Von Daten zu Stochastischen Modellen

Publiziert in *Elemente der Mathematik*, 2000, Vol. 55, pp. 1–18

Peter Bühlmann
Seminar für Statistik
Departement Mathematik, ETH Zentrum
CH-8092 Zürich

März 1999

1 Einleitung

Ein wohl allen empirischen Wissenschaften gemeinsames Ziel ist das Schliessen von Daten der realen Welt auf abstrakte Modelle. Ein Modell ist strikte genommen einfach eine Menge, welche mit Verknüpfungsregeln für deren Elemente versehen ist. In den Anwendungen wird dann dessen *Interpretation*, welche eine allgemeine vereinfachende Beschreibung des Beobachteten liefert, oder dessen *Prognose-Potenzial* wichtig sein. Wir beschränken uns hier auf stochastische Modelle.

Der Vorgang, wie man von Daten auf stochastische Modelle schliessen kann, gehört zum Kern der Statistik. Er ist massgebend unterstützt von der induktiven Logik: da das Schliessen von endlich vielen Beobachtungen (dem *Besonderen*) auf ein Modell (dem *Allgemeinen*) nicht mit Sicherheit möglich ist, benützt man die Wahrscheinlichkeitstheorie, um mit gewissen (typischerweise grossen) Wahrscheinlichkeiten immer noch Aussagen über das Allgemeine zu machen. Wir werden auf diese grundlagentheoretischen Aspekte nicht näher eingehen. Auch ist der Prozess 'von Daten zu stochastischen Modellen' häufig von einer interessierenden Fragestellung beeinflusst. Noch bevor man zu den Daten kommt, sollte idealerweise die Fragestellung im Zentrum stehen: zuerst Fragestellung, dann Daten und schliesslich stochastische Modell-Bildung. Wir werden aber auch diesen ersten Schritt, welcher zum Beispiel die Planung eines Experimentes beinhaltet, nicht weiter diskutieren.

Vielmehr möchten wir, vorwiegend exemplarisch, einen kleinen Aspekt der stochastischen Modell-Bildung diskutieren: er beinhaltet eine sehr beschränkte Auswahl von Problemstellungen. Das statistische Testen von Hypothesen wird kurz angeschnitten, hauptsächlich wird aber auf das in gewissem Sinne komplementäre Problem von optimalen Vorhersage-Modellen eingegangen. Dabei streifen wir Methoden der quantitativen Bestrafung für komplexe Modelle, Optimalität bei vielfältigen, riesigen Modell-Klassen und ausblickend einen modernen Ansatz der Mittelung von komplexen Modellen. Alle realen Daten-Beispiele handeln von zeitlich abhängigen Beobachtungen.

2 Zwei Ansätze für Modellwahl

Verschiedenste Methoden für die Wahl eines stochastischen Modells können vom Verwendungszweck her grundsätzlich in zwei Klassen eingeteilt werden: der *strukturelle* Ansatz, wo die Struktur eines Modells interessiert, oder der *entscheidungstheoretische* Ansatz, welcher als Ziel ein optimales Vorhersage-Potenzial des Modells hat. Wie wir sehen werden, kann der entscheidungstheoretische Ansatz auch interessante strukturelle Informationen liefern, umgekehrt erhält man aber im allgemeinen mit dem strukturellen Ansatz keine vollständige Information für eine optimale Prognose.

2.1 Struktureller Ansatz

Es geht hier darum, signifikante (oder populärer ausgedrückt: relevante) Struktur oder zumindest einige signifikante strukturelle Komponenten der Daten zu entdecken. Dafür benützen wir den Formalismus des statistischen Tests, welcher auf einem Falsifizierungs-Argument beruht: es kann bloss eine (Null-)Hypothese probabilistisch verworfen, aber nicht bewiesen werden. Einer (Null-)Hypothese ist immer eine Alternative entgegengesetzt. Dieser Formalismus kann folgendermassen konkretisiert werden. Als Grundlage ist ein allgemeines Basis-Modell spezifiziert, so dass die (Null-)Hypothese ein Spezialfall dieses allgemeinen Basis-Modells ist. Die Alternative ist dann das Komplement der (Null-)Hypothese *bezüglich des Basis-Modells.*

Das folgende einfache 'Spielzeug-Problem' illustriert den grundlegenden Gedanken bei der strukturellen Modellwahl, welcher auch bei realen Anwendungen in viel komplizierteren Situationen prägend ist. Abbildung 1 zeigt $n = 100$ simulierte Daten

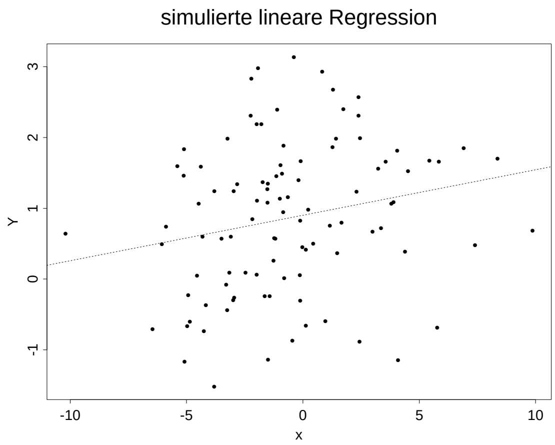

Abbildung 1: 100 simulierte Datenpunkte einer einfachen linearen Regression mit geschätzter Regressionsgerade (Kleinste-Quadrate Schätzung).

$(x_1, Y_1), \ldots, (x_n, Y_n)$. Es scheint vernünftig, als Basis-Modell ein einfaches lineares Regressions-Modell anzunehmen,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ (i = 1, \ldots, n),$$

wobei $\varepsilon_1, \ldots, \varepsilon_n$ unabhängig und identisch verteilt (i.i.d.) sind mit Erwartungswert $\mathbb{E}[\varepsilon_i] = 0$ und Varianz $\mathrm{Var}(\varepsilon_i) = \sigma^2 < \infty$. Als Nullhypothese ($H_0$) betrachten wir den Spezialfall des Basis-Modells,

$$H_0 : \beta_1 = 0,$$

welches besagt, dass es keinen Einfluss von erklärenden Variablen x_i auf Y_i ($i = 1, \ldots, n$) gibt. Die Frage ist jetzt, ob die Nullhypothese H_0 oder das Basis-Modell mit $\beta_1 \neq 0$ adäquater zur Beschreibung der Daten ist. Rein visuell, siehe Abbildung 1, kommt man zu keinem klaren, eindeutigen Schluss. Auch nicht mit der Punktschätzung $\hat{\beta}_1 = 0.064 \neq 0$ für die Steigung der Regressionsgeraden, da diese Schätzung bloss wegen den zufälligen Rauschtermen ε_i ($i = 1, \ldots, n$) von Null verschieden sein könnte. Mit Hilfe des klassischen t-Tests findet man aufgrund der Daten, dass die Nullhypothese auf dem 5% Test-Niveau [1] verworfen wird. Der wahre Wert, welchen wir hier ja bei diesem simulierten Beispiel kennen, ist $\beta_1 = 0.05$, und der statistische t-Test entscheidet also bei diesen Daten richtig.

Wir möchten im Folgenden kurz auf ein in der Praxis interessierendes reales Daten-Beispiel eingehen, wo ein viel komplizierteres strukturelles Problem vorliegt.

