Entwicklung eines Prognosekonzepts mit bedingten multivariaten Wahrscheinlichkeitsverteilungen

Anwendungen aus der Automobilindustrie

Zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften

(Dr. rer. pol.)

von der Fakultät für
Wirtschaftswissenschaften
der Universität Fridericiana zu Karlsruhe

genehmigte

DISSERTATION

von

Diplom-Wirtschaftsmathematiker
Eric Alexander Stützle

Tag der mündlichen Prüfung: 18. November 2003
Referent: Prof. Dr. Georg Bol
Korreferent: Prof. Dr. Karl-Heinz Waldmann
Karlsruhe 2003

Vorwort

Seit langer Zeit sind präzise, aussagekräftige und interpretierbare Vorhersagen von Ereignissen von großem Interesse. Die vorliegende Arbeit beinhaltet die Entwicklung, Validierung und Anwendung einer Prognosemethodik, die motiviert durch praktische Problemstellungen, auf unterschiedlichen Wissensgebieten basiert.

Es ist ein Prognosesystem entwickelt, das die Vorhersage von multivariaten Wahrscheinlichkeitsverteilungen abhängig von erklärenden Variablen ermöglicht. Generelle Verteilungsklassen, wie etwa hyperbolische oder stabile Verteilungen können in diesem Konzept integriert werden. Die Schätzung der bedingten Wahrscheinlichkeitsverteilung basiert auf der numerischen Minimierung der Cross-Entropie. Zur Optimierung wird der globale Suchalgorithmus „Multi-Level Single-Linkage„ verwendet. Nichtlineare Abhängigkeiten der bedingten Verteilungsparameter können durch generelle funktionale Approximatoren, wie das Multi-Layer Perzeptron modelliert werden.

Oftmals ist die Prognose neuer Objekte problematisch. Dieses Konzept der bedingten Verteilungen erlaubt durch die Konstruktion eines allgemeinen attributbasierten Prognosesystems die Berechnung zukünftiger Zielgrößen mit neuen Attributkombinationen.

Zur Entscheidungsunterstützung bei praktischen Problemstellungen ist die Kenntnis der bedingten Wahrscheinlichkeitsverteilung von Vorteil. Sie ermöglicht eine quantitative Aussage über die Zuverlässigkeit der Prognosewerte. Ebenso dient sie der optimalen Entscheidung bei Existenz asymmetrischer Kostenverhältnisse. Dies ist am Fallbeispiel des zu prognostizierenden Ersatzteilebedarfs illustriert. Anhand empirischer Untersuchungen zeigt sich das Konzept der bedingten Verteilungen dominant gegenüber ausgewählten statistischen Verfahren. Die Verringerung des Prognosefehlers auf Grund von nicht-normalen Verteilungsklassen ist am Beispiel der Absatzprognose von Nutzfahrzeugen präsentiert.

Diese Arbeit entstand während meiner Tätigkeit als Doktorand am Forschungszentrum der DaimlerChrysler AG in Ulm in Zusammenarbeit mit dem Institut für Statistik und Mathematische Wirtschaftstheorie der Universität

Karlsruhe (TH).

Ganz herzlich bedanke ich mich bei meinem Doktorvater Herrn Prof. Dr. Georg Bol für seine Unterstützung und konstruktive Kritik während der gesamten Zeit. Herrn Prof. Dr. Karl-Heinz Waldmann gilt mein Dank für die Übernahme des Korreferats und die unkomplizierte Zusammenarbeit. Mein besonderer Dank gilt auch Herrn Prof. Dr. Gholamreza Nakhaeizadeh, der mir die Arbeit am Forschungszentrum Ulm ermöglicht hat.

Die fachliche Betreuung und Unterstützung am Forschungszentrum in Ulm durch Herrn Dr. Tomas Hrycej war für meine Arbeit von unschätzbarem Wert. Die zahlreichen Diskussionen und die intensive Zusammenarbeit sind Bereicherungen sowohl in fachlicher als auch in menschlicher Hinsicht. Des Weiteren danke ich Rainer Schu und Pia Leister für die kritischen Korrekturen meiner Arbeit.

Ich möchte diese Möglichkeit nutzen, meinen Eltern zu danken, da sie mir jederzeit zur Seite standen und mich unterstützt haben. Ein spezielles Dankeschön gilt auch meinen einzigartigen Freunden, die mein Leben in der Platzgasse 19 und meine Persönlichkeit bereichert haben.

Nicht zuletzt danke ich ganz besonders meiner Freundin Esther für ihr Verständnis und ihre Unterstützung während dieser arbeitsintensiven Zeit.

Meinem Opa

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

Seit langer Zeit sind möglichst präzise, aussagekräftige und interpretierbare Vorhersagen bzw. Prognosen[1] von Ereignissen von großem Interesse.

Kirchgraber und Ruf bemerken: „*Prognosen sind häufig falsch. Wer sich darüber mokiert, unterschätzt den Anspruch, den er stellt.*“[2]

1.1 Vorbemerkung

Das tägliche Leben beinhaltet unzählige Quellen der Unsicherheit. Im Unterbewusstsein trifft man sehr häufig Entscheidungen, die in der Zukunft liegen und gerade aus diesem Grund mit mehr oder weniger viel Unsicherheit behaftet sind. Man versucht die Unsicherheit durch Erfahrungswerte oder entwickelte Strategien abzuwägen und abzuschätzen. Bei Betrachtung dieser Heuristiken oder Strategien mit kritischer Skepsis gewinnt man die Einsicht, dass sie nicht immer die Güte und Aussagekraft besitzen, die von ihnen erwartet wird.

Natürlich ist Unsicherheit nicht ausschließlich auf das private Leben begrenzt. Auch öffentliche Institutionen oder wirtschaftliche Unternehmen sehen sich ständig gewissen Unsicherheiten und Risiken ausgesetzt. Niemand

[1] Das Wort *Prognose* stammt von dem griechischen Begriff „*prognosis*“ ab, was übersetzt das Vorherwissen bedeutet.

[2] Siehe (Kirchgraber and Ruf, 1997).

ist bislang in der Lage, zukünftige Geschehnisse dieser Welt exakt und mit Sicherheit vorauszusagen. Durch Planungen des Menschen in die Zukunft ist das Thema der Prognostik, womit sich diese Arbeit befasst, aufgegriffen.

Die Prognostik ist im engeren Sinne, nicht nur wegen des breiten theoretischen Hintergrundes der Schätztheorie, der Regressions- und Zeitreihenanalyse, sondern zusätzlich aufgrund der vielzähligen Anwendungsgebiete im wissenschaftlichen Feld der Statistik angesiedelt.

Die Existenz der Computertechnik und damit der Neuroinformatik nahm unter anderem entscheidenden Einfluss auf die Entwicklung von Prognosemethoden. Als Beispiel wäre das wissenschaftliche Gebiet der neuronalen Netze zu nennen, die durch ihre potentielle Generalisierungsfähigkeit eine Alternative zur nichtlinearen Regression bieten. Armstrong bezeichnet die Prognostik am Ende der 90er Jahre als eine eigene Wissenschaft.[3]

Diese Arbeit ist aus vielen Gründen sehr eng mit einigen Methoden aus der Statistik verwandt. Jedoch ist das im Folgenden vorgestellte Prognosekonzept nicht ganzheitlich in das weite Gebiet der Statistik einzuordnen. Der Charakter eines quantitativen Prognosemodells ist diesem Konzept und den statistischen Methoden jedoch uneingeschränkt gemein. Obwohl Prognosekonzepte, wie die Regressions- oder Zeitreihenanalyse, auf einem tiefen und ausgeprägten theoretischen Fundament stehen, stoßen sie häufig an Grenzen. Ebenso wie die klassischen statistischen Methoden ist auch das in dieser Abhandlung dargestellte Konzept stark durch praktische Anwendungen motiviert - eine weitere sinnvolle Gemeinsamkeit.

Durch die Motivation aus der realen Praxis ist die im Folgenden entwickelte, validierte und auf verschiedene Probleme angewendete Methodik unter den Aspekten der Bedingtheit, der flexiblen Verteilungsklassen, der Aussagekraft, der Entscheidungstheorie und der universellen Anwendungsfähigkeit konzipiert.

Abbildung 1.1 zeigt die Struktur und den Inhalt dieser Arbeit und skizziert die Vereinigung der Wissensgebiete Neuroinformatik, numerische Mathematik und Stochastik zur zielorientierten Anwendung in der Praxis. Das

[3]Siehe (Armstrong, 1989).

Abbildung 1.1: Struktur und Inhalt der Arbeit

im Anschluss vorgestellte Prognosekonzept stellt unter anderem aufgrund dieser Eigenschaft der „Methodenvereinigung“ ein Konzept dar, das klassische Verfahrensweisen umfasst.

Die Kernthemen dieser Abhandlung sind die Entwicklung, Validierung und Anwendung eines Prognosekonzepts, das den Zusammenhang zwischen bekannten exogenen Einflüssen und der zugrunde gelegten Wahrscheinlichkeitsverteilung der Zielgrößen modelliert. Es wird nicht auf datentechnische Gesichtspunkte, wie etwa die Datenvorverarbeitung eingegangen. Folglich sind klassische Themen der Zeitreihenanalyse, wie die Transformation auf Stationarität, nicht behandelt. Diese datenbezogenen Verarbeitungsschritte sind bei allen Modellierungsmethoden nahezu identisch.[4]

1.2 Ziel der Arbeit

Im folgenden Abschnitt sind die wesentlichen Aussagen und Ziele der Arbeit formuliert, wobei besonderes Augenmerk auf dem wissenschaftlichen Mehrwert der Abhandlung liegt.

Das primäre Ziel liegt in der Entwicklung eines Prognosekonzepts, das - motiviert durch Anforderungen aus praktischen Problemstellungen - in der Lage ist, präzise, aussagekräftige und interpretierbare Vorhersagen zu berechnen, und dadurch den Entscheidungsträgern eine optimale Hilfe bereitstellt. Was bei dieser abstrakten Formulierung die Adjektive präzise, aussagekräftig und interpretierbar bedeuten und wie diese Eigenschaften optimal ausgenutzt werden können, ist ebenfalls im Folgenden zu klären.

Die Redensart „*Der Weg ist das Ziel*“ trifft in gewisser Weise auf die Entstehung und Vorgehensweise dieser Arbeit zu. Zu Beginn werden Ziele aus praktischen Motivationsgründen definiert, die Mittel zur Zielerreichung vorgestellt und schließlich der Weg aufgezeigt, auf dem die Vielfalt der Hilfs-

[4] Im bekannten **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (CRISP-DM), dargestellt in (Wirth and Hipp, 2000) und (Chapman et al., 2000), ist der Konzeptteil dieser Arbeit eindeutig im Modellierungsschritt angesiedelt. Dennoch sind die übrigen Elemente für die Umsetzung praktischer Problemstellungen, wie etwa in den Kapitel 9 und 10, unumgänglich.

mittel und die theoretischen Eigenschaften zu einer Einheit verschmelzen. Am Ende des Weges steht die Anforderung, formulierte Motivationsaspekte umzusetzen und erfolgreich in praktischen Anwendungen zu agieren und zu überzeugen.

Neben der reinen Entwicklung einer Prognosemethodik ist es notwendig, die Analogien zu verschiedenen klassischen Prognosemethoden herauszuarbeiten und darzustellen, um den Bezug zur Welt der bekannten Prognosetechniken herzustellen. Es sind hierzu Methoden aus der Literatur in ausreichender Kürze darzustellen und in das hier entwickelte Konzept einzubetten. Dadurch sind die Erweiterungen, erste Vorteile und die wesentlichen Verallgemeinerungen des dargestellten Konzepts klar abgegrenzt und theoretisch skizziert. Um die theoretischen Aussagen beispielhaft zu unterstreichen, sind Prognoseergebnisse auf realen Datensätzen zu berechnen und zu vergleichen.

Der wissenschaftliche Mehrwert dieser Abhandlung ist erreicht, wenn es gelingt, die aufgeführten Anforderungen zu erfüllen.

> Die Konzeption eines Prognosesystems, das in der Lage ist, den funktionalen Zusammenhang zwischen exogenen Einflüssen und Verteilungsparametern unter den folgenden Voraussetzungen zu identifizieren.
>
> Die zugrunde liegende Verteilungsklasse ist nicht zwingend die Normalverteilung, sondern besitzt flexiblere Eigenschaften.
>
> Das Prognosesystem nutzt durch einen regressionsähnlichen Ansatz die zur Verfügung stehenden Informationen als exogene Variablen so gut wie möglich aus und operiert daher nicht ausschließlich auf einzelnen Zeitreihen.
>
> Die zu entwickelnde Methodik umfasst bekannte Techniken aus der Statistik.
>
> Ferner besteht die Prognosetechnik die Prüfung auf künstlich generierten Datensätzen, indem sie die erwünschten Ergebnisse produziert.
>
> Schließlich runden exzellente Prognosewerte der Methodik auf einer realen Datengrundlage die erfüllten Anforderungen ab.

1.3 Inhalt der Arbeit

Im Folgenden wird eine kurze inhaltliche Vorschau auf die vorliegende Abhandlung gegeben. Die Kapitel und Abschnitte sind knapp skizziert und ein logischer Zusammenhang wird aufgezeigt.

Zu Beginn stellt ein zum Thema „*Prognose*“ hinführendes Kapitel das in dieser Arbeit zugrunde gelegte Verständis von Prognostik dar. Dies geschieht durch skizzierte Schwächen klassischer Prognosemethoden, die es aufgrund der praktischen Relevanz zu beheben gilt. Es sind unterschiedliche Motivationsgründe aufgezeigt, die sich schießlich als die Anforderungen des zu entwickelnden Prognosekonzepts entpuppen. Durch die Einführung von Bezeichnungen und die Festlegung von Vereinbarungen wird formal das Prognoseumfeld notiert, auf welchem die Abhandlung basiert.

Der erste Hauptteil dieser Arbeit ist der Entwicklung, Charakterisierung und Abgrenzung des Prognosemodells gewidmet, das in einem zweiten Schritt anhand von künstlich generierten Daten validiert wird. Abschließend werden die Vorteile dieses Prognosekonzepts mittels realer Anwendungen aus der Automobilindustrie unter Beweis gestellt.

Der konzeptionelle Teil der Arbeit stellt die Mittel zur Verwirklichung der definierten Ziele bereit und fügt diese zu einer Einheit zusammen. Dieser Teil der Abhandlung ist in folgende Kapitel untergliedert:

Kapitel 3 beschreibt die allgemeine Basis des Prognosekonzepts. Die Generierung von Wahrscheinlichkeitsverteilungen einer bestimmen Verteilungsfamilie bildet den Schwerpunkt dieses Abschnitts. Es wird weiterhin ein erster Ausblick auf die Vielzahl von flexiblen Verteilungsklassen gegeben, die in einer so konzipierten Methodik Verwendung finden können.

Da es gilt, funktionale Zusammenhänge zwischen den Verteilungsparametern und exogenen Einflussgrößen zu identifizieren, stellt Kapitel 4 mögliche Approximatoren dar, die entweder über eine lineare Abbildung oder ein neuronales Netz realisiert werden können. Hierbei sind die neuronalen Netze, speziell das Multi-Layer Perzeptron als universeller Approximator, etwas ausführlicher beschrieben. Des Weiteren verlangt das Prognosekonzept in dieser Phase nach einer „*technischen Transformation*“ der reellen Ausgänge des

funktionalen Approximators auf die gesuchten Verteilungsparameter.

Das nachfolgende Kapitel 5 beinhaltet die Herleitung der Kostenfunktion. Ausgehend von dem Anpassungsmaß der Cross-Entropie kristallisiert sich die Zielfunktion als die bekannte negative Log-Likelihood Funktion heraus.

Nach der Herleitung einer im Allgemeinen nicht konvexen Kostenfunktion wird das erforderliche globale Optimierungsverfahren „*Multi-Level Single-Linkage*“ beschrieben, dessen Eigenschaften kurz diskutiert werden. Da die Umsetzung und Verwendung von globalen Optimierungsalgorithmen selten in der Praxis anzutreffen sind, finden einige technische Details Erwähnung.

Die bereits angedeutete Vielfalt möglicher anzunehmender Verteilungsfamilien ist in Kapitel 6 demonstriert. Hierdurch wird die Breite und Flexibilität dieses Ansatzes ein weiteres Mal unter Beweis gestellt. Die unterschiedlichen Verteilungsklassen sind kurz beschrieben und es wird in manchen Fällen auf gewisse Charakteristiken eingegangen. Hiermit ist das Prognosekonzept zu einer Einheit zusammengeführt und es verbleibt die Abgrenzung bzw. Inklusion bekannter klassischer Prognosemethoden aus der Literatur.

Dieser Thematik ist das letzte Kapitel 7 des Konzeptteils gewidmet. Es zeigt sich, dass einige klassische Prognosemethoden aus der Regressions- und Zeitreihenanalyse im hier vorgestellten Konzept beinhaltet. Dieser Abschnitt stellt sich jedoch nicht der Aufgabe, sämtliche theoretischen Aussagen dieser fundierten Methoden in das entwickelte Verteilungskonzept zu übertragen, vielmehr zeigt sich, dass die Modellierungsart klassischer Konzepte problemlos abgebildet werden kann.

Im zweiten Hauptteil dieser Abhandlung gilt es, das entwickelte Prognosekonzept zu validieren. Da dies anhand von realen Daten kaum vollständig möglich ist, wird es vor dem praktischen Einsatz auf künstlich generierten Datensätzen getestet und somit die Fähigkeit der Verteilungsidentifikation geprüft.

Im dritten Teil der Arbeit findet eine tatsächliche Berechnung von Prognosen in zwei realen Anwendungen statt. In beiden Beispielen wird die konzeptionelle und kalibrierende Vorgehensweise klar herausgestellt, die bei der Durchführung praktischer Aufgaben zu erbringen ist.

In Kapitel 9 wird der Bedarf von Ersatzteilen in den nächsten Jahren

vorhergesagt und die Stärken der Verteilungsprognose im Vergleich zu alternativen Verfahren, die klare Defizite aufweisen, dargestellt. Im zweiten Anwendungsfall wird ein Nachfragemodell für Nutzfahrzeugen erstellt. Hierbei liegt der Schwerpunkt auf der Annahme von flexiblen Wahrscheinlichkeitsverteilungen. Es werden die Ergebnisse Benchmark-Berechnungen gegenüber gestellt, um den Mehrwert zu verdeutlichen.

Der finale Teil gibt eine kompakte Zusammenfassung der erreichten Ziele der Arbeit und einen Ausblick auf potentielle Weiterentwicklungen, Verbesserungen und Anwendungen.

Kapitel 2

Problemstellung: Prognose

Im folgenden Kapitel wird basierend auf dem herkömmlichen Verständnis von Prognostik und den zugehörigen Techniken der dieser Arbeit zugrunde gelegte Blickwinkel für die Thematik definiert, motiviert und begründet. In den unterschiedlichen Motivationsgründen, die explizit aufgeführt sind, liegt das Fundament und die Inspiration dieser Arbeit.

2.1 Prognose in der klassischen Literatur

Prognosemethoden werden häufig in der Literatur in mindestens zwei Klassen unterteilt.[1]

Die *qualitativen* oder *subjektiven* Techniken, die hier nicht weiter betrachtet werden, sind intuitive Schätzungen zukünftiger Ereignisse, die auf Daten beruhen können, aber nicht müssen. In vielen Fällen kann diese Art der Prognose nicht nachvollzogen werden, da in den meisten Fällen unbekannt ist, wie die zur Verfügung stehenden Informationen in den angegebenen Prognosewert einfließen. Dennoch ist diese Vorgehensweise in manchen Situationen die einzig mögliche.[2] Ein paar wenige Beispiele für qualitative Prognosetechniken sind die Kombination von Expertenmeinungen[3], die DELPHI-

[1] Siehe u.a. (Abraham and Ledolter, 1983).

[2] Vgl. (Abraham and Ledolter, 1983).

[3] In diesem Fall berechnet sich der Prognosewert y zum Zeitpunkt t als:
$y_t = (\text{Prognose(Manager1)} + \text{Prognose(Manager2)} + \text{Prognose(Manager3)})\ /3$

Methode[4] oder historische Analysen[5].

Prognosemethoden wie die hier behandelten, die auf mathematischen und statistischen Sachverhalten basieren, werden *quantitativ* genannt. Falls das Prognosemodell identifiziert und berechnet ist, sind die Prognosewerte automatisch bestimmt und völlig reproduzierbar.

Des Weiteren können quantitative Techniken in *deterministische* und *probabilistische* Modelle unterteilt werden. Probabilistische Methoden werden auch oft *stochastisch* oder *statistisch* genannt.[6]

Im Falle von deterministischen Modellen ist der Zusammenhang zwischen der Zielvariablen Y und den erklärenden Variablen $X_1, \ldots, X_m$ funktional exakt gegeben. Dies lässt sich formal beschreiben durch

$$Y = f_\omega(X_1, \ldots, X_m),$$

wobei $\omega = (\omega_1 \ldots, \omega_P)$ den freien Parametervektor der Funktion f bezeichnet. Die Funktion f und die Koeffizienten ω sind im deterministischen Fall mit Sicherheit bekannt.[7]

In unsicheren Situationen sind die Zusammenhänge stochastischer Natur. Dies führt zu probabilistischen Modellen der Form

$$Y = f_\omega(X_1, \ldots, X_m) + \epsilon,$$

wobei ϵ ein zufälliges Rauschen darstellt, das in konkreten Modellen die Realisation einer Zufallsvariablen beschreibt, die einer gewissen Wahrscheinlichkeitsverteilung gehorcht. Die Zielvariable Y besitzt aufgrund des funktionalen Zusammenhangs zu ϵ dieselbe Wahrscheinlichkeitsverteilung.

Oftmals sind die funktionale Form von f und die Koeffizienten ω nicht

[4]Die DELPHI-Methode ist ein iterativer Prozess, bei welchem Experten Fragebögen beantworten.

[5]Als historische Analyse wird die Suche nach ähnlichen Ereignissen in der Vergangenheit bezeichnet, um dadurch auf die derzeitige Situation zu schließen, wie etwa die Einführung neuer Produkte, vgl. etwa (Abraham and Ledolter, 1983).

[6]Vgl. (Abraham and Ledolter, 1983).

[7]Als Beispiele hierfür seien die bekannten Gesetze der Physik genannt, vgl. (Abraham and Ledolter, 1983).

bekannt und müssen daher anhand von bekannten Informationen aus der Vergangenheit identifiziert werden.

Falls die Daten in zeitlicher Ordnung zur Verfügung stehen, spricht man von einer Zeitreihe. In der klassischen Literatur werden aufgrund einer solchen Datengrundlage häufig Techniken aus der so genannten Zeitreihenanalyse vorgeschlagen.

Bei einer weiteren Prognosemethodik aus der Statistik, der Regressionsanalyse, kommen typischerweise exogene Variablen in Betracht, um die Stochastik der Zielvariablen zu beschreiben. Die unabhängigen Variablen sind dann nicht ausschließlich Vergangenheitswerte der Zielgröße im Gegensatz zu den meisten Methoden der Zeitreihenanalyse. Sogar durch diese knappen Bemerkungen wird deutlich, dass Zeitreihenmodelle häufig als spezielle Regressionsmodelle interpretierbar sind.

Die oben angesprochenen Methoden der Zeitreihen- oder Regressionsanalyse zur Vorhersage von Ereignissen besitzen ein stark ausgeprägtes theoretisches Fundament, das in einer Vielzahl von Schriftstücken ausführlich niedergeschrieben ist.[8] Daher wird in dieser Arbeit nicht auf das theoretische Gedankengut der Methoden eingegangen. An entsprechenden Stellen, wie etwa in Kapitel 7 dieser Arbeit, sind einige Charakteristika der Methoden vorgestellt und Hinweise notiert, in denen eine Abhandlung der theoretischen Hintergründe präsentiert ist.

Der Fokus liegt im Folgenden auf der Entwicklung, Validierung und Anwendung eines Prognosekonzepts, das in vielen Punkten zwar Gemeinsamkeiten mit klassischen Techniken aufweist, jedoch einerseits in vielerlei Hinsicht etliche Verallgemeinerungen bietet und andererseits durch spezielle Eigenschaften entscheidende Vorteile besitzt.

Die meisten Autoren klassischer Literatur weisen auf den iterativen und interaktiven Prozess der Modellierung hin und erklären ihn als unumgänglich.[9] Ebenso wird die notwendige Testphase der identifizierten Modelle be-

[8] Vgl. etwa (Hamilton, 1994) und (Seber, 1977).

[9] Siehe etwa (Chatfield, 1996), (DeLurgio, 1998), (Levenbach and Cleary, 1981), (Levenbach and Cleary, 1984) oder (Jain, 1988).

tont. Diese technischen Details, die etwa auch im CRISP-DM Prozess[10] abgebildet sind, werden hier als bekannt vorausgesetzt. Chatfield teilt außerdem die verbreitete Meinung, dass komplexe Modelle dazu tendieren, zwar eine bessere Anpassung aufzuweisen, allerdings schlechtere ex-ante Vorhersagen zu berechnen als einfachere Modelle.[11] Um dieses Dilemma zu vermeiden das berechnete Prognosemodell in jedem Fall auf unbekannten Testdatensätzen evaluiert wird.

Eine ganze Menge von Autoren sind der Überzeugung, jeder Aufgabe ist eine optimale Prognosemethodik zuzuordnen, die eine beste Lösung erbringt. Diese Ansicht widerspricht dem im Folgenden vorgeschlagenen Konzept gänzlich. Es wird ein einziges Prognosekonzept entwickelt und dargestellt, das abhängig von der Informationsgrundlage spezifisch definiert und kalibriert werden kann und daher eine Vielzahl von klassischen Methoden umfasst.[12]

Klassische Literatur aus der Statistik und Ökonometrie setzt zwar oftmals unterschiedliche Schwerpunkte und stellt daher verschiedene Themen ausführlicher dar als andere, begrenzt jedoch dadurch die Prognostik auf die herkömmlichen Methoden aus der Zeitreihen- und Regressionsanalyse mit deren einschränkenden Annahmen.[13]

2.2 Motivation der Arbeit

Wie bereits bemerkt, ist ein weit verbreitetes Verständnis von Prognose, ungeachtet ob es sich um eine deterministische oder stochastische Zielgröße handelt, die Angabe einer einzigen Zahl. Als Beispiel gibt eine Prognosemethode aus dem Gebiet der Zeitreihenanalyse den zukünftigen Wert der Inflationsrate für das nächste Jahr als einen bestimmten Prozentsatz an. Die Prognose eines Aktienkurses für den kommenden Tag oder die folgende Woche wird durch die meisten Methoden auf einen bestimmten Wert oder eine

[10] Vgl. etwa (Wirth and Hipp, 2000) und (Chapman et al., 2000).
[11] Siehe (Chatfield, 1996).
[12] Vgl. Kapitel 7.
[13] Siehe etwa (DeLurgio, 1998), (Farnum and Stanton, 1989), (Holden et al., 1990), (Hüttner, 1986), (Mittelhammer et al., 2000) oder (Newbold and Bos, 1993).

bestimmte Veränderung geschätzt. Es werden teilweise Korridore angegeben, die jedoch jegliches, auf der Vergangenheit begründetes, mathematisch statistisches Fundament vermissen lassen. Der zumeist genannte und bekannteste einzelne Prognosewert ist der geschätzte bedingte Erwartungswert der stochastischen Zielgröße. Diebold formuliert: *„In fact, the bulk of the literature focuses on point forecasts, while conspicuously smaller sub-literatures interval forecasts and probability forecasts.“*[14] Diese Ansicht ist etwa auch in (Tay and Wallis, 2000) wiederzufinden.

Aus der Entscheidungstheorie ist bekannt, dass die Betrachtung der gesamten Wahrscheinlichkeitsverteilung, im Gegensatz zu der Beschränkung auf den bedingten Erwartungswert oder die bedingte Varianz einer Zufallsgröße, vonnöten ist. Timmermann schreibt sehr passend zu den Motivationsaspekten dieser Arbeit in einem Editorial: „*A decision maker whose loss function depends asymmetrically on the outcome of future values of possibly non-Gaussian variables will generally want to know not only the conditional mean or variance but also the full conditional density of the variables.*“[15]

Im Gegensatz zu der vereinfachten Sichtweise der ausschließlichen Prognose des bedingten Erwartungswerts impliziert der statistische Blickwinkel, dass die zu prognostizierende stochastische Zielgröße einer gewissen bedingten Wahrscheinlichkeitsverteilung gehorcht. Y bezeichnet einen Zufallsvektor, dessen Verteilung bedingt unter dem erklärenden deterministischen Variablenvektor x formal durch

$$d(y|x) \tag{2.1}$$

dargestellt ist.[16]

Vorausgesetzt es sind Entscheidungen über zukünftige Aktionen zu treffen, kann jeder Aktion a unter der Annahme einer Situation y ein Wert zugeordnet werden. Unter einer Verteilung der Ereignisse y kann eine optimale Entscheidung in einem gewissen Sinn getroffen werden. Das Optimalitätskriterium könnte etwa der minimale erwartete Verlust darstellen. Sei der Verlust

[14] Siehe (Diebold et al., 1998).
[15] Siehe (Timmermann, 2000).
[16] Vgl. (Stuart et al., 1999).

für die Entscheidung a und dem Ereignis y, bedingt unter der erklärenden Variablen x definiert als $L(a, y, x)$. Dann ergibt sich die optimale Entscheidung als

$$\arg\min \int_y L(a, y, x) d(y|x) dx, \tag{2.2}$$

die offensichtlich nicht ohne Kenntnis der bedingten Wahrscheinlichkeitsverteilung $d(y|x)$ berechenbar ist.

Die entscheidungstheoretischen Vorteile der Kenntnis der gesamten bedingten Verteilung motivieren aus unterschiedlichen Gründen die Entwicklung des hier präsentierten Prognosekonzepts. Dieser Aspekt wird in Abschnitt 2.2.4 ausführlich aufgegriffen, um die aufwendige Vorgehensweise unter diesem Konzept zu begründen und zu rechtfertigen.

Leider ist es nicht auf einfachem Wege möglich, das allgemeine Gesetz zu modellieren, das den Zusammenhang zwischen der Wahrscheinlichkeitsverteilung und dem Inputvektor x beschreibt. Gäbe es eine Menge von Beobachtungen zu demselben Attributvektor mit unterschiedlichen Realisationen, so wäre es möglich, die Parameter der Verteilung über klassische Schätzmethoden direkt zu bestimmen. Diese Situation kommt jedoch in der Praxis sehr selten vor. In den meisten Fällen liegen paarweise Datensätze (y, x) vor, die annähernd in einem kontinuierlichen Raum bezüglich beider Dimensionen auftreten, so dass kaum ein Attributvektor x doppelt oder sogar mehrfach auftaucht. Aus diesem Grund ist eine Anforderung an dieses Konzept, die Wahrscheinlichkeitsverteilung als Funktion des Inputvektors x auf einem weniger direkten Weg mit Hilfe der vorliegenden Datengrundlage zu identifizieren.

Die klassische Statistik legt bislang das Augenmerk auf die analytisch durchgehend begründbaren Techniken, deren Aussagen unter gewissen vereinfachten Annahmen Gültigkeit besitzen. Leider beeinträchtigen diese einschränkenden Annahmen die Allgemeingültigkeit und dadurch die Anwendbarkeit von Prognoseergebnissen als Entscheidungsunterstützung unter den Voraussetzungen, Umständen und Situationen der realen Welt.

Die beliebteste und häufigste Annahme ist hierbei die Gauß'sche Normalverteilung der Residuen. Ferner werden die Varianzen der Residuen dabei als

identisch und unabhängig von den erklärenden Inputvariablen vorausgesetzt. Im Fall einer einzigen Zielgröße kommt dies der Annahme von Homoskedastizität gleich und führt daher zu der bekannten Regressionsformel

$$y = \bar{Y}^T \bar{X} (\bar{X}^T \bar{X})^{-1} x, \tag{2.3}$$

wobei $\bar{X}$ und $\bar{Y}$ die historischen Datenmatrizen darstellen (in der Terminologie des maschinellen Lernens bzw. von neuronalen Netzen auch als Trainingsdaten bezeichnet).[17]

Das allgemeine lineare Modell[18] lässt zwar heteroskedastische Zielgrößen zu, verlangt jedoch a priori eine Bestimmung der Kovarianzmatrix Σ, d.h. die Kovarianzmatrix muss vor der Identifikation des Prognosemodells bekannt sein. Weiterhin ist der Einfachheit halber in den meisten Fällen eine diagonale Kovarianzmatrix angenommen, so dass keine Abhängigkeit zwischen den Zielgrößen modelliert wird. Die Prognose berechnet sich in diesem Fall aus der allgemeinen Regressionsformel

$$y = \bar{Y}^T \Sigma^{-1} \bar{X} (\bar{X}^T \Sigma^{-1} \bar{X})^{-1} x. \tag{2.4}$$

Diese Charakteristika von linearen Modellen sei an dieser Stelle lediglich einführend und zielführend dargestellt. Ausführliche Beschreibungen und der klare Zusammenhang zu dem hier entwickelten Prognosekonzept wird in Abschnitt 7.1.1 diskutiert.

Zeitreihenmodelle, die eine heteroskedastische Modellierung ermöglichen, sind etwa die bekannten Methoden ARCH[19] oder GARCH[20] und ihre zahlreichen Erweiterungen.[21] Diese Klasse von Prognosetechniken sind jedoch nur in der Lage, spezielle funktionale Zusammenhänge der Kovarianzmatrix und den erklärenden Variablen abzubilden.[22]

[17] Homoskedastische bekannte Zeitreihenmethoden, wie etwa ARMA- oder ARIMA-Modelle, fallen ebenfalls in diese Klasse der Regressionsmodelle.
[18] Siehe (Aitken, 1935).
[19] Siehe (Engle, 1982).
[20] Siehe (Bollerslev, 1986).
[21] Vgl. (Bera and Higgins, 1993).
[22] Siehe dazu die Abschnitte 7.1.4.6 und 7.1.4.7.

Diese einschränkenden Annahmen sind in vielen praktischen Anwendungen nicht zu vertreten. Etwa im Fall der Bedarfsprognose von Ersatzteilen oder der Nachfrageprognose von Nutzfahrzeugen, die in den Kapiteln 9 und 10 dargestellt sind, werden einige Defizite und die damit verbundenen Prognosefehler bzw. Kosten deutlich. In den jeweiligen Abschnitten ist diese Behauptung detailliert begründet und anhand von Ergebnissen verifiziert.

Ein weiteres klassisches, jedoch nicht sehr verbreitetes Prognosekonzept stellen die verallgemeinerten linearen Modelle[23] dar. Diese lockern zwar die oben angeführten strengen Annahmen etwas auf, lassen jedoch andersartige Probleme vermuten. Das verallgemeinerte lineare Modell ist ein Gerüst, das es zulässt lineare Abhängigkeiten der Varianzen von exogenen Einflussgrößen zu modellieren. Jedoch ist leider nicht einmal für diese lineare Abhängigkeit eine geschlossene analytische Lösung bekannt.

Aus diesen bisherigen Erläuterungen ergeben sich die ersten Defizite klassischer Prognosemethoden, die sich dadurch gleichzeitig als Herausforderungen eines zu entwickelnden Konzepts präsentieren.

Ein weiterer motivierender Aspekt ist die Verwendbarkeit unterschiedlicher Verteilungsklassen, wie etwa multivariater Normalverteilungen, multivariater t-, stabiler[24] und hyperbolischer[25] Wahrscheinlichkeitsverteilungen. Ebenso ist das Konzept für nichtparametrische Approximationsverteilungen, wie Mixturverteilungen[26] obiger Klassen tauglich.

Zusammenfassend ist das klar definierte Ziel des zu entwickelnden Prognosekonzepts die Schätzung einer multivariaten bedingten Wahrscheinlichkeitsverteilung

$$d(y|x).$$

[23] Siehe (McCullagh and Nelder, 1989) oder (Fahrmeir and Tutz, 1994).
[24] Siehe etwa (Rachev and Mittnik, 2000).
[25] Siehe etwa (Eberlein and Prause, 2000).
[26] Siehe etwa (Titterington et al., 1985) oder (Redner and Walker, 1984).

2.2.1 Motivation durch den Aspekt der Attribut-Basiertheit

Wie bereits erwähnt soll im Folgenden eine quantitative Prognosemethodik entwickelt werden, die aufgrund der zur Verfügung stehenden Daten in der Lage ist, den funktionalen Zusammenhang zwischen den Parametern der zugrunde gelegten Wahrscheinlichkeitsverteilung und den exogenen Inputvektoren abzubilden.

Durch diese Charakterisierung entsteht eine gewisse Verwandtschaft zu den bekannten Regressionsmodellen aus der Statistik, die jedoch ausschließlich den Zusammenhang zwischen exogenen Variablen und dem bedingten Lokationsparameter identifizieren, wobei die Zielvariable als normalverteilt angenommen wird. Diese attributbasierte bedingte Vorgehensweise besitzt jedoch viele Vorteile gegenüber den klassischen Techniken aus der Zeitreihenanalyse, die auf einzelnen Datensätzen operieren.

Die Qualität jedes Prognosemodells ist von den zugrunde liegenden Daten abhängig. Sind die Datenzeitreihen kurz und lückenhaft, wie es in der Praxis häufig vorkommt, sind herkömmliche Zeitreihenmethoden meist nicht in der Lage, sinnvolle Prognosemodelle zu repräsentieren.

Stehen jedoch eine Vielzahl von „ähnlichen“ Zeitreihen[27] zur Verfügung, haben attributbasierte Modelle durch das Lernen über alle Datensätze klare Vorteile. Durch die größere Datenmenge ist es viel eher möglich, die Zusammenhänge zwischen den Verteilungsparametern und den exogenen Variablen zu identifizieren und eine Generalisierungsfähigkeit des Prognosemodells zu erreichen.

Falls jedoch, etwa durch Messreihen unter gleichen Bedingungen, für einen Attributvektor mehrere Beobachtungen zur Verfügung stünden, könnte die Verteilung direkt geschätzt werden. Da dies in der Praxis sehr selten vorkommt, ist eine bedingte Modellierung erforderlich.

Sind die Datensätze so heterogen, dass unterschiedliche funktionale Zu-

[27] Als ähnliche Zeitreihen werden in diesem Zusammenhang Datensätze verstanden, für die ähnliche funktionale Zusammenhänge zwischen den unabhängigen und abhängigen Variablen angenommen werden können.

sammenhänge zwischen den exogenen Variablen und den Zielgrößen angenommen werden müssen, sind Cluster mit ausreichend vielen homogenen Trainingsdaten zu formieren, mit Hilfe derer die dazu korrespondierenden Modelle erstellt werden können. Der bedingte Ansatz extrahiert somit Informationen aus der Gesamtheit der Datensätze und vermeidet das Dilemma der zu kurzen und informationsschwachen Zeitreihen.

Ein weiterer sehr wichtiger Vorteil von Regressionsmodellen ist die auch auf kürzeren Zeitreihen mögliche Generalisierungsfähigkeit. Durch diese Eigenschaft kann der gelernte funktionale Zusammenhang sogar auf Zielvariablen angewendet werden, die keine Vergangenheit besitzen, d.h. es können etwa mit den optimierten Prognosemodellen zukünftige Absatzzahlen von neuen Produkten anhand ihrer Eigenschaften errechnet werden.

2.2.2 Motivation durch die Annahme flexibler Verteilungsklassen

Die Prognostik beschäftigt sich im Wesentlichen mit der Identifikation von Wahrscheinlichkeitsverteilungen stochastischer Prozesse bzw. Zufallsgrößen, um durch das Erlernte aus der Vergangenheit auf den Fortgang der Zielvariablen in der Zukunft zu schließen.

Da unterschiedlichen Zufallsgrößen verschiedene stochastische Prozesse unterliegen und sie somit verschiedenen Verteilungen gehorchen, ist es für eine adäquate Modellierung vonnöten, der Prognosetechnik eine flexible Stochastik zur Hand zu geben. Es ist in der Regel erforderlich, Zufallsvariablen zu modellieren, die etwa schief verteilt sind, ausschließlich auf der positiven reellen Achse leben oder verstärkt Ausreißer aufweisen.

Konkret bedeutet dies, dass es für die Modellierung von entscheidendem Vorteil sein kann, neben der Normalverteilung alternative Wahrscheinlichkeitsverteilungen zur Verfügung zu haben, die in der Lage sind, die Prozesse realistischer abzubilden.[28]

[28]Die ungenaue Approximation der Entwicklung von Finanzderivaten durch die Normalverteilung wurde schon in (Mandelbrot, 1963) beanstandet. Auch in (Anscombe, 1967) steht über die Normalverteilung geschrieben: „*...is too good to be true*“.

Mögliche Ansätze bieten die Familien der t-, stabilen, hyperbolischen oder Mixtur-Verteilungen, die Eigenschaften, wie Schiefe oder „heavy tails“ besitzen. Basiert ein Prognosekonzept auf flexiblen Verteilungsklassen, wie die oben genannten, so ist eine realistischere Modellierung im Gegensatz zur Normalverteilungsannahme offensichtlich.

2.2.3 Motivation durch Variabilitätsaussagen

Die primären Anforderungen an ein Prognosesystem sind nicht die automatische Durchführung von Aktionen oder das autarke Entscheiden, vielmehr werden Prognoseergebnisse in erster Linie von Verantwortlichen als entscheidungsunterstützend zu Rate gezogen. Dadurch ist es um so wichtiger, dass die Prognose interpretierbar ist und das Prognosemodell sogar eigenständig Güteaussagen über die gelieferten Ergebnisse vorschlägt.

Ein Beispiel für eine gute Interpretierbarkeit der Ergebnisse liefert die Kenntnis der bedingten Wahrscheinlichkeitsverteilung des zu prognostizierenden Vektors von Zielgrößen. Etwa die Standardabweichung als Maß der Streuung einer stochastischen Zufallsgröße kann zur Bewertung der Unsicherheit eingesetzt werden, da sie die Häufigkeit des Auftretens von extremen Werten zum Ausdruck bringt. Ist etwa die prognostizierte Standardabweichung eines Produktes mit geringem Preis höher als die eines teuren, so kann von einer zuverlässigeren Prognose des kostenaufwendigeren Artikels ausgegangen werden.

Gleichzeitig stellt die geschätzte Standardabweichung eine Bewertung des prognostizierten Mittelwerts dar. Ist sie groß, so ist es gefährlich sich auf die Erwartungswertprognose zu stützen, da vom Mittelwert weit entfernte Werte relativ wahrscheinlich sind.

Eine eindeutige Ursache für eine Vergrößerung der Standardabweichung ist die Erhöhung des Prognosehorizonts. Je weiter der Prognosezeitpunkt von der Gegenwart entfernt liegt, d.h. je weniger sichere Informationen zur Verfügung stehen und je mehr sich die Umwelt in der Zwischenzeit verändern kann, desto stärker wächst die Unsicherheit und damit die prognostizierte Varianz. Der Vergleich unterschiedlicher Streuungen kann daher nur in Relation

zum Prognosehorizont gültig sein. Ein prognostizierter Wert ist daher erst durch die zusätzliche Kenntnis der Streuung oder allgemein der bedingten Wahrscheinlichkeitsverteilung interpretierbar. Ohne diese Zusatzinformation ist die Aussage rein zufällig und für Entscheidungsträger nahezu wertlos.

Des Weiteren ist die Kenntnis der einzelnen Varianzen ebenfalls aus anwendungsorientierter Sicht von großem Interesse. Ist es etwa erforderlich, aus zwei oder mehreren einzelnen Zielgrößen die Summe zu prognostizieren, ist dies ohne die Kenntnis des Korrelationskoeffizienten i.Allg. nur ungenau möglich.[29]

2.2.4 Motivation aus der Entscheidungstheorie

Weshalb die Unsicherheit in der realen Welt eine große Rolle spielt und daher vor allem in der Prognostik betrachtet und quantifiziert werden muss, liegt unter anderem in der Entscheidungstheorie begründet.[30] Morgan und Henrion schreiben: *„Ein gängiger Ansatz mit Unsicherheit umzugehen ist sie zu ignorieren.“*[31] Quade bezeichnet diese Strategie der Ignoranz als *„a chronic disease of planers“*.[32] Um ehrlich zu sein, auch in heutiger Zeit ist diese Krankheit noch nicht ausgestorben.

Natürlich gibt es Stimmen, die behaupten, dass durch Verwendung des „besten Schätzers“ oft genug das Ziel einer optimalen Entscheidung, im Sinne des maximalen Nutzens, erreicht werden kann. Dieser Anspruch ist jedoch zu gering, um realistisch zu sein.

Die Entscheidungstheorie liefert das tatsächliche und unumstößliche Argument, dass es unbedingt notwendig ist, die Unsicherheit einer Prognoseaussage zu quantifizieren, um optimale Entscheidungen treffen zu können. Es seien im Folgenden einige entscheidungstheoretische Begründungen dargestellt.[33]

---

[29] Durch den Zusammenhang $\sigma_{Y_1+Y_2} = \sqrt{\sigma^2_{Y_1} + \sigma^2_{Y_2} + 2\rho_{Y_1,Y_2}\sigma_{Y_1}\sigma_{Y_2}}$ ist die Abhängigkeit der Varianz der Summe von dem Korrelationskoeffizienten ρ ersichtlich.

[30] Vgl. (Morgan and Henrion, 1992), S. 43 ff.

[31] Siehe (Morgan and Henrion, 1992).

[32] Siehe (Quade, 1975).

[33] Vgl. für die folgenden Ausführungen (Morgan and Henrion, 1992), S. 43 ff.

- Jeder Mensch besitzt eine individuelle Risikoaversion.

- Falls sich eine Entscheidung aus unterschiedlichen unsicheren Informationen ergibt, so sollten die einzelnen Unsicherheiten die Gewichtung der Kombination beeinflussen.

- Grundsätzlich wird die Begrenzung des zusätzlichen Aufwands, neue Informationen zu beschaffen abhängig von der vorherrschenden Unsicherheit beeinflusst. Je größer die Unsicherheit ist, desto größer ist der erhoffte Mehrwert weiterer relevanter Informationen.

Zusätzlich zu diesen bekannten Begründungen wird in (Henrion, 1982) ein weiteres entscheidungstheoretisches Argument angeführt, das die Behandlung und Berücksichtigung von Unsicherheit fordert. Es wird durch die mögliche Asymmetrie der Verlustfunktion um die optimale Entscheidung argumentiert.[34] Dass dieses Phänomen in der Praxis Auftritt, zeigt etwa die Bedarfsprognose von Ersatzteilen in Kapitel 9, wobei von unterschiedlichen Kostensätzen der Nachbestellung gegenüber der Lagerhaltung ausgegangen werden muss.

Dieses Argument wird auch unter anderem in (Diebold et al., 1998) und (Tay and Wallis, 2000) aufgegriffen und dient als Motivation des nachfolgenden Konzepts der Verteilungsprognose.

Ist nach der Identifikation der bedingten Wahrscheinlichkeitsverteilung eine Entscheidung über die tatsächlich durchzuführende Aktion zu treffen, so ist nicht garantiert, dass der am wahrscheinlichsten auftretende Wert eine optimale Entscheidung repräsentiert. Dies ergibt sich direkt aus Gleichung (2.2), die hier wiederholend formuliert wird

$$\arg\min \int_y L(a, y, x) d(y|x) dx. \tag{2.5}$$

Der einzige Fall, für welchen der bedingte Erwartungswert eine optimale Entscheidung darstellt, ergibt sich bei einer symmetrischen Verlustfunktion L um

[34] Das in diesem Zusammenhang entwickelte Maß, bezeichnet als „*expected value of including uncertainty*“ ist z.B. in (Morgan and Henrion, 1992) dargestellt.

das Ereignis y. Sobald sich diese Funktion als asymmetrisch[35] erweist, muss mit Hilfe der Kenntnis über die geschätzte Wahrscheinlichkeitsverteilung in einem zusätzlichen Schritt eine optimale Entscheidung berechnet werden.[36]

2.2.5 Motivation durch universelle Anwendbarkeit

Die Eigenschaft der universellen Anwendbarkeit von Prognosemethoden wird immer wieder von Anwendern gefordert und gleichzeitig bezweifelt. Viele Autoren sind überzeugt, dass es ein universelles Prognosekonzept nicht geben kann und es daher erforderlich ist, für jede Aufgabe eine spezielle statistische Methodik auszuwählen und ein Prognosemodell neu zu konzipieren.

Im Gegensatz dazu meint die universelle Anwendbarkeit in diesem Zusammenhang die Flexibilität und mögliche Anpassung des Konzepts auf eine Fülle von Prognoseaufgaben und Situationen aus der Praxis.

Der spezifische Prozess der Modellkonzeption und Kalibrierung ist bei jeder neuen Aufgabe natürlich unumgänglich. Es sind bei jeder praktischen Anwendung datenvorverarbeitende Schritte durchzuführen oder Analysen etwa bezüglich der Relevanz einzelner Einflussgrößen notwendig. Diese Prozessstufen sind jedoch, wie bereits erwähnt, nicht Inhalt und Thema dieser Arbeit.

Prognoseaufgaben treten sowohl in vielen unterschiedlichen Situationen des täglichen Lebens als auch bei einer Vielzahl von betriebswirtschaftlichen Fragestellungen auf. Finanzmärkte beschäftigen sich täglich, sogar stündlich oder minütlich mit der Frage, wie sich wohl die Derivate entwickeln werden. Produzierende Industrieunternehmen oder private und öffentliche Dienstleistungsunternehmen versuchen in regelmäßigen Abständen die Umsatzzahlen, die Kosten, den erwarteten Gewinn und vieles mehr zu prognostizieren, um Planungen für die Zukunft entwickeln zu können. Im aktuellen Themengebiet des Supply Chain Managements sind Prognosen die Basis des kompletten Konzepts. Aufgrund von Vorhersagen sollen die Informationsflüsse zwischen den einzelnen „Mitspielern“ des Netzwerkes verbessert werden, um gewisse

[35] Eine asymmetrische Verlustfunktion ist in Gleichung (9.4) aufgeführt. Es können jedoch in der Realität beliebig komplexe Funktionen auftreten.

[36] Zur Herleitung einer optimalen Entscheidung siehe Abschnitt 9.8.

Ziele zu optimieren.[37]

All diese Aufgabenstellungen basieren auf der identischen Fragestellung, deren Kern die Identifikation eines Prognosemodells aus bekannten und zur Verfügung stehenden Informationen darstellt. Es wird daher ein weiteres Mal die Nützlichkeit eines präzisen, interpretierbaren und aussagekräftigen Prognosesystems deutlich.

Wie bereits erwähnt beschäftigen sich Finanzmärkte ebenfalls intensiv mit der Zukunft. Die Modellierung des Wertverlaufes von Derivaten mit Hilfe von Wahrscheinlichkeitsverteilungen dient jedoch nicht ausschließlich dem Zweck des reinen Handels, sondern ebenso der Risikoanalyse und des Risikocontrollings, was besonders die Aussagekraft der Ergebnisse fordert. Bekannte und aktuelle Risikoanalysetechniken, wie etwa die „Value at Risk“ (VaR) Analysen benötigen die Schätzung der zugrunde liegenden Verteilung als Basis ihrer Berechnungen. Diese a posteriori Anwendung zeigt darüber hinaus den Zusatznutzen der Kenntnis der gesamten bedingten Wahrscheinlichkeitsverteilung.

Nach der ausführlichen Motivation eines Prognosekonzepts, das die funktionalen Zusammenhänge zwischen den exogenen Variablen und den Verteilungsparametern modelliert, geht der folgende Abschnitt auf die tatsächliche Formulierung des beschriebenen Prognoseverständnisses detailliert ein.

2.3 Formulierung der Prognoseaufgabe

Für den Zweck der Formalisierung von Prognoseaufgaben werden folgende Bezeichnungen eingeführt und Vereinbarungen festgelegt.

Sei eine Datenbasis bestehend aus K Paaren (x_k, y_k) gegeben, so ergibt sich die Datengrundlage der Prognoseaufgabe wie in Tabelle 2.1 dargestellt.

In Matrixschreibweise ergeben sich die Beobachtungen x_k und y_k als Zeilen der Matrizen $\bar{X} \in I\!R^{K \times m}$ und $\bar{Y} \in I\!R^{K \times n}$. Die Komponenten der Matrix $\bar{Y}$ heißen abhängige (Response-, Ziel-) Variablen, wohingegen die Einträge der Matrix $\bar{X}$ die unabhängigen, erklärenden Variablen repräsentieren. Diese

[37] Vgl. etwa (Christopher, 1998) oder (Petrovic et al., 1998).

Datensatz	unabhängige Variablen	abhängige Variablen
1	$x_{11}, \ldots, x_{1m}$	$y_{11}, \ldots, y_{1n}$
2	$x_{21}, \ldots, x_{2m}$	$y_{21}, \ldots, y_{2n}$
$\vdots$	$\vdots$	$\vdots$
K	$x_{K1}, \ldots, x_{Km}$	$y_{K1}, \ldots, y_{Kn}$

Tabelle 2.1: Allgemeine Datenbasis bei Prognoseaufgaben

werden auch häufig als Inputvariablen bezeichnet.

Wie bereits erwähnt, wäre durch eine Vielzahl von Beobachtungen y_k zu jedem Attributvektor x_k ein direkte Parameterschätzung möglich. So wäre es aufgrund der theoretisch fundierten Schätztheorie mit Hilfe eines adäquaten Punktschätzers möglich, die Verteilungsparameter direkt zu bestimmen. Eine realistische Datenbasis beinhaltet jedoch in der Regel zu jedem Attributvektor genau ein Beobachtungstupel y_k. Diese Situation beschreibt die übliche Grundlage der Regressionsanalyse und sogar generell einer Prognoseaufgabe.

Die Zielgrößen y_k sind bekannte Realisationen der Zufallsvariablen Y_k, deren Verteilung zu bestimmen ist. Die Datenbasis stellt daher eine empirische bedingte Dichtefunktion dar, die es durch numerische Verfahren zu approximieren gilt. Die unterschiedlichen Datensätze, d.h. die einzelnen Beobachtungen, werden i.Allg. als unabhängig angenommen. Bei einem multivariaten Ansatz beschreibt die Anzahl n der Komponenten des Vektors y_k gleichzeitig die Dimension der zu bestimmenden Wahrscheinlichkeitsverteilung. Werden sukzessive die einzelnen Randverteilungen geschätzt und anschließend die Korrelationsstruktur, so handelt es sich um n univariate Schätzprobleme.

Die Definition und Konzipierung des unabhängigen Inputvektors x_k beeinflusst entscheidend die Charakteristik der Modellierung. Prinzipiell kann der Inputvektor aus beliebig vielen unterschiedlichen Informationsquellen bestehen. Aus diesem Grund sind sowohl Vergangenheitswerte der Zielgröße Y_k als Inputvariablen denkbar als auch exogene Einflussgrößen, die den Umweltzustand ausdrücken oder die Zielgröße charakterisieren. Zeitabhängige Variablen, die etwa die Saisonalität oder schlicht den Prognosezeitpunkt beschreiben, sind ebenso als beeinflussende exogene Größen möglich. Weiterhin

können diese Inputvariablen beliebig transformiert in die Modellierung involviert sein.

Die Gemeinsamkeit dieser Inputgrößen ist deren Bekanntheit zum Prognosezeitpunkt. Sie werden im Moment der Modellberechnung als deterministisch angenommen. Unsichere Einflüsse erfordern eine realisierbare Erweiterung der Methodik, auf die hier nicht eingegangen wird.

Schon an dieser Stelle kommt zum Ausdruck, dass durch die Wahl und Definition der Inputvariablen des Prognosemodells Techniken abgebildet werden können, die in der Literatur unterschiedlich behandelt sind. Haben etwa ausschließlich vergangene Werte der Zufallsvariablen Einfluss auf den zu prognostizierenden Lokationsparameter, so liegt der Fall eines einfachen homoskedastischen Zeitreihenmodells vor.[38]

Wie in obigen Abschnitten ausführlich dargestellt, ist das wesentliche Ziel einer Prognoseaufgabe, den funktionalen Zusammenhang zwischen den unabhängigen Variablen x und der Verteilung der Zielvariablen $d(y|x)$ zu identifizieren, um mit gegenwärtig bekannten Einflussgrößen eine Vorhersage des Zielvektors berechnen zu können.

Die bedingte multivariate Wahrscheinlichkeitsdichte des k-ten Responsevektors sei, wie schon in Gleichung (2.1) eingeführt, mit

$$d_\theta(y_k|x_k)$$

bezeichnet. Weiterhin sei die Wahrscheinlichkeitsdichte durch den bedingten Parametervektor $\theta(x) \in I\!R^P$ eindeutig bestimmt.

Die Prognoseaufgabe ergibt sich daher als die Suche nach dem funktionalen Zusammenhang zwischen den unabhängigen Variablen x und den Parametern der bedingten Wahrscheinlichkeitsverteilung θ. Formal lässt sich diese Identifikationsaufgabe mit folgender Gleichung ausdrücken

$$\theta = f_\omega(x). \tag{2.6}$$

[38] Die Bildung zeitlicher Fenster der Datensätze ermöglicht eine Erfassung von Autokorrelationen im Zeitreihenkontext. Eine Autokorrelation n-ter Ordnung erfordert n vergangene Werte der Zeitreihe als Inputgrößen x.

Hierbei bezeichnet ω die freie Parametrisierung des funktionalen Approximators f.

Am folgenden Beispiel der bekannten Normalverteilung sei die Identifikationsaufgabe für eine spezielle Verteilungsklasse illustriert.

Beispiel 2.3.1 (Normalverteilung)
Sei Y eine $N(\mu(x), \Sigma(x))$-verteilte Zufallsvariable - bedingt durch den erklärenden Vektor $x \in \mathbb{R}^m$ - mit der Wahrscheinlichkeitsdichte

$$d_{(\mu,\Sigma)}(y|x) = \frac{1}{\sqrt{2\pi|\Sigma(x)|}} e^{1/2(y-\mu(x))^T \Sigma(x)^{-1}(y-\mu(x))},$$

so ergibt sich der zu identifizierende funktionale Zusammenhang zwischen den Verteilungsparametern μ und Σ und der exogenen Einflussgröße x als

$$\theta(x) = (\mu(x), \Sigma(x)) = f_\omega(x).$$

□

Für Methoden, die sich ausschließlich auf die Identifikation des funktionalen Zusammenhangs zwischen den Inputvariablen und dem Lokationsparameter der Verteilung beschränken, ergibt sich in diesem speziellen Fall die Gleichung (2.6) zu

$$\theta(x) = E[Y|X = x] = f_\omega(x). \tag{2.7}$$

Es wird in Kapitel 7 gezeigt, dass gewisse Methoden aus der Regressions- und Zeitreihenanalyse dieser Modellierungsart entsprechen und somit einen Spezialfall der Gleichung (2.6) darstellen.

Zusammenfassend ist eine Prognoseaufgabe in diesem Sinne die Identifikation eines funktionalen Zusammenhangs zwischen den Charakteristika der bedingten Wahrscheinlichkeitsverteilung des Zielgrößenvektors Y und dem gegebenen Einflussvektor x.

Teil I

Prognosekonzept: Identifikation von multivariaten bedingten Wahrscheinlichkeitsverteilungen

Einführend zu folgenden konzeptionellen Darstellungen und ausgehend von der Herausforderung, eine Prognosemethodik, basierend auf bedingten multivariaten Wahrscheinlichkeitsverteilungen zu entwickeln, werden die zur Realisierung erforderlichen Teilaufgaben und die daraus resultierenden Ziele definiert. Mit Hilfe des Schemas aus Abbildung 2.1 wird die anschließende konzeptionelle Vorgehensweise illustriert. Hieraus ergibt sich eine Gliederung in die Kapitel und Abschnitte des folgenden Konzeptteils. Die Gesamtheit des ersten Hauptteils formt das in dieser Arbeit präsentierte Prognosesystem.

Wie schon mehrmals betont und sowohl durch praktische als auch theoretische Motivationsaspekte untermauert, ist das vorwiegende Ziel dieses Konzepts die datenbasierte Identifikation der „besten“ bedingten multivariaten Wahrscheinlichkeitsverteilung einer Zielvariablen in Abhängigkeit von beobachtbaren Einflussgrößen. Zu Beginn der Konzeption eines derartigen Prognosesystems stellt sich die Frage.

Wie lassen sich Klassen von Wahrscheinlichkeitsverteilungen generieren? Basierend auf einer „einfachen kanonischen Dichtefunktion“ konstruiert das hier gewählte Konzept über eine lineare Transformation die restlichen Elemente der angenommenen Klasse von Wahrscheinlichkeitsverteilungen. Die an dieser Stelle erforderliche Verteilungsannahme stellt keine wesentliche Einschränkung der Allgemeinheit dar, da sehr unterschiedliche Verteilungsklassen durch diese Strategie Verwendung finden können.

Die Wahl von konkreten parametrischen Verteilungen macht dieses Konzept im engeren Sinne parametrisch. Da jedoch Mixturverteilungen in diesem Konzept ebenfalls vorausgesetzt werden können und somit jede beliebige Form einer nichtparametrischen stetigen Dichtefunktion durch die Approximation über parametrische Funktionen erzeugt werden kann, ist das vorliegende Konzept im erweiterten Sinn „nichtparametrisch“.[39]

Dieser Prozess lässt sich problemlos auf bedingte Verteilungsklassen übertragen, das Ziel aus Abschnitt 2.2.1 ist also erreicht.

Eine Beschreibung der Generierung von Verteilungsklassen und eine erste

[39]Die Modellierung von bedingten Dichtefunktionen wird etwa bei (Bishop, 1995b) in die drei Kategorien parametrisch, nicht-parametrisch und semi-parametrisch eingeteilt. Im eigentlichen Sinn sind jedoch alle parametrische Approximationen.

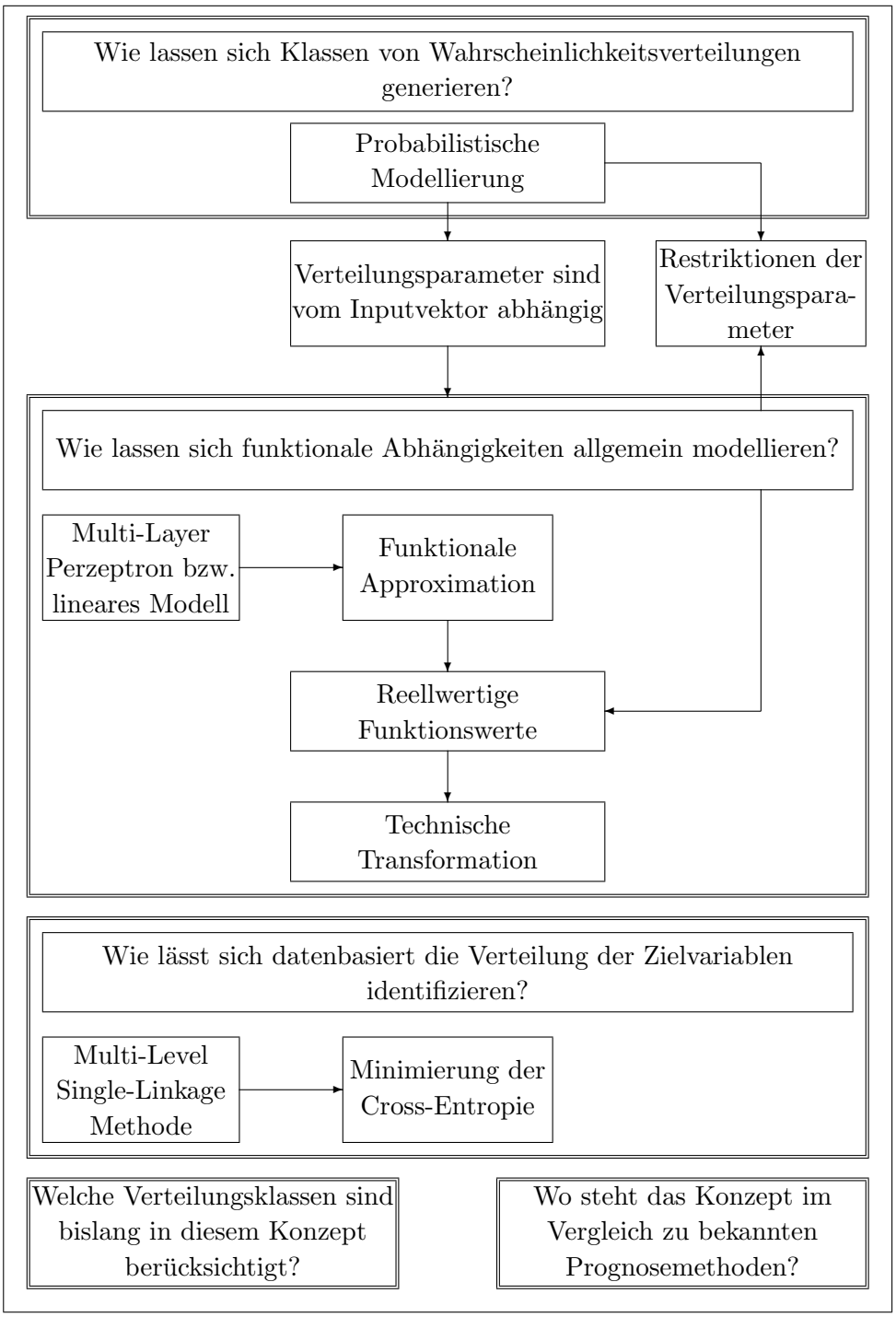


Abbildung 2.1: Skizze der Konzeptstruktur

kurze Übersicht zu den hier behandelten Verteilungsklassen ist Inhalt des folgenden Kapitels 3, das den Grundstein des Konzepts bildet.

