Stochastische Prozesse in der statistischen Modellierung

Gerhard Winkler [1]

GSF - Forschungszentrum für Umwelt und Gesundheit GmbH

Postfach 1129, D-85758 Oberschleißheim

IBB - Institut für Biomathematik und Biometrie

Oktober 2000

[1] gwinkler@GSF.de, http://www.gsf.de/ibb/

Ein Vorwort

Stochastische Prozesse gewannen in letzter Zeit wachsende Bedeutung in statistischer Modellierung und Inferenz. Ideen aus Wahrscheinlichkeitstheorie und stochastischer Analysis, aber auch aus den Ingenieurwissenschaften wurden in die Statistik übernommen; weitere tiefliegende Theorien wurden so für die Statistik erschlossen. Zudem erwiesen sich Sampling und Optimierungsmethoden wie Markov-Chain-Monte-Carlo Methoden als außerordentlich nützlich für die Behandlung statistischer Modelle; auch sie beruhen auf der Theorie stochastischer Prozesse und Felder. Die zugehörige statistische Inferenz ist noch nicht abschließend untersucht und stellt ein neues anspruchsvolles Forschungsgebiet dar.

Stochastische Prozesse sind Modelle für zufällige zeitliche Abläufe, insbesondere dann, wenn zeitliche Abhängigkeiten vorliegen, wie bei fast jeder praktischen Anwendung. Bei Verallgemeinerung auf zufällige Felder können auch raum-zeitliche Abhängigkeiten modelliert werden. Die zugehörige Inferenz erweist sich als bedeutend schwieriger als im Falle von unabhängigen Variablen. Neben parametrischen Verfahren spielen auch nichtparametrische Methoden eine wachsende Rolle.

In diesem Text ist nichts neu. Es werden vielmehr einige wichtige repräsentative Beispiele auf elementarem Niveau vorgestellt. Es geht dabei in erster Linie um eine verständliche und zugleich saubere Einführung der Begriffe. Für die weitergehende Theorie wird auf die Fachliteratur verwiesen. Als Schwerpunkte wurden Markov-Chain-Monte-Carlo Methoden, konditionale und dynamische Modelle sowie Ereignis- und Survivalanalyse ausgewählt.

Erster Anstoß für diesen Report waren Notizen zur Vorlesung ‘Angewandte stochastische Prozesse II’, die ich als Vertreter des Lehrstuhles Weichselberger am Institut für Statistik der Ludwig-Maximilians Universität München im Wintersemester 1998/99 gehalten habe. Sie bot die Chance, die genannten wichtigen Konzepte noch einmal für sich selber zu überdenken. Der Be-

richt wurde am Institut für Biomathemathik und Biometrie am Forschungszentrum für Umwelt und Gesundheit erstellt. Die Arbeit stand im Zusammenhang mit einer Kooperation mit dem Institut für Statistik der Ludwig-Maximilians Universität im Rahmen des Sonderforschungsbereiches 386: 'Statistische Analyse diskreter Strukturen'. In diesem Zusammenhang möchte ich mich bei Prof. Ludwig Fahrmeir für die Zusammenarbeit bedanken. Besonderer Dank gebührt Andreas Martin vom IBB, der den Text sorgfältig durchgesehen hat.

Gerhard Winkler im Oktober 2000

Inhaltsverzeichnis

Kapitel 1

Markov-Chain-Monte-Carlo Methoden

Diese MCMC-Methoden haben bereits einen festen Platz in praktisch allen statistischen Anwendungen. Sie dienen der Simulation komplexer zufälliger Größen sowie der suboptimalen Berechnung von Schätzern in hochdimensionalen Parameterräumen. Eine besonders wichtige Rolle spielen sie bei der Untersuchung Bayesscher Modelle. Sie beruhen im wesentlichen auf der Konvergenz der Marginalverteilungen von (leicht simulierbaren) Markovketten, unter welchen die interessierende Verteilung stationär ist. Wir schicken eine elementare Einführung in endliche Markovketten voraus, die gleichzeitig als Grundlage für spätere Kapitel dient.

1.1 Endliche Markovketten

Viele zeitliche Abläufe (Prozesse), die in der Statistik betrachtet werden, weisen Abhängigkeiten von ihrer Vergangenheit auf. Etwa die Verkaufszahlen eines Ladens in einer Woche werden von der Zufriedenheit der Kunden mit den Einkäufen der Vorwoche abhängen, die heutige Luftverschmutzung von der gestrigen, die morgigen Börsenkurse von den heutigen, die Länge einer Warteschlange um drei Uhr von der Länge um zwei Uhr usw.

Die Modellierung voller Abhängigkeiten mag ein realistisches Abbild der Wirklichkeit liefern, sie wird aber zu großen Schwierigkeiten bei der wahrscheinlichkeitstheoretischen Behandlung und der statistischen Inferenz führen. Man sucht also nach Modellen, die genügend Abhängigkeit erlauben und gleichzeitig statistisch behandelbar sind. Eine reichhaltige und geeignete Klasse sind die Markovprozesse . Bei ihnen wird nur die Abhängigkeit vom letzten Zeitpunkt modelliert; das ist ein akzeptabler Kompromiß zwischen der Abhängigkeit von der gesamten Vergangenheit und völliger Unabhängigkeit. Sie liegen vielen der nachfolgenden Beispielen und Anwendungen zugrunde oder sind methodische oder didaktische Voraussetzung für allgemeinere Modelle. Weil sie so wichtig für diese Vorlesung sind, wiederholen wir die Grundlagen und geben einige Ableitungen.

Zusätzlich werden wir illustrieren, wie Markovprozesse in Form stochastischer Algorithmen als wirksame Hilfsmittel für Computersimulationen sowie für die Optimierung komplizierter Funktionen eingesetzt werden können.

1.1.1 Vorbereitungen

Ein (zeitdiskreter) *stochastischer Prozeß* ist eine Folge $\xi_0, \xi_1, \xi_2, \ldots$ von Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \boldsymbol{P})$. Dabei bezeichnet Ω die Menge der *Elementarereignisse*, $\mathcal{F}$ die σ-Algebra der in betracht zu ziehenden *Ereignisse* und $\boldsymbol{P}$ ein *Wahrscheinlichkeitsmaß* auf $\mathcal{F}$. [1] Im folgenden betrachten wir Prozesse, bei denen die Zufallsvariablen ξ_t nur Werte in einem endlichen Raum E annehmen können; o.E. können wir $E = \{1, \ldots, N\}$

[1] Elementarereignisse sind die 'kleinsten' Ereignisse ω, die wir formulieren können, zum Beispiel $\omega =$'es fällt die 0 beim Münzwurf', $\omega =$'es fällt die 3 beim Würfel', '$\omega =$'ich messe 2,72 Volt mit dem Voltmeter' oder $\omega =$'der Weg, den ein ganz bestimmtes Molekül im Lauf der Zeit nimmt' oder $\omega =$'der genaue Weg, wie der Skifahrer mit der Nummer 3 den Slalom nimmt'. Ereignisse A sind Mengen von Elementarereignissen, die von Interesse sind, z.B. A_0='der Würfel fällt auf eine gerade Zahl', $A_1 =$'die Messung liegt zwischen 2,5 und 2,9 Volt', $A_2 =$'der Slalomfahrer nimmt das erste Tor regelgemäß'. Die Menge der Ereignisse $\mathcal{F}$ folgt elementaren logischen Regeln. So ist $\Omega \in \mathcal{F}$, d.h. Ω ist ein Ereignis; das Ereignis Ω bedeutet, daß irgendeines der möglichen Elementarereignisse eintritt. Ist A ein Ereignis, so ist auch $\Omega \backslash A$ ('A tritt nicht ein') ein Ereignis. Damit folgt daß $\emptyset = \Omega \backslash \Omega \in \mathcal{F}$; das Ereignis $\emptyset$ bedeutet, daß nichts passiert. Sind A und B Ereignisse, so auch $A \cup B$. Z.B. wäre beim Slalom $A_1 =$'der Slalomfahrer nimmt das erste Tor' und $A_2 =$'der Slalomfahrer nimmt das zweite Tor', so ist $A_1 \cup A_2 =$'er nimmt das erste oder das zweite Tor' ein Ereignis. Aus technischen Gründen fordert man, daß die 'oder'-Beziehung nicht nur für endlich viele, sondern sogar für abzählbar viele Ereignisse gilt, d.h. sind $A_i \in \mathcal{F}$, $i = 1, 2, \ldots$, so ist auch $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ ein Ereignis. Das Wahrscheinlichkeitsmaß, eine nichtnegative Funktion auf $\mathcal{F}$ ($\boldsymbol{P} \geq 0$) spiegelt diese logische Struktur wider; es gilt $\boldsymbol{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \boldsymbol{P}(A_i)$ falls die A_i paarweise disjunkt sind; damit gilt z.B. $\boldsymbol{P}(\emptyset) = \boldsymbol{P}(\Omega) - \boldsymbol{P}(\Omega) = 0$, was unter der Prämisse, daß wir ein Zufallsexperiment durchführen wollen und somit irgendetwas herauskommen muß vernünftig ist. Desweiteren wählt man die (willkürliche aber praktische) Normierung $\boldsymbol{P}(\Omega) = 1$. Daraus folgt $0 \leq \boldsymbol{P}(A) \leq 1$ für alle $A \in \mathcal{F}$.

wählen. Damit erfassen wir z.B. binäre Zustände wie schwarz - weiß bei Bildern, krank - gesund usw., ordinale Merkmale und (endlich viele) numerische Größen. Die Indizes $t \in \{0, 1, 2, \ldots\} = \mathbb{N}_0$[2] interpretieren wir meist als (diskrete) Zeitpunkte, die Werte $x \in E$ nennen wir *Zustände* .

Wir wollen konditionale Modelle betrachten, d.h. solche bei denen der (zufällige) gegenwärtige Zustand ξ_t von den vergangenen $\xi_0, \ldots, \xi_{t-1}$ abhängt. Der einfachste Fall ist, daß ξ_t nur vom *letzten* Zustand ξ_{t-1} abhängt. Solche Prozesse lassen sich mathematisch noch recht gut handhaben. Das allgemeinere Modell läßt sich sogar formal auf diesen Fall zurückführen, allerdings hat man dann zu jedem Zeitpunkt t einen anderen Raum E_t.

Bemerkung 1.1.1 Bei physikalischen Anwendungen hat man im allgemeinen nur die Abhängigkeit vom letzten Zeitpunkt. Ob ein Luftmolekül zur Zeit t im Volumen V ist, hängt nur von seinem Ort zur Zeit $t-1$ ab und nicht von seinem vorherigem Schicksal. Ebenso hängt beim Roulette mein Kapital jeweils nur von der vorherigen Runde ab. Bei der Spracherkennung muß man natürlich mehrere Buchstaben in Folge kennen, um den nächsten vorherzusagen: bei 'h' ist der nächste Buchstabe nur schwer vorherzusagen, bei 'heut' wird es eher 'e' sein. Auch 'i' oder 'z' wären denkbar für 'heutig' oder 'heutzutage'. In solchen Situationen werden also Abhängigkeiten von mehreren vergangenen Zeitpunkten eine Rolle spielen. Schaltet man einen spannenden Krimi erst am Ende ein, so wird man die Story kaum mehr verstehen können. Man braucht die Kenntnis von Anfang an, um späteres vorhersagen zu können.

Wir wollen nun eine Markovsche (zufällige) zeitliche Abfolge von Zuständen modellieren. Zunächst soll der Startzustand zufällig sein. Dies drücken wir durch eine Startverteilung ν auf E aus, gemäß derer der Startzustand x_0 zufällig gewählt wird[3]. Um die Abhängigkeit darzustellen, brauchen wir *Übergangswahrscheinlichkeiten* $P_t(i,j) = p_{ij}^t$; sie geben an, mit welcher Wahrscheinlichkeit zur Zeit t der Zustand j eintritt, wenn zur Zeit $t-1$ der Zustand i vorlag. Für jedes i ist also der Vektor $(P_t(i,1), \ldots, P_t(i,N))$ ein Wahrscheinlichkeitsvektor, d.h. es gilt

$$P_t(i,j) \geq 0,\ i,j \in E,\ \sum_{j=1}^{N} P_t(i,j) = 1,\ i \in E.$$

[2] $\mathbb{N}_0$ ist die Menge der natürlichen Zahlen inklusive der Null

[3] d.h. durch Zahlen $\nu(i) \geq 0, i \in E$ mit $\sum_{i \in E} \nu(i) = 1$.

Man kann diese Vektoren in einer *Übergangsmatrix* , auch *Markovkern* genannt anordnen:

$$P_t = \begin{pmatrix} p_{11}^t & p_{12}^t & \cdots & p_{1N}^t \\ p_{21}^t & p_{22}^t & \cdots & p_{2N}^t \\ \vdots & \vdots & \vdots & \vdots \\ p_{N1}^t & p_{N2}^t & \cdots & p_{NN}^t \end{pmatrix}.$$

P_t ist eine *stochastische Matrix* , d.h. alle Einträge sind nichtnegativ und die Zeilensummen sind eins.

Wir zeigen nun, daß es stochastische Prozesse mit dieser Abhängigkeitsstruktur überhaupt gibt. Genauer: wir konstruieren einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \boldsymbol{P})$ und eine Folge $(\xi_t)_{t\in T}$ von Zufallsvariablen darauf, so daß gilt:

$$\boldsymbol{P}(\xi_0 = x) = \nu(x), \;\; \boldsymbol{P}(\xi_t = x|\xi_{t-1} = y) = P_t(x, y), \tag{1.1}$$

d.h. ν ist die Startverteilung und die Verteilung zur Zeit t bedingt durch den vorherigen Zeitpunkt ist durch die Übergangswahrscheinlichkeit P_t gegeben.

Eine Folge $(x_t)_{t\geq 0} = (x_0, x_1, \ldots)$ von Zuständen heißt *Pfad*. Er repräsentiert eine zeitliche Abfolge von (zufälligen) Zuständen, d.h. eine Realisierung des Prozesses. Die Menge aller Pfade ist der unendliche Produktraum E^T, wobei $T = \mathbb{N}_0$. Wir definieren nun eine Verteilung $\boldsymbol{P}$ auf $\Omega = E^T$. Als σ-Algebra wählen wir die Produkt-σ-Algebra $\mathcal{F}$; die einzelnen Komponenten $E^t = E$ von E^T sind mit der Potenzmenge als σ-Algebra versehen. Dann ist die *Projektion* ξ_t der Pfade auf den Zustand zur Zeit t, also

$$\xi_t : \Omega \longrightarrow E, \; (x_s)_{s\in T} \longmapsto x_t, \tag{1.2}$$

meßbar. Wir setzen

$$\boldsymbol{P}_t(\xi_0 = x_0, \ldots, \xi_t = x_t) = \nu(x_0) P_1(x_0, x_1) \cdots P_t(x_{t-1}, x_t), \tag{1.3}$$

wobei t und $(x_0, \ldots, x_t)$ beliebig sind. Dies ist eine Zähldichte auf der Menge $E^{\{0,\ldots,t\}}$ aller Pfade der Länge t und somit ist eine Wahrscheinlichkeit $\boldsymbol{P}_t$ auf E^t definiert. Die Folge $(\boldsymbol{P}_t)_{t\geq 0}$ von Wahrscheinlichkeiten ist *konsistent* und somit existiert ein Wahrscheinlichkeitsmaß $\boldsymbol{P}$ auf $(\Omega, \mathcal{F})$, so daß stets

$$\boldsymbol{P}(\xi_0 = x_0, \ldots, \xi_t = x_t) = \boldsymbol{P}_t(\xi_0 = x_0, \ldots, \xi_t = x_t) \tag{1.4}$$

gilt[4]. Damit haben wir aus ν und den P_t einen stochastischen Prozeß konstruiert.