Beispiel 1: Wasserstand des Rio Negro in Manaus (Brasilien).
Die Daten bestehen aus täglichen Messungen von 1903-1992, welche insgesamt 32874 Werte ergeben. Natürlich weisen diese starke saisonale Schwankungen auf, welche hier nicht primär interessieren. In Abbildung 2 sind deshalb korrigierte Wasserstandswerte gegeben, so dass die Saison-Effekte verschwinden sollten, siehe Brillinger (1997). Die uns interessierende Fragestellung lautet hier: Gibt es eine systematische Erhöhung des Wasserstandes im Verlaufe der Zeit? Diese Fragestellung wird oft im Zusammenhang mit der Abholzung im Einzugsgebiet des Rio Negro diskutiert, da ein Baum mit seinen Wurzeln ein natürliches Aufsauge-Potenzial für Regenwasser besitzt.

Als allgemeines stochastisches Basis-Modell betrachten wir

$$X_t = m_t + \varepsilon_t \ (t = 1, \ldots, n = 32874),$$

wobei X_t den Saison-korrigierten Wasserstand, m_t einen deterministischen Trend und ε_t einen Rausch-Term zum Zeitpunkt t bezeichnen. Spezifischer nehmen wir an, dass $(m_t)_t$ eine schwach monoton wachsende Trend-Folge ist und $(\varepsilon_t)_t$ farbige stationäre Rausch-Terme darstellen, d.h. alle mit Erwartungswert $\mathbb{E}[\varepsilon_t] = 0$, aber im Gegensatz zu weissem Rauschen sind ε_s und ε_t korreliert für $s \neq t$. Die Annahme von farbigem Rauschen ist von den Daten her motiviert, die eine Zeitreihe bilden. Die ursprünglich interessierende Fragestellung, übersetzt in die Sprache der Modell-Welt, kann wie im obigen 'Spielzeug-Problem' mit einer Nullhypothese formalisiert werden: $m_t \equiv m$ für alle t. Aus Gründen einer vernünftigen Asymptotik zum Testen dieser Nullhypothese, betrachtet man anstelle einer Folge indiziert mit $\mathbb{N}$ ein reskaliertes Kurven-Problem. Bei Stichprobengrösse n sei $m_t = m(t/n)$ ($t = 1, \ldots, n$), wobei $m(.) : [0, 1] \to \mathbb{R}$ eine schwach monoton wachsende Kurve ist. Bei zunehmendem

[1] Das heisst, die Wahrscheinlichkeit für einen Fehler 1. Art ist 5%. Ein Fehler 1. Art bedeutet, dass die Nullhypothese fälschlicherweise vom statistischen Test verworfen wird.

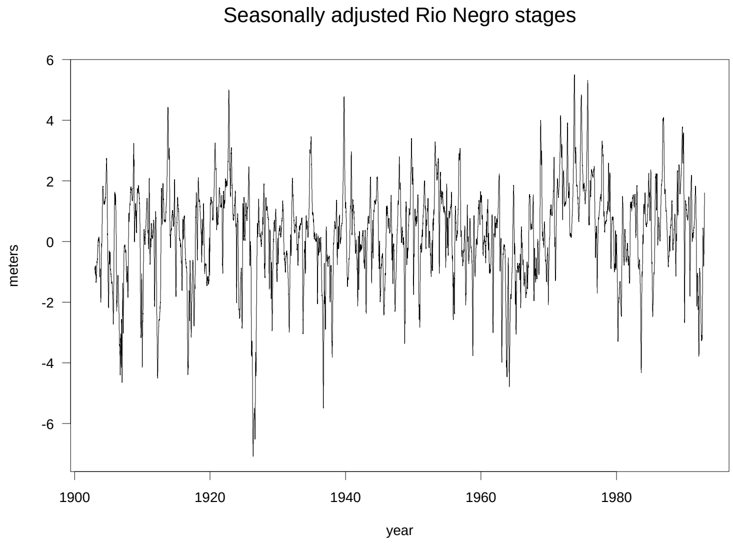

Abbildung 2: Saison-korrigierte Wasserstände des Rio Negro, zentriert um Null.

n beobachtet man die Kurve $m(.)$ also an immer dichter liegenden Punkten. Als Nullhypothese formulieren wir dann,

$$H_0: \ m(x) \equiv m \text{ für alle } x \in [0, 1].$$

welche ein Untermodell des allgemeinen Basis-Modells darstellt.

Aufgrund der Daten kann diese Nullhyothese getestet werden, obschon wir es mit einem schwierigen, nichtparametrischen Problem zu tun haben: die Rausch-Terme sind farbig (abhängig) und die Trend-Kurve $m(.)$ ist unendlich-dimensional. Die folgenden gefundenen Resultate sind von Brillinger (1997). Abbildung 3 zeigt eine

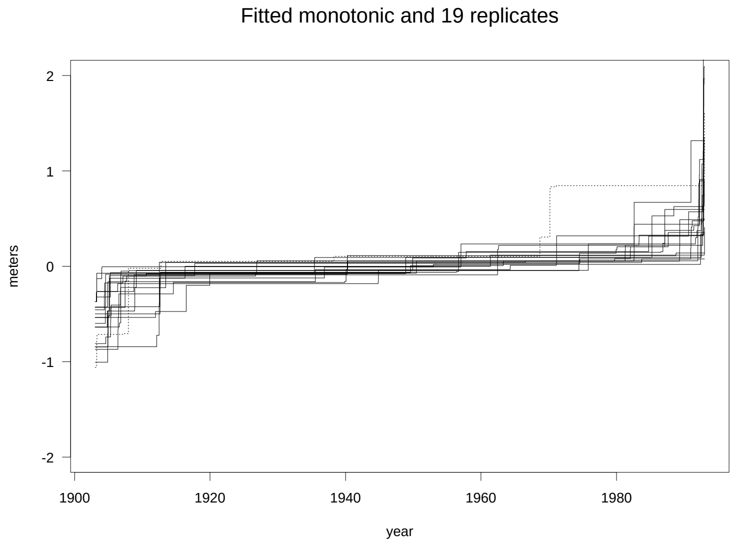

Abbildung 3: Geschätzte Trendkurve (gepunktet) und 19 Replikate dieser Schätzung unter H_0 (ausgezogen), zentriert um Null.

Schätzung $\hat{m}(.)$ der Trend-Kurve $m(.)$, basierend auf dem 'Pool-Adjacent-Violator'-Algorithmus, siehe Friedman & Tibshirani (1984). Die Frage ist dann, ob sich die Sprünge in der Schätzung, wo also die Trend-Kurve strikt monoton wächst, bloss aufgrund der Rausch-Effekte zeigen? Um dies zu beantworten werden 19 Replikate dieser Schätzung unter der Nullhypothese H_0 erzeugt. Dafür wird im zentrierten Fall $m_t \equiv 0$ gesetzt und die Schwierigkeit liegt dann in einer geeigneten Simulation des unbekannten Prozesses $(\varepsilon_t)_t$, welche mit einer Resampling-Technik durchgeführt wird. Man kann zeigen, dass so erhaltene Replikate des Kurven-Schätzers die statistische Variation von $\hat{m}(.)$ asymptotisch korrekt beschreiben, falls die Nullhypothese stimmt. Schätzer und Replikate ergeben zusammen 20 Kurven: auf dem 5% Test-Niveau fragt man, ob die geschätzte Kurve die Extremste unter allen zwanzig ist. Gemäss Abbildung 3 ist dies der Fall und wir schliessen, dass H_0 verworfen wird. Brillinger (1997) beschreibt das Resultat als 'there is a soupçon of an increasing trend'.

Wir hatten es hier also mit einem Beispiel zu tun, wo es um die *'wahre' Struktur* eines Modells geht. Der Begriff 'Wahrheit' ist dabei selbstverständlich bloss bezüglich eines postulierten allgemeinen Basis-Modells zu verstehen.

2.2 Entscheidungstheoretischer Ansatz

Hier interessiert die Güte eines geschätzten Modells, gemessen mit einer Verlustfunktion. Oft ist das Ziel in einer Anwendung, eine Prognose-Güte zu maximieren.