Aufgrund der Wahl eines bedingten Ansatzes stellt sich die in der folgenden Frage formulierte Herausforderung, den funktionalen Zusammenhang zwischen den Verteilungsparametern und den Inputvektoren zu modellieren:

Wie lassen sich funktionale Abhängigkeiten allgemein modellieren? Da die Verteilung der zu prognostizierenden Variablen in dem hier präsentierten Konzept von deren Vergangenheit oder von exogenen Einflüssen abhängt, d.h. eine bedingte Verteilung angenommen ist, gilt es, einen Approximator zu wählen und zu identifizieren, der in der Lage ist, den funktionalen Zusammenhang abzubilden. Es werden in Kapitel 4 zwei Methoden zur funktionalen Approximation dargestellt. Da die neuronalen Netze sehr viel mächtigere Hilfsmittel als die linearen Approximationen an die Hand geben, sind diese mit ihren Eigenschaften ausführlich behandelt. Speziell die Fähigkeit der universellen Approximationsfähigkeit des Multi-Layer Perzeptrons wird in diesem Zusammenhang beschrieben und von illusorischen Erwartungen abgegrenzt.

Abschließend zu dieser Thematik taucht das Dilemma der restringierten Verteilungsparameter und den i.Allg. reellwertigen Ausgängen der funktionalen Approximatoren auf. Dieses wird über parameterspezifische „technische Transformationen“ gelöst.

Neuronale Netze und lineare Abbildungen beschreiben daher das erste wesentliche technische Hilfsmittel zur Umsetzung des gewünschten Prognosesystems. Um diese funktionalen Approximatoren optimal an die Datengrundlage anzupassen, bedarf es weiteren Maßnahmen aus dem Gebiet der numerischen Optimierung.

Wie lässt sich datenbasiert die Verteilung der Zielvariablen identifizieren? Die Aufgabenstellung, basierend auf realen Daten ein „optimales“ Prognosemodell zu identifizieren, muss mittels Techniken der numerischen Mathematik gelöst werden. Zu Beginn ist eine Kostenfunktion zu definieren. Anschließend an die Herleitung der Zielfunktion über die Anforderung der minimalen Cross Entropie wird das numerische globale Optimierungsverfahren beschrieben. Da die Kostenfunktion i.Allg. nicht konvex ist, reicht

eine lokale Optimierungsprozedur nicht aus, so dass mit Hilfe einer globalen Optimierungsmethode, das in diesem Sinn optimal an die Daten angepasste Modell identifiziert werden kann.

Die Ausführungen des Abschnitts 2.2.2 geben Hinweise, dass eine Modellierung aufgrund der bekannten Normalverteilung i.Allg. nicht ausreicht. Es ergibt sich daher folgende Fragestellung.

Welche Verteilungen sind bisher in diesem Konzept berücksichtigt? Das den Wahrscheinlichkeitsverteilungen gewidmete Kapitel stellt die bislang im vorgestellten Prognosekonzept theoretisch berücksichtigten Verteilungsfamilien dar und zeigt damit die Allgemeinheit und Breite des vorliegenden Ansatzes. Die unterschiedlichen Verteilungsklassen sind kurz beschrieben und ihre Nützlichkeit kommt durch die Darstellung ihrer Flexibilität speziell im Vergleich mit der Normalverteilung zum Ausdruck. Es werden einige Verteilungsklassen besonders betont, die aus gewissen Gründen in einer praktischen Implementierung umgesetzt sind. Die erhoffte potentielle Qualitätsverbesserung durch eine realistischere Modellierung über flexible Verteilungen ist folglich durch diesen Ansatz ermöglicht.

Um die Tauglichkeit und Effizienz des Systems unter Beweis zu stellen, muss es sich vergleichen lassen, und es ist daher die folgende kritische Bemerkung gerechtfertigt:

Wo steht das Konzept im Vergleich mit bekannten Prognosemethoden? Die Abgrenzung der hier beschriebenen Prognosemethodik zu den bekannten und herkömmlichen Prognoseverfahren wird in Kapitel 7 ausführlich erläutert. Die Inklusion von Verfahren aus der Zeitreihenanalyse, der linearen Modelle und von jüngeren Modellen aus der Neuroinformatik ist ein weiterer Beweis für die Mächtigkeit des Prognosekonzepts von bedingten multivariaten Wahrscheinlichkeitsverteilungen. Es sei an dieser Stelle wiederholend bemerkt, dass es in diesem Zusammenhang nicht notwendig ist, sich mit der Übertragbarkeit sämtlicher theoretischer Aussagen zu befassen. Es wird gezeigt, dass die funktionale Modellierung einiger bekannter Prognosemethoden aus der Statistik über das hier vorgestellte Konzept abgebildet werden kann.

Obige Ausführungen liefern neben den skizzierten Anworten zu den for-

mulierten Fragen zusätzlich eine Aussage über die mögliche Umsetzung der Motivationsaspekte aus den Abschnitten 2.2.3, 2.2.4 und 2.2.5. Letztendlich wird allerdings das Erreichen der Ziele erst auf realen Daten ersichtlich, wenn ein direkter Vergleich unterschiedlicher Methoden möglich ist. Dies geschieht in den Kapiteln 9 und 10.

Kapitel 3

Probabilistische Modellierung

Das folgende Kapitel setzt den Grundstein eines Prognosesystems, das im Wesentlichen durch Motivationsaspekte aus der Praxis inspiriert ist. Als Fazit der Formulierung von Prognoseaufgaben stellte sich die Identifikation einer bedingten multivariaten Wahrscheinlichkeitsverteilung heraus. Die Erzeugung solcher Verteilungen sind Inhalt der nachfolgenden Abschnitte.

3.1 Generierung von Verteilungsklassen

In den letzten Jahrzehnten gab es in der Statistik verstärkte Bemühungen, univariate Verteilungen auf ihre multivariaten Pendants zu erweitern. Die Artikel (Kotz, 1975) und (Goodman and Kotz, 1981) geben eine übersichtliche Klassifikation dieser Methoden basierend auf Kriterien wie Abhängigkeitstypus, Analogie der mathematischen Form und anderen Charakteristika. Es entstanden unterschiedliche Möglichkeiten, multivariate Klassen von Wahrscheinlichkeitsverteilungen zu konstruieren. Das in (Fang et al., 1990) vorgestellte und favorisierte Konzept *„By means of stochastic decomposition“* für die Generierung von symmetrischen multivariaten Verteilungen dient der folgenden Vorgehensweise als inspirierendes Prinzip.

Bei vielen Klassen von Wahrscheinlichkeitsverteilungen ist es möglich, eine „kanonische Dichtefunktion“ zu spezifizieren, die als Repräsentant der gesamten Verteilungsklasse dient. Die multivariate kanonische Form erlaubt

eine Interpretation als Basis der Verteilungsklasse, da durch lineare Transformationen alle in dieser Familie existierenden Wahrscheinlichkeitsverteilungen erzeugt werden können. Die Eigenschaft der kanonischen Dichte, nur wenige Verteilungsparameter zu besitzen oder sogar parameterfrei zu sein, vereinfacht den folgenden Ansatz wesentlich.

Diese Vorgehensweise, die im Folgenden als *Verteilungsgenerierung*[1] bezeichnet ist, wird erläutert. Gleichzeitig wird der Bezug zur hiesigen Konzeptentwicklung hergestellt.

Sei $x \in I\!R^m$, wie in Kapitel 2 eingeführt, ein deterministischer Inputvektor und Z ein n-dimensionaler Zufallsvektor mit kanonischer Wahrscheinlichkeitsdichte[2]

$$g_v(z). \tag{3.1}$$

Hierbei bezeichnet v die Parameter der kanonischen Dichtefunktion, falls diese existieren. Da sie die Form der Dichtefunktion erheblich beeinflussen können, wird dieser Typ von Verteilungsparametern von nun an als *Formparameter* bezeichnet.[3]

Die Familie der Normalverteilungen ist ein Beispiel für Verteilungsklassen, die frei von Formparametern sind. Als kanonische parameterfreie Dichtefunktion für die Normalverteilungsklasse lässt sich die bekannte multivariate Gauß'sche Standardnormalverteilung anführen:

$$g(z) = \frac{1}{\sqrt{(2\pi)^n}}\, e^{-\frac{1}{2}z^T z}.$$

Sie besitzt den Erwartungswertvektor Null und die Einheitsmatrix I_n als Kovarianzmatrix. Für weitere Ausführungen zu der hierdurch erzeugten Verteilungsklasse sei auf den Abschnitt 6.1.1.1 verwiesen.

Ein solcher Repräsentant als Basis für eine Verteilungsfamilie kann nicht

[1] Vgl. etwa (Stützle and Hrycej, 2002a).

[2] Es sei bemerkt, dass die multivariate kanonische Dichte g einer Verteilungsklasse nicht generell als Produkt von univariaten Wahrscheinlichkeitsverteilungen derselben Familie darstellbar ist, so z.B. bei der Familie der generalisiert hyperbolischen Wahrscheinlichkeitsverteilungen.

[3] In der englischen Literatur wird diese Art von Verteilungsparametern auch als *shape parameters* bezeichnet.

nur im Gauß'schen Fall identifiziert werden. Ebenso besitzen Verteilungsklassen, wie elliptische, stabile oder hyperbolische Verteilungen parameterfreie oder parameterarme kanonische Dichtefunktionen die das Fundament ihrer Verteilungsfamilie bilden. In Kapitel 6 sind mögliche Verteilungsklassen mit ihren Eigenschaften ausführlich dargelegt.

Als Hilfsmittel für die Erzeugung der Verteilungsfamilie wird folgende lineare Transformation verwendet. Über eine nichtsinguläre Matrix $A = (\alpha_{ij})_{i,j=1,\ldots,n}$ und einen Lageparametervektor $\mu = (\mu_1, \ldots, \mu_n)^T$ lässt sich eine lineare Transformation

$$y = \mu + A^{-1}z \tag{3.2}$$

formulieren, welche den Vektor von Zufallsvariablen $Y = (Y_1, \ldots, Y_n)^T$ erzeugt und folgende Beziehung impliziert:

$$z = A(y - \mu). \tag{3.3}$$

Offensichtlich beeinflusst der Vektor μ wesentlich die Lage der stochastischen Variablen Y und spielt daher die Rolle eines *Lokations-* oder *Lageparameters*. Ferner spiegelt sich implizit in der Matrix A die Skalierungs- und Abhängigkeitsstruktur des Zufallsvektors Y wider. A wird daher im weiteren Verlauf dieser Arbeit als *Struktur-* oder *Skalierungsmatrix* bezeichnet.

Der Transformationssatz für Lebesque-Integrale[4] gibt die Form der Wahrscheinlichkeitsdichte des generierten Zufallsvektors Y an:

$$d(y) = |A| g_v(z). \tag{3.4}$$

Hierbei bezeichnet $|A| \neq 0$ die Determinante der nichtsingulären Matrix A. Die Wahrscheinlichkeitsdichte d aus Gleichung (3.4) repräsentiert somit die ganze Breite der Verteilungsfamilie, deren kanonische Dichte $g_v(z)$ in Gleichung (3.1) eingeführt wurde.

Im Fall von bedingten Wahrscheinlichkeitsverteilungen, den es hier zu

[4] Vgl. (Bauer, 2002), Seite 127.

behandeln gilt, hängen die Verteilungsparameter vom unabhängigen Inputvektor $x \in I\!R^m$ ab. Daher ergibt sich die lineare Transformation (3.3) zu folgender Gleichung:

$$z = A(x)(y - \mu(x)). \tag{3.5}$$

Die bedingte Wahrscheinlichkeitsdichte der i.Allg. multivariaten Zufallsvariablen Y besitzt nach Gleichung (3.4) und unter Verwendung des Zusammenhangs (3.5) die Form

$$d(y|x) = |A(x)|\ g_{v(x)}(A(x)(y - \mu(x))). \tag{3.6}$$

3.2 Mögliche Verteilungsklassen

Eine Vielzahl von Verteilungsklassen können durch dieses Prinzip im Konzept der multivariaten bedingten Wahrscheinlichkeitsverteilungen Verwendung finden. Eine sehr große Untermenge von betroffenen Verteilungen sind die *elliptischen symmetrischen Verteilungen*[5], oder genauer solche, für welche eine analytische Dichtefunktion existiert.

Elliptische symmetrische Verteilungen - oder kurz elliptische Verteilungen - besitzen eine derartige Dichte, dass sich durch waagerechte Schnitte der Dichtefunktion Ellipsen ergeben. Jede elliptische Verteilung $d(y)$ kann über die Transformation (3.2) einer *sphärischen Verteilung* $g(z)$ erreicht werden. Die sphärische Verteilung übernimmt hierbei die Rolle der kanonischen Dichtefunktion (3.1). Die Dichte von sphärischen Verteilungen kann in der Form

$$h(u) := h(z^T z) \tag{3.7}$$

ausgedrückt werden.[6] Hierbei bezeichnet h eine skalare Funktion, die als *Dichtegenerator* bezeichnet wird.

Die breite Klasse der elliptischen Verteilungen beinhaltet z.B. die folgenden multivariaten Verteilungen:

[5] Siehe etwa (Kelker, 1970).
[6] Vgl. (Fang et al., 1990).

- multivariate Gauß-Verteilungen
- symmetrische generalisiert hyperbolische Verteilungen
- symmetrische Kotz-type Verteilungen
- symmetrische Pearson-type VII Verteilungen (einschließlich $t-$ und Cauchy-Verteilungen)
- symmetrische Pearson-type II Verteilungen
- Bessel Verteilungen
- logistische Verteilungen
- symmetrische stabile Verteilungen

Neben elliptischen Verteilungen, die von Natur aus symmetrisch sind, können ebenso ausgewählte asymmetrische Verteilungen in diesem Konzept Verwendung finden. Hierunter lassen sich Verteilungsfamilien, wie etwa

- generalisiert hyperbolische Verteilungen
- asymmetrische stabile Verteilungen
- Gauß'sche Mixtur-Verteilungen

aufzählen.

Die hier aufgeführten Verteilungsfamilien sind die zum derzeitigen Entwicklungsstand des Konzepts betrachteten Klassen von Wahrscheinlichkeitsverteilungen. Kapitel 6 ist diesen unterschiedlichen Verteilungsklassen gewidmet. Die Dichtefunktionen bzw. kanonische Dichten und einige Eigenschaften dieser konkreten Wahrscheinlichkeitsverteilungen sind dort dargestellt.

Kapitel 4

Funktionale Approximation

Ein weiteres Hilfsmittel zur Umsetzung eines bedingten Prognosekonzepts ist Inhalt dieses Kapitels.

Die bedingte Wahrscheinlichkeitsdichte $d_{(\mu,A,v)}(y|x)$ aus Gleichung (3.6) ist vollständig über die lineare Transformation (3.5), repräsentiert durch $\mu(x)$, $A(x)$, und die kanonische Dichte $g_{v(x)}$, bestimmt.

Sind, wie im Folgenden angenommen, keine Bedingungen an eine spezielle Form der Abbildungen $\mu(x)$ und $A(x)$ geknüpft, so können diese durch einen adäquaten generellen funktionalen Approximator dargestellt werden. Die Abbildungen $\mu(x)$ und $A(x)$ seien durch die Vektoren ω_1 und ω_2 parametrisiert und daher mit $\mu_{\omega_1}(x)$ und $A_{\omega_2}(x)$ bezeichnet.

In einigen Klassen von Wahrscheinlichkeitsverteilungen existieren zusätzlich Formparameter der kanonischen Dichte $g_{v(x)}$, wie etwa $\alpha(x)$, $\beta(x)$ und $\lambda(x)$ bei der generalisiert hyperbolischen Verteilungsklasse. In diesen Fällen erweitern sich die parametrischen Funktionen um $v_{\omega_3}(x)$.

Da i.Allg. funktionale Approximatoren, wie z.B. lineare Abbildungen oder neuronale Netze, auf reelle Räume abbilden, die Parameter der bedingten Wahrscheinlichkeitsverteilung jedoch gewissen Restriktionen unterliegen, erfordert dies eine zusätzliche Transformation τ, um die reellen Funktionswerte in die Verteilungsparameter zu überführen. Die konkrete Abbildungsvorschrift dieser erforderlichen Transformation τ wird in Abschnitt 4.3 ausführlich hergeleitet und beschrieben.

Es ergibt sich daher die vollständige zu identifizierende Abbildung durch

$$(\mu, A, v)^T = (\tau \circ \bar{f}_\omega)(x) =: f_\omega(x). \tag{4.1}$$

Hierbei bezeichnet $\bar{f}_\omega$ den *funktionalen Approximator* und $\tau = (\tau_\mu, \tau_A, \tau_v)^T$ repräsentiert den Operator, der die reellen Funktionswerte von $\bar{f}_\omega$ in den Vektor der Verteilungsparameter $(\mu, A, v)^T$ transformiert. Die Abbildung τ wird *technische Transformation*[1] genannt. Die Unterteilung der Transformation in drei Gruppen resultiert aus den verschiedenen Verteilungsparametern und den zugehörigen üblicherweise unterschiedlichen Restriktionen. Diese Gliederung gilt analog für die gesamte Abbildung $f_\omega(x)$.

Als *genereller funktionaler Approximator* entsteht daher ebenso eine aus drei Teilen bestehende Abbildung, die durch

$$f_\omega(x) := (\mu_{\omega_1}(x), A_{\omega_2}(x), v_{\omega_3}(x))^T \tag{4.2}$$

beschrieben ist.

In Abbildung 4.1 ist das allgemeine dreistufige Gesamtkonzept skizziert. Die unterschiedliche Typen von Verteilungsparametern sind durch (μ, A, v) präsentiert.

Die Funktionswerte des funktionalen Approximators $\bar{f}$ werden als *Zwischenparameter*

$$\bar{p} = (\bar{p}_\mu, \bar{p}_A, \bar{p}_v)^T =: \bar{f}_\omega(x) \tag{4.3}$$

bezeichnet. Weiter sei $\bar{P}$ die Dimension des Vektors der Zwischenparameter $\bar{p}$, d.h. für den funktionalen Approximator gilt:

$$\bar{f}_\omega(x) = \bar{p} : I\!R^m \to I\!R^{\bar{P}}. \tag{4.4}$$

Analog zum generellen funktionalen Approximator f aus Gleichung (4.2) werden die Teilabbildungen des funktionalen Approximators $\bar{f}$ mit

$$\bar{f}_\omega(x) := (\bar{\mu}_{\omega_1}(x), \bar{A}_{\omega_2}(x), \bar{v}_{\omega_3}(x))^T$$

[1] Bei Eindeutigkeit der Bezeichnung wird τ schlicht Transformation genannt.

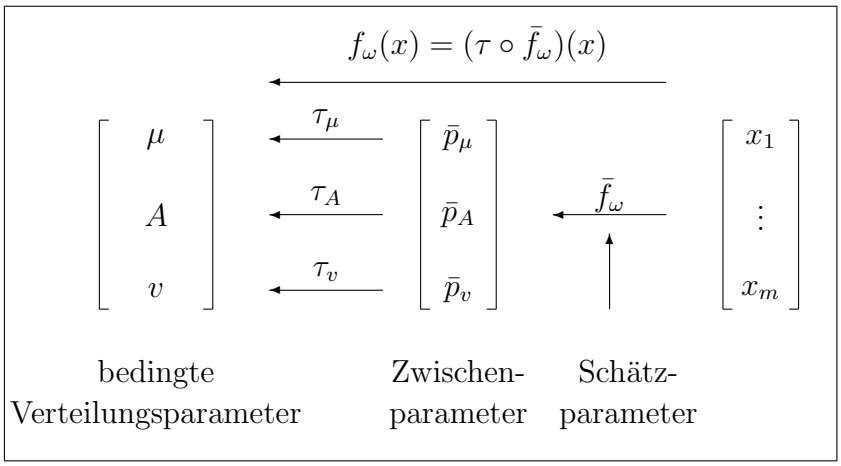


Abbildung 4.1: Prognosekonzept der bedingten multivariaten Wahrscheinlichkeitsverteilungen

bezeichnet. Die konkreten Teilabbildungen ergeben sich wie folgt:

- Der erste Teil des Approximators $\bar{f}_\omega$

$$\bar{\mu}_{\omega_1}(x) = \bar{p}_\mu : I\!R^m \rightarrow I\!R^n \tag{4.5}$$

  erzeugt einen n-dimensionalen Zwischenparametervektor, der über die Transformation τ_μ in den Lokationsparametervektor μ überführt wird. $\bar{p}_\mu$ wird als *Lokations-Zwischenparametervektor* bezeichnet.

- Der zweite und wesentliche Teil der Funktion $\bar{f}_\omega(x)$, dargestellt als

$$\bar{A}_{\omega_2}(x) = \bar{p}_A : I\!R^m \rightarrow I\!R^{n \times n}, \tag{4.6}$$

  erzeugt Funktionswerte, die über die Transformation τ_A mit der Strukturmatrix A korrespondieren. $\bar{p}_A$ wird aus diesem Grund *Vektor der Struktur-Zwischenparameter* genannt. Es sei an dieser Stelle vorweggenommen, dass die Annahme einer oberen Dreiecksmatrix keine Einschränkungen für die Erzeugung der betrachteten Verteilungsklassen mit sich bringt. Diese Behauptung wird zu Beginn des Abschnitts 4.3

gezeigt. Durch diese Vereinfachung ergibt sich die Funktion (4.6) zu

$$\bar{A}_{\omega_2}(x) = \bar{p}_A : I\!R^m \rightarrow I\!R^{1/2n(n+1)}. \tag{4.7}$$

- Bei der Existenz von Formparametern in der kanonischen Dichtefunktion ergibt sich der dritte Teil des funktionalen Approximators $\bar{f}$ zu:

$$\bar{v}_{\omega_3}(x) = \bar{p}_v : I\!R^n \rightarrow I\!R^u, \tag{4.8}$$

  wobei der Bildbereich $I\!R^u$ eine niedrige Dimension besitzt. Die Funktionswerte $\bar{p}_v$ repräsentieren Zwischenparameter, die mit Hilfe der Transformation τ_v in die Formparameter der Wahrscheinlichkeitsverteilung überführt werden. Die Komponenten von $\bar{p}_v$ heißen *Form-Zwischenparameter*.

In den Abschnitten 4.1 und 4.2 werden einerseits die lineare Approximation und andererseits die neuronalen Netze als adäquate funktionale Approximatoren $\bar{f}$ vorgestellt und ihre Eigenschaften beschrieben. Die Herleitung der technischen Transformation τ, um den gestellten Anforderungen an die Verteilungsparameter gerecht zu werden, wird in Abschnitt 4.3 diskutiert.

Die Komponenten des freien Parametervektors ω der Funktionen $f_\omega(x)$ bzw. $\bar{f}_\omega(x)$, den es durch numerische Optimierung zu bestimmen gilt, werden als *Schätzparameter* bezeichnet. Die verwendete Schätzmethode wird in Kapitel 5 hergeleitet und beschrieben.

4.1 Lineare Approximation

Eine analytisch interpretierbare Form für funktionale Approximatoren stellt die Klasse der linearen Abbildungen dar. Der lineare Approximator $\bar{f}_\omega(x) : I\!R^m \rightarrow I\!R^{\bar{P}}$ lässt sich wie folgt notieren:

$$\bar{f}_\omega(x) := Cx + c,$$

mit

$$C = \begin{pmatrix} \omega_{11} & \dots & \omega_{1m} \\ \vdots & \ddots & \vdots \\ \omega_{\bar{P}1} & \dots & \omega_{\bar{P}m} \end{pmatrix} \quad \text{und} \quad c = \begin{pmatrix} \omega_{\bar{P}m+1} \\ \vdots \\ \omega_{\bar{P}m+\bar{P}} \end{pmatrix}.$$

Da sich die Ausgänge als Linearkombinationen der unabhängigen Einflüsse ergeben, lassen sich die funktionalen Zusammenhänge analysieren und interpretieren. In Kapitel 4.3 wird sich zeigen, dass die Lokations-Zwischenparameter direkt als Lageparameter Verwendung finden können, d.h. die Transformation durch die identische Abbildung realisiert wird. Aus diesem Grund lassen sich speziell für lineare Modelle die prognostizierten Lageparameter bzw. bedingten Erwartungswerte sehr einfach und offensichtlich in Abhängigkeit der Inputgrößen interpretieren, was besonders für betriebswirtschaftliche Fragestellungen von Vorteil sein kann.

Verwendet man eine lineare Abbildung zur Approximation der Beziehung zwischen den Eingangsgrößen und den Zwischenparametern, so ergibt sich der Optimierungsraum der Schätzparameter ω zu

$$\omega \in I\!R^{\bar{P}(m+1)}. \tag{4.9}$$

Das Gesamtkonzept unter Verwendung dieser Funktionsform ist in Abbildung 4.2 illustriert.

Formal ergeben sich bei einer Verwendung von linearen Approximatoren die Zwischenparameter als:

$$\bar{p}_i(x) = \sum_{j=1}^{m} \omega_{ij} x_j + \omega_{\bar{P}m+i}, \quad i = 1, \dots, \bar{P}. \tag{4.10}$$

Da der berechtigte Anspruch gestellt werden sollte, dass ein Prognosekonzept ebenfalls für nichtlineare Zusammenhänge zwischen unabhängigen Inputgrößen und den Prognosewerten geeignet sein sollte, werden in Abschnitt 4.2 neuronale Netze mit ihrer Eigenschaft als universelle Approximatoren vorgestellt und beschrieben.

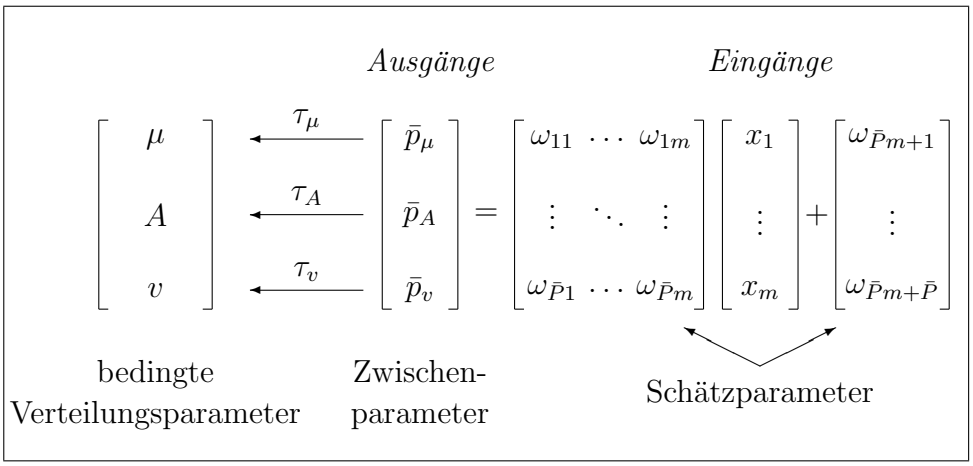


Abbildung 4.2: Lineare Abbildungen als funktionale Approximatoren

4.2 Neuronale Netze

In den letzten Jahren haben sich neuronale Netze als leistungsstarke Methode für eine Vielzahl von Anwendungen bewährt. Sie werden mittlerweile erfolgreich in den verschiedensten Bereichen der Ökonomie, der Physik oder der Medizin eingesetzt. Einleitend wird ein kritischer Aspekt zum Themengebiet der neuronalen Netze formuliert, um den Mythos der „black box“ aus dieser Arbeit zu vertreiben.[2]

Die Entwicklung neuronaler Netze entsprang dem Versuch, die Leistungsfähigkeit biologischer Nervensysteme nachzubilden, um die Schranken der sequentiell arbeitenden Rechner zu durchbrechen. In Anlehnung an das biologische Vorbild Gehirn wurde eine Vielzahl einfacher Recheneinheiten miteinander verbunden - in der Hoffnung, komplexe Phänomene wie „Intelligenz“ oder „Lernfähigkeit“ nachbilden zu können.

Solche dem Menschlichen entliehene Attribute sind dafür verantwortlich, dass neuronale Netze häufig als Verfahren angepriesen werden, welche in der Lage sind, selbständig Aufgaben zu lösen, die sogar von Experten nur unzureichend bewältigt werden können.

Die immer wieder formulierten Erwartungen, wie „neuronale Netze sind

[2]Vgl. die Sichtweise in (Anders, 1996).

intelligent, können selbständig generalisieren, sind fehlertolerant oder sind herkömmlichen Verfahren grundsätzlich überlegen“ sind sicherlich überhöht.

Ein wirklich intelligentes Verhalten, wie das Erlernen und eigenständige Abstrahieren eines Zusammenhangs, können neuronale Netze schon allein wegen der dazu benötigten Komplexität bei weitem nicht erlangen.

Damit stellen sich nun die Fragen, was neuronale Netze tatsächlich leisten können, wann sich ihr Einsatz lohnt und was man von ihnen realistischerweise erwarten darf.

Es ist gerechtfertigt, neuronale Netze als Methode zur Approximation nichtlinearer funktionaler Zusammenhänge zu verwenden. Eine Begründung lässt sich infolge dessen formulieren, da sich jedes feed-forward Netz durch eine Funktion der abhängigen Variablen und den Netzwerkgewichten beschreiben lässt. Spätestens mit dieser Darstellung wird offensichtlich, dass neuronale Netzwerke keine wirkliche oder auch nur nachgebildete Intelligenz besitzen können. Sie sind nichts anderes als nichtlineare Funktionen.

Aus oben genannten Gründen findet die Verwendung neuronaler Netze im vorliegenden Zusammenhang ihre Rechtfertigung. Die in dieser Arbeit betrachtete Problemstellung erhebt einzig und allein den Anspruch auf Funktionsapproximation und Generalisierungsfähigkeit nach adäquatem Training.[3]

Im Folgenden wird auf die Entstehung, die mathematische Modellierung und wesentliche Eigenschaften neuronaler Netze eingegangen. Abschließend stellt die Einbettung der neuronalen Netze in das Gesamtkonzept der Verteilungsprognose den inhaltlichen Zusammenhang her.

4.2.1 Das mathematische Modell

Inspiriert durch die Neurophysiologie wird, durch Abstraktion der Struktur des Nervensystems zu einem stark vereinfachten Modell, versucht, die Funktionsweise dieses Systems nachzubilden.

Es wurden Modelle konstruiert, die aus einzelnen Bausteinen, den so genannten Neuronen, bestehen. Die nachempfundenen Nervenzellen geben Im-

[3]Diese Eigenschaft der universellen Approximationsfähigkeit stetiger Funktionen ist Inhalt der beiden in Abschnitt 4.2.3 aufgeführten Sätze 4.2.1 und 4.2.2.

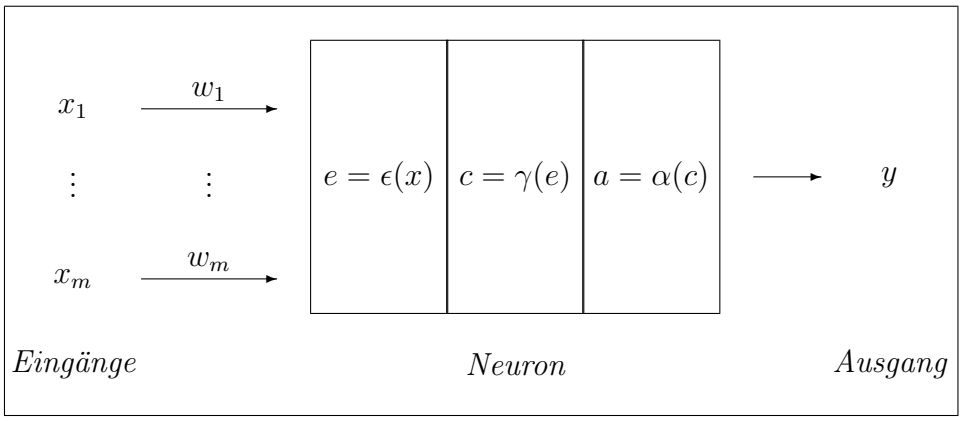


Abbildung 4.3: Die innere Struktur eines künstlichen Neurons

pulse weiter, falls die ankommenden Informationen ausreichend ausgeprägt sind, d.h. einen vorgegebenen Schwellenwert überschreiten. Die innere Struktur eines künstlichen Neurons ist in Abbildung 4.3 graphisch verdeutlicht.[4]

Die Verarbeitungseinheit „Neuron“ in einem künstlichen neuronalen Netz ist also die formalisierte Entsprechung der Sinneszelle in biologischen Systemen und kann durch die Angabe des Aktivierungszustands, der verwendeten Propagierungsfunktion, der Aktivierungsfunktion und der Ausgabefunktion vollständig beschrieben werden.

Das künstliche Neuron besitzt m Eingänge $x_1, \ldots, x_m$, wobei es die Informationen entweder von anderen Neuronen, den sogenannten Vorgängerneuronen, erhält oder Impulse direkt von der Umwelt erlangt. Den m Verbindungen zwischen den Vorgängerneuronen und dem Empfängerneuron sind Gewichte $w_1, \ldots, w_m$ zugeordnet.

Hieraus ergibt sich der effektive Eingangswert e, der sich über die Propagierungsfunktion oder Eingabefunktion ϵ berechnet.

Die *Propagierungsfunktion* beschreibt also das Verhalten eines Neurons im Hinblick auf seine Informationsverarbeitung. In der Regel werden die Eingaben der vorgeschalteten Neuronen entsprechend den Gewichtungen der Eingangskanten aufsummiert. Es ist jede beliebige Funktion als Eingabefunktion zur Berechnung der effektiven Eingabe $e = \epsilon(x)$ möglich. So auch die folgen-

[4] Vgl. (Stützle, 1999).

den vier naheliegenden Beispiele:

- Summation der gewichteten Eingänge: $e = \sum_{i=1}^{m} \omega_i x_i$
- Maximalwert der gewichteten Eingänge: $e = \max_{i=1,\ldots,m} \omega_i x_i$
- Produkt der gewichteten Eingänge: $e = \prod_{i=1}^{m} \omega_i x_i$
- Minimalwert der gewichteten Eingänge: $e = \min_{i=1,\ldots,m} \omega_i x_i$

Der *Aktivierungszustand* $c(t)$ definiert den aktuellen Zustand eines Neurons und ergibt sich aus dem eben errechneten effektiven Eingang mittels der Aktivierungsfunktion.

Die Aktivierungsfunktion γ legt fest, wie ein Aktivierungszustand $c(t-1)$ zum Zeitpunkt $t-1$ in einen Aktivierungszustand $c(t)$ zum Zeitpunkt t überführt wird, unter Berücksichtigung eventueller Einwirkungen auf das Neuron. Diese Transformation lässt sich also formal schreiben als:

$$c(t) = \gamma(e) = \gamma(\epsilon(x)). \tag{4.11}$$

In der Praxis relevant sind die folgenden drei Aktivierungsfunktionsklassen:

- die linearen Aktivierungsfunktionen,
- die Schwellenwertfunktionen und
- die sigmoiden Funktionen.

Es sei an dieser Stelle noch eine weitere bekannte Klasse von Aktivierungsfunktionen erwähnt, die so genannten *radial basis Funktionen*. Dies sind Funktionen der Form:

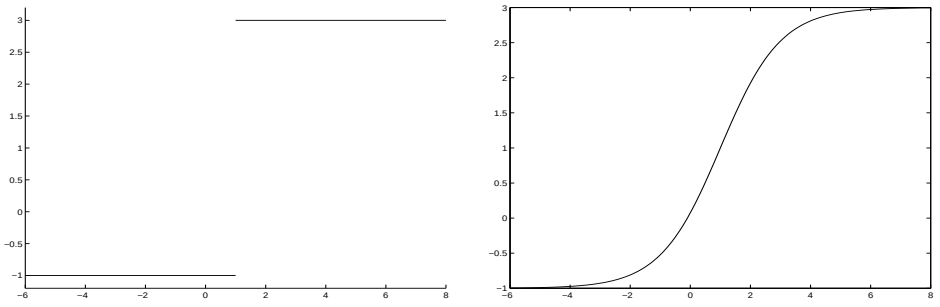
$$b(x) = h(\|x - c\|) = h\left(\sqrt{\sum_{i=1}^{n}(x_i - c_i)^2}\right),$$

wobei $\|\cdot\|$ die Euklidische Norm, c das Zentrum der radial basis Funktion und h eine stetige Funktion bezeichnen.[5]

[5] Zur Beschreibung einiger Charakteristika bzw. Vor- und Nachteile sei etwa auf (Hrycej, 1997) verwiesen.

Die konkrete Auswahl der Aktivierungsfunktion ist netzspezifisch bzw. hängt von dem konkret vorliegenden Anwendungsproblem ab.

Im einfachsten Fall werden lineare Aktivierungsfunktionen verwendet. Man benutzt diese Klasse von Aktivierungsfunktionen sinnvollerweise nur dann, wenn das Netz keine Zwischenschichten aufweist und die Abhängigkeit zwischen Ein- und Ausgabedaten linear ist bzw. hinreichend gut linear approximiert werden kann. Es ist induktiv leicht zu zeigen, dass mehrschichtige Netzwerke mit linearen Aktivierungsfunktionen in ein einstufiges Netzwerk überführt werden können. Das Produkt der Zwischenschichtmatrizen ergibt die Gewichtungsmatrix des äquivalenten einstufigen Netzwerkes.



(a) Schwellenwertfunktion mit $\Theta = 1$, $m = -1$, $M = 3$

(b) Allgemeine Fermi-Funktion mit $\Theta = 1$, $m = -1$, $M = 3$ und $\rho = 1/4$

Abbildung 4.4: Vergleich einer unstetigen Schwellenwertfunktion mit einer stetigen Fermi-Funktion

Die natürlich motivierte Aktivierungsfunktion ist sicherlich die in Abbildung 4.4(a) dargestellte Schwellenwertfunktion. Lässt man für den Ausgang nur die Werte „0“ (Ruhezustand) und „1“ (Feuern) zu, so lautet die Aktivierungsfunktion:

$$\gamma_\Theta(e) = \begin{cases} 1, & \text{falls } e > \Theta \\ 0, & \text{falls } e \leq \Theta. \end{cases} \tag{4.12}$$

Die Zahl Θ bezeichnet die Schwelle. Andere Werte sind als Neuronenausgänge ebenso zulässig. Die allgemeine Form der Schwellenwertfunktion ergibt sich

daher als:

$$\gamma_\Theta(e) = \begin{cases} M, & \text{falls } e > \Theta \\ m, & \text{falls } e \leq \Theta, \end{cases}$$

wobei m den minimalen und M den maximalen Funktionswert darstellt.

Der Nachteil dieser Funktion ist allerdings die Unstetigkeit an der Stelle Θ. Da die Schwellenwertfunktion also nicht auf dem ganzen Definitionsbereich differenzierbar ist, verbietet sich ihre Anwendung in sämtlichen Optimierungsverfahren, welche die erste oder höhere Ableitungen beinhalten.

Die in der Praxis am häufigsten verwendeten Aktivierungsfunktionen sind die aus der Klasse der sigmoiden Funktionen. Sie spielen eine große Rolle bei der Modellierung nichtlinearer Zusammenhänge zwischen Ein- und Ausgabedaten. Zudem erfüllen sie die Anforderungen der Stetigkeit, der Differenzierbarkeit und der Monotonie auf dem ganzen Definitionsbereich. In Analogie zu obiger Schwellenwertfunktion sei die allgemeine Fermi-Funktion, dargestellt in Abbildung 4.4(b), als Beispiel einer sigmoiden Funktion genannt. Ein weiterer Parameter ρ charakterisiert die Steigung der Fermi-Funktion. Offensichtlich nähert sich also die Fermi-Funktion für größer werdende ρ der Schwellenwertfunktion. Formal ergibt sich die allgemeine Fermi-Funktion zu

$$\gamma_{\Theta,m,M}(e) = m + \frac{M-m}{1 + exp(-4\rho\frac{e-\Theta}{M-m})}. \tag{4.13}$$

Der Tangenshyperbolicus wird ebenfalls häufig in der Literatur als Beispiel für eine logistische Aktivierungsfunktion angegeben. Dies ist jedoch ein Sonderfall der allgemeinen Fermi-Funktion (4.13) mit $\Theta = 0$, $m = -1$, $M = 1$ und $\rho = 1$.

Die *Ausgabefunktion* α legt den Wert y eines Neurons in Abhängigkeit des aktuellen Aktivierungszustands $c(t)$ fest. In der Regel wird hier jedoch die Identität verwendet, da die nötigen Charakteristika über die Aktivierungsfunktion modelliert werden können:

$$\alpha(c) = y. \tag{4.14}$$

Zusammengefasst lässt sich schließlich der Ausgabewert y in Abhängigkeit

von der Eingabe x mit Hilfe der Gleichungen (4.11) und (4.14) wie folgt formulieren:

$$y = \alpha(c) = \alpha(\gamma(e)) = \alpha(\gamma(\epsilon(x))).$$

4.2.2 Feed-forward Netzwerke

Da in dieser Arbeit ausschließlich feed-forward Netzwerke zur Anwendung kommen, wird in diesem Abschnitt diese Klasse von neuronalen Netzen gesondert betrachtet. Der Grund für die Auswahl dieser Art von neuronalen Netzen ist ihre Eigenschaft, universeller Approximator von stetigen Funktionen zu sein. Dies wird in Abschnitt 4.2.3 erläutert.

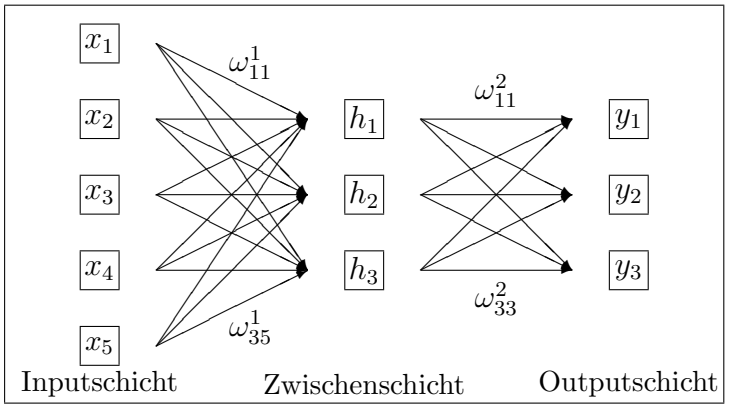


Abbildung 4.5: Ein 5-3-3 feed-forward Netz

Bei einem feed-forward Netzwerk, das in Abbildung 4.5 beispielhaft dargestellt ist, sind die Neuronen schichtweise angelegt. Die Neuronen direkt aufeinander folgender Schichten sind miteinander verbunden, d.h. es gibt keine Verbindungen zwischen entfernteren Schichten oder innerhalb einer Schicht. Eine weitere Eigenschaft von feed-forward Netzwerken ist das Versenden von Signalen in nur eine Richtung im Netz, nämlich von der Eingangs- zur Ausgangsschicht.

Um einige Eigenschaften der feed-forward Netze mathematisch zu beschreiben, wird die folgende Notation eingeführt. Es sei

l	:	Anzahl der Schichten
n_k	:	Anzahl der Neuronen in der k-ten Schicht
Θ_i^k	:	Grundaktivierung des i-ten Neurons in der k-ten Schicht
y_i^k	:	Ausgangsaktivierung des i-ten Neurons in der k-ten Schicht
ω_{ij}^k	:	Gewicht der Verbindung von Neuron j in der $(k-1)$-ten Schicht zu Neuron i in der k-ten Schicht,

wobei die Größen Θ_i^k, y_i^k und ω_{ij}^k folgende Dimension besitzen:

$$\Theta^k \in I\!R^{n_k} \quad k = 1, \ldots, l$$

$$y_i^k \in I\!R^{n_k} \quad k = 1, \ldots, l$$

$$(\omega_{ij}^k)_{i=1,\ldots,n_k\,;\,j=1,\ldots,n_{k-1}} =: W^k \in I\!R^{n_k \times n_{k-1}} \quad k = 1, \ldots, l$$

Nachfolgend wird der funktionale Zusammenhang zwischen der Eingangs- und der Ausgangsschicht formal beschrieben. Der Einfachheit halber wird die Summation der gewichteten Eingänge als Propagierungsfunktion verwendet.

Es zeigt sich, dass speziell diese Art von neuronalen Netzen Vektorfunktionen darstellen, da der Ausgangsvektor y^l nur in Abhängigkeit des Eingangsvektors y^0 erklärbar ist. Um dies nun formal zu begründen, wird in einem ersten Schritt durch Gleichung (4.15) die Berechnung der Ausgangsaktivierung von Neuron i in der k-ten Schicht durch Zuflüsse der Neuronen aus der $(k-1)$-ten Schicht formuliert.

$$y_i^k = \gamma_i\Big(\sum_{j=1}^{n_{k-1}} \omega_{ij}^k y_j^{k-1} + \Theta_i^k\Big), \quad i = 1, \ldots, n_k, \quad k = 1, \ldots, l \tag{4.15}$$

In Matrixschreibweise ergibt sich Gleichung (4.15) zu:

$$y^k = \gamma(W^k y^{k-1} + \Theta^k), \quad k = 1, \ldots, l, \tag{4.16}$$

wobei die Aktivierungsfunktion wie folgt abbildet:

$$\gamma : I\!R^{n_{k-1}} \to I\!R^{n_k}.$$

Um schließlich die Ausgangsaktivierungen y^k der Neuronen in der k-ten

Schicht in Abhängigkeit von den Impulsen der Eingangsschicht y^0 zu formulieren, kann nach Gleichung (4.16) folgender rekursiver Zusammenhang entwickelt werden:

$$\begin{aligned} y^k &= \gamma(W^k y^{k-1} + \Theta^k) \\ &= \gamma(W^k \gamma(W^{k-1} y^{k-2} + \Theta^{k-1}) + \Theta^k) \\ &\vdots \\ &= \gamma(W^k \gamma(W^{k-1} \gamma(W^{k-2} \dots \gamma(W^1 y^0 + \Theta^1) \cdots + \Theta^{k-2}) + \Theta^{k-1}) + \Theta^k) \end{aligned} \tag{4.17}$$

Bemerkung 1 (Die Funktion der Zwischenschicht)
Bevor in Abschnitt 4.2.3 speziell auf die Eigenschaft eingegangen wird, dass ein Multi-Layer Perzeptron ein universeller Approximator für beliebige stetige Funktionen ist, sollte zuvor das Augenmerk auf die Zwischenschicht gerichtet werden. Es stellt sich die Frage, wie sich die Wahl der Neuronenanzahl in den Zwischenschichten auf die Approximationseigenschaft des neuronalen Netzes auswirkt. In (Scherer, 1997) wird das Beispiel der Approximation einer kubischen Grundfunktion aufgeführt. Aufgrund von leichten stochastischen Störungen können jedoch im Einzelfall mehr oder weniger starke Abweichungen in den Daten vorhanden sein. Kennt man wie in diesem Beispiel die zugrunde liegende Funktionsart, so ist klar, dass ein kubisches Polynom die beste Approximation beschreibt. Es muss daher das Polynom nur noch durch „geschickte“ Wahl der Koeffizienten optimal justiert werden.

Wählt man jedoch ein Polynom höheren Grades, so beschreibt man immer mehr das stochastische Rauschen, d.h., das mögliche Oszillieren von Polynomen höheren Grades stellt sich aufgrund der verrauschten Daten ein und man verliert sowohl den funktionalen Zusammenhang als auch die Generalisierungsfähigkeit.

Einen ähnlichen Effekt, der Overfitting *genannt wird, hat die Erhöhung der Neuronenanzahl in den Zwischenschichten zur Folge. Zwar erhöht sich mit steigender Anzahl auch die Flexibilität und die Anpassungsgüte für die Trainingsdatensätze, jedoch geht durch Modellieren des stochastischen Rauschens die Generalisierungsfähigkeit verloren.*

Aus obigen Gründen muss zu Beginn der Modellierung die zu wählende

Architektur des neuronalen Netzes an die Datengrundlage und die praktische Problemstellung sinnvoll angepasst werden. Auch aus numerischer Sicht kann dies entscheidend zur Verbesserung der Modellgüte beitragen, da sich bei jeder Erhöhung der Neuronenanzahl der Rechenaufwand erheblich erhöhen kann. Bei Hinzunahme eines Neurons erhöht sich die Dimension der Optimierungsaufgabe um die Anzahl der Verbindungen, die in dem Neuron ankommen zuzüglich der vom Neuron abgehenden.

4.2.3 Das Multi-Layer Perzeptron als universeller Approximator

Das Multi-Layer Perzeptron[6] ist ein spezielles neuronales Netz aus der Klasse der feed-forward Netze, die in Abschnitt 4.2.2 eingeführt wurden. Charakterisiert ist das Multi-Layer Perzeptron durch eine spezielle Aktivierungsfunktion, die bekannte Fermi-Funktion des Abschnitts 4.2.1.

Im Detail ist ein Multi-Layer Perzeptron definiert als feed-forward Netzwerk mit

- Propagierungsfunktion: Summation der gewichteten Eingänge

- Aktivierungsfunktion: Fermi-Funktion (häufig mit $m = 0$ und $M = 1$)

- Ausgabefunktion: Identität.

Die rekursive Herleitung der Gleichungen (4.17) zeigt, dass ein Multi-Layer Perzeptron bei fest gewählten Gewichten eine reellwertige Vektorfunktion darstellt.

Als Beispiel sei im Folgenden die funktionale Form eines Multi-Layer Perzeptrons mit einer Zwischenschicht ($l = 1$) und den Parametern $M = 1$, $m = 0$, $\Theta = 0$ und $\rho = 1/4$ der Fermi-Funktion gewählt. Es ergibt sich eine geschlossene Darstellung der Ausgangsneuronen in Abhängigkeit der

[6] Siehe (Rosenblatt, 1962).

Eingänge wie folgt:

$$y_{i_3} = f_{i_3}^{NN}(x) = \sum_{i_2=1}^{n_2} \omega_{i_3 i_2}^2 \frac{1}{1+\exp(-\sum_{i_1=1}^{n_1} \omega_{i_2 i_1}^1 x_{i_1})}, \quad i_3 = 1, \ldots, n_3.$$

Funahashi zeigt, dass jede stetige Funktion durch ein Multi-Layer Perzeptron mit mindestens einer Zwischenschicht und sigmoiden Aktivierungsfunktionen beliebig genau approximiert werden kann.[7] Die präzise Formulierung dieser Behauptung für den univariaten Fall mit einer versteckten Schicht ist im folgenden Satz 4.2.1 dargestellt. Für die Verallgemeinerung auf vektorwertige Funktionen und mehrschichtige Netzwerke sei auf den Satz 4.2.2 verwiesen. Der induktive Beweis des Satzes 4.2.2 ist anschließend vervollständigend aufgeführt.[8]

Wie bereits erwähnt, werden der Einfachheit halber zunächst reellwertige Funktionen betrachtet, in denen das neuronale Netz ein einziges Neuron in der Ausgangsschicht besitzt. Zur weiteren Vereinfachung der Notation werden Gewichte unterschiedlicher Schichten mit eindeutigen Bezeichnungen v und w gekennzeichnet.

Satz 4.2.1
Sei $\gamma : \mathbb{R}^n \to \mathbb{R}$ eine nicht konstante, beschränkte und monoton wachsende Aktivierungsfunktion. Des Weiteren sei $K \subset \mathbb{R}^n$ kompakt und $f : K \to \mathbb{R}$ stetig.
Dann existieren für ein beliebiges $\epsilon > 0$ eine natürliche Zahl N_2 und reelle Konstanten v_i, Θ_i $(i = 1, \ldots, N_2)$ und w_{ij} $(i = 1, \ldots, N_2;\ j = 1, \ldots, n)$, so dass

$$\tilde{f}(x) := f^{NN}(x) = \sum_{i=1}^{N_2} v_i \gamma(\sum_{j=1}^{n} w_{ij} x_j - \Theta_i)$$

[7] Siehe (Funahashi, 1989). Ein weiterer Beweis zu dieser Behauptung wurde zur gleichen Zeit in (Hornik et al., 1989) geliefert.

[8] Der allgemeine Fall für eine beliebige Anzahl von versteckten Schichten kann sodann über das Induktionsprinzip bewiesen werden. In (Funahashi, 1989) wird zudem für den Spezialfall eines 4-Schicht Perzeptrons ein alternativer Beweis über den Satz von Kolmogorov-Arnold-Sprecher angegeben, auf den hier verzichtet wird.

folgende Ungleichung erfüllt:

$$\max_{x \in K} \|f(x) - \tilde{f}(x)\| < \epsilon.$$

Die Euklidische Norm von $y \in \mathbb{R}^n$ ist definiert durch $\|y\| := \sqrt{\sum_{i=1}^{n} y_i^2}$.

Auf den ausführlichen Beweis dieser Aussage sei an dieser Stelle verzichtet und auf den Originaltext verwiesen.[9]

Da sich die Formulierung des Satzes 4.2.1 auf reellwertige Funktionen beschränkt, wird in Satz 4.2.2 eine Verallgemeinerung zu vektorwertigen Funktionen angegeben.[10]

Satz 4.2.2 (Eine Verallgemeinerung zu Satz 4.2.1)
Sei $\gamma : \mathbb{R}^n \to \mathbb{R}$ eine nicht konstante, beschränkte und monoton wachsende stetige Funktion. $K \subset \mathbb{R}^n$ sei kompakt und $l \geq 3$ ($l \in \mathbb{N}$) beliebig, aber fest. Dann existiert für jede stetige Funktion $f : K \to \mathbb{R}^m$ und ein beliebiges $\epsilon > 0$ ein l-Schicht Netzwerk[11], dessen Input-Output Zusammenhang gegeben ist durch $\tilde{f} : K \to \mathbb{R}^m$, die sigmoide bzw. lineare Aktivierungsfunktion der Zwischenschichten bzw. der Eingangs- und Ausgangsschichten, derart dass

$$\max_{x \in K} \|f(x) - \tilde{f}(x)\| < \epsilon.$$

Neben den bereits aus (Funahashi, 1989) und (Hornik et al., 1989) zitierten Beweisen, existieren eine Reihe von alternativen Veröffentlichungen. So kann die Behauptung aus Satz 4.2.1 über das Argument der Untersummen, die das Integral der zu approximierenden Funktion annähern verifiziert werden.[12] Dabei wird entscheidend ausgenutzt, dass abzählbar unendlich viele Stützstellen zur Verfügung stehen. Es ist bei allen Argumentationsweisen in-

[9] Vgl. (Funahashi, 1989). Eine skizzenhafte Darstellung mit den wesentlichen Argumentationsschritten findet sich etwa in (Stützle, 1999).

[10] Dieser Beweis wurde ebenfalls in (Funahashi, 1989) durch Anwendung obigen Satzes 4.2.1 geliefert.

[11] *Vgl. Gleichung (4.17).*

[12] Vgl. etwa (Kruse et al., 1996).

tuitiv einsichtig, dass die Beweisführung zum Erfolg führt, was in Bemerkung 2 (s.u.) deutlich wird.

Abschließend folgen einige Anmerkungen zu der universellen Approximationsfähigkeit. Es werden Eigenschaften des Multi-Layer Perzeptrons aufgeführt, um die Verwendung als funktionaler Approximator zu verdeutlichen. Weiterhin sei auf etwaige Probleme hingewiesen, um möglichen Missverständnissen vorzubeugen.

Bemerkung 2

i) Häufig verwendete Aktivierungsfunktionen, wie die sigmoiden Funktionen[13]*, erfüllen die Voraussetzungen, nicht konstant, beschränkt, monoton wachsend und stetig zu sein.*

ii) Die obigen Darstellungen basieren im Wesentlichen auf dem Text (Funahashi, 1989). Parallel dazu wird in (Hornik et al., 1989) ein Beweis geliefert, der zwar für ein 3-Schicht Netzwerk formuliert ist, jedoch auch nicht stetige Aktivierungsfunktionen zulässt. Dennoch haben beide Texte keine expliziten Aussagen über die Anzahl der Neuronen in der/den Zwischenschichten formuliert. Hrycej bemerkt, dass Schranken bezüglich der Anzahl von Zwischenschichtneuronen noch nicht entdeckt wurden.[14] *Außerdem wird auf Erfahrungswerte hingewiesen, die die Anzahl der Neuronen in der versteckten Schicht auf* $\min[\max(n_0, n_l), 6]$ *datieren, wobei* n_0 *und* n_l *die Anzahl der Neuronen in der Eingangs- bzw. Ausgangsschicht sind. Dies entspricht der in dieser Arbeit angewendeten Strategie.*

iii) Es hat sich also gezeigt, dass schon 3-Schicht Netzwerke ausreichen, um jede stetige Funktion beliebig genau zu approximieren. Weiterhin wird in (Funahashi, 1989) darauf hingewiesen, dass es von Interesse sein könnte zu untersuchen, ob $(k > 3)$*-Schicht Netzwerke Approximatoren sind, die weniger Kosten verursachen, ob man also durch Hinzufügen einer oder mehrerer Zwischenschichten eine Reduktion der Neuronen*

[13] *Ein Beispiel einer sigmoiden Funktion ist etwa* $\gamma(x) = 1/(1 + e^{-x})$.
[14] *Siehe (Hrycej, 1997).*

bzw. der Verbindungen erreichen könnte.

iv) In den bislang bekannten Beweisen aus der Literatur werden beliebig viele Neuronen in den Zwischenschichten zugelassen. Vor allem der Beweis in (Kruse et al., 1996) verdeutlicht, dass die Eigenschaft solcher neuronaler Netze als universelle Approximatoren für stetige Funktionen intuitiv und klar ist. Dennoch ist die Aussage der Sätze 4.2.1 und 4.2.2 der wesentliche Grund dafür, dass neuronale Netze in einer Vielzahl von praktischen Anwendungen zum Einsatz kamen. Allerdings ist die Qualität der Modellierung essenziell von der verwendeten Optimierungsmethode[15] *abhängig.*

4.2.4 Multi-Layer Perzeptron bei bedingten Wahrscheinlichkeitsverteilungen

Die reellwertige Approximationsfunktion $\bar{f}_\omega$ aus Gleichung (4.4) bildet vom Raum der Inputvariablen $x \in \mathbb{R}^m$ auf die reellwertigen Zwischenparameter $\bar{p} \in \mathbb{R}^{1/2n(n+3)+u} = \mathbb{R}^{\bar{P}}$ ab, wobei u die Anzahl der Formparameter bezeichnet.

Die funktionale Approximation wird mit Hilfe eines multi-layer Perzeptrons, das n_2 versteckte Neuronen in der einzigen Zwischenschicht besitzt, vorgenommen. Die Einbettung der neuronalen Netze in das Gesamtkonzept ist in Abbildung 4.6 illustriert.

Die Dimension des Optimierungsraums der freien Schätzparameter, die durch numerische Methoden zu identifizieren sind, beläuft sich auf

$$\omega \in \mathbb{R}^{mn_2+n_2\bar{P}}. \tag{4.18}$$

Wählt man in der Praxis die Netzarchitektur mit einer maximalen Anzahl der versteckten Neuronen von

$$n_2 = \left[\frac{m\bar{P}+\bar{P}}{m+\bar{P}}\right],$$

[15] *Vgl. hierzu Kapitel 5.*

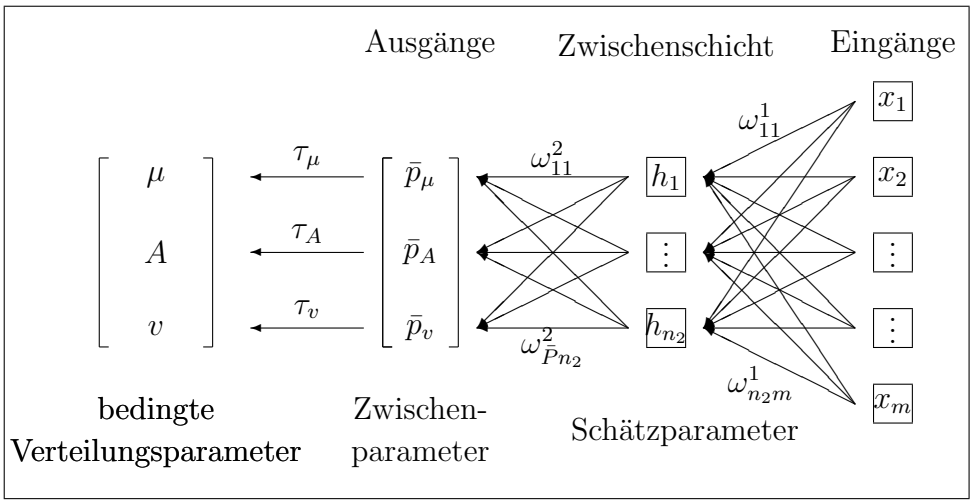


Abbildung 4.6: Feed-forward Netze als funktionale Approximatoren

wobei $[\,\cdot\,]$ die Gauß-Klammer bezeichnet, so überschreitet aufgrund der Beziehung zwischen Gleichung (4.9) und (4.18) die Dimension des Schätzparameterraums im Fall von neuronalen Netzen nie die von linearen Approximatoren aus Abschnitt 4.1, was jedoch keine Obergrenze darstellt. Für die Netzwerkarchitektur sei wiederholend auf Bemerkung 2 verwiesen.[16]

Formal ergibt sich formal das Multi-Layer Perzeptron mit dem gewichteten Durchschnitt als Propagierungsfunktion, der sigmoiden Fermi-Funktion als Aktivierungsfunktion, der Identität als Ausgabefunktion und n_2 versteckten Neuronen in der einzigen Zwischenschicht als:

$$\bar{p}_{i_3}(x) = \sum_{i_2=1}^{n_2} \omega^2_{i_3 i_2} \frac{1}{1+\exp(-\sum_{i_1=1}^{m} \omega^1_{i_2 i_1} x_{i_1})}, \quad i_3 = 1, \ldots, \bar{P}. \tag{4.19}$$

[16]Bei den in dieser Arbeit beschriebenen praktischen Anwendungen wurde die aus (Hrycej, 1997) zitierte Empfehlung $\min[\max(n_0, n_l), 6]$ verwendet.

4.3 Transformationen reeller Funktionswerte auf Parameter der Wahrscheinlichkeitsverteilungen

In den Abschnitten 4.1 und 4.2 wurden lineare Abbildungen und neuronale Netze als Optionen für den funktionalen Approximator $\bar{f}$ dargestellt. Da die Bildbereiche reellwertige Räume sind, die Parameter der Wahrscheinlichkeitsklassen jedoch gewissen Restriktionen unterliegen können, ist vielfach eine Transformation τ zur Gewährleistung dieser Einschränkungen erforderlich.

Inhalt des folgenden Abschnitts ist die Entwicklung und Formulierung der erforderlichen Transformationen τ, mittels der die mit Restriktionen behafteten Parameter der Wahrscheinlichkeitsverteilungen erzeugt werden können. Im Zusammenhang der Formulierung des ganzheitlichen Prognosekonzepts fand diese Abbildung, die als *technische Transformation* bezeichnet wird, bereits zu Beginn des Kapitels 4 Erwähnung.

Seien die Komponenten der Transformation τ, die auf die Zwischenparameter angewendet werden, wie folgt bezeichnet:

- Der Funktionswert von τ_{μ_i} mit $i = 1, \ldots, n$ korrespondiert zum Vektor der Lageparameter,

- der Funktionswert von $\tau_{\alpha_{ii}}$ mit $i = 1, \ldots, n$ korrespondiert zu den Diagonalelementen der Strukturmatrix A,

- der Funktionswert von $\tau_{\alpha_{ij}}$ mit $i > j = 1, \ldots, n$ korrespondiert zu den nichtdiagonalen Elementen der Strukturmatrix A und

- der Funktionswert von τ_{v_o} mit $o = 1, \ldots, u$ korrespondiert zu den Formparametern der Wahrscheinlichkeitsverteilung.

Die Indizierung gilt analog für den Vektor der Zwischenparameter $\bar{p}$. Diese Bezeichnungen sind konsistent mit den Darstellungen des Gesamtkonzepts der Abbildungen 4.1, 4.2 und 4.6.

4.3.1 Transformation der Lokations-Zwischenparameter

Die Lageparameter der betrachteten Verteilungen[17] unterliegen keinen spezifischen Restriktionen, d.h. der Vektor der Lokationsparameter μ ist beliebig reellwertig zulässig. Aus diesem Grund ist für die Funktionswerte $\bar{p}_\mu$ des funktionalen Approximators $\bar{\mu}_{\omega_1}(x)$ aus Gleichung (4.5) keine explizite zusätzliche Transformation erforderlich. Es ergibt sich daher:

$$\mu = \tau_\mu(\bar{p}_\mu) = \bar{p}_\mu. \tag{4.20}$$

Der Zwischenparametervektor $\bar{p}_\mu$ repräsentiert direkt den Lokationsvektor μ.[18] Die Transformation τ_μ in Gleichung (4.20) kann daher trivialerweise als identische Abbildung gewählt werden.

4.3.2 Transformation der Form-Zwischenparameter

Die Form-Zwischenparameter $\bar{p}_v$, die zu dem Vektor der Formparameter v korrespondieren, sind abhängig von der Verteilungsklasse zu transformieren. Ein beschränktes Intervall $[a; b]$ könnte etwa den Zulässigkeitsbereich eines Formparameters darstellen. Eine Komponente der Transformation $\tau_v(\bar{p}_v)$ würde folglich vom reellen Raum $\mathbb{R}$ in das beschriebene Intervall $v_o \in [a; b]$ abbilden. Mögliche Transformationen, die diese Eigenschaft erfüllen, sind die sigmoiden Funktionen.[19] Wählt man in diesem Beispiel

$$\tau_{v_o}(\bar{p}_{v_o}) = \frac{(b-a)}{1+e^{-\bar{p}_{v_o}}} + a, \tag{4.21}$$

so ergibt sich der gewünschte Wertebereich $[a; b]$ für die o-te Komponente des Formvektors v.

[17] Siehe Abschnitt 3.1 und Kapitel 6.

