Modellierung nichtlinearer dynamischer Systeme

Diplomarbeit

vorgelegt von

Jörg D. Wichard

aus

Ziegenhain

angefertigt im

Dritten Physikalischen Institut
der Georg–August–Universität zu Göttingen

1997

Inhaltsverzeichnis

Kapitel 1

Einleitung

> *"Dann aber muß der Zufall doch etwas anderes sein als ein Name für unsere Unwissenheit, dann müssen wir die Erscheinungen, deren Ursache wir nicht kennen, in verschiedene Klassen einteilen; wir müssen die zufälligen Erscheinungen, über welche uns die Wahrscheinlichkeitsrechnung vorläufig unterrichtet, unterscheiden von denjenigen, die nicht zufällig sind, und über die wir nichts aussagen können, ehe wir nicht die für sie geltenden Gesetze bestimmt haben."*
>
> *Henri Poincaré, Wissenschaft und Methode (1908)*

Dieser Ausspruch H. Poincarés gelangte über ein halbes Jahrhundert später zu neuer Beachtung, als man (wieder) entdeckte, daß physikalische Systeme mit wenigen Freiheitsgraden, die nichtlinearen Gleichungen genügen, bereits irreguläres, unvorhersagbares Verhalten zeigen können. Derartige Phänomene werden unter dem Begriff *Chaos* zusammengefaßt, um sie gegen stochastische Systeme mit einer Vielzahl von Freiheitsgraden abzugrenzen [31].
Durch den zunehmenden Einsatz leistungsfähiger Rechenanlagen wurde es möglich, anstelle von linearisierten Systemen die originalen, nichtlinearen Systemgleichungen zu untersuchen, und neue Ansätze zur Beschreibung chaotischer Systeme zu entwickeln. Trotzdem ist die numerische Behandlung nichtlinearer Gleichungssysteme nach wie vor sehr schwierig und aufwendig, deshalb versucht man, die nichtlinearen Probleme umzuformulieren, um sich des weitentwickelten Apparates der linearen Approximationstheorie

bedienen zu können. Die in dieser Arbeit vorgestellten „pseudolinearen Modelle“ seien hier als Beispiel genannt.

Oft ist es unmöglich, alle Systemvariablen gleichzeitig zu messen, und in vielen Fällen liegt nur die Messung einer einzigen Variablen vor. Dennoch kann man unter bestimmten Voraussetzungen Aussagen über die Dynamik im Phasenraum des Systems machen, obwohl nur die Messung einer Systemvariablen vorliegt. Dies ist eine Folgerung der Einbettungstheoreme von Takens [55] (1981) und von Sauer et al. [48] (1991), die zu den fundamentalen Ergebnissen der Chaosphysik zählen. Diese Einbettungstheoreme erlauben es, ein diffeomorphes Abbild des Phasenraumes zu konstruieren, das zur Modellierung der Systemdynamik herangezogen werden kann. Dieser Ansatz liegt allen in dieser Arbeit untersuchten Methoden der Modellbildung zugrunde, dabei wird unter einem Modell eine mathematische Beschreibung des Systemverhaltens verstanden, d.h. eine Abbildung des rekonstruierten Phasenraums auf die gemessene Zeitreihe des Systems.

Eine wesentliche Eigenschaft chaotischer Systeme ist die sensitive Abhängigkeit der Dynamik von den Anfangsbedingungen. Da es einerseits unmöglich ist, die Anfangsbedingungen eines Systems beliebig genau zu bestimmen, und andererseits die Fehler in den Anfangsbedingungen exponentiell wachsen, ist es bereits nach kurzer Zeit nicht mehr möglich, den aktuellen Systemzustand vorherzusagen.
Dieser Effekt hat natürlich Auswirkungen auf die Modellierung chaotischer Systeme, denn ein Modell kann sich deshalb nicht mehr durch die Vorhersage des Systemverhaltens über lange Zeiträume bewähren[1]. In der nichtlinearen Physik müssen deshalb andere Verfahren zur Validierung eines Modells herangezogen werden. Im Rahmen dieser Arbeit werden zwei dieser Verfahren vorgestellt. Dabei wird sich herausstellen, daß die Einsatzfähigkeit dieser Verfahren entscheidend von der Wahl der Basisfunktionen abhängt, die zur Modellbildung herangezogen werden.
Das folgende Kapitel behandelt zunächst die Grundlagen der nichtlinearen Dynamik im Hinblick auf die Modellierung nichtlinearer Systeme.
Im Anschluß daran werden wir verschiedene Ansätze der nichtlinearen Modellbildung vorstellen und die in dieser Arbeit untersuchten Verfahren eingehend beschreiben.
Die Darstellung der Ergebnisse gliedert sich in zwei Teile. Zunächst sollen die vorgestellten Modellierungsverfahren im Hinblick auf die benötigten Rechenzeiten miteinander verglichen werden. Im Anschluß daran werden zwei Möglichkeiten zur Selektion der wesentlichen Modellparameter vorgestellt und miteinander verglichen. Dabei handelt es sich um das aus der Stati-

[1] So galt z.B. die Vorhersage der Bahn des Halleyschen Kometen als Verifizierung der Newtonschen Himmelsmechanik [3].

stik bekannte Verfahren der *Cross-Validation* und dem informationstheoretischen Ansatz der *Minimum-Description-Length*. Ein zentrales Ergebnis dieser Arbeit ist es, eine hinreichende Bedingung angeben zu können, unter der das letztgenannte Verfahren ein eindeutiges Ergebnis liefert.
Das abschließende Kapitel faßt dann die Resultate dieser Arbeit zusammen.

Kapitel 2

Theoretische Grundlagen

2.1 Der Begriff des Modells

Ein Modell ist ein Objekt, das auf Grundlage einer Struktur-, Funktions- oder Verhaltensanalogie zu einem entsprechenden Original von einem Beobachter eingesetzt und genutzt wird, um eine bestimmte Aufgabe lösen zu können, deren Durchführung mittels direkter Operationen am Original zunächst überhaupt nicht möglich bzw. unter gegebenen Bedingungen zu aufwendig ist [29].
Jedes Modell dient zur Erfassung entweder der Struktur, der Funktion oder des Verhaltens des Originals.
Das Modell muß sich vom Original qualitativ in jenen Merkmalen unterscheiden, die eine direkte Bearbeitung des Originals verhindern (z.B. leichtere Beschreibung durch einfache Modellgleichungen), zugleich soll das Modell dem Original in jenen Eigenschaften ähnlich sein, die speziell Gegenstand seiner Nutzung sind und dem Zweck der Modellierung entsprechen.
Eine wesentliche Aufgabe jeder Modellbildung besteht darin, diese beiden, sich in gewissem Sinne widersprechenden Bedingungen gegeneinander abzuwägen. Zur Lösung dieses Problems wird in dieser Arbeit die aus der Informationstheorie stammende Methode der Minimum-Description-Length (MDL) herangezogen.

Der rechnerische Aufwand beim Erstellen und bei der späteren Nutzung des Modells sinkt mit zunehmender Vereinfachung desselben, andererseits geht die Aussagekraft des Modells, die Übertragbarkeit der gewonnenen Informationen, mit wachsender Vereinfachung und Verkleinerung verloren.
Da das Modell sowohl vom Original verschiedene als auch gleiche Eigenschaften hat, müssen die am Modell gewonnenen Erkenntnisse wieder am Original überprüft werden. Ein gängiges Verfahren besteht darin, einen Teil der zur Verfügung stehenden Daten von der Modellbildung auszuschließen,

um an diesem „Testdatensatz" z.B. die Kurzzeitvorhersagen des Modells zu überprüfen.

2.2 Dynamische Systeme

Unter einem dynamischen System wird im allgemeinen ein deterministisches System in seiner zeitlichen Entwicklung verstanden.

Im weiteren Verlauf stehen Systeme im Vordergrund, bei welchen jeder Zustand eindeutig und vollständig durch einen Punkt $\vec{x} \in \mathbb{R}^d$ in einem endlichdimensionalen reellen Vektorraum beschrieben wird. Die Zeitentwicklung des Systems wird im kontinuierlichen Fall gegeben durch einen Satz gewöhnlicher Differentialgleichungen

$$\dot{\vec{x}} = F(\vec{x}), \tag{2.1}$$

dabei soll in einem Gebiet $G \subset \mathbb{R}^d$ gelten

$$F \in C^1(G),$$

wodurch für alle $\vec{x} \in G$ die Existenz und Eindeutigkeit einer Lösung von (2.1) gesichert ist [23].

Die zeitliche Entwicklung eines Punktes $\vec{x}_0(t)$ nennt man die *Trajektorie* oder Bahnkurve von $\vec{x}_0$. Die Gesamtheit aller Trajektorien nennt man das Phasenbild des Systems.

Im zeitdiskreten Fall wird ein dynamisches System beschrieben durch Differenzengleichungen der Form

$$\vec{x}_n = f(\vec{x}_{n-1}), \tag{2.2}$$

dabei soll $f \in C^0(G)$ gelten.

Beide Fälle definieren einen *Fluß* im Phasenraum, der im allgemeinen gegeben wird durch eine stetig differenzierbare Abbildung

$$\begin{aligned} \varphi : \mathbb{R}^d \times \mathbb{K} &\rightarrow \mathbb{R}^d \\ (\vec{x}, t) &\mapsto \varphi(\vec{x}, t) \end{aligned} \tag{2.3}$$

mit den Eigenschaften

$$\begin{aligned} \varphi(t = 0, \vec{x}) &= \vec{x} \\ \varphi(s, \varphi(t, \vec{x})) &= \varphi(s + t, \vec{x}) \qquad \forall s, t \in \mathbb{K} \;\; \forall \vec{\mathrm{x}} \in \mathbb{R}^{\mathrm{d}}. \end{aligned} \tag{2.4}$$

Für kontinuierliche dynamische Systeme ist $\mathbb{K} = \mathbb{R}$, für diskrete Systeme gilt $\mathbb{K} = \mathbb{Z}$, und der Fluß ist durch die Abbildung 2.2 selbst gegeben.
Den Zusammenhang zwischen Flüssen und Differentialgleichungen erhält man, indem man $\varphi(t, \vec{x}_0)$ als Lösung von Gleichung (2.1) interpretiert, die zur Zeit $t = 0$ durch den Punkt $\vec{x}_0$ geht. Auf diese Weise korrespondiert jeder Fluß mit einer Differentialgleichung, die Umkehrung gilt im allgemeinen nicht [5].

Ein System ist *dissipativ*, wenn sich das d-dimensionale Volumen eines gegebenen Phasenraumgebietes unter Einwirkung des Flusses im zeitlichen Mittel kontrahiert. Als Folgerung aus dem Liouvilleschen Satz der klassischen Mechanik ergibt sich somit

$$\mathrm{div} F(\vec{x}) \quad < \quad 0$$

als Bedingung für Dissipation [1, 2].

Von besonderem Interesse sind die Gebiete des Phasenraumes, die unter der Wirkung des Flusses invariant bleiben. Auf ihnen spielt sich das Langzeitverhalten eines dissipativen Systems ab, nachdem die Transienten abgeklungen sind [32].

Im folgenden seien U und V zwei offene Mengen des Phasenraums. Eine Menge $A \subset \mathbb{R}^d$ nennt man *Attraktor*[1] mit der *fundamentalen Nachbarschaft* U, wenn folgende Bedingungen erfüllt sind:

- A ist eine kompakte Menge

- A ist invariant bzgl. des Flusses

$$\forall t \in \mathbb{R} \; : \; \varphi(A, t) = A$$

- A ist eine *anziehende Menge*

$$\forall V \supset A \;\; \exists t \in \mathbb{R} \; : \; \varphi(U, t) \subset V$$

- Die Menge A läßt sich nicht in zwei nichtleere invariante Mengen zerlegen

Das *Bassin* des Attraktors A wird durch alle Punkte des Phasenraumes definiert, für die A eine anziehende Menge ist [46].

[1] Für diesen fundamentalen Begriff gibt es leider keine allgemein anerkannte Definition, in diesem Fall ist die von D. Ruelle vorgeschlagene Version verwendet worden [46].

2.3 Das invariante Maß

Ein wichtiges Hilfsmittel bei der Behandlung chaotischer Systeme ist das invariante natürliche Maß. Wenn langfristige Vorhersagen des Systemverhaltens nicht mehr möglich sind, kann man mit Hilfe des invarianten Maßes immerhin statistische Aussagen über den möglichen Zustand des Systems machen. Allerdings kann das invariante Maß selten explizit angegeben werden, und man ist auf Näherungsverfahren angewiesen [19, 38].
Wenn nach dem Abklingen der Einschwingvorgänge die Bewegung des Systems im Phasenraum auf dem Attraktor A verläuft, dann gibt es ein Wahrscheinlichkeitsmaß μ auf dem d-dimensionalen Borelschen Meßraum $(\mathbb{R}^d, \mathcal{B}^d)$ mit den Eigenschaften

$$\begin{aligned} \mu(A) &= 1 && (2.5)\\ \forall B \subseteq A;\ \forall t > 0\ : \mu(B) &= \mu(\varphi(B, -t)). && (2.6) \end{aligned}$$

Ein solches Maß heißt *φ-invariantes Maß*, und die Existenz folgt aus der Kompaktheit des Trägers und der Stetigkeit des Flusses [59].
Es kann für ein System mehrere invariante Maße geben, in der Physik ist es deshalb üblich, das von Kolmogorov vorgeschlagene *natürliche invariante Maß* zu betrachten. Dieses ist das letzte invariante Maß, das übrig bleibt, wenn das System gestört wird [16].
Ein invariantes Maß heißt *ergodisch*, wenn sich der Phasenraum des Systems nicht in zwei disjunkte, meßbare und jeweils flußinvariante Teilmengen mit positivem Maß zerlegen läßt [4]. Systeme mit einem ergodischen Maß werden *ergodische Systeme* genannt. In ergodischen Systemen kann man das Zeitmittel einer Funktion durch das Phasenraummittel ersetzen.
Sei $\mathrm{g} : \mathbb{R}^\mathrm{n} \to \mathbb{R}$ eine μ-meßbare Funktion. Dann wird das *Zeitmittel* von g definiert durch

$$\langle \mathrm{g} \rangle_\mathrm{t} = \lim_{\mathrm{T}\to\infty} \frac{1}{\mathrm{T}} \int_0^\mathrm{T} \mathrm{g}(\varphi(\vec{\mathrm{x}}_0, \mathrm{t}))\, \mathrm{dt} \tag{2.7}$$

und das *Phasenraummittel* durch

$$\langle \mathrm{g} \rangle_\mu = \int_\mathrm{A} \mathrm{g}(\vec{\mathrm{x}})\, \mathrm{d}\mu(\vec{\mathrm{x}}). \tag{2.8}$$

In einem ergodischen System gilt dann für fast alle Startwerte $\vec{x}_0 \in A$ (d.h. bis auf eine Nullmenge bzgl. μ)

$$\langle \mathrm{g} \rangle_\mathrm{t} = \langle \mathrm{g} \rangle_\mu. \tag{2.9}$$

Für einige diskrete dynamische System, wie z.B. die Zeltabbildung [51], läßt sich Ergodizität explizit nachweisen. In den meisten Fällen bleibt die Ergodizität des Systems jedoch eine unbewiesene Hypothese.

2.4 Verallgemeinerte Dimensionen

Der klassische Dimensionsbegriff hat seinen Ursprung in der euklidischen Geometrie und deren Weiterentwicklung zur analytischen Geometrie mit der Einführung eines Koordinatensystems durch René Descartes (1596-1650).
Die Verallgemeinerung der auf diese Weise geschaffenen Objekte führt zum Begriff des Vektorraums. Unter der Dimension eines Vektorraums versteht man die Mächtigkeit seiner Basis.
Die von Brouwer entdeckte Invarianz der Dimension unter homöomorphen Abbildungen erlaubte eine Verallgemeinerung des Dimensionsbegriffes für beliebige, zum euklidischen Raum lokal homöomorphe Mengen.
Um die zerklüftete (fraktale) Struktur bestimmter Attraktoren quantitativ zu beschreiben, wurden dem klassischen Dimensionsbegriff Konzepte verallgemeinerter Dimensionen beiseitegestellt, die invariant gegenüber stetig differenzierbaren Koordinatentransformationen sind.

2.4.1 Die Hausdorffdimension

Hausdorff untersuchte im Jahre 1919 die Cantormenge im Hinblick auf das 1914 von Carathéodory eingeführte äußere Maß und fand für diese Menge einen nicht-ganzzahligen Wert des äußeren Maßes [17].
Seitdem ist Hausdorffs Name mit diesem Maß und dem weiterführenden Dimensionsbegriff verbunden, der später von Mandelbrot als fraktale Dimension bezeichnet wurde [33].
Als „fraktale Mengen“ (auch kurz „Fraktale“) bezeichnet Mandelbrot z.B. solche Mengen mit gebrochener Hausdorffdimension.

Für eine Menge $U \subseteq \mathbb{R}^d$ mit $U \neq \emptyset$ ist der Durchmesser von U definiert als

$$d(U) := \sup\{\|x - y\| \text{ mit } x, y \in U\}.$$

Sei $A \subset \bigcup_i U_i$ und $0 < d(U_i) \leq \epsilon$ für alle i, dann heißt $\{U_i\}_i$ ϵ-Überdeckung von A.

Als äußeres α-Hausdorffmaß einer Menge $A \subseteq \mathbb{R}^d$, $\alpha \geq 0$ bezeichnet man nun den Grenzwert

$$H^\alpha(A) := \lim_{\epsilon \to 0} \inf \left\{ \sum_i d(U_i)^\alpha \mid d(U_i) \leq \epsilon \text{ und } \{U_i\}_i \text{ } \epsilon\text{-Überdeckung von } A \right\}$$

und als *Hausdorffdimension* von A den eindeutigen Wert

$$\dim_H(A) := \inf\{\alpha | H^\alpha(A) = 0\} = \sup\{\alpha | H^\alpha(A) = \infty\} .$$

2.4.2 Die Kapazitätsdimension

Sei $N(\epsilon, A)$ die Anzahl der offenen Kugeln mit dem Radius ϵ, die benötigt wird, um die Menge A zu überdecken. Die *Kapazitätsdimension* $\dim_K(A)$ wird dann definiert durch

$$\dim_K(A) \quad := \quad \lim_{\epsilon\to 0}\sup \frac{\log(N(\epsilon, A))}{\log(1/\epsilon)}. \tag{2.10}$$

Für die klassischen Objekte der euklidischen Geometrie (z.B. die platonischen Körper) nehmen Hausdorffdimension und Kapazitätsdimension nur ganzzahlige Werte an, d.h., sie stimmen mit der oben erwähnten klassischen Dimension überein [17].

2.4.3 Die Rényi-Dimensionen

Dieser Dimensionsbegriff wurde von Rényi in die Informationstheorie eingeführt [42].
Sei $\{U_i\}_i$ eine ϵ-Überdeckung von A und $p_i = \mu(U_i)$ die Aufenthaltswahrscheinlichkeit des Systems in der Menge U_i des Attraktors A. Dann definiert man daß Maß q-ter Ordnung der Information als

$$I_q(\epsilon, \mu) \quad := \quad \frac{1}{1-q}\log_2(\sum_i p_i^q) \qquad q \in \mathbb{R}\ , \tag{2.11}$$

und die verallgemeinerte *Rényi-Dimension q-ter Ordnung* als den Grenzwert

$$D_q(A, \mu) \quad := \quad \lim_{\epsilon\to 0} \frac{I_q(\epsilon, \mu)}{\log_2(1/\epsilon)}. \tag{2.12}$$

Die Verwendung des Logarithmus zur Basis 2 anstelle des natürlichen Logarithmus ist in der Informationstheorie üblich. Die verallgemeinerte Rényi-Dimension q-ter Ordnung ist jedoch von der Wahl der Basis des Logarithmus unabhängig. Im Grenzfall $q \to 0$ erhält man die Kapazitätsdimension. Für $q \to 1$ entspricht (2.11) der Shannon-Information [52], und man nennt die zugehörige verallgemeinerte Dimension auch die Informationsdimension [32].
Unter Umständen ist mit der Bestimmung der Rényi-Dimensionen ein hoher numerischer Aufwand verbunden, besonders in hochdimensionalen Systemen ist die Partitionierung des Phasenraumes sehr speicherintensiv.
Der von Grassberger vorgeschlagene Algorithmus [21] erlaubt eine weniger aufwendige Berechnung der ganzzahligen Rényi-Dimensionen, besonders vereinfacht sich die Berechnung für den Fall $q = 2$, die sogenannte *Korrelationsdimension.*
Allgemein gilt für die Rényi-Dimensionen die Ungleichung

$$D_q(A, \mu) \le D_p(A, \mu) \qquad \text{für} \quad q > p\ . \tag{2.13}$$

2.5 Lyapunov-Exponenten

Bereits H. Poincaré machte die Beobachtung, daß in bestimmten Systemen kleine Störungen in den Anfangsbedingungen zu langfristig unvorhersagbarem Systemverhalten führen [39] und somit scheinbar „zufällige" Erscheinungen durchaus deterministischen Ursprungs sein können.
Diese Einsichten wurden wieder aktuell, als man die Entwicklung turbulenter Strömungen untersuchte (siehe z.B. Ruelle [47]) und dabei *seltsame Attraktoren*[2] fand.
Die sensitive Abhängigkeit von den Anfangsbedingungen kann als wesentliches Merkmal chaotischer Dynamik angesehen werden.
Um dieses Phänomen quantitativ zu beschreiben, bedient man sich des Konzeptes der Lyapunov-Exponenten (auch charakteristische Exponenten genannt), die Aufschluß über die mittlere Divergenzrate zweier benachbarter Trajektorien im Phasenraum geben.
Um die Lyapunov-Exponenten für kontinuierliche Systeme zu definieren, betrachtet man zunächst die Linearisierung des Flusses $T^n_{\vec{x}}$ im Punkt $\vec{x}$.
Die Matrix der partiellen Ableitungen der Komponeten des Flusses $\varphi(\vec{x}, 1)$ an der Stelle $\vec{x}$ wird definiert durch

$$T_{\vec{x}} \quad := \quad D_{\vec{x}}(\varphi(\vec{x}, 1)) = (\partial\varphi_i(\vec{x}, 1)/\partial x_i), \tag{2.14}$$

und für $T^n_{\vec{x}}$ ergibt sich

$$T^n_{\vec{x}} \quad = \quad D_{\vec{x}}\varphi(\vec{x}, n). \tag{2.15}$$

Man definiert nun die symmetrische Matrix $\Lambda_{\vec{x}}$ durch

$$\Lambda_{\vec{x}} \quad := \quad \lim_{n\to\infty} (T^{n\dagger}_{\vec{x}} \cdot T^n_{\vec{x}})^{1/2n} \,, \tag{2.16}$$

wobei $T^{n\dagger}_{\vec{x}}$ die Transponierte der Matrix $T^n_{\vec{x}}$ bezeichnet.
Nach dem multiplikativen Ergodentheorem von Oseledec existiert der Grenzwert (2.16) für ein ergodisches Maß μ auf einem kompakten Träger und ist für μ-fast alle $\vec{x}$ konstant [46].
Die Matrix (2.16) definiert eine lineare Abbildung auf dem Tangentialraum im Punkt $\vec{x}$, deren Eigenwerte sich als die „Streckungsfaktoren" der Abbildung entlang der entsprechenden Eigenräume interpretieren lassen. Die logarithmischen Eigenwerte dieser Matrix nennt man die Lyapunov-Exponenten (oder auch das Lyapunov-Spektrum) des Systems, und man bezeichnet sie mit $\{\lambda_1, \ldots, \lambda_d\}$.
Man kann das Lyapunov-Spektrum zur Klassifizierung verschiedener Attraktortypen benutzen, wobei ein Attraktor *chaotisch* genannt wird, wenn mindestens ein Lyapunov-Exponent echt positiv ist.