Beispiel 2: Helligkeit des 'White Dwarf' Sternes PG 1159-035
Die Daten bestehen aus 2000 Messungen von 'Helligkeit' (Lichtintensitäten), welche in Abständen von jeweils 10 Sekunden gemacht wurden, siehe Abbildung 4. Interessierende Fragestellungen sind unter anderen Periodizitäten und Schwingungen

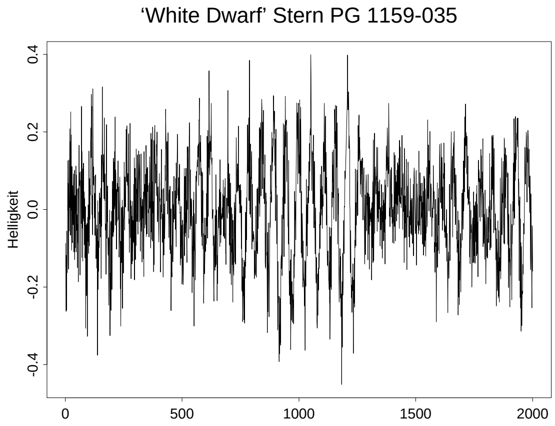

Abbildung 4: Helligkeit des 'White Dwarf' Sternes PG 1159-035.

der Lichtintensitäten, welche oft geeigneter im Frequenzbereich (Fourier-Bereich) analysiert werden. Der Einfachheit halber beschränken wir uns jedoch hier auf den Zeitbereich und betrachten ein stochastisches autoregressives Modell der Ordnung

p, abgekürzt mit $AR(p)$,

$$X_t = \theta_1 X_{t-1} + \ldots + \theta_p X_{t-p} + \varepsilon_t \ (t = 1, \ldots, n = 2000), \tag{1}$$

wobei $\varepsilon_1, \ldots, \varepsilon_n$ unabhängig und identisch verteilt (i.i.d.) sind mit Erwartungswert $\mathbb{E}[\varepsilon_t] = 0$, Varianz $\mathrm{Var}(\varepsilon_t) = \sigma^2 < \infty$ und X_t unabhängig von $\{\varepsilon_s; s < t\}$ ist; $\theta_1, \ldots, \theta_p$ sind die unbekannten Parameter. Natürlich weiss man a priori nicht, wie gross die Ordnung p gewählt werden soll. Deshalb betrachtet man die ganze Klasse von $AR(p)$-Modellen mit $0 \le p < \infty$,

$$\mathcal{M}_{AR} = \bigcup_{p=0}^{\infty} \{M;\ M \text{ ein } AR(p) \text{ mit Parameter } \theta_M \in \Theta_M \subset \mathbb{R}^{dim(M)}\}. \tag{2}$$

Hierbei (und im folgenden) bezeichnet M eine Modell-*Struktur*, θ_M der zur Struktur zugehörige Parametervektor (typischerweise unbekannt) im Parameterraum Θ_M und $dim(M) = dim(\theta_M)$ die Dimension des Modells (der Modell-Struktur). In (2) ist $dim(M) = p_M$ gerade die Ordnung des AR-Modells mit Struktur M. Das parametrische Modell ist dann vollständig beschrieben durch das Paar (M, θ_M).

Die Schätzung von den unbekannten θ_M in einer gegebenen Modell-Struktur M kann hier mit der berühmten Kleinste-Quadrate Methode durchgeführt werden, welche auf Legendre und Gauss zurückgeht. Die übliche Notation dafür ist $\hat{\theta}_M$,

$$\hat{\theta}_M = \mathrm{argmin}_{\theta_M = (\theta_1, \ldots, \theta_{p_M})'} \sum_{t=p_M+1}^{n} (X_t - \theta_1 X_{t-1} - \ldots - \theta_{p_M} X_{t-p_M})^2,$$

wobei p_M die Ordnung der AR-Modell-Struktur M bezeichnet. Mit dieser Schätzung konstruiert man sich eine geschätzte Prognose im Modell mit Struktur M für die nächste unbekannte Zufallsvariable X_{n+1},

$$\hat{\mu}_{M;n+1} = \hat{\theta}_1 X_n + \ldots + \hat{\theta}_{p_M} X_{n-p_M+1}.$$

Dieser gegenübergestellt steht die wahre Prognose (eines Orakels),

$$\mu^*_{n+1} = \theta^*_1 X_n + \ldots + \theta^*_{p^*} X_{n-p^*+1},$$

wobei $\theta^*_M = (\theta^*_1, \ldots, \theta^*_{p^*})'$ den wahren Parametervektor in dem wahren $AR(p^*)$ bezeichnet. (Wir nehmen hier an, dass das wahre Modell in $\mathcal{M}_{AR}$ liegt. Solch eine Restriktion wird in Kapitel 3 aufgehoben). Das Ziel ist nun, ein Risiko, zum Beispiel den erwarteten quadratischen Verlust [2], zu minimieren,

$$R(M) = \mathbb{E}[(\hat{\mu}_{M;n+1} - \mu^*_{n+1})^2].$$

Bezüglich einer solchen Risiko-Funktion definiert man die optimale Modell-Struktur in der Klasse $\mathcal{M}_{AR}$ als,

$$M_{opt} = \mathrm{argmin}_{M \in \mathcal{M}_{AR}} R(M).$$

[2] Es kann gezeigt werden, dass diese Risiko-Funktion auch vernünftig ist für die Schätzung der Fourier-Transformierten in AR-Modellen (Shibata, 1981); die interessierenden Fragestellungen im Frequenzbereich können also auch mit der beschriebenen Technik analysiert werden.

Das folgende erstaunliche Phänomen beschreibt jetzt aber die Andersartigkeit zwischen wahrer und optimaler Modell-Struktur. Auch falls die wahre Struktur $M_{wahr} \in \mathcal{M}_{AR}$, so ist im Allgemeinen

$$M_{opt} \neq M_{wahr}.$$

Die Frage nach der Optimalität einer Modell-Struktur ist also grundlegend anders als die Frage nach der wahren Struktur! Für die entscheidungstheoretische Modellwahl, oder die Modellwahl bezüglich der besten Prognose, nützt unter Umständen die Kenntnis der wahren Struktur wenig.

Wir wollen vorerst eine intuitive Erklärung für dieses, auf den ersten Blick doch paradoxe Phänomen geben. Das Risiko einer Modell-Struktur beinhaltet implizit die Schätzungen der unbekannten Parameter. Die Ungenauigkeiten bei diesen Schätzungen addieren sich mit jedem Parameter auf. Deshalb kann manchmal eine Modell-Struktur mit wenig unbekannten Parametern einer hoch-dimensionalen, wahren Struktur vorgezogen werden; obwohl man dann natürlich einen systematischen Fehler mit dem unwahren niedrig-dimensionalen Modell machen wird. Diese Intuition kann mathematisch quantifiziert werden. Das Risiko für eine Struktur M kann folgendermassen zerlegt werden,

$$R(M) = E[(\hat{\mu}_{M;n+1} - \mu^*_{n+1})^2] \approx E[(\mu^*_{M;n+1} - \mu^*_{n+1})^2] + \sigma^2 \frac{dim(M)}{n},$$

wobei $\mu^*_{M;n+1}$ die beste lineare Prognose in Modell-Struktur M für X_{n+1} ist (bezüglich Risiko $R(.)$). Der erste Term auf der rechten Seite der approximativen Gleichung beschreibt den systematischen Fehler (Bias) zwischen der besten Prognose $\mu^*_{M;n+1}$ in der Struktur M (diese hat nichts mit Schätzung zu tun) und der wahren Prognose μ^*_{n+1}, der zweite Term beschreibt den Schätzfehler, welcher sich aus den Varianzen der einzelnen Parameterschätzungen ergibt. Interessanterweise wächst dieser zweite Varianz-Term *linear* in der Dimension der Modell-Struktur. Insbesondere, falls $dim(M_{wahr})$ gross ist, so kann dieser für die wahre Struktur einen dominierenden Negativ-Effekt auf die Prognose-Güte (Risiko $R(.)$) haben.