[4]Dies garantiert der Konsistenzsatz von Kolmogorov, vgl. [3], Satz 35.3.

Definition 1.1.2 *Der in (1.2), (1.3) und (1.4) konstruierte Koordinatenprozeß heißt* kanonischer Prozeß *mit Startverteilung* ν *und Übergangswahrscheinlichkeiten* P_t.

Wir müssen nur noch nachweisen, daß (1.1) gilt. Auf dem Weg dorthin rechnen wir einige wichtige Verteilungen aus. Bei unseren Rechnungen nehmen wir stets an, daß die Nenner der vorkommenden Brüche strikt positiv sind. Sonst werden, wie üblich, alle bedingten Wahrscheinlichkeiten gleich null gesetzt.

Wir berechnen zuerst die *eindimensionalen Marginalverteilungen* ν_t, d.h. die Verteilungen $\nu_t = \boldsymbol{P} \circ \xi_t^{-1}$ der einzelnen Variablen ξ_t. Es ergibt sich

$$\begin{aligned} \nu_t(x) &= \boldsymbol{P}(\xi_t = x) = \sum_{(x_0,\dots,x_{n-1})\in E^{n-1}} \boldsymbol{P}((x_0,\dots,x_{n-1},x)) \qquad (1.5) \\ &= \sum_{(x_0,\dots,x_{t-1})\in E^{n-1}} \nu(x_0)P_1(x_0,x_1)\cdot\ldots\cdot P_t(x_{t-1},x) = \nu P_1\cdots P_t(x), \end{aligned}$$

wobei $P_1\cdots P_t$ als Matrizenprodukt interpretiert und ν aus Dimensionsgründen als Zeilenvektor aufgefaßt wird.

Sei nun $s < t$. Die *zweidimensionale Marginalverteilung* $\nu_{st} = \boldsymbol{P}\circ(\xi_s,\xi_t)^{-1}$ berechnet sich zu

$$\begin{aligned} &\nu_{st}(x,y) = \boldsymbol{P}(\xi_s = x, \xi_t = y) \\ &= \sum_{x_0,\dots,x_{s-1}} \sum_{x_{s+1},\dots,x_{t-1}} \boldsymbol{P}((x_0,\cdots,x_{s-1},x,x_{s+1},\cdots,x_{t-1},y)) \\ &= \nu P_1\dots P_s(x)P_{s+1}\dots P_t(x,y). \end{aligned}$$

Ähnlich wie oben lassen sich auch die *mehrdimensionalen Marginalverteilungen* $\boldsymbol{P}\circ(\xi_{t_1},\dots,\xi_{t_k})^{-1}$ berechnen; sie sind gegeben durch Ausdrücke der Gestalt

$$\boldsymbol{P}(\xi_{t_1} = x_1,\dots,\xi_{t_k} = x_k).$$

Damit ist unsere ursprüngliche Frage, ob wir den richtigen Prozeß konstruiert haben, leicht zu beantworten:

Theorem 1.1.3 *Der kanonische Prozeß* (ξ_t) *aus (1.2), (1.3) und (1.4) erfüllt (1.1).*

Beweis. Wir rechnen einfach nach, indem wir (1.5)und (1.6) benutzen:

$$\begin{aligned}\boldsymbol{P}(\xi_t = y|\xi_{t-1} = x) &= \frac{\boldsymbol{P}(\xi_{t-1} = x, \xi_t = y)}{\boldsymbol{P}(\xi_{t-1} = x)} = \frac{\nu_{t-1,t}(x,y)}{\nu_{t-1}(x)} \\ &= \frac{\nu P_1 \dots P_{t-1}(x) P_t(x,y)}{\nu P_1 \dots P_{t-1}(x)} = P_t(x,y).\end{aligned}$$

□

Wir haben den Prozeß absichtlich so konstruiert, daß er nur Abhängigkeiten vom letzten Zeitpunkt widerspiegelt. In der Tat ist für den gegenwärtigen Zustand nur der letzte relevant und nicht die gesamte Vorgeschichte. Formal bedeutet das:

Theorem 1.1.4 *Für den kanonischen Prozeß zu ν und (P_t) gilt*

$$\boldsymbol{P}(\xi_t = y|\xi_0 = x_0, \dots, \xi_{t-1} = x) = P_t(x,y) = \boldsymbol{P}(\xi_t = y|\xi_{t-1} = x). \tag{1.6}$$

Beweis. Dies ist wieder eine elementare Rechnung

$$\begin{aligned}&\boldsymbol{P}(\xi_t = y|\xi_0 = x_0, \dots, \xi_{t-1} = x) \\ &\quad = \frac{\boldsymbol{P}(\xi_0 = x_0, \dots, \xi_{t-1} = x, \xi_t = y)}{\boldsymbol{P}(\xi_0 = x_0, \dots, \xi_{t-1} = x)} \\ &\quad = \frac{\nu(x_0) P_1(x_0, x_1) \dots P_{t-1}(x_{t-2}, x) P_t(x,y)}{\nu(x_0) P_1(x_0, x_1) \dots P_{t-1}(x_{t-2}, x)} \\ &\quad = P_t(x,y) = \boldsymbol{P}(\xi_t = y|\xi_{t-1} = x).\end{aligned}$$

□

Weil dieser Begriff unsere obigen Forderungen zusammenfaßt, erhält er einen Namen, nämlich den des Erfinders:

Definition 1.1.5 *Ein stochastischer Prozeß (ξ_t) (auf irgendeinem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \boldsymbol{P})$) mit der Eigenschaft*

$$\boldsymbol{P}(\xi_t = y|\xi_0 = x_0, \dots, \xi_{t-1} = x) = P_t(x,y) = \boldsymbol{P}(\xi_t = y|\xi_{t-1} = x) \tag{1.7}$$

heißt Markovkette *oder Markovprozeß mit Übergangswahrscheinlichkeiten P_t. Die Identität (1.7) definiert die* Markoveigenschaft*.*

Wir können den letzten Satz nun umformulieren:

Theorem 1.1.6 *Der kanonische stochastische Prozeß zu ν und (P_t) ist ein Markovprozeß.*

Geht es nur um die Verteilung eines Prozesses, so können wir stets mit dem kanonischen Prozeß arbeiten.

Wir geben eine zweite Konstruktion an, die zweierlei Vorteile hat. Einerseits benötigen wir keine starke Version des Konsistenzsatzes von Kolmogorov - sondern nur die Existenz einer Folge von i.i.d. Variablen[5], bzw.des Produktmaßes $\boldsymbol{P} = \lambda^T$ auf $(0,1]^T$, wobei λ das Lebesgue-Maß auf $(0,1]$ bezeichnet[6]. Zum zweiten ergibt sich daraus unmittelbar ein Programm zur Computer-Simulation des Prozesses.

Wir simulieren zuerst eine einzelne Zufallsvariable mit endlich vielen Zuständen. Im Prinzip werden alle simulierbaren Zufallsvariablen mit Hilfe von auf dem Einheitsintervall $(0,1]$ gleichverteilten Zufallsvariablen U erzeugt. Diese stehen in jedem Computerprogramm als Prozedur zur Verfügung.

Simulation 1.1.7 Wir wollen eine Zufallsvariable η simulieren, welche Werte in der Menge $\{1,\ldots,N\}$ mit jeweils den Wahrscheinlichkeiten $p_1,\ldots,p_N$ annimmt. Dazu generieren wir eine Zufallsvariable U, welche im Einheitsintervall $(0,1]$ gleichverteilt ist. Dann partitionieren wir $(0,1]$ in N sukzessive Intervalle $I_k = (a_{k-1}, a_k]$, $k = 1,\ldots,N$ der Längen $p_k = a_k - a_{k-1}$. Wir setzen also $a_0 = 0, a_1 = p_1$, so daß $\lambda(a_0, a_1] = p_1$, $a_2 = p_1 + p_2$, so daß $\lambda(a_1, a_2] = p_2$, usw. und schließlich $a_{N-1} = \sum_{i=1}^{N-1} p_k$, $a_N = 1$, so daß $\lambda((a_{N-1}, 1]) = p_N$ ist. Definieren wir nun

$$\eta : (0,1] \mapsto \{1,\ldots,N\}$$

durch

$$\eta(u) = k \iff U(u) \in I_k,$$

so gilt

$$\boldsymbol{P}(\eta = k) = \lambda(I_k) = p_k$$

und η hat die gewünschte Verteilung. Formal hat η die Gestalt

$$\eta = \sum_{k=1}^{N} k \cdot \chi_{\left(\sum_{i=0}^{k-1} a_i, \sum_{i=0}^{k} a_i\right]}(U). \tag{1.8}$$

[5]Die Abkürzung i.i.d. bedeutet 'independent, identically distributed', also unabhängig, identisch verteilt.

[6]gegeben durch $\lambda((a,b]) = b - a$, das natürliche euklidische Maß

Ein Pseudocode für ein Programm hätte z.B. die Gestalt

```
Procedure Uniform;
{Liefert eine in (0,1] gleichverteilte Zufallsvariable U;
bei wiederholtem Aufruf sind die Werte unabhängig.}

Procedure DZV (Input: p[1],...,p[N]; Output: η);
{liefert eine diskrete Zufallsvariable DZV mit der
Zähldichte p[k]}
BEGIN
U:=Uniform; i:=1; CDF:=p[1];
  WHILE (CDF<U) DO
   BEGIN i:=SUCC(i); CDF:=CDF+p[i] END;
  DZV:=i
END;{DZV}
```

Damit können wir nun leicht eine gesamte endliche Markovkette simulieren. Die folgende Konstruktion liefert gleichzeitig einen *zweiten Existenzbeweis* für Markovketten mit vorgegebenen Übergangswahrscheinlichkeiten .

Simulation 1.1.8 Um den Markovprozeß zu simulieren, erzeugt man zunächst die Startvariable ξ_0 gemäß 1.1.7 mit $p_k = \nu(k)$. Dann ruft man weitere i.i.d. U's auf, und erzeugt die Variablen ξ_t, wobei man als $p[1], \dots, p[N]$ die Wahrscheinlichkeiten $p[j] = P_t(\xi_{t-1}, j),\ 1 \le j \le N$, übergibt. In Programmform sieht das so aus:

```
Procedure MarkovChain {INPUT:ν,(P_t)_t ∈ T; OUTPUT: (ξ_t)_{t∈T}}
BEGIN
 ξ_0:=DZV(ν[1],...,ν[N]);
 FOR t=1 TO ∞ DO ξ_t:=DZV(P_t(ξ_{t-1},1),...,P_t(ξ_{t-1},N))
END;{MarkovChain}
```

Formal entspricht das folgendem: Man konstruiert zunächst ξ_0 gemäß (1.8) zu ν. Sodann definiert man

$$f(i,u,t) = \sum_{k=0}^{N} k \cdot \chi_{(\sum_{i=0}^{k-1} p_{ik}^t, \sum_{i=0}^{k} p_{ik}^t]}(u),$$

so daß $f(i,u,t) = k$ genau dann wenn

$$u \in \left(\sum_{i=0}^{k-1} p_{ik}^t, \sum_{i=0}^{k} p_{ik}^t \right].$$

Dann setzt man für $t = 0, 1, \ldots$ rekursiv

$$\eta_t = f(\eta_{t-1}, U_t, t),$$

wobei $(U_t)_t$ eine i.i.d.-Folge von im Einheitsintervall gleichverteilten Zufallsvariablen ist. Dieser Prozeß hat die richtige Verteilung und die Formeln entsprechen obiger Prozedur. Die i.i.d. Variablen U_t erzeugt man durch wiederholtes Aufrufen des Zufallsgenerators, wobei man nur zu beachten hat, daß man nicht immer dieselbe Saat (seed) benutzt.

1.1.2 Der Kontraktionskoeffizient

Der Kontraktionskoeffizient ist ein wichtiges Hilfsmittel zur Untersuchung des asymptotischen Verhaltens von Markovketten. In diesem Abschnitt bezeichnen E eine endliche Menge, P eine Übergangswahrscheinlichkeit oder einen Markovkern auf E und μ, ν, ϱ usw. Verteilungen (Wahrscheinlichkeitsmaße, Wahrscheinlichkeitsvektoren) auf E. Um die Konvergenz zu messen, benötigen wir eine Metrik. $\mathcal{P}$ bezeichne die Menge aller Wahrscheinlichkeitsmaße auf E. Für $\mu, \nu \in \mathcal{P}$ sei

$$\|\mu - \nu\| = \sum_{x \in E} |\mu(x) - \nu(x)|$$

die (Norm der) *totale*(n) *Variation*. Dies ist einfach der l_1-Abstand.

Lemma 1.1.9 *Es gilt*

$$\begin{aligned} \|\mu - \nu\| &= 2 \sum_x (\mu(x) - \nu(x))^+ \\ &= 2 \left(1 - \sum_x \min\{\mu(x), \nu(x)\} \right) \\ &= \max \left\{ \left| \sum_x h(x)(\mu(x) - \nu(x)) \right| : |h| \le 1 \right\}. \end{aligned}$$

Beweis. Elementare Rechnung liefert:

$$\begin{aligned}\|\mu-\nu\| &= \sum_x (\mu(x)-\nu(x))^+ + \sum_x (\mu(x)-\nu(x))^- \\ &= \sum_{x:\mu(x)\geq\nu(x)} (\mu(x)-\nu(x)) + \sum_{x:\mu(x)<\nu(x)} (\nu(x)-\mu(x)).\end{aligned}$$

Die Differenz der beiden Summen verschwindet und deshalb sind sie gleich. Damit gilt

$$\|\mu-\nu\| = 2\cdot\left(\sum_x (\mu(x)-\nu(x))^+\right)$$

und somit die erste Identität. Die zweite folgt aus:

$$\begin{aligned}&\|\mu-\nu\| \\ &= \sum_{x:\mu(x)\geq\nu(x)} \mu(x) - \sum_{x:\mu(x)\geq\nu(x)} \nu(x) - \sum_{x:\mu(x)<\nu(x)} \mu(x) + \sum_{x:\mu(x)<\nu(x)} \nu(x) \\ &= \sum_x \mu(x) + \sum_x \nu(x) - 2\left(\sum_{x:\mu(x)<\nu(x)} \mu(x) + \sum_{x:\mu(x)\geq\nu(x)} \nu(x)\right) \\ &= 2\left(1-\sum_x \mu(x)\wedge\nu(x)\right).\end{aligned}$$

Die Ungleichung

$$\begin{aligned}\|\mu-\nu\| &= \sum_x |\mu(x)-\nu(x)| \\ &\geq \max\left\{\left|\sum_x h(x)(\mu(x)-\nu(x))\right| : |h|\leq 1\right\}\end{aligned}$$

ist klar. Um die Gleichheit zu zeigen, setzt man

$$h(x) = \operatorname{sgn}(\mu(x)-\nu(x))$$

ein. □

Die Mischungseigenschaft eines Markovkernes läßt sich durch den (Dobruschinschen) *Kontraktionskoeffizienten* $c(P)$ messen:

$$c(P) = \frac{1}{2}\max_{x,y} \|P(x,\cdot)-P(y,\cdot)\|.$$

Bemerkung 1.1.10 Es gilt offensichtlich

$$0 \leq \|\mu - \nu\| \leq 2,$$

wobei $\|\mu(x) - \nu(x)\| = 2$, genau dann, wenn μ und ν *orthogonal* sind, d.h. disjunkte Träger haben. Somit gilt

$$0 \leq c(P) \leq 1,$$

wobei $c(P) = 0$ genau dann, wenn alle $P(x; \cdot)$, $x \in E$, gleich sind, und $c(P) = 1$, wenn wenigstens zwei der $P(x, \cdot)$ orthogonal sind.