[18] Dies würde im Fall von Verteilungen, wie etwa der Beta Verteilung nicht zutreffen, für welche der Lokationsvektor einen beschränkten Zulässigkeitsbereich besitzt.

[19] Vgl. die Beschreibung der sigmoiden Funktionen in Abschnitt 4.2.1 und Abbildung 4.4(b).

Es seien nun einige Transformationen der wesentlichen Verteilungsklassen aufgezeigt.

Der Formparameter m der Klasse der t-Verteilungen: Um die Restriktion $m > 0$ für den Freiheitsgrad einer t-verteilten Zufallsvariablen zu erreichen, kann folgende Transformation für den reellen Zwischenparameter $\bar{p}_v$ vorgeschlagen werden:

$$m = \tau(\bar{p}_{v_m}) = e^{\bar{p}_{v_m}}.$$

Die Formparameter α, β der stabilen Verteilungsklasse: Betrachtet man den theoretischen Zulässigkeitsbereich der Formparameter von univariaten stabilen Verteilungen, so muss gewährleistet sein, dass $0 < \alpha < 2$ und $-1 < \beta < 1$ gilt. Die Restriktionen für α und β können jeweils über die sigmoide Funktion (4.21) mit $a = 0$ und $b = 2$ bzw. $a = -1$ und $b = 1$ erreicht werden.

Die Formparameter α, β und λ der generalisiert hyperbolischen Verteilungsklasse: Die theoretischen Bedingungen, die in der Familie der generalisiert hyperbolischen Verteilungen einzuhalten sind, lassen sich formal durch $\|\beta\| < 1$, $\alpha > 0$ und $\lambda \in I\!R$ ausdrücken. Um diesen Restriktionen Rechnung zu tragen, kann etwa die Bedingung

$$\tau(\bar{p}_{v_\beta}) := C_2 \prod_{i=1}^{n} \frac{\bar{p}_{v_{\beta_i}}}{\sqrt{\sum_{i=1}^{n} \bar{p}^2_{v_{\beta_i}}}} \left[\frac{2}{1 + \exp(-\sqrt{\sum_{i=1}^{n} \bar{p}^2_{v_{\beta_i}}})} - 1 \right] \in [0\ ;\ C_2]$$

für den Symmetrieparametervektor β Anwendung finden. C_2 bezeichnet hierbei eine Regulierungskonstante, die in der Praxis i.Allg. $C_2 < 1$ gewählt wird. Die Randbereiche des Parameterraums für die Parameterschätzung sind generell wenig relevant.[20]

Ferner sind aus ähnlichen praktischen Gründen zusätzlich schärfere Restriktionen an den Formparameter α und den Generalisierungsparameter λ geknüpft. Um die Ränder der Zulässigkeitsbereiche auszuschließen, können

[20] Vgl. etwa (Eberlein and Prause, 1999) und (Blæsild and Jensen, 1981). Es hat sich bei den hier implementierten Beispielen etwa $C_2 = 0,6$ bewährt.

etwa die Transformationen wie folgt definiert werden:

$$\lambda = \tau(\bar{p}_{v_\lambda}) := \frac{1}{2}\, n + \exp(\frac{4}{1+\exp(-\bar{p}_{v_\lambda})} - 2) \in [1/2n + e^{-2}\ ;\ \frac{1}{2}n + e^2]$$

und

$$\alpha = \tau(\bar{p}_{v_\alpha}) := \exp(\frac{4}{1+\exp(-\bar{p}_{v_\alpha})} - 2) \in [e^{-2}; e^2] = [0,13\ ;\ 7,39].$$

Die Formparameter $q_i, i = 1, \ldots, R$ der Gauß'schen Mixtur: Eine Verteilung aus der Klasse der Gauß'schen Mixturverteilungen besitzt R Formparameter q_i die der Bedingung

$$\sum_{i=1}^{R} q_i = 1 \quad \text{mit } q_i > 0,\ i = 1, \ldots, R$$

unterliegen. Diese Restriktionen können etwa durch die Transformation einer *softmax*-Funktion[21]

$$q_i = \frac{e^{\bar{p}_{v_i}}}{\sum_{r=1}^{R} e^{\bar{p}_{v_r}}} \tag{4.22}$$

erreicht werden. Für den hier betrachteten Fall einer binären Mixtur reicht allerdings die sigmoide Funktion mit $a = 0$ und $b = 1$ aus Gleichung (4.21), da $q_2 = 1 - q_1$ gilt.

Für Formparameter, die keinen Restriktionen unterliegen, verhalten sich die zugehörigen Transformationen analog zu denen der Lokations-Zwischenparameter und sind identische Abbildungen. Obige Transformationen sind in der Praxis bewährte und getestete Funktionen, die keinen Anspruch auf Allgemeingültigkeit besitzen.

[21] Vgl. etwa (Bridle, 1990) oder (Jacobs et al., 1991).

4.3.3 Transformationen der Struktur-Zwischenparameter

Dieser Abschnitt behandelt die Transformation der Funktionswerte des funktionalen Approximators $\bar{A}_{\omega_2}(x)$ aus Gleichung (4.6) bzw. (4.7), dem für die Strukturmatrix A zuständigen Teil des funktionalen Approximators.

Abhängig von der Form und den Eigenschaften der Strukturmatrix A sind die reellen Ausgänge des funktionalen Approximators transformiert.

Motiviert durch die Cholesky-Zerlegung wird $A \in I\!R^{n \times n}$ als obere Dreiecksmatrix mit positiven Diagonalelementen angenommen. Die Begründung und Rechtfertigung ergibt sich aus folgender Argumentation.

In einem ersten Schritt wird gezeigt, dass trotz der einschränkenden Annahme einer oberen Dreiecksmatrix die komplette elliptische Verteilungsfamilie über die lineare Transformation (3.2) erreicht werden kann. Ferner ist diese Annahme keine Einschränkung für die Generierung der hier betrachteten asymmetrischen Verteilungsklassen, wie der generalisiert hyperbolischen Verteilung.

Gleichung (3.7) zeigt, dass die betrachtete breite Familie der sphärischen Verteilungen im Wesentlichen den Zufallsvektor z in der Form $z^T z$ beinhaltet. Aufgrund des Zusammenhangs aus Gleichung (3.5) ergibt sich die quadratische Form:

$$z^T z = (y - \mu(x))^T A(x)^T A(x)(y - \mu(x)) > 0. \tag{4.23}$$

Da über die Cholesky-Zerlegung bekannt ist, dass jede positiv definite Matrix eine eindeutige Zerlegung der Form $A^T A$ besitzt, ist die ganze elliptische Verteilungsklasse erreicht.

Auch für die asymmetrischen Verteilungsfamilien, wie der generalisiert hyperbolischen Verteilung aus Abschnitt 6.3, die Terme wie $\beta^T z$ beinhalten, bleibt das Konzept gültig. Der Schiefeparameter β wird lediglich in transformierter Form

$$\beta^T A(x)(y - \mu(x)) := \hat{\beta}^T (y - \mu(x)) \tag{4.24}$$

geschätzt, so dass auch hier jedes beliebige $\hat{\beta} \in I\!R^n$ repräsentiert wird, um die ganze Verteilungsfamilie darzustellen. Durch die Rücktransformati-

on $\beta = A^{-T}\hat{\beta}$ kann bei Bedarf auf den ursprünglichen Verteilungsparameter geschlossen werden.

Wie Gleichung (4.24) zeigt, behält die Vorgehensweise aus Abschnitt 3.1 für ausgewählte asymmetrische Verteilungsklassen, wie generalisierte hyperbolische Verteilungen, ebenfalls seine Gültigkeit.

Mit dem Ziel der Trennung von Skalierungsparametern und Elementen, die reine Strukturparameter repräsentieren, wird die Skalierungsmatrix A durch

$$A = UD \tag{4.25}$$

faktorisiert.[22] Hierbei ist U eine obere Dreiecksmatrix mit $u_{ii} = 1$, $i = 1, \ldots, n$ und D besitzt Diagonalgestalt.

Für den dreidimensionalen Fall ($n = 3$) bedeutet dies:

$$UD = \begin{pmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} d_{11} & 0 & 0 \\ 0 & d_{22} & 0 \\ 0 & 0 & d_{33} \end{pmatrix} = \begin{pmatrix} d_{11} & d_{22}u_{12} & d_{33}u_{13} \\ 0 & d_{22} & d_{33}u_{23} \\ 0 & 0 & d_{33} \end{pmatrix}.$$

Wird auf diese weitere Zerlegung verzichtet, so ergibt sich die Kovarianzmatrix $\Sigma = (A^T A)^{-1}$ komponentenweise als:

$$\begin{aligned}
\sigma_{11} &= \frac{\alpha_{12}^2\alpha_{23}^2 + \alpha_{12}^2\alpha_{33}^2 + \alpha_{22}^2\alpha_{13}^2 + \alpha_{22}^2\alpha_{33}^2 - 2\alpha_{12}\alpha_{13}\alpha_{22}\alpha_{23}}{\alpha_{11}^2\alpha_{22}^2\alpha_{33}^2} \\
\sigma_{22} &= \frac{\alpha_{23}^2 + \alpha_{33}^2}{\alpha_{22}^2\alpha_{33}^2} \\
\sigma_{33} &= \frac{1}{\alpha_{33}^2} \\
\sigma_{12} &= \sigma_{21} = -\frac{\alpha_{12}\alpha_{23}^2 + \alpha_{12}\alpha_{33}^2 - \alpha_{13}\alpha_{22}\alpha_{23}}{\alpha_{11}\alpha_{22}^2\alpha_{33}^2} \\
\sigma_{13} &= \sigma_{31} = \frac{\alpha_{12}\alpha_{23} - \alpha_{13}\alpha_{22}}{\alpha_{11}\alpha_{22}\alpha_{33}^2} \\
\sigma_{23} &= \sigma_{32} = -\frac{\alpha_{23}}{\alpha_{22}\alpha_{33}^2}
\end{aligned}$$

Es zeigt sich, dass sich die Varianzen und Kovarianzen als gemischte Aus-

[22] Vgl. (Stützle and Hrycej, 2001).

drücke von Elementen aus A ergeben. Die Skalierung der einzelnen Inputvariablen steht jedoch in quadratischem Zusammenhang mit den Varianzen und in linearem Zusammenhang mit den Kovarianzen, was sich in dieser direkten Zerlegung nicht widerspiegeln kann.

Wird jedoch die Cholesky-Zerlegung nach Gleichung (4.25) erweitert, erhält man eine Darstellung der Kovarianzmatrix

$$\Sigma = ((UD)^T UD)^{-1} = D^{-1}U^{-1}(U^T)^{-1}D^{-1}$$

für $n = 3$ als:

$$\begin{aligned}
\sigma_{11} &= \frac{1 + u_{12}^2 + (u_{12}u_{23} - u_{13})^2}{d_{11}^2} \\
\sigma_{22} &= \frac{u_{23}^2 + 1}{d_{22}^2} \\
\sigma_{33} &= \frac{1}{d_{33}^2} \\
\sigma_{12} &= \sigma_{21} = -\frac{u_{12} + (u_{12}u_{23} - u_{13})u_{23}}{d_{22}d_{11}} \\
\sigma_{13} &= \sigma_{31} = \frac{u_{12}u_{23} - u_{13}}{d_{11}d_{33}} \\
\sigma_{23} &= \sigma_{32} = -\frac{u_{23}}{d_{33}d_{22}}.
\end{aligned}$$

Offensichtlich erscheinen die strikt positiven Diagonalelemente d_{ii} ausschließlich im Nenner der Komponenten von Σ. In diesem Fall lässt sich die Kovarianzmatrix mit $S = U^{-1}(U^T)^{-1}$ und $V = D^{-1} = diag(\frac{1}{d_{11}}, \frac{1}{d_{22}}, \frac{1}{d_{33}})$ als

$$\Sigma = VSV$$

formulieren, wobei

$$S = \begin{pmatrix}
1 + u_{12}^2 + (u_{12}u_{23} - u_{13})^2 & -u_{12} - (u_{12}u_{23} - u_{13})u_{23} & u_{12}u_{23} - u_{13} \\
-u_{12} - (u_{12}u_{23} - u_{13})u_{23} & u_{23}^2 + 1 & -u_{23} \\
u_{12}u_{23} - u_{13} & -u_{23} & 1
\end{pmatrix}$$

gilt. Diese Zerlegung zeigt, dass die Diagonalelemente die Rolle der Skalie-

rungsfaktoren spielen und sich in den Elementen von U die stochastische Abhängigkeitsstruktur bzw. die Korrelationsstruktur abbildet.

Über diese Separation in Skalierungs- und Abhängigkeitsmatrix wurde eine signifikante Verbesserung der Anpassung an die realen Daten im Vergleich zur klassischen Cholesky-Zerlegung erzielt.[23]

Um die positiven Diagonalelemente $d_{ii} > 0,\ i = 1, \ldots, n$ und dadurch ebenfalls $\alpha_{ii} > 0,\ i = 1, \ldots, n$ zu erzwingen, kann etwa die Exponentialfunktion verwendet werden. Die Transformation τ_A bildet schließlich die Komponenten der zugehörigen Struktur-Zwischenparameter auf die Komponenten der Strukturmatrix $A = UD$ wie folgt ab:

$$\begin{aligned} \alpha_{ii} &= d_{ii} = \tau_{\alpha_{ii}}(\bar{p}_{d_{ii}}) = e^{\bar{p}_{d_{ii}}},\ i = 1. \ldots . n \\ \alpha_{ij} &= d_{ii} u_{ij} = \tau_{\alpha_{ij}}(\bar{p}_{\alpha_{ij}}) = e^{\bar{p}_{\alpha_{ii}}} \bar{p}_{\alpha_{ij}}, \\ &\quad i = 1 \ldots n-1;\ j = 2 \ldots n;\ j > i. \end{aligned}$$

Durch die angenommene Dreiecksgestalt der Strukturmatrix A ergibt sich deren Determinate $|A|$, die in der Log-Likelihood Gleichung (5.7) Verwendung findet, als einfacher Ausdruck der Struktur-Zwischenparameter $\bar{p}_A$. Über den Determinantenmultiplikationssatz lässt sich die Determinante von A mit der Formel

$$|A| = |UD| = \prod_{i=1}^{n} d_{ii} = \prod_{i=1}^{n} e^{\bar{p}_{d_{ii}}}.$$

berechnen.

[23] Für den Normalverteilungsfall lässt sich hierzu (Williams, 1996) vergleichen. Diese Gegenüberstellung ist in Kapitel 9 (Tabelle 9.6) zu finden.

Kapitel 5

Parameterschätzung

Nach den Darstellungen von funktionalen Approximatoren gilt es ein Prinzip zu präsentieren, welches das Prognosemodell optimal im Sinne einer Datenanpassung identifiziert. Zu diesem Zweck sind Hilfsmittel aus der numerischen Mathematik notwendig, da es i.Allg. analytisch nicht möglich ist, eine optimale Lösung zu bestimmen.

Die Herleitung der Maximum-Likelihood Zielfunktion über die Minimierung der Cross Entropie[1] zwischen der empirischen Verteilung und der berechneten Verteilung ist die erste aufgegriffene Thematik dieses Kapitels. Die zur Bestimmung der optimalen freien Schätzparameter verwendete globale Optimierungsmethode ist im Anschluss motiviert und erläutert.

5.1 Minimierung der Cross Entropie

Da es, wie schon häufig erwähnt, das Ziel ist, die Wahrscheinlichkeitsverteilung zu identifizieren, die in gewissem Sinn die empirische Verteilung der Daten möglichst exakt widerspiegelt, ist eine Kostenfunktion zu bestimmen, die diese Anforderung abbildet.

Eine allgemeine Methode, die Ähnlichkeit zwischen zwei Dichtefunktionen von Wahrscheinlichkeitsverteilungen $d^*(y)$ und $d(y)$ zu messen, ist die so

[1] Der Begriff *Entropie* wurde von (Shannon, 1948) in die Informationstheorie eingeführt.

genannte *Kullback Entropie*[2]

$$H_K(d, d^*) := \int_y d^*(y) \log \frac{d^*(y)}{d(y)} dy = \int_y d^*(y) \log d^*(y) dy - \int_y d^*(y) \log d(y) dy. \tag{5.1}$$

Die Kullback Entropie ist Null genau dann, wenn die beiden zu vergleichenden Verteilungen identisch sind. In dem hier vorliegenden Kontext ist dieses Maß nützlich, um die Ähnlichkeit zwischen der „wahren“ bzw. empirischen Verteilung $d^*(y)$ und ihrer Approximation $d(y)$ zu bestimmen.

Aus diesem Grund ist die Minimierung der Kullback Entropie durch Variation der geschätzten Verteilung $d(y)$ eine mögliche Vorgehensweise, die beste Approximation einer Verteilung zu bestimmen.

In diesem Fall ist der erste Term der Differenz in Gleichung (5.1) invariant zur Variation der geschätzten Dichtefunktion $d(y)$ und kann daher vernachlässigt werden. Es ergibt sich der vereinfachte Term

$$H(d, d^*) = -\int_y d^*(y) \log d(y) dy, \tag{5.2}$$

der *Cross Entropy* genannt wird.

Dieser Ausdruck (5.2) kann offensichtlich nicht ohne die Kenntnis der Dichtefunktion $d^*(y)$ berechnet werden. Steht jedoch eine Menge von K unabhängigen Beobachtungen zur Verfügung[3], so kann die empirische Verteilung durch

$$d_e^*(y) = \frac{1}{K} \sum_{k=1}^{K} \delta(y - y_k) \tag{5.3}$$

approximiert werden. δ bezeichnet hierbei die Dirac-Funktion.[4]

Wird in Gleichung (5.2) $d^*(y)$ durch den Schätzer $d_e^*(y)$ aus Gleichung (5.3) ersetzt, so folgt die geschätzte Cross Entropie über die Linearität und

---

[2] Siehe (Kullback, 1959).

[3] Vgl. hierzu die Annahmen des Abschnitts 2.3.

[4] Die Dirac-Funktion ist definiert als $\delta(x) := \begin{cases} 1, \text{ falls } x > 0 \\ 0, \text{ falls } x \leq 0 \end{cases}$.
Weiterhin ist $d_e^*(y)$ ein erwartungstreuer Schätzer für die Verteilung $d^*(y)$, vgl. etwa (Parzen, 1962a) oder (Grabec and Sachse, 1997).

die Berechnung des Integrals als

$$\begin{aligned} H(d, d^*) &\approx -\int_y d_e^*(y) \log d(y) dy \\ &= -\int_y \frac{1}{K} \sum_{k=1}^{K} \delta(y - y_k) \log d(y) dy \\ &= -\frac{1}{K} \sum_{k=1}^{K} \int_y \delta(y - y_k) \log d(y) dy \\ &= -\frac{1}{K} \sum_{k=1}^{K} \log d(y_k). \end{aligned} \tag{5.4}$$

Obige Herleitung zeigt, dass die Gleichung (5.4) bis auf eine Konstante $1/K$ mit der negativen Log-Likelihood Funktion äquivalent ist.

Die Integration des Maximum Likelihood Prinzips in das Gesamtkonzept, motiviert durch die Minimierung der Cross Entropie, wird in Abbildung 5.1 illustriert.

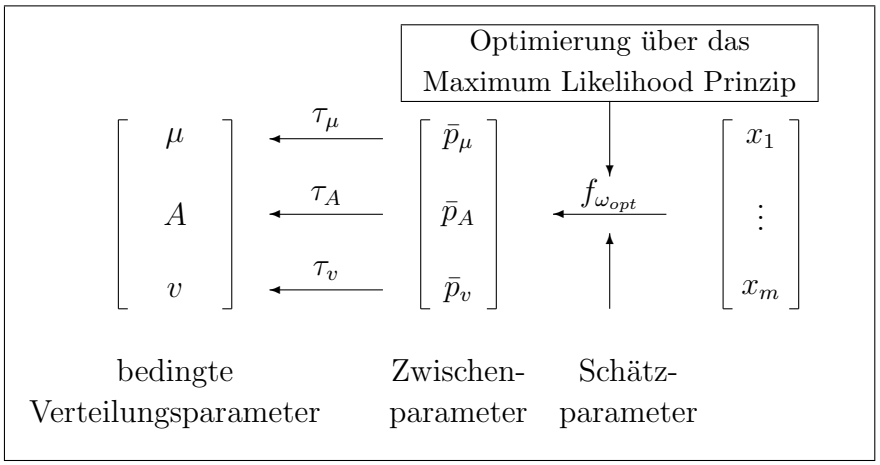


Abbildung 5.1: Optimierungsprinzip: Maximum Likelihood

Diese Argumentation kann analog auf den hier betrachteten Fall von bedingten Wahrscheinlichkeitsverteilungen $d^*(y|x)$ und $d(y|x)$ übertragen werden. Aus Gleichung (5.4) ergibt sich die negative Log-Likelihood Funktion

im bedingten Fall zu:

$$L(X,Y) = -\sum_{k=1}^{K} log(d(y_k|x_k)). \tag{5.5}$$

Aus dem Transformationssatz in Gleichung (3.6) ist die Beziehung zwischen der kanonischen und der zu identifizierenden Dichtefunktion bekannt. Die negative Log-Likelihood Zielfunktion aus (5.5) ergibt sich folglich zu:

$$L(X,Y) = -\sum_{k=1}^{K} log(|A(x_k)|) - \sum_{k=1}^{K} log(g(z_k)). \tag{5.6}$$

Die Parametrisierung der negativen Log-Likelihood Funktion L, die sich unter Verwendung der Transformation (3.5) aus der Darstellung (5.6) ergibt, zeigt der Ausdruck

$$L(X,Y,\omega) = -\sum_{k=1}^{K} log(|A_{\omega_2}(x_k)|) - \sum_{k=1}^{K} log(g_{\omega_3}(A_{\omega_2}(x_k)(y - \mu_{\omega_1}(x_k)). \tag{5.7}$$

Nach der Entwicklung einer Zielfunktion für das Prognosesystem steht nun die numerische Anpassung des Modells an die zugrunde liegenden Daten im Vordergrund.

5.2 Numerische Optimierung

In Gleichung (5.7) ist die negative Log-Likelihood Funktion $L(X,Y,\omega)$ über die parametrischen Abbildungen $\mu_{\omega_1}(x)$, $A_{\omega_2}(x)$ und $v_{\omega_3}(x)$ eindeutig bestimmt.

Durch die Minimierung der negativen Log-Likelihood Funktion bezüglich der Parametermenge $(\omega_1, \omega_2, \omega_3) = \omega \in \Omega$ wird eine Maximierung des Likelihoods erzielt. Es ergibt sich konkret die Optimierungsaufgabe

$$\min_{\omega \in \Omega} L(X,Y,\omega)$$

mit $\Omega = I\!R^{mn_2+n_2\bar{P}}$ für den Fall des 3-Schicht Perzeptrons aus Gleichung (4.19) als funktionalen Approximator bzw. $\Omega = I\!R^{(m+1)\bar{P}}$ bei Verwendung eines linearen Approximators, wie in Gleichung (4.10) beschrieben. Da die Log-Likelihood Zielfunktion i.Allg. nicht als konvex angenommen werden kann, ist eine globale numerische Optimierungsmethode[5] erforderlich. Der am häufigsten verwendete Lernalgorithmus für das Multi-Layer Perzeptron ist bislang noch immer der backpropagation Algorithmus[6], der ein steilstes Abstiegsverfahren mit konstanten Schrittweiten darstellt. Es ist jedoch bekannt, dass diese Verfahren keine Garantie der Konvergenz in einen stationären Punkt der Zielfunktion gewährleisten. Aus diesem Grund fand eine globale Optimierungsmethode in der Implementierung des hier vorgestellten Prognosekonzepts Verwendung.

Die globalen Optimierungsmethoden können in deterministische oder stochastische Verfahrensklassen aufgeteilt werden. Bei der deterministischen Gruppe von Algorithmen müssen gewisse Bedingungen an die Zielfunktion gestellt werden, um die Optimalität in einer endlichen Anzahl von Iterationen zu garantieren. Eine häufig verwendete Annahme ist die Kenntnis der Lipschitz-Konstanten, die allerdings in der Praxis meistens nur schwer zu berechnen ist. Auf diese Art von Optimierungsmethoden wird nicht weiter eingegangen.

Die stochastischen globalen Methoden bieten hingegen eine „asymptotische Garantie des Erfolgs“, d.h. mit schwachen Bedingungen an die Zielfunktion konvergiert bei anwachsender Stichprobenzahl die Wahrscheinlichkeit, das globale Minimum zu identifizieren gegen 1.[7] Die meisten stochastischen Verfahren lassen sich durch eine generelle Vorgehensweise charakterisieren, die in zwei Phasen eingeteilt werden kann.[8]

1. In einer *globalen Phase* werden die Zielfunktionswerte einer gewissen

[5] Im Kontext neuronaler Netze spricht man auch von einem Lernalgorithmus.

[6] Siehe etwa (Rumelhart et al., 1986).

[7] Aus technischen Gründen nehmen alle stochastischen globalen Optimierungsverfahren bzw. Lernalgorithmen die Konvexität und Kompaktheit des Parameterraums Ω an, wobei das globale Minimum im Innern $\mathring{\Omega}$ von Ω liegt, siehe etwa (Boender et al., 1982).

[8] Vgl. (Pardalos and Romeijn, 2002).

Anzahl von zufällig ausgewählten Punkten (Stichproben) berechnet.

2. In einer *lokalen Phase* werden die Stichproben transformiert, um mit Hilfe eines geeigneten lokalen Suchalgorithmus Kandidaten für das globale Minimum zu erhalten.

Wie bereits erwähnt gilt für stochastische Methoden, dass ausschließlich probabilistische Aussagen über die gefundene globale Lösung des Optimierungsproblems möglich sind. Unter den Voraussetzungen einer Gleichverteilungsannahme auf Ω und der Stetigkeit von L kann gezeigt werden, dass die Stichprobe mit kleinstem Zielfunktionswert mit Wahrscheinlichkeit 1 bei wachsender Iterationsanzahl gegen das globale Minimum konvergiert.[9]

Aufgrund dieser asymptotischen Garantie, erzeugt durch die globale Phase, wird eine grundsätzliche Zuverlässigkeit der stochastischen Methoden sichergestellt. Aus Effizienzgründen ist die lokale Phase jedoch unumgänglich. Im idealen Fall wird eine optimale Verwendung der lokalen Optimierungsmethode genau dann erreicht, wenn so viele lokale Optimierungsprozeduren gestartet werden wie Attraktoren im Wertebereich der Zielfunktion existieren.[10]

Umfangreiche Vergleichsstudien bestärken aus praktischer Sicht die Wahl der Multi-Level Single-Linkage Methode als globaler Optimierungsalgorithmus.[11]

Nach obiger Motivation der Methode seien in den folgenden Abschnitten die Entstehung dieser Optimierungsprozedur beschrieben sowie einige ihrer Eigenschaften.

---

[9] Diese Behauptung ist etwa in (Rinnooy Kan and Timmer, 1987) verifiziert.

[10] Attraktoren beschreiben Regionen, in denen das lokale Minimum über Abstiegsverfahren erreicht wird.

[11] Anhand unterschiedlicher Optimierungsaufgaben werden in (Ružička and Kober, 1992) die Algorithmen „singlestart“, „Bayesian reduced multistart“, „combined tunneling function random search“, „density clustering“, „single linkage clustering“, „multi-level single-linkage“ und „multi-level mode analysis“ gegenübergestellt.

5.3 Multi-Level Single-Linkage Methode

Anhand des einfachsten aller stochastischen Optimierungsverfahren, dem „Pure Random Search“[12], lassen sich die Schwächen von globalen Optimierungsmethoden, die ausschließlich aus einer globalen Phase bestehen, eindrucksvoll illustrieren. Diese Methode besteht nur aus einem einzigen Schritt.

Pure Random Search

> Berechne die Funktionswerte der Zielfunktion L in N Punkten, welche aus der Menge Ω, wo eine Gleichverteilung herrscht, zufällig gezogen wurden. Der Punkt mit dem kleinsten Zielfunktionswert ist die Lösung dieses Optimierungsverfahrens.

Es ist erstaunlich, dass sich im Vergleich dieses einfachen stochastischen Ansatzes zu einem ähnlich simplen deterministischen Zugang, dem „Grid Search“, ein Vorteil zugunsten des Pure Random Search Algorithmus herausstellen kann.[13] Das Grid-Search Verfahren berechnet die Zielfunktionswerte an allen Gitterpunkten eines gleichmäßigen Netzes über Ω.[14]

Dennoch stellt Pure Random Search für die praktische Implementierung offensichtlich keine ernst zu nehmende Methode dar. Daher wird im Folgenden zusätzlich eine lokale Phase im Optimierungsalgorithmus berücksichtigt. Die einfachste Art und Weise, Gebrauch von lokalen Optimierungsprozeduren P zu machen, ist die so genannte Multistart-Methode.

Multistart

> *Schritt 1:* Man zieht zufällig einen Punkt ω aus der Parametermenge Ω, für die eine Gleichverteilung angenommen wird.
>
> *Schritt 2:* Verwende die neue Stichprobe ω als Startwert und führe die lokale Suchprozedur P durch.

[12] Diese Methode geht auf (Brooks, 1958) und (Anderssen, 1972) zurück.

[13] Siehe zu diesen Versuchen (Ivanov, 1972) und (Anderssen and Bloomfield, 1975).

[14] In der ursprünglichen Version des Grid-Search Verfahrens werden äquidistante Stützstellen gewählt.

Schritt 3: Ein Kriterium gibt an, ob das Verfahren abgebrochen oder zu Schritt 1 zurückgegangen werden soll.[15]

Das lokale Minimum mit dem kleinsten Zielfunktionswert ist die Lösung des Optimierungsverfahrens.

Offensichtlich besitzt die Multistart-Methode beträchtliche Schwächen in Bezug auf Effizienzbetrachtungen, da in Schritt 2 möglicherweise ein und dasselbe Minimum wiederholt berechnet wird. Um diese überflüssigen zeitaufwendigen lokalen Optimierungsverfahren zu vermeiden, sollte idealerweise P nur einmal pro Einzugsgebiet[16] R_{ω^*} eines Minimums ω^* gestartet werden.

Versuche, die Problematik der unzureichenden Effizienz von Multistart zu lösen, wurden auf unterschiedlichste Weise unternommen.[17] Diese Methoden kommen bei der Suche nach dem globalen Minimum jedoch selten zum Erfolg, selbst dann, wenn eine Stichprobe aus dem Einzugsgebiet R_{ω^*} von ω^* gezogen wird. Dies liegt darin begründet, dass nicht unbedingt ein lokales Optimierungsverfahren im Einzugsgebiet des globalen Minimums gestartet wird.

Eine weitaus bessere Modifikation des Multistart stellen die Cluster-Verfahren bereit, die im nächsten Unterabschnitt beschrieben werden.

5.3.1 Cluster-Methoden

Die Grundidee der Cluster-Methoden lässt sich wie folgt skizzieren. Durch Ziehung einer gleichverteilten Stichprobe aus der Parametermenge Ω werden Gruppen gebildet. Die sogenannten Cluster, von gegenseitig „nah beieinander liegenden“ Punkten, bilden idealerweise Punktmengen, die den Einzugsgebieten der lokalen Minima entsprechen. Daher kann in jeder einzelnen Gruppe

[15] Im Artikel (Rinnooy Kan and Timmer, 1987) wird ein Kriterium vorgestellt, das einen Bayes-Schätzer sowohl für die Anzahl der lokalen Minima als auch für die relative Größe jedes Einzugsgebiets $R(\omega^*)$ zugrunde legt. Eine ausführliche Darstellung zu Bayes'schen Abbruchkriterien wird in (Boender, 1984) vorgestellt.

[16] Als Einzugsgebiet R_{ω^*} des Minimums ω^* wird die Menge der Punkte aus Ω definiert, von denen aus die lokale Suchprozedur P zu ω^* konvergiert.

[17] Einen der ersten Vorschläge unterbreitet (Hartman, 1973). Diese Methode startet eine lokale Suchprozedur P nur dann, wenn eine Stichprobe ω aus Ω gezogen wird, deren Funktionswert $L(X, Y, \omega)$ kleiner als der des bisher kleinsten lokalen Minimums ist.

ein lokales Optimierungsverfahren P gestartet werden und somit wären alle lokalen Minima identifiziert.

Es seien die beiden Hauptprobleme der Cluster-Methoden vorweggenommen:

1. Fehlerart (Qualitätsproblem): Die erzeugten Cluster beinhalten mehrere Einzugsgebiete, so dass das globale Minimum verfehlt werden könnte, da nur eine lokale Suchprozedur pro Cluster durchgeführt wird.

2. Fehlerart (Effizienzproblem): Mehrere Cluster könnten in einem Einzugsgebiet enthalten sein, so dass ein lokales Minimum häufiger berechnet wird.

Es sei an dieser Stelle wiederholend an den Unterschied zwischen Cluster C und Einzugsgebieten R erinnert. Cluster sind eine Vereinigung von Punkten, die nach gewissen Vorschriften entwickelt werden, wohingegen das Einzugsgebiet eines Minimums die Menge von Punkten ist, von der aus die lokale Suchprozedur zu diesem Minimum konvergiert.

Die in dieser Arbeit betrachteten iterativen Clustermethoden können als Standardtechniken angesehen werden. Sie gehorchen dem folgenden Grundprinzip:

Grundprinzip von Cluster-Methoden

Die Menge der Stichproben S und der stationären Punkte X^* sei als leer initialisiert.

Schritt 1 (globale Phase):
Ziehe N Punkte aus der Parametermenge Ω, deren Elemente als gleichverteilt angenommen werden.
Werte die Zielfunktion an diesen Punkten aus und füge sie der Stichprobenmenge S hinzu.

Schritt 2 (lokale Phase):
Wähle nach einer geeigneten Vorschrift[18] eine Teilmenge der Stichprobenmenge S aus, woraufhin jeder Punkt dieser Teilmenge als Startwert

[18]Zwei Konzepte der Transformation von S sind die Reduktion und die Konzentration.

einer lokalen Optimierungsmethode verwendet wird.
Füge die neu berechneten stationären Punkte der Menge X^* hinzu.

Schritt 3 (Abbruchkriterium):
Ein Kriterium gibt an, ob abgebrochen oder zu Schritt 1 zurück gegangen werden soll.

Das Element der Menge X^* mit kleinstem Zielfunktionswert ist die Lösung des Optimierungsverfahrens.

In *Schritt 2* dieser skizzierten Vorgehensweise werden die Cluster gebildet, wobei jedes Cluster um einen so genannten *Kernpunkt* entsteht. Der nicht zugeordnete Punkt mit kleinstem Zielfunktionswert oder das durch eine lokale Suchprozedur mit diesem Startwert gewonnene lokale Minimum ist z.B. als Kernpunkt denkbar.

Falls in *Schritt 2* die Menge der ausgewählten Punkte identisch ist mit der Stichprobenmenge S ist, so reduziert sich das Grundprinzip auf die Multistart Methode.

Drei mögliche Ansätze, die obigem Grundkonzept entsprechen, sind etwa:[19]

- Density Clustering
- Single-Linkage Clustering
- Mode Analysis

Da sich in der bereits erwähnten Vorstudie[20] das Single-Linkage Clustering als das geeignetste Verfahren herauskristallisiert hat, wird in dieser Arbeit vorwiegend auf diese Clustermethode eingegangen. Neben dem geringeren Rechenaufwand liegt der Vorteil der Single-Linkage Methode in der

Das Prinzip der Reduktion vernachlässigt zwischenzeitlich einen gewissen Prozentsatz der Stichprobe S. Im Zuge der Konzentration werden auf alle Stichproben ein oder mehrere steilste Abstiegsschritte angewendet. Nach (Rinnooy Kan and Timmer, 1987) liefert die Technik der Reduktion bessere Ergebnisse, da die Gleichverteilung auf der reduzierten Stichprobe weiterhin bestehen bleibt. Aus diesem Grund wird in der vorliegenden Abhandlung nur auf diese Art von Clustermethoden eingegangen.

[19] Vgl. (Rinnooy Kan and Timmer, 1987).

[20] Siehe (Ružička and Kober, 1992).

Variabilität ihrer Cluster im Gegensatz zu der Starrheit der Density-Cluster. Die Form der Cluster sollte mit wachsender Iterationszahl gegen die Vereinigung der Niveaumengen-Komponenten $L_{\omega^*}(y_k^{\gamma kN})$[21] der lokalen Minima ω^* konvergieren.

5.3.1.1 Single-Linkage Clustering

Die Single-Linkage Methode formt die Cluster sukzessive, wobei jede Gruppierung durch einen Kernpunkt geboren wird. Nach der Gründung eines Clusters C wird der Punkt mit minimalem Abstand dem Cluster hinzugefügt. Zur Aufnahme in das Cluster C wird daher jenes ω ausgewählt, das die Vorschrift

$$d(\omega, C) := \min_{\tilde{\omega} \in C} \|\omega - \tilde{\omega}\| \ ,$$

erfüllt, wobei $\|\cdot\|$ die Euklidische Norm bezeichnet. Diese Vorgehensweise wird so lange wiederholt, bis die Distanz $d(\omega, C)$ einen kritischen Abstand r_k überschreitet.

Um den Algorithmus formal exakt darzustellen, müssen einige geringfügige Bedingungen an die Vorschrift für die Wahl der Kernpunkte gestellt werden. Der Grund hierfür sind Schwierigkeiten der lokalen Suchalgorithmen in der Region nahe am Rand von S und in der Nachbarschaft stationärer Punkte.[22] Um dennoch theoretische Aussagen zu ermöglichen, wird vereinbart, dass keine lokalen Suchprozeduren in diesen Regionen stattfinden.[23]

---

[21] Die Niveaumengen-Komponenten sind wie folgt definiert: Mit y_k^i sei der i-t kleinste Zielfunktionswert der Stichprobe S in der k-ten Iteration bezeichnet.
Weiter sei $E(y_k^i) := \{\omega \in S : L(\omega) \leq y_k^i\}$ die (y_k^i)-Niveaumenge von L.
Dann ist $E_\omega(y_k^i)$ die (zusammenhängende) Niveaumengen-Komponente von $E(y_k^i)$, die ω beinhaltet.

[22] Zur Begründung der folgenden technischen Einschränkungen siehe etwa Ausführungen in (Rinnooy Kan and Timmer, 1987).

[23] Die Menge S wird durch folgende Annahme neu beschrieben. Es seien alle lokalen Minima Elemente der Menge S_τ, die wie folgt definiert ist:

$$S_\tau = S \backslash Q_\tau,$$

wobei $Q_\tau := \{x \in S : d(x, S \backslash \overset{\circ}{S}) < \tau\}$ gilt ($\overset{\circ}{S}$ bedeutet hier das Innere von S). Q_τ beschreibt daher eine Menge von Punkten aus S, die kleineren Abstand als τ zum Rand von S besitzen. Mit anderen Worten bedeutet dies, dass S_τ ein „verkleinertes“ Inneres von

Mit Hilfe dieser Annahmen und Bezeichnungen lässt sich nun die $k-$te Iteration der Single-Linkage Methode in Algorithmusschreibweise formulieren.

Single-Linkage Methode

Sei X^1 die Menge der lokalen Minima in X^*, M die Mächtigkeit von X^1, $\gamma \in (0;1]$ und $j := 1$.

Schritt 1 (Bestimmung der reduzierten Stichprobe):
Bestimme die γkN Stichproben mit kleinsten Zielfunktionswerten.[24]

Schritt 2 (Bestimmung des Kernpunktes):
STOPP, falls alle Punkte der reduzierten Stichprobe einem Cluster zugeordnet sind.
Falls $j \leq M$: Wähle $j-$tes lokales Minimum in X^1 als nächsten Kernpunkt; gehe zu Schritt 3.
Sonst: Bestimme nächsten Kernpunkt $\bar{x}$ als Punkt mit kleinstem Zielfunktionswert unter den noch nicht zugeordneten Punkten der reduzierten Stichprobe.
Falls $\bar{x} \notin Q_\tau \cup X^*_\nu$ wende lokale Suchprozedur P mit $\bar{x}$ als Startwert an und füge alle vorkommenden stationären Punkte der Menge X^* hinzu. Passe gegebenenfalls X^1 und M an.[25]

Schritt 3 (Bildung des Clusters):
Bilde ein Cluster initiiert durch einen Kernpunkt (aus Schritt 2) folgendermaßen:

S darstellt.

Des Weiteren sei die zweite kritische Menge X^*_ν definiert durch:

$$X^*_\nu := \{x \in S : \|x - x^*\| < \nu \; \forall x^* \in X^*\},$$

wobei $\nu > 0$ (ν klein) ist. X^*_ν beinhaltet also diejenigen Punkte von S, die näher als ein klein gewählter Abstand ν an den in X^* notierten stationären Punkten liegen.

[24] I.Allg. lässt sich nur schwer eine Aussage bezüglich der optimalen Wahl der Konstanten γ formulieren. In (Rinnooy Kan and Timmer, 1987) wird für praktische Anwendungen ein Wert zwischen 10% und 20% angegeben.

[25] Es sei hierbei die notwendige und hinreichende Optimalitätsbedingung beachtet.

Füge die Punkte aus der reduzierten Stichprobe, welche innerhalb der Distanz r_k zum Cluster liegen, dem Cluster hinzu.
Setze $j := j + 1$ und gehe zu Schritt 2.

Mit der geeigneten Wahl des kritischen Abstands r_k ist es möglich, die Wahrscheinlichkeit einer der oben erwähnten Fehlermöglichkeiten zu minimieren.[26] Mit der Festlegung von

$$r_k = \pi^{-1/2}\left(\Gamma\left(1+\frac{n}{2}\right)|S|\,\frac{\kappa\log(kN)}{kN}\right)^{1/n} \tag{5.8}$$

lassen sich Aussagen in Abhängigkeit der Konstanten κ formulieren. Hierbei bedeutet N die Anzahl der insgesamt gezogenen Stichproben aus dem $n-$dimensionalen Parameterraum Ω, k bezeichnet den Iterationsindex und $|S|$ die Mächtigkeit von S. Zur Charakterisierung der Single-Linkage Methode und Abschätzung ihrer Fähigkeiten seien folgende Behauptungen formuliert.[27]

Satz 5.3.1
Es sei der kritische Abstand r_k wie in Gleichung (5.8) definiert.

i) Für die Werte $\kappa > 2$ konvergiert die Wahrscheinlichkeit, dass eine lokale Suche durch die Single-Linkage Methode gestartet wird bei wachsender Iterationszahl k gegen Null.

ii) Für die Werte $\kappa > 4$ ist die Gesamtanzahl der gestarteten lokalen Suchalgorithmen mit Wahrscheinlichkeit 1 endlich, selbst wenn das Ziehen der Stichproben nie endet.

Satz 5.3.2
Falls die kritische Distanz r_k des Single-Linkage Verfahrens bei wachsender Iterationszahl k gegen Null konvergiert, findet die Methode mit Wahrscheinlichkeit 1 nach endlich vielen Iterationsschritten ein lokales Minimum in jeder

[26] Vgl. (Boender et al., 1982).
[27] Die beiden Sätze 5.3.1 und 5.3.2 orientieren sich an Theoremen aus (Boender et al., 1982).

Niveaumengen-Komponente von $E(y_\gamma)$, aus welcher eine Stichprobe gezogen wurde.

Auf die Verifikation obiger Sätze 5.3.1 und 5.3.2 sei an dieser Stelle verzichtet.[28]

Obige Ausführungen verbinden die Idee der Stichprobenreduktion mit einer Clustermethode. Zu Beginn wurde die Fehlerart des Qualitätsproblems skizziert, die durch das Verfehlen eines lokalen Minimums charakterisiert ist. Da nicht immer vermieden werden kann, dass ein Cluster bzw. eine Niveaumengen-Komponente mehrere Einzugsgebiete lokaler Minima beinhaltet, kann diese Fehlermöglichkeit auch bei Clustermethoden nicht ausgeschlossen werden. Dieses Defizit haben alle drei oben genannten Clustermethoden gemein.[29] Daher kann auch die in obigem Abschnitt dargestellte Single-Linkage Methode diese Fehlerart theoretisch nicht zufriedenstellend beheben.

Abschnitt 5.3.2 wird sich speziell mit diesem Problem beschäftigen und eine Methode vorstellen, die bei wachsender Iterationszahl mit Wahrscheinlichkeit 1 genau einmal in jedem Einzugsgebiet eine lokale Suchprozedur anwendet.

5.3.2 Multi-Level Methoden

Wie bereits zu Ende des vorigen Abschnitts 5.3.1 erwähnt, wird nun das Ziel sein, eine Methode vorzustellen, welche die Genauigkeit von Multistart mit so wenig wie möglich Anwendungen von lokalen Suchprozeduren garantiert.

Ein entscheidender Mangel bei der Betrachtung der oben erläuterten Methode ist die unzureichende Nutzung der Zielfunktionswerte der Stichproben. Die Information der Funktionswerte wird ausschließlich für die Bildung der reduzierten Stichprobe S in der globalen Phase berücksichtigt. Die Single-Linkage Methode konzentriert sich in der lokalen Phase einzig und allein auf die Lage der reduzierten Stichprobenpunkte. Folglich unterscheidet die Methode nicht zwischen verschiedenen Einzugsbereichen, die in der gleichen

[28] Vgl. hierzu etwa die ausführlichen Darstellungen in (Rinnooy Kan and Timmer, 1987).
[29] Vgl. (Rinnooy Kan and Timmer, 1987).

Niveaumengen-Komponente von $E(y_k^{(\gamma kN)})$ liegen. Dies ist der Grund für das verbleibende Risiko von Qualitätsfehlern.

Der Funktionswert einer Stichprobe ω könnte bei der Frage, zu welchem Einzugsgebiet ω gehört, von Bedeutung sein. Da die lokale Suchmethode, die das Einzugsgebiet definiert, als striktes Abstiegsverfahren angenommen wird, kann ω nicht zu einem Einzugsgebiet eines lokalen Minimums ω^* gehören, falls kein Abstiegsweg von ω nach ω^* existiert. Ein Abstiegsweg ist definiert als eine Folge von Punkten mit monoton fallenden Zielfunktionswerten.

Offensichtlich ist es unmöglich, alle Abstiegspfade, die von ω ausgehen, zu betrachten und zu untersuchen. Daher wird ein Hilfsmittel, die sogenannte r_k-Abstiegsfolge, herangezogen.[30]

Hinführend zur endgültigen Multi-Level Single-Linkage Methode sei die k-te Iteration des folgenden Algorithmus angegeben.

Vorbereitender Algorithmus:

Es sei $\bar{M}$ die Anzahl der bisher bekannten Minima.

Schritt 1: Eröffne $\bar{M}$ verschiedene Cluster, bestehend aus den bisher bekannten lokalen Minima.

Schritt 2: Ordne die Stichproben nach den Zielfunktionswerten,
d.h. $L(\omega_i) < L(\omega_{i+1}),\ 1 \leq i \leq kN-1$.
Setze $i := 1$.

Schritt 3: Teile den Stichprobenpunkt ω_i jedem Cluster zu, welches ein Element, näher als die kritische Distanz r_k beinhaltet.
Falls ω_i keinem Cluster zugeteilt werden kann, führe eine lokale Suchprozedur mit Startpunkt ω_i durch, um so ein lokales Minimum ω^* zu erhalten.
Falls $\omega^* \notin X^*$, füge ω^* zu X^* hinzu, setze $\bar{M} := \bar{M} + 1$ und eröffne ein Cluster mit ω^*.
Füge ω_i dem durch ω^* eröffneten Cluster hinzu.

[30] Eine r_k-Abstiegsfolge ist eine Folge von Stichprobenpunkten, so dass alle sukzessiv aufeinander folgenden Punkte eine Distanz besitzen, die kleiner als r_k ist. Ferner haben die Folgenelemente monoton fallende Zielfunktionswerte.

Schritt 4: Falls $i = kN$, STOPP
Sonst setze i:=i+1 und gehe zu Schritt 3.

Es wird angenommen, dass in Schritt 3 keine unnötige lokale Suchprozedur gestartet und folglich der Startwert zum Cluster des neu berechneten lokalen Minimums hinzugefügt wird. Unter dieser Annahme kann die Existenz einer r_k-Abstiegsfolge zwischen ω und ω^* gefolgert werden. Es sei an dieser Stelle bemerkt, dass die Stichprobe ω möglicherweise in mehrere Cluster geordnet wird, falls von ω zu mehreren lokalen Minima r_k-Abstiegsfolgen existieren.

Obwohl es eine r_k-Abstiegsfolge von ω nach ω^* gibt, lässt sich nicht sicherstellen, dass ω im Einzugsgebiet von ω^* liegt. Die Begründung liegt in der Möglichkeit anderer Abstiegspfade, denen die Suchmethode gefolgt sein kann. Dies ist jedoch ausgeschlossen, falls sich ω im Inneren des Einzugsgebiets von ω^* befindet und r_k klein genug ist.

Diese Aussage kann skizzenhaft wie folgt begründet werden. Da zwei absolute Minima immer von einem Gebiet mit höheren Zielfunktionswerten getrennt sind, wird der „vorbereitende Algorithmus“ jedes Minimum in der Nachbarschaft einer Stichprobe auffinden, falls nur r_k klein genug ist.

Die hier unterlassene Reduzierung der Stichprobe ist ein weiterer Unterschied zur Single-Linkage Methode. Da die Zielfunktionswerte in obigem Algorithmus explizit verwendet werden, kann die Methode auf die komplette Stichprobe angewendet werden.

Nach diesen Erläuterungen wäre es durchaus möglich, den vorbereitenden Algorithmus mit „Multi-Level Single-Linkage Methode“ zu bezeichnen. Es lässt sich jedoch ein noch einfacherer Algorithmus formulieren, der ebenso wie die obige Prozedur für genügend kleines r_k die gleichen Minima identifiziert.

Es sei jedoch auf die eigentliche Überflüssigkeit der Clusterbildung hingewiesen, da die Entscheidung, ob eine lokale Suchprozedur P gestartet wird, nicht von einer Clusterzugehörigkeit abhängt. Eine Suchmethode wird nur dann durchgeführt, falls in einer Umgebung von ω mit Radius r_k keine Stichprobe z existiert, deren Zielfunktionswert kleiner ist als der von ω.

Nachfolgend wird ein Algorithmus vorgestellt, der auf die überflüssige Clusterbildung verzichtet. Ebenso wie bei der Single-Linkage Methode ist für theoretische Betrachtungen die Annahme erforderlich, dass P nicht von einem

Element aus der Menge $Q_\tau \cup X_\nu^*$ gestartet wird. Es sei die $k-$te Iteration betrachtet.

Multi-Level Single-Linkage Methode:

> Wende P genau dann auf jede Stichprobe ω_i $\forall i = 1, \ldots, kN$ an, falls $\omega_i \notin Q_\tau \cup X_\nu^*$ und keine Stichprobe z existiert mit $L(z) < L(\omega)$ und $\|z - \omega_i\| \leq r_k$. Füge neue stationäre Punkte, die während der lokalen Suche auftreten, X^* hinzu.

Abschließend werden in diesem Kapitel zwei Aussagen zur Problematik der 1. und 2. Fehlermöglichkeit angegeben. Wiederum lassen sich über die Wahl des kritischen Abstands r_k aus Gleichung (5.8) die folgenden beiden Aussagen in Abhängigkeit des Parameters κ formulieren:[31]

Satz 5.3.3 (Aussage zur 2. Fehlerart)
Es sei der kritische Abstand r_k wie in Gleichung (5.8) definiert.

i) Falls $\kappa > 0$ und ω eine beliebige Stichprobe ist, konvergiert die Wahrscheinlichkeit, dass P auf ω angewendet wird, für größer werdendes k gegen Null.

ii) Falls $\kappa > 2$ ist, konvergiert die Wahrscheinlichkeit, dass P in der $k-$ten Iteration angewendet wird, für größer werdendes k gegen Null.

iii) Falls $\kappa > 4$ ist, bleibt die Gesamtzahl der gestarteten lokalen Suchprozeduren P endlich mit Wahrscheinlichkeit 1, obwohl das Ziehen der Stichproben nie endet.

Satz 5.3.4 (Aussage zur 1. Fehlerart)
Falls r_k bei wachsendem k gegen Null konvergiert, wird jedes lokale Minimum ω^* von der Multi-Level Single-Linkage Methode mit Wahrscheinlichkeit 1 in einer endlichen Anzahl von Iterationen gefunden.

[31] Vgl. (Rinnooy Kan and Timmer, 1987).

Durch die Vermeidung der expliziten Clusterbildung kann durch die Verwendung spezieller Sortierungs- und Suchalgorithmen lineare Laufzeit erreicht werden, d.h. die erwartete Rechenzeit bis zur k-ten Iteration beträgt $O(k)$.[32]

Bislang wurde zur Art der verwendeten lokalen Suchprozedur keine Aussage gemacht, obwohl diese Wahl entscheidenden Einfluss auf die Effizienz der gesamten Methode hat. In Abschnitt 5.4 werden einige ausgewählte lokale Optimierungsalgorithmen vorgestellt. Dabei beschränken sich die folgenden Darstellungen auf die so genannten Quasi-Newton Verfahren.

5.4 Lokale Optimierungsverfahren

Es existieren eine Vielzahl von unterschiedlichen numerischen lokalen Optimierungsverfahren.[33]

Die wohl in der Praxis am häufigsten verwendeten Algorithmen sind

- die Gradienten-Methoden,
- die konjugierten Gradienten-Methoden,
- die Newton Methoden und
- die Quasi-Newton Methoden.

Für detaillierte Ausführungen sei auf einige Lehrbücher verwiesen, die nichtlineare lokale Optimierungsmethoden ausführlich behandeln.[34]

Numerische lokale Optimierungsverfahren zweiter Ordnung sind bezüglich ihrer schnellen Konvergenz[35] und ihrer Robustheit im Fall von schlecht konditionierten Problemen bekannt. Aus diesen Gründen und motiviert durch praktische Ergebnisse[36] wurden vielzählige experimentelle Studien[37] dieser

[32] Vgl. etwa die Darstellungen in (Ružička and Kober, 1992).

[33] Das UFO System in (Lukšan et al., 1992) beinhaltet eine Vielzahl von numerischen Methoden für unrestringierte Optimierungsaufgaben.

[34] Siehe etwa (Polak, 1971), (Gill et al., 1981), (Dennis and Schnabel, 1983), (Fletcher, 1987) oder (Press et al., 1992).

[35] Die Konvergenzraten sind i.Allg. superlinear oder sogar quadratisch.

[36] Siehe etwa (Lukšan, 1990).

[37] Siehe etwa (Ružička and Kober, 1992).

Methoden durchgeführt.

In der hier verwendeten globalen Multi-Level Single-Linkage Optimierungsmethode[38] wurde das Quasi-Newton Verfahren mit dem Update von Broyden, Fletcher, Goldfarb und Shanno (BFGS) implementiert. Daher wird ausschließlich auf diesen lokalen Suchalgorithmus eingegangen.

Die numerischen lokalen Optimierungsverfahren zweiter Ordnung basieren im Wesentlichen auf der lokalen quadratischen Approximation der Zielfunktion. Die Taylor-Entwicklung der Zielfunktion $L(\omega)$ um einen beliebigen Punkt ω_0 im Parameterraum besitzt folgende Gestalt:

$$\tilde{L}(\omega) = L(\omega_0) + (\omega - \omega_0)^T g + \frac{1}{2}(\omega - \omega_0)^T H(\omega - \omega_0), \tag{5.9}$$

wobei g den Gradient und H die Hessematrix der Zielfunktion L, ausgewertet an der Stelle ω_0, bezeichnen.[39]

Die korrespondierende Approximation des Gradienten der Zielfunktion ergibt sich aus der Taylor-Entwicklung (5.9) direkt als:

$$\nabla\tilde{L} = g + H(\omega - \omega_0). \tag{5.10}$$

Aus dieser Annäherung des Gradienten im Punkt ω_0 lässt sich ein iteratives Verfahren zweiter Ordnung ableiten, das formal wie folgt beschrieben werden kann:

$$\omega^+ = \omega + \alpha s, \tag{5.11}$$

wobei die Suchrichtung s das Gleichungssystem $Hs = -g$ erfüllt. Hierbei bezeichnen ω und ω^+ den alten und neuen veränderlichen Parametervektor. Wird die explizite Hesse-Matrix H verwendet, so bezeichnet

$$s = -H^{-1}g \tag{5.12}$$

die *Newtonrichtung* oder den *Newtonschritt* und Gleichung (5.11) beschreibt

[38] Siehe (Ružička and Kober, 1992).

[39] Dies entspricht der Approximation der Zielfunktion L durch eine quadratische Funktion $\tilde{L}$ im Punkt ω_0.

eine Iteration des klassischen Newton-Verfahrens.[40]

Um die Schrittweite α zu bestimmen, gibt es wiederum unterschiedliche Strategien.[41] Die hier verwendete und aus der Praxis bekannte Schrittweitenbestimmung ist der so genannte *line search*. Die Schrittweite α wird hierbei so gewählt, dass folgende Gleichungen erfüllt sind:

$$L^+ - L \leq \epsilon_1 \alpha s^T g \quad \text{und} \quad s^T g^+ \geq \epsilon_2 s^T g,$$

wobei $0 < \epsilon_1 < 1/2$ und $\epsilon_1 < \epsilon_2 < 1$ gilt. Hierbei bezeichnen L bzw. L^+ den alten bzw. neuen Zielfunktionswert und g bzw. g^+ den alten bzw. neuen Gradienten der Zielfunktion.

Offensichtlich bestehen bei der direkten Verwendung des Newton-Verfahrens numerische Herausforderungen, die bei großen industriellen Problemstellungen nicht auf akzeptable Weise gelöst werden können. Erstens erfordert die exakte Berechnung der Hesse-Matrix $O(K|\Omega|^2)$ Rechenoperationen und zweitens benötigt die Inversion der Hesse-Matrix weitere $O(|\Omega|^3)$ Rechenoperationen.

Aus diesen Gründen liegen die so genannten *Quasi-Newton Verfahren* nahe, da sie nach jeder Iteration eine verbesserte Approximation der inversen Hessematrix erzeugen. In den anfänglichen Iterationsschritten ist eine grobe Approximation der Hessematrix ohne bedeutende Effizienz- und Genauigkeitsverluste in Kauf zu nehmen. Zur Berechnung des iterativen Updates B^+ wird im Wesentlichen die Information erster Ordnung, d.h. des Gradienten der Zielfunktion benötigt. Probleme, die auftreten, falls die Hessematrix nicht positiv definit ist, werden mit Hilfe einer Initialisierung durch eine positiv definite Matrix, etwa der Einheitsmatrix, und einer Sicherstellung der positiven Definitheit der folgenden Hessematrixapproximationen B^+ beseitigt.

Die iterativen Approximationen der Hessematrix werden über die so ge-

[40] Dieses iterative Optimierungsverfahren wird in der Literatur auch unter der Bezeichnung *Newton-Raphson-Verfahren* aufgeführt.

[41] Diese Methoden zur Schrittweitenbestimmung werden etwa in (Dennis and Schnabel, 1983) als Modifikationen des Newton-Verfahrens zur globalen Konvergenz bezeichnet und ausführlich behandelt.

nannte *Quasi-Newton Bedingung* konstruiert:

$$s^+ - s = -H^{-1}(g^+ - g). \tag{5.13}$$

Diese Bedingung entsteht aus den Newtonrichtungen (5.12) von zwei aufeinander folgenden Iterationen.

Die Iterationsvorschrift des *BFGS-Updates*[42] ist, basierend auf Gleichung (5.13), definiert als:

$$B^+ := \frac{1}{\gamma}\left(B + \frac{\gamma}{\rho b}\bar{g}\bar{g}^T - \frac{1}{c}B\bar{\omega}(B\bar{\omega})^T + \frac{\beta}{c}\left(\frac{c}{b}\bar{g} - B\bar{\omega}\right)\left(\frac{c}{b}\bar{g} - B\bar{\omega}\right)^T\right),^{43} \tag{5.14}$$

mit

$$\begin{aligned} \bar{\omega} &= \omega^+ - \omega = \alpha s \\ \bar{g} &= g^+ - g \end{aligned}$$

und

$$\begin{aligned} b &= \bar{g}^T\bar{\omega} \\ c &= \bar{\omega}B\bar{\omega}. \end{aligned}$$

Es lässt sich zeigen, dass der BFGS-Update aus Gleichung (5.14) die Quasi-Newton Bedingung (5.13) erfüllt.[44]

Wird als Initialisierungsmatrix, wie oben bereits erwähnt, die Einheitsmatrix verwendet, so ist der erste Newtonschritt äquivalent mit der Richtung des negativen Gradienten. Außerdem ist in jedem Iterationsschritt garantiert, dass $-Bg$ eine Abstiegsrichtung darstellt, da die positive Definitheit aller BFGS-Updates sichergestellt ist.

Neben der einleitend erwähnten Robustheit und der schnellen Konvergenz

[42] Der BFGS-Update geht auf Broyden, Fletcher, Goldfrab und Shanno zurück und erhielt daher seine Bezeichnung.

[43] Für die Wahl der Parameter $\rho > 0$ (Bigg's Parameter), $\gamma > 0$ (Oren's Parameter) und β sei auf (Ružička and Kober, 1992) und (Lukšan, 1990) verwiesen.

[44] Vgl. etwa (Luenberger, 1984).

der lokalen Optimierungsverfahren zweiter Ordnung haben die Quasi-Newton Verfahren einen weiteren entscheidenden Vorteil gegenüber den konjugierten Gradientenmethoden, da die Schrittweitenbestimmung bei Quasi-Newton Verfahren bei weitem nicht so sensibel genau zu bestimmen ist.[45]

Dagegen ergeben sich gegebenenfalls durch die Speicherung der Matrix B bei einer großen Anzahl von freien Schätzparametern Probleme. Es sind $O(|\Omega|^2)$ Einträge zu speichern, wohingegen die Optimierungsverfahren erster Ordnung $O(|\Omega|)$ Einträge an Speicherplatz benötigen.[46]

Der Gradient g der Maximum-Likelihood Zielfunktion L lässt sich über die Kettenregel der Differentialrechnung bestimmen als:

$$\begin{aligned} g = \frac{\partial L(X, Y, \omega)}{\partial \omega} &= \sum_{k}^{K} \frac{\partial L(x_k, y_k, \omega)}{\partial \omega} \\ &= \sum_{k}^{K} \frac{\partial L(x_k, y_k, \omega)}{\partial \bar{p}(x_k, \omega)} \frac{\partial \bar{p}(x_k, \omega)}{\partial \omega} \end{aligned}$$

Der erste Term der Summation kann für die meisten hier betrachteten Verteilungsklassen analytisch berechnet werden.[47] Der zweite Term errechnet sich über die bekannten backpropagation Formeln.[48]

[45] Vgl. (Shanno, 1978).

[46] Es sei an dieser Stelle auf die *limited memory BFGS Quasi-Newton Algorithmen* verwiesen, die ebenso wie das konjugierte Gradientenverfahren nur $O(|\Omega|)$ Einträge speichern.

[47] Für den Fall der generalisiert hyperbolischen Verteilungsklasse vgl. etwa (Prause, 1999). Die weiteren Verteilungsfamilien ergeben keine wesentlichen Schwierigkeiten. Lediglich bei stabilen Verteilungen sind Hilfsmittel erforderlich, siehe hierzu Abschnitt 6.2.

[48] Vgl. (Rumelhart et al., 1986).

Kapitel 6

Klassen von Wahrscheinlichkeitsverteilungen

Um den allgemeinen Anspruch der Methode, wie in Abschnitt 2.2.2 motivierend formuliert, zu belegen, werden Klassen von Wahrscheinlichkeitsverteilungen präsentiert, die im vorgestellten Konzept Verwendung finden können.

Häufig nehmen traditionelle Prognosemethoden, wie klassische lineare Modelle oder herkömmliche Zeitreihenmodelle, normalverteilte Residuen an.

Die Motivation, nichtnormale Klassen von Wahrscheinlichkeitsverteilungen in Betracht zu ziehen, erklärt sich aus der bekannten Tatsache, dass realistische Zufallsgrößen aus der Praxis oftmals nicht annähernd der Normalverteilung gehorchen. Einige bekannte Beispiele sind bei der Modellierung von Finanzdaten aufgetreten. Pagan und zuvor Mandelbrot zeigen, dass Finanzwerte existieren, die mindestens "semi-heavy tails" besitzen, d.h. die Kurtosis größer als die Null-Kurtosis der Normalverteilung ist.[1] Insbesondere für Risikoschätzungen etwa durch Value-at-Risk[2] Analysen ist die Modellierung der Tails von besonderem Interesse. Zweifellos wird die Verteilung stochastischer Zielvariablen in einigen Fällen gut durch die Gauß'sche Normalverteilung approximiert, dennoch existiert in den meisten praktischen Problemstellungen keine begründete Rechtfertigung eines Präferierens der

[1] Siehe (Pagan, 1996) und (Mandelbrot, 1963).
[2] Value-at-Risk wird in Kurzform auch als VaR bezeichnet.

Normalverteilungsannahme. Diese Einschätzung impliziert nicht, dass die in der Realität auftretenden Zufallsgrößen exakt den hier alternativ vorgestellten Wahrscheinlichkeitsverteilungen gehorchen, sondern dass durch flexible Verteilungen eine bessere Approximation der vorliegenden Prozesse möglich wird.

Auch ist bei den hier betrachteten Anwendungen der Bedarfsprognose von Ersatzteilen und der Absatzprognose von Nutzfahrzeugen, die in den Kapiteln 9 und 10 dargestellt sind, die Verwendung einer nichtnormalen Verteilungsklasse berechtigt. Dies liegt im Wesentlichen im asymmetrischen Verhalten und den existierenden heavy tails der zufälligen Zielgrößen begründet.