[2] Ein seltsamer Attraktor ist nach Ruelle der Attraktor eines Systems mit empfindlicher Abhängigkeit von den Anfangsbedingungen [46].

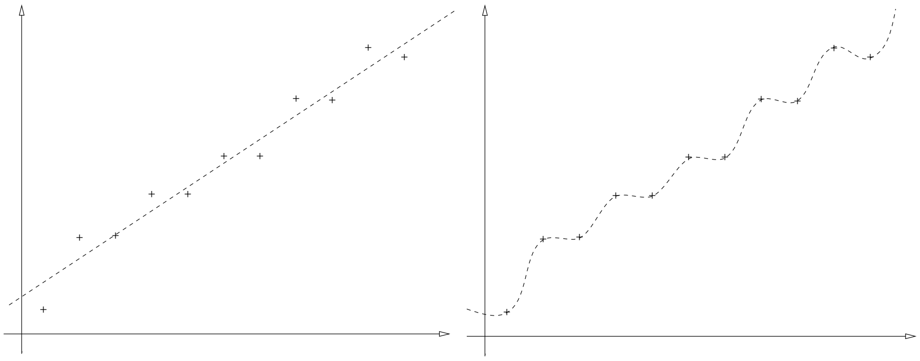


Abbildung 2.1: Ein lineares Modell zur Beschreibung von 11 Meßwerten (links). Die Abweichungen werden als Meßfehler interpretiert. Ein Polynom 11-ten Grades wurde durch dieselben 11 Meßwerte gelegt (rechts).

Die experimentelle Bestimmung der Lyapunov-Exponenten wird in der Regel durch den Einfluß von Rauschen und durch die Begrenzung der Meßgenauigkeit extrem erschwert [32].

2.6 Modellvalidierung

Eine wichtige Frage in der Modellbildung ist die Bestimmung der optimalen Größe des Modells, die in der Regel durch die Anzahl der freien Parameter gegeben ist.
In diesem Sinne wird die Verkleinerung des Modells zu einer schlechteren Beschreibung der Daten führen, aber dafür wird das Modell für den Beobachter „überschaubarer". Zum Beispiel könnte man versuchen, einen Datensatz von 11 Meßwerten, der sich nicht exakt auf einer Geraden unterbringen läßt, zu beschreiben, indem man jeden Wert zur Stützstelle einer Basisfunktion macht. Die Übereinstimmung von Modell und Daten wäre perfekt, aber niemand würde dieses Modell für sinnvoll halten, sondern eher einen linearen Zusammenhang annehmen und für die Abweichungen Meßfehler verantwortlich machen. In Abbildung 2.1 wurde dieses Beispiel veranschaulicht. Hat man keine andere Information über das System als diese Messung, dann scheint der lineare Zusammenhang „natürlicher" zu sein als ein Polynom mit 11 Stützstellen. Dieses Bestreben nach „Vereinfachung" läßt sich nicht aus den Grundlagen der Wahrscheinlichkeitsrechnung ableiten, es ist vielmehr eine Denknotwendigkeit, die von außen an die beschreibende Theorie herangetragen werden muß [29]. In der angelsächsischen Literatur bezeichnet man dieses Prinzip als *Ockham's Razor*[3], und es besagt, daß man unter allen Mo-

[3] Der Franziskanermönch William of Ockham (1285-1347) wandte das Motto *„Non sunt multiplicanda entia praeter necessitatem"* mit solcher Rasiermesserschärfe an, daß sich die Bezeichnung *Ockham's Razor* bis heute gehalten hat (s. Encyclopaedia Britannica).

dellen, die einen Sachverhalt beschreiben, stets das einfachste wählen sollte [58]. Ein Modell ist in diesem Sinne „einfach“, wenn es mit wenigen Grundannahmen auskommt, oder wenn die mathematische Beschreibung nicht sehr aufwendig ist. Zur Optimierung der Modellgröße sollen im folgenden zwei verschiedene Methoden vorgestellt werden.

2.6.1 Minimum-Description-Length (MDL)

Die von J.Rissanen [43, 44, 45] eingeführte *Minimum-Description-Length* (MDL) dient zur Festlegung der Modellgröße.
Angenommen, man hat zu einem Satz Meßdaten $x = \{x_1, \dots, x_N\}$ ein Modell $F(x, \theta)$ zur Beschreibung dieser Daten aufgestellt, wobei $\theta = (\theta_1, \dots, \theta_k)$ die Parameter dieses Modells sein sollen.
Der übliche Weg, die Parameter θ abzuschätzen, wird durch die *Maximum-Likelihood-Methode* (ML) gegeben. Dazu definiert man die *Likelihoodfunktion* $L_x(\theta)$, die jedem θ die Wahrscheinlichkeit $P_\theta(x)$ der Beobachtung von x zuordnet. Die Idee hinter diesem Schätzer ist, daß Werte der Parameter θ, für die $P_\theta(x)$ groß ist, plausibler erscheinen als solche, für die $P_\theta(x)$ klein ist [30]. Oft ist es bequemer, die logarithmische Likelihoodfunktion $\mathcal{L}_x(\theta)$ zu verwenden, um eine Schätzung für θ zu erhalten. Hat man ein θ^* gefunden, welches $\mathcal{L}_x(\theta)$ maximiert, so nennt man θ^* die ML-Schätzung von θ. Leider erhält man auf diese Weise keinen Aufschluß über die Größe des Modells. Es ist klar, daß eine Vergrößerung des Parameterraumes zu einer besseren Beschreibung der Daten führt, aber das Modell wird dadurch unnötig aufgebläht (man spricht in diesem Fall von *Overfitting* [50]).

Die Idee der MDL ist die Einführung eines Terms in $\mathcal{L}_x(\theta)$, der die „Modellkosten“ berücksichtigt. Der von Rissanen [44] gegebene Ausdruck für diese Kostenfunktion lautet

$$L(x) \quad := \quad -\ln\left(P_\theta(x)\right) + \frac{k}{2}\ln(N) \ , \tag{2.17}$$

dabei ist N die Länge des Datensatzes und k die Anzahl der Modellparameter. Nun werden θ^* und k^* durch das Minimum der Funktion $L(x)$ bestimmt. Vergrößert man die Anzahl der Modellparameter, so verkleinert sich $-\ln\left(P_\theta(x)\right)$, aber es erhöhen sich die Kosten $\frac{k}{2}\ln(N)$ und vice versa.

Die Ergebnisse Rissanes lassen sich auch von einem informationstheoretischen Standpunkt aus interpretieren. Dabei geht es darum, die ideale Codelänge zur Übertragung der Daten und der Modellparameter zu ermitteln. Diese Herangehensweise wird in Kapitel 5 im Detail vorgestellt.

2.6.2 Cross-Validation

Bei dieser Methode geht man davon aus, daß sich ein Modell besonders bei der Vorhersage zukünftiger (unbekannter) Daten auszeichnen sollte.
Um die Vorhersagefähigkeit des Modells zu testen, teilt man den Datensatz in zwei Teile:

- Einen Trainingsdatensatz, auf dessen Grundlage das Modell erstellt wird
- Einen Testdatensatz, an welchem man den Vorhersagefehler unter Verwendung verschiedener Modellterme bestimmt

Dabei ist wesentlich, daß der Testdatensatz von der Modellbildung ausgeschlossen wird, und somit eigentlich ein Test am zu modellierenden System gemacht wird[4]. Man startet mit einem „leeren“ Modell und fügt iterativ neue Modellterme hinzu. Dabei wird der Vorhersagefehler am Testdatensatz sinken, solange man die für die Systemdynamik wesentlichen Modellterme hinzunimmt. Die Artefakte des Overfitting sollten jedoch den Vorhersagefehler wieder anwachsen lassen, da sie im allgemeinen nicht die Systemdynamik beschreiben.

2.7 Rekonstruktion des Attraktors

Führt man an einem dynamischen System eine Messung durch, so mißt man normalerweise eine einzige Systemgröße durch zeitdiskrete Abtastung mit einer festen Abtastzeit r. Die erhaltene Liste der Meßwerte nennt man die *Zeitreihe*[5] des Systems. Die Einbettungstheoreme von Takens [55] und von Sauer et al. [48] eröffnen einen Weg, den Attraktor eines Systems aus einer Zeitreihe zu rekonstruieren.
Beide Theoreme sind Verallgemeinerungen des Whitneyschen Einbettungssatzes der Differentialtopologie, der besagt, daß sich jede d-dimensionale differenzierbare Mannigfaltigkeit abgeschlossen in den $\mathbb{R}^{2d+1}$ einbetten läßt [6].
Sei φ der Fluß eines dynamischen Systems auf der Mannigfaltigkeit M und $\tau \in \mathbb{R}$. Die Messung einer Systemgröße soll durch eine stetig differenzierbare Funktion $h : M \to \mathbb{R}$ beschrieben werden.
Die Idee bei der Einführung sog. Delay-Koordinaten ist die Konstruktion eines Vektors aus den jeweils um τ zeitverschobenen Werten der Zeitreihe.

[4] In der Statistik werden derartige Verfahren als *Bootstrap-Methoden* bezeichnet. Diese Bezeichnung rührt daher, daß man versucht, sich an der „Stiefelschlaufe“ (engl.: bootstrap) aus dem Sumpf mangelnder Kenntnis des Systems zu ziehen [30].

[5] In der mathematischen Zeitreihenanalyse definiert man Zeitreihen dieser Art als diskrete Realisierung eines stochastischen Prozesses, siehe z.B. [30, 7].

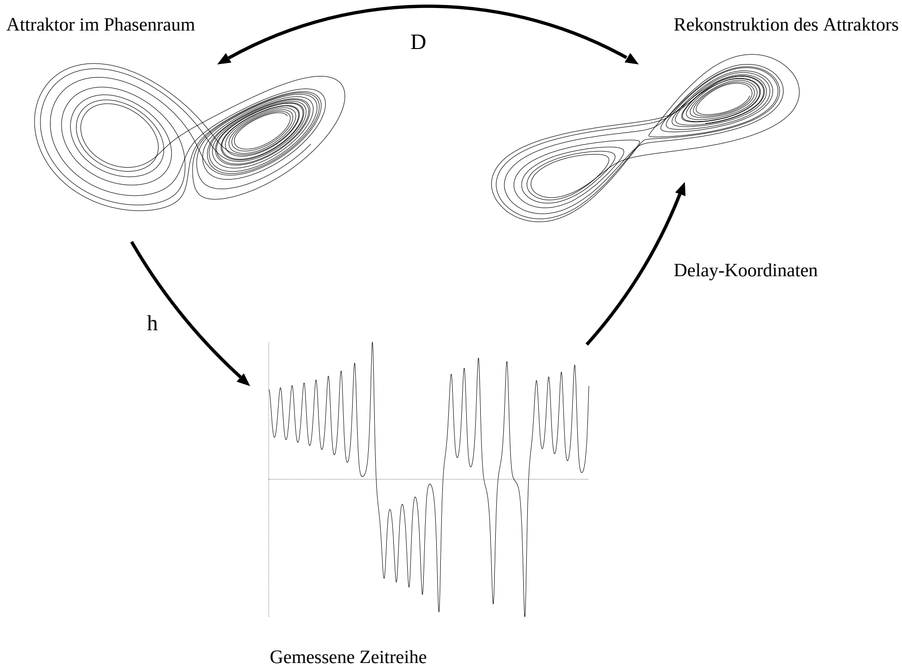


Abbildung 2.2: Schema der Delay-Einbettung

Die n-dimensionale Delay-Koordinaten-Abbildung $D(h, \varphi, \tau) : M \rightarrow \mathbb{R}^n$ mit der Delay-Zeit τ wird dann definiert durch

$$D(h, \varphi, \tau)(\vec{x}) = (h(\vec{x}), h(\varphi(\vec{x}, -\tau)), \ldots, h(\varphi(\vec{x}, -(n-1)\tau))).$$

Die Verwendung der Systemvergangenheit zur Beschreibung des aktuellen Zustandes geschah in Anlehnung an die lineare Zeitreihenanalyse, in der sich dieses Verfahren seit langem bewährt hat [7]. In Abbildung 2.2 ist dieses Verfahren schematisch dargestellt.

Sauer et al. [48] konnten die Ergebnisse von Takens verallgemeinern. Sie zeigten, daß unter weniger starken Voraussetzungen als beim Einbettungstheorem von Takens die Delay-Rekonstruktion des Attraktors diffeomorph zum Original-Attraktor im Phasenraum ist.
Für eine offene Menge $U \subset M$ und eine kompakte (φ-invariante) Menge $A \subset U$ mit der Kapazitätsdimension $\dim_K(A) = d$, in der $\varphi(\vec{x}, t)$ keine periodischen Orbits der Periode τ und 2τ, nur endlich viele Fixpunkte und periodische Orbits der Periode $3\tau, 4\tau, \ldots, n\tau$ besitzt und ferner die Linearisierung des Flusses entlang der periodischen Orbits verschiedene Eigenwerte hat, ist für fast jede stetig differenzierbare Funktion $h : U \rightarrow \mathbb{R}$ die Delay-Koordinaten-Abbildung $D(h, \varphi, \tau) : U \rightarrow \mathbb{R}^n$ eine eindeutig umkehrbare Abbildung und für jede kompakte Teilmenge einer in A enthaltenen glatten Mannigfaltigkeit auch eine Einbettung, falls $n > 2d$ gilt.

Die Möglichkeit, ein diffeomorphes Abbild der Dynamik im Phasenraum durch Messung einer (oder weniger) Systemgrößen zu erhalten, erlaubt insbesondere die Bestimmung wichtiger C^1-Invarianten des Systems (z.B. Rényi-Dimensionen, Lyapunov-Exponenten, Entropien). Darüberhinaus bietet dieses Verfahren auch die Möglichkeit der Visualisierung verschiedener Attraktoren aus gemessenen Zeitreihen.

2.8 Identifikation nichtlinearer deterministischer Systeme

Oftmals werden Systeme beobachtet, für die noch keine konsistente Theorie vorliegt, die Aufschluß über die inneren Zusammenhänge der beobachtbaren Größen liefern könnte, oder die so komplex sind, daß sie keine Reduktion auf „Laborbedingungen" erlauben (ökonomische Daten oder medizinische Meßreihen, wie z.B. Herzrhythmusstörungen im EKG).
Seit der erfolgreichen Einführung von nichtlinearen deterministischen Modellen (Farmer und Sidorovich (1987) [18], Crutchfield und McNamara (1987) [15], Cremers und Hübler (1987) [14] und Casdagli (1989) [9]) vermutete man in vielen Systemen mit irregulärem Zeitverhalten „deterministisches Chaos" auf niedrigdimensionalen Attraktoren, wobei sich die Hoffnung, einen strengen Determinismus zu finden, oft als trügerisch erwies.
In vielen Fällen erhebt sich die Frage, ob eine gemessene Zeitreihe überhaupt von einem deterministischen System stammt, oder ob dem zugrundeliegenden Determinismus noch eine stochastische Komponente überlagert ist (z.B. Meßfehler oder Rauschen).
Um nichtlineare deterministische Systeme von stochastischen Systemen zu trennen, wurden verschiedene statistische Methoden vorgeschlagen. Eine Möglichkeit bietet die von Theiler et al. vorgeschlagene Methode der Surrogatdaten [57].
Dazu formuliert man eine Hypothese über den möglichen Ursprung des beobachteten Prozesses (z.B.: „Es handelt sich um weißes Rauschen") und erzeugt sich nun mehrere, im Hinblick auf die Hypothese gleichwertige Datensätze, die sogenannten Surrogatdaten (in diesem Beispiel also Zeitreihen mit dem Mittelwert und der Varianz der gemessenen Daten). Dann berechnet man für die verschiedenen Datensätze einige charakteristische Größen, z.B. Lyapunovexponenten, Korrelationsdimensionen, Vorhersagefehler eines lokal linearen Modells usw., und vergleicht die jeweiligen Werte der Surrogatdaten mit den Originaldaten. Treten dabei signifikante Abweichungen zwischen den Surrogatdaten und den Originaldaten auf, so kann man die Hypothese auf einem vorgegebenen Signifikanzniveau verwerfen. Diese Methode hat natürlich auch einige Tücken, wenn man sie vorbehaltlos anwendet. Zu nennen wäre hier z.B. der Einfluß digitaler Filter auf die nichtlineare Struktur der Zeitreihe [56, 41].

Kapitel 3

Modellierung der Dynamik

Im folgenden wird angenommen, daß sich das System nach dem Abklingen der Einschwingvorgänge auf einem Attraktor $A \subset M$ befindet, wobei M eine differenzierbare Mannigfaltigkeit sein soll. Die Zeitentwicklung eines Zustandes $\vec{x}_z \in A$ wird beschrieben durch den Fluß $\varphi(\vec{x}_z, t)$. Die Messung einer Systemgröße geschieht in äquidistanten Zeitschritten mit der festen Samplingrate $r \in \mathbb{R}$ und wird beschrieben durch eine glatte Meßfunktion $h : M \to \mathbb{R}$. Auf diese Weise erhält man die Zeitreihe $\{x_1, x_2, \dots, x_N\}$ durch die Abbildung

$$x_i \quad = \quad h(\varphi(\vec{x}_z, r \cdot i)). \tag{3.1}$$

Ferner soll eine Delay-Abbbildung $D(h, \varphi, \tau) : M \to \mathbb{R}^d$ existieren, die ein diffeomorphes Bild des Attraktors im Rekonstruktionsraum liefert.
Im folgenden sei $\vec{x}_n = (x_n, x_{n-\tau}, \dots, x_{n-(d-1)\tau})$ ein Vektor im Rekonstruktionsraum des Systems.
Alle hier vorgestellten Arten der Modellierung beziehen sich auf den Rekonstruktionsraum des Systems. Der Fluß im Rekonstruktionsraum $\psi(\vec{x}, t)$ wird dann gegeben durch

$$\psi(\vec{x}, t) \quad = \quad D(\varphi(D^{-1}(\vec{x}), t)). \tag{3.2}$$

Die Hauptaufgabe der Modellierung besteht nun in der Approximation des Flusses $\psi(\vec{x}, t)$ im Rekonstruktionsraum. Dazu müssen zu Beginn drei grundsätzliche Entscheidungen getroffen werden:

- Die Wahl einer geeigneten Menge von Basisfunktionen, mit denen die Approximation durchgeführt werden soll
- Die Wahl eines Verfahrens zur Bestimmung der Parameter der Approximation und zur Auswahl der wesentlichen Modellterme
- Die Entscheidung für ein lokales oder für ein globales Modell

Bei der Auswahl der Basisfunktionen ist es sinnvoll, auf nichtlineare Funktionen zurückzugreifen, um der nichtlinearen Natur des Systems gerecht zu werden [10].
Das Verfahren zur Bestimmung der Modellparameter hängt im allgemeinen von der Wahl der Basisfunktionen ab.
Dabei kann man zwischen zwei Ansätzen auswählen:

- Modelle, die linear in den Parametern sind, so daß sich die Bestimmung der Parameter wieder auf ein lineares Optimierungsproblem zurückführen läßt
- Modelle, die nichtlinear in den Parametern sind

In dieser Arbeit sollen nur Modelle betrachtet werden, die linear in den Parametern[1] sind, da sich für diesen Fall der numerische Aufwand der Modellierung in einem erträglichen Rahmen hält.