Die beiden Ungleichungen im folgenden Lemma sind fundamental.

Lemma 1.1.11 *Auf $\mathcal{P}$ gelten stets*

$$\begin{aligned} \|\mu P - \nu P\| &\leq c(P)\|\mu - \nu\| \\ c(PQ) &\leq c(P)c(Q). \end{aligned}$$

Insbesondere gelten

$$\begin{aligned} \|\mu P - \nu P\| &\leq \|\mu - \nu\|, \\ \|\mu P - \nu P\| &\leq 2c(P). \end{aligned}$$

Beweis. Wir beweisen die erste Ungleichung. Sei f eine reelle Funktion auf E und

$$d = (\max_x f(x) + \min_x f(x))/2.$$

Dann gilt

$$\max_x |f(x) - d| = (1/2) \max_{x,y} |f(x) - f(y)|.$$

Wir schreiben kurz $\mu(f)$ für $\sum_x f(x)\mu(x)$ und folgern

$$\begin{aligned} |\mu(f) - \nu(f)| &= |\mu(f - d) - \nu(f - d)| \leq \max_x |f(x) - d| \cdot \|\mu - \nu\| \\ &= (1/2) \max_{x,y} |f(x) - f(y)| \cdot \|\mu - \nu\|. \end{aligned} \tag{1.9}$$

Für eine Funktion h auf E ist Ph definiert durch

$$Ph(x) = \sum_y h(y)P(x, y).$$

Einsetzen von Ph für f ergibt

$$\begin{aligned}
\|\mu P - \nu P\| &= \max\{|(\mu P)h - (\nu P)h| : |h| \leq 1\} \\
&= \max\{|\mu(Ph) - \nu(Ph)| : |h| \leq 1\} \\
&\leq \max\left\{(1/2)\max_{x,y}|Ph(x) - Ph(y)| : |h| \leq 1\right\}\|\mu - \nu\| \\
&= (1/2)\max_{x,y}\max\{|Ph(x) - Ph(y)| : |h| \leq 1\}\|\mu - \nu\| \\
&= c(P)\|\mu - \nu\|.
\end{aligned}$$

Damit ist die erste Ungleichung bewiesen. Die zweite Ungleichung folgt aus

$$\begin{aligned}
c(PQ) &= (1/2)\max_{x,y}\|PQ(x.\cdot) - PQ(y,\cdot)\| \\
&= (1/2)\max_{x,y}\|P(x,\cdot)Q - P(y,\cdot)Q\| \\
&\leq c(P)c(Q).
\end{aligned}$$

Die übrigen Ungleichungen folgen aus den bewiesenen weil $c(P) \leq 1$ und $\|\mu - \nu\| \leq 2$ ist. Damit ist der Beweis vollständig. □

Schließlich benötigen wir noch eine einfache Abschätzung.

Lemma 1.1.12 *Für jeden Markovkern auf einem endlichen Raum E gilt*

$$c(Q) \leq 1 - |E|\min\{Q(x,y) : x, y \in E\} \leq 1 - \min\{Q(x,y) : x, y \in E\}$$

. Insbesondere gilt für strikt positives Q, daß $c(Q) < 1$.

Beweis. Mit Lemma 1.1.9 folgt

$$\|\mu - \nu\|/2 = 1 - \sum_x \min\{\mu(x), \nu(x)\}$$

für Verteilungen μ und ν. Also gilt

$$c(Q) = 1 - \min\left\{\sum_z Q(x,z) \wedge Q(y,z) : x, y \in E\right\}$$

woraus die erste Ungleichung folgt. Der Rest ist klar. □

1.1.3 Homogene Markovketten

Eine homogene Markovkette ist gegeben durch eine Startverteilung ν und einen Markovkern P; d.h. daß in obiger Notation $P_t = P$ für alle $t = 1, 2, \ldots$. Die n-te Marginalverteilung ist

$$\nu_n = \nu P^n,$$

insbesondere gilt also $\nu_0 = \nu$. Wir interessieren uns für die Konvergenz der ν_n. Gute Kandidaten für Grenzverteilungen

$$\nu_\infty = \lim_{n\to\infty} \nu_n$$

sind *invariante Verteilungen* μ mit $\mu P = \mu$, denn die Ketten konvergieren zumindest für das Startmaß $\nu_0 = \mu$:

$$\mu P^n = (\mu P)P^{n-1} = \mu \longrightarrow \mu, \; n \to \infty.$$

Damit diese Konvergenz auch für den Start in anderen Verteilungen stattfindet, muß P den Raum E ‘durchmischen‘. Dies ist der Fall bei strikt positiven P: man kommt in einem Schritt von jedem x zu jedem y mit positiver Wahrscheinlichkeit. Etwas allgemeiner sind *primitive Markovkerne* für die es ein $\tau \in \mathsf{N}$ gibt, so daß P^τ strikt positiv ist.

Lemma 1.1.13 *Für jeden strikt positiven Markovkern Q ist $c(Q) < 1$.*

Beweis. Dies folgt unmittelbar aus Lemma 1.1.12. □

Daraus folgt schon der Konvergenzsatz.

Theorem 1.1.14 *Sei P primitiv mit invarianter Verteilung μ. Dann gilt*

$$\nu P^n \longrightarrow \mu, \; n \longrightarrow \infty,$$

gleichmäßig in allen $\nu \in \mathcal{P}$.

Beweis. Es gilt für große n, daß

$$\begin{aligned} \|\nu P^n - \mu\| &= \|\nu P^n - \mu P^n\| \leq \|\nu - \mu\| c(P^n) \\ &\leq \; 2c(P^{k\cdot\tau})c(P^{n-k\cdot\tau}) \leq 2(c(P)^\tau)^k \longrightarrow 0, \end{aligned}$$

wodurch der Satz bewiesen ist. □

Theorem 1.1.15 *Jeder primitive Markovkern besitzt genau eine invariante Verteilung. Diese ist strikt positiv.*

Beweis. Da die Zeilensummen von Markovkernen gleich 1 sind, hat P den Eigenwert 1 mit Rechtseigenvektoren $(a,\ldots,a)^*$. Da 1 auch Eigenwert der Transponierten P^* ist, gibt es $\nu^* \in \mathbb{R}^E \backslash \{0\}$ (Das Symbol $\mathbb{R}$ bezeichnet die Menge der reellen Zahlen) mit $P^*\nu^* = \nu^*$ und somit $\nu P = \nu$. Sei nun $P > 0$. Hätte ν eine strikt positive und eine strikt negative Komponente, dann würden Invarianz, $P > 0$ und $\sum_y P(x,y) = 1$ implizieren, daß

$$\sum_y |\nu(y)| = \sum_y \left| \sum_x \nu(x) P(x,y) \right| < \sum_x \sum_y |\nu(x)| P(x,y) = \sum_x |\nu(x)|.$$

Dies ist ein Widerspruch und wir dürfen $\nu(x) \geq 0$ für jedes x annehmen. Da $\nu \neq 0$, gilt

$$\nu(y) = \sum_x \nu(x) P(x,y) > 0 \text{ für jedes } y.$$

Schließlich liefert Normalisierung von ν eine invariante strikt positive Verteilung μ. Die Eindeutigkeit folgt schon aus Satz 1.1.14.
Sei P schließlich primitiv mit $P^\tau > 0$. Wieder hat P den Eigenwert 1 und es gibt $\nu \in \mathbb{R}^E \backslash \{0\}$, so daß $\nu P = \nu$ und somit auch $\nu P^\tau = \nu$. Da nun $P^\tau > 0$, hat P^τ eine eindeutig bestimmte invariante Verteilung. Wegen der Eindeutigkeit gilt $\mu = \nu$. Somit ist μ auch die eindeutig bestimmte invariante Verteilung von P und sie ist überdies strikt positiv. Ist $\tilde{\nu}$ eine weitere invariante Verteilung von P, so ist $\tilde{\nu}$ auch P^τ-invariant und deshalb gleich ν.
□

Die Approximation von Erwartungswerten beruht auf einem Gesetz der großen Zahlen. Für eine Startverteilung ν und einen Markovkern P sei $(\xi_i)_{i \geq 0}$ die zum induzierten Markovprozeß gehörige Folge der zufälligen Zustände zu den Zeitpunkten $i = 0, 1, \ldots$. Der Erwartungswert $\sum_x f(x)\mu(x)$ einer Funktion f auf E bezüglich einer Verteilung μ sei mit $\boldsymbol{E}_\mu(f)$ bezeichnet.

Theorem 1.1.16 (Schwaches Gesetz der großen Zahlen) *Sei E eine endliche Menge und sei P ein Markovkern auf E mit invarianter Verteilung μ und $c(P) < 1$. Dann gilt für jede Startverteilung ν und jede Funktion f auf E, daß*

$$\frac{1}{n} \sum_{i=1}^n f(\xi_i) \longrightarrow \boldsymbol{E}_\mu(f)$$

in $L^2(\boldsymbol{P}_\nu)$. Darüber hinaus gilt für jedes $\varepsilon > 0$, daß

$$\boldsymbol{P}_\nu\left(\left|\frac{1}{n}\sum_{i=1}^n f(\xi_i) - \boldsymbol{E}_\mu(f)\right| > \varepsilon\right) \leq \frac{13\|f\|_2^2}{(1-c(P))n\varepsilon^2}$$

wobei $\|f\|_2^2 = \sum_x |f(x)|^2$.

Für i.i.d. Zufallsvariablen ξ_i hängen die Größen $P(x,y)$ nicht von x ab; somit sind alle Zeilen der zu P gehörigen Matrix gleich und es ist $c(P) = 0$. In diesem Fall erhält man das übliche schwache Gesetz der großen Zahlen.

Beweis. Für jedes $x \in E$ sei $g_x = (1/n)\sum_{i=1}^n \chi_{\{\xi_i = x\}} - \mu(x)$. Dann gilt

$$\begin{aligned} \boldsymbol{E}_\nu\left(\left(\frac{1}{n}\sum_{i=1}^n f(\xi_i) - \boldsymbol{E}_\mu(f)\right)^2\right) &= \boldsymbol{E}_\nu\left(\left(\sum_x f(x) g_x\right)^2\right) \\ \leq \boldsymbol{E}_\nu\left(\left(\sqrt{\sum_x f(x)^2}\sqrt{\sum_x g_x^2}\right)^2\right) &= \|f\|_2^2 \sum_x \boldsymbol{E}_\nu(g_x^2). \end{aligned} \tag{1.10}$$

Man berechnet

$$\begin{aligned} &\sum_x \boldsymbol{E}_\nu\left(g_x^2\right) \\ &= \sum_x \boldsymbol{E}_\nu\left(\left(\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{\xi_i = x\}} - \mu(x)\right)^2\right) \\ &= \sum_x \frac{1}{n^2}\sum_{i,j=1}^n \boldsymbol{E}_\nu((\chi_{\{\xi_i = x\}} - \mu(x))(\chi_{\{\xi_j = x\}} - \mu(x))) \\ &= \frac{1}{n^2}\sum_{i,j=1}^n \sum_x \big((\nu_{ij}(x,x) - \mu(x)^2) \\ &\quad -(\mu(x)\nu_j(x) - \mu(x)^2) - (\mu(x)\nu_i(x) - \mu(x)^2)\big). \end{aligned}$$

Drei Mittel sind zu schätzen. Schwierig ist nur das erste. Da $\mu P = \mu$, gelten für $i, k > 0$ die folgenden Abschätzungen:

$$\sum_x |\nu P^i(x)\varepsilon_x P^k(x) - \mu(x)^2|$$

$$
\begin{aligned}
&\le \sum_x |\nu P^i(x)\varepsilon_x P^k(x) - \mu P^i(x)\varepsilon_x P^k(x)| + |\mu(x)\varepsilon_x P^k(x) - \mu(x)\mu P^k(x)| \\
&\le \sum_x |\nu P^i(x) - \mu P^i(x)| + |\varepsilon_x P^k(x) - \mu P^k(x)| \\
&\le \|(\nu-\mu)P^i\| + \|(\varepsilon_x - \mu)P^k\| \le 2\cdot(c(P)^i + c(P)^k).
\end{aligned}
$$

Daraus folgt

$$
\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_x |\nu_{ij}(x,x) - \mu(x)^2| &\le \frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(c(P)^i + c(P)^{j-i}\right) \\
&\le \frac{4}{n}\sum_{i=0}^{\infty} c(P)^i = \frac{4}{n(1-c(P))}.
\end{aligned}
$$

Die gleiche Abschätzung gilt für das Mittel über Paare (i,j) von Indizes $j < i$. Wir setzen $\nu_{ii}(x,x) = \nu P^i(x)$ und $\nu_{ii}(x,y) = 0$ wenn $x \neq y$. Die Summe über die entsprechenden Terme ist durch n beschränkt und deshalb gilt

$$
\frac{1}{n^2}\sum_{i,j=1}^{n}\sum_x |\nu_{ij}(x,x) - \mu(x)^2| \le \frac{9}{n(1-c(P))}.
$$

Mit Hilfe von (1.10) können das zweite und dritte Mittel abgeschätzt werden:

$$
\begin{aligned}
&\frac{\mu(x)}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_x |\nu_j(x) - \mu(x)| \\
&\le \frac{1}{n}\sum_{j=1}^{n} \|\nu P^j - \mu P^j\| \\
&\le \frac{2}{n}\sum_{i=0}^{\infty} c(P)^j \le \frac{2}{n(1-c(P))}.
\end{aligned}
$$

Also gilt

$$
\sum_x \boldsymbol{E}(g_x^2) \le \frac{13}{n(1-c(P))}.
$$

Zusammen mit (1.10) ist damit der erste Teil der Behauptung bewiesen. Der zweite Teil folgt daraus mit der Markovschen Ungleichung. $\square$

1.1.4 Inhomogene Markovketten

In Verallgemeinerung der homogenen Markovketten ist eine *inhomogene Markovkette* gegeben durch eine Startverteilung ν und eine Folge von Markovkernen $P_i,\ i \geq 1$. Die n-te Marginalverteilung ist

$$\nu P_1 P_2 \cdots P_n.$$

Wir beweisen wieder einen Konvergenzsatz für die Marginalverteilungen.

Lemma 1.1.17 *Seien μ_n, $n \geq 1$, Verteilungen auf E, so daß*

$$\sum_n \|\mu_{n+1} - \mu_n\| < \infty$$

. Dann gibt es eine Verteilung μ_∞, so daß $\mu_n \to \mu_\infty$ in $\|\cdot\|$, wenn $n \to \infty$.

Da E endlich ist, fallen punktweise und L^1-Normkonvergenz bzgl. $\|\cdot\|$ zusammen.