Um diese Besonderheiten aufzugreifen und ihnen Rechnung zu tragen, wurden im Laufe der letzten Jahrzehnte eine Reihe zur Normalverteilung alternativer Verteilungsfamilien in unterschiedlichsten Anwendungsgebieten vorgeschlagen. Hierzu gehören sicherlich die Bemühungen in der Welt der Finanzwirtschaft. Dieses Anwendungsgebiet sei hier beispielhaft verwendet, die immer stärker werdende Relevanz und Akzeptanz von alternativen Verteilungsklassen zu illustrieren.

Bereits Anfang der sechziger Jahre postulierten (Mandelbrot, 1963), (Fama, 1965), (Mandelbrot and Taylor, 1967) die sogenannten stabilen Verteilungen, die als Spezialfall die Gauß'sche Normalverteilung enthalten.[3]

Etwa in (Blattberg and Goneges, 1974) wurde die t-Verteilung als Alternative zur Modellierung von Finanzmarktrenditen vorgeschlagen.

Zudem finden sich zahlreiche Vorschläge, die auf Kombinationen unterschiedlicher Verteilungen beruhen. So wurde beispielsweise in (Bishop, 1994) die Dichtefunktion der Renditen von Finanzmarktderivaten durch eine lineare Kombination zweier Normalverteilungen modelliert.

Weiterhin basieren die publizierten Arbeiten (Eberlein and Keller, 1995) und (Prause, 1999) auf der Familie der verallgemeinert hyperbolischen Verteilungen im Kontext von Finanzmarktdaten.

Es ist daher ein Ziel der Arbeit, diese Klassen von Wahrscheinlichkeitsverteilungen in ein einheitliches allgemeines Prognosekonzept zu integrieren

[3]Zu weiteren Darstellungen siehe etwa (Rachev and Mittnik, 1997) und (Rachev and Mittnik, 2000).

und abzubilden.[4]

	Verteilungsklasse	Unterklassen (1) - (18)
(1)	Sphärisch	(2) - (15)
(2)	Elliptisch	(3) - (15)
(3)	- Kotz-type	(4)
(4)	- - Normalverteilung	-
(5)	- Pearson-type VII	(6), (7)
(6)	- - t	(7)
(7)	- - - Cauchy	-
(8)	- Pearson-type II	-
(9)	- logistisch	-
(10)	- Bessel	(11)
(11)	- - Laplace	-
(12)	- sym. stabil	(4), (7)
(13)	- sym. hyperbolisch	(4)
(14)	- sym. gen. hyperbolisch	(4), (13)
(15)	- unimod. Mixtur von Normalvtlg.	(4)
(16)	Stabil	(4), (6), (7), (12)
(17)	Hyperbolisch	(4), (13)
(18)	Gen. hyperbolisch	(4), (13), (14), (15), (17)

Tabelle 6.1: Inklusionsrelationen einiger Klassen von Wahrscheinlichkeitsverteilungen

Wie bereits in Kapitel 3 erwähnt, lässt sich eine Vielfalt von Verteilungsklassen in obiges Konzept integrieren. In den folgenden Abschnitten sind daher Verteilungsklassen aufgeführt, die im Zusammenhang mit dem hier entwickelten und vorgestellten Prognosekonzept betrachtet sind. Tabelle 6.1 zeigt den Zusammenhang der unterschiedlichen Verteilungsfamilien auf und begründet die Schwerpunkte der folgenden Darstellungen. Neben der Normalverteilung wird den elliptischen, generalisiert hyperbolischen und den stabilen Verteilungen als auch den Gauß'schen Mixturverteilungen aufgrund ihrer Flexibilität besonderes Augenmerk geschenkt. Die elliptischen Verteilungen werden anhand ihrer sphärischen Form definiert, da diese die Rolle der kanonischen Dichte spielt. Die Gauß'sche Mixtur stellt eine gewisse Ausnahme

[4]Vgl. den Motivationsaspekt aus Abschnitt 2.2.2.

gegenüber den übrigen Verteilungsfamilien dar, da sie im erweiterten Sinn als eine parametrische Approximation von nichtparametrischen Verteilungen verstanden werden kann.

Die Repräsentation der Wahrscheinlichkeitsfamilien in den folgenden Abschnitten über ihre kanonische Form wird die Anwendungsfähigkeit für das entwickelte Prognosekonzept verdeutlichen.

6.1 Elliptische Wahrscheinlichkeitsverteilungen

In diesem Kapitel werden Untermengen der breiten elliptischen Verteilungsfamilie behandelt. Auf detaillierte Untersuchungen bezüglich charakteristischen Funktionen, Randverteilungen oder Momente der hier vorgestellten Verteilungsfamilien sei jedoch innerhalb dieser Betrachtungen verzichtet.[5] Wie in Tabelle 6.2 ersichtlich, sind Verteilungsfamilien wie die Normalverteilung oder die t-Verteilung Elemente dieser Klasse.

Eine Auflistung von Untermengen und Spezialfällen von elliptischen Verteilungen wird ebenfalls in Tabelle 6.2 präsentiert.[6] Hierbei sind entweder der Dichtegenerator h, die Dichtefunktion f oder die charakteristische Funktion (c.f.) angegeben.

Aus den Tabellen 6.1 und 6.2 wird ersichtlich, dass symmetrische Spezialfälle aus anderen Verteilungsklassen, wie generalisiert hyperbolische oder stabile Verteilungen, in der Klasse der elliptischen Verteilungen liegen. In diesen Fällen sind die Verteilungen elliptisch, falls der Symmetrieparameter β als Null fixiert ist. Die allgemeinen hyperbolischen und stabilen Verteilungsklassen werden im späteren Verlauf dieser Arbeit behandelt.[7]

[5] Ausführliche Darstellungen sind etwa in (Fang et al., 1990) gegeben. In (Schmidt, 2001) werden theoretische Aussagen zu den Tail-Abhängigkeiten von elliptischen Verteilungen untersucht.

[6] Diese Auflistung ist angelehnt an (Jensen, 1985). $c_i, i = 1, \ldots, 9$ stellen unterschiedliche Normalisierungskonstanten dar.

[7] Die symmetrischen generalisiert hyperbolischen Verteilungen sind in Abschnitt 6.3 als Spezialfall der generalisiert hyperbolischen Verteilungen behandelt.

Verteilung	Dichtegenerator $h(u)$, charakteristische Funktion (c.f.) $\phi(t)$
Kotz-type	$h(u) = c_1(u)^{N-1}\exp(-ru^s)$, $r,s>0, \quad 2N+n>2$
- Multinormal	$h(u) = c_2\exp(-1/2u)$
Pearson-type VII	$h(u) = c_3(1+u/s)^{-N}$, $N > n/2, s>0$
- t	$h(u) = c_4(1+u/s)^{-(n+m)/2}, m \in I\!N$
- Cauchy	$h(u) = c_5(1+u/s)^{-(n+1)/2}, s>0$
Pearson-type II	$h(u) = c_6(1-u/s)^m, m>0$
logistisch	$h(u) = c_7\exp(-u)/[1+\exp(-u)]^2, u \geq 0$
Bessel	$h(u) = c_8(u^{1/2}/\gamma)^a K_a(u^{1/2}/\gamma)$, $a > -n/2, \gamma > 0$, wobei $K_a(\cdot)$ die modifizierte Bessel Funktion dritter Art mit Index a bezeichnet
sym. stabil	$\phi(t) = \exp(r(t^Tt)^{\alpha/2})$, $0 < \alpha \leq 2, r < 0$
sym. gen. hyperb.	$h(u) = c_9 K_{\lambda-n/2}(\alpha\sqrt{1+u})/(1+u)^{n/4-\lambda/2}$, wobei $\lambda, \alpha \in I\!R$ und $K_{\lambda-n/2}(\cdot)$ die modifizierte Bessel Funktion dritter Art mit Index a bezeichnet.

Tabelle 6.2: Teilmengen n-dimensionaler sphärischer Wahrscheinlichkeitsverteilungen

6.1.1 Symmetrische Kotz-type Verteilung

Als erste Klasse der elliptischen Wahrscheinlichkeitsverteilungen sind die *symmetrischen Kotz-type Verteilungen* aufgeführt.[8] Diese Verteilungsklasse fand, wie viele andere Verteilungsfamilien, ihre Motivation in der ungenauen Approximation der Gauß'schen Normalverteilung für die Modellierung gewisser stochastischer Prozesse.[9] Der Dichtegenerator h aus Gleichung (3.7), über den sich eine elliptische Verteilung definiert, ist in diesem Fall beschrieben durch:

$$h(u) = C_n(u)^{N-1}\exp(-ru^s), \quad r,s>0, \quad 2N+n>2, \tag{6.1}$$

[8] Die symmetrischen Kotz-type Verteilungen wurden ursprünglich in (Kotz, 1975) vorgestellt.

[9] Vgl. etwa (Koutras, 1986).

wobei sich

$$C_n = \frac{s\Gamma(n/2)}{\pi^{n/2}\Gamma((2N+n-2)/2s)} r^{(2N+n-2)/2s},$$

die normalisierende Konstante, aus Eigenschaften der sphärischen Verteilungen ableitet.[10]

Die kanonische Dichtefunktion einer multivariaten symmetrischen Kotz-type Verteilung ergibt sich folglich zu:

$$g(z) = C_n[z^T z]^{N-1} \exp(-r[z^T z]^s). \tag{6.2}$$

Die Restriktionen $r > 0$, $s > 0$ und $2N + n > 2$ gelten ebenso.

Die ursprüngliche Kotz-type Verteilung ist durch den Spezialfall $s = 1$ beschrieben.

6.1.1.1 Normalverteilung

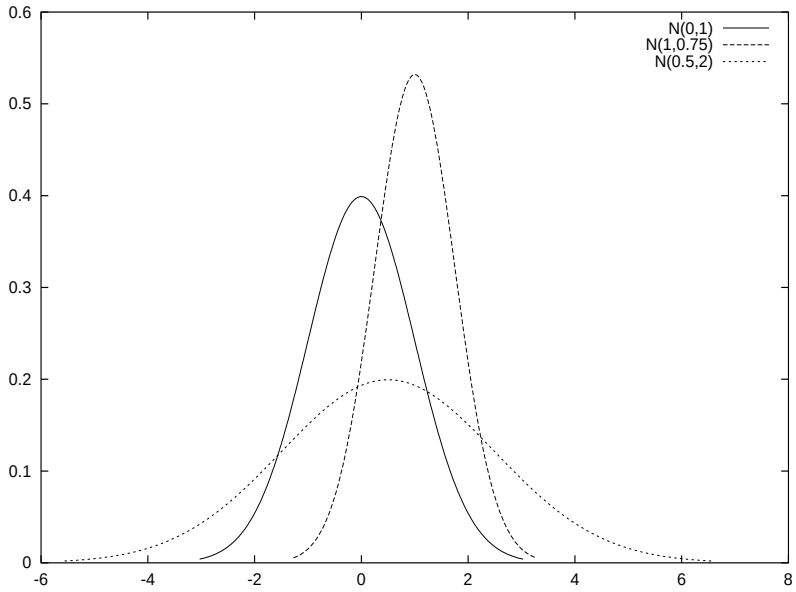


Abbildung 6.1: Beispiele für Dichtefunktionen aus der Klasse der univariaten Gauß'schen Normalverteilungen

Ein weiterer Spezialfall der symmetrischen Kotz-type Verteilung ist die

[10] Siehe dazu (Fang et al., 1990), Seite 26 ff.

wohl bekannteste Verteilungsklasse, die *Gauß'sche Normalverteilung* oder kurz *Normalverteilung*. Diese ergibt sich für eine Parameterkombination von $N = 1$, $s = 1$ und $r = 1/2$. Aufgrund der hohen Praxisrelevanz wird diese Verteilungsfamilie im Folgenden etwas ausführlicher behandelt.

Die Normalverteilung definiert sich über den Dichtegenerator im Sinne der elliptischen Verteilungen durch:

$$h(u) = \frac{1}{\sqrt{(2\pi)^n}} \exp(-\frac{1}{2}u). \tag{6.3}$$

Die parameterfreie kanonische Form der Verteilungsfamilie lässt sich daher sofort formulieren als:

$$g(z) = \frac{1}{\sqrt{(2\pi)^n}} \exp(-\frac{1}{2}z^T z). \tag{6.4}$$

Über den Zusammenhang der Zufallsvariablen Z und Y aus Gleichung (3.3), den Transformationssatz für Lebesque-Integrale[11] und die Bezeichnung $\Sigma^{-1} := A^T A$ ergibt sich die bekannte Dichtefunktion zu:

$$d(y) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp(-\frac{1}{2}[(y-\mu)^T \Sigma^{-1}(y-\mu)]). \tag{6.5}$$

Der Vektor μ und die positiv definite Matrix Σ stellen dabei die bekannten Verteilungsparameter *Erwartungswert* und *Kovarianzmatrix* dar. Offensichtlich existieren bei der Gauß'schen Normalverteilung keine weiteren Formparameter. Daher ist bei der Bezeichnung von normalverteilten Zufallsvariablen eine Vernachlässigung der Generatorfunktion gerechtfertigt. Es wird üblicherweise kurz $Y \sim N_n(\mu, \Sigma)$ geschrieben.

Der direkte und offensichtliche Zusammenhang zwischen dem Dichtegenerator und der kanonischen Form von elliptischen Verteilungen sei an dieser Stelle repräsentativ für die folgenden elliptischen Klassen erläutert. Fortan sind elliptische Verteilungsfamilien daher ausschließlich über ihre kanonische Form präsentiert.

[11] Siehe etwa (Bauer, 2002).

6.1.2 Symmetrische Pearson-type VII Verteilung

Ein n-dimensionaler Zufallsvektor $Y \sim PVII_n(\mu, \Sigma, h)$ ist *Pearson-type VII* verteilt, falls Y eine kanonische Dichtefunktion der Form

$$g(z) = C_n(1 + (z^T z)/m)^{-N} \tag{6.6}$$

besitzt, wobei $N > n/2$, $m > 0$ gilt und C_n folgende Gestalt besitzt:

$$C_n = \frac{\Gamma(N)}{(\pi m)^{n/2}\Gamma(N - n/2)}.$$

Die Familie der Pearson-type VII Verteilungen beinhaltet weitere wichtige Verteilungsklassen. So sind etwa die t-Verteilung und die Cauchy Verteilung Spezialfälle, die in den folgenden Abschnitten aufgeführt sind.

6.1.2.1 t-Verteilung

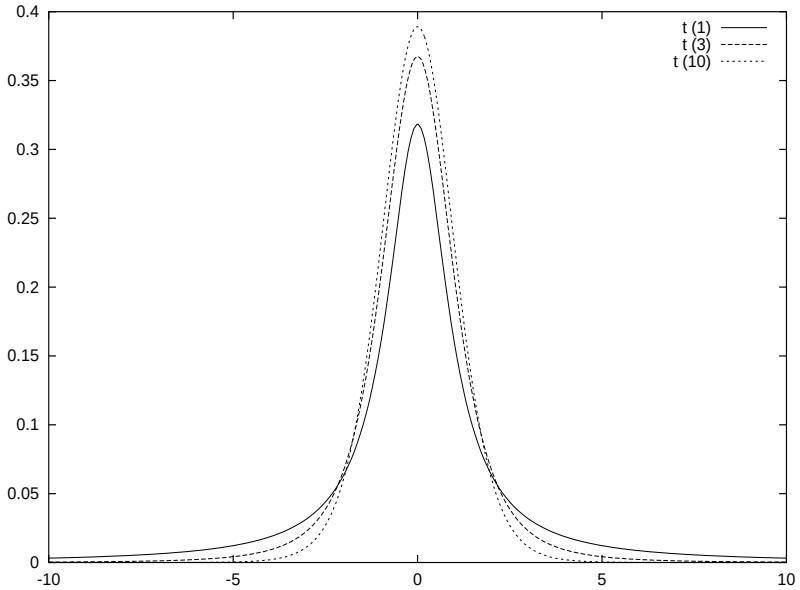


Abbildung 6.2: Beispiele für Dichtefunktionen aus der Klasse der univariaten t-Verteilungen

Falls in Gleichung (6.6) $N = 1/2(n + m)$ und $m \in \mathbb{R}_+$ gilt, so erhält man eine korrespondierende Verteilung, die *multivariate t-Verteilung* genannt

wird. Formal ergibt sich die kanonische Dichtefunktion zu[12]

$$g(z) = \frac{\Gamma((n+m)/2)}{(\pi m)^{n/2}\Gamma(m/2)}(1 + z^T z/m)^{-(n+m)/2}. \tag{6.7}$$

6.1.2.2 Cauchy Verteilung

Die Verteilung $t_n(1, \mu, \Sigma)$ ist auch bekannt als *Cauchy-Verteilung* und wird mit $C_n(\mu, \Sigma)$ bezeichnet. Da die Cauchy-Verteilung ein Spezialfall der t-Verteilung ist, haben die beiden Verteilungsklassen einige Eigenschaften gemein. Es sei an dieser Stelle jedoch bemerkt, dass die Momente der multivariaten Cauchy-Verteilung nicht existieren.[13] Die Darstellung der multivariaten Cauchy-Verteilung durch deren kanonische Dichtefunktion ergibt sich durch:

$$g(z) = \frac{\Gamma((n+1)/2)}{n^{(n+1)/2}}(1 + z^T z)^{-(n+1)/2}. \tag{6.8}$$

6.1.3 Symmetrische Pearson-type II Verteilung

Die kanonische Dichtefunktion der *symmetrischen Pearson-type II Verteilung* als Untermenge der elliptischen Verteilungsklasse besitzt die Gestalt:

$$g(z) = \frac{\Gamma(n/2 + m + 1)}{\pi^{n/2}\Gamma(m+1)}(1 + z^T z)^m, \tag{6.9}$$

wobei $\|z\| \in [0; 1]$ und für m die Restriktion $m > -1$ gilt.[14]

6.1.4 Bessel Verteilung

Eine weitere Teilmenge der elliptischen Verteilungen sind die so genannten *Bessel-Verteilungen*. Die kanonische Dichtefunktion g besitzt bei dieser Ver-

[12] Vgl. etwa (Cornish, 1954), (Dunnett and Strobel, 1954) oder (Laurent, 1955). Für eine alternative Beschreibung eines $t_n(m, \mu, \Sigma)$-verteilten Zufallsvektors Y aus den unabhängigen Zufallsgrößen $T \sim N_n(0, \Sigma)$ und $S \sim \chi_m$ sei etwa auf (Johnson and Kotz, 1972) verwiesen.

[13] Siehe (Fang et al., 1990), Seite 88.

[14] Diese Verteilungsfamilie wurde definiert in (Kotz, 1975). Eine detaillierte Diskussion ist im Artikel (Johnson, 1987) zu finden.

teilungsklasse die Form:

$$g(z) = \frac{(\|z\|/\gamma)^a K_a(\|z\|/\gamma)}{2^{a+n-1}\pi^{n/2}\gamma^n\Gamma(a+n/2)}, \tag{6.10}$$

wobei $a > -n/2$ und $\gamma > 0$ gilt. $K_a(\cdot)$ bezeichnet hierbei die modifizierte Bessel-Funktion 3. Art.[15]

Setzt man $a = 0$ und $\gamma = \delta/\sqrt{2}$ mit $\delta > 0$, so erhält man den Spezialfall einer Bessel-Verteilung, der *Laplace-Verteilung* genannt wird.

6.1.5 Logistische Verteilung

Im Sammelwerk (Kotz et al., 1985) sind unter anderem zwei weitere elliptische Verteilungsklassen notiert. Einerseits die *elliptisch logistische Verteilungsklasse* und andererseits die symmetrisch stabile Verteilungsklasse, die ebenso wie die symmetrischen generalisierten hyperbolischen Verteilungen als Spezialfall ihrer eigenen Klasse behandelt wird.

Falls

$$g(z) = C_n \frac{\exp(-z^T z)}{(1+\exp(-z^T z))^2} \tag{6.11}$$

die kanonische Dichtefunktion einer Zufallsvariablen Z darstellt, so ist diese elliptisch logistisch verteilt. Die Normalisierungskonstante errechnet sich in diesem Fall aus:

$$C_n = \frac{\pi^{n/2}}{\Gamma(n/2)} \int_0^\infty y^{n/2-1} \frac{e^{-y}}{(1+e^{-y})^2} dy.$$[16]

6.2 Stabile Verteilung

Stabile Verteilungen stellen ein breites und interessantes Feld in der Wahrscheinlichkeitstheorie und der Modellierung stochastischer Prozesse dar. Unter anderem fand diese Verteilungsfamilie Anwendung in den Gebieten der

[15] Siehe hierzu etwa (Abramowitz and Stegun, 1972).

[16] Unter der Verwendung von unterschiedlichen Definitionen wurden in (Gumbel, 1962), (Malik and Abraham, 1973) und (Fang and Xu, 1989) multivariate logistische Verteilungen studiert.

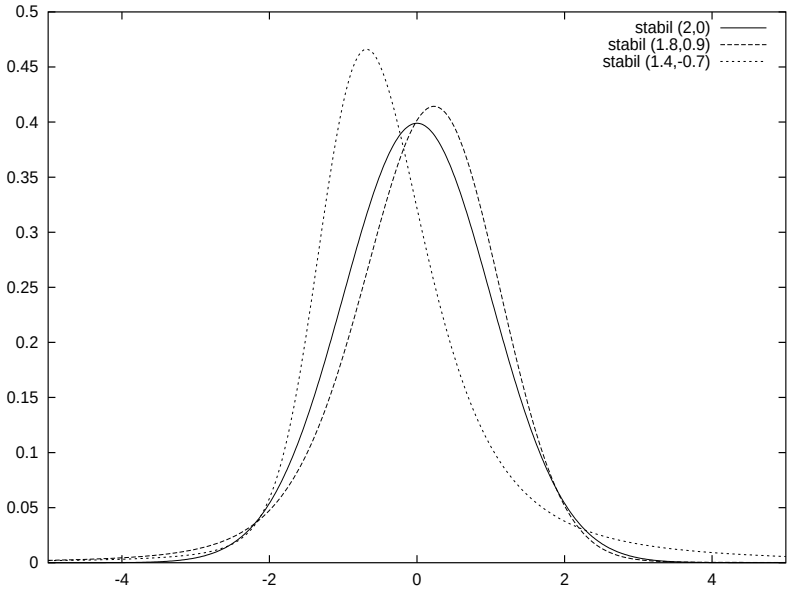


Abbildung 6.3: Beispiele für Dichtefunktionen aus der Klasse der univariaten stabilen Verteilungen

Physik, der Astronomie und der Ökonometrie.[17] Die Klasse der symmetrischen stabilen Verteilungen geht auf (Cauchy, 1853) zurück, wodurch die Cauchy-Verteilung als Spezialfall ihren Namen erhielt. Die Klasse der stabilen Verteilungen wurde später in (Lévy, 1923; Lévy, 1925) ausführlich studiert, weshalb sie in der Literatur auch als *Levy- oder Paretianverteilung* bekannt wurde.[18]

Die Klasse der stabilen Verteilungen erlangte in den letzten Jahren weiteres Ansehen durch die Modellierung von Finanzmarktdaten. Als Mandelbrot eine positive Kurtosis in unterschiedlichen Wertpapierverläufen zeigte, stieß dies das Verwenden nichtnormaler Verteilungsklassen weiter an.[19]

Zum besseren Verständnis werden die univariaten stabilen Verteilungen

[17] Stellvertretend für eine Vielzahl von Anwendungen aus der Ökonometrie seien an dieser Stelle die frühen Arbeiten (Press, 1968), (Roll, 1968) und (Teichmoeller, 1971) zitiert. Für weitere Anwendungen sei auf die Übersicht (Holt and Crow, 1973), Seite 150 verwiesen.

[18] Weitere theoretische Aussagen sind etwa in (Gnedenko and Kolmogorov, 1954) und (Feller, 1966) veröffentlicht.

[19] Siehe (Mandelbrot, 1963).

betrachtet.[20]

Die Klasse der stabilen Verteilungen besitzt vier Parameter. Neben den Lage- und Skalierungsparametern μ und σ existieren zwei weitere Formparameter. Ein erster Formparameter, der den Abfall der Dichtefunktion maßgeblich beeinflusst, wird mit α bezeichnet. In der Literatur wird dieser Parameter oftmals in die Bezeichnung der Verteilungsklasse aufgenommen, so dass die Klasse auch ($\alpha-$)*stabil* genannt wird. Dieser Formparameter nimmt die Werte des Intervalls $(0; 2]$ an. Je kleiner dieser Parameter ist, desto mehr Masse liegt in den Tails der Wahrscheinlichkeitsdichte, d.h. desto dicker sind die Tails der Verteilung und desto wahrscheinlicher das Auftreten von extremen Werten.

Es lässt sich zeigen, dass für den Fall $\alpha \neq 2$ ein asymptotisch polynomialer Abfall vorliegt. Im Fall $\alpha = 2$ und $\beta = 0$ entspricht die stabile Verteilung der Normalverteilung und der Abfall ist folglich exponentiell. Dieser Formparameter wird in der Literatur auch *Stabilitätsindex* oder *charakteristischer Exponent* genannt.

Ein weiterer Formparameter der stabilen Verteilung ist $\beta \in [-1; 1]$, der den Grad der Schiefe bzw. Symmetrie angibt. Die Wahrscheinlichkeitsdichte ist genau dann symmetrisch um den Mittelwert, wenn der Symmetrieparameter $\beta = 0$ ist.[21]

Es lässt sich die ($\alpha-$)stabile Verteilungsklasse wie folgt definieren:[22]

Definition 6.2.1
Eine Zufallsvariable $Y \in \mathbb{R}$ besitzt eine (quasi-)stabile Verteilung, falls für beliebige positive Zahlen A, B und C eine Zahl D existiert, so dass gilt:

$$AY_1 + BY_2 \stackrel{d}{=} CY + D,$$ [23]

wobei Y_1 und Y_2 unabhängige Zufallsvariablen vom gleichen Verteilungstyp

[20] Für den multivariaten Fall sei etwa auf (Press, 1972) und (Rachev and Mittnik, 2000) verwiesen.

[21] Siehe (Holt and Crow, 1973), Seite 148.

[22] Diese Definition ist in Anlehnung an (Samorodnitsky and Taqqu, 1994) formuliert.

[23] $\stackrel{d}{=}$ *bedeutet identisch nach Verteilung.*

wie Y sind.[24] *Die Zufallsvariable Y heißt stabil, falls in Definition 6.2.1 $D = 0$ gilt. Für $D \neq 0$ heißt sie quasi-stabil. Eine (quasi-)stabile Zufallsvariable sei im Folgenden mit*

$$Y \sim S_\alpha(\mu, \sigma, \beta)$$

bezeichnet.

Stabile Verteilungen sind stetig und besitzen eine Dichtefunktion, die jedoch i.Allg. nur aus Termen von unendlichen Reihen ausgedrückt werden kann. Ein wesentlicher Nachteil im Vergleich zu den bislang vorgestellten Verteilungsklassen ist daher, dass die Dichtefunktion einer stabilverteilten Zufallsvariablen i.Allg. keine geschlossene Gestalt im reellen Raum besitzt.

Ein dienlicher Weg zur praktischen und theoretischen Behandlung dieser Klasse von Wahrscheinlichkeitsverteilungen ist die Betrachtung ihrer charakteristischen Funktionen.[25]

Mit Hilfe charakteristischer Funktionen $\phi(t)$ wird eine Familie von Verteilungen genau dann als stabil bezeichnet, wenn sich $\phi(t)$ schreiben lässt als

$$\log \phi(t) = i\mu t - \sigma|t|^\alpha \left[1 + i\beta sign(t) w(t, \alpha)\right], \tag{6.12}$$

wobei

$$sign(t) = \begin{cases} 1, & \text{falls} \quad t > 0 \\ 0, & \text{falls} \quad t = 0 \\ -1, & \text{falls} \quad t < 0 \end{cases}$$

und

$$w(t, \alpha) = \begin{cases} \tan \frac{\pi\alpha}{2}, & \text{falls} \quad \alpha \neq 1 \\ \frac{2}{\pi} \tan |t|, & \text{falls} \quad \alpha = 1 \end{cases}$$

gilt.

[24] *Eine Verteilungsklasse heißt also stabil, wenn sie gegenüber linearen Transformationen abgeschlossen ist, d.h. Linearkombinationen von Zufallsvariablen mit gemeinsamer Verteilungsklasse sind vom gleichen Verteilungstypus.*

[25] Die charakteristische Funktion einer stetigen Zufallsvariable Y ist die Fourier-Stieltjes Transformation der Dichtefunktion und daher definiert als: $\phi(t) = E[e^{itY}] = \int_{-\infty}^{\infty} e^{ity} f(y)\, dy$, wobei $f(y)$ die Dichtefunktion von Y bezeichnet und $i \equiv \sqrt{-1}$ gilt.

Die kanonische Form der Dichte $g(z)$ lässt sich mit Hilfe der Inversionsformel[26] über die charakteristische Funktion (6.12) formulieren als

$$g(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itz} \phi(t)\, dt. \tag{6.13}$$

Gleichung (6.13) zeigt die Notwendigkeit einer *Fast Fourier Transformation* als Hilfsmittel zur Approximation der kanonischen Dichte. Die diskretisierte kanonische Dichtefunktion wird daher für praktische Anwendungen tabellarisch abgespeichert.[27]

Aufgrund obiger Eigenschaft und der beschränkten technischen und zeitlichen Ressourcen ist es nicht möglich, den charakteristischen Exponenten α und den Symmetrieparameter β in das bedingte Konzept aufzunehmen. Sie werden daher für den Fall von stabilen Verteilungen als unbedingt angenommen und direkt geschätzt.

Die folgende Auflistung zeigt die wesentlichen Eigenschaften des Lokationsparameters μ, des Skalenparameters σ und des Symmetrieparameters β von stabilen Verteilungen.[28] Für den Verteilungsparameter α sei auf obige verbale Beschreibung und die anschließende Bemerkung bezüglich der polynomial abfallenden Tails verwiesen.

- Verschiebung:
  Sei $a \in I\!R$ und $Y \sim S_\alpha(\mu, \sigma, \beta)$, dann gilt:

$$Y + a \sim S_\alpha(\sigma, \beta, \mu + a).$$

- Skalierung:
  Sei $a \in I\!R$, $a \neq 0$ und $Y \sim S_\alpha(\mu, \sigma, \beta)$, dann gilt:

$$aY \sim \begin{cases} S_{a\mu,\alpha}(|a|\sigma, sign(a)\beta), & falls \;\; \alpha \neq 1, \\ S_1((a\mu - 2a)/(\pi(ln|a|)\sigma\beta), |a|\sigma, sign(a)\beta), & falls \;\; \alpha = 1. \end{cases}$$

[26] Vgl. etwa (Bauer, 2002).
[27] Vgl. etwa Tabellen in (Holt and Crow, 1973).
[28] Kurze Beweise zu einigen Eigenschaften der Verteilungsparameter sind etwa in (Samorodnitsky and Taqqu, 1994) notiert.

- Symmetrieeigenschaft:
  Mit beliebigem $0 < \alpha < 2$ gilt:

$$Y \sim S_\alpha(0, \sigma, \beta) \Leftrightarrow -Y \sim S_\alpha(0, \sigma, -\beta).$$

Wie schon zu Beginn dieses Abschnitts erwähnt, ist der Abfall der Dichtefunktion einer stabilen Zufallsvariable asymptotisch polynomial. Um diese Eigenschaft formal zu begründen, wird das Verhalten der (α-)stabilen Verteilungen an den Tails $P(Y > \bar{y})$ und $P(Y < -\bar{y})$ für $\bar{y} \to \infty$ untersucht.

Eine Verteilung besitzt asymptotisch polynomial abfallende Tails, falls mit einer langsam variierenden Funktion $L(\bar{y})$[29], einer Konstanten $c \in I\!R$ und $\alpha \in I\!R_+$ folgende Beziehung gilt:[30]

$$P(Y > \bar{y}) \sim c\bar{y}^{-\alpha} L(\bar{y}).^{31}$$

Für die Tails einer stabilen Verteilung gilt mit $0 < \alpha < 2$:[32]

$$\begin{aligned} \lim_{\bar{y}\to\infty} \bar{y}^\alpha P(Y > \bar{y}) &= \frac{C_\alpha(1+\beta)}{2\sigma^\alpha} \\ \lim_{\bar{y}\to\infty} \bar{y}^\alpha P(Y < -\bar{y}) &= \frac{C_\alpha(1-\beta)}{2\sigma^\alpha}, \end{aligned}$$

wobei

$$C_\alpha = \left(\int_0^\infty \bar{y}^{-\alpha} \sin \bar{y}\, d\bar{y} \right)^{-1} = \begin{cases} \frac{(1-\alpha)}{\Gamma(2-\alpha)\cos(\frac{\pi\alpha}{2})}, & \text{für} \quad \alpha \neq 1 \\ \frac{2}{\pi}, & \text{für} \quad \alpha = 1. \end{cases}$$

Offensichtlich ist für $L(\bar{y}) \equiv 1$ und $c = \frac{C_\alpha(1+\beta)}{2\sigma^\alpha}$ bzw. $c = \frac{C_\alpha(1-\beta)}{2\sigma^\alpha}$ die asymptotische Polynomialität für die Klasse der stabilen Verteilungen gezeigt.

Auf weitere Eigenschaften der stabilen Verteilungsklasse wie unendliche

[29] $L(\bar{y})$ ist eine langsam variierende Funktion, falls gilt: $\lim_{\bar{y}\to\infty} \frac{L(t\bar{y})}{L(\bar{y})} = 1 \ \ \forall t > 0$.
[30] Vgl. (Embrechts et al., 2001).
[31] $\sim$ bedeutet in diesem Zusammenhang asymptotisch gleich.
[32] Siehe (Samorodnitsky and Taqqu, 1994).

Teilbarkeit, Unimodalität und die Berechnung von Momenten sei an dieser Stelle verzichtet, da diese für die hier vorliegenden Zwecke keine direkte Relevanz besitzen.[33]

6.3 Generalisiert hyperbolische Verteilung

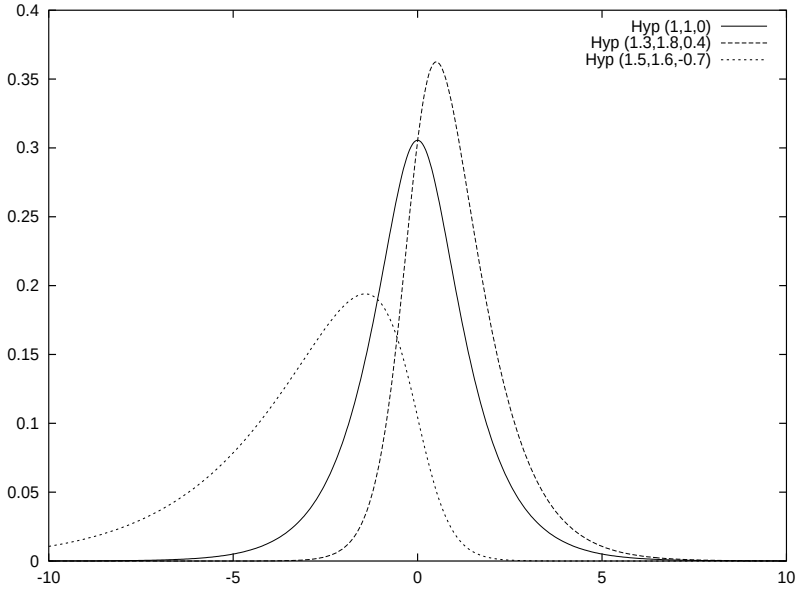


Abbildung 6.4: Beispiele für Dichtefunktionen aus der Klasse der univariaten generalisiert hyperbolischen Verteilungen

Im Jahr 1977 wurden von Barndorff-Nielsen die hyperbolischen Verteilungen vorgestellt.[34] Die multivariate hyperbolische Verteilungsklasse ergibt sich als Varianz-Erwartungswert-Mixtur von n-dimensionalen Normalverteilungen mit Lageparameter $\bar{\mu} \in I\!R^n$, Drift $\bar{\beta}$ und Strukturmatrix $\bar{\Sigma}$. Die Mischverteilung ist die generalisiert inverse Gauß'sche Verteilung.[35]

[33] Siehe hierzu etwa (Holt and Crow, 1973).

[34] Siehe (Barndorff-Nielsen, 1977).

[35] Sei Y eine n-dimensionale Zufallsvariable. Für ein gegebenes $u \geq 0$ folgt Y einer n-dimensionalen Normalverteilung mit Kovarianzmatrix $u\bar{\Sigma}$ und Erwartungswertvektor $\bar{\mu} + u\bar{\beta}$, wobei $\bar{\Sigma}$ eine positiv definite Matrix mit Determinate 1 bezeichnet und $\bar{\mu} \in I\!R^n$ und $\bar{\beta} \in I\!R^n$ konstante Vektoren sind. Sei ferner angenommen, dass $U = u$ eine Verteilungsfunktion F auf $[0; \infty)$ besitzt. Dann heißt die Verteilung von Y *normale Varianz-Erwartungswert-*

Da die Klasse der multivariaten hyperbolischen Verteilungen nicht gegen die Bildung von Randverteilungen und regulären affinen Transformationen abgeschlossen ist, präsentierte Barndorff-Nielsen eine diesbezüglich abgeschlossene Erweiterung, die generalisierten hyperbolischen Verteilungen mit dem zusätzlichen Generalisierungsparameter λ.[36] Die endlich dimensionalen Randverteilungen der generalisiert hyperbolischen Verteilungsfamilie verbleiben in derselben Klasse. Für den Spezialfall $\lambda = 1$ heißen die eindimensionalen Randverteilungen hyperbolisch.

Der Name der Verteilungsfamilie rührt von der Form der logarithmierten Dichtefunktion, da waagerechte Schnitte Hyperboloide darstellen.[37] Die theoretischen mathematischen Eigenschaften der generalisiert hyperbolischen Verteilungsklasse sind in der Literatur detailliert ausgearbeitet und präsentiert.[38]

Selbst wenn die Klasse der hyperbolischen Verteilungen, mit ihren flexiblen Eigenschaften noch keinen sehr großen Bekanntheitsgrad besitzt, so erzielte sie in vielen unterschiedlichen wissenschaftlichen Bereichen doch schon beträchtliche Erfolge. Die hyperbolische Verteilungsklasse wurde bereits zur Modellierung der Partikelgröße von Sandablagerungen und Luftturbulenzen angewendet.[39]

Während der letzten Jahre fand die Klasse der hyperbolischen Verteilungen, aufgrund ihrer flexiblen Eigenschaften, in zahlreichen finanzmathematischen Arbeiten Verwendung.[40]

Ein n-dimensionaler Zufallsvektor Y heißt generalisiert hyperbolisch verteilt mit Lokation $\mu \in I\!R^n$ und Skalierung $\Sigma \in I\!R^{n \times n}$, falls der stochastische

Mixtur mit Lageparameter $\bar{\mu}$, Drift $\bar{\beta}$, Strukturmatrix $\bar{\Sigma}$ und Mischverteilung F. Zur detaillierten Herleitung sei auf (Barndorff-Nielsen et al., 1982) verwiesen.

[36] Siehe (Barndorff-Nielsen, 1977).

[37] Im Vergleich formen die log-normalen Dichtefunktionen einen Paraboloid und die waagerechten Schnitte der Dichtefunktion von elliptischen Verteilungsfamilien bilden Ellipsen.

[38] Vgl. etwa (Barndorff-Nielsen and Blæsild, 1981).

[39] Siehe (Barndorff-Nielsen, 1977), (Barndorff-Nielsen et al., 1985) und (Barndorff-Nielsen et al., 1989). Weitere praktische Anwendungen sind etwa in (Bagnold and Barndorff-Nielsen, 1980) veröffentlicht.

[40] Siehe etwa (Eberlein and Keller, 1995), (Rydberg, 1996), (Keller, 1997), (Prause, 1997) (Eberlein et al., 1998), (Eberlein and Prause, 1999), (Eberlein, 1999), (Prause, 1999), (Eberlein and Prause, 2000) und (Blingham et al., 2002).

Zusammenhang $Y \stackrel{d}{=} \mu + A^{-1}Z$ aus Gleichung (3.2) bzw. $Z \stackrel{d}{=} A(Y - \mu)$ gilt und Z die folgende Dichtefunktion besitzt:[41]

$$g(z) = a_n \frac{K_{\lambda - n/2}(\alpha\sqrt{1 + z^T z})}{\sqrt{1 + z^T z}^{n/2-\lambda}} e^{\alpha\beta^T z}. \tag{6.14}$$

a_n bezeichnet die Normalisierungskonstante und besitzt die Form

$$a_n = \frac{\sqrt{\alpha^2 - \beta^T\beta}^{\lambda}}{(2\pi)^{n/2} K_\lambda(\sqrt{\alpha^2(1 - \beta^T\beta)})\alpha^{n/2-\lambda}},$$

wobei K_ν die modifizierte Besselfunktion dritten Grades mit dem Index ν darstellt.[42] Hierbei sind die Verteilungsparameter restringiert durch $\|\beta\| < 1$, $\alpha > 0$ und $\lambda \in I\!R$. Es gilt die Bezeichnung $Z \sim Hyp(\alpha, \beta, \lambda)$ bzw. $Y \sim Hyp(\mu, \Sigma, \alpha, \beta, \lambda)$.

Diese Parametrisierung hat die positive Eigenschaft, invariant gegenüber affinen Transformationen zu sein. Weiterhin beinhaltet die breite Klasse der generalisiert hyperbolischen Verteilungen eine Vielzahl von bekannten Verteilungen.[43]

- Für den Fall $\beta = (0, \ldots, 0)^T$ zählen die symmetrischen multivariaten hyperbolischen Verteilungen zu der oben behandelten Klasse der elliptischen Verteilungen. Die Dichtefunktion von Y kann sodann dargestellt werden durch

  $$f(y) = |\Sigma|^{-1/2} h((y - \mu)^T \Sigma^{-1}(y - \mu)).$$

  Die beliebige Funktion $h : I\!R_0^+ \to I\!R_0^+$ heißt *Dichtegenerator*.[44]

- Falls der Generalisierungsparameter λ den Wert $(n+1)/2$ annimmt, ist die Familie der multivariaten hyperbolischen Verteilungen dargestellt.

---

[41] $\stackrel{d}{=}$ bedeutet identisch nach Verteilung und A bezeichnet die obere Dreiecksmatrix der Cholesky-Zerlegung von $\Sigma = A^T A$.

[42] Zur Diskussion der Bessel-Funktion, auch bekannt als Hankelfunktion oder MacDonald'sche Funktion, sei etwa auf (Abramowitz and Stegun, 1972) verwiesen.

[43] Siehe zum Teil Tabelle 6.1.

[44] Vgl. Gleichung 3.7 in Abschnitt 3.2.

- Die Darstellung von multivariaten inversen Gaußverteilungen ergibt sich durch $\lambda = -1/2$.[45]

- Im univariaten Fall ergibt sich die Normalverteilung als eine Grenzverteilung der generalisiert hyperbolischen Verteilung.[46]

Die multivariate generalisiert hyperbolische Verteilungsklasse ist häufig in der Literatur für $y \in I\!R^n$ über die Lebesque-Dichte wie folgt definiert:[47]

$$d(y) = a_n \frac{K_{\bar{\lambda}-n/2}(\bar{\alpha}\sqrt{\bar{\delta}^2 + (y-\bar{\mu})^T\bar{\Sigma}^{-1}(y-\bar{\mu})})}{(\bar{\alpha}^{-1}\sqrt{\bar{\delta}^2 + (y-\bar{\mu})^T\bar{\Sigma}^{-1}(y-\bar{\mu})})^{n/2-\bar{\lambda}}} e^{\bar{\beta}^T(y-\bar{\mu})}$$

mit

$$a_n = \frac{(\sqrt{\bar{\alpha}^2 - \bar{\beta}^T\bar{\Sigma}\bar{\beta}}/\bar{\delta})^{\bar{\lambda}}}{(2\pi)^{n/2}K_{\bar{\lambda}}(\bar{\delta}\sqrt{\bar{\alpha}^2 - \bar{\beta}^T\bar{\Sigma}\bar{\beta}})},$$

wobei die Verteilungsparameter den folgenden Restriktionen unterliegen: $\bar{\lambda} \in I\!R$, $\bar{\alpha} \in I\!R^+$, $\bar{\beta}, \bar{\mu} \in I\!R^n$, $\bar{\delta} > 0$, $\bar{\beta}^T\bar{\Sigma}\bar{\beta} < \bar{\alpha}^2$ und $\bar{\Sigma} \in I\!R^{n\times n}$ bezeichnet eine positiv definite Matrix mit Determinante 1 ($|\bar{\Sigma}| = 1$).[48] K_ν repräsentiert wiederum die modifizierte Besselfunktion dritten Grades mit dem Index ν.

Diese Parametrisierung enthält einen weiteren Skalierungsfaktor $\bar{\sigma}$, wodurch die Invarianz gegenüber affinen Transformationen verloren geht. Eine weitere alternative Parametrisierung kann über den folgenden Zusammenhang $\tilde{\lambda} = \bar{\lambda}$, $\tilde{\mu} = \bar{\mu}$, $\tilde{\beta} = \bar{\beta}$, $\tilde{\chi} = \bar{\delta}^2$, $\tilde{\psi} = \bar{\alpha}^2 - \bar{\beta}^T\bar{\Sigma}\bar{\beta}$ und $\tilde{\Sigma} = \bar{\Sigma}$ definiert werden.[49]

[45] Vgl. (Eberlein and Prause, 2000); zur formalen Darstellung der inversen Gaußverteilung sei z.B. auf (Prause, 1999) verwiesen.

[46] Weitere Spezialfälle und Grenzverteilungen sind z.B. die multivariate t-Verteilung oder die Cauchy-Verteilung. Für detaillierte Ausführungen bezüglich der erforderlichen Mischverteilung und der korrespondierenden Parameterkombination sei auf (Barndorff-Nielsen, 1978), (Barndorff-Nielsen et al., 1982) und die Tabelle 1.1 in (Prause, 1999) verwiesen. In der Arbeit von (Prause, 1999) wird die Klasse der generalisiert hyperbolischen Verteilungen ausführlich diskutiert. Daher wird hier auf weitere Charakterisierungen verzichtet. In (Prause, 1999) finden sich unter anderem Berechnungen der Momente, der charakteristischen Funktion, der logarithmierten Likelihoodfunktion und deren partiellen Ableitungen.

[47] Vgl. etwa (Eberlein and Keller, 1995).

[48] Bei dieser Darstellung werden die Grenzverteilungen, die an den Rändern des Parameterraums entstehen, vernachlässigt; siehe z.B. (Blæsild and Jensen, 1981).

[49] Diese Parametrisierung ist etwa in (Atkinson, 1982) für den dort entwickelten Zufallsgenerator verwendet.

Die bijektive Abbildung obigen Parametervektors $(\tilde{\mu}, \tilde{\Sigma}, \tilde{\chi}, \tilde{\psi}, \tilde{\beta}, \tilde{\lambda})$ bzw. der herkömmlichen Parametrisierung $(\bar{\mu}, \bar{\delta}, \bar{\Sigma}, \bar{\alpha}, \bar{\beta}, \bar{\lambda})$ auf die Darstellung obiger Definition $(\mu, \Sigma, \alpha, \beta, \lambda)$ ist definiert durch

$$\begin{aligned}
\mu &= \bar{\mu} = \tilde{\mu} \\
\Sigma &= \bar{\delta^2}\bar{\Sigma} = \tilde{\chi}\tilde{\Sigma}, \quad \delta^2 = \tilde{\chi} \\
\alpha &= \bar{\alpha}\bar{\delta} = \sqrt{(\tilde{\psi} + \tilde{\beta}^T\tilde{\Sigma}\tilde{\beta})\tilde{\chi}} \\
\beta &= 1/\bar{\alpha}A^T\bar{\beta} = 1/\sqrt{\tilde{\psi} + \tilde{\beta}^T\tilde{\Sigma}\tilde{\beta}}A^T\tilde{\beta}, \quad A^TA = \tilde{\Sigma} \\
\lambda &= \bar{\lambda} = \tilde{\lambda}.
\end{aligned}$$

Über die Parametrisierung der Darstellung (6.14) und die vereinfachende Definition $R_{\lambda,i}(x) := \frac{K_{\lambda+i}(x)}{x^i K_\lambda(x)}$ ergeben sich der Mittelwertvektor und die Kovarianzmatrix einer multivariaten generalisiert hyperbolischen Zufallsvariablen Y zu:[50]

$$E[Y] = \mu + R_{\lambda,1}(\sqrt{\alpha^2(1-\beta^T\beta)})\ \Sigma\alpha A^{-T}\beta$$

und

$$\begin{aligned}
Cov[Y] &= R_{\lambda,1}(\sqrt{\alpha^2(1-\beta'\beta)})\ \Sigma \\
&+ \left[R_{\lambda,2}(\sqrt{\alpha^2(1-\beta'\beta)}) - R^2_{\lambda,1}(\sqrt{\alpha^2(1-\beta'\beta)})\right] \\
&\quad \frac{\Sigma A'^{-1}\beta(A'^{-1}\beta)'\Sigma}{1-\beta'\beta}
\end{aligned}$$

Für den speziellen symmetrischen Fall ($\beta = 0$) mit dem Generalisierungsparameter $\lambda = 1$ ergeben sich der Erwartungswertvektor und die Kovarianzmatrix zu $E[Y] = 0$ und $Cov[Y] = K_2(\alpha)/(\alpha K_1(\alpha))\ \Sigma$.

6.4 Endliche Mixturverteilung

Eine weitere und seit langem bekannte Möglichkeit, flexible Verteilungsfamilien mit asymmetrischer Form und einer Kurtosis ungleich Null zu erhalten,

[50] Zum Beweis sei auf (Schmidt et al., 2003) verwiesen.

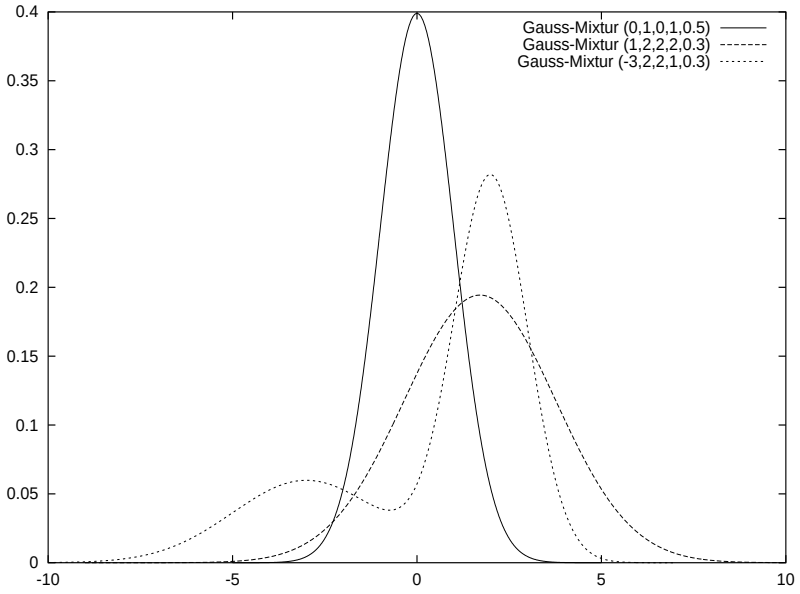


Abbildung 6.5: Beispiele für Dichtefunktionen aus der Klasse der univariaten binären Gauß'schen Mixtur-Verteilungen

ist die Bildung von Mixturverteilungen.[51] Schon vor über einhundert Jahren studierte Pearson die Anfänge der Mixturverteilungen, indem er zwei Normalverteilungen linear kombinierte.[52] Kombinationen von Exponential-, Poisson-, Weibull- und Binomialverteilungen wurden etwas später behandelt.[53]

Wie bereits erwähnt sind Mixturdichtefunktionen parametrische Approximationen von beliebigen Wahrscheinlichkeitsdichten, so dass durch diese Klasse die nichtparametrischen Verteilungen in diesem Konzept repräsentiert sind.

Da die herkömmlichen statistischen Analysen nicht generell unmittelbar auf die Mixturverteilungen anzuwenden sind und oftmals keine geschlossene Form für die Schätzer der Verteilungsparameter existiert, sind die Mixturverteilungen bis zum Zeitalter der Computer und numerischen Lösungsverfahren

[51] Vgl. etwa (Redner and Walker, 1984), (Titterington et al., 1985) oder (McLachlan and Basford, 1988).

[52] Siehe (Pearson, 1894). Gauß'sche Mixturverteilungen gehören zu der in Abschnitt 6.1 behandelten Klasse der elliptischen Verteilungen, siehe (Fang et al., 1990), Seite 48.

[53] Siehe unter anderem (Rider, 1962), (Cohen Jr., 1964) und (Blischke, 1955).

in den Hintergrund gerückt.[54] Die Renaissance der Mixturverteilungen ist jedoch in den letzten Jahren nicht zu verbergen. Es sind zu viele praktische Anwendungen, in denen die Mixturverteilungen zum Einsatz kamen, um sie alle aufzuzählen.[55] Es sei an dieser Stelle das Anwendungsfeld der Finanzmathematik ein weiteres Mal gesondert betont, nicht nur, weil Mixturverteilungen in vielen Modellierungsansätzen Verwendung finden, sondern da die aktuellen Forschungsarbeiten in diesem Bereich zu dem thematisch angrenzenden Gebiet dieser Arbeit zählen.[56]

Es sei eine Definition der Mixturverteilung vorgestellt, bevor die in diesem Zusammenhang für die Praxis relevante Form der binären Gauß'schen Mixturverteilung spezifiziert wird.

Sei Y ein n-dimensionaler reellwertiger Zufallsvektor, dessen Verteilung durch die Wahrscheinlichkeitsdichte

$$d(y) = q_1 d_1(y) + \cdots + q_R d_R(y) \tag{6.15}$$

repräsentiert ist, wobei

$$\sum_{i=1}^{R} q_i = 1 \quad \text{mit } q_i > 0,\ i = 1, \ldots, R \tag{6.16}$$

und

$$d_i \geq 0, \quad \int_{I\!R^n} d_i(y)\, dy = 1, \quad \text{für } i = 1, \ldots, R \tag{6.17}$$

gilt. Unter diesen Voraussetzungen besitzt Y eine *endliche Mixturverteilung*

[54] Vgl. (Titterington et al., 1985), Seite ix.

[55] Ein sehr ausführlicher Überblick ist etwa in (Titterington et al., 1985) Kapitel 2 gegeben.

[56] Es sei an dieser Stelle in aller Kürze auf einige Literaturstellen des letzten Jahrzehnts verwiesen. Die Artikel (Hamilton, 1991), (Jacobs et al., 1991), (Nowlan, 1991), (Bishop, 1994), (Nabney et al., 1995), (Ormoneit and Neuneier, 1996), (Schittenkopf et al., 1998), (Vlassis et al., 1999), (Vlassis and Kröse, 1999), (Schittenkopf et al., 1999), (Schittenkopf et al., 2000), (Vlassis et al., 2000), (Bartlmae and Rauscher, 2000) und (Weigend and Shi, 2000), die dem Wissenschaftsgebiet der Neuroinformatik zuzuordnen sind, befassen sich mit der Modellierung von Finanzwerten unter Zuhilfenahme von Gauß'schen Mixturverteilungen. Details zu diesen Ansätzen sind in Abschnitt 7.2 ausgeführt. Es werden dort außerdem die einschränkenden Annahmen und Parallelen zu der hier entwickelten Methodik diskutiert und in den Tabellen 7.2, 7.3 und 7.4 zusammengefasst.

und die Funktion $d(y)$ heißt *endliche Mixturdichte*. Die Parameter $q_1, \ldots, q_R$ werden *Mixturkoeffizienten* und $d_1, \ldots, d_R$ *Komponenten der Mixturdichte* genannt. Offensichtlich erfüllt $d(y)$ als gewichtete Summe von Wahrscheinlichkeitsdichten durch die Bedingungen (6.16) und (6.17) die Anforderungen einer Wahrscheinlichkeitsdichte. Überdies können die Mixturkoeffizienten durch die Restriktion (6.16) als a priori Wahrscheinlichkeiten interpretiert werden.

Die Komponenten $d_i, i = 1, \ldots, R$ stellen beliebige Wahrscheinlichkeitsdichten dar. Im folgenden soll jedoch von parametrischen Formen der Komponenten ausgegangen werden.

Für die Motivation von Mixturverteilungen existieren in der Literatur zwei Interpretationen, die beide ihre Rechtfertigung besitzen. Diese werden mit *direkter* und *indirekter Anwendung* bezeichnet.[57]

Im Fall der direkten Anwendungen liegt die Annahme zugrunde, dass die Beobachtung x zu einer von R Kategorien gehört bzw. einem von R a priori definierten Cluster zugewiesen werden kann. In Folge dieser Interpretation bezeichnet d_i die Wahrscheinlichkeitsdichte von X unter der Voraussetzung, dass die Beobachtung x der i-ten Kategorie angehört. In diesem Fall bezeichnet q_i die Wahrscheinlichkeit, dass die Beobachtung x ein Element dieses i-ten Clusters ist.[58]

Bei indirekten Anwendungen handelt es sich um die Modellierung von Zufallsgrößen mit Hilfe von Mixturverteilungen, bei denen ein „mathematischer Kunstgriff“[59] ausgenutzt wird, um bestimmte flexible statistische Eigenschaften der Verteilungsklasse zu erhalten. Die hier schon häufig dargestellte Sichtweise und Motivation[60] zu alternativen Verteilungsklassen folgt daher der zweiten Art der Anwendung von Mixturverteilungen.

Ein in der Literatur häufig diskutiertes Thema ist die zu wählende Kom-

[57] Vgl. etwa (Titterington et al., 1985).

[58] In (McLachlan and Basford, 1988) ist die Motivation und Begründung für die Verwendung von Mixturverteilungen stark auf dieser Art der Interpretation aufgebaut.

[59] Vgl. (Titterington et al., 1985).

[60] Vgl. die Eigenschaft der heavy tails und einer Kurtosis ungleich Null als einen Motivationsgrund für generalisiert hyperbolische oder stabile Verteilungen und der Darstellung in Abschnitt 2.2.2. Es seien jedoch hier die exponentiell abfallenden Tails im Gegensatz zu den hyperbolischen Verteilungen genannt.

ponentenanzahl der Mixturverteilung. Diese Thematik ist nah mit der aus der Clusteranalyse bekannten Wahl der Clusteranzahl verwandt. Dies rechtfertigt jedoch eine geringe Aufmerksamkeit für diese Problematik wegen der unterschiedlichen Anwendungsabsicht. Die Suche nach der Anzahl von Komponenten ist vergleichbar mit der Festlegung der lag-Parameter in der Zeitreihenanalyse oder der Netzarchitektur von neuronalen Netzen bzw. der Festlegung der polynomialen Form bei der Funktionsapproximation. Die Gefahr des „Overfittings“[61] ist an dieser Stelle nicht zu vernachlässigen. Bei der Bestimmung der Komponentenanzahl wird häufig auf das Mittel der statistischen Tests verwiesen.[62] Es sei jedoch gleichzeitig auf die damit verbundenen Probleme hingewiesen.[63]

Für diese Arbeit soll die Beschränkung auf den in der Praxis am häufigsten verwendeten Fall der binären Gauß'schen Mixtur betrachtet werden. Hierbei werden zwei i.Allg. unterschiedliche Gauß'sche Normalverteilungen $N(\mu_1, \Sigma_1)$ und $N(\mu_2, \Sigma_2)$ mit dem Gewichtungsparameter q angenommen.[64] Die Wahrscheinlichkeitsdichte ergibt sich in diesem Spezialfall ($R = 2$) zu:

$$d(y) = qd_1(y) + (1 - q)d_2(y),$$

wobei mit der bekannten Dichtefunktion aus Gleichung (6.5)

$$d_i(y) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_i|}} \exp(-\frac{1}{2}[(y - \mu_i)^T \Sigma_i^{-1} (y - \mu_i)]), \quad i = 1, 2$$

gilt. Die kanonische Dichtefunktion folgt auf analoge Weise mit Hilfe der Darstellung (6.4).

[61] Dieser Begriff stammt aus der künstlichen Intelligenz und bezeichnet die funktionale Abbildung des Zufalls und den damit verbundenen Verlust der Generalisierungsfähigkeit, vgl. die ähnlichen Erläuterungen in Abschnitt 4.2.2.

[62] Siehe (Titterington et al., 1985).

[63] Neuere Arbeiten (z.B. (Vlassis et al., 2000)) greifen diese Methodik auf und formulieren statistische, auf der geschätzten Kurtosis basierende Tests. Diese Vorgehensweise ist jedoch für die in der Praxis häufig auftretenden - hier behandelten - Fälle von bedingten multivariaten Verteilungsklassen nicht anwendbar.

[64] Mixturverteilungen aus den oben beschriebenen Verteilungsklassen lassen sich analog behandeln. Mixturverteilungen aus diskreten Verteilungsklassen wurden schon in (Davis, 1952), (Aitchison, 1955), (Cohen Jr., 1965) oder (Ashton, 1971) behandelt.

Eine Eigenschaft der Mixturverteilungen, im Gegensatz zu den bisher betrachteten Verteilungsklassen, ist die mögliche Bi- oder Multimodalität.[65] Für die hier vorliegenden Betrachtungen reicht eine Beschränkung auf die eingipfligen Gauß'schen Mixturdichten, da Prognoseaufgaben normalerweise keine Clusteraufgaben darstellen. Hierdurch lassen sich sinnvolle Aussagen über die Momente der Verteilung formulieren.

Die Linearitätseigenschaft des Integrals ermöglicht es, den Erwartungswert einer binären Gauß'schen Mixturverteilung μ_{res} direkt zu formulieren als gewichtete Summe der einzelnen Komponenten:

$$\mu_{res} = q\mu_1 + (1-q)\mu_2.$$

Die Varianz hingegen hängt nicht ausschließlich von den einzelnen Varianzen der Verteilungskomponenten ab, sondern vergrößert sich, je weiter die Mittelwerte voneinander entfernt liegen. Die resultierende Varianz einer mixturverteilten Zufallsvariablen σ_{res} folgt der Vorschrift:

$$\sigma_{res}^2 = q\sigma_1^2 + (1-q)\sigma_2^2 + q(1-q)(\mu_1-\mu_2)^2. \tag{6.18}$$

Abrundend zur Beschreibung von Mixturverteilungen wird die erwähnte Eigenschaft der Schiefe und des Exzesses von univariaten unimodalen Gauß'schen Mixturverteilungen formal charakterisiert.

Die Schiefe s einer binären univariaten Gauß'schen Mixturverteilung errechnet sich aus der Gleichung:

$$s = \frac{q(1-q)(\mu_1-\mu_2)[3(\sigma_1^2-\sigma_2^2)+(1-2q)(\mu_1-\mu_2)]}{(\sqrt{q\sigma_1^2+(1-q)\sigma_2^2+(q-q^2)(\mu_1-\mu_2)^2})^3}. \tag{6.19}$$

Für identische Erwartungswerte $\mu_1 = \mu_2$ folgt offensichtlich eine symmetrische Verteilungsform. Falls jedoch $\mu_1 \neq \mu_2$ gilt, ergibt sich eine symmetrische Verteilung genau dann, falls $\sigma_1 = \sigma_2$ und $q = 1/2$ gilt, da $0 < q < 1$ vor-

[65] Bi- bzw. Multimodalität liegt vor, falls die Wahrscheinlichkeitsdichte zwei bzw. mehrere lokale Maxima besitzt. Besitzt die Dichtefunktion ein Maximum, so heißt sie unimodal. In (Titterington et al., 1985) ist die Bi- bzw. Multimodalität von Mixturverteilungen ausführlich diskutiert.

ausgesetzt werden kann.[66] In allen weiteren Fällen liegt eine Schiefe ungleich Null vor.

Hingegen ist der Exzess einer univariaten unimodalen Gauß'schen Mixturverteilung genau dann gleich Null, falls identische Varianzen der einzelnen Komponenten vorliegen. Über die zentralen Momente[67] lässt sich der Exzess einer Gauß'schen mixturverteilten Zufallsvariablen durch folgenden Ausdruck formalisieren:

$$\begin{aligned} e &= \frac{3(q\sigma_1^4 - (1-q)\sigma_2^4) + 6(\mu_1-\mu_2)^2(\sigma_1^2-\sigma_2^2)(q^2-q^3)}{q(\sigma_1^2-\sigma_2^2)+\sigma_2^2+(\mu_1-\mu_2)^2(1-q)} \\ &+ \frac{(q-4q^2+6q^3-3q^4)(\mu_1-\mu_2)^4}{q(\sigma_1^2-\sigma_2^2)+\sigma_2^2+(\mu_1-\mu_2)^2(1-q)}. \end{aligned} \tag{6.20}$$

Sei $\mu_1 = \mu_2$ und $\sigma_1 = 1$, so vereinfacht sich Gleichung (6.20) zu:

$$e = \frac{3q(1-q)(1-\sigma_2^2)^2}{(q+(1-q)\sigma_2^2)^2}$$

und es stellt sich für jedes $\sigma_2 \neq 1$ ein Exzess ungleich Null ein.

[66] Es sei ohne Einschränkung der Allgemeinheit $0 < q < 1$, da für $q = 0$ oder $q = 1$ eine klassische Gaußverteilung vorliegt.

[67] Siehe z.B. (Cohen Jr., 1965).