Dieses Modellwahl-Phänomen tritt in Situationen mit allgemeinen Modellklassen $\mathcal{M}$ und Risikofunktionen auf, siehe auch Kapitel 3 und 4. Zusammenfassend können wir also festhalten, dass die optimale Modell-Struktur ganz allgemein einen *Bias-Varianz Trade-off* berücksichtigt.

Natürlich ist unsere vorhin betrachtete Risikofunktion $R(.)$ unbekannt und daher auch die optimale Modell-Struktur M_{opt}. Schätzungen von $R(.)$ aus den Daten sind aber bekannt, so zum Beispiel der 'Final Prediction Error' (FPE) von Akaike (1969) oder Mallows C_p (Mallows, 1973). Eine Modellwahl für den konkreten Datensatz in Beispiel 2 wird im nächsten Kapitel 3 mit einer allgemeineren Methode durchgeführt.

3 Modellwahl mit Akaike's Kriterium

Wir diskutieren hier eine noch viel universeller anwendbare Methode, um entscheidungstheoretische Modellwahl durchzuführen.

Bevor wir spezifischere Annahmen über ein Modell machen, bezeichnen wir mit P^* die (wahre) Wahrscheinlichkeitsverteilung der Daten $X_1, \ldots, X_n$. Der Werteraum von X_t sei $\mathcal{X}$ $(t = 1, \ldots, n)$, zum Beispiel $\mathcal{X} = \mathbb{R}$. Man möchte dieses wahre P^* schätzen, so dass man die gesamte stochastische Kenntnis zumindest approximativ besitzt.

Dazu benützt man häufig eine möglichst geeignete parametrische Modellklasse,

$$\begin{aligned} &\mathcal{M} \text{ eine diskrete Menge,} \\ &\mathcal{P} = \bigcup_{M \in \mathcal{M}} \{P_{\theta_M},\ \theta_M \in \Theta_M \subset \mathbb{R}^{dim(M)}\}. \end{aligned} \tag{3}$$

Die Menge $\mathcal{M}$ besitzt als Elemente alle interessierenden Modell-Strukturen M, $\mathcal{P}$ ist dann die zugehörige Klasse von Wahrscheinlichkeitsverteilungen P_{θ_M}, indiziert mit einem unbekannten Parameter θ_M $(M \in \mathcal{M})$. Beispiele dafür sind die Klasse in (2), aber auch die Klasse in (6) in Kapitel 4, welche kategorielle Daten beschreibt. Letztere ist auch ein Beispiel dafür, wo das in Kapitel 2.2 vorgestellte Risiko $R(.)$ mit dem erwarteten quadratischen Verlust keinen Sinn macht, da bei kategoriellen Daten keine Ordnung vorhanden ist.

Die Schätzung des unbekannten Parameter-Vektors θ_M in Modell-Struktur M kann im allgemeinen mit der Maximum-Likelihood Methode ausgeführt werden,

$$\hat{\theta}_M = \text{argmin}_{\theta_M \in \Theta_M} - \log\left(dP_{\theta_M}(X_1, \ldots, X_n)\right),$$

wobei $dP_{\theta_M}(X_1, \ldots, X_n)$ die Wahrscheinlichkeitsdichte (oder Wahrscheinlichkeit) der Daten $X_1, \ldots, X_n$ im Modell mit Struktur M und Parametervektor θ_M bezeichnet. Der Maximum-Likelihood Schätzer $\hat{\theta}_M$ gibt also im Modell mit Struktur M maximale Wahrscheinlichkeit für die beobachteten Daten. Falls das Modell P_{θ_M} für unabhängige und identisch normalverteilte Zufallsvariablen steht, so fällt der Maximum-Likelihood Schätzer mit dem Kleinste-Quadrate Schätzer zusammen.

Wie bereits oben erwähnt möchten wir eine Risikofunktion, welche universeller anwendbar ist als diejenige von Kapitel 2.2 basierend auf quadratischem Verlust. Wir betrachten hier die sogenannte Kullback-Leibler Information für eine Modell-Struktur M,

$$\text{KLI}(M) = I_n(P^*, P_{\hat{\theta}_M}) = \int_{\mathcal{X}^n} \log\left(\frac{dP^*(x)}{dP_{\hat{\theta}_M}(x)}\right) dP^*(x).$$

Dieses Risiko ist auch bekannt als relative Entropie von $dP^*/dP_{\hat{\theta}_M}$ bzgl. $d\hat{P}_{\hat{\theta}_M}$. Eine alternative, mehr wahrscheinlichkeitstheoretische Form dafür ist,

$$\text{KLI}(M) = C + \mathbb{E}[-\log\left(dP_{\hat{\theta}_M}(Y_1, \ldots, Y_n)\right)], \tag{4}$$

wobei $C = \log(dP^*(Y_1, \ldots, Y_n))$ eine Konstante bezüglich Modellwahl ist (keine funktionelle Abhängigkeit von M) und $Y_1, \ldots, Y_n$ sind Zufallsvariablen, welche unabhängig von den Daten $X_1, \ldots, X_n$ sind, jedoch dieselbe Wahrscheinlichkeitsverteilung P^* haben. Diese Variablen $Y_1, \ldots, Y_n$ können als sogenannter *'Test-Set'* interpretiert werden: die aus den Daten $X_1, \ldots, X_n$ geschätzte Verteilung $P_{\hat{\theta}_M}$ wird an den neuen, von den Daten unabhängigen, Test-Variablen $Y_1, \ldots, Y_n$ evaluiert. Minimierung von $\mathrm{KLI}(M)$ bezüglich M ist äquivalent zu Minimierung von $\mathbb{E}[-\log\left(dP_{\hat{\theta}_M}(Y_1, \ldots, Y_n)\right)]$, der geschätzten negativen log-likelihood, evaluiert an und danach gemittelt über den 'Test-Set'.

Analog zu Kapitel 2.2 definieren wir die optimale Modell-Struktur,

$$M_{opt} = \mathrm{argmin}_{M \in \mathcal{M}} \mathrm{KLI}(M).$$

Auch hier gilt im allgemeinen $M_{opt} \neq M_{wahr}$. Die Erklärung dafür liefert wiederum eine Bias-Varianz Zerlegung,

$$\mathrm{KLI}(M) = I_n(P^*, P_{\theta^*_M}) + \int_{\mathcal{X}^n} \log\left(\frac{dP_{\theta^*_M}(x)}{dP_{\hat{\theta}_M}(x)}\right) dP^*(x)$$

und unter Regularitätsbedingungen,

$$\mathrm{KLI}(M) \approx I_n(P^*, P_{\theta^*_M}) + \frac{1}{2} dim(M). \quad (5)$$

Hier ist $P_{\theta^*_M}$ die Wahrscheinlichkeitsverteilung, welche zur Struktur M gehört und am nächsten zu P^* ist (bezüglich KLI(.)), das heisst $\theta^*_M = \mathrm{argmin}_{\theta_M \in \Theta_M} I_n(P^*, P_{\theta_M})$. Die Illustration in Abbildung 5 schematisiert die Formel. Der erste Term auf der