Beweis. Für $m < n$ gilt

$$\|\mu_n - \mu_m\| \leq \sum_{k \geq m} \|\mu_{k+1} - \mu_k\|.$$

Der letztere Ausdruck konvergiert mit $m \to \infty$ gegen 0. Deshalb ist (μ_n) eine Cauchyfolge im kompakten Raum $\{\mu \in \mathbb{R}^E : \mu \geq 0, \sum_x \mu(x) = 1\}$ und hat somit einen Grenzwert μ_∞ darin. $\square$

Theorem 1.1.18 *Seien P_n, $n \geq 1$, Markovkerne und jedes P_n besitze eine invariante Verteilung μ_n. Ferner seien die folgenden Bedingungen erfüllt:*

$$\sum_n \|\mu_n - \mu_{n+1}\| < \infty, \tag{1.11}$$

$$\lim_{n\to\infty} c(P_i \dots P_n) = 0 \quad \textit{für jedes} \quad i \geq 1. \tag{1.12}$$

Dann existiert eine Verteilung μ_∞, für die

$$\nu P_1 \dots P_n \longrightarrow \mu_\infty \quad \textit{für} \quad n \to \infty.$$

gleichmäßig in allen Startverteilungen ν gilt.

Beweis. Die Existenz des Grenzwertes μ_∞ wurde im Lemma nachgewiesen. Seien nun $i \geq 1$ und $k \geq 1$. Die Invarianz liefert

$$\mu_i P_i \dots P_{i+k} = \mu_i P_{i+1} \dots P_{i+k}.$$

Außerdem gilt

$$\sum_{j=1}^{k} (\mu_{i-1+j} - \mu_{i+j}) P_{i+j} \dots P_{i+k} = \mu_i P_{i+1} \dots P_{i+k} - \mu_{i+k}.$$

Zum Beweis: der erste Term in der Summe ist

$$\mu_i P_{i+1} \dots P_{i+k},$$

der letzte ist gleich

$$-\mu_{i+k} P_{i+k} = \mu_{i+k}$$

und die Kombination von Paaren aufeinanderfolgender Terme liefert

$$\begin{aligned} &-\mu_{i+j} P_{i+j} \dots P_{i+k} + \mu_{i+j} P_{i+j+1} \dots P_{i+k} \\ &= \; -\mu_{i+j} P_{i+j+1} \dots P_{i+k} + \mu_{i+j} P_{i+j+1} \dots P_{i+k} = 0. \end{aligned}$$

Damit gilt

$$\begin{aligned} &\mu_\infty P_i \dots P_{i+k} - \mu_\infty \\ &= \; (\mu_\infty - \mu_i) P_i \dots P_{i+k} + \mu_i P_{i+1} \dots P_{i+k} - \mu_\infty \\ &= \; (\mu_\infty - \mu_i) P_i \dots P_{i+k} + \sum_{j=1}^{k} (\mu_{i-1+j} - \mu_{i+j}) P_{i+j} \dots P_{i+k} \\ &\quad + \mu_{i+k} - \mu_\infty. \end{aligned}$$

Daraus folgt

$$\|\mu_\infty P_i \dots P_{i+k} - \mu_\infty\| \leq 2 \cdot \sup_{n \geq i} \|\mu_\infty - \mu_n\| + \sum_{n \geq i} \|\mu_n - \mu_{n+1}\|. \tag{1.13}$$

Verwendet wurden dabei Lemma 1.1.11 und die Beschränktheit des Kontraktionskoeffizienten durch 1. Mit Bedingung (1.11) und weil μ_∞ existiert, wird für großes i der Ausdruck auf der rechten Seite klein. Für $2 \leq i \leq n$ fahren wir fort mit

$$\begin{aligned} &\|\nu P_i \dots P_n - \mu_\infty\| \\ &= \|(\nu P_i \dots P_n - \mu_\infty P_i \dots P_n) + \mu_\infty P_i \dots P_n - \mu_\infty\| \\ &\leq 2 \cdot c(P_i \dots P_n) + \|\mu_\infty P_i \dots P_n - \mu_\infty\|. \end{aligned} \tag{1.14}$$

Für große n wird der erste Term wegen (1.12) klein. Damit ist der Beweis vollständig. □

Der Satz heißt auch Satz von DOBRUSHIN (DOBRUSHIN (1956)). Einige einfache Kriterien für die Gültigkeit der Voraussetzungen sind:

Lemma 1.1.19 *Für Verteilungen μ_n, $n \geq 1$ auf einem endlichen Raum E gilt Bedingung (1.11) falls jede der Folgen $(\mu_n(x))_{n\geq 1}$ schließlich fällt oder wächst..*

Beweis. Nach Lemma 1.1.10 gilt

$$0 \leq \sum_n \|\mu_{n+1} - \mu_n\| = 2 \sum_x \sum_n (\mu_{n+1}(x) - \mu_n(x))^+.$$

Wegen der Monotonie gibt es n_0 so daß entweder $(\mu_{n+1}(x) - \mu_n(x))^+ = 0$ für alle $n \geq n_0$ und somit $\sum_{n\geq n_0}(\mu_{n+1}(x) - \mu_n(x))^+ = 0$ oder $(\mu_{n+1}(x) - \mu_n(x))^+ = \mu_{n+1}(x) - \mu_n(x)$ und somit

$$\sum_{n=n_0}^{N} (\mu_{n+1}(x) - \mu_n(x))^+ = \mu_{N+1}(x) - \mu_{n_0}(x) \leq 1$$

für alle großen N. Deshalb ist die Doppelsumme endlich und Bedingung (1.11) ist erfüllt. Damit ist der Beweis vollständig. □

Lemma 1.1.20 *Bedingung (1.12) ist erfüllt, falls*

$$\prod_{k\geq i} c(P_k) = 0 \quad \textit{für jedes} \quad i \geq 1. \tag{1.15}$$

oder falls

$$c(P_n) > 0 \quad \textit{für jedes } n \textit{ und} \quad \prod_{k\geq 1} c(P_k) = 0. \tag{1.16}$$

Beweis. Bedingung (1.15) impliziert (1.12) wegen der zweiten Regel in Lemma 1.1.11, und (1.16) impliziert offensichtlich (1.15). □

1.2 Sampling und Annealing

Im Kapitel Sampling und Annealing[7] befassen wir uns mit der Simulation und der Schätzung von Parametern aus Verteilungen Π, die einer direkten Behandlung nicht zugänglich sind.

Wir beschränken uns auf strikt positive Verteilungen $\Pi > 0$. Diese können stets in der *Gibbsschen Form*

$$\Pi(x) = \frac{1}{Z}\exp(-H(x)),\ Z = \sum_{z\in E}\exp(-H(z)),$$

geschrieben werden. Π heißt dann *Gibbsverteilung* zur *Energiefunktion* H; die Normierung Z heißt *Zustandssumme* . Man kann z.B. immer setzen

$$H(x) = -\ln\Pi(x) - \ln Z,$$

wobei Z irgendeine positive Zahl ist. Dann gilt nämlich

$$\exp(-H(x)) = \Pi(x)\cdot Z.$$

Wir stellen uns folgende Aufgaben:

1. Simuliere gemäß Π, d.h. ziehe Stichproben, die gemäß Π verteilt sind;
2. Schätze Parameter in Π;
3. Finde eine Maximalstelle x^* von $x \mapsto \Pi(x)$.

Die erste Aufgabe werden wir mit Hilfe des Konvergenzsatzes für homogene Markovketten lösen und die zweite mit dem schwachen Gesetz der großen Zahlen für homogene Markovketten. Die daraus abgeleiteten Algorithmen heißen *Sampler*. Für die dritte Aufgabe müssen wir entsprechende Aufgaben für inhomogene Markovketten lösen; sie hat mit der Bestimmung von Maximumlikelihood-Schätzern zu tun. Dies behandeln wir später.

Zunächst formulieren wir noch einige Bemerkungen, die die Notwendigkeit solcher Verfahren begründen sollen.

In der Analyse von Signalen, Bildern oder örtlichen Verteilungen in der Epidemiologie hat man es mit sehr großen Grundräumen zu tun. Diese sind

[7]Für Kapitel 2 und 3 verweisen wir auf G. WINKLER: Image Analysis, Random Fields and Dynamic Monte Carlo Methods. A Mathematical Introduction. Springer Verlag (1995), [17].

oft von Produktgestalt $E = \prod_{s \in S} E_s$. Bei Bildern z.B. wäre S ein Pixelraster und E_s die Menge der möglichen Grauwerte in s. In der Epidemiologie wäre S vielleicht die Menge von Landkreisen und E_s eine Fallzahl im Landkreis s usw. Aufgrund der hohen Zahl der Komponenten von $x \in E$ stehen analytische Hilfsmittel zur Bestimmung von x^* im kontinuierlichen Fall nur in Spezialfällen zur Verfügung. Sind die Komponenten von E diskret, so ist die Mächtigkeit von E leicht in der Größenordnung $256^{256 \times 256}$; für allgemeine H sind keine diskreten Optimierungsmethoden bekannt. Die Erwartungswerte für (2) können ebenfalls nicht direkt berechnet werden, da die Zustandssumme Z über $|E|$ Terme gebildet wird. Einen möglichen Ausweg bilden stochastische Verfahren, die auf bedingten Verteilungen beruhen. Diese sind nämlich für eine große Klasse praktisch relevanter Energiefunktionen handhabbar.

1.2.1 Bedingte Verteilungen

Wir berechnen bedingte Verteilungen für eine praktisch wichtige Klasse von Gibbsmaßen.

Beispiel 1.2.1 (Potentiale) Sei $E = \Pi_{x \in S} E_s$ und H von der Gestalt

$$H(x) = \sum_{C \in \mathcal{C}} V_C(x),$$

wobei V_C nur von der Restriktion x_C von x auf C abhängt und $\mathcal{C}$ aus kleinen Teilmengen C von S besteht. Dann gilt für die bedingte Verteilung

$$\Pi(x_s | x_{S \setminus s})$$

von $\Pi \propto \exp(-H(x))$, daß

$$\begin{aligned}
\Pi(x_s | x_{S\setminus s}) &= \frac{Z^{-1} \exp(-H(x_s x_{S\setminus s}))}{Z^{-1} \sum_{z_s} \exp(H(z_s x_{S\setminus s}))} \\
&= \frac{\exp(-\sum_{C \in \mathcal{C}, C \ni s} V_C(x_s x_{S\setminus s})) \exp(-\sum_{C \in \mathcal{C}, C \not\ni s} V_C(x_C))}{\sum_{z_s} \exp(-\sum_{C \in \mathcal{C}, C \ni s} V_C(z_s x_{S\setminus s})) \exp(-\sum_{C \in C, C \not\ni s} V_C(x_C))} \\
&= \frac{\exp(-\sum_{C \in \mathcal{C}, C \ni s} V_C(x_s x_{S\setminus s}))}{\sum_{z_s} \exp(-\sum_{C \in \mathcal{C}, C \ni s} V_C(z_s x_{S\setminus s}))}.
\end{aligned}$$

Nun sind nur noch die - für kleine C harmlosen Werte - $V_C(x)$ mit $C \ni s$ zu berechnen.

Beispiel 1.2.2 (Ising Modelll) Im Ising Modell

$$H(x) = -\sum_{s\sim t} x_s x_t, \;\; x_s = \pm 1,$$

hat man $V_{st}(x) = -x_s x_t$ für $s \sim t$ (und $V_C \equiv 0$ sonst) und somit

$$\Pi(x_s | x_{S\setminus s}) = \frac{\exp(-\sum_{s\sim t} x_s x_t)}{\exp(-\sum_{s\sim t} x_t) + \exp(\sum_{s\sim t} x_t)},$$

d.h.

$$\Pi(x_s = 1 | x_{S\setminus s}) = \frac{1}{1 + \exp(2 \cdot \sum_{s\sim t} x_t)},$$
$$\Pi(x_s = -1 | x_{S\setminus s}) = \frac{1}{1 + \exp(-2 \cdot \sum_{s\sim t} x_t)}.$$

Die Idee ist nun, aus diesen bedingten Verteilungen Markov-Prozesse zu konstruieren, deren Marginalverteilungen gegen Π konvergieren, bzw. gegen andere nützliche Verteilungen. So verschiebt man die Raumkomplexität in die Zeit.

1.2.2 Gibbs und Metropolis Sampler

Unser Ziel war, Stichproben gemäß einem Gibbsmaß Π zu ziehen (*sampling*). Im Hinblick auf die Grenzwertsätze für Markovketten ist also eine homogene Markovkette mit invarianter Verteilung Π zu konstruieren, die rechentechnisch handhabbar ist. *Eine* Möglichkeit der Konstruktion basiert auf den bedingten Verteilungen zu Π; der entsprechende Algorithmus heißt *Gibbs Sampler*. Er wird später in Konkurenz zu leichter simulierbaren Ketten – wie dem Metropolis Sampler – stehen. Die Theorie des Gibbs Samplers ist jedoch besonders elegant und durchsichtig.

Im folgenden sei E ein Produkt endlicher Räume E_s, $s \in S$, mit endlicher Indexmenge S. Für jedes $I \subset S$ ist ein Markovkern auf E durch folgende Vorschrift definiert:

$$\Pi_I(x, y) = \begin{cases} Z_I^{-1} \exp(-H(y_I x_{S\setminus I})) & \text{falls } y_{S\setminus I} = x_{S\setminus I} \\ 0 & \text{sonst} \end{cases} \tag{1.17}$$
$$Z_I = \sum_{z_I} \exp(-H(z_I x_{S\setminus I})).$$

Diese Markovkerne nennt man *lokale Charakteristiken* von Π. Sampeln aus $\Pi_I(x,\cdot)$ ändert x höchstens auf I. In diesem Text enthält I meist nur ein Pixel, d.h. I ist von der Gestalt $I = \{s\}$ mit $s \in S$. Die Gibbs Verteilung Π ist invariant für Π_I. Dies folgt aus einer stärkeren, aber leichter nachzuweisenden Eigenschaft.

Lemma 1.2.3 *Die Gibbs Verteilung Π und ihre lokalen Charakteristiken Π_I erfüllen die* detailed balance *Gleichung, d.h. für alle $x, y \in E$ und $I \subset S$ gilt*

$$\Pi(x)\Pi_I(x,y) = \Pi(y)\Pi_I(y,x).$$

Allgemeiner erfüllen die Verteilung μ und der Markovkern P die *detailed balance Gleichung*, wenn

$$\mu(x)P(x,y) = \mu(y)P(y,x)$$

für alle x und y. Dies bedeutet, daß die zu μ und P gehörige homogene Markovkette in der Zeit umkehrbar, also *reversibel* ist.

Beweis. (von Lemma 1.2.3) Entweder beide Seiten der Identität verschwinden oder es gilt $y_{S\setminus I} = x_{S\setminus I}$. Da $x = x_I y_{S\setminus I}$ und $y = y_I x_{S\setminus I}$, gilt weiter

$$\exp(-H(x))\frac{\exp(-H(y_I x_{S\setminus I}))}{\sum_{z_I}\exp(-H(z_I x_{S\setminus I}))} = \exp(-H(y))\frac{\exp(-H(x_I y_{S\setminus I}))}{\sum_{z_I}\exp(-H(z_I y_{S\setminus I}))}.$$

Daraus folgt die detailed balance Gleichung. □

Stationarität folgt daraus leicht.

