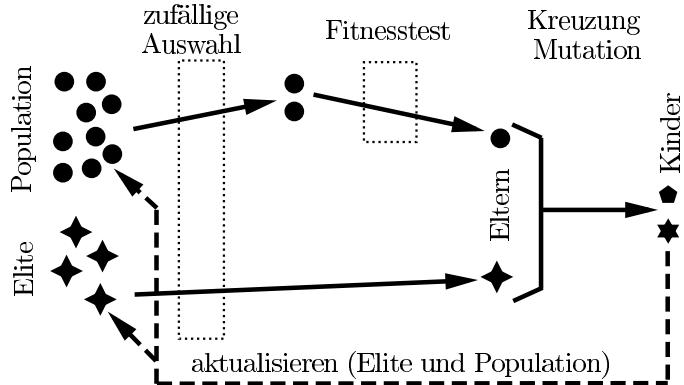


Abb. 10.29 Pareto-Grenze mit Randbedingungen

ε -MOEA

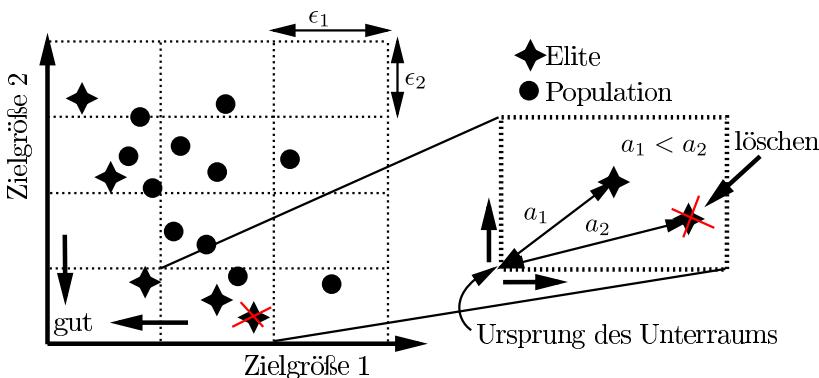
Neben Verfahren wie NSGA-II, die mit einer einzelnen Gruppe von Individuen arbeiten, welche während des Optimierungsprozesses kontinuierlich verbessert werden, existieren Algorithmen wie zum Beispiel das ε -MOEA [18], die zwei separate Gruppen von Individuen zur Optimierung einsetzen. Neben der Populationsgruppe, welche Individuen der aktuellen Generation beinhaltet, enthält die zweite Gruppe nur Individuen der aktuellen Pareto-Grenze und wird als Elite Gruppe bezeichnet. Die Elite Gruppe besitzt im Gegensatz zur Populationsgruppe keine vorgegebene Größe und kann je nach gefundenen Individuen anwachsen oder schrumpfen. Grundsätzlich entspricht dieses dem dargestellten Ansatz zur Erweiterung naturanaloger Optimierungsverfahren für mehrdimensionale Optimierungsaufgaben aus Kapitel 10.5.6. Die Elitegruppe ist dabei mit dem externen Archiv vergleichbar. Der Optimierungsablauf von ε -MOEA ist in Abbildung 10.30 schematisch dargestellt. Zur Erzeugung zweier Eltern werden aus der Populationsgruppe zwei zufällige Individuen ausgewählt und verglichen. Ist ein Individuum fitter (dominant) wird dieses als erstes Elternteil gewählt. Sollte kein Individuum dominieren so wird ein zufälliges der Beiden gewählt. Das zweite Elternteil wird zufällig aus der Elitegruppe gewählt, wobei hier sichergestellt ist, dass es sich um ein besonders fittes Individuum handelt, da es auf der aktuellen Pareto-Grenze liegt. Durch Kreuzung und Mutation werden aus den Eltern neue Individuen (Kinder) berechnet und anschließend jeweils in die Populations- und Elitegruppe einsortiert. Die Populationsgröße bleibt konstant, wobei die Elitegruppe alle Pareto-optimalen Elemente enthält und somit eine variable Größe aufweist.

Werden als Faktoren reelle Zahlen verwendet, können unendlich viele Ergebnisse auf der Pareto-Grenze gefunden werden. Während der Optimierung würde dadurch die Individuenanzahl der Elitegruppe kontinuierlich ansteigen. Eine uneingeschränkte Steigerung der Individuenanzahl würde jedoch den Optimierungsprozess deutlich verlangsamen, da die gesamte Pareto-Grenze und somit jeder Punkt der Elitegruppe gleichzeitig optimiert werden muss. In der Praxis sind Lösungen für ein Problem gesucht, die sich merklich voneinander unterscheiden, so dass eine gezielte Auswahl der 'besten' Pareto-optimalen Lösung stattfinden kann, was gegen

Abb. 10.30 ϵ -MOEA

eine uneingeschränkte Anzahl an Individuen in der Elite Gruppe spricht (siehe auch Kapitel 10.5.6).

Der ϵ -MOEA Algorithmus verwendet zur Begrenzung der Elitegruppe eine Unterteilung des Zielgrößenraums in Schichten der Breite ϵ_i , wie es in Abbildung 10.31 dargestellt ist. In jedem durch diese Unterteilung gebildeten Unterraum darf nur *ein* Individuum der Elitegruppe existieren. Sind mehrere Individuen in einem Unterraum vorhanden, so wird nur das Individuum in die Elitegruppe übernommen, welches den geringsten Abstand zum Ursprung des n_z -dimensionalen Unterraums aufweist (Abbildung 10.31, rechts). Dieses zusätzliche Gütekriterium wird eingeführt, da alle Individuen aus der Elitegruppe auf der Pareto-Grenze liegen und somit über die Dominanz ohne die Crowding Distance (siehe Seite 360) kein Individuum bevorzugt werden kann.

Abb. 10.31 ϵ -MOEA Gitter

Abbildungen 10.32 und 10.33 zeigen den Optimierungsverlauf der Aufgabe aus Gleichung 10.45. Beginnend mit einer weit streuenden Startpopulation wer-

den schnell Individuen in der Nähe der wahren Pareto-Grenze ermittelt. Bereits in Generation 6 (nach 280 Auswertungen der Funktion) wird die Pareto-Grenze gut abgebildet. Die Vergrößerung der Schichtweite von $\epsilon = 0.1$ auf $\epsilon = 0.5$ weist eine deutliche Reduktion der Individuenanzahl in der Elitegruppe auf, wobei eine gute Gleichverteilung durch beide Einstellungen erzielt wird (z.B. von 26 auf 8 in Generation 14).

Bei steigender Anzahl der betrachteten Zielgrößen weist ϵ -MOEA schnell große Elitegruppen auf, so dass die geschickte Wahl der ϵ Werte an Bedeutung zunimmt. Abhilfe können hier zusätzliche Randbedingungen für die Zielgrößen (Maximalwerte) schaffen, die sich zum Beispiel automatisch an die aktuell gefundenen Bereiche der Zielgrößen anpassen und bei steigender Anzahl von Individuen in der Elitegruppe den Zielgrößenbereich weiter einschränken.

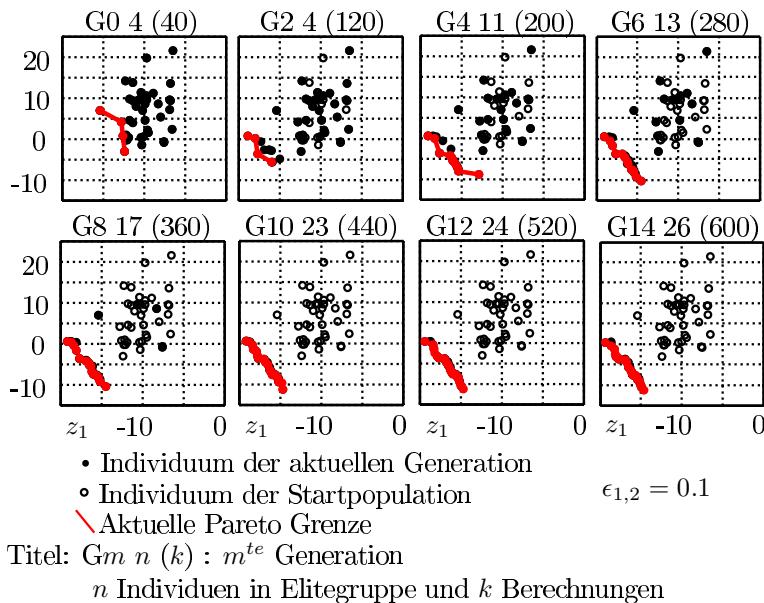


Abb. 10.32 ϵ -MOEA: Beispiel mit $\epsilon = 0.1$

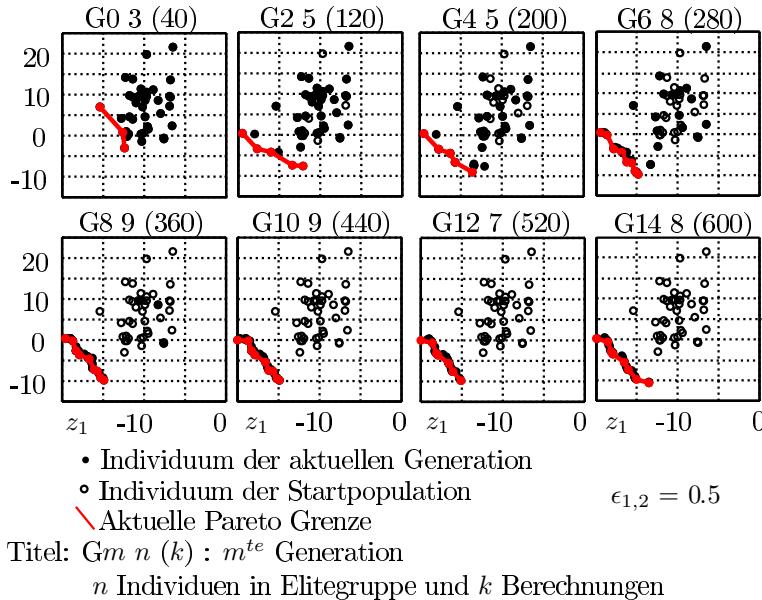


Abb. 10.33 ϵ -MOEA: Beispiel mit $\epsilon = 0.5$

10.7 Qualität multidimensionaler Pareto-Grenzen

Zur Beurteilung verschiedener Optimierungsalgorithmen werden unterschiedliche Kriterien herangezogen, die in drei Hauptgruppen einteilbar sind.

1. Qualität des Ergebnisses bzw. der Pareto Front
2. Rechen- und Zeitaufwand
3. Robustheit des Ergebnis

Gerade die Beurteilung der Qualität einer mehrdimensionalen Pareto Front ist deutlich schwerer als die einer einzelnen Zielgröße, bei der lediglich ein einzelner skalarer Wert die Qualität des Ergebnisses bestimmt. Zur Qualitätsbeurteilung einer berechneten Pareto-Grenze schlägt zum Beispiel Zitzler et al. folgende Unterkriterien vor [57, 26]

1. Der Abstand zwischen der gefundenen Pareto-Grenze und der idealen Grenze soll minimal sein, wobei die ideale Pareto-Grenze bekannt sein muss.
2. Eine *gute* Verteilung der Ergebnisse auf der Pareto Front. In den meisten Fällen ist dies gleichbedeutend mit einer Gleichverteilung. Weist die ideale Grenze jedoch Lücken oder Sprünge auf, so wird die Annahme einer Gleichverteilung nur bedingt sinnvolle Ergebnisse liefern.
3. Der von der Pareto-Grenze abgedeckte Bereich in jeder Zieldimension soll maximal sein.
4. Eine hohe Anzahl von nicht redundanten Pareto-optimalen Lösungen.

In vielen Fällen werden mehrdimensionale Optimierungsaufgaben nach unterschiedlichen Rechenzeiten oder Funktionsaufrufen untersucht, um dann zu entscheiden, ob sich bereits ein stabiles Ergebnis eingestellt hat. Ist dies der Fall, wird der Optimierungsprozess beendet. Die Beurteilung ist dabei von der persönlichen Einschätzung des Anwenders abhängig. Mit einer robusten Qualitätsbeurteilung von Pareto-Grenzen wäre es möglich, ein Konvergenzkriterium zum Abbruch von Optimierungen einzuführen. Verschiedene Ansätze sind in unterschiedlichen Quellen dargestellt, wobei eine Übersicht der Verfahren in [31] zu finden ist:

1. S-Metrik: Größe des dominierten Bereichs einer Pareto-Grenze [56]
2. C-Metrik: Überdeckung zweier Pareto-Grenzen [56]
3. D-Metrik: Unterschied der Überdeckung zweier Pareto-Grenzen [56]
4. Generationenabstand (*Generational Distance*): Durchschnittlicher Abstand zwischen aktueller und wahrer Pareto-Grenze [47]
5. Fehlerverhältnis (*Error Ratio*): Anteil der approximierten Pareto-optimalen Punkte, die in der wahren Pareto-Grenze enthalten sind [47]
6. Maximaler Pareto Front Fehler: Maximaler Minimalabstand zwischen Punkten zweier Pareto-Grenzen [47]
7. ONVG (Overall Nondominated Vector Generation): Anzahl der gefundenen Pareto-optimalen Punkte [47]
8. ONVGR (Overall Nondominated Vector Ratio): ONVG im Verhältnis zur wahren Pareto-Grenze [47]
9. Crowding Distance: Beurteilt die Packungsdichte der Pareto-optimalen Punkte (Kapitel 9) [12]
10. $R_{1,2,3}$ Indikatoren [27]

Weitere Abstandskriterien finden sich in [45] und [11]. Als Beispiel werden hier die verschiedenen R-Indikatoren [47] sowie die S-, C- und D-Metriken [56] genauer dargestellt.

10.7.1 R-Indikatoren

Am Ende einer typischen Optimierung mit mehreren Zielgrößen liegen unterschiedliche Pareto-optimale Lösungen vor und ein Entscheidungsträger DM (*decision-maker*) muss eine endgültige Lösung aus der Gruppe der Pareto-optimalen Ergebnisse auswählen. Diese Wahl wird durch die persönlichen Präferenzen des Entscheidungsträger beeinflusst, so dass die Entscheidung je nach Anwender unterschiedlich ausfallen kann. Grundsätzlich können jedoch verschiedene Annahmen und Aussagen über den Auswahlprozess getroffen werden. Jede Präferenz des Anwenders kann durch eine Nutz- oder Qualitätsfunktion u (*utility function*) beschrieben werden, welche jedem Punkt im Zielgrößenraum einen Qualitätswert zuordnet [47].

$$u : \Re^{n_z} \rightarrow \Re \quad (10.47)$$

Ziel der Auswahlprozesses des Entscheidungsträgers ist es im Endeffekt den Qualitätswert u zu optimieren. Ein verwendeter Qualitätswert u ist kompatibel mit dem allgemeinen Dominanzkriterium (Kapitel 10.2), wenn und nur wenn gilt:

$$\forall z^1, z^2 \in \Re^{n_z} \quad z^1 \prec z^2 \Rightarrow u(z^1) \leq u(z^2) \quad (10.48)$$

Strikt kompatibel mit dem Dominanzkriterium sind alle Qualitätsfunktionen, wenn und nur wenn sie hingegen folgende Bedingung erfüllen:

$$\forall z^1, z^2 \in \Re^{n_z} \quad z^1 \prec z^2 \Rightarrow u(z^1) < u(z^2) \quad (10.49)$$

Kompatible und strikt kompatible Qualitätsfunktionen werden in den Gruppen U_c und U_{sc} zusammengefasst.

Mit der Hilfe von *parametrischen* Qualitätsfunktionen können Gruppen von Qualitätsfunktionen ($u(\mathbf{z}, \mathbf{r})$, $\mathbf{r} \in D(\mathbf{r}) \subseteq \Re^n$) definiert werden, wobei \mathbf{r} ein Parametervektor ist und $D(\mathbf{r})$ der dazugehörige Definitionsbereich. Eine Gruppe von parametrischen Qualitätsfunktionen ist somit definiert durch [27]:

$$U(\mathbf{r}) = \{u(\mathbf{z}, \mathbf{r}) \mid \mathbf{r} \in D(\mathbf{r})\} \quad (10.50)$$

Eine mögliche und häufig verwendete Gruppe U_p von parametrischen Qualitätsfunktionen basiert auf der gewichteten L_p Norm:

$$L_p(\mathbf{z}^1, \mathbf{z}^2, \Lambda) = \left(\sum_{j=1}^{n_z} \lambda_j |z_j^1 - z_j^2|^p \right)^{1/p} \quad (10.51)$$

$$\begin{aligned} \text{mit } p \in \{1, 2, \dots\} + \{\infty\} \text{ und } \Lambda = [\lambda_1, \dots, \lambda_{n_z}], \lambda_j \geq 0 \\ \implies u_p(\mathbf{z}, \mathbf{z}^*, \Lambda, p) = - \left(\sum_{j=1}^{n_z} \lambda_j |z_j^* - z_j|^p \right)^{1/p} \end{aligned} \quad (10.52)$$

Dabei steht z für einen zu beurteilenden Punkt einer approximierten Pareto-Grenze, $z^{1,2}$ für zwei beliebige Punkte und z^* für einen ideal angenommenen Punkt im Zierraum. Für die Sonderfälle $p = \infty$ und $p = 1$ ergeben sich folgende gewichtete Qualitätsfunktionen [27, 58]:

$$\text{Chebychev : } u_\infty(\mathbf{z}, \mathbf{z}^*, \Lambda) = - \max_j \{ \lambda_j |z_j^* - z_j| \} \text{ mit } j \in [1, \dots, n_z] \quad (10.53)$$

$$\text{gewichtete Summe : } u_1(\mathbf{z}, \mathbf{z}^*, \Lambda) = - \sum_{j=1}^{n_z} \lambda_j |z_j^* - z_j| \text{ mit } j \in [1, \dots, n_z] \quad (10.54)$$

$$\text{Kombination : } u_k(\mathbf{z}, \mathbf{z}^*, \Lambda) = u_\infty(\mathbf{z}, \mathbf{z}^*, \Lambda) + \rho u_1(\mathbf{z}, \mathbf{z}^*, \Lambda), \rho \geq 0 \quad (10.55)$$

Sei nun A eine Approximation der gesuchten Pareto-Grenze und $u^*(A)$ der Maximalwert der verwendeten Qualitätsfunktion:

$$u^*(A) = \max_{\mathbf{z} \in A} \{u(\mathbf{z})\} \quad (10.56)$$

Als Beispiel seien nun zwei Approximationen A und B einer Pareto-Grenze gegeben, wobei der Entscheidungsträger Approximation A durch seine Präferenzen bevorzugt. In diesem Fall findet der Entscheidungsträger eine spezielle Lösung in A , die einen besseren Kompromiss aller Zielgrößen aufweist als alle Lösungen aus B und gleichzeitig jede Lösung in A nicht schlechter ist als alle Lösungen aus B . Sei nun $U(A < B) \subseteq U$ eine Gruppe von Qualitätsfunktionen für die A besser als B ist:

$$U(A < B) = \{u \in U \mid u^*(A) > u^*(B)\} \quad (10.57)$$

Bei einem Vergleich zweier approximierter Grenzen werden von Jaszkiewicz verschiedene *Outperformance*-Beziehungen einer vorgegebenen Gruppe U von Qualitätsfunktionen definiert [29]:

Outperformance Beziehung

Eine Approximation A outperforms B bezüglich einer Gruppe U von Qualitätsfunktionen, wenn $U(A < B) \neq \emptyset$ und $U(B < A) = \emptyset$.

Es existiert eine Gruppe von Qualitätsfunktionen, welche in A bessere Ergebnisse liefert als in B , wobei das Gegenteil nicht zutrifft. Unabhängig von den Präferenzen des Entscheidungsträgers kann davon ausgegangen werden, dass er niemals ein dominiertes Ergebnis als besten Kompromiss auswählen würde [29]. Das bedeutet, dass bei weiteren Analysen lediglich nicht-dominierte Punkte $ND(A \cup B)$ aus der gemeinsamen Menge $A \cup B$ betrachtet werden müssen.

Schwache Outperformance Beziehung

Eine Approximation A schwach (*weak*) outperforms B [$A O_w B$], wenn $A \neq B$ und $ND(A \cup B) = \emptyset$.

Dies ist der Fall, wenn jeder Punkt von B ebenfalls in A enthalten ist und mindestens ein Punkt von A nicht in B enthalten ist (z.B. B und A in Abbildung 10.34)

Starke Outperformance Beziehung

Eine Approximation A stark (*strong*) outperforms B [$A O_s B$], wenn $ND(A \cup B) = A$ und $B \setminus ND(A \cup B) \neq \emptyset$.

Dies ist der Fall, wenn für jeden Punkt in B ein Punkt in A existiert, der gleich ist oder der den Punkt in B dominiert und gleichzeitig mindestens ein Punkt in B von einem Punkt in A dominiert wird (z.B. C und A in Abbildung 10.34). Die Differenz von B und $ND(A \cup B)$ ist somit nicht leer.

Komplette Outperformance Beziehung

Eine Approximation A komplett (*completely*) outperforms B [$A O_c B$], wenn $ND(A \cup B) = A$ und $B \cap ND(A \cup B) = \emptyset$.

Dieses ist der Fall, wenn jeder Punkt aus B von mindestens einem Punkt in A dominiert wird (z.B. C und A in Abbildung 10.34).

Es gilt $O_c \subset O_s \subset O_w$, so dass die komplette Outperformance Relationen die starke Outperformance enthält und die stark Outperformance die schwache Outperformance. Mit diesen Relationen können leider nicht alle Pareto Approximationen mitein-

ander eindeutig verglichen werden und es ist lediglich ein qualitativer Vergleich möglich.

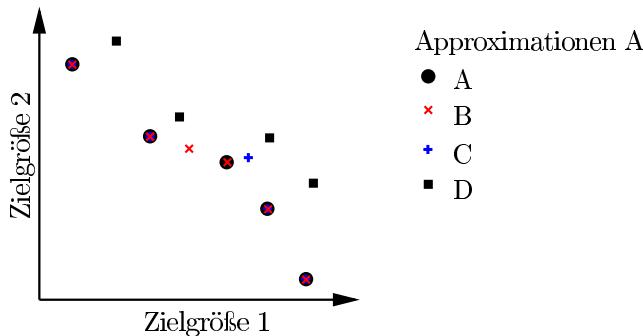


Abb. 10.34 Beispiel für verschiedene Outperformance Stufen: BO_wA , CO_sA , DO_cA

Daher sei nun eine Gruppe U von Qualitätsfunktionen gegeben, bei der jede (der Qualitätsfunktionen) eine Wahrscheinlichkeit besitzt, genau die des Entscheidungsträgers zu sein. Wenn nun zwei Approximationen der Pareto-Grenze verglichen werden, ist es sinnvoll, die Approximation als besser zu beurteilen, bei der mit hoher Wahrscheinlichkeit die Mehrzahl der gewählten Qualitätsfunktionen ein besseres Ergebnis liefert. Die Wahrscheinlichkeit der einzelnen Qualitätsfunktionen sei durch eine beliebige Wahrscheinlichkeitsverteilung $p(u)$ festgelegt. Weiterhin wird eine qualitative Bewertungsfunktion C eingeführt, die Eins ist, wenn nach der Qualitätsfunktion u^* Approximation A besser ist als B :

$$C = \begin{cases} 1 & \text{für } u^*(A) > u^*(B) \\ 0.5 & \text{für } u^*(A) = u^*(B) \\ 0 & \text{für } u^*(A) < u^*(B) \end{cases} \quad (10.58)$$

Eine Integration der Bewertungsfunktion C über die berücksichtigte Qualitätsfunktionen aus U liefert die Beurteilungsgröße R_1 :

$$R_1(A, B, U, p) = \int_{u \in U} C(A, B, u) p(u) du \quad (10.59)$$

Die Approximation A wird als besser gegenüber B bewertet, wenn $R_1(A, B, U, p) > 0.5$ ist und somit die Qualitätsfunktionen aus U häufiger Approximation A bevorzugen. Vereinfacht kann gesagt werden: „Wenn schon nicht bekannt ist, welche Präferenzen der Entscheidungsträger hat, nehme die Approximation, welche bei den meisten denkbaren Präferenzen am häufigsten besser abschneidet“. Für den R_1 -Indikator gilt allgemein:

$$R_1(A, B, U, p) = 1 - R_1(B, A, U, p) \quad (10.60)$$

Zur Veranschaulichung seien zwei Approximationen ($A, B = R$) und der optimale, aber nicht erreichbare Zielpunkt O gegeben.

$$\begin{aligned} A &= -\{[3, 10], [5, 7], [9, 7]\} \\ B = R &= -\{[2, 9], [5, 6], [10, 6]\} \\ O &= -[10, 10] \end{aligned}$$

Unterschiedliche Prioritäten zwischen den Zielgrößen werden im einfachsten Fall durch Gewichte $\lambda_i \in [0, 1]$ mit $\sum \lambda_i = 1$ berücksichtigt. Im zweidimensionalen Fall werden alle Gewichte über $\Lambda(t) = [t, t-1]$, $t \in [0, 1]$ definiert. Im vorliegenden Fall ergibt sich für die Gruppe der Chebychev-Qualitätsfunktionen folgende Form:

$$U(t) = \{u_\infty(\mathbf{z}, \mathbf{z}^*, \Lambda(t))\} = -\max\{|t| - 10 - z_1|, (1-t)| - 10 - z_2|\}, \quad t \in [0, 1] \quad (10.61)$$

Bei den gegebenen Daten ergeben sich für die Pareto-Grenze A und $t = 0.25$ folgende Qualitätswerte:

$$U(A, 0.25) = \begin{bmatrix} -\max\{0.25| -10 + 3|; 0.75| -10 + 10|\} \\ -\max\{0.25| -10 + 5|; 0.75| -10 + 7|\} \\ -\max\{0.25| -10 + 9|; 0.75| -10 + 7|\} \end{bmatrix} = \begin{bmatrix} -1.75 \\ -2.25 \\ -2.25 \end{bmatrix} \quad (10.62)$$

$$\Rightarrow u^*(A, 0.25) = \max\{U(A, 0.25)\} = -1.75 \quad (10.63)$$

Abbildung 10.35 zeigt, dass die Qualitätsfunktion für B lediglich bei $t \geq 0.8$ größer ist als für A , was bedeutet, dass für den überwiegenden Teil der angenommenen Qualitätsfunktionen Approximation A besser abschneidet. Die Integration der Bewertungsfunktion C bei konstantem $p = 1$ führt zu $R_1 \approx 0.8 > 0.5$, was zur Bevorzugung der Grenze A führt.

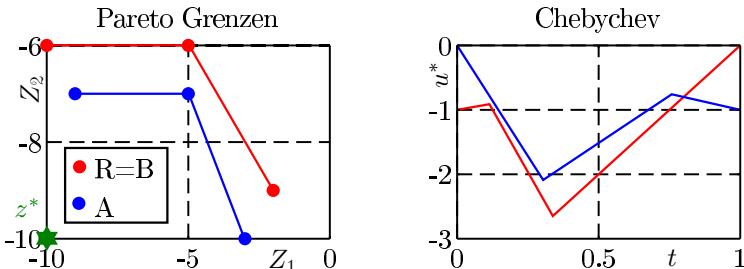


Abb. 10.35 Beispiel für die Berechnung des R_1 -Indikators (Chebychev)

HANSEN [27] zeigt, dass beim Vergleich von mehr als zwei Approximationen mit Hilfe der Bewertungsgröße R_1 in einigen Fällen keine eindeutig beste Approximation bestimmbar ist. So kann beispielsweise bei drei Approximationen A, B und C die Approximation B besser als A , C besser als B und A besser als C sein. Abhilfe schafft hier die Einführung einer fixen Referenz R , zu der alle Approximationen verglichen

werden. Dies führt zur erweiterten Bewertungsgröße R_{1R} :

$$R_{1R}(A, U, p) = R_1(A, R, U, p) \quad (10.64)$$

Zu beachten ist, dass das Kriterium R_{1R} noch nicht mal kompatibel mit der „kompletten Outperformance“ Beziehung O_c ist. Wenn zum Beispiel $A O_c B$ und $R O_c A$ ist, dann folgt daraus, dass beide Kriterien Null werden $R_{1R}(A, U, p) = R_{1R}(B, U, p) = 0$ und somit keine Unterscheidung möglich ist. Die verwendete Referenz R des R_1 Kriteriums muss immer in einem sinnvollen, erreichbaren Rahmen gewählt werden [27].

Zur Verbesserung des R_1 Kriteriums wird die Bewertungsgröße R_2 eingeführt, welche den zu erwartenden besten Punkt im Zielraum stärker berücksichtigt [27].

$$R_2(A, B, U, p) = E(u^*(A)) - E(u^*(B)) \quad (10.65)$$

$$= \int_{u \in U} u^*(A)p(u)du - \int_{u \in U} u^*(B)p(u)du \quad (10.66)$$

$$= \int_{u \in U} u^*(A) - u^*(B)p(u)du \quad (10.67)$$

Die in diesem Kriterium verwendeten Mittelwerte der Qualitätsfunktionen ergeben sich zu $u^*(A) \approx -1.174 > u^*(B) \approx -1.387$, was zum Kennwert $R_2 = 0.213$ führt und somit die Approximation A bevorzugt. Mit der Kenngröße R_2 ist im Gegensatz zu R_1 ein eindeutiges Ranking verschiedener Approximationen der Pareto-Grenze möglich. Für R_2 gilt immer die Beziehung:

$$R_2(A, B, U, P) = -R_2(B, A, U, P) \quad (10.68)$$

Identisch zu R_1 kann auch für R_2 ein Referenz-Qualitätskriterium erstellt werden. Dazu wird lediglich die zweite Approximation durch eine konstante Referenz R ersetzt.

In einigen Anwendungsfällen ist ein Relativwert aussagekräftiger als eine absolute Differenz, so dass alternativ auch die dritte Variante des Qualitätskriterium (R_3) eingesetzt werden kann [27]. Auch dieses Kriterium kann durch Einführung einer konstanten Referenz R zu R_{3R} erweitert werden.

$$R_3(A, B, U, p) = E\left(\frac{u^*(A) - u^*(B)}{u^*(B)}\right) = \int_{u \in U} \frac{u^*(A) - u^*(B)}{u^*(B)} p(u)du \quad (10.69)$$

Die Qualitätskriterien R_1 , R_2 und R_3 sind nicht unabhängig von der Skalierung der einzelnen Zielgrößen, so dass eine Normierung der Zielgrößen sinnvoll ist. Dazu wird neben dem bereits eingeführten, idealen Punkt O ein zweiter, schlechterer Punkt S eingeführt. Die Normierung erfolgt dann mittels der Spannbreite der beiden Punkte in jeder Faktordimension separat:

$$|z_j^* - z_j| \rightarrow \left| \frac{z_j^* - z_j}{o_j - s_j} \right| \quad (10.70)$$

Sollen mehrere Zielgrößen n_z mit gleicher Stufenanzahl n_s und der Nebenbedingung $\sum \lambda_i = 1$ gewichtet werden, so bestehen $\binom{n_s+n_z-2}{n_z-1}$ mögliche Gewichtungskombinationen. Bei zum Beispiel $n_z = 3$ und $n_s = 4$ würden sich $\binom{5}{2} = \frac{5!}{2!(5-2)!} = 10$ Elemente für Λ ergeben.

$$\lambda_1 \in \left[0, \frac{1}{3}, \frac{2}{3}, 1\right] \Rightarrow \Lambda = \begin{bmatrix} 0 & 0 & 1 \\ \frac{1}{3} & 0 & \frac{2}{3} \\ \frac{2}{3} & 0 & \frac{1}{3} \\ \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{2}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{2}{3} & \frac{1}{3} \\ 0 & \frac{2}{3} & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad (10.71)$$

Ein typischer idealisierter Verlauf einer R -Metrik während des Optimierungsverlaufs ist in Abbildung 10.36 gezeigt. Wenn die R -Metrik einen nahezu konstanten Wert erreicht hat, kann davon ausgegangen werden, dass ein stabiles Ergebnis für die Pareto-Grenze vorliegt, so dass es sinnvoll ist, den Optimierungsvorgang zu beenden.

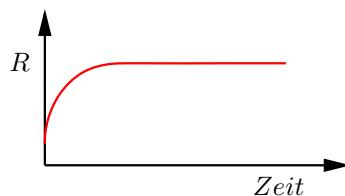


Abb. 10.36 R -Metrik im Verlauf einer Optimierung

10.7.2 Hypervolumen

Neben den R -Indikatoren aus Kapitel 10.7.1 sind in der Literatur häufig die S -, C - und D -Metriken von ZITZLER zu finden [56]. Die S -Metrik misst den Anteil des Zielraumes, der durch eine Pareto-Grenze dominiert wird. Abbildung 10.37 zeigt beispielhaft das Hypervolumen einer zweidimensionalen Minimierungsaufgabe [56]. Das Ergebnis der S -Metrik hängt dabei von der Wahl des optimalen Referenzpunktes ab (hier im Koordinatenursprung) und kann nach Veldhuizen [47] bei

konvexen Pareto-Grenzen zu Fehlinterpretationen führen. Mittels der S -Metrik ist es weiterhin nicht möglich festzustellen, ob eine Approximation A eine andere B komplett dominiert.

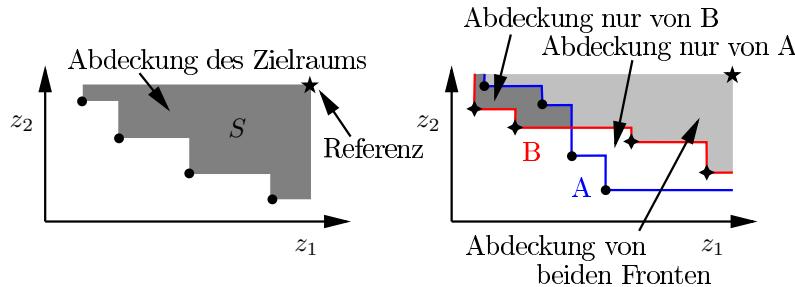


Abb. 10.37 Hypervolumen

Zum direkten Vergleich zweier Approximationen A und B wurde daher die C -Metrik eingeführt, welche die Anzahl der Punkte aus B , welche durch A schwach dominiert werden, in Relation zur Gesamtzahl der Punkte in B setzt.

$$C = (A, B) = \frac{|\{b \in B | \exists a \in A : a \preceq b\}|}{|B|} \quad (10.72)$$

Ein Wert von $C = 1$ bedeutet, dass alle Elemente aus B von A schwach dominiert werden und der Wert $C = 0$ bedeutet, dass kein Element aus B von A schwach dominiert wird. Es gilt nicht zwangsläufig, dass $C(A, B)$ gleich $1 - C(A, B)$ ist. Auch die C -Metrik führt in einigen Fällen nicht zu einem sinnvollen beziehungsweise eindeutigen Vergleichen zweier Approximationen, so dass zur Verbesserung des Kriteriums die D -Metrik eingeführt wurde. Die D -Metrik bestimmt die Differenz der Überdeckung zweier Approximationen.

$$D(A, B) = S(A + B) - S(B) \quad (10.73)$$

D berechnet den schwach dominierten Zielraum von A , welcher von B nicht schwach dominiert wird. Ein Beispiel ist in Abbildung 10.37 zu finden.

Im zweidimensionalen Raum ist die Berechnung der S -Metrik noch übersichtlich und einfach durchzuführen. In höheren Dimensionen wird die Berechnung jedoch kompliziert und aufwendig. Verschiedene Arbeiten befassen sich daher mit optimierten Berechnungsmöglichkeiten [21, 7, 2, 3]. Bei einem abgegrenzten Zielraum, bei dem ein idealer Punkt O und ein schlechtester Punkt S gegeben sind (siehe Kapitel 10.7.1), wird der Zielraum zur Approximation der S -Metrik mit zufälligen (Monte Carlo) Punkten M gefüllt (Abbildung 10.38). Anschließend wird das Verhältnis der Punkte die von Approximation A dominierten werden zur Gesamtzahl der Punkte gebildet.

$$\hat{S} = \frac{|\{m \in M | \exists a \in A : a \preceq m\}|}{|M|} \quad (10.74)$$

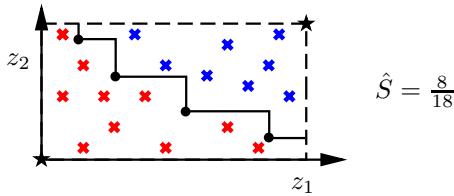


Abb. 10.38 Approximation der S-Metrik

10.8 Zusammenfassung

Typische Optimierungsaufgaben technischer Systeme weisen mehrere Zielgrößen beziehungsweise Qualitätsmerkmale (z_1, \dots, z_{n_z}) auf, die gleichzeitig verbessert werden sollen, sich jedoch teilweise widersprechen. Das optimierte System wird daher ein Kompromiss zwischen allen betrachteten Zielgrößen sein.

Klassische Optimierungsalgorithmen, wie zum Beispiel das Gradientenverfahren, sind lediglich in der Lage eine Zielgröße zu berücksichtigen, so dass durch eine mathematische Kombination der Zielgrößen eine übergeordnete Zielgröße $z^* = f(z_1, \dots, z_{n_z})$ erzeugt wird, welche anstelle der einzelnen Zielgrößen optimiert wird. Der Kompromiss zwischen den einzelnen Zielgrößen wird dabei bereits vor der eigentlichen Optimierung durch die gewählte Kombination der Zielgrößen zur übergeordneten Zielgröße festgelegt. Das Ergebnis der Optimierung ist genau die eine Lösung, welche die übergeordnete Zielgröße bei gegebenem Startwert und Optimierungsalgorithmus optimiert.

In der Praxis ist jedoch vor der Optimierung meistens nicht bekannt, wie die einzelnen Zielgrößen am geschicktesten kombiniert werden und welche Kompromisse zwischen den Zielgrößen eingegangen werden sollen oder können. Sinnvoll ist deshalb eine gleichzeitige Betrachtung jeder einzelnen Zielgröße und Ermittlung von 'optimalen' Lösungen bei denen eine beliebige Zielgröße nur verbessert werden kann, wenn mindestens eine der anderen Zielgrößen verschlechtert wird. Diese Lösungen stellen die Menge aller Kompromisslösungen dar, welche bei separater Betrachtung der Zielgrößen gefunden werden kann. Die Lösungen werden *Pareto-optimal* genannt und bilden die *Pareto-Grenze*. Im Anschluss an den Optimierungsprozess ist es dem Anwender möglich, eine sinnvolle Lösung zu wählen, da er *alle* Kompromisse zwischen den Zielgrößen plus den zugehörigen Faktoreinstellungen (zum Beispiel Fertigungsaufwand oder Kosten) kennt.

Zur Bestimmung der Pareto-optimalen Lösungen beziehungsweise der Pareto-Grenze haben sich in vielen praktischen Anwendungen genetische Optimierungs-

algorithmen für mehrere Zielgrößen bewährt. Sie besitzen nur eine geringe Wahrscheinlichkeit in lokale Optima zu verharren, können mit komplexen, nicht-linearen und nicht-stetigen Funktionszusammenhängen arbeiten, sind flexibel an unterschiedlichste Problemstellungen anpassbar und das Auftreten nicht definierter Faktorbereiche ist unproblematisch. In der Praxis zeigt sich, dass bereits die heute vorhandenen Algorithmen für fast alle Anwendungen ausreichend sind und hauptsächlich Unterschiede in der Konvergenzgeschwindigkeit auftreten.

Da im Gegensatz zu den klassischen Verfahren mit nur einer Zielgröße und jeweils einer aktuellen, optimalen Lösung gleichzeitig mehrere Zielgrößen und dadurch eine mehrdimensionale Pareto-Grenze mit mehreren Lösungen optimiert wird, sind häufig deutlich mehr Berechnungen durchzuführen. Daher ist es sinnvoll komplexe Simulationsmodelle mit langen Rechenzeiten durch Metamodelle (Kapitel 9) zu ersetzen.

Weiterentwicklungen im Bereich der multidimensionalen Optimierungsverfahren gliedern sich in zwei Bereiche. Einerseits die Entwicklung allgemeiner Algorithmen, die robust bei einer Vielzahl unterschiedlicher Aufgaben einsetzbar sind und andererseits hochspezialisierte Algorithmen, die für ein einziges oder wenige Optimierungsaufgaben die beste Performance liefern.

Literaturverzeichnis

1. Alvarez-Benitez, J.E., Everson, R.M., Fieldsend, J.E.: *A MOPSO algorithm based exclusively on pareto dominance concepts*. In: Evolutionary Multi-Criterion Optimization, pp. 459–473. Springer (2005) 348, 349, 350, 351
2. Bader, J., Zitzler, E.: *HypE: An Algorithm for Fast Hypervolume-Based Many-Objective Optimization*. TIK Report 286, Computer Engineering and Networks Laboratory (TIK), ETH Zürich (2008) 375
3. Bader, J., Zitzler, E.: *A Hypervolume-Based Optimizer for High-Dimensional Objective Spaces*. In: Conference on Multiple Objective and Goal Programming (MOPGP 2008), Lecture Notes in Economics and Mathematical Systems. Springer (2009) 375
4. Ballester, P., Carter, J.: *Real-Parameter Genetic Algorithms for Finding Multiple Optimal Solutions in Multi-modal Optimization*. In: Proceedings of the 2003 Genetic and Evolutionary Computation Conference (GECCO-03), pp. 706–717. Springer (2003) 355
5. Ballester, P., Carter, J.: *An Effective Real-Parameter Genetic Algorithm with Parent Centric Normal Crossover for Multimodal Optimisation*. In: Proceedings of the Adaptive Computing in Design and Manufacture, pp. 901–913. Springer (2004) 355
6. Bingham, D., Surjanovic, S.: *Optimization Test Problems* (2015). URL <http://www.sfu.ca/~ssurjano/optimization.html>. (abgerufen 11/2016) 339
7. Bowen, T.J.: *HASO: Hypervolume approximation based on slicing objectives*. Ph.D. thesis, School of Computer Science and Software Engineering, The University of Western Australia (2010) 375
8. Cheng, M.Y., Prayogo, D.: *Symbiotic Organisms Search: A new metaheuristic optimization algorithm*. Computers and Structures **139**, pp. 98–112 (2014) 347
9. Chittka, L., Thomson, J., Waser, N.: *Flower constancy, insect psychology, and plant evolution*. Naturwissenschaften **86**, pp. 361–177 (1999) 344
10. Corne, D., Knowles, J., Oates, M.: *The Pareto envelope-based selection algorithm for multiobjective optimization*. In: Proceeding of the sixth International Conference on Parallel Problem Solving from Nature VI (PPSN VI), pp. 839–848 (2000) 358

11. Czyzak, P., Jaszkiewicz, A.: *Pareto simulated annealing – a metaheuristic technique for multiple-objective combinatorial optimization*. Journal of Multi-Criteria Decision Analysis **7**(1), pp. 34–47 (1998) 368
12. Deb, K.: *A Fast and Elitist Multi-Objective Genetic Algorithm: NSGA-II*. Tech. rep., Kanpur Genetic Algorithms Laboratory (2000). URL <http://www.iitk.ac.in/kangal/papers/tech2000001.ps.gz>. (abgerufen 11/2016) 350, 358, 361, 368
13. Deb, K.: *Software Developed at KanGAL* (2014). URL <http://www.iitk.ac.in/kangal/codes.shtml>. (abgerufen 11/2016) 350, 358
14. Deb, K., Agrawal, R.: *Simulated binary crossover for continuous search space*. Complex Systems **9**, pp. 115–148 (1995) 355
15. Deb, K., Agrawal, S., Pratap, A., Meyarivan, T.: *A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II*. Lecture Notes in Computer Science **2000**, pp. 849–858 (2000) 350
16. Deb, K., Anand, A., Joshi, D.: *A computationally efficient evolutionary algorithm for real-parameter optimization*. Evolutionary Computation **10**(4), pp. 371–395 (2002) 350, 355
17. Deb, K., Kumar, A.: *Real-coded Genetic Algorithms with Simulated Binary Crossover: Studies on Multimodal and Multiobjective Problems*. Complex Systems **9**, pp. 115–148 (1995) 355
18. Deb, K., Mohan, M., Mishra, S.: *A Fast Multi-Objective Evolutionary Algorithm for Finding Well-Spread Pareto-Optimal Solutions*. Tech. rep., Kanpur Genetic Algorithms Laboratory (2003). URL <http://www.iitk.ac.in/kangal/papers/k2003002.pdf>. (abgerufen 11/2016) 358, 364
19. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: *A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II*. IEEE Transactions on Evolutionary Computation **6**, pp. 182–197 (2002) 358, 359
20. Eshelman, L.J., Schaffer, J.D.: *Real-coded genetic algorithms and interval-schemata*. In: D.L. Whitley (ed.) Foundation of Genetic Algorithms 2, pp. 187–202. Morgan Kaufmann., San Mateo, CA (1993) 355
21. Fleischer, M.: *The Measure of Pareto Optima Applications to Multi-objective Metaheuristics*. In: C. Fonseca, P. Fleming, E. Zitzler, L. Thiele, K. Deb (eds.) Evolutionary Multi-Criterion Optimization, *Lecture Notes in Computer Science*, vol. 2632, pp. 519–533. Springer Berlin Heidelberg (2003) 375
22. Fonseca, C.M., Fleming, P.J.: *Genetic Algorithm for multiobjective optimization: Formulation, Discussion and Generalization*. In: Proceedings of the 5th International Conference on Genetic Algorithms, pp. 416–423 (1993) 358
23. Fonseca, C.M., Fleming, P.J.: *Muliobjective Optimization and multiple Constraint Handling with Evolutionary Algorithms Part I: A unified Formulation*. In: IEEE Transactions on Systems, Man and Cybernetics, vol. 28, pp. 26–37 (1998) 328, 330, 358
24. Fonseca, C.M., Fleming, P.J.: *Muliobjective optimization and multiple constraint handling with evolutionary algorithms. II. Application example*. IEEE Transactions on Systems **28**(1), pp. 38–47 (1998) 328, 330
25. Gigerenzer, G., Todd, P.M., Group, A.R.G.A.R.: *Simple heuristics that make us smart*. Oxford University Press (2000) 325
26. Grosan, C., Oltean, M., Dumitrescu, D.: *Performance metrics for multiobjective optimization evolutionary algorithms*. In: Proceedings of the Conference on Applied and Industrial Mathematics (CAIM '03) (2003) 367
27. Hansen, M.P., Jaszkiewicz, A.: *Evaluating the quality of approximations to the non-dominated set*. IMM, Department of Mathematical Modelling, Technical University of Denmark (1998) 368, 369, 372, 373
28. Horn, J., Nafpliotis, N.: *Multiobjective optimization using the niched pareto genetic algorithm*. Tech. rep., University of Illinois USA (1993) 358
29. Jaszkiewicz, A.: *Multiple objective metaheuristic algorithms for combinatorical optimization*. Ph.D. thesis, Poznan University of Technology (2001) 370
30. Kennedy, J., Eberhart, R.: *Particle swarm optimization*. In: IEEE International Conference on Neural Networks, vol. 4, pp. 1942–1948 (1995) 336

31. Knowles, J.D.: *Local-Search and Hybrid Evolutionary Algorithms for Pareto Optimization*. Ph.D. thesis, The University of Reading, Department of Computer Science (2002) 368
32. Knowles, J.D., Corne, D.W.: *Approximating the non-dominated front using the Pareto archived evolution strategy*. Evolutionary Computation **8**, pp. 149–172 (2000) 358
33. Kursawe, F.: *A Variant of Evolution Strategies for Vector Optimization*. In: Proceedings of the 1st Workshop on Parallel Problem Solving from Nature, pp. 193–197. Springer-Verlag, London (1990) 361
34. Lin, C.Y., Hajela, P.: *Genetic search strategies in multicriterion optimal design*. Structural Optimization **4**, pp. 99–107 (1992) 358
35. Lozano, M., Herrera, F., Krasnogor, N., Molina, D.: *Real-coded memetic algorithms with crossover hill-climbing*. Evolutionary Computation Journal **12**, pp. 273–302 (2004) 355
36. Lygoe, R.J.: *Complexity Reduction in High-Dimensional Multi-Objective Optimisation*. Ph.D. thesis, University of Sheffield (2010) 328, 330
37. Miettinen, K.: *Nonlinear Multiobjective Optimization, International Series in Operations Research and Management Science*, vol. 12. Kluwer Academic Publishers, Dordrecht (1999) 335
38. Mondrzyk, D.: *Entwurf und Analyse metaheuristischer Verfahren zum Einsatz im Bereich der Motoroptimierung*. Master's thesis, RWTH Aachen (2013) 338, 339
39. Ono, I., Kobayashi, S.: *A Real Coded Genetic Algorithm for Function Optimization Using Unimodal Normal Distributed Crossover*. In: Proceedings of the 7th International Conference on Genetic Algorithms, pp. 246–253 (1997) 355
40. Padhye, N.: *Comparison of Archiving Methods in Multi-objective particle Swarm Optimization (MOPSO): Empirical Study*. In: Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation, GECCO '09, pp. 1755–1756. ACM (2009) 348
41. Pavlyukevich, I.: *Lévy flights, non-local search and simulated annealing*. Journal of Computational Physics **226**(2), pp. 1830 – 1844 (2007) 345
42. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press (2007) 199, 203, 220, 254, 265, 276, 277, 281, 285, 307, 334, 381, 398, 400
43. Raghuwanshi, M., Kakde, O.: *Survey on Multiobjective Evolutionary and Real Coded Genetic Algorithms*. Complexity International **11**, pp. 150–161 (2005) 358
44. Schaffer, J.: *Multi objective optimization with vector evaluated genetic algorithms*. Ph.D. thesis, Vanderbilt University, Nashville, USA (1984) 358
45. Schott, J.R.: *Fault tolerant design using single and multicriteria genetic algorithm optimization*. Ph.D. thesis, Massachusetts Institute of Technology, Dept. of Aeronautics and Astronautics (1995) 368
46. Takahashi, O., Kobayashi, S.: *An adaptive neighboring search using crossover-like mutation for multi modal function optimization*. In: IEEE International Conference on Systems, Man and Cybernetics, vol. 1, pp. 261–267 (2001) 355
47. van Veldhuizen, D.A.: *Multiobjective Evolutionary Algorithms: Classifications, Analyses and New Innovations*. Ph.D. thesis, Graduate School of Engineering of the Air Force Institute of Technology (1999) 368, 374
48. Voigt, H., Mühlenbein, H., Cvetkovic, D.: *Fuzzy Recombination for the Breeder Genetic Algorithm*. In: Proc. of the Sixth International Conference on Genetic Algorithms, pp. 104–111 (1995). URL <http://www.muehlenbein.org/fuzzy95.pdf>. (abgerufen 11/2016) 355
49. Wikipedia: *Test functions for optimization*. URL http://en.wikipedia.org/w/index.php?title=Test_functions_for_optimization. (abgerufen 11/2016) 339
50. Xie, J., Zhou, Y., Chen, H.: *A Novel Bat Algorithm Based on Differential Operator and Lévy Flights Trajectory*. Computational Intelligence and Neuroscience **2013**, p. 13 (2013) 342
51. Yang, X.S.: *Firefly Algorithms for Multimodal Optimization*. In: O. Watanabe, T. Zeugmann (eds.) Stochastic Algorithms: Foundations and Applications, Lecture Notes in Computer Science, vol. 5792, pp. 169–178. Springer Berlin Heidelberg (2009) 340

52. Yang, X.S.: *Nature-Inspired Metaheuristic Algorithms*. Evolutionary Computation (2010) 341, 342
53. Yang, X.S.: *A New Metaheuristic Bat-Inspired Algorithm*. Nature Inspired Cooperative Strategies for Optimization (NICSO 2010); Studies in Computational Intelligence **284**, pp. 65–74 (2010) 342
54. Yang, X.S.: *Flower pollination algorithm for global optimization*. Unconventional Computation and Natural Computation 2012, Lecture Notes in Computer Science **7445**, pp. 240–249 (2012) 344, 345
55. Yang, X.S.: *Nature-Inspired Optimization Algorithms*. Elsevier Ltd, Oxford (2014) 341, 342, 344
56. Zitzler, E.: *Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications*. Ph.D. thesis, Institut für Technische Informatik und Kommunikationsnetze Computer Engineering and Networks Laboratory (1999) 358, 368, 374
57. Zitzler, E., Deb, K., Thiele, L.: *Comparison of Multiobjective Evolutionary Algorithms: Empirical Results*. Evolutionary Computation **8**, pp. 173–195 (2000) 367
58. Zitzler, E., Knowles, J., Thiele, L.: *Quality Assessment of Pareto Set Approximations*. In: J. Branke, K. Deb, K. Miettinen, R. Slowinski (eds.) Multiobjective Optimization, Lecture Notes in Computer Science, vol. 5252, pp. 373–404. Springer Berlin, Heidelberg (2008) 369
59. Zitzler, E., Laumanns, M., Thiele, L.: *SPEA2: Improving the Strength Pareto Evolutionary Algorithm*. TIK Report 103, Computer Engineering and Networks Laboratory (TIK), ETH Zürich (2001). URL <http://www.tik.ee.ethz.ch/sop/publicationListFiles/zlt2001a.pdf>. (abgerufen 11/2016) 358
60. Zitzler, E., Thiele, L.: *An evolutionary algorithm for multiobjective optimization: the strength Pareto approach*. Tech. rep., Computer Engineering and Networks Laboratory (TIK), ETH Zürich (1998) 358
61. Zitzler, E., Thiele, L.: *Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach*. IEEE Transactions on Evolutionary Computation **3**, pp. 257–271 (1999) 358
62. Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C.M., da Fonseca, V.G.: *Performance Assessment of Multiobjective Optimizers: An Analysis and Review*. Trans. Evol. Comp **7**(2), pp. 117–132 (2003) 327

Kapitel 11

Korrelationsanalyse

Die Korrelation beschreibt den mathematischen Zusammenhang zwischen zwei statistisch verteilten Variablen. Eine kausale Beziehung zwischen den beiden Variablen muss in Wirklichkeit jedoch nicht bestehen.