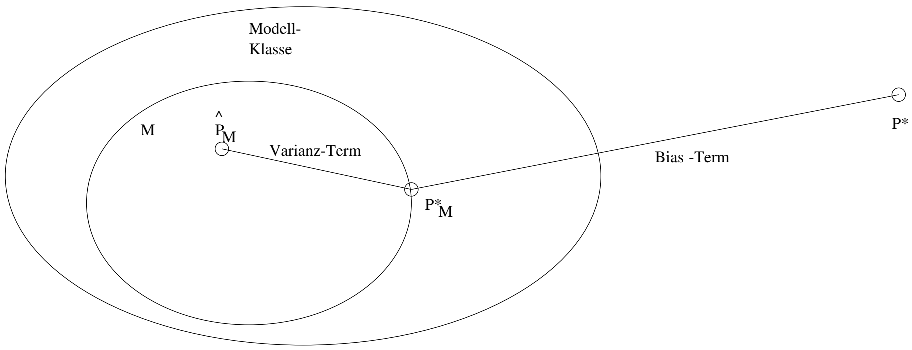

Abbildung 5: Bias-Varianz Zerlegung der Kullback-Leibler Information $I_n(P^*, P_{\hat{\theta}_M})$: $\hat{P}_M$ und P^*_M sind dabei Abkürzungen für $P_{\hat{\theta}_M}$, respektive $P^*_{\theta_M}$.

rechten Seite von (5) beschreibt den Bias (systematischer Fehler), der zweite den Varianz-Term (verursacht durch die Schätzung von unbekannten Parametern), welcher *linear* in der Anzahl Parameter wächst. Falls M 'komplexer' wird, das heisst in einer aufsteigenden Folge $M_1 \prec M_2 \prec \ldots$ ($M_i \prec M_{i+1}$ bedeutet M_i ist Untermodell von M_{i+1}) mit wachsender Anzahl Parameter wird der Bias-Term kleiner. Die wahre Wahrscheinlichkeitsverteilung der Daten P^* ist nicht notwendigerweise ein Element der Modellklasse $\mathcal{M}$. Trotzdem, $\mathrm{KLI}(M)$ kann bis auf die irrelevante Konstante C in (4) vernünftig und sehr einfach geschätzt werden. Diese Erkenntnis von Akaike (1973) gilt heute als der 'Breakthrough in Statistics Nr. 19.' (deLeeuw, 1991). Eine Schätzung von $2\mathrm{KLI}(M) - 2C$ ist

$$AIC(M) = -2\log(dP_{\hat{\theta}_M}(X_1, \ldots, X_n)) + 2dim(M).$$

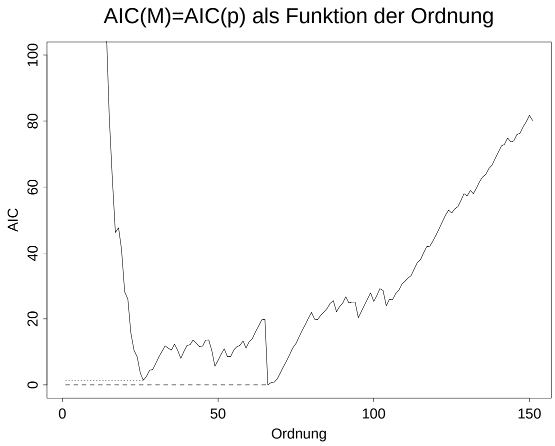

Abbildung 6: AIC Kriterium für Datensatz in Beispiel 2 und Modellklasse in (2). Optimale geschätzte Ordnung ist 65 (gestrichene Linie), sub-optimale Ordnung ist 25 (gepunktete Linie).

Sie trägt den Namen des Erfinders (Akaike, 1973) und heisst 'Akaike Information Criterion'. Der erste Term auf der rechten Seite ist ein Gütemass für den sogenannten Fit des geschätzten Modells für die Daten. Als ein konkretes Beispiel, wo das Modell P_{θ_M} für n unabhängige, identisch normalverteilte Zufallsvariablen steht, so ist das Gütemass eine Residuenquadratsumme. Es sagt aber *nichts* über das Vorhersage-Potenzial eines geschätzten Modells aus; insbesondere wird mit 'komplexerer' Modell-Struktur M dieses Gütemass kleiner. Der zweite Term auf der rechten Seite ist ein Bestrafungsterm, welcher 'komplexe' Strukturen linear in der Dimensionalität bestraft. Die Schätzung von KLI(M) ist also bis auf die für die Modellwahl irrelevante Konstante C in (4) ein Gütemass für den Fit plus ein Bestrafungsterm.

Eine geschätzte optimale Modell-Struktur ist nun gegeben durch

$$\hat{M}_{opt} = \text{argmin}_{M \in \mathcal{M}} AIC(M)$$

Falls die Klasse $\mathcal{M}$ von unendlicher Kardinalität ist, so restringiert man die Suche und minimiert bezüglich $\mathcal{M}_n = \{M;\ M \in \mathcal{M}, dim(M) \le c_n\}$, zum Beispiel $c_n = \sqrt{n}$.

Zusammenfassend halten wir fest, dass Modellwahl mit dem AIC-Kriterium universell anwendbar ist: die Kullback-Leibler Information KLI(.) als Risikofunktion ist auch für nichtnormalverteilte oder kategorielle Daten sinnvoll, und das AIC-Kriterium ist auch dann eine vernünftige Risikoschätzung, falls die wahre Wahrscheinlichkeitsverteilung P^* nicht in der betrachteten Modellklasse $\mathcal{M}$ liegt (was realistischerweise ja der Fall sein wird).

Wir analysieren nun den Datensatz von Beispiel 2 und benützen das AIC Kriterium um die optimale Modell-Struktur in der Klasse in (2) zu schätzen. Das Resultat ist in Abbildung 6 beschrieben. [3] Das AIC ist gross bei kleinen Ordnungen (grosser Bias-Term) und bei grossen Ordnungen (grosser Varianz-Term). Wir verfolgen

[3] Wir nehmen dabei an, dass die Rauschterme ε_t normalverteilt sind.

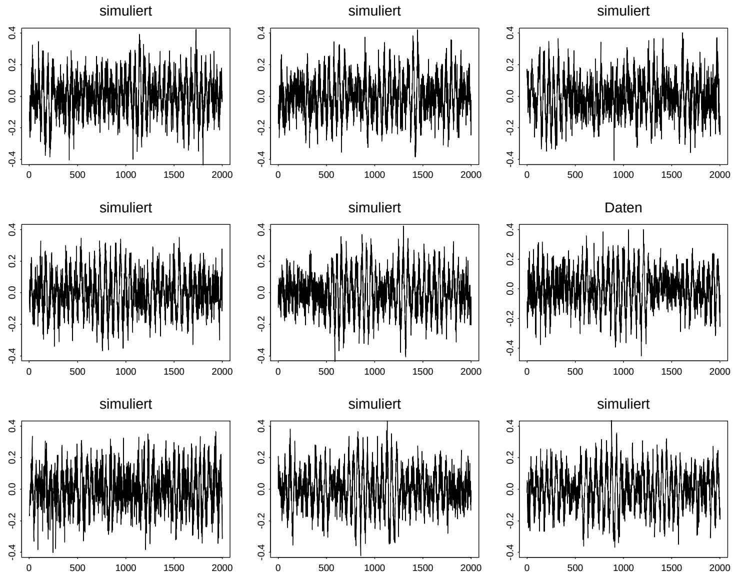

Abbildung 7: Acht simulierte Datensätze und wahrer Datensatz von Beispiel 2.

nun noch kurz die sub-optimale Lösung mit Ordnung 25, da der Verlust bezüglich AIC gegenüber dem Optimum mit Ordnung 65 klein ist und einfachere Modelle bei gleicher Güte prinzipiell vorzuziehen sind. Abbildung 7 zeigt den wahren Datensatz und 8 simulierte Datensätze des Modells in (1) mit $p = 25$ und normalverteilten Rauschtermen ε_t. Es ist visuell praktisch unmöglich den wahren von den simulierten Datensätzen zu unterscheiden. Dies ist eher ein Glücksfall: die Daten in Beispiel 2 lassen sich sehr gut durch ein extrem einfaches, nämlich lineares und Gauss'sches Modell beschreiben.