Theorem 1.2.4 *Falls μ und P die detailed balance Gleichung erfüllen, so ist μ invariant für P. Insbesondere sind Gibbs Verteilungen invariant bezüglich ihrer lokalen Charakteristiken.*

Beweis. Summiere beide Seiten der detailed balance Gleichung über x. □

Eine Abzählung $S = \{s_1, \ldots, s_\sigma\}$, $\sigma = |S|$, von S heißt *Besuchsschema*. Zur Vereinfachung schreiben wir $S = \{1, \ldots, \sigma\}$. Ein Markovkern ist definiert durch

$$P(x,y) = \Pi_{\{1\}} \ldots \Pi_{\{\sigma\}}(x,y). \tag{1.18}$$

Man beachte, daß die Verknüpfungen in (1.18) Matrixmultiplikationen sind. Die homogene Markovkette zum Kern P entspricht dem folgenden Algorithmus:

- Ziehe Startkonfiguration x gemäß Startverteilung ν (z.B. $\nu = \delta_x$).
- Update x in Pixel Nummer eins durch die Intensität y_1, welche zufällig gemäß $\Pi_{\{1\}}(y,x)$ gezogen wird Dies ergibt eine neue Konfiguration $y = y_1 x_{S\setminus\{1\}}$, die dann in der zweiten Komponente upgedated wird usw. Nach dem Schritt Nummer σ ist ein *Sweep* beendet.
- Führe nun (viele) weitere Sweeps durch.

Dieser Algorithmus wird durch folgende Aussage gerechtfertigt:

Theorem 1.2.5 *Für jedes $x \in E$ gilt*

$$\lim_{n\to\infty} \nu P^n(x) = \Pi(x)$$

gleichmäßig in allen Startverteilungen ν.

Beweis. Die Gibbsverteilung $\mu = \Pi$ ist invariant bzgl. ihrer lokalen Charakteristiken nach Satz 1.2.4 und deshalb auch bzgl. deren Komposition P. Außerdem ist $P(x,y)$ strikt positiv, da in jedem $s \in S$ die Wahrscheinlichkeit, ein y_s zu sampeln strikt positiv ist. Deshalb ist der Satz ein Spezialfall von Satz 1.1.14. □

Man kann die Pixel auch in zufälliger Reihenfolge besuchen: Sei G eine Verteilung auf S. Man ersetze die lokalen Charakteristiken (1.17) in (1.18) durch Kerne

$$\tilde{\Pi}(x,y) = \begin{cases} G(s)\Pi_{\{s\}}(x,y) & \text{wenn } y_{S\setminus\{s\}} = x_{S\setminus\{s\}} \text{ für ein } s \in S \\ 0 & \text{sonst} \end{cases} \tag{1.19}$$

und setze $\tilde{P} = \tilde{\Pi}^\sigma$. G ist eine *Vorschlags-* oder *Explorationsverteilung*, meist die Gleichverteilung auf S.

Theorem 1.2.6 *Sei G eine strikt positive Verteilung auf S. Dann gilt*

$$\lim_{n\to\infty} \nu \tilde{P}^n(x) = \Pi(x)$$

für alle $x \in E$.

Beweis. Π und $\tilde{P}$ erfüllen die detailed balance Gleichung und somit ist Π invariant für $\tilde{P}$. Weiter folgt aus der strikten Positivität von G, daß $\tilde{P}$ strikt positiv ist und die Konvergenz ergibt sich nach Satz 1.1.14. □

Für die MMS Schätzung benötigen wir ein Gesetz der großen Zahlen. Die Zahl

$$\delta_s = \max\{|H(x) - H(y)| : x_{S\setminus\{s\}} = y_{S\setminus\{s\}}\}$$

ist die *Oszillation* von H in s und

$$\Delta = \max\{\delta_s : s \in S\}$$

ist die *maximale lokale Oszillation* von H. Schließlich bezeichne (ξ_i) eine Folge von Zufallsvariablen, deren Verteilung durch den Markovprozeß zur Startverteilung ν und zum Markovkern P gegeben ist. Die durch den Prozeß auf $E^{\mathbb{N}_0}$ induzierte Verteilung sei mit $\mathbf{P}$ bezeichnet.

Theorem 1.2.7 *Sei (ξ_i) durch (1.18) oder (1.19) induziert. Dann gilt für jede Funktion f auf E, daß*

$$\frac{1}{n}\sum_{i=0}^{n-1} f(\xi_i) \longrightarrow \boldsymbol{E}_{\Pi}(f)$$

in $L^2(\mathbf{P})$ und in Wahrscheinlichkeit. Für jedes $\varepsilon > 0$ gilt

$$\mathbf{P}\left(\left|\sum_{i=0}^{n-1} f(\xi_i) - \mathsf{E}_{\Pi}(f)\right| \geq \varepsilon\right) \leq \frac{c}{n\varepsilon^2} e^{\sigma\Delta}$$

wobei $c = 13\|f\|^2$ für (1.18) und $c = 13\|f\|^2 \min_S G(s)^{-\sigma}$ für (1.19).

Beweis. Der Markovkern P in (1.18) ist strikt positiv und wenn dessen Kontraktionskoeffizient echt kleiner eins ist, folgt die Behauptung aus Satz 1.1.16. Für $x \in E$ sei z_s eine Minimalstelle in s, d.h.

$$H(z_s x_{S\setminus\{s\}}) = m_s = \min\{H(v_s x_{S\setminus\{s\}}) : v_s \in E_s\}.$$

Dann gilt

$$\Pi_{\{s\}}(x, y_s x_{S\setminus\{s\}}) = \frac{\exp\left(-\left(H(y_s x_{S\setminus\{s\}}) - H(z_s x_{S\setminus\{s\}})\right)\right)}{\sum_{v_s \in \mathbf{X}_s} \exp\left(-\left(H(v_s x_{S\setminus\{s\}}) - m_s\right)\right)} \geq |E_s|^{-1} \mathrm{e}^{-\delta_s}$$

und somit

$$\min_{x,y} P(x,y) \geq \prod_{s=1}^{\sigma} \left(|E_s|^{-1} e^{-\sigma_s} \right) = |E|^{-1} \mathrm{e}^{-\Delta\sigma}.$$

Die Abschätzung in Lemma 1.1.12 liefert

$$c(P) \leq 1 - |E| \cdot \min_{x,y} P(x,y) \leq 1 - \mathrm{e}^{-\Delta\sigma}. \tag{1.20}$$

Damit folgt das Gesetz der großen Zahlen im Fall (1.18). Der Beweis für (1.19) unterscheidet sich nur in offensichtlichen Details. □

Das Gesetz der großen Zahlen impliziert, daß der Algorithmus nicht terminiert (mit positiver Wahrscheinlichkeit). Sei für jedes $x \in E$

$$A_{x,n} = \frac{1}{n} \sum_{i=0}^{n-1} \chi_{\{x\}}(\xi_i)$$

die relative Häufigkeit der Besuche in x während der ersten $n-1$ Schritte. Da

$$\boldsymbol{E}_{\Pi}(\chi_{\{x\}}) = \Pi(x)$$

folgt aus dem Satz:

Proposition 1.2.8 *Unter den Voraussetzungen von Satz 1.2.7 gilt $A_{x,n} \longrightarrow \Pi(x)$ in Wahrscheinlichkeit.*

Insbesondere besucht der Gibbs Sampler jeden Zustand unendlich oft.

Die Konvergenzsätze für den Gibbs Sampler lassen sich in besonders eleganter Weise herleiten. Deshalb wurde er für den theoretischen Teil benutzt. In der Praxis erweisen sich oft andere Verfahren als vorteilhaft oder leichter zu programmieren. Wir skizzieren nun die vielleicht bekannteste Alternative, den Metropolis Sampler (Metropolis, Rosenbluth, Teller und Teller (1953)). Sei H eine Energiefunktion (möglicherweise ersetzt durch die parametrisierte Version βH) und sei x die Konfiguration, die gerade aktualisiert werden soll. Das Updating erfolgt in zwei Schritten:

1. Der Vorschlag. Eine neue Konfiguration y wird erzeugt durch Sampeln aus einer Verteilung $G(x,\cdot)$ auf E.

2. Der Annahmeschritt.

(a) Falls $H(y) \leq H(x)$ so wird y als neue Konfiguration akzeptiert.

(b) Falls $H(y) > H(x)$ so wird y akzeptiert mit Wahrscheinlichkeit

$$\exp(H(x) - H(y)).$$

(c) Wird y nicht akzeptiert, so wird x beibehalten.

Die Matrix (der Markovkern) G heißt *Vorschlags-* oder *Explorationsmatrix*.

Eine neue Konfiguration y, die weniger vorteilhaft als x ist, wird also nicht automatisch zurückgewiesen, sondern mit einer Wahrscheinlichkeit akzeptiert, die mit dem Energiezuwachs $H(y) - H(x)$ fällt.

Beispiel 1.2.9 Bei Bildern ist ein natürlicher Vorschlag, ein Pixel rein zufällig zu wählen und dort einen zufälligen Grauwert einzusetzen. Genauer heißt das

$$G(x,y) = \begin{cases} \frac{1}{\sigma(N-1)} & \text{wenn } x_s \neq y_s \text{ für genau ein } s \in S \\ 0 & \text{sonst} \end{cases} \tag{1.21}$$

wobei σ die Zahl der Pixel und N die Zahl der Grauwerte ist (wir nehmen $|E_s| = N$ für alle s an). Solche Algorithmen heißen auch *Single Flip Algorithmen*.

Er hat explizit die Gestalt

$$\pi(x,y) = \begin{cases} G(x,y)\exp(-(H(y) - H(x))^+) & \text{wenn } x \neq y \\ 1 - \sum_{z \in E \setminus \{x\}} \pi(x,z) & \text{wenn } x = y \end{cases} \tag{1.22}$$

Definition 1.2.10 *Eine stochastische Matrix* $(G(x,y)_{x,y\in X}$ *heißt* irreduzibel *, wenn es zu jedem x und y in E eine Folge $x = x_0, x_1, \ldots, x_{n(x,y)-1}, x_{n(x,y)} = y$ gibt mit $G(x_{i-1}, x_i) > 0$ für alle $1 \leq i \leq n(x,y)$.*

Bemerkung 1.2.11 Primitive stochastische Matrizen sind irreduzibel. Irreduzibilität erzwingt *nicht* die Existenz eines n, das für alle Paare gemeinsam obige Kettenbedingung erfüllt (wie bei primitiven Matrizen). Existiert ein solcher Index, so ist die Matrix irreduzibel und *aperiodisch*.

Der entsprechende Konvergenzsatz lautet

Theorem 1.2.12 *Sei E eine endliche Menge, H eine nichtkonstante Funktion auf E und Π die Gibbs Verteilung zu H. Die Vorschlagsmatrix G sei symmetrisch und irreduzibel. Dann gilt:*
(a) Für jedes $x \in E$ und jede Startverteilung ν auf E gilt

$$\nu\pi^n(x) \longmapsto \Pi(x) \quad \textit{für} \quad n \to \infty.$$

(b) Für jede Startverteilung ν und jede Funktion f auf $\boldsymbol{X}$ gilt

$$\frac{1}{n}\sum_{i=1}^{n} f(\xi_i) \longrightarrow \boldsymbol{E}_\Pi(f) \quad \textit{für} \quad n \to \infty$$

in $L^2(\mathbf{P})$ und in Wahrscheinlichkeit.

1.2.3 Annealing

Die Berechnung der MAP Schätzer für Gibbs Verteilungen ist äquivalent zur Minimierung von Energiefunktionen. Eine einfache Modifikation des Gibbs Samplers findet die Minima – wenigstens in der Theorie.

Sei H eine Funktion auf E. Die Funktion βH hat für große β dieselben Minima wie H, jedoch sind diese tiefer.

Für eine Funktion H und eine reelle positive Zahl β ist die Gibbs Verteilung zur *inversen Temperatur* β gegeben durch

$$\Pi^\beta(x) = (Z^\beta)^{-1}\exp(-\beta H(x)), Z^\beta = \sum_z \exp(-\beta H(z)).$$

Sei M die Menge der globalen Minimalstellen von H.

Proposition 1.2.13 *Sei Π eine Gibbs Verteilung mit Energiefunktion H. Dann gilt*

$$\lim_{\beta\to\infty} \Pi^\beta(x) = \begin{cases} \frac{1}{|M|} & \textit{falls } x \in M \\ 0 & \textit{sonst} \end{cases}.$$

Für $x \in M$ wächst die Funktion $\beta \to \Pi^\beta(x)$ und für $x \notin M$ fällt sie schließlich.

Gelingt es also für hohes β aus Π^β zu Sampeln, so hat man (approximativ) Minimalstellen von H.

Beweis. Sei m der minimale Wert von H. Dann gilt

$$\Pi^\beta(x) = \frac{\exp(-\beta H(x))}{\sum_z \exp(-\beta H(z))}$$
$$= \frac{\exp(-\beta(H(x)-m))}{\sum_{z:H(z)=m}\exp(-\beta(H(z)-m)) + \sum_{z:H(z)>m}\exp(-\beta(H(z)-m))}$$

Ist x oder z Minimum, so verschwindet der entsprechende Exponent und der Summenterm ist eins. Die anderen Exponenten sind strikt negativ und die entsprechenden Terme konvergieren für $\beta \to \infty$gegen 0. Deshalb wächst der Ausdruck monoton gegen $|M|^{-1}$ falls x Minimalstelle ist und er konvergiert gegen 0 sonst. Sei nun $x \notin M$ und $a(y) = H(y) - H(x)$. Wir schreiben $\Pi^\beta(x)$ in der Form

$$\frac{1}{|\{y : H(y) = H(x)\}| + \sum_{a(y)<0}\exp(-\beta a(y)) + \sum_{a(y)>0}\exp(-\beta a(y))}.$$

Wir zeigen, daß der Nenner schließlich wächst. Differentiation nach β liefert

$$\sum_{y:a(y)<0}(-a(y))\exp(-\beta a(y)) + \sum_{y:a(y)>0}(-a(y))\exp(-\beta a(y)).$$

Der zweite Term geht gegen 0 und der erste gegen Unendlich, wenn $\beta \to \infty$. Deshalb wird die Ableitung schließlich positiv und $\beta \mapsto \Pi^\beta(x)$ fällt schließlich. Damit ist der Beweis vollständig. $\square$

Bemerkung 1.2.14 Mit $\beta \to 0$ konvergiert die Gibbs Verteilung Π^β gegen die Gleichverteilung auf E, denn in

$$\Pi^\beta(x) = \frac{\exp(-\beta H(x))}{\sum_y \exp(-\beta H(y))}$$

strebt jeder Exponent gegen eins.

Wir wählen ein Besuchsschema und bezeichnen es mit $S = \{1, \ldots, \sigma\}$. Ein *Abkühlschema* ist eine wachsende Folge $\beta(n)$ positiver Zahlen. Für jedes $n \geq 1$ ist ein Markovkern definiert durch

$$P_n(x,y) = \Pi_{\{1\}}^{\beta(n)} \ldots \Pi_{\{\sigma\}}^{\beta(n)}(x,y),$$

wobei $\Pi_{\{k\}}^{\beta(n)}$ die lokale Charakteristik von $\Pi^{\beta(n)}$ in k ist. Zusammen mit einer Startverteilung definieren diese Kerne eine inhomogene Markovkette.

Theorem 1.2.15 *Sei $(\beta(n))_{n\geq 1}$ ein Abkühlschema, welches gegen Unendlich wächst, so daß schließlich*

$$\beta(n) \leq \frac{1}{\sigma\Delta} \ln n.$$

Dann gilt

$$\lim_{n\to\infty} \nu P_1 \dots P_n(x) = \begin{cases} |M|^{-1} & \text{wenn } x \in M \\ 0 & \text{sonst} \end{cases}$$

gleichmäßig in allen Startverteilungen ν.