11.1 Pearson Korrelation

Die wohl am häufigsten eingesetzte Kennzahl zur Beschreibung der Abhängigkeit zweier Variablen x und y ist die lineare Korrelation nach *Pearson*. Bei dieser Kennzahl wird ein linearer Zusammenhang sowie eine Normalverteilung beider Variablen vorausgesetzt. Die Pearson-Korrelation r wird durch Gleichung 11.1 berechnet [12, 15]:

$$r = \frac{\sum_{i=1}^{n_r} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n_r} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n_r} (y_i - \bar{y})^2}} \quad (11.1)$$

Mittels der empirischen Kovarianz K_{xy} und der Stichprobenvarianz $\sigma_{x,y}^2$ der Variablen x und y ¹ kann die Korrelation auch nach Gleichung 11.2 berechnet werden.

¹ mit dem Vorfaktor $\left(\frac{1}{n_r}\right)$ als unkorrigierte oder mit $\left(\frac{1}{n_r-1}\right)$ als korrigierte Varianz

$$\begin{aligned}
 r &= \frac{K_{xy}}{\sigma_x \sigma_y} \\
 K_{xy} &= \frac{1}{n_r} \sum_{i=1}^{n_r} (x_i - \bar{x})(y_i - \bar{y}) \\
 \sigma_x^2 &= \frac{1}{n_r} \sum_{i=1}^{n_r} (x_i - \bar{x})^2 \text{ und} \\
 \sigma_y^2 &= \frac{1}{n_r} \sum_{i=1}^{n_r} (y_i - \bar{y})^2
 \end{aligned} \tag{11.2}$$

Der Wert der Pearson-Korrelation liegt immer im Bereich von $r \in [-1, 1]$. Ein Korrelationswert von $|r| = 1$ steht dabei für eine perfekte lineare Korrelation und ein Wert von $r = 0$ für keine lineare Korrelation zwischen den Variablen. Abbildung 11.1 zeigt für einige einfache Beispiele die Pearson-Korrelation sowie einige alternative Korrelationswerte. Hierbei handelt es sich um sogenannte Rangkorrelationen, die in Kapitel 11.4 beschrieben werden (Spearman und Kendall). In Klammern ist jeweils die Signifikanz (p-Wert) des Korrelationswerts angegeben (Kapitel 11.3). Je kleiner p ist desto sicherer ist eine Korrelation vorhanden. Die Pearson-Korrelation ist weit

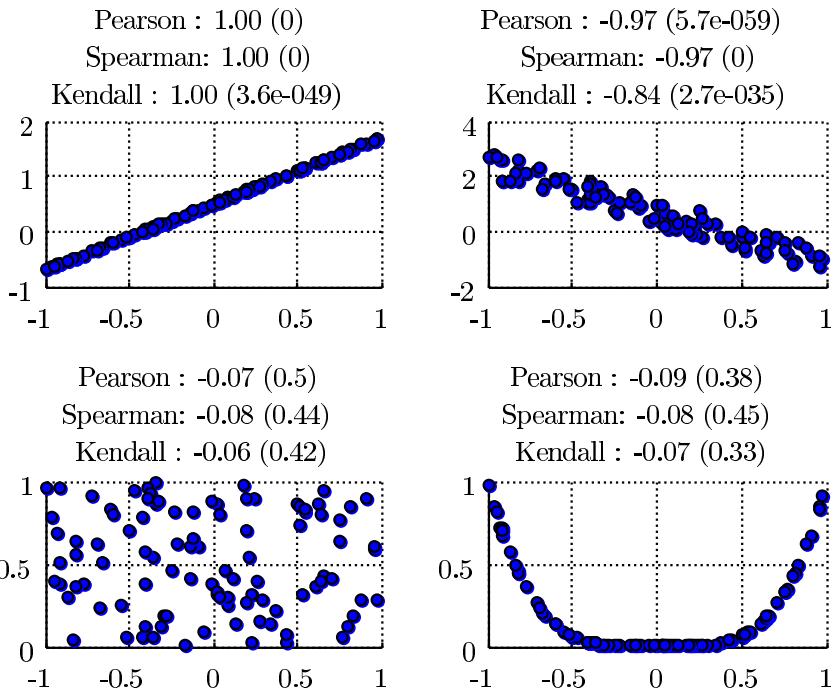


Abb. 11.1 Pearson-, Spearman- und Kendall's τ Korrelation (p-Wert): a) perfekte lineare Korrelation b) perfekte lineare Korrelation mit überlagertem Rauschen c) zufälliger unkorrelierter Zusammenhang d) biquadratischer Zusammenhang $y = x^4$

verbreitet, kann jedoch besonders bei Messausreißern (Abbildung 11.3) und nicht normalverteilten Daten zu falschen Ergebnissen führen [4, 9]. In diesen Fällen sind Methoden der Rangkorrelation (Kapitel 11.4) hilfreich.

11.2 Scheinkorrelation und verdeckte Korrelation

Bei der Analyse von Korrelationen muss auf Schein- und verdeckte Korrelationen geachtet werden, um voreilige Schlüsse aus den Daten zu vermeiden.

Scheinkorrelation

Scheinkorrelationen sind Korrelationen bei denen zwar ein signifikanter r -Wert berechnet wird, jedoch kein kausaler Zusammenhang zwischen den beiden analysierten Variablen besteht. Die eigentliche Korrelation ist in diesen Fällen meist durch eine dritte nicht betrachtete Variable verursacht. Ein Beispiel ist die starke Korrelation zwischen finziellem Reichtum und Anzahl der Kopfhaare bei Männern. Die eigentliche korrelierende Variable zwischen den Variablen ist das Alter der betrachteten Personen. Ältere Männer haben im Durchschnitt mehr Geld gespart als junge Männer und haben mit hoher Wahrscheinlichkeit auch weniger Haare.

Verdeckte Korrelation

Bei verdeckten Korrelationen werden Korrelationen nicht erkannt, da sich die Korrelationen innerhalb von Untergruppen, welche nicht separat analysiert werden, gegenseitig auslöschen. Beispielhaft können wir hier die Korrelation zwischen der Zufriedenheit von Kunden mit einem Mittagessen und der zugeführten Menge an Chili betrachten. Die Testpersonen kommen dabei aus zwei verschiedenen Ländern, wobei in der einen Gruppe die Zufriedenheit mit erhöhter Schärfe des Gerichts ansteigt und bei der zweiten Gruppe im gleichen Maße abnimmt. Werden beide Gruppen separat analysiert, so wird die Korrelation eindeutig festgestellt. Bei gemeinsamer Analyse beider Gruppen wird keine Korrelation festgestellt, da die gegenläufige Tendenzen sich verdecken.

11.3 Signifikanz einer Korrelation

Ob eine gefundene Korrelation wirklich statistisch relevant ist hängt neben dem eigentlich Korrelationswert r ebenfalls von der verwendeten Stichprobengröße n_r

ab. Je kleiner die Stichprobengröße ist, desto größer muss r werden, um sicher zu stellen, dass es sich um eine signifikante Korrelation handelt. Mittels eines Hypothesentests bei Verwendung der *Student t*-Verteilung sowie der Nullhypothese, dass zwischen den Variablen (Faktoren) kein linearer Zusammenhang besteht ($r = 0$) kann das Signifikanzniveau t bestimmt werden:

$$t = r \sqrt{\frac{dF}{1 - r^2}} \text{ mit } dF = n_r - 2 \text{ (Freiheitsgrad, degree of freedom)} \quad (11.3)$$

Als Beispiel sei eine Menge mit $n_r = 5$ Datenpaaren und einem Korrelationswert von $r = 0.99$ gegeben. Das dazugehörige Signifikanzniveau ergibt sich dann zu $t = 0.99 \sqrt{\frac{5-2}{1-0.99^2}} \approx 12.16$. Der Vergleichswert der zweiseitigen t -Quantile für einen Freiheitsgrad von $dF = 3$ und einem Vertrauensbereich von 98% ist $t_{(3,98\%)} \approx 4.54$ [13, 2]. Dieser Wert ist deutlich kleiner als $t = 12.16$, so dass von einem statistisch gesicherten (signifikanten) Zusammenhang ausgegangen werden kann. Der meist verwendete Signifikanzwert p^2 kann mittels der (regulierten) unvollständigen Betafunktion I_x bestimmt werden (*betainc*, in Matlab [10] oder Octave [11]) und würde in unserem Beispiel eine p-Wert von $p = 0.0012$ ergeben.

$$\begin{aligned} p &= I_{\frac{dF}{dF+r^2}} \left(\frac{dF}{2}, \frac{1}{2} \right) \text{ mit} \\ I_x(a, b) &= \frac{\text{Beta}(x, a, b)}{\text{Beta}(1, a, b)} \text{ und} \\ \text{Beta}(x, a, b) &= \int_0^x t^{a-1} (1-t)^{b-1} dt \end{aligned} \quad (11.4)$$

11.3.1 Permutationstest

Neben der Berechnung der Signifikanz mittels eines Hypothesentests wird immer häufiger, gerade bei dem Einsatz von Computern, eine Berechnung durch Permutation der Variablen eingesetzt (Permutationstest). Hierzu werden die Variablen mehrmals zufällig permutiert und dann die Korrelation der permutierten Variablen bestimmt, wodurch es auf einfache Weise möglich ist die Wahrscheinlichkeitsverteilung der Korrelationswerte bei zufälliger Wahl der Variablen zu ermitteln. Die Abschätzung der Signifikanz basiert dann auf der Permutationsanzahl, die einen größeren bzw. kleineren Korrelationswert als die Originalverteilung der Variablen aufweist. Die einseitige p_1 beziehungsweise zweiseitige p_2 Signifikanz wird dann nach Gleichung 11.5 bestimmt.

² Kleine p -Werte stehen für hohe Signifikanz

$$p_1 = \begin{cases} \frac{\#(r_p > r)}{n_p} & \text{wenn } r \geq 0 \\ \frac{\#(r_p < r)}{n_p} & \text{wenn } r < 0 \end{cases} \quad (11.5)$$

$$p_2 = \frac{\#(r_p > |r|)}{n_p} + \frac{\#(r_p < -|r|)}{n_p}$$

In diesen Gleichungen bezeichnet n_p die Anzahl der Permutationen und $\#(r_p > r)$ die Anzahl der Permutationen mit $r_p > r$. Abbildung 11.2 zeigt auf der linken Seite einen annähernd quadratischen Zusammenhang, welcher mit einem zufälligen gleichverteilten Rauschen überlagert ist. Das rechte Bild zeigt das Histogramm der Korrelationswerte aller Permutationen sowie in rot den Korrelationswert $\pm|r|$ des Originaldatensatzes. Die beidseitige Signifikanz entspricht typischerweise dem berechneten Wert des Hypothesentests. Je kleiner der p-Wert ausfällt, desto statistisch abgesicherter ist die Korrelation. Typisch gewählte Signifikanzgrenzen liegen bei $\alpha = 0.01$ und $\alpha = 0.05$.

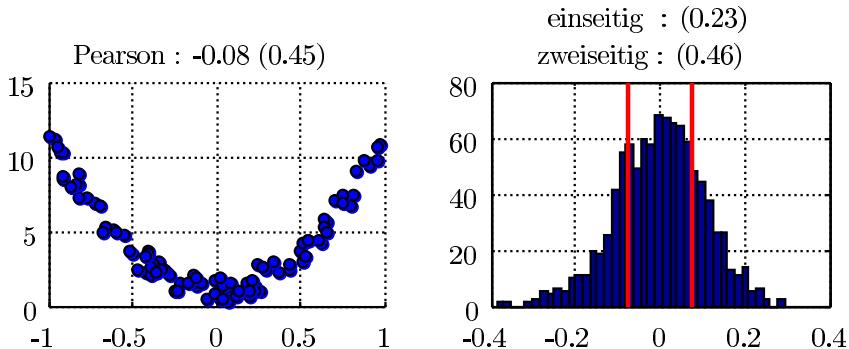


Abb. 11.2 Permutationstest zur Bestimmung der Signifikanz

11.4 Rangkorrelation

Die Rangkorrelation beschreibt die Korrelation zwischen zwei Variablen, wobei keine Annahme über die statistische Verteilung der Variablen vorausgesetzt wird und die Korrelation nicht linear sondern lediglich monoton steigend oder fallend angenommen wird. Die Methoden der Rangkorrelation sind besonders robust gegenüber Ausreißer, verlieren jedoch auch einige Informationen der Originaldaten, wie zum Beispiel die wahre Verteilungsfunktion.

Abbildung 11.3 zeigt die bereits aus Kapitel 11.1 bekannten Korrelationsbeispiele, jedoch mit jeweils einem Ausreißer, der den Wert $y = 100$ aufweist. Die *Pearson* Korrelation (siehe Kapitel 11.1) kann die vorhandenen Zusammenhänge in den oberen beiden Beispielen nicht mehr erkennen, wobei die beiden Rangkorrelationen

Spearman und *Kendall's Tau* dieses ohne Probleme richtig interpretieren.

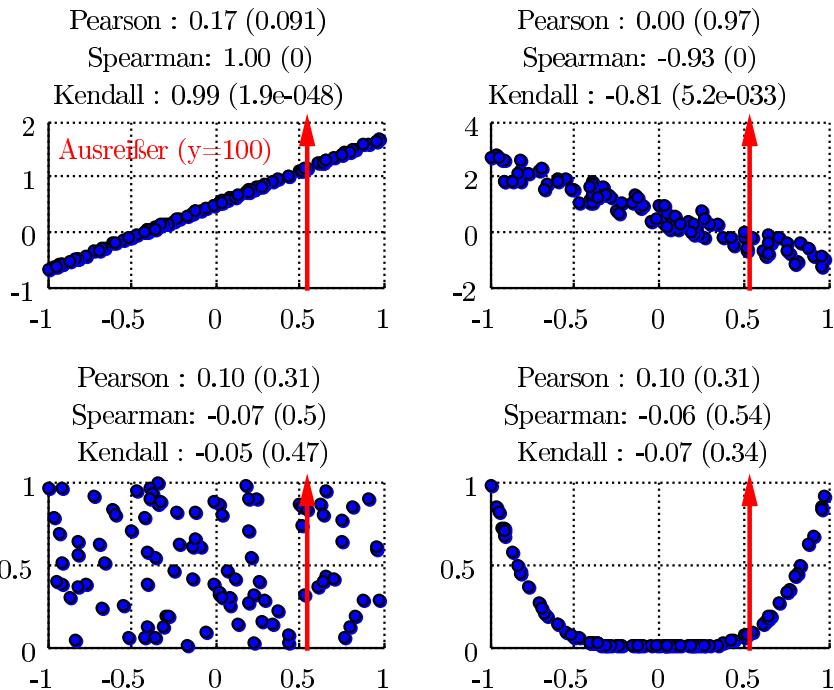


Abb. 11.3 Pearson-, Spearman- und Kendall's τ Korrelation mit Ausreißer [$y=100$] (p-Wert): a) perfekte lineare Korrelation b) perfekte lineare Korrelation mit überlagertem Rauschen c) zufälliger unkorrelierter Zusammenhang d) biquadratischer Zusammenhang $y = x^4$

Zur Bestimmung der Rangkorrelation werden alle Variablenwerte x_i aufsteigend sortiert und durch ihren Rang (bzw. ihre Position) im Verhältnis zu allen anderen Variablen $x_{j \neq i}$ ersetzt. Sind keine doppelten x_i Werte vorhanden, so entstehen dadurch n_r Ränge, wobei diese den Zahlen 1 bis n_r entsprechen. Falls einige x_i Werte identisch sind, wird Ihnen der Mittelwert der Ränge zugeordnet, welche sie bekommen hätten, wenn alle x_i Werte leicht unterschiedlich gewesen wären. Die Zusammenfassung von x_i Werten wird Bindung (*tie*) genannt [16, 8], wobei auch nicht ganzzahlige Ränge entstehen können. Zur Kontrolle wird die Summe aller Ränge berechnet, welche immer $\frac{n_r(n_r+1)}{2}$ entsprechen muss. Im Anschluss wird ebenfalls die zweite Variable y durch die dazugehörigen Ränge ersetzt. Tabelle 11.1 zeigt ein einfaches Beispiel zur Bestimmung der Ränge einer Variablen x .

x	x_{sort}	$Rang_{x,sort}$ ohne Bindung	$Rang_{x,sort}$ mit Bindung	$Rang_x$	x
1.5	1	1	1	2.5	1.5
1.5	1.5	2	2.5	2.5	1.5
4	1.5	3	2.5	5	4
3	3	4	4	4	3
1	4	5	5	1	1
5	5	6	7	7	5
5	5	7	7	7	5
9.5	5	8	7	9	9.5
5	9.5	9	9	7	5

Tabelle 11.1 Beispiel zur Berechnung von Rangwerten mit Bindungen

11.4.1 Spearman

Die Spearman Rangkorrelation ist eine Pearson Korrelation (Kapitel 11.1) bei der die Variablenwerte x_i, y_i durch Ihre Ränge r_{x_i}, r_{y_i} ersetzt wurden.

$$r_s = \frac{\sum_{i=1}^{n_r} (r_{x_i} - \bar{r}_x)(r_{y_i} - \bar{r}_y)}{\sqrt{\sum_{i=1}^{n_r} (r_{x_i} - \bar{r}_x)^2} \sqrt{\sum_{i=1}^{n_r} (r_{y_i} - \bar{r}_y)^2}} \quad (11.6)$$

Im Fall dass keine Bindungen (Wiederholungen der Variablenwerte) auftreten, kann zur Berechnung eine vereinfachte Form auf Basis der Rangunterschiede zwischen r_{x_i} und r_{y_i} verwendet werden ($d_i = r_{x_i} - r_{y_i}$) [16]:

$$r_{s_{\text{ohne Bindung}}} = 1 - \frac{6 \sum_{i=1}^{n_r} d_i^2}{n_r (n_r^2 - 1)} \quad (11.7)$$

Sollten Bindungen vorliegen ist es hingegen möglich die Rangkorrelation r_s mit Gleichung 11.8 zu ermitteln [16].

$$r_{s_{\text{mit Bindung}}} = \frac{2 \cdot \frac{n_r^2 - n_r}{12} - T_x - T_y - \sum_{i=1}^{n_r} d_i^2}{2 \sqrt{\left(\frac{n_r^3 - n_r}{12} - T_x \right) \left(\frac{n_r^3 - n_r}{12} - T_y \right)}} \quad (11.8)$$

Dabei ist T_x (T_y entsprechend) wie in Gleichung 11.9 definiert, wobei $t_{x,k}$ die Anzahl der Beobachtungen der Variable x mit gleichem Rang k darstellt und n_R die Anzahl der unterschiedlichen Ränge von x .

$$T_x = \sum_{k=1}^{n_R} [(t_{x,k}^3 - t_{x,k})/12] \quad (11.9)$$

11.4.2 Kendalls τ

Kendalls τ betrachtet im Gegensatz zu Spearman das Verhältnis der Ränge und nicht die genaue Differenz. In der Literatur finden sich zwei gleichwertige Berechnungsmethoden für τ [16, 14, 1]. Im ersten Fall werden die Wertepaare $xy_i = (x_i, y_i)$ nach aufsteigenden x_i sortiert und anschließend alle möglichen Wertepaarkombinationen (xy_i, xy_j) mit $i < j$ verglichen. Dabei wird das Auftreten der Verhältnisse aus Tabelle 11.2 gezählt.

Tabelle 11.2 Kendalls τ Verhältnisse

C	$x_i < x_j$	$y_i < y_j$
D	$x_i < x_j$	$y_i > y_j$
T_x	$x_i \neq x_j$	$y_i = y_j$
T_y	$x_i = x_j$	$y_i \neq y_j$
T_{xy}	$x_i = x_j$	$y_i = y_j$

Mit diesen Kennzahlen ist Kendalls τ mittels Gleichung 11.10 zu berechnen.

$$\tau = \frac{C - D}{\sqrt{(C + D + T_x)(C + D + T_y)}} \quad (11.10)$$

Sind die Ränge der Variablen x und y bereits bekannt, so kann alternativ die Berechnungsformel aus Gleichung 11.11 verwendet werden.

$$\tau = \frac{C - D}{\sqrt{(n_0 - n_1)(n_0 - n_2)}} \quad (11.11)$$

mit

$$\begin{aligned} n_0 &= \frac{n_r(n_r - 1)}{2} \\ n_1 &= \sum_i^{n_{R_x}} \frac{t_i(t_i - 1)}{2} \\ n_2 &= \sum_j^{n_{R_y}} \frac{u_j(u_j - 1)}{2} \end{aligned}$$

C : Anzahl gleichsinniger Paare

D : Anzahl gegensinniger Paare

t_i : Anzahl Bindungen des i^{ten} Rangs von x

u_j : Anzahl Bindungen des j^{ten} Rangs von y

Die Differenz $[C - D]$ ist ebenfalls mit der vereinfachten Gleichung 11.12 auf Basis der Vorzeichenfunktion (sign) bestimmbar.

$$C - D = \sum_{i,j=1, i < j}^{n_r} sign(x_i - x_j) sign(y_i - y_j) \quad (11.12)$$

Zur Veranschaulichung wird die Korrelation zwischen den beiden folgenden Variablen berechnet, wobei sich ein Kendall $\tau_b = 0.7247$ ergibt.

$$\begin{aligned} x &= [2, 4, 1, 2, 12, 22, 3, 5, 33] \\ y &= [1, 3, 1, 6, 10, 12, 5, 6, 11] \end{aligned} \quad (11.13)$$

$$\begin{aligned} C &= 29 \\ D &= 4 \\ T_x &= 2 \\ T_y &= 1 \\ \frac{T_{xy}}{n_0} &= 0 \\ n_0 &= 36 \\ n_1 &= 1 \\ n_2 &= 2 \end{aligned}$$

11.5 Nichtlineare Korrelation

Nichtlineare beziehungsweise nicht monotone Korrelationen werden durch klassische Verfahren wie Pearson Korrelation (Kapitel 11.1), Spearman Rangkorrelation (Kapitel 11.4.1) oder Kendalls τ (Kapitel 11.4.2) nicht erkannt. CHEN et al. [3] schlagen deshalb eine Methode vor, welche auf Basis lokaler Korrelationen, Aussagen über nichtlineare Korrelationen erstellt. Im ersten Schritt werden dazu die zu untersuchenden Variablen x und y in Ränge r_x und r_y umgewandelt, wodurch gleichverteilte Daten erzeugt werden (Kapitel 11.4.). Im folgenden Schritt werden diese zusätzlich auf den Bereich $r_{x,y}^* \in [0, 1]$ normiert.

$$r_{x_i}^* = \frac{r_{x_i} - \min(r_x)}{\max(r_x) - \min(r_x)} \quad (11.14)$$

Das Korrelationsintegral I , welches den prozentualen Anteil von Datenpunkten beschreibt, welcher kleiner oder gleich einem vorgegebenen Radius r ist, wird durch Gleichung 11.15 bestimmt.

$$\hat{I}(r) = \frac{1}{n_r^2} \sum_{i,j=1}^{n_r} \# \left(\frac{d_{i,j}}{\max(d)} \leq r \right) \quad (11.15)$$

Hierbei beschreibt $d_{i,j}$ den euklidischen Abstand zwischen zwei bivariaten Variablen (r_i^*, r_j^*) und $\sum \#$ steht für die Anzahl der Elemente, bei denen die angegebene Bedingung erfüllt ist.

$$d_{i,j} = \sqrt{\left(r_{x_i}^* - r_{x_j}^*\right)^2 + \left(r_{y_i}^* - r_{y_j}^*\right)^2} \quad (11.16)$$

Der Abstand $d_{i,j}$ wird mit Hilfe des maximal auftretenden Abstands $\max(d)$ ebenfalls auf den Bereich $d_{i,j} \in [0 \dots 1]$ normiert. Das approximierte Korrelationsintegral \hat{I} , wie es beispielhaft in Abbildung 11.4 dargestellt ist, hat die Eigenschaften einer kumulativen Verteilungsfunktion (CDF: *cumulative distribution function*).

Die Ableitung des Korrelationsintegrals \hat{I} liefert die passende Dichtefunktion \hat{D} , welche die Änderung der Anzahl von Beobachtungen (Datenpunkte) in der Nachbarschaft eines bestimmten Radius r beschreibt. Im Folgenden wird diese Funktion als Nachbarschaftsdichte (*neighbor density*) bezeichnet [3].

$$\hat{D} = \frac{\Delta \hat{I}}{\Delta r} \quad (11.17)$$

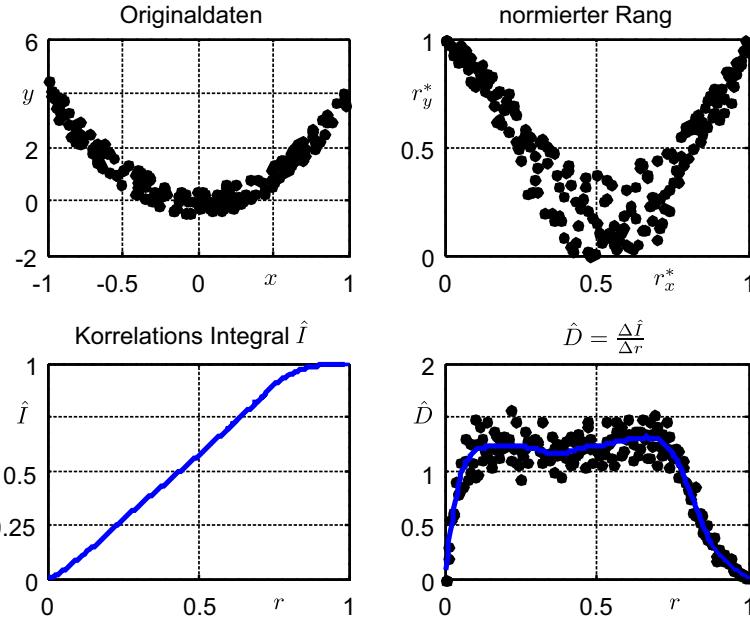


Abb. 11.4 Nichtlineare Korrelation: quadratischer Zusammenhang mit Rauschen, Korrelationsintegral und Dichtefunktion

Die Nachbarschaftsdichte weist folglich die Eigenschaften einer Wahrscheinlichkeitsdichtefunktion (pdf: *probability density function*) auf und wird an vorgegebene Gitterpunkte mit einem konstanten Abstand Δr analysiert $r = [0, \Delta r, 2\Delta r, \dots, 1]$. In vielen Fällen wird $\Delta r = \frac{1}{n_r}$ gewählt. Wie Abbildung 11.4 deutlich zeigt, muss die Nachbarschaftsdichte \hat{D} vor einer weiteren Analyse geglättet werden, wobei ein be-

liebiger Glättungsalgorithmus verwendet werden kann. CHEN et al. schlagen beispielsweise einen erweiterten Whittaker-Algorithmus vor [6, 7]. Die lokale Korrelation $lc(r)$ kann im Anschluss als Abweichung der Nachbarschaftsdichte $\hat{D}(r)$ zu einer Dichte $\hat{D}_0(r)$ von unkorrelierten Daten aufgefasst werden.

$$lc(r) = \hat{D}(r) - \hat{D}_0(r) \quad (11.18)$$

Die Dichte einer unkorrelierten Datenmenge wird durch ein Permutationsexperiment abgeschätzt. Dazu werden beide Variablen x und y mehrmals (n_P -mal) zufällig angeordnet (permutiert). Für alle Permutationen werden dann die dazugehörigen Nachbarschaftsdichten \hat{D}_b mit $b = 1, \dots, n_P$ bestimmt. Der Mittelwert oder Median aller Dichten ergibt anschließend eine Approximation der unkorrelierten Nachbarschaftsdichte \hat{D}_0 der gegebenen Daten.

$$\hat{D}_0(r) \approx \frac{1}{n_P} \sum_{b=1}^{n_P} \hat{D}_b(r) \quad (11.19)$$

Abbildung 11.5 (oben, links) zeigt für das gegebene quadratische Beispiel die Nachbarschaftsdichte $\hat{D}(r)$ (blau) und die approximierte Dichte $\hat{D}_0(r)$ (rot), welche aus n_P Permutationen (schwarz) ermittelt wurde. Bereits hier ist ein deutlicher Unterschied zwischen \hat{D} und \hat{D}_0 erkennbar, was darauf hinweist, dass die Originaldaten eine Korrelation bzw. Abhängigkeit aufweisen. Auf der rechten Seite von Abbildung 11.5 ist die lokale Korrelation $lc(r)$ der Originaldaten (blau) und der n_P Permutationen (schwarz) abgebildet. Die Signifikanz der lokalen Korrelation (rot) wird mit Hilfe der Nullhypothese H_0 (keine Korrelation zwischen den Variablen) und einem Permutationstest, welcher auf den bereits vorhandenen n_P Permutationen beruht, bestimmt [3, 5].

$$H_0 : lc(r) = 0 \quad \text{für } r = 0, \frac{1}{\Delta r}, \dots, 1 \quad (11.20)$$

Die Signifikanz wird nun durch die Anzahl der lokalen Korrelationen der Permutationen lc_b größer der lokalen Korrelation lc im Verhältnis zur Gesamtzahl der Permutationen n_P abgeschätzt.

$$p(lc, r) \approx \frac{1}{n_P} \sum_{b=1}^{n_P} \# \{ |lc_b(r)| > |lc(r)| \} \quad (11.21)$$

Zur Abschätzung der globalen Korrelation, welche in den meisten Anwendungsfällen benötigt wird, kann die maximale (M) oder die mittlere (T) lokale Korrelation betrachtet werden.

$$\begin{aligned} M &= \max(|lc_i|), i = 1, \dots, n_G \\ T &= \frac{1}{n_G} \sum_{i=1}^{n_G} |lc_i| \quad \text{mit } n_G = \text{Anzahl Gitterpunkte} \end{aligned} \quad (11.22)$$

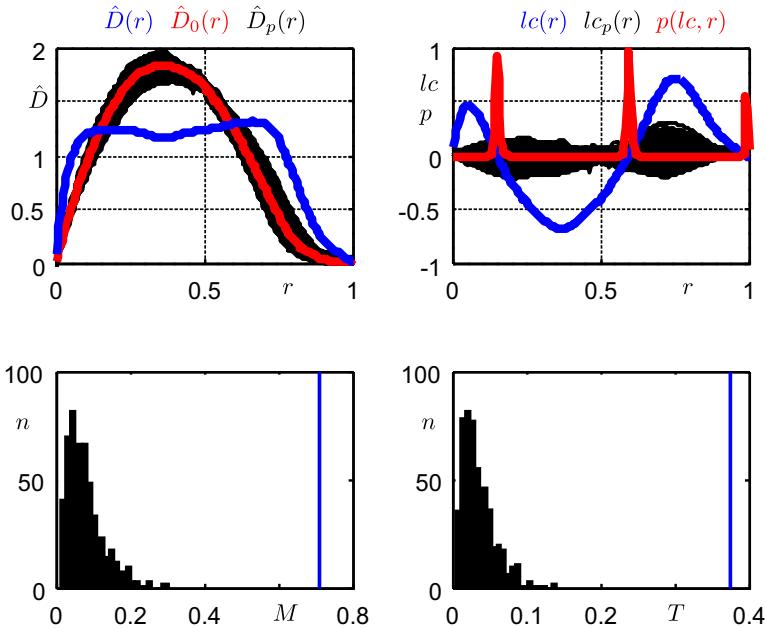


Abb. 11.5 Nichtlineare Korrelation: quadratischer Zusammenhang mit Rauschen, lokale Korrelation und Signifikanz

Abbildung 11.5 zeigt dazu jeweils ein Histogramm (schwarz) der Permutationen sowie den Wert des zu analysierenden Datensatzes in blau. Die Signifikanz wird über einen Permutationstest mit der Nullhypothese (H_0 : keine Korrelation) abgeschätzt und nach Gleichung 11.23 berechnet.

$$\begin{aligned} p(M) &\approx \frac{1}{n_P} \sum_{b=1}^{n_P} \# \{|M_b^*| > |M|\} \\ p(T) &\approx \frac{1}{n_P} \sum_{b=1}^{n_P} \# \{|T_b^*| > |T|\} \end{aligned} \quad (11.23)$$

Da im vorhandenen Beispiel kein M_b beziehungsweise T_b der Permutationen größer ist als der jeweilige Wert der Originaldaten, ist hier $p = 0$. Kleine Werte für $p < 0.05$ weisen dabei auf einen signifikanten Zusammenhang in den Daten. Im Vergleich zur nicht-linearen Korrelation des quadratischen Beispiels zeigen Abbildung 11.6 und 11.7 die Auswertung für zwei nicht korrelierte Variablen. Die Analyse von M und T führt dabei zu einem p-Wert von $p(M) \approx 0.3$ und $p(T) \approx 0.4$ was typischerweise nicht reicht um die Nullhypothese zu verwerfen, wodurch korrekterweise keine Signifikanz gezeigt werden kann.

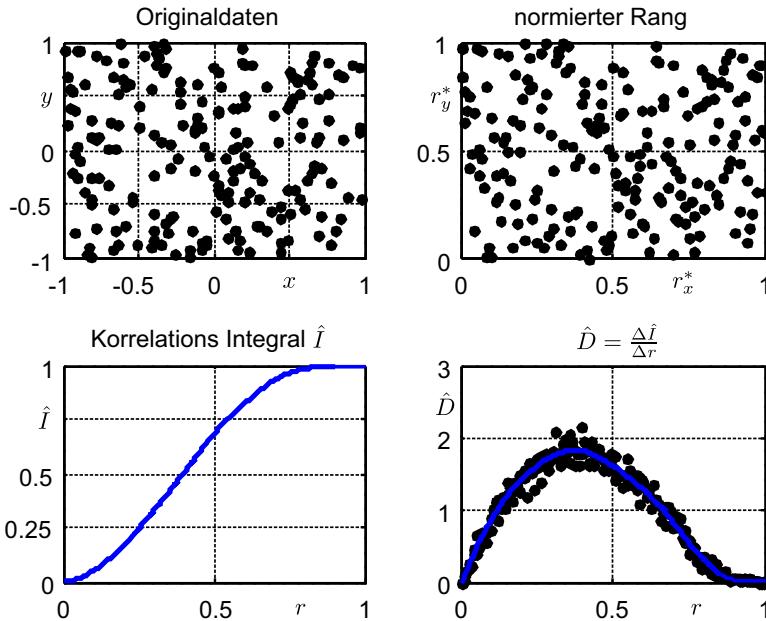


Abb. 11.6 Nichtlineare Korrelation: unkorrelierte Zufallszahlen, Korrelationsintegral und Dichtefunktion

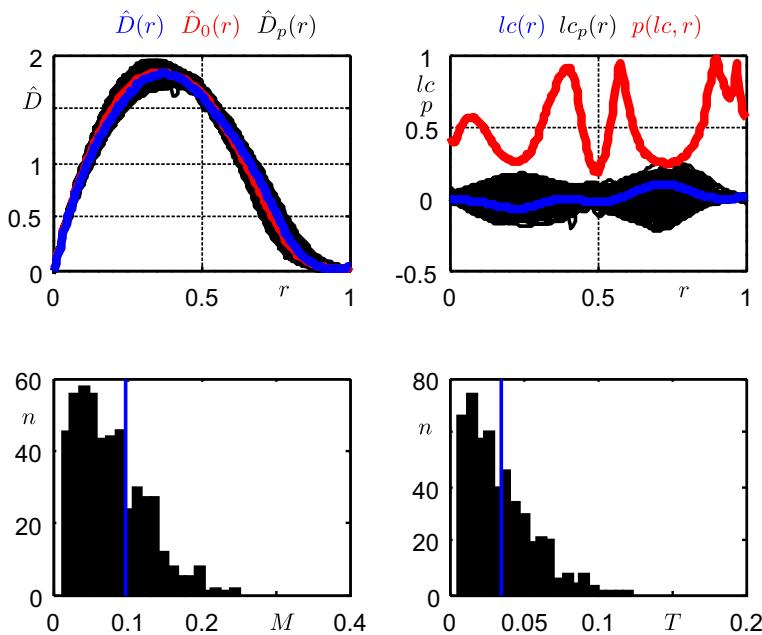


Abb. 11.7 Nichtlineare Korrelation: unkorrelierte Zufallszahlen, lokale Korrelation und Signifikanz

Literaturverzeichnis

1. Agresti, A.: *Analysis of Ordinal Categorical Data*. John Wiley and Sons, NJ (2010) 388
2. Andreß, H.J.: *Students t-Verteilung Berechnung*. URL <http://eswf.uni-koeln.de/glossar/surfstat/t.htm>. (abgerufen 03.2017) 384
3. Chen, Y.A., Almeida, J., Richards, A., Müller, P., Carroll, R.J., Rohrer, B.: *A nonparametric approach to detect nonlinear correlation in gene expression*. Journal of Computational and Graphical Statistics **19**(3), pp. 552–568 (2010) 389, 390, 391
4. Devlin, S., Gnanadesikan, R., Kettenring, J.: *Robust Estimation and Outlier Detection with Correlation Coefficients*. Biometrika **62**(3), p. 531–545 (1975) 383
5. Efron, B., Tibshirani, R.J.: *An Introduction to the Bootstrap*, 1 edn. Chapman and Hall/CRC (1994) 391
6. Eilers, P.H.C.: *A Perfect Smoother*. Analytical Chemistry **75**(14) (2003) 391
7. Eilers, P.H.C.: *Supporting Material to: A Perfect Smoother* (2009) 391
8. Fahrmeir, L., Künstler, R., Pigeot, I., Tutz, G.: *Statistik – Der Weg zur Datenanalyse*. Springer, Berlin (2004) 386
9. Huber, P., Ronchetti, E.M.: *Robust Statistics*. Wiley (2009) 239, 383
10. Mathworks: *Matlab Dokumentation* (2015). URL <https://de.mathworks.com/help/matlab/>. (abgerufen 03/2017) 236, 239, 240, 241, 384
11. Octave: *betainc - Compute the regularized incomplete Beta function*. URL <https://octave.sourceforge.io/octave/function/betainc.html>. (abgerufen 2017) 384
12. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press (2007) 199, 203, 220, 254, 265, 276, 277, 281, 285, 307, 334, 381, 398, 400
13. Surfstat: *t-distribution calculator*. URL <https://surfstat.anu.edu.au/surfstat-home/tables/t.php>. (abgerufen 03.2017) 384
14. Wikipedia: *Kendall tau rank correlation coefficient*. URL http://en.wikipedia.org/wiki/Kendall_tau_rank_correlation_coefficient. (abgerufen 2012,2016) 388
15. Wikipedia: *Pearson product-moment correlation coefficient*. URL http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient. (abgerufen 2012,2016) 381
16. Wikipedia: *Rangkorrelationskoeffizient*. URL <http://de.wikipedia.org/wiki/Rangkorrelationskoeffizient>. (abgerufen 2012,2016) 386, 387, 388

Kapitel 12

Komponentenanalyse

12.1 Einleitung

Stehen zur Analyse eines technischen Systems lediglich Datensätze zur Verfügung, bei denen die Variablen (Faktoren) miteinander korrelieren, ähnliche Zusammenhänge beschreiben oder vermischt redundante Informationen enthalten, ist eine aussagekräftige Analyse nur schwer möglich. Zur Strukturierung, Vereinfachung oder zum besseren Verständnis der Daten ist es dann sinnvoll, die Originaldaten durch neue Variablen abzubilden, welche sich aus einer Kombination der ursprünglichen Daten ermitteln lassen. Sinnvoll ist dieses zum Beispiel wenn die Anzahl der benötigten Variablen zur Abbildung der vorhandenen Datenstruktur deutlich reduziert werden kann, da in einer folgenden Modellbildung weniger und unkorrelierte Faktoren berücksichtigt werden müssen. Grundsätzlich wird zwischen Verfahren zur Hauptkomponentenanalyse (oder Hauptachsenanalyse) und Faktorenanalyse unterschieden [21, 22]. Beide Verfahren stellen lineare Modelle zwischen den Variablen mittels der Kovarianzmatrix auf.

Die **Hauptkomponentenanalyse** ist ein deskriptives-exploratives Verfahren, welches einen niedrig-dimensionalen Unterraum sucht, der die vorhandenen Daten am besten beschreibt, wobei die neuen Hauptkomponenten orthogonal zueinander stehen und die Originaldaten nach absteigender Varianz abbilden. Die Hauptkomponenten haben dadurch eine klare Reihenfolge und je nach gewünschter Qualität der Abbildung können im Abschluss mehr oder weniger Komponenten berücksichtigt werden.

Die **Faktorenanalyse** dient dazu, aus empirischen Beobachtungen verschiedener manifesten Variablen auf meist wenige zugrunde liegende latente Variablen (Faktoren) zu schließen. Es ist ein modellbasiertes Verfahren und bildet die vorhandene Kovarianzmatrix durch ein lineares Modell bestmöglich ab. Alle Komponenten sind dabei gleichberechtigt und die Anzahl muss bereits vor der Berechnung festgelegt werden.

12.2 Hauptkomponentenanalyse (PCA)

Die Hauptkomponentenanalyse (Hauptachsenanalyse, Principal Component Analyses [PCA]) wurde bereits 1901 von Pearson eingeführt (siehe auch Kapitel 11.1) [15]. PCA transformiert vorhandenen Daten mit n_f Dimensionen (Faktoren) in ein neues Koordinatensystem mit $n_{pc} \leq n_f$ Dimensionen, wobei die größte Varianz der Daten auf die erste Koordinate (erste Hauptkomponente) projiziert wird. Die zweitgrößte Varianz wird anschließend auf die zweite Hauptkomponente projiziert, wobei diese orthogonal zur ersten ist. Jede weitere Hauptachse wird durch absteigende Varianzen definiert, wobei sie orthogonal zu allen vorherigen Hauptachsen sind. Bei der Umwandlung geht die Hauptkomponentenanalyse immer von normalverteilten Daten aus. Abbildung 12.1 zeigt beispielhaft eine Koordinatentransformation (ohne Dimensionsreduktion) zweier korrelierter Variablen x und y in zwei unkorrelierte Variablen h_1 und h_2 . Da die Hauptkomponentenanalyse vorhandene Daten auf

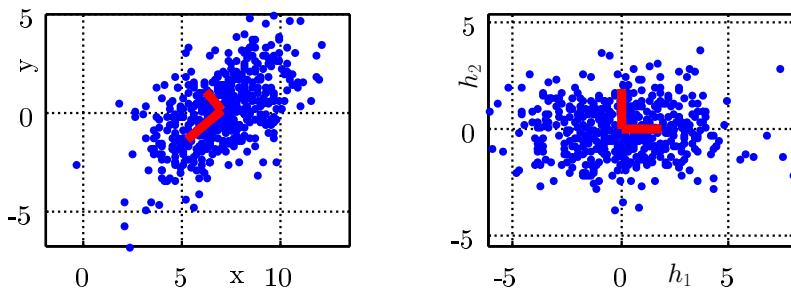


Abb. 12.1 Hauptkomponentenanalyse in 2 Dimensionen ohne Achsenreduktion

Koordinaten mit absteigender Varianz projiziert, ist eine einfache Reduktion der benötigten beziehungsweise verwendeten Dimensionen möglich. Dabei werden Komponenten (Dimensionen) mit insignifikanter Varianz gelöscht. Abbildung 12.2 zeigt dazu ein Beispiel, bei dem eine Reduktion von einem zweidimensionalen Fall in eine einzelne Dimension erfolgt. Die Hauptkomponentenanalyse ist ohne Probleme auch in höherdimensionalen Fällen einsetzbar wie Abbildung 12.3 illustriert.

Zur Berechnung der Hauptkomponenten finden sich in der Literatur verschiedene Ansätze, wobei eine verbreitete Variante auf die Singulärwertzerlegung (singular value decomposition [SVD]) beruht [20, 1]. In einem ersten Schritt wird dazu jede Variable (Faktor) der Datenmatrix $X_{n_r \times n_f}$ mit n_r Datenpunkten und n_f Faktoren (Variablen, Dimensionen) zentriert, was zu einer neuen zentrierten Datenmatrix

$$X_z = (x_{ij}^*)_{\substack{i=1,\dots,n_r \\ j=1,\dots,n_f}} \text{ führt.}$$

$$x_{ij}^* = x_{ij} - \frac{1}{n_r} \sum_{k=1}^{n_r} x_{kj} = x_{ij} - \bar{x}_j \quad (12.1)$$

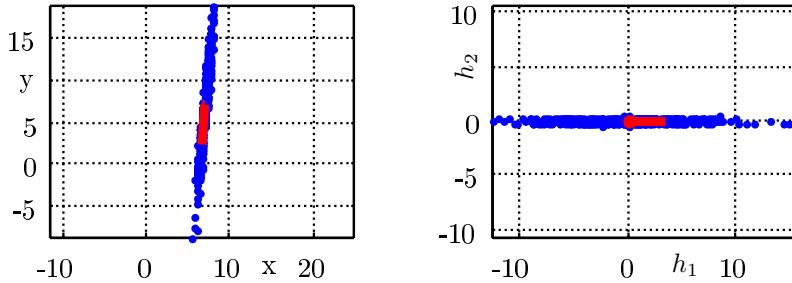


Abb. 12.2 Hauptkomponentenanalyse in 2 Dimensionen mit möglicher Achsenreduktion

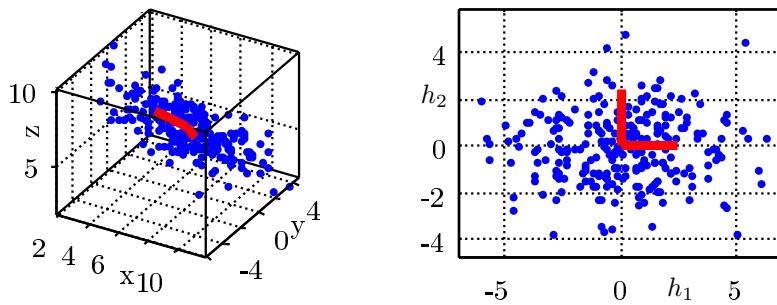
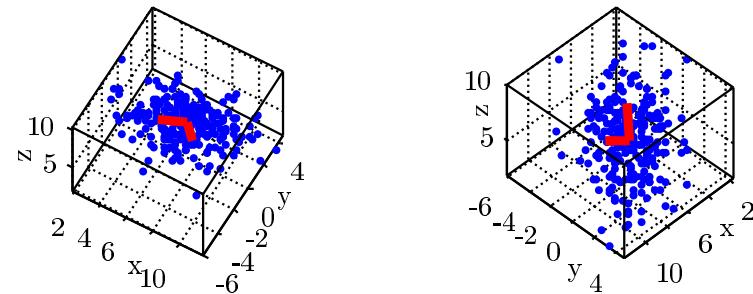


Abb. 12.3 Hauptkomponentenanalyse in einem 3 dimensionalen Raum

Die Hauptkomponenten, welche die Koordinatenachsen mit absteigender Varianz beschreiben, sind durch die Eigenvektoren mit absteigenden Eigenwerten der Kovarianzmatrix C beschrieben. Die empirische Kovarianz zweier Variablen \mathbf{x}_i und \mathbf{x}_j ist dabei durch Gleichung 12.2 gegeben.

$$C(x_j, x_k) = \frac{\sum_{i=1}^{n_r} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{n_r - 1} \quad (12.2)$$

Da die Daten bereits vor der Analyse zentriert wurden, ist der jeweilige Mittelwert Null ($\bar{x}_j = \bar{x}_k = 0$), was die Berechnung der kompletten Kovarianzmatrix auf Gleichung 12.3 reduziert.

$$C = \frac{X'_z X_z}{n_r - 1} \quad (12.3)$$

Die Berechnung der Eigenwerte und Eigenvektoren von C ist somit gleichbedeutend mit der Bestimmung der Eigenwerte und Eigenvektoren von $X'_z X_z$, was mittels der *ökonometrischen* Singulärwertzerlegung (SVD) ausgeführt wird. Die ökonomische Singulärwertzerlegung ist dabei wie in Gleichung 12.4 definiert [16, 26].

$$X_z = U_o \Sigma_o V' \quad (12.4)$$

mit

V ist eine $n_f \times n_f$ unitäre Matrix mit reellen oder komplexen Einträgen

U_o ist eine $n_r \times n_f$ unitäre Matrix

Σ_o ist eine $n_f \times n_f$ Matrix mit nichtnegativen reellen Zahlen auf der Diagonalen

$$\Sigma_o = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_{n_f} \end{pmatrix} \quad (12.5)$$

mit $\sigma_1 \geq \dots \geq \sigma_{n_f} > 0$

Weiterhin gilt, dass:

$$\begin{aligned} XX' &= U \Sigma^2 U' \\ X'X &= V \Sigma^2 V' \end{aligned} \quad (12.7)$$

Die Spalten von V bilden die Eigenvektoren (EV) von $X'X$ beziehungsweise C und die Eigenwerte von C sind proportional zu den quadrierten Diagonalkomponenten von Σ .

$$EV_k = (V_{ik})_{i=1 \dots n_f} \quad (12.8)$$

Die zentrierten Daten aus dem Originalkoordinatensystem X werden anschließend mit n_h ausgewählten Eigenvektoren in die neuen Hauptachsen transformiert.

$$X_h = X_z EV_h \text{ mit } i = 1 \dots n_h \quad (12.9)$$

Tabelle 12.1 zeigt zur Veranschaulichung die Hauptkomponententransformation an einem Beispiel mit 10 Datenpunkten in 3 Dimensionen.