3.1 Lokale Modelle

Von einem *lokalen Modell* spricht man, wenn man nur die Zustände in einem kleinen Gebiet des Rekonstruktionsraums zur Approximation des Flusses in dem entsprechenden Gebiet verwendet. Die ersten lokalen Modelle wurden 1987 von Farmer und Sidorovich [18] vorgeschlagen, bei dem Versuch die Zeitreihen niedrigdimensionaler chaotischer Systeme (Rayleigh-Bénard-Konvektion und Taylor-Couette-Strömung) vorrauszusagen. Grundlegend für diese Art der Modellierung ist die Stetigkeit des Flusses, aus der folgt, daß sich die Entwicklung eines Systemzustandes für einen kleinen Zeitraum nicht wesentlich von der Entwicklung eng benachbarter Zustände im Phasenraum (und somit auch im Rekonstruktionsraum) unterscheidet. Man versucht deshalb eine Abbildung zu finden, die den Fluß in der Nachbarschaft eines Zustandes $\vec{x}_n$ approximiert[2], um mit Hilfe dieser Abbildung die zukünftigen Systemzustände $\vec{x}_{n+i}, \quad i = 1, 2, \ldots$ vorauszusagen.
Dazu sucht man sich die k nächsten Nachbarn $\{\vec{x}_j\}_{j=1,\ldots,k}$ des Punktes $\vec{x}_n$ bezüglich einer geeigneten (z.B. der euklidischen) Metrik und betrachtet die Bilder $\{\vec{x}_{j+i}\}_{j=1,\ldots,k}$ dieser benachbarten Punkte unter der Wirkung des Flusses

$$\vec{x}_{j+i} = \psi(\vec{x}_j, i) \quad 1 \leq j \leq k. \tag{3.3}$$

[1] Man spricht in diesem Fall auch von „pseudolinearen“ Modellen, denn man führt die Approximation mit nichtlinearen Basisfunktionen auf ein lineares Optimierungsproblem zurück [27].

[2] Diese Abbildung wird bei M. Casdagli *„local predictor“* genannt, und die Approximation des Flusses setzt sich dann aus diesen lokalen Funktionen zusammen [9].

Die Methoden zur Approximation von $\psi(\vec{x}_j, i)$ sind im wesentlichen die gleichen, die auch zur globalen Modellierung herangezogen werden. Diese sollen im nächsten Abschnitt näher erläutert werden. Der einzige Unterschied liegt in der zugrundegelegten Menge der Daten, da man bei der globalen Modellierung alle rekonstruierten Zustände verwendet. Die Approximation des gesamten Flusses setzt sich dann aus den einzelnen lokalen Approximationen zusammen. Es ist also nicht möglich, diese Funktion in einer „geschlossenen Form“ anzugeben.
Der Vorteil der lokalen Modelle ist der geringere numerische Aufwand bei der Approximation des Flusses, da die zugrundegelegten Datensätze wesentlich kleiner sind als im globalen Fall.
Je länger die verwendeten Zeitreihen werden, desto besser wird in der Regel die lokale Approximation des Flusses, denn die Hinzunahme neuer Datenpunkte verringert im Mittel die Abstände der Punkte im Rekonstruktionsraum, was einen kleineren Fehler in der Vorhersage des nächsten Systemzustandes zur Folge hat [18].
Daraus folgt umgekehrt, daß lokale Modelle bei extrem kurzen Zeitreihen ihre Vorhersagefähigkeit einbüßen.
Die Suche der nächsten Nachbarn kann bei langen Zeitreihen sehr aufwendig werden. Zu dieser Problematik finden sich bei Schreiber [49, 50] Vorschläge für effiziente Suchmethoden.

3.2 Globale Modelle

Das Ziel der globalen Modellierung ist die Approximation des Flusses im Rekonstruktionsraum unter Berücksichtigung aller rekonstruierten Zustände. Im Gegensatz zu den lokalen Modellen ist es möglich, die gefundenen Modellgleichungen in „geschlossener Form“ anzugeben. Da sowohl die Delay-Koordinaten-Abbildung $D(h, \varphi, \tau)$ als auch die Meßfunktion $h : M \to \mathbb{R}$ stetig differenzierbar sind, folgt die Existenz einer glatten Abbildung

$$\begin{aligned} f : \mathbb{R}^d &\to \mathbb{R} \\ f(\vec{x}_n) &= x_{n+i} \end{aligned} \tag{3.4}$$

der rekonstruierten Zustände $\vec{x}_n$ auf die Werte der Zeitreihe x_{n+i}.
Für diese Funktion $f(\vec{x}_n)$ sucht man nun eine Näherung $\tilde{f}(\vec{x}_n)$ in einer geeigneten Funktionenbasis. Dabei stehen verschiedene Ansätze zur Auswahl.

3.2.1 Polynomiale Basisfunktionen

Nach dem Approximationssatz von Weierstraß läßt sich jede stetige Funktion auf einer kompakten Menge gleichmäßig durch eine Folge von Polynomen

annähern [24].
Dies rechtfertigt, einen Ansatz der Form

$$\tilde{f}(\vec{x}) = \sum_i a_i p_i(\vec{x}) \tag{3.5}$$

zu wählen, wobei $p_i(\vec{x})$ ein Polynom in d Veränderlichen ist. Die Bestimmung der Parameter läßt sich dabei auf ein lineares Optimierungsproblem zurückführen.
Polynome haben den Vorteil, daß sie sich einfach mit dem Computer generieren, speichern und behandeln lassen, was wohl die große Beliebtheit dieser Modelle erklärt.
Ein großes Problem polynomialer Modelle ist die Anzahl der Parameter, die in höheren Dimensionen und bei höherem Grad rapide wächst.

3.2.2 Radiale Basisfunktionen

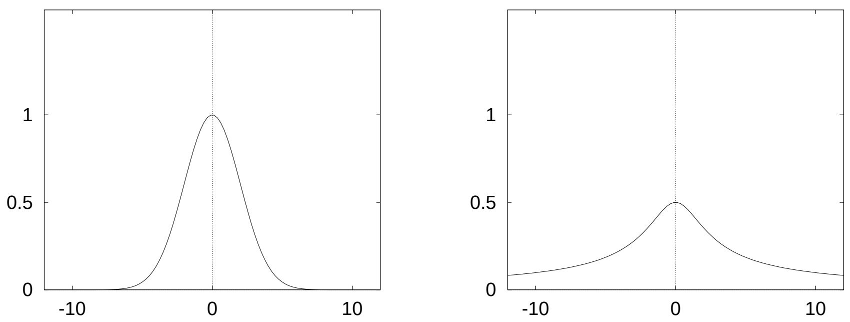


Abbildung 3.1: Auf der linken Seite ist die glockenförmige Basisfunktion aus Gleichung (3.7) für $\beta = 2$ dargestellt. Die Glockenkurve ist nur in der Nähe ihres Zentrums nennenswert von Null verschieden. Auf der rechten Seite ist die inverse multiquadratische Funktion aus Gleichung (3.9) dargestellt, ebenfalls für $\beta = 2$.

Radiale Basisfunktionen (RBF) sind skalare Funktionen $\phi(r)$, die von einem Argument $r \in \mathbb{R}$ abhängen. Um die Funktion $f(\vec{x})$ zu approximieren, wählt man k Zentren $\{\vec{y}_i\}_{i=1,\dots,k}$ auf dem rekonstruierten Attraktor aus und macht einen Ansatz der Form

$$\tilde{f}(\vec{x}) = \lambda_0 + \sum_i^k \lambda_i \phi_i(\|\vec{x} - \vec{y}_i\|). \tag{3.6}$$

Die Bestimmung der Koeffizienten λ_i kann man wieder auf ein lineares Verfahren zurückführen [12, 53].
Als RBF stehen mehrere Funktionstypen zur Auswahl, z.B. Glockenkurven

$$\phi(r) = \exp\left(-\frac{r^2}{2\beta^2}\right), \tag{3.7}$$

multiquadratische Funktionen

$$\phi(r) = (r^2 + \beta^2)^{\frac{1}{2}} \tag{3.8}$$

oder inverse multiquadratische Funktionen

$$\phi(r) = \frac{1}{(r^2 + \beta^2)^{\frac{1}{2}}} . \tag{3.9}$$

Dabei ist β eine reelle Konstante.
Die Glockenkurven und die inversen multiquadratischen Funktionen beeinflussen die Approximation nur lokal, in der Nähe ihres Zentrums. In Abbildung 3.1 ist dieses Verhalten an einem Beispiel veranschaulicht worden.

3.3 Die Methode der Singulärwertzerlegung

Um die Funktion in Gleichung (3.4) zu approximieren, kann man einen Superpositionsansatz der Art

$$\tilde{f}(\vec{x}) = \sum_{i=1}^{k} a_i \phi_i(\vec{x}) \tag{3.10}$$

wählen, nachdem man sich auf eine Klasse von Basisfunktionen $\phi_i(\vec{x})$ festgelegt hat. Die Bestimmung der Parameter $\{a_i\}_{i=1,\dots,k}$ wird dann zu einem linearen Problem.
Man sucht nach einem Parametersatz, der die Summe der quadrierten Fehler

$$\chi^2 = \sum_{n=1}^{N} \left(x_{n+1} - \sum_{i=1}^{k} a_i \phi_i(\vec{x}_n) \right)^2 \tag{3.11}$$

minimiert. Dazu definiert man die $(N \times k)$-Matrix $\mathbf{A}$ der Basisfunktionen

$$\mathbf{A} := \begin{pmatrix} \phi_1(\vec{x}_1) & \phi_2(\vec{x}_1) & \cdots & \phi_k(\vec{x}_1) \\ \phi_1(\vec{x}_2) & \phi_2(\vec{x}_2) & \cdots & \phi_k(\vec{x}_2) \\ & \ddots & & \\ & & \ddots & \\ \phi_1(\vec{x}_N) & \phi_2(\vec{x}_N) & \cdots & \phi_k(\vec{x}_N) \end{pmatrix} \tag{3.12}$$

und den Koeffizientenvektor

$$\vec{a} := (a_1, a_2, \dots, a_k)^\dagger \tag{3.13}$$

sowie den Bildvektor

$$\vec{y} := (x_{n+1}, x_{n+2}, \dots, x_{N+1})^\dagger . \tag{3.14}$$

Gleichung (3.11) läßt sich dann schreiben als

$$\chi^2 = \|\vec{y} - \mathbf{A}\vec{a}\|^2 , \tag{3.15}$$

und man sucht nun einen Vektor $\vec{a}$, der χ^2 minimiert. Dazu bedient man sich der Methode der Singulärwertzerlegung einer reellen Matrix [22].
Die $(N \times k)$-Matrix $\mathbf{A}$ läßt sich zerlegen in das Produkt einer zeilenweise orthogonalen $(N \times k)$-Matrix U, einer diagonalen $(k \times k)$-Matrix W und einer orthogonalen $(k \times k)$-Matrix $V^\dagger$, also

$$\mathbf{A} = U\,W\,V^\dagger \tag{3.16}$$

mit

$$U^\dagger\,U = E_k, \qquad V^\dagger\,V = E_k$$

und

$$W = \begin{pmatrix} w_1 & & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & & w_k \end{pmatrix}.$$

Die von Null verschiedenen Diagonalelemente der Matrix W nennt man die *singulären Werte* von $\mathbf{A}$ [22]. Die Anzahl der singulären Werte ist die Dimension des Bildes der zur Matrix $\mathbf{A}$ gehörenden Abbildung.
Der Vektor $\vec{a}$, für den Gleichung (3.15) das Minimum annimmt, wird gegeben durch

$$\vec{a} = V\,W^{-1}\,U^\dagger\,\vec{y}\,,$$

wobei man die Kehrwerte der verschwindenden Diagonalelemente aus W in der Matrix W^{-1} auf Null setzt [40].
Die Varianz der Schätzung des Koeffizienten a_j erhält man aus

$$\sigma^2(a_j) = \sum_{i=1}^{k} \left(\frac{V_{ji}}{w_i}\right)^2 , \tag{3.17}$$

wobei der Ausdruck (3.17) für verschwindende Diagonalelemente in W wieder auf Null gesetzt wird [40].
Eine anschauliche Interpretation der Gleichung (3.17) erhält man, wenn man sich die Spalten der Matrix V als die Einheitsvektoren in Richtung der Hauptachsen des Fehlerparaboloides[3] zum Koeffizientenvektor $\vec{a}$ vorstellt. Die Kehrwerte der singulären Werte aus der Matrix W sind dann die Längen der entsprechenden Hauptachsen, und einem großen Singulärwert entspricht eine zuverlässige Schätzung des zugehörigen Koeffizienten.
Die Algorithmen zur Singulärwertzerlegung, die in dieser Arbeit Verwendung fanden, wurden aus *Numerical Recipes in C* [40] entnommen.

[3]Ein Fehlerparaboloid ist das mehrdimensionale Analogon zum Konfidenzintervall im eindimensionalen Fall der Methode der kleinsten Quadrate.

3.4 Gradientenverfahren

Für Modelle, die linear in den Parametern sind, wurde die Suche nach den Koeffizienten der Basisfunktionen durch Minimierung des mittleren quadratischen Fehlers auf eine Singulärwertzerlegung der Matrix **A** zurückgeführt (siehe Gleichung (3.15)).
Diese Methode ist numerisch stabil und liefert zudem eine Schätzung für die Varianz der Koeffizienten, die zur Auswahl der wesentlichen Modellterme verwendet werden kann. Leider ist dieses Verfahren für große Matrizen und lange Zeitreihen sehr speicherintensiv, und es erhebt sich in solchen Fällen die Frage nach weniger aufwendigen Alternativen.
Eine Möglichkeit besteht in der Minimierung der Summe der quadratischen Fehler

$$\chi^2(\vec{a}) = \sum_{n=1}^{N} \left(y_n - \sum_{i=1}^{k} a_i \phi_i(\vec{x}_n) \right)^2 \tag{3.18}$$

mittels eines Gradientenverfahrens.
Dabei faßt man $\chi^2(\vec{a})$ als eine reellwertige Funktion der k Koeffizienten der Basisfunktionen auf

$$\begin{aligned} \chi^2 : \mathbb{R}^k &\to \mathbb{R}^+ \\ \vec{a} &\mapsto \chi^2(\vec{a}) \end{aligned} \tag{3.19}$$

und versucht einen optimalen Koeffizientenvektor $\vec{a}$ zu finden, indem man sich auf der k-dimensionalen Fehlerlandschaft in absteigender Richtung bewegt.

Zur Vereinfachung der Notation werden die folgenden Ausdrücke definiert:

$$\begin{aligned} \widetilde{\phi_i\phi_j} &= \sum_{n=1}^{N} \phi_i(\vec{x}_n)\phi_j(\vec{x}_n) \qquad & \widetilde{y^2} &:= \sum_{n=1}^{N} (y_n)^2 \\ \tilde{\phi_i} &:= \sum_{n=1}^{N} \phi_i(\vec{x}_n) \qquad & \widetilde{y\phi_i} &:= \sum_{n=1}^{N} y_n \, \phi_i(\vec{x}_n) \,. \end{aligned}$$

Ferner definiert man die symmetrische $(k \times k)$- Matrix Φ und den Vektor $\vec{\Upsilon}$ als

$$\Phi := \begin{pmatrix} \widetilde{\phi_1\phi_1} & \widetilde{\phi_1\phi_2} & \dots & \widetilde{\phi_1\phi_k} \\ & \ddots & & \\ & & \ddots & \\ \widetilde{\phi_k\phi_1} & \widetilde{\phi_k\phi_2} & \dots & \widetilde{\phi_k\phi_k} \end{pmatrix} \qquad \vec{\Upsilon} := \begin{pmatrix} \widetilde{y\phi_1} \\ \widetilde{y\phi_2} \\ \vdots \\ \widetilde{y\phi_k} \end{pmatrix} . \tag{3.20}$$

Gleichung (3.18) läßt sich dann umformen zu

$$\begin{aligned}
\chi^2(\vec{a}) &= \sum_{n=1}^{N} \left(y_n - \sum_{i=1}^{k} a_i \phi_i(\vec{x}_n) \right)^2 \\
&= \sum_{n=1}^{N} \left[(y_n)^2 - 2 \left(y_n \sum_{i=1}^{k} a_i \phi_i(\vec{x}_n) \right) + \left(\sum_{i=1}^{k} a_i \phi_i(\vec{x}_n) \right) \cdot \left(\sum_{j=1}^{k} a_j \phi_j(\vec{x}_n) \right) \right] \\
&= \widetilde{y^2} \; - \; 2\langle \vec{a} | \vec{\Upsilon} \rangle \; + \; \langle \vec{a} | \Phi \vec{a} \rangle \;, \qquad (3.21)
\end{aligned}$$

dabei bezeichnet $\langle \cdot | \cdot \rangle$ das Standardskalarprodukt im $\mathbb{R}^k$.
Nach Gleichung (3.21) läßt sich $\chi^2(\vec{a})$ als eine positiv semidefinite quadratische Form in $\{a_i\}_{i=1,\ldots,k}$ auffassen, wodurch die Existenz eines eindeutigen globalen Minimums gesichert ist.

3.4.1 Einfacher Gradientenabstieg

Es liegt nahe, sich dem Minimum zu nähern, indem man die Fehlerlandschaft in negativer Gradientenrichtung durchwandert.
Zur Berechnung des Fehlergradienten $\vec{\nabla}\chi^2(\vec{a})$ bildet man die partiellen Ableitungen

$$\begin{aligned}
\frac{\partial \chi^2(\vec{a})}{\partial a_j} &= \sum_{n=1}^{N} -2 \, \phi_j(\vec{x}_n) \left(y_n - \sum_{i=1}^{k} a_i \phi_i(\vec{x}_n) \right) \\
&= 2 \left[\sum_{i=1}^{k} a_i \, \left(\sum_{n=1}^{N} \phi_j(\vec{x}_n) \phi_i(\vec{x}_n) \right) - \sum_{n=1}^{N} y_n \phi_j(\vec{x}_n) \right] \\
&= 2 \left[\sum_{i=1}^{k} a_i \widetilde{\phi_i \phi_j} - \Upsilon_j \right] \qquad (3.22)
\end{aligned}$$

und erhält als Gradienten

$$\vec{\nabla}\chi^2(\vec{a}) \;\; = \;\; 2\,[\Phi \vec{a} - \Upsilon] \,. \qquad (3.23)$$

Nach Wahl eines Startvektors $\vec{a}_0$ bewegt man sich nun nach der Iterationsvorschrift

$$\begin{aligned}
\vec{a}_{n+1} &= \vec{a}_n - \lambda \, \vec{\nabla}\chi^2(\vec{a}_n) \\
&= \vec{a}_n - 2\lambda \, [\Phi \vec{a}_n - \Upsilon] \qquad (3.24)
\end{aligned}$$

in Richtung des Fehlerminimums, wobei λ der Parameter zur Steuerung der Schrittweite des Abstiegs ist. Es ist klar, daß von der Wahl dieses Parameters die Konvergenzgeschwindigkeit des Verfahrens abhängen wird, jedoch ist von vornherein nicht klar, wie man diese Wahl zu treffen hat.

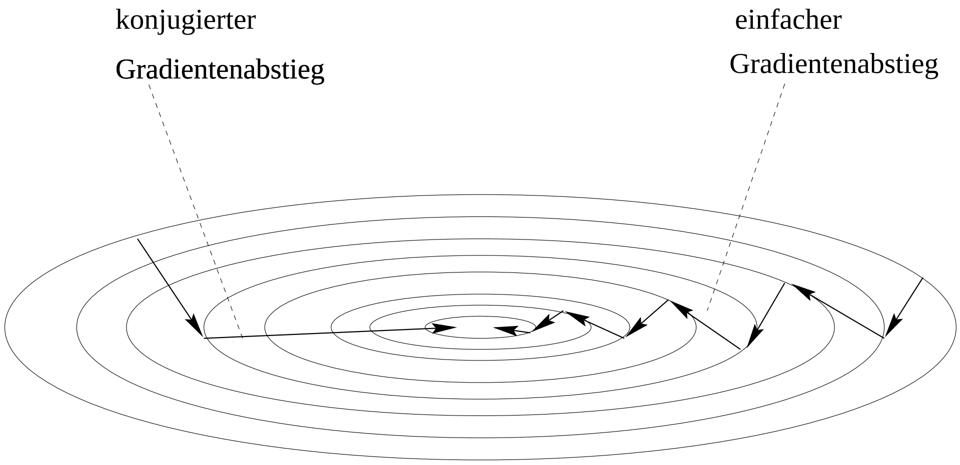


Abbildung 3.2: Schema des einfachen bzw. konjugierten Gradientenverfahrens auf den Höhenlinien einer Funktion in zwei Variablen

Eine Möglichkeit besteht darin, sich bis zum Minimum in Gradientenrichtung zu bewegen (dann ist der Gradient gerade parallel zu den Höhenlinien), um von dort aus wieder orthogonal zur alten Bewegungsrichtung zu starten (siehe Abbildung 3.2).
In manchen Fällen kann diese Vorgehensweise dazu führen, daß man sich in vielen kleinen Schritten um die „optimale" Abstiegsrichtung herumbewegt.

3.4.2 Konjugierter Gradientenabstieg

Die Idee bei diesem Verfahren ist die Konstruktion sogenannter *konjugierter Richtungen*, in denen eine Minimierung der Funktion durchgeführt werden soll. Dabei nennt man zwei Vektoren $\vec{h}_i$ und $\vec{h}_j$ *konjugiert bezüglich* Φ, wenn gilt

$$\langle \vec{h}_i | \Phi \vec{h}_j \rangle \quad = \quad 0\,.$$

Angenommen, man hat sich entlang der Richtung $\vec{h}_i$ in das Minimum dieser Richtung $\vec{a}_{i+1}$ bewegt. Die neue Richtung $\vec{h}_{i+1}$ erhält man nun, indem man zu dem Gradienten $\vec{g}_{i+1}$ im Punkt $\vec{a}_{i+1}$ die alte Richtung $\vec{h}_i$ skaliert mit dem Faktor γ_i addiert.