4 Vielfalt einer Modellklasse

Wir motivieren das Thema dieses Kapitels mit einem weiteren Beispiel.

Beispiel 3: Ein-dimensionale DNA von Drosophila.
Abbildung 8 zeigt einen Ausschnitt einer DNA-Sequenz von Drosophila, welche 25000 Zeichen lang ist. Interessierende Fragestellungen sind unter anderen die Ähnlichkeit zu anderen DNA Sequenzen oder Lokalisierung der 'kodierenden' Teilstücke in der Sequenz. Ein naheliegendes Modell ist eine stationäre Markov-Kette der Ordnung p (abgekürzt mit $MC(p)$) mit Werteraum $\mathcal{X} = \{A, T, G, C\}$, welche durch folgende Übergangswahrscheinlichkeiten charakterisiert ist:

$$\mathbb{P}[X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots] = \mathbb{P}[X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_{t-p} = x_{t-p}],$$

für alle $x_{t-1}, x_{t-2}, \ldots$ Dabei ist $p \in \mathbb{N}$ minimal. Dem Index t kommt wegen der Stationarität keine spezielle Bedeutung zu. Diese Wahrscheinlichkeiten können in einem Parametervektor $\theta_M \in (0,1)^{dim(M)}$ der Dimension $dim(M) = 3 \cdot 4^p$ zusammengefasst werden. Bei unbekannter Ordnung p betrachtet man oft die dazugehörende

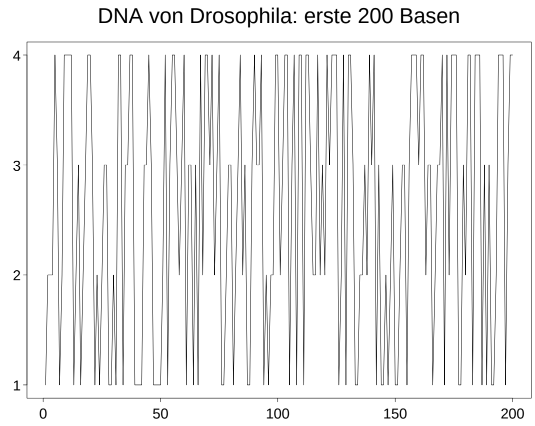

Abbildung 8: DNA von Drosophila. Die Zahlen 1,2,3,4 entsprechen den Basen A,T,G,C.

Modellklasse,

$$\mathcal{M}_{MC} = \bigcup_{p=0}^{\infty} \{M : \ M \text{ eine } MC(p) \text{ mit Übergangsw.'keit } \theta_M \in (0,1)^{dim(M)}\} \qquad (6)$$

Obwohl diese probabilistisch eine sehr schöne und natürliche Klasse bildet, ist sie in statistischem Sinne zu simpel und strukturell zu wenig reichhaltig.

Um dies zu illustrieren, betrachten wir die Dimension (Anzahl Parameter) von Struktur M als Funktion der Ordnung $p = p_M$, das heisst die Funktion $Dim(p) = 3 \cdot 4^p$.

p	0	1	2	3	4	5	10
Dim	3	12	48	192	768	3072	$\approx 3.1 \cdot 10^6$

Aus dieser Tabelle entnimmt man, dass es keine 'Zwischen-Modelle' gibt: die Dimension vervierfacht sich bei jeder zusätzlichen Ordnung und es gibt somit nur sehr 'sprunghafte' Erhöhungen der Dimension. Dies impliziert insbesondere auch, dass man oft zu wenig Flexibilität hat um den vorher diskutierten Bias-Varianz Trade-off gut zu berücksichtigen. Überdies hat man sehr schnell (zu) viele Parameter um Abhängigkeiten der Ordnung $p > 3$ zu modellieren. Solches wird auch in Braun & Müller (1998) bei der statistischen Analyse von DNA Sequenzen diskutiert.

Ein erfolgreicher Ausweg aus dieser zu simplen Modellklasse bilden die sogenannten Variable Length Markov Chains ($VLMC$), welche wohl zuerst in der Informationstheorie Fuss gefasst haben. Die Idee dabei ist, dass eine stationäre $VLMC$ ein Gedächtnis von *variabler* Länge hat. Eine $VLMC(p)$ ist charakterisiert durch die folgenden Übergangswahrscheinlichkeiten:

$$\mathbb{P}[X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots] = \mathbb{P}[X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_{t-\ell} = x_{t-\ell}]$$
$$0 \le \ell = \ell(x_{t-1}, x_{t-2}, \ldots) \le p, \ \ p \in \mathbb{N} \text{ minimal},$$

für alle $x_{t-1}, x_{t-2}, \ldots$. Das Gedächtnis weist eine variable Länge ℓ auf, welche selbst eine Funktion der Vergangenheit $x_{t-1}, x_{t-2}, \ldots$ ist. Wiederum kommt wegen der Stationarität dem Index t keine spezielle Bedeutung zu. Ein solches Gedächtnis von variabler Länge, oder äquivalent die Funktion $\ell(.)$, kann als Baum dargestellt werden. Auf der linken Seite in Abbildung 9 ist das Gedächtnis einer vollen MC(2) darge-

Abbildung 9: Baum-Darstellung von VLMC-Gedächtnis.

stellt mit $4^2 = 16$ Zuständen (Endknoten). Rechts in Abbildung 9 das Gedächtnis einer $VLMC(2)$ mit bloss 6 Zuständen: die Vergangenheiten $X_{t-1} = A$; $X_{t-1} = T$; $X_{t-1} = G$; $X_{t-1} = C$ und $X_{t-2} = A$; $X_{t-1} = C$ und $X_{t-2} \in \{T, G\}$; $X_{t-1} = C$ und $X_{t-2} = C$ (Repräsentation mit 5 Endknoten und einem inneren Knoten). Das variable Gedächtnis liest sich wie folgt: $\ell(A, \ldots) = \ell(T, \ldots) = \ell(G, \ldots) = 1$, $\ell(C, A, \ldots) = 2$, $\ell(C, T, \ldots) = \ell(C, G, \ldots) = 1$, $\ell(C, C, \ldots) = 2$.

Mit Hilfe der obigen Baum-Darstellungen wird schnell einmal klar, dass die Klasse der $VLMC$'s gute 'Zwischen-Modelle' hat und dass gewisse lange Abhängigkeiten mit wenig Zuständen modelliert werden können; bei beiden sind Bäume mit dünnen Ästen gefragt. Die Klasse aller $VLMC$'s besitzt also, im Gegensatz zu $\mathcal{M}_{MC}$ in (6), eine grosse Vielfalt oder Reichhaltigkeit.