Der Satz stammt von S. und D. Geman (1984), [11]. Deren Beweis ist nicht leicht nachzuvollziehen. Die Methode der Kontraktionskoeffizienten hingegen erlaubt eine äußerst transparente Beweisführung. Wir schicken dem Beweis ein Lemma voraus.

Lemma 1.2.16 *Für die reellen Folgen (a_n) und (b_n) gelte $0 \leq a_n \leq b_n \leq 1$. Dann folgt aus $\sum_n a_n = \infty$, daß $\prod_n (1 - b_n) = 0$.*

Beweis. Aus der bekannten Ungleichung

$$\ln x \leq x - 1, \quad 0 < x,$$

folgt

$$\ln(1 - b_n) \leq \ln(1 - a_n) \leq -a_n.$$

Da die Summen divergieren, gilt

$$\sum_n \ln(1 - b_n) = -\infty,$$

was zu

$$\prod_n (1 - b_n) = 0$$

äquivalent ist. □

Beweis. des Satzes. Wir müssen $\beta(n)$ so wählen, daß die Voraussetzungen von Satz 1.1.18 mit obigen P_n und $\mu_n = \Pi^{\beta(n)}$ erfüllt sind. Daraus und mit Satz 1.2.13 folgt dann die Behauptung.
Das Gibbsfeld μ_n ist invariant bezüglich der Kerne P_n nach Satz 1.2.4. Da $(\beta(n))$ wächst, sind die Folgen $(\mu_n(x))$, $x \in E$ schließlich monoton wegen

Proposition 1.2.13 und deshalb gilt (1.11) wegen Lemma 1.1.19. Wie in (1.20) gilt

$$c(P_n) \leq 1 - \mathrm{e}^{-\beta(\mathrm{n})\Delta\sigma}.$$

Wir leiten nun die hinreichende Bedingung (1.15) d.h. $\prod_{k \geq i} c(P_k) = 0$ für alle i her. Nach Lemma 1.2.16 gilt dies, falls $\exp(-\beta(n)\Delta\sigma) \geq a_n$ für $a_n \in [0,1]$ mit divergenter unendlicher Reihe. Eine natürliche Wahl ist $a_n = n^{-1}$ und damit ist

$$\beta(n) \leq \frac{1}{\sigma\Delta} \ln n$$

für schließlich alle n hinreichend. Damit ist der Beweis vollständig. □

Das logarithmische Abkühlschema ist willkürlich gewählt, da die Bedingung

$$\sum_n \exp(-\beta(n)\Delta\sigma) = \infty.$$

entscheidend ist. Oft benutzt man stückweise konstante Schemata.

Der Satz gilt auch für zufällige Besuchsschemata in (1.19). Dann setzt man

$$P_n = \left(\tilde{\Pi}^{\beta(n)}\right)^\sigma.$$

Es ist

$$c(P_n) \leq 1 - \gamma e^{-\beta(n)\Delta\sigma}$$

mit $\gamma = \min_s G(s)^\sigma$. Wenn G strikt positiv ist, dann ist $\gamma > 0$ und es gilt

$$\gamma \exp(-\beta(n)\Delta\sigma) \geq \gamma n^{-1}.$$

Da(γn^{-1}) eine divergente unendliche Reihe induziert, ist die Behauptung damit bewiesen.

Eine einfache Version des entsprechenden Konvergenzsatzes für Annealing auf der Grundlage des Metropolis Samplers sei ohne Beweis zitiert. Näheres findet man in G. Winkler (1995), [17], Kapitel 8.2.

Theorem 1.2.17 *Seien E eine endliche Menge, H eine nichtkonstante Funktion auf E und G eine symmetrische irreduzible Vorschlagsmatrix. Sei ferner $\beta(n)$ ein Abkühlschema, welches gegen Unendlich wächst und die Bedingung*

$$\beta(n) \approx \frac{1}{\sigma\Delta} \ln n$$

schließlich erfüllt. Dann konvergieren für jede Startverteilung ν auf E die eindimensionalen Randverteilungen

$$\nu\pi^{\beta(1)} \ldots \pi^{\beta(n)}$$

gegen die Gleichverteilung auf den Minimalstellen von H.

Die Minimalstellen von H sind natürlich wieder die Moden der assoziierten Gibbsverteilung.

1.3 Anwendungen von MCMC Methoden

Sampling, Annealing und verwandte Methoden wurden im englischen Sprachraum unter dem Stichwort zusammengefaßt. Wie üblich, wurde dieser Terminus alsbald ins Deutsche übernommen.

In diesem Abschnitt illustrieren wir in unsystematischer Weise (d.h. wir folgen didaktischen und nicht systematischen Gesichtspunkten) die Anwendbarkeit dieser Methoden. Ihr genereller Vorteil ist, daß sie ungeheuer flexibel und leicht zu implementieren sind; die Programme sind somit problemlos an neue Bedürfnisse anzupassen. Als Nachteile sollten teilweise lange Laufzeiten (z.B. in der Bildanalyse) und ein zu schwaches theoretisches Fundament genannt werden. Somit sind sie vor allem bei heuristischen Untersuchungen sowie in der explorativen Datenanalyse nützlich. Details werden im jeweiligen Beispiel besprochen.

1.3.1 Der Metropolis Algorithmus in der kombinatorischen Optimierung

Annealing als Methode in der kombinatorischen Optimierung wurde in KIRKPATRICK, GELLATT and VECCHI (1982), BONOMI and LUTTON (1984) and ČERNY (1985) vorgeschlagen. In der kombinatorischen Optimierung ist der Zustandsraum typischerweise nicht von Produktgestalt wie in der Bildanalyse. Das klassische Beispiel, vielleicht weil es so einfach zu beschreiben ist, ist das 'Travelling Salesman Problem'. Es ist eines der am besten untersuchten NP-schweren Probleme. Es kann als gutes Beispiel dienen, wie dynamische Monte Carlo Methoden in der kombinatorischen Optimierung eingesetzt werden können.

Beispiel 1.3.1 (Travelling Salesman Problem) Ein Handlungsreisender soll N Städte genau einmal besuchen und anschließend heimkehren. Gegeben sind die Abstände zwischen je zwei Städten (die nicht unbedingt symmetrisch sein müssen). Gesucht ist eine kürzeste Route.

Dieses Problem hat viele, praktisch wichtige äquivalente Formulierungen.

Beispiel 1.3.2 Ein winziger Lötkolben soll automatisch eine Reihe von Lötstellen auf einem Chip setzen. Die Ausfallrate der Chips wächst mit der Weglänge, die der Lötkolben durchläuft. Deshalb sollte die totale Weglänge minimiert werden.

Wir bleiben der Formulierung das Travelling Salesman Problems. Sei N die Anzahl der Städte und somit o.E. $\mathbf{C} = \{1, \ldots, N\}$ die Menge der Städte . Der Abstand zwischen Stadt i und j sei $d(i,j) \geq 0$ (er muß nicht unbedingt symmetrisch sein). Eine 'Tour' ist eine Abbildung $\varphi : \mathbf{C} \mapsto \mathbf{C}$, so daß $\varphi^k(i) \neq i$ für alle $k = 1, \ldots, N-1$ und $\varphi^N(i) = i$ für alle i, d.h eine zyklische Permutation von $\mathbf{C}$. Die Menge E aller Touren hat $(N-1)!$ Elemente. Die Kosten einer Tour entsprechen der totalen Länge

$$H(\varphi) = \sum_{i \in \mathbf{C}} d(i, \varphi(i)).$$

Der Einfachkeit halber nehmen wir $d(i,j) = d(j,i)$ an, d.h. wir betrachten das *symmetrische Travelling Salesman Problem.* Das generelle Vorgehen beim Annealing mit dem Metropolis Algorithmus ist klar. Es ist nur noch der Vorschlagsschritt zu definieren. Bei der einfachsten Variante werden lediglich zwei Städte ausgetauscht, welche gleichverteilt gezogen werden.Eine Tour ψ ist Nachbar von Tour φ, wenn sie auf folgende Weise aus φ hervorgeht: Man repräsentiere φ als gerichteten Graphen mit den Städten als Knoten und Pfeilen entsprechend der Tour. Man entfernt zwei nicht aufeinanderfolgende Pfeil mit Anfangspunkten p und $\varphi^{-1}(q)$, und ersetzt sie durch die Pfeile von p nach$\varphi^{-1}(q)$ und von $\varphi(p)$ nach q und dreht schließlich die Pfeile zwischen $\varphi(p)$ und $\varphi^{-1}(q)$ um. Dies liefert eine neue Tour. Diese Generierung heißt Zweiertausch (Abbildung 1.1).

Im symmetrischen Travelling Salesman Problem läßt sich die Energiedifferenz $H(\psi) - H(\varphi)$ leicht ausrechnen, da nur zwei Terme der Summe durch neue ersetzt werden.

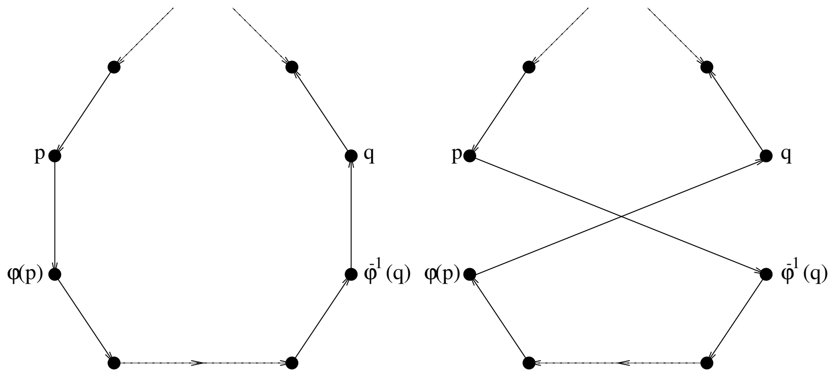
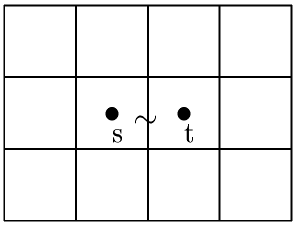


Abbildung 1.1: Zweiertausch beim Travelling Salesman Problem

1.3.2 Simulation von (binären) Mustern

Hier wird im wesentlichen Gibbs oder Metropolis Sampling aus gewissen Gibbsmaßen eingesetzt. Wir definieren zunächst das Modell und geben dann einfache Beispiele.

Wir erklären zunächst die Situation: Gegeben ist ein Pixelraster S wie in Abb. 1.2. Zu jedem Pixel $s \in S$ gibt es einen endlichen Vorrat E_s von Intensitäten. Ein Muster ist eine Konfiguration $(x_s)_{s \in S} \in \prod E_s =: E$. Auf E ist ein Gibbsmaß Π gegeben, das mehr oder weniger typische, d.h. wahrscheinliche Muster charakterisiert. Simulation typischer Muster und ihrer Charakteri-



s, t	:	Pixel
$s \sim t$	:	Nachbarn
x_s	:	Idealwert in s
y_s	:	Beobachtung in s

Abbildung 1.2: Pixelraster und einfache Nachbarschaftsbeziehung

stika erfolgt dann über Sampling. Wir geben einfache Modellklassen an.

Beispiel 1.3.3 Das *Ising Modell* ist gegeben durch

$$H(x) = -\beta \sum_{s \sim t} x_s x_t, \beta > 0, \ \ x_s = \pm 1.$$

Das Modell stammt aus der Physik, wo es die Gestalt

$$H(x) = -\frac{1}{kT}\left(J\sum_{s\sim t} x_s x_t - mB\sum_s x_s\right)$$

hat. J ist eine Materialkonstante, T die absolute Temperatur, k die Boltzmannkonstante, m hängt wieder vom Material ab und B quantifiziert ein äußeres magnetische Feld. $J > 0$ treibt die x_s in dieselbe, $J < 0$ in entgegengesetzte Richtung. Interpretiert man die x_s als Richtung von Spins - oder Elementarmagneten - in einem Kristallgitter S, so entspricht $J > 0$ einem Ferromagneten und $J < 0$ einem Antiferromagneten.
Das Modell ist nach (dem Deutschen) E. Ising benannt, der es in seiner Doktorarbeit 1925 bei W. Lenz untersucht hat (und teilweise falsche Schlüsse von Dimension eins auf höhere Dimensionen zog; dadurch wurde die statistische Physik um Jahre zurückgeworfen). In der Musteranalyse dient es der Modellierung (stückweise) glatter Muster.

Beispiel 1.3.4 (Das Potts Modell) Bei endlichem Grauwertevorrat G ist eine weitere natürliche Verallgemeinerung des homogenen Ising Modells das *Potts Modell*

$$H(x) \quad = \quad -\beta\sum_{s\sim t}\chi_{\{x_s=x_t\}} = \beta\sum_{s\sim t}\varphi(x_s - x_t) = -\beta|s\sim t : x_s = x_t|$$

mit

$$\varphi(u) = \begin{cases} -1 & \text{falls } u = 0 \\ 0 & \text{sonst.} \end{cases}$$

Beispiel 1.3.5 (Allgemeine Binärmodelle) 'Glattheit' kann man jetzt verallgemeinern: Im Ising Modell bedeutet es die Nähe zu den zwei konstanten Mustern $x \equiv 1$ oder $x \equiv -1$. Diese können wir durch beliebige Muster (und ihre Inverse) ersetzen, indem wir die Vorzeichen der Terme $x_s x_t$ geeignet wählen:

$$H(x) = \sum_{s\sim t} v_{st} x_s x_t, \;\; v_{st} = \pm 1.$$

Noch allgemeiner kann man die Stärke der Ähnlichkeit lokal steuern durch

$$H(x) = \sum_{s\sim t} a_{st} x_s x_t, \;\; a_{st} \in \mathbb{R}.$$

oder noch allgemeiner

$$H(x) = \sum_{s \sim t} a_{st} x_s x_t + \sum_s a_s x_s.$$

Durch solche Modelle lassen sich beliebige binäre Muster charakterisieren.

1.3.3 Bayessche Bildanalyse am Beispiel der glatten Regression mit Sprüngen

Wir betrachten nun ein Beispiel aus der Bildanalyse (das im Fall einer Dimension auch endliche Zeitreihen umfaßt).

In unserem Beispiel geht es um die Rekonstruktion eines i.a. unbekannten Idealbildes x aus einem Datensatz y. Dabei stellt man sich vor, daß die Daten eine gestörte Version von x darstellen. Die Zuordnung eines x zu y stellt ein *inverses Problem* dar, das i.a. nicht eindeutig lösbar ist. Um die Lösungsmenge einzuschränken, stellt man Zusatzbedingungen, die ein gewisses Vorwissen über oder Erwartungen an die Rekonstruktion widerspiegeln. Eine Möglichkeit dies zu tun, ist die Angabe einer Gütefunktion für x.

Offensichtlich hängt die Rekonstruktion nun von der (z.B. subjektiv) gewählten Gütefunktion ab. Der Wahl dieser Funktion entspricht im klassischen Fall die (ebenfalls subjektive) Auswahl von Filtern.

Der natürliche Rahmen, um verschiedene Datenquellen, Vorinformationen und Regularitätsbedingungen zu verknüpfen ist das *Bayessche Paradigma* [8]. In seinem Rahmen können weitere sinnvolle Schätzer für x formuliert und begründet werden.

Die a priori Verteilung wird so aufgebaut, daß Bilder x mit unerwünschten Eigenschaften niedrige und solche mit erwünschten Eigenschaften hohe Wahrscheinlichkeit haben.