Die neue Richtung wird somit gegeben durch

$$\vec{h}_{i+1} \quad := \quad \vec{g}_{i+1} + \gamma_i \vec{h}_i\,, \tag{3.25}$$

wobei man den Skalar γ_i entweder nach der Methode von Polak-Ribiere bestimmt zu

$$\gamma_i \quad := \quad \frac{\langle \vec{g}_{i+1} | \vec{g}_{i+1} \rangle}{\langle \vec{g}_i | \vec{g}_i \rangle} \tag{3.26}$$

oder sich der modifizierten Methode von Fletcher-Reeves bedient:

$$\gamma_i \quad := \quad \frac{\langle (\vec{g}_{i+1} - \vec{g}_i) | \vec{g}_{i+1} \rangle}{\langle \vec{g}_i | \vec{g}_i \rangle} \; . \tag{3.27}$$

Die so konstruierten Vektoren genügen den Orthogonalitätsbedingungen

$$\langle \vec{g}_i | \vec{g}_j \rangle = 0 \quad \langle \vec{g}_i | \vec{h}_j \rangle = 0 \qquad i \neq j \tag{3.28}$$

sowie der Bedingung

$$\langle \vec{h}_i | \Phi \vec{h}_j \rangle \; = \; 0 \qquad i \neq j \; , \tag{3.29}$$

welche die oben angesprochene Konjugation der Richtungen zum Ausdruck bringt [40].
Wenn man Gleichung (3.29) geometrisch interpretiert, dann handelt es sich dabei um die Orthogonalität der beiden Richtungsvektoren $\vec{h}_i$ und $\vec{h}_j$ bezüglich der symmetrischen Bilinearform $\langle \cdot \mid \Phi \, \cdot \rangle$ und des dazugehörigen *Vektorraumes mit geometrischer Struktur* [34].
Der konjugierte Gradientenabstieg erfolgt dann optimal bezüglich der Geometrie der Fehlerlandschaft.

3.5 Measure-Based-Modeling

Die grundlegende Idee des Measure-Based-Modeling ist die Approximation der Abbildung $f : \mathbb{R}^d \rightarrow \mathbb{R}$ durch Polynome, die eine Orthonormalbasis bzgl. eines diskreten Skalarproduktes bilden, das durch die Zeitreihe definiert wird.
In diesem Abschnitt soll das zur Berechnung einer Orthonormalbasis benötigte Gram-Schmidt-Verfahren vorgestellt werden, das von Gioana et al. [20] vorgeschlagen wurde. Zu diesem klassischen Gram-Schmidt-Verfahren gibt es eine modifizierte, numerisch stabilere Variante, die im zweiten Teil erläutert wird.
Zunächst betrachtet man die d-dimensionalen Rekonstruktionsvektoren $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\}$, die man aus der Zeitreihe gewonnen hat. Für diese Vektoren $\vec{x} = (x_1, x_2, \ldots, x_d)$ definiert man nun Polynome in d Veränderlichen. Zur mathematischen Behandlung dieser Polynome führt man Multiindizes in Form von d-Tupeln natürlicher Zahlen ein.
Sei $\mathcal{I} \in \mathbb{N}^d$ ein solcher Index mit $\mathcal{I} := (i_1, i_2, ..., i_d)$ und $|\mathcal{I}| := \sum_{\nu=1}^{d} i_\nu$ der Betrag von $\mathcal{I}$.
Die Summe zweier Indizes $\mathcal{I}$ und $\mathcal{J}$ wird komponentenweise definiert durch

$$\mathcal{I} + \mathcal{J} \quad := \quad (i_1 + j_1, i_2 + j_2, \ldots, i_d + j_d) \;\; . \tag{3.30}$$

Ein Monom in $\vec{x}$ wird dann gegeben durch

$$\vec{x}^{\mathcal{I}} \quad := \quad x_1^{i_1} x_2^{i_2} ... x_d^{i_d}, \tag{3.31}$$

wobei $|\mathcal{I}|$ der Grad des Monoms genannt wird. Für das Produkt zweier Monome erhält man

$$\vec{x}^{\mathcal{I}} \cdot \vec{x}^{\mathcal{J}} \quad = \quad \vec{x}^{\mathcal{I}+\mathcal{J}} .$$

Für eine Approximation von $\tilde{f}(\vec{x})$ in d Dimensionen vom Grad g in der Form

$$\tilde{f}(\vec{x}) = \sum_{\mathcal{I}} p^{(\mathcal{I})} \pi^{\mathcal{I}}(\vec{x})$$

hat man

$$k = \frac{(g+d)!}{g!\, d!} \tag{3.32}$$

Koeffizienten $p^{(\mathcal{I})}$ zu bestimmen.
Um die Polynome in einer fortlaufenden Notation bezeichnen zu können, wird eine Ordnungsrelation auf dem $\mathbb{N}^d$ eingeführt. Für einen vorgegebenen Approximationsgrad g faßt man ein $\mathcal{I} \in \mathbb{N}^d$ als die d-stellige g-adische Entwicklung einer natürlichen Zahl auf, sortiert die Indizes nach der Größe und numeriert sie von 1 bis k.
Für den Fall $g = 2$ und $d = 3$ erhält z.B. die Reihenfolge:

$$(0,0,0) \leq (0,0,1) \leq (0,0,2) \leq (0,1,0) \leq (0,1,1) \leq (0,2,0) \leq (1,0,0) \ldots$$

Das Polynom $\pi^{\mathcal{I}_\nu}(\vec{x})$ hat dann die Gestalt

$$\pi^{\mathcal{I}_\nu}(\vec{x}) \quad = \quad \sum_{q=1}^{\nu} a_{(\nu,q)} \vec{x}^{\mathcal{I}_q} . \tag{3.33}$$

Dabei ist $a_{(\nu,q)}$ der Koeffizient des q-ten Monoms im ν-ten Basispolynom, und man bezeichnet $|\mathcal{I}_\nu|$ als den Grad von $\pi^{\mathcal{I}_\nu}(\vec{x})$.
Für die weitere Behandlung ist es vorteilhaft, die $\mathcal{I}$-ten Momente $M_{\mathcal{I}}$ der Zeitreihe als

$$M_{\mathcal{I}} \quad := \quad \frac{1}{N} \sum_{i=1}^{N} \vec{x}_i^{\mathcal{I}} \tag{3.34}$$

zu definieren.
Das Skalarprodukt bzgl. des ergodischen Maßes μ auf dem Attraktor A des Systems definiert man für zwei Funktionen $\pi^{\mathcal{I}}(\vec{x}), \pi^{\mathcal{J}}(\vec{x})$ durch

$$\left\langle \pi^{\mathcal{I}}(\vec{x}) \,\middle|\, \pi^{\mathcal{J}}(\vec{x}) \right\rangle_\mu \quad := \quad \int_A \pi^{\mathcal{I}}(\vec{x}) \pi^{\mathcal{J}}(\vec{x}) d\mu(\vec{x}) \quad , \tag{3.35}$$

ferner wird das diskrete Zeitmittel des Produktes der Funktionen $\pi^{\mathcal{I}}, \pi^{\mathcal{J}}$ definiert als

$$\left\langle \pi^{\mathcal{I}}(\vec{x}) \cdot \pi^{\mathcal{J}}(\vec{x}) \right\rangle_t \quad := \quad \frac{1}{N} \sum_{i=1}^{N} \pi^{\mathcal{I}}(\vec{x}_i) \pi^{\mathcal{J}}(\vec{x}_i) \quad . \tag{3.36}$$

Wegen der vorausgesetzten Ergodizität des Systems stimmt das Phasenraummittel aus Gleichung (3.35) mit dem diskreten Zeitmittel (3.36) überein, und dadurch ergibt sich für das diskrete Skalarprodukt

$$\left\langle \pi^{\mathcal{I}}(\vec{x}) \;\middle|\; \pi^{\mathcal{J}}(\vec{x}) \right\rangle_\mu \;=\; \frac{1}{N}\sum_{i=1}^{N} \pi^{\mathcal{I}}(\vec{x}_i)\pi^{\mathcal{J}}(\vec{x}_i)\;. \tag{3.37}$$

An dieser Stelle sei angemerkt, daß man ein diskretes Skalarprodukt wie in Gleichung (3.37) auch ohne den Formalismus des ergodischen Maßes einführen kann. An den folgenden Betrachtungen ändert sich dadurch nichts. Da dieser Formalismus von Giona et al. [20] eingeführt wurde, und wir uns in dieser Arbeit auf diesen Ansatz beziehen, soll die entsprechende Notation beibehalten werden[4].

Unter Verwendung von Definition (3.34) läßt sich das Skalarprodukt $\left\langle \vec{x}^{\mathcal{I}} \;\middle|\; \vec{x}^{\mathcal{J}} \right\rangle_\mu$ als Moment der Zeitreihe schreiben :

$$\left\langle \vec{x}^{\mathcal{I}} \;\middle|\; \vec{x}^{\mathcal{J}} \right\rangle_\mu \;=\; M_{(\mathcal{I}+\mathcal{J})}. \tag{3.38}$$

3.5.1 Das klassische Gram-Schmidt-Verfahren

Als Ausgangsbasis für das Gram-Schmidt-Verfahren wählt man die Standardbasis $\left\{\vec{x}^{\mathcal{I}_\nu}\right\}_{\nu=1,\ldots,k}$ der Polynome in d Veränderlichen. Man startet nun mit $\pi^{\mathcal{I}_1}(\vec{x}) = 1$ und bildet iterativ die neue Basis:

$$\begin{aligned}
\tilde{\pi}^{\mathcal{I}_\nu}(\vec{x}) &= \vec{x}^{\mathcal{I}_\nu} - \sum_{n=1}^{\nu-1} \left\langle \pi^{\mathcal{I}_n}(\vec{x}) \;\middle|\; \vec{x}^{\mathcal{I}_\nu} \right\rangle_\mu \pi^{\mathcal{I}_n}(\vec{x}) \\
&= \vec{x}^{\mathcal{I}_\nu} - \sum_{n=1}^{\nu-1} \left[\left\langle \sum_{m=1}^{n} a_{(n,m)} \vec{x}^{\mathcal{I}_m} \;\middle|\; \vec{x}^{\mathcal{I}_\nu} \right\rangle_\mu \sum_{k=1}^{n} a_{(n,k)} \vec{x}^{\mathcal{I}_k} \right] \\
&= \vec{x}^{\mathcal{I}_\nu} - \sum_{n=1}^{\nu-1} \left[\sum_{m=1}^{n} a_{(n,m)} \left\langle \vec{x}^{\mathcal{I}_m} \;\middle|\; \vec{x}^{\mathcal{I}_\nu} \right\rangle_\mu \sum_{k=1}^{n} a_{(n,k)} \vec{x}^{\mathcal{I}_k} \right] \\
&= \vec{x}^{\mathcal{I}_\nu} - \sum_{n=1}^{\nu-1} \left[\sum_{m=1}^{n} a_{(n,m)} M_{(\mathcal{I}_m+\mathcal{I}_\nu)} \sum_{k=1}^{n} a_{(n,k)} \vec{x}^{\mathcal{I}_k} \right].
\end{aligned} \tag{3.39}$$

Ist das Polynom $\tilde{\pi}^{\mathcal{I}_\nu}(\vec{x})$ gegeben durch den Ausdruck

$$\tilde{\pi}^{\mathcal{I}_\nu}(\vec{x}) \;=\; \sum_{q=1}^{\nu} \tilde{a}_{(\nu,q)} \vec{x}^{\mathcal{I}_q},$$

[4] Der Ansatz von Giona et al. hat explizit Eingang in die Literatur gefunden, so z.B. bei Brown et al. [8].

dann erhält man die zugehörigen Koeffizienten $\tilde{a}_{(\nu,q)}$ durch Koeffizientenvergleich in Gleichung (3.39) in der Form :

$$\tilde{a}_{(\nu,q)} = \begin{cases} 1 & : \quad q = \nu \\ -\sum\limits_{n=q}^{\nu-1} \left[\sum\limits_{k=1}^{n} a_{(k,n)} M_{(\mathcal{I}_\nu + \mathcal{I}_k)} \right] a_{(n,q)} & : \quad q < \nu \end{cases} \tag{3.40}$$

Der Normierungsfaktor $\mathcal{N}_\nu$ des Polynoms $\tilde{\pi}^{\mathcal{I}_\nu}(\vec{x})$ ergibt sich zu:

$$\begin{aligned} \mathcal{N}_\nu &= \left[\left\langle \tilde{\pi}^{\mathcal{I}_\nu}(\vec{x}) \;\middle|\; \tilde{\pi}^{\mathcal{I}_\nu}(\vec{x}) \right\rangle_\mu \right]^{\frac{1}{2}} \\ &= \left[\left\langle \sum_{n=1}^{\nu} \tilde{a}_{(\nu,n)} \vec{x}^{\mathcal{I}_n} \;\middle|\; \sum_{m=1}^{\nu} \tilde{a}_{(\nu,m)} \vec{x}^{\mathcal{I}_m} \right\rangle_\mu \right]^{\frac{1}{2}} \\ &= \left[\sum_{n=1}^{\nu} \tilde{a}_{(\nu,n)} \sum_{m=1}^{\nu} \tilde{a}_{(\nu,m)} \left\langle \vec{x}^{\mathcal{I}_n} \;\middle|\; \vec{x}^{\mathcal{I}_m} \right\rangle_\mu \right]^{\frac{1}{2}} \\ &= \left[\sum_{n=1}^{\nu} \tilde{a}_{(\nu,n)} \sum_{m=1}^{\nu} \tilde{a}_{(\nu,m)} M_{(\mathcal{I}_n + \mathcal{I}_m)} \right]^{\frac{1}{2}} , \end{aligned} \tag{3.41}$$

und das neue Basispolynom $\pi^{\mathcal{I}_\nu}(\vec{x})$ erhält man dann durch Normierung des Polynoms

$$\pi^{\mathcal{I}_\nu}(\vec{x}) = \frac{1}{\mathcal{N}_\nu} \tilde{\pi}^{\mathcal{I}_\nu}(\vec{x}) \ .$$

3.5.2 Das modifizierte Gram-Schmidt-Verfahren

Das klassische Gram-Schmidt-Verfahren ist bei schlecht konditionierten Problemen anfällig gegen Rundungsfehler, welche die Orthogonalität der neuen Basisvektoren zerstören können [54]. Um dieses Verfahren zu verbessern, wurde von A. Björck eine Modifikation vorgeschlagen, die im wesentlichen auf einer Änderung in der Reihenfolge bei der Berechnung der ONB beruht [13]. Der Rechenaufwand ist bei beiden Verfahren gleich groß.
Beim klassischen Gram-Schmidt-Verfahren erhält man das ν-te Basispolynom, indem man das ν-te Polynom der Ausgangsbasis zu den bereits berechneten $\nu - 1$ Polynomen orthogonalisiert (siehe Gleichung (3.39)). Beim modifizierten Gram-Schmidt-Verfahren bedient man sich wieder der Ausgangsbasis $\left\{ \vec{x}^{\mathcal{I}_\nu} \right\}_{\nu=1,\dots,k}$ der Polynome in d Veränderlichen. Ferner definiert man sich „Hilfspolynome“ $\sigma_n^{\mathcal{I}_\nu}(\vec{x})$ gemäß Definition (3.33) zur Erfassung der Zwischenergebnisse im n-ten Schritt des Verfahrens.
Die gewünschte Orthogonalbasis erhält man, indem man im n-ten Schritt die Polynome $\left\{ \sigma_n^{\mathcal{I}_\nu}(\vec{x}) \right\}_{\nu=n+1,\dots,k}$ bezüglich des Polynoms $\pi^{\mathcal{I}_n}(\vec{x})$ orthogonalisiert, das neu erhaltene Polynom $\sigma_n^{\mathcal{I}_{n+1}}(\vec{x})$ normiert und als $(n+1)$-tes

Basispolynom definiert.
Dadurch erhält man die Iterationsvorschrift

$$\sigma_n^{\mathcal{I}_\nu}(\vec{x}) \;=\; \sigma_{n-1}^{\mathcal{I}_\nu}(\vec{x}) - \left\langle \sigma_{n-1}^{\mathcal{I}_\nu}(\vec{x}) \;\middle|\; \pi^{\mathcal{I}_n}(\vec{x}) \right\rangle \pi^{\mathcal{I}_n}(\vec{x}); \quad n < \nu \leq k$$

für die Hilfspolynome im n-ten Verfahrensschritt.
Durch Normierung des Polynoms $\sigma_n^{\mathcal{I}_{n+1}}(\vec{x})$ erhält man das $(n+1)$-te Basispolynom

$$\pi^{\mathcal{I}_{n+1}}(\vec{x}) \;=\; \frac{1}{\mathcal{N}_{n+1}}\, \sigma_n^{\mathcal{I}_{n+1}}(\vec{x}) \;,$$

dabei ist

$$\mathcal{N}_{n+1} \;=\; \left[\left\langle \sigma_n^{\mathcal{I}_{n+1}}(\vec{x}) \;\middle|\; \sigma_n^{\mathcal{I}_{n+1}}(\vec{x}) \right\rangle_\mu \right]^{\frac{1}{2}}$$

der Normierungsfaktor, der sich analog zu Gleichung (3.41) in Momenten der Zeitreihe ausdrücken läßt. Mit den neuen orthogonalen Basisfunktionen $\left\{\pi^{\mathcal{I}_\nu}(\vec{x})\right\}_{\nu=1,\dots,k}$ läßt sich nun die Approximation der Abbildung in Angriff nehmen.

3.5.3 Approximation

Hat man eine ONB nach einem der beiden Verfahren erzeugt, so ist das nächste Ziel nun eine Approximation der Abbildung

$$\begin{aligned} f : \mathbb{R}^d &\;\rightarrow\; \mathbb{R} \\ f(\vec{x}_n) &\;=\; x_{n+1} \end{aligned} \tag{3.42}$$

in der Form

$$\tilde{f}(\vec{x}) = \sum_{\nu=1}^{k} p^{(\nu)} \pi^{\mathcal{I}_\nu}(\vec{x}) \;, \tag{3.43}$$

wobei die Polynome der Orthonormalitätsbedingung

$$\left\langle \pi^{\mathcal{I}_\nu}(\vec{x}) \;\middle|\; \pi^{\mathcal{I}_\kappa}(\vec{x}) \right\rangle_\mu \;=\; \delta_{\nu\kappa}$$

genügen.
Der mittlere quadratische Fehler der Approximation wird definiert als

$$\chi^2 \;=\; \frac{1}{N} \sum_{i=1}^{N} \left[\sum_{\nu=1}^{k} p^{(\nu)} \pi^{\mathcal{I}_\nu}(\vec{x}_i) - x_{i+1} \right]^2 . \tag{3.44}$$

Die Koeffizienten der Entwicklung in Gleichung (3.43) erhält man durch die Projektion der Funktion $f(\vec{x})$ auf die Orthonormalbasis

$$p^{(\nu)} = \left\langle f(\vec{x}) \mid \pi^{\mathcal{I}_\nu}(\vec{x}) \right\rangle_\mu . \tag{3.45}$$

Zur weiteren Behandlung des Problems definiert man sich das *$\mathcal{I}$-te funktionale Moment* $\Gamma_\mathcal{I}$ der Zeitreihe in der Form

$$\begin{aligned} \Gamma_\mathcal{I} &:= \left\langle f(\vec{x}) \mid (\vec{x})^\mathcal{I} \right\rangle_\mu \\ &= \int\limits_A f(\vec{x})\ (\vec{x})^\mathcal{I} d\mu(\vec{x}) , \end{aligned} \tag{3.46}$$

welches sich wegen der Ergodizität des Systems zusammen mit Gleichung (3.42) schreiben läßt als

$$\Gamma_\mathcal{I} = \frac{1}{N}\sum_{i=1}^{N} x_{i+1}\ \vec{x}_i^\mathcal{I} \ . \tag{3.47}$$

Wird das $\mathcal{I}_\nu$-te Basispolynom gegeben durch

$$\pi^{\mathcal{I}_\nu}(\vec{x}) = \sum_{q=1}^{\nu} a_{(\nu,q)}\vec{x}^{\mathcal{I}_q} , \tag{3.48}$$

so erhält man für die Entwicklungskoeffizienten in Gleichung (3.43) die Darstellung

$$\begin{aligned} p^{(\nu)} &= \left\langle f(\vec{x}) \mid \pi^{\mathcal{I}_\nu}(\vec{x}) \right\rangle_\mu \\ &= \left\langle f(\vec{x}) \;\middle|\; \sum_{q=1}^{\nu} a_{(\nu,q)}\vec{x}^{\mathcal{I}_q} \right\rangle_\mu \\ &= \sum_{q=1}^{\nu} a_{(\nu,q)} \left\langle f(\vec{x}) \mid \vec{x}^{\mathcal{I}_q} \right\rangle_\mu \\ &= \sum_{q=1}^{\nu} a_{(\nu,q)}\ \Gamma_{\mathcal{I}_q} . \end{aligned} \tag{3.49}$$

Faßt man die obigen Betrachtungen zusammen, so läßt sich die Approximation der Abbildung $f(\vec{x})$ im wesentlichen auf zwei Schritte zurückführen:

1. Die Berechnung der Momente der Zeitreihe $\{M_{\mathcal{I}_\nu}\}_{\nu=1,\dots,2k}$ zur Entwicklung der orthogonalen Basis

2. Die Berechnung der funktionalen Momente der Zeitreihe $\{\Gamma_{\mathcal{I}_\nu}\}_{\nu=1,\dots,k}$ zur Bestimmung der Entwicklungskoeffizienten

Für hochdimensionale Systeme und hohe Grade der Approximation ist dieses Verfahren sehr rechenintensiv. Nach Gleichung (3.32) wächst die Anzahl k der Basispolynome g-ten Grades in d Dimensionen mit $k = (g+d)!/(g!\,d!)$. Somit hat man k Entwicklungskoeffizienten $\left\{p^{(\nu)}\right\}_{\nu=1,\dots,k}$ zu bestimmen und dadurch auch die $k(k+1)/2$ Polynomkoeffizienten $\left\{a_{(j,i)}\right\}_{1\leq i\leq j\leq k}$.
Nun betrachtet man die Änderung von χ^2 bei der Wegnahme der m-ten Basisfunktion:

$$
\begin{aligned}
\chi^2 &= \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{\nu=1}^{k} p^{(\nu)}\pi^{\mathcal{I}_\nu}(\vec{x}_i) - x_{i+1}\right]^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[p^{(m)}\pi^{\mathcal{I}_m}(\vec{x}_i) + \left(\sum_{\substack{\nu=1\\ \nu\neq m}}^{k} p^{(\nu)}\pi^{\mathcal{I}_\nu}(\vec{x}_i) - x_{i+1}\right)\right]^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\left(p^{(m)}\pi^{\mathcal{I}_m}(\vec{x}_i)\right)^2 - 2p^{(m)}\pi^{\mathcal{I}_m}(\vec{x}_i)\left(\sum_{\substack{\nu=1\\ \nu\neq m}}^{k} p^{(\nu)}\pi^{\mathcal{I}_\nu}(\vec{x}_i) - x_{i+1}\right)\right. \\
&\quad \left. + \left(\sum_{\substack{\nu=1\\ \nu\neq m}}^{k} p^{(\nu)}\pi^{\mathcal{I}_\nu}(\vec{x}_i) - x_{i+1}\right)^2\right] \\
&= \left(p^{(m)}\right)^2 \left\langle \pi^{\mathcal{I}_m}(\vec{x}) \mid \pi^{\mathcal{I}_m}(\vec{x})\right\rangle_\mu - 2p^{(m)}\sum_{\substack{\nu=1\\ \nu\neq m}}^{k} p^{(\nu)}\left\langle \pi^{\mathcal{I}_m}(\vec{x}) \mid \pi^{\mathcal{I}_\nu}(\vec{x})\right\rangle_\mu \\
&\quad -2p^{(m)}\left\langle \pi^{\mathcal{I}_m} \mid f(\vec{x})\right\rangle_\mu + \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{\substack{\nu=1\\ \nu\neq m}}^{k} p^{(\nu)}\pi^{\mathcal{I}_\nu}(\vec{x}_i) - x_{i+1}\right]^2 \\
&= \left(p^{(m)}\right)^2 - 2\left(p^{(m)}\right)^2 + \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{\substack{\nu=1\\ \nu\neq m}}^{k} p^{(\nu)}\pi^{\mathcal{I}_\nu}(\vec{x}_i) - x_{i+1}\right]^2 \\
&= -\left(p^{(m)}\right)^2 + \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{\substack{\nu=1\\ \nu\neq m}}^{k} p^{(\nu)}\pi^{\mathcal{I}_\nu}(\vec{x}_i) - x_{i+1}\right]^2 \qquad (3.50)
\end{aligned}
$$

Die Gleichung (3.50) zeigt, daß χ^2 monoton wächst, wenn man die Anzahl der Basispolynome erniedrigt. Die Summe in Gleichung (3.50) ist der mittlere quadratische Fehler der Approximation, wenn man das m-te Basispolynom aus dem Modell entfernt hat.