Kapitel 7

Inklusion bekannter Prognosekonzepte aus der Statistik und der Neuroinformatik

Nachdem in den vorangegangenen Kapiteln ein Prognosekonzept mit allen benötigten Hilfsmitteln methodisch präsentiert werden konnte, das den einzelnen Anforderungen und Motivationsaspekten entspricht, gilt es nun diese Methodik gegenüber bekannten Prognoseverfahren abzugrenzen. Der Inhalt dieses Kapitels wird jedoch weniger eine klassische theoretische Abgrenzung sein, als vielmehr die Inklusion bekannter Verfahren aus der Statistik und der Neuroinformatik, um den abgedeckten Umfang der Verteilungsmethodik darzustellen und gleichzeitig den Mehrwert des vorgestellten Systems zu betiteln und zu würdigen.

Es wird daher gezeigt, dass die in dieser Arbeit entwickelte Methodik sowohl Verfahren aus der klassischen Statistik wie lineare Modelle oder Zeitreihenmodelle als auch Prognosemethoden aus der künstlichen Intelligenz, wie neuronale Netze und deren stochastische Erweiterungen umfasst.

Es wird nicht auf die breiten theoretischen Hintergründe der einzelnen statistischen Prognosemodelle bzw. -methoden eingegangen. Eine Darstellung

der für die Inklusion relevanten Eigenschaften ist für den hier vorliegenden Zweck suffizient.

Durch diese Inklusion und Abgrenzung existierenden Prognosemethoden werden die wissenschaftlichen und praktischen Zielsetzungen des hier entwickelten Konzepts deutlich. An einigen Stellen werden Erweiterungen und wissenschaftliche Erkenntnisse explizit formuliert.

7.1 Klassische statistische Modelle

Unter klassischen statistischen Prognosemodellen bzw. Prognosemethoden werden in diesem Zusammenhang lineare Modelle und Zeitreihenmodelle verstanden. Obwohl viele Methoden in der Literatur strikt den Zeitreihenmodellen zugeordnet sind, können diese als Regressionsansätze verstanden und interpretiert werden.[1] Es sei an dieser Stelle betont, dass in dem vorliegenden Zusammenhang die funktionale Modellierung im Mittelpunkt der Betrachtungen steht, d.h. der aus realen Daten zu identifizierende funktionale Zusammenhang zwischen den exogenen und endogenen Variablen. Der Fokus liegt daher ebenso wenig auf „datenvorverarbeitenden“ Schritten, da diese im wesentlichen bei allen Ansätzen identisch vorgenommen werden könnten, als auf den theoretischen Hintergründen der Verfahren.

Das Verständnis von Zeitreihenmodellen als Regressionsansätze wird an den entsprechenden Stellen erwähnt und vereinfacht die Argumentation der Subsummierung dieser Methoden durch das in dieser Arbeit vorgestellte Prognosekonzept.

Im Folgenden wird in ausreichender Kürze das Gerüst von linearen Modellen bzw. Regressionsmodellen erklärt, um darauf aufsetzend die entscheidenden Annahmen und Charakteristiken der einzelnen Modelle zu notieren und diese in das entwickelte Verteilungskonzept abzubilden.

[1] Vgl. u.a. (Abraham and Ledolter, 1983), (Chatfield, 1989) und (Hamilton, 1994).

7.1.1 Lineare Modelle

In diesem Abschnitt werden in einem ersten Schritt die linearen Modelle in das Prognosekonzept der multivariaten bedingten Wahrscheinlichkeitsprognose eingebettet.

Lineare Modelle gehören zu den bekanntesten Verfahren der uni- und multivariaten Statistik und sind wohl daher die verbreitetste Art, funktionale Zusammenhänge zwischen stochastischen und deterministischen Variablen zu modellieren.[2] Die Theorie der linearen Modelle ist seit Jahren in der Literatur fundiert zugänglich. Es existieren infolgedessen unter anderem vielseitige Möglichkeiten der Ergebnisinterpretation und -analyse.[3]

Im Kontext von linearen Modellen wird die abhängige Variable[4] y_k als Funktion von m Inputvariablen[5] $x_{k1}, \ldots, x_{km}$ und einem zufälligen Fehlervektor[6] ϵ_k betrachtet. Der Index k ist in diesem Zusammenhang zur Notation der k-ten Beobachtung verwendet.[7] Die Beobachtung kann sich sowohl auf das k-te Prognoseobjekt als auch auf den k-ten Zeitpunkt der Zielgröße beziehen. Formal lässt sich daher der beschriebene Zusammenhang formulieren als:

$$y_k = r_\beta(x_k) + \epsilon_k, \tag{7.1}$$

wobei β den Vektor der unbekannten Modell- oder Schätzparameter bezeichnet.[8] Hierbei spielt die nicht beobachtbare Fehlervariable ϵ_k die Rolle des stochastischen Einflusses, wodurch y_k die Realisation einer Zufallsvariablen Y_k bezeichnet. Die Abbildung $r_\beta : \mathbb{R}^m \to \mathbb{R}$ ist eine beliebige parametrische Funktion mit linearen Schätzparametern β. Falls, wie häufig angenommen, r eine lineare Abbildung darstellt, ergibt sich das lineare Modell (7.1) für die

[2] Vgl. (Fahrmeir and Hamerle, 1984).
[3] Vgl. etwa (Eckey et al., 2002).
[4] y_k wird auch Zielgröße, endogene Variable oder Regressand genannt.
[5] x_k heißt auch Regressor oder unabhängige Variable.
[6] ϵ_k wird auch häufig Störvariable genannt.
[7] Vgl. Abschnitt 2.3 und Tabelle 2.1.
[8] Man beachte, dass die Nomenklatur im vorgestellten Verteilungskonzept der multivariaten bedingten Wahrscheinlichkeitsverteilungen analog gewählt ist.

k-te Beobachtung explizit zu

$$y_k = \beta_0 + \beta_1 x_{k1} + \cdots + \beta_m x_{km} + \epsilon_k \quad k = 1, \ldots, K. \tag{7.2}$$

Es sei an dieser Stelle betont, dass sich das Adjektiv *linear* im Zusammenhang mit linearen Modellen auf die Modellparameter $\beta = (\beta_0, \ldots, \beta_m)$ bezieht und nicht auf die unabhängigen Variablen x_k. So ist etwa

$$y_k = \beta_0 + \beta_1 x_{k1} + \beta_2 x_{k2}^2 + \epsilon_k \tag{7.3}$$

ein lineares Modell. Dagegen stellt in diesem Zusammenhang

$$y_k = \beta_0 + \beta_1^2 x_{k1} + \epsilon_k \text{[9]} \tag{7.4}$$

ein nichtlineares Modell dar.

Es bezeichnet r bei linearen Modellen eine Linearkombination von so genannten *Basisfunktionen* $r_0, \ldots, r_m$, wodurch sich Gleichung (7.2) formulieren lässt als:

$$y_k = r_\beta(x_k) = \beta_0 + \beta_1 r^1(x_{k1}) + \cdots + \beta_m r^m(x_{km}) + \epsilon_k \quad k = 1, \ldots, K. \tag{7.5}$$

Es ergibt sich in Matrixschreibweise:

$$y = R\beta + \epsilon, \tag{7.6}$$

mit

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_K \end{pmatrix}, R = \begin{pmatrix} 1 & r^1(x_{11}) & \ldots & r^m(x_{1m}) \\ 1 & r^1(x_{21}) & \ldots & r^m(x_{2m}) \\ \vdots & \vdots & & \vdots \\ 1 & r^1(x_{K1}) & \ldots & r^m(x_{Km}) \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}$$

und $\epsilon = (\epsilon_1, \ldots, \epsilon_K)^T$, wobei die erste Spalte von R in der Literatur häufig als

[9] Diese Beispiele sind an (Stuart et al., 1999) angelehnt.

Scheinvariable $x_0 = 1_K$ bezeichnet wird, um einen konstanten Term in den Regressionsgleichungen zu erzeugen. Die Matrix R korrespondiert über die Basisfunktionen direkt mit der so genannten *Design-Matrix* X der zugrunde liegenden Daten.

Da in den hier vorliegenden Fällen quantitative Prognoseaufgaben mit quantitativen Inputattributen zu behandeln sind und die Design-Matrix[10] X mit vollem Rang angenommen wird, heißt das lineares Modell auch Regressionsanalyse.[11]

Die Literatur unterteilt häufig die linearen Modelle grob in die folgenden drei Klassen, welche hier, zum Zweck der Übersichtlichkeit dieses Abschnitts, analog übernommen werden:[12]

- klassische lineare Modelle
- allgemeine lineare Modelle
- verallgemeinerte lineare Modelle.

Die genannten Modelle unterscheiden sich hauptsächlich durch die getroffenen Verteilungsannahmen der stochastischen Zielgrößen, was durch die folgenden Ausführungen zum Ausdruck kommt.

7.1.1.1 Klassische lineare Modelle

Wie bereits beschrieben sind die Gleichungen (7.1)-(7.6) stochastische Modelle, da der Fehlervektor ϵ eine nicht beobachtbare Zufallsvariable ist. Beim *klassischen linearen Modell (KLM)* werden folgende Annahmen[13] über die stochastischen Eigenschaften der Zufalls-/Störvariable ϵ getroffen:

[10] Zur Darstellung der Design-Matrix eines autoregressiven Regressionsmodells vgl. etwa (Hartung, 1999), S.735 ff.

[11] X besitzt vollen Rang bedeutet, dass $rg(X) = m + 1$, d.h. $K \geq m + 1$. Der Fall $rg(X) < m + 1$ wird z.T. in der Varianz- und Kovarianzanalyse behandelt (siehe etwa (Fahrmeir and Hamerle, 1984)).

[12] Vgl. etwa (Fahrmeir and Hamerle, 1984) und (Stuart et al., 1999).

[13] Die Annahmen sind in Anlehnung an (Fahrmeir and Hamerle, 1984) und (Abraham and Ledolter, 1983) formuliert, jedoch in vielzähligen Stellen der Literatur so zu finden.

- Der Erwartungswertvektor und die Varianzen von $\epsilon \in I\!R^K$ sind konstant, d.h.

$$E[\epsilon_k] = 0 \quad \text{und} \quad Var[\epsilon_k] = \sigma^2 \quad k = 1, \ldots, K. \tag{7.7}$$

- Die Komponenten des Störvektors ϵ sind unkorreliert, d.h.

$$Cov[\epsilon_i, \epsilon_j] = 0 \quad i, j = 1, \ldots, K, \quad i \neq j. \tag{7.8}$$

- Der Störvektor stammt aus der Klasse der multivariaten Normalverteilungen, d.h.

$$\epsilon \sim N(0, \sigma^2 I_K). \tag{7.9}$$

Für alle Beobachtungen $k = 1, \ldots, K$ ist durch die Stochastik des Fehlers ϵ_k die abhängige Zielgröße y_k eine Realisation der Zufallsvariablen Y_k. Da die endogene Variable y_k durch das lineare Modell von der Inputvariablen x_k abhängt, kann Gleichung (7.1) gleichbedeutend als Ausdruck des bedingten Erwartungswerts von Y_k beschrieben werden:

$$E[Y_k | X_k = x_k] = r_\beta(x_k).^{14} \tag{7.10}$$

Um die Äquivalenz der Prognosemodelle vergleichen zu können, wird die Notation der **bedingten** Verteilungen aufgenommen. Es lassen sich die Annahmen in diesem Zusammenhang wie folgt formulieren:

- Der **bedingte** Erwartungswert hängt von der unabhängigen Inputvariablen x_k ab. Hingegen sind die Varianzen vom Inputvektor x_k unabhängig, d.h.

$$E[Y_k | X_k = x_k] = r_\beta(x_k) \quad \text{und} \quad Var[Y_k | X_k = x_k] = \sigma^2 \quad k = 1, \ldots, K. \tag{7.11}$$

[14] Vgl. etwa (Stuart et al., 1999) oder (Abraham and Ledolter, 1983), die vorwiegend den Spezialfall $r_\beta(x_k) = x_k^T \beta$ formulieren.

- Die Komponenten des Zufallsvektors Y sind unkorreliert, d.h.

$$Cov[Y_i, Y_j] = 0 \quad i, j = 1, \ldots, K, \quad i \neq j. \tag{7.12}$$

- Y ist aus der Klasse der multivariaten **bedingten** Normalverteilungen mit Mittelwertvektor $(r_\beta(x_1), \ldots, r_\beta(x_K))^T$ und Kovarianzmatrix $\sigma^2 I_K$, d.h.

$$Y \sim N(r_\beta(x), \sigma I_K). \tag{7.13}$$

Die angenommene Eigenschaft gleicher Varianzen σ^2 aller Zufallsvariablen ϵ_k bzw. $Y_k, k = 1, \ldots, K$ in (7.7) bzw. (7.11) wird als *Homoskedastizität* bezeichnet. Im Kontext von Zeitreihen bedeutet dies, dass in allen Zeitpunkten die Streuungen der zu prognostizierenden Variablen identisch sind.

Mit der Annahme des vollen Rangs der Matrix R ergibt sich nach der gewöhnlichen *Kleinsten-Quadrate Methode*[15]

$$\min_\beta \sum_{i=1}^{K} \epsilon_i^2 \Leftrightarrow \min_\beta (y - R\beta)^T (y - R\beta) \tag{7.14}$$

ein Schätzer $\hat{\beta}$ für β, so dass die Fehlerquadrate minimal werden. Der bekannte KQ-Schätzer $\hat{\beta}$ lässt sich formal als

$$\hat{\beta} = (R^T R)^{-1} R^T Y \tag{7.15}$$

notieren. Hierbei bezeichnet Y die gegebene Datenmatrix und R die in Gleichung (7.6) eingeführte Matrix der Basisfunktionen.

Zur Berechnung eines Prognosewerts $\bar{y}$ zu vorgegebenen Inputvektoren $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_m)^T$ ergibt sich die Regressionsformel bzw. das Prognosemodell für klassische lineare Modelle zu

$$\bar{y} = Y^T R (R^T R)^{-1} \bar{x}. \tag{7.16}$$

[15] Auch als *(KQ)-Methode* bezeichnet.

Es sei der Vollständigkeit halber erwähnt, dass als Schätzer für σ^2 oftmals die gemittelte Residuenquadratsumme

$$\hat{\sigma} = \frac{1}{K-m-1} \sum_{k=1}^{K} (y_k - x_k^T \hat{\beta})^2$$

Verwendung findet.

Unter obiger Verteilungsannahme der bedingten Normalverteilung ist die Methode der KQ-Schätzung äquivalent mit dem in Abschnitt 5.1 vorgestellten Maximum-Likelihood Prinzip.[16]

Obige Annahmen (7.11) und (7.13) implizieren, dass der bedingte Erwartungswert von Y_k eine Funktion des Inputvektors x_k darstellt. Diese Beziehung ist jedoch nicht deterministisch, da für jedes feste x_k die korrespondierenden y_k um ihren Erwartungswert $E[Y_k|X_k = x_k]$ streuen. Diese Streuung hängt jedoch weder von der k-ten Beobachtung noch von x_k ab. Weiterhin kann aufgrund der Annahme (7.8) der Fehler ϵ_k nicht aus anderen Fehlern bestimmt oder vorhergesagt werden.

Eine Projektion der hier beschriebenen klassischen linearen Modelle auf das Prognosekonzept der multivariaten bedingten Wahrscheinlichkeitsverteilungen ergibt sich durch die Spezifikation des generellen funktionalen Approximators aus Gleichung (4.2). Man wählt

$$\mu_{\omega_1}(x_k) := r_\beta(x_k) \tag{7.17}$$

und

$$A_{\omega_2}(x_k) := \frac{1}{\sigma} (= const.), \quad \forall x_k \in I\!R^m. \tag{7.18}$$

Der dritte Teil des generellen funktionalen Approximators $v_{\omega_3}(x_k)$ ist in diesem Fall durch die Verteilungsannahme der Gauß'schen Normalverteilung nicht existent, da die Dichtefunktion (6.5) bzw. der Dichtegenerator (6.3) der Normalverteilung keine weiteren Formparameter besitzt.

In Anlehnung an die Generierung von Wahrscheinlichkeitsverteilungen in Abschnitt 3.1 und wegen der speziellen Form der Dichtefunktion einer univa-

[16] Vgl. etwa (Hartung and Elpelt, 1995).

riaten Normalverteilung aus Abschnitt 6.1.1 wird durch Definition (7.18) die Unabhängigkeit der Varianzen der einzelnen Beobachtungen $Y_1, \ldots, Y_k$ vom Inputvektor x_k erreicht, d.h.

$$Var[Y_k|X_k = x_k] = \sigma^2.$$

Folglich ergibt sich, dass jede Komponente der Zufallsvariablen Y normalverteilt ist mit bedingtem Erwartungswert $r_\beta(x_k)$ und unbedingter Varianz σ^2, d.h.

$$Y_k \sim N(r_\beta(x_k), \sigma^2) \quad k = 1, \ldots, K. \tag{7.19}$$

Für die gemeinsame Verteilung aller unkorrelierter Beobachtungen[17] folgt

$$Y \sim N(r_\beta(x), \sigma^2 I_K), \tag{7.20}$$

da $\sigma_1^2 = \cdots = \sigma_K^2 = \sigma^2$ und somit für die Kovarianzmatrix der Beobachtungen $\Sigma = diag(\sigma_1^2, \ldots, \sigma_K^2) = \sigma^2 I_K$ gilt.

Mit Hilfe einer geeigneten Definition des allgemeinen funktionalen Approximators ist daher die Äquivalenz der Prognosekonzepte gezeigt, da die Erfüllung aller charakterisierender Annahmen (7.11) bis (7.13) des klassischen linearen Modells gegeben ist.

7.1.1.2 Allgemeine lineare Modelle

Die *allgemeinen linearen Modelle (ALM)*, vorgestellt in (Aitken, 1935), entstehen durch eine sinnvolle, in der Praxis häufig auftretende Relaxation der strengen Annahmen der klassischen linearen Modelle bezüglich der homoskedastischen Kovarianzmatrix.[18]

Die allgemeinen linearen Modelle stellen, durch die Annahme einer allgemeinen symmetrisch positiv definiten Kovarianzmatrix, einen ersten verallgemeinernden Schritt hin zum hier vorgestellten Verteilungskonzept dar.

[17] Es sei wiederholend erwähnt, dass die K Beobachtungen stets als unabhängig angenommen werden, vgl. Kapitel 2.

[18] Als eine von unzähligen Anwendungen der allgemeinen linearen Modelle sei (Adams et al., 1991) genannt, der den Elektrizitätsbedarf in Abhängigkeit der Zeit prognostiziert.

Analog zu den Annahmen (7.7) bis (7.9) der klassischen linearen Modelle lässt sich über die Verteilung der Störvariable ϵ_k für den erweiterten Fall der allgemeinen linearen Modelle folgendes formulieren:

- Der Erwartungswertvektor ist konstant, die Varianzen der Komponenten von $\epsilon \in I\!R^K$ können jedoch variieren, d.h.

$$E[\epsilon_k] = 0 \quad \text{und} \quad Var[\epsilon_k] = \sigma_k^2 \quad k = 1, \ldots, K. \tag{7.21}$$

- Die Komponenten des Störvektors ϵ können korreliert sein, d.h.

$$Cov[\epsilon_i, \epsilon_j] = \Sigma \quad i, j = 1, \ldots, K. \tag{7.22}$$

- Der Störvektor stammt aus der Klasse der multivariaten Normalverteilungen, d.h.

$$\epsilon \sim N(0, \Sigma). \tag{7.23}$$

Der wesentliche Unterschied zu den klassischen linearen Modellen, bei denen eine diagonale Kovarianzmatrix $\Sigma = diag(\sigma^2, \ldots, \sigma^2) = \sigma I_K$ mit konstanten Diagonalelementen angenommen wird, ist die Verallgemeinerung auf die bekannte Form einer symmetrisch positiv definiten Kovarianzmatrix für den k-dimensionalen Fehlervektor ϵ:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & Cov[\epsilon_1, \epsilon_2] & \ldots & Cov[\epsilon_1, \epsilon_K] \\ Cov[\epsilon_2, \epsilon_1] & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ Cov[\epsilon_K, \epsilon_1] & \ldots & \ldots & \sigma_K^2 \end{pmatrix}.$$

Der direkte Zusammenhang[19] zwischen den stochastischen Fehlervariablen ϵ_k und den Zielvariablen Y_k lässt eine bedingte Formulierung der Verteilungsannahmen zu:

[19] Hier sei auf die Argumentation über die Regressionsgleichung (7.10) für das klassische lineare Modell verwiesen.

- Der **bedingte** Erwartungswert hängt von der Inputvariablen x_k ab. Die Varianzen der unterschiedlichen Beobachtungen Y_k können ebenfalls (abhängig von x_k) variieren, d.h.

$$E[Y_k|X_k = x_k] = r_\beta(x_k) \quad \text{und} \quad Var[Y_k|X_k = x_k] = \sigma_k^2 \quad k = 1, \ldots, K. \tag{7.24}$$

- Die Komponenten des Zufallsvektors Y können korreliert sein, d.h.

$$Cov[Y_i, Y_j] = \Sigma \quad i, j = 1, \ldots, n, \tag{7.25}$$

  wobei für Σ keine Diagonalgestalt gefordert wird.

- Der Zufallsvektor Y ist multivariat **bedingt** normalverteilt mit Erwartungswert $r_\beta(x) := (r_\beta(x_1), \ldots, r_\beta(x_K))^T$ und Kovarianzmatrix $\Sigma \in I\!R^{K\times K}$, d.h.

$$Y \sim N(r_\beta(x), \Sigma). \tag{7.26}$$

Die Eigenschaft (7.24) der unterschiedlichen Varianzen σ_k der Beobachtungen Y_k wird als *Heteroskedastizität* bezeichnet, falls die Komponenten Y_k von Y unkorreliert sind, d.h. Σ Diagonalgestalt mit unterschiedlichen Diagonalelementen besitzt.

Trotz der angenommenen allgemeinen Form der Kovarianzmatrix besitzt Σ oftmals eine spezielle Struktur.[20] Falls, wie oben erwähnt, unterschiedliche Objekte die Komponenten der Zielvariable Y darstellen, kann in den meisten Fällen Diagonalgestalt angenommen werden, da dies mit der Annahme der Unabhängigkeit der Beobachtungen äquivalent ist.

Ist es etwa möglich, die Folge der Störvariablen durch einen stochastischen *autoregressiven* Prozess zu beschreiben, so führt dies zu einer speziell strukturierten Fehlerkovarianzmatrix Σ. Liegt etwa ein autoregressiver Prozess erster Ordnung vor, so sind ausschließlich die Diagonale und ersten Nebendiagonalen mit Werten ungleich Null besetzt.[21]

[20] Vgl. etwa (Fahrmeir and Hamerle, 1984).

[21] Zur Definition von Heteroskedastizität und Autokorrelation siehe z.B. (Hamilton, 1994).

Das Pendant zur gewöhnlichen KQ-Schätzung ist für die allgemeinen linearen Modelle die in der Literatur als allgemeine KQ-Schätzung bezeichnete Schätzmethode.[22] Die Berechnungsformel des Prognosewerts $\bar{y}$ aus dem Inputvektor $\bar{x}$ lässt sich über dieses Schätzprinzip formal notieren als:

$$\bar{y} = Y^T \Sigma^{-1} R (R^T \Sigma^{-1} R)^{-1} \bar{x}, \tag{7.27}$$

da der allgemeine (KQ)-Schätzer $\hat{\beta}$[23] des Parametervektors β folgende Gestalt besitzt:

$$\hat{\beta} = (R^T \Sigma^{-1} R)^{-1} R^T \Sigma^{-1} Y.^{24} \tag{7.28}$$

R und Y sind analog zu Gleichung (7.16) bezeichnet. Der Schätzer $\hat{\beta}$ ist hierbei die Lösung der allgemeinen (KQ)-Optimierungsaufgabe

$$\min_{\beta \in I\!R^m} (y - R\beta)^T \Sigma^{-1} (y - R\beta). \tag{7.29}$$

Besitzt Σ Diagonalgestalt, so spricht man auch von *gewichteter* Kleinster-Quadrate Schätzung.[25]

Bei Verwendung des Maximum-Likelihood Schätzprinzips ergibt sich unter obigen Annahmen auch für allgemeine lineare Modelle ein zum KQ-Schätzer $\hat{\beta}$ identischer Schätzer.[26]

Offensichtlich ist die Bestimmung der Kovarianzmatrix Σ, die die Varianzen und die Kovarianzen zwischen den einzelnen Beobachtungen ausdrückt, vor der Bestimmung des Prognosemodells erforderlich.

Wie in den Annahmen (7.7) und (7.11) der klassischen linearen Modelle spiegelt sich auch in den Bedingungen (7.21) und (7.24) die Eigenschaft wider, dass der Erwartungswert von Y_k eine Funktion der unabhängigen Inputvariablen x_k darstellt. Die Streuung um diesen Wert $E[Y_k | X_k = x_k]$ ist

[22] Vgl. z.B. (Fahrmeir and Hamerle, 1984).

[23] $\hat{\beta}$ wird auch als Aitken-Schätzer bezeichnet.

[24] Zur Herleitung siehe etwa (Stuart et al., 1999).

[25] Vgl. (Fahrmeir and Hamerle, 1984) und beachte den Zusammenhang zur exponentiellen Glättung im Zeitreihenkontext.

[26] Es lässt sich direkt zeigen, dass sich die Optimierungsaufgabe nach dem Maximum-Likelihood Prinzip identisch zur Minimierung (7.29) darstellt.

allerdings unter diesen Modellannahmen nicht mehr konstant, sondern kann zwischen den einzelnen Beobachtungen $Y_k = y_k$ variieren.

Die Inklusion der allgemeinen linearen Modelle in das Konzept der bedingten multivariaten Wahrscheinlichkeitsverteilungen ergibt sich daher durch nachfolgende Definitionen. Man wählt in diesem Fall die Funktionen $\mu_{\omega_1}(x_k)$ und $A_{\omega_2}(x_k)$ des generellen funktionalen Approximators (4.2) als:

$$\mu_{\omega_1}(x_k) := r_\beta(x_k) \tag{7.30}$$

und

$$A_{\omega_2}(x_k) := \frac{1}{\sigma(x_k)}. \tag{7.31}$$

Analog zur Argumentation für die klassischen linearen Modelle existiert auch für die allgemeinen linearen Modelle kein dritter Teil $v_{\omega_3}(x_k)$ des generellen funktionalen Approximators, da unverändert die Annahme der Gauß'schen Verteilungsklasse besteht und diese, wie schon mehrmals erwähnt, keine reinen Formparameter besitzt.

Um die Eigenschaft einer heteroskedastischen Kovarianzmatrix der allgemeinen linearen Modelle zu erreichen, wird in Definition (7.31) eine nichtkonstante Abbildung $1/\sigma(x_k)$ gewählt. Hierdurch werden unterschiedliche Varianzen σ_k für die Beobachtungen Y_k zugelassen. Es ergibt sich daher für die bedingten Varianzen der einzelnen Beobachtungen $Y_1, \ldots, Y_k$

$$Var[Y_k|X_k = x_k] = \sigma^2(x_k) \quad k = 1, \ldots, K.$$

Die Beobachtungen Y_k sind also bedingt normalverteilt mit bedingtem Erwartungswert $r_\beta(x_k)$ und bedingter Varianz $\sigma^2(x_k)$, d.h.

$$Y_k \sim N(r_\beta(x_k), \sigma^2(x_k)) \quad k = 1, \ldots, K. \tag{7.32}$$

Für die gemeinsame Verteilung aller unkorrelierter Beobachtungen folgt demnach

$$Y \sim N(r_\beta(x), \Sigma(x)), \tag{7.33}$$

mit $\Sigma(x) = diag(\sigma_1^2(x_1), \ldots, \sigma_K^2(x_K))$.

Durch die geeignete Wahl des allgemeinen funktionalen Approximators (4.2) sind alle charakteristischen Verteilungsannahmen (7.24) bis (7.26) des allgemeinen linearen Modells erfüllt. Daher ist das allgemeine lineare Prognosemodell über das in dieser Arbeit beschriebene Konzept abgebildet.

7.1.1.3 Verallgemeinerte lineare Modelle

Die in (Nelder and Wedderburn, 1972) eingeführten *verallgemeinerten linearen Modelle (GLM)*[27] stellen eine weitere Verallgemeinerung der klassischen linearen Modelle dar.[28]

Zur Herleitung der verallgemeinerten linearen Modelle wird in der Literatur stets auf die klassischen linearen Modellen zurückgegriffen. Es sei wiederholend erwähnt, dass die klassischen linearen Modelle von einer konstanten Varianz aller Beobachtungen Y_k ausgehen. Bei der Formulierung folgender Annahmen der verallgemeinerten linearen Modelle wird dieser Bezug deutlich.

- Zu gegebenem Input x_k sind die Zufallsvariablen $Y_k,\ k = 1, \ldots, K$ (bedingt) unabhängig und die bedingte Verteilung von Y_k gehört einer Exponentialfamilie an mit bedingtem Erwartungswert $E[Y_k|X_k = x_k] = \mu_k$ und gegebenenfalls einem allgemeinen Skalierungsparameter ϕ, der von der k-ten Beobachtung unabhängig ist.

- Der bedingte Erwartungswert μ_k ist über eine eineindeutige und hinreichend oft differenzierbare *Link-Funktion* $g : \mathbb{R} \to \mathbb{R}$ mit dem linearen Schätzer $\eta_k = z_k^T \beta$ über

$$\mu_k = h(\eta_k) = h(z_k^T \beta) \quad \text{bzw.} \quad \eta_k = g(\mu_k)$$

  verbunden, wobei h die inverse Funktion von g ist. Der Vektor β be-

[27] Generalized Linear Models

[28] In (Dempster, 1971) sind verallgemeinerte lineare Modelle für natürliche Link-Funktionen vorgestellt. Die hier formulierte kurze Einführung ist hauptsächlich an (Fahrmeir and Tutz, 1994) und (McCullagh and Nelder, 1989) angelehnt. Für weitere Eigenschaften und theoretische Hintergründe von verallgemeinerten linearen Modellen sei ebenfalls auf diese Werke verwiesen.

zeichnet die Schätz-/Modellparameter und z_k stellt den Design-Vektor dar.

Das verallgemeinerte lineare Modell unterscheidet sich vom klassischen linearen Modell in den folgenden beiden Punkten.

Einerseits verbindet eine Link-Funktion, die i.Allg. ungleich der Identität angenommen wird, den linearen Schätzer η_k und den Erwartungswert μ_k. Diese Konstruktion ist aufgrund der breiteren Verteilungsannahme notwendig. Falls etwa eine Poisson-Verteilung, die auf der positiven reellen Achse lebt, angenommen wird, so kann eine Link-Funktion gewählt werden, die den linearen Schätzer η auf die positiven reellen Zahlen abbildet.[29]

Die zweite Verallgemeinerung gegenüber den klassischen linearen Modellen ist die bereits angedeutete breitere Verteilungsannahme einer Exponentialfamilie für die unabhängigen Fehlervariablen ϵ_k. Die Verteilungsklasse der Exponentialfamilie kann über die Dichtefunktion

$$f_{\theta_n,\phi}(y_n) = c(y_n) \exp\left[\theta_n y_n - b(\theta_n)\right]/a(\phi)$$

beschrieben werden, wobei ϕ ein von k und dadurch von x_k unabhängiger weiterer Skalierungsparameter ist. So ist etwa σ^2 im Falle der Normalverteilung ein derartiger Parameter. Hierbei sind $c : I\!R \to I\!R_0^+$ und $b : I\!R \to I\!R$ beliebige Funktionen.[30] Bei spezieller Wahl von c und b ergeben sich bekannte Verteilungen, die dieser Klasse der Exponentialfamilie angehören, wie etwa die Normalverteilung, die Bernoulliverteilung, die Poissonverteilung, die Gammaverteilung und die inverse Gauß-Verteilung.

Da keine unendlich große Anzahl von Messwiederholungen zu jeder erklärenden Variablen x_k vorliegt[31], existiert i.Allg. keine geschlossene Form eines Schätzers für θ. Natürlich ist in diesem Fall eine iterative Maximum-Likelihood Schätzung anwendbar.[32]

[29] Für geeignete Link-Funktionen zu unterschiedlichen Verteilungsannahmen sei etwa auf (McCullagh and Nelder, 1989) verwiesen.

[30] Vgl. die Definition von Exponentialfamilien etwa in (Fahrmeir and Hamerle, 1984).

[31] In diesem Fall könnte ein gewichteter KQ-Schätzer auch für verallgemeinerte lineare Modelle definiert werden, vgl. (Fahrmeir and Hamerle, 1984).

[32] Vgl. etwa (Fahrmeir and Tutz, 1994).

Für diese Ausführungen genügt eine Beschränkung auf die *natürlichen Link-Funktionen.*[33] Wählt man die Teilabbildung

$$\mu_{\omega_1} := g(z_k^T \beta) = g(r_\beta(x_k))$$

des funktionalen Approximators (4.2), so ist im Konzept der multivariaten bedingten Wahrscheinlichkeitsverteilungen den Zusammenhang zwischen dem linearen Schätzer $\eta = z_k^T \beta$ und dem bedingten Erwartungswert μ_k konstruiert.

Der Parameter ϕ wird je nach Verteilungsannahme über die Funktion $A_{\omega_3}(x_k)$ oder $v_{\omega_3}(x_k)$ des generellen funktionalen Approximators konstant abgebildet, weil er nach obiger Voraussetzung nicht von der Inputvariablen x_k abhängt.[34]

Da, wie oben angedeutet, die Maximum-Likelihood Schätzung der Vorgehensweise des Konzepts der multivariaten bedingten Wahrscheinlichkeitsverteilungen entspricht, ist die Inklusion der verallgemeinerten linearen Modelle beschrieben.

7.1.2 Multivariate lineare Modelle

Die Erweiterung der *multivariaten linearen Modelle (MLM)* im Gegensatz zu den oben behandelten univariaten linearen Modellen besteht darin, dass die unabhängigen Inputvariablen $x_1, \ldots, x_m$ zur Erklärung nicht nur einer skalaren Variablen y, sondern mehrerer Zielvariablen $(y_1, \ldots, y_n)^T = y$ dienen, wobei die Zufallsvariablen $(Y_1, \ldots, Y_n)^T = Y$ korreliert sind.[35]

Analog zu Gleichung (7.6) ergibt sich das multivariate lineare Modell zu:

$$Y = RB + E, \tag{7.34}$$

wobei $Y \in I\!R^{K \times n}$ die Beobachtungsmatrix und $R \in I\!R^{K \times (m+1)}$ die über die

[33] Damit wird der Fall bezeichnet, in dem die Linearkombination $z_k^T \beta$ gleich dem natürlichen Parameter θ ist. Nach (Fahrmeir and Hamerle, 1984) besitzt dieser Fall in Theorie und Praxis besondere Bedeutung.

[34] Vgl. Gleichung 7.18 der klassischen linearen Modelle.

[35] Vgl. (Fahrmeir and Hamerle, 1984).

Basisfunktionen zur Design-Matrix X korrespondierende Matrix darstellen. Die Modellparametervektoren $\beta_{(j)}, j = 1, \ldots, m+1$[36] sind in der Modellparametermatrix $B \in I\!R^{(m+1)\times n}$ und die Fehlervariablen $\epsilon_{(i)}, i = 1, \ldots, n$ in der Matrix $E \in I\!R^{K\times n}$ zusammengefasst.

Für jede Spalte $j = 1, \ldots, n$ gilt also ein univariates lineares Modell:

$$y_{(j)} = X\beta_{(j)} + \epsilon_{(j)} \quad j = 1, \ldots, n,$$

für das ein klassisches lineares Modell angenommen wird.[37] Die Zeilen, d.h. die Beobachtungen sind daher untereinander unkorreliert, die Elemente jeder Zeile dagegen untereinander korreliert. Es ergibt sich daher formal:

$$Cov[Y_{(j)i}, Y_{(j)k}] = 0 \quad i, k = 1, \ldots, K,\ j = 1, \ldots, n \tag{7.35}$$

und

$$Cov[Y_{ki}, Y_{kj}] = \sigma_{ij} I_n \quad i, j = 1, \ldots, n,\ k = 1, \ldots, K. \tag{7.36}$$

Unter der Normalverteilungsannahme gilt schließlich für alle unkorrelierten Beobachtungen

$$Y_k = (Y_1, \ldots, Y_n)_k^T \sim N(x_k B, \Sigma) \quad k = 1, \ldots, K.$$

Es wird also über alle Beobachtungen die identische und daher vom Inputvektor x_k unabhängige Kovarianzmatrix angenommen.

Die unbekannten zu schätzenden Modellparameter werden analog zu den klassischen linearen Modellen über das (KQ)-Schätzprinzip bestimmt und es ergibt sich ein Schätzer mit der folgenden Gestalt:

$$\hat{B} = (R^T R)^{-1} R^T Y,$$

wodurch sich das Prognosemodell als

$$\bar{y} = Y^T R (R^T R)^{-1} \bar{x}$$

[36] Die Indizierung $\beta_{(j)}$ bezeichnet die j-te Spalte einer Matrix.
[37] Vgl. (Fahrmeir and Hamerle, 1984).

ergibt.

Es kann gezeigt werden, dass $\hat{B}$ unter der Normalverteilungsannahme ein Maximum-Likelihood Schätzer ist.

Der Schätzer der Kovarianzmatrix Σ ergibt sich wiederum aus der *Residuenmatrix* zu

$$\hat{\Sigma} = \frac{1}{K-m-1} Y^T [I - R(R^T R)^{-1} R^T] Y.$$

Durch die offensichtliche Analogie zu den univariaten klassischen linearen Modellen erreicht man die Inklusion der multivariaten linearen Modelle auf direktem Weg. Die Wahl einer nicht konstanten Funktion

$$\mu_{\omega_1}(x_k) := r_B(x_k) \tag{7.37}$$

und einer konstanten Abbildung

$$A_{\omega_2}(x_k) := \frac{1}{\Sigma^{1/2}} (= const.), \quad \forall x_k \in I\!R^m \tag{7.38}$$

des allgemeinen funktionalen Approximators gewährleistet die Modellannahmen der multivariaten linearen Modelle.

Es sei an dieser Stelle auf weitere theoretische Ausführungen, wie etwa auf die Schätz- und Testtheorie von linearen Modellen verzichtet.[38]

7.1.3 Nichtlineare Modelle

Vervollständigend werden in diesem Abschnitt *nichtlineare Modelle (NLM)*

$$y_k = \bar{r}_\beta(x_k) + \epsilon_k, \quad E[\epsilon_k] = 0 \tag{7.39}$$

[38]Für die ausführliche Behandlung der univariaten Regressionsanalyse sei etwa auf (Drapper and Smith, 1966), (Searle, 1971) oder (Seber, 1977) verwiesen. Die Thematik der multivariaten Regressionsanalyse ist z.B. in (Anderson, 1958), (Mardia et al., 1979), (Dillon and Goldstein, 1984), (Johnson and Wichern, 1999) oder (Krzanowski and Marriott, 1994) und (Krzanowski and Marriott, 1995) detailliert behandelt. Diese Literaturangaben sind Beispiele zu unzähligen Werken mit der Thematik „Regressionsanalyse“.

betrachtet. Im Gegensatz zu den linearen Modellen gehen hier die Modell-/Schätzparameter meist nichtlinear in die Regressionsgleichung ein.[39]

Im Prognosekonzept der multivariaten bedingten Wahrscheinlichkeitsverteilungen werden nichtlineare Modellparameter durch die Verwendung von neuronalen Netzen als funktionaler Approximator realisiert.[40]

Nichtlineare Modelle sind ein weites Gebiet in der Statistik. Daher sei zur ausführlichen theoretischen Diskussion etwa auf (Gallant, 1987) und (Seber and Wild, 1989) verwiesen.

7.1.4 Stochastische Zeitreihenmodelle

„Prinzipiell nennt man jede zeitliche Folge quantitativer Beobachtungswerte zu einem bestimmten Vorgang eine Zeitreihe.“ [41]

In den vorangegangenen Abschnitten wurden die linearen und nichtlinearen Modelle in das Prognosekonzept der multivariaten bedingten Wahrscheinlichkeitsverteilungen eingebettet. Im Folgenden wird dies auf analoge Weise für bekannte Zeitreihenmodelle erreicht.

Im Kontext der Zeitreihenanalyse ist in diesem Zusammenhang weder eine Beschreibung des zeitlichen Vorgangs selbst, sprich des stochastischen Prozesses, von Interesse noch soll die zeitliche Beobachtung der Kontrolle dienen, was einem frühzeitigen Erkennen von „abnormalen“ Veränderungen gleichkommt.[42]

Da eine Vielzahl der Zeitreihenmodelle mit linearen und nichtlinearen Modellen aus obigen Abschnitten äquivalent sind, ist für diese Methoden eine Inklusion bereits gezeigt.

Die Beschreibung und die Identifikation[43] des zu Grunde liegenden stochastischen Prozesses ist nicht Thema dieses Abschnitts, da dies mit der In-

[39] Siehe Beispiel 7.4.

[40] In (Anders, 1996) werden die Parallelen von neuronalen Netzen und statistischen Prognoseverfahren herausgearbeitet und detailliert dargestellt.

[41] Aus (Hartung, 1999).

[42] Obwohl dieses Konzept ebenfalls für die Entwicklung eines Frühwarnsystems herangezogen werden kann und in vereinfachter Form in (Hrycej and Stützle, 2001) zum praktischen Einsatz kam, ist dieser Sachverhalt hier nicht Bestandteil der Ausführungen.

[43] Etwa durch die Box-Jenkins Methode.

terpretation der Daten vor dem Identifikationsprozess des Prognosemodells in Verbindung steht.[44]

Ein stochastischer Prozess kann als Ansammlung von Zufallsvariablen verstanden werden, die nach der Zeit geordnet und auf einer Menge von Zeitpunkten definiert sind, die diskret oder stetig ist.[45] Es wird hier ausschließlich der Fall von diskreten stochastischen Prozessen Y_t betrachtet.[46]

Eine Zeitreihe kann als eine spezielle Realisation eines stochastischen Prozesses verstanden werden. Im Zusammenhang mit Zeitreihenbetrachtungen ist daher die Konstruktion der Datenmatrix X nach der Definition des Modelltyps erforderlich. Zur Konstruktion von Trainingsdatensätzen durch Fensterbildung sei auf das Kapitel 2 verwiesen.

Bei Zeitreihendaten existieren im Wesentlichen drei verschiedene Informationsquellen für die Prognose des Zukunftswerts, die in geeigneter Form als exogene Inputvariablen verwendet werden können:

- die Zeitreihe aus der Vergangenheit,

- der Zeitindex und

- exogene Einflussgrößen.

Anhand der unterschiedlichen Ausnutzung dieser Informationsquellen lassen sich die bekanntesten Zeitreihenmodelle klassifizieren. Um die vergangenen Realisationen der Zufallsvariablen, den Zeitbezug, bestimmte Saisonalität und exogene Einflussgrößen, für eine Prognose zum Zeitpunkt t_0 zu berücksichtigen, wählt man den Inputvektor etwa als:

$$x = (y_{-q}, \ldots, y_0, t_0, sin(1/6\pi t_0), cos(1/6\pi t_0), \bar{x}_1, \ldots, \bar{x}_l). \tag{7.40}$$

[44] Die folgenden Ausführungen zur Zeitreihenanalyse orientieren sich unter anderem an (Hamilton, 1994), (Chatfield, 1989), (Abraham and Ledolter, 1983), (Brockwell and Davis, 2002) und (Hartung, 1999). Für weitere theoretische Ausführungen zu stochastischen Prozessen sei etwa auf (Durett, 1999), (Parzen, 1962b) oder (Tsay, 2001) verwiesen. Für Vergleichsstudien von unterschiedlichen klassischen Zeitreihenmethoden seien etwa die Artikel (Makridakis et al., 1982) oder (Martin and Witt, 1989) genannt.

[45] Vgl. (Chatfield, 1989).

[46] Y_t kann jedoch völlig analog zu den bisherigen Zufallsgrößen Y_k verstanden werden, vgl. Abschnitt 2.3.

Tabelle 7.1 stellt die unterschiedlichen Informationsquellen mit ihren Zeitbezügen dar.

	bekannt		unbekannt
	Vergangenheit	Gegenwart	Zukunft
Zeitreihe :	$y_{-q}, \ldots, y_{-1}$	y_0	$y_1, \ldots, y_n$
Zeitindex :	$t_{-q}, \ldots, t_{-1}$	$t_0, t_1, \ldots, t_n$	
exogene Einflüsse :	$\bar{x}_1, \ldots, \bar{x}_l$		

Tabelle 7.1: Informationsquellen bei der Zeitreihenanalyse

Zeitreihenmodelle unterscheiden sich, wie bereits erwähnt, maßgeblich durch die Wahl des Inputvektors einerseits und bezüglich Homo- und Heteroskedastizität andererseits.

Die hier betrachteten Zeitreihenmodelle stellen, analog zu den linearen Modellen, formal den Zusammenhang zwischen den Inputvariablen und der zu prognostizierenden Variable wie folgt dar:

$$y = s_\beta(x, \epsilon). \tag{7.41}$$

Da die zu modellierende abhängige Variable $Y = (Y_1, \ldots, Y_n)^T$ über die Abbildung s von einer nicht beobachtbaren Zufallsvariablen ϵ abhängt, ist sie stochastisch und Gleichung (7.41) lässt sich analog zur Darstellung (7.10) als bedingter Erwartungswert formulieren:

$$E[Y_t|X = x] = s_\beta(x) \quad t = 1, \ldots, n.^{47} \tag{7.42}$$

Eine Möglichkeit stochastische Prozesse zu beschreiben, ist die Angabe der gemeinsamen Wahrscheinlichkeitsverteilung von $Y_t,\ t = 1, \ldots, n$. Vorwiegend beschränkt sich die Praxis auf eine Darstellung über die Momente des stochastischen Prozesses, die *Erwartungswert-*, *Varianz-* und *Autokovarianzfunktion* genannt werden.[48]

[47] Vgl. (Hamilton, 1994).

[48] Vgl. etwa (Chatfield, 1989). Häufig stehen speziell das erste und zweite Moment im Vordergrund der Betrachtungen.

Die Erwartungswertfunktion $\mu(t)$ ist definiert als

$$\mu(t) = E[Y_t]. \tag{7.43}$$

Die Varianzfunktion $\sigma^2(t)$ ist definiert als

$$\sigma^2(t) = Var[Y_t]. \tag{7.44}$$

Die Autokovarianzfunktion $\gamma(t_1, t_2)$ ist definiert als

$$\gamma(t_1, t_2) = E\{[Y_{t_1} - \mu(t_1)][Y_{t_2} - \mu(t_2)]\}. \tag{7.45}$$

Ein stochastischer Prozess heißt *schwach stationär*[49], falls weder die Erwartungswertfunktion $\mu(t)$ noch die Autokovarianzfunktion $\gamma(t_1, t_2)$ von den Zeitpunkten t, t_1, t_2 abhängt.[50]

Stationarität ist selten eine aus der Praxis begründete Annahme als vielmehr eine vereinfachende Voraussetzung, die aus Gründen der leichteren Identifikation des stochastischen Prozesses getroffen wird.

Die hier betrachteten schwach stationären Zeitreihenmodelle AR, MA und ihre Mischform ARMA sind über sich selbst und einen Prozess ϵ_t erklärt, der ein so genanntes „weißes Rauschen“[51] darstellt.

7.1.4.1 Weißes Rauschen

Ein diskreter stochastischer Prozess ϵ_t heißt *Weißes Rauschen*, falls die Erwartungswertfunktion konstant gleich Null und die Varianzfunktion konstant gleich σ^2 ist, d.h.

$$E[\epsilon_t] = 0 \quad \text{und} \quad Var[\epsilon_t] = \sigma^2 \quad \forall t,$$

[49] Schwache Stationariät wird in der Literatur auch als Stationarität zweiter Ordnung bezeichnet.

[50] Im Fall einer konstanten Varianzfunktion spricht man von *Homoskedastizität*, vgl. (Hartung, 1999).

[51] Dieser Prozess wird auch häufig *white noise* genannt.

und falls zusätzlich die Zufallsvariablen ϵ_t alle über die Zeit unkorreliert sind

$$\gamma(\epsilon(t_1), \epsilon(t_2)) = 0 \quad \text{für} \quad t_1 \neq t_2.$$

Ohne Einschränkung der Allgemeinheit kann hier der vom Mittelwert bereinigte Prozess betrachtet werden.[52]

7.1.4.2 Autoregressiver Prozess der Ordnung p

Sei ϵ_t ein weißes Rauschen, dann ist Y_t ein *autoregressiver Prozess der Ordnung p*, falls

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \epsilon_t, \quad \forall t \tag{7.46}$$

gilt und die Schätzparameter $\beta_1, \ldots, \beta_p$ ungleich Null sind.[53]

Die Darstellung über den bedingten Erwartungswert

$$E[Y_t | Y_{t-1} = y_{t-1}, \ldots, Y_{t-p} = y_{t-p}] = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} \tag{7.47}$$

zeigt bei Normalverteilungsannahme die Analogie zu den klassischen linearen Modellen. Die Inputvariablen sind in diesem Fall Vergangenheitswerte der Zielgröße und keine klassischen exogenen Einflüsse. Die Inklusion des autoregressiven Zeitreihenmodells in das Konzept der multivariaten bedingten Wahrscheinlichkeitsverteilungen ist somit gezeigt.

Diese Sichtweise der Verbindung von Zeitreihen- und Regressionsmodellen findet sich in der Literatur auch unter den so genannten *autoregressiven Regressionsmodellen* wieder.[54]

---

[52] Vgl. hierzu (Hartung, 1999).

[53] Damit der AR(p)-Prozess schwach stationär ist, muss zusätzlich gefordert werden, dass die Lösungen z der Gleichung $1 - \beta_1 z - \beta_2 z^2 - \cdots - \beta_p z^p = 0$, die auch komplex sein können, außerhalb des Einheitskreises liegen, d.h. $|z| > 1$, siehe etwa (Hartung, 1999), Seite 678.

[54] Vgl. (Hartung, 1999).

7.1.4.3 Moving Average Prozess der Ordnung q

Sei ϵ_t ein weißes Rauschen, dann ist Y_t ein *Moving Average Prozess der Ordnung* q, falls

$$Y_t = \beta_1\epsilon_{t-1} + \beta_2\epsilon_{t-2} + \cdots + \beta_q\epsilon_{t-q} + \epsilon_t, \quad \forall t \tag{7.48}$$

und die Schätzparameter $\beta_1, \ldots, \beta_p$ ungleich Null sind.[55]

Bei einer Beobachtungslänge von n Zeitpunkten ergibt sich, beispielsweise für $q = 1$, über die Rekursion

$$\begin{aligned}
Y_1 &= \epsilon_1 \\
Y_2 &= \beta\epsilon_1 + \epsilon_2 = \beta Y_1 + \epsilon_2 \\
Y_3 &= \beta\epsilon_2 + \epsilon_3 = \beta(Y_2 - \beta Y_1) + \epsilon_3 = \beta Y_2 - \beta^2 Y_1 + \epsilon_3 \\
Y_4 &= \beta\epsilon_3 + \epsilon_4 = \beta(Y_3 - (\beta Y_2 - \beta^2 Y_1)) + \epsilon_4 = \beta Y_3 - \beta^2 Y_2 + \beta^3 Y_1 + \epsilon_4 \\
&\vdots
\end{aligned}$$

folgender Zusammenhang:

$$E[Y_t|Y_{t-1} = y_{t-1}, \ldots, Y_{t-n} = y_{t-n}] = \sum_{i=1}^{n-1}(-1)^{n-i-1}\beta^{n-i}y_i + \epsilon_n. \tag{7.49}$$

Hierbei ist β der einzige freie Modellparameter. Diese Rekursion lässt sich analog für den Fall $q > 1$ verallgemeinern. Es ergeben sich sodann q zu identifizierende Schätz-/Modellparameter. Durch Gleichung (7.49) zeigt sich die Äquivalenz zu den nichtlinearen Modellen. Offensichtlich ist es nicht trivial die freien Schätzparameter zu bestimmen. MA-Modelle verlangen daher iterative Verfahren.[56] Die MA(q)-Modellierung hat jedoch den Vorteil, dass der polynomiale Zusammenhang bekannt ist. Es besteht daher die Möglichkeit μ_{ω_1} des allgemeinen funktionalen Approximators wie bereits bekannt über

[55] Zur Eindeutigkeit fordert man, dass die Lösungen z der Gleichung $1 + \beta_1 z + \beta_2 z^2 + \cdots + \beta_q z^q = 0$, die auch komplex sein können, außerhalb des Einheitskreises liegen, siehe etwa (Hartung, 1999) Seite 681.

[56] Vgl. (Chatfield, 1989), Seite 58.

die Anwendung von neuronalen Netzen zu approximieren oder direkt die polynomiale Form zu modellieren.

Die Autokovarianzen, insbesondere die Varianzen $Var[Y_t] = \sigma^2 \sum_{i=0}^{q} \beta_i^2$ mit $\beta_0 = 1$, eines MA(q)-Prozesses ergeben sich aufgrund unkorrelierter Fehlervariablen als konstant und damit unabhängig vom Zeitpunkt t, so dass der zweite Teil des allgemeinen funktionalen Approximators, analog zu Gleichung (7.18), als konstante Abbildung gewählt werden kann.

Im Gegensatz zu der autoregressiven Zeitreihenmethode benötigt eine Modellierung mit Hilfe des Moving Average Prozesses theoretisch eine unendlich lange Zeitreihe. Da diese in der Praxis nie zur Verfügung steht, wird die Modellidentifizierung mit einer endlichen Anzahl von Zeitreihendaten durchgeführt. Für genügend lange Zeitreihen kann ohne Bedenken auf diese Weise vorgegangen werden, da die unendliche Reihe aus Gleichung (7.49) für $\beta < 1$ konvergiert.[57]

7.1.4.4 ARMA-Prozesse

Sei ϵ_t ein weißes Rauschen, dann ist der stochastische Prozess Y_t, der sich aus einem AR(p)- und einem MA(q)-Prozess zusammensetzt, ein *gemischter autoregressiver moving average Prozess der Ordnung p und q*, falls

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \tilde{\beta}_1 \epsilon_{t-1} + \tilde{\beta}_2 \epsilon_{t-2} + \cdots + \tilde{\beta}_q \epsilon_{t-q} + \epsilon_t, \quad \forall t \tag{7.50}$$

und die Schätzparameter $\beta_1, \ldots, \beta_p$ und $\tilde{\beta}_1, \ldots, \tilde{\beta}_q$ ungleich Null sind.[58]

Durch die Linearität des bedingten Erwartungswerts und die Homoskedastizität ist klar, dass sich die Abbildungen μ_{ω_1} und A_{ω_2} des allgemeinen funktionalen Approximators sofort einerseits als Summe aus den obigen Gleichungen (7.47) und (7.49) und andererseits als konstante Funktion zur Schätzung der Varianz ergeben.

[57] Es existieren eine Vielzahl von alternativen Schätzmethoden, wie etwa in (Box et al., 1994) oder in (Kendall et al., 1983) über die Dualität zwischen den AR und MA-Modellen.

[58] Zur Gewährleistung der schwachen Stationarität und der Eindeutigkeit werden die Bedingungen an die Schätzparameter von den AR(p)- und MA(q)-Prozessen übernommen.

7.1.4.5 ARIMA-Prozesse

Die bislang aufgeführten Zeitreihenmodelle haben die Eigenschaft der schwachen Stationarität gemein. Nicht schwach stationäre Prozesse lassen sich in einigen Fällen etwa durch lineare Filter in schwach stationäre Prozesse überführen.

Zu Beginn sei der so genannte *back-shift Operator vom Grad j* wie folgt definiert:

$$B^j Y_t := Y_{t-j}, \quad \forall j.$$

Die Differenzenmethode[59] bietet eine Möglichkeit, Trend- und Saisonkomponente bei konstanter Saisonfigur zu eliminieren. Die Bereinigung eines stochastischen Prozesses Y_t von einem polynomialen Trend vom Grad d kann durch Bildung der d-ten Differenzen

$$Y_t^* = \nabla^d Y_t = \nabla^{d-1}(\nabla Y_t), \quad \text{für} \quad t = d+1, \ldots, n \tag{7.51}$$

erfolgen. Hierbei gilt $\nabla Y_t = Y_t - Y_{t-1}$. Mit Hilfe des back-shift Operators lassen sich die d-ten Differenzen äquivalent zu Gleichung (7.51) wie folgt formulieren:

$$Y_t^* = (1-B)^d Y_t.^{60} \tag{7.52}$$

Ist Y_t ein nicht schwach stationärer Prozess derart, dass Y_t^* einen schwach stationären ARMA(p,q)-Prozess bildet, $p, q \geq 0$, so heißt der Prozess Y_t *integrierter ARMA(p,q)-Prozess* (kurz: ARIMA(p,d,q)).

Nach Bildung der d-ten Differenzen ist die Situation der ARMA(p,q)-Prozesse rekonstruiert und die Argumentation abgeschlossen.[61]

Die bislang vorgestellten Zeitreihenmethoden weisen allesamt die starke Annahme der Homoskedastizität auf. Da diese Annahme, wie bereits des Öfteren bemerkt, in den seltensten Fällen haltbar ist, wurde in (Engle, 1982) ein heteroskedastischer Zeitreihenansatz, die so genannten autoregressiven

[59] Vgl. etwa (Gosset, 1914).
[60] Vgl. (Chatfield, 1989).
[61] Abschließend sei auf den Artikel (Makridakis et al., 1982) verwiesen, der anhand von 1111 Zeitreihen 24 klassische Zeitreihenmethoden vergleicht.

bedingten heteroskedastischen (kurz: ARCH) Prozesse, entwickelt. Es wird hierbei nicht nur der Lokationsparameter der Zielgrößenverteilung als bedingt angenommen, sondern, ähnlich den allgemeinen linearen Modellen, werden die Varianzen der Beobachtungen nicht als konstant vorausgesetzt.

Bollerslev stellt mit dem generalisierten autoregressiven bedingten heteroskedastischen (kurz: GARCH) Modell eine wesentliche Verallgemeinerung zum ARCH-Modell vor.[62] Hierbei wurde durch die zusätzliche Abhängigkeit der Varianz von den vergangenen Varianzen eine Generalisierung erreicht.[63]

Die Literatur der letzten zehn Jahre zu diesem Thema ist zu umfangreich, um sie an dieser Stelle ausführlich zu erwähnen.[64] Speziell das breite Anwendungsfeld der Finanzmärkte bot dieser Methodik fruchtbaren Boden. Es entstanden eine Vielzahl von Erweiterungen der heteroskedastischen Modellierung unter anderem durch unterschiedliche Verteilungsannahmen der Störterme.[65]

7.1.4.6 ARCH-Prozesse

Ein ARCH(q)-Prozess kann auf unterschiedliche Weise definiert werden. In (Bera and Higgins, 1993) wird der ARCH-Prozess im Kontext der Fehlerverteilung eines dynamischen linearen Regressionsmodells definiert, was die zu zeigende Identifikation mit dem hier entwickelten Konzept vermuten lässt und vereinfacht. Daher wird auf diese Sichtweise zurückgegriffen, die keinerlei Einschränkungen zu der ursprünglichen Notation in (Engle, 1982) mit sich bringt.

Sei die stochastische Zielvariable durch

$$Y_t = x_t^T \beta + \epsilon_t \quad t = 1, \ldots, T, \tag{7.53}$$

[62] Siehe (Bollerslev, 1986).

[63] Es ist eine Analogie der Erweiterung von MA-Modellen zu den ARMA-Prozessen erkennbar, vgl. (Bera and Higgins, 1993).

[64] Es sei auf die Übersichtsartikel (Engle and Bollerslev, 1986), (Bollerslev et al., 1992), (Bera and Higgins, 1993) und (Bollerslev et al., 1994) verwiesen, die wiederum mehrere hundert Werke zitieren.

[65] Hier sei etwa auf (Engle, 1982), (Bollerslev, 1987), (Pagan and Schwert, 1990), (Bollerslev et al., 1992), (Pagan, 1996), (Granger and Sin, 2000) und (Hauser and Kunst, 2001) verwiesen.

generiert, wobei $x_t = (1, x_1, \ldots, x_m) \in I\!R^{m+1}$ dem Vektor von Einflussgrößen entspricht, der, wie etwa in Gleichung (7.40) und Tabelle 7.1 dargestellt, sowohl vergangene Werte der abhängigen Variablen als auch exogene Inputvariablen beinhalten kann. $\beta = (\beta_1, \ldots, \beta_{m+1})$ bezeichnet wiederholend den Vektor der Schätz-/Modellparameter. Die ARCH-Modelle setzen eine Verteilung des stochastischen Fehlers ϵ_t voraus, die von den realisierten Inputvariablen der Informationsmenge $\Phi_{t-1} = \{y_{t-1}, x_{t-1}, y_{t-2}, x_{t-2}, \ldots\}$ abhängt.

Speziell werden im ursprünglichen ARCH-Modell[66] normalverteilte Fehlervariablen angenommen, so dass

$$\epsilon_t | \Phi_{t-1} \sim N(0, h_t),$$

wobei

$$h_t = \tilde{\beta}_0 + \tilde{\beta}_1 \epsilon_{t-1}^2 + \cdots + \tilde{\beta}_q \epsilon_{t-q}^2 \tag{7.54}$$

mit $\tilde{\beta}_0 > 0$ und $\tilde{\beta}_i \geq 0, i = 1, \ldots, q$ gilt, um die Positivität der bedingten Varianzen zu gewährleisten. Da $\epsilon_{t-i} = y_{t-i} - x_{t-i}^T \beta$ für $i = 1, \ldots, q$ gilt, ist h_t eine Funktion der Elemente von Φ_{t-1} und stellt die Varianzfunktionen des stochastischen Prozesses dar. Neben einer Linearkombination der quadratischen Störterme können weitere Varianzfunktion formuliert werden, die eine verallgemeinernde Stoßrichtung zum ursprünglichen ARCH-Modell bilden.[67]

Es ergibt sich also die Normalverteilung mit bedingtem Erwartungswert $x_t^T \beta$ und bedingter Varianz h_t für die stochastische Zielvariable Y_t:

$$Y_t \sim N(x_t^T \beta, h_t).$$

Wählt man daher

$$\mu_{\omega_1} := x_t^T \beta$$

und

$$A_{\omega_2} := \frac{1}{\sqrt{h_t}},$$

[66] Siehe (Engle, 1982).

[67] Siehe (Geweke, 1986) und (Milhoj, 1987). Unter Verwendung von $\log(h_t) = \tilde{\beta}_0 + \tilde{\beta}_1 \log(\epsilon_{t-1}^2) + \cdots + \tilde{\beta}_q \log(\epsilon_{t-q}^2)$ entsteht etwa das so genannte log ARCH Modell.

so bildet sich das ARCH-Modell im Sinne des Konzepts der bedingten multivariaten Wahrscheinlichkeitsverteilungen ab.

7.1.4.7 GARCH-Prozesse

Das bekannte und in der Finanzwelt weit verbreitete GARCH-Modell ist eine Verallgemeinerung des ARCH-Ansatzes über die Modifikation der Varianzfunktion.[68]

Bollerslev schlug die Gestalt der Varianzfunktion aus Gleichung (7.54) wie folgt vor:[69]

$$h_t = \tilde{\beta}_0 + \tilde{\beta}_1 \epsilon_{t-1}^2 + \cdots + \tilde{\beta}_q \epsilon_{t-q}^2 + \beta_1 h_{t-1} + \cdots + \beta_p h_{t-p}, \tag{7.55}$$

wobei wiederum die Ungleichungen

$$\begin{aligned} \tilde{\beta}_0 &> 0 \\ \tilde{\beta}_i &\geq 0 \quad \text{für} \quad i = 1, \ldots, q \\ \beta_j &\geq 0 \quad \text{für} \quad j = 1, \ldots, p \end{aligned}$$

die Positivität der Varianzfunktion garantieren.[70]

Offensichtlich lässt sich diese Verallgemeinerung analog zu den ARCH-Modellen mittels des hier präsentierten Konzepts der multivariaten bedingten Wahrscheinlichkeitsverteilungen abbilden.

Auf die ausführliche Behandlung der vielzähligen Modifikationen und Erweiterungen der ARCH- und GARCH- Modelle sei an dieser Stelle verzichtet, da sich das Prinzip der Inklusion im Wesentlichen übertragen lässt.[71]

[68] Als eines von vielen Beispielen einer Anwendung des GARCH-Modells auf Finanzmarktdaten sei der Artikel (Mittnik and Paolella, 2000) genannt.

[69] Siehe (Bollerslev, 1986).

[70] Unabhängig davon präsentierte (Taylor, 1986) ein identisches Modellierungskonzept.

[71] Erweiterungen des ARCH-Modells sind etwa AARCH (Tsay, 1987), ARCD (Hansen, 1992) und (Hansen, 1994), ARFIMA-ARCH (Hauser and Kunst, 2001), EGARCH (Nelson, 1991), GARCH (Bollerslev, 1986), NARCH (Higgins and Bera, 1992), PNP ARCH (Engle and Ng, 1991), QARCH (Sentana, 1991), QTARCH (Gourieriuox and Monfort, 1992), TARCH (Zakoian, 1990) uvm.

7.2 Methoden aus der Neuroinformatik

Prognoseaufgaben sind seit jeher das klassische Anwendungsgebiet von neuronalen Netzen. Aufgrund ihrer Generalisierungsfähigkeit als nichtlineare Approximatoren werden neuronale Netze häufig als Prognosemodelle verwendet (siehe Abschnitt 4.2.3).[72]

In diesem Abschnitt werden daher ausschließlich Arbeiten aufgeführt und kategorisiert, die sich mit der Prognose von Wahrscheinlichkeitsverteilungen beschäftigen, da neuronale Netze im Sinne nichtlinearer Modelle bereits in Abschnitt 4.2 behandelt sind. Die folgende Darstellung wählt als Kriterium der Auflistung die getroffene Verteilungsannahme der Konzepte.