Soll die Anzahl der Dimensionen reduziert werden $n_h < n_f$, so kann mit Hilfe der Eigenwerte $(\sigma_1, \dots, \sigma_{n_f})$ der prozentuale Anteil p_e jeder Hauptkomponente an der Gesamtvarianz bestimmt werden.

Tabelle 12.1 kleines PCA Beispiel

Originaldaten			zentrierte Daten			Hauptachsen		transformierte Daten	
x_1	x_2	x_3	x_{z1}	x_{z2}	x_{z3}	h_1	h_2	x_{h1}	x_{h2}
8,29	1,72	4,86	1,618	0,88	-0,845	-0,804...	0,531...	-0,698...	1,892...
4,07	2,91	4,26	-2,602	2,07	-1,445	0,585...	0,627...	3,456...	0,742...
7,77	-1,75	6,29	1,098	-2,59	0,585	-0,105...	-0,571...	-2,459...	-1,375...
9,24	0,86	5,15	2,568	0,02	-0,555			-1,995...	1,692...
6,99	1,27	5,21	0,318	0,43	-0,495			0,0479...	0,721...
6,49	0,26	6,06	-0,182	-0,58	0,355			-0,230...	-0,663...
4,66	1,45	6,5	-2,012	0,61	0,795			1,891...	-1,139...
8,17	-0,92	6,15	1,498	-1,76	0,445			-2,281...	-0,562...
5,51	2,14	5,75	-1,162	1,3	0,045			1,690...	0,173...
5,53	0,46	6,82	-1,142	-0,38	1,115			0,579...	-1,480...

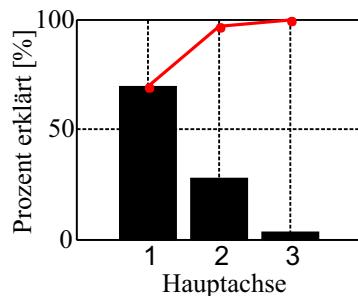
$$p_e(h_i) = \frac{\sigma_i^*}{\sum_{j=1}^{n_f} \sigma_j^*} \quad (12.10)$$

mit $\sigma_i^* = \left(\frac{\Sigma_{i,i}}{\sqrt{n_r - 1}} \right)^2$

Tablle 12.2 und Abbildung 12.4 zeigen für das aktuelle Beispiel, dass bereits die ersten zwei Hauptkomponenten 97 Prozent der gesamten Varianz erklären und in vielen Fällen auf die dritte Hauptkomponente verzichtet werden kann.

Tabelle 12.2 Erklärte Varianz mittels der Hauptkomponenten

h_i	$p_e [\%]$	$\sum p_e [\%]$
1	69.3	69.3
2	27.48	96.78
3	3.22	100

**Abb. 12.4** Erklärte Varianz mittels Hauptkomponenten

Wird ein System im Folgenden mittels der Hauptkomponenten analysiert und opti-

miert, so ist die Interpretation der Ergebnisse in einigen Fällen schwierig und eine Rücktransformation in das Originalkoordinatensystem sinnvoll. Die Rücktransformation wird dabei durch die inverse Eigenvektormatrix EV^{-1} ermöglicht. Hierbei ist zu beachten, dass es sich um eine Rücktransformation in das zentrierte Koordinatensystem handelt und im Anschluss die abgezogenen Mittelwerte aus Gleichung 12.1 addiert werden müssen. Im Fall einer Dimensionsreduktion $n_h < n_f$ muss eine Pseudoinverse berechnet werden [25]. Da in diesem Fall Informationen aus den vernachlässigten Achsen fehlen, können die Originaldaten jedoch nicht exakt ermittelt werden.

$$X_z = X_h EV^{-1} \quad (12.11)$$

12.3 Kernel-Hauptkomponentenanalyse (kPCA)

Die klassische Hauptkomponentenanalyse basiert auf einer linearen Achsentransformation und funktioniert nur bedingt bei Daten, die keine linearen Zusammenhänge aufweisen. Verschiedene Arbeiten zeigen, dass nicht linear separierbare Daten aus einem Faktorraum \mathcal{R}^{n_f} in einen anderen Faktorraum $\mathcal{R}_{n_f^*}$ (*Eigenschaftsraum*) transformiert werden können, in dem die Daten linear separierbar sind, wobei n_f^* größer als n_f sein kann [18, 17, 19, 23]. Obwohl eine Erhöhung der Variablenanzahl zur Analyse oder Modellbildung nicht gewünscht ist, hat sich in vielen Fällen gezeigt, dass die Modellbildung für die transformierten Daten einfacher ist. Zur Transformation zwischen den Faktorräumen wird im ersten Schritt eine allgemeine Funktion Φ eingeführt.

$$\begin{aligned} \Phi : \mathcal{R}^{n_f} &\rightarrow \mathcal{R}^{n_f^*} \\ x &\rightarrow \Phi(x) \end{aligned} \quad (12.12)$$

Die willkürliche aber meist komplexe Funktion Φ muss glücklicherweise während der Kernel-Hauptkomponentenanalyse (kPCA) niemals explizit berechnet werden. Stattdessen wird eine Kernelmatrix $K_{n_r \times n_r}$ berechnet, in der jede Spalte das Skalarprodukt jedes transformierten Punktes mit allen anderen transformierten Punkten darstellt.

$$K = k(X, Y) = \Phi(X)' \Phi(Y) \quad (12.13)$$

Typische Kernelfunktionen finden sich in vielen Literaturstellen, wobei Tabelle 12.3 einige typische Definitionen darstellt in denen a, \dots, d konstante Parameter zum anpassen der Funktionen bezeichnen [16, 3]:

Anstelle Hauptkomponenten im transformierten Raum zu berechnen, wird lediglich die Projektion auf die Hauptachsen ermittelt. Die Projektion eines Punktes \mathbf{x}_i im Eigenschaftsraum $\Phi(\mathbf{x})$ auf die k^{te} Hauptachse V_k wird dazu nach Gleichung 12.14 berechnet.

$$\begin{aligned}
\text{Linear: } k(\mathbf{x}_i, \mathbf{x}_j) &= \mathbf{x}_i \cdot \mathbf{x}_j \\
\text{Power: } k(\mathbf{x}_i, \mathbf{x}_j) &= (\mathbf{x}_i \cdot \mathbf{x}_j)^d \\
\text{Polynom: } k(\mathbf{x}_i, \mathbf{x}_j) &= (a\mathbf{x}_i \cdot \mathbf{x}_j + b)^d \\
\text{Sigmoidal: } k(\mathbf{x}_i, \mathbf{x}_j) &= \tanh(a\mathbf{x}_i \cdot \mathbf{x}_j + b) \\
\text{Gauß: } k(\mathbf{x}_i, \mathbf{x}_j) &= \exp(-\frac{1}{2}|\mathbf{x}_i - \mathbf{x}_j|^2/\sigma^2)
\end{aligned}$$

Tabelle 12.3 Kernelfunktionen

$$V_k \cdot \Phi(\mathbf{x}_i) = \sum_{j=1}^{n_r} \alpha_{k,i} \langle \Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x}_i) \rangle = \sum_{j=1}^{n_r} \alpha_{k,i} k(\mathbf{x}_j, \mathbf{x}_i) \quad (12.14)$$

$\langle \Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x}_i) \rangle$ bildet dabei die Kernelmatrix K und es muss lediglich der Vektor α_k berechnet werden, was durch die Lösung des Eigenwertproblems $n_r \lambda \alpha = K\alpha$ mit $\alpha_i \geq 0$ und $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_{n_r}$ erfolgt [17]. Zur Normalisierung von α_k muss $\lambda_k (\alpha_k \cdot \alpha_k) = 1$ gelten, wobei im Vergleich zur klassischen Hauptkomponentenanalyse (Kapitel 12.2), grundsätzlich λ_i den Eigenwerten und α_i den Eigenvektoren entspricht.

Da eine Zentrierung der Daten im Ereignisraum $\mathcal{R}^{n_f^*}$ durch zentrierte Daten im Ursprungsraum \mathcal{R}^{n_f} (siehe Kapitel 12.2) nicht garantiert werden kann und die Daten sowieso nicht explizit mit Φ transformiert werden, wird lediglich die Kernelmatrix „zentriert“ [17].

$$K^* = K - 1_{n_r} K - K 1_{n_r} + 1_{n_r} K 1_{n_r} \quad (12.15)$$

1_{n_r} ist dabei eine $n_r \times n_r$ Matrix, bei der jedes Element den Wert $\frac{1}{n_r}$ aufweist. Im Anschluss wird eine klassische Hauptkomponentenanalyse von K^* durchgeführt (Kapitel 12.2) und die Eigenvektoren EV_k ermittelt. Da K^* eine $n_r \times n_r$ Matrix ist können n_r Hauptkomponenten (Eigenvektoren) ermittelt werden. Die Eigenvektoren werden im Anschluss mit den dazugehörigen Eigenwerten und Gleichung 12.16 normiert.

$$\alpha_k = \frac{EV_k}{\sqrt{\sigma_k}} \quad (12.16)$$

Die eigentliche Projektion eines Datenpunktes $\mathbf{x}^\#$ auf die k^{te} Hauptkomponenten α_k im Eigenschaftsraum wird mit Hilfe der *nicht* zentrierten Kernelfunktion $k(\mathbf{x}^\#, \mathbf{x}_i)$ und α_k durchgeführt.

$$\mathbf{x}_{pc_k}^\# = K^\# \alpha_k \quad (12.17)$$

$K^\# = k(\mathbf{x}^\#, \mathbf{x}_i)$ mit $i = 1, \dots, n_r$ beschreibt dabei die Beziehung des Datenpunkts $\mathbf{x}^\#$ zu allen Trainingspunkten aus X . Zur Verdeutlichung zeigt Abbildung 12.5 ein auf Arbeiten von Schölkopf basierendes Beispiel [18], bei dem eine verrauschte 2-dimensionale quadratische Funktion mittels dem Power Kernel ($d = 2$) transformiert wird. Die ersten drei Darstellungen zeigen die Datenpunkte im Originalkoordinatensystem sowie Höhenlinien des projizierten Originalkoordinatensystems x_1, x_2 in

jeweils eine der ersten drei Hauptkomponenten h_1 bis h_3 . Bereits die erste Hauptkomponente zeigt eine gute Anpassung an die vorhandene nichtlineare Datenstruktur. Passend dazu zeigt die Analyse der erklärten Varianz (Eigenwerte), dass auch im Eigenschaftsraum bereits mit zwei Hauptkomponenten nahezu 100% der Varianz erklärbar wird (Abbildung 12.6). Passend dazu zeigt das letzte Bild unten rechts (Abbildung 12.5) die Projektion der Daten auf die ersten zwei Hauptachsen.

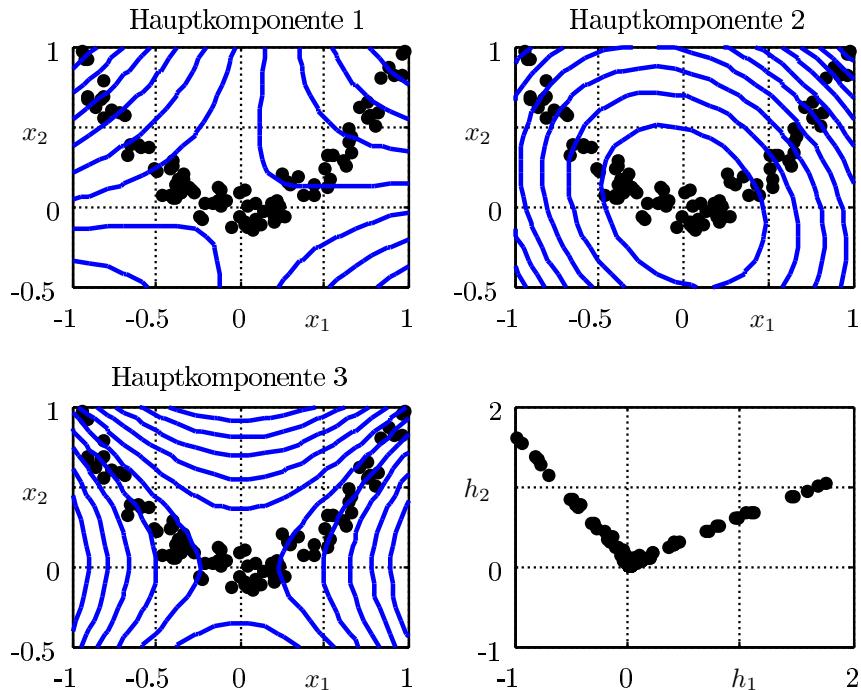


Abb. 12.5 Kernel Hauptkomponentenanalyse (kPCA) für ein quadratisches Beispiel

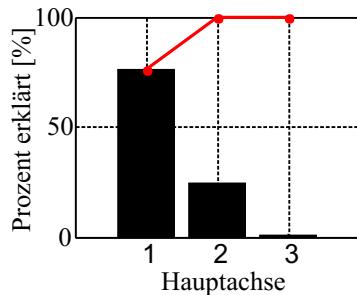


Abb. 12.6 Erklärte Varianz einer kPCA für ein quadratisches Beispiel

12.4 Unabhängige Komponentenanalyse (ICA)

Das grundsätzliche Ziel einer Komponentenanalyse ist es eine aussagekräftige Abbildung multivariater Daten mittels einer linearen Transformation zu finden. Die Unabhängige Komponentenanalyse (Independent Component Analysis: ICA) oder auch Unabhängigkeitsanalyse versucht hingegen eine lineare Abbildung von *nicht*-normalverteilten Daten zu finden, so dass diese so statistisch unabhängig wie möglich sind [8, 9, 11]. Zur Verdeutlichung sollen hier aus der Literatur gebräuchliche und einfache Beispiele dienen. Abbildung 12.7 zeigt das sogenannte *Cocktail-Party* Problem, bei dem zwei unabhängige Tonsignale, wie zum Beispiel zwei Personen auf einer Party oder zwei Geräuschquellen eines technischen Systems, von zwei unterschiedlichen Personen wahrgenommen beziehungsweise an zwei unterschiedlichen Positionen gemessen werden. Die Originalsignale (rot und blau) können dabei nicht separat ermittelt werden, so dass lediglich zwei Messsignale gegeben sind, die eine unbekannte Mischung aus beiden Originalsignalen enthalten. Das Ziel der unabhängigen Komponentenanalyse ist es nun aus den gemischten Signalen die unabhängigen Originalsignale zu extrahieren. Mit diesen meist aussagekräftigeren Daten können dann anschließend sinnvollere Metamodelle aufgebaut und analysiert werden, als auf Basis der gemischten Signale. Mathematisch wird die lineare Vermischung der Signale wie in Gleichung 12.18 abgebildet.

$$\begin{aligned} x_1 &= a_{11}s_1 + a_{12}s_2 \\ x_2 &= a_{21}s_1 + a_{22}s_2 \end{aligned} \quad (12.18)$$

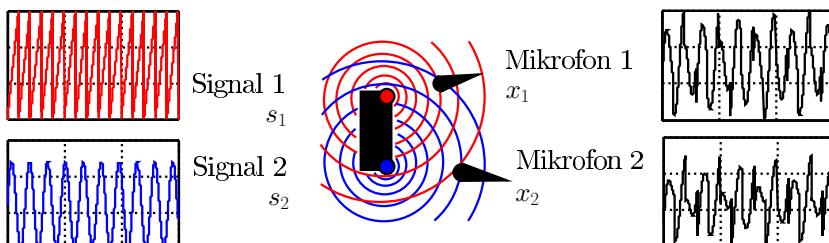


Abb. 12.7 Cocktail-Party Problem

Die Unabhängigkeitsanalyse ist nicht auf zeit-basierte Daten beschränkt, sondern kann auch auf andere Messdaten, wie zum Beispiel aus einem Design of Experiment, angewendet werden. Dazu zeigen die Abbildungen 12.8 und 12.9 eine Unabhängigkeitsanalyse zweier verrauschter nicht-normalverteilter und unabhängiger Variablen s_1 und s_2 .

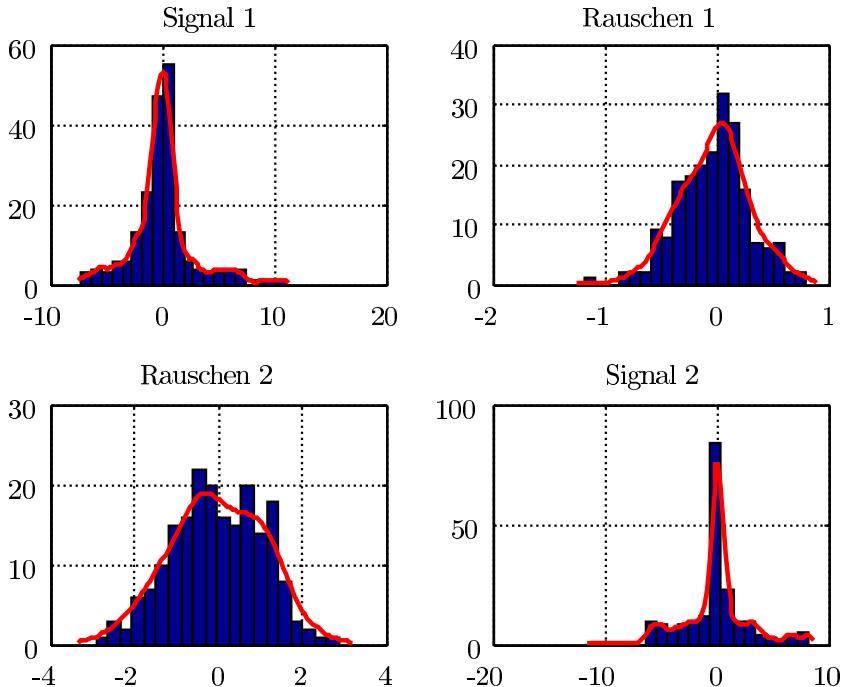


Abb. 12.8 Originalsignale mit nicht-normaler Verteilung

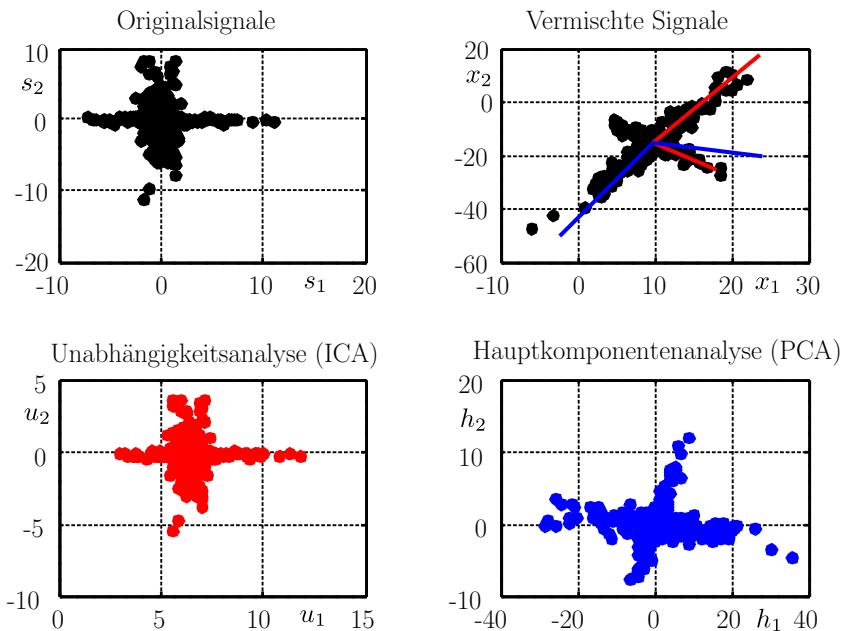


Abb. 12.9 original, vermischte und analysierte Variablen (ICA, PCA)

Die beiden Signale werden zu den abhängigen Variablen $x_{1,2}$ mittels Gleichung 12.19 vermischt.¹

$$\begin{aligned}x_1 &= 0.8s_1 + 1.3s_2 \\x_2 &= -s_1 + 3s_2 \\X = SA \text{ mit } A &= \begin{pmatrix} 0.8 & -1 \\ 1.3 & 3 \end{pmatrix} \text{ und } S = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \\S &= XA^{-1} = XW\end{aligned}\tag{12.19}$$

Im Gegensatz zur klassischen Hauptkomponentenanalyse (PCA), welche orthogonale Hauptkomponenten ermittelt, die in Richtung der größten Varianzen zeigen (blaue Achsen, Abbildung 12.9), ermittelt die unabhängige Komponentenanalyse (ICA) die unabhängigen Achsen entlang der nicht-normalverteilten Daten (rote Achsen). Lediglich die Unabhängigkeitsanalyse kann somit aus den vermischten Daten die grundsätzlichen Originalsignale separieren.

Ohne Einschränkung der Allgemeingültigkeit der Unabhängigkeitsanalyse wird davon ausgegangen, dass alle Originaldaten $s_{1,2}$ und vermischten Daten $x_{1,2}$ zentriert sind, also einen Mittelwert von Null aufweisen. Weiterhin kann nach HYVÄRINEN ohne Beeinträchtigung der Allgemeingültigkeit davon ausgegangen werden, dass alle Daten aus S eine Varianz von eins aufweisen ($E\{s_k^2\} = 1$) [8].

Da beide Matrizen S und A unbekannt sind, ist es nicht möglich die absolute Varianz (Energie) der unabhängigen Komponenten zu bestimmen. Jeder skalare Faktor einer Spalte k von S kann durch eine Division der k^{ten} Spalte von A mit dem gleichen skalaren Wert ausgeglichen werden. Somit ist es leicht möglich jede unabhängige Komponente auf eine Varianz von eins zu normieren und A entsprechend zu adaptieren. Das Vorzeichen der Komponenten ist nicht eindeutig bestimmt, was in der Praxis jedoch in den überwiegenden Fällen ohne Bedeutung ist. Die Reihenfolge der Komponenten ist im Gegensatz zur PCA (Kapitel 12.2) nicht bestimmbar, so dass jede unabhängige Komponente gleichberechtigt ist. Der Grund liegt ebenfalls in der Tatsache, dass A und S nicht absolut bekannt sind, so dass jede Permutation der Spalten von S durch eine Permutationsmatrix P mit der inversen Permutationsmatrix P^{-1} welche auf A angewendet wird ausgeglichen werden kann [8].

$$SPP^{-1}A = SA\tag{12.20}$$

12.4.1 Unabhängigkeit

Zwei Variablen $x_{1,2}$ werden als unabhängig bezeichnet, wenn beliebige Informationen einer Variable keine Information der anderen Variable liefern. Mathematisch

¹ In der Literatur finden sich meistens Herleitungen, in denen die Daten nicht in Zeilen sondern Spalten gespeichert werden, so dass sich dadurch entsprechende Gleichungen ergeben. $X = AS$ und $S = A^{-1}X = WX$

kann dieses Verhalten über Dichtefunktionen beschrieben werden. Sei $p(x_1, x_2)$ die multivariate Dichtefunktion (joint probability density function) der beiden Variablen x_1 und x_2 sowie $p_1(x_1) = \int p(x_1, x_2) dx_2$ die Dichtefunktion, wenn x_1 alleine betrachtet wird. x_1 und x_2 sind genau dann linear unabhängig, wenn gilt [8]:

$$p(x_1, x_2) = p_1(x_1) p_2(x_2) \quad (12.21)$$

Basierend auf dieser Unabhängigkeitsdefinition kann die wichtige Eigenschaft abgeleitet werden, dass für beliebige Funktionen f_1 und f_2 folgende Bedingung erfüllt sein muss [8]:

$$E\{f_1(x_1)f_2(x_2)\} = E\{f_1(x_1)\}E\{f_2(x_2)\} \quad (12.22)$$

Die Unabhängigkeit ist eine stärkere Bedingung als die Unkorreliertheit, so dass unabhängige Daten auch immer unkorreliert sind, was jedoch umgekehrt nicht der Fall sein muss. Variablen gelten als unkorreliert, wenn gilt:

$$E\{x_1x_2\} = E\{x_1\}E\{x_2\} \quad (12.23)$$

Veranschaulicht wird der Unterschied durch ein Beispiel von HYVÄRINEN [8]:

$$X = \begin{pmatrix} 0 & 1 \\ 0 & -1 \\ 1 & 0 \\ -1 & 0 \end{pmatrix}$$

$$f_1(x) = f_2(x) = x^2 \quad (12.24)$$

$$\begin{aligned} E\{x_1x_2\} &= E\{x_1\} = E\{x_2\} = 0 \\ E\{f_1(x_1)f_2(x_2)\} &= E\{x_1^2x_2^2\} = 0 \\ E\{f_1(x_1)\} &= E\{f_2(x_2)\} = E\{x_1^2\} = E\{x_2^2\} = \frac{1}{2} \\ \Rightarrow E\{x_1x_2\} - E\{x_1\}E\{x_2\} &= 0 \rightarrow \text{unkorreliert} \\ \Rightarrow E\{x_1^2x_2^2\} &= 0 \neq \frac{1}{4} = E\{x_1^2\}E\{x_2^2\} \rightarrow \text{nicht unabhängig} \end{aligned} \quad (12.25)$$

Da unabhängige Daten auch immer unkorreliert sind, kann die Suche von unabhängigen Komponenten auf unkorrelierte Komponenten beschränkt werden, was die Anzahl der freien Parameter im Bestimmungsalgorithmus einschränkt.

12.4.2 Nicht-Normalverteilt

Die Unabhängigkeitsanalyse ist im Gegensatz zu den meisten klassischen statistischen Analyseverfahren, welche von normalverteilten Daten ausgehen, auf nicht-normalverteilte Daten beschränkt. Zur Veranschaulichung zeigt Abbildung 12.10 zwei unabhängige Variablen mit Normalverteilung und jeweils einer Varianz von

eins. Die Verteilung der Datenpunkte ist symmetrisch, so dass keine Informationen zur Berechnung der Vermischungsmatrix extrahiert werden können. Es kann gezeigt werden, dass jede orthogonale Transformation der beiden Variablen x_1 und x_2 immer eine symmetrische Verteilung ergibt. Aus diesem Grund ist die Unabhängigkeitsanalyse lediglich möglich wenn *maximal* eine der betrachteten Variablen eine Normalverteilung aufweist.

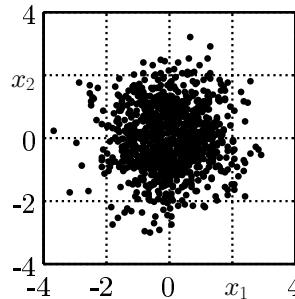


Abb. 12.10 Verteilung zweier unabhängiger und normalverteilter Variablen

Der zentrale Grenzwertsatz besagt, dass die Verteilungsfunktion der Summe zweier unabhängiger Variablen näher an einer Normalverteilung liegt, als die beiden Variablen alleine [28]. Daraus lässt sich folgern, dass die gesuchten unabhängigen nicht-normalverteilten Komponenten eine deutlichere nicht-Normalverteilung aufweisen als die gemessenen vermischten Signale oder auch jede beliebige Vermischung der unabhängigen Komponenten. Bei der unabhängigen Komponentenanalyse (ICA) muss somit die nicht-Normalverteilung von S maximiert werden. Dieses Optimierungsproblem weist $2n_f$ lokale Maxima in einem n_f dimensional Raum auf [8]. Jede unabhängige Komponente kann dabei jeweils zweimal gleichberechtigt bei \mathbf{s}_j und $-\mathbf{s}_j$ auftreten. Zur Beurteilung des Grades der nicht-Normalverteilung einer zentrierten Variable (Mittelwert = 0) mit der Varianz von eins wurden verschiedene Kriterien entwickelt [8].

Kurtosis

Klassischerweise wird die Kurtosis k (Wölbung, Exzess) verwendet, welche durch Gleichung 12.26 definiert ist.

$$k(x) = \frac{E\{x^4\}}{(E\{x^2\})^2} - 3 \quad (12.26)$$

Für eine normalverteilte Variable x mit dem Mittelwert $\bar{x} = 0$ und einer Varianz von $Var(x) = 1$ ist die Kurtosis $k = 0$, da $\frac{E\{x^4\}}{(E\{x^2\})^2} = 3$. Für nahezu alle nicht-normalverteilten Variablen (es gibt nur wenige Ausnahmen) ist die Kurtosis un-

gleich Null und kann somit als Maß für den Grad der nicht-Normalverteilung verwendet werden. Bei einer fest vorgegebenen Varianz von eins wird die Kurtosis vereinfacht wie in Gleichung 12.27 berechnet.

$$k(x) = E\{x^4\} - 3 \quad (12.27)$$

Obwohl die Kurtosis oder sein Absolutwert in vielen Fällen zur Unabhängigkeitsanalyse verwendet wird, weist sie gerade bei Verwendung von physikalischen Messdaten einige Nachteile auf. So regt die Kurtosis sehr empfindlich auf Ausreißer und kann somit von einigen irrelevanten Messdaten am Rand der Verteilungsfunktion abhängen [8, 4].

Negentropie

Das alternative Kriterium Negentropie (negative Entropie) basiert auf der Informationstheorie und der differentiellen Entropie [8]. Die Entropie einer Variablen wird als Grad der Information interpretiert, welche in den Daten enthalten ist. Umso unvorhersehbarer (zufälliger) die Variabel ist, desto größer ist die dazugehörige Entropie H . Für eine zufällige diskrete Variable x ist die Entropie wie in Gleichung 12.28 definiert [8, 2, 14].

$$H(x) = - \sum_i P(x = a_i) \log_2 \{P(x = a_i)\} \quad (12.28)$$

mit a_i : alle möglichen Werte von x

P : Wahrscheinlichkeit

Bei gegebener Verteilungsfunktion $f(x)$ für eine Variable x ist die Entropie somit durch Gleichung 12.29 definiert [2, 14, 8]:

$$H(\mathbf{x}) = - \int f(\mathbf{x}) \log_2 \{f(\mathbf{x})\} d\mathbf{x} \quad (12.29)$$

Eine normalverteilte Zufallsvariable x weist im Vergleich zu allen anderen Zufallsvariablen mit der gleichen Varianz die höchste Entropie auf. Das bedeutet, dass die Entropie als Maß für den Grad der nicht-Normalverteilung J eingesetzt werden kann.

$$J(x) = H(x_{normal}) - H(x) \geq 0 \quad (12.30)$$

x_{normal} ist dabei eine normalverteilte Zufallsvariable mit der gleichen Kovarianzmatrix wie x . J ist immer größer Null und nur genau dann Null wenn x normalverteilt ist. Da zur expliziten Berechnung der Negentropie die genaue Verteilungsfunktion bekannt sein muss, ist die genaue Berechnung in den meisten Fällen kompliziert, so dass zur Vereinfachung verschiedene Approximationen eingesetzt werden. Ein typisches Verfahren verwendet dazu Momente höherer Ordnung [10, 8].

$$J(x) \approx \frac{1}{12} E\{x^3\}^2 + \frac{1}{48} k(x)^2 \quad (12.31)$$

Da hierbei auch die Kurtosis $k(x)$ verwendet wird, ist diese Approximation ebenfalls nicht robust, so dass HYVÄRINEN eine weitere Approximationsmethode auf Basis der maximalen Entropie einführt [5, 8].

$$J(x) \approx \sum_{i=1}^p c_i [E\{G_i(x)\} - E\{G_i(v)\}]^2 \geq 0 \quad (12.32)$$

In dieser Gleichung sind c_i positive Konstanten und G_i beliebige aber nicht quadratische Funktionen. v ist eine normalverteilte Variable mit einem Mittelwert von Null und einer Varianz von eins. Auch wenn die Approximation in einigen Fällen nicht exakt ist, konnte gezeigt werden, dass sie als Maß für die nicht-Normalverteilung einsetzbar ist. Der Wert ist immer größer gleich Null und nur genau dann Null, wenn x eine Normalverteilung aufweist [8]. Zur Vereinfachung wird lediglich eine Funktion G_i verwendet, so dass sich die Berechnung auf folgende Form reduzieren lässt.

$$J(x) \propto [E\{G(x)\} - E\{G(v)\}]^2 \quad (12.33)$$

Typische Funktionen für $G(x)$ sind dabei:

$$\begin{aligned} G(x) &= \frac{1}{a_1} \log \{\cosh(a_1 x)\} \text{ mit } 1 \leq a_1 \leq 2 \\ G(x) &= -e^{-\frac{x^2}{2}} \end{aligned} \quad (12.34)$$

Weitere Methoden zur Unabhängigkeitsanalyse basieren auf der Berechnung von Transinformation (mutual information) oder der Maximal-Likelihood Methode, wobei diese hier nicht näher betrachtet werden [8, 27, 24].

12.4.3 Datenvorbereitung

Zur Vereinfachung der Berechnung und besseren Konditionierung gegebener Daten werden einige Umformungen eingeführt. Im ersten Schritt ist, wie bereits beschrieben, eine Zentrierung der Daten notwendig.

$$X_z = X - E\{X\} \quad (12.35)$$

Im nächsten Schritt werden die Daten X_z in einer Weise linear transformiert, so dass diese (X_w) unkorreliert sind und die Kovarianzmatrix der Einheitsmatrix I entspricht.

$$E\{X'_w X_w\} = I \quad (12.36)$$

Dieses Vorgehen ist als *Whiten* bekannt und wird entweder durch eine Eigenwertzerlegung oder Singulärwertzerlegung (SVD) (siehe Kapitel 12.2) der Kovarianzmatrix von X_z ermöglicht [8]. Sind die Eigenvektoren E und Eigenwerte D der Kovarianzmatrix $K = \frac{X_z' X_z}{n_p}$ bekannt, so wird die *whiten* Matrix M_w und *dewhitened* Matrix M_{dw} zur Rücktransformation sowie die transformierten Daten X_w nach den Gleichungen 12.37 berechnet.

$$\begin{aligned} M_w &= \left[\sqrt{D} \right]^{-1} E' \\ M_{dw} &= E \sqrt{D} \\ X_w &= X_z M_w \end{aligned} \tag{12.37}$$

Durch das Whitening wird sichergestellt, dass die zu berechnende Vermischungsmatrix A_w orthogonal ist.

$$X_w = M_w A S = A_w S \tag{12.38}$$

Das Whitening reduziert die Anzahl unbekannter Komponenten von n_f^2 der Originalmatrix A auf $\frac{n_f(n_f-1)}{2}$ der orthogonalen Matrix A_w [8].

12.4.4 FastICA

Von HYVÄRINEN wurde ein effizienter Algorithmus zur Unabhängigkeitsanalyse (FastICA) vorgeschlagen [6, 8, 9, 11]. Der Algorithmus ermittelt einen Richtungsvektor \mathbf{w} , so dass die Projektion $X_w \mathbf{w}$ die Nicht-Normalverteilung, welche durch die Negentropie $J(X_w \mathbf{w})$ abgeschätzt wird, maximiert. Die Bedingung, dass $X_w \mathbf{w}$ die Einheitslänge aufweisen muss, wird durch das vorherige Whiten der Daten und die Normierung von \mathbf{w} auf eins erfüllt $\mathbf{w}^* = \frac{\mathbf{w}}{\|\mathbf{w}\|}$. Das grundlegende Vorgehen von FastICA zur Ermittlung eines Richtungsvektors \mathbf{w} wird durch Algorithmus 19 beschrieben [6, 8, 9, 11, 12, 7].

- 1 Daten zentrieren $X_z = X - E\{X\}$
- 2 Whiten der zentrierten Daten: $X_w = X_z M_w$ erstelle zufälligen normalisierten Startvektor \mathbf{w}
- 3 solange \mathbf{w} nicht konvergiert tue
 - 4 korrigiere $\mathbf{w} \leftarrow E\{g(\mathbf{w} X'_w)\} X_w - E\{g'(\mathbf{w} X'_w)\} \mathbf{w}$, wobei g und g' die ersten zwei Ableitungen von G sind
 - 5 normalisiere von \mathbf{w} mit $\mathbf{w} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$
- 6 Ende
- 7 $\mathbf{w} = M_w \mathbf{w}'$

Algorithmus 19 : FastICA

Konvergenz bedeutet in diesem Fall, dass der alte und neue \mathbf{w} Vektor die gleiche Richtung aufweisen und deren Skalarprodukt nahezu eins ist. \mathbf{w} muss nicht in den

selben Punkt konvergieren, da das Vorzeichen irrelevant ist (\mathbf{w} und $-\mathbf{w}$ definieren grundsätzlich die gleiche Richtung). Der vorgeschlagene Algorithmus verwendet die Funktion G nicht direkt, sondern lediglich deren ersten zwei Ableitungen. Tabelle 12.4 zeigt einige typische Funktionen [12].

Tabelle 12.4 Verschiedene verwendete Funktionen G für FastICA

Name	$G(x)$	$g(x)$	$g'(x)$
Skew	$\frac{1}{3}x^3$	x^2	$2x$
Pow3	$\frac{1}{4}x^4$	x^3	$3x^2$
Gauß	$-e^{-\frac{x^2}{2}}$	$xe^{-\frac{x^2}{2}}$	$e^{-\frac{x^2}{2}} - x^2 e^{-\frac{x^2}{2}}$
Tanh	$\frac{1}{a} \log \cosh(ax)$	$\tanh(ax)$	$-a(\tanh(ax)^2 - 1)$
Pearson	—	$\frac{x-a}{b+cx+dx^2}$	$\frac{1}{dx^2+cx+b} + \frac{(a-x)(c+2dx)}{(dx^2+cx+b)^2}$
PolyL	$\sum_{i=0}^L a_i x^i$	$\sum_{i=1}^L i a_i x^{i-1}$	$\sum_{i=2}^L i(i-1) a_i x^{i-2}$

Nach den Kuhn-Tucker Bedingungen [13, 8] kann das Optimum von $E\{G(\mathbf{w}X'_w)\}$ unter der Randbedingung $E\{(\mathbf{w}X'_w)^2\} = \|\mathbf{w}\|^2 = 1$ an folgender Stelle gefunden werden:

$$E\{X'_w g(\mathbf{w}X'_w)\} - \beta \mathbf{w} = 0 \quad (12.39)$$

Die Lösung der Gleichung mittels der Newton Methode führt zur FastICA Iteration aus Algorithmus 19 [8].

$$\mathbf{w} \leftarrow E\{g(\mathbf{w}X'_w) X_w\} - E\{g'(\mathbf{w}X'_w)\} \mathbf{w} \quad (12.40)$$

Der dargestellte Algorithmus berechnet lediglich *eine* der unabhängigen Komponenten, so dass er für jede weitere Komponente erneut durchgeführt werden muss. Damit der Algorithmus nicht zu einer bereits gefundenen Komponente konvergiert, werden die gefundenen Komponenten $\mathbf{w}_1 X'_w, \mathbf{w}_2 X'_w, \dots$ nach jeder Iteration de-korrielt. HYVÄRINEN schlägt dazu unterschiedliche Methoden vor [8].

Deflation

Bei dem Verfahren *Deflation* wird jede Komponente nacheinander bestimmt. Wenn beispielsweise p unabhängige Komponenten $\mathbf{w}_1, \dots, \mathbf{w}_p$ bereits gefunden wurden, wird am Anfang jeder Iteration (Algorithmus 19) jede Projektionen $\mathbf{w}_{p+1} \mathbf{w}'_k \mathbf{w}_k, k = 1, \dots, p$ von \mathbf{w}_{p+1} subtrahiert und im Anschluss \mathbf{w}_{p+1} erneut normiert.

$$\begin{aligned} \mathbf{w}_{p+1} &= \mathbf{w}_{p+1} - \mathbf{w}_{p+1} W'_{1p} W_{1p} \\ \mathbf{w}_{p+1} &= \frac{\mathbf{w}_{p+1}}{\|\mathbf{w}_{p+1}\|} \\ \text{mit } W_{1p} &= [\mathbf{w}_1 \mathbf{w}_2 \dots \mathbf{w}_p] \end{aligned} \quad (12.41)$$

Symmetrische Dekorrelation

Im symmetrischen Verfahren (Algorithmus 20) werden alle unabhängigen Komponenten gleichzeitig berechnet. Sei W_a die Matrix, welche alle unabhängigen Komponenten (Vektoren) \mathbf{w} enthält. Die symmetrisch Dekorrelation wird dann durch Gleichung 12.42 berechnet.

$$W_a^* = \sqrt{(W_a W_a')^{-1}} W_a \quad (12.42)$$

Die zufällige Startmatrix W_a muss dabei so gewählt werden, dass alle Komponenten orthogonal sind, was zum Beispiel durch eine Singulärwertzerlegung einer Zufallsmatrix erzielt werden kann (siehe Kapitel 12.2).

- 1 W_a = unitäre $n_f \times n_{IC}$ Matrix der SVD einer Zufallsmatrix der Größe $n_f \times n_{IC}$
- 2 solange \mathbf{W}_a nicht konvergiert tue
- 3 $W_a = \sqrt{(W_a W_a')^{-1}} W_a$
- 4 $W_a \leftarrow E\{g(W_a X_w') X_w\} - E\{g'(W_a X_w')\} W_a$
- 5 Ende
- 6 $W = M_w W_a'$

Algorithmus 20 : Symmetrische Dekorrelation

Die Projektion der Daten $X = X_z + E\{X\}$ auf die n_{IC} unabhängigen Komponenten wird im Anschluss durch Gleichung 12.43 ermöglicht.

$$X_{IC} = X_z W' + E\{X\} W' \quad (12.43)$$

Literaturverzeichnis

1. Berrar, D.P., Dubitzky, W., Granzow, M. (eds.): *A Practical Approach to Microarray Data Analysis*. Springer (2002) 396
2. Cover, T.M., Thomas, J.: *Elements of information theory*. Wiley, Wiley Series in Telecommunications and Signal Processing (2006) 408
3. Gunn, S.R.: *Support Vector Machines for Classification and Regression*. Tech. rep., University of Southampton (1998). URL <http://users.ecs.soton.ac.uk/srg/publications/pdf/SVM.pdf>. (abgerufen 11/2016) 268, 400
4. Huber, P.: *Projection pursuit*. The Annals of Statistics **13**(2), pp. 435–475 (1985) 408
5. Hyvärinen, A.: *New approximations of differential entropy for independent component analysis and projection pursuit*. In: Proceedings of the 1997 conference on Advances in Neural Information Processing Systems 10, NIPS '97, pp. 273–279. MIT Press, Cambridge, MA, USA (1998) 409
6. Hyvärinen, A.: *Fast and Robust Fixed-Point Algorithms for Independent Component Analysis*. IEEE Transactions on Neural Networks **10**(3), pp. 626–634 (1999) 410
7. Hyvärinen, A., Oja, E.: *A fast fixed point algorithm for independent component analysis*. Neural Computation **9**(7), pp. 1483–1492 (1997) 410
8. Hyvärinen, A., Oja, E.: *Independent Component Analysis: Algorithms and Applications*. Neural Networks **13**(4-5), pp. 411–430 (2000) 403, 405, 406, 407, 408, 409, 410, 411
9. Hyvärinen, A.: *FastICA: Publications* (2013). URL <http://www.cs.helsinki.fi/u/ahyvarin/papers/fastica.shtml>. (abgerufen 11/2016) 403, 410
10. Jones, M., Sibson, R.: *What is projection Pursuit?* Journal of the Royal Statistical Society **150**, pp. 1–36 (1987) 408
11. Karhunen, J., Hyvärinen, A., Oja, E.: *Independent Component Analysis (ICA) and Blind Source Separation (BSS)* (2013). URL <http://research.ics.aalto.fi/ica/newindex.shtml>. (abgerufen 11/2016) 403, 410
12. Keralapura, M., Pourfathi, M., Sirkeci-Mergen, B.: *Impact of Contrast Functions in Fast-ICA on Twin ECG Separation*. IAENG International Journal of Computer Science **38** (2011) 410, 411
13. Luenberger, D.G.: *Optimization by Vector Space Methods*. Wiley Professional (1969) 411
14. Papoulis, A.: *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill (1990) 408
15. Pearson, K.: *On Lines and Planes of Closest Fit to Systems of Points in Space*. Philosophical Magazine **2**(11), p. 559–572 (1901). URL <http://stat.smmu.edu.cn/history/pearson1901.pdf>. (abgerufen 11/2016) 396
16. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press (2007) 199, 203, 220, 254, 265, 276, 277, 281, 285, 307, 334, 381, 398, 400
17. Schölkopf, B., Müller, K.R., Smola, A.: *Nonlinear component analysis as a kernel eigenvalue problem*. Neural Computation **10**, pp. 1299–1319 (1998) 400, 401
18. Schölkopf, B., Müller, K.R., Smola, A.: *Lernen mit Kernen*. Informatik Forschung und Entwicklung **14**, pp. 154–163 (1999) 400, 401
19. Schölkopf, B., Smola, A.J., Müller, K.R.: *Kernel Principal Component Analysis*. In: B. Schölkopf, C.J.C. Burges., A.J. Smola (eds.) *Advances in Kernel Methods*, pp. 327–352. MIT Press, Cambridge, MA, USA (1999) 400
20. Wall, M.E., Rechtsteiner, A., Rocha, L.M.: *Singular Value Decomposition and Principal Component Analysis*. A Practical Approach to Microarray Data Analysis (2002). URL <http://arxiv.org/abs/physics/0208101>. (abgerufen 11/2016) 396
21. Wikipedia: *Faktorenanalyse*. URL <http://de.wikipedia.org/wiki/Faktorenanalyse>. (abgerufen 2012,2016) 395
22. Wikipedia: *Hauptkomponentenanalyse*. URL <http://de.wikipedia.org/wiki/Hauptkomponentenanalyse>. (abgerufen 2012,2016) 395

23. Wikipedia: *Kernel Principal Component Analysis*. URL http://en.wikipedia.org/wiki/Kernel_principal_component_analysis. (abgerufen 2012,2016) 400
24. Wikipedia: *Maximum Likelihood Methode*. URL <http://de.wikipedia.org/wiki/Maximum-Likelihood-Methode>. (abgerufen 2012,2016) 409
25. Wikipedia: *Pseudoinverse*. URL <http://de.wikipedia.org/wiki/Pseudoinverse>. (abgerufen 2012,2016) 400
26. Wikipedia: *Singular value decomposition*. URL http://en.wikipedia.org/wiki/Singular_value_decomposition. (abgerufen 2012,2016) 398
27. Wikipedia: *Transinformation*. URL <http://de.wikipedia.org/wiki/Transinformation>. (abgerufen 2013,2016) 409
28. Wikipedia: *Zentraler Grenzwertsatz*. URL http://de.wikipedia.org/wiki/Zentraler_Grenzwertsatz. (abgerufen 2013,2016) 407

Kapitel 13

Sensitivitätsanalyse

13.1 Einleitung

Mathematische Modelle von physikalischen, medizinischen oder anderen Systemen basieren meist auf einer Vielzahl von komplexen, nichtlinearen und gekoppelten Gleichungssystemen. Ein Grundpfeiler für die Analyse dieser Systeme ist das Verständnis von Einfluss der Varianzen aller Faktoren \mathbf{x} (Eingangsvariablen) auf die Varianz der betrachteten Systemantwort y (Ausgangsgröße). Unter dem Begriff *Sensitivitätsanalyse* (SA) werden Verfahren zusammengefasst, die genau für diesen gesuchten Zusammenhang Kenngrößen ermitteln. Grundsätzlich wird dabei zwischen drei Bereichen der Sensitivitätsanalyse unterschieden [12]:

Faktor Screening: Durch ein Faktor Screening wird der *qualitative* Einfluss von Faktoren (Eingangsvariablen) auf eine Ausgangsvariable ermittelt. Dieses Verfahren wird hauptsächlich zur Unterscheidung von signifikanten und nicht signifikanten Faktoren eingesetzt, wobei keine quantitativen Kenngrößen ermittelt werden. Da in der Praxis meist qualitative Vergleiche der Faktoren benötigt werden, wird im Weiteren keine spezielle Betrachtung dieses Bereichs dargestellt.

Lokale Sensitivitätsanalyse: Die lokale Sensitivitätsanalyse untersucht den Einfluss von Faktoren bei einem bestimmten Funktionswert der Ausgangsvariable y beziehungsweise bei spezieller Kombination der Eingangsvariablen \mathbf{x}_i (zum Beispiel beim einem lokalen Optimum). Grundsätzlich wird dabei untersucht, welche Auswirkungen *kleine* Änderungen der Faktoreinstellungen auf die Ausgangsvariablen haben. Mit diesen Verfahren werden beispielsweise Stabilitätsanalysen (Robustheit) für ausgewählte Faktorkombinationen ermöglicht.

Globale Sensitivitätsanalyse: Die globale Sensitivitätsanalyse ermittelt den Einfluss von Faktoren bei Variation über ihren gesamten Definitionsbereich. Die hierzu eingesetzten Verfahren haben sich in der Praxis besonders zum besseren Verständnis der Signifikanz einzelner Faktoren in einem Modell sowie dem direkten Vergleich verschiedener Faktoren bewährt. Der grundsätzliche Ablauf einer globalen Sensitivitätsanalyse ist in Abbildung 13.1 dargestellt [12]. Verschiedene Faktoren (x_1, \dots, x_{n_f}) weisen gleiche oder unterschiedliche Verteilungen auf. Die Varianz

führt in Abhängigkeit vom verwendeten deterministischen Modell zu einer Varianz der Ausgangsvariable y . Durch die globale Sensitivitätsanalyse wird der Anteil der Varianz v_j von y bestimmt, welcher durch den Faktor x_j verursacht wird. Damit ist ein Vergleich der Signifikanz verschiedener Faktoren für die Ausgangsvariable y möglich.