Dabei hat man sich aber ein beträchtliches Problem eingehandelt: die Klasse ist enorm gross. So ist zum Beispiel die Anzahl $VLMC$ Untermodelle von MC(p) gleich $11(2^{4^{p-1}} + \sum_{k=1}^{p-2}(2^{4^k} - 1)) + 1$. Die folgende Tabelle verdeutlicht die astronomischen Grössen,

p	1	2	3	4
# Untermod.	12	176	721062	$\approx 2.0 \cdot 10^{20}$

Auch für die modernsten Computer der nächsten Generationen sind diese Zahlen zu gross. Falls man zum Beispiel für die Berechnung eines Modells eine Sekunde Rechenzeit brauchen würde, so müsste man für $p = 3$ bereits ungefähr 8.3 Tage und für $p = 4$ ungefähr $6.4 \cdot 10^{10}$ Jahrtausende rechnen! Eine globale Suche nach Optimalität wie zum Beispiel in Kapitel 3 mit $\hat{M}_{opt} = \text{argmin}_M AIC(M)$ ist für $p > 2$ nicht mehr möglich.

Eine grosse Innovation von Rissanen (1983) hilft aber diesem Problem ab. Man kann eine Suche nach der optimalen Modell-Struktur *'lokal'* statt 'global' durchführen. Die Idee dabei ist, eine Entscheidung für oder gegen ein Untermodell mittels Inspektion an *einzelnen* Endknoten von Bäumen zu machen. Dies kann in dem sogenannten Context-Algorithmus implementiert werden, welcher mit $O(n \log(n))$ wesentlichen Operationen arbeitet. Neuere Resultate zeigen, dass dieser Context-Algorithmus asymptotisch den wahren unterliegenden minimalen Zustandsraum (Baum) findet und statistisch effizient (asymptotisch optimal) ist.

Wir analysieren nun damit den DNA-Datensatz von Beispiel 3. Aus molekularbiologischen Gründen wird oft eine binäre Sequenz gebildet. Wir identifizieren wie folgt: $g = (G, C)$, $t = (T, A)$. Alle in diesem Kapitel gemachten Überlegungen machen wir jetzt für den binären Fall. Eine binäre $VLMC$ wird mit dem Context-Algorithmus auf dem Exon-Teil [4] der Sequenz (25000 Basen) angepasst. Das angepasste $VLMC$-Modell ist von Ordnung 6 und hat 26 Parameter. Interessanterweise kommen im Baum bloss Äste der Länge 0,3 und 6 vor: man kann also eine Darstellung in Triplets geben, siehe Abbildung 10. [5] Diese Darstellung ist interessant,

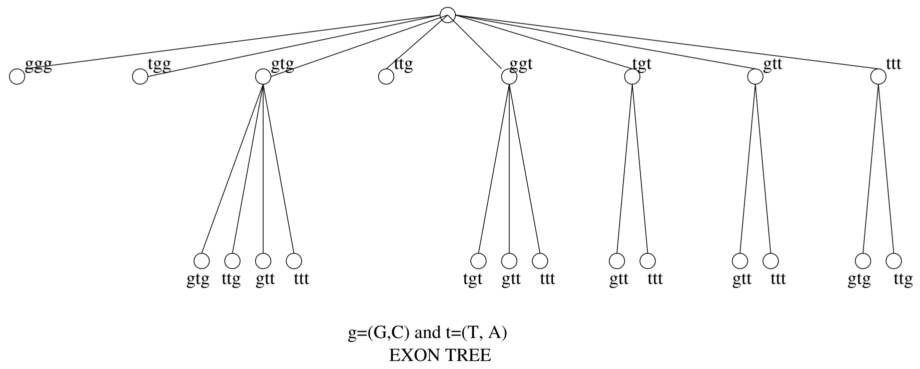

Abbildung 10: Triplet-Baum Darstellung von Gedächtnis einer angepassten $VLMC$ für binärisierte DNA-Sequenz in Beispiel 3. [4]

weil man von der Molekularbiologie weiss, dass Aminosäuren von Triplets der DNA kodiert werden. Diese Triplet Baum-Darstellung hat also eine schöne Interpretation in der Molekularbiologie. Obschon die Modellwahl mit dem Context-Algorithmus primär darauf abzielt, gute Vorhersage-Modelle zu finden, haben wir es hier mit einem Beispiel zu tun, bei dem ein gewähltes Modell zusätzlich eine sehr schöne strukturelle Interpretation hat. Es ist also ein Beispiel, wo möglicherweise $M_{opt} \approx M_{wahr}$, was gemäss dem früher formulierten Modellwahl-Phänomen nicht die Regel ist.

Die Idee von $VLMC$'s kann auch mit einigem Aufwand auf das schwierigere Problem von stationären $\mathbb{R}$-wertigen Zeitreihen übertragen werden, wir nennen diese Modelle dann verallgemeinerte $VLMC$'s. Analog zu vorhin erhält man auch in diesem Falle eine vielfältige Modellklasse. Wir illustrieren nochmals an einem Beispiel.

Beispiel 4: Tägliche Returns von BMW Aktien.
Abbildung 11 zeigt 1000 tägliche Return-Daten des BMW Aktienpreises, das heisst $X_t = \log(P_t/P_{t-1})$ mit dem Aktienpreis P_t. Die interessierende Fragestellung ist oft direkt von mathematischer Natur: die Wahrscheinlichkeitsverteilung eines Returns der Zukunft, gegeben die Werte von heute und der Vergangenheit. Kenntnis davon ermöglicht die Konstruktion von diversen Risiko-Massen im sogenannten Risk-Management.

[4] Dies ist der Teil der DNA-Sequenz, wo man a-priori weiss, dass er 'kodiert'.

[5] Zum Beispiel beschreibt der zweite Knoten von links in Tiefe 2 die Übergangswahrscheinlichkeiten als $\mathbb{P}[X_t = x_t|(X_{t-1}, \ldots, X_{t-6}, \ldots) = (g, t, g, t, t, g, \ldots)] = \mathbb{P}[X_t = x_t|(X_{t-1}, \ldots, X_{t-6}) = (g, t, g, t, t, g)]$.

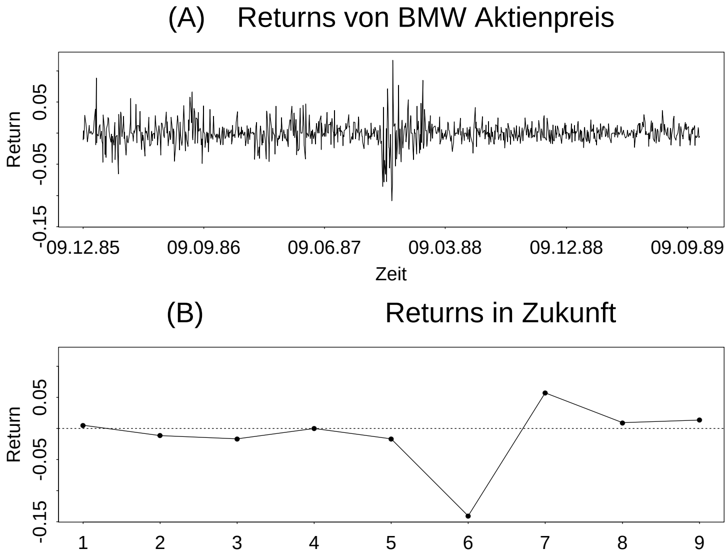

Abbildung 11: (A): 1000 tägliche Returns. (B): Die nächsten neun täglichen Returns der darauffolgenden Tage. Tag 6 ist der Montag nach dem Wiedervereinigungs-Wochenende in Deutschland.