Annahme der univariaten Normalverteilung: Die ersten Arbeiten, die neuronale Netze zur Prognose von Wahrscheinlichkeitsverteilungen verwendeten, waren (White, 1989), (White, 1991), (White, 1992) und (White, 1994).

In (Nix and Weigend, 1994) findet ebenfalls ein neuronales Netz zur Schätzung univariater Normalverteilungen unter der Annahme von Homoskedastizität Verwendung. Zur Erzeugung der erforderlichen Positivität der Varianz werden, wie in den meisten hier zitierten Methoden, die korrespondierenden Ausgänge des Multi-Layer Perzeptrons mit Hilfe der Exponentialfunktion oder der quadratischen Funktion transformiert. Bis heute erscheinen Texte aus dem Forschungsgebiet der Neuroinformatik, wie etwa (de-la Vega et al., 2002) und (Qi, 2002), die noch immer die strenge Einschränkung der Homoskedastizität akzeptieren.

Annahme der univariaten Gauß'schen Mixturverteilung: In (Nabney et al., 1995), (Bishop, 1996) und später auch (Weigend and Shi, 2000) werden stochastische Zielvariablen mit Hilfe von univariaten Gauß'schen Mixturverteilungen modelliert. Allerdings nehmen auch sie die Unabhängigkeit der Varianz vom Inputvektor x an, so dass die starke Annahme der Homos-

[72] In (Adya and Collopy, 1998) wird ein Überblick neuroinformatischer Arbeiten der letzten Jahre gegeben. Weiterhin notiert (Hill et al., 1994) Literaturstellen, die den Zusammenhang von neuronalen Netzen und den Methoden aus der klassischen Statistik verdeutlichen. Als empirische Arbeiten seien etwa (Stock and Watson, 1998), (Moshiri and Cameron, 2000) oder (Qi, 2002) genannt.

kedastizität bestehen bleibt.

Der Gedanke, die univariaten Gauß'schen Mixturverteilungen als Verteilungsklasse zugrunde zu legen, wurde jedoch schon zuvor in den Publikationen (Nowlan, 1991), (Jacobs et al., 1991) und (Hamilton, 1991) formuliert und umgesetzt.

Annahme der univariaten bedingten Gauß'schen Mixturverteilung: Die Modellierung von bedingten Verteilungen, speziell von Finanzmarktdaten, mit Hilfe neuronaler Netze erregte Mitte der 90-er Jahre starkes Interesse. Die Verwendung Gauß'scher Mixturverteilungen nahm zu. Es wurde akzeptiert, dass die Annahme der Homoskedastizität generell zu einschränkend ist. Daher wurden in einigen Artikeln, wie etwa (Bishop, 1994), (Bishop and Legleye, 1995), (Bishop and Nabney, 1996), (Neuneier et al., 1994), (Neuneier, 1995), (Ormoneit and Neuneier, 1996), (Neuneier, 1998), (Schittenkopf et al., 1998), (Schittenkopf et al., 1999), (Schittenkopf and Dorffner, 2000), (Schittenkopf et al., 2000)[73] und (Bartlmae and Rauscher, 2000) die Varianz als vom Input abhängiger Ausgang des neuronalen Netzes modelliert. Dennoch bleibt in diesen Arbeiten der univariate Blickwinkel bestehen.

Annahme der multivariaten bedingten Gauß'schen Mixturverteilung mit unabhängigen Zielvariablen: Ein weit breiteres Konzept stellt Bishop in seinem Buch vor.[74] Er präsentiert einen multivariaten Ansatz der Schätzung von Gauß'schen Mixturverteilungen, wobei jedoch die endogenen Zielvariablen als unabhängig angenommen werden, d.h. es wird lediglich die Diagonale der Kovarianzmatrix geschätzt.[75] Ormoneit bezeichnet diese Art der Neuronalen Netze als *Mixture Density Networks* (MDN).[76] Die Arbeiten (Vlassis et al., 1999), (Vlassis and Kröse, 1999) und (Vlassis et al., 2000) können dieser Kategorie von Prognosemethoden zugeordnet werden, obwohl deren Fokus auf der zu bestimmenden Anzahl der Mixtur-

[73] Eine Verallgemeinerung der Mixture Density Networks zum rekurrenten MDN, um so Dynamik zu modellieren.

[74] Siehe (Bishop, 1995b). Etwas früher präsentiert er die Arbeit (Bishop, 1994).

[75] In (Khaikine and Holthausen, 2000) sind weitere theoretische Hintergründe dieses Ansatzes beschreiben.

[76] Siehe (Ormoneit, 1998).

komponenten liegt.

Annahme der multivariaten bedingten Normalverteilung mit abhängigen Zielvariablen: Während die oben zitierte Schule eng miteinander verknüpft und verbunden ist, entwickelte Williams ebenfalls ein Konzept der Schätzung multivariater bedingter Normalverteilungen, das tatsächlich die Korrelationsstruktur, d.h. die gesamte Kovarianzmatrix der multivariaten Zielvariablen abhängig vom exogenen Inputvektor identifiziert.[77]

7.3 Zusammenfassung

Abschließend erfolgt in den Tabellen 7.2, 7.3 und 7.4 eine Darstellung der unterschiedlichen Charakteristika der vorgestellten Prognosemethoden in Bezug auf einige ausgewählte Attribute. Diese hat, wie die obigen Ausführungen, keinen Anspruch auf Vollständigkeit, repräsentiert jedoch die hier aufgeführten und in der Praxis meist verwendeten Prognosemethoden aus der klassischen Statistik, der Zeitreihenanalyse und der Neuroinformatik.

Die Darstellung und Zusammenfassung der Tabellen 7.2, 7.3 und 7.4 stellt wiederholend die Verallgemeinerung des präsentierten Prognosesystems von bedingten multivariaten Wahrscheinlichkeitsverteilungen gegenüber den aus der Literatur bekannten Methoden heraus. Es gelang, diese Verfahren in das vorgestellte Konzept einzubinden und sie dadurch als Untermenge zu identifizieren. Die Tauglichkeit des entwickelten und technisch umgesetzten Prognosesystems wird im Folgenden geprüft, bevor es den Test auf realen Daten zu bestehen hat. Bei den Anwendungsfällen der Bedarfsprognose von Ersatzteilen und der Absatzprognose von Nutzfahrzeugen werden einige hier eingebettete Methoden Benchmarkergebnisse liefern, um die Performanz des Verteilungskonzepts zu beurteilen.

Der konzeptionelle Hauptteil dieser Arbeit ist mit diesem Kapitel abgeschlossen. Als Zwischenfazit kann die Entwicklung eines Prognosesystems festgehalten werden, das orientiert an den motivierenden Aspekten aus Kapi-

[77] Siehe (Williams, 1996). In (Williams, 1999) wird später das Bindeglied der „technischen Transformation“ zwischen neuronalem Netz und Verteilungsparametern erweitert.

Modell	Annahme der Verteilung	linear / nichtlinear	Verteilungsparameter bedingt / nicht bedingt	Dimension der Zielvariablen
KLM	Normal-Verteilung	linear	Lokation: bedingt Kovarianzstruktur: nicht bedingt Form: $\emptyset$	univariat
ALM	Normal-Verteilung	linear	Lokation: bedingt Kovarianzstruktur: nicht bedingt (heteroskedastisch) Form: $\emptyset$	univariat
GLM	Exponential-familie	linear	Lokation: bedingt Kovarianzstruktur: nicht bedingt Form: $\emptyset$	univariat
MLM	Normal-Verteilung	linear	Lokation: bedingt Kovarianzstruktur: nicht bedingt Form: $\emptyset$	multivariat korreliert
NLM	Normal-Verteilung	nichtlinear	Lokation: bedingt Kovarianzstruktur: nicht bedingt Form: $\emptyset$	univariat

Tabelle 7.2: Klassische statistische Modelle und ihre Eigenschaften

Modell	Annahme der Verteilung	linear / nichtlinear	Verteilungsparameter bedingt / nicht bedingt	Dimension der Zielvariablen
AR, MA, ARMA, ARIMA	Normal-Verteilung	linear	Lokation: bedingt Kovarianzstruktur: nicht bedingt Form: $\emptyset$	univariat
ARCH	Normal-Verteilung	linear	Lokation: bedingt Kovarianzstruktur: bedingt Form: $\emptyset$	univariat
GARCH	Normal-Verteilung	linear	Lokation: bedingt Kovarianzstruktur: bedingt Form: $\emptyset$	univariat

Tabelle 7.3: Stochastische Zeitreihenmodelle und ihre Eigenschaften

tel 2, einige Defizite herkömmlicher Verfahren kompensiert und Eigenschaften besitzt, die qualitativ hochwertige und interpretierbare Prognoseergebnisse vermuten und erwarten lassen.

Verteilungs-annahme	linear / nichtlinear	Verteilungsparameter bedingt / nicht bedingt	Dimension der Zielvariablen
Normal-Verteilung etwa (White, 1989)	nichtlinear	Lokation: bedingt Kovarianzstruktur: nicht bedingt Form: $\emptyset$	univariat
Gauß'sche Mixtur-Verteilung etwa (Nowlan, 1991)	nichtlinear	Lokation: bedingt Kovarianzstruktur: nicht bedingt Form: nicht bedingt	univariat
Gauß'sche Mixtur-Verteilung etwa (Bishop, 1994)	nichtlinear	Lokation: bedingt Kovarianzstruktur: bedingt Form: nicht bedingt	univariat
Gauß'sche Mixtur-Verteilung etwa (Bishop, 1995b)	nichtlinear	Lokation: bedingt Kovarianzstruktur: bedingt Form: nicht bedingt	multivariat unkorreliert
Normal-Verteilung etwa (Williams, 1996)	nichtlinear	Lokation: bedingt Kovarianzstruktur: bedingt Form: $\emptyset$	multivariat korreliert

Tabelle 7.4: Methoden aus der Neuroinformatik und ihre Eigenschaften

Teil II

Validierung des Prognosemodells

Kapitel 8

Validierung des Prognosekonzepts mit Hilfe synthetisch generierter Daten

Dieses Kapitel befasst sich mit der Validierung der Prognosemethode bedingter multivariater Verteilungsklassen anhand von künstlich generierten Daten[1]. Diese Verifikation wird für die hier betrachteten Verteilungsklassen aus Kapitel 6 durchgeführt. Hierzu zählen die Gauß'schen Normalverteilungen, die t-Verteilungen, die stabilen Verteilungen, die generalisiert hyperbolischen Verteilungen und die Gauß'schen Mixturverteilungen. Die Vorgehensweise für die unterschiedlichen Verteilungsfamilien erfolgt, bis auf die ausführlicheren Darstellungen im Normalverteilungsfall, analog. Abschließend rechtfertigt eine Quer-Validierung die Betrachtung der ganzen Palette von Verteilungsklassen.

Dieser Schritt der Verifikation des Prognosesystems auf synthetischen Daten ist unumgänglich, da die bedingte Wahrscheinlichkeitsverteilung realer Daten nie bekannt ist, d.h. auch in der Vergangenheit keine Sollgröße zur Bewertung der berechneten Ergebnisse existiert.

[1] Künstlich erzeugte Daten werden häufig auch als synthetische Daten bezeichnet.

8.1 Validierung von Verteilungsklassen

Die Schätzung von multivariaten bedingten Wahrscheinlichkeitsverteilungen kann durch experimentell generierte Datensätze validiert werden. Es wird ein funktionaler Zusammenhang zwischen den Parametern der Wahrscheinlichkeitsverteilung und einer exogenen Einflussgröße durch die nachfolgend beschriebene Vorgehensweise simuliert.

Die künstlichen Daten werden aus einer gemeinsamen Verteilung, definiert als $d(y|x)h(x)$, von $Y \times X$ gezogen. Hierbei beschreibt $d(y|x)$ die bedingte zu identifizierende Verteilung und $h(x)$ die Dichtefunktion der Gleichverteilung von X auf einem festgelegten Intervall $[a; b]$. Meist sind in den folgenden Versuchen jeweils 10000 künstliche Datensätze generiert.[2]

8.1.1 Gauß'sche Normalverteilung

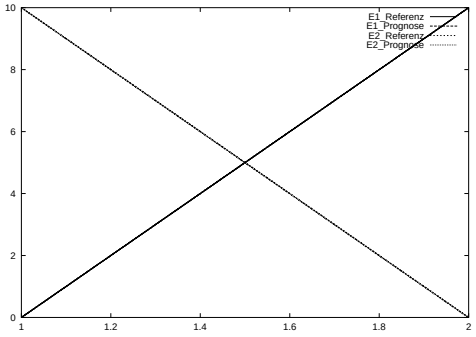
Das erste Experiment testet die Fähigkeit der Methodik, lineare Abhängigkeiten des Mittelwertvektors und der Standardabweichungen einer bivariaten Gauß'schen normalverteilten Zufallsvariablen vom exogen gegebenen Einfluss x zu identifizieren. Die Zufallsgröße X ist in diesem Fall auf dem Intervall $[1; 2]$ gleichverteilt. Der Korrelationskoeffizient wurde unabhängig von der exogenen Variablen x auf den Wert $\rho_{ref} = 0,5$ festgelegt.

Die Generierung wurde derart gewählt, dass sich der zu identifizierende lineare Zusammenhang zwischen dem Lokationsvektor und den Skalierungsparametern wie folgt ergibt:

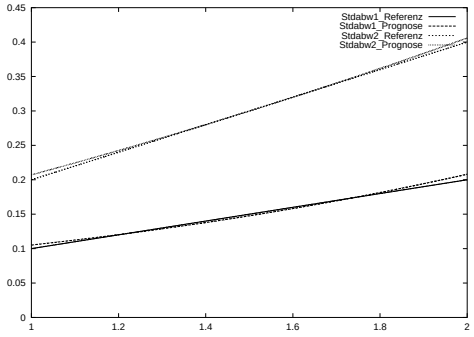
$$\begin{aligned} \mu_1(x) &= 10\ x - 10 & \qquad \sigma_1^2(x) &= 0,1\ x \\ \mu_2(x) &= -10\ x + 20 & \qquad \sigma_2^2(x) &= 0,2\ x. \end{aligned}$$

Der Vergleich zwischen den synthetisch erzeugten Referenzwerten der Mittelwerte bzw. der Standardabweichungen und den vom Prognosemodell berechneten bedingten Parametern der Normalverteilung ist in den Abbil-

[2] In einigen explizit ausgewiesenen Fällen wurden mehr als 10.000 Datensätze zum Zweck der höheren Genauigkeit und besseren Illustration generiert.



(a) Linearer funktionaler Zusammenhang der bedingten Erwartungswerte.



(b) Linearer funktionaler Zusammenhang der bedingten Standardabweichungen.

Abbildung 8.1: Referenz- und Prognoseverlauf sowohl der bedingten Erwartungswerte als auch der bedingten Standardabweichungen einer bivariaten „linear-bedingten“ normalverteilten Zufallsvariablen

dungen 8.1(a) bzw. 8.1(b) illustriert.[3] Simultan wurde der Korrelationskoeffizient von der Prognosemethodik auf den Wert $\hat{\rho} = 0,50082$ geschätzt, was einem relativen Fehler von $0,164\%$ entspricht. Die Übereinstimmung der Geraden zeigt die sehr hohe Güte der Schätzung.

Da die nichtlineare Modellierung einen aus der Praxis geforderten Anspruch darstellt, wird ausschließlich der allgemeine Fall der nichtlinearen Bedingtheit betrachtet.

Die Arbeiten von (Weigend and Nix, 1994) und (Williams, 1996) diskutieren unter anderem künstlich generierte Daten einer univariaten Normalverteilung. Die dort definierten funktionalen Abhängigkeiten

$$\begin{aligned} \mu_1(x) &= \sin(2,5\,x)\ \sin(1,5\,x) \\ \sigma_1^2(x) &= 0,01\ +\ 0,25\ [1-\sin(2,5\,x)]^2 \end{aligned} \tag{8.1}$$

stellen einen geeigneten und anspruchsvollen Richtwert für die Validierung

[3] In Abbildung 8.1 sind die Referenzwerte mit $E1\_Referenz$, $E2\_Referenz$ bzw. $Stdabw1\_Referenz$, $Stdabw2\_Referenz$ bezeichnet. Alle prognostizierten Kurvenverläufe sind mit $E1\_Prognose$ und $E2\_Prognose$ bzw. $Stdabw1\_Prognose$ und $Stdabw2\_Prognose$ benannt.

der Methodik im Falle einer bedingten univariaten Normalverteilung dar.

Die aus der Generierung resultierende Menge der Trainingsdatensätze ist in Abbildung 8.2 präsentiert.[4] Dass die Varianz ebenso wie der Mittelwert als bedingter Verteilungsparameter angenommen ist, spiegelt sich in der unterschiedlichen Streuung der Datenpunkte um den Mittelwertverlauf über das betrachtete Intervall $[0; \pi]$ wider.

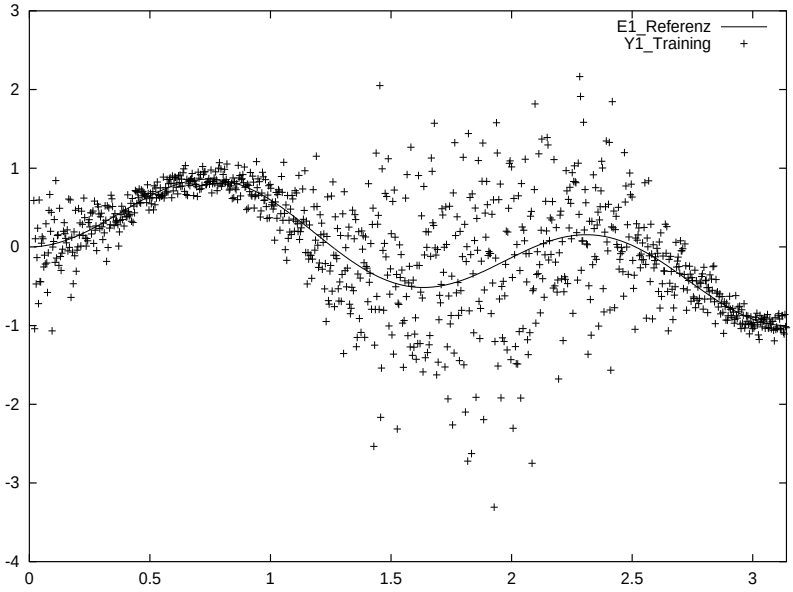


Abbildung 8.2: Künstlich erzeugte Trainingsmenge von univariaten normalverteilten Daten um die Erwartungswertfunktion $\mu(x) = \sin(2,5\ x)\sin(1,5\ x)$ mit Varianzverlauf $\sigma^2(x) = 0,01 + 0,25[1 - \sin(2,5\ x)]^2$ im Intervall $[0; \pi]$

Abbildung 8.3 zeigt die sehr gute Anpassung des Modells an den in Gleichung 8.1 definierten und vorgegebenen nichtlinearen Zusammenhang der Verteilungsparameter und der endogenen Einflussgröße x. Die geringen Ungenauigkeiten im mittleren Bereich des Intervalls $[0; \pi]$ stehen offensichtlich mit der in diesem Abschnitt existierenden hohen Standardabweichung in direktem Zusammenhang.

Schon die Trainingsdaten in Abbildung 8.2 illustrieren die große Streuung

[4] Zur besseren Illustration ist in Abbildung 8.2 nur ein Zehntel aller Datenpunkte abgetragen.

im Bereich zwischen 1 und 2. Eine Verbesserung der Anpassung lässt sich jederzeit durch die Erhöhung der Anzahl der Traingsdatensätze erreichen, was im folgenden Fall der Modellierung einer „nichtlinear-bedingten“ bivariat normalverteilten Zufallsvariablen zum Ausdruck kommt.

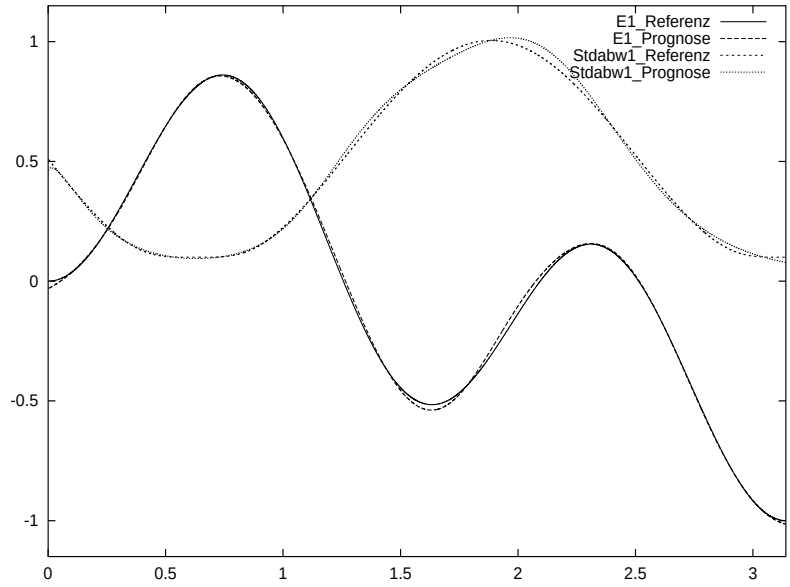


Abbildung 8.3: Referenz- und Prognoseverlauf des bedingten Erwartungswerts und der bedingten Standardabweichung einer univariaten „nichtlinear-bedingten“ normalverteilten Zufallsvariablen

Die Ergebnisse eines erweiterten Experiments mit nichtlinearen funktionalen Abhängigkeiten und bivariaten normalverteilten Zufallsvariablen wird in den Abbildungen 8.4 und 8.5 illustriert. Der vorgegebene Korrelationskoeffizient von $0,5$ wurde auch in diesem Versuch mit einem geringen relativen Fehler von $0,064\%$ mit einem Wert von $0,50032$ bestimmt. Die Ergebnisse basieren auf einer Datengrundlage von 100.000 Datensätzen, da anspruchsvolle funktionale Verläufe zu modellieren sind. Die zu identifizierenden Funktionsvorschriften wurden in diesem Beispiel für den bedingten Mittelwertvektor durch

$$\begin{aligned} \mu_1(x) &= \sin(2,5\ x)\ \sin(1,5\ x) \\ \mu_2(x) &= \cos(3,5\ x)\ \cos(0,5\ x) \end{aligned}$$

und für die bedingten Standardabweichungen durch

$$\begin{aligned} \sigma_1(x) &= \sqrt{0,01 \ + 0,25 \ [1 - \sin(2,5 \ x)]^2} \\ \sigma_2(x) &= \sqrt{0,01 \ + 0,25 \ [1 - \sin(3,5 \ x)]^2} \end{aligned}$$

definiert.[5]

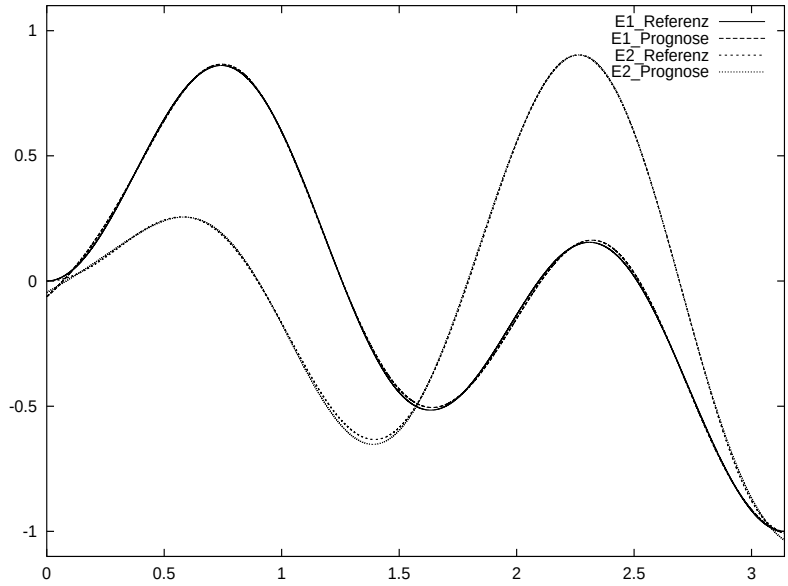


Abbildung 8.4: Referenz- und Prognoseverlauf der bedingten Erwartungswerte einer bivariaten „nichtlinear-bedingten“ normalverteilten Zufallsvariablen

Bei den hier durchgeführten Validierungsexperimenten und Anwendungen wird nicht auf unterschiedliche Architekturen des neuronalen Netzes zurückgegriffen. Es wurden alle Versuche mit der in Bemerkung 2 definierten Anzahl[6] versteckter Neuronen in einem 3-Schicht Perzeptron durchgeführt.[7]

[5]Die Arbeit von (Williams, 1996) dient hierbei wiederum als Vorgabe des funktionalen Zusammenhangs, wodurch ein direkter Vergleich der Performanz möglich ist. Vgl. (Williams, 1996), Seite 849.

[6]Die Anzahl der Neuronen in der versteckten Schicht des Multi-Layer Perzeptrons folgt der Vorschrift $\min[\max(n_0, n_l); 6]$, wobei n_0 und n_l die Anzahl der Neuronen in der Eingangs- bzw. Ausgangsschicht sind.

[7]In (Williams, 1996) stellte sich ebenfalls ein 3-Schicht Netzwerk mit 6 versteckten Neuronen in der einzigen Zwischenschicht als das Geeignete heraus.

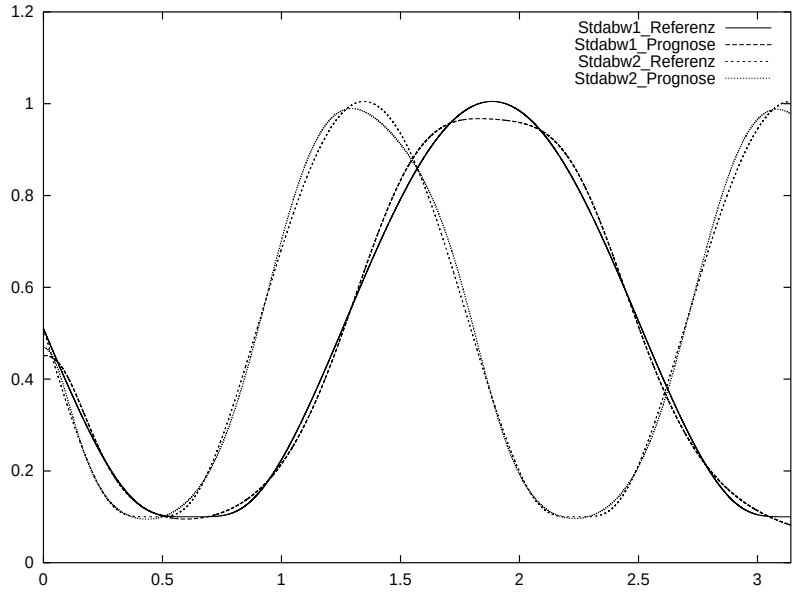


Abbildung 8.5: Referenz- und Prognoseverlauf der bedingten Standardabweichungen einer bivariaten „nichtlinear-bedingten“ normalverteilten Zufallsvariablen

Ferner ist durch die Verwendung der in Kapitel 5.3 dargestellten und hier implementierten globalen Multi-Level Single-Linkage Optimierungsmethode eine genauere Untersuchung von lokalen Minima hinfällig.[8]

Vergleichbar mit der Modellierung von univariaten bedingten Zielgößen zeigt sich eine natürliche Ungenauigkeit in den Bereichen, in denen eine hohe Standardabweichung vorherrscht. Im Gegensatz zum univariaten Fall zeigt sich ein geringer Schätzfehler in der Varianzprognose. Trotz der größeren Optimierungsaufgabe tritt diese Erscheinung jedoch nicht, wie erwartet werden könnte, ausgeprägter auf.

8.1.2 t-Verteilung

Die zweite Verteilungsklasse, die in dieser Versuchsreihe validiert wird, sind die in Abschnitt 6.1.2.1 als Spezialfall der elliptischen Verteilungen eingeführ-

[8]Etwa in (Williams, 1996) wird durch die Verwendung von lokalen Optimierungsmethoden eine Mittelung über unterschiedliche lokale Optima zur Verbesserung der Schätzergebnisse durchgeführt.

ten t-Verteilungen.

Wie bereits erwähnt wird im Folgenden lediglich die allgemeine bivariate „nichtlinear-bedingte“ Modellierung betrachtet. Aufgrund der echten Inklusion der linearen und univariaten Modelle kann die Gültigkeit dieser Teilmengen gefolgert werden. Zur Illustration einer Reihe von unterschiedlichen funktionalen Zusammenhängen sowohl im univariaten als auch im multivariaten Fall sei auf den Abschnitt 6.1.1.1 verwiesen.

Eine geeignete Generierung von bivariaten „nichtlinear-bedingten“ t-verteilten Zufallsvariablen ergibt die folgenden funktionalen Abhängigkeiten des Mittelwertvektors zu:

$$\begin{aligned} \mu_1(x) &= 1 + \frac{9}{1 + e^{-10(x-1,5)}} \\ \mu_2(x) &= 10 - \frac{9}{1 + e^{-10(x-1,5)}} \end{aligned} \tag{8.2}$$

Die nichtlineare Bedingtheit der Standardabweichungen liegt in den Daten wie folgt zugrunde:

$$\begin{aligned} \sigma_1(x) &= 1 + \frac{1}{1 + e^{-10(x-1,5)}} \\ \sigma_2(x) &= 1 + \frac{1}{2 + e^{-10(x-1,5)}} \end{aligned} \tag{8.3}$$

In diesem Beispiel ist keine periodische Abhängigkeit der Verteilungsparameter vom Einfluss x vorausgesetzt. Die sigmoide Funktion beschreibt eine ansteigende Streuung bei wachsendem exogenem Einfluss. Die beiden Erwartungswerte werden als gegenläufig angenommen.

Der zusätzliche Formparameter der Freiheitsgrade wurde für die Generierung der synthetischen Daten auf den Wert 15 festgelegt.

Die Abbildungen 8.6 und 8.7 stellen die Anpassung der Modelle an die vorgegebenen funktionalen Zusammenhänge dar. Vergleichbar mit den Ergebnissen aus Abschnitt 6.1.1.1 ist auch in diesem Fall eine geringe, unwesentliche, Ungenauigkeit im Bereich von vorherrschend hohen Varianzen zu erkennen. Der Approximationsfehler des Mittelwertvektors ist für eine Überschätzung der Varianz verantwortlich, da diese zusätzliche Unsicherheit

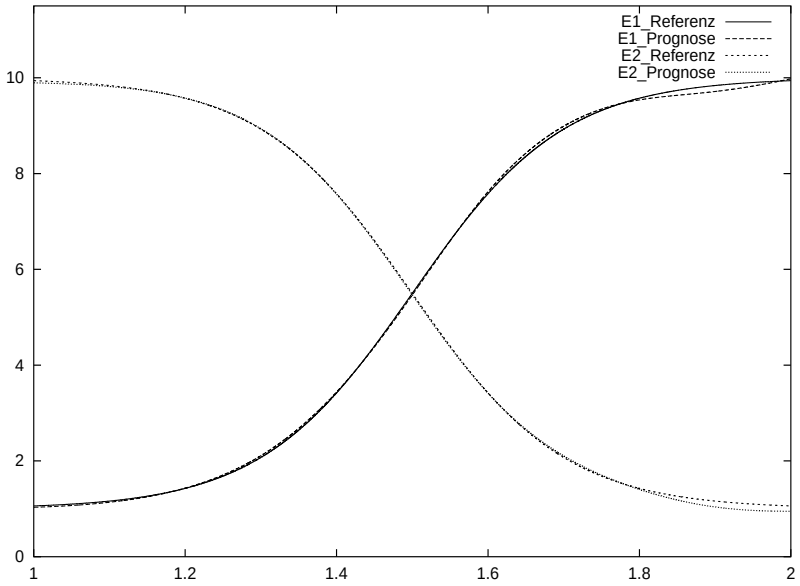


Abbildung 8.6: Referenz- und Prognoseverlauf der bedingten Erwartungswerte einer bivariaten „nichtlinear-bedingten“ t-verteilten Zufallsvariablen

in der Schätzung berücksichtigt wird. Durch eine Erhöhung der Datengrundlage könnte diesem Phänomen entgegengewirkt werden. Zusammenfassend zeigt die Schätmethode auch für die Klasse der t-Verteilungen eine sehr gute Anpassung.

Schließlich wurde der als unbedingt gewählte Formparameter durch den Wert von $13,455$ identifiziert. Dieser Unterschied wirkt sich jedoch auf die Form der Wahrscheinlichkeitsdichte nur unwesentlich aus.[9]

8.1.3 Stabile Verteilung

Eine weitere Klasse von Wahrscheinlichkeitsverteilungen, die in dem hier vorgestellten Konzept Verwendung finden können, ist die in Kapitel 6.2 vorgestellte Familie der stabilen Verteilungen. In bekannter Weise wird nun die Validierung der Prognosemethodik für diesen Verteilungstypus durchgeführt.

Der zugrunde liegende funktionale Zusammenhang des Lokations- und

[9]Es sei ein Ungenauigkeitsproblem bei einer Datenmenge von zehntausend Datensätzen für die Schätzung von bedingten t-Verteilungen mit Freiheitsgraden kleiner als 5 erwähnt.

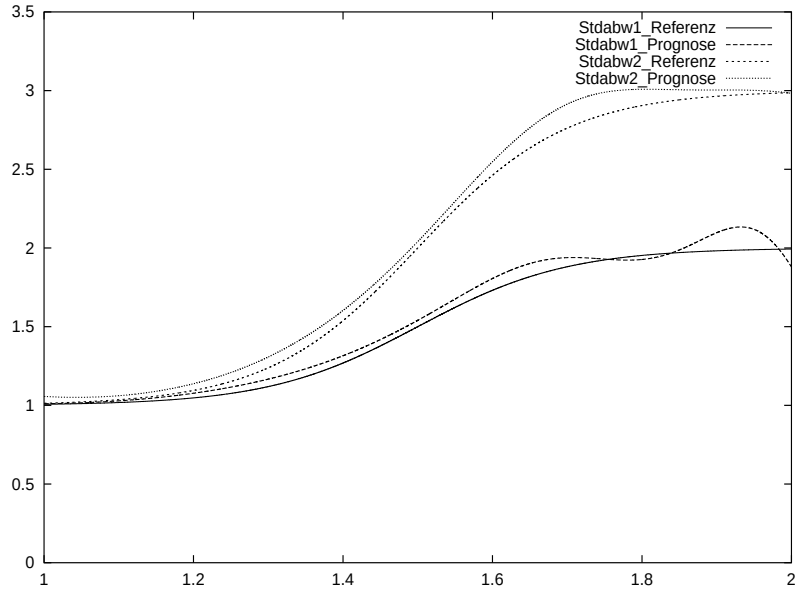


Abbildung 8.7: Referenz- und Prognoseverlauf der bedingten Standardabweichungen einer bivariaten „nichtlinear-bedingten“ t-verteilten Zufallsvariablen

Skalierungsparameters entspricht den in den Gleichungen (8.2) und (8.3) formulierten sigmoiden Abhängigkeiten. Die als unbedingt angenommenen Formparameter sind durch die Werte $\alpha_{ref} = 1,7$ und $\beta_{ref} = 0,8$ definiert, was eine Asymmetrie und eine positive Kurtosis bewirkt.

Da die Formparameter der stabilen Verteilungen von den Lokations- und Skalierungsparametern, die hier je nach exogenen Einflussgröße x variieren, nicht unabhängig sind, ergeben unterschiedliche Parameterkombinationen sehr ähnliche Verläufe der Wahrscheinlichkeitsdichten. Zum Beispiel kann ein wachsender Stabilitätsindex α durch einen größer gewählten Skalierungsparameter näherungsweise kompensiert werden. Dieses Phänomen tritt bei den im folgenden Abschnitt 8.1.4 behandelten generalisiert hyperbolischen Verteilungen sogar verstärkt auf, da diese einen weiteren generalisierenden Formparameter besitzen. Aus diesem Grund werden in den Abbildung 8.8, 8.9 und 8.10 die zu schätzenden und identifizierten Dichtefunktionen direkt verglichen. Die Dichten sind für einige repräsentative Werte der Inputvariablen x abgebildet.

Trotz der oben beschriebenen möglichen „Parametersubstitution“ wur-

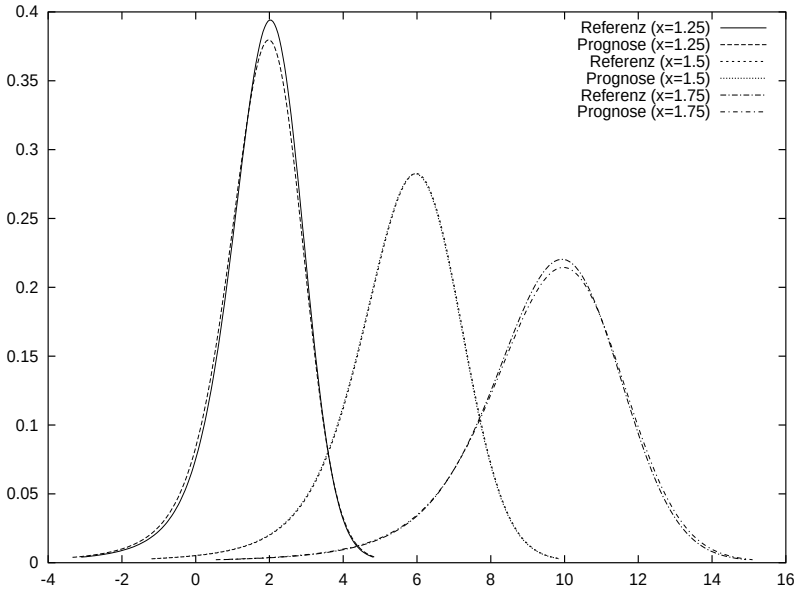


Abbildung 8.8: „Nichtlinear-bedingte“ Dichtefunktionen stabil-verteilter Zufallsvariablen

den die Werte der Formparameter $\alpha_{ref} = 1,65$ und $\beta_{ref} = 0,78$ durch das Prognosemodell mit geringen Abweichungen identifiziert. Die geringe Unterschätzung von α wird durch eine Überschätzung von σ kompensiert, was insgesamt die sehr gute Anpassung der Dichtefunktionen an die empirische Verteilung begründet.

Ebenso wie bei der Klasse der t-Verteilungen können bei der Modellierung von stabilen Verteilungen Ungenauigkeitsprobleme auftreten. Wählt man etwa einen Stabilitätsindex $\alpha < 1,2$, so ergibt sich eine Form der Dichtefunktion, die kaum mit einer angemessenen Anzahl von Trainingsdatensätzen identifizierbar ist.

8.1.4 Generalisiert hyperbolische Verteilung

Als letzte klassische Familie von Wahrscheinlichkeitsverteilungen wird im Folgenden die Modellierungsfähigkeit von generalisiert hyperbolischen Verteilungen validiert.[10]

[10] Vgl. (Stützle and Hrycej, 2003).

Es kann gezeigt werden, dass die hyperbolischen Verteilungen ausreichen, um die Klasse der generalisiert hyperbolischen Verteilungen mit ausreichender Genauigkeit zu repräsentieren, da die Dichtefunktion mit einem Generalisierungsparameter $\lambda \neq 1$ durch eine hyperbolische Dichte (λ konstant als 1 fixiert) hinreichend genau dargestellt werden kann.[11]

Aus diesem Grund reicht sowohl für die folgenden praktischen Anwendungen als auch für die durchgeführte Validierung der Prognosemethodik die Betrachtung der Klasse von reinen hyperbolischen Verteilungen aus.

Wie bereits erwähnt stellt sich die Eigenschaft der Formparameter, von den Lokations- und Skalenparametern nicht unabhängig zu sein, bei der Klasse der hyperbolischen Verteilungen ebenso ein wie bei den stabilen Verteilungen. Ein wachsender Formparameter α lässt sich auch für diesen Verteilungstyp durch einen vergrößerten Skalenparameter ausgleichen. Abbildung 8.9 zeigt analog zu Abbildung 8.8 einige ausgewählte Vergleiche der Referenz- zu der prognostizierten Dichtefunktion.

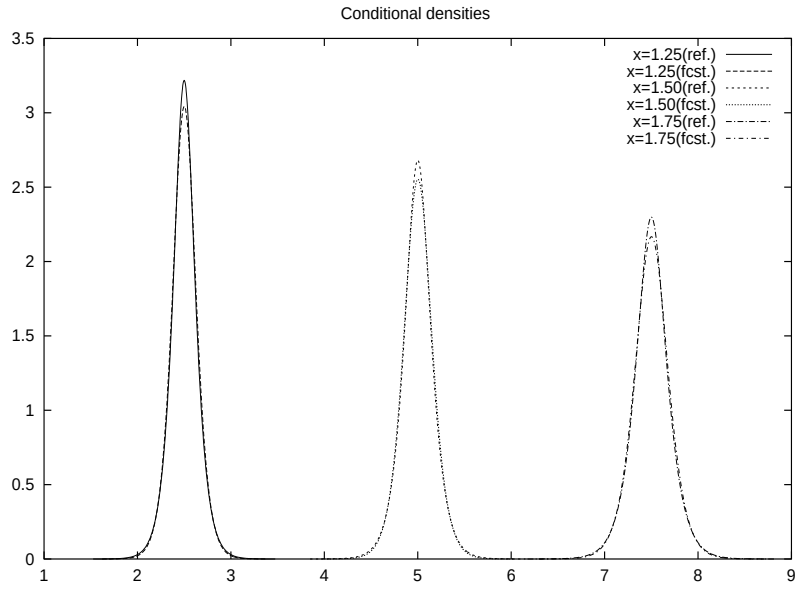


Abbildung 8.9: „Nichtlinear-bedingte“ Dichten generalisiert hyperbolischverteilter Zufallsvariablen

Die Formparameter der generierten hyperbolischen Verteilung wurden in

[11] Vgl. dazu (Schmidt et al., 2003).

diesem Experiment unbedingt als $\lambda = 1$, $\alpha = 1,7$ und $\beta = 0,3$ gewählt. Weiterhin wurde eine lineare Bedingtheit der Lokations- und Skalierungsparameter vorausgesetzt. Erschwerend hierbei ist die Modellierung einer bivariaten hyperbolischen Verteilung. Die Abhängigkeit der beiden Zufallsvariablen ist in gleicher Weise wie bei der Gauß'schen Normalverteilung in Abschnitt 6.1.1.1 gewählt.[12] In Abbildung 8.9 sind die eindimensionalen Randverteilung abgebildet.

8.1.5 Binäre Gauß'sche Mixturverteilung

Abschließend zur Modellvalidierung für unterschiedliche Verteilungsklassen wird in diesem Abschnitt die Fähigkeit der Schätzmethodik geprüft, eine Mischung aus zwei normalverteilten Zufallsvariablen zu identifizieren. Hierfür wird zwar nicht der extreme Fall einer bimodularen Wahrscheinlichkeitsdichte betrachtet, dennoch wurde eine schiefe Wahrscheinlichkeitsdichte zur Prüfung gewählt, die zusätzlich eine Kurtosis ungleich Null aufweist.[13]

Die Ungenauigkeiten der in Abbildung 8.10 gegenübergestellten Dichtefunktionen resultieren aus der stark streuend generierten künstlichen Datenmenge von 10.000 Datensätzen. Für $x = 1,75$ ergibt sich etwa nach der Formel (6.18) eine Varianz der Mixtur-Verteilung von $\sigma^2_{res} = 0,2 \cdot 9 + 0,8 \cdot 4 + 0,2 \cdot 0,8 \cdot 1 = 5,16$.

Der Gewichtungsparameter, der in den synthetischen Daten auf den Wert $0,2$ eingestellt wurde, konnte von der Methode als $0,16$ identifiziert werden. Aus bekannten Gründen der „Parametersubstitution“ werden jedoch auch bei diesem Verteilungstyp wiederum die in Abbildung 8.10 dargestellten Wahrscheinlichkeitsdichten direkt miteinander verglichen, um eine Aussage über die Güte der Schätzung treffen zu können. Trotz der linear anwachsenden großen Varianzen wird eine gute Approximation der vorgegebenen Dichtefunktionen erreicht.

[12] Allerdings ist für die hyperbolische Verteilungsfamilie das Konzept der Korrelationsmatrix nicht von der Klasse der Normalverteilungen zu übernehmen.

[13] Die Modellierung bimodularer Verteilungen weist eine ähnlich gute Performanz auf, ist für die hier vorliegenden Anwendungen jedoch nicht relevant.

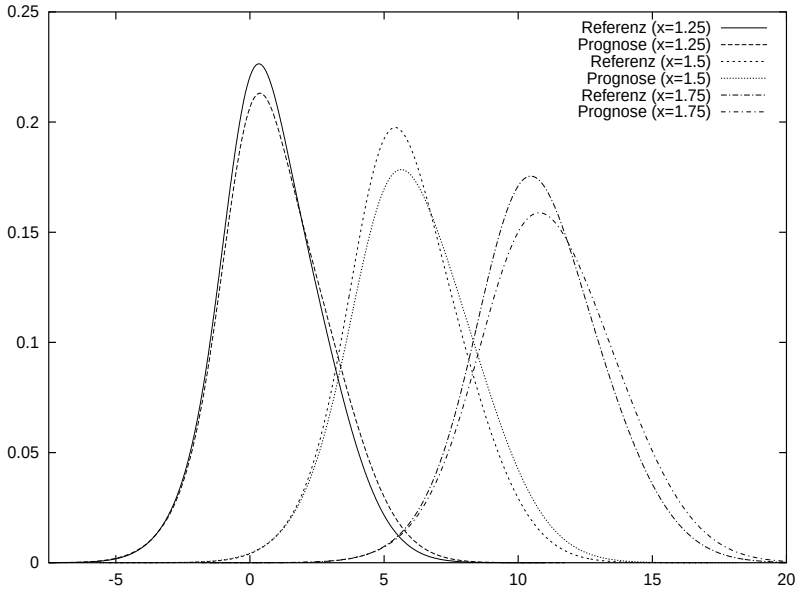


Abbildung 8.10: Vergleich „nichtlinear-bedingter“ Dichtefunktionen mixturverteilter Zufallsvariablen

8.2 Quer-Validierung

Dieser Abschnitt ist einer weiteren Validierungsstufe der Prognosemethodik von bedingten Wahrscheinlichkeitsverteilungen gewidmet. Schon in den einleitenden Sätzen einiger Abschnitte des Kapitels 6 kristallisiert sich die Entstehung und Formulierung unterschiedlicher Klassen von Wahrscheinlichkeitsverteilungen als nicht ausschließlich in der Wahrscheinlichkeitstheorie begründet heraus. Intensive Bemühungen, alternative Verteilungsfamilien zu definieren, äußerten sich daher nicht zuletzt in der Bestrebung, stochastische Prozesse adäquat zu modellieren.[14] Die Frage, ob sich durch die Vielzahl von unterschiedlichen Verteilungsklassen ein tatsächlicher Mehrwert im Zusammenhang mit der hier entwickelten Methodik ergibt, lässt sich durch die folgenden Darstellungen beantworten.

Auf vier unterschiedlich generierten künstlichen Datenmengen mit jeweils 25.000 Datensätzen werden die wesentlichen Klassen von Wahrscheinlichkeitsverteilungen aus Kapitel 6 geschätzt. Das Maß der Cross-Entropie aus

[14] Vgl. etwa (Mandelbrot, 1963).

Abschnitt 5.1 bietet die Möglichkeit, die Güte der Anpassung an die synthetischen Daten zu beurteilen. Es sollte daher in jedem Fall diejenige Verteilung die beste Approximation bieten, mit welcher die Trainingsdaten generiert wurden.

Die zu modellierenden Datensätze bestehen aus einer univariaten Zufallsvariablen. Zur Vereinfachung ist der Lokations- und Skalierungsparameter ebenso von einer exogenen Größe x unabhängig wie die Formparameter der jeweiligen Verteilung. Eine konträre Annahme würde die Aussagen dieses Abschnitts weiter bestärken, da dann die Anpassung an die zugrunde liegende empirische bedingte Verteilung sogar diffiziler wäre. Die generierten Wahrscheinlichkeitsverteilungen in den Trainingsdaten drücken jedoch spezielle Charakteristika, wie etwa Schiefe oder „heavy tails“ der bestimmten Verteilungsklasse aus, um potenziell den quantitativen Unterschied der Modellierungsfähigkeit klar identifizieren zu können.

Es wurden Datenmengen mit den folgenden konkreten Wahrscheinlichkeitsverteilungen synthetisch erzeugt:[15]

- eine Standardnormalverteilung ($N(0,1)$),
- eine t-Verteilung mit 15 Freiheitsgraden,
- eine generalisiert hyperbolische Verteilung mit den Formparametern $\alpha = 1,3$, $\beta = 0,7$ und $\lambda = 1$ und
- eine binäre Gauß'sche Mixturverteilung aus den Normalverteilungen ($N_1(1,2)$) und ($N_2(2,3)$) und einem Gewichtungsparameter $q = 0,2$.

Tabelle 8.1 präsentiert die Maßzahlen der Cross-Entropien spaltenweise. Da jeder Datensatz eine spezifische Untergrenze dieses Informationsmaßes besitzt, sind die unterschiedlichen Spalten der Tabelle nicht zu vergleichen.

Die erste Spalte zeigt, dass die Gauß-Verteilung ein Spezialfall der betrachteten Verteilungen ist, da alle Approximationen ein ähnliches Anpassungsmaß aufweisen. Dass sich hierbei die hyperbolische Verteilungsklasse

[15] Eine als stabil verteilte Zufallsvariable wurde nicht in die Betrachtungen involviert, da die Verteilung in der generierten Datenmenge zu ungenau ist. Es ergab sich bei einer Datenmenge von 25.000 Datensätzen $\int_{-\infty}^{\infty} f(y)dy \approx 0,8$.

Generierung Schätzung	Gauß	t	hyperbolisch	Mixtur
Gauß	**1,81132111**	2,57632887	2,39938789	2,09237671
t	**1,81132111**	**2,57280064**	2,37196699	2,06007213
stabil	**1,81132111**	*2,57423291*	2,36070265	2,05357244
hyperbolisch	**1,81168392**	**2,57279545**	**2,35434019**	*2,05032543*
Mixtur	**1,81117575**	**2,57272390**	**2,35662838**	**2,04732685**

Tabelle 8.1: Vergleich der Anpassungsgüte von unterschiedlichen Wahrscheinlichkeitsverteilungen

als die, mit geringem Abstand, schlechteste Approximation der zugrunde liegenden normalverteilten Daten präsentiert, liegt an der größeren Anzahl von zu identifizierenden Parametern.

Fett markierte Maßzahlen stellen die besten Approximationen an die generierten Trainingsdaten dar. Eine kursive Betonung zeigt zwar ein gewisses Defizit der Entropien gegenüber der bzw. den besten Anpassungen. Dennoch liegen diese Modellierungsgüten noch in einem Akzeptanzbereich im Gegensatz zu den restlichen Ergebnissen.

Schon bei t-verteilten synthetischen Datensätzen weist die Approximation durch die Normalverteilung klare Schwächen auf und grenzt sich gegenüber den anderen Modellen negativ ab. Diese Unterschiede in der Güte kommen im Wesentlichen durch die schlechte Anpassung in den Tails zustande. Die Klasse der stabilen Verteilungen liegt zwar noch im Toleranzbereich, zeigt bei diesem Datensatz jedoch ebenfalls erste Nachteile gegenüber den hyperbolischen Verteilungen und der Gauß'schen Mixturverteilung.

Obwohl die stabilen Verteilungen aus theoretischer Sicht eine sehr flexible Klasse von Wahrscheinlichkeitsverteilungen repräsentiert, können sie in diesem direkten Vergleich nicht überzeugen. Dies liegt vor allem in der schwierig zu berechnenden Wahrscheinlichkeitsdichte begründet. Ein weiterer klarer Vorteil der hyperbolischen Wahrscheinlichkeitsverteilungen zeigt die Modellierung von schiefen Verteilungen. Die in den hyperbolischen und mixturverteilten künstlichen Daten definierten Eigenschaften sind Asymmetrie und eine Kurtosis ungleich Null.

Bei einer Generierung von sehr großen Datenmengen mit einer breiten Streuung und langsam abfallenden Tails zeigen die hyperbolischen Verteilungen jedoch durch den polynomialen Abfall ihrer Tails einen klaren Vorteil gegenüber der Klasse der Gauß'schen Mixturverteilungen, die ein exponentielles Verhalten in den Tails aufweisen, was in Abschnitt 6.4 gezeigt wird.

Die Anpassung an mixturverteilte Daten ist für klassische Dichtefunktionen besonders anspruchsvoll, da sie sehr atypische Funktionsverläufe aufweisen können. Daher ist durch die kursiv gedruckte Entropie der hyperbolischen Verteilung bei mixturverteilter Datengrundlage eine gute Qualität gekennzeichnet.

Der große und entscheidende Nachteil einer Modellierung mit Hilfe der Gauß'sche Mixtur ist die mit der Dimension quadratisch ansteigende Anzahl der Schätzparameter. Da in diesem Experiment eindimensionale Verteilungen zu schätzen waren, kam dieser Nachteil, im Gegensatz zur Validierung in Abschnitt 8.1.5, jedoch nicht zum Ausdruck.

Als Fazit lässt sich nach diesen Untersuchungen die Klasse der hyperbolischen Wahrscheinlichkeitsverteilungen klar präferieren und als die stabilste Verteilungsfamilie identifizieren. Sie weist einen guten „trade-off“ zwischen Flexibilität durch ihre Formparameter und akzeptabler Anzahl von Schätzparametern auf. Zusätzlich ist die Dichtefunktion geschlossen formulierbar, so dass im Gegensatz zu den stabilen Verteilungen nicht auf zusätzliche mathematische Hilfsmittel zur Berechnung zurückgeriffen werden muß.

Teil III

Ausgewählte Anwendungen

Zur empirischen Untersuchung der hier entwickelten und validierten Prognosemethodik der bedingten Wahrscheinlichkeitsverteilungen werden zwei ausgewählte praktische Anwendungen präsentiert.

Zum einen stellt sich die Prognoseaufgabe des Ersatzteilebedarfs in der Automobilindustrie aus vergangenen Bedarfszahlen und verschiedenen Charakteristika der einzelnen Teile für die nächsten 15 Jahre. Da i.Allg. ein asymmetrisches Kostenverhältnis vorliegt, kann hierbei die Schätzung der gesamten Verteilung zur kostenoptimalen Entscheidung genutzt werden.

Zum anderen ist die Absatzprognose von Lastkraftwagen eine Herausforderung. Basierend auf vergangenen Verkäufen, bereits bestehenden Auftragszahlen und beschreibenden Eigenschaften der Produkte soll eine Vorhersage für die kommenden Monate erstellt werden.

Die Zielsetzungen der beiden konkreten praktischen Anwendungen sind teilweise unterschiedlicher Art. Für die erste Untersuchung stellt sich primär die Aufgabe der Kostenreduzierung. Es sind die Aufwendungen einer Unter- bzw. Überschätzung zu minimieren. Aus praktischen Gründen wird hierbei anhand der gängigen Normalverteilungsklasse und der Familie der log-Normalverteilungen argumentiert. Der Prognosehorizont erstreckt sich hierbei, wie bereits erwähnt, über einen sehr langen Zeitraum von bis zu 15 Jahren. Zusätzlich ist eine Datenmenge von zirka 700.000 Ersatzteilen mit einer Historie von 14 Jahren zu bewältigen.

Bei der Bedarfsprognose von Nutzfahrzeugen beläuft sich der Prognosezeitraum auf eine sehr viel kürzere Zeitspanne, da die Nachfrage für die nächsten sechs Monate von vorrangiger Bedeutung ist. Diese Untersuchungen schöpfen hierbei die Breite der hier vorgestellten Verteilungsklassen aus und zeigen das damit verbundene Verbesserungspotential bzgl. der Prognosequalität. Hierbei stehen nur etwa drei Jahreszyklen an Datenmaterial zur Verfügung.

Die Verwendung der bedingten Verteilungen und das Lernen der funktionalen Zusammenhänge aus der Gesamtheit aller Datensätze zeigt in beiden praktischen Anwendungen die klaren Vorteile und die Überlegengheit gegenüber alternativen Prognosemodellen aus sowohl der Regressionsanalyse und der Zeitreihentheorie als auch individuell implementierten Verfahren des

maschinellen Lernens.

Die im Folgenden gewählte allgemeine experimentelle Vorgehensweise ist in Abbildung 8.11 skizziert.

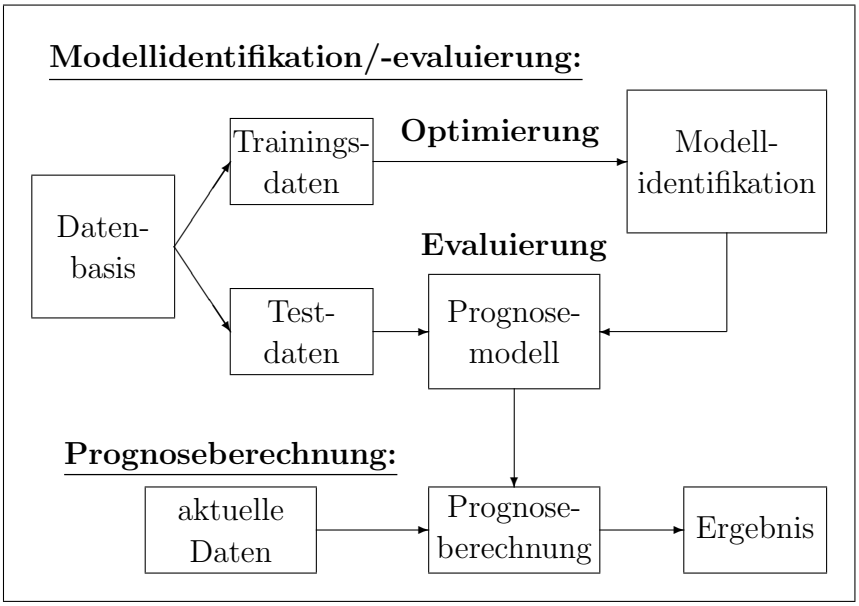


Abbildung 8.11: Allgemeine experimentelle Vorgehensweise der Modellidentifikation und realen Prognose

Es können zwei unabhängige Schritte definiert werden. Der auf bekannten und existierenden Daten basierende Modellierungsprozess setzt sich aus der Modelloptimierung und Modellevaluierung zusammen. Die konkrete Prognoseberechnung erfolgt erst in den operativen Planungsschritten.

Im Modellierungsprozess ist die herkömmliche Trennung von Trainings- und Testdatensätzen eine bekannte Vorgehensweise im Fall einer ausreichend vorhandenen Datenmenge.[16] Auf den Trainingsdaten wird über ein Optimierungsverfahren[17] das Prognosemodell identifiziert. Der funktionale Kern des Modells kann etwa durch ein lineares Modell[18] oder ein neuronales Netz[19] repräsentiert werden.

[16]Diese Prozedur ist unter anderem vom CRISP-Prozess, vgl. (Wirth and Hipp, 2000), bekannt.
[17]Vgl. Kapitel 5.
[18]Vgl. Abschnitt 4.1.
[19]Vgl. Abschnitt 4.2.

Anschließend folgt die Evaluierung auf den bislang unbeachteten Testdaten. Es wird eine künstliche Zukunft simuliert, indem die Daten zu einem bestimmten Zeitpunkt getrennt sind. Eine zufällige Stichprobe aus der gesamten Datenmenge ergibt ebenso eine verifizierende Testmenge.

Kann auf den Trainingsdaten keine zufrieden stellende Güte der Anpassung erreicht werden, ist die ursprüngliche Modellkonzeption zu verändern und anzupassen. Es muss jedoch nicht prinzipiell eine sinnvolle Modellanpassung existieren. Dies tritt etwa ein, falls die unabhängigen Variablen für die Prognosegrößen irrelevant sind und somit keine Aussagekraft besitzen.

Ist das Prognosemodell hinreichend genau und stabil, so kann basierend auf aktuellen Daten eine reale Prognose durchgeführt werden. Dieser abschließende Schritt wird in den folgenden Darstellungen unbeachtet bleiben, da hierbei keinerlei Überprüfung und Bewertung der Ergebnisse möglich ist. Die endgültige Prognoseberechnung wird ausschließlich in operativen Planungsszenarien durchgeführt.

Kapitel 9

Prognose des Ersatzteilebedarfs

Dieses Kapitel beschreibt die Anwendung des Konzepts der Verteilungsschätzung auf die Bedarfsprognose von Ersatzteilen in der Automobilindustrie. Es stellt sich die Aufgabe, die Nachfrage von zirka 700.000 einzelnen Ersatzteilen für die nächsten 15 Jahre zu prognostizieren.[1]

Für die Planung kurz- und langfristiger Zeiträume ist eine qualitativ hochwertige kostenoptimale Prognose erforderlich, da die Bevorratung und Bereitstellung von Ersatzteilen eine sehr sensible Problematik sowohl für das Service-Management als auch für die Kostenplaner der Lagerstätten bedeutet.

Die Bereitstellung von Ersatzteilen, auch für ältere Fahrzeuge bis zu Oldtimern, ist Bestandteil des Serviceangebots und Kundenmanagements renommierter Automobilhersteller, wodurch Ersatzteile sehr lange Zeit auf Vorrat eingelagert werden müssen. Die Zulieferindustrie wird die Produktion auslaufender Teile lange Zeit vor dem tatsächlichen Verschwinden vom Markt einstellen. Daher ist ein derart langer Prognosehorizont von bis zu 15 Jahren für ältere bzw. auslaufende Ersatzteile erforderlich.

Die wirtschaftliche Zwischenbevorratung von vier bis sechs Jahren ist neben der Minimierung der Lagerhaltungskosten eine weitere potentielle Kostenersparnis. Eine größere Bestellung komplexer Teile bietet dem Zulieferer eine optimale Ausnutzung von Fertigungssystemen, wodurch günstige Ein-

[1] Vgl. etwa (Stützle and Hrycej, 2002c).

kaufspreise ermöglicht werden.

Es existieren etliche Gründe für das Bestreben nach einer qualitativ hochwertigen Vorhersage von Ersatzteilen, die es an dieser Stelle nicht alle aufzuzählen gilt. Etwa sind durch qualifizierte Planungen gezielte Verkaufs- oder Werbeaktionen kostengünstig zu realisieren. Ebenso wirkt sich die Verfügbarkeit von Ersatzteilen direkt auf die, zwar schwer zu messende aber strategisch wichtige Kundenzufriedenheit aus, was sich im Image der Gesellschaft widerspiegelt.

Zusammenfassend ist daher eine Prognose der bedingten Wahrscheinlichkeitsverteilungen, sowohl für die aktuelle operative als auch für eine langfristig strategische Planung erforderlich, um kostenoptimale Entscheidungen für die Bedarfsmeldung aller Ersatzteile einzeln berechnen zu können.

9.1 Datengrundlage

Die Datengrundlage besteht im Wesentlichen aus historischen Zeitreihen von Ersatzteilbedarfen. Die meisten hier präsentierten empirischen Untersuchungen basieren auf einer Datenbasis von lediglich 14 bzw. 15 vergangenen Jahren. Es sind Bedarfe von PKW-Teilen seit 1987 bzw. von LKW-Teilen seit 1988 verfügbar.

Unter diesen Voraussetzungen ist sicherlich fraglich, ob eine Methodik aus der klassischen Zeitreihenanalyse für eine qualitativ hochwertige Prognose der nächsten 15 Jahre geeignet sein kann. Die Verwendung von attributbasierten Verfahren wie der linearen Regression oder dem hier vorgestellten Konzept der bedingten Wahrscheinlichkeitsverteilungen ist daher erforderlich. Diese Methoden nutzen für die Modellidentifikation Informationen aus allen Datensätzen gleichzeitig und operieren nicht ausschließlich auf einzelnen Zeitreihen. Die Fähigkeit dieser Ansätze, Informationen über funktionale Abhängigkeiten von exogenen Variablen aus der gesamten Datenmasse zu extrahieren, spricht daher für ihre Verwendung.

Auch wenn Bedarfe systematisch zwischen 1987 bzw. 1988 bis 2001 erfasst wurden, gibt es Teile, die nicht über den gesamten Zeitraum existier-

ten.[2] Dies betrifft Ersatzteile, die noch keine 14 bzw. 15 Jahre alt sind oder schon im betrachteten Zeitraum ausgelaufen sind. Für solche Teile sind die entsprechenden Zeitreihen sogar kürzer als 15 Jahre. Tabelle 9.1 zeigt einige beliebige Beispiele für Zeitreihen von Ersatzteilbedarfen.

1987	...	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
100838	...	181870	153580	110326	7705	1387	0	0	0	0	0
766	...	2114	2748	3285	4218	6478	7576	9249	11108	13548	12373
0	...	0	22	184	280	329	364	304	153	75	10
1	...	0	1	1	0	4	2	4	5	2	0
0	...	0	0	0	0	0	39	29	131	112	148
0	...	237	146	271	254	150	272	138	0	0	0

Tabelle 9.1: Beispiele für Zeitreihen von Ersatzteilbedarfen

Des weiteren ist die Ausnutzung lückenhafter Zeitreihen lediglich unter zusätzlichem Aufwand möglich. Ein Modell mit unvollständigen Inputs oder Outputs kann nur mit Hilfe von numerisch und algorithmisch aufwendigen Methoden erstellt werden.[3]

Neben den einzelnen Teilebedarfen der letzten 15 Jahre sind die individuellen Preise der Ersatzteile bekannt. Zusätzlich steht jedoch für die folgenden Experimente keine weitere Datenquelle zur Verfügung. Ferner können Informationen lediglich aus der eindeutig identifizierenden Teilenummer extrahiert werden.[4]

Es liegen daher folgende Größen als Grundlage für potentielle exogen erklärende Variablen vor:

- die gesamte Ersatzteilnachfrage der letzten 14 bzw. 15 Jahre
- der Preis der einzelnen Ersatzteile
- wesentliche Charakteristika des jeweiligen Ersatzteils, die durch die Teilenummern verschlüsselt sind.