Bei den Umformungen hat man folgende Identitäten benutzt:

$$\begin{aligned} \left\langle \pi^{\mathcal{I}_m} \mid f(\vec{x}) \right\rangle_\mu &= p^{(m)} \\ \left\langle \pi^{\mathcal{I}_m}(\vec{x}) \mid \pi^{\mathcal{I}_m}(\vec{x}) \right\rangle_\mu &= 1\;. \end{aligned}$$

Entfernt man iterativ jeweils die Basisfunktion, für die $\left(p^{(m)}\right)^2$ am kleinsten ist, so wächst χ^2 nicht nur monoton, sondern $\chi^2(k)$ ist dann eine konvexe Funktion in k. Dies wird sich bei den in Kapitel 5 untersuchten Methoden der Modellvalidierung als großer Vorteil des MB-Modeling erweisen.

Kapitel 4

Vergleich der Modellierungsverfahren

4.1 Vergleich der benötigten Rechenzeiten

In diesem Abschnitt werden Gradientenverfahren, MB-Modeling und die Methode der Singulärwertzerlegung (SVD) im Hinblick auf die Laufzeit der Programme untersucht. Dabei werden Polynome als Basisfunktionen verwendet und die von den einzelnen Verfahren benötigte CPU-Zeit zum Auffinden der Modellparameter in Abhängigkeit von der Größe des Modells und der Länge der Zeitreihe miteinander verglichen.
Ein polynomiales Modell vom Grad g, welches die Dynamik in einem d-dimensionalen Rekonstruktionsraum beschreibt, besteht aus $k = \frac{(g+d)!}{g!\,d!}$ Basispolynomen.
Für das Gradientenverfahren und für die Methode der Singulärwertzerlegung werden die Basispolynome gegeben durch Monome in der Form $\vec{x}^{\mathcal{I}_\nu}$; $1 \leq \nu \leq k$, analog zu Definition (3.31).

Die Ausgangsbedingungen ergeben sich wie folgt:

- Die Matrix A, die der Singulärwertzerlegung unterzogen wird, hat k Spalten und N Zeilen

- Beim Gradientenverfahren findet die Suche nach dem Minimum der Fehlerfunktion in einem k-dimensionalen Raum statt. Die Länge der Zeitreihe geht im wesentlichen bei der Berechnung der Matrix Φ ein (siehe Gleichung (3.20))

- Beim MB-Modeling sind zur Berechnung der k Basispolynome insgesamt $d^{2(g+1)}$ Momente der Zeitreihe und $d^{(g+1)}$ funktionale Momente der Zeitreihe zu berechnen, wobei jedesmal über die gesammte Zeitreihe summiert werden muß

In den Abbildungen 4.1 und 4.2 sind die Ergebnisse des Zeitvergleichs dargestellt. Die Methode der SVD erweist sich dabei als das überlegene Verfahren, nur bei dem für die Anwendung wenig relevanten Fall $d = 1$ ist die Methode des MB-Modeling schneller. Die benötigte CPU-Zeit wächst bei allen drei Methoden im wesentlichen exponentiell mit der Einbettungsdimension und dem Approximationsgrad. Dies läßt sich aus dem näherungsweise linearen Anstieg der logarithmierten CPU-Zeiten für MB-Modeling und SVD erkennen. Die Anzahl der benötigten Rechenschritte ist bei beiden Verfahren durch die Ausgangsbedingungen festgelegt. Beim Gradientenverfahren ist dies nicht der Fall, da man nicht genau sagen kann, wieviele Iterationen bis zum Erreichen des Minimums durchgeführt werden müssen. Auf diese Weise erklären sich die Abweichungen der CPU-Zeit vom erwarteten Wert. Das Gradientenverfahren schneidet beim Zeitvergleich am ungünstigsten ab. Den vergleichsweise geringen Speicherbedarf des Verfahrens bezahlt man dabei mit längeren Rechenzeiten.

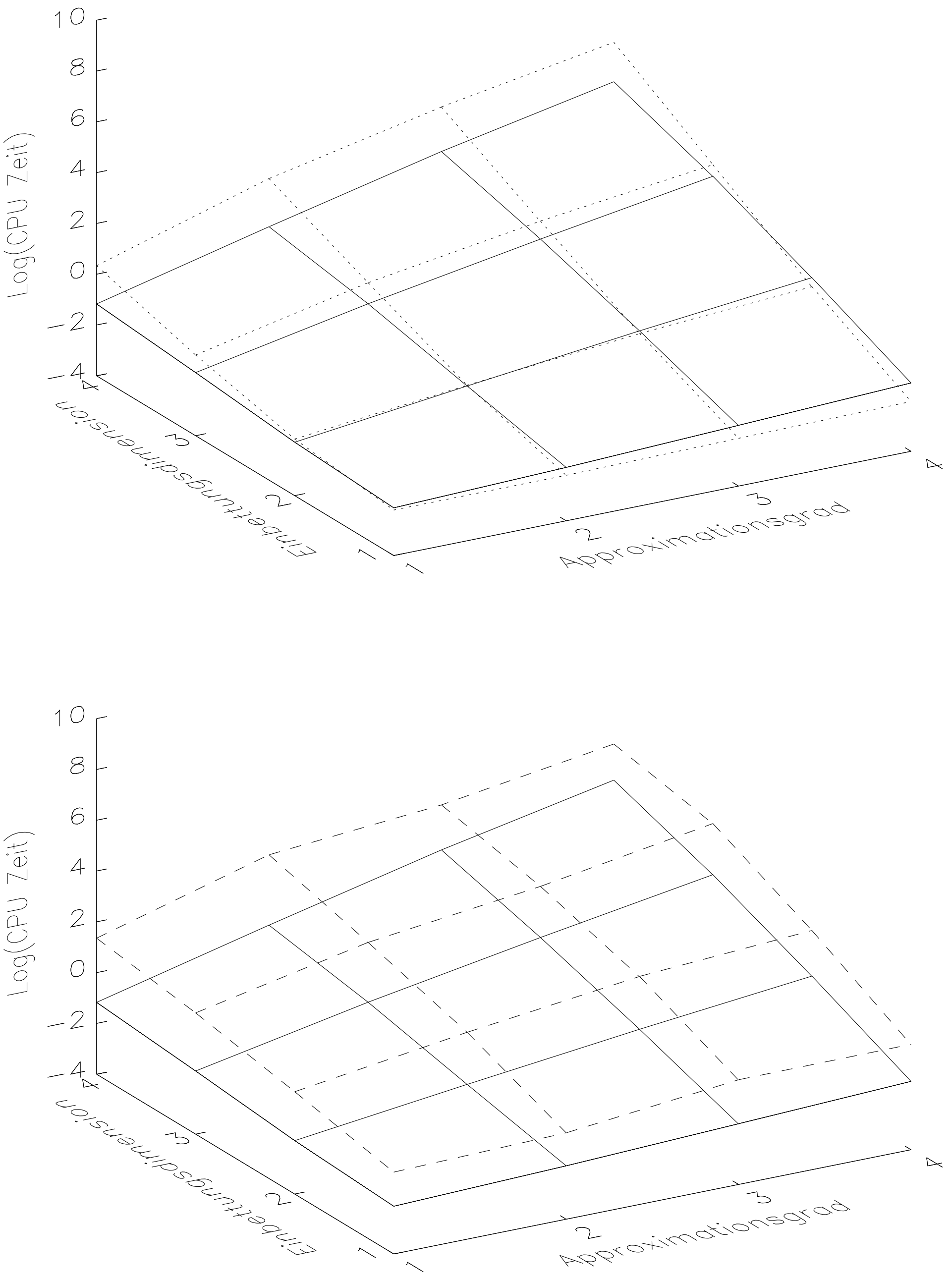


Abbildung 4.1: Vergleich der benötigten CPU-Zeiten zwischen dem MB-Modeling (···), dem Gradientenverfahren (- - -) und der SVD (—) als Funktion von Einbettungsdimension und Approximationsgrad. Die zugrundegelegte Zeitreihe entstammte einem Rössler-System und hatte eine Länge von 3000 Werten. Die Rechnungen wurden auf einer SUN Ultra-1-Workstation durchgeführt. Die CPU-Zeit wurde in Sekunden gemessen.

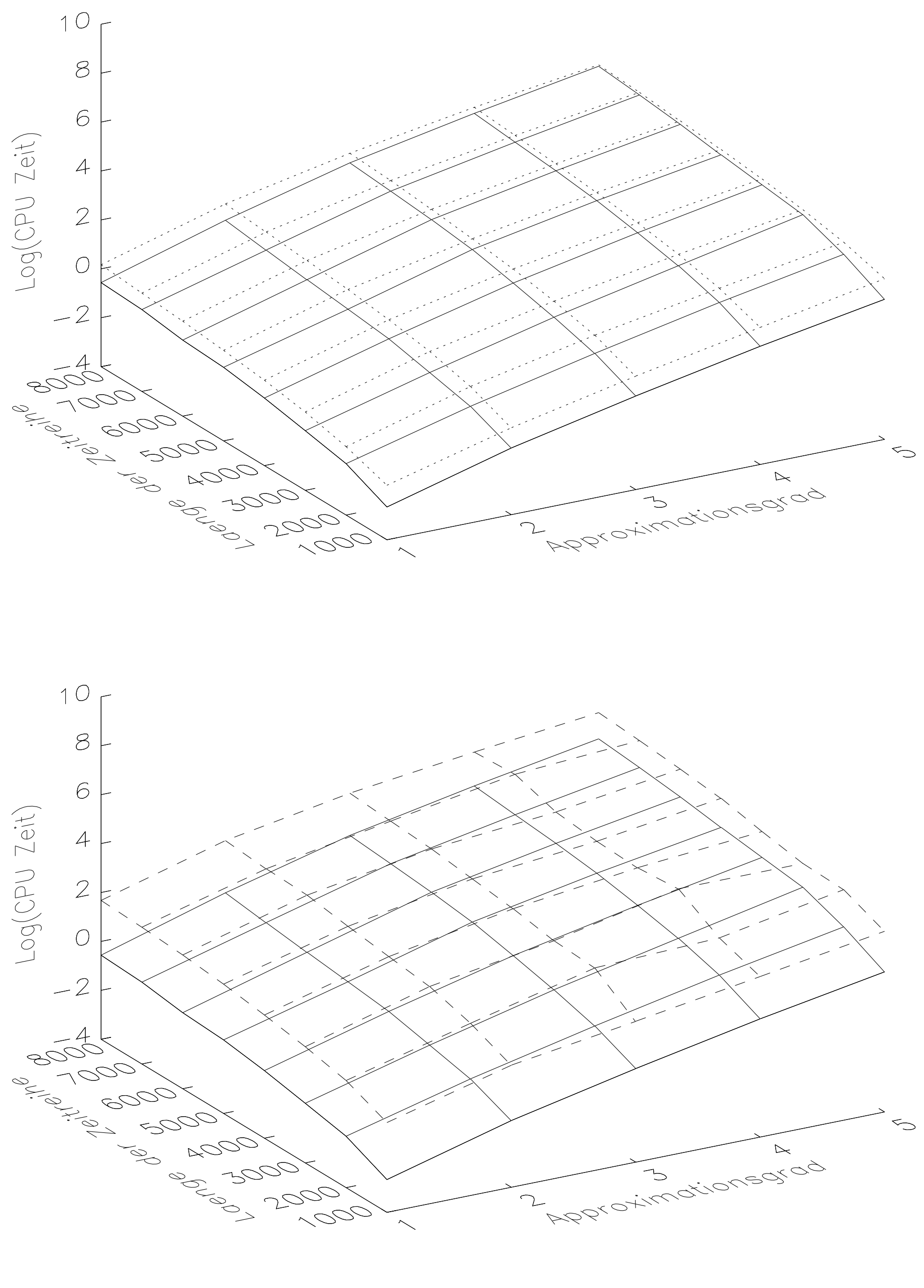


Abbildung 4.2: Vergleich der benötigten CPU Zeiten zwischen dem MB-Modeling (···), dem Gradientenverfahren (- - -) und der SVD (—–) als Funktion des Approximationsgrades und der Länge der Zeitreihe. Die zugrundegelegte Zeitreihe enstammte einem Rössler-System und wurde in den $\mathbb{R}^3$ eingebettet. Die Rechnungen wurden auf einer SUN Ultra-1-Workstation durchgeführt. Die CPU-Zeit wurde in Sekunden gemessen.

Kapitel 5

Selektion der Modellparameter

Zu Beginn dieses Kapitels sollen die Formeln, die zur Berechnung der MDL verwendet wurden, erläutert werden.
Dazu ist es notwendig, eine kurze Einführung in die Informationstheorie zu geben, um die grundlegenden Begriffe verwenden zu können. Danach soll die MDL nach Rissanen vorgestellt werden. Dabei wird im wesentlichen der Übersichtsartikel aus der *Encyclopedia of Statistical Sciences* zugrunde gelegt [44]. Danach werden wir die MDL in der Interpretation von K. Judd und A. Mees [27] vorstellen und auf die dort auftauchenden Probleme eingehen.

5.1 Grundlagen der Codierung

Zur Einleitung wird eine Definition des Codes gegeben, wie sie z.B. in Jablonski [26] zu finden ist.
Sei $\mathcal{X}$ eine endliche, nichtleere Menge und $\mathcal{A} = 0, 1, ..., k-1$ ein *Alphabet* mit $k \geq 2$. Ein *Code* ist eine injektive Abbildung von $\mathcal{X}$ in die Menge der nichtleeren endlichen Folgen von Elementen aus $\mathcal{A}$.
Ein Element aus $\mathcal{A}$ heißt *Buchstabe* des Alphabets, eine Folge von Buchstaben nennt man ein *Codewort* (bezeichnet mit a), dabei ist die Länge $L(a)$ des Codewortes durch die Anzahl seiner Buchstaben definiert.
Wenn kein Codewort das Anfangsstück eines anderen Codewortes ist, dann nennt man diesen Code einen *Präfixcode*.
Für einen Präfixcode mit den Wortlängen $n_1, n_2, \ldots, n_M$ zur Alphabetlänge k gilt die *Ungleichung von Kraft* :

$$\sum_{i=1}^{M} k^{-n_i} \leq 1 \; . \tag{5.1}$$

Kennt man die Wahrscheinlichkeitsverteilung $P(a)$ der Codewörter, die man zur Codierung eines Datensatzes benötigt, so kann man die mittlere Codewortlänge $\mathcal{L}(a)$ abschätzen durch

$$\mathcal{L}(a) = \sum_a P(a)L(a) \ . \tag{5.2}$$

Hat man zwei Wahrscheinlichkeitsverteilungen $P(x)$ und $Q(x)$ auf einer Menge $\mathcal{X}$ definiert, so gilt:

$$-\sum_{x\in\mathcal{X}} P(x)\log_2 P(x) \quad \leq \quad -\sum_{x\in\mathcal{X}} P(x)\log_2 Q(x) \ , \tag{5.3}$$

wobei Gleichheit genau im Fall $P = Q$ gilt [30].

Für das Wahrscheinlichkeitsmaß P heißt die Größe

$$H(P) \quad := \quad -\sum_{x\in\mathcal{X}} P(x)\log_2 P(x)$$

die *Entropie* von P.

Will man nun die mittlere Codewortlänge $\mathcal{L}(a)$ minimieren, so muß nach Gleichung (5.3) gelten

$$L(a) \quad = \quad -\log_2 P(a) \ . \tag{5.4}$$

Diesen Ausdruck nennt man die *ideale Codelänge* von a.
Ein zentrales Ergebnis der Informationstheorie ist der Quellencodierungssatz, der besagt, daß die untere Schranke für die mittlere Codewortlänge durch die Entropie gegeben ist [30].

5.2 Die MDL nach Rissanen

Im folgenden soll der Zusammenhang zwischen der MDL und der Maximum-Likelihood-Methode dargestellt werden.
Zu einem beobachteten Satz von Daten $x = \{x_1, \ldots, x_N\}$ stellt man ein Modell $F(x, \theta)$ zur Beschreibung dieser Daten auf und erhält durch die Betrachtung der Modellierungsfehler eine parametrisierte Wahrscheinlichkeitsverteilung $P_\theta(x)$, die jedem Parametersatz $\theta = (\theta_1, \ldots, \theta_k)$ des Modells die Wahrscheinlichkeit der Beobachtung von x zuordnet.

Nun stellt man folgendes Gedankenexperiment an: Angenommen, man kennt den „wahren" Parametersatz θ^* des Modells und hat alle möglichen beobachtbaren Datensätze der Länge N zur Verfügung. Will man den beobachteten Datensatz codieren und wählt dazu eine Codesequenz der Länge $-\log_2 P_\theta(x)$, dann gilt für die mittlere Codewortlänge die Ungleichung

$$-\sum P_{\theta^*}(x)\log_2 P_{\theta^*}(x) \quad \leq \quad -\sum P_{\theta^*}(x)\log_2 P_\theta(x) \ , \tag{5.5}$$

wobei über alle möglichen Datensätze summiert wird. Die ideale Codewortlänge erhält man für $\theta = \theta^*$. Um die mittlere Codewortlänge zu minimieren, versucht man, den Ausdruck

$$L(x|\theta) \quad := \quad -\log_2 P_\theta(x) \tag{5.6}$$

bezüglich θ zu minimieren, und dies ist dann auch der Maximum-Likelihood-Schätzer für θ [44].
Die ideale Codelänge $L(x, \theta)$ zur Codierung von Daten und Parametern wird nach Rissanen beschrieben durch den Ausdruck

$$L(x, \theta) \quad = \quad L(x|\theta) + L(\theta) \ , \tag{5.7}$$

wobei $L(\theta)$ die ideale Codewortlänge zur Codierung des Parametersatzes ist. Diese Größe wird nach Rissanen [44] in erster Näherung gegeben durch

$$L(\theta) \quad = \quad \frac{k}{2} \log_2(N) \ .$$

Die *Description-Length* $L(x)$ für Daten und Modellparameter ergibt sich dann zu

$$L(x) \quad = \quad -\log_2 P_\theta(x) + \frac{k}{2} \log_2(N) \ , \tag{5.8}$$

und die Werte für $\theta(x)$ und $k(x)$, die diesen Ausdruck minimieren, heißen die *MDL-Schätzer* für θ und k [44, 45].

5.3 Die MDL von Judd und Mees

Nun soll der Ansatz von K. Judd und A. Mees vorgestellt werden, der in dieser Arbeit Verwendung findet. Gleichzeitig sollen auch die Probleme dargestellt werden, die bei der Realisierung der MDL auftreten.