Seien $\boldsymbol{X}$ und $\boldsymbol{Y}$ Meßräume, Π ein Wahrscheinlichkeitsmaß auf $\boldsymbol{X}$ und $P(x, y)$ eine Übergangswahrscheinlichkeit von $\boldsymbol{X}$ nach $\boldsymbol{Y}$.

Wir nehmen zunächst an, daß beide Räume endlich sind und Π strikt positiv ist. Dann wird durch

$$\Pi(x, y) = \Pi(x) P(x, y), \;\; x \in \boldsymbol{X}, \;\; y \in \boldsymbol{Y},$$

eine gemeinsame Verteilung auf $\boldsymbol{X} \times \boldsymbol{Y}$ definiert mit $\Pi(y|x) = P(x, y)$ und

$$\Pi(x|y) = \frac{\Pi(x, y)}{\Pi(E \times \{y\})}.$$

[8] *Paradigma*: Beispielhafte Vorgehensweise

Für Verteilungen mit Dichten ist der Zugang analog.

Der am meisten verwendete Schätzer für den wahren Parameter x^* ist der maximum a posteriori Schätzer, also der Modus der a posteriori Verteilung. Er wird im allgemeinen durch *Simulated Annealing* berechnet.

Eine einfache a priori Gibbsverteilung für stückweise glatte Binärbilder ist durch das Ising Modell gegeben.

Beispiel 1.3.6 (Das Ising Modell) Das *Ising Modell ohne äußeres Feld* hat die Energiefunktion

$$H(x) = -\beta \sum_{s \sim t} x_s x_t, \ \beta > 0, \ \ x_s = \pm 1.$$

Offensichtlich hat H zwei Minima $x_s^* \equiv 1$ und $x_s^* \equiv -1$ mit dem Funktionswert:

$$H(x^*) = -\beta |\{(s,t) \in S \times S : s \sim t\}|.$$

Jedes Nachbarpaar $s \sim t$ mit $x_s \neq x_t$ verschlechtert die Bilanz um 2β. Somit ist

$$(H(x) - H(x^*))/2\beta$$

die Zahl der ungleichen Nachbarn, also die Konturlänge . Konfigurationen mit niedrigem $H(x)$ haben also kurze Konturen , ähnlich wie im Modell von Blake & Zisserman. Wir hatten erwähnt, daß dies eine gewisse Regularität der Konturen erzwingt. Der Parameter β steuert den Grad der Regularität. An diesem Beispiel sieht man ein typisches Phänomen dieser Modelle: Neben den Minima gibt es eine Vielzahl lokaler Minima, die sich im Wert nur wenig von den Minima unterscheiden - in diesem diskreten Fall nennen wir x ein *lokales Minimum*, wenn die Änderung in einem beliebigen Pixel - d.h. einer Koordinate von $\boldsymbol{X} = \{-1, 1\}^S$ - den Wert von H nicht verschlechtert. So haben in einem $n \times n$ Gitter S alle Bilder x, welche durch eine senkrechte oder waagerechte Gerade in zwei gleichfarbige Teile getrennt sind, denselben Wert

$$a = H(x^*) + 2\beta n.$$

Flippt man das äußerste Pixel auf der Trennlinie, so verschlechtert sich der Wert um 2β. Fräst man die Linie sukzessive ab, so bleibt H gleich, um dann beim letzten Pixel wieder um 2β auf a herunterzuspringen. Qualitativ betrachtet hat die Energiefläche große Plateaus mit vielen flachen lokalen Minima und zwei flachen (globalen) Minima. Die Beziehung zu

$$\tilde{H}(x) = \gamma \sum_{s \sim t} (x_s - x_t)^2$$

ist klar. Das Quadrat liefert nur die Werte 0 und 4. Also wird $\tilde{H}$ durch geeignete Wahl von γ und Addition einer Konstante in H übergeführt (Addition einer Konstante ändert Π nicht). Ebenso ist äquivalent:

$$\tilde{H}(x) = \gamma \sum_{s\sim t} \varphi(x_s - x_t), \;\; \varphi(u) = \begin{cases} 0 & \text{falls } u = 0 \\ 1 & \text{sonst.} \end{cases}$$

Typisches (multiplikatives) Rauschen auf Binärbildern ist

Beispiel 1.3.7 (Kanalrauschen) Wir betrachten Binärbilder x mit $x_s = \pm 1$. In den Pixeln werden die Zustände unabhängig geflippt:

$$y_s = x_s \cdot \eta_s; \;\; \eta_s = \begin{cases} -1 & \text{mit Wahrscheinlichkeit } p \\ \\ 1 & \text{mit Wahrscheinlichkeit } 1-p \end{cases}, \quad \eta_s \text{ i.i.d.},$$

Dann gilt

$$P(x,y) = p^{|\{s:y_s=-x_s\}|}(1-p)^{|\{s:y_s=x_s\}|}$$

und

$$-\ln P(x,y) = -|\{s : y_s = -x_s\}| \ln p - |\{s : y_s = x_s\}| \ln(1-p).$$

Wegen

$$\chi_{\{x_s=y_s\}} = \frac{x_s y_s}{2} + \frac{1}{2}$$

ergibt sich

$$-\ln P(x,y) = \frac{1}{2} \ln\left(\frac{p}{1-p}\right) \sum_s x_s y_s + \frac{|s|}{2} \ln\left(\frac{1}{p(1-p)}\right).$$

Glatte Bilder mit vielen Grauwerten werden durch folgende a priori Verteilung charakterisiert:

Beispiel 1.3.8 Sei die a priori Verteilung gegeben durch

$$\Pi(x) = \frac{1}{Z} \exp\left(-\sum_{s\sim t}(x_s - x_t)^2\right), \;\; Z = \sum_z \exp\left(-\sum_{s\sim t}(z_s - z_t)^2\right).$$

Das Rauschen sei additiv, weiß und gaußisch, also

$$P(x,dy) = \frac{1}{\sqrt{2\pi\sigma^{2|S|}}} \exp\left(-\frac{1}{2\sigma^2}\sum_s (y_s - x_s)^2\right) dy.$$

Dann gilt für die a posteriori Verteilung

$$\Pi(dx|y) \propto \exp\left(-\left\{\gamma \sum_{s\sim t}(x_s - x_t)^2 + \sum_s (y_s - x_s)^2\right\}\right)\, dx.$$

Allgemein ist die Gibbssche Form der a posteriori Verteilung gegeben durch:

Proposition 1.3.9 *Ist $P(x, dy)$ eine beliebige strikt positive (Zähl-) Dichte $f_x(y)\,dy$ und $\Pi(x) \propto \exp(-H(x))$ ein Gibbsmaß, so gilt*

$$\Pi(x|y) \propto \exp(-\{H(x) - \ln(f_x(y))\})$$

Beweis. Man multipliziere aus. $\square$

Die a priori Verteilung aus Beispiel 1.3.8 führt zu Überglättung; Kanten werden verschmiert. Deshalb bauen wir ein Modell für stückweise, d.h. kantenerhaltende Glättung. Dazu führt man Schaltervariablen b_{st} ein, welche die Glättung aus- und den Bruchterm einschalten: Für benachbarte Pixel s, t gemäß der Skizze1.2 definieren wir die *Mikrokante* $s \sim t$ als das Element des dualen Gitters zwischen s und t. Für die Mikrokanten definieren wir die Schaltervariablen

$$b_{st} = \begin{cases} 0 & \text{keine Kante zwischen } s \text{ und } t \\ 1 & \text{Kante zwischen } s \text{ und } t \end{cases} \qquad \text{für } s \sim t.$$

Die a priori Verteilung ist dann

$$\Pi(x) = \frac{1}{Z} \exp\left(-\sum_{s\sim t} (\lambda^2 (x_s - x_t)^2 (1 - b_{st}) + \alpha b_{st})\right),$$

wobei Z wie üblich die Summe von $\Pi(z)$ über den Zustandsraum ist und die a posteriori Verteilung ist

$$\Pi(dx|y) \propto \exp\left(-\left\{H^{post}(x, b) + \sum_s (y_s - x_s)^2\right\}\right)\, dx$$

mit

$$H^{post}(x, b) = \sum_{s\sim t} \underbrace{\underbrace{\lambda^2 (x_s - x_t)^2}_{\text{Glättung}} \underbrace{(1 - b_{st})}_{\text{ein/aus}} + \underbrace{\alpha b_{st}}_{\text{Penalty}}}_{h_{\lambda\alpha}}.$$

Wir diskutieren: Ist $d = \lambda^2(x_s - x_t)^2 > \alpha$, so lohnt es sich, die Strafe α für b_{st} zu bezahlen, d.h. der Glättungsterm wird ausgeschaltet. Ob dieser Fall sich lohnt, hängt wiederum vom Datenterm $(y_s - x_s)^2$ ab. Wird umgekehrt von diesem Term ein kleines d favorisiert, so lohnt sich der Bruch nicht und d wirkt wie im letzten Beispiel.

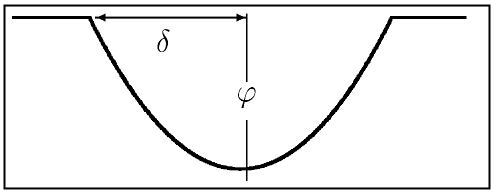
Wir werden an dieser Stelle noch nicht auf die Minimierung von H eingehen. Der erste Schritt dorthin ist jedoch von eigenständigem Interesse.

Für eine Minimumstelle (x^*, b^*) von H gilt

$$\begin{aligned}
&\mathbf{D}(x^*, b^*) + \sum_{s\sim t} h_{\lambda\alpha}(x_s^* - x_t^*, b_{st}^*)\\
&= \left(\min_{x,b} \mathbf{D}(x) + \sum_{s\sim t} h_{\lambda\alpha}(x_s - x_t, b_{st})\right)\\
&= \left(\min_{x} \mathbf{D}(x) + \sum_{s\sim t} \min_{b_{st}=0,1} h_{\lambda,\alpha}(x_s - x_t, b_{st})\right)\\
&= \left(\min_{x} \mathbf{D}(x) + \sum_{s\sim t} \min\{\lambda^2(x_s - x_t)^2, \alpha\}\right)\\
&= \left(\min_{x} \mathbf{D}(x) + \sum_{s\sim t} \varphi(x_s - x_t)\right);
\end{aligned}$$

die Funktion φ ist dabei gegeben durch

$$\varphi(\Delta) = \min\{\lambda^2\Delta^2, \alpha\}. \tag{1.23}$$



Um x^* zu berechnen genügt es also, eine Minimierung nur in der Komponente x durchzuführen. Die zweite Komponente b^* der Minimalstelle (x^*, b^*) ist dann eindeutig rekonstruierbar:

$$b_{st}^* = 1 \Longleftrightarrow |x_s - x_t| > \delta(= \sqrt{\alpha}/\lambda).$$

Wir landen insgesamt bei einer a priori Verteilung mit Energie

$$H(x) = \sum_{s \sim t} \varphi(x_s - x_t)$$

und einer a posteriori Verteilung mit Energie

$$H^{post}(x) = \sum_{s \sim t} \varphi(x_s - x_t) + \mathbf{D}(x).$$

1.4 Hepatitis B: Eine Fallstudie mit MCMC Methoden

Dieser Abschnitt ist eine verkürzte Nacherzählung eines Artikels von D.J. Spiegelhalter, N.G. Best, W.R. Gilks und H. Inskip (1996), [15]. Das Beispiel soll illustrieren, wie MCMC Methoden bei Problemen der statistischen Inferenz praktisch eingesetzt werden können.

1.4.1 Die Problemstellung

Hepatitis B (HB) kommt in vielen Teilen der Welt vor; in Teilen Westafrikas ist praktisch jeder infiziert. Etwa 20 der Infizierten verlieren die Infektion nie und sind ansteckend. Sie tragen ein erhöhtes Risiko für Lebererkrankungen; Leberkrebs ist eine der Haupttodesursachen. Die Gambian Hepatitis Intervention Study (GHIS) ist ein Impfprogramm gegen HB. Ihre Wirkung beruht auf der Dauer der Immunität nach der Impfung. Deshalb wurde eine Kohortenstudie durchgeführt.

Blutproben wurden in periodischen Zeitabständen jedem Kind entnommen und die Menge der Antikörper darin bestimmt. Diese Messung heißt Anti-HB-Titer[9]. Die Einheit heißt mIU[10]. Eine ähnliche Studie im benachbarten Senegal legte nahe, daß für alle Kinder gilt:

$$\text{anti-HB Titer} \propto \frac{1}{t},$$

[9]Titer: u.a. Menge eines Antikörpers (meist bezogen auf die Verdünnungsstufe der zu untersuchenden Lösung). Smoothers for discontinuous signals G. Winkler, V. Liebscher and V. Aurich

[10]milli-International-Units

wobei t die Zeit nach dem Abschluß des Impfverfahrens bezeichnet. Dies ist äquivalent zu einem linearen Zusammenhang von log-Titer und log-Zeit:

$$y = \alpha_i - 1 \cdot \log t,$$

wobei y den log anti-HB Titer für Kind i bezeichnet und α_i nach der letzten Impfung für jedes Kind konstant ist.

Es soll geprüft werden, ob das Modell korrekt ist, insbesondere ob der Wert minus eins der Ableitung stimmt. Damit hätte man eine einfache Methode um den individuellen Impfschutz vorherzusagen.

1.4.2 Die Modell- und Verteilungsannahmen

Das Bild zeigt die Daten für 106 Kinder aus der Studie. Jedes Kind hat eine Basismessung nach der letzten Impfung und mindestens zwei Folgemessungen. Das Kind ($*$) ist ein 'Ausreißer'.