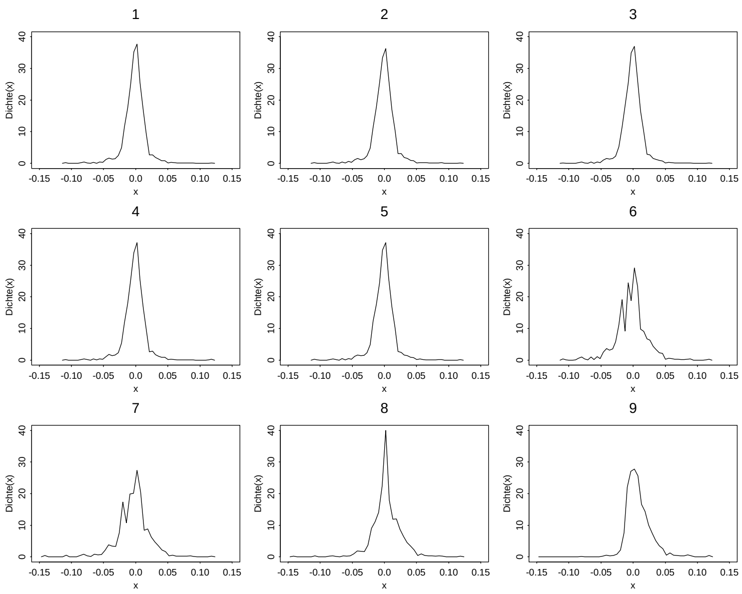

Abbildung 12: 1-Tage-Vorhersagedichten für die neun Tage in Abbildung 11 (B).

Für eine angepasste verallgemeinerte $VLMC$ zeigt Abbildung 12 die 1-Tage Vorhersagedichten für die 9 zukünftigen Tage von Abbildung 11 (B): das dabei benützte Modell hat 65 Parameter, was viel mehr Komplexität aufweist als die für Finanzzeitreihen oft benützten GARCH-Modelle. Die grossen und kleinen Returns an den Tagen 6 und 7 werden im Sinne von grosser Varianz bei den Vorhersagedichten korrekt prognostiziert. Die Vorzeichen der Returns können (natürlich) nicht vernünftig vorhergesagt werden. Erstaunlicherweise werden aber die 'Switches im Regime' an den Tagen 6 und 8 richtig vorhergesagt. Wie man solche 'Switches im Regime' quantitativ aus den Daten geschätzt hat, kann zum Beispiel folgendermassen formuliert werden. Man beobachtet eine relative bedingte Häufigkeit von 11.1% für das Unterschreiten des 5%-Quantils der gesamten Reihe, gegeben die Vergangenheit ähnlich wie bei Tag 6 (die Ähnlichkeit ist mit Hilfe des benützten Modells definiert). Dies entspricht mehr als einer Verdoppelung zum unabhängigen Fall, wo zukünftige Werte nicht von der Vergangenheit abhängen. Fairerweise fügen wir an, dass verallgemeinerte $VLMC$-Vorhersage für 'Switches im Regime' nicht immer so phantastisch wie in diesem Beispiel funktioniert.

5 Komplexe stochastische Modelle

In Vorhersage-Problemen werden öfters komplexe stochastische Modelle verwendet wie Neuronale Netze, 'Finite-State Machines' oder auch Methoden aus dem Gebiet von Pattern Recognition. Die in Kapitel 4 vorgestellten $VLMC$'s sind Spezialfälle von 'Finite-State Machines', welche vor allem in der Informationstheorie entwickelt wurden. In allen Fällen konstruiert man eine (für die entsprechende Anwendung) vielfältige Modellklasse.

Das Finden von passenden komplexen, stochastischen Prognose-Modellen ist dann äquivalent zum Problem der entscheidungstheoretischen Modellwahl in einer Klasse wie in (3). Insbesondere müssen zunehmend komplexe Modelle stärker bestraft werden. Wie wir in Kapitel 4 gesehen haben, kann man bei vielfältigen Modell-Klassen aus rechentechnischen Gründen keine globale Modell-Struktur Suche mit zum Beispiel dem AIC Kriterium durchführen. Der Context-Algorithmus (Rissanen, 1983), welcher in Kapitel 4 erwähnt wurde, erweist sich im Falle von $VLMC$'s als ein vernünftiges Verfahren, welches lokal sucht. In diesem Spezialfall vereinfachen der endliche Werteraum der Variablen X_t und die hierarchische Struktur in der Modell-Klasse, nämlich dass das Gedächtnis einer $VLMC$ immer noch aus zeitlich aufeinanderfolgenden Variablen besteht, das Problem beträchtlich. Viel schwieriger wird es bei Werteräumen wie $\mathbb{R}^d$ mit $1 \le d < \infty$, zum Beispiel bei den in Beispiel 4 verwendeten verallgemeinerten $VLMC$'s, und bei nicht-hierarchischen Modellen.

Wir möchten nun noch mit einer etwas erweiternden und ausblickenden Sicht schliessen. In der angewandten Statistik betrachtet man oft, insbesondere bei komplexen Problemen, mehrere 'gute' Modelle um vielleicht ein vollständigeres Bild zu erhalten. Warum überhaupt *das* optimale Modell? Interessanterweise wird auch diese Frage bei reinen 'Black-Box' Verfahren und komplexen stochastischen Vorhersage-Systemen wieder aufgegriffen. Rein experimentell ist eine beträchtliche Evidenz vorhanden, dass Mittelbildung über mehrere verschiedene Prognosen in komplexen Modellen letztendlich eine bessere Vorhersage liefert. Solch gemittelte Prognosen lassen sich sehr einfach mit Hilfe von zufälliger Perturbation implementieren. Man nimmt die einem persönlich am meisten zusagende komplexe Modell-Klasse mit deren Schätz-Algorithmus und mittelt dann über Prognosen, von denen jede jeweils durch zufällige Perturbation der Daten (des Inputs des Systems) zustandegekommen ist, siehe zum Beispiel Breiman (1996). Die theoretischeren Gründe dieses simplen, aber oft effektiven Tricks, sind bis heute weitgehend unerforscht.

Bemerkung. Dieser Artikel basiert auf meiner Einführungsvorlesung an der ETH Zürich vom Juni 1998. Mein spezieller Dank richtet sich an Prof. Urs Stammbach von der ETH Zürich für die Einladung zur Veröffentlichung, an Dr. Werner Stahel von der ETH Zürich für konstruktive Kommentare zum Manuskript und an Prof. David Brillinger von der University of California in Berkeley für die Aufarbeitung der Abbildungen 2 und 3. Letztere sind Modifikationen aus [5], welche hier mit der Genehmigung von John Wiley & Sons Limited abgedruckt sind.

Literatur

[1] Akaike, H. (1969). Fitting autoregressive models for prediction. Annals of the Institute of Statistical Mathematics **21**, 243-247.

[2] Akaike, H. (1973). Information theory and the maximum likelihood principle. In 2nd International Symposium on Information Theory (Eds. B.N. Petrov and F. Csàki), pp. 267-281. Akademiai Kiàdo, Budapest.

[3] Braun, J.V. & Müller, H.-G. (1998). Statistical methods for DNA sequence. Statistical Science **13**, 142-162.

[4] Breiman, L. (1996). Bagging predictors. Machine Learning **26**, 123-140.

[5] Brillinger, D.R. (1997). Random process methods and environmental data: the 1996 Hunter Lecture. Environmetrics **8**, 269-281.

[6] deLeeuw, J. (1991). Introduction to Akaike (1973) Information theory and an extension of the maximum likelihood principle.In Breakthroughs in Statistics, Vol. I, Foundations and Theory (Eds. N.L. Johnson & S. Kotz), pp. 599-609. Springer, New York.

[7] Friedman, J. & Tibshirani, R. (1984). The Monotone Smoothing of Scatterplots. Technometrics **26**, 243-250.

[8] Mallows, C.L. (1973). Some comments on C_p. Technometrics **15**, 661–675.

[9] Rissanen, J. (1983). A universal data compression system. IEEE Transactions on Information Theory **IT-29**, 656-664.

[10] Shibata, R. (1981). An optimal autoregressive spectral estimate. Annals of Statistics **9**, 300-306.