In diesem Ansatz wird ein Modell interpretiert als eine Möglichkeit, Daten zu komprimieren, um sie von einem (virtuellen) Sender zu einem (ebenfalls virtuellen) Empfänger zu übertragen. Ein Modell ist in diesem Sinne optimal, wenn die Länge des Codes zur Übertragung der modellierten Daten und der Modellparameter möglichst klein ist.
Die ideale Codewortlänge zur Codierung von Daten und Modellparametern ist wieder gegeben durch den Ausdruck

$$L(x, \theta) \quad = \quad L(x|\theta) + L(\theta) \ . \tag{5.9}$$

Jeder Modellparameter θ_j muß als binäre Gleitkommazahl im Computer dargestellt werden. Wenn der j-te Modellparameter mit der relativen Genauigkeit δ_j übertragen werden soll, so werden die ersten $\log_2(\delta_j)$ bit der

Mantisse von θ_j übertragen. Der auf diese Weise „verkürzte" Modellparameter wird mit $\bar{\theta}_j$ bezeichnet. Judd und Mees zeigten nun, daß die ideale Codewortlänge zur Codierung der Parameter $\bar{\theta}_j; \quad j = 1, \ldots, k$ näherungsweise gegeben wird durch

$$L(\bar{\theta}) \approx \sum_{j=1}^{k} \ln \frac{\gamma}{\delta_j} \quad , \tag{5.10}$$

dabei ist die Konstante γ festgelegt durch die Maschinengenauigkeit. Dabei wird der natürliche Logarithmus anstelle des Logarithmus zur Basis 2 verwendet. Die Einheit der Codewortlängen wird dann in sogenannten *nats* [1] angegeben [27].

Zur Bestimmung der idealen Codelänge betrachtet man die Verteilung der Modellierungsfehler (auch *Residuen* genannt). Die zu codierenden Daten seien gegeben durch N *Urbild-Bild-Paare* der Form $\{(\vec{x}_n, x_{n+1})\}_{n=1,\ldots,N}$, und ferner sei $F(\bar{\theta}, \vec{x}_n)$ das zugehörige Modell mit den Parametern $\bar{\theta}$.

Zur Vereinfachung der Schreibweise definiert man die Größe χ^2 als die Summe der quadrierten Modellierungsfehler

$$\chi^2 = \sum_{n=1}^{N} (F(\bar{\theta}, \vec{x}_n) - x_{n+1})^2. \tag{5.11}$$

Nimmt man an, daß die Residuen des Modells einer Normalverteilung $\mathcal{N}(0, \sigma^2)$ genügen, dann wird die Dichte $f(x|\bar{\theta})$ von $P_{\bar{\theta}}(x)$ gegeben durch

$$f(x|\bar{\theta}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^N \exp\left(-\frac{\chi^2}{2\sigma^2} \right) \quad , \tag{5.12}$$

wobei für den ML-Schätzer der Varianz $\hat{\sigma}^2$ gilt:

$$\hat{\sigma}^2 = \frac{\chi^2}{N} \; . \tag{5.13}$$

Der Ausdruck für $L(x|\bar{\theta})$ ergibt sich nach Gleichung (5.6) und der Definition der Wahrscheinlichkeitsdichte $f(x|\bar{\theta})$ zu

$$L(x|\bar{\theta}) = -\ln f(x|\bar{\theta}) \; . \tag{5.14}$$

Judd und Mees zeigten, daß sich der Ausdruck für die Description-Length $L_k(x)$ unter Verwendung von Gleichung (5.10) und Gleichung (5.14) als

$$\begin{aligned} L_k(x) = {} & \left(\frac{N}{2} - 1 \right) \ln \left(\frac{\chi^2}{N} \right) + (k+1) \left(\frac{1}{2} + \ln \gamma \right) \\ & - \sum_{j=1}^{k} \ln \delta_j + \frac{1}{2} \left[N + N \ln 2\pi + \ln \left(\frac{N}{2} \right) \right] \end{aligned} \tag{5.15}$$

schreiben läßt [27].

[1] Ein *nat* ist nach Mees das Analogon zur üblichen Einheit *bit* der Information. Die Einheiten unterscheiden sich nur durch einen positiven Faktor.

5.4 Die Probleme der MDL

Im folgenden soll auf die Probleme bei der Deutung und Anwendung der MDL eingegangen werden.

Zuerst stellt sich ein rein technisches Problem. Der Ausdruck in Gleichung (5.15) kann in manchen Fällen negative Werte annehmen, was in der Interpretation der Description-Length als ideale Codewortlänge keinen Sinn ergibt. Für $(\chi^2/N) \leq 1$ kann der Logarithmus des ersten Terms in Gleichung (5.15) beliebig große negative Werte annehmen. Je besser das Modell die Daten beschreibt (also je kleiner χ^2 wird), desto stärker fällt dieses Phänomen auf.
In Abbildung 5.1 ist dieses Verhalten für die Hénon-Abbildung (siehe Definition (5.20)) veranschaulicht worden.

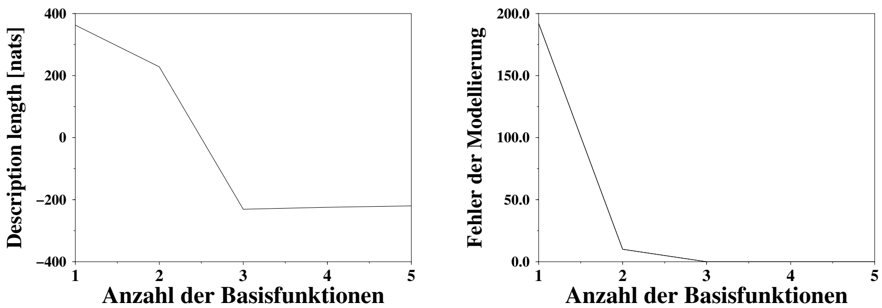


Abbildung 5.1: Ein Beispiel für negative Werte der Description-Length. Ein polynomiales Modell wurde für die Hénon-Abbildung erstellt. Die 200 Werte der Zeitreihe waren mit gaußverteiltem Rauschen unterlegt, das SNR betrug 60 dB.

Die Erklärung liegt in der Verwendung der Wahrscheinlichkeitsdichte $f(x|\bar{\theta})$ der Normalverteilung in Gleichung (5.14) anstelle einer diskreten Wahrscheinlichkeitsverteilung, die der Diskretisierung des Datensatzes Rechnung trägt, denn die Daten können nur mit einer begrenzten Auflösung Δx gemessen, bzw. im Computer dargestellt werden.
Für die Wahrscheinlichkeitsverteilung ergibt sich dann näherungsweise der Ausdruck

$$P(x|\bar{\theta}) \approx f(x|\bar{\theta})\,(\Delta x)^N \; . \tag{5.16}$$

Auf die ML-Schätzung der Parameter hat dies keinen wesentlichen Einfluß, da der Wert von $\bar{\theta}$, der $f(x|\bar{\theta})$ maximiert, in erster Näherung auch $P(x|\bar{\theta})$ maximieren wird [30].
Verwendet man die Wahrscheinlichkeitsverteilung aus Gleichung (5.16) zur

Berechnug der Description-Length, so erhält man eine zusätzliche additive Konstante, die an der Lage des Minimums jedoch nichts ändert [36].

Das zweite Problem bei der Interpretation der MDL betrifft das Modell selbst. Ein Einwand, der oft gegen die MDL ins Feld geführt wird, ist die Tatsache, daß die Struktur des Modells unberücksichtigt bleibt. So ist es z.B. möglich, daß ein sehr kompliziertes Modell[2] nur wenige Parameter benötigt, um die Daten sehr gut zu beschreiben, ein vergleichsweise „einfaches“ Modell jedoch mehr Parameter benötigt, um die Daten mit der gleichen Güte zu beschreiben.
Diesem Einwand kann man begegnen, indem man die MDL als ein Auswahlkriterium auffaßt, das innerhalb einer fest gewählten Modellklasse das optimale Modell bestimmt, d.h., wenn man sich auf eine Menge von Basisfunktionen festgelegt hat, dann liefert die MDL ein Kriterium zur Auswahl der entsprechenden Basisfunktionen. Dadurch verliert die MDL an Allgemeingültigkeit, denn es ist a priori nicht möglich, auf der Grundlage der MDL grundsätzlich verschiedene Modelle miteinander zu vergleichen. In einer späteren Arbeit deuten Judd und Mees dieses Problem an [37].

Ein weiteres Problem ist die Frage, welche zusätzlichen Informationen neben den Parametern des Modells übertragen werden müssen, um das zugehörige Modell zu spezifizieren. Judd und Mees versuchen dieses Problem zu lösen, indem sie eine „Markierung“ übertragen, die das benutzte Modell kennzeichnet. Dieser Ansatz kann zu einigen Schwierigkeiten führen, die an zwei Beispielen veranschaulicht werden sollen.
Wählt man ein Modell auf der Grundlage von radialen Basisfunktionen und überträgt nur die Koeffizienten, so fehlt eine wesentliche Information, nämlich die jeweiligen Stützstellen der Basisfunktionen. Diese Information ist nicht a priori gegeben, da sie im zu übertragenden Datensatz enthalten ist, denn die Stützstellen wählt man sinnvollerweise aus dem Datensatz aus (siehe Definition (3.6)). Eine Markierung, die im Vorfeld zwischen Sender und Empfänger vereinbart worden ist, sollte diese Information nicht enthalten, denn sie setzt bereits die Kenntnis des Datensatzes voraus.
Strenggenommen muß man noch den Aufwand zur Codierung der Stützstellen in Rechnung stellen, da diese eigentlich auch Modellparameter sind. Auf diese Weise würden sich die „Modellkosten“ vervielfachen, bei einer d-dimensionalen Approximation schätzungsweise um den Faktor $(d+1)$.

Bei polynomialen Modellen verhält es sich ähnlich. Überträgt man nur die Koeffizienten der Basispolynome, so müssen diese Koeffizienten noch den

[2] Die „Komplexität“ des Modells könnte man in diesem Fall als die *Kolmogorov-Chaitin-Komplexität* des erstellten Computerprogramms definieren. Leider gibt es im allgemeinen keinen Algorithmus, der über die *KC-Komplexität* dieses Programms entscheiden könnte [11, 47].

richtigen Basispolynomen zugeordnet werden. Auch in diesem Fall ist die „Markierung“ des Modells ein Problem, wie folgende Überlegung zeigt:
Im allgemeinen Fall wären der Einbettungsdimension und dem Approximationsgrad keine oberen Grenzen gesetzt, die Markierungen könnten dann beliebig lang werden. Wir wollen deshalb einen Fall betrachten, der noch auf einem Computer zu realisieren ist, dabei sollen Polynome 10-ten Grades in einem 10-dimensionalen Rekonstruktionsraum möglich sein. Nach Gleichung (3.32) gibt es somit

$$k \quad = \tfrac{(10+10)!}{10!\,10!} = \quad 184756$$

verschiedene Basispolynome. Dann hat man durch die Kombination dieser Basispolynome

$$\sum_{n=1}^{k} \binom{k}{n} \quad = \quad 2^k - 1$$

verschiedene Modelle zur Auswahl, die mit einer Markierung versehen werden müssen. Die Übertragung dieser Markierung ist offensichtlich keine Nebensächlichkeit mehr.
Um die eindeutige Zuordnung von Koeffizienten und Basisfunktionen zu gewährleisten, sollte jedem Koeffizienten noch der Exponent des entsprechenden Monoms mitgeliefert werden.

Ein großes technisches Problem bei der Anwendung der MDL ist die Frage nach der Eindeutigkeit des Minimums. In dieser Arbeit soll ein hinreichendes Kriterium für die Eindeutigkeit der MDL angegeben werden.
Dazu faßt man in Gleichung (5.15) die konstanten Terme zusammen und erhält für die Description-Length den Ausdruck

$$L_k(x) \quad = \quad \alpha \ln(\chi^2) + \beta k + \omega \ , \tag{5.17}$$

dabei sind α, β und ω Konstanten. Der Term βk wächst monoton in k und repräsentiert die ideale Codewortlänge zur Codierung der Parameter. Der Term $\alpha \ln(\chi^2)$ steht für die ideale Codewortlänge zur Codierung der Daten. Wenn χ^2 als Funktion von k monoton fällt und konvex ist, dann hat die Description-Length ein eindeutiges Minimum. In Abbildung 5.2 ist dieses Verhalten schematisch dargestellt.
Ein hinreichendes Kriterium für die Monotonie von χ^2 ist die Wahl einer orthonormalen Funktionenbasis bei der Approximation des Flusses. Im folgenden seien die Matrix $\mathbf{A}$, der Koeffizientenvektor $\vec{a}$ und der Bildvektor $\vec{y}$

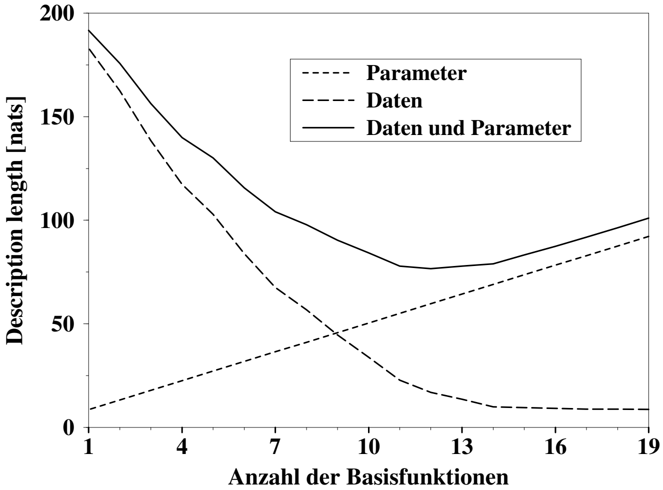


Abbildung 5.2: Schematische Darstellung der Description-Length $L_k(x)$, aufgeschlüsselt in den Anteil der Parameter und den Anteil der Daten, siehe Gleichung (5.17).

definiert durch

$$\mathbf{A} \;:=\; \begin{pmatrix} \phi_1(\vec{x}_1) & \phi_2(\vec{x}_1) & \cdots & \phi_k(\vec{x}_1) \\ \phi_1(\vec{x}_2) & \phi_2(\vec{x}_2) & \cdots & \phi_k(\vec{x}_2) \\ & \ddots & & \\ & & \ddots & \\ \phi_1(\vec{x}_N) & \phi_2(\vec{x}_N) & \cdots & \phi_k(\vec{x}_N) \end{pmatrix}$$

$$\vec{a} \;:=\; (a_1, a_2, \ldots, a_k)^\dagger$$

$$\vec{y} \;:=\; (x_{n+1}, x_{n+2}, \ldots, x_{N+1})^\dagger \;.$$

Mit $\phi_m(\vec{x})$ wird die m-te Zeile von $\mathbf{A}$ bezeichnet. Für die Basisfunktionen soll die Orthonormalitätsbedingung

$$\langle \phi_i(\vec{x}) \mid \phi_j(\vec{x}) \rangle \;:=\; \sum_{n=1}^{N} \phi_i(\vec{x}_n)\phi_j(\vec{x}_n) \;=\; \delta_{i,j} \tag{5.18}$$

gelten. Durch die *Normalengleichungen* ist dann die Beziehung

$$a_m \;=\; \langle \phi_m(\vec{x}) \mid \vec{y}\, \rangle$$

gegeben [40].

Mit diesen Bezeichnungen läßt sich χ^2 schreiben als

$$\begin{aligned}\chi^2 &= \| \mathbf{A}\vec{a} - \vec{y} \|^2 \\ &= \left\| \sum_{i=1}^{k} a_i \phi_i(\vec{x}) - \vec{y} \right\|^2 .\end{aligned}$$

Nun betrachtet man die Änderung von χ^2, wenn die m-te Basisfunktion entfernt wird:

$$\begin{aligned}
\chi^2 &= \left\| \sum_{i=1}^{k} a_i \phi_i(\vec{x}) - \vec{y} \right\|^2 \\
&= \left\| a_m \phi_m(\vec{x}) + \sum_{\substack{i=1 \\ i\neq m}}^{k} a_i \phi_i(\vec{x}) - \vec{y} \right\|^2 \\
&= \left\| a_m \phi_m(\vec{x}) + \sum_{\substack{i=1 \\ i\neq m}}^{k} a_i \phi_i(\vec{x}) - \vec{y} \right\|^2 \\
&= \left\langle a_m \phi_m(\vec{x}) + \left(\sum_{\substack{i=1 \\ i\neq m}}^{k} a_i \phi_i(\vec{x}) - \vec{y} \right) \middle| \, a_m \phi_m(\vec{x}) + \left(\sum_{\substack{i=1 \\ i\neq m}}^{k} a_i \phi_i(\vec{x}) - \vec{y} \right) \right\rangle \\
&= a_m^2 + 2a_m \left\langle \phi_m(\vec{x}) \middle| \left(\sum_{\substack{i=1 \\ i\neq m}}^{k} a_i \phi_i(\vec{x}) - \vec{y} \right) \right\rangle + \left\| \sum_{\substack{i=1 \\ i\neq m}}^{k} a_i \phi_i(\vec{x}) - \vec{y} \right\|^2 \\
&= a_m^2 - 2a_m \langle \phi_m(\vec{x}) \mid \vec{y} \rangle + \left\| \sum_{\substack{i=1 \\ i\neq m}}^{k} a_i \phi_i(\vec{x}) - \vec{y} \right\|^2 \\
&= a_m^2 - 2a_m^2 + \left\| \sum_{\substack{i=1 \\ i\neq m}}^{k} a_i \phi_i(\vec{x}) - \vec{y} \right\|^2 \\
&= -a_m^2 + \left\| \sum_{\substack{i=1 \\ i\neq m}}^{k} a_i \phi_i(\vec{x}) - \vec{y} \right\|^2 \qquad (5.19)
\end{aligned}$$

Durch Gleichung (5.19) ist nun sichergestellt, daß χ^2 monoton wächst, wenn die m-te Basisfunktion entfernt wird. Entfernt man jeweils die Basisfunktion, für die χ^2 am kleinsten wird, dann ist χ^2 eine konvexe Funktion in m, und das Minimum der Description-Length ist eindeutig bestimmt.
Zusammenfassend ergeben sich bei der MDL also folgende Schwierigkeiten:

- Die MDL ist nur bis auf eine additive Konstante eindeutig bestimmt

- Die MDL läßt die Struktur des Modells unberücksichtigt und eignet sich deshalb nur als Kriterium innerhalb einer fest gewählten Modellklasse
- Es ist nicht klar, welcher Aufwand betrieben werden muß, um das ausgewählte Modell eindeutig zu kennzeichnen
- Die MDL ist nicht für alle Modelle eindeutig bestimmt. Ein hinreichendes Kriterium für die Eindeutigkeit ist die Wahl einer orthonormalen Funktionenbasis

5.5 Optimierung des Modells

In diesem Abschnitt sollen die beiden vorgestellten Verfahren zur Bestimmung der optimalen Modellgröße (MDL und Cross-Validation) an verschiedenen Datensätzen getestet werden.

Zur Berechnung der Description-Length wurde Gleichung (5.15) verwendet:

$$\begin{aligned} L_k(x) &= \left(\frac{N}{2}-1\right)\ln\left(\frac{\chi^2}{N}\right)+(k+1)\left(\frac{1}{2}+\ln\gamma\right) \\ &\quad -\sum_{j=1}^{k}\ln\delta_j+\frac{1}{2}\left[N+N\ln 2\pi+\ln\left(\frac{N}{2}\right)\right] . \end{aligned}$$

Für die Konstante γ wurde der von Judd und Mees vorgeschlagene Wert $\gamma = 32$ gewählt. Ferner wurde angenommen, daß alle Modellparameter mit der gleichen relativen Genauigkeit $\delta_i = 16;\ i = 1, \ldots, k$ übertragen werden sollen.

Bei der Durchführung der Cross-Validation wurde der Datensatz in zwei Hälften geteilt. Die erste Hälfte der Daten, die sogenannten *Trainingsdaten*, wurden der Modellierung zugrundegelegt. Die zweite Hälfte der Daten, die *Testdaten* wurden beim Erstellen des Modells unberücksichtigt gelassen. An den Testdaten wurde danach der mittlere quadratische 1-Schritt-Vorhersagefehler berechnet.

Es wurden drei verschieden Modelle verwendet:

- Glockenförmige radiale Basisfunktionen gemäß Definition (3.7), für die Konstante β wurde die Varianz der Zeitreihe gewählt. Die Koeffizienten wurden mit der Methode der SVD aus Abschnitt 3.3 berechnet
- Polynomiale Basisfukntionen, ebenfalls via SVD erstellt

- Ein Modell mit orthonormalen Polynomen als Basisfunktionen, die auf der Grundlage des MB-Modeling berechnet wurden (siehe Abschnitt 3.5)

Bei der Auswahl der Basisfunktionen wurde mit dem „ganzen Modell“ gestartet und dann folgende Schritte durchgeführt:

1. Es wurde jeweils eine Basisfunktion aus dem Modell entfernt und für das „reduzierte Modell“ χ^2 berechnet.

2. Die Basisfunktion, für die χ^2 den kleinsten Wert annahm, wurde aus dem Modell endgültig entfernt.

3. Mit dem neuen Modell wurde die Berechnung der Description-Length durchgeführt, bzw. es wurde der mittlere quadratische 1-Schritt-Vorhersagefehler am Testdatensatz berechnet.

4. Die Prozedur 1.-3. wurde mit dem neuen Modell wiederholt.

Dieses Verfahren funktioniert auch in umgekehrter Richtung, wenn man von einem „leeren“ Modell startet und dann nach und nach neue Basisfunktionen hinzunimmt, um χ^2 zu verringern [35].

5.5.1 NH_3-FIR-Laser

Der untersuchte Datensatz wurde an einem Far-infrared (FIR) Laser gemessen. Eine ausführliche Beschreibung des Versuchsaufbaus und der theoretischen Grundlagen findet sich in dem Artikel *„Lorenz-Like Chaos in NH_3-FIR Lasers“* von Hübner et al. in Referenz [25].
Die Zeitreihe hatte eine Länge von 1000 Werten und wurde in den $\mathbb{R}^3$ eingebettet. Es wurden $k = 38$ radiale Basisfunktionen zur Approximation verwendet. Bei den beiden polynomialen Modellen wurde eine Approximation 2-ten Grades gewählt.