[2]Zum Zeitpunkt der meisten hier präsentierten empirischen Untersuchungen lagen Ersatzteilbedarfe bis zum Jahr 2001 vor. Die Experimente sind jedoch ohne Einschränkung auf das Jahr 2002 erweiterbar. In (Maunz, 2003) sind Daten einschließlich den Bedarfszahlen von 2002 verwendet. Einige Ergebnisse werden an geeigneter Stelle aufgeführt.

[3]Ein Beispiel für diese Art von Methoden ist der EM-Algorithmus. Auf detaillierte Ausführungen wird an dieser Stelle verzichtet, da die Datenmenge auch bei Vernachlässigung solcher Datensätze ausreichend umfangreich ist.

[4]Die Wahl von exogenen Variablen ist unter anderem Thema des Abschnitts 9.2.

Da der Teilebedarf mit hoher Wahrscheinlichkeit aus mehreren quantitativen Faktoren als ausschließlich der historischen Bedarfe und der teilespezifischen Ausfallraten resultiert, könnte durch eine Berücksichtigung etwa des Fahrzeugbestands oder dem Alter der Fahrzeuge eine Verbesserung der Prognosequalität erzielt werden. Leider standen zum Zeitpunkt der hier präsentierten empirischen Versuchsreihen diese Informationen nicht zur Verfügung und waren auch nicht mit akzeptablen Aufwand zu beschaffen. Dieser Aspekt verbirgt jedoch für zukünftige Weiterentwicklungen dieses Anwendungsfalls eine Möglichkeit der Ergebnisverbesserung.

9.2 Konzeption der Modellierung

Dieses Kapitel beschreibt unterschiedliche Schritte der Modellidentifikation und mögliche Varianten der Modellierung. Es werden Grundannahmen getroffen und erste Begründungen für eine Umsetzung des Konzepts der bedingten multivariaten Wahrscheinlichkeitsverteilungen geliefert.

9.2.1 Bildung von Generationen

Da die Nachfrage nach alten und neuen Teilen, im Sinne der verstrichenen Zeit seit der Markteinführung, als unterschiedlich beobachtet werden kann, sind spezielle Modelle für einzelne *„Generationen“* sinnvoll. So sind etwa unterschiedliche Modelle für Teile gebildet, die am Prognosezeitpunkt 10 oder 9 Jahre in Fahrzeugen verbaut waren. So bekommen unterschiedlich alte Ersatzteile die Möglichkeit verschiedenen funktionalen Abhängigkeiten und Gesetzmäßigkeiten zu gehorchen. Tabelle 9.2 zeigt die Definition von Generationen zu unterschiedlichen Prognosezeitpunkten. Die Bildung diverser Modelle auf Basis von Generationen ist durch die große Gesamtdatenmenge möglich.

Formal stützt sich eine Prognose für N zukünftige Jahre im Wesentlichen auf die Bedarfsentwicklung von M vergangenen Jahren. Um die Information über zukünftige Bedarfe mit dem Wissen aus der Vergangenheit assoziieren zu können, werden daher für die Modellidentifikation Zeitverläufe aus $M+N$

Markteinführung	...	Prognosezeitpunkt	Generation
2001		2001	0
1988	...	2001	7
1984	...	1997	7
1977	...	2001	24

Tabelle 9.2: Bildung von Generationen

zusammenhängenden Jahren benötigt.

Gegen die volle Ausnutzung der Zeitreihenlänge[5] spricht einerseits, dass die zeitlich kürzeren Datensätze nicht auf einfache Weise verwendet werden können. Andererseits werden durch kürzere Zeitreihen und dadurch sukzessive Verschiebung des Zeitfensters zusätzliche Trainingsdatensätze gewonnen. So ergeben sich bei P Zeitreihen mit einer Länge von 14 bzw. 15 Jahren eine Anzahl von $(Q+1)P$ Zeitreihen mit $14-Q$ bzw. $15-Q$ Elementen. Durch höhere Datensatzanzahl wird eine höhere Sicherheit bei der Schätzung der Modellparameter erreicht.

Es sprechen jedoch einige Gründe für eine möglichst große Anzahl sowohl der Vergangenheitswerte M als auch der Prognosewerte N. Mit der Verwendung von Werten aus der Vergangenheit trägt eine längere Historie einen größeren Informationsgehalt. Der Informationsbeitrag nimmt jedoch mit dem Zeitabstand vom Zeitpunkt der Prognose ab. So ist für die Vorhersage des Ersatzteilebedarfs von 1999 die Kenntnis des vergangenen Jahres von mehr Bedeutung als der Bedarf aus früheren Jahren, wie etwa 1991.

Die Modellidentifikation und Evaluierung wird aufgrund der Datengrundlage für einen Prognosehorizont von 5 Jahren durchgeführt. Eine Erweiterung des Konzepts auf eine Vorhersage der nächsten 15 Jahre kann - basierend auf dem 5-jährigen Prognosekonzept - durch rekursive Verwendung der prognostizierten Verteilung unter der Normalverteilungsannahme erreicht werden. Im Falle der log-Normalverteilung sind gewisse Approximationen erforderlich, unter deren Verwendung jedoch eine analoge rekursive Vorgehensweise möglich ist. Unter Verwendung eines linearen funktionalen Approximators kann sichergestellt werden, dass die resultierende 15-dimensionale Log-

[5] In diesem Fall wäre $M+N=14$ bzw. $M+N=15$.

Normalverteilung durch rekursive Verwendung von 5-Jahresprognosen in ihrer Verteilungsklasse verbleibt.[6]

Mit der Festlegung eines Prognosehorizonts von 5 Jahren ($N = 5$) und einer verwendeten Historie von ebenfalls 5 Jahren ($M = 5$) werden aus der existierenden Datenmenge 10-jährige Zeitfenster für jedes einzelne Ersatzteil gebildet. Dies erhöht die informative Datenmenge und unterstützt das weitere Vorgehen. Die entstehenden Datensätze eines im Jahr 1977 im Markt eingeführten Ersatzteils ist in Tabelle 9.3 aufgezeigt ($N = 5$, $M = 5$, $P = 1$ und $Q = 5$). Es ergeben sich daher $(Q+1)P = 6$ Datensätze mit einer Länge von $15 - Q = 10$ Elementen.

1987	1988	1989	...	1997	1998	1999	2000	2001		
766	785	1193	...	7576	9249	11108	13548	12373	Zeitreihe	
									Generation	Datensatz
766	785	1193	...	7576					15	1
	785	1193	...	7576	9249				16	2
		1193	...	7576	9249	11108			17	3
			...	7576	9249	11108	13548		18	4
			...	7576	9249	11108	13548	12373	19	5

Tabelle 9.3: Bildung von Datensätzen unter den Vereinbarungen $N = 5$ und $M = 5$

In der Gesamtheit wurden 25 altersspezifische Modelle (vgl. Tabelle 9.2) gebildet und dadurch die gemeinsame Wahrscheinlichkeitsverteilung der Bedarfe für die bevorstehenden 5 Jahre prognostiziert. Jedes Prognosemodell ist für die gesamte Anzahl der ca. 700.000 Teile gültig und differenziert die einzelnen Teile mit Hilfe der oben erwähnten exogen bedingenden Einflussgrößen.

Es könnte durchaus auch das Alter der Teile als differenzierendes Merkmal in Frage kommen. Dann hätte ein einziges Schätzmodell genügt. Da unterschiedlich alte Teile jedoch verschiedene funktionale Abhängigkeiten besitzen

[6]Zur detaillierten Herleitung der Erweiterung des Prognosekonzepts für einen 15-jährigen Prognosehorizont des Ersatzteilebedarfs sei auf die formalen Ausführungen in (Hrycej and Stützle, 2001) verwiesen. Unter anderem basiert die Begründung auf Eigenschaften der Normalverteilung, die alternative Wahrscheinlichkeitsverteilungen, wie die Klasse der hyperbolischen Verteilungen nicht aufweisen. Dadurch ist diese Erweiterung unter Verwendung eines linearen funktionalen Approximators bei den hier betrachteten Verteilungen ausschließlich für die Normalverteilung möglich.

können und die große Datenmenge eine Aufteilung in Submodelle erlaubt, sind Verbesserungen hierdurch nicht zu erwarten.

9.2.2 Verteilungsannahme

Die im Folgenden dargestellten Ergebnisse sind meist unter der Normalverteilungsannahme der Residualterme auf den logarithmierten Datensätzen erreicht. Die Suche nach einer log-Normalverteilung in den hier vorliegenden Daten ist im Wesentlichen durch zwei Argumente gestützt. Da die Bedarfszahlen ausschließlich positiv sind, liegt eine Verteilung nahe, die auf der positiven reellen Achse lebt. Des Weiteren ist, wie bereits erwähnt, eine rekursive Erweiterung der 5-jährigen Modellierung auf einen Prognosehorizont von 15 Jahren unter der Normal- bzw. log-Normalverteilung gültig, so dass die resultierende Verteilung der 15 Zielvariablen wiederum in der Klasse der Normal- bzw. Log-Normalverteilungen liegt.

9.2.3 Extraktion von Testdatensätzen

Da lediglich 14 bzw. 15 Jahre an historischen Bedarfszahlen zur Verfügung stehen, ist es bedenklich, einen ausreichend umfangreichen zusammenhängenden Testzeitraum aus der Datengrundlage zu extrahieren, auf dem eine Generalisierungsfähigkeit des Modells geprüft werden kann. Die Testdaten zur Evaluierung des Prognosemodells werden daher willkürlich nach dem Zufallsprinzip aus der Gesamtheit der Ersatzteilumfangs gewählt. Diese Daten sind gegenüber dem Prozess der Modelloptimierung isoliert (s. Abbildung 8.11). Nach der Berechnung des Prognosemodells wird dieses auf den unbeachteten Testdaten schließlich evaluiert. Diese bekannte und übliche Vorgehensweise fand bereits mehrmals Erwähnung.

9.2.4 Abhängigkeit der Varianz

Generell lässt sich für Ersatzteile mit großer Nachfrage eine kleinere Varianz beobachten als für seltene Teile. Dies bestärkt den hier verwendeten bedingten Ansatz. Diese Information kann etwa über klassische lineare Modelle, die

eine konstante Varianz annehmen, nicht extrahiert und abgebildet werden. Tabelle 9.4 zeigt am Beispiel einer ausgewählten Generation, dass das Prognosemodell eindeutig für Teile mit kleineren Bedarfen eine größere Varianz ausweist und umgekehrt.

Die direkte Folgerung ist die Zuverlässigkeit der Prognose von großen Stückzahlen, die weitaus höher ist als die von kleineren - ein in der Statistik bekanntes Phänomen.[7]

Bedarf	Anzahl der Teile	Mittlere Varianz	Minimale Varianz	Maximale Varianz	Streuung der Varianz
> 1.000.000	17	2,96e-03	8,12e-04	6,01e-03	1,61e-03
> 100.000	170	8,86e-03	8,12e-04	0,154	1,40e-02
> 10.000	1.140	0,322	8,12e-04	76,2	3,50
> 1.000	4.877	0,793	8,12e-04	265	6,84
> 100	13.018	2,38	8,12e-04	914	22,2
> 10	22.540	8,08	8,12e-04	2,36e+03	67,6
> 1	24.722	13,2	8,12e-04	3,21e+03	104
< 10	1.818	63,4	4,09e-02	3,21e+03	267
< 100	11.661	25,3	1,27e-02	3,21e+03	149
< 1.000	19.840	16,3	4,31e-03	3,21e+03	116
< 10.000	23.617	13,8	2,91e-03	3,21e+03	106
< 100.000	24.552	13,3	1,34e-03	3,21e+03	104
< 1.000.000	24.705	13,2	1,13e-03	3,21e+03	104
<10.000.000	24.722	13,2	8,12e-04	3,21e+03	104

Tabelle 9.4: Varianzeigenschaften bei unterschiedlich häufigen Ersatzteilen an einer repräsentativen Teilmenge des gesamten Ersatzteilbestandes

Es hat sich des Weiteren eine starke Abhängigkeit der prognostizierten Varianzen der Zielgrößen vom Preis der einzelnen Teile herausgestellt. Ebenso wie bei den Bedarfen ergibt sich hierbei ein Zusammenhang, dass die kostenaufwendigeren Teile eine kleinere Varianz besitzen, als die günstigen Kleinteile.

Ebenso kristallisiert sich eine starke Korrelation von $\rho \approx 0,5$ zwischen den fünf individuellen Prognosejahren heraus.[8] Aus der zeitreihentheoretischen Sichtweise wird dies als Autokorrelation bezeichnet. Jedoch ist hierbei

[7] Diese Phänomen wird als das *Gesetz der großen Zahlen* bezeichnet.

[8] ρ bezeichnet hierbei den Korrelationskoeffizienten.

im Gegensatz zu der Varianz kein exogener charakterisierender Einfluss zu vermuten und zu erkennen.

9.2.5 Definition der exogenen Variablen

Ein wesentlicher Schritt bei der Modellidentifikation ist die Bestimmung der exogenen Inputvariablen des Prognosemodells. Es könnten die zur Verfügung stehenden Daten direkt als Einflussgrößen herangezogen werden. Allerdings ist es in den meistens praktischen Anwendungen hilfreich geeignete Transformationen der ursprünglichen Daten als Inputdaten des Prognosesystems zu verwenden.

Schließlich stellen sich im vorliegenden Fall der Ersatzteilprognose folgende Abhängigkeiten des Lokationsvektors und der Strukturmatrix als bedeutend heraus.[9] Zur Begründung dieser Aussage sei auf das Kapitel 9.5 verwiesen.

Aus den Bedarfszeitreihen des 10-jährigen Zeitfensters

$$\bar{x}_{-4}, \bar{x}_{-3}, \ldots, \bar{x}_0, \bar{x}_1, \ldots, \bar{x}_5,$$

wobei x_0 den letzten bekannten Bedarf bezeichnet, gehen

- die Bedarfsanteile $x_{-4} = \bar{x}_{-4}/s, \ldots, x_0 = \bar{x}_0/s$ der letzten fünf Jahre

als exogene Inputvariablen des Erwartungswertvektor in die Modellierung ein. Hierbei stellt $s = \sum_{i=-4}^{0} \bar{x}_i$ den gesamten Teilebedarf der letzten fünf Jahre dar und spielt die Rolle eines Normierungsfaktors. Eine Normierung ist speziell bei Datensätzen notwendig, deren Einträge sich in einigen Größenordnungen unterscheiden. Wie in den Tabellen 9.1 und 9.4 ersichtlich ist, existiert eine breite Spanne - 1 bis 10 Millionen Stück - von Bedarfen unterschiedlicher Ersatzteile.

Um jedoch von den Prognosewerten zu realen Bedarfszahlen zu gelangen, ist es erforderlich, die durch das Modell bestimmten zukünftigen Werte mit

[9] Im Fall der Normalverteilungsannahme sind die Bezeichnungen Lokationsvektor und Strukturmatrix äquivalent zum Vektor der Erwartungswerte und der Kovarianzmatrix.

den entsprechenden Normierungsfaktoren zu verrechnen. Dies ist in Abbildung 9.1 skizziert.

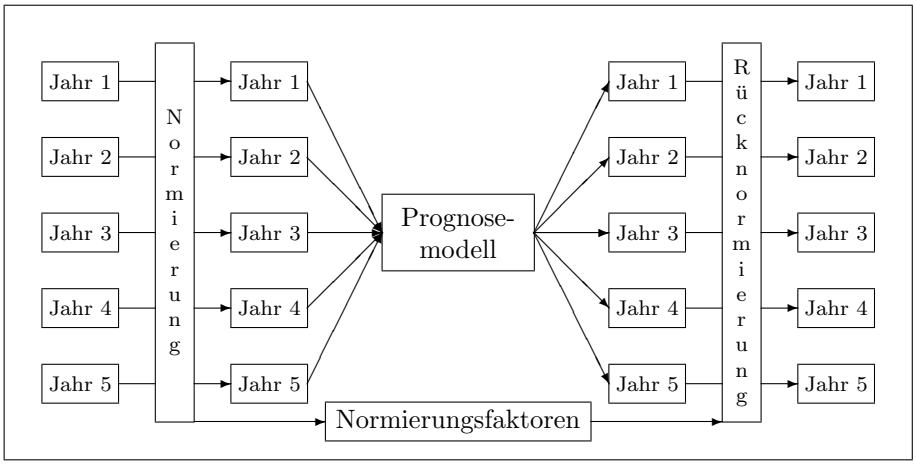


Abbildung 9.1: Normierung und Rücknormierung zur Berechnung zukünftigen Ersatzteilbedarfe

Die Datensätze aus Tabelle 9.3 sind normiert in Tabelle 9.5 abgebildet. Hierbei ist ersichtlich, dass sich für jeden Datensatz ein anderer Normierungsfaktor ergeben kann.

1987	1988	1989	...	1997	1998	1999	2000	2001		
766	785	1193	...	7576	9249	11108	13548	12373	Zeitreihe	
									$\sum_{i=-4}^{0} \bar{x}_i$	Datensatz
0.133	0,136	0,206	...	1,311					5778	1
	0,110	0,167	...	1,063	1,298				7126	2
		0,131	...	0,834	1,018	1,222			9089	3
			...	0,678	0,827	0,993	1,212		11181	4
			...	0,538	0,657	0,789	0,962	0,879	14080	5

Tabelle 9.5: Normierung der Datensätze zur Verwendung als exogene Inputvariablen

Wie bereits in Abschnitt 9.1 angedeutet, sind nicht ausschließlich vergangene Bedarfe eine Informationsquelle für das Prognosemodell zur Berechnung zukünftiger Ersatzteile.

Binäre Attribute sind aus dem eindeutigen sogenannten Marketing-Code, im Folgenden auch als Teilenummer bezeichnet, eines Ersatzteils extrahiert und neben den Bedarfsanteilen als exogene Einflüsse des Lokationsvektors

definiert.[10]

Die bedeutendste Charakteristik eines Ersatzteils lässt sich aus der ersten Stelle des Marketing-Codes ablesen, wodurch

- eine Zuordnung zu Personenwagen oder Nutzfahrzeugen möglich ist.

Weitere wichtige differenzierende Eigenschaften der einzelnen Teile bietet die zweite und dritte Ziffer der Teilenummer. Über diese Klassifizierung lassen sich die Ersatzteile in folgende Gruppen unterteilen:

- Teile und Tauschteile,
- Teile und Tauschteile für Oldtimer,
- Aggregate und Tauschaggregate,
- Aggregate und Tauschaggregate für Oldtimer,
- Zubehör,
- Boutique-Artikel,
- Zubehör mit A-Teile-Nummer und
- Sonderwerkzeuge.

Die Tatsache, dass alle Ersatzteile unter anderem in knapp 100 so genannte Hauptgruppen eingeteilt werden können, bietet eine weitere Information, die ebenfalls als Inputvariablen in die Modellierung des Lokationsparameters einfließen und eine erklärende Eigenschaft versprechen.[11] Es sind aus dieser Kategorisierung 28 binäre Attribute abgeleitet, da diese mit ausreichender Stückzahl besetzt sind.[12] Die restlichen werden nicht unterschieden und somit durch den selben binären Vektor charakterisiert.

[10] Der Marketing-Code eines Keilriemens für Personenwagen ist etwa 1 10 150 und für Nutzfahrzeuge 2 10 150.

[11] Die Zuordnung zu den Hauptgruppen ergibt sich ebenfalls aus der Teilenummer.

[12] So beinhaltet die Hauptgruppe 54 (elektrische Ausstattung und Installation) die meisten Teile mit einer Anzahl von 28861.

Es ergibt sich nach Bündelung der binären Attribute schließlich eine Anzahl von 38 exogenen Inputvariablen des Prognosemodells für den Erwartungswertvektor.

Nach ausführlichen Untersuchungen[13] stellen sich für die Varianz drei wesentliche Einflüsse als charakterisierend heraus:

- die logarithmierte Summe der Ersatzteilbedarfe der letzten fünf Jahre ($\log s = \log \sum_{i=-4}^{0} \bar{x}_i$)
- die Entropie bezüglich der letzten fünf Jahresanteile ($\sum_{i=-4}^{0} x_i \log x_i$)
- der logarithmierte Preis des entsprechenden Ersatzteils ($\log p$).

In Kapitel 9.5 ist die bedingte heteroskedastische Modellierung und die Wahl dieser exogenen Inputvariablen durch Ergebnisse aus empirischen Experimenten begründet.

9.3 Anpassungsgüte an empirische Daten

Zu Beginn der experimentellen Untersuchung werden zwei alternative Zerlegungen bzw. Repräsentationen der inversen Strukturmatrix verglichen. Die Modellierung ist analog zu obigen Beschreibungen erfolgt. Die Größe des Zielfunktionswertes, d.h. des Werts der Maximum-Likelihood Funktion, lässt eine Beurteilung der Anpassung an die empirischen Daten zu. Je kleiner der Maximum-Likelihood Zielfunktionswert ist, desto kleiner ist die Differenz zwischen empirischer und geschätzter Dichtefunktion, d.h. desto besser konnte die empirische Dichtefunktion durch das Prognosemodell angenähert werden. Nicht zu beurteilen ist hierdurch die Modellgüte bzw. die Prognosegüte dieses Modells.

Tabelle 9.6 zeigt die Überlegenheit der in Abschnitt 4.3.3 vorgeschlagenen Zerlegung der inversen Strukturmatrix in $(UD)^T UD$, die eine Separation der Skalierung und der Korrelationen bewirkt, gegenüber der klassischen Cholesky-Dekomposition $A^T A$.[14] Es ist ein entscheidender Vorteil sowohl auf

[13] Tabelle 9.7 beinhaltet die wesentlichen Ergebnisse dieser Untersuchungen.

[14] Die klassische Cholesky-Zerlegung findet etwa in (Williams, 1996) Verwendung.

Zerlegung der Strukturmatrix	Training	Test
A^TA	2,8008	0,7976
$(UD)^TUD$	8,7902	8,8269

Tabelle 9.6: Likelihood-Zielfunktionswerte

den Trainings- als auch auf den Testdatensätzen zu erkennen und somit ist eine eindeutig bessere Anpassung an die empirische Verteilung erreicht.

Zu den formalen Begründungen dieser Darstellung bzw. Zerlegung der inversen Strukturmatrix sei auf die Ausführungen des Abschnitts 4.3.3 verweisen.

9.4 Kostenberechnung von Unter- bzw. Überschätzung

Für die erste Beurteilung der Güte eines Prognosemodells existiert eine Vielzahl von unterschiedlichen Fehlermaßen.[15] Die Güte, der Nutzen und der praktische Mehrwert eines Prognoseergebnisses kann bei den meisten praktischen Aufgaben lediglich mit Hilfe eines Maßes beurteilt werden, in welchem die Zielsetzung berücksichtigt ist und zum Ausdruck kommt.[16] Aus diesem Grund wird ein Fehlermaß formuliert, das die verursachten Kosten einer Unter- bzw. Überschätzung berücksichtigt, da es in erster Linie gilt diese zu minimieren.

Den nachfolgenden Experimenten liegt daher eine Kostenformel zugrunde, die sich wie folgt beschreiben lässt.

Sie berechnet die aus der Prognose entstandenen Gesamtkosten K_{ges} aus der Abweichung zwischen den prognostizierten Bedarfen und der tatsächli-

[15] Beispiele für herkömmliche Fehlermaße sind der mittlere absolute/relative Fehler, der mittlere prozentuale Fehler, der mittlere quadratische Fehler oder der Theil'sche Ungleichheitskoeffizient.

[16] So muss laut (Schwarze, 1980) zur Beurteilung eines Prognosemodells:„in erster Linie ein Fehlermaß verwendet werden, das dem Optimierungs- bzw. Anpassungskriterium des Modells entspricht."

chen Nachfrage der Teile. Durch Multiplikation dieser Differenz mit dem entsprechenden Kostenfaktor k_1^k oder k_2^k - hier unabhängig von t - und Kumulation über alle betrachteten Teile ergibt sich schließlich der totale Kostenbetrag. Formal lassen sich die Gesamtkosten formulieren als:

$$K_{ges} = \sum_{k=1}^{K} \sum_{t=1}^{T} \delta(y_t^k - \hat{y}_t^k) k_1^k (y_t^k - \hat{y}_t^k) + \delta(\hat{y}_t^k - y_t^k) k_2^k (\hat{y}_t^k - y_t^k), \tag{9.1}$$

wobei $\hat{y}_t^k$ den geschätzen Bedarf des k-ten Teils zum Zeitpunkt t und δ die Dirac-Funktion[17] beschreibt.

Basierend auf den Gesamtkosten kann das bei diesen Experimenten vorwiegend verwendete Gütemaß bestimmt werden. Die Kostenquote ergibt sich im Wesentlichen auf Basis von Gleichung (9.1). Es werden die durch die Fehlprognose entstandenen Kosten in Relation zum existierenden Umsatz des jeweiligen Ersatzteils ausgedrückt, so dass sich die sogenannte *Kostenquote* schreiben lässt als:

$$K_{quote} = \sum_{k=1}^{K} \frac{p^k \sum_{t=1}^{T} |y_t^k - \hat{y}_t^k|}{U_k}, \tag{9.2}$$

wobei $U_k = p^k \sum_{t=1}^{T} y_t^k$ den Umsatz des k-ten Ersatzteils während des Prognosezeitraums von T Jahren bezeichnet. Die Wahl des Preises p^k des k-ten Ersatzteils als Kostenfaktor ist von Experten vorgegeben, jedoch theoretisch durch jede andere Maßzahl ersetzbar. Die Gewichtung durch den Preis bewirkt, dass die umsatzstarken Teile, die sogenannten Kostentreiber, geeignet verstärkt in die Bewertung einfließen. Dieses Maß findet Verwendung, falls die jährlichen Fehlschätzungen innerhalb des Prognosehorizonts relevant sind, da sich jährliche Unter- und Überschätzungen während des Prognosezeitraums nicht ausgleichen können.

Liegt hingegen das Hauptaugenmerk auf der Endbevorratung eines Ersatzteils, so ist lediglich der Fehlbestand am Ende des betrachteten Zeitraums von Interesse. Es ist in diesem Fall irrelevant, ob zu einem früheren Zeitpunkt

[17] Siehe zur Definition der Dirac-Funktion Abschnitt 5.1.

eine Überschätzung vorlag, wenn im Laufe des Prognosezeitraums diese Teile aus dem Lagerbestand abverkauft und gleichzeitig unterschätzt werden. Hierzu ist ein Fehlermaß formuliert, das diese Anforderungen erfüllt:

$$K_{EBR} = \sum_{k=1}^{K} |\frac{p^k \sum_{t=1}^{T} y_t^k - \hat{y}_t^k}{U_k}|. \tag{9.3}$$

Im Folgenden ist jedoch, falls nicht explizit bemerkt, die Kostenquote aus Gleichung (9.2) als Fehlermaß verwendet.

9.5 Exogene Inputvariablen der Verteilungsprognose

Die Ergebnisse dieses Kapitels verifizieren die Definition der exogenen Inputvariablen aus Abschnitt 9.2.5. Die Bestimmung der unabhängigen Variablen und die Annahme der Log-Normalverteilung beim Ansatz der Verteilungsprognose für die Problematik der Ersatzteilprognose ist durch Tabelle 9.7 begründet.

Die optionalen Inputgrößen sind in Kapitel 9.2 aus den zur Verfügung stehenden Informationen abgeleitet und beschrieben. Analog zu den dortigen Ausführungen bedeutet in Tabelle 9.7 die Beschriftung

- 5: die Verwendung der letzten fünf Jahresbedarfe
- 9: die Verwendung der Inputs, die aus den Marketing-Codes generiert sind
- 38: die Verwendung der Inputs, die aus der Zuordnung zu den Teile-Hauptgruppen entstehen
- 3: die drei exogenen Einflüsse $\log \sum_{i=-4}^{0} \bar{x}_i$, $\sum_{i=-4}^{0} x_i \log x_i$ und log p, ebenfalls beschrieben in Kapitel 9.2.

Die Datenbasis dieser Untersuchung besteht, begründet durch beschränkte Ressourcen, aus Ersatzteilen, die zum Prognosezeitpunkt bereits 20 Jahre

		Erwartungswert					
		5+0		5+9		5+38	
		Normal	Log-Normal	Normal	Log-Normal	Normal	Log-Normal
V	0	36,27/36,10	35,97/34,92	36,19/35,92	35,88/34,78	35,39/34,95	35,01/34,43
A	3	37,64/37,21	33,74/33,85	36,10/36,71	33,66/33,75	35,86/37,23	**33,21/33,63**
R	3+9	35,63/35,74	33,64/33,84	35,78/36,11	33,58/33,72	35,44/36,58	33,04/33,73
	3+38	35,81/35,97	33,72/33,88	35,87/36,33	33,64/33,83	35,47/36,95	33,06/34,12

Tabelle 9.7: Verifikation der Wahl der exogenen Inputvariablen für die Verteilungsprognose [in %]

existierten, d.h. eine spezielle „Generation von Teilen“. Diese repräsentative Untermenge umfasst 30.791 Zeitreihen.

Die beiden Ziffern in jeder Spalte beziehen sich auf die Kostenquote bzgl. der Trainings- und Testmenge. Als Testmenge wurde in diesem Fall zufällig ein Drittel der Datensätze aus der gesamten Datengrundlage entnommen.

Auf Grund dieser Untersuchung wird für die folgenden Experimente der Erwartungswert mit allen zur Verfügung stehenden Inputvariablen bedingt, wohingegen die Varianz heteroskedastisch unter drei exogenen Einflüssen modelliert wird.

Auffällig ist bereits bei dieser Untersuchung, dass im Fall der Log-Normalverteilung ein eindeutiger Unterschied in der Prognosequalität zwischen einer homoskedastischen (Varianz: 0) und heteroskedastischen (Varianz: > 0) Modellierung vorliegt. Dass sich dies für den normalverteilten Fall so nicht widerspiegelt, liegt an der schlechten Approximation des Prognosemodells an die empirische Verteilung. Speziell für seltene Teile kann die Normalverteilung, die auf der kompletten reellen Achse lebt, keine akzeptable Prognose liefern. Offensichtlich weist die Modellierung basierend auf der Log-Normalverteilung über alle betrachteten Versionen eine bessere Prognosegüte auf.

9.6 Benchmark-Methoden aus der Praxis

Bevor die Ergebnisse aus empirischen Untersuchungen präsentiert und beschrieben werden, stellen die beiden folgenden Abschnitte Ansätze vor, die von Fach-Experten als Benchmark-Methoden bezeichnet werden. Diese Verfahren bilden daher aus praktischer Sicht einen Maßstab, an dem sich jede Methode zur Ersatzteilplanung zumindest zu messen hat.

9.6.1 Lineare Approximation mit determinierten Koeffizienten

Das bislang in der Praxis verwendete Verfahren hat sich durch Expertenwissen im Laufe der Zeit entwickelt. Der Grundgedanke hinter diesem Verfahren ist die aus der Statistik bekannte lineare Regression. Es wurde während der letzten Jahre durch sukzessive Veränderung der Regressionskoeffizienten eine Gleichung definiert, auf welcher die Prognose des operativen Geschäfts basiert. Das Prognosemodell besteht aus einer gewichteten Summe der letzten Jahresbedarfe, die für alle Teile gleichermaßen angewendet wird.

Die letzten fünf Jahresbedarfe werden folglich i. Allg. durch unterschiedliche Faktoren gewichtet. Je weiter der Jahresbedarf vom Prognosezeitpunkt entfernt liegt, desto weniger wird dieser berücksichtigt. Konkret zeigt Tabelle 9.8 die Berechnungsvorschriften für unterschiedlich alte Ersatzteile.[18]

Teilealter (Jahre)	Berechnung des Prognosewerts
≥ 5	$y_t = \frac{1}{14}(8y_{t-1} + 2,5y_{t-2} + 1,5y_{t-3} + y_{t-4} + y_{t-5})$
4	$y_t = \frac{1}{13}(8y_{t-1} + 2,5y_{t-2} + 1,5y_{t-3} + y_{t-4})$
3	$y_t = \frac{1}{12}(8y_{t-1} + 2,5y_{t-2} + 1,5y_{t-3})$
2	$y_t = \frac{1}{10,5}(8y_{t-1} + 2,5y_{t-2})$
1	$y_t = y_{t-1}$

Tabelle 9.8: Berechnungsvorschriften einer in der Praxis angewendeten Prognosemethode

Die Berechnung von späteren Jahresbedarfe $y_{t+i}, i = 1, \ldots, T$ ergibt sich sodann sukzessive aus realen und prognostizierten Werten.

9.6.2 Clustering-Methode

Eine weitere in der Praxis entwickelte und verwendete Methodik zur Ersatzteilprognose basiert auf dem Prinzip der Clusterbildung aus der künstlichen Intelligenz. Dieses Verfahren unterscheidet die zu prognostizierenden Ersatzteile nach den in Abschnitt 9.1 beschriebenen Generationen. Es wird

[18] Für eine ausführliche Beschreibung und praktische Umsetzung dieser Methodik sei auf (Maunz, 2003) verwiesen.

angenommen, dass sich die Bedarfe gleichaltriger Ersatzteile in der Zukunft „ähnlicher“ verhalten werden, als Teile die älter oder jünger sind. Daher wird für jede Generation ein eigenes Prognosemodell erstellt.[19] Die Trainingsdatensätze einer Generation werden nach einem festgelegten Abstandsmaß[20] in 100 Cluster aufgeteilt.[21] Charakterisierend hierfür sind ausschließlich die vergangenen Bedarfe.

Fällt nun das zu prognostizierende Teil in ein solches Cluster, so ist der Median dieses Clusters als Prognosewert des jeweiligen Ersatzteils für das kommende Jahr definiert.[22]

Nachfolgende Jahre werden ebenso wie beim Verfahren aus Abschnitt 9.6.1 sukzessive aus den bekannten und vorhergesagten Bedarfen auf analoge Weise bestimmt.

Ein offensichtliches Defizit dieser Methodik ist, dass es nicht auf junge Teile anzuwenden ist. Ersatzteile, die keine Historie von mindestens fünf Jahren aufweisen, können zum momentanen Stand der Implementierung nicht mit diesem Verfahren vorhergesagt werden. Der Grund hierfür ist, dass zur Bestimmung der Zugehörigkeit eines Teils zu einem Cluster, die Kenntnis der letzten fünf Vergangenheitswerte notwendig ist.

9.6.3 Vergleich der Benchmark-Methoden aus der Praxis

Der sehr stark von der Praxis getriebene Vergleich zwischen den Benchmark-Methoden der Abschnitte 9.6.1 und 9.6.2 und einem auf Basis des Cluster-Algorithmus nachgeahmten Verfahren ist im Folgenden präsentiert.

Der Vergleichsmaßstab für die Gegenüberstellungen der Methoden basiert auf Annahmen, die sich zwar so üblicherweise in der Praxis kaum wieder-

[19] Diese Vorgehensweise entspricht der für die Verteilungsprognose, vgl. Kapitel 9.2.

[20] Das in diesem Fall gewählte Abstandsmaß ist die quadratische Distanz $d = \sum_{t=1}^{T}(y_t - \bar{y}_t)^2$, wobei $\bar{y}$ den sogenannten Kern des Clusters bezeichnet.

[21] Die Anzahl der Cluster wurde bei der Implementierung dieser Methodik festgelegt. Es hat sich bislang kein Indiz gezeigt, dass eine Änderung verbesserte Ergebnisse mit sich bringen würde.

[22] Eine detaillierte Beschreibung dieses Verfahrens befindet sich in (Maunz, 2003).

finden, jedoch durch die eingeschränkten Fähigkeiten dieser Verfahren unumgänglich sind. So wird etwa ein symmetrisches Kostenverhältnis vorausgesetzt. Aus diesem Grund wird der Preis p^k des jeweiligen Ersatzteils sowohl als Kostensatz einer Über- als auch einer Unterdeckung angenommen. Dies bedeutet formal $k_1^k = k_2^k = p^k$.

Ein erster Vergleich der Benchmark-Methoden und einem kommerziellen Software-Paket ist in Tabelle 9.9 dargestellt. Als bewertendes Fehlermaß findet hierbei die Kostenquote aus Gleichung (9.2) Verwendung.

Methodik	Kostenquote
Clustering-Methode	38,98%
nachgebildeter Cluster-Algorithmus	40,02%
Lineare Approximation	45,11%

Tabelle 9.9: Kostenquoten für Benchmark-Prognosemethoden aus der Praxis

Der Cluster-Algorithmus erzielt in diesem Vergleich das beste Ergebnis. Die, mit Hilfe einer kommerziell erhältlichen alternativen Data-Mining-Software, nachgebildete Cluster-Methode erreicht jedoch eine ähnliche Güte von knapp über 40%. Die lineare Approximation weist mit über 45% das schlechteste Ergebnis auf.

Auf Grund dieser Prognoseergebnisse wird im Folgenden die Clustering-Methode im Vergleich zu weiteren Verfahren betrachtet.

9.6.4 Weitere Untersuchungen anhand der Verteilungsprognose

Aus der Vielzahl von Experimenten und Analysen in (Maunz, 2003) seien im Folgenden die wesentlichen Ergebnisse zusammengefasst. Die Arbeit (Maunz, 2003), eine Beauftragung des Anwendungsbereichs, hat den Vergleich obiger Benchmark-Methoden und der hier entwickelten Verteilungsprognose unter verschiedenen Gesichtspunkten zum Fokus.

9.6.4.1 Methodenvergleich

Basierend auf einer Datengrundlage von Jahresbedarfen zwischen 1987 und 2002 wurde eine strikte zeitliche Trennung von Trainings- und Testzeitraum vorgenommen. Als bekannte Vergangenheit sind die Jahre 1987 bis 1995 gewählt, wodurch 1995 die simulierte Gegenwart darstellt. Die restlichen Jahre zwischen 1996 und 2002 sind als Testzeitraum reserviert. Insgesamt besteht die Trainings- und Testmenge in diesem Zeitraum aus etwa 230.000 Teilen, etwa der Hälfte aller momentan geführten 470.000 Teile.[23] Es sind schließlich etwa 156.000 Teile in den folgenden Experimenten berücksichtigt, da die Clustering-Methode für die restlichen ca. 74.000 Ersatzteile keine Ergebnisse liefern würde.[24]

Die Wahl eines Testzeitraums bei ohnehin sehr kurzen Zeitreihen, bezogen auf den Prognosehorizont, ist sicherlich bedenklich, hat jedoch den Vorteil, dass alle Ersatzteile in der Testmenge enthalten sind und daher Analysen auch auf Teilegruppen durchführbar sind.[25]

In Tabelle 9.10 sind sowohl die Kostenquote als auch die absoluten Kosten der Unter- und Überdeckung der Benchmark-Methoden und der Verteilungsprognose gegenübergestellt.

Methodik	Kostenquote	Kosten der Unter- bzw. Überdeckung
Clustering-Methode	39%	6,0 Mrd. Euro
Lineare Approximation	48%	7,4 Mrd. Euro
Verteilungsprognose	36%	5,5 Mrd. Euro

Tabelle 9.10: Vergleich von Benchmark-Prognosemethoden aus der Praxis und der Verteilungsprognose

[23] Die Ausschlusskriterien garantieren die Existenz des Teils ab 1995 und die Aktivität des Teils im Zeitraum von 1987 bis 2002.

[24] Das Defizit der Clustering-Methode, nicht auf junge Teile anwendbar zu sein, wurde bereits im Abschnitts 9.6.2 erwähnt.

[25] Bzgl. Auswertungen auf einzelnen Teilegruppen nach Umsatzanteil, Bestellhäufigkeit, Absatzmenge, Verkaufspreis, Volumen (Klein-, Mittel- und Sperrigteilen), Alter oder Sparten (PKW oder LKW) sei auf (Maunz, 2003) verwiesen.

Bei einem Gesamtumsatz von 15,4Mrd. Euro in den Jahren 1996 bis 2002, ergeben sich beträchtliche Kosten, die durch Fehleinschätzungen der Verfahren begründet sind. Das Potential der Kostenersparnis von knapp 2Mrd. Euro bzw. 0,5Mrd. Euro in 7 Jahren wird hierbei deutlich.

Des Weiteren zeigt Tabelle 9.11 die natürliche Verschlechterung der Prognosegüte im Laufe der Zeit, die bei allen drei Verfahren gleichermaßen zu erkennen ist.

Methodik	1996	1997	1998	1999	2000	2001	2002
Clustering-Methode	26%	33%	40%	46%	54%	61%	65%
Lineare Approximation	30%	39%	47%	55%	65%	75%	84%
Verteilungsprognose	**24%**	**31%**	**36%**	**43%**	**50%**	**58%**	**63%**

Tabelle 9.11: Kostenquoten der Benchmark-Prognosemethoden aus der Praxis und der Verteilungsprognose einzelner Prognosejahre, siehe (Maunz, 2003), Seite 64

Als Fazit kann aus den Untersuchungen in (Maunz, 2003) eindeutig geschlossen werden, dass die Verteilungsprognose bzgl. des Fehlermaßes der Kostenquote den beiden Benchmark-Methoden überlegen ist.

9.6.4.2 Varianzschätzung als Sicherheitsaussage

Im Zuge der Analysen in (Maunz, 2003) wurde ein Vergleich zwischen der Varianzschätzung und der Kostenquote durchgeführt, der die Aussagekraft der Varianz als Unsicherheitsmaß einer Prognose bestärkt. Die Ergebnisse sind in Tabelle 9.12 zitiert.

Es zeigt sich eindeutig, dass Ersatzteile mit einer durchschnittlich größer geschätzten Varianz ebenfalls eine höhere Kostenquote aufweisen. Umgekehrt formuliert wird bei Ersatzteile, deren Zukunftswerte sehr präzise vorhergesagt wurden, eine relativ kleine Varianz prognostiziert. Die von der Verteilungsprognose zusätzlich berechnete Information - die Varianz - erbringt daher einen echten Mehrwert zur Beurteilung des Prognosewerts. Sie sollte daher zur optimalen Entscheidung ausgenutzt werden.

Varianz	Kostenquote	Anzahl der Teile	Anteil
$\sigma^2 \geq 2$	55%	55.651	35,7%
$1,9 \leq \sigma^2 > 2$	50%	47.982	30,8%
$1,8 \leq \sigma^2 > 1,9$	38%	18.094	11,6%
$1,7 \leq \sigma^2 > 1,8$	34%	12.375	8,0%
$1,6 \leq \sigma^2 > 1,7$	30%	5742	3,7%
$1,5 \leq \sigma^2 > 1,6$	28%	13.877	8,9%
$1,4 \leq \sigma^2 > 1,5$	25%	1.369	0,9%
$1,3 \leq \sigma^2 > 1,3$	20%	464	0,3%
$\sigma^2 < 1,3$	19%	92	0,1%

Tabelle 9.12: Vergleich von Varianzschätzung und Kostenquote, siehe (Maunz, 2003), Seite 76

In (Maunz, 2003) wird etwa eine Ampelbewertung der prognostizierten Teile auf Basis ihrer Varianz vorgeschlagen, um die unsicheren Prognosen von den sicheren zu trennen und für die kritischen Vorhersagen entsprechende Maßnahmen einleiten zu können.

Die geschätzte Varianz kann jedoch auf jede andere Weise zur Beurteilung der Prognose oder zur Berechnung einer optimalen Entscheidung herangezogen werden. So wird etwa in Kapitel 9.8, auf Basis einer allgemeinen Verlustfunktion, eine Vorschrift zur Bestimmung der optimalen Bestellmenge unter Kenntnis der bedingten Wahrscheinlichkeitsverteilung vorgestellt und angewendet.

Es zeigt sich an einem praktischen Beispiel, dass die Varianzaussage der Verteilungsprognose als Maß der Unsicherheit einen echten und direkten Mehrwert gegenüber herkömmlichen Methoden erzeugt. Zusätzlich ist hierdurch verifiziert, dass die in Abschnitt 2.2.3 motivierend formulierte Anforderung an ein Prognosesystem, die Fähigkeit einer Variabilitätsbeurteilung zu besitzen, umgesetzt ist.

9.7 Vergleich alternativer Prognosemethoden

Es genügt nicht dem Anspruch eines alternativen Verfahrens lediglich gegenüber den in der Praxis verwendeten Methoden zu bestehen. Zur allgemeinen Beurteilung einer Methodik ist es notwendig, diese mit anderen wissenschaftlichen Ansätzen zu vergleichen.

In diesem konkreten Beispiel versprechen sämtliche Methoden aus der Zeitreihenanalyse auf Grund der vorliegenden Datenbasis wenig Erfolg. Sie sind daher aus dieser Betrachtung ausgeschlossen.[26]

Es existieren Methoden aus der klassischen Statistik und der künstlichen Intelligenz, die sich zum Vergleich mit dem hier präsentierten Prognosekonzept der bedingten multivariaten Wahrscheinlichkeitsverteilungen und der in der Praxis verwendeten Clustering-Methode eignen.

Die für einen Vergleich gewählte Methode aus der Statistik ist der bekannte und bewährte Ansatz der linearen Regression.[27] Die unabhängigen Variablen sind somit die vergangenen fünf Jahresbedarfe.

Das in der Neuroinformatik klassische Werkzeug zur Berechnung von Prognosen sind neuronale Netze. Daher wird ein Multi-Layer Perzeptron für einen Vergleich herangezogen.[28] Es ist hierbei ein Multi-Layer Perzeptron mit einer Zwischenschicht, die aus 5 Neuronen besteht, verwendet. Die Eingangsschicht beinhaltet die bekannten fünf vergangenen Jahresbedarfe.

Der Vergleich zwischen Prognosemethoden aus der klassischen Statistik (Lineare Regression), der künstlichen Intelligenz (Multi-Layer Perzeptron) und der Verteilungsprognose aus dieser Arbeit auf Basis des gewichteten Kostenquotienten (9.2) ist in Tabelle 9.13 aufgeführt.[29]

Die in den Clustering-Algorithmen verwendete Zielgröße des Medians lässt sich analytisch schwer behandeln und ist daher im Wesentlichen nur

[26] Im nachfolgenden Anwendungsfall, der Nachfrageprognose von Nutzfahrzeugen, zeigt sich die schlechte Prognosegüte von Zeitreihenmethoden auf zu kurzen Datensätzen.

[27] Siehe Kapitel 7.1.

[28] Siehe Kapitel 4.2.

[29] Wegen der Gewichtung durch den Umsatz ist die Kostenquote der Clustering-Methode in den Tabellen 9.9 und 9.13 nicht identisch.

Methodik	Kostenquote bei Gewichtung durch den Umsatz	Prozentuale Verbesserung
Clustering-Methode	44,10%	
Lineare Regression	39,82%	9,7%
Multi-Layer Perzeptron	42,92%	2,6%
Verteilungsprognose	**35,82%**	18,8%

Tabelle 9.13: Kostenquoten für Varianten klassischer Prognosemethoden aus der Statistik und der künstlichen Intelligenz

im Zusammenhang mit Cluster-Methoden verwendbar. Daher wurde in der erweiterten Studie ein Ersatz gesucht, der sich als Mittel der Begrenzung des negativen Einflusses der Daten-Ausreißer bewährt.

Dass sich die Gewichtung der Datensätze durch den Umsatz auszahlt, liegt an dem Gesetz der großen Zahlen. Da Ausreißer insbesondere bei niedrigen Stückzahlen auftreten, erhalten sie geringe Gewichte und haben dadurch geringen Einfluss auf das gesuchte Prognosemodell. Gleichzeitig trägt diese Gewichtung dem höheren Kosteneinfluss umsatzstarker Teile Rechnung.

Obwohl eine Gewichtung der Datensätzen einer Reduzierung der Datenmenge gleichkommt, ist der Vorteil der Gewichtung durch den Umsatz eindeutig ersichtlich. Es stellt sich heraus, dass ohne eine derartige Gewichtung die Varianz durch die hoch variierenden Kleinteile im Durchschnitt etwa 10-fach überschätzt wird.

Erstaunlich an den Ergebnissen ist, dass alle wissenschaftlichen Ansätze eine bessere Prognosequalität aufweisen als die in der operativen Planung eingesetzte Clustering-Methode.

Die größte Verbesserung gegenüber der Clustering-Methode erbringt die Verteilungsprognose. Sie steigert die Prognosegüte um ca. $18,8\%$. Die nichtlineare Prognosemethode - das Multi-Layer Perzeptron - bringt in diesem Zusammenhang keine wesentliche Verbesserung. Lediglich $2,6\%$ Unterschied ergeben sich beim Vergleich zur Clustering-Methode. Die lineare Regression erweist sich zwar als etwas zuverlässigere Methode, indem sie eine Steigerung der Prognosequalität von 9,7% erzeugt. Dennoch liegt auch sie noch weit vom Niveau der Verteilungsprognose entfernt.

Weiterhin ist zu den einzelnen Prognosemethoden die Dimension der zu bewerkstelligenden numerischen Optimierungsaufgabe zu bemerken. Die freien Parameter der linearen Regression belaufen sich bei einem Prognosehorizont von 5 Jahren auf 30, bei einem Multi-Layer Perzeptron mit 5 Neuronen in der Zwischenschicht auf 180 und bei einem Cluster-Algorithmus mit 100 Clustern auf 3.000. Die Verteilungsprognose verwendet in diesem Beispiel ein lineares Modell als funktionalen Approximator und benötigt daher 100 freie Schätzparameter. Tabelle 9.14 zeigt die Anzahl freier (zu optimierender) Parameter der unterschiedlichen Methoden.

Methodik	Anzahl der freien Schätzparameter
Clustering-Methode	3.000
Lineare Regression	30
Multi-Layer Perzeptron	180
Verteilungsprognose	100

Tabelle 9.14: Freie Parameter der Optimierungsaufgaben bei unterschiedlichen Prognosemethoden

9.8 Entscheidung unter asymmetrischen Kosten

Wie einleitend bereits erwähnt, ist die Kostenasymmetrie der Unter- bzw. Überschätzung eine wichtige Eigenschaft dieser praktischen Prognoseproblematik im Ersatzteilemanagement. Im Falle einer Überschätzung, d.h. einer zu hohen Teilenachfrage bei den Zulieferern, verbleiben die Teile in den Lagern und erzeugen Kosten, die hier mit k_1 bezeichnet sind. Kostenquellen in diesem Fall sind etwa Lager- oder Kapitalbindungskosten.

Falls gegenteilig eine Bestellung weniger Teile beinhaltet als die tatsächliche Nachfrage erfordert, sind die betreffenden Teile durch kostenaufwendige Nachbestellungen zu beschaffen. Dieser Kostensatz sei mit k_2 bezeichnet. Diese Kostenart ist leider schwer greifbar, da hierzu etwa teure Sonderbestel-

lungen in Auftrag gegeben werden müssen. In Extremfällen sind die nachbestellten Teile sogar in aufwendiger Einzelarbeit zu fertigen.

Eine Kostenasymmetrie liegt also für den Fall $k_1 \neq k_2$ vor. Dieser realistische Fall fordert daher eine individuelle Bestimmung der optimalen Bestellmenge, da der geschätzte Erwartungswert in jedem kostenasymmetrischen Fall eine suboptimale Entscheidung bedeutet.

Weiterhin kann jedes einzelne Ersatzteil ein individuelles Kostenverhältnis besitzen, das a priori von Experten oder durch ökonomische Verfahren festzulegen ist. Dies ist mit diesem Konzept ohne zusätzlichen Aufwand abbildbar.

Um also die optimale Bestellmenge zu bestimmen, ist die Kenntnis der bedingten Nachfrageverteilung erforderlich, wie bereits in Gleichung (2.2) und (2.5) motivierend dargestellt ist. O'Hagan schlug eine bilineare Kostenfunktion

$$L(a,y,x) = \begin{cases} (a-y)k_1, & \text{falls} \quad y < a \\ (y-a)k_2, & \text{falls} \quad y > a \\ 0, & sonst, \end{cases} \tag{9.4}$$

vor, die in diesen Experimenten Anwendung findet.[30] Es ist konzeptionell jede andere beliebige Funktion der Kosten vorstellbar und umsetzbar.

Die Kostenfaktoren k_1 und k_2 seien von der endogenen Einflussgröße unabhängig. Die Kosten ergeben sich daher zu:

$$K(a,y,x) = \int_{-\infty}^{a} d(y|x)(a-y)k_1 dy + \int_{a}^{\infty} d(y|x)(y-a)k_2 dy, \tag{9.5}$$

wobei $d(y|x)$ die geschätzte bedingte Wahrscheinlichkeitsdichte beschreibt.

Unter der hier angenommenen Gauß'schen Normalverteilung ergibt sich eine kostenoptimale Entscheidung durch Minimierung der Kostenfunktion (9.5) nach der Vorschrift:

$$\hat{y}_{opt} = \mu(x) + \sigma(x) G^{-1}\left(\frac{k_2}{k_1 + k_2}\right), \tag{9.6}$$

[30] Siehe (O'Hagan, 1994).

wobei G^{-1} die Inverse der standardnormalen Verteilungsfunktion bezeichnet. An der Entscheidungsregel (9.6) ist offensichtlich, dass für die Berechnung der kostenoptimalen Bestellmenge $\hat{y}_{opt}$ sowohl der bedingte Erwartungswert $\mu(x)$ als auch die bedingte Standardabweichung $\sigma(x)$ der Prognose benötigt werden.

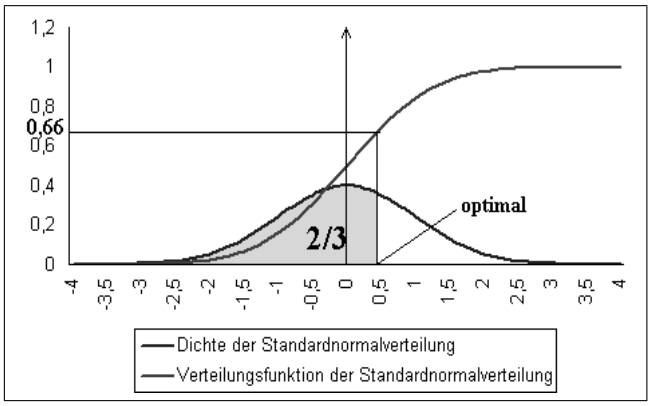


Abbildung 9.2: Bestimmung der kostenoptimalen Bestellmenge eines Ersatzteils unter dem Kostenverhältnis $k_1 : k_2 = 1 : 2$

Abbildung 9.2 zeigt beispielhaft für das Kostenverhältnis $k_1 : k_2 = 1 : 2$ eines speziellen Ersatzteils die Verschiebung der kostenoptimalen Entscheidung nach rechts bezogen auf den Erwartungswert. Da in diesem Fall eine Unterschätzung des tatsächlichen Bedarfs den doppelten Kostensatz einer Überschätzung verursachen würde, ist eine Absicherung, d.h. einen Sicherheitsbestand im Lager zu halten, die sofortige und logische Schlussfolgerung. Unter der Annahme einer korrekten Verteilungsschätzung ist daher eine kostenoptimale Entscheidung möglich.

9.9 Kostenoptimale Entscheidung mit Hilfe der Verteilungsprognose

Mit dem Ziel der Bestimmung einer kostenoptimalen Bestellmenge werden bei unterschiedlichen Kostenverhältnissen die Berechnungen der Clustering-

Methode und der Verteilungsprognose gegenübergestellt.[31] Zur Illustration genügt für diese Untersuchung eine repräsentative Datenmenge von 4.450 Teilen.[32] Eine Multiplikation mit dem Faktor 100 entspricht in etwa der realen Situation aller aktiven Ersatzteile. Die dargestellten Kosten der Unter- und Überdeckung beziehen sich in den folgenden Ausführungen und Tabellen auf den Prognosezeitraum von 5 Jahren.

Für den folgenden Vergleich wird die Maßzahl der Gesamtkosten aus Gleichung (9.1) verwendet, um die Kostenersparnis der einzelnen Prognose quantifizieren zu können.

Methodik	reales Kostenverhältnis $k_1 : k_2$		
	1:1	1:3	1:5
Clustering-Methode	43.735	76.489	109.243
Verteilungsprognose	35.377	70.607	100.142

Tabelle 9.15: Gesamtkosten der Über- und Unterdeckung unterschiedlicher Prognosemethoden [in 1.000 Euro]

Tabelle 9.15 gibt die resultierenden Gesamtkosten der in der Praxis eingesetzten Clustering-Methode im Vergleich zur multivariaten bedingten Verteilungsprognose bei unterschiedlichen Kostenverhältnissen wieder. Schon bei einem symmetrischen Kostenverlauf ($k_1 : k_2 = 1 : 1$) erzielt die Verteilungsprognose eine Reduzierung der Gesamtkosten von zirka 19,1%. Bezogen auf alle momentan aktiven Ersatzteile wäre dies eine Kostenersparnis von ca. 835Mio. Euro über den Zeitraum von 5 Jahren. Dies entspricht etwa dem Ergebnis (18,8%) aus Tabelle 9.13.

Weiterhin sind in Tabelle 9.16 die berechneten Gesamtkosten für drei unterschiedliche Kostenverhältnisse $k_1 : k_2$ dargestellt. Sowohl die Optimalität der Entscheidungsregel (9.6) als auch der finanzielle Vorteil kommen hierbei zum Ausdruck. Da die geringsten Gesamtkosten bei jedem realen Kostenverhältnis durch die entsprechende Entscheidungsregel erzielt wurden, wird

[31] Siehe Tabelle 9.13.

[32] Zufällig wurde ein Prozent der Ersatzteile ausgewählt.

Entscheidungs-regel	reales Kostenverhältnis $k_1 : k_2$ 1:1	1:3	1:5
1:1	**35.377**	72.295	109.212
1:3	39.172	**70.607**	102.043
1:5	41.279	70.710	**100.142**

Tabelle 9.16: Gesamtkosten der Verteilungsprognose bei unterschiedlichen Entscheidungsregeln und realen Kostenverhältnissen [in 1.000 Euro]

die Optimalität der Entscheidung deutlich. Die dadurch resultierende Ersparnis leitet sich aus den Differenzen zu den anderen Entscheidungsregeln ab. Würde etwa bei einem realen Kostenverhältnis von 1 : 3 eine symmetrische Entscheidungsregel angenommen werden, so würden bereits bei einem Prozent der Ersatzteile im Vergleich zur korrekten Entscheidungsregel Mehrkosten von zirka $1,688$ Mio. Euro entstehen.

Zusätzlich evaluieren die Ergebnisse die Verteilungsmethodik, da nach der Anwendung der optimalen Entscheidungsregel (9.6) die Diagonalen der Tabelle die geringsten Gesamtkosten in den jeweiligen Spalten aufweisen.

Kapitel 10

Absatzprognose von Nutzfahrzeugen

Das Schätzprinzip von multivariaten bedingten Wahrscheinlichkeitsverteilungen fand neben der Bestimmung des zukünftigen Ersatzteilebedarfs bei weiteren Prognoseaufgaben Anwendung. Im Folgenden wird die Identifikation eines Verteilungsmodells, das den Absatz von Nutzfahrzeugen in den nächsten Monaten berechnet, beschrieben und mit herkömmlichen Verfahren verglichen.[1]

Motiviert durch die Zielsetzung der kostenminimalen Produktions- und Absatzplanung sowie den Herausforderungen des Supply Chain Managements ist die Bestimmung der zukünftigen Verkaufszahlen eine fundamentale Aufgabenstellung.

Der Absatzprognose der entsprechenden Konsumgüter, in diesem Beispiel von Nutzfahrzeugen, liegt sowohl eine kostenoptimale Produktions- und operative Planung als auch ein zu verbessernder Informationsfluss zwischen den produzierenden und den zuliefernden Akteuren sowie den Endverbrauchern eines Liefernetzwerkes motivierend zugrunde. Dies lässt sich auf die meisten Güter, die Teil eines Liefernetzwerkes sind, verallgemeinern. Die Basis für eine mehrwertschöpfende Informationspolitik ist unter anderem die Güte, aber auch die Zuverlässigkeit und der Informationsgehalt von Prognoseaussagen.

[1] Vgl. etwa (Stützle and Hrycej, 2002b).

So ist auch in der Automobilbranche, die traditionell ein ausgeprägtes Zuliefernetzwerk besitzt, die Qualität der zu kommunizierenden Prognose ein mit großen Bemühungen verfolgtes Ziel. Ist es etwa möglich, die Nachfrage von Nutzfahrzeugen für die kommenden Monate mit dem zusätzlichen Sicherheitsmaß der Streuung zu bestimmen, so besitzt dies sowohl für die Zulieferindustrie als auch für die integrierte Produktionsprogrammplanung der Unternehmung ein enormes Potential der Kostenersparnis.

Dass der Markt die Fabrik steuert, ist im Zeitalter der großen Konkurrenz, der alle Organisationen der Automobilindustrie ausgesetzt sind, seit langer Zeit bekannt. Der integrierte Programmplanungsprozess erfordert eine Prognose auf den unterschiedlichsten Aggregationsebenen und für verschiedenste Prognosehorizonte. So stellen

- Aggregationstiefe:

  Gesamtbetrachtung, Baureihe (BR), Baumustergruppe (BMG),

  Fahrzeugbaumuster (FBM), Code und Teile

- Prognosehorizont:

  Jahre, Monate, Tage

- Vertriebsregionen:

  Gesamtbetrachtung, Länder, Vertriebsbereiche

die wesentlichen Dimensionen des realen Planungsprozesses dar. Hierbei treten die genannten Ausprägungen in unterschiedlichen Kombinationen auf.

Die in diesem Kapitel vorgestellten Prognosemodelle sind letztendlich jedoch vor allem begründet und definiert durch die zur Verfügung stehenden Datengrundlagen. Die folgenden Darstellungen repräsentieren daher eine gewisse Auswahl dieser vielzähligen Planungsaufgaben. Analoge Prognosemodelle lassen sich jedoch mit entsprechenden spezifischen Anpassungen für andere Aggregatebenen und Zeiträume bestimmen. Abhängig von etwa dem Prognosehorizont und der Produkttiefe sind unterschiedliche exogene Einflüssse zu vermuten und daher ist eine modifizierte Modellierung erforderlich, auch falls gewisse Informationen übergreifend wirken können. Dennoch

ist die Verfügbarkeit der Daten grundlegend und unabdingbar für die praktische Erstellung eines Prognosemodells.

10.1 Datengrundlage

Bevor die Konzeption des Prognosemodells konkretisiert wird, sei die zur Verfügung stehende Datengrundlage skizziert. Die den präsentierten Experimenten zugrunde liegenden Informationen sind die nachfolgend beschriebenen Zeitreihen:

- die vergangenen Monatsverkäufe x_t von Nutzfahrzeugen der Jahre 1998 bis 2001

- die aus dem Markt bekannten realen Auftragszahlen der zukünftigen 6 Monate

- ausgewählte makroökonomische Indikatoren, die von Experten als relevant eingeschätzt werden, wie etwa für den deutschen Markt: schwedische Krone, 3-Monatszins, 10-jährige Rendite, Indikatoren für die Industrieproduktion und die Kapazitätsauslastung, das IFO-Geschäftsklima, die Einzelhandelsumsätze und die Aufträge im Metall- und Baugewerbe

- eine durch Experten definierte Tabelle, die jedes Fahrzeugbaumuster einem der Segmente Bau-, Sonder-, Fern- oder Verteilerverkehr zuordnet

Neben der bekannten zeitlichen Zuordnung der vergangenen Absatzzahlen sind zusätzlich implizite Informationen aus der eindeutig zugeordneten Fahrzeugidentifikationsnummer (FIN) zu extrahieren. Es lassen sich charakteristische Fahrzeugattribute identifizieren, wie etwa:

- die Baureihe: Actros, Atego-Leicht, Atego-Schwer oder Econic,

- die Fahrzeugart: Pritschenwagen, Kipper, Betonmischer, Sattelzugmaschine, Kommunal- oder Feuerwehrfahrzeug,

- die Motorleistung in PS,
- die Achseneigenschaften: Anzahl der Achsen, Anzahl der antreibenden Achsen und Anzahl der lenkbaren Achsen,
- die Tonnage in t,
- die Federungsart: Luft- oder Stahlfederung und
- der Radstand in mm.

Diese Informationsquellen werden nun bezüglich ihrer Relevanz geprüft und schrittweise in die Modellierung aufgenommen, um ihren Mehrwert zu überprüfen und darzustellen.

10.2 Konzeption des Prognosemodells

Die folgenden Abschnitte sind im Wesentlichen der Frage gewidmet, auf welche Weise die Absatzprognose von einzelnen Nutzfahrzeugen über mehrere Monate optimal unter der Verwendung realer Datengrundlagen stattfinden kann.

Eine besondere Eigenschaft dieser Aufgabe, die eine statistisch fundierte Prognose erschwert, ist die Variantenvielfalt der Produkte in der Nutzfahrzeugbranche. Wegen der sehr konkreten und hohen Anforderungen von Kunden und die spezifischen Einsatzgebiete der Fahrzeuge werden derzeit im Durchschnitt pro Tag kaum identische Nutzfahrzeuge pro Produktionsstätte angefertigt. Zusätzlich dazu ist das Flottengeschäft mit identischen Fahrzeugen eine weitere erschwerende Eigenschaft des Nutzfahrzeuggeschäfts, das Schwankungen der Nachfrage bewirken kann.