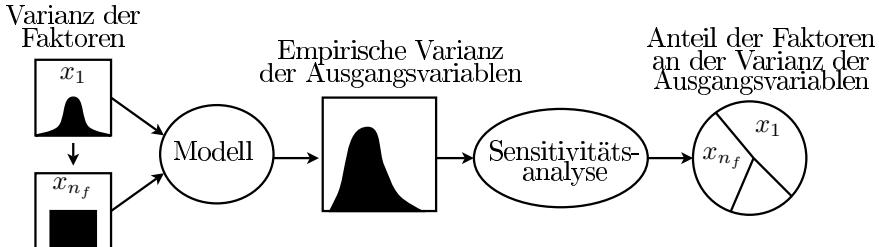


Abb. 13.1 Globale Sensitivitätsanalyse

13.2 Sensitivitätsanalyse bei Linearen Modellen

In vielen Bereichen werden auch heute noch lineare Approximationsmodelle der folgenden Grundform eingesetzt:

$$y_i = b_0 + \sum_{j=1}^{n_f} b_j x_{ij} + \varepsilon_i \quad (13.1)$$

Für diese Modelltypen gibt es eine Vielzahl von Methoden, den Einfluss eines Faktors x_j auf die Ausgangsvariable y zu ermitteln. Die im Folgenden dargestellten Sensitivitätsanalysen basieren grundsätzlich auf reellen Werten für x_{ij} und y . Zur Stabilisierung der Algorithmen (zum Beispiel bei nicht Gauß-verteilten Daten oder bei Ausreißern) können die Daten durch Ihren *Rang* ersetzt werden. Dazu werden die Daten im ersten Schritt aufsteigend nach ihrer Größe sortiert und anschließend durchgehend nummeriert. Beispiel: $y_3 \leq y_1 \leq y_4 \dots \rightarrow y_3 = 1, y_1 = 2, y_4 = 3 \dots$. Ob die Verwendung des Rangs sinnvoll ist, muss in jedem Anwendungsfall separat entschieden werden, da durch dieses Verfahren Informationen (z.B. die Verteilungsform) verloren gehen.

13.2.1 Normierte Regressionskoeffizienten

Ein- und Ausgangsvariablen eines Systems können mittels der jeweiligen Standardabweichung (σ_y, σ_{x_j}) und des Mittelwerts (\bar{y}, \bar{x}_j) nach Gleichung 13.2 normiert wer-

den.

$$x_{ij,norm} = \frac{\bar{x}_j - x_{ij}}{\sigma_{x_j}}, \quad y_{i,norm} = \frac{\bar{y} - y_i}{\sigma_y}, \quad i = 1, \dots, n_r, \quad j = 1, \dots, n_f \quad (13.2)$$

Die Koeffizienten des Regressionsmodells, welche auf den normierten Variablen basieren, werden als normierte Regressionskoeffizienten (*SRC: standardized regression coefficient*) bezeichnet und sind ein Maß für den Einfluss jedes Faktors x_j auf die Ausgangsvariable y . Je größer der Betrag des SRC-Werts ist desto größer ist der Einfluss des zugehörigen Faktors. Liegt bereits ein Regressionsmodell basierend auf nicht normierten Variablen vor, so können die normierten Regressionskoeffizienten direkt mit Gleichung 13.3 ermittelt werden [6]:

$$b_{j,norm} = b_j \frac{\sigma_{x_j}}{\sigma_y} \quad (13.3)$$

13.2.2 Partialsumme der Quadrate

Die klassische ANOVA (siehe Kapitel 4.5) unterteilt die Varianz des linearen Regressionsmodells in die folgenden Bestandteile (*SST: total sum of squares, SSE: error sum of squares, SSR: regression sum of squares*, siehe Kapitel 9.3.1) :

$$SST = SSR + SSE \quad (13.4)$$

$$SST = \sum_{i=1}^{n_r} (y_i - \bar{y})^2, \quad SSE = \sum_{i=1}^{n_r} (y_i - \hat{y}_i)^2, \quad SSR = \sum_{i=1}^{n_r} (\hat{y}_i - \bar{y})^2 \quad (13.5)$$

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} \quad (13.6)$$

Der Determinationskoeffizient R^2 ist dabei ein Maß für die Güte des Modells. In einigen Ausnahmefällen, wenn zum Beispiel der konstante Regressionsterm vernachlässigt wird, können negative R^2 berechnet werden [1]. In diesen Fällen ist es sinnvoller mit der Pearson Korrelation zu arbeiten (Kapitel 11.1). $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}$ seien zwei Untergruppen der betrachteten Faktoren, wobei kein Faktor in beiden Gruppen gleichzeitig enthalten ist. $SSR(\mathbf{x}_{(1)})$ bzw. $SSR(\mathbf{x}_{(1)}, \mathbf{x}_{(2)})$ geben jeweils den SSR-Wert für Regressionsmodelle an, welche nur die Faktoren aus der Untergruppe $\mathbf{x}_{(1)}$ bzw. aus beiden Untergruppen enthalten. Der Wert $SSR(\mathbf{x}_{(2)}|\mathbf{x}_{(1)})$ misst die Verringerung des Modellfehlers für den Fall, dass einem Regressionsmodell, basierend auf den Faktoren der Gruppe $\mathbf{x}_{(1)}$, die Faktoren aus der zweiten Gruppe $\mathbf{x}_{(2)}$ hinzugefügt werden. Dabei gilt:

$$SSR(\mathbf{x}_{(2)}|\mathbf{x}_{(1)}) = SSR(\mathbf{x}_{(1)}, \mathbf{x}_{(2)}) - SSR(\mathbf{x}_{(1)}) = SSE(\mathbf{x}_{(1)}) - SSE(\mathbf{x}_{(1)}, \mathbf{x}_{(2)}) \quad (13.7)$$

Besteht nun die Gruppe $\mathbf{x}_{(2)}$ lediglich aus einem Faktor x_j und die Gruppe $\mathbf{x}_{(2)}$ aus allen restlichen Faktoren (\mathbf{x}_{-j}) kann durch

$$p_j = \text{SSR}(x_j | \mathbf{x}_{-j}) = \text{PSS}_j \quad (13.8)$$

die Wichtigkeit des Faktors x_j auf die Ausgangsvariable y bestimmt werden, wobei p_j als '*partial sum of squares*' (PSS) bezeichnet wird [6].

13.2.3 Partieller Determinationskoeffizient

Basierend auf den Gleichungen aus Kapitel 13.2.2 wird der partielle Determinationskoeffizient (*CPD: coefficient of partial determination*) nach Gleichung 13.9 definiert:

$$\text{CPD}_{\mathbf{x}_{(2)}|\mathbf{x}_{(1)}} = \frac{\text{SSR}(\mathbf{x}_{(2)}|\mathbf{x}_{(1)})}{\text{SSE}(\mathbf{x}_{(1)})} = \frac{\text{SSE}(\mathbf{x}_{(1)}) \text{SSE}(\mathbf{x}_{(1)}, \mathbf{x}_{(2)})}{\text{SSE}(\mathbf{x}_{(1)})} \quad (13.9)$$

Der CPD ist ebenfalls ein Maß für die Wichtigkeit einer Faktorgruppe $\mathbf{x}_{(2)}$ für die Ausgangsvariable y .

13.2.4 Predictive Error Sum of Squares

Mit dem Begriff *Predictive Sum of Squares* (PRESS) wird in der Literatur der Kennwert aus Gleichung 13.10 bezeichnet, welcher ein Maß für die Güte des verwendeten Modells darstellt (siehe auch Kapitel 9.3.1).

$$\text{PRESS} = \sum_{i=1}^{n_r} (y_i - \hat{y}_{-i})^2 \quad (13.10)$$

Dabei steht \hat{y}_{-i} für die Vorhersage von y_i , wenn ein Regressionsmodell verwendet wird, welches ohne den i^{ten} Datenpunkt, also mit $n_r - 1$ Datenpunkten trainiert wird. Wird nun der PRESS ohne den Faktor x_j also mit $n_f - 1$ Faktoren erstellt, so kann der Kennwert PRESS_{-j} als Maß für die Wichtigkeit des Faktors x_j für die Ausgangsvariable y interpretiert werden. Dabei steht ein höherer PRESS_{-j} Wert für wichtige Faktoren, da durch die Vernachlässigung eines wichtigen Faktors die Güte des Regressionsmodells stärker abnimmt als bei Vernachlässigung eines unwichtigen Faktors [6].

13.2.5 Partielle Korrelationsfaktoren

Die Korrelation zwischen y und einem Faktor x_k wird durch den *Partial Correlation Coefficient* (PCC) beschrieben. Zur Berechnung dieses Kennwerts werden Regressionsmodelle für y und x_k basierend auf allen anderen Faktoren x_{-k} erstellt.

$$\hat{y}_i = c_o + \sum_{j \neq k} c_j x_{ij} \quad \text{und} \quad \hat{x}_{ik} = b_o + \sum_{j \neq k} b_j x_{ij}, \quad i = 1, \dots, n_f \quad (13.11)$$

Die Korrelation zwischen $(y_i - \hat{y}_i)$ und $(x_{ik} - \hat{x}_{ik})$ wird als Maß für die Korrelation zwischen x_k und y interpretiert [6]. Eine stärkere Korrelation steht dabei für einen wichtigeren Faktor x_k bezüglich y .

13.3 Sensitivitätsanalyse bei nichtlinearen Modellansätzen

Die Bedeutung von nichtlinearen Modellen, wie sie durch Neurale Netzwerke oder Support Vector Machines (Kapitel 9) erzeugt werden, nimmt stetig zu und die Anwendung ist in unterschiedlichen Bereichen bereits unverzichtbar geworden. Verfahren zur globalen Sensitivitätsanalyse, die auf linearen Modellen basieren, können bei nichtlinearen Zusammenhängen nicht sinnvoll eingesetzt werden, so dass auf Rangkorrelationen oder *Varianz basierten* Sensitivitätsanalysen zurückgegriffen werden muss [12].

13.3.1 Korrelationsverhältnis

Die Varianz der Ausgangsvariable y bezüglich eines Faktors x_j besteht aus zwei Bestandteilen [9, 11, 12]:

$$V(y) = V[E(y|x_j)] + E[V(y|x_j)] \quad (13.12)$$

Dabei wird der erste Term als Varianz bedingte Erwartung (*VCE: variance conditional expectation*) bezeichnet und ist ein brauchbares Maß für den Zusammenhang zwischen x_j und y . Basierend auf dem VCE wird das Korrelationsverhältnis (KV) (*CR: correlation ratio*) definiert:

$$KV_j = \frac{V[E(y|x_j)]}{V(y)} = \frac{VCE}{V(y)} ; \quad 0 \leq KV_j \leq 1 \quad (13.13)$$

Der Wert des Korrelationsverhältnisses für einen Faktor x_j (KV_j) steigt mit der Wichtigkeit von x_j auf die Varianz der Ausgangsvariable y . Bei Verwendung eines linearen Modells ist das Korrelationsverhältnis gleich dem Determinationskoeffizient aus Kapitel 13.2.2 ($R_j^2 = KV_j$).

Zur Abschätzung des Korrelationsverhältnisses schlägt MCKAY [9, 10] ein Verfahren basierend auf *Latin Hypercube Sampling* (LHS) aus Kapitel 8.3.3 vor. Dabei wird ein $wLHS_{n_r}$ verwendet, welches aus n_r Versuchsläufen und w Wiederholungen besteht. Die w Wiederholungen können dabei durch einfache Permutationen der einzelnen Spalten des ersten LHS erzeugt werden [12, 9]. Eine Abschätzung für $V(a)$ wird dann mit Gleichung 13.14 berechnet:

$$\widehat{V(y)} = \frac{1}{n_r w} \sum_{i=1}^{n_r} \sum_{k=1}^w (y_{ik} - \bar{y})^2 \text{ mit } \bar{y} = \frac{1}{n_r w} \sum_{i=1}^{n_r} \sum_{k=1}^w y_{ik} \quad (13.14)$$

MCKAY konnte zeigen, dass eine Abschätzung für KV durch Gleichung 13.15 gegeben ist[9]:

$$\widehat{KV(x_j)} = \frac{w \sum_{i=1}^{n_r} (\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{n_r} \sum_{k=1}^w (y_{ik} - \bar{y})^2} \quad (13.15)$$

\bar{y}_i bezeichnet dabei den Mittelwert aller Systemantworten y , bei dem der Faktor x_j einen bestimmten (*i – ten*) der n_r Werte aufweist. In jeder Wiederholung $k \in [1, \dots, n_w]$ wird damit genau ein Versuch gewählt und alle gewählten Faktorkombinationen weisen den gleichen x_j Wert auf. Die Abschätzung konvergiert mit steigender Anzahl der Wiederholungen zum wahren Wert des Korrelationsverhältnisses. Als Beispiel wird die Funktion in Gleichung 13.16 betrachtet, welche folgende approximierten CR Werte liefert: $CR \approx [0.385, 0.385, 0.103, 0]$

$$y = x_1 + x_2 + x_3^2 + x_2 x_4 \text{ mit } x_j \in [-1 \dots 1] \quad (13.16)$$

Importance Measure

HORA und IMAN führen zur Beurteilung des Einflusses von Faktor x_j auf die Ausgangsvariable y den Kennwert *Importance Measure* I_j ein [2, 9]:

$$I_j = \sqrt{V(y) - E[V(y|x_j)]} = \sqrt{U_j - [E(y)]^2} \quad (13.17)$$

Dabei ist U_j durch Gleichung 13.18 definiert:

$$U_j = \int [E(y|x_j)]^2 PDF(x_j) dx_j \quad (13.18)$$

PDF ist dabei die Wahrscheinlichkeitsdichte (*probability density function*) des Faktors x_j . Da der Term $E(y)$ konstant ist, kann die Wichtigkeit des Faktors x_j lediglich durch U_j beurteilt werden, wobei U_j bei steigender Bedeutung von x_j wächst. Zwischen KV und I_j besteht ein direkter Zusammenhang:

$$KV = \frac{I_j^2}{V(y)} \quad (13.19)$$

U_j kann durch Gleichung 13.20 abgeschätzt werden [9, 13]:

$$\widehat{U_j} = \frac{1}{n_r} \sum_{i=1}^{n_r} y_i (x_{-j}^a, x_j) y_i (x_{-j}^b, x_j) \quad (13.20)$$

Zur Berechnung von $y_i(x_{-j}^a, x_j)$ und $y_i(x_{-j}^b, x_j)$ werden dabei zwei unterschiedliche Testfelder verwendet, bei denen nur die Werte des Faktors x_j identisch sind. Die zur Berechnung des Korrelationsverhältnisses benötigten Funktionswerte $E(y)$ und $V(y)$ werden durch eine einfache Monte-Carlo-Simulation (Kapitel 8.3.1) oder direkt durch die vorhandenen Daten zur Berechnung von U_j abgeschätzt. Für die Approximation des Korrelationsverhältnisses gilt dann:

$$\widehat{KV}_j = \frac{\widehat{U}_j - \widehat{E}(y)}{\widehat{V}(y)}^2 \quad (13.21)$$

$$\text{mit } \widehat{E}(y) = \frac{1}{2n_r} \sum_{i=1}^{n_r} \left[y_i(x_{-j}^a, x_j) + y_i(x_{-j}^b, x_j) \right] = \bar{y} \quad (13.22)$$

$$\widehat{V}(y) = \frac{1}{2n_r} \left\{ \sum_{i=1}^{n_r} \left[y_i(x_{-j}^a, x_j) - \bar{y} \right]^2 + \left[y_i(x_{-j}^b, x_j) - \bar{y} \right]^2 \right\} \quad (13.23)$$

Mittlere Stichprobe

Eine weitere Berechnungsmethode des Korrelationsverhältnisses stellt REEDIJK vor (Mittlere Stichprobe) [9, 12]. Dabei werden im ersten Schritt $n_r^* = rn_r$ unterschiedliche Versuchsläufe durchgeführt. Anschließend werden die Versuchsläufe nach aufsteigenden Faktorstufen des zu untersuchenden Faktors x_j sortiert und in r gleiche Gruppen der Größe n_r aufgeteilt. Für jede Gruppe $k = 1 \dots r$ wird anschließend der Erwartungswert $E_k(y|x_j)$ ermittelt.

$$E_k(y|x_j) = \frac{\sum_{i=1}^{n_r} y_{j(k)}}{n_r} = E_k^{(j)} \quad (13.24)$$

Mit steigendem n_r konvergiert die Varianz der r Erwartungswerte zur *Varianz bedingten Erwartung* (VCE), mit dem das Korrelationsverhältnis (Gleichung 13.13) berechnet wird.

$$\widehat{VCE}_j = \frac{1}{r} \sum_{k=1}^r \left(E_k^{(j)} - \overline{E^{(j)}} \right)^2 \quad (13.25)$$

$$\widehat{KV}_j = \frac{\widehat{VCE}_j}{\widehat{V}(y)} \quad (13.26)$$

Robustheit und Ausreißer

Die Berechnung des Korrelationsverhältnisses wird stark durch Ausreißer beeinflusst. Zur Verbesserung der Robustheit schlagen HORA und IMAN [7] vor, die Ausgangsvariable y durch $\log(y)$ zu ersetzen. Damit folgt für das KV_{log} :

$$KV_{log_j} = \frac{V[E(\log(y)|x_j)]}{V(\log(y))} \text{ mit } 0 \leq KV_{log_j} \leq 1 \quad (13.27)$$

Eine Interpretation des Ergebnisses bezüglich der ursprünglichen Ausgangsvariable y ist schwierig. Eine Steigerung der Robustheit kann beim Korrelationsverhältnis durch die Verwendung des Rangs anstelle des Funktionswerts y erreicht werden. Hier gilt entsprechend zu Kapitel 13.2, dass nicht der *wahre* Anteil der Varianz, sondern lediglich ein qualitativer Vergleich der Varianzanteile verschiedener Faktoren bestimmt wird.

13.3.2 Sobol's Kennzahl

Zur Bestimmung der Sensitivität eines Faktors x_j oder mehrerer Faktoren $x_{j_1 \dots j_s}$ auf eine Systemantwort y wird diese nach *Sobol* im ersten Schritt in der ANOVA-Form (Gleichung 13.28) dargestellt. In dieser Form wird der Zusammenhang zwischen \mathbf{x} und y mittels separater Funktionen g für jede mögliche Kombination der Faktoren x_1, \dots, x_{n_f} sowie einem konstanten Term g_0 abgebildet [16]. Dabei steht g_{jkl} zum Beispiel für einen Summanden, der nur die Faktoren x_j, x_k und x_l enthält.

$$\begin{aligned} y(x_1, \dots, x_{n_f}) &= g_0 + \sum_{j=1}^{n_f} g_j(x_j) + \sum_{1 \leq j < k \leq n_f} g_{jk}(x_j, x_k) + \dots \\ &\quad \dots + g_{1 \dots n_f}(x_j, \dots, x_{n_f}) \\ g_0 &= \int_0^1 y(x_1, \dots, x_{n_f}) dx_1, \dots, x_{n_f} \end{aligned} \quad (13.28)$$

Zur Vereinfachung wird im Folgenden von einem normierten Faktorraum $[0, 1]^{n_f}$ ausgegangen. Weiterhin gilt in der ANOVA-Form, dass jede Funktion g bei Integration über den Faktorbereich Null ergibt (Gleichung 13.29). Somit sind alle Elemente von Gleichung 13.28 orthogonal, wodurch jeder Term g mittels eines Integrals über $y(\mathbf{x})$ dargestellt werden kann [16].

$$\int_0^1 g_{j_1 \dots j_s}(x_{j_1, \dots, j_s}) dx_k = 0, \text{ mit } k = j_1, \dots, j_s \quad (13.29)$$

$$\int y(\mathbf{x}) d\mathbf{x} = g_0 \quad (13.30)$$

$$\int y(\mathbf{x}) \prod_{k \neq i} dx_k = g_0 + g_i(x_i) \quad (13.31)$$

$$\int y(\mathbf{x}) \prod_{k \neq i,j} dx_k = g_0 + g_i(x_i) + g_j(x_j) + g_{ij}(x_i, x_j) \quad (13.32)$$

...

Die Varianz D auf Grund aller Faktoren $\mathbf{x} = [x_1, \dots, x_{n_f}]$ sowie $D_{j_1 \dots j_s}$ einer festen Kombination von s Faktoren ist dann wie in Gleichung 13.33 und 13.34 definiert [16].

$$D = \int y(\mathbf{x})^2 d\mathbf{x} - g_0^2 \quad (13.33)$$

$$D_{j_1 \dots j_s} = \int g_{j_1 \dots j_s} dx_{j_1}, \dots, dx_{j_s} \quad (13.34)$$

Grundsätzlich entspricht die Summe aller möglichen $D_{j_1 \dots j_s}$ genau D .

$$D = \sum D_{j_1 \dots j_s} \quad (13.35)$$

Der globale Sensitivitätsindex $S_{j_1 \dots j_s}$ der s Faktoren x_{j_1} bis x_{j_s} ist definiert durch [16].

$$S_{j_1 \dots j_s} = \frac{D_{j_1 \dots j_s}}{D} \quad (13.36)$$

$S_{j_1 \dots j_s}$ ist nicht negativ und die Summe aller globalen Sensitivitätsindizes ergeben Eins.

$$\sum S_{j_1 \dots j_s} = 1 \quad (13.37)$$

Im weiteren soll nicht nur der Sensitivitätsindex der festen Kombination von s Faktoren bestimmt werden, sondern der Sensitivitätsindex der gesamten Gruppe der s Faktoren, welche ebenfalls alle Untergruppen beziehungsweise Kombinationen der s Faktoren X_{j_1} bis X_{j_s} berücksichtigt. Die betrachtete Teilgruppe mit s Elementen ($1 \leq s < n_f$) wird zur Vereinfachung mit $\mathbf{u} = [x_{j_1}, \dots, x_{j_s}]$ bezeichnet. Alle anderen Faktoren, die nicht zu \mathbf{u} gehören werden in der Teilgruppe \mathbf{w} zusammengefasst. Die Varianz D_u der Teilgruppe \mathbf{u} wird durch die Summe aller Varianzen ermittelt, die eine Kombination der Faktoren aus \mathbf{u} enthalten (Gleichung 13.38).

$$D_u = \sum_{i_1 \dots i_k \in \mathbf{u}} D_{i_1 \dots i_k} \quad (13.38)$$

Die totale Varianz D_u^{tot} berücksichtigt hingegen alle Faktorkombinationen bei denen mindestens ein Faktor von \mathbf{u} enthalten ist und wird nach Gleichung 13.39 bestimmt.

$$D_u^{tot} = D - D_w \quad (13.39)$$

Die Sensitivitätsindizes S_u und S_u^{tot} sind dann durch die Gleichungen 13.40 und 13.41 definiert und es gilt $0 \leq S_u \leq S_u^{tot} \leq 1$ sowie $S_u^{tot} = 1 - S_w$.

$$S_u = \frac{D_u}{D} \quad (13.40)$$

$$S_u^{tot} = \frac{D_u^{tot}}{D} = \frac{D - D_w}{D} \quad (13.41)$$

Ein Wert von $S_u = S_u^{tot} = 0$ bedeutet, dass $y(\mathbf{x})$ nicht von den Faktoren aus \mathbf{u} abhängt und im Gegenzug $S_u = S_u^{tot} = 1$, dass $y(\mathbf{x})$ komplett durch die Faktoren aus \mathbf{u} beschrieben wird. Sobol zeigt in [16] ein anschauliches Beispiel mit drei Faktoren x_1 bis x_3 zum besseren Verständnis.

$$\begin{aligned} \mathbf{u} &= \{x_1\} \text{ und } \mathbf{w} = \{x_2, x_3\} \\ S_u &= S_1 \\ S_u^{tot} &= S_1 + S_{12} + S_{13} + S_{123} = 1 - S_{23} \end{aligned} \quad (13.42)$$

$$\begin{aligned} \mathbf{u} &= \{x_1, x_2\} \text{ und } \mathbf{w} = \{x_3\} \\ S_u &= S_1 + S_2 + S_{12} \\ S_u^{tot} &= S_1 + S_2 + S_{12} + S_{13} + S_{23} + S_{123} = 1 - S_3 \end{aligned}$$

Zur Berechnung der Varianzen und Sensitivitätsindizes können im einfachsten Fall zwei Datensätze $X^{(1)}$ und $X^{(2)}$ mit n_r zufällig (MonteCarlo) verteilten Faktorkombinationen $\mathbf{x}_j^{(1)}$ und $\mathbf{x}_j^{(2)}$ verwendet werden. Wenn n_r groß genug ist, gilt für die Approximation der Varianzen:

$$D \approx \frac{1}{n_r} \sum_{j=1}^{n_r} y(\mathbf{x}_j)^2 - \bar{y}^2 \quad (13.43)$$

$$D_u^{tot} \approx \frac{1}{2n_r} \sum_{j=1}^{n_r} \left[y(\mathbf{x}_j) - y\left(\mathbf{x}_{j(u)}^{(1)}, \mathbf{x}_{j(w)}^{(2)}\right) \right]^2 \quad (13.44)$$

$$D_w^{tot} \approx \frac{1}{2n_r} \sum_{j=1}^{n_r} \left[y(\mathbf{x}_j) - y\left(\mathbf{x}_{j(u)}^{(2)}, \mathbf{x}_{j(w)}^{(1)}\right) \right]^2 \quad (13.45)$$

$$D_u = D - D_w^{tot} \quad (13.46)$$

$$D_w = D - D_u^{tot} \quad (13.47)$$

$y\left(\mathbf{x}_{j(u)}^{(1)}, \mathbf{x}_{j(w)}^{(2)}\right)$ bedeutet dabei, dass die Faktoren der Teilgruppe \mathbf{u} aus dem Datensatz $X^{(1)}$ und die Faktoren der Teilgruppe \mathbf{w} aus dem Datensatz $X^{(2)}$ entnommen werden. Alternativ können auch folgende Gleichungen verwendet werden:

$$g_0 = \bar{y} \approx \frac{1}{n_r} \sum_{j=1}^{n_r} y\left(\mathbf{x}_j^{(1)}\right) \quad (13.48)$$

$$D_u \approx \frac{1}{n_r} \sum_{j=1}^{n_r} \left[y\left(\mathbf{x}_j^{(1)}\right) y\left(\mathbf{x}_{j(u)}^{(1)}, \mathbf{x}_{j(w)}^{(2)}\right) \right] - g_0^2 \quad (13.49)$$

$$D_w \approx \frac{1}{n_r} \sum_{j=1}^{n_r} \left[y\left(\mathbf{x}_j^{(1)}\right) y\left(\mathbf{x}_{j(u)}^{(2)}, \mathbf{x}_{j(w)}^{(1)}\right) \right] - g_0^2 \quad (13.50)$$

$$D \approx \frac{1}{n_r} \sum_{j=1}^{n_r} y\left(\mathbf{x}_j^{(1)}\right)^2 - g_0^2 \quad (13.51)$$

Die Verwendung von Sobol's Kennzahlen ergibt einen guten Einblick in die Zusammenhänge von Faktoren und Ausgangsvariablen bei nichtlinearen und linearen Systemen. Bei steigender Faktorenanzahl steigt der Rechenaufwand stark an.

13.3.3 FAST (Fourier Amplitude Sensitivity Test)

Der *Fourier Amplitude Sensitivity Test* (FAST) wurde zuerst in den 70^{er} Jahren von CUKIER et al. dargestellt [3, 4, 5, 8]. Gegenüber Sobol's Kennzahlen (Kapitel 13.3.2) zeichnet es sich besonders durch einen effizienteren Rechenalgorithmus aus. Die mehrdimensionale Integration im Verfahren von Sobol wird durch eine eindimensionale Analyse ersetzt. Dazu wird der n_f -dimensionale Faktorraum entlang einer einzelnen vorgegebenen Kurve analysiert. Die Kurve wird dabei durch eine einzelne Laufvariable s eindeutig definiert. Die ursprünglichen Faktorstufen werden dazu durch die Gleichung 13.52 ersetzt [14]:

$$x_j = G_j(\sin[\omega_j s]) \quad -\infty \leq s \leq \infty \quad (13.52)$$

ω_j sind unterschiedliche und *geschickt* ausgewählte Frequenzen, die den Faktoren $1, \dots, n_f$ zugeordnet sind. G_j sind vorgegebene Transformationsfunktionen, auf die später genauer eingegangen wird. Wenn nun s variiert wird, ändern sich alle Faktoren x_1, \dots, x_{n_f} gleichzeitig entlang einer Kurve durch den Faktorraum. Jeder Faktor oszilliert dabei periodisch entsprechend seiner zugeordneten Frequenz ω_j . Die zu analysierende Ausgangsvariable y zeigt dabei je nach Abhängigkeit von den verschiedenen Faktoren unterschiedlich starke periodische Oszillationen bei den gewählten Frequenzen ω_j . Bei einer starken Abhängigkeit zwischen dem Faktor x_j und y würde eine Frequenzanalyse der Ausgangsvariable y eine stärkere Amplitude bei der Frequenz ω_j und deren harmonischen Schwingungen aufweisen als bei Frequenzen unwichtiger Faktoren. Die Amplitude der Frequenzen und deren harmonische Schwingungen kann dadurch als Maß für die Sensitivität zwischen den Faktoren und y verwendet werden.

Eine genaue Berechnung der Sensitivität ist nur dann möglich, wenn die verwendete Kurve, mit der der Faktorraum durchlaufen wird, nahe an jedem möglichen Punkt

des Faktorraums entlang läuft. Dies wird nur dann erreicht, wenn für jede gewählte Frequenz ω_j gilt, dass sie nicht durch eine Linearkombination (mit ganzen Zahlen) der anderen verwendeten Frequenzen dargestellt werden kann. Ist dies der Fall ist die Analyse der Ausgangsvariable y durch eine Auswertung entlang der vorgegebenen Kurve durchgeführt. Es gilt beispielsweise für den globalen mit r potenzierten Mittelwert von y :

$$\bar{y}^r = \lim_{t \rightarrow \infty} \frac{1}{2t} \int_{-t}^t y^r(x_1(s), \dots, x_{n_f}(s)) ds = \lim_{t \rightarrow \infty} \frac{1}{2t} \int_{-t}^t y^r(s) ds \quad (13.53)$$

CUKIER et al. [3] zeigt, dass bei Verwendung von positiven ω_j eine Betrachtung der Funktion $y(s)$ zwischen $-\pi$ und π ausreicht. Daraus folgt für die Berechnung von \bar{y} und \hat{D} :

$$\bar{y}^r = \frac{1}{2\pi} \int_{-\pi}^{\pi} y^r(s) ds \quad (13.54)$$

$$\hat{D} = \frac{1}{2\pi} \int_{-\pi}^{\pi} y^2(s) ds - \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} y(s) ds \right]^2 \quad (13.55)$$

Die Funktion $y(s)$ wird im folgenden Schritt als Fourier Reihe dargestellt:

$$y(s) = \sum_k [A_k \cos(ks) + B_k \sin(ks)] \quad k \in \mathbb{Z} = \{-\infty, \dots, -1, 0, 1, +\infty\} \quad (13.56)$$

$$\begin{aligned} \text{mit } A_k &= \frac{1}{2\pi} \int_{-\pi}^{\pi} y(s) \cos(ks) ds \\ \text{und } B_k &= \frac{1}{2\pi} \int_{-\pi}^{\pi} y(s) \sin(ks) ds \end{aligned} \quad (13.57)$$

Das Spektrum der Fourier Reihe ist nun folgendermaßen definiert: $\Lambda_k = A_k^2 + B_k^2$

Da $y(s)$ eine reelle Funktion ist, gilt weiterhin $A_k = A_{-k}$, $B_k = B_{-k}$ und $\Lambda_k = \Lambda_{-k}$. Die gesuchten Werte für \hat{D} und \hat{D}_j werden im Anschluss durch die Analyse des Frequenzspektrums über den gesamten Frequenzbereich $[\hat{D}]$ beziehungsweise nur an der Frequenz ω_j und den harmonischen Frequenzen $k\omega_j$ ($k = 1, 2, 3, \dots$) $[\hat{D}_j]$ bestimmt.

$$\hat{D} = \sum_{k \in \mathbb{Z}_{-0}} \Lambda_k = 2 \sum_{k=1}^{+\infty} \Lambda_k \quad \text{mit } \mathbb{Z}_{-0} = \mathbb{Z} - \{0\} \quad (13.58)$$

$$\hat{D}_j = \sum_{k \in \mathbb{Z}_{-0}} \Lambda_{k\omega_j} = 2 \sum_{k=1}^{+\infty} \Lambda_{k\omega_j} \quad (13.59)$$

Die Sensitivität ist entsprechend der Sobol Kennzahlen (Kapitel 13.3.2) definiert und entspricht der Sensitivität erster Ordnung nach Sobol, welche den Haupteffekt von Faktor x_j auf die Ausgangsvariable y beschreibt [14, 15]:

$$\hat{S}_j^{FAST} = \frac{\hat{D}_j}{\hat{D}} \quad (13.60)$$

eFAST

Zur Ermittlung des totalen Sensitivitätsindex erweitert SALTELLI FAST zum erweiterten (extended) FAST (eFAST) [14]. Betrachtet werden dazu alle Frequenzen, die nicht zu den Frequenzen ω_j und den dazugehörigen harmonischen Frequenzen $k\omega_j$ gehören ($k \in \mathbb{N}$ mit $k > 0$). Diese Frequenzen enthalten alle Abhängigkeiten, die nicht den Haupteffekten der Faktoren x_1, \dots, x_{n_f} zugeordnet werden können. Folglich können diese Frequenzen nur durch Interaktionen der Faktoren erzeugt werden ($D - \sum D_j$).

Zur Abschätzung des Gesamteffekts eines Faktors x_j wird ω_j eine deutlich höhere Frequenz zugewiesen als allen anderen Faktoren [14]. Die Amplituden der harmonischen Frequenzen $k\omega_j$ streben bei steigendem $k = 1, 2, 3, 4, \dots$ schnell gegen Null. Für die Abschätzung des Effekts eines Faktors sind nur die ersten harmonischen Frequenzen zu berücksichtigen. Da die Frequenz ω_j deutlich größer ist als die Frequenzen ω_{-j} , kann der Effekt von allen Faktoren ohne x_j durch Betrachtung der Frequenzen bis $\omega_j/2$ abgeschätzt werden. Bis $\omega_j/2$ können Interferenzen mit dem Faktor x_j , also der Frequenz ω_j , ausgeschlossen werden. Allgemein wird der Effekt aller Faktoren jeder Ordnung ohne x_j mit Gleichung 13.61 abgeschätzt [14, 12].

$$\hat{D}_{-j} = 2 \sum_{k=1}^{\omega_j/2} \Lambda_k \quad (13.61)$$

Der totale Sensitivitätsindex ist dann durch Gleichung 13.62 definiert.

$$\widehat{S}_{T_j}^{FAST} = 1 - \frac{\hat{D}_{-j}}{\hat{D}} \quad (13.62)$$

Transformationsfunktion

Die Wahl der Transformationsfunktionen G_j und der Frequenzen ω_j kommt eine hohe Bedeutung bei der Bestimmung der Sensitivitätswerte zu, wobei in der Literatur unterschiedliche Funktionen für G_j vorgeschlagen werden [3, 4, 8, 14]. SALTELLI verwendet die folgende Funktion, die im Bereich $[-\pi, \pi]$ definiert ist und den Bereich $[0, 1]$ oszillierend durchläuft.

$$x_j = \frac{1}{2} + \frac{\arcsin(\sin[\omega_j s])}{\pi} \quad (13.63)$$

Durch die zwei Frequenzen $\{\omega_1, \omega_2\} = \{11, 31\}$ wird ein zwei-dimensionaler Raum wie in den Abbildung 13.2 durchlaufen. Zur Analyse der Sensitivität werden entlang der Kurve beliebig viele Datenpunkte gewählt (rechts).

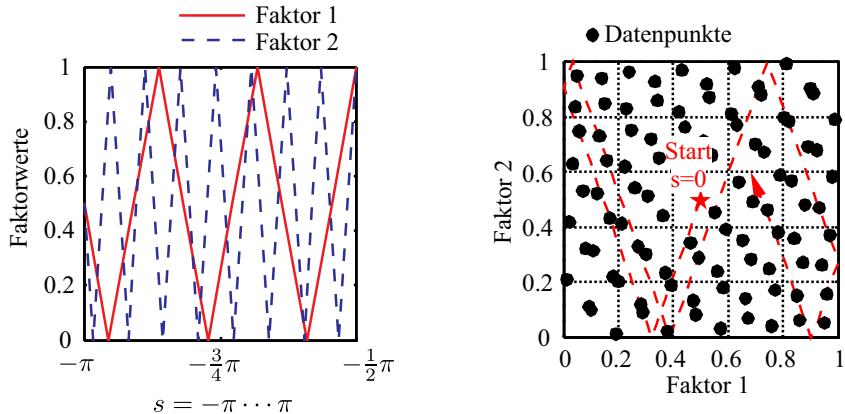


Abb. 13.2 Faktortransformation, $\omega_{1,2} = (11, 31)$

Wahl der Frequenzen

Zur Auswahl der Frequenzen und zur Bestimmung von $S_{T_j}^{FAST}$ werden folgende Regeln vorgeschlagen [14].

$$\omega_j = \left\lfloor \frac{n_r - 1}{2M} \right\rfloor \quad (13.64)$$

Die Gaußklammer $\lfloor \cdot \rfloor$ bezeichnet dabei die größte ganze Zahl, die kleiner oder gleich der Zahl in der Gaußklammer ist. Weiterhin ist M eine ganze Zahl größer Null und gibt die Anzahl der berücksichtigten harmonischen Schwingungen an. Typische Werte für M liegen bei $M \geq 4$. Die Frequenzen der übrigen Faktoren werden mit einer der folgenden Berechnungsvorschriften bestimmt (Gleichung 13.65 und 13.66).

$$\omega_{-j} = \begin{cases} 1 + \left\lfloor \frac{\omega_{\max}}{n_f - 2} k \right\rfloor & \text{wenn } \omega_{\max} \geq n_f - 1 \\ (k \bmod \omega_{\max}) + 1 & \text{wenn } \omega_{\max} < n_f - 1 \end{cases} \quad \text{mit } k = 0 \dots (n_f - 2) \quad (13.65)$$

$$\omega_{-j} = \begin{cases} \left\lfloor 1 + \frac{\omega_{\max}}{n_f - 2} k \right\rfloor & \text{wenn } \omega_{\max} \geq n_f - 1 \\ (k \bmod \omega_{\max}) + 1 & \text{wenn } \omega_{\max} < n_f - 1 \end{cases} \quad \text{mit } k = 0 \dots (n_f - 2) \quad (13.66)$$

Als Beispiel wird auch hier die Funktion aus Gleichung 13.16 analysiert. x_1 und x_3 weisen nur Haupteffekte auf, während x_2 und x_4 lediglich eine Interaktion aufweisen. Zur Untersuchung der Sensitivität wird $n_r = 10000$ und $M = 4$ gewählt, wodurch sich eine Frequenz von $\omega_j = 1249$ für den zu untersuchenden Faktor x_j ergibt. Für die übrigen Frequenzen werden folgende Frequenzen ermittelt: $\omega_{-j} = \{1, 78, 156\}$. Die Sensitivitätsanalyse mittels eFAST ergibt die in Abbildung 13.3 dargestellten Haupt- und Interaktionseffekte, wonach x_1 und x_2 den größten Einfluss auf die Varianz von y aufweisen. Durch die Begrenzung der Faktoren auf den Bereich $-1 \leq x_j \leq 1$ ist der Einfluss von x_3 durch seinen geradzahligen Exponenten bereits deutlich kleiner.

Die linke Seite der Abbildung 13.3 zeigt die FFT (Fast Fourier Transformation) der Analyse des Faktors x_2 . $\frac{\hat{D}_2}{2}$ und $\frac{\hat{D}_{-2}}{2}$ berechnen sich durch die Summe aller Amplituden beziehungsweise durch die Summe der Frequenzen bis $\omega = \frac{\omega_j}{2}$. Der Wert von $\frac{\hat{D}_2}{2}$ wird hingegen durch die Summe der Amplituden bei $\omega_2 = 1249$ und der drei folgenden harmonischen Frequenzen ($M = 4$) bestimmt. Zur Berechnung aller Sen-

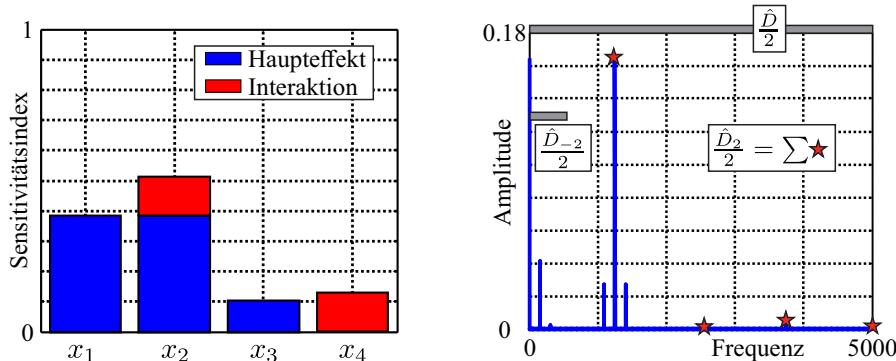


Abb. 13.3 eFAST Beispiel

sitivitätswerte inklusive der Haupt- und Interaktionskennwerte werden mit eFAST lediglich $n = n_r n_f$ Berechnungen des Modells benötigt. Sind nur die Effekte erster Ordnung gesucht, reduziert sich der Aufwand auf n_r Berechnungen.

13.4 Zusammenfassung

Die Sensitivitätsanalyse bietet bei vielen Analysen hilfreiche Informationen zum Verständnis und zur Weiterentwicklung von Simulationsmodellen und technischen Systemen. Dabei werden allgemein die drei Bereiche *Faktor Screening*, *lokale*- und *globale Sensitivitätsanalyse* unterschieden. In allen Bereichen wird nach Zusammenhängen zwischen der Varianz von Eingangsfaktoren und Systemantworten gesucht, die als Basis für weitere Entscheidungen dienen. Neben klassischen Verfahren, die für lineare Modelle entwickelt worden sind (Normierte Regressionskoeffizienten, PSS, PRESS, PCC, CPD), sind verschiedene Algorithmen vorhanden, die auch bei Verwendung nichtlinearer Modelle aussagekräftige Ergebnisse liefern (Korrelationsverhältnis, Sobol, FAST, eFAST). Dabei ermöglichen die Verfahren FAST und eFAST, unabhängig vom gewählten Metamodellansatz, mit geringem Rechenaufwand genaue und aussagekräftige Ergebnisse.

Literaturverzeichnis

1. Albers, S., Skiera, B.: *Marktforschung. Grundlagen - Methoden - Anwendungen*, chap. Regressionsanalyse, pp. 205–236. C. Homburg (1999) 417
2. Chan, K., Saltelli, A., Tarantola, S.: *Sensitivity analysis of model output: variance-based methods make the difference*. In: S. Andradttir, K.J. Healy, D.H. Withers, B.L. Nelson (eds.) Proceedings of the 1997 Winter Simulation Conference (1997) 420
3. Cukier, R.I., Fortuin, C.M., Shuler, K.E., Petschek, A.G., Schaibly, J.H.: *Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory*. Journal of Chemical Physics **59**, pp. 3873–3878 (1973) 425, 426, 427
4. Cukier, R.I., Levine, H.B., Shuler, K.E.: *Nonlinear Sensitivity Analysis of Multiparameter Model Systems*. Journal of Computational Physics **26**, pp. 1–42 (1978) 425, 427
5. Cukier, R.I., Schaibly, J.H., Shuler, K.E.: *Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. III Analysis of the approximations*. The Journal of Chemical Physics **63**, pp. 1140–1149 (1975) 425
6. Fang, K.T., Li, R., Sudjianto, A.: *Design and Modeling for Computer Experiments (Computer Science & Data Analysis)*. Chapman & Hall/CRC (2005) 192, 193, 194, 195, 197, 209, 210, 212, 213, 214, 219, 238, 254, 255, 417, 418, 419
7. Iman, R.L., Stephen, C.H.: *A Robust Measure of Uncertainty Importance for Use in Fault Tree System Analysis*. Risk Analysis **10**(3), pp. 401–406 (1990) 421
8. Koda, M., McRae, G.J., Seinfeld, J.H.: *Automatic Sensitivity Analysis of Kinetic Mechanisms*. International J. Chemical Kinetics **11**, pp. 427–444 (1979) 425, 427
9. McKay, D.M.: *Evaluating prediction uncertainty*. The Commission, Washington DC (1995) 419, 420, 421
10. McKay, D.M.: *Nonparametric variance-based methods of assessing uncertainty importance*. Reliability engineering & system safety **57**, pp. 267–279 (1997) 419
11. Pearson, K.: *Mathematical contributions to the theory of evolution*. In: Proceedings of the Royal Society of London, vol. 71, pp. 288–313 (1903) 419
12. Reedijk, C.: *Sensitivity Analysis of Model Output: Performance of various local and global sensitivity measures on reliability problems*. Master's thesis, Delft University of Technology (2000) 415, 419, 421, 427
13. Saltelli, A.: *Sensitivity analysis of model output: An investigation of new techniques*. Computational statistics & data analysis (1993) 420
14. Saltelli, A., Tarantola, S., Chan, K.P.S.: *A quantitative model-independent method for global sensitivity analysis of model output*. Technometrics **41**(1), pp. 39–56 (1999) 425, 426, 427, 428
15. Sobol, I.M.: *Sensitivity analysis for non-linear mathematical models*. Mathematical Modelling Computational Experiment **1**, pp. 407–414 (1993) 426
16. Sobol, I.M.: *Global Sensitivity Indices for Nonlinear Mathematical Models and Their Monte Carlo Estimates*. Math. Comput. Simul. **55**(1-3), pp. 271–280 (2001) 422, 423, 424

Kapitel 14

Strategie

14.1 Einleitung

Statistische Versuchsplanung hat viel mit Mathematik zu tun. Letztlich sind aber oft die nicht-mathematischen Dinge erfolgsentscheidend. Zum Beispiel hilft der strukturierte Ablauf, ein Problem zielgerichtet anzugehen. In sehr vielen Fällen sind ganze Arbeitsgruppen an Vorbereitung und Durchführung einer Versuchsreihe beteiligt, mitunter sogar abteilungsübergreifend. Die DoE hat in gewisser Weise einen teambildenden Character, da sie nicht nur einen favorisierten Faktor nach dem anderen untersucht, sondern von vornherein auf die gleichzeitige Analyse mehrerer Faktoren ausgerichtet ist. Damit wird das Projekt zu einer Sache gemeinsamen Interesses. Nachdem die theoretischen Grundlagen geklärt sind, soll dieses Kapitel dem Leser einige praktische Hinweise vermitteln und so den Einstieg in die Methode erleichtern.

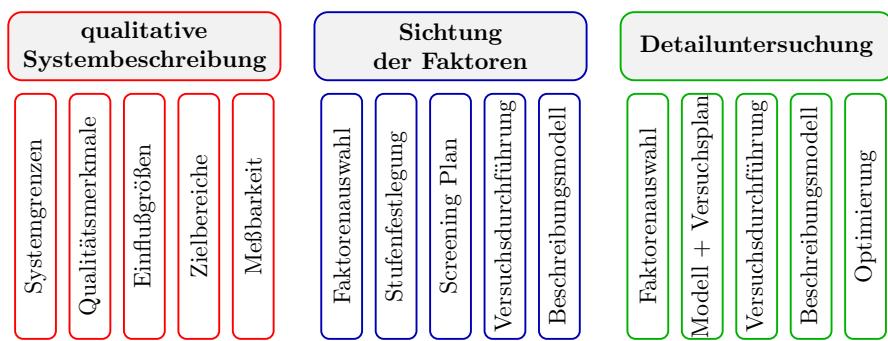


Abb. 14.1 Schematischer Ablauf einer DoE-Anwendung. Die drei Phasen werden nacheinander durchlaufen und bauen aufeinander auf.

14.2 Qualitative Systembeschreibung

Die qualitative Systembeschreibung mit Hilfe eines Parameterdiagramms ist der Dreh- und Angelpunkt jeder DoE. Fehler die hier gemacht werden, kann kein Statistiker jemals wieder korrigieren. Bei einer guten qualitativen Systembeschreibung kommt man mit fast jedem Versuchsplan zum Ziel, bei einer schlechten überhaupt nicht.

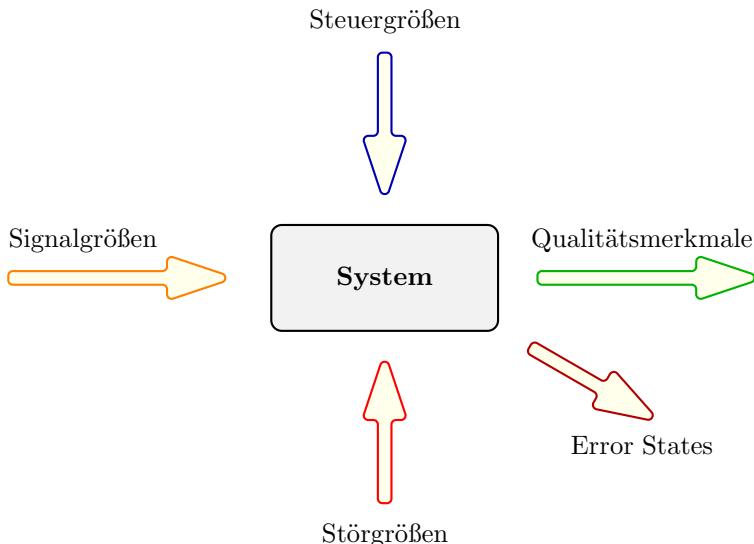


Abb. 14.2 Parameterdiagramm. Steuer-, Signal- und Störgrößen nehmen Einfluss auf das System. Ein Teil der Ergebnisse ist erwünscht, die Qualitätsmerkmale. Error States kennzeichnen die unerwünschten Ergebnisse. Im Abschnitt *Robustes Design* wurde dieses Diagramm im Detail vorgestellt.

Zunächst müssen Systemgrenzen und Qualitätsmerkmale definiert werden. In der Praxis kann sich hierbei eine längere Diskussion entwickeln. Diese ist hilfreich und sollte auf keinen Fall abgewürgt werden. Hierzu einige Kontrollfragen, quasi als Checkliste:

Welches System wird untersucht?

Wo sind die Systemgrenzen?

Ist das Team in Anbetracht der so definierten Systemgrenzen vollständig?

Welche Aufgabe hat das System?

Was unterscheidet ein gutes von einem schlechten System?

Ist das Qualitätsmerkmal messbar?

Wie gut lässt sich das Qualitätsmerkmal messen?

Gibt es eine messbare Ersatzgröße?

Was sollte man sicherheitshalber zusätzlich aufzeichnen?

Gibt es mehrere Qualitätsmerkmale?

Welchen Wert des Qualitätsmerkmals wollen wir jeweils erreichen?

Welches Qualitätsmerkmal hat Vorrang?

Sind wir mit dem System vollständig zufrieden, wenn die gewählten Qualitätsmerkmale in den gewünschten Bereich kommen?

Im nächsten Schritt geht es um die Festlegung der Faktoren für den Versuchsplan. Hierzu sollte man zunächst etwas Abstand von den möglicherweise vorhandenen Einschränkungen gewinnen. Es ist immer besser, aus einer großen Liste einige Favoriten auszuwählen, als sich vorschnell auf die erstbesten Größen festzulegen. Clevere DoE-Berater walzen diesen Teil der Gruppenaktivität kräftig aus, durchaus im Sinne des Gesamterfolgs. Bewährt haben sich Metaplan Karten und Pinnwand, damit keine Gedanken verloren gehen und eine nachfolgende Sortierung möglich ist.

Welche Parameter beeinflussen das System? (Brain-Storming mit Papp-Karten)

Gehören sie wirklich zum System (Steuergrößen)?

Falls nein, sollten sie als Störgrößen betrachtet werden ?