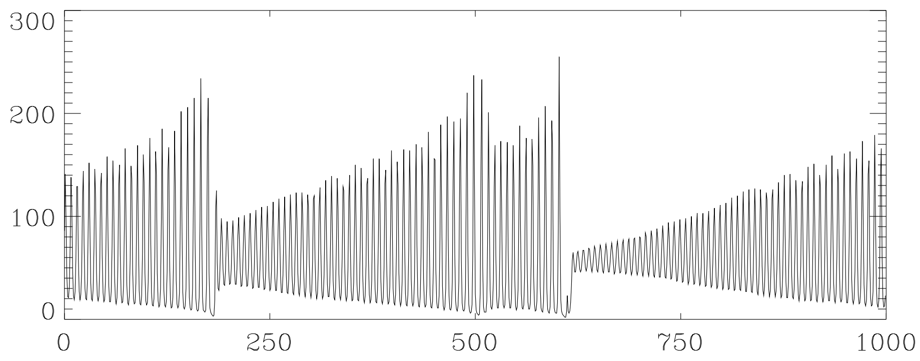


Abbildung 5.3: Zeitreihe des NH_3-FIR-Lasers.

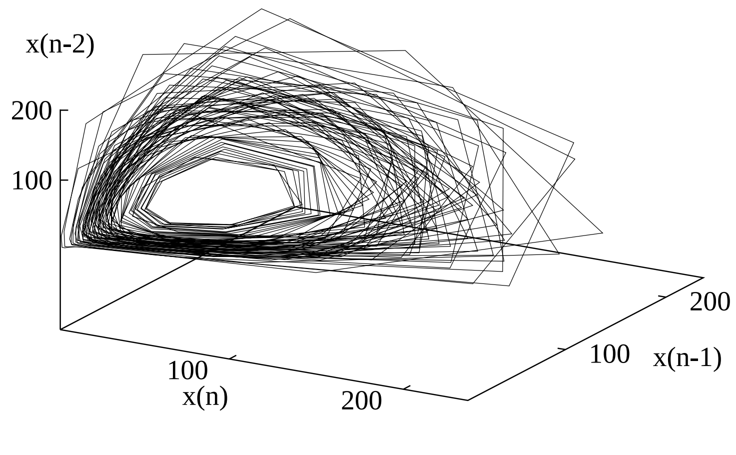
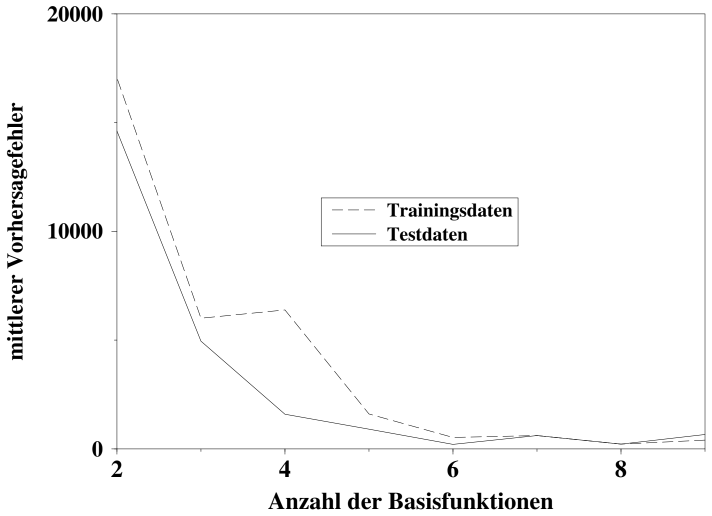


Abbildung 5.4: Einbettung der Zeitreihe des NH_3-FIR-Lasers in den $\mathbb{R}^3$ mit der Delayzeit $\tau = 1$.



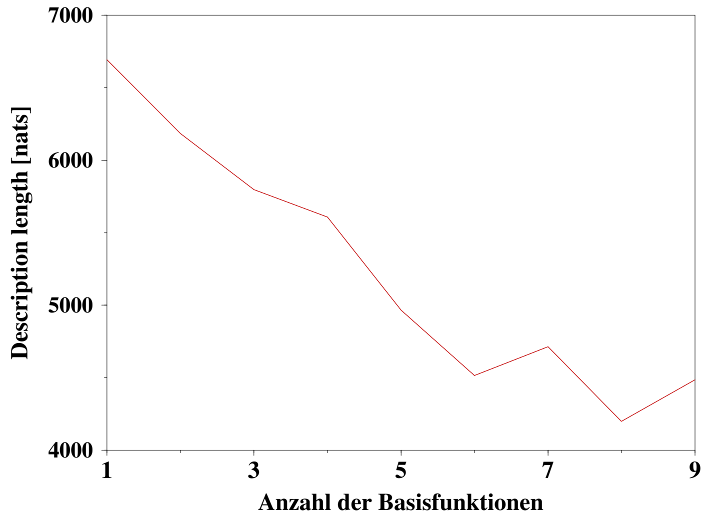


Abbildung 5.5: Cross-Validation (oben) und MDL (unten) für ein polynomiales Modell der NH_3-FIR-Laser-Daten, welches mit dem SVD-Verfahren erstellt wurde.

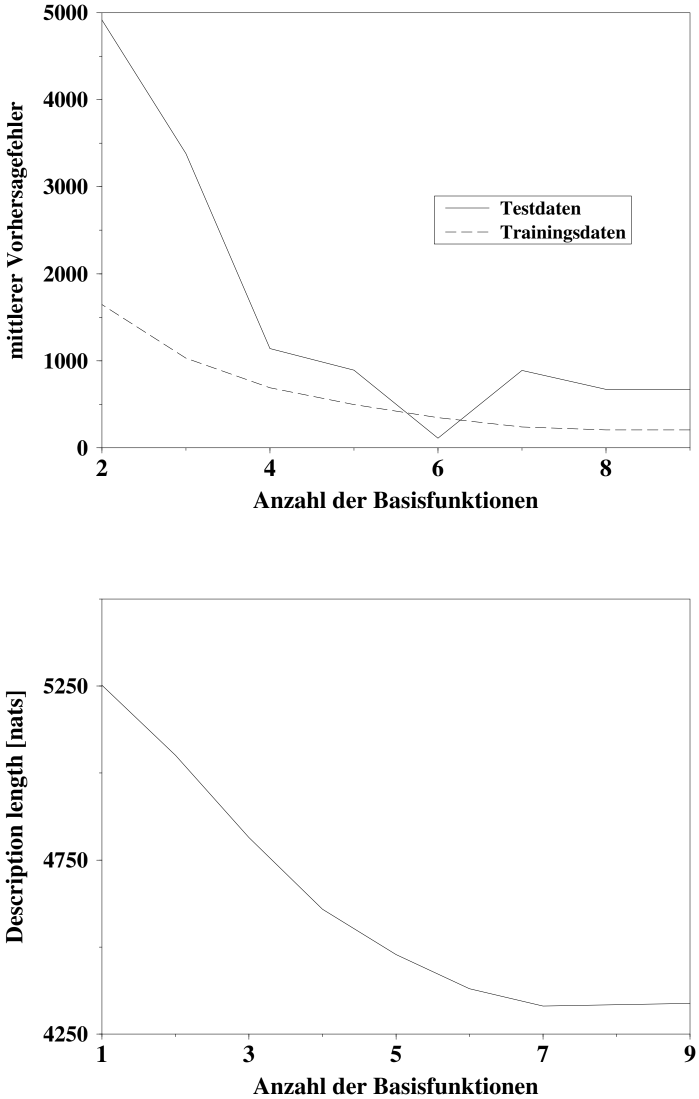


Abbildung 5.6: Cross-Validation (oben) und MDL (unten) für ein Modell der NH_3-FIR-Laser-Daten auf der Grundlage von orthonormalen Polynomen. Die Description-Length hat ihr Minimum bei $k = 7$. Die Cross-Validation zeigt das optimale Modell bereits für $k = 6$ an.

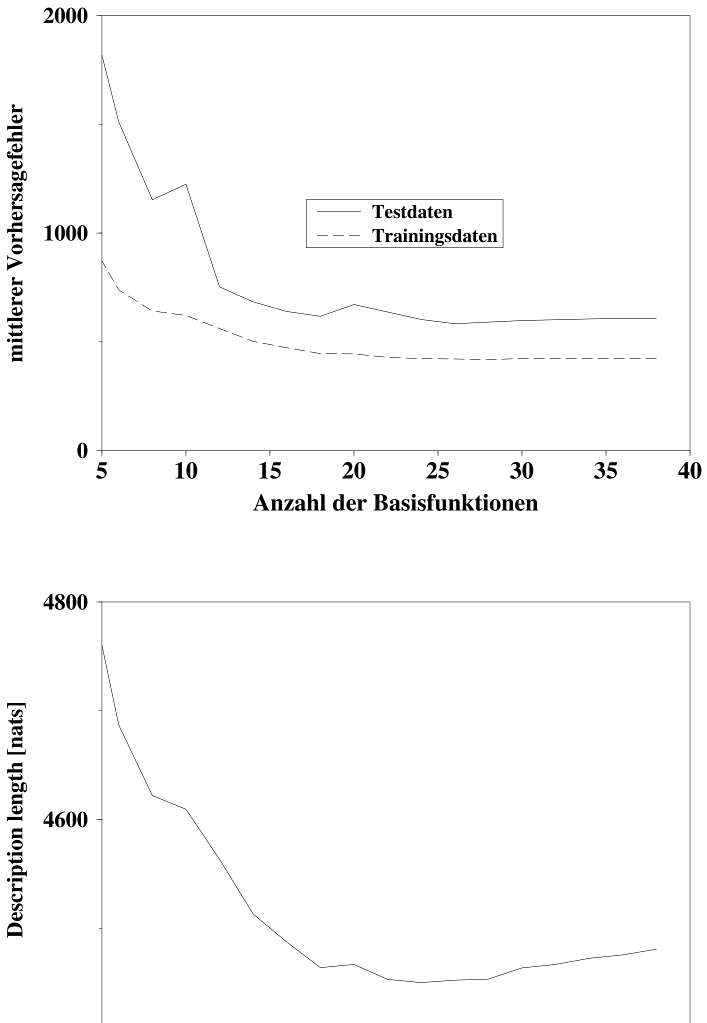


Abbildung 5.7: Cross-Validation (oben) und MDL (unten) für ein Modell der NH_3-FIR-Laser-Daten mit radialen Basisfunktionen. Die MDL erhält man bei $k = 24$, die Cross-Validation liefert das optimale Modell für $k = 26$.

Bei der Betrachtung der Abbildung 5.6 bemerkt man, daß der mittlere Vorhersagefehler der Trainingsdaten beim MB-Modeling monoton steigt, wenn man die Anzahl der Basisfunktionen verringert. Das entspricht dem erwarteten Ergebnis. Im Gegensatz dazu schwankt der mittlere Vorhersagefehler bei den einfachen polynomialen Basisfunktionen in Abbildung 5.5. Die Description-Length nimmt zwei Minima an, ebenso wie der mittlere Vorhersagefehler bei der Cross-Validation. Das optimale Modell läßt sich nicht mehr eindeutig bestimmen. Man hat die Wahl zwischen einem Modell mit sechs und einem Modell mit acht Basisfunktionen.

Bei dem Modell auf Grundlage der radialen Basisfunktionen fällt der Vorhersagefehler für die Trainingsdaten monoton, wie man in Abbildung 5.7 sieht. Die Description-Length hat ein globales Minimum für $k = 24$ und ein lokales Minimum für $k = 17$. Der mittlere Vorhersagefehler der Cross-Validation erreicht sein Minimum für $k = 26$.

Am Beispiel der NH_3-FIR-Laser-Daten zeigt sich, daß nur die orthonormalen Polynome ein eindeutiges globales Minimum der Description-Length liefern. Die Methode der Cross-Validation liefert in diesem Fall ein kleineres Modell mit $k = 6$ Basisfunktionen.

In Abbildung 5.8 ist eine Vorhersage über 50 Werte der Zeitreihe dargestellt, die mit dem optimalen Modell orthonormaler Polynome im Sinne der MDL gemacht wurde. Dabei wurde jeder vorhergesagte Wert zur Originalzeitreihe hinzugefügt, so daß er bei der nächsten Vorhersage als neues Urbild zur Verfügung stand. Der Vergleich mit der Vorhersage eines lokal konstanten Modells in Abbildung 5.9 zeigt, daß das polynomiale Modell keine Verschiebung gegenüber der Originalzeitreihe zeigt[3]. Das zur Vorhersage verwendete Polynom $P(x_n, x_{n-1}, x_{n-2})$ hatte folgende Gestalt:

$$\begin{aligned} P(x_n, x_{n-1}, x_{n-2}) &= 38.7844 - 0.9221 x_n + 0.1615 x_{n-1} + 1.6588 x_{n-2} \\ &\quad + 0.0050 x_n^2 - 0.0075 x_n x_{n-1} + 0.0068 x_{n-1}^2 \\ &\quad + 0.00426 x_n x_{n-2} - 0.0165 x_{n-1} x_{n-2} \end{aligned}$$

[3] Diese Verschiebung gegenüber der Originalzeitreihe ist ein Artefakt des lokalen Modells. Der vorhergesagte Wert bewegt sich mit seinem nächsten Nachbarn durch den Rekonstruktionsraum. Die Verschiebung ist proportional zum Abstand des vorhergesagten Wertes zum nächsten Nachbarn.

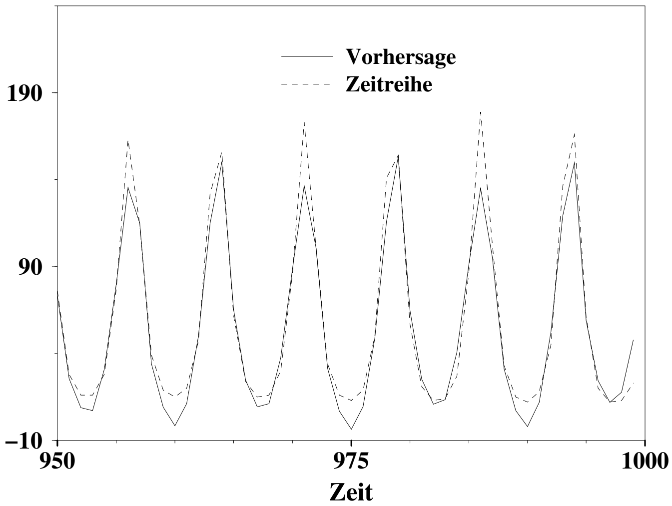


Abbildung 5.8: Vorhersage der Zeitreihe des NH_3-FIR-Lasers mit dem optimierten Modell aus orthonormalen Polynomen. Es wurden 7 Basispolynome verwendet, entsprechend dem MDL-Kriterium (siehe Abb. 5.6).

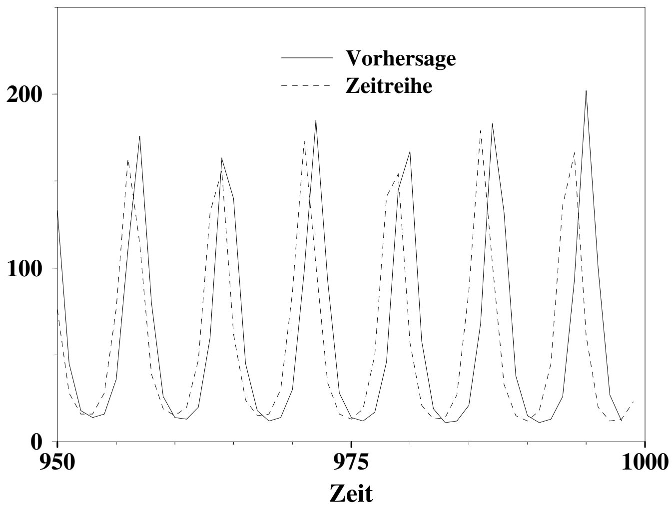


Abbildung 5.9: Vorhersage der Zeitreihe des NH_3-FIR-Lasers mit einem lokal konstanten Modell und einem nächsten Nachbarn.

5.5.2 Der Rössler-Oszillator

Der Rössler-Oszillator wird beschrieben durch ein System gekoppelter nichtlinearer Differentialgleichungen. In dem hier betrachteten Fall liegt folgendes Differentialgleichungssystem zugrunde:

$$\begin{aligned} \dot{x}_1 &= 2 + x_1(x_2 - 4) \\ \dot{x}_2 &= -x_1 - x_3 \\ \dot{x}_3 &= x_2 + 0.45x_3 \end{aligned}$$

Diese Differentialgleichung wurde auf einem Analogrechner RAT 700 von Telefunken integriert. Die Zeitreihen wurden von einer Workstation mit eingebautem A/D-Wandler mit 16 Bit bei 1KHz abgetastet [28]. Die Zeitreihe der x_2-Komponente des Systems wurde in den $\mathbb{R}^3$ eingebettet mit der Delayzeit $\tau = 15$. Das Ergebnis dieser Einbettung ist in Abb. 5.10 dargestellt. Es wurden $k = 37$ radiale Basisfunktionen zur Approximation verwendet. Bei den beiden polynomialen Modellen wurde eine Approximation 3-ten Grades gewählt.
Die Abbildungen 5.11, 5.12 und 5.13 zeigen die Überlegenheit der orthonormalen Polynome bezüglich der Möglichkeit einer Parameterselektion. Sowohl für die Cross-Validation als auch für die MDL existiert ein eindeutiges globales Minimum. Fatalerweise sind diese Minima verschieden. Um die beiden Verfahren gegeneinander zu testen, wurde die Vorhersagefähigkeit der durch diese Verfahren optimierten Modelle verglichen, indem einhundert Werte der Zeitreihe vorausgesagt wurden. Das Ergebnis dieser Vorhersage ist in den Abbildungen 5.14 und 5.15 dargestellt. Das Modell, das auf der Grundlage der MDL erstellt wurde, zeigt sich dabei als überlegen.

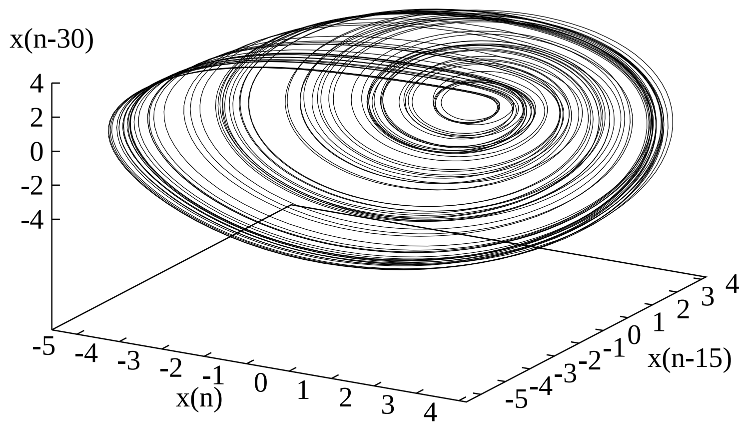


Abbildung 5.10: Einbettung der Zeitreihe des Rössler-Oszillators in den $\mathbb{R}^3$ mit der Delayzeit $\tau = 15$.

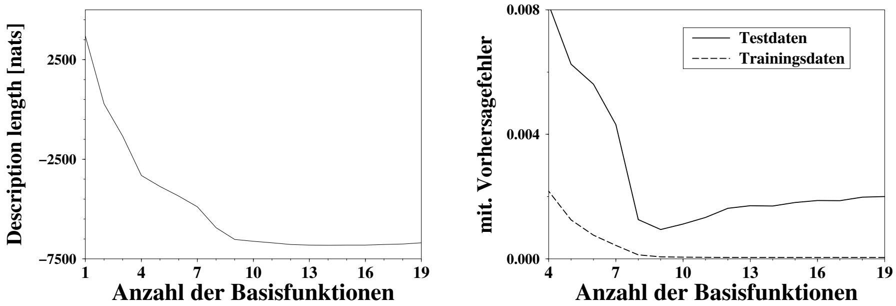


Abbildung 5.11: Description-Length und Cross-Validation für die orthonormalen Polynome. Der Vorhersagefehler der Trainingsdaten fällt monoton mit k. Das Minimum des Vorhersagefehlers der Testdaten liegt bei $k = 9$. Das Minimum der Description-Length liegt bei $k = 16$.

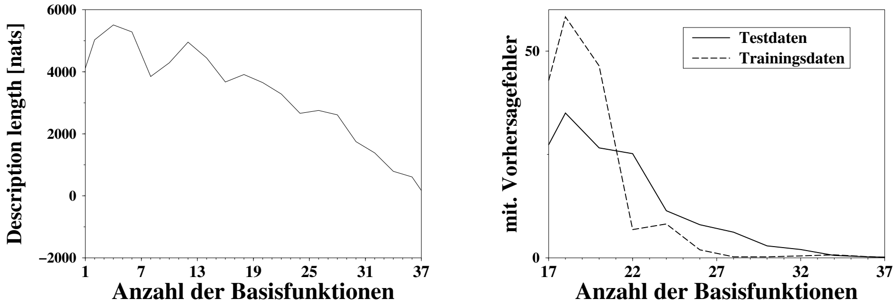


Abbildung 5.12: Description-Length und Cross-Validation für die radialen Basisfunktionen. Keines der beiden Kriterien gibt eine Antwort auf die Frage nach dem optimalen Modell.

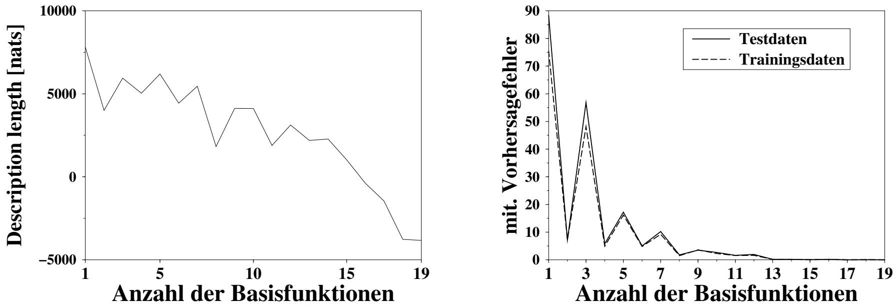


Abbildung 5.13: Description-Length und Cross-Validation für das polynomiale Modell, das mit dem SVD-Verfahren erstellt wurde. Hier ist ebenfalls keine Aussage mehr über ein optimales Modell im Sinne dieser beiden Kriterien zu machen.

