

Statistische und numerische Methoden der Datenanalyse

Von Prof. Dr. rer. nat. Volker Blobel
und Prof. Dr. rer. nat. Erich Lohrmann
DESY und Universität Hamburg



B. G. Teubner Stuttgart · Leipzig 1998

Prof. Dr. rer. nat. Volker Blobel

Studium in Braunschweig und Hamburg mit Promotion 1968, wiss. Tätigkeit an der Universität Hamburg und am Deutschen Elektronen-Synchrotron DESY, seit 1977 Professor an der Universität Hamburg, zwischen 1979 und 1996 verschiedene Forschungsaufenthalte am CERN (European Laboratory for Particle Physics) in Genf.
Arbeitsgebiet: experimentelle Hochenergiephysik.

Prof. Dr. rer. nat. Erich Lohrmann

Studium in Stuttgart, Promotion 1956 mit einer Arbeit über Prozesse sehr hoher Energie in der kosmischen Strahlung, Arbeiten über kosmische Strahlung und Hochenergiephysik an den Universitäten Bern, Frankfurt und Chicago, seit 1961 am Deutschen Elektronen-Synchrotron, seit 1976 Professor an der Universität Hamburg, Mitglied des CERN Direktoriums 1976–78, Zuständigkeit Forschung und Datenverarbeitung.
Gegenwärtiges Arbeitsgebiet: experimentelle Hochenergiephysik.

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Blobel, Volker:

Statistische und numerische Methoden der Datenanalyse / von Volker

Blobel ; Erich Lohrmann. – Stuttgart ; Leipzig : Teubner, 1998

(Teubner-Studienbücher : Physik)

ISBN 978-3-519-03243-4 ISBN 978-3-663-05690-4 (eBook)

DOI 10.1007/978-3-663-05690-4

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Das gilt besonders für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

© 1998 B. G. Teubner Stuttgart · Leipzig

Vorwort

Der Umfang des Datenmaterials in Wissenschaft und Technik nimmt immer schneller zu; seine Auswertung und Beurteilung erweisen sich zunehmend als die eigentliche Schwierigkeit bei vielen wichtigen Problemen. Dem steht auf der anderen Seite ein seit Jahren ungebrochenes Anwachsen der Rechnerleistung und die zunehmende Verfügbarkeit mächtiger Algorithmen gegenüber, aber es ist oft nicht einfach, von diesen Hilfsmitteln den richtigen und professionellen Gebrauch zu machen. Dieses Buch, entstanden aus der Praxis der Verarbeitung großer Datenmengen, will eine Einführung und Hilfe auf diesem Gebiet geben.

Viele der Probleme sind statistischer Natur. Hier ist es sprichwörtlich leicht, Fehler zu machen. Deshalb sind der Erklärung und der kritischen Durchleuchtung statistischer Zusammenhänge auch im Hinblick auf die Praxis ein angemessener Raum gewidmet und ebenso den Monte Carlo-Methoden, welche heute einen verhältnismäßig einfachen Zugang zu vielen statistischen Problemen bieten.

Werkzeuge für die Organisation und Strukturierung großer Datenmengen bilden ein weiteres wichtiges Thema. Dazu gehören auch effiziente Verfahren zum Sortieren und Suchen, welche oft Teil größerer Algorithmen sind.

Die Verarbeitung großer Datenmengen hat oft die Extraktion verhältnismäßig weniger Parameter zum Ziel. Hier sind Verfahren wie die Methoden der kleinsten Quadrate und der Maximum-Likelihood wichtig, in Verbindung mit effektiven Optimierungsalgorithmen. Ein weiteres Problem, welches oft unterschätzt wird, ist die Rekonstruktion ursprünglicher Verteilungen aus fehlerbehafteten Messungen durch Entfaltung.

Mit der Verfügbarkeit mathematischer Bibliotheken für Matrixoperationen können viele Probleme elegant in Matrixschreibweise formuliert und auf Rechnern gelöst werden. Deswegen werden auch die einfachen Grundlagen der Matrixalgebra behandelt.

Das Buch führt in diese Dinge ein, es ist gedacht für Naturwissenschaftler und Ingenieure, die mit Problemen der Datenverarbeitung befaßt sind. Neben Algorithmen dienen zahlreiche einfache Beispiele der Verdeutlichung.

Um der Unsitte entgegenzuwirken, mehr oder weniger gut verstandene Rezepte schwarze-schachtelartig anzuwenden, wird auf Erklärungen und Beweise innerhalb eines vernünftigen Rahmens großer Wert gelegt, ebenso auf eine kurze Einführung des Informationsbegriffs.