Gibt es sonstige Störgrößen?

Gibt es Signalgrößen, die den Betriebsbereich des Systems beschreiben?

Gab es in der Vergangenheit einen bislang noch nicht genannten Parameter, der Einfluss hatte? Gibt es Parameter, bei denen wir uns nicht sicher sind?

Kann man die Parameter reproduzierbar im Versuch einstellen?

Sind die Parameter voneinander unabhängig?

Gibt es kritische Kombinationen der Parameter untereinander?

Ist eine Veränderung des Parameters in der Serienproduktion durchsetzbar?

Was wissen wir bereits über den Parameter?

Wenn man alle Fragen beantworten kann, weiss man schon recht viel über sein System. Selten fängt man bei Null an, daher ist es immer gut, die Vorgeschichte systematisch zusammenzutragen. Es würde an ein Wunder grenzen (oder deutet auf ein müdes Team hin), wenn auf Anhieb alle Parameter in einem Versuchsplan Platz finden. Ausdünnen gehört zur Übung, sollte aber behutsam vonstatten gehen. Die Metaplan Karten lassen sich auf einer Pinnwand wunderbar sortieren. Doppelnenngungen sollte man nicht abhängen (wertet ab), sondern elegant durch Karten gleicher Nennung abdecken. Motto: Alles ist richtig. Oft redet sich das Expertenteam bei der Parametersuche warm und ist sich am Ende einig, dass eine neue Versuchsreihe unbedingt nötig ist. Dann sind alle im Boot (vor allem, wenn der DoE-Berater darauf achtet, dass jede Schlüsselfigur seinen Parameter unterbringen darf).

14.3 Versuchsdurchführung und Auswertung

Bei der nachfolgenden DoE werden aus der Fülle der Parameter die Faktoren des Versuchsplans ausgewählt. In der Regel ist ein zweistufiges Vorgehen empfehlenswert. Zunächst wird man viele Faktoren einem Screening unterziehen und dann wenige Faktoren im Detail untersuchen. Dies hat viele Vorteile. Zum einen sollte man nicht gleich am Anfang die Zahl der Faktoren unnötig beschränken. Jeder nicht beachtete Parameter geht verloren und man weiß letztlich nie so genau, ob er nicht vielleicht doch einen großen Einfluss hat. Ein einziger riesig großer Versuchsplan scheitert in der Praxis sehr oft. Besser sind Versuchspläne in vernünftiger Größe, die in endlicher Zeit ein brauchbares Zwischenergebnis liefern. Welche Größe passend ist, hängt von den Randbedingungen ab. Neben Kosten und Versuchsaufwand spielt die verfügbare Zeit eine große Rolle. In Krisensituationen kann es sinnvoll sein, bewusst auf kleinere Felder auszuweichen, um schnell an erste Ergebnisse zu kommen. Möglicherweise denkt man nach der ersten Versuchsreihe auch anders über die Stufeneinstellungen und kann diese in der zweiten Versuchsreihe passend korrigieren. Bei der Screening-DoE wird man auf große Kontraste hinarbeiten, also möglichst weite Stufenabstände wählen. Die nachfolgende Optimierung arbeitet möglicherweise mit einem enger abgesteckten Faktorraum.

Bieten sich einige Stufeneinstellungen von vornherein an?

Wo sind die technisch sinnvollen Grenzen? (Bauraum, Kosten, Verfügbarkeit)

Passen die Stufenabstände der unterschiedlichen Faktoren zueinander?

Sind einzelne Faktoren durch besonders weite Abstände bevorzugt?

Könnte es als Folge von Nichtlinearitäten im mittleren Einstellungsbereich völlig andere Ergebnisse geben?

Screening bedeutet: sichten und ausdünnen. Zunächst sollte man also möglichst viele Parameter als Faktoren in den Versuchsplan aufnehmen. Vorversuche können notwendig sein, um zu testen, ob kritische Kombinationen im Versuchsplan realisierbar sind. Im Zweifelsfall liefert ein großes Screening-Feld mehr Informationen als mehrere kleine Felder. Außerdem ist es besser, zusätzliche Kombinationen zu testen, als die gleiche Kombination mehrfach zu wiederholen. Blockbildung und Randomisierung sollten eingesetzt werden, wo es sinnvoll ist. Aus all diesen Erwägungen erfolgt nun die Auswahl des passenden Screening-Versuchsplans. In der Regel bleibt man bei zweistufigen Versuchsplänen, wobei sich dicht besetzte Pläne nach dem Yates-Standard und vor allem Plackett-Burman Pläne anbieten.

Steht der Versuchsplan, kommt die Logistik in's Spiel. Alle beteiligten Personen müssen die notwendigen Informationen bekommen. Material muss besorgt werden. Prüfstände werden belegt. Nicht selten scheitern gute Pläne an einer mangelhaften Ausführung. Oft ist nicht einmal böser Wille dabei, klare Darstellung ist Trumpf. Der "Operator" (also das Prüfstandspersonal) braucht einen Versuchsplan im Klartext und kann mit – oder + nichts anfangen. Es lohnt sich immer, vor Ort ein wenig Zeit zu investieren und die Beteiligten in die Grundprinzipien der DoE einzubringen. Die gesamte Testreihe sollte möglichst in einem Zug durchgeführt werden. Dies er-

fordert eine lückenlose Materialbeschaffung (Stückliste) mit entsprechender Vorbereitung (Zeitplan, Budget,...). Normalerweise sind gute Prüfstände voll ausgelastet. Jede Versuchsreihe muss also vorher korrekt eingesteuert werden (Umfang, Rüstzeiten, etc.). Wer den Betrieb aufhält (zum Beispiel weil das Material fehlt) fliegt vom Prüfstand. Eine DoE bietet in dieser Situation Vor- und Nachteile. Vorteilhaft ist die Möglichkeit, ohne Verzögerung in einem Zug mehrere Faktoren untersuchen zu können. Problematisch ist eine abgebrochene Messreihe, weil sie in der Regel nicht mehr auszuwerten ist. Dann ist entweder alles verloren oder man muss die Aktion mit entsprechendem Aufwand wiederholen, zumindest die fehlenden Versuche. Wiederholungen und Bestätigungs läufe gehören dazu, also sollten sie von Anfang an mit eingeplant werden. Wichtig, bereiten Sie die Entscheidungsträger darauf vor, dass es nicht bei der Screening DoE bleiben wird. Allen sollte das geplant zweistufige Vorgehen bekannt sein, damit es nicht nach einer verpatzten und dann auf andere Weise wiederholten Versuchsreihe aussieht. Normalerweise ist es kein Problem, diesen Punkt zu kommunizieren, weil jede andere Strategie mehr Versuche beinhalten würde.

Nach der Versuchsdurchführung steht die statistische Modellbildung an. Hier müssen Spezialisten aus den Fachabteilungen und Statistiker Hand in Hand arbeiten.

Lassen sich die abgeleiteten Schlussfolgerungen fachlich erklären?

Ist das Beschreibungsmodell hinreichend genau? (Fehlen wichtige Faktoren?)

Welche Faktoren sind signifikant?

Gibt es Faktoren, bei denen sich bereits eine optimale Einstellung gezeigt hat?

(Diese Faktoren bleiben im weiteren Verlauf bei dieser Einstellung.)

Liegen möglicherweise starke Wechselwirkungen vor?

Liegen möglicherweise starke Nichtlinearitäten vor?

Gibt es unterstützende CAE-Modelle oder Komponententests?

Die zweite Runde verläuft ähnlich. Nach der Faktorenauswahl folgt die Festlegung des Versuchsplans. Bei Bedarf sind neue Stufenabstände einstellbar. Möglicherweise ändert sich auch die Grundeinstellung des Systems, indem einige der im Screening untersuchten Faktoren auf eine neue Einstellung gebracht werden. Mit Hilfe der verfügbaren Ergebnisse muss eine Entscheidung getroffen werden, welches Beschreibungsmodell für die Detailuntersuchung angestrebt wird. Wechselwirkungen und Nichtlinearitäten spielen dabei eine Rolle. Im Zweifelsfall ist die schlichte Variation eines einzelnen Parameters eine wertvolle Zusatzinformation. Versuchsplan und Beschreibungsmodell sind miteinander verbunden, von zwei Ausnahmen abgesehen. 1. Monte-Carlo Versuchspläne und deren Derivate (Hypercubes, Space-Filling Design) beinhalten keine Bindung an ein Modell. 2. Man kann zweistufige Versuchspläne der Auflösungsstufe V bei Bedarf nachträglich erweitern, um auch quadratische Effekte zu erfassen. Die zweite Versuchsreihe geht auf den Prüfstand und die zweite Auswertung erfolgt. Diesmal sind natürlich die Anforderung an die Modellgenauigkeit höher. Außerdem geht es um die Festlegung der vorgeschlagenen Systemeinstellung, was immer mit Bestätigungs läufen verbunden sein sollte. Die Besprechung der Ergebnisse sollte wieder im Team erfolgen. Zum einen ist es wichtig, die Schlussfolgerungen zu hinterfragen und passende technische Er-

klärungen zu finden. Zum anderen bildet dies auch einen würdigen Abschluss der Studie, für die sich oft viele Leute eingesetzt haben.

14.4 CAE

Die DoE Anwendungen für Computermodelle laufen im Grundsatz gleich ab, allerdings meistens mit kleinerer Besetzung. Viele Berechnungsingenieure sind quasi Solisten. Die DoE gibt ihnen die Chance, aus der Isolation auszubrechen. Das klingt zunächst vermessener oder gar zynisch, aber bei genauerem Hinsehen wird die Aussage klar. Die qualitative Systembeschreibung kann auch hier im Team erfolgen. Auch wenn sie das Berechnungsverfahren nicht verstehen, so können Kollegen an dieser Stelle trotzdem einen wertvollen Beitrag leisten. Insofern gilt dieser Abschnitt des Strategiekapitels für Test und CAE in gleichem Maße. Die Ausführung der CAE-Studie verbleibt in der Regel bei Einzelnen. Immer häufiger werden allerdings derartige Studien auch fremdvergeben. Dann müssen beide Seiten wissen, worauf es ankommt. Logistik ist auch hier nicht unbedeutend, denn man muss die Läufe organisieren und die Rechnerkapazität vorhalten. Ratsam sind ebenfalls Rechnungen mit extremen Kombinationen, um zu prüfen ob das gesamte Feld abgearbeitet werden kann.

Oft überspringt man die Screening Phase, was aber nicht die Regel sein muss, denn gerade der Umgang mit vielen Eingangsgrößen ist eine der Stärken der Berechnungsverfahren. Bei der statistischen Modellbildung spielt Streuung normalerweise keine Rolle, dafür liegt das Augenmerk auf den Nichtlinearitäten und Wechselwirkungen. Viele Versuchspläne sind fest mit einem Beschreibungsmodell verbunden. Hier sollte man das Beschreibungsmodell auch mit neuen Kombinationen überprüfen, um darauf hin zu entscheiden, ob ein neuer Versuchsplan nötig ist. Hier kann der Statistiker helfen. An dieser Stelle sei jedoch davor gewarnt, über das Ziel hinaus zu schießen. Jedes Modell ist nur eine Näherung, kann aber für den Anwendungszweck völlig ausreichend sein. Auch hier sollte der Berechnungsingenieur externe Meinungen einholen: Wie genau muss es eigentlich sein? Ist das Ergebnis kommunizierbar?

Die Kommunikation der Ergebnisse ist sehr wichtig und erschließt dem Berechnungsingenieur weitere Quellen der Erfahrung. Mitunter laufen Test und CAE nebeneinander und die Berechner sehen die physischen Tests nicht mehr. Oft zeigen sich im Test Phänomene, die noch nicht im CAE-Modell abgebildet wurden, dann ist die Jagd nach Zehnteln ohnehin Makulatur. In den meisten Fällen werden die Kollegen dankbar sein, wenn der Berechnungsingenieur das Verhalten seines Modells nachvollziehbar dokumentiert. Nicht ohne Grund hat man schließlich mit der Berechnung angefangen. Oft wird leider der Nutzen der CAE-Methoden unterschätzt. Hier leistet die DoE einen Beitrag mit gut darstellbaren Beschreibungsmodellen.

Bei Detailuntersuchungen und Optimierungen kann es sinnvoll sein, das Metamodell mit einem anderen Verfahren aufzubauen, zum Beispiel Radial Basis Function, Neuronale Netze oder Splines. Ausschlaggebend ist die erreichbare Genauigkeit

des Metamodells im Vergleich zur angestrebten Verbesserung des Qualitätsmerkmals.

14.5 Software

Mittlerweile bieten sehr viele Programme die Erstellung von Versuchsplänen und deren Auswertung an. Zum Teil handelt es sich um Spezialprogramme, die ausschließlich für diesen Zweck entwickelt wurden. Hinzu kommen aber auch viele Statistikprogramme mit DoE-Zusatzmodul, die ebenfalls einen beachtlichen Leistungsumfang bieten. Die benötigten mathematischen Verfahren sind schon seit vielen Jahren bekannt und nicht sonderlich anspruchsvoll in Bezug auf die Rechtleistung. Im Vergleich zu den Kosten einer Versuchsreihe oder im Vergleich zu den Lizenzkosten kommerzieller Berechnungsprogramme ist die Software nicht teuer, erschließt aber dafür die Welt der DoE. Also lohnt sich die Anschaffung auf jeden Fall.

Bastellösungen mit Tabellenkalkulationsprogrammen führen nicht weit, binden Ressourcen bei der Programmierung und bleiben immer erklärbungsbedürftig für neue Anwender. Einige Nischenprodukte sind kommerzialisierte Softwarelösungen mit sehr begrenztem Leistungsumfang, ursprünglich offenbar eher für den Eigenbedarf entwickelt. Davon ist abzuraten, weil die Beschränkungen bei ernsthafter Anwendung zum Hindernis werden. Bei Berechnungsprogrammen wird mittlerweile oft ein DoE-Paket zur Ansteuerung der Modelle angeboten. Diesen Zweck erfüllen die DoE-Pakete in der Regel tadellos, jedoch ist die Auswertung oft nicht besonders weit entwickelt. Lückenhaft kann auch die Qualitätskontrolle der Versuchspläne sein.

Im Wesentlichen erfüllen die Auswerteprogramme drei Aufgaben:

1. Auswahl eines passenden Versuchsplans

- Auswahl aus vorkonfektionierten Standard-Versuchsplänen
- Unterstützung bei der Konstruktion eines individuellen Versuchsplans
- Möglichkeit zum Import benutzerdefinierter Felder
- Analyse des Versuchsplans in Bezug auf das geplante Beschreibungsmodell

2. Datenanalyse

- Import der Versuchsdaten
- Lösung des Gleichungssystems für jedes Qualitätsmerkmal
- statistische Analysen (ANOVA, Konfidenzintervalle etc.)
- Interaktion mit dem Beschreibungsmodell (Planspiele)
- Verknüpfung der Qualitätsmerkmale (MRO, PCA)
- Vorhersage einer optimalen Einstellung
- Export der Daten zur Weiterverarbeitung

3. Darstellung der Ergebnisse

- Interaktive Darstellung für Planspiele
- standardisierte Darstellungen im Programm
- editierbarer Export der Darstellungen für Berichte

Die Softwareanbieter setzen individuelle Akzente in den oben genannten Aufgabenbereichen. Es gibt viele sehr gute Lösungen, aber das Programm muss zur geplanten Anwendung passen. Je nach Anwendung ergeben sich unterschiedliche Leistungsanforderungen.

Bei Anwendungen in der Verfahrenstechnik reichen standardisierte Versuchspläne oft nicht aus. Wenn die Software keine benutzerdefinierten Felder erlaubt und die Erstellung spezieller Felder nicht richtig unterstützt, ist sie in vielen Fällen überhaupt nicht einsetzbar. Auch die Qualitätskontrolle der Felder ist dann wichtig und muss in jedem Fall auf das geplante Beschreibungsmodell abgestimmt sein.

CAE-Anwendungen gestatten häufig eine große Zahl von Faktoren, Stufen und Qualitätsmerkmalen. Die Leistungsgrenze der Auswerteprogramme verschiebt sich jedoch ständig in Richtung aufwendiger Modelle und einer höheren Zahl von Faktoren. Automatisierter Datenimport wird bei CAE-Anwendungen schnell zum Thema, wenn die Erstellung eines Metamodells routinemäßig geplant ist.

Statistische Analysen der Daten und standardisierte Berechnungen (Effekte, Wechselwirkungen) bieten alle Programme dieser Kategorie an, daraus ergibt sich kein Unterscheidungsmerkmal. Es sei denn, im Unternehmen existiert ein spezieller Standard, der nicht von allen Programmen unterstützt wird. Die Option der Multiple-Response-Optimisation ist immer empfehlenswert, da es selten bei einem Qualitätsmerkmal bleibt. Eine gute Interaktion mit dem Beschreibungsmodell ist für die erfolgreiche Analyse wichtig. Diese "Planspiele" bringen ein gutes Systemverständnis und müssen in der industriellen Anwendung schnell vonstatten gehen, weil sie nicht selten in Besprechungen mit den jeweiligen Fachabteilungen verlangt werden.

Im Großen und Ganzen ergibt sich folgende Fallunterscheidung:

1. Durchschnittlicher Anwender

Statistikprogramme mit DoE-Zusatzmodul decken den normalen Anwendungsbereich gut ab und bieten darüber hinaus viele Möglichkeiten weiterer statistischer Analysen, unabhängig von der statistischen Versuchsplanaufstellung. In vielen Fällen werden die lokalen DoE-Experten im Unternehmen zur Anlaufstelle für Statistikfragen aller Art oder umgekehrt. Ein Universalwerkzeug wie "Statgraphics[®]", "JMP[®]", "StatisticaTM" oder "MinitabTM" ist dann sehr praktisch. Wer andere Statistikanwendungen nicht braucht, fährt auch mit einem spezialisierten DoE-Programm sehr gut, zum Beispiel "Design Expert[®]".

2. Anwender mit Bedarf an speziellen Versuchsplänen

In der Verfahrenstechnik und der chemischen Industrie allgemein wird der Anwender mit vorkonfektionierten Versuchsplänen nicht zureckkommen. Hier ist in der Regel ein spezialisiertes DoE-Programm die beste Wahl. "Modde[®]", "StavexTM" oder "Design Expert[®]" seien als Beispiel genannt. Ein sorgfältiger Vergleich ist an dieser Stelle sinnvoll, denn die Programme unterscheiden sich in

Bezug auf die Leistungsfähigkeit bei der Erstellung maßgeschneiderter Versuchspläne. "JMP®" ist ebenfalls enorm leistungsfähig, man muss allerdings etwas Zeit in die Scriptsprache investieren, um sich eine passende Lösung zusammenzustecken. Die DoE-Module der Statistikprogramme holen auf und erreichen mittlerweile fast die Leistungsfähigkeit der spezialisierten Programme, jedoch leidet mitunter die Bedienbarkeit an den nachträglich angestückelten Features.

3. CAE-Poweruser

Während normale CAE-Anwender noch mit üblichen Statistikprogrammen alle Versuchspläne bearbeiten können, wird der Poweruser mit seinen großen Metamodellen schnell die Kapazitätsgrenzen sprengen. Im Laufe der Zeit verschieben sich wohlgerne die Kapazitätsgrenzen der Auswerteprogramme und in vielen Fällen reicht dann die Leistungsfähigkeit der Standardlösungen völlig aus. Der Poweruser kennt sich in der Regel mit Programmierung aus und schrekt nicht vor einer scriptorientierten Ansteuerung zurück. In diesen Fällen steht die Option offen, gleich zur Multivariaten Datenanalyse mit Regressionsverfahren zu greifen. Dies geht sogar zum Nulltarif mit spartanischen aber leistungsfähigen Freeware-Programmen, zum Beispiel "R". Für Matlab werden Zusatzpakete angeboten, die alle gängigen Verfahren zur Multivariaten Datenanalyse beinhalten, zum Beispiel die "SUMO-toolbox" (SURrogate MOdeling). Immer populärer wird eine neue Kategorie von Software, die primär zur automatisierten Ablaufsteuerung von Computermodellen entwickelt wurde und massiv in die multivariate Datenanalyse einsteigt. "modeFRONTIER™", "Isight", "optiSLang", "HyperStudy" und "Optimus" seien exemplarisch genannt. Vorteilhaft ist hier die reichhaltige Auswahl an Solvieren und Näherungsverfahren, in Verbindung mit weit gesteckten Kapazitätsgrenzen. Hier gibt es allerdings deutliche Unterschiede, daher ist ein genauer Abgleich mit den individuellen Anforderungen erforderlich.

In jedem Fall ist es sinnvoll, mehrere Testanwendungen zu erstellen und mit den Demoversionen der in Frage kommenden Programme zu bearbeiten. Die Benutzeroberflächen sind sehr unterschiedlich, auch bei Programmen einer Kategorie. Dies liegt zum Teil an der historischen Entwicklung der Programme. Viele sind bereits seit Jahrzehnten auf dem Markt und wurden von Großrechnern auf PCs portiert. Die aus heutiger Sicht exotische Benutzerführung wird mitunter nur schrittweise umgestellt. Die Entwicklungsziele der Programme sind teilweise auch völlig unterschiedlich. DoE befindet sich aus der Sicht des Softwareherstellers im Zentrum des Interesses, ist ein Randgebiet der Statistik, wird als eine unter mehreren Methoden für eine multivariate Datenanalyse angeboten oder als kleine Zugabe für ein Programm zur automatisierten Ansteuerung von Computermodellen. Darüber hinaus verfolgt jeder Entwickler sein eigenes Konzept, also wird es auch in Zukunft bei der Bedienung und der Darstellung der Ergebnisse Unterschiede geben. Der praktische Nutzwert für den Anwender hängt nicht nur von der objektiven Leistungsfähigkeit des Programms ab, sondern auch vom individuellen Aufwand für Einarbeitung und Gebrauch. Hier gibt es keinen einheitlichen Maßstab.

Ein Software-Ranking wäre immer subjektiv gefärbt und müsste fairerweise in kurzen Zeitintervallen aktualisiert werden, um den jeweils neuesten Versionen

Rechnung zu tragen. Daher die Empfehlung: Vertrauen Sie auf Ihr eigenes Urteil, nachdem Sie an praktischen Beispielen die Demoversionen getestet haben. Innerhalb einer Programmkatgorie kann dann eigentlich nichts schief gehen, denn der Marktdruck nivelliert die Leistungsunterschiede¹

¹ Mit Ausnahme der Programme zum Metamodelling, also den CAE front ends. Hier muss man bis auf Weiteres mit deutlichen Unterschieden in der Funktionalität rechnen.

Kapitel 15

Strategie für komplexe Systeme

Für Analysen komplexer Systeme lässt sich in Abgrenzung zu klassischen DoE basierten Analysen kein festes beziehungsweise ideales Vorgehen definieren. Grundsätzlich ist der Prozess jedoch in mehrere Schritte einteilbar, welche in Abbildung 15.1 abgebildet sind.

- Vorbereitung und Planung
- Erstellung eines Versuchsplans
- Experiment (Messung und Rechnung)
- Qualitätsanalyse der Daten
- Erzeugung von Metamodellen
- Qualitätsanalyse der Metamodelle
- Analyse der Daten und Metamodelle
- Optimierung
- Prüfung der Analyse und Optimierungsergebnisse
- Dokumentation

Die genauen Inhalte jedes Prozessschrittes werden dabei an die jeweilige Aufgabe und das zu untersuchende System angepasst. Im Folgenden sind zu jedem der Schritte einige grundlegende Strategien beziehungsweise Denkanstöße dargestellt, die sich in vielen Projekten als hilfreich erwiesen haben.

15.1 Vorbereitung und Planung

Eine ausführliche Vorbereitung und Planung ist die Grundlage für eine erfolgreiche Durchführung aller Prozessschritte und somit des übergeordneten Projekts. Die Vorbereitung ist meist wichtiger als langwierige Optimierungen und Auswahlverfahren für den besten Metamodellansatz inklusive Modellparameter oder Analysealgorithmus.

In der Praxis erweist sich die Vernachlässigung der Vorbereitung als häufigste Ur-

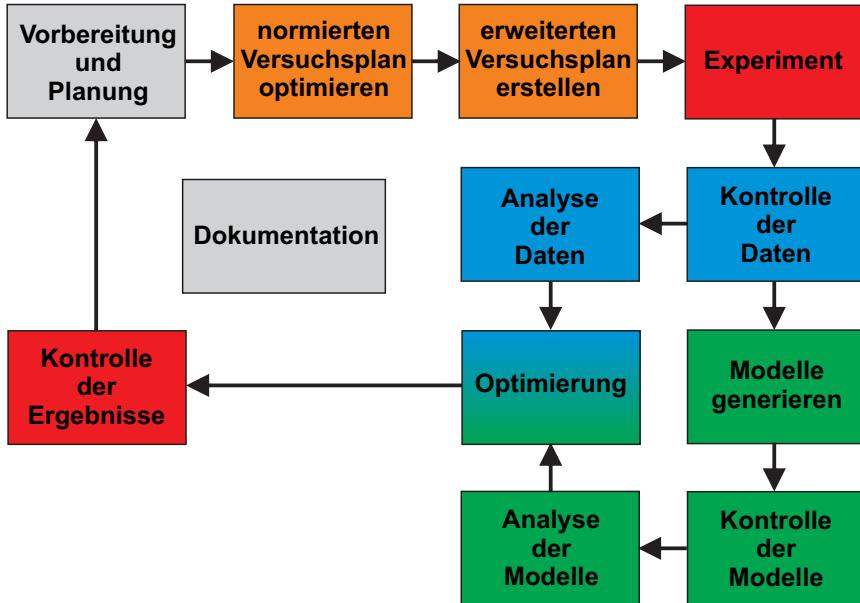


Abb. 15.1 Analyse komplexer Systeme: Flussdiagramm

sache für das Scheitern eines Projektes. Zu schnell lassen sich Argumente zur Verkürzung oder zum Überspringen der Vorbereitungsphase finden.

- Wir müssen schnellstmöglich Ergebnisse oder Fortschritte liefern.
- Das Projekt steht auf der Stelle und es muss dringend weitergehen.
- Wir wissen was zu tun ist und daher können wir direkt anfangen.
- Wir sind bereits jetzt zeitlich hinter dem Projektplan.
- Das ganze Gerede bringt uns nicht im geringsten weiter.
- Es ist schon alles vorbereitet und wir sollten anfangen.
- Darüber können wir uns auch später unterhalten.
- Offene Punkte klären wir beim nächsten Mal.
- Wir warten auf einen Testplan.

Bei diesen oder ähnlichen Argumenten ist es nicht immer leicht, den ersten Prozessschritt in ausreichendem Umfang zu bearbeiten und alle Beteiligten davon zu überzeugen, dass ein hoher Preis für eine schlechte Vorbereitung im Laufe des Projekts zu zahlen ist. Neben einem nicht idealen, verspäteten oder keinem Ergebnis, ergeben sich dann schnell Erkenntnisse wie:

- Das haben wir nicht bedacht.
- Dafür waren die Messung und Analyse leider nicht vorgesehen.
- Dann bleibt uns nichts anderes übrig als einen Großteil der Messungen mit neuen Faktorgrenzen zu wiederholen.
- Das Qualitätskriterium haben wir leider nicht in unserem Datensatz.

- Die ganze Methode ergibt keine sinnvollen Ergebnisse.

Aus diesem Grund ist es ratsam folgende grundlegende Fragen im Vorfeld zu klären, wobei es sich hierbei um erste Anregungen handelt und die Liste keinen Anspruch auf Vollständigkeit beansprucht.

1. Sind alle notwendigen Personen und Arbeitsgruppen im Projekt involviert?

Diese Frage umfasst Personen aller beteiligten Arbeitsbereiche des Projektes, vom Prüfstandsfahrer oder CAE-Simulant, der für die eigentliche Erzeugung der Projektdaten verantwortlich ist, über Ingenieure, die die Daten analysieren und zur Optimierung eines Systems verwenden, bis hin zum (Projekt-)Manager, der für eine reibungslose Durchführung oder die strategische Einordnung des Projekts verantwortlich ist. Zu berücksichtigen sind dabei auch potenzielle spätere Anwender der Daten, Modelle und Ergebnisse, die bislang noch nicht aufgetreten sind. Dazu ist eine grundlegende Analyse weiterer Anwendungsmöglichkeiten notwendig. Häufig können mit kleinen Änderungen im Versuchsplan oder der Modellerstellung viele weitere Aufgaben bearbeitet werden und es ist nicht unwahrscheinlich, dass nach den ersten Ergebnispräsentationen Ideen zur weiteren Anwendungen entstehen und geklärt werden muss, warum nicht bereits vorher daran gedacht wurde.

“Könnten diese Informationen nicht auch zur Klärung der folgenden Frage oder des folgenden Problems verwendet werden?”

2. Was sind die genauen Ziele des Projekts?

Für die meisten Beteiligten scheint diese Frage die einfachste zu sein, wobei sich in der Praxis zeigt, dass die typischen Antworten sehr vage formuliert und daher nicht verwendbar sind.

- Wir wollen die Leistung des Produkts verbessern.
- Es wäre schön, wenn der Fertigungsaufwand reduziert würde.
- Es sollten keine Zusatzkosten auftreten.
- Wir benötigen das Ergebnis schnellstmöglich.

Wichtig ist hier das Ziel zu konkretisieren und im besten Fall als messbare Größe zu definieren. Ein einfaches Beispiel wäre dazu:

Die Leistungsabgabe des Verbrennungsmotors eines Fahrzeugs mit der Typenbezeichnung 313 muss sich mindestens von 20 PS auf 25 PS erhöhen. Die momentanen Fertigungskosten von 313 Taler und die Montagezeit von 313 Minuten dürfen nicht überschritten werden. Die maximalen Einbaumaße des Grundmotors inklusive aller Anbauteile sind auf die jetzigen Maße limitiert und die Anschlusslemente sind gleich zu halten. Zugehörige Detailinformationen befinden sich in Dokument DD313. Abschluss des Projekts mit einem lauffähigen Prototyp ist am 3.1.3000.

Die Beschreibung beinhaltet messbare Qualitätsgrößen mit quantitativen Zielwerten und Randbedingungen sowie eine Beschreibung des zu einem definierten Zeitpunkts terminierten Ergebnisses.

3. Welche Ressourcen stehen dem Projekt zur Verfügung?

Ein sinnvoller Projektplan kann nur entwickelt werden, wenn die vorhandenen

Ressourcen berücksichtigt werden. Dabei sind neben zur Verfügung stehende Mitarbeiter, Finanzierung und Zeit ebenfalls vorhandene Prüfstands- und Computerkapazitäten sowie Messtechnik und der mögliche Automatisierungsgrad aller Prozessschritte zu beachten. Der beste theoretische Plan wird scheitern, wenn er durch falsche Ressourcenplanung nicht umsetzbar ist.

4. Mit welchen Qualitätsmerkmalen kann das System beurteilt werden?

Hierbei ist zu klären, mit welchen quantitativen Größen das System zu bewerten ist und wie diese Größen zu bestimmen sind. Oft liegen in diesem Zusammenhang lediglich qualitative Aussagen vor, die konkretisiert werden müssen.

“Der Anstieg muss steiler werden.”

Neben der Umformulierung in eine quantitative Aussage mit genau definierten Zielen (siehe oben), sollte die Bestimmungsmethode spezifiziert werden, um spätere Missverständnisse zu vermeiden. Abbildung 15.2 zeigt dazu einen Kurvenverlauf mit variablem Anstieg. Das Qualitätsmerkmal *Anstieg* könnte dabei zur Vergleichbarkeit durch zwei Punkte (20% und 80%) definiert werden, welche durch das Erreichen von 20 und 80 Prozent des Maximalwerts bestimmt werden.

Neben den ersichtlichen Qualitätsmerkmalen, welche die Optimierungsgrößen beschreiben, müssen weitere Systemantworten berücksichtigt werden, die beispielsweise zur Beurteilung der Messqualität (Varianz des gemessenen Signals) oder zur Definition zusätzlichen Randbedingungen (z.B. maximale Temperatur oder Lautstärke) dienen.

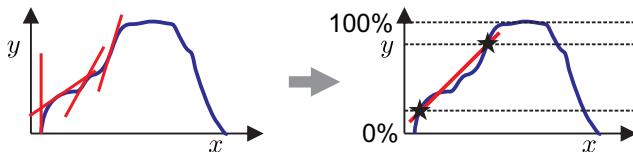


Abb. 15.2 Bestimmung einer Qualitätsgröße

5. Welche Faktoren und Faktorgrenzen sind zu verwenden?

Bei dieser Frage ist zu definieren mit welchen Faktoren das System (die Qualitätsgrößen) beeinflusst wird und in welchen Grenzen diese zu verändern sind. Zwei grundsätzliche Tendenzen sind dabei abzuwägen: Einerseits der Wunsch, große Faktorbereiche zu erfassen um beispielsweise ein Faktorscreening durchzuführen, oder ein grobes Verständnis des Systems in weiten Grenzen zu erhalten. Andererseits sind oft kleine Variationsgrenzen sinnvoll, wenn komplexe Zusammenhänge in einem abgegrenzten Bereich des möglichen Faktorraums hoch aufgelöst abgebildet werden müssen. In einigen Fällen ist eine kontinuierliche Verkleinerung des Faktorraums in sequenziellen Systemanalysen hilfreich, wobei Messungen sowie Simulationsergebnisse aus vorherigen Analysen, welche in den verkleinerten Faktorraum fallen, wiederverwendet werden können.

Wichtig ist dabei immer, dass die Faktoren in einem *sinnvollen* und zu einander abgestimmten Bereich variiert werden. Ist beispielsweise der Bereich eines einzelnen Faktors unsinnig *groß* im Verhältnis zu allen andern gewählt, so wird

wahrscheinlich der Effekt dieses einen Faktors auf die Qualitätsgrößen so groß sein, dass er alle anderen Effekte überschattet. Bei einem für die Praxis und die Aufgabe sinnvoll gewählten Bereich wären die Effekte hingegen in einer vergleichbaren Größenordnung.

6. Sind die gewählten Faktoren unabhängig und sind alle Kombinationen einsetzbar?

Bei der Erstellung eines normierten Testplans wird davon ausgegangen, dass alle Faktoren beliebig und unabhängig in den gewählten Grenzen einstellbar sind, was nicht immer der Fall ist. Manchmal werden ausgewählte Faktoren direkt durch andere Faktoren festgelegt und sind daher nicht unabhängig wählbar. Als Beispiel dient ein Stahlzylinder mit Durchmesser D , Höhe L , Oberfläche O und Gewicht G . Grundsätzlich können alle Parameter als Faktoren verwendet werden, aber wenn z.B. D und L definiert wurden, sind auch O und G festgelegt. In einem anderen Beispiel soll die Leistung sowie die Temperatur eines Verbrennungsmotors verstellt werden. Grundsätzlich sind beide Faktoren frei einstellbar aber es zeigt sich, dass bei höheren Leistungen die Kühlung des Motors nicht mehr ausreicht und dann die Temperatur mit Erhöhung der Leistung ebenfalls ansteigt und dadurch beide Faktoren stark korrelieren. In anderen Fällen sind Faktoren in gewissen Kombinationen aus zum Beispiel mechanischen Gründen nicht einstellbar. Nehmen wir beispielhaft an, dass zwei Bauelemente auf einer vorgegebenen Strecke hintereinander montiert werden (Abbildung 15.3). Wenn die Faktoren die absolute Position der Bauteile angeben, können diese zwar unabhängig gewählt werden, aber die Bedingung, dass Bauteil 2 immer hinter Bauteil 1 platziert werden muss, verhindert 50% der Faktorkombinationen. Dies kann akzeptiert werden, muss aber bei der Modellerstellung und den späteren Prozessschritten berücksichtigt werden. Alternative Definitionen der Faktoren können Abhilfe schaffen, erzeugen aber teilweise andere Schwierigkeiten. Eine erste Variante ist die Definition der zweiten Position als Entfernung zur ersten Position (Abbildung 15.3). Dadurch ist sichergestellt, dass die zweite Position immer hinter der ersten Position liegt. Vorteilhaft ist diese Definition zum Beispiel, wenn der Abstand zwischen den Positionen der eigentlich signifikante Einflussfaktor auf die Qualitätsgröße ist und nicht unbedingt die absolute Position. Zu berücksichtigen ist jedoch, dass sich der maximale absolute Wert der zweiten Position vergrößert und eventuell einer Maximalgrenze überschreitet. Ist eine feste maximale Position vorgegeben könnte alternativ die zweite Position als prozentualer Abstand zwischen der ersten Position und der maximalen Position definiert werden. Dies verhindert, dass die zweite Position vor der ersten Position oder hinter der Maximalposition liegt. Die absolute Variationsbreite des zweiten Faktors wird dadurch in Abhängigkeit vom ersten Faktor verändert, was eine einfache Interpretation der Faktoreinstellung für Position zwei erschwert, da gleiche prozentuale Unterschiede nicht identische reale Unterschiede beschreiben. Weiterhin wird der neue Faktor eine deutliche Interaktion mit Position eins aufweisen, die später zusätzlich durch das Metamodell erkannt und abgebildet werden muss. Neben der Kombinierbarkeit von Faktoren ist zu klären, ob die Faktoren in beliebigen Stufen eingestellt oder eingesetzt werden können. Wenn eine Material-

dicke nur in den Stufen 1,2,4 und 6 mm lieferbar ist, macht eine Stufe von 3.2 mm wenig Sinn, was in der Versuchsplanerstellung berücksichtigt werden muss. Ähnliches gilt zum Beispiel auch für die Anzahl von Schrauben. In einer Berechnung ist der Wert 6.3 Schrauben eventuell noch sinnvoll zu verarbeiten, jedoch wird diese Einstellung im physikalischen Experiment schwer umzusetzen sein.

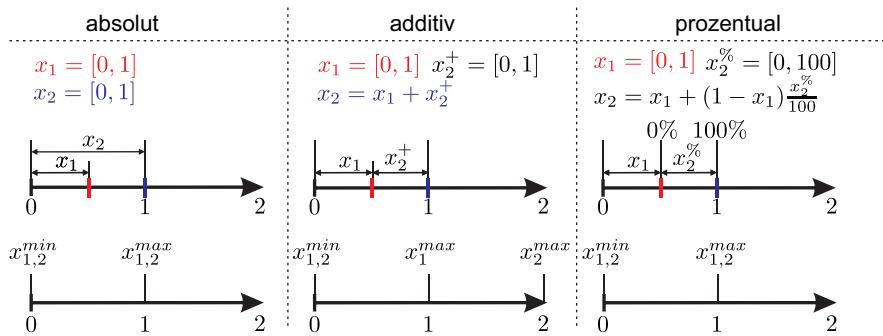


Abb. 15.3 definition von Faktoren

7. Welche Störgrößen beeinflussen die Datenbestimmung und wie groß ist ihr Effekt?

In den seltensten Fällen kann von einer idealen ungestörten Datenbestimmung ausgegangen werden. So existieren eventuell nicht kontrollierbare Parameter, welche das Systemverhalten beeinflussen. Störgrößen können dabei Umwelteinflüsse, wie Temperatur oder Luftfeuchtigkeit sein, aber auch unterschiedliche Hardware, wie Prüfstände oder Messgeräte sowie unterschiedliche Prüfstandsfahrer. Zeitlicher Drift von Prüfständen ist ebenfalls zu berücksichtigen, der im Laufe einer oder zwischen Messungen auftritt. Im Bereich von Computerexperimenten ist im Simulationsmodell absichtlich integriertes und nicht reproduzierbares Rauschen oder numerisches Rauschen zu prüfen. So ist beispielsweise bei einigen komplexen CFD Simulationen ein numerisches Rauschen durch kleine Faktoränderungen beobachtbar, die das Qualitätsmerkmal um einen kontinuierlichen Trend streuen lassen. Weiterhin können auch hier Hardwareänderungen (32bit, 64bit) zu unterschiedlichen Genauigkeiten und Ergebnissen führen. Werden Qualitätsgrößen am Ende einer nicht eingeschwungenen Simulation bestimmt, so bewirken kleine Änderungen in den Faktoren, dass einmal im Tal und ein anderes Mal am Berg einer Schwingung das Ergebnis bestimmt wird, was ebenfalls zu einer Streuung um den eigentlich gewünschten Qualitätswert führt.

8. Sind Faktoren und Qualitätsfunktionen zu tauschen?

Obwohl bereits in mehreren Punkten die Fragen zum eigentlichen Ziel des Projekts, den Einsatz der Metamodelle und die Qualitätsgrößen gestellt wurde, sollte hier geklärt werden, ob ein Tausch einzelner Faktoren und Qualitätsgrößen Vorteile liefern. Die Entscheidung, ob ein Parameter ein Faktor oder eine Qualitätsgröße ist, variiert in Abhängig vom Einsatz. Zur Erläuterung betrachten wir

den bereits erwähnten Stahlzylinder und das fiktive Projektziel, den gefertigten Zylinder später mittels eines vorgegebenen LKWs transportieren zu können. Der Transport wird dabei durch die Limitierung der Größen Gewicht G und Durchmesser D begrenzt. Mögliche Parameter und Qualitätsgrößen sind diesmal Höhe H , Durchmesser D und Gewicht G . Wird der Zylinder als Plateau für verschiedene Baukräne verwendet ist neben der Höhe H der Durchmesser D als Faktor zu wählen, da der Zylinder an die Standfläche der Kräne angepasst werden muss. Das Gewicht G würde zwangsläufig zum Qualitätskriterium. In einer anderen Anwendung werden die Zylinder als Gegengewicht für Baukräne verwendet. Dabei ist in erster Linie das Gewicht G der Zylinder entscheidend, so dass das Gewicht G als zweiter Faktor neben H gewählt und der Durchmesser D zur Qualitätsgröße wird.

9. Wie ist die benötigte Approximationsgenauigkeit im Vergleich zur gegebenen Datengenauigkeit?

Durch diese Frage sollen Informationen über die zu erwartende Mess- und Simulationsgenauigkeit gesammelt und in Relation zu benötigten Approximationsgenauigkeiten gestellt werden. Schnell werden in den gegebenen Daten und den daraus entwickelten Metamodellen Änderungen und Effekte gesucht, die weit unterhalb der gegebenen Mess- und Simulationsgenauigkeit liegen. Besonders am Ende eines Projektes werden die optimierten Parametersätze am Prüfstand oder in der Simulation nachgemessen und es wird eine Abweichung zum berechneten Ergebnis festgestellt. Schnell wird dies auf ein ungenaues Metamodell oder eine schlechte Analyse zurückgeführt. In vielen Fällen kann die Abweichung jedoch auf Messungenauigkeiten oder Systemveränderungen zurückgeführt werden. Die weitläufige Vorstellung, dass eine Messung die *Wahrheit* darstellt, erweist sich oft als falsch, was schnell durch eine strukturierte Projekt-durchführung mittels Versuchsplanung und Messdatenanalyse aufgedeckt wird. Um spätere Diskussionen zu umgehen, sollte dies mit einigen Tests vor Projektbeginn dokumentiert werden.

Neben allgemeinen Messdatenanalysen zur Detektion von Ausreißern, Systemveränderungen (z.B. Systemdrift) oder Robustheit von Faktoreinstellungen (Kapitel 8.7) kann bereits eine einfache Wiederholung eines Versuchsplans und die direkte Gegenüberstellung der real verwendeten Faktoreinstellungen und Messergebnisse einen ersten Hinweis auf die Messgenauigkeit liefern.

Simulationen sind wie bereits erläutert im ersten Schritt als deterministisch anzusehen. Gleiche Faktoreinstellungen liefern immer gleiche Qualitätsgrößen. Dieses ändert sich zum Beispiel, wenn eine über die aktuelle Zeit initialisierte Zufallsgröße in das Modell integriert wird. Bei instabilen, schwingenden oder nicht konvergierenden Simulationen treten jedoch bereits bei kleinen (insignifikanten) Faktoränderungen starke Streuungen in den Qualitätsgrößen auf. Neben kleinen Faktoränderungen zur Analyse der Ergebnissstabilität sind alternativ kleine kontinuierliche Änderungen der Faktoren innerhalb des Faktorraums sinnvoll. Durch kontinuierliche Änderungen ändern sich die Qualitätsgrößen in den meisten Anwendungsfällen ebenfalls kontinuierlich oder sie weisen lediglich vereinzelte Sprünge auf. Ist das Simulationsmodell instabil oder haben eingeführte

Rauschtermen einen signifikanten Einfluss, zeigen die Ergebnisse Rauscheffekte um einen mittleren Trend. Die Wahrscheinlichkeit, dass es sich dabei um wahre Effekte handelt, ist gering und sollte im Zweifelsfall näher untersucht werden. Die klassische Messmittelanalyse ist natürlich ebenfalls eine Möglichkeit umfangreiche Informationen über die erreichbare Messgenauigkeit und störende Einflussgrößen zu erhalten. Auf eine genauere Betrachtung der Messmittelanalyse wird hier verzichtet.

Die benötigte Datengüte hängt stark vom Anwendungsfall ab. Es ist zum Beispiel zu klären, ob die Approximation von Absolutwerten nötig oder die Vorhersage von Trends ausreichend ist. Werden zum Beispiel die "besten" Faktoreinstellungen für eine Optimierungsaufgabe gesucht oder soll ein grundlegendes Systemverständnis aufgebaut werden, ist lediglich das grundlegende Verhalten wichtig. Absolutwerte, welche im Anschluss der Analyse ermittelt werden, weisen dann eventuell Abweichungen auf, jedoch befindet sich das System mit den gewählten Faktoren im gewünschten Zustand. Ein letztes Feintuning der gewählten Faktoreinstellungen und eine Kontrolle des Systems sollte immer ein Bestandteil des Projektplans sein und ist somit keine Zusatzarbeit.

Eine genauso wichtige Aufgabe, wie die Ermittlung der gegebenen Datenqualität, ist die verständliche Aufbereitung dieser Information für alle Beteiligten, da nur so eine gesunde Balance zwischen erwarteter und realer lieferbare Modellbeziehungsweise Analysegenauigkeit ermöglicht wird. In einigen Fällen wird den Beteiligten zum ersten Mal bewusst, welche reale Daten- und Messgenauigkeit vorliegt, da vorher die benötigte Datenbasis zur sinnvollen Beurteilung nicht vorlag. Die Datenbasis, welche mittels der statistischen Versuchsplanung ermittelt wird ist somit eine ideale Grundlage, um neben den eigentlichen Projektzielen ebenfalls ohne zusätzlichen Messaufwand notwendige Qualitätsuntersuchungen durchzuführen.

10. Können Experimente an einem Tag oder in einem Durchlauf abgearbeitet werden?

Dieser Frage ist bei einigen physikalischen Experimenten kritisch, da sich das zu untersuchende System nach einem Neustart oder nach einer gewissen Wartezeit verändert und danach eine Absolutwertverschiebung aufweist (siehe auch Messmittelanalyse und Datenqualität). In diesen Fällen ist es vorteilhaft, die Messungen in einem ununterbrochenen Experiment durchzuführen, so dass eine saubere Abbildung des grundlegenden Systemverhaltens (Trend) möglich ist. Wird die Messung hingegen auf mehrere Teilabschnitte mit zufälliger Verschiebung des Systemmittelwerts aufgeteilt, so ist nicht nur die Vorhersage eines genauen Absolutwerts sondern ebenfalls des grundlegenden Systemverhaltens schwierig.

11. Sind bereits Vorkenntnisse vorhanden?

In den wenigsten Fällen wird ein komplett unbekanntes System analysiert. Eine Aufgabe ist es, vorhandenes Wissen zu sammeln und in die Versuchsplanung und anschließenden Analysen zu integrieren. Informationen aus allen beteiligten Bereichen sind dabei hilfreich.

- Existieren bereits Daten und Analysen aus vorherigen Projekten?

- Welche Faktorkombination und -bereiche sind nicht sinnvoll oder vorteilhaft?
- Sind bereits grundlegende Zusammenhänge zwischen Faktoren und Qualitätsgrößen bekannt?
- Gibt es besonders komplexe oder interessante Bereiche, die feiner aufgelöst werden müssen?
- Haben sich spezielle Analyse- und Optimierungsalgorithmen für ähnliche Aufgaben bewährt?

Die meisten hier dargestellten und in der Praxis eingesetzten Verfahren gehen davon aus, dass keine Vorkenntnisse über das zu analysierende System vorhanden sind. Liegen jedoch Vorkenntnisse vor, können die allgemeinen Verfahren an die gegebenen Informationen und Randbedingungen angepasst werden, was die Effizienz erhöht. Jede Änderung muss dabei genau durchdacht und kommuniziert sowie in allen folgenden Prozessschritten berücksichtigt und eventuell integriert werden. Wird zum Beispiel ein Bereich eines Faktorraums herausgeschnitten, muss dieser mittels einer zusätzlichen Randbedingung im Analyse- und Optimierungsschritt ausgeschlossen werden. Weiterhin sind allgemeine Algorithmen, die eine Variation über den gesamten Faktorraum voraussetzen (z.B. globale Sensitivitätsanalyse), nur bedingt aussagekräftig und anzupassen.

12. Ist ein iteratives Vorgehen sinnvoll?

Im Rahmen dieser Frage ist zu klären, ob alle Projektaufgaben mit einem Versuchsplan und den darauf aufbauenden Analysen zu bearbeiten sind. Dieses beinhaltet zum Beispiel die Frage, ob alle Randbedingungen bereits soweit eingegrenzt werden können, dass ein Versuchsplan sinnvoll aufgebaut werden kann. Verschiedene Projektaufgaben wie Screening über einen großen Faktorbereich und eine hohe Faktanzahl sowie eine Optimierung in einem kleinen Faktorbereich mit ausgewählten Hauptfaktoren sollten nicht mit einem Versuchsplan und einer Analyse bearbeitet werden. Hier ist es sinnvoller einen groben Screeningplan zu verwenden und Teilergebnisse für eine zweite detaillierte Analyse zu verwenden. In anderen Fällen ist die wahre Komplexität des zu untersuchenden Systems nicht bekannt, so dass es sinnvoller ist, einen weit gefassten aber groben Versuchsplan aufzustellen, mit dem erste Analysen und Optimierungen durchgeführt und am realen System geprüft werden. Werden dabei besonders interessante Bereiche oder wichtige Regionen mit schlechter Approximationsgenauigkeit erkannt, kann in diesen Faktorunterräumen die Datenmenge durch weitere Stützstellen erhöht werden. Weitere Analysen werden dann auf einer je nach Anwendung sinnvollen Kombination aus vorhandenen und neuen Daten durchgeführt.