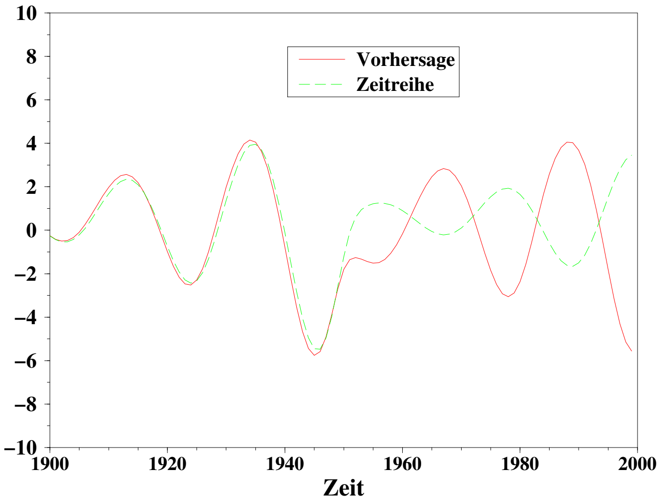


Abbildung 5.14: Vorhersage der Zeitreihe des Rössler-Oszillators mit dem nach der Cross-Validation optimierten Modell aus aus orthonormalen Polynomen. Es wurden 9 Basispolynome verwendet.

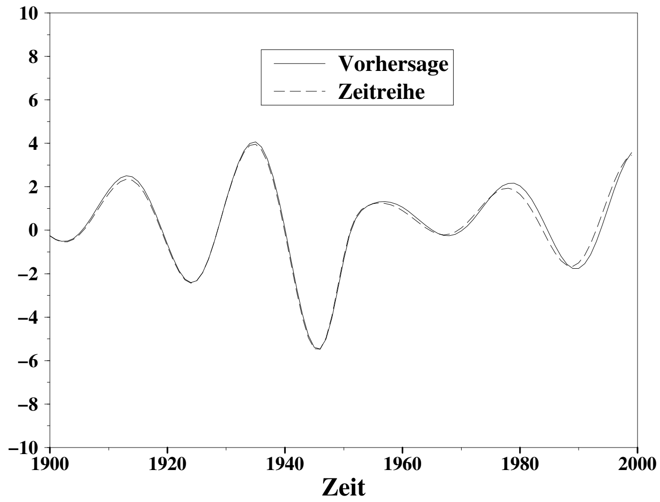


Abbildung 5.15: Vorhersage der Zeitreihe des Rössler-Oszillators mit dem nach der MDL optimierten Modell aus orthonormalen Polynomen. Es wurden 16 Basispolynome verwendet. Die Vorhersage ist deutlich besser als im Beispiel von Abbildung 5.14.

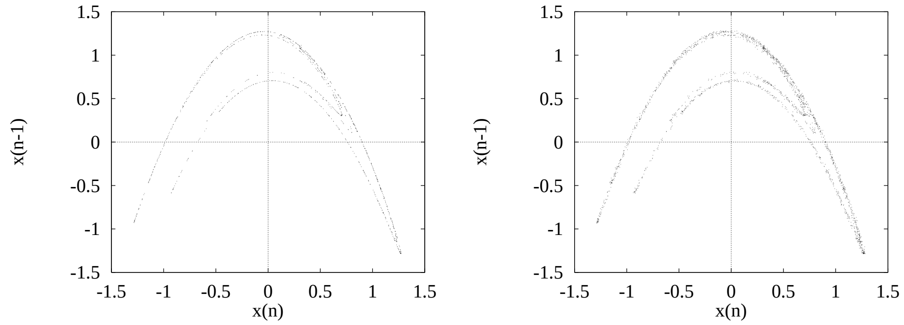


Abbildung 5.16: Die Delay-Einbettung mit der Delay-Zeit $\tau = 1$ für die Hénon-Abbildung im ungestörten Fall (links) und verrauscht (SNR: 40 dB rechts).

5.6 Der Einfluß von Rauschen auf die MDL

In diesem Abschnitt wird die MDL unter Einfluß von Rauschen untersucht. Dazu wurde die Zeitreihe des Hénon-Systems mit gaußverteiltem Rauschen überlagert und ein Modell mit polynomialen Basisfunktionen für die verrauschten Zeitreihen erstellt. Die Zeitreihe hatte in allen Fällen eine Länge von 200 Werten.

Die Hénon-Abbildung ist definiert durch die rekursiven Gleichungen

$$x_{n+1} \quad = \quad 1 - a x_n^2 + b x_{n-1} \tag{5.20}$$

mit den Parametern $a = 1.4$ und $b = 0.3$.
Zu der Zeitreihe wurde gaußverteiltes Rauschen ξ_i mit definiertem Signal-Rausch-Verhältnis addiert, so daß die verrauschte Zeitreihe $\{\tilde{x}_i\}_{i=1,\ldots,200}$ gegeben war durch

$$\tilde{x}_i \quad = \quad x_i + \xi_i. \tag{5.21}$$

In Abbildung 5.16 ist die Delay-Einbettung für die Hénon-Abbildung dargestellt.

In Abbildung 5.17 ist das Ergebnis der Simulation dargestellt. Die Werte für die Koefizienten a, b und den konstanten Wert wurden bei der Description-Length $k = 3$ entnommen, dabei wurden immer die drei entscheidenden Koeffizienten gefunden, wie die nachstehende Tabelle zeigt:

SNR [dB]	const. Term	a	b
60	1.0005	1.4001	0.3001
40	1.0058	1.4003	0.3007
30	1.0048	1.3932	0.3004
20	1.0447	1.3148	0.2907

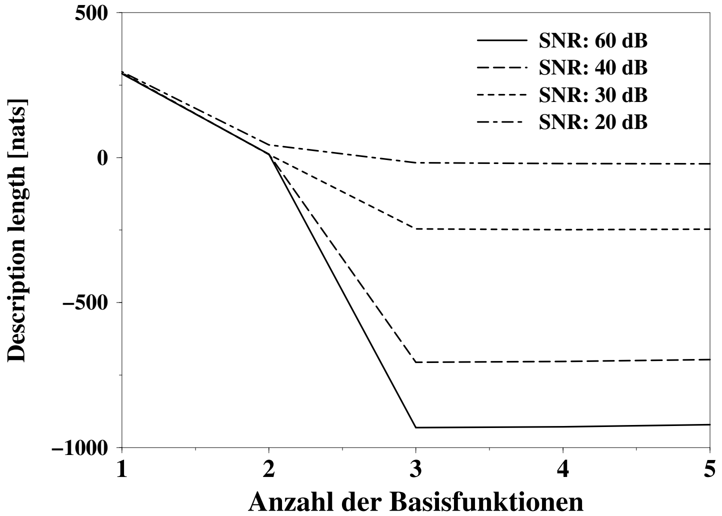


Abbildung 5.17: Das Verhalten der MDL bei verrauschten Daten, am Beispiel der Hénon-Abbildung.

Aus Abbildung 5.17 entnimmt man, daß die MDL empfindlich gegen additives Rauschen ist. Für ein SNR von 60 dB wird das Minimum an der erwarteten Stelle gefunden, für höhere Werte flachen die Kuven zusehends ab, und das Minimum verschiebt sich nach rechts.

5.7 Fazit

In diesem Kapitel wurde die Methode der MDL beschrieben und an Beispielen vorgestellt.
Es konnte gezeigt werden, daß der Einsatz dieser Methode nicht unproblematisch ist.
Das in Abschnitt 5.4 gewonnene, theoretische Ergebnis der Eindeutigkeit der MDL für orthonormale Basisfunktionen konnte für die orthonormalen Polynome an zwei Beispielen verifiziert werden.
Weiterhin stellte sich heraus, daß die Methode der MDL für nichtorthonormale Basisfunktionen zu unbefriedigenden Ergebnissen führen kann. Beim Vergleich mit der Cross-Validation zeigte sich im allgemeinen keine Übereinstimmung zwischen den, durch die beiden Verfahren ermittelten optimalen Modellen.

Kapitel 6

Zusammenfassung und Ausblick

Zu Beginn dieser Arbeit wurde eine kurze Einführung in die theoretischen Grundlagen der nichtlinearen Dynamik gegeben. Im Anschluß daran wurden die in dieser Arbeit untersuchten Methoden der Modellierung eingehend vorgestellt. Das Hauptaugenmerk war dabei auf die Entwicklung einer orthonormalen Basis von Polynomen gerichtet, die zur Approximation der Dynamik im Rekonstruktionsraum verwendet werden konnte.
In Kapitel 4 wurden die Modellierungsverfahren bzgl. der benötigten Rechenzeiten miteinander verglichen. Dabei zeigte sich, daß der Rechenaufwand bei der Approximation des Flusses mit orthonormalen Basispolynomen im allgemeinen höher ist als die Approximation in der polynomialen Standardbasis.
Im nachfolgenden Kapitel wurden zwei Methoden zur Modellvalidierung an experimentellen Zeitreihen miteinander verglichen, insbesondere wurde die *Minimum-Description-Length* (MDL) ausführlich dargestellt und die Probleme bei der Interpretation und beim Einsatz dieser Methode im Detail diskutiert. Dabei wurden zwei offenen Fragen der MDL aufgedeckt:

- Welcher Aufwand muß betrieben werden, um ein Modell eindeutig zu beschreiben ?
- Unter welchen Bedingungen ist das Minimum der Description-Length eindeutig bestimmt ?

Ein Ergebnis dieser Arbeit ist es, die Orthonormalität der Basisfunktionen als hinreichende Bedingung für die Existenz eines eindeutigen Minimums angeben zu können. Die orthonormalen Basispolynome sind in diesem Sinne vorteilhaft bei der Modellierung nichtlinearer Dynamik, denn sie liefern ein eindeutiges Ergebnis der MDL und damit ein Entscheidungskriterium bei der Wahl der optimalen Größe des Modells. Die Frage, ob diese Bedingung auch notwendig für die Existenz eines eindeutigen Minimums der MDL ist, scheint noch offen zu sein[1] und stellt einen interessanten Ansatzpunkt zur Weiterentwicklung der Theorie dar.

In dieser Arbeit wurden Polynome zur Konstruktion einer orthonormalen Funktionenbasis verwendet. Das in Kapitel 3 vorgestellte Verfahren läßt sich auch auf andere Funktionenklassen übertragen, so wurde z.B. von Mees ein Orthogonalisierungsverfaren für radiale Basisfunktionen untersucht [35]. In Verbindung mit der MDL stellen orthonormale Funktionsbasen ein taugliches Werkzeug zur Bildung nichtlinearer Modelle dar, um im Sinne H. Poincarés *die Gesetze der nicht zufälligen Erscheinungen zu bestimmen.*

[1] In der verwendeten Literatur [44, 45, 43, 37, 27, 35] ist zur Beantwortung dieser Frage zumindest kein Hinweis gefunden worden.

Anhang A

Notation

Erläuterung der verwendeten Symbole

$\mathcal{B}^d$	Die Borelsche σ-Algebra im $\mathbb{R}^d$
$C^n(G)$	Der Raum der n-mal stetig differenzierbaren Fkt. auf dem Gebiet G
$D_{\vec{x}}f$	Die Jacobimatrix der Funktion f an der Stelle $\vec{x}$
E_n	Die $(n \times n)-$Einheitsmatrix
$\|\vec{x}\|$	Die euklidische Norm des Vektors $\vec{x}$
$\langle \cdot \mid \cdot \rangle$	Das Standardskalarprodukt
$\langle \cdot \mid \cdot \rangle_\mu$	Das Skalarprodukt bzgl. des ergodischen Maßes
$\mathcal{L}_x(\theta)$	Die logarithmische Maximum-Likelihoodfunktion des Parameters θ zur Beobachtung x
$\mathcal{N}(\mu, \sigma^2)$	Die Normalverteilung mit dem Mittelwert μ und der Varianz σ^2
$M^\dagger$	Die transponierte Matrix zur Matrix M

Literaturverzeichnis

[1] J. Argyris, G. Faust, und M. Haase. *Die Erforschung des Chaos.* Vieweg, 1994.

[2] V.I. Arnold. *Mathematical Methods of classical Mechanics.* Springer Verlag, 1978.

[3] V.I. Arnold. *Huygens and Barrow, Newton and Hook.* Birkhäuser Verlag, Basel, 1990.

[4] V.I. Arnold und A. Avez. *Problèmes Ergodiques de la Mécanique Classique.* Gauthier-Villars, 1967.

[5] D.K. Arrowsmith und C.M. Place. *Dynamische Systeme.* Spektrum Lehrbuch. Spektrum Akademischer Verlag, Heidelberg, 1990.

[6] T. Bröcker und K. Jänich. *Einführung in die Differentialtopologie.* Springer Verlag, 1973.

[7] P.J. Brockwell und R.A. Davis. *Time Series : Theory and Methods.* Springer Series in Statistics. Springer Verlag, New York, Inc., zweite Auflage, 1991.

[8] R. Brown, N.F. Rulkov, und E.R. Tracy. Modeling and synchronizing chaotic systems from time-series data. *Physical Review E*, 49(5):3784–3800, 1994.

[9] M. Casdagli. Nonlinear prediction of chaotic time series. *Physica D*, 35:335 – 356, 1989.

[10] M. Casdagli, D. Des Jardins, S. Eubank, J.D. Farmer, J. Gibson, N. Hunter, und J. Theiler. Nonlinear modeling of chaotic time series: Theory and applications. Preprint, 1991.

[11] G.J. Chaitin. On the length of programs for computing finite binary sequences. *Journal of the ACM*, 13:547–569, 1966.

[12] S. Chen, S.A. Billings, C.F.N. Cowan, und P.M. Grant. Practical identification of narmax models using radial basis functions. *Int. J. Control*, 52(6):1327–1350, 1990.

[13] S. Chen, S.A. Billings, und W. Luo. Orthogonal least squares methods and their application to non-linear system identification. *Int. J. Control*, 50(5):1873–1896, 1989.

[14] J. Cremers und A. Hübler. Construction of differential equations from experimental data. *Z. Naturforsch.*, 42(a):797–802, 1987.

[15] J.P. Crutchfield und B.S. McNamara. Equations of motion from a data series. *Complex Systems*, 1:417 – 452, 1987.

[16] J.-P. Eckmann und D. Ruelle. Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.*, 57(3):617 – 656, 1985.

[17] K.J. Falconer. *The geometry of fractal sets.* Cambridge University Press, 1985.

[18] J.D. Farmer und J.J. Sidorowich. Predicting chaotic time series. *Phys. Rev. Lett.*, 59(8):845 – 848, 1987.

[19] G. Froyland, K. Judd, A.I. Mees, D. Watson, und K. Murao. Constructing invariant measures from data. *Int. J. Bif. Chaos*, 5(4):1181–1192, 1995.

[20] M. Giona, F. Lentini, und V. Cimagalli. Functional reconstruction and local prediction of chaotic time series. *Phys. Rev. A*, 44(6):3496 – 3502, 1991.

[21] P. Grassberger. Generalized dimensions of strange attractors. *Phys. Lett. A*, 1983.

[22] G. Hämmerlin und K.H. Hoffmann. *Numerische Mathematik.* Springer-Verlag, Berlin, vierte Auflage, 1994.

[23] H. Heuser. *Gewöhnliche Differentialgleichungen.* Teubner, Stuttgart, 1989.

[24] H. Heuser. *Lehrbuch der Analysis 2.* Teubner, Stuttgart, 1990.

[25] U. Hübner, C.O. Weiss, N.B. Abraham, und D. Tang. Lorenz-Like Chaos in NH_3-FIR Lasers. In A.S. Weigend und N.A. Gershenfeld, Herausgeber, *Times Series Prediction: Forecasting the Future and Understanding the Past.* Addison-Wesley, 1993.

[26] S.W. Jablonski und O.B. Lupanow. *Diskrete Mathematik und mathematische Fragen der Kybernetik.* Akademie-Verlag, Berlin, 1980.

[27] K. Judd und A. Mees. On selecting models for nonlinear time series. *Physica D*, 82:426–444, 1995.

[28] L. Junge. *Synchronisation nichtlinearer dynamischer Systeme*. Diplomarbeit, Georg-August-Universität, Göttingen, Januar 1997.

[29] G. Klaus und M. Buhr. *Philosophisches Wörterbuch*. das europäische buch, Berlin, 1972.

[30] U. Krengel. *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. Vieweg Verlag, 1991.

[31] W. Lauterborn. Methoden der Chaosphysik. *Kleinheubacher Berichte*, 32:585–598, 1989.

[32] R.W. Leven, B.P. Koch, und B. Pompe. *Chaos in dissipativen Systemen*. Akademie Verlag, Berlin, 1994.

[33] B. Mandelbrot. *Die fraktale Geometrie der Natur*. Birkhäuser Verlag, Basel, 1987.

[34] E. Maus. *Lineare Algebra 1*. Mathematisches Institut der Georg-August-Universität, Göttingen, 1991. Skript zur Vorlesung Wintersemester 1990/91.

[35] A. Mees. Parsimonous dynamical reconstruction. *Int. J. Bif. Chaos*, 3(3):669–675, 1993.

[36] A. Mees. Persönliche Mitteilung, Dezember 1996.

[37] A. Mees und K. Judd. Parsimony in dynamical modeling. Preprint, January 1995. Centre for Applied Dynamics and Optimization, The University of Western Australia.

[38] G.J. Ortega. Invariant measures as lagrangian variables: Their application to time series analysis. *Phys. Rev. Lett.*, 77(2):259–261, 1996.

[39] H. Poincaré. *Wissenschaft und Methode*. Teubner, Leipzig, 1914.

[40] W.H. Press, S.A. Teukolsky, W.T. Vetterling, und B.P. Flannery. *Numerical Recipes in C*. Cambridge University Press, zweite Auflage, 1992.

[41] P.E. Rapp, A.M. Albano, I.D. Zimmermann, und M.A. Jimenez-Montano. Phase-randomized surrogates can produce spurious identifications of non-random structure. *Phys. Lett. A*, 192:27–33, 1994.

[42] A. Rényi. *Wahrscheinlichkeitsrechnung mit einem Anhang über Informationstheorie*. VEB Deutscher Verlag der Wissenschaften, 1962.

[43] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[44] J. Rissanen. Minimum description lenght principle. In S. Kotz und N.L. Johnson, Herausgeber, *Encyclopedia of Statistical Sciences*. Wiley, New York, 1985.

[45] J. Rissanen. Stochastic complexity and modeling. *The Annals of Statistics*, 14(3):1080–1100, 1986.

[46] D. Ruelle. *Chaotic evolution and strange attractors*. Cambridge University Press, 1989.

[47] D. Ruelle. *Zufall und Chaos*. Springer Verlag, Berlin, zweite Auflage, 1994.

[48] T. Sauer, J.A. Yorke, und M. Casdagli. Embedology. *J.Stat.Phys.*, 65, 1991.

[49] T. Schreiber. Efficient neighbor searching in nonlinear time series analysis. *Int. J. Bif. Chaos*, 5(2):349–358, 1995.

[50] T. Schreiber und H. Kantz. *Nonlinear Time Series Analysis*. Preprint, 1996.

[51] H.G. Schuster. *Deterministic Chaos*. VCH Verlagsgesellschaft mbH, Weinheim, 1988.

[52] C.E. Shannon und W. Weaver. *The mathematical theory of communication*. University of Ill. Press, Urbana, 1949.

[53] J. Stark. Recursive prediction of chaotic time series. *Nonl. Science*, 3:197–223, 1993.

[54] J. Stoer und R. Bulirsch. *Numerische Mathematik 2*. Springer-Verlag, Berlin, dritte Auflage, 1990.

[55] F. Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence*, Lect. Notes Math. Springer-Verlag, Berlin, 1981.

[56] J. Theiler und S. Eubank. Don't bleach chaotic data. Preprint, Los Alamos.

[57] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, und J.D. Farmer. Testing for nonlinearity in time series: the method of surrogate data. *Physica D*, 58:77 – 94, 1992.

[58] W.M. Thorburn. The Myth of Occam's Razor. *Mind*, 27:345–353, 1918.

[59] P. Walters. *An introduction to ergodic theory*. Springer-Verlag, 1982.

Danksagung

Herzlich bedanken möchte ich mich bei Prof. Dr. W. Lauterborn, der diese Diplomarbeit ermöglicht hat, ebenso bei Dr. U. Parlitz, der stets mit Rat und Tat zur Seite stand.

Allen Mitgliedern der Arbeitsgruppe Nichtlineare Dynamik bin ich sehr zu Dank verpflichtet für viele anregende Diskussionen, die zur Fertigstellung dieser Arbeit beigetragen haben, insbesondere Dipl.-Phys. C.D. Ohl und Dipl.-Phys. M. Wiesenfeldt für die zahlreichen Hilfen bei der Handhabung des Computers, sowie cand. phys. L. Junge, der mir die Analogrechnerdaten zur Verfügung stellte.

Meinen Freunden Dipl.-Math. M. Keßeböhmer, cand. math. J. Thirase und cand. phys. F. Euchner danke ich für viele anregende Diskussionen.

Auch unserer Fotografin, den Angestellten im Büro und unseren Soft- und Hardwarespezialisten möchte ich für die freundliche Hilfe bei den vielen kleinen und großen Problemen danken.

Bei allen Institutsangehörigen, besonders bei den Mitgliedern der Institutsband, bedanke ich mich für die freundliche und offene Atmosphäre im Haus.