Programmcodes von im Text erwähnten Algorithmen sind, sofern sie einen solchen Umfang haben, daß Kopieren von Hand nicht ratsam ist, im Internet zugänglich gemacht (<http://www.desy.de/~blobel>). Sie sind von den Autoren benutzt und überprüft worden, trotzdem kann für ihre Korrektheit und für die Abwesenheit von Nebeneffekten keine Garantie übernommen werden. Kurze Programme sind meist in Fortran, in einigen Fällen in C formuliert. Sie sollen

zur Präzisierung der Algorithmen dienen. Sie sind kurz genug, so daß sie jeder in seine Lieblingssprache umschreiben kann.

Vielen Kollegen, vor allem Dr. F. Gutbrod und Dr. D. Haidt sind wir dankbar für wertvolle Diskussionen, Informationen und Anregungen. Frau A. Blobel und Frau U. Rehder danken wir für ihre Hilfe bei der Erstellung des Manuskripts, Herrn Dr. P. Spuhler vom Teubner-Verlag danken wir für die gute Zusammenarbeit.

Hamburg, im Juli 1998

V. Blobel und E. Lohrmann

Inhaltsverzeichnis

1	Datenbehandlung und Programmierung	11
1.1	Information	11
1.2	Codierung	13
1.3	Informationsübertragung	17
1.4	Analogsignale – Abtasttheorem	18
1.5	Repräsentation numerischer Daten	20
1.6	Programmorganisation	23
1.7	Programmprüfung	24
2	Algorithmen und Datenstrukturen	28
2.1	Algorithmen und ihre Analyse	28
2.2	Datenstrukturen	31
2.2.1	Datenfelder	31
2.2.2	Abstrakte Datentypen	32
2.3	Sortieren	36
2.4	Suchen	49
2.5	Weitere Algorithmen	52
3	Methoden der linearen Algebra	57
3.1	Vektoren und Matrizen	57
3.2	Symmetrische Matrizen	62
3.3	Vertauschungs-Algorithmus	63
3.4	Dreiecksmatrizen	68
3.5	Allgemeine LR-Zerlegung	71
3.6	Cholesky-Zerlegung	75
3.7	Inversion durch Partitionierung	78
3.8	Diagonalisierung symmetrischer Matrizen	81
3.9	Singulärwert-Zerlegung	89
4	Statistik	93
4.1	Einleitung	93
4.2	Wahrscheinlichkeit	93
4.3	Verteilungen	98
4.3.1	Grundbegriffe	98

4.3.2	Erwartungswerte und Momente	100
4.3.3	Charakterisierung von Wahrscheinlichkeitsdichten	102
4.4	Spezielle diskrete Verteilungen	103
4.4.1	Kombinatorik	103
4.4.2	Binomialverteilung	105
4.4.3	Poisson-Verteilung	107
4.5	Spezielle Wahrscheinlichkeitsdichten	111
4.5.1	Gleichverteilung	111
4.5.2	Normalverteilung	111
4.5.3	Gammaverteilung	116
4.5.4	Charakteristische Funktionen	118
4.5.5	χ^2 -Verteilung	119
4.5.6	Cauchy-Verteilung	123
4.5.7	t -Verteilung	124
4.5.8	F -Verteilung	125
4.6	Theoreme	128
4.6.1	Die Tschebyscheff-Ungleichung	128
4.6.2	Das Gesetz der großen Zahl	129
4.6.3	Der Zentrale Grenzwertsatz	129
4.7	Stichproben	131
4.7.1	Auswertung von Stichproben.	131
4.7.2	Gewichte	134
4.7.3	Numerische Berechnung von Stichprobenmittel und -Varianz	135
4.8	Mehrdimensionale Verteilungen	136
4.8.1	Zufallsvariable in zwei Dimensionen	136
4.8.2	Zweidimensionale Gauß-Verteilung	138
4.8.3	Mehrdimensionale Wahrscheinlichkeitsdichten	142
4.8.4	Mehrdimensionale Gauß-Verteilung	143
4.9	Transformation von Wahrscheinlichkeitsdichten	145
4.9.1	Transformation einer einzelnen Variablen	145
4.9.2	Transformation mehrerer Variabler und Fehlerfortpflanzung	146
4.9.3	Beispiele zur Fehlerfortpflanzung	148
4.10	Faltung	151

5	Monte Carlo-Methoden	154
5.1	Einführung	154
5.2	Zufallszahlengeneratoren	155
5.3	Zufallszahlen für beliebige Verteilungen	162
5.4	Zufallszahlen für spezielle Verteilungen	165
5.4.1	Zufallswinkel und -vektoren	165
5.4.2	Normalverteilung	166
5.4.3	Poisson-Verteilung	168
5.4.4	χ^2 -Verteilung	169
5.4.5	Cauchy-Verteilung	170
5.4.6	Binomialverteilung	170
5.5	Monte Carlo-Integration	170
5.5.1	Integration in einer Dimension	170
5.5.2	Varianzreduzierende Methoden	171
5.5.3	Vergleich mit numerischer Integration	175
5.5.4	Quasi-Zufallszahlen	176
6	Schätzung von Parametern	178
6.1	Problemstellung und Kriterien	178
6.2	Robuste Schätzung von Mittelwerten	179
6.3	Die Maximum-Likelihood-Methode	183
6.3.1	Prinzip der Maximum-Likelihood	183
6.3.2	Methode der kleinsten Quadrate	188
6.4	Fehler der Parameter	189
6.5	Anwendungen der Maximum-Likelihood-Methode	191
6.5.1	Beispiele	191
6.5.2	Histogramme	194
6.5.3	Erweiterte Maximum-Likelihood-Methode	197
6.6	Eigenschaften der Maximum-Likelihood-Methode	198
6.6.1	Konsistenz	198
6.6.2	Erwartungstreue	199
6.6.3	Asymptotische Annäherung an die Gauß-Funktion	199
6.7	Konfidenzgrenzen	201
6.8	Bayes'sche Statistik	204
6.9	Systematische Fehler	210