13. Wie hoch ist der Automatisierungsgrad?

Ein hoher Automatisierungsgrad ist in allen Prozessphasen sinnvoll. Gerade die automatisierte Mess- oder Simulationsdurchführung sowie Datenaufnahme und -kontrolle ist entscheidend, da dadurch bereits viele Fehlerquellen ausgeschlossen werden können. Fehlerhafte Daten, die durch falsche Faktoreinstellungen oder fehlerhaftes Ablesen von Signalen entstehen, lassen sich später nur schwer finden und führen schnell zu falschen Schlussfolgerungen. Eine umfangreiche

zeitliche und finanzielle Investition in einen hohen Automatisierungsgrad zahlt sich typischerweise im Laufe eines oder während später folgenden Projekten aus, da fehlerhafte Daten zu umfangreichen Wiederholungsmessungen oder sogar zum Scheitern eines Projektes führen. Eine Automatisierung der Modellgenerierung, Analyse und Optimierung ist ebenfalls voranzutreiben, da auch hier manuelle Fehler fatale Folgen haben und später nur schwer nachzuvollziehen sind. Die investierten Ressourcen zahlen sich auch hier schnell aus, da notwendige Wiederholungen einzelner Prozessschritte zwar Zeit benötigen, diese dann aber automatisiert und ohne große Personalressourcen durchgeführt werden. Erneute Analysen sind beispielsweise notwendig, wenn neue oder erweiterte Daten vorhanden sind, neue Projektziele und Randbedingungen zum Einsatz kommen oder sich logische Fehler in den Analyseprozess eingeschlichen haben, die in einem automatisierten Prozess deutlich schneller zu ändern sind als bei erneuter manuellen Analyse.

Bei allen Vorteilen der Automatisierung besteht die Gefahr, dass den Ergebnissen irgendwann blind vertraut wird. Jedes Ergebnis muss weiterhin mit Fachwissen und einfachen logischen Tests kritisch hinterfragt werden.

14. Ist eine ausreichende Dokumentation des Prozesses, der Daten und der Ergebnisse sichergestellt?

In einigen Projekten wird das Vorgehen nur unzureichend dokumentiert und die verwendeten Daten unstrukturiert gespeichert. Dies macht es schwierig oder unmöglich zu einem späteren Zeitpunkt den Prozess nachzuvollziehen oder zu beurteilen sowie die Daten für ein weiteres Projekt sinnvoll zu verwenden. Grundsätzlich sollten alle Prozessschritte, verwendete Algorithmen und Daten so dokumentiert werden, dass diese auch zu einem späteren Zeitpunkt von fachkundigen Personen nachvollziehbar und Algorithmen sowie Daten wiederverwendbar sind. Daher ist es notwendig, die Form der Dokumentation und Datenspeicherung frühzeitig zu besprechen. Idealerweise wird die Dokumentation direkt in den automatisierten Prozess implementiert. In der Praxis zeigt sich immer wieder, dass ein guter Vorsatz, die Dokumentation am Ende des Projekts durchzuführen, nur unzureichend oder nicht umgesetzt wird. Entweder sind nicht mehr alle Informationen vorhanden beziehungsweise bereits in Vergessenheit geraten oder eine neue und wichtigere Aufgabe steht an.

15. Ist eine ausreichende Mess- und Arbeitsbeschreibung vorhanden?

In die Versuchsplan- und Modellestellung sowie Analyse und Optimierung wird meistens viel Zeit investiert. Auf eine detaillierte Beschreibung der Versuchsdurchführung und Messaufgaben wird hingegen oft verzichtet. In der Praxis führt dieses zu Problemen, da unterschiedliche Personen für Projekt- und Versuchsplanning sowie Versuchsvorbereitung und -durchführung verantwortlich sind. Eine einfache Temperaturbestimmung eines Heizkörpers führt dadurch schnell zu nicht vergleichbaren oder unerwarteten Daten. Es ist zum Beispiel nicht klar, ob eine durchschnittliche Temperatur (Ort bzw. Zeit) oder an einer bestimmten Position am Heizkörper gemessen werden soll. Weiterhin kann die Oberflächentemperatur mit oder ohne Lackierung oder die Temperatur im Material selbst gewünscht sein. Das Messmittel ist ebenfalls nicht spezifiziert, so können unter-

schiedliche Messgeräte verschiedene Messungenauigkeiten oder Referenzpunkte aufweisen. Es ist oft erstaunlich, auf wie viele verschiedene Arten ein Messwert aufgenommen werden kann, obwohl ein Auftraggeber der Meinung ist, dass die Messung genau spezifiziert wurde. Dadurch treten schnell unerklärliche Phänomene und Ergebnisse auf, die fälschlicherweise auf verschiedenen Prüfständen oder Simulationsmodelle zurückgeführt werden.

16. Haben Begriffe unterschiedliche Bedeutungen?

Verwirrungen und Zeitverzögerungen treten auf, wenn Mitarbeiter unter gleichen oder ähnlichen Begriffen unterschiedliche Dinge verstehen. Ein "Modell" kann für den einen ein komplexes Simulationsmodell sein oder ein Metamodell, welches auf Basis von Simulationsdaten erstellt wurde. Andere bezeichnen damit hingegen ein vereinfachtes Steuerungsmodell eines Prüfstands. Zum reibungslosen Ablauf eines Projekts sollten daher die Begriffsdefinitionen geklärt werden.

17. Sind Personen zu trainieren?

Zur Durchführung eines Projekts sind meist viele Personen notwendig, von denen sich einige noch nie oder nur selten mit Versuchsplanung und Analysen beschäftigt haben. Zur erfolgreichen und reibungsfreien Durchführung aller Arbeiten ist es jedoch notwendig, dass alle beteiligten Personen das grundlegende Vorgehen und alle kritische Details kennen und verstehen. Aus diesem Grund sollten im Vorfeld Trainings für beteiligte Personen durchgeführt werden, wobei sich Inhalte und Detaillierungsgrad jeweils an den speziellen Aufgaben orientieren. So ist für einen Prüfstandbetreiber oder Simulant wichtiger zu verstehen, welche Auswirkungen zum Beispiel vertauschte Faktoreinstellungen auf den Analyseprozess haben als für Projektleiter, die eher den Gesamtprozess und Ressourcen koordinieren müssen.

Nachdem grundlegende Informationen in der Vorbereitungsphase gesammelt und diskutiert wurden, können die folgenden Phasen zielstrebig bearbeitet werden.

15.2 Versuchsplan erstellen

Mit den Informationen über Faktoren und Qualitätsgrößen wird ein geeigneter Versuchsplan erstellt oder gewählt. Zuerst muss jedoch entschieden werden, ob ein klassischer Versuchsplan oder eher ein raumfüllendes Design eingesetzt werden soll. Bei Verwendung von Computerexperimenten (Kapitel 7), potentiell insignifikanten Faktoren (Kapitel 7.2.2) oder komplexen Systemzusammenhängen ist ein raumfüllendes Design meist vorteilhaft. Unabhängig von der gewählten Form wird im ersten Schritt ein normierter Versuchsplan erstellt, der in einem späteren Schritt mittels definierter Faktorgrenzen in reale Faktoreinstellungen umgewandelt wird. Die Optimierung von normierten Versuchsplänen hat den positiven Nebeneffekt, dass diese auch in spätere Projekte einsetzbar sind. Die eingesetzte Qualitätsgröße zur Optimierung der Versuchspläne weist eine untergeordnete Rolle auf (Kapitel 8). Raumfüllende Versuchspläne weisen keine Faktorkombinationen in den Ecken des Faktorraums auf. Ist bereits bekannt, dass diese extremen Einstellungen für eine sinnvolle

Analyse notwendig sind, sollten diese Faktorkombinationen dem Versuchsplan manuell hinzugefügt werden.

Gerade bei physikalischen Experimenten und raumfüllenden Versuchsplänen ist zu berücksichtigen, dass nicht immer alle Faktoreinstellungen sinnvoll sind und eventuell alternative Einstellungen notwendig sind oder eine Verringerung der Stufenzahl. Rohrleitungen sind zum Beispiel nur in bestimmten Durchmessern lieferbar, so dass es wenig Sinn macht, einen Rohrdurchmesser mit $d = 5.23433234\text{ mm}$ festzulegen, wenn dieser nur in 5 und 6 mm zu liefern ist.

Zur Beurteilung der Messstabilität ist ein zusätzlicher wiederkehrender Referenzpunkt in den Versuchsplan einzubringen (Kapitel 8.7). Weiterhin werden zur späteren Validierung von Metamodellen zufällige Faktorkombinationen in den Versuchsplan eingestreut, die nicht zum Training der Metamodelle verwendet werden (Kapitel 9.18). Wenn die Voranalyse der Faktoren Besonderheiten, wie nicht einstellbare Faktorkombinationen ergeben, müssen diese im Testplan berücksichtigt werden oder geklärt werden, wie diese im späteren Versuchsablauf behandelt werden.

Eine allgemeingültige Empfehlung für die benötigte Versuchsanzahl ist im Gegensatz zur klassischen DoE nur schwer möglich. Erste Hinweise ergeben sich jedoch bereits aus der klassischen Versuchsplanung, wobei hier eher einfache Modellzusammenhänge betrachtet werden. Bei komplexeren Zusammenhängen sind diese Größen eher als untere Grenze zu betrachten. Auch die Anzahl der maximal pro Qualitätsgröße signifikanten Faktoren beeinflusst die ideale Versuchsplangröße (Kapitel 7.2.2). Die zur Verfügung stehende Bearbeitungszeit sowie die Finanz- und Prüfstandressourcen erweisen sich oft als die treibenden Parameter für die maximale Versuchsplangröße, wobei die dadurch eventuell auftretenden Einschränkungen zum Beispiel in der erwartenden Modellkomplexität frühzeitig an alle Beteiligten kommuniziert werden muss. Neben der Erzeugung eines Versuchsplans, der im gesamten Faktorbereich bereits eine maximale Informationsmenge liefert, existieren Anwendungen, bei denen nach und nach die Datenmenge im gesamten oder in Unterbereichen des Faktorraums verbessert werden sollen. Kann die Gesamtanzahl der benötigten Versuchsläufe nicht sicher bestimmt werden, so können Sequenzen zur Erstellung eines ersten Plans verwendet werden (Kapitel 8.3.1). Zeigt sich im Laufe des Projekts, dass die Datenmenge im gesamten Faktorraum nicht ausreichen, können einfach weitere Testkombinationen der Sequenz verwendet werden. Soll hingegen im ersten Schritt eine grobe Übersicht des zu untersuchenden Systems ermittelt werden und im zweiten Schritt eine Verfeinerung in interessante oder schlecht abgebildete Faktorbereiche, so sind mehrere Pläne mit unterschiedlichen Faktorbereichen zu kombinieren. Im Idealfall werden die neuen Pläne so ausgelegt, dass sich die neuen Faktorkombinationen so in den ursprünglichen einfügen, dass weiterhin eine gute Gleichverteilung erreicht wird. Sind bereits durch die Planung Faktorbereiche mit höherer Komplexität bekannt, ist es sinnvoll, bereits im ersten Versuchsplan diesen Bereich mit einer höheren Testpunktedichte auszustatten.

Umso mehr Informationen über die Faktoren und Qualitätsgrößen vorhanden sind, desto besser können die Versuchspläne an die spezielle Aufgabe angepasst werden, wobei diese Anpassungen in den späteren Prozessschritten berücksichtigt werden müssen. In den meisten Fällen ist es jedoch ausreichend, Standardpläne zu verwen-

den, die insgesamt vielleicht eine etwas geringere Qualität liefern, aber dafür viel Anpassungsarbeit ersparen. In einigen Fällen ist es für die Versuchsdurchführung einfacher, von der zufälligen Reihenfolge der Faktorkombinationen abzuweichen und einzelne Faktoren in Gruppen (kleine und große Werte) zusammenzufassen oder zu sortieren. Dieses ist in jedem Fall zu verhindern, da dadurch zeitliche Veränderungen des Prüfstands (Drift) nicht mehr von Effekten des sortierten Faktors unterscheidbar sind.

15.3 Experiment (Messung und Rechnung)

In diesem Prozessschritt werden das zu untersuchende System mit vorgegebenen Faktorkombinationen betrieben und die Qualitätsmerkmale bestimmt. Durch eine ausreichende Vorbereitung und Planung sollte hier bereits die zu erwartende Datenqualität und eine detaillierte Definition der Datenaufnahme bekannt sein. Eine stabile Automatisierung der Messung und Simulation sowie der Datenspeicherung ist ein entscheidender Erfolgsfaktor. Gerade bei physikalischen Experimenten ist neben der automatischen Speicherung der gemessenen Qualitätsgrößen die Messung und Speicherung der real verwendeten Faktoreinstellungen notwendig, da diese zur weiteren Analyse verwendet werden müssen.

15.4 Kontrolle der Daten

Bevor ermittelte Daten in weitere Prozessschritte eingesetzt werden, müssen diese auf ihre Verwendbarkeit geprüft werden. Im ersten Schritt ist dabei zu klären, ob wirklich alle Messpunkte aufgenommen und die gewünschten Faktoreinstellungen eingehalten wurden. Insbesondere wenn die Faktorgrenzen sehr groß werden, sind einige Kombinationen nicht einstellbar. Alternativ können dann einstellbare Faktorkombinationen in der Nähe der ursprünglichen Faktorkombination gewählt werden, was bei einer späteren Analyse Informationen über die Einstellgrenzen liefert. In anderen Fällen ist es nicht möglich Faktoren exakt einzustellen. Unabhängig von den Ursachen für eine Faktoränderung ist es notwendig die realen Einstellungen zu kennen. Es ist dann zu klären, ob die Daten noch immer ausreichen, um die gewünschte Aufgabe zu bearbeiten. Dabei ist die Korrelation der Faktoren erneut zu prüfen und dass alle notwendigen Faktorbereiche enthalten sind. Parallel dazu sind die Qualitätsgrößen zu betrachten. Fehlen beispielsweise in einigen Datensätzen einzelne Qualitätsgrößen, muss entschieden werden, ob der Datensatz gelöscht oder der Versuch wiederholt werden muss. Alternativ können auch unterschiedliche Datensätze für verschiedenen Gruppen der Qualitätsgrößen erzeugt werden, was bei einer späteren parallelen Analyse aller Qualitätsgrößen eventuell Berücksichtigung finden muss. Einfache Scatterplots, bei denen die Qualitätsgrößen über die Versuchsreihenfolge aufgetragen werden, ermöglichen bereits eine einfache erste Beurteilung

der Datenqualität. Die Variation der Qualitätsmerkmale muss dabei über den Versuchszeitraum auf Grund der zufälligen Versuchsreihenfolge einigermaßen konstant bleiben. Weiterhin werden Messungen mit stark abweichenden Werten schnell erkannt (Kapitel 8.7). Bevor auffällige Daten manipuliert oder gelöscht werden, muss immer die Ursache für das auffällige Verhalten geklärt werden. Stark abweichende Absolutwerte von vereinzelten Messpunkten können neben Fehlmessungen ebenfalls von speziellen Faktorkombinationen erzeugt werden. Eine voreilige Löschung des Datensatzes würde dann einen besonders interessanten Faktorbereich aus den Daten entfernen. Stellt sich bei einer genaueren Betrachtung heraus, dass es sich zwar um keine Fehlmessung handelt, aber die spezielle Faktorkombination nicht für die Analyse relevant ist, kann die Messung zur Stabilisierung der nachfolgenden Analyse und Modellbildung entfernt werden. Dabei muss die neue Struktur des Datensatzes in den folgenden Schritten berücksichtigt werden. Eine Approximation im entfernten Faktorbereich ist dann nicht mehr zulässig. Wurde ein wiederkehrender Referenzpunkt in den Versuchsplan eingebracht (Kapitel 8.7), kann durch dessen Streuung eine einfache Abschätzung der Messvarianz und der Prüfstandsdrift ermittelt werden.

15.5 Erzeugung von Metamodellen

Auf Basis der geprüften Daten werden Metamodelle zur Abbildung der Qualitätsgrößen erzeugt. Dabei werden nicht die geplanten sondern die im Experiment real eingesetzten Faktoreinstellungen verwendet. Die Wahl des passenden Metamodellsatzes hängt von verschiedenen Randbedingungen ab. Zuerst müssen ausreichend und passende Daten zur Erzeugung eines gewählten Modelltyps vorhanden sein. Ein quadratisches Modell ist zum Beispiel nicht mit zwei Stützstellen zu trainieren. Zusätzlich kann ein komplexer Zusammenhang in einem Faktorbereich nur soweit aufgelöst werden, wie beschreibende Daten/Stützstellen vorhanden sind. Es ist zu klären, ob die benötigte Modellkomplexität sinnvoll abschätzbar ist. Sind nur einfache grundlegende Zusammenhänge abzubilden, sollte sich auch der verwendete Modellsatz daran orientieren. Flexible adaptive Modelle können sich zwar ebenfalls an einfache Zusammenhänge anpassen, bergen aber das Risiko, sich an das Rauschen anzupassen (Overfitting, Kapitel 9.16) oder komplexe Zusammenhänge in die Daten zu finden, die nicht vorhanden sind (Kapitel 9.16 neuronale Netzwerke Beispiel). In einigen Anwendungsfällen, bei denen die Komplexität im Vorfeld nicht bekannt war, hat sich ein gestufter Ansatz als vorteilhaft erwiesen. Im ersten Schritt wird ein einfaches oder robust parametrisiertes Modellverfahren verwendet, um eine Abbildung des grundlegenden Systemverhaltens zu ermöglichen. Wenn sich bei weiterer Analyse herausstellt, dass die Modellgüte nicht ausreichend ist und basierend auf den gegebenen Daten und deren Qualität eine Steigerung möglich ist, wird die Flexibilität des Modellsatzes erhöht. Die Approximationsqualität der meisten adaptiven Modellsätze (Kapitel 9), welche *keine* festen Vorgaben über die abzubildenden Zusammenhänge benötigen, ist bei richtiger Parameterwahl

ausreichend. Auf den Arbeitsaufwand, das absolut beste Modell zu ermitteln, kann meistens verzichtet werden. Die gewonnene Zeit sollte besser für einzelne problematische schwer abzubildende Fälle aufgespart werden oder wenigstens auf wichtige Bereiche des Faktorraums oder Qualitätsraum reduziert werden. Ist zum Beispiel die Leistungsaufnahme eines Systems zu reduzieren, so ist lediglich die Vorhersagequalität eines Metamodells bei kleinen Leistungen wichtig, da es bei hohen Leistungsaufnahmen irrelevant ist, ob die Aufnahme zwanzig oder vierzig mal höher ist als bei einer optimalen Systemeinstellung.

15.6 Kontrolle der Metamodelle

Die Qualitätsanalyse von Metamodellen geht Hand in Hand mit deren Erstellung und liefert grundlegende Informationen über die zu erwartende Approximationsqualität des Metamodells. Die Approximationsgüte eines Metamodells ist eine Kombination aus gegebener Datenqualität (Mess- und Simulationsqualität) und des mathematischen Abbildungsverfahrens. Die Approximationsgenauigkeit von Trainings- und Validierungsdaten sollte dabei nicht die gegebene Datenqualität überschreiten, da sich das Modell dann zum Beispiel an das Messrauschen anpasst. Die Messgenauigkeit ist bei späterer Prüfung optimierter Faktoreinstellungen am untersuchten System zu erwarten. Abweichungen zwischen realen Messwerten und zugehörigen Modellvorhersagen im Bereich der Messgenauigkeit sind nicht auf die Güte des mathematischen Abbildungsverfahrens zurückzuführen, welches sinnvollerweise Messrauschen aus der Approximation entfernt.

Die Approximationsgenauigkeit adaptiver Metamodellverfahren ist nur schwer durch klassische Verfahren, wie Analyse der Residuen von Trainingsdaten, sinnvoll zu beurteilen. Eine ungewollte Anpassung an zum Beispiel Messrauschen kann nicht ausgeschlossen werden. Aus diesem Grund wurden im Versuchsplan verschiedene zufällige Faktorkombinationen zur Validierung eingefügt, welche nicht zum Training der Metamodelle verwendet wurden. Ist die Approximationsabweichung der Validierungsdaten in einer ähnlichen Größenordnung wie die der Trainingsdaten, ist dies ein erstes Indiz für ein stabiles Modell. Falls die Abweichung der Validierungsdaten deutlich größer ist, so kann davon ausgegangen werden, dass hier ein fehlerhaftes Metamodell erzeugt wurde. Es hat sich an Messrauschen angepasst oder es lagen nicht genügend Trainingsdaten vor, um alle Faktorbereiche ausreichend genau abzubilden. Die Approximationsgenauigkeit sollte dabei immer in Relation zur vorliegenden Datengenauigkeit beurteilt werden, da sich diese in der Residuenanalyse wiederfinden wird.

15.7 Analyse der Daten und Metamodelle

Nachdem kontrollierte Daten und Metamodelle vorliegen, können diese umfangreich ausgewertet werden. In einigen Fällen sind bereits die vorliegenden Daten vollkommen ausreichend, um erste oder alle notwendigen Analysen durchzuführen. Somit ist manchmal eine Modellerstellung überhaupt nicht mehr notwendig oder es können nur noch vereinzelte gezieltere Modelle erzeugt werden. Die vorliegenden Daten sind im Idealfall im gesamten interessanten Faktorraum erzeugt worden und beschreiben das System bereits sehr gut. Sind beispielsweise nur einige vorteilhafte Faktorkombinationen für gewählte Gütekriterien gesucht, so können diese aus den vorhandenen Daten ohne Modellbildung entnommen werden. Weiterhin ist es möglich, grundsätzliche Korrelations- oder Sensitivitätsanalysen zwischen Faktoren und Qualitätsgrößen durchzuführen. Nicht einstellbare oder kritische Faktorbereiche für das zu untersuchende System lassen sich ebenfalls analysieren, was eine Beschreibung der wahren Faktorgrenzen ermöglicht. Diese sollten dann in den folgenden Prozessschritten berücksichtigt werden. Erste Aussagen und Abschätzungen der Messqualität und Systemstabilität sind möglich.

Werden Metamodelle aus den Daten erzeugt, werden diese typischerweise für umfangreiche Detailanalysen verwendet. Hierzu zählt die Bestimmung von Haupt- und Interaktionseffekten oder lokale und globale Sensitivitätsanalysen. Durch die geringe Berechnungszeit der Metamodelle im Bereich von Millisekunden bis Sekunden werden mit geringem Rechenaufwand verschiedene Faktoreinstellungen verglichen oder der Einfluss von Änderungen direkt (online) untersucht. Gerade die Möglichkeit einer direkten Visualisierung der Qualitätsänderungen auf Basis zugehöriger Faktoränderungen ermöglicht ein detailliertes Verständnis des untersuchten Systems. Zur Beurteilung der lokalen Systemstabilität (Robustheit) werden im einfachsten Fall die Faktoren in kleinen Grenzen um die zu untersuchende Faktoreinstellung variiert und die Änderung der Qualitätsgrößen analysiert.

Alle Extrapolationen außerhalb der Trainingsgrenzen sind kritisch und dürfen nur mit erhöhter Aufmerksamkeit verwendet werden. Sie geben eine Abschätzung, wie sich das System verhalten könnte, wobei die Qualität der Approximationen geprüft oder separat beurteilt werden muss. Konnten nicht alle Faktorkombinationen gemessen werden und hat sich dadurch der betrachtete Faktorraum verkleinert, ist häufig die Berechnung einer *konvexen* Hülle zur Bestimmung des erlaubten Faktorraums sinnvoll. Mit dieser Hülle ist für neue Faktorkombinationen zu beurteilen, ob es sich um eine Extrapolation handelt. Ist die eigentliche Hülle um die Faktoren nicht konvex sondern in einigen Bereichen konkav, wird dieses nicht alle kritische Bereiche aufdecken können. Die konvexe Hülle ist jedoch im Gegensatz zur konkaven Hülle eindeutig und einfach zu bestimmen und meist zur ersten Abschätzung ausreichend. Bei allen Analysen mit Metamodellen ist immer die Approximationsgenauigkeit der Modelle zu beachten, welche oft durch das vorhandene Rauschniveau und der Drift von Messungen bestimmt wird. Treten zum Beispiel in der Analyse unlogische Ergebnisse auf, wie beispielsweise negative Abstände, müssen diese in Relation zur Modellgenauigkeit betrachtet werden. Die Vorhersage einer Distanz von $d = -0.01\text{ mm}$ ist zwar nicht sinnvoll, aber in Relation zu einer Messgenauigkeit

der Trainingsdaten von vielleicht 1 mm irrelevant. Mit diesem Ergebnis ist einfach ein Wert um Null innerhalb der Approximationsgenauigkeit zu erwarten.

15.8 Optimierung

Neben einer Analyse des Systems ist in Projekten häufig eine Optimierung verschiedener Qualitätsgrößen notwendig. Sind mehrere Qualitätsgrößen zu optimieren ist es sinnvoll, einen Algorithmus für mehrere Qualitätsgrößen zu verwenden. Jede Zusammenfassung der Qualitätsgrößen zu einer globalen Qualitätsgröße beinhaltet eine im Vorfeld definierte Gewichtung oder Abhängigkeit der Qualitätsgrößen, die das Optimierungsergebnis beeinflusst. Die Zusammenfassung verringert die Pareto-Grenze auf eine einzige Lösung. Da im Vorfeld der Optimierung die Bandbreite der Pareto-optimalen Lösungen und die dazugehörigen Faktoreinstellungen nicht bekannt sind, kann eine sinnvolle Lösung bei mehreren Qualitätsgrößen erst nach Bestimmung der gesamten Pareto-Grenze ermittelt werden. In einigen Fällen finden sich auf der Pareto-Grenze Lösungen, die sich nur geringfügig in den Qualitätsgrößen unterscheiden, aber durch deutlich unterschiedliche Faktoreinstellung ermöglicht werden. In diesen Fällen hat der Anwender nur dann die Möglichkeit die Lösung mit *besseren* Faktoreinstellungen zu wählen, wenn die Lösung auf der Pareto-Grenze bekannt ist.

Im ersten Schritt der Optimierung sollte vergleichbar mit der Vorgehensweise bei der Metamodellerzeugung ein stabiler Algorithmus verwendet werden, der mit vielen Optimierungsproblemen zurechtkommt. Gute Erfahrungen liegen mit abgewandelten NSGA2 Algorithmen vor, die Ergebnisse in meist akzeptabler Rechenzeit liefern. Meist ist nicht entscheidend, dass ein Ergebnis in minimaler Rechenzeit erzielt wird, sondern eher, dass die wahre Pareto-Grenze bestimmt wird und der Algorithmus nicht in lokalen Optima stecken bleibt. Hohe Steigerungen in den Algorithmeffizienzen lassen sich besonders dann erzielen, wenn spezielle Eigenschaften des Optimierungsproblems berücksichtigt werden. Diese Spezialalgorithmen, welche nur noch sinnvoll für das dazugehörige Problem zu verwenden sind, können Rechenzeiten von mehreren Stunden auf Minuten reduzieren. Dieser Aufwand ist jedoch nur sinnvoll, wenn es sich um eine häufig wiederkehrende Aufgabe handelt oder das Optimierungsproblem mit Standardalgorithmen nicht zu lösen ist.

15.9 Prüfung der Analyse und Optimierungsergebnisse

Jedes Analyse- und Optimierungsergebnis muss schlussendlich am untersuchten System geprüft werden. Erst damit wird die Richtigkeit der Ergebnisse endgültig bewiesen. Abweichungen beim Abgleich zwischen Approximationsergebnissen und Prüfung der Ergebnisse sind aus bereits erläuterten Gründen zu erwarten. Neben Rauschen oder Drift in den Trainingsdaten ist ein vom Algorithmus verursachter

Approximationsfehler möglich. Die Kontrolle am untersuchten System wird ebenfalls durch Rauschen gestört, was zu merklichen Abweichungen führen kann. Je nach Projektziel und abhängig von der vorausschauend durchgeführten Planungsphase ist dieses mehr oder weniger problematisch. Sind zum Beispiel die besten Faktoreinstellung für eine Qualitätsgröße gesucht, ist der Absolutwert nicht entscheidend, da die Faktoreinstellungen noch immer das beste Gesamtergebnis für das analysierte System liefern. Das Ergebnis liegt lediglich auf einem anderen Absolutniveau. Soll innerhalb des Projekts hingegen ein grundlegendes Systemverständnis aufgebaut werden und das Optimierungsergebnis als Basis für ein anschließendes manuelles Feintuning dienen, so sind auch hier kleine Abweichungen keine entscheidenden Einschränkung. In Fällen, bei denen Absolutwerte kritisch sind, werden zur Sicherheit bei der Optimierung Sicherheitsmargen im Bereich der Vorher sagegenauigkeit eingeplant. Sollte eine Überprüfung zeigen, dass Bereiche des Faktorraums nicht ausreichend genau abgebildet werden, können die neuen Datensätze zur Verbesserung der vorhandenen Metamodelle eingesetzt oder der Bereich mit weiteren Messungen lokal verfeinert werden.

15.10 Dokumentation

Die Dokumentation wird wie die Planung am häufigsten vernachlässigt, was sich später im Projekt oder in Folgeprojekten negativ auswirkt. Grundsätzlich sollten alle Prozessschritte von Planung, Testplangenerierung, Messung, Analyse, Modellbildung, Optimierung bis hin zur Ergebnisüberprüfung umfangreich und wenn möglich automatisch dokumentiert werden. Dieses ermöglicht alle Daten und Ergebnisse auch nach einer gewissen Zeit nachzuvollziehen oder wiederzuverwenden. Sehr unangenehm ist es, wenn zu einem späteren Zeitpunkt nicht mehr glaubhaft dargestellt werden kann, wie präsentierte Ergebnisse entstanden sind oder vorhandene Daten und Modelle wegen fehlender Dokumentation nicht in neuen Projekten eingesetzt werden können. In diesen Fällen werden viele Arbeiten unnötigerweise wiederholt und Ressourcen verschwendet. Gerade bei physikalischen Experimenten sind die zu untersuchenden Systeme oft nicht mehr vorhanden, so dass eine neue Datenerzeugung nicht mehr möglich ist.

Sind die Daten, Modelle und Analysen nach den hier dargestellten Verfahren und Prozessen generiert, sollten diese in einer strukturierten Art vorliegen, so dass bei einer angemessenen Dokumentation einer Wiederverwendung oder Erweiterung nichts entgegen spricht.

15.11 Schlusswort

Die dargestellten Verfahren und Algorithmen sind kein Allheilmittel für jede Projektaufgabe und sollten nie zum Selbstzweck eingesetzt werden. Häufig ist der Einsatz

aller oder ausgewählter Verfahren aber sinnvoll und liefert umfangreiche und deutlich mehr Ergebnisse bei begrenzten Ressourcen als beim Einsatz unstrukturierter Analyseverfahren.

Der Methodeneinsatz sollte dabei nicht zum Automatismus werden auch wenn das in einigen Fällen angestrebt wird (Ergebnis auf Knopfdruck). Alle Ergebnisse und Prozessschritte sollten wie bei jeder sinnvollen Projektdurchführung hinterfragt und kurz geprüft werden. Ein Anwender der Methoden muss grundsätzlich wissen was er tut, auch wenn es nicht nötig ist, jedes mathematische Detail zu verstehen. Er sollte daher die Grundlagen und die Unterschiede zwischen den Verfahren kennen sowie Auswirkungen und Bedeutungen der Algorithmus-Parameter beurteilen können.

Die Kapitel dieses Buches sind so konzipiert, dass Grundlagen der Verfahren kurz dargestellt werden, auch wenn sie beim ersten Durchlesen manchmal komplex erscheinen. Sie bieten jedoch eine Grundlage, mit weiterführender Literatur tiefer in die Materie einzusteigen. Dabei wird versucht, einen Kompromiss zwischen mathematisch korrekter Darstellung und einer grundsätzlichen Verständlichkeit für Nicht-mathematiker bereitzustellen. Wenn möglich finden sich in jedem Kapitel Gleichungen und Pseudo-Algorithmen, die es ermöglichen sollten, mit überschaubarem Aufwand in beliebigen Programmiersprachen, welche mathematische Basispakete enthalten, grundlegende Algorithmen abzubilden und zu testen. Die dargestellten Verfahren zeigen einen Ausschnitt der Möglichkeiten und erheben keinen Anspruch auf Vollständigkeit oder die besten Ergebnisse für jede Anwendung zu liefern. Viele Verfahren haben sich jedoch in der Praxis bewährt und sollten für einen fundamentalen Einstieg und die meisten Anwendungen ausreichen. Weiterhin bieten sie eine Grundlage, kommerzielle Softwarepakete zu verstehen und gezielt zu bedienen.

Anhang A

Berechnungsmodell zum Fallbeispiel Rasensprenger

An warmen Sommertagen gesellt sich zur Bewässerungsfunktion des Rasensprengers noch die Nebenfunktion der Kinderbelustigung. Bei der dann angestrebten langen Nutzungsdauer gelangt zu viel Wasser auf den Rasen. Insgesamt lassen sich drei unabhängige Qualitätsmerkmale identifizieren: große Reichweite, hohe Drehzahl und geringer Wasserverbrauch. Betrachtet wird das System *Rasensprenger* ab Zuleitung hinter dem Absperrhahn.

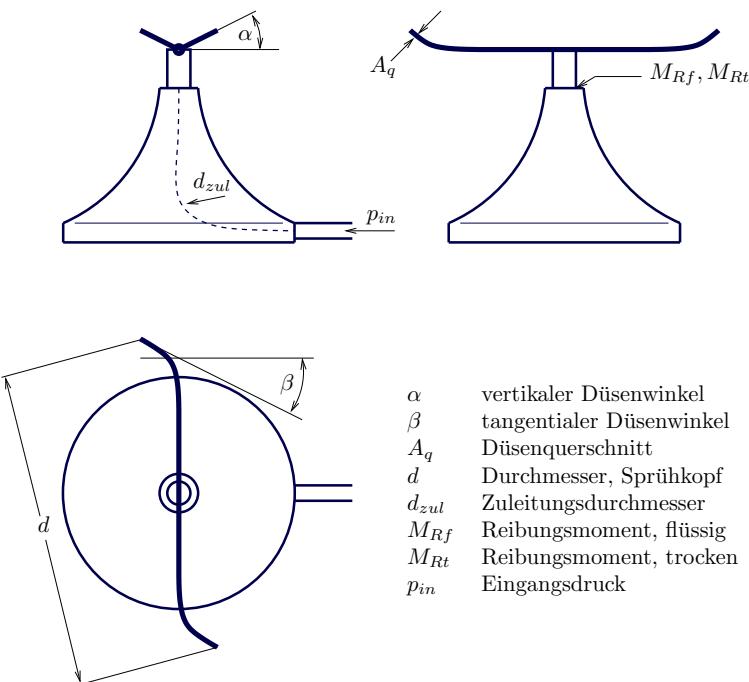


Abb. A.1 Schematische Darstellung eines Rasensprengers.

A.1 Nomenklatur

α	vertikaler Düsenwinkel
β	tangentialer Düsenwinkel
η_L	dynamische Viskosität der Luft
λ	Reibungskoeffizient
v	kinematische Viskosität
ζ	Widerstandskoeffizient
φ	Winkel des Geschwindigkeitsvektors, bezogen auf die Horizontale
ρ, ρ_L, ρ_W	Dichte
ω	Winkelgeschwindigkeit des Rasensprengerarms
a, a_h, a_v	Beschleunigung, mit horizontaler und vertikaler Komponente
A_q	Düsenquerschnitt
A_{zul}	Querschnitt des Zulaufs
c_v	Widerstandsbeiwert, Druckverlust in der Zuleitung
d	Durchmesser des Rasensprengers
d_{Tr}	Tropfendurchmesser
d_{zul}	Zuleitungsdurchmesser, von Wasseranschluß bis Düse
F_W	Kraft auf einen Tropfen durch den Luftwiderstand
m_{Tr}	Tropfenmasse
M_A	Antriebsmoment
M_R, M_{Rf}, M_{Rt}	Reibungsmoment, mit Anteilen von flüssiger und trockener Reibung
n	Drehzahl des Rasensprengers
p_{en}	effektiv treibender Druck
p_{in}	Eingangsdruck
$\Delta p_v, \Delta p_{zul}$	Druckverluste durch Reibung
\dot{Q}	gesamter Volumenstrom
R	Radius eines Rasensprengerarms
Re	Reynoldszahl
s_h, s_v, s_{h0}, s_{v0}	hor. und vert. Komponente der Tropfenposition, mit Startwerten
v_h, v_v, v_{h0}, v_{v0}	hor. und vert. Komponente der Geschwindigkeit, mit Startwerten
$v_a, v_{at}, v_{ar}, v_{av}$	Absolutgeschwindigkeit, mit tang., radialer, vert. Komponente
$v_r, v_{at}, v_{rr}, v_{rv}$	Relativgeschwindigkeit, mit tang., radialer und vert. Komponente
v_{zul}	Strömungsgeschwindigkeit im Zulauf

A.2 Berechnung

Das bereits im Kapitel *Auswertung* verwendete Fallbeispiel wird hier näher erläutert, um dem Leser die Möglichkeit zu geben, es bei Bedarf selbst für eigene Studien zu benutzen, sozusagen als Erstanwendung. Insgesamt gibt es acht voneinander unabhängige Parameter mit dem in der Tabelle vorgeschlagenen Einstellbereich. Das

zugehörige *Octave / Matlab* Modell ist numerisch recht stabil und gestattet auch einen größeren Einstellbereich.

Parameter	Einheit	Einstellung		
		-	0	+
α	°	15	30	45
β	°	0	15	30
A_q	mm ²	2	3	4
d	mm	100	150	200
M_{Rl}	Nm	0,01	0,015	0,02
M_{Rf}	Nm/s	0,01	0,015	0,02
p_{in}	bar	1	1,5	2
d_{zul}	mm	5	7,5	10

Eine konstante Drehzahl stellt sich ein, wenn Reibungsmoment und Antriebsmoment im Gleichgewicht stehen. Das Reibungsmoment besteht aus einem konstanten Anteil, der trockenen Reibung und einem drehzahlabhängigen Anteil, der flüssigen Reibung.

$$M_R = M_{Rl} + n \cdot M_{Rf} \quad (\text{A.1})$$

Das Antriebsmoment kommt durch den Impuls der Wassertröpfchen zustande.

$$M_A = \rho_W \dot{Q} v_{at} \cdot R \quad (\text{A.2})$$

$$= 2\rho_W v_r A_q v_{at} \cdot R \quad (\text{A.3})$$

Die Absolutgeschwindigkeit der Tröpfchen folgt aus dem Energiesatz,

$$v_a = \sqrt{\frac{2p_{en}}{\rho}} \quad (\text{A.4})$$

wobei die treibende Druckdifferenz auch die Druckverluste berücksichtigt.

$$P_{en} = P_{in} - \Delta p_v - \Delta p_{zul} \quad (\text{A.5})$$

Die Komponentenzerlegung der Absolutgeschwindigkeit in tangentiale, radiale und vertikale Komponente ist nicht trivial, da die beiden Düsenwinkel die relative Ausrichtung des Wasserstrahls in Bezug zum rotierenden Rasensprenger angeben. Mit Kenntnis der Düsengeschwindigkeit lässt sich jedoch die Relativgeschwindigkeit zunächst betragsmäßig berechnen und anschließend in Komponenten zerlegen. Durch vektorielle Addition mit der Düsengeschwindigkeit folgt daraus dann die gesuchte Absolutgeschwindigkeit als vollständig bestimmter Vektor. Sobald die Relativgeschwindigkeit ermittelt ist, lässt sich auch der Volumenstrom angeben.

$$v_D = \omega R \quad \text{mit} \quad R = \frac{d}{2} \quad (\text{A.6})$$

$$v_a^2 = v_r^2 + v_D^2 - 2v_r \cos \alpha \cos \beta \cdot v_D \quad (\text{A.7})$$

$$v_r = v_D \cos(\alpha) \cos(\beta) + \sqrt{v_a^2 - v_D^2 (\cos^2(\alpha) \cos^2(\beta) - 1)} \quad (\text{A.8})$$

$$\dot{Q} = 2v_r A_q \quad (\text{A.9})$$

$$\mathbf{v}_r = \begin{bmatrix} v_{rt} \\ v_{rr} \\ v_{rv} \end{bmatrix} = v_r \cdot \begin{bmatrix} \cos(\alpha) \cos(\beta) \\ \cos(\alpha) \sin(\beta) \\ \sin(\alpha) \end{bmatrix} \quad (\text{A.10})$$

$$\mathbf{v}_a = \begin{bmatrix} v_{at} \\ v_{ar} \\ v_{av} \end{bmatrix} = \mathbf{v}_r - \begin{bmatrix} \omega \cdot R \\ 0 \\ 0 \end{bmatrix} \quad (\text{A.11})$$

Damit sind alle Größen der Momentengleichungen bis auf ω verfügbar. ω ergibt sich iterativ aus der Forderung nach einem Gleichgewicht der Drehmoment. Ausgehend von einem konservativ abgeschätzten Startwert für ω fährt der Rasensprenger in der Simulation an und erreicht beim Momentengleichgewicht eine Grenzdrehzahl. Die Reibungsenergie wird dem System entzogen und äußert sich durch einen Druckverlust, da es außer dem Wasserstrahl keine weitere Energiequelle gibt.

$$\Delta p_v = \frac{M_R \omega}{\dot{Q}} \quad (\text{A.12})$$

Hinzu kommt der Druckverlust der querschnittsarmen Zuleitung zur Düse innerhalb des Rasensprengers. Nur bei großen Volumenströmen in Verbindung mit einem engen Zuleitungsquerschnitt spielt dies eine Rolle. Näherungslösungen sind also an dieser Stelle unkritisch. Aus tabellierten Daten folgt für glatte Rohre mit der Länge 300mm mit einem Durchmesser von 5mm bis 10mm und einem Volumenstrom zwischen 1 l/min und 10 l/min eine kompakte Näherung für den Verlustkoeffizienten.

$$\Delta p_{zul} = c_v \dot{Q}^2 \quad (\text{A.13})$$

$$c_v = 60000^2 \cdot 10^{5,0704 - 0,579413 d_{zul} + 0,196432 d_{zul}^2} \quad (\text{A.14})$$

Der Vorfaktor ergibt sich durch die Umrechnung von $\frac{l}{\text{min}}$ in $\frac{m^3}{s}$, da die Berechnung im Gegensatz zu den tabellierten Daten durchgängig SI-Einheiten benutzt. Erst wenn ω gegen einen stabilen Wert konvergiert, erreichen auch die Druckverluste ihren Endwert. Die Simulation beginnt daher mit einem geeigneten Startwert für v_a , der korrigiert wird, sobald für die Druckverluste bessere Werte vorliegen.

Die Flugbahn der Wassertröpfchen wäre eine Wurfparabel, wenn es keinen Luftwiderstand gäbe. Dieser ist jedoch nicht vernachlässigbar und bremst die Tröpfchen in Abhängigkeit von Ihrer Größe und Fluggeschwindigkeit. Vereinfachend wurde

eine bewährte Formulierung von SCHADE, KUNZ übernommen [1], die auf Arbeiten von ABRAHAMS basiert und für einen weiten Bereich der Reynoldszahl gilt.

$$Re = \frac{v_a d_{Tr} \rho_L}{\eta_L} \quad (\text{A.15})$$

$$\zeta = \frac{24}{Re} \cdot \left(1 + 0,11\sqrt{Re} \right)^2 \quad (\text{A.16})$$

Als Konstanten wurden folgende Werte angesetzt:

$$\rho_L = 1,25 \frac{\text{kg}}{\text{m}^3} \quad (\text{A.17})$$

$$\eta_L = 1,82 \cdot 10^{-5} \frac{\text{kg}}{\text{ms}} \quad (\text{A.18})$$

$$\rho_w = 1000 \frac{\text{kg}}{\text{m}^3} \quad (\text{A.19})$$

Tropfendurchmesser und Tropfenmasse richten sich bei kleinen Düsenquerschnitten nach der Größe der Austrittsöffnung.

$$d_{Tr} = \sqrt{\frac{4A_q}{\pi}} \quad (\text{A.20})$$

$$m_{Tr} = \frac{\pi}{6} d_{Tr}^3 \rho_L \quad (\text{A.21})$$

Aus dem Luftwiderstandsbeiwert ergeben sich die Bremskraft und die Tropfenverzögerung. Die Richtung der Tropfenverzögerung hängt von der momentanen Tropfengeschwindigkeit ab und wird in jedem Zeitschritt neu ermittelt. Die Berechnung der Flugbahn ist abgeschlossen, wenn ihre vertikale Komponente den Wert Null erreicht hat.

$$F_w = \frac{\rho_L}{2} v_a^2 A_q \zeta \quad \text{mit} \quad a = \frac{F_w}{m_{Tr}} \quad (\text{A.22})$$

$$s_{h0} = 0 \quad s_{v0} = 1\text{mm} \quad (\text{A.23})$$

$$v_{h0} = v_a \cos(\alpha) \quad v_{v0} = v_a \sin(\alpha) \quad (\text{A.24})$$

$$a_h = a \cos(\varphi) \quad a_v = a \sin(\varphi) \quad (\text{A.25})$$

$$\cos(\varphi) = \frac{v_h(t)}{v(t)} \quad \sin(\varphi) = \frac{v_v(t)}{v(t)} \quad (\text{A.26})$$

$$v_h = v_{h0} + \int_0^t a_h dt \quad v_v = v_{v0} + \int_0^t a_v dt \quad (\text{A.27})$$

$$s_h = s_{h0} + \int_0^t v_h dt \quad s_v = s_{v0} + \int_0^t v_v dt \quad (\text{A.28})$$

A.3 Erweiterungen

Basierend auf der dargestellten Basisvariante des Berechnungsprogramms wurden verschiedene optionale Erweiterungen eingeführt, um größere Stufenbreiten rechnen zu können. Dadurch entstehen stark nichtlineare Zusammenhänge zwischen den Eingangsgrößen und den Qualitätsmerkmalen. Für die Verdeutlichung der aufwendigen Verfahren (Kriging, Radial Based Functions, neuronale Netze etc.) war dies erforderlich.