Nach ausführlichen Vergleichen unterschiedlicher Modellvarianten nehmen prinzipiell für die folgenden Experimente dieses Abschnitts die aus den existierenden Datenmengen extrahierten unabhängigen exogenen Größen

- Absatzzahlen der letzten 6 Monate $(x_{-5}, \ldots, x_0)$,
- 12 Saisonalitätsfrequenzen $(\sin \frac{\pi i t}{6}, \cos \frac{\pi i t}{6}, i = 1, \ldots, 6)$,

- maximal 23 Produktattribute sowie

- maximal 9 makroökonomische Indikatoren für den deutschen Markt

Einfluss auf den Lokationsparameter der Wahrscheinlichkeitsverteilung. Ferner ist die Skalierung abhängig von

- der logarithmierten Summe der Verkaufszahlen der letzten 6 Monate ($\log s = \log \sum_{i=-5}^{0} x_i$),

- der Entropie bezüglich der letzten 6 Monatsabsätze ($\sum_{i=-5}^{0} x_i \log x_i$) sowie

- der Entropie bezüglich der letzten 3 Jahresumsätze ($\sum_{j=-2}^{0} a_i \log a_i$)

modelliert, wobei die nichtdiagonalen Elemente der Strukturmatrix unbedingt geschätzt werden.

Die exogenen Einflussgrößen, wie etwa die Produktattribute und makroökonomischen Indikatoren, sind bei den nachfolgend beschriebenen Versuchen auf unterschiedliche Weise variiert, um dadurch deren Relevanz zu testen und zu überprüfen.

Das hierbei verwendete Fehlermaß ist der relative gewichtete geometrische Prognosefehler

$$err_g = \left(\prod_{k=1}^{K} e^{w_k |y_k - \hat{y}_k| / y_k} \right)^{1/\sum_{k=1}^{K} w_k}, \tag{10.1}$$

wobei w_k die absolute Gesamtstückzahl des $k-ten$ Fahrzeugbaumusters der letzten sechs Monate bezeichnet. Mit Hilfe einer Gewichtung durch die absoluten Stückzahlen wird der bekannte geometrische Durchschnitt angepasst. Formal wird hierdurch lediglich der Tatsache Rechnung getragen, dass größere Stückzahlen mehr Einfluss auf das Gütemaß haben sollten als geringe. Der geometrische Durchschnitt ist dem arithmetischen Mittel aufgrund der geringeren Empfindlichkeit gegenüber Ausreißern zu bevorzugen.

Die erste konzeptionelle Entscheidung besteht in der Wahl des Prognosehorizonts, der im folgenden Abschnitt kurz behandelt ist, falls dieser nicht exogen vorgegeben wird.

10.2.1 Prognosehorizont

Mehrmonatige Prognosen können entweder getrennt für einzelne Monate oder simultan über den gesamten Prognosehorizont erstellt werden. Für das zweite Vorgehen spricht die Möglichkeit, die Korrelationen zwischen einzelnen Prognosemonaten zu schätzen und dadurch zusätzliche Informationen zu gewinnen. Da diese Option zu den wichtigen Merkmalen und Vorteilen der vorgestellten Verteilungsmethode zählt, wird im Folgenden die simultane Prognose aller Monate des jeweils betroffenen Zeithorizonts berechnet. Durch die relativ kurze Zeitspanne, für die Absatzdaten verfügbar sind, ist zunächst ein Prognosehorizont von 6 Monaten gewählt. Eine Erweiterung auf einen längeren Horizont ist geradlinig, wobei gegebenenfalls eine natürliche Verschlechterung der Prognosegüte in Kauf genommen werden muss.

10.2.2 Modelle für Produktgruppen

Eine zu entscheidende Frage der Modellkonzeptionierung ist die Wahl der Produktgruppenebene, auf welcher individuelle Prognosemodelle zu formulieren sind. Mit dem Ziel der optimalen Informationsextraktion und des Umstandes der sehr kurzen Zeitreihen wird eine Modellierung auf den unteren Hierarchiestufen, auf denen etwa Zeitreihenmethoden operieren, wenig Potential für qualitativ hochwertige Prognoseaussagen haben.

Ein einziges Gesamtmodell könnte hingegen aus allen Datensätzen, d.h. allen Fahrzeugbaumustern, gleichzeitig funktionale Zusammenhänge identifizieren. Dies ist unter anderem eine Stärke der attributbasierten Regressionsansätze, zu denen ebenfalls die Methode der bedingten multivariaten Wahrscheinlichkeitsverteilungen zählt.

Abbildung 10.1 zeigt die potentiellen Produktgruppierungen, für welche eine Modellierung aufgrund der Datengrundlage möglich ist. So ist jedes Baumuster eindeutig einer Baureihe zugeordnet, die wiederum einem Segment zugeteilt werden kann.

Eine Modellierung auf der unteren Ebene der einzelnen Fahrzeugbaumuster mit einer zugehörigen Motorart entspricht der Vorgehensweise klassischer Methoden aus der Zeitreihenanalyse. In diesem Fall wären 4.248

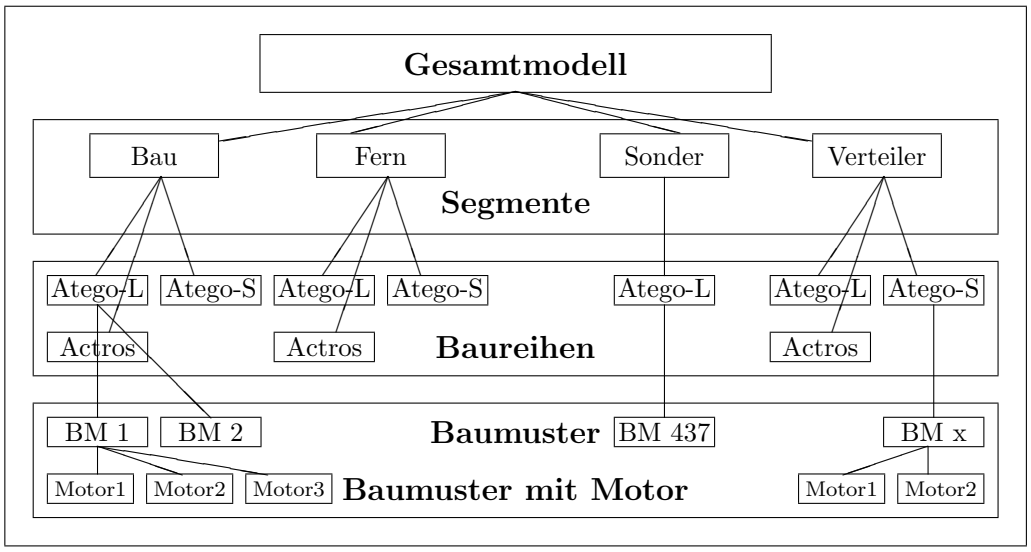


Abbildung 10.1: Hierarchieebenen von Produktgruppen

Prognosemodelle zu berechnen. Alternativ kann jedoch bei attributbasierten Verfahren, wie den linearen Modellen oder der Verteilungsprognose, ein gemeinsames Modell für eine ganze Produktgruppe formuliert und identifiziert werden. Dabei bietet sich die Möglichkeit, durch Einbeziehung der Produktattribute spezifisches Verhalten der einzelnen Fahrzeugbaumuster zu modellieren. Somit ist wiederum der Vorteil einer sehr viel größeren Datenmenge erreicht und die Generalisierungsfähigkeit des Prognosemodells verbessert. Bei der Gruppierung in Kombinationen von Segment und Baureihen ergeben sich 10 zu identifizierende Prognosemodelle. Die 4 Segmente stellen nach der Gesamtebene die am stärksten aggregierte Hierarchiestufe dar. Es wird im Folgenden vorwiegend mit Hilfe dieser Modelle argumentiert.

10.3 Vergleich der Verteilungsprognose mit klassischen Prognosemethoden

In empirischen Versuchen wird die Prognosemethodik von bedingten Wahrscheinlichkeitsverteilungen einigen klassischen Vorgehensweisen aus der Zeitreihen- und Regressionsanalyse gegenübergestellt.

Bei der Anwendung der relativ aufwendigen Prognosemethode der bedingten Verteilungen stellt sich immer wieder die Frage, ob ähnliche Ergebnisse nicht auch mit einfacheren klassischen Mitteln etwa aus der Zeitreihen- oder Regressionsanalyse möglich sind.

Auf der Datengrundlage, die in Kapitel 10.1 dargestellt ist, wurde die Prognoseaufgabe von Nutzfahrzeugabsätzen sowohl mit der Verteilungsprognose als auch mit einem einfachen FBM-spezifischen Zeitreihenmodell, das Trend und Saisonalität abbildet, und mit einem homoskedastischen autoregressiven Regressionsmodell durchgeführt.

Modellierung	Funktion	Param.	Fehler(Train.)	Fehler
Zeitreihenanalyse	linear	81	0,0952	0,9365
lineares Modell	linear	135	0,3393	0,4278
Verteilungsprognose	linear	153	0,3305	0,3815

Tabelle 10.1: Vergleich der Verteilungsprognose mit klassischen Prognosemodellen

Im Folgenden sei nun die hier vorgestellte Prognosemethodik der bedingten multivariaten Wahrscheinlichkeitsverteilungen klassischen Methoden gegenübergestellt. Tabelle 10.1 zeigt, dass eine einzelne Zeitreihe zu kurz ist, um eine Generalisierungsfähigkeit auf Testdaten zu erreichen. Dadurch erklärt sich der enorme Prognosefehler des Zeitreihenmodells auf den Testdatensätzen, da es keinerlei Generalisierungsfähigkeit besitzt.

Ein homoskedastisches Regressionsmodell ist stets von schlechterer Prognosequalität als die heteroskedastische Verteilungsprognose, was speziell auf der Testdaten zum Ausdruck kommt.

Insgesamt ist jedoch die Prognosegenauigkeit enttäuschend. Der relative Fehler liegt auf der Testmenge zwischen 35% und 40%.

10.4 Modellierung mit flexiblen Wahrscheinlichkeitsverteilungen

Die Möglichkeit der optimalen Ausnutzung der zur Verfügung stehenden Informationen sowie der Qualitätssprung der Prognosegüte aufgrund der Verwendung flexibler Verteilungsklassen bestimmen den Inhalt dieses Abschnitts.

Wie bereits erwähnt sind die zum Prognosezeitpunkt bekannten Auftragseingänge eine weitere Möglichkeit, die Prognosequalität entscheidend zu verbessern.

Aus diesem Grund wird der Inputvektor des Prognosemodells um diese Variablen ergänzt und es ergeben sich für den Lokationsparameter die Größen:

- die Absatzzahlen der letzten 6 Monate,
- 12 Saisonalitätsfrequenzen sowie
- die realen und bekannten Auftragszahlen der zukünftigen 3 Monate

als bedingende Einflüsse. Die Strukturmatrix bleibt nach vergleichenden Experimenten von den existierenden Aufträgen unabhängig. Die Modellierung nimmt eine Abhängigkeit der Skalierung von folgenden Inputvariablen an:

- der logarithmierten Summe der Verkaufszahlen der letzten 6 Monate,
- der Entropie bezüglich der letzten 6 Monatsabsätze sowie
- der Entropie bezüglich der vergangenen 3 Jahresumsätze.

Tabelle 10.2 zeigt den relativen durchschnittlichen Prognosefehler einer 6-monatigen Vorhersage unterschiedlicher Prognosemodelle unter Verwendung verschiedener Informationsquellen. Die verwendeten exogenen Variablen für den Lokationsparameter variieren zwischen der Nachfrage der letzten sechs Monate (past demand: pd), der Saisonalität (season: s) und den bekannten Auftragszahlen der nächsten drei Monate (existing orders: eo). Der gewichtete geometrische Durchschnittsfehler der Prognose zeigt eine $22,8\%$-ige

Modellierung	Verteilung	Information	Prognosefehler
homoskedastisch	normal	pd	0,3514
homoskedastisch	normal	pd + s	0,3393
homoskedastisch	normal	pd + s + eo	0,2409

Tabelle 10.2: Fehler der Sechs-Monatsprognose bei unterschiedlichen Informationen auf Basis der Trainingsmenge

Verbesserung durch die Verwendung der existierenden Aufträge und der Saisonalität als Informationsquelle.

Die homoskedastische Modellierung entspricht klassischen linearen Modellen, da ausschließlich der Lokationsparameter der Normalverteilung als bedingt unter exogenen Variablen angenommen wird (vgl. Tabelle 10.1).

Die Gefahrenquelle von Inputgrößen - wie den bekannten Aufträgen für zukünftige Verkäufe - ist ihre „Unsicherheit“ des Eintretens. Im Zeitraum zwischen Januar 1999 und Juli 2001 lagen jedoch bei nur 4 unmittelbar folgenden Monaten die Auftragsstände um mehr als 25% von den späteren Verkäufen entfernt, so dass sich diese Monatsstände als Störeinflüsse entpuppten.

Modellierung	Verteilung	Prognosefehler	
		6 Monate	1 Monat
homoskedastisch	normal	0,240	0,161
heteroskedastisch	normal	0,235	0,127
heteroskedastisch	stabil	0,223	0,039
heteroskedastisch	hyperbolisch	0,212	0,025

Tabelle 10.3: Qualitätsverbesserung der Monats- bzw. 6-Monatsprognose durch Verwendung alternativer Verteilungsklassen und des heteroskedastischen Verteilungsansatzes.

Abschließend sei die Prognosequalität zusätzlich für den ersten und in der Praxis wichtigsten Prognosemonat verglichen. Je näher der Prognosewert an der Gegenwart liegt, desto bedeutender ist er für die momentane operative Planung, d.h. desto mehr Kosten werden durch eine Fehlprognose verursacht

bzw. durch eine exzellente Vorhersage eingespart. Ebenso wird im Konzept des Supply Chain Managements den zeitlich nahen Prognosewerten mehr Bedeutung zugemessen als den noch weit entfernten, da diese natürlicherweise mit sehr viel mehr Unsicherheit behaftet sind.

Tabelle 10.3 zeigt eindrucksvoll die unterschiedliche Güte homoskedastischer[2] und heteroskedastischer Prognosemodelle unter verschiedenen Verteilungsannahmen. Der relative gewichtete geometrische Fehler der Monatsprognose des homoskedastischen Normalverteilungsmodells wird durch die heteroskedastische Modellierung um 21% verbessert. Des Weiteren kann eine 76%-ige Steigerung der Prognosequalität durch die Verwendung der stabilen Verteilungsfamilie anstelle der Normalverteilungsklasse erreicht werden. Schließlich erzielt jedoch die Annahme einer heteroskedastischen hyperbolischen Verteilung eine 84%-ige Steigerung.

Diese Ergebnisse zeigen den Mehrwert bezüglich der Prognosegüte durch die Verwendung der entwickelten Verteilungsmethodik. Abschließend kann daher dem präsentierten validierten Prognosesystem bedingter Wahrscheinlichkeitsverteilungen die Funktionsfähigkeit und die Generierung eines signifikanten Mehrwerts attestiert werden.

10.5 Prognosequalität einzelner Modellvarianten unter verschiedenen Informationsquellen

Die Entscheidung, für welche Produktgruppen eigenständige Prognosemodelle erstellt werden, welche Informationen letztendlich für eine erfolgreiche Modellierung maßgeblich sind und ob eine Verallgemeinerung, wie die nichtlineare Modellierung, einen Mehrwert erbringt, ist Bestandteil dieses Abschnitts.

[2]Homoskedastizität bedeutet in diesem Zusammenhang, wie auch in Tabelle 10.2, dass ausschließlich die Lokation von den exogenen Variablen abhängig gewählt wurde. Im Fall von Heteroskedastizität ist zusätzlich die Strukturmatrix bedingt unter exogenen Variablen.

Stellt sich heraus, dass ein Gesamtmodell über alle Fahrzeugbaumuster mit beschreibenden binären Attributen, die Informationen über das Segment und die Baureihe des Baumusters tragen, gibt es wenig Gründe, eine alternative Modellierung zu bevorzugen und obige Untersuchen sind hinreichend.

Weiterhin wird die Relevanz und Aussagekraft der einzelnen, bereits genannten Informationsquellen verifiziert bzw. falsifiziert.

Aufgrund alternativ erstellter Prognosemodelle mit variierender Wahl der exogenen Variablen kann daher der Informationsgehalt einzelner Quellen bewertet werden. Die Prognoseleistung auf der Trainings- und Testmenge gibt sodann Aufschluss über die Güte des zusätzlichen Informationsbeitrags.

Der Vollständigkeit halber sei vorweg erwähnt, dass der funktionale Zusammenhang zwischen den Verteilungsparametern und den exogenen Variablen während der folgenden Versuchsreihe durch eine lineare Abbildung beschrieben ist. Die Normalverteilung ist die im Folgenden zugrunde liegende Verteilungsannahme.

Modell	Attribute	Indikatoren	Param.	Train.	Fehler	Test	Fehler
Gesamt-	nein	nein	153	4.248	0,3305	152	0,3815
Gesamt-	ja	nein	291	4.248	0,3262	152	0,3905
Gesamt-	ja	ja	345	4.248	0,3232	152	0,4857
Bau-	nein	nein	153	1.007	0,3265	33	0,4232
Bau-	ja	nein	291	1.007	0,3272	33	0,4470
Bau-	ja	ja	**345**	1.007	0,3238	33	**0,9332**
Fern-	nein	nein	153	1.426	0,3113	49	0,3952
Fern-	ja	nein	291	1.426	0,3048	49	0,4030
Fern-	ja	ja	345	1.426	0,2865	49	0,7090
Verteiler-	nein	nein	153	1.766	0,3218	68	0,3535
Verteiler-	ja	nein	291	1.766	0,3197	68	0,3688
Verteiler-	ja	ja	345	1.766	0,3158	68	0,5590

Tabelle 10.4: Informationsgehalt auf unterschiedlichen Hierarchieebenen

In Tabelle 10.4 zeigt sich, dass alle Informationsquellen nur einen relativ geringen Anteil von Nachfrageschwankungen erklären. Dies äußert sich bereits auf der Trainingsmenge durch einen durchschnittlichen relativen Pro-

gnosefehler von zirka 30%. Wie zu erwarten ist, verbessert sich zwar die Anpassung des Prognosemodells auf der Trainingsmenge mit erhöhter Anzahl von Schätzparametern, was jedoch keine allgemeine Aussage über die Güte und Generalisierungsfähigkeit des Modells zulässt.

Nach Anwendung des optimierten Prognosemodells auf die bislang unbeachteten Testdatensätze zeigt sich eine extreme Verschlechterung des relativen geometrischen Prognosefehlers. Da dieses negative Phänomen ebenfalls auf der großen Datenmenge des Gesamtmodells auftritt, kann von störendem unsicherem Einfluss der exogenen Variablen ausgegangen werden. Das Modell besitzt kaum die Fähigkeit der Generalisierung auf neuen Datensätzen. Die Verwendung der Produktattribute als bedingende Einflussgrößen verändern die Güte aller Modelle nicht positiv. Darüber hinaus verbessern die makroökonomischen Indikatoren lediglich die Anpassung des Modells für die Trainingsmenge, führen jedoch auf den Testdatensätzen zu stark zufälligen Prognosen, was sich speziell bei der Modellierung des Bausegments zeigt.

Es ist daher festzustellen, dass weder die Verwendung von Produktattribute noch von makroökonomischen Indikatoren eine Verbesserung der Prognosequalität nach sich ziehen.

Die Tabelle 10.4 enthält den Vergleich zwischen segmentspezifischen Modellen und einem Gesamtmodell für alle einzelnen Fahrzeugbaumuster. Das Segment Sonderfahrzeuge ist in sämtlichen Betrachtungen vernachlässigt, da lediglich in dem zur Verfügung stehenden Zeitraum 13 Fahrzeuge diesem Segment zugeordnet werden konnten. Die Spalten „Train.“ und „Test“ beinhalten die Anzahl der Trainings- und Testdatensätze dieses Experiments. Der relative geometrische Prognosefehler des Gesamtmodells auf den Testdaten weist keine signifikant schlechtere Güte auf als der Gesamtfehler über alle segmentspezifischen Modelle, der sich ohne Berücksichtigung von Produktattributen und makroökonomischen Indikatoren zu zirka 38% ergibt. Die Modellierung des bedingten Lokationsparameters mit Hilfe der Produktattribute weist einen Gesamtfehler über alle Segmentmodelle von zirka $39,46\%$ auf.

10.5.1 Vergleich von linearer und nichtlinearer Modellierung

Der Inhalt des folgenden Abschnitts ist der Vergleich zwischen der linearen und nichtlinearen Modellierung des funktionalen Zusammenhangs zwischen den exogenen Einflussvariablen und den Verteilungsparametern. Das Ziel hierbei ist die Erörterung, ob sich der höhere Aufwand einer nichtlinearen Modellierung durch neuronale Netze gegenüber der parameterärmeren linearen Variante auszahlt.

Modell	Parameteranzahl	Fehler(Train.)	Fehler(Test)
linear	153	0,3305	0,3815
nichtlinear	317	0,3248	0,3995

Tabelle 10.5: Vergleich von linearen und nichtlinearen Prognosemodellen

Tabelle 10.5 prüft daher konkret eine Qualitätsverbesserung der Prognosemodelle durch eine nichtlineare Modellierung der Abhängigkeit zwischen den Lokations- und Skalierungsparametern der prognostizierten Wahrscheinlichkeitsverteilung und den exogenen Einflussvariablen. Die Approximation des nichtlinearen funktionalen Zusammenhangs basiert, wie in Kapitel 4.2 dargestellt, auf einem Multi-Layer Perzeptron.

Hierbei lag ein einziges Gesamtmodell über alle Fahrzeugbaumuster zugrunde und somit ist die Anzahl der Datensätze der Trainings- (4.248) und Testdaten (152) identisch mit denen aus Tabelle 10.4 und hier nicht explizit notiert.

Die Erhöhung der zu identifizierenden Schätzparameter des nichtlinearen Prognosemodells bewirkt, wie auch die Berücksichtigung einer größeren Anzahl von Inputgrößen, eine bessere Anpassung an die Trainingsdaten. Leider spiegelt sich diese Verbesserung nicht in den Testdaten wider. Daher kann i.Allg. nicht von einer Steigerung der Prognosequalität durch nichtlineare Modellierung ausgegangen werden.

Der relative Prognosefehler in Tabelle 10.5 zeigt keine Verbesserung der Prognosequalität durch eine nichtlineare Modellierung.

10.5.2 Varianzprognose und Aggregatmodelle

Die generell schlechte Prognosegüte der Absatzzahlen von Nutzfahrzeugen kann vor allem in den geringen Stückzahlen einzelner Fahrzeugbaumuster begründet liegen. Geringe Zahlen sind statistisch weniger signifikant und unterliegen daher größeren Schwankungen als deren Summen. Es wäre also möglich, dass aggregierte Verkaufszahlen genauer vorhergesagt werden können als die einzelner Fahrzeugbaumuster.

Zunächst sei jedoch beispielhaft anhand zweier Baumuster gezeigt, dass große Stückzahlen relativ genauer geschätzt werden als geringe. Tabelle 10.6

FBM	Monat	Soll	Prog.	Stdabw.	rel.Fehler	Stdabw/Prog	Diff.
1	1	23	22,67	0,46	$-1,4\%$	$2,03\%$	$0,63\%$
2		178	175,34	2,85	$-1,5\%$	$1,63\%$	$0,13\%$
1	2	23	21,9	0,99	$-4,8\%$	$4,52\%$	$0,28\%$
2		179	183,47	5,2	$2,5\%$	$2,83\%$	$0,33\%$
1	3	20	14,78	1,49	$26,10\%$	$10,10\%$	$16,00\%$
2		111	110,79	5,68	$0,2\%$	$5,13\%$	$4,93\%$
1	4	16	16,5	1,68	$3,1\%$	$10,20\%$	$7,10\%$
2		116	122,61	7,16	$5,7\%$	$5,80\%$	$0,10\%$
1	5	15	13,86	1,59	$-7,6\%$	$11,50\%$	$3,90\%$
2		106	104,24	6,29	$-1,7\%$	$6,00\%$	$4,30\%$
1	6	30	16,06	1,71	$-46,5\%$	$10,60\%$	$35,90\%$
2		143	118,47	6,28	$-17,2\%$	$5,30\%$	$11,90\%$

Tabelle 10.6: 6-monatige Varianzprognose am Beispiel zweier Baumuster

stellt den relativen Prognosefehler dem Quotienten aus Standardabweichung und Prognosewert gegenüber. Es zeigt sich, wie vermutet und statistisch begründet, die relativ zur Stückzahl kleiner geschätzte Standardabweichung des Baumusters mit größerer Nachfrage. Diese Untersuchungen basieren auf einem Gesamtmodell für alle Fahrzeugbaumuster.

Im Folgenden wird die zweite Hypothese durch die Betrachtung und Modellierung von summierten Absatzzahlen verfolgt.

Aggregat	Anzahl	Fehler(Train.)	Fehler(Test)
BR	3	0,0911	0,1020
Segment+BR	10	0,1098	0,1930
FBM	317	0,1912	0,3256
FBM+Motor	4.248	0,2396	0,3396

Tabelle 10.7: Prognose unterschiedlicher Produktgruppenaggregate

Tabelle 10.7 zeigt ausführlich die bessere Prognosequalität bei Prognosen von aggregierten Absatzzahlen. Tabelle 10.7 bestätigt die obigen Aussagen und Vermutungen und stellt die Fehlerwerte für unterschiedliche Produktgruppenebenen dar.

Trotz dieser sehr volatilen Zeitreihen sind die Vorteile des Prognosekonzepts der multivariaten bedingten Verteilungen klar zu erkennen. Auch wenn sich in diesem Beispiel durch eine nichtlineare Modellierung keine weitere Verbesserung ergab, so ist eine Steigerung der Prognosequalität durch das Verteilungskonzept im Vergleich zur Regressions- und Zeitreihenanalyse unbestreitbar.

In beiden präsentierten Anwendungen der Kapitel 9 und 10 kann von der Realisierung der Motivationsaspekte aus Kapitel 2.2 profitiert werden. Speziell die Attribut-Bedingtheit und die flexiblen Verteilungsklassen trugen zu einer überzeugenden Prognosequalität im Vergleich zu herkömmlichen Verfahren bei.

Teil IV

Zusammenfassung und Ausblick

Kapitel 11

Zusammenfassung

In dieser Abhandlung wird ein neues Konzept der probabilistischen Modellierung und Prognostik vorgestellt. Mit Hilfsmitteln aus der Stochastik, der numerischen Mathematik und der Neuroinformatik entstand ein Prognosesystem, das in der Lage ist, eine bedingte multivariate Wahrscheinlichkeitsverteilung für eine stochastische Zielgröße zu bestimmen. Diese Eigenschaft dient unter anderem der optimalen Entscheidungsfindung bei praktischen Problemstellungen.

Konkret ermöglicht hierbei die geeignete Kombination aus Minimierung der Cross-Entropie, funktionaler Approximation durch neuronale Netze und globaler numerischer Optimierung die Identifikation der zugrunde liegenden multivariaten Wahrscheinlichkeitsverteilung in Abhängigkeit von den erklärenden Inputvariablen.

Für die Entwicklung des Prognosesystems sind sukzessive die formulierten Anforderungen und Motivationsaspekte berücksichtigt und umgesetzt. Die motivierenden Ziele finden sich im Laufe der Arbeit immer wieder. Die Bedingtheit etwa wurde von Beginn an als grundlegender Konzeptbaustein identifiziert und konnte für alle hier betrachteten Wahrscheinlichkeitsverteilungen umgesetzt werden. Durch den Ansatz eines regressionsähnlichen Modells ist diese Eigenschaft automatisch gesichert. Die Einführung einer technischen Transformation gewährleistet schließlich die notwendige Abbildung auf die restringierten Verteilungsparameter.

Der attributbasierte Ansatz, wie in Abschnitt 2.2.1 motivierend beschrie-

ben, zeigt im Vergleich zu Methoden aus der herkömmlichen Zeitreihenanalyse in beiden vorgestellten Anwendungen klare Vorteile und macht eine sinnvolle Prognose auf der gegebenen Datengrundlage möglich.

Im vorliegenden Konzept sind als Regressionsfunktionen sowohl lineare Modelle als auch neuronale Netze zur nichtlinearen Approximation möglich. Beide Optionen sind in die Konzeption integriert. Zur Anpassung der funktionalen Approximatoren an die empirischen Daten wird ein globales Optimierungsverfahren verwendet, dessen Vorteile geegnüber herkömmlichen lokalen Algorithmen erläutert werden.

Das allgemeine Konzept, das auf der Transformation von kanonischen Verteilungen basiert, kann für eine Fülle von Verteilungsklassen Verwendung finden. Hierzu zählen Normalverteilungen, t-, stabile oder hyperbolische Verteilungen ebenso wie die approximative Gauß'sche Mixtur für nichtparametrische Verteilungsklassen. Das seit langer Zeit erkannte Defizit der Modellierungsansätze, die ausschließlich auf der Normalverteilung basieren, empirische Daten unzureichend zu approximieren, wird dadurch entscheidend abgeschwächt.

Das in Abschnitt 2.2.2 formulierte Ziel der flexiblen Verteilungsklassen ist dadurch erreicht. Diese Fähigkeit macht sich in praktischen Anwendungen durch eine enorme Verbesserung der Prognosequalität offensichtlich bezahlt.

Allerdings ist nicht ausschließlich die Qualitätsverbesserung ein eindeutiger Mehrwert dieses Systems. Aufgrund der Kenntnis der geschätzten bedingten Wahrscheinlichkeitsverteilung kann diese Prognosemethodik auf unterschiedliche Weise eine optimale Entscheidung unterstützen. Etwa durch zusätzliche Variabilitätsaussagen ist es möglich, die Qualität und Zuverlässigkeit des Prognosewerts zu beurteilen und zu bewerten. Dieses gesteckte Ziel aus Sektion 2.2.1 ist somit erreicht und liefert eine wertvolle Zusatzinformation zum reinen Prognosewert.

Falls des Weiteren eine definierte Verlustfunktion existiert, kann mit Hilfe der bedingten Verteilung eine in diesem Sinne optimale Entscheidung hergeleitet werden. Diese in der Praxis häufig auftretende Situation, die einleitend in Abschnitt 2.2.4 motiviert wurde, kann im vorliegenden Konzept profitierend umgesetzt werden. Bei der Bedarfsprognose von Ersatzteilen wird dies

verdeutlicht.

Das in dieser Abhandlung entwickelte Konzept löst im Speziellen die strenge Annahme von unabhängigen Zielvariablen auf und schätzt alle Komponenten der Strukturmatrix, die im Normalverteilungsfall direkt mit der Kovarianzmatrix korrespondiert. Diese Art der probabilistischen Modellierung erzeugt daher nicht ausschließlich einen Schätzer für die Zuverlässigkeit der vorhergesagten Werte, sondern liefert zusätzlich Aussagen über die Abhängigkeiten zwischen den prognostizierten Variablen.

Der konzeptionellen Darstellungen des Prognosesystems dieser Arbeit folgend, kann gezeigt werden, dass das präsentierte Prognosemodell eine Vielzahl von bekannten Prognosetechniken aus der Statistik und der Neuroinformatik umfasst und daher eine Erweiterung und Verallgemeinerung dieser Konzepte ist.

Im Anschluss an die Entwicklung und Abgrenzung des Verteilungskonzepts wird die Tauglichkeit auf synthetischen Datensätzen verifiziert. Datenmengen, in denen unterschiedliche empirische Verteilungen abgebildet sind, fanden Verwendung und dienten somit als empirische Soll-Verteilungen. Es ist gezeigt, dass dieses System in der Lage ist, alle hier verwendeten Verteilungsklassen eindeutig zu identifizieren.

Schließlich findet die identifizierte bedingte Wahrscheinlichkeitsverteilung auf unterschiedliche Weise praktische Verwendung. Einerseits dient die Verteilung der Berechnung einer optimalen Entscheidung unter asymmetrischen Kosten durch die Minimierung der gesamten Aufwände für die Ersatzteillogistik. Andererseits zeigt das Konzept der Verteilungsprognose klare Vorteile gegenüber klassischen Prognosemethoden bei schwierigen Aufgaben der Absatzprognose variantenreicher Nutzfahrzeuge.

Da diese empirischen Untersuchungen lediglich beispielhaft die Anwendbarkeit dieses Prognosekonzepts verdeutlichen und die Konzeption auf eine Vielzahl von Problemen aus der Prognostik anwendbar ist, konnte die in Abschnitt 2.2.5 geforderte universelle Anwendbarkeit ebenfalls erzielt werden.

Kapitel 12

Ausblick

Eine Stoßrichtung, um das hier entwickelte Prognosekonzept weiter zu entwickeln, wäre die Integration und Implementierung weiterer Klassen von Wahrscheinlichkeitsverteilungen. Zum Beispiel ist für die Modellierung von Lagerbeständen im Supply Chain Management - durch Kapazitätsrestriktionen der Lager - eine Verteilung von Nöten, die lediglich auf einem abgeschlossenen Intervall existiert. Es würden sich daher eventuell Verteilungsklassen, wie Beta- oder Gammaverteilungen anbieten.

Falls - etwa durch Anforderungen aus der Praxis - eine multivariate Verteilung benötigt wird, erfordert dies die Konzeption von Kopulas. Da keine multivariaten Beta- und Gammaverteilungen existieren, sind die univariaten Randverteilungen und die Kopula zur Abbildung der Abhängigkeitsstruktur zu schätzten. Die resultierende multivariate Verteilung besitzt sodann univariate Randverteilungen, die aus der Klasse der Beta- bzw. Gammaverteilungen stammen und deren Abhängigkeitsstruktur über die Kopula repräsentiert ist.

Eine weitere Möglichkeit der Weiterentwicklung dieses Konzepts ist die Betrachtung von unsicheren Einflüssen. Bislang wurde von deterministischen exogenen Größen als Inputvariablen ausgegangen. Falls jedoch unsichere Inputs auf die Zielvariablen wirken, könnte etwa mittels Mixturen eine weitere Verallgemeinerung erreicht werden.

Schließlich soll als letzte hier erwähnte Generalisierung des Prognosesystems der Bayes-Ansatz Erwähnung finden. Zur Umsetzung des Bayes-

Ansatzes ist die Festlegung oder Kenntnis der a priori Verteilung der Schätzparameter notwendig, was sich in der Realität als sehr schwierig gestaltet. Werden etwa kleine Werte für die Gewichte im neuronalen Netz mit höherer Wahrscheinlichkeit angenommen als große Werte, bedeutet dies eine Glättung der nichtlinearen Funktion. In Werken, wie etwa (Williams, 1999), (Bishop, 1995a) oder (Hamilton, 1991), wurde dies bereits ansatzweise beleuchtet.

Literaturverzeichnis

Abraham, B. and Ledolter, J. (1983). *Statistical Methods for Forecasting*. Wiley & Sons, Inc.

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical functions*. Dover Publications, New York.

Adams, G., Allen, P. G., and Morzuch, B. J. (1991). Probability distribution of short-term electricity peak load forecasts. *International Journal of Forecasting*, 7:283–297.

Adya, M. and Collopy, F. (1998). How effective are neural networks at forecasting and prediction? *Journal of Forecasting*, 17:481–495.

Aitchison, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistic Association*, 50:901–908.

Aitken, A. C. (1935). On least squares and linear combinations of observations. *Proceedings of the Royal Statistical Society*, 55:42–48.

Anders, U. (1996). *Statistische Neuronale Netzwerke*. Dissertation, Universität Karlsruhe.

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. Wiley, New York.

Anderssen, R. (1972). Global optimization. *University of Queensland Press*, pages 1–15.

Anderssen, R. and Bloomfield, P. (1975). Properties of the random search in global optimization. *Journal of Optimization Theory and Applications*, 16:383–398.

Anscombe, F. J. (1967). Topics in the investigation of linear relations fitted by the method of least squares (with discussion). *Journal of the Royal Statistical Society*, B 29:1–52.

Armstrong, J. S. (1989). Reflections on forecasting in the 1980's. *International Journal of Forecasting*, 5:467–468.

Ashton, W. D. (1971). Distribution for gaps in road traffic. *Journal of the Institution of Mathematical Applications*, 7:37–46.

Atkinson, A. C. (1982). The simulation of generalized inverse gaussian and hyperbolic random variables. *SIAM Journal of Scientific Statistical Computation*, 3(4):502–515.

Bagnold, R. A. and Barndorff-Nielsen, O. E. (1980). The pattern of natural size distribution. *Sedimentology*, 27:199–207.

Barndorff-Nielsen, O. E. (1977). Exponentially decreasing distributions for the logarithm of particle size. *Proceedings of Royal Society London A, 353*, pages 401–419.

Barndorff-Nielsen, O. E. (1978). Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of Statistics*, 5:151–157.

Barndorff-Nielsen, O. E. and Blæsild, P. (1981). Hyperbolic distributions and ramifications: Contributions to theory and application. *In C. Taillie, G. Patil, and B. Baldessari (Eds.), Statistical Distributions in Scientific Work*, 4:19–44. Dordrecht: Reidel.

Barndorff-Nielsen, O. E., Blæsild, P., Jensen, J. L., and Sørensen, M. (1985). The fascination of sand. *In A. C. Atkinson and S. E. Fienberg, editors, A Celebration of Statistics, Springer, New York*, pages 57–87.

Barndorff-Nielsen, O. E., Blæsild, P., Jensen, J. L., and Sørensen, M. (1989). Wind shear and hyperbolic distributions. *Meteorology*, 49:417–431.

Barndorff-Nielsen, O. E., Kent, J., and Sørensen, M. (1982). Normal variance-mean mixtures and z distributions. *International Statistical Review*, 50:145–159.

Bartlmae, K. and Rauscher, F. A. (2000). Measuring dax market risk: A neural network volatility mixture approach. DaimlerChrysler AG.

Bauer, H. (2002). *Wahrscheinlichkeitstheorie*, volume 5. Walter de Gruyter.

Bera, A. K. and Higgins, M. L. (1993). ARCH models: Properties, estimation and testing. *Jourrnal of Econometric Surveys*, 7(4):305–366.

Bishop, C. M. (1994). Mixture density networks. Technical Report NCRG/94/004, Neural Research Group Aston University, Birmingham B4 7ET, UK.

Bishop, C. M. (1995a). Bayesian methods for neural networks. Technical Report NCRG/95/009, Neural Research Group Aston University, Birmingham B4 7ET, UK.

Bishop, C. M. (1995b). *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

Bishop, C. M. (1996). Neural networks: A pattern recognition perspective. Technical Report NCRG/96/001, Neural Research Group Aston University, Birmingham B4 7ET, UK.

Bishop, C. M. and Legleye, C. (1995). Estimating conditional probability densities for periodic variables. Technical report, Neural Research Group Aston University, Birmingham B4 7ET, UK.

Bishop, C. M. and Nabney, I. T. (1996). Modelling conditional probability distributions for periodic variables. *Neural Computation*, 8:1123–1133.

Blæsild, P. and Jensen, J. L. (1981). Multivariate distributions of hyperbolic type. *In C. Taillie, G. Patil, and B. Baldessari (Eds.), Statistical distributions in scientific work*, 4:45–66. Dordrecht: Reidel.

Blattberg, R. C. and Goneges, N. J. (1974). A comparison of the stable and student distributions as statistical models for stock prices. *Journal of Business*, 47:244–280.

Blingham, N. H., Kiesel, R., and Schmidt, R. (2002). Semi-parametric models in finance: Economic applications.
www.mathematik.uni-ulm.de/finmath/.

Blischke, W. R. (1955). Estimating the parameters of mixtures of binomial distributions. *Journal of the American Statistic Association*, 59:510–528.

Boender, C. (1984). *The generalized multinomial distribution: A Bayesian analysis and applications*. PhD thesis, Erasmus Universiteit Rotterdam (Centrum voor Wiskunde en Informatice), Amsterdam.

Boender, C., Rinnooy Kan, A., Timmer, G., and Stougie, L. (1982). A stochastic method for global optimization. *Mathematical Programming*, 22(2):125–140.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327.

Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *The Review of Economics and Statistics*, 69:542–547.

Bollerslev, T., Chou, R. Y., and Kroner, K. F. (1992). ARCH modeling in finance: A review of the theory and empirical evidence. *Journal of Econometrics*, 52:5–59.

Bollerslev, T., Engle, R. F., and Nelson, D. B. (1994). ARCH models. *Handbook of Econometrics*, IV:2959–3083.

Box, G., Jenkins, G., and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*. Prentice Hall, Inc., 3^{rd} edition.

Bridle, J. S. (1990). *In F. Fogelman Soulié and J. Hérault (Eds.). Neurocomputing: Algorithms, Architecures and Applications*, chapter Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition, pages 227–236. Springer-Verlag, New York.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to Time Series and Forecasting*. Springer Verlag, 2^{nd} edition.

Brooks, S. (1958). A discussion of random methods for seeking maxima. *Operations Research*, 6:291–308.

Cauchy, A. (1853). Sur les résultats moyens d'observation de même nature, et sur les résultats les probables, 94-104; sur la probabilité des erreurs qui affectent des résultats moyens d'observation de même nature, 104-114. *Gauthier-Villars, Paris (1900)*. (Originals in C.R. Acad. Sci. Paris 37, 198 (Aug. 8, 1853) and 264 (Aug. 15, 1853)).

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). CRISP-DM 1.0. http://www.crisp-dm.org/.

Chatfield, C. (1989). *The Analysis of Time Series: An Introduction*. Chapman and Hall, 4^{th} edition.

Chatfield, C. (1996). Model uncertainty and forecast accuracy. *Journal of Forecasting*, 15:495–508.

Christopher, M. (1998). *Logistics and Supply Chain Management*. Financial Times; Prentice Hall International, 2^{nd} edition.

Cohen Jr., A. C. (1964). Estimation in mixtures of poisson and mixtures of exponential distributions. Technical report, NASA Technical Memorandum, NASA TM X-53245, George C. Marshall Space Flight Center, Huntsville, Alabama.

Cohen Jr., A. C. (1965). Estimation in mixtures of discrete distributions. *In Classical and Contagious Discrete Distributions (ed. G. P. Patil)*, pages 373–378.

Cornish, E. A. (1954). The multivariate t-distribution associated with a set of normal sample deviates. *Australian Journal of Physics*, 7:531–542.

Davis, D. J. (1952). An analysis of some failure data. *Journal of the American Statistic Association*, 47:113–150.

de-la Vega, M.-D. C., Pino-Majias, R., Pascual-Acosta, A., and Munoz-Garcia, J. (2002). Building neural network forecasting models from time series ARIMA models: A procedure and a comparative analysis. *Intelligent Data Analysis*, 6:53–65.

DeLurgio, S. A. (1998). *Forecasting Principles and Applications*. Irwin / McGraw-Hill, 1^{st} edition.

Dempster, A. P. (1971). An overview of multivariate data analysis. *Journal of Multivariate Analysis*, pages 316–346.

Dennis, J. and Schnabel, R. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, NJ.

Diebold, F., Gunther, T., and Tay, A. (1998). Evaluating density forecasts with applications in financial risk management. *International Economic Review*, 39:863–883.

Dillon, W. R. and Goldstein, M. (1984). *Multivariate Analysis: Methods and Applications*. John Wiley & Sons, Inc.

Drapper, N. R. and Smith, H. (1966). *Applied regression analysis*. Wiley, New York.

Dunnett, C. W. and Strobel, M. (1954). A bivariate generalization of student's t-distribution, with tables for certain special cases. *Biometrika*, 41:153–169.

Durett, R. (1999). *Essentials of stochastic processes*. Springer-Verlag New York, Inc.

Eberlein, E. (1999). Application of generalized hyperbolic lévy motions to finance. Freiburg Center for Data Analysis and Modelling, preprint no. 64, University of Freiburg.

Eberlein, E. and Keller, U. (1995). Hyperbolic distributions in finance. *Bernoulli: Official Journal of the Bernoulli Society of Mathematical Statistics and Probability*, pages 281–299.

Eberlein, E., Keller, U., and Prause, K. (1998). New insights into smile, mispricing and value at risk: the hyperbolic model. *Journal of Business*, 71:371–405.

Eberlein, E. and Prause, K. (1999). The generalized hyperbolic model: financial derivatives and risk measures. Freiburg Center for Data Analysis and Modelling, preprint no. 56, University of Freiburg.

Eberlein, E. and Prause, K. (2000). The generalized hyperbolic model: financial derivatives and risk measures. *Mathematical Finance*.

Eckey, H.-F., Kosfeld, R., and Rengers, M. (2002). *Multivariate Statistik*. Gabler Verlag.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (2001). *Modelling Extremal Events for Insurance and Finance*. Springer.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007.

Engle, R. F. and Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Econometric Reviews*, 5:1–87.

Engle, R. F. and Ng, V. K. (1991). Measuring and testing the impact of news on volatility. Technical report, Mimeo, Department of Econometrics, University of California, San Diego.

Fahrmeir, L. and Hamerle, A. (1984). *Multivariate statistische Verfahren*. Walter de Gruyter.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag New York, Inc.

Fama, E. (1965). The behavior of stock market prices. *Journal of Business*, 38:34–105.

Fang, K.-T., Kotz, S., and Ng, K.-W. (1990). *Symmetric Multivariate and Related Distributions*. Chapman and Hall.

Fang, K. T. and Xu, J. L. (1989). A class of multivariate distributions including the multivariate logistic. *Journal of Math. Research and Exposition*, 9:91–100.

Farnum, N. R. and Stanton, L. W. (1989). *Quantitative Forecasting Methods*. PWS-KENT Publishing Co.

Feller, W. (1966). *An Introduction to Probability Theory and its Applications*. Wiley.

Fletcher, R. (1987). *Practical Methods of Optimization*. Wiley & Sons, New York.

Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3):183–192.

Gallant, A. R. (1987). *Nonlinear Statistical Models*. Wiley, New York.

Geweke, J. (1986). Modelling the persistence of conditional variances: comment. *Econometric Reviews*, 5:57–61.

Gill, P. E., Murray, W., and Wright, M. H. (1981). *Practical Optimization*. Academic Press, London.

Gnedenko, B. V. and Kolmogorov, A. N. (1954). *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley.

Goodman, I. R. and Kotz, S. (1981). Hazard rate based on isoprobability contours. *Statistical Distributions in Scientific Work*, 5:289–309. (eds G.P. Patil et al.) Boston/London: D. Reidel.

Gosset, W. S. (1914). The elimination of spurious correlation due to position in time or space. *Biometrika*, 10:179–180.

Gourieriuox, C. and Monfort, A. (1992). Qualitative threshold ARCH models. *Journal of Econometrics*, 52:159–199.

Grabec, I. and Sachse, W. (1997). *Synergetics of Measurement, Prediction and Control*. Springer-Verlag, Berlin.

Granger, C. W. J. and Sin, C.-Y. (2000). Modelling the absolute returns of different stock indices: Exploring the forecastility of an alternative measure of risk. *Journal of Forecasting*, 19(4):277–298.

Gumbel, E. J. (1962). Bivariate logistic distribution. *Journal of American Statistic Association*, 56:335–349.

Hamilton, J. D. (1991). A quasi-bayesian approach to estimating parameters for mixtures of normal distributions. *Journal of Business and Economic Statistics*, 9(1):27–39.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hansen, B. (1994). Autoregressive conditional density estimation. *International Economic Review*, 35:705–730.

Hansen, B. E. (1992). Autoregressive conditional density estimation. Working Paper 332, Rochester Center for Economic Research.

Hartman, J. K. (1973). Some experiments in global optimization. *Naval Research Logistics Quarterly*, 20:569–576.

Hartung, J. (1999). *Statistik*, volume 12. R. Oldenbourg Verlag.

Hartung, J. and Elpelt, B. (1995). *Multivariate Statistik*, volume 5. R. Oldenbourg Verlag.

Hauser, M. A. and Kunst, R. M. (2001). Forecasting high-frequency financial data with the ARFIMA-ARCH model. *Journal of Forecasting*, 20:501–518.

Henrion, M. (1982). *The value of knowing how little yu know: the advantages of probabilistic treatment of uncertainty in policy analysis*. PhD thesis, Carnegie Mellon University, Pittsburgh.

Higgins, M. L. and Bera, A. K. (1992). A class of nonlinear ARCH models. *International Economic Review*, 33:137–158.

Hill, T., Marquez, L., O'Connor, M., and Remus, W. (1994). Artificial neural network models for forecasting and decision making. *International Journal of Forecasting*, 10(1):5–15.

Holden, K., Peel, D., and Thompson, J. (1990). *Economic forecasting: an introduction*. Cambridge University Press.

Holt, D. R. and Crow, E. L. (1973). Tables and graphs of the stable probability density functions. *Journal of Research of the National Bureau of Standards - B. Mathematical Sciences*, 778:143–198.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Network*, 2(5):359–366.

Hrycej, T. (1997). *Neurocontrol: towards an industrial control methodology*. John Wiley & Sons, Inc., New York.

Hrycej, T. and Stützle, E. A. (2001). A novelty filter based intelligent early warning system for quality control. In *Proc. 2001 IASTED Int. Conf. on Artificial Intelligence and Soft Computing*, Cancun, Mexico.

Hrycej, T. and Stützle, E. A. (2001). Konzept für die 15-Jahres-Prognose des Ersatzteilebedarfs. Arbeitspapier, DaimlerChryser AG, Forschung & Technologie, FT3/AD.

Hüttner, M. (1986). *Prognoseverfahren und ihre Anwendung*. Walter de Gruyter, Berlin/New York.

Ivanov, V. V. (1972). On optimal algorithms of minimization in the class of functions with Lipschitz condition. *Information Processing 2*, pages 1324–1327.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3:79–87.

Jain, C. (1988). *Understanding Business Forecasting*. Graceway Publishing Company, Inc.

Jensen, D. R. (1985). Multivariate distributions. *Encyclopedia of Statistical Sciences*, 6:pp. 43–55. (eds. S. Kotz, N.L. Johnson and C.B. Read), Wiley.

Johnson, M. E. (1987). *Multivariate Statistical Simulation*. John Wiley & Sons, Inc., New York.

Johnson, N. L. and Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. John Wiley & Sons, Inc.

Johnson, R. A. and Wichern, D. W. (1999). *Applied Multivariate Statistical Analysis*, volume 4. Prentice Hall.

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter. *Sankhya, A*, 32:419–430.

Keller, U. (1997). *Realistic modelling of financial derivatives*. Dissertation, Universität Freiburg.

Kendall, M., Stuart, A., and Ord, J. (1983). *The Advanced Theory of Statistics*, volume 3. Griffin, London, 4^{th} edition.

Khaikine, M. and Holthausen, K. (2000). A general probability estimation approach for neural computation. *Neural Computation*, 12:433–450.

Kirchgraber, U. and Ruf, U. (1997). Von Modellen und Prognosen oder warum manche Mathematiker nicht ungern über das Wetter reden. *Mathematikunterricht*, 43(5):30–36.

Kotz, S. (1975). Multivariate distributions at a cross-road. *Statistical Distributions in Scientific Work*, 1. (eds G.P. Patil et al.) Boston/London: D. Reidel.

Kotz, S., Johnson, N. L., Read, C. B., and Kotz, S. I. (1985). *Encyclopedia of Statistical Sciences: Multivariate Analysis to Plackett and Burman Designs*, volume 6. Wiley, New York.

Koutras, M. (1986). On the generalized noncentral chi-squared distribution induced by an elliptical gamma law. *Biometrika*, 73:528–532.

Kruse, R., Nauck, D., and Klawonn, F. (1996). *Neuronale Netze und Fuzzy-Systeme*. Vieweg-Verlag.

Krzanowski, W. J. and Marriott, F. H. C. (1994). *Multivariate analysis, part 1: distributions, ordination and inference*. Arnold, London.

Krzanowski, W. J. and Marriott, F. H. C. (1995). *Multivariate analysis, part 2: classification, covariance structures and repeated measures*. Arnold, London.

Kullback, S. (1959). *Information theory and statistics*. Wiley & Sons, New York.

Laurent, A. G. (1955). Distribution d'échantillon et de charactéristiques d'échantillons quand la population de référence est laplace-gaussienne de paramètres inconnus. *Journal de la Société de Statistique de Paris*, 96:262–296.

Levenbach, H. and Cleary, J. P. (1981). *The Beginning Forecaster: The Forecasting Process Through Data Analysis*. Lifetime Learning Publications.

Levenbach, H. and Cleary, J. P. (1984). *The Modern Forecaster: The Forecasting Process Through Data Analysis*. Lifetime Learning Publications.

Lévy, P. (1923). Théorie des erreurs. La loi de Gauss et les lois exceptionelles. *Bulletin Société Mathématique France*, 52:49–85.

Lévy, P. (1925). *Calcul des probabilités*. Gauthier Villars.

Luenberger, D. G. (1984). *Linear and Nonlinear Programming*. Addison-Wesley, 2^{nd} edition.

Lukšan, L. (1990). Computational experience with improved variable metric methods for unconstrained minimization. *Kybernetika*, 26(5):415–431.

Lukšan, L., Šiška, M., and Tuma, M. (1992). Interactive system for universal functional optimization (ufo). Technical Report 529, ICIS of CAS, Prague.

Makridakis, S., Anderson, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., and Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 1:111–153.

Malik, H. J. and Abraham, B. (1973). Multivariate logistic distribution. *Ann. Statistics*, 1:588–590.

Mandelbrot, B. B. (1963). The variation of certain speculative prices. *Journal of Business*, 36:519–530.

Mandelbrot, B. B. and Taylor, H. (1967). On the distribution of stock prices differences. *Operations Research*, 15:1057–1062.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, Inc.

Martin, C. and Witt, S. F. (1989). Forecasting tourism demand: A comparison of the accuracy of several quantitative methods. *International Journal of Forecasting*, 5:7–19.

Maunz, A. (2003). Untersuchung verschiedener verfahren für die Mittel- und Langfristplanung von Ersatzteilbedarfen unter Berücksichtigung ihrer Einsetzbarkeit und Wirtschaftlichkeit. Diplomarbeit, Universität Karlsruhe.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall, New York.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.

Milhoj, A. (1987). A multiplicative parameterization of ARCH models. Research Report 101, Institute of Statistics, University of Copenhagen.

Mittelhammer, R. C., Judge, G. G., and Miller, D. J. (2000). *Econometric Foundations*. Cambridge, University Press.

Mittnik, S. and Paolella, M. S. (2000). Conditional density and value-at-risk prediction of asian currency exchange rates. *Journal of Forecasting*, 19(4):313–333.

Morgan, M. G. and Henrion, M. (1992). *Uncertainty*. Cambridge University Press.

Moshiri, S. and Cameron, N. (2000). Neural networks versus econometric models in forecasting inflation. *Journal of Forecasting*, 19:201–217.

Nabney, I. T., Bishop, C. M., and Legleye, C. (1995). Modelling conditional probability distributions for periodic variables. Technical Report NCRG/95/010, Neural Computing Research Group, Aston University, UK.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistic Society*, A 135:370–384.

Nelson, D. B. (1991). Conditional heteroscedasticity in asset returns: a new approach. *Econometrica*, 59:347–370.

Neuneier, R. (1995). Optimal strategies with density estimating neural networks. *ICANN '95*.

Neuneier, R. (1998). *Optimale Investitionsentscheidungen mit neuronalen Netzen*. Dissertation, Universität Kaiserslautern.

Neuneier, R., Hergert, F., Finnhoff, W., and Ormoneit, D. (1994). Estimation of conditional densities: A comparison of neural network approaches. *ICANN94-Proceedings of the International Conference on Artificial Neural Networks*, pages 689–692.

Newbold, P. and Bos, T. (1993). *Introductory Business and Economic Forecasting*. South-Western Publishing Co., 2^{nd} edition.

Nix, D. and Weigend, A. (1994). Estimating the mean and variance of the target probability distribution. *World Congress of Neural Networks, Lawrence Erlbaum Associates*.

Nowlan, S. J. (1991). *Soft Competitive Adaption: Neural Network Learning Algorithms based on Fitting Statistical Mixtures*. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh.

O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics, Volume II B: Bayesian Inference*. Arnold, London.

Ormoneit, D. (1998). *Probability Estimating Neural Networks*. Shaker Verlag Aachen.

Ormoneit, D. and Neuneier, R. (1996). Experiments in predicting the german stock index dax with density estimating neural networks. *Conference on Computational Intelligence for Financial Engineering, New York*, pages 66–71.

Pagan, A. R. (1996). The econometrics of financial markets. *Journal of Empirical Finance*, 3:15–102.

Pagan, A. R. and Schwert, G. W. (1990). Alternative models for conditional stock volatility. *Journal of Econometrics*, 45:267–290.

Pardalos, P. M. and Romeijn, H. E. (2002). *Handbook of global optimization*. Kluwer Academic Publications.

Parzen, E. (1962a). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 35:1065–1076.

Parzen, E. (1962b). *Stochastic Processes*. Holden-Day, San Francisco.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society*, 185:71–110.

Petrovic, D., Roy, R., and Petrovic, R. (1998). Modelling and simulation of a supply chain in an uncertain environment. *European Journal of Operational Research*, 109:299–309.

Polak, E. (1971). *Computational Methods in Optimization*. Academic Press, New York.

Prause, K. (1997). Modelling financial data using generalized hyperbolic distributions. Freiburg Center for Data Analysis and Modelling, preprint no. 48, University of Freiburg.

Prause, K. (1999). *The generalized hyperbolic model: Estimation, financial derivatives, and risk measures*. PhD thesis, University of Freiburg. http://www.freidok.uni-freiburg.de/volltexte/15.

Press, S. J. (1968). A compound events model for security prices. *Journal of Business*, 40:317–335.

Press, S. J. (1972). *Applied multivariate analysis*. Holt, Reinhart and Winston, Inc.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C*, volume Cambridge. Cambridge University Press.

Qi, M. (2002). Forecasting real time financial series. In *Proc. 2001 IEEE Int. Joint Conference on Neural Networks*, Honolulu, Hawaii.

Quade, E. S. (1975). *Analysis for public decisions*. Elsevier, New York.

Rachev, S. and Mittnik, S. (1997). Econometric modeling in the presence of heavy-tailed innovations: a survey of some recent advances. *Stochastic Models*, 13:841–866.

Rachev, S. and Mittnik, S. (2000). *Stable Paretian Models in Finance*. Wiley & Sons, Inc.

Redner, R. and Walker, H. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26.

Rider, P. R. (1962). Estimating the parameters of mixed poisson, binomial and weibull distributions by the method of moments. *Bulletin de l'Institut International de Statistique*, 39.

Rinnooy Kan, A. and Timmer, G. (1987). Stochastic global optimization methods, part i: Clustering methods, part ii: Multi-level methods. *Mathematical Programming*, 39(1):26–78.

Roll, R. (1968). *The efficient market model applied to U.S. Treasury bill rates*. PhD thesis, Graduate School of Business, University of Chicago.

Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington, DC.

Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning internal representations by error propagation. In Rumelhart, D. and McClelland, J., editors, *Parallel Distributed Processing*, volume 1, chapter 8. MIT Press, Cambridge MA.

Ružička, P. and Kober, R. (1992). Neural network based controllers. Technical report, Forschungsinstitut für anwendungsorientierte Wissensverarbeitung, Ulm, Germany.

Rydberg, T. H. (1996). The normal inverse gaussian lévy process: Simulation and approximation. Research Report 344, University of Aarhus, Departement of Theoretical Statistics.

Samorodnitsky, G. and Taqqu, M. S. (1994). *Stable Non-Gaussian Random Processes*. Chapman & Hall, London.

Scherer, A. (1997). *Neuronale Netze: Grundlagen und Anwendungen*. Vieweg (Computational intelligence).

Schittenkopf, C. and Dorffner, G. (2000). Risk-neutral density extraction from option prices: Improved pricing with mixture density networks. Technical Report 47, Adaptive Information Systems and Modelling in Economics and Management Science, Vienna University of Economics and Business Administration.

Schittenkopf, C., Dorffner, G., and Dockner, E. J. (1998). Volatility prediction with mixture density networks. *ICANN98-Proceeding of the International Conference on Artificial Neural Networks*, pages 929–934.

Schittenkopf, C., Dorffner, G., and Dockner, E. J. (1999). Forecasting time-dependent conditional densities: A semi-nonparametric neural network approach. Technical Report 36, Adaptive Information Systems and Modelling in Economics and Management Science, Vienna University of Economics and Business Administration.

Schittenkopf, C., Dorffner, G., and Dockner, E. J. (2000). Forecasting time-dependent conditional densities: A neural network approach. *Journal of Forecasting*, 19(4):355–374.

Schmidt, R. (2001). Tail dependences for elliptically contoured distributions. Abteilung Zahlentheorie und Wahrscheinlichkeitstheorie, University of Ulm.

Schmidt, R., Hrycej, T., and Stützle, E. A. (2003). Integrative multivariate distribution models with generalized hyperbolic margins. Submitted in Journal of Computational Statistics, Springer-Verlag.

Schwarze, J. (1980). *Angewandte Prognoseverfahren*. Herne-Verlag, Berlin. Arbeitsgruppe Prognoseverfaren der deutschen Arbeitsgruppe für Operation Research, Neue Wirtschaftsbriefe.

Searle, S. R. (1971). *Linear models*. Wiley, New York.

Seber, G. A. F. (1977). *Linear regression analysis*. Wiley, New York.

Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear Regression*. Wiley, New York.

Sentana, E. (1991). Quadratic ARCH models: a potential re-interpretation of ARCH models. Technical report, Mimeo, Departement of Economics and Financial Markets Group, London, School of Economics.

Shanno, D. F. (1978). Conjugate gradient methods with inexact searches. *Mathematics of Operations Research*, 3(3):244–256.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423 and 623–656.

Stock, J. H. and Watson, M. W. (1998). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. *NBER Working Paper Series - National Bureau of Economic Research*, 6607.

Stuart, A., Ord, K., and Arnold, S. (1999). *Kendall's Advanced Theory of Statistics, Volume 2A: Classical Inference and the Linear Model*. Arnold, London, 6^{th} edition.

Stützle, E. A. (1999). Systemidentifikation mit Hilfe von neuronalen Netzen und der Multi-Level Single-Linkage Methode. Diplomarbeit, Universität Trier.

Stützle, E. A. and Hrycej, T. (2001). Forecasting of conditional distributions - an application to the spare parts demand forecast. In *Proc. 2001 IASTED Int. Conf. on Artificial Intelligence and Soft Computing*, Cancun, Mexico.

Stützle, E. A. and Hrycej, T. (2002b). Estimating multivariate conditional distributions - an application to the truck demand forecast. In *Proc. OR2002*, Klagenfurt, Austria, Springer Verlag.

Stützle, E. A. and Hrycej, T. (2002a). Estimating multivariate conditional distributions via neural networks and global otimization. In *Proc. 2002 IEEE Int. Joint Conference on Neural Networks*, Honolulu, Hawaii.

Stützle, E. A. and Hrycej, T. (2002c). Modelling future demand by estimating the multivariate conditional distribution via the maximum likelihood

principle and neural networks. In *Proc. 2002 IASTED Int. Conf. on Modelling, Identification and Control*, Innsbruck, Austria.

Stützle, E. A. and Hrycej, T. (2003). Numerical method for estimating multivariate conditional distributions. Submitted in Journal of Computational Statistics, Springer-Verlag.

Tay, A. S. and Wallis, K. F. (2000). Density forecasting: A survey. *Journal of Forecasting*, 19(4):234–254.

Taylor, S. J. (1986). *Modelling Financial Time Series*. John Wiley, Chichester, UK.

Teichmoeller, J. A. (1971). A note on the distribution of stock price changes. *Journal of the American Statistic Association*, 66:282–284.

Timmermann, A. (2000). Editorial: Density forecasting in economics and finance. *Journal of Forecasting*, 19(4):231–234.

Titterington, D. M., Smith, A. F. M., and Markov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester: John Wiley.

Tsay, R. S. (1987). Conditional heteroscedastic time series models. *Journal of the American Statistical Association*, 82:590–604.

Tsay, R. S. (2001). *Analysis of Financial Time Series*. John Wiley & Sons Inc., 1^{st} edition.

Vlassis, N. A. and Kröse, B. (1999). Mixture conditional density estimation with the em algorithm. In *Proc. ICANN'99, 9th International Conference on Artificial Neural Networks*, Edinburgh, Scotland.

Vlassis, N. A., Likas, A., and Kröse, B. (2000). A multivariate kurtosis-based approach to gaussian mixture modelling. Technical report, Intelligent Autonomous Systems, technical series, nr. IAS-UVA-00-04, University of Amsterdam.

Vlassis, N. A., Papakonstantinou, G., and Tsanakas, P. (1999). Mixture density estimation based on maximum likelihood and sequential test statistics. *Neural Processing Letters*.

Weigend, A. S. and Nix, D. A. (1994). Predictions with confidence intervals (local error bars). In *Proc. of the International Conference on Neural Information Processes*, pages 847–852, Seoul, Korea.

Weigend, A. S. and Shi, S. (2000). Predicting daily probability distributions of s&p500 returns. *Journal of Forecasting*, 19(4):375–392.

White, H. (1989). Learning in artificial neural networks: A statistical perspective. *Neural Computation*, pages 425–464.

White, H. (1991). Nonparametric estimation of conditional quantiles using neural networks. In *Proc. of the Twenty-Second Symposium on the Interface*, pages 190–199, New York: Springer-Verlag.

White, H. (1992). Parametrical statistical estimation with artificial neural networks. Technical report, University of California, San Diego.

White, H. (1994). Parametrical statistical estimation with artificial neural networks. *In V. Cherkassky, J. Friedman, and H. Wechsler, editors, From Statistics to Neural Networks: Theory and Pattern Recognition Applications, Springer-Verlag*, pages 127–146.

Williams, P. M. (1996). Using neuronal networks to model conditional multivariate densities. *Neural Computation*, 8:843–854.

Williams, P. M. (1999). Matrix logarithm parametrizations for neural network covariance models. *Neural Networks*, 12:299–308.

Wirth, R. and Hipp, J. (2000). CRISP-DM: Towards a standard process modell for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, Manchester, UK.

Zakoian, J.-M. (1990). Threshold heteroscedastic model. Technical report, Mimeo, INSEE, Paris.