7 Methode der kleinsten Quadrate	212
7.1 Einleitung	212
7.2 Lineare kleinste Quadrate	214
7.2.1 Bestimmung von Parameterwerten	214
7.2.2 Normalgleichungen im Fall gleicher Fehler	215
7.2.3 Matrixschreibweise	215
7.2.4 Kovarianzmatrix der Parameter	217
7.2.5 Quadratsumme der Residuen	218
7.2.6 Korrektur der Datenwerte	218
7.3 Lösungseigenschaften	219
7.3.1 Erwartungstreue	219
7.3.2 Das Gauß-Markoff-Theorem	220
7.3.3 Erwartungswert der Summe der Residuenquadrate	220
7.4 Der Fall unterschiedlicher Fehler	222
7.4.1 Die Gewichtsmatrix	222
7.4.2 Lösung im Fall der allgemeinen Kovarianzmatrix	223
7.5 Kleinste Quadrate in der Praxis	224
7.5.1 Normalgleichungen für unkorrelierte Daten	224
7.5.2 Geradenanpassung	224
7.5.3 Reduktion der Matrixgröße	228
7.6 Nichtlineare kleinste Quadrate	230
7.6.1 Linearisierung	230
7.6.2 Konvergenz	231
7.7 Kleinste Quadrate mit Nebenbedingungen	232
7.7.1 Einleitung	232
7.7.2 Nebenbedingungen ohne ungemessene Parameter	234
7.7.3 Der allgemeine Fall	236
8 Optimierung	239
8.1 Einleitung	239
8.1.1 Optimierungsprobleme und Minimierung	239
8.1.2 Minimierung ohne Nebenbedingungen	240
8.1.3 Allgemeine Bemerkungen zu Methoden der Minimierung	244
8.2 Eindimensionale Minimierung	245
8.2.1 Suchmethoden	245
8.2.2 Die Newton-Methode	248

8.2.3	Kombinierte Methoden	251
8.3	Suchmethoden für den Fall mehrerer Variabler	252
8.3.1	Gitter- und Monte Carlo-Suchmethoden	253
8.3.2	Einfache Parametervariation	254
8.3.3	Die Simplex-Methode	255
8.4	Minimierung ohne Nebenbedingungen	259
8.4.1	Die Newton-Methode als Standardverfahren	260
8.4.2	Modifizierte Newton-Methoden	268
8.4.3	Bestimmung der Hesse-Matrix	269
8.4.4	Numerische Differentiation	271
8.5	Gleichungen als Nebenbedingungen	275
8.5.1	Lineare Nebenbedingungen	276
8.5.2	Nichtlineare Nebenbedingungen	279
8.6	Ungleichungen als Nebenbedingungen	280
8.6.1	Optimierungsbedingungen	281
8.6.2	Schranken für die Variablen	283
9	Prüfung von Hypothesen	285
9.1	Prüfung einer einzelnen Hypothese	285
9.1.1	Allgemeine Begriffe	285
9.1.2	χ^2 -Test	286
9.1.3	Studentscher t -Test	290
9.1.4	F -Test	291
9.1.5	Kolmogorov-Smirnov-Test	291
9.2	Entscheidung zwischen Hypothesen	292
9.2.1	Allgemeine Begriffe	292
9.2.2	Strategien zur Wahl der Verwurfsregion	294
9.2.3	Minimax-Kriterium	296
9.3	Allgemeine Klassifizierungsmethoden	298
9.3.1	Fishersche Diskriminantenmethode	299
9.3.2	Neuronale Netze	300
10	Parametrisierung von Daten	305
10.1	Einleitung	305
10.2	Spline-Funktionen	308
10.3	Orthogonale Polynome	320
10.4	Fourierreihen	328

11 Entfaltung	330
11.1 Problemstellung	330
11.2 Akzeptanzkorrekturen	332
11.3 Entfaltung in zwei Intervallen	333
11.4 Entfaltung periodischer Verteilungen	334
11.5 Diskretisierung	337
11.6 Entfaltung ohne Regularisierung	338
11.7 Entfaltung mit Regularisierung	341
Literaturverzeichnis	349
Index	353