Druckverlust

Die Berechnung des Druckverlustes in der Zuführleitung kann neben der Basisvariante ebenfalls durch den in Gleichung A.29 dargestellten Ansatz ermittelt werden. Durch die separate Berechnung des Reibungskoeffizienten λ können neben laminaren ebenfalls turbulente Strömungsverluste berücksichtigt werden, worauf im Rahmen dieser Arbeit verzichtet wird. Weiterhin wird im Gegensatz zur Basisvariante nicht ein Zulauf vor dem Rasensprenger angenommen, sondern die Arme selbst als Zulauf betrachtet, so dass neben dem Durchmesser d_{zul} der Leitung ebenfalls der Radius R des Rasensprengers einen direkten Einfluss auf den Druckverlust aufweist.

$$\Delta p_{zul} = \lambda \frac{\rho}{2} \frac{R}{d_{zul}} v_{zul}^2$$

mit $v_{zul} = \frac{\dot{Q}}{A_{zul}}$ und $\lambda_{\text{laminar}} = \frac{64}{Re} = 64 \frac{v}{d_{zul} v_{zul}}$ (A.29)

Flugweite

Soll der Radius R des Rasensprengers in der Flugweitenbestimmung berücksichtigt werden, so kann die Flugweite s^* mit der Basisflugweite s_h bestimmt werden.

$$s^* = \sqrt{[R + \sin(\beta) s_h]^2 + [\cos(\beta) s_h]^2} \quad (\text{A.30})$$

Haftriebung

Werden die Faktoren α , β und M_{Rt} in großen Bereichen variiert, so treten Faktorkombinationen auf, bei denen das Antriebsmoment des Wasserstrahls geringer ist als das Reibmoment M_{Rt} , so dass keine Rotation des Rasensprengers auftritt ($n = 0$). Da der implementierte Lösungsalgorithmus in diesen speziellen Fällen keinen Gleichgewichtszustand findet, wird vor der Volumenstromberechnung geprüft, ob die Haftriebung bei $n = 0$ überwunden wird. Ist dieses nicht der Fall, so wird die Flugweite mit dem berechneten Volumenstrom \dot{Q}_0 und der Drehzahl $n = 0$ ermittelt.

$$M_A = \dot{Q}_0 \rho R v_t$$

$$v_t = \frac{\dot{Q}_0}{2A_q \cos(\alpha) \cos(\beta)} \quad (\text{A.31})$$

$$\text{Energieerhaltung: } p_{in} - \Delta p_{zul} = \frac{\rho}{2} v_{aus}^2 \quad (\text{A.32})$$

$$\begin{aligned}
 \text{Basisansatz :} & \Delta p_{zul} = c_v \dot{Q}_0^2 \\
 \Rightarrow & p_{in} - c_v \dot{Q}_0^2 = \frac{\rho}{2} v_{aus}^2 = \frac{\rho}{2} \left(\frac{\dot{Q}_0}{2A_q} \right)^2 \\
 \Rightarrow & \dot{Q}_0 = \sqrt{\frac{p_{in}}{c_v + \frac{\rho}{8A_q^2}}}
 \end{aligned} \tag{A.33}$$

$$\begin{aligned}
 \text{Erweiterter Ansatz :} & \Delta p_{zul} = 64 \frac{v}{d_{zul} v_{zul}} \frac{\rho}{2} \frac{R}{d_{zul}} v_{zul}^2 \\
 \Rightarrow & p_{in} - 64 \frac{v}{d_{zul} v_{zul}} \frac{\rho}{2} \frac{R}{d_{zul}} v_{zul}^2 = \frac{\rho}{2} \left(\frac{\dot{Q}_0}{2A_q} \right)^2 \\
 \Rightarrow & p_{in} - 64 \frac{v}{2d_{zul}^2} \rho R \frac{\dot{Q}_0}{A_{zul}} = \frac{\rho}{2} \left(\frac{\dot{Q}_0}{2A_q} \right)^2 \\
 \Rightarrow & \frac{2}{\rho} p_{in} - 64 \frac{v}{2d_{zul}^2 A_{zul}} R \dot{Q}_0 = \left(\frac{\dot{Q}_0}{2A_q} \right)^2 \\
 \Rightarrow & \dot{Q}_0^2 + \frac{32R4A_q^2 v}{d_{zul}^2 A_{zul}} \dot{Q}_0 - \frac{2A_q^2 4}{\rho} p_{in} = 0 \\
 \text{mit } & c_{zul} = \frac{128RA_q^2 v}{d_{zul}^2 A_{zul}} \\
 \Rightarrow & \dot{Q}_0 = -\frac{c_{zul}}{2} + \sqrt{\left(\frac{c_{zul}}{2} \right)^2 + \frac{8A_q^2}{\rho} p_{in}}
 \end{aligned} \tag{A.34}$$

Kodierung

Neben der Standardkodierung $[-1; 1]$ sind insgesamt folgende Faktorkodierungen implementiert worden:

- keine Kodierung
- $[-1; 1]$
- $[0; 1]$
- $[1, 2, 3, \dots, n_s]$

Variablen-Übergabe

Zur Erleichterung einer Automatisierung verschiedener Versuchsläufe können der Berechnungsfunktion folgende Variablen übergeben werden:

- *FileName*: Name der Eingabedatei (z.B. '1128.inp')
- *Fak*: Faktoreinstellungen [Spalte der Eingabedatei, Min, Max]

Ein Beispiel für eine mögliche Faktoreinstellung ist:

```

Fak=[1, 0      , 90      % alpha  [ °]
      2, 0      , 90      % beta   [ °]
      3, 2e-6   , 4e-6   % Aquer  [mm^2]
      4, 0.1    , 0.2    % Durchmesser [m]
      5, 0.01   , 0.02   % Mtrocken [Nm]
      6, 0.01   , 0.02   % Mfluessig [Nm/s]
      7, 1      , 2       % Druck   [bar]
      8, 7      , 10     % Durchmesser Zuleitung [mm]
  ];

```

A.4 Quellcode

Der folgende Quellcode ist unter Octave und Matlab lauffähig. Erforderlich ist eine Eingabedatei mit dem gewünschten Versuchsplan passend zu der gewählten Kodierung (zum Beispiel -1 bis 1). Jede Zeile entspricht einer separaten Berechnung mit der entsprechenden Faktoreneinstellung. Der gesamte Versuchsplan wird per Stapelverarbeitung mit einem Aufruf abgearbeitet.

```
function Rasensprenger (FileName,Fak)
% Rasensprengerversuch
% (c) 2002 - 2009
% Hauptteil: Karl Siebertz
% Erweiterung: David van Bebber
%
% Grundlegende Einstellungen
Kodierung = 1; % OKeine 1[-1,+1] 2[0,+1] 3[1,2,...,ns]
dpzuVariante = 1; % 1Basis 2Variation
sflVariante = 1; % 1Basis 2Variation
%
% Konstanten
g=10; pi=3.141592654; rho=1000; dynVis=1;
kinVis=dynVis/rho; MaxFehler=0.005;
% Datei mit Parameterbelegung einlesen
if nargin < 1
    FileName = '1128.inp';
end
s=load(FileName);
[kzei, kspa] = size(s);
if nargin >= 2
    % Umrechnungsfaktoren für kodierte Daten
    PF = 1; % Position des Faktors
    [aspa,amin,aplu]=SetD(Fak(PF,1),Fak(PF,2),Fak(PF,3));PF=PF+1;
    [bspa,bmin,bplu]=SetD(Fak(PF,1),Fak(PF,2),Fak(PF,3));PF=PF+1;
    [cspa,cmin,cplu]=SetD(Fak(PF,1),Fak(PF,2),Fak(PF,3));PF=PF+1;
    [dsp,a,min,dplu]=SetD(Fak(PF,1),Fak(PF,2),Fak(PF,3));PF=PF+1;
    [espa,emin,eplu]=SetD(Fak(PF,1),Fak(PF,2),Fak(PF,3));PF=PF+1;
    [fspa,fmin,fplu]=SetD(Fak(PF,1),Fak(PF,2),Fak(PF,3));PF=PF+1;
    [gspa,gmin,gplu]=SetD(Fak(PF,1),Fak(PF,2),Fak(PF,3));PF=PF+1;
    [hspa,hmin,hplu]=SetD(Fak(PF,1),Fak(PF,2),Fak(PF,3));PF=PF+1;
else
    [aspa,amin,aplu]=SetD(1, 0 , 90 ); % alpha [°]
    [bspa,bmin,bplu]=SetD(2, 0 , 90 ); % beta [°]
    [cspa,cmin,cplu]=SetD(3, 2e-6, 4e-6); % Aquer [mm^2]
    [dsp,a,min,dplu]=SetD(4, 0.1 , 0.2 ); % Durchmesser [m]
    [espa,emin,eplu]=SetD(5, 0.01, 0.02); % Mtrucken [Nm]
    [fspa,fmin,fplu]=SetD(6, 0.01, 0.02); % Mfluessig [Nm/s]
    [gspa,gmin,gplu]=SetD(7, 1 , 2 ); % Druck [bar]
    [hspa,hmin,hplu]=SetD(8, 7 , 8 ); % Durchm. Zuleitung [mm]
end
% Ausgabedateien
ido0=fopen('d-kompl.dat','w');idol=fopen('d-qm1.dat','w');
ido2=fopen('d-qm2.dat' , 'w');ido3=fopen('d-qm3.dat','w');
```

```

ido4=fopen('d-qm.dat'      , 'w');ido5=fopen('d-par.dat','w');
% große Schleife
ominc=0;
for j=1:kzei
    n=0;sfl=0;qp=0;
% Auswahl der Kodierungsfunktion
switch Kodierung
    case 0
        NormFunc = @Norm0;
    case 1
        NormFunc = @Norm1;
    case 2
        NormFunc = @Norm2;
    case 3
        NormFunc = @Norm3;
    otherwise
        break
end
% Berechnung der Parameter
alpha= feval(NormFunc,amin,aplu,s,j,aspa);
beta = feval(NormFunc,bmin,bplu,s,j,bspa);
A    = feval(NormFunc,cmin,cplu,s,j,cspa);
d    = feval(NormFunc,dmin,dplu,s,j,dspa);
mt   = feval(NormFunc,emin,eplu,s,j,espa);
mf   = feval(NormFunc,fmin,fplu,s,j,fspa);
pin  = feval(NormFunc,gmin,gplu,s,j,gspa);
dzul = feval(NormFunc,hmin,hplu,s,j,hspa);
pin  = pin * 1e5;
h    = pin * 1e-4;
R    = d/2;
sina = sin(alpha*pi/180); cosa = cos(alpha*pi/180);
tana = tan(alpha*pi/180); sinb = sin( beta*pi/180);
cosb = cos(beta *pi/180); tanb = tan( beta*pi/180);
% Interpolation gültig für d von 5mm bis 10mm
cvzul = 10^(5.0704 -0.579413*dzul+0.0196432*dzul^2);
cvzul = (cvzul*60000^2);
% Startwerte
m0    = 2*rho*A*R*2*g*h*cosa*cosb;
n1    = 0.1*abs(m0-mt) / (mf+5.0e-4);
omega = 2*pi*n1;
msoll = mt+omega*mf;
mdiff = m0;
va0   = sqrt(2*pin/rho);
deltap= abs(msoll*omega)/(A*va0); % Verlustleistung Startwert
dzul = dzul * 1e-3;
Azul  = pi/4*dzul^2;
if dpzulVariante == 1
    dpzul=cvzul*(A*va0)^2;
else
    dpzul=0.1*pin;
end
it=0; va=0; vr=0; m=m0;
% Durchfluss bei n = 0
if dpzulVariante == 1

```

```

qp= sqrt(pin/(cvzul+rho/8/A^2));
else
    c = 128*R*A^2*kinVis/(dzul^2*Azul);
    qp= -c/2+sqrt((c/2)^2+8*rho*pin*A^2);
end
va = qp/2/A;
vr = va;
vrt= va*cosb*cosa;
vat= vrt;
m = rho*qp*R*vat;
if m>mt % Haftreibung überschritten?
% Iteration bis zum Momentengleichgewicht
while abs(mdif) > MaxFehler*abs(m)
    n = omega/2/pi;
    msoll= mt+n*mf;
    varm = omega*R;
    % Energiebilanz des gesamten Rasensprengers
    pen = pin-deltap-dpzul;
    if(pen < 0.01*pin)
        fprintf('Fehler: pen < 0.01*pin\n');
        pin,deltap,dpzul,msoll
        m0,m,mdif,ominc,vr,varm,va
        va=0;
        break;
    end
    va=sqrt(2*pen/rho);
    if(va^2+varm^2*(cosa^2*cosb^2-1) < 0 )
        fprintf('Fehler: va^2+varm^2*(cosa^2*cosb^2-1)<0\n');
        va,vr,varm,vak
        break;
    end
    vr = varm*cosa*cosb;
    vr = vr+sqrt(va^2+varm^2*(cosa^2*cosb^2-1));
    vrt = vr*cosb*cosa;
    vrr = vr*cosa*sinb;
    vrv = vr*sina;
    vat = vrt-omega*R;
    var = vrr;
    vav = vrv;
    % Kontrolle der Komponentenzerlegung
    vak = sqrt(vat^2+var^2+vav^2);
    m = 2*rho*vr*A*R*vat;
    mdif = m-msoll;
    ominc = 0.1*min(abs(mdif/m),(0.5*pen/pin));
    % variable Schrittweite
    omega = omega*(1+ominc)^sign(mdif);
    qp = 2*vr*A;
    % Verlustleistung in Druck umgerechnet
    deltap= abs(msoll*omega)/qp;
    if dpzulVariante == 1
        dpzul= cvzul*qp^2;
    else
        vzul = qp/2/Azul; % qp/2 durch einen Arm
        Re = abs(dzul*vzul/kinVis);
    
```

```

dpzul= 64/Re*rho/2*R/dzul*vzul^2;
end
it=it+1;
if it > 10000
    fprintf('Fehler: it > 10000\n');
    it,msoll,mdiff,ominc,alpha,beta
    A,d,mt,mf,vr,va,vrt,varm,vat,omega
    break;
end
if(omega < 0.0062 )
    fprintf('Fehler: omega < 0.0062\n');
    it,omega
    n=0;
    break;
end
else
    omega=0; n=0;
end
% Flugbahn
ddropf = sqrt(4*A/pi);
etaluft= 1.82e-5;
nyluft = etaluft/1.25;
v      = va;
z      = 1.0e-3;
sfl    = 0.0;
vh     = va*cosa;
vv     = va*sina;
deltat = 0.005;
mtr   = pi/6*dtropf^3*rho;
while z > 0
    if(va<0.01)
        break;
    end
    Re  = va*dtropf/nyluft;
    % Abraham, The Physics of Fluids 13, S.2194
    zeta= 24/Re*(1+0.11*sqrt(Re))^2;
    Fwid= 1.25/2*va^2*pi/4*dtropf^2*zeta;
    atr = Fwid/mtr;
    sfl = sfl+vh*deltat;
    z   = z+vv*deltat;
    vh  = vh-atr*cosa*deltat;
    vv  = vv-g*deltat-atr*sina*deltat;
    va  = sqrt(vh^2+vv^2);
    cosa= vh/va;
    sina= vv/va;
end;
if sflVariante == 1
    sfl=sfl;
else
    sfl=sqrt((R+sinb*sfl)^2+(cosb*sfl)^2);
end
qp    = 2*vr*A*60000;
pverh= deltap/(rho*g*h);

```

```
% Ausgabe
Amm2 = 1000000*A;
dmm = 1000*d;
mtmm = mt*1000;
mfmm = mf*1000;
dzulmm= dzul*1e3;
fprintf(ido0,'%6.2e %6.2e %6.2e %6.2e ',alpha,beta,Amm2,dmm);
fprintf(ido0,'%6.2e %6.2e %6.2e %6.2e ',mtmm,mfmm,h,dzulmm);
fprintf(ido0,'%10.8e %10.8e %10.8e \n',n,sfl,qp);
fprintf(ido1,'%10.4f \n',n);
fprintf(ido2,'%10.4f \n',sfl);
fprintf(ido3,'%10.8f \n',qp);
fprintf(ido4,'%10.8f %10.8f %10.8f \n',n,sfl,qp);
fprintf(ido5,'%6.2e %6.2e %6.2e %6.2e ',alpha,beta,Amm2,dmm);
fprintf(ido5,'%6.2e %6.2e %6.2e %6.2e \n',mtmm,mfmm,h,dzulmm);
end;
fclose(ido0);fclose(ido1);fclose(ido2);
fclose(ido3);fclose(ido4);fclose(ido5);
% Hilfsfunktionen
function Value=Norm0(MinVal,MaxVal,data,row,col) % ohne Kodierung
if size(data,2) < col || size(data,1) < row
    Value = (MinVal+MaxVal)/2;
else
    Value = data(row,col);
end
function Value=Norm1(MinVal,MaxVal,data,row,col) % [-1;1]
if size(data,2) < col || size(data,1) < row
    Value = (MinVal+MaxVal)/2;
else
    Value = MinVal+(MaxVal-MinVal)*(data(row,col)+1)/2;
end
function Value=Norm2(MinVal,MaxVal,data,row,col) % [0;1]
if size(data,2) < col || size(data,1) < row
    Value = (MinVal+MaxVal)/2;
else
    Value = MinVal+(MaxVal-MinVal)*data(row,col);
end
function Value=Norm3(MinVal,MaxVal,data,row,col) % [1,2,...,ns]
if size(data,2) < col || size(data,1) < row
    Value = (MinVal+MaxVal)/2;
else
    minStufe=min(data(:,col));data(:,col)=data(:,col)-minStufe+1;
    maxStufe=max(data(:,col));diffStufe=maxStufe-1;
    if diffStufe == 0
        Value=(MinVal+MaxVal)/2;
    else
        Value=MinVal+(MaxVal-MinVal)*(data(row,col)-1)/diffStufe;
    end
end
function [spalte,minus,plus]=SetD(Spalte,Minimal,Maximal)
spalte=Spalte;minus=Minimal;plus=Maximal; % Set Factor Data
```

Literaturverzeichnis

1. Schade, H., Kunz, E.: *Strömungslehre*. Walter de Gruyter, Berlin New York (1980) 465

Anhang B

Computer-Experiment

B.1 Rasensprenger mit erweitertem Faktorraum

Zur Veranschaulichung der in den Kapiteln 8 bis 13 dargestellten Verfahren aus dem Bereich *Computer-Experimente* und *Multivariate Datenanalyse*

- Testfelder
- Metamodelle
- Optimierung
- Sensitivitätsanalyse

wird eine Analyse des im Anhang A dargestellten Rasensprengerbeispiels mit erweitertem Faktorraum durchgeführt.

Faktor	Einheit	Min	Max	Faktor	Einheit	Konstant
α	°	0	120	$M_{trocken}$	Nm	0.015
β	°	0	90	$M_{flüssig}$	Nm/s	0.015
A	mm^2	2	4	$Druck$	bar	1.5
d_{Arm}	mm	100	200	$d_{Leitung}$	mm	8

Tabelle B.1 Faktoren

Der Haupteffekt jedes Faktors x_j wird im Vorfeld mittels 5000 Monte-Carlo-Simulationen (Kapitel 8.3.1) ausreichend genau abgeschätzt und im Anschluss zur Validierung der verschiedenen Verfahren verwendet. Der jeweils zu analysierende Faktor x_j wird dabei, im Gegensatz zu allen anderen Faktoren, während der Monte-Carlo-Simulation konstant gehalten $x_{j\text{konst}}$. Jede Monte-Carlo-Simulation liefert dann eine Approximation des Mittelwerts \bar{y} bei der gegebener Faktorstufe $x_{j\text{konst}}$, wenn alle anderen Faktoren gleichmäßig beziehungsweise nach ihrer wahren Verteilung variiert werden.

$$\hat{\bar{y}}_{x_{j\text{konst}}} = \frac{1}{n_r} \sum_{i=1}^{n_r} g(\mathbf{x}_{i,-j}, x_{j\text{konst}}) \quad (\text{B.1})$$

$\mathbf{x}_{i,-j}$ gibt dabei die Menge aller Faktoren ohne x_j an, welche durch die Monte-Carlo-Simulation zufällig variiert werden. Mittels Wiederholungen der Monte-Carlo-Simulation mit unterschiedlichen Faktorstufen für x_j wird der komplette Haupteffekt des Faktors x_j abgeschätzt.

Zur Bestimmung des Interaktionseffekts zweier unabhängiger Faktoren x_j und x_k werden bei gleicher Vorgehensweise beide Faktoren während der jeweiligen Monte-Carlo-Simulation auf gewählten Faktorstufen konstant gehalten.

$$\widehat{\bar{y}}_{x_{j\text{konst}}, x_{k\text{konst}}} = \frac{1}{n_r} \sum_{i=1}^{n_r} g(\mathbf{x}_{i,-\{j,k\}}, x_{j\text{konst}}, x_{k\text{konst}}) \quad (\text{B.2})$$

Zur Bestimmung der Haupt- und Interaktionseffekte werden im Folgenden die Faktoren im Bereich $x \in [-1; 1]$ in den Stufen $\Delta x_{\text{Haupt}} = 0.1$ und $\Delta x_{\text{Inter}} = 0.2$ variiert. Die Abbildungen B.1 bis B.4 zeigen die Haupt- und Interaktionseffekte bei Verwendung des Simulationsmodells aus Anhang A, wobei die Berechnung in der verwendeten Ausführung mehrere Stunden beansprucht.

Drehzahl

Bei steigendem Winkel α und β sinkt das Antriebsmoment des Wasserstrahls so weit ab, dass ab einem bestimmten Winkel die Trockenreibung nicht überwunden wird und sich somit keine Rotation einstellt ($n = 0 \frac{1}{s}$). Es bildet sich dadurch ein deutliches Plateau an den Rändern des Faktorraums aus (Abbildung B.1, links). Die stärksten Interaktionen zeigen sich bei der Drehzahl zwischen den Winkeln α und β (Abbildung B.2, oben links).

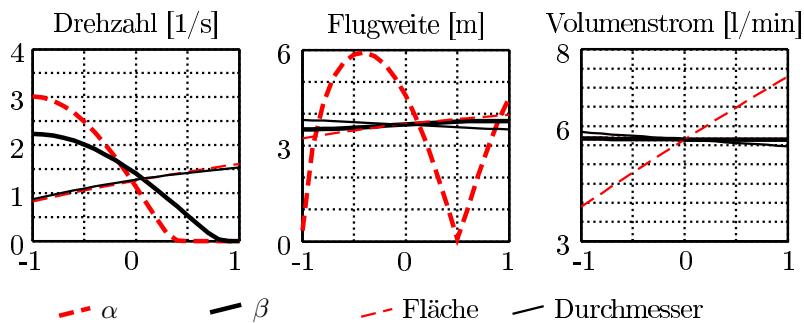


Abb. B.1 Rasensprenger, Haupteffekte, Originalmodell

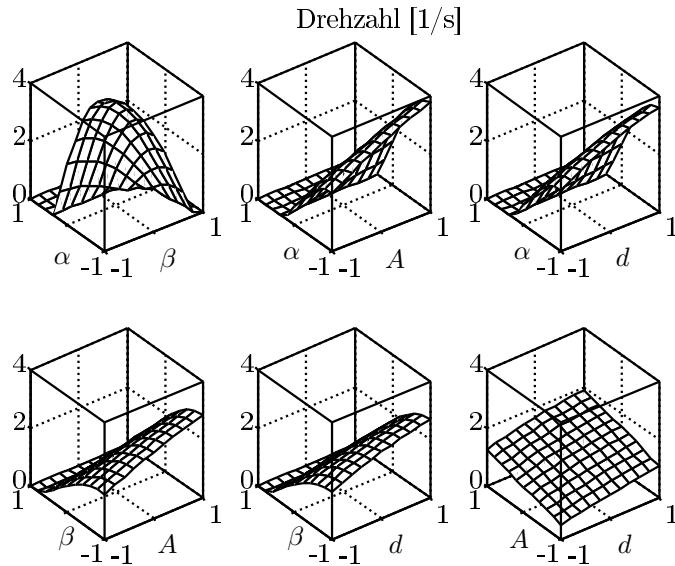


Abb. B.2 Rasensprenger, Interaktion, *Drehzahl*, Originalmodell

Flugweite

Den größten Haupteffekt auf die Flugweite weist der Winkel α auf, wobei bei steigendem Winkel die Flugweite schnell ansteigt und später wieder abfällt (Abbildung B.1, Mitte und Abbildung B.3). Durch einen direkten Zusammenhang zwischen Flugweite und der Rotationsgeschwindigkeit des Rasensprengers bildet sich die maximale Flugweite nicht bei $\alpha = 45^\circ$ ($x_\alpha = -0.25$) sondern bereits bei ca. $\alpha = 36^\circ$ ($x_\alpha = -0.4$). Bei $\alpha = 90^\circ$ und 0° ($x_\alpha = 0, 0.5$) wird hingegen ein lokales Minimum der Flugweite erreicht.

Volumenstrom

Der Wasserverbrauch ist überwiegend vom Querschnitt A und nur leicht vom Durchmesser d des Rasensprengers abhängig (Abbildung B.1, rechts und Abbildung B.4).

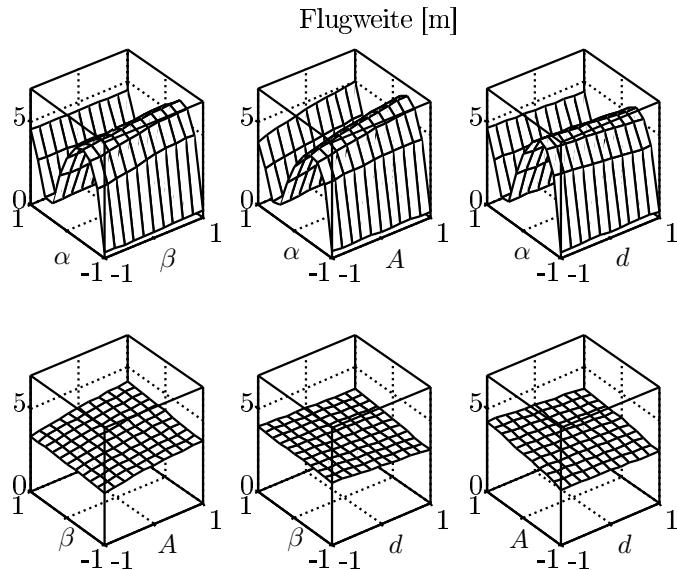


Abb. B.3 Rasensprenger, Interaktion, *Flugweite*, Originalmodell

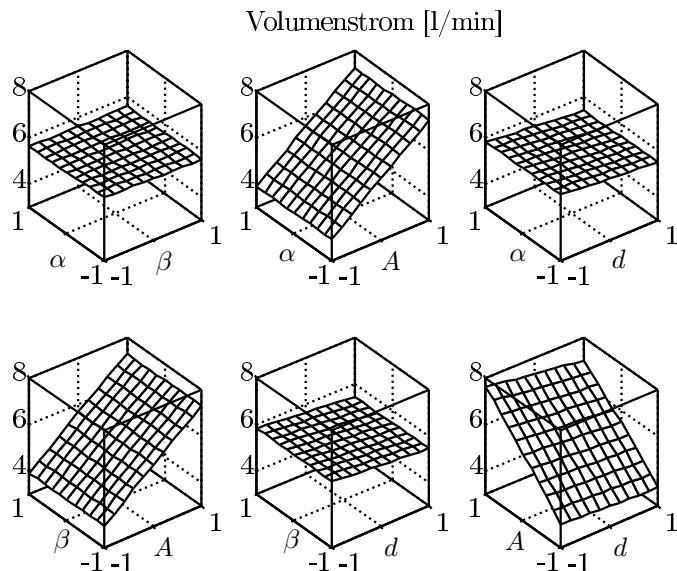


Abb. B.4 Rasensprenger, Interaktion, *Volumenstrom*, Originalmodell

B.2 Testfelder und Metamodelle

Auf Basis verschiedener Testfelder werden unterschiedliche Metamodelle erzeugt, wobei folgende Felder und Modellverfahren beispielhaft berücksichtigt werden.

Testfelder

- Vollfaktorplan ($n_s = 2, n_r = 2^4 = 16$)
- Vollfaktorplan ($n_s = 3, n_r = 3^4 = 81$)
- Latin Hypercube, optimiert mit zentrierter Diskrepanz ($n_s = n_r = 81$)

Metamodelle

- Linare Regression
- Quadratisches Regressionsmodell (Response Surface)
- Spline Regression
- Künstliches Neuronales Netzwerk

Vollfaktorplan ($n_s = 2, n_r = 2^4 = 16$)

Der Vollfaktorplan mit 2 Stufen ($n_s = 2, n_r = 2^4 = 16$) kann strukturbedingt nur lineare Terme und einfache Interaktionen berücksichtigen, so dass lediglich die Kombination mit einem linearen Regressionsmodell sinnvoll ist. Andere Modelltypen sind zwar einsetzbar, ergeben aber keinen Vorteil, da einfach nicht mehr Informationen in den gegebenen Daten vorhanden sind. Abbildung B.5 zeigt die mit dem linearen Regressionsmodell ermittelten Haupteffekte für Drehzahl, Flugweite und Volumenstrom, wobei der Volumenstrom bereits durch das lineare Modell ausreichend genau abgebildet wird. Für die Drehzahl kann jedoch lediglich eine grundsätzliche Tendenz für die Faktoren ermittelt werden. Die geschwungenen Formen, die in Abbildung B.1 zu erkennen sind, sind nicht darstellbar. Eine sinnvolle Vorhersage der Flugweite ist hingegen nicht möglich, da beide Extremwerte nicht abgebildet werden. Einerseits fehlen dazu durch die Verwendung von lediglich zwei Faktorstufen die benötigten Informationen im Inneren des Faktorraums und andererseits ist das lineare Regressionsmodell nicht in der Lage, die Komplexität des Zusammenhangs darzustellen. Gleiche Ergebnisse liefert die Analyse der Interaktionen, welche in Abbildungen B.6 bis B.8 dargestellt sind.

Vollfaktorplan ($n_s = 3, n_r = 3^4 = 81$)

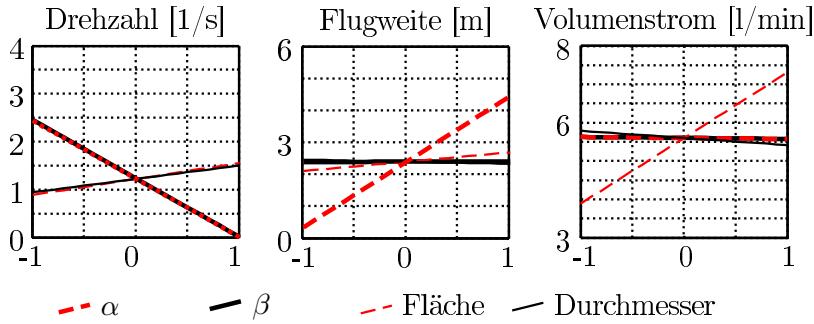
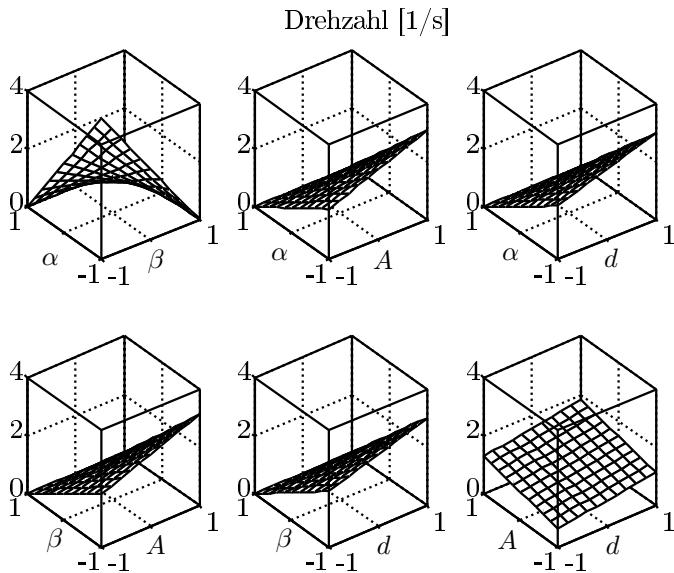
Aus der Analyse des 2-stufigen Vollfaktorplans wird deutlich, dass eine Erhöhung der Stufenanzahl zur Abbildung der Funktionszusammenhänge notwendig ist. Als Beispiel wird hier ein einfacher und nicht optimaler 3-stufiger Vollfaktorplan mit $n_r = 81$ Versuchsläufen verwendet. Basierend auf den erzeugten Daten dieses Versuchsplans wird ein Regressionsmodell mit (maximal) quadratischen Termen und

eine Spline Regression erstellt. Der Volumenstrom wird von beiden Modellen wie bereits im 2-stufigen Vollfaktorplan ausreichend genau abgebildet und wird in folgenden Analysen nicht weiter betrachtet, da keine deutliche Steigerung der Genauigkeit durch bessere Testfelder oder Metamodelle zu erwarten ist. Die Haupteffekte bezüglich der Drehzahl zeigen bereits eine bessere Abbildung des wahren Verlaufs als beim Plan mit zwei Stufen. Die Plateaus am Rand und der geschwungene Zusammenhang zwischen Drehzahl und den Winkeln α und β können jedoch trotz der hohen Anzahl an Versuchsläufen nicht abgebildet werden. Die Vorhersage der Flugweite ist nicht sinnvoll einsetzbar, da der komplexe Verlauf nicht abgebildet werden kann. Der Einsatz des Metamodells für komplexe Zusammenhänge (Spline Regression) bringt in Kombination mit dem 3-stufigen Vollfaktorplan keine Vorteile, da noch immer benötigte Daten zur Abbildung des Verlaufs im inneren Faktorbereich fehlen. Die Abbildungen B.10 bis B.13 zeigen die Interaktionen für Drehzahl und Flugweite, welche mit den beiden Metamodellen (Quadratische- und Spline Regression) ermittelt werden. Auch hier bestätigen sich die Schlussfolgerungen, die bei der Betrachtung der Haupteffekte bereits gefunden wurden.

Latin Hypercube, optimiert mit zentrierter Diskrepanz

$(n_s = n_r = 81)$

Zum direkten Vergleich mit dem 3-stufigen Vollfaktorplan wird ein symmetrischer Latin Hypercube mit 81 Versuchsläufen verwendet, welcher mit dem Gütekriterium *zentrierte Diskrepanz* (Kapitel 8.2.3) optimiert wurde. Auf Basis der ermittelten Daten werden die drei Metamodelle *Künstliches Neuronales Netzwerk* (KNN, $n_{vE1} = 3$, $n_{vE2} = 1$), *Spline Regression* (RS) und *Quadratische Regression* (QR) erstellt. Das Künstliche Neuronale Netzwerk und die Spline Regression nutzen die durch das LHC Testfeld ermittelten Informationen weitestgehend aus und bilden die Haupteffekte der Drehzahl und Flugweite mit ausreichender Genauigkeit ab (Vergleiche Abbildung B.1 und B.14). Neben den geschwungenen Verläufen und den Plateaus der Drehzahl kann ebenfalls die Flugweite von beiden Metamodellen richtig abgebildet werden. Eine Analyse und Optimierung des Systems ist somit mit beiden Metamodellen, im Gegensatz zu den vorherigen Metamodellen, erstmals möglich. Wird im Vergleich ein quadratisches Regressionsmodell verwendet, so können die vorhandenen Informationen nicht sinnvoll verwendet werden, da das Metamodell durch seinen vorgegebenen Funktionszusammenhang nicht den komplexen Zusammenhang darstellen kann. Ein Großteil der vorhandenen Informationen geht dadurch verloren und steht zur weiteren Analyse nicht zur Verfügung. Identische Schlussfolgerungen können aus den Darstellungen der Interaktionen in Abbildungen B.15 bis B.20 ermittelt werden.

**Abb. B.5** Vollfaktor, 2 Stufen, Haupteffekte**Abb. B.6** Vollfaktor, 2 Stufen, Interaktion (Drehzahl)

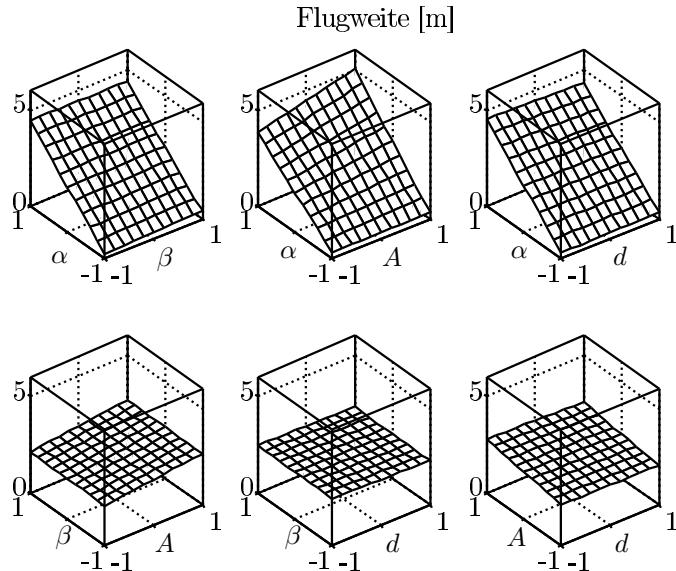


Abb. B.7 Vollfaktor, 2 Stufen, Interaktion (Flugweite)

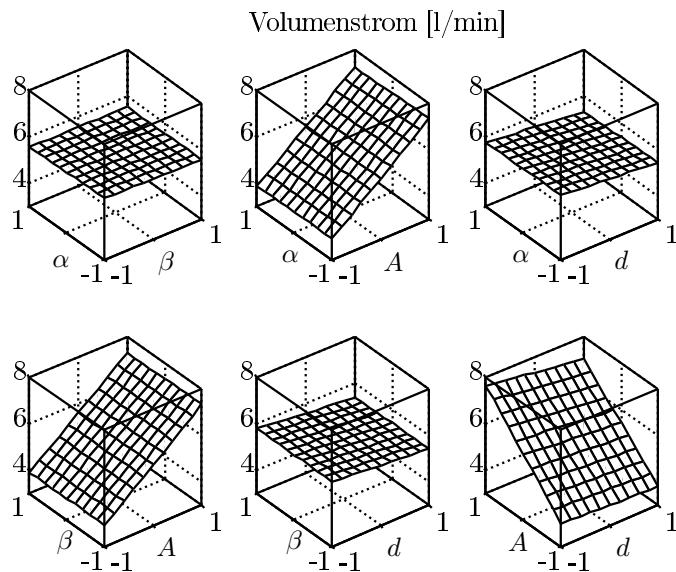
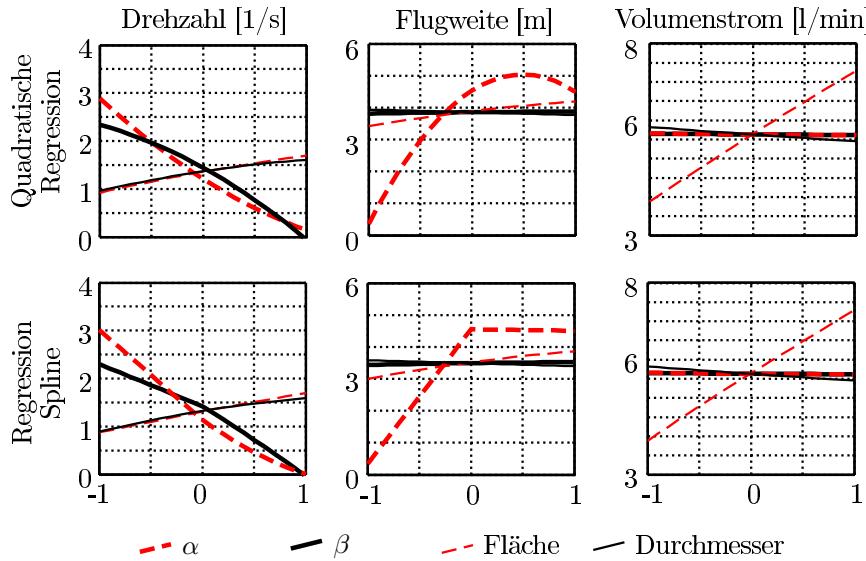
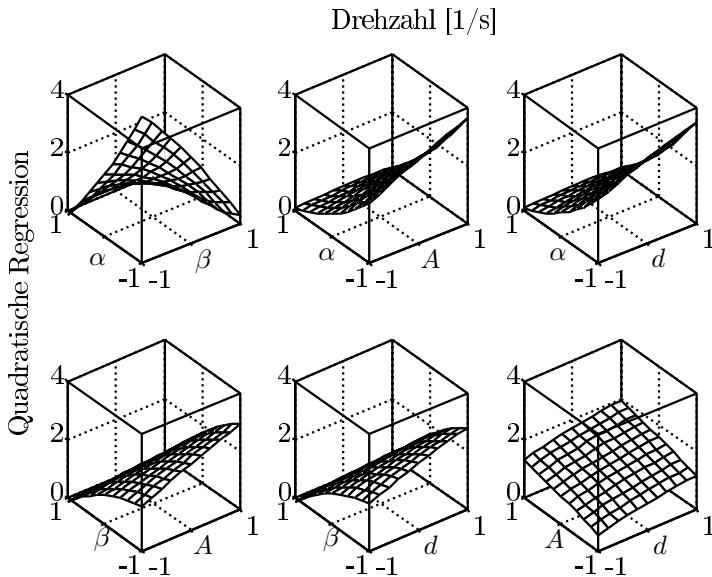


Abb. B.8 Vollfaktor, 2 Stufen, Interaktion (Volumenstrom)

**Abb. B.9** Vollfaktor, 3 Stufen, Haupteffekte**Abb. B.10** Vollfaktor, 3 Stufen, Interaktion (Drehzahl), Quadratische Regression

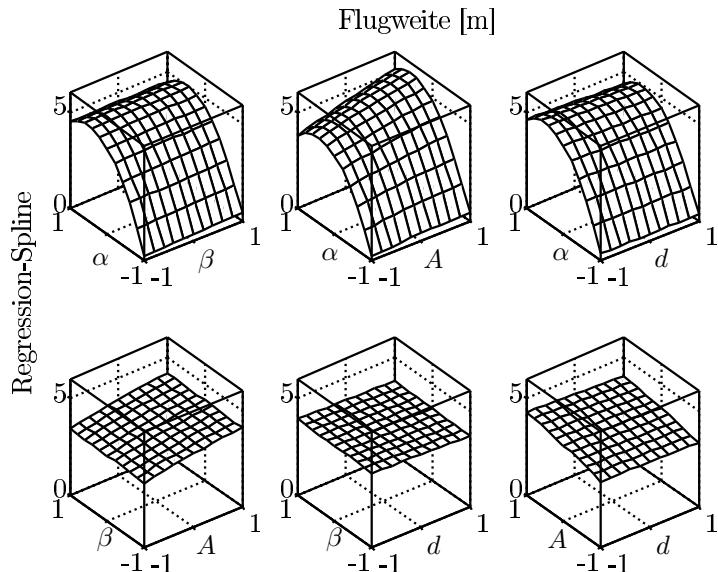


Abb. B.11 Vollfaktor, 3 Stufen, Interaktion (Flugweite), Quadratische Regression

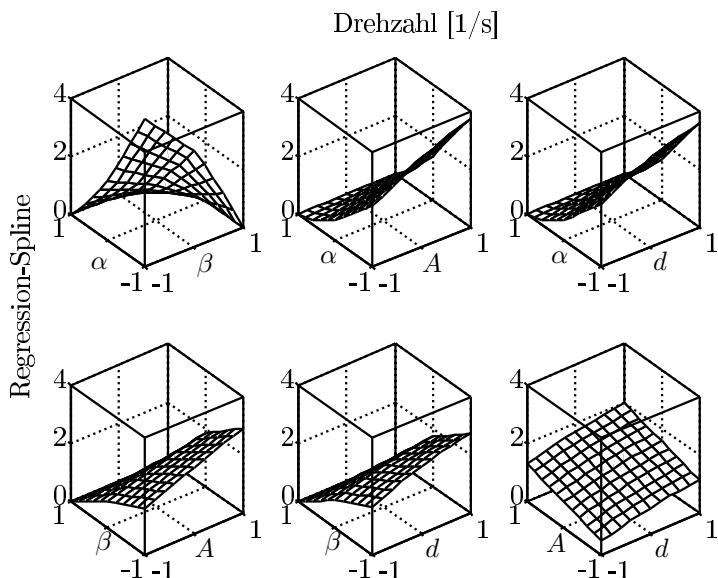


Abb. B.12 Vollfaktor, 3 Stufen, Interaktion (Drehzahl), Spline Regression

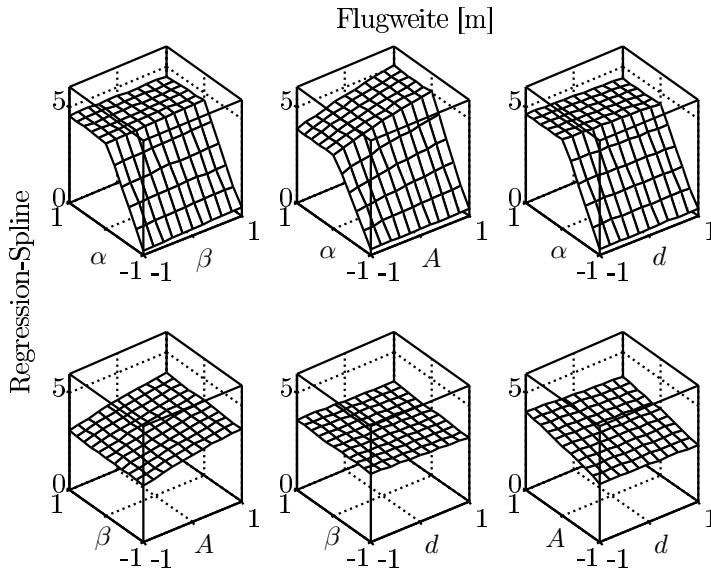


Abb. B.13 Vollfaktor, 3 Stufen, Interaktion (Flugweite), Spline Regression

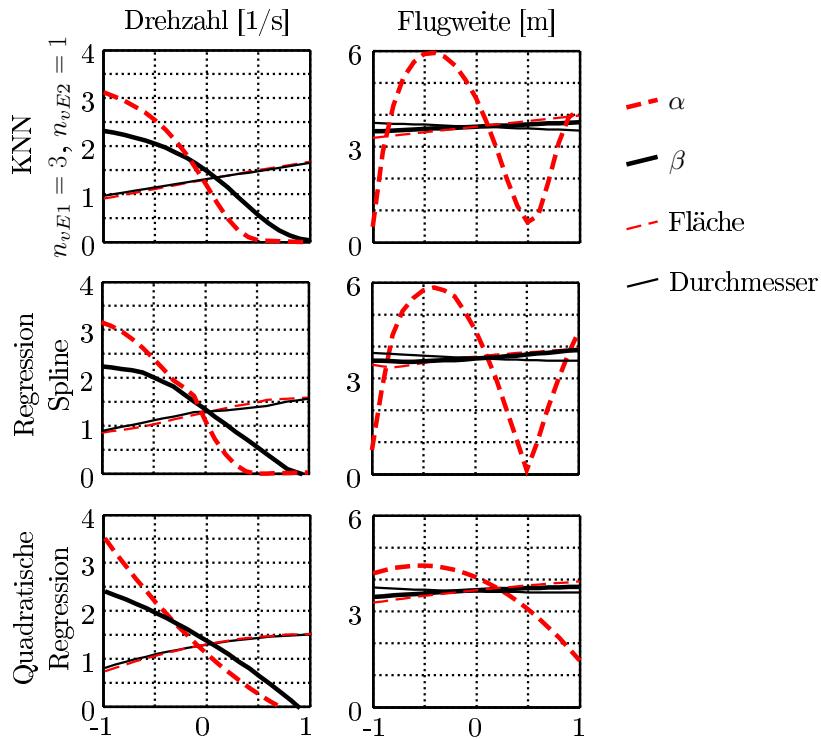


Abb. B.14 Symmetrisches LHC (zentrierte Diskrepanz) , Haupteffekte

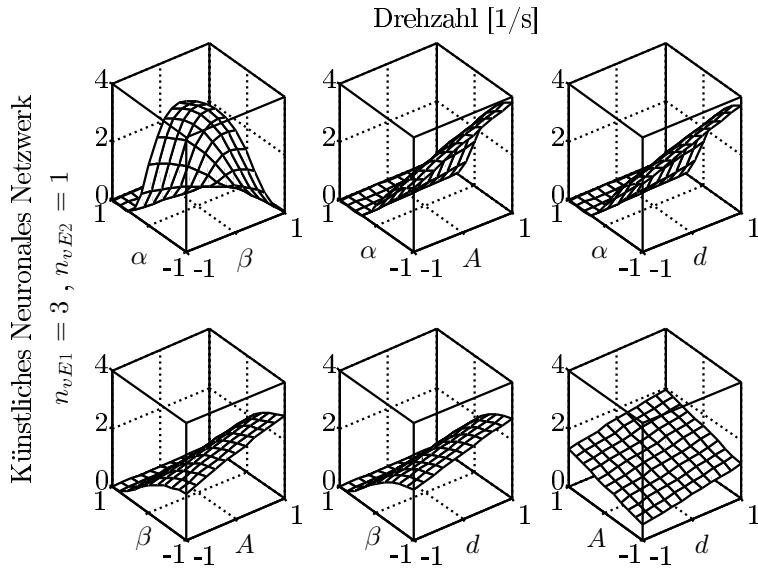


Abb. B.15 Symmetrisches LHC (zentrierte Diskrepanz), Interaktion (*Drehzahl*), KNN ($n_{vE1} = 3$, $n_{vE2} = 1$)

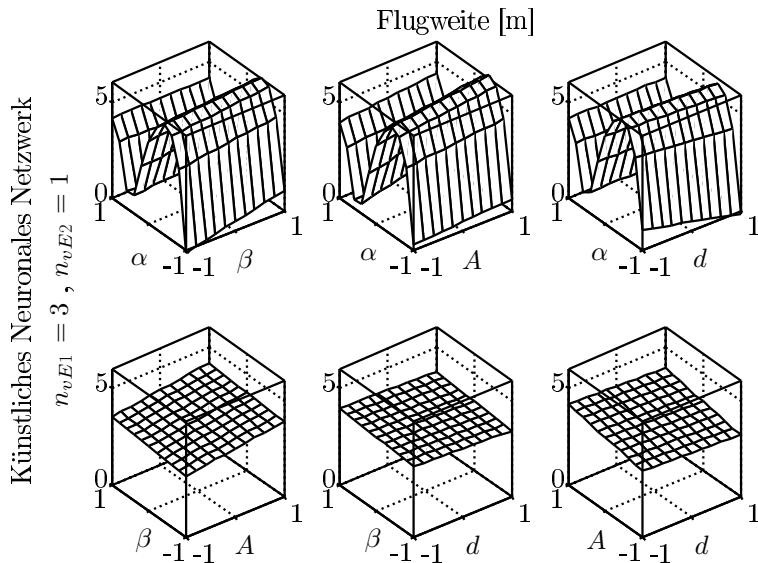


Abb. B.16 Symmetrisches LHC (zentrierte Diskrepanz), Interaktion (*Flugweite*), KNN ($n_{vE1} = 3$, $n_{vE2} = 1$)

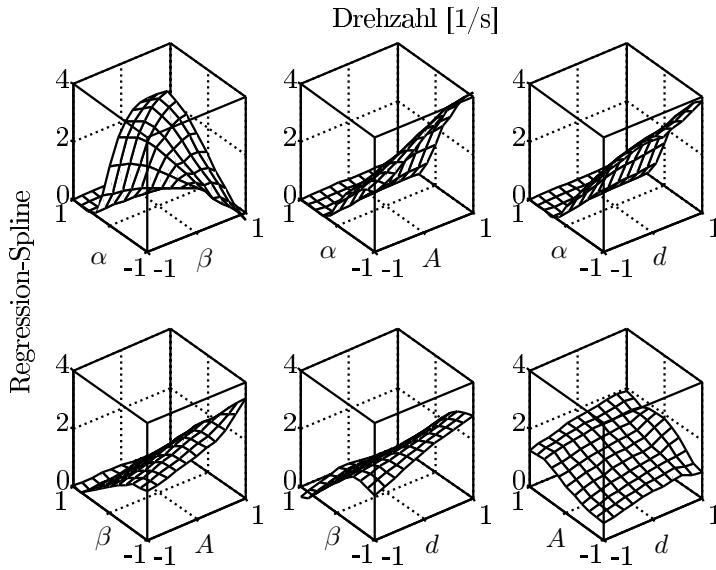


Abb. B.17 Symmetrisches LHC (zentrierte Diskrepanz), Interaktion (*Drehzahl*), Spline Regression

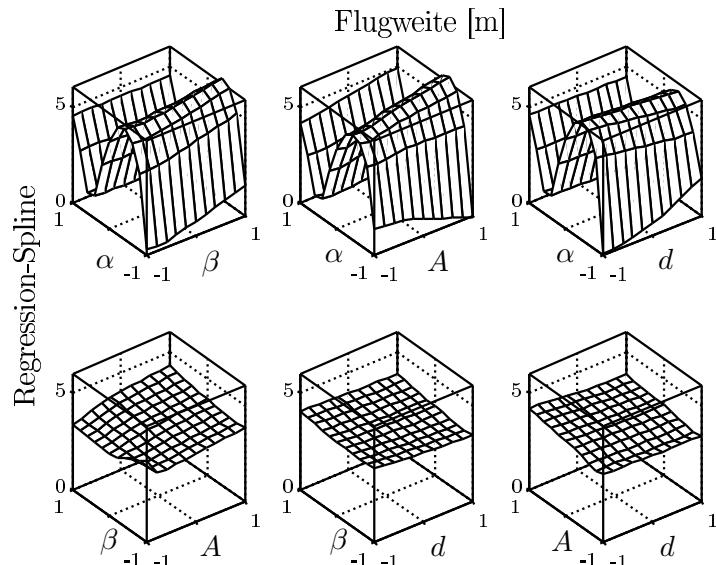


Abb. B.18 Symmetrisches LHC (zentrierte Diskrepanz), Interaktion (*Flugweite*), Spline Regression

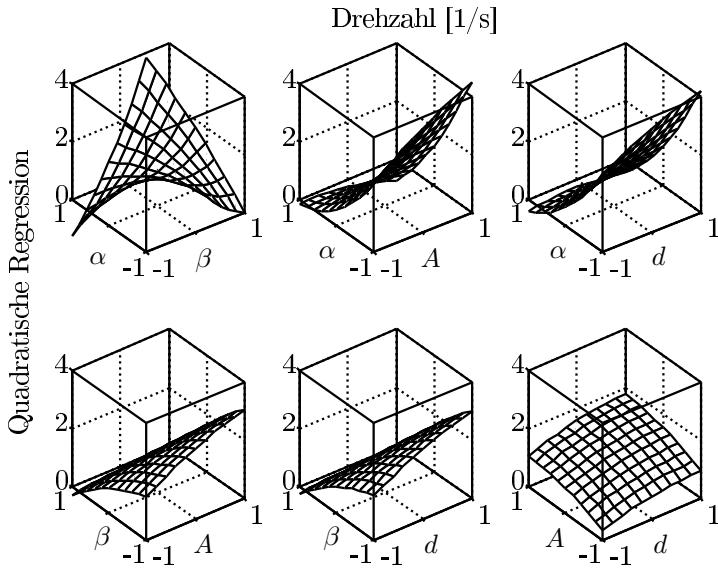


Abb. B.19 Symmetrisches LHC (zentrierte Diskrepanz), Interaktion (*Drehzahl*), Quadratische Regression

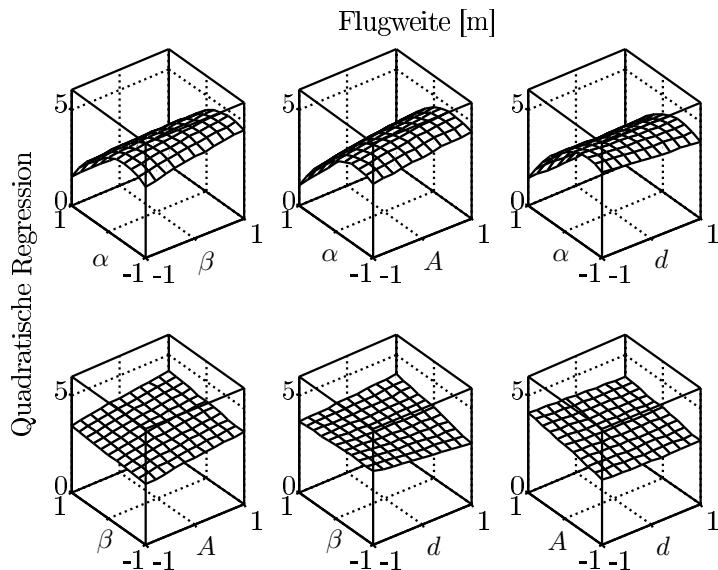


Abb. B.20 Symmetrisches LHC (zentrierte Diskrepanz), Interaktion (*Flugweite*), Quadratische Regression

Qualität der Metamodelle

Zur Qualitätsprüfung der Metamodelle wird die Approximationsgenauigkeit an 30 zufällig im Faktorraum verteilten Testpunkten mit Hilfe des *Mean Square Error* (MSE) analysiert (Kapitel 9.18). Abbildung B.21 zeigt für die Ausgangsvariable Drehzahl eine deutliche Reduktion des MSE, wenn eine sinnvolle Kombination von Testfeld (LHC) und Metamodellansatz (KNN und RS) verwendet wird. Zusätzlich ist zum MSE ein direkter Vergleich von Originalmodell (x-Achse) und Approximation der Metamodelle (y-Achse) für die Testdaten dargestellt. Bei Betrachtung der Drehzahl weist das KNN in diesem Beispiel einen leicht besseren MSE als die Spline Regression auf.

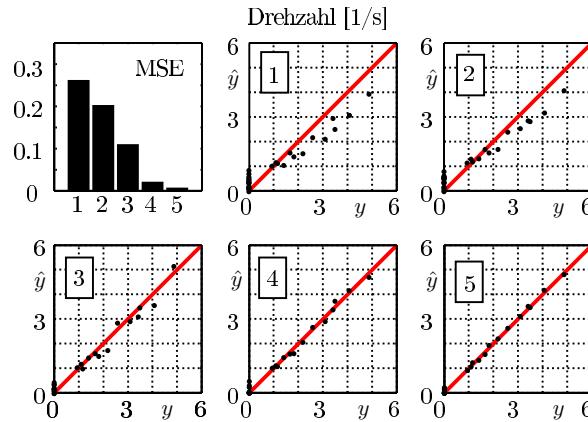
Abbildung B.22 zeigt zum Vergleich die Analyse für die gleichen Testpunkte und die Vorhersage der Flugweite. Durch den komplexeren Zusammenhang ist der Vorteil der flexiblen Metamodelle (KNN und RS) in Kombination mit dem LHC gegenüber klassischen Verfahren deutlicher erkennbar. Im Gegensatz zur Drehzahl ist diesmal das Metamodell basierend auf die Spline Regression etwas besser als das KNN. Dieses zeigt, dass keine allgemein gültige Empfehlung für ein spezielles Metamodellverfahren gegeben werden kann, da die erreichbare Qualität von dem jeweiligen zu approximierenden Zusammenhang inklusive eventueller Messfehler sowie dem aktuell verwendeten Testfeld abhängt. Es ist jedoch bereits erkennbar, dass flexible Metamodellansätze gute Approximationenmodelle bei ausreichender Informationsmenge (hier durch gleichverteiltes Testfeld) erzeugen.

Resümee

Die Betrachtung der verschiedenen Metamodellergebnisse zeigt, dass klassische Testpläne mit wenig Faktorstufen häufig keine ausreichenden Datenmengen liefern, um komplexe Zusammenhänge zwischen Faktoren und Ausgangsvariablen abzubilden. Auch eine drastische Erhöhung der Testpunkte führt dabei nur zu einer geringen Verbesserung der Approximationsgenauigkeit, wenn die Stufenanzahl nicht deutlich erhöht wird. Werden die Testpunkte jedoch gleichförmig im Faktorraum verteilt, so ist es möglich, Daten zu ermittelt, die zur Abbildung komplexer Zusammenhänge ausreichen.

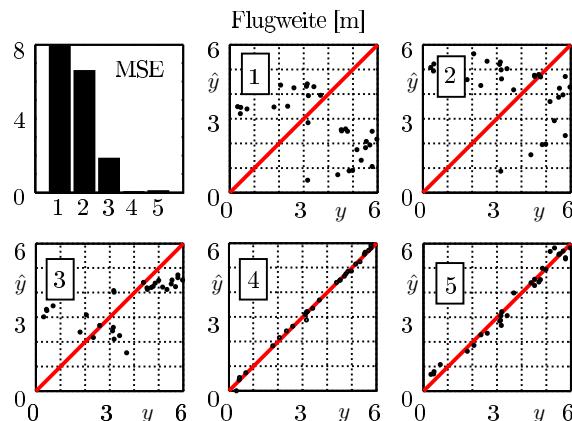
Klassische Metamodelle mit fest vorgegebenen Zusammenhängen können diese Informationen meistens nicht vollständig verwerten, so dass die Zusammenhänge auch bei ausreichenden Informationen nicht abgebildet werden können. Erst durch die Kombination von passenden Testfeldern und flexiblen Metamodellansätzen werden ausreichende Informationsmengen ermittelt und bei der Erstellung von Metamodellen vollständig berücksichtigt, wodurch komplexe Zusammenhänge ausreichend genau abgebildet werden.

Die in Kapitel 8 und Kapitel 9 dargestellten Versuchspläne und Metamodellansätze liefern in Abhängigkeit von der Analyseaufgabe zwar leicht unterschiedliche Approximationen, jedoch sind die Unterschiede bei richtigem Einsatz gering und für die überwiegenden Anwendungen vernachlässigbar.



- 1: Vollfaktorplan, 2 Stufen, Lineare Regression
- 2: Vollfaktorplan, 3 Stufen, Quadratische Regression
- 3: Symmetrisches LHC (ZD), Quadratische Regression
- 4: Symmetrisches LHC (ZD), Regression-Spline
- 5: Symmetrisches LHC (ZD), Künstliches Neuronales Netzwerk

Abb. B.21 Mean Square Error (MSE), *Drehzahl*



- 1: Vollfaktorplan, 2 Stufen, Lineare Regression
- 2: Vollfaktorplan, 3 Stufen, Quadratische Regression
- 3: Symmetrisches LHC (ZD), Quadratische Regression
- 4: Symmetrisches LHC (ZD), Regression-Spline
- 5: Symmetrisches LHC (ZD), Künstliches Neuronales Netzwerk

Abb. B.22 Mean Square Error (MSE), *Flugweite*

B.3 Sensitivitätsanalyse

Eine Sensitivitätsanalyse wird mit dem in Kapitel 13.3.3 beschriebenen *Extended Fourier Amplitude Sensitivity Test* (eFAST) durchgeführt. Mit den Randbedingungen $M = 4$ und $n_r = 10000$ ergeben sich die Analysefrequenzen $\omega = \{1249, 1, 78, 156\}$. Abbildung B.23 und Tabellen B.2 bis B.4 zeigen die Ergebnisse der durchgeführten Analyse. Die bereits durch die qualitative Betrachtung der Haupt- und Interaktionseffekte gefundenen Abhängigkeiten zwischen Faktoren und Ausgangsvariablen werden hier quantitativ bestätigt. Die beiden Winkel α und β weisen den größten Effekt auf die Drehzahl auf, wobei sie ebenfalls einen deutlichen Interaktionsanteil beinhalten. Die Flugweite wird vom Winkel α beeinflusst, wobei lediglich ein kleiner Interaktionseffekt vorhanden ist. Der Volumenstrom hingegen ist nahezu ausschließlich und direkt von der Querschnittsfläche A abhängig. In weiteren Analysen könnte durch dieses Ergebnis beispielsweise auf die Betrachtung des Durchmessers verzichtet werden und der Volumenstrom unabhängig von der Drehzahl und der Flugweite optimiert werden.

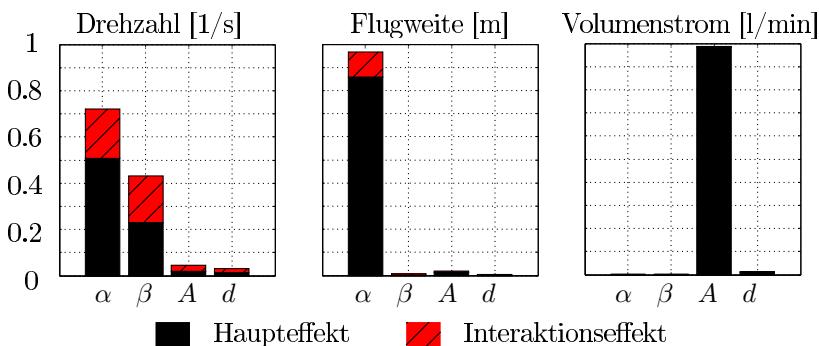


Abb. B.23 Sensitivitätsanalyse mit eFAST

Faktor	Haupt	Total
α	0.50926	0.72281
β	0.23257	0.43107
A	0.019988	0.044047
d	0.014103	0.030613

Tabelle B.2 Drehzahl: eFAST ($M = 4, n_r = 10000, \omega = \{1249, 1, 78, 156\}$)

Faktor	Haupt	Total
α	0.85696	0.96732
β	0.0029497	0.0076774
A	0.014052	0.019266
d	0.0024516	0.0045866

Tabelle B.3 Flugweite: eFAST ($M = 4, n_r = 10000, \omega = \{1249, 1, 78, 156\}$)

Faktor	Haupt	Total
α	0.00018667	0.00082598
β	0.00010181	0.00069857
A	0.98417	0.988
d	0.011448	0.015213

Tabelle B.4 Volumenstrom: eFAST ($M = 4, n_r = 10000, \omega = \{1249, 1, 78, 156\}$)

B.4 Optimierung

Zur Optimierung des Rasensprengers soll in diesem Beispiel die Drehzahl sowie die Flugweite maximiert werden. Die Nebenkosten (Wasserverbrauch) und somit der Volumenstrom ist jedoch zu minimieren.

Abbildung B.24 zeigt 1000 Ergebnisse für zufällige Faktorkombinationen, welche eine große Variation im Zielgrößenraum aufweisen (rot). Weiterhin ist die ange-näherte Pareto-Grenze dargestellt, welche durch eine NSGA2 Optimierung (Kapi-tel 10) ermittelt wurde. Sie zeigt deutlich den erreichbaren Kompromiss zwischen den drei gewählten Zielgrößen.

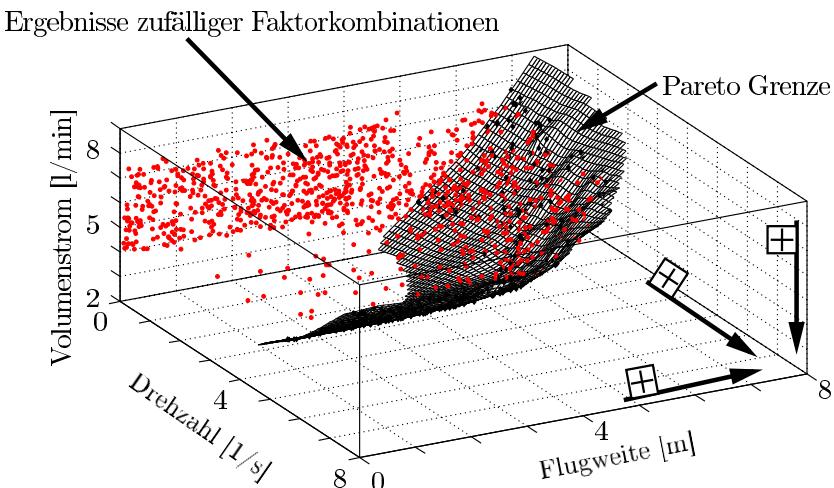


Abb. B.24 NSGA2 Optimierung (Rasensprenger)

Nachdem die erreichbaren Drehzahlen, Flugweiten und Volumenströme bekannt

sind kann nun ein akzeptabler Kompromiss mit angepassten Randbedingungen für die Zielgrößen ermittelt werden:

$$\begin{aligned} \text{Drehzahl} &\geq 4 [1/s] \\ \text{Flugweite} &\geq 4 [m] \\ \text{Volumenstrom} &\leq 5 [l/min] \end{aligned}$$

Auf der ermittelten Pareto-Grenze finden sich in diesem Beispiel 18 Faktoreinstellungen, die diese Randbedingungen erfüllen. Die Histogramme der 18 Faktorkombinationen

Drehzahl [1/s]	Flugweite [m]	Volumenstrom [l/min]	α [°]	β [°]	A [mm ²]	d [m]
4.09	4.77	4.06	28.66	6.01	2.12	0.20
4.70	4.28	4.47	17.95	10.09	2.35	0.19
4.27	4.66	4.12	25.44	5.02	2.15	0.20
4.02	4.09	3.95	15.42	5.18	2.04	0.17
5.05	4.23	4.70	18.17	2.59	2.48	0.20
4.36	4.74	4.27	26.18	5.46	2.24	0.20
4.22	4.62	4.09	23.95	11.38	2.14	0.20
4.22	5.03	4.77	24.32	4.71	2.49	0.16
4.45	4.38	4.20	19.85	12.13	2.20	0.20
4.16	4.68	4.11	24.26	11.11	2.14	0.19
4.48	4.59	4.34	22.27	6.19	2.27	0.19
5.38	4.19	5.00	18.17	0.10	2.65	0.20
4.11	5.20	4.74	30.50	9.40	2.49	0.17
4.82	4.84	4.99	23.48	2.27	2.63	0.18
5.15	4.03	4.78	15.97	5.07	2.52	0.19
4.52	4.51	4.34	21.16	5.89	2.27	0.19
4.86	4.34	4.71	18.10	8.27	2.48	0.19
4.38	4.46	4.16	20.95	12.13	2.17	0.20
4.67	4.36	4.45	18.95	8.96	2.33	0.19
4.12	5.14	4.66	31.05	14.58	2.46	0.19
5.19	4.09	4.81	16.70	3.45	2.54	0.20

Tabelle B.5 Ausgewählte Punkte der Pareto-Grenze

nationen in Abbildung B.25 zeigen, dass die gewählten Pareto-optimalen Ergebnisse kleine β Winkel und Düsenquerschnitte A sowie große Durchmesser d aufweisen. Für den Winkel α führt der gewählte Kompromiss zu Winkel um 21°.

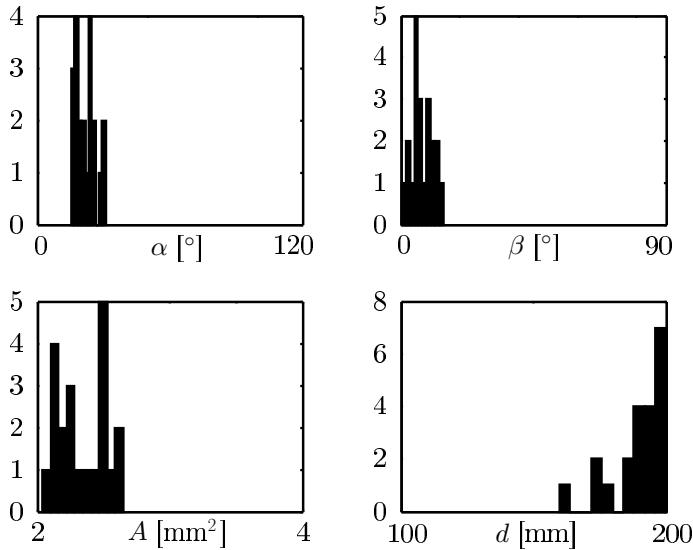


Abb. B.25 Histogramm ausgewählter Pareto-optimaler Faktorkombinationen

Im weiteren ist es möglich zusätzliche Auswahlkriterien bezüglich der Faktoreinstellungen zu berücksichtigen. Ist zum Beispiel eine kostengünstige Fertigung von Winkeln $\beta \geq 10^\circ$ und Querschnitten $A \leq 2.2\text{mm}^2$ möglich, so reduzieren sich die in Frage kommenden Pareto-optimalen Punkte auf die in der Tabelle B.6 aufgeführten. Im weiteren Entwicklungsprozess kann nun eine oder mehrere dieser Lösungen weiter verfolgt beziehungsweise genauer analysiert werden.

Drehzahl [1/s]	Flugweite [m]	Volumenstrom [l/min]	$\alpha [^\circ]$	$\beta [^\circ]$	$A [\text{mm}^2]$	$d[\text{m}]$
4.22	4.62	4.09	23.95	11.38	2.14	0.20
4.45	4.38	4.20	19.85	12.13	2.20	0.20
4.16	4.68	4.11	24.26	11.11	2.14	0.19
4.38	4.46	4.16	20.95	12.13	2.17	0.20

Tabelle B.6 Ausgewählte Punkte der Pareto-Grenze mit Berücksichtigung der Faktoreinstellungen

Nomenklatur

\bar{v}	Mittelwert der Variablen v
\hat{v}	Approximation der Variablen v
ε	Fehler
ε	Grenze (Zielgröße), Gitterbreite
σ	Standardabweichung
$\hat{\sigma}$	geschätzte Standardabweichung
$\hat{\sigma}_{(-i)}$	geschätzte Standardabweichung, ohne Berücksichtigung des Testwertes von Versuchslauf i
η^2	Korrelationsverhältnis
ω_j	Frequenz des Faktors j (FAST und eFAST)
b	Regressionskoeffizient (siehe auch c)
B	Basisfunktion
c	Konstante oder Regressionskoeffizient
\mathbf{c}	Vektor der Regressionskoeffizienten
c_0	Mittelwert (totaler)
c_j	Halber Haupteffekt von Faktor $j \hat{=} E_j/2$
c_{jk}	Halber Interaktionseffekt von Faktor j und $k \hat{=} E_{jk}/2$
c_{jj}	Halber quadratischer Effekt von Faktor $j \hat{=} E_{jj}/2$
D	Gesuchte Funktion
D	Menge von Abständen zwischen Testpunkten (MaxiMin)
D	Diskrepanz (Gleichverteilung)
d	Abstand zweier Testpunkte (eng: distance)
$DFITS$	Difference in Fits
E	Effekt oder Erwartungswert
E_j	Effekt von Faktor j
E_{jj}	Quadratischer Effekt von Faktor j
E_{jk}	Interaktionseffekt von Faktor j und k
f	Wahre Funktion des Gütekriteriums oder der Ausgangsvariablen
g	geometrischer Mittelwert der Versuchsergebnisse y_1 bis y_{n_r}
G_j	Transformationsfunktion für Faktor j (FAST und eFAST)
G_i	i^{te} Generation (genetische Optimierung)

H	Hutmatrix
h_{ii}	Hauptdiagonalelement der Hutmatrix
i	Laufvariable (häufig für Versuchslauf)
j	Laufvariable (häufig für Faktor)
k	Laufvariable
k	Kind-Individuum (genetische Optimierung)
J_k	Anzahl gleicher Abstände d_k (MaxiMin)
n_a	Anzahl der Ausgangsvariablen (Künstliches Neurales Netzwerk)
n_b	Anzahl der Basisfunktionen, Anzahl Gewichte oder Anzahl Bits
n_c	Anzahl der Regressionskoeffizienten
n_e	Anzahl der Neuronen
n_{E1}	Anzahl der Neuronen in versteckter Ebene 1
n_f	Anzahl der Faktoren
n_g	Anzahl der Individuen in einer Generation (Genetische Optimierung)
n_l	Anzahl der Level $\hat{=} n_s$
n_m	Anzahl der Modellkonstanten
n_q	Anzahl der Qualitätsmerkmale
n_r	Anzahl der Versuchsläufe
n_s	Anzahl der Stufen $\hat{=} n_l$
n_y	Anzahl der Ausgangsgrößen eines Künstlichen Neuronalen Netzwerks
n_z	Anzahl der Zielgrößen
p	Anzahl Modellparameter
q_i	Rampenfunktion der i^{ten} Ausgangsvariable
R	Korrelationsmatrix
r	Residuum $y_i - \hat{y}_i$
S	Schwellwert
T	Testfeld
U	Cholesky-Zerlegung $R = U'U$
u	Zufallszahl
x	Eingangsvariable(n), Faktor(en)
x_{ij}	Faktor j aus Versuchslauf i
X	Matrix der Eingangsvariablen $X = \begin{pmatrix} x_{11} & \cdots & x_{1n_s} \\ \vdots & \ddots & \vdots \\ x_{n_r1} & \cdots & x_{n_rn_s} \end{pmatrix}$
V	Varianz
\hat{V}	Schätzwert der Varianz
VIF	Varianz-Inflations-Faktor
w	Gewichtungsfaktor
y	Ausgangsvariable, Gütekriterium, Qualitätsmerkmal
\mathbf{y}	Vektor der Ausgangsvariable für alle Versuchsläufe
\bar{y}	Mittelwert der Ausgangsvariable für alle Versuchsläufe
y_i	Ausgangsvariable (gemessen) des i^{ten} Versuchslauf
\hat{y}	Approximierte Ausgangsvariable
	Beispiel: $\hat{y} = c_0 + c_1 \cdot x_1 + c_2 \cdot x_2 + c_{11} \cdot x_1 + c_{22} \cdot x_2 + c_{12} \cdot x_1 \cdot x_2$

\hat{y}_i	Approximierte Ausgangsvariable für den Versuchslauf i
$\hat{y}_{i(-i)}$	Approximierte Ausgangsvariable für den Versuchslauf i , ohne Berücksichtigung des gemessenen Wertes y_i
$\hat{y}_{j(-i)}$	Approximierte Ausgangsvariable für den Versuchslauf j , ohne Berücksichtigung des gemessenen Wertes y_i von Versuchslauf i .
Y	Vektor der Ausgangsvariablen $Y = (y_0, \dots, y_n)'$
z	Zielgröße
z	transformiertes Qualitätsmerkmal (Box-Cox)

Abkürzungen und Markennamen

ABFC	Adaptive Basis Function Construction
AIC	Akaike Informationskriterium
ANOVA	Analysis of Variance
ARD	Automatic Relevance Determination
BBD	Box Behnken Design
BIC	Bayesian information criterion
CCD	Central Composite Design
CD	Crowding Distance
CE	Computer Experiment
Cov	Kovarianz (covariance)
CPD	Coefficient of partial determination
CR	Correlation ratio
CV	Kreuzvalidierung (Cross-Validation)
Design Expert®	Statistikprogramm der Firma StatEase Inc.
DIFFITS	Difference in Fits
DoE	Design of Experiment
DOE	Design of Experiment
FAST	Fourier Amplitude Sensitivity Test
FDS	Fraction of Design Space
FFT	Fast Fourier Transformation
ED	Einhüllende Diskrepanz (=WD)
eFAST	Erweiterter Fourier Amplitude Sensitivity Test
GLP	Good Lattice Point Method
GPM	Gauß Prozess Modell
GGM	Gute Gitter Punkt Methode (=GLP)
GGT	Größter Gemeinsamer Teiler
IMSE	Integrated Mean Squares Error
JMP®	“John’s Macintosh Project”, Statistikprogramm der Firma SAS
KNN	Künstliches Neuronales Netzwerk
KV	Korrelationsverhältnis
LASSO	Least absolute shrinkage and selection operator

LCLS	Links Zyklisches LS (left cycle latin square)
LHC	Latin Hypercube
LHD	Latin Hypercube Design
LHS	Latin Hypercube Sampling
LS	Latin Square
LS	Lokale Suche (local search)
MARS®	Multi Adaptive Regression Splines
Minitab™	Statistikprogramm der Firma Minitab Inc.
MMSE	Minimum Mean Square Error
MSE	<i>mean squared error</i> (mittlere quadratische Approximationsfehler)
mod	Modulo (Eine Funktion, die den Rest aus der Division zweier ganzer Zahlen angibt.)
NOA	Nearly Orthogonal Array
OA	Orthogonal Array
OALHD	Orthogonal Array basierendes Latin Hypercube Design
OLHD	Orthogonal Latin Hypercube Design
PCC	Partial Correlation Coefficient
PDF	Probability Density Function
PRESS	Predictive Sum of Squares
PSS	partial sum of squares
RAND	Zufallszahl
RBF	Radial Basis Funktionen
RIC	Risiko Informationskriterium
RMSE	<i>root mean squared error</i> (\sqrt{MSE})
RSS	Residual Sum of Squares = <i>SSE</i>
SA	Simuliertes Abkühlen (simulated annealing)
SBS	Sequential Backward Selection
SE	Stochastisches Evolutionsmethode (stochastic evolutionary algorithm)
SFFS	Seuqential Floating Forward Selection
SFS	Sequential Forward Selection
SI	Sensitivitätsindex
SI	Système International d'unités
SRC	Standardized regression coefficient
SSE	Error sum of squares
SSR	Regression sum of squares
SST	Total sum of squares
Statgraphics®	Statistikprogramm der Firma StatPoint Technologies Inc.
TA	Threshold Accepting
VCE	Variance Conditional Expectation
VIF	Varianz-Inflations-Faktor (variance inflation factor)
Vol	Volumen
WD	Wrap Around Discrepancy (=ED)
ZD	Zentrierte L_2 Diskrepanz

Sachverzeichnis

- D_2 -Diskrepanz, 195
 D_∞ Diskrepanz, 195
 D_p Diskrepanz, 194
 F -Verteilung, 119, 127, 133
 R^2 , 123–125, 136
 R^2_{adj} , 124, 125
 R^2_{adj} , 237
 α -Risiko, 102, 104, 106–108, 110, 120, 123, 128, 129, 132
 β -Risiko, 102, 106, 108
 ε -MOEA, genetische Optimierung, 364
 v -SVR-Verfahren, 269
 p -Wert, 119, 120, 133–136
(t,m,s) Netz, 204
(t,s)-Sequenz, 204
- ABFC: Adaptive Basis-Funktion Konstruktion, 241
Abstand
 Testpunkte, 191
Abweichungen vom Versuchsplan, 62
Achsentransformation, 24
Adaptive Basis-Funktion Konstruktion, 241
Akaike's Informationskriterium, 237
Aktivierung eines Neurons, 303
Akzeptanzschwelle, 103, 104, 106
Alternativhypothese, 111
Anderson-Darling-Test, 134
ANOVA, 73, 154, 158
Approximationsmodell, 231
ARD, 300
Auflösung, 31, 33, 40
Auflösungsstufe III, 64
Auflösungsstufe IV, 64
Auswertung, 9
automatic relevance determination, 300
- Backpropagation, 305
Bartlett-Test, 135
Bat Algorithm, 342
Bauteiltoleranzen, 153
Bayessches Informationskriterium, 237
Bedingte Verteilung, 291
Beschreibungsmodelle, 28
Bi-Plot, 146
Binäre Kodierung, 353
 Kreuzung, 355
 Mutation, 357
Binomialverteilung, 103, 106, 110, 127
Blütenbestäubung-Optimierung, 344
Blockbildung, 87, 90, 96, 99, 120–122
Blockschaubild, 3
Box, George, 88, 89, 91, 92, 95
Box-Behnken-Design, 42
Box-Behnken-Designs, 92
Box-Cox Plot, 83
Box-Cox Transformation, 83
Box-Cox- Transformation, 92
- CCD, 40
Center for Quality and Productivity
 Improvement, 91, 92
center point, 40, 84
Central-Composite Design, 169
Central-Composite-Design, 40
Coefficient of partial determination, 418
Computer-Experiment, 179
Computermodelle, 179
Computersimulationen, 39
Cook-Distanz, 82
corporate knowledge, 136
Correlation ratio, 419
Cross Validation, 315

- cross validation
 - Regressionsbaum, 252
- Crowding Distance, NSGA-II, 360
- Cumulative Normal Distribution, 219
- D-optimal, 55
- Daniel-Plot, 69
- Datenkontrolle, 92
- Deming, W.E., 92
- desirability function, 141, 173
- Determinationskoeffizient - partieller, 418
- DFITS, 81
- Diagnose, 61
- Diagnoseinstrument, 64
- Discrepancy, 194
- Diskrepanz, 194
- Diskrepanz, D_2 , 195
- Diskrepanz, D_∞ , 195
- Diskrepanz, D_p , 194
- Diskrepanz, einhüllende, 197
- Diskrepanz, modifiziert, 196
- Diskrepanz, zentrierte, 196
- Dokumentation der Ergebnisse, 136
- Dominanz, 326
 - Grenzwert, 328
 - Priorität, 328
- Draper-Lin, 41
- Dreher, 63
- Dreifachwechselwirkung, 19, 29
- eFAST, 427
 - totaler Sensitivitätsindex, 427
- Effekt, 2, 12, 99, 127, 163
- Effekt, echter, 87
- Effekt, relevanter, 128
- Effekt, signifikanter, 127, 128, 133, 136
- Effekt, wahrer, 88
- Effekt, zufälliger, 117
- Effekt-Diagramm, 13
- Effekte
 - scheinbare, 68
 - wahre, 72
- Effektstärke, 64
- Effizienz, 28
- Eigenwerte, 146
- Eingangsgrößen, 3
- Einhüllende Diskrepanz, 197
- enhanced stochastic evolutionary algorithmus, 216
- Entropie, 193
- Erklärungskraft eines Modells, 122, 123, 125, 136
- error states, 148
- Error sum of squares, 417
- Ersatzmodell, 231
- Euklidische Norm, 191
- Euler'sches Quadrat, 58
- Extended FAST, 427
- extended Fourier Amplitude Sensitivity Test, 427
- Extrapolationen, 24
- F-Verhältnis, 75
- F-Verteilung, 133
- Faktor, 2
 - kategorial, 156
- Faktoreffekt, 90, 94, 99, 106, 111, 113, 117, 125
- Faktorelimination, 122, 123
- Faktoren, 5
- Faktorstufen, 98, 116, 128, 132
- Faktorwahl, 234
- Fallstudie, 9
- Faltung, 35
- FAST, 425
- Fast Orthogonal Arrays, 204
- FAST, extended, 427
- FastICA, 410
- Faure-Sequenz, 202
- FDS-Plot, 66
- Feedforward Netzwerk, 303
- Fehler 1. Art, 102
- Fehler 2. Art, 102
- Fehlerrückführung, 305
- Feld
 - äußeres, 150
 - inneres, 150
- FireFly, 340
- Fisher, R.A., 89–96, 103, 117
- Fledermaus-Optimierung, 342
- Flower pollination Algorithm, 344
- fold over, 35
- Fourier Amplitude Sensitivity Test, 425
- Fraction of Design Space, 66
- fractional factorial designs, 28
- fraktionelle faktorielle Versuchspläne, 28
- Freiheitsgrad, 114, 115, 119, 120, 127
- Freiheitsgrade, 76
- G-optimal, 56
- Gauß Prozess Modelle, 290
 - ARD , 300
 - Kovarianzfunktionen, 295
- Gauß-Verteilung, 219
- Gemischte Orthogonale Arrays, 204
- Generalized Faure Sequence, 202
- Generator, 32
- Genetische Optimierung, 353

- Gesamtvarianz, 73
Gewichteter Abstand, 275
Glühwürmchen, 340
Gleichungssystem, 29
Gleichverteilte Testfelder, 209
Gleichverteilung, 194
Gosset, W.S., 91
GPM, 290
Greedy, 290
Grenzwertsatz, Zentraler, 96
Grundbegriffe, 2
- half normal plot, 117
Half-Normal-Plot, 69
Halton-Sequenz, 199
Hammersley-Sequenz, 201
Hat Matrix, 237
Hat Matrix H , 237
Hauptachsenanalyse, 396
Hauptachsentransformation, 146
Haupteffekt, 100, 113–115, 123
Hauptkomponentenanalyse, 396
 Datenvorbereitung, 405, 406, 409
 Kurtosis, 407
 Negentropie, 408
Hebelwert, 67
Heuristik, 325
Hunter, W.S., 91, 92
Hut Matrix, 237
Hutmatrix, 67
Hybrid-Quasi-Monte-Carlo, 204
Hypervolumen, 374
Hypothese, 102–104, 106, 107, 110, 117, 127
Hypothesentest, 68, 119
- ICA, 403
 Datenvorbereitung, 409
 Kurtosis, 407
 Negentropie, 408
 Nicht-Normalverteilt, 406
 Unabhängigkeit, 405
Importance Measure, 420
Impurity
 Regressionsbaum, 250
Independent Component Analyses, 403
Individuum, 327
Informationskriterium, 237
 R^2_{adj} , 237
 Aikaike's Informationskriterium, 237
 Bayessches Informationskriterium, 237
 Mallows C_p , 237
Integrität, hierarchische, 123
Interaktion, 88, 122
Inverse Verteilungsfunktion, 219
- Inverse-Distance-Weighting Method, 275
Inversen-Abstands-Gewichtungs Methode, 275
irreguläre Felder, 33
- k-fold Kreuzvalidierung, 316
Künstliche Neuronale Netzwerke, 302
Kendalls τ , 388
Kernel, 265
 Support Vector Machines, 265
Kernel Principal Component Analyses, 400
Kernel Regression, 243
Kernel-Hauptachsenanalyse, 400
Kernel-Hauptkomponentenanalyse, 400
Klassifikation
 nicht komplett linear separierbarer Daten, 262
 Support Vector Machines, 258
Kleinster Abstand, 275
Kodierung, 6
Kodierungskette, 33
Koksma-Hlawka Ungleichung, 196
Kolmogoroff-Smirnoff-Test, 134
Komponentenanalyse, 395
 FastICA, 410
 Hauptachsen, 396
 Hauptkomponenten, 396
 ICA, 403
 Kernel-Hauptkomponenten, 400
 kPCA, 400
 PCA, 396
 Unabhängige, 403
 Unabhängigkeitsanalyse, 403
Konstruktionsmethoden
 Testfelder, 198
Konsumentenrisiko, 103
Kontrastmethode, 12
Kontrollverfahren, 61, 87
Korrelation
 Kendalls τ , 388
 Pearson, 381
 Rangkorrelation, 385
 Scheinkorrelation, 383
 Signifikanz, 383
 Spearman, 387
 verdeckte, 383
Korrelationsanalyse, 381
Korrelationsmatrix, 64
Korrelationsverhältnis, 419
Kosten-Nutzen-Analyse, 153
Kovarianz, 381
Kovarianzfunktionen, 295
kPCA, 400
Kreuzung
 Binäre Kodierung, 355

- Reelle Faktoren, 355
- Kreuzung, genetische Optimierung, 355
- Kreuzvalidierung, 315
 - Regressionsbaum, 252
 - Kreuzvalidierung, k-fold, 316
 - Kreuzvalidierung, Leave-One-Out, 316
 - Kriging, 276
 - universal, 283
 - Kronecker-Sequenz, 202
 - Kurtosis, 407
- L12, 34
- L16, 32
- L8, 32
- L9, 58
- LASSO, 238
- Latin Hypercube, 44, 205
- Latin Hypercube Design, 205
- Latin Hypercube Sampling, 205
- Latin Square, 57
- Latin-Hypercube-Design, 46
- Least Square Support Vector Regression, 272
- Leave-One-Out Kreuzvalidierung, 316
- Lebensdauer, 88
- Level, 6
- Levene-Test, 135
- leverage, 67
- Lineare Regression, 232
- lineare Straffunktion (SVR), 266
- lineares Beschreibungsmodell, 21
- lineares Gleichungssystem, 28
- logarithmische Transformation, 50
- Lokale Polynom Regression, 243
- lokale Suche, local search, 215
- Low Discrepancy Procedures, 199
- Mallows C_p , 237
- MaxiMin, 191
- Mean Squared Error, 314
- Mean Squares, 115, 116
- Menschenverstand, gesunder, 135
- Messmittelfähigkeitsanalyse, 154
- Messrauschen, 87, 90, 98, 99, 122
- Messungenauigkeit, 90
- Metaheuristik, 325
- Metamodell
 - Regressionsbaum, 248
 - Splines, 254
- Metamodelle, 180, 231
 - Adaptive Basis-Funktion Konstruktion, 241
 - Faktorwahl, 234
 - Gauß Prozess Modelle, 290
 - Gewichteter Abstand, 275
 - Inverse-Distance-Weighting Method, 275
- Inversen-Abstands-Gewichtungs Methode, 275
- Künstliche Neuronale Netzwerke, 302
- Kleinster Abstand, 275
- Kriging, 276
- Lineare Regression, 232
- Modellkombination, 313
- Multivariate Adaptive Regression Splines, 256
- Nächster Punkt, 275
- Nearest Point, 275
- Polygon Verfahren, 275
- Polynome, 234
- Qualität, 314
- Radial Basis Funktion, 284
- Ridge Regression, 273
- Robuste Regression, 239
- Support Vector Machines, 258
- Universal Kriging, 283
- Metamodelle, Kernel Regression, 243
- Metamodelle, Lokale Polynom-Regression, 243
 - Method of least squares, 233
 - Methode der kleinsten Fehlerquadrate, 233
 - MiniMax, 191
 - Mischungspläne, 52
 - Mittelwert, 88, 93, 96, 97, 111, 118, 119, 127
 - Mittlerer quadratischer Approximationsfehler, 314
 - Mixed Orthogonal Arrays, 204
 - Modelleffizienz, 125
 - Modellfehler, 119, 120, 133, 134
 - Modellkombination, 313
 - Modellkonstanten, 22, 28, 39
 - Modifizierte Diskrepanz, 196
 - Monte-Carlo, 199
 - Monte-Carlo, Hybrid-Quasi, 204
 - Monte-Carlo, Quasi, 199
 - Monte-Carlo-Verfahren, 44
 - MSE, 237
 - Multiple-Response-Optimisation, 139, 151, 175
 - Multivariate Adaptive Regression Splines, 256
 - Mutation
 - Binäre Kodierung, 357
 - Reelle Faktoren, 357
 - Mutation, genetische Optimierung, 357
 - Nächster Punkt, 275
 - Naturanaloge Optimierung, 336
 - Nearest Point, 275
 - Nearly Orthogonal Arrays, 204
 - Negentropie, 408
 - Neuron

- Aktivierung, 303
- Nicht-Normverteilt, 406
- NichtLineareKorrelation, 389
- Nichtlinearitäten, 84
- Norm, euklidische, 191
- Normal Probability Plot, 134, 135
- Normalverteilung, 96, 118, 134, 219
- normierte Regressionskoeffizienten, 416
- NSGA-II
 - genetische Optimierung, 358
- Nullhypothese, 102–104, 108, 110, 111, 117–119, 126, 127
- One-Of-N Kodierung, 308
- Optimierung, 325
 - ϵ -MOEA, 364
 - Bat, 342
 - binäre Kodierung, 353
 - Blütenbestäubung, 344
 - Dominanz, 326
 - FireFly, 340
 - Fledermaus, 342
 - Flower Pollination, 344
 - genetisch, 353
 - Glühwürmchen, 340
 - Heuristik, 325
 - Hypervolumen, 374
 - Individuum, 327
 - mehrere Zielgrößen, 348
 - Metaheuristik, 325
 - Naturanaloge, 336
 - NSGA-II, 358
 - Pareto-Grenze, 326
 - Partikelschwarmoptimierung, 336
 - PROB, 351
 - PURE RANDOM, 350
 - R-Indikator, 368
 - Randbedingungen, 332
 - RANDOM, 350
 - Rang, 359
 - ROUNDS, 351
 - Symbiotic Organisms Search, 346
 - Symbiotische Organismen, 346
 - Testfelder, 214
 - Zielgrößen, 325
- Optimierung, *Entropie*, 217
- Optimierung, *MaxiMin_p*, 217
- Optimierung, *ZD₂*, 217
- Optimierung, Crowding Distance (NSGA-II), 360
- Optimierung, enhanced stochastic evolutionary algorithmus, 216
- Optimierung, Kreuzung, 355
- Optimierung, local search, 215
- Optimierung, lokale Suche, 215
- Optimierung, Mutation, 357
- Optimierung, simulated annealing, 216
- Optimierung, simuliertes Abkühlen, 216
- Optimierung, stochastische Evolutionsverfahren, 216
- Optimierung, Suchalgorithmus, 215
- Optimierung, threshold accepting, 215
- Optimierung, verbessertes stochastisches Evolutionsverfahren, 216
- Optimierungsrechnung, 140
- orthogonal, 7, 62
- Orthogonal array-based Latin Hypercube, 206
- Orthogonal Design Tables, 205
- Orthogonale Arrays, 204
- Orthogonale Design Tabellen, 205
- Orthogonales Latin Hypercubes, 207
- over-fit, 76, 169
- Overfitting, 309
- p value, 110, 119, 127
- p-Wert, 75
- Parameter, 5
- Parameterbereich, 6
- Parameterdesign, 139, 147
- Parameterdiagramm, 147, 160, 165
- Pareto-Grenze, 186, 326
 - Qualität, 367
- Pareto-optimal, 187
- Partial Correlation Coefficient, 418
- Partial sum of squares, 418
- Partialsumme der Quadrate, 417
- Partielle Korrelationsfaktoren, 418
- Partieller determinationskoeffizient, 418
- Partikelschwarmoptimierung, 336
- PCA, 146, 171, 396
 - Kernel, 400
- Pearson Korrelation, 381
- Permutationstest, 384
- Plackett-Burman, 33, 37, 162
- Plot, 93, 94
- Polygon Verfahren, 275
- Polynome, 51, 234
- Power, 106, 108, 110, 127, 128, 130, 132
- Predictive Error Sum of Squares, 418
- Predictive Sum of Squares, 418
- Principal Component Analyses, 396
- Principal Component Analysis, 146, 164, 171
- PROB, 351
- Produzentenrisiko, 103
- Pruning, 251
- Pseudo-Zufallszahlen, 199
- Pseudofaktoren, 57
- PURE RANDOM, 350

- quadratische Straffunktion (SVR), 270
- quadratisches Beschreibungsmodell, 39
- Qualität
 - Entropie, 193
 - MaxiMin, 191
 - MiniMax, 191
 - Qualität, Gleichverteilung, 194
 - Qualität, Metamodellen, 314
 - Qualität, Uniformity, 194
 - Qualität, Vergleich, 197
 - Qualitätskriterium
 - Testfelder, 190
 - Qualitätsmerkmal, 4, 139, 162, 166
 - Quantilschritte, 69
 - quartic, 50
 - Quasi-Monte-Carlo, 199
 - Quasi-Monte-Carlo, Hybrid, 204
- R-Indikator, 368
- Radial Basis Funktion, 284
- Rampenfunktionen, 140
- Randbedingungen, 7
- Randbedingungen, Optimierung, 332
- RANDOM, 350
- Randomisierung, 87, 90, 94, 96, 99, 100
- Randomized orthogonal arrays Latin
 - Hypercubes, 206
- Randverteilung, 291
- Rang
 - NSGA-II, 359
 - Optimierung, 359
- Rangkorrelation, 385
 - Kendalls τ , 388
 - Spearman, 387
- Rasensprenger, 10
- Rauschen, 96, 111, 113, 117, 118, 120, 128, 129, 133
- Reduktionsstufe, 32
- Reelle Faktoren
 - Kreuzung, 355
 - Mutation, 357
- Regression
 - Adaptive Basis-Funktionen, 241
 - lineare, 232
 - robuste, 239
- Regression sum of squares, 417
- Regressionsanalyse, 56
- Regressionsbaum, 248
 - cross validation, 252
 - Geschnitten, 251
 - Impurity, 250
 - Kreuzvalidierung, 252
 - Pruned, 251
 - Verunreinigung, 250
- Regressionsverfahren, 24
- reguläre Felder, 32
- Replikation, 90
- Reproduzierbarkeit, 98
- Residualplot, 78
- Residuenanalyse, 87, 123, 133, 136
- resolution, 32
- Ridge Regression, 273
- Robuste Regression, 239
- robustes Design, 139
- robustness, 139
- Rothamsted, 90
- ROUNDS, 351
- Rundungsfehler, 68
- SBS, 236
- Scheinkorrelation, 383
- Schulphysik, 7
- Schutzplanke, 160
- Scree-Plot, 146
- screening designs, 28
- Screening Versuchspläne, 28, 28, 160
- Sensitivitätsanalyse, 415
 - eFAST, 427
 - extended Fourier Amplitude Sensitivity Test, 427
 - Faktor Screening, 415
 - FAST, 425
 - Fourier Amplitude Sensitivity Test, 425
 - globale, 415
 - Importance Measure, 420
 - Kendalls τ , 388
 - Korrelationsverhältnis, 419
 - lineare Modelle, 416
 - lokale, 415
 - nichtlineare Modelle, 419
 - normierte Regressionskoeffizienten, 416
 - Partialsumme der Quadrate, 417
 - Partielle Korrelationsfaktoren, 418
 - Partieller determinationskoeffizient, 418
 - Predictive Error Sum of Squares, 418
 - Sobol's Kennzahl, 422
 - Spearman, 387
- Sequential Backward Selection, 236
- Sequential Forward Selection, 235
- Sequentielle Gleitende Vorwärts Selektion, 236
- Sequentielle Rückwärts Selektion, 236
- Sequentielle Vorwärts Selektion, 235
- Sequenz
 - Halton, 199
 - Hammersley, 201
 - Kronecker, 202
 - Sobol, 203

- Sequenz, (t,s), 204
Sequenz, Faure, 202
Sequenz, Van-der-Corput, 199
Sequenz, Verallgemeinerte Faure, 202
Sequential Floating Forward Selection, 236
SFFS, 236
SFS, 235
Shewart, W.A., 92
Sicherheitsbedürfnis, 109
signal to noise ratio, 152
Signalgrößen, 148, 165
signifikant, 87, 93, 98, 100–102, 106, 107, 110, 111, 113, 114, 117, 119, 120, 122, 123, 126–129, 132, 134–136
Signifikanz, 115–117, 119–121, 128, 129, 133, 136
Permutationstest, 384
Signifikanz einer Korrelation, 383
Signifikanztest, 100, 114
Simplex-Centroid-Design, 53
Simplex-Lattice-Design, 52
Simplexgitterplan, 52
simulated annealing, 216
Simulationsmodelle, 179
Simulierte Abkühlen, 216
Six Sigma, 91, 92
skalare Bewertungsfunktion, 140
Sobol's Kennzahl, 422
Sobol-Sequenz, 203
Space-Filling-Design, 45
Spearman, 387
Splines, 254
SSE, 237
SSR, 237
SST, 237
Störgrößen, 148, 165
Standard Normal Distribution, 219
Standardabweichung, 125, 128–132
Statistik, 87, 90–92
Steuergrößen, 148
Stichprobenumfang, 96, 97, 108
stochastische Evolutionsverfahren, 216
Strategie
komplexe Systeme, 441
Streuung, 68, 88, 98, 99, 110, 116, 126, 133
Stufen, 6, 28, 39, 154
Stufenabstand, 6, 40
Stufenmittelwerte, 14
Suchalgorithmus, 215
Sum Of Squares Between Groups, SSB, 111, 113
Sum Of Squares Within Groups, SSW, 111
Support Vector Machines, 258
Klassifikation, 258
nicht linear, 264
Regression, 266
Support Vector Regression, 266
v-SVR-Verfahren, 269
Least Square, 272
lineare Straffunktion, 266
quadratische Straffunktion, 270
Support Vetor Machines
Kernel, 265
Surrogatemodell, 231
Symbiotic Organisms Search, 346
Symbiotische Organismen, 346
Symmetrisches Latin Hypercube, 207
Systemgrenzen, 3
t-Test, 91
t-Verteilung, 91
Taguchi, 34, 139
Teifaktorplan, 37
teilfaktorielle Versuchspläne, 28
Teilfaktorpläne, 28
Teilfaktorplan, 39
Teilvarianzen, 73
Test, statistischer, 102, 134
Testen, statistisches, 87, 90, 93, 101, 110, 111, 117, 126, 127, 135
Testfelder, 198
Bereich entfernen, 220
Erweiterung, 222
Gleichverteilte, 209
Konstruktionsmethoden, 198
Latin Hypercube, 205
Latin Hypercube Design, 205
Latin Hypercube Sampling, 205
Optimierung, 214
Qualitätskriterium, 190
ungleichverteilt, 219
Uniform Design, 209
Testfelder, Fast Orthogonale Arrays, 204
Testfelder, Gemischte Orthogonale Arrays, 204
Testfelder, Mixed Orthogonal Arrays, 204
Testfelder, Nearly Orthogonal Arrays, 204
Testfelder, Orthogonal array-based Latin Hypercube, 206
Testfelder, Orthogonal Design Tables, 205
Testfelder, Orthogonale Arrays, 204
Testfelder, Orthogonale Design Tabellen, 205
Testfelder, Orthogonales Latin Hypercube, 207
Testfelder, Randomized orthogonal arrays
Latin Hypercubes, 206
Testfelder, Symmetrisches Latin Hypercubes, 207
Testfelder, Zentriertes Latin Hypercube, 205

- Threshold Accepting, 215
- Toleranzdesign, 139, 153
- Torus Algorithmus, 202
- Total sum of squares, 417
- Total Sum Of Squares, TTS, 111
- Totaler Sensitivitätsindex, 427
- Transferfunktion, 231
- Transferfunktionen, 180
- Transformationsfunktion G_j , 427
- Unabhängige Komponentenanalyse, 403
- Unabhängigkeit, 405
- Unabhängigkeitsanalyse, 403
- Uniform Design, 209
- Uniformity, 194
- Universal Kriging, 283
- Van-der-Corput-Sequenz, 199
- Variabilität, 89, 90, 93, 111, 113–115, 117, 121, 123, 126
- Variance Conditional Expectation, 419
- Varianz, 96, 111, 118, 119, 124, 133–136
- Varianz-Inflations-Faktor, 65
- Varianzanalyse, 73, 87, 90, 91, 100, 110, 111, 117, 119, 122, 124, 126–128, 133, 135
- Ventiltrieb, 165
- Verallgemeinerte Faure-Sequenz, 202
- verbessertes stochastisches Evolutionsverfahren, 216
- verdeckte Korrelation, 383
- Verfahrenstechnik, 1, 52
- vermengt, 29, 64
- Versuchsplan, 27, 99, 100, 114, 121, 132, 136
 - komplexe Zusammenhänge, 189
 - nichtlineare Zusammenhänge, 189
- Versuchsplan, randomisierter, 90, 94, 100
- Versuchsplanung, statistische, 87, 89, 92, 93, 101, 117
- Versuchsstreuung, 12, 68, 89, 96–99
- Versuchsumfang, 104, 108, 110, 127, 132
- Versuchsvorbereitung, 63
- Versuchswiederholung, 87, 98
- Versuchswiederholungen, 130
- Versuchszahl(en), 91, 101, 103, 106, 108, 128
- Verteilungsfunktion, 219
- Verteilungsfunktion, inverse, 219
- Verunreinigung
 - Regressionsbaum, 250
- VIF, 65
- Vollfaktorplan, 6, 88, 94, 99, 112, 121, 130, 167
- Vorhersagemodell, 87
- Wahrscheinlichkeit, 96, 101, 102, 106, 108, 110, 117, 119, 127, 128
- Wahrscheinlichkeitsdichte, 219
- Wahrscheinlichkeitsrechnung, 90, 96, 103
- Wahrscheinlichkeitsverteilung, 90
- Wechselwirkung, 2, 15, 100, 113–115, 123, 133, 135, 136, 151
 - abschwächend, 19
 - verstärkend, 19
- Wechselwirkungsdiagramm, 20
- Wechselwirkungseffekt, 17
- Wechselwirkungsspalten, 19
- Wiederholung (von Versuchen), 89, 90, 96, 98
- Wrap Around Discrepancy, 197
- Yates-Standard, 32
- Zentrierte Diskrepanz, 196
- Zentriertes Latin Hypercube, 205
- Zielgröße, übergeordnete, 333
- Zielgrößen
 - kombinieren, 333
- Zielgrößen, Optimierung, 325
- Zufallszahl, pseudo, 199
- zusammenfassende Ersatzgröße, 145
- Zweifachwechselwirkungen, 29
- zyklische Vertauschung, 33