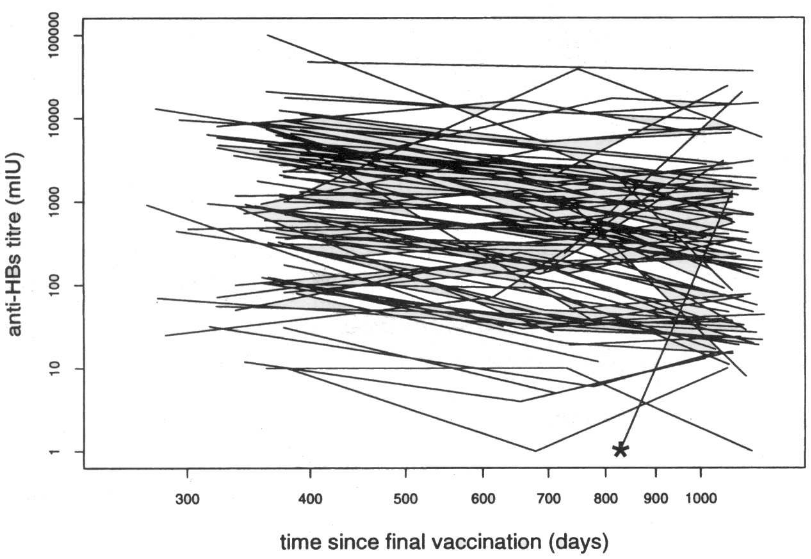


Abbildung 1.3: Rohdaten; Einzelmessungen durch Geradenstücke verbunden. Aus [12] mit freundlicher Erlaubnis von W.R. Gilks *et al.*

Zunächst wurde für jedes Kind eine Gerade gefittet (nach der Methode der kleinsten Quadrate):

$$\boldsymbol{E}(y_{ij}) = \alpha_i + \beta_i(\log t_{ij} - \log 730).$$

Dabei ist i der Index für das Kind und j der Index für die j-te Messung, so daß ij für die j-te Messung an Kind i steht. Wegen der Zentrierung auf $\log 730$ ist der Interzept oder Achsenabschnitt α_i der Regressionsgerade der geschätzte log-Titer nach zwei Jahren. Das Bild 1.4 zeigt die Verteilung der 106 unabhängig für jedes Kind gemäß üblicher linearer Regression geschätzten Parameter $\hat{\alpha}_i$ und $\hat{\beta}_i$; ferner sind $\hat{\alpha}_i$ und $\hat{\beta}_i$ gegeneinander aufgetragen. Die Verteilung der $\hat{\alpha}_i$ sieht halbwegs gaußisch aus, ausgenommen der verein-

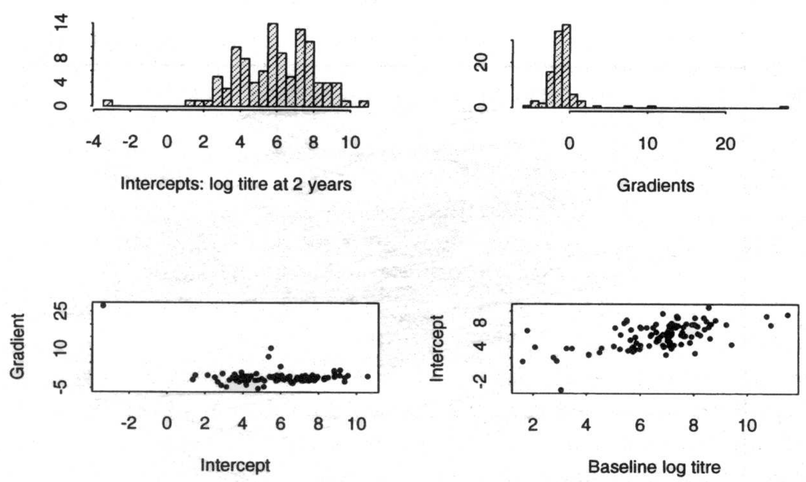


Abbildung 1.4: Verteilungen der unabhängig für jedes Kind geschätzten Parameter $\hat{\alpha}_i$ und $\hat{\beta}_i$, Plot von $\hat{\alpha}_i$ gegen $\hat{\beta}_i$ und von Basistiter y_{i0} gegen $\hat{\alpha}_i$. Aus [12] mit freundlicher Erlaubnis von W.R. Gilks *et al.*

zelte stark negative Wert für das Kind $(*)$. Für die Verteilung der geschätzten Ableitungen $\hat{\beta}_i$ gilt dasselbe, ausgenommen einige sehr hohe Schätzer, insbesondere für Kind $(*)$. Der Plot der geschätzten Interzepts gegen die geschätzten Gradienten legt - wieder den offensichtlichen Ausreißer $(*)$ ausgenommen - Unabhängigkeit von α_i und β_i nahe. Der letzte Plot stellt den Basistiter y_{i0} dem Interzept α_i gegenüber; hoher Basistiter scheint höhere Werte bei nachfolgenden Messungen zu fördern.

Bei der Studie geht es nicht um die 106 untersuchten Kinder selbst sondern um die Gesamtbevölkerung. Unabhängige lineare Regression für jedes einzelne Kind liefert keine Basis für statistische Inferenz über die Bevölkerung. Deshalb sollen Annahmen über die zugrundeliegende Population in das Modell eingebaut werden. Dies geschieht über a priori Annahmen bezüglich der Verteilungen der α_i und β_i.

Wir haben drei verschiedene Komponenten bei der Konstruktion eines vollen (Bayesschen) wahrscheinlichkeitstheoretischen Modelles, die auf die speziellen Daten der Studie angewandt werden sollen:

- Die Variablen des Modells müssen spezifiziert werden; ihre qualitative Abhängigkeitsstruktur muß festgelegt werden. Dafür sind sogenannte *graphische Modelle* besonders geeignet.

- Die parametrische Form der Abhängigkeiten muß spezifiziert werden; dies gibt uns die Likelihoodterme des Modelles. Die a priori Verteilungen und die bedingten Verteilungen mögen jeweils für sich Standardform haben; die Verknüpfung durch die spezifizierte 'bedingte Abhängigkeitsstruktur' erlaubt dennoch beliebig komplexe Konstruktionen.

- Von besonderem Interesse ist natürlich auch die Wahl der a priori Verteilungen.

1.4.3 Strukturelle Modellierung durch ein graphisches Modell

Wir machen folgende (Un)-Abhängigkeitsannahmen, die wir anschließend durch eine graphische Darstellung veranschaulichen. Die y_{ij} sind unabhängig konditional auf ihren Mittelwert μ_{ij} und einen Parameter σ der den Fehler in der Stichprobe wiedergibt. Jeder Mittelwert μ_{ij} ist eine deterministische Funktion der Zeit t_{ij} und des Interzepts α_i sowie der Ableitung β_i.

Die α_i werden unabhängig aus einer Verteilung mit Erwartung α_0 und Varianz σ_α gezogen, während die β_i unabhängig aus einer Verteilung mit entsprechenden Parametern β_0 und σ_β gezogen werden. Diese - zunächst verteilungsfreie Abhängigkeitsstruktur kann man anschaulich mit Hilfe eines Graphen darstellen:

Es ergibt sich ein gerichteter azyklischer Graph[11]. Die durchgezogenen Pfeile symbolisieren stochastische, die gestrichelten deterministische Abhängigkeit; z.B. ergibt sich die bedingte Verteilung von μ_{ij} gegeben $\alpha_i, \beta_i, t_{ij}$ einfach als Bildmaß des Produktmaßes der Verteilungen der Einflußgrößen.

[11] 'gerichtet', weil jede Verbindung von Knoten ein Pfeil ist (der eine Abhängigkeit symbolisiert); 'azyklisch', weil man - der Pfeilrichtung folgend - von keinem Knoten entlang der Pfeile wieder zu diesem Knoten zurück kommt.

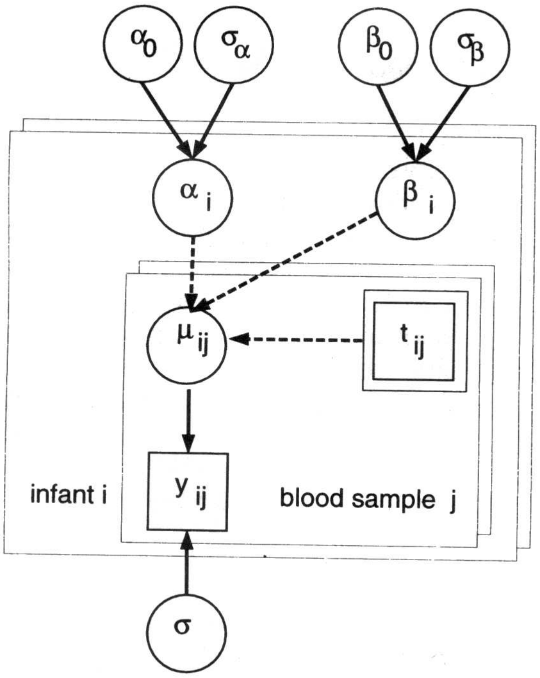


Abbildung 1.5: Graphisches Modell für die Hepatitis B Daten. Aus [12] mit freundlicher Erlaubnis von W.R. Gilks *et al.*

Was uns eigentlich interessiert ist die volle Verteilung Π auf dem Raum aller Parameter

$$\alpha_0, \sigma_\alpha, \beta_0, \sigma_\beta, \alpha_i, \beta_i, \mu_{ij}, t_{ij}, y_{ij}, \sigma.$$

Aus dem Graphen 1.5, kann man ablesen, daß sich die gemeinsame Verteilung aller Variablen als Produkt bedingter Verteilungen der Knoten gegeben die Eltern schreiben lassen, wobei die 'Eltern' eines Knotens v diejenigen Knoten sind, von denen ein durchgezogener Pfeil zum Knoten v geht. Gestrichelte Pfeile werden dabei weggelassen. Die deterministische Abhängigkeit schlägt sich einfach in einer deterministischen Transformation der (bedingten) Dichten nieder.

Lemma 1.4.1 *Im Graphen (1.5) gilt: Für eine Parameterkonstellation* $V \in \mathbb{R}^2 \times \mathbb{R}_+^3$ *(d.h. einen globalen Zustand auf dem Graphen) gilt*

$$\boldsymbol{P}(V) = \prod_{v \text{ Einzelparameter}} \boldsymbol{P}(v|\text{Eltern von } v). \tag{1.24}$$

Statt eines formalen Beweises argumentieren wir anschaulich:

Beispiel 1.4.2 Die gemeinsame Verteilung von β_0, σ_β und β_i ist

$$\begin{aligned} &\boldsymbol{P}(\beta_0 \in A, \sigma_\beta \in B, \beta_i \in C) \\ &\quad = \boldsymbol{P}(\beta_i \in C | \beta_0 \in A, \sigma_\beta \in B) \cdot \boldsymbol{P}(\beta_0 \in A, \sigma_\beta \in B) \\ &\quad = \boldsymbol{P}(\beta_i \in C | \beta_0 \in A, \sigma_\beta \in B) \boldsymbol{P}(\beta_0 \in A) \boldsymbol{P}(\sigma_\beta \in B) \end{aligned}$$

für β_0 unabhängig von σ_β. Ähnlich gilt mit der Unabhängigkeitsannahme von σ_0 und σ_β daß

$$\boldsymbol{P}(\alpha_0 \in A, \sigma_\alpha \in B, \alpha_i \in C) = \boldsymbol{P}(\alpha_i \in C | \sigma_0 \in A, \sigma_\alpha \in B) \boldsymbol{P}(\sigma_0 \in A) \boldsymbol{P}(\sigma_\alpha \in B).$$

Dies ergibt die Produktform dieser dreidimensionalen Marginalverteilungen, d.h. die Behauptung für die Teilgraphen mit den Knoten β_0, σ_β und β_i bzw. $\alpha_0, \sigma_\alpha, \alpha_i$. Indem man den Graphen weiter durchforstet, ergibt sich zwangsweise die Behauptung.

Wir sehen, daß diese einfache Argumentation daran liegt, daß der Graph keine Schleifen hat.

1.4.4 Wahrscheinlichkeitstheoretische Modellierung

Wir benötigen Likelihood- und a priori Modellannahmen. Zunächst die Likelihood-Annahmen:

$$\begin{aligned} y_{ij} &\sim \mathcal{N}(\mu_{ij}, \sigma^2) && (1.25) \\ \mu_{ij} &= \alpha_i + \beta_i(\log t_{ij} - \log 730) && (1.26) \\ \alpha_i &\sim \mathcal{N}(\alpha_0, \sigma_\alpha^2) && (1.27) \\ \beta_i &\sim \mathcal{N}(\beta_0, \sigma_\beta^2). && (1.28) \end{aligned}$$

Damit sind die bedingten Verteilungen spezifiziert. Das Modell enthält nun noch unabhängige Parameter $\sigma, \sigma_\alpha, \sigma_\beta, \alpha_0$ und β_0. Auf den Raum dieser Parameter wird nun die a-priori Verteilung aufgesetzt. Es wurden benutzt:

$$\begin{aligned} \alpha_0, \beta_0 &\sim \mathcal{N}(0, 10\,000) \\ \sigma, \sigma_\alpha, \sigma_\beta &\sim \Gamma(0,01; 0,01), \end{aligned}$$

wobei $\Gamma(a,b)$ eine Gamma-Verteilung mit Erwartungswert a/b und Varianz a/b^2 ist[12]. Somit ist die Varianz gleich $10^{-2}/10^{-4} = 100$. (Die Verteilungen sind also relativ flach, so daß sie das Ergebnis nicht allzu sehr beeinflussen.

1.4.5 Die bedingten Verteilungen

Wir erinnern uns: Wir leben auf einem Parameterraum, der wie beim Gibbs Sampler von Produktgestalt ist:

$$\begin{aligned}
& (\alpha_0; \sigma_\alpha; \beta_0; \sigma_\beta; \sigma; \\
& \alpha_i : i = 1, \ldots, 106; \beta_i : i = 1, \ldots, 106; \;\; \mu_{ij}, i = 1, \ldots, 106; j = 1, \ldots, n_i; \\
& y_{ij}, i = 1, \ldots, 106; j = 1, \ldots, n_i) \\
\in \;\; & \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}^{106} \times \mathbb{R}^{106} \times \mathbb{R}^{106 \times \max n_i} \times \mathbb{R}^{106 \times \max n_i} \\
=: \;\; & \prod_{i=1}^{N} \boldsymbol{X}_i
\end{aligned}$$

Der Gibbs Sampler zieht nun sukzessive Stichproben aus den bedingten Verteilungen der einzelnen Variablen ('Koordinaten', Zustände der Knoten im Graphen), gegeben die restlichen Variablen. Aus der Produktform der Gesamtverteilung ergibt sich für einen Knoten v, daß

$$\begin{aligned}
\boldsymbol{P}(v|\text{Rest}) &\propto \boldsymbol{P}(v, \mathit{Rest}) \qquad (1.29)\\
&\propto \text{Terme in } \boldsymbol{P}(V) \text{ die } v \text{ enthalten} \\
&\propto \boldsymbol{P}(v|\text{Eltern}(v)) \times \prod_{w \in \text{Kinder}(v)} \boldsymbol{P}(w|\text{Eltern}(w).
\end{aligned}$$

Damit können die für das Sampling benötigten bedingten Verteilungen bestimmt werden. Für die Interzepts α_i z.B. ergibt sich

$$\boldsymbol{P}(\alpha_i) \propto \exp\left(-\frac{(\alpha_i - \alpha_0)^2}{2\sigma_\alpha^2}\right) \times \prod_{j=1}^{n_i} \exp\left(\frac{(y_{ij} - \alpha_i - \beta_i(\log t_{ij} - \log 730))^2}{2\sigma^2}\right)$$

[12]Die Dichte der Gammaverteilung mit Parametern a, b ist

$$f_{a,b}(u) = \frac{1}{\Gamma(b)} a^b u^{b-1} \exp(-au), \;\; u > 0.$$

Erwartungswert und Varianz sind a/b und a/b^2. Die χ_k^2-Verteilung hat die Dichte $f_{1/2,k/2}$. Allgemeiner sind endliche Summen $\sum |\xi_i|^\alpha$ mit standardnormalverteilten ξ_i gammaverteilt.

Löst man das Quadrat auf, so sieht man, daß eine Normalverteilung vorliegt mit Erwartungswert

$$\frac{\frac{\alpha_0}{\sigma_\alpha^2} + \frac{1}{\sigma^2}\sum_{j=1}^{n_i} y_{ij} - \beta_i(\log t_{ij} - \log 730)}{\frac{1}{\sigma_\alpha^2} + \frac{n_i}{\sigma^2}}$$

und Varianz

$$\frac{1}{\frac{1}{\sigma_\alpha^2} + \frac{n_i}{\sigma^2}}.$$

Analog behandelt man β_i, α_0 und β_0; auch sie sind normalverteilt.

Ähnlich verhält es sich mit dem Präzisionsparameter $\tau_\alpha = \sigma_\alpha^{-2}$. Die allgemeine Vorschrift (1.29) besagt, daß die volle bedingte Verteilung von τ_α proportional zum Produkt der a priori Verteilung für τ_α und den Likelihoodtermen, gegeben durch 1.27 ist. Dies sind die Likelihoodterme für τ_α weil die Parameter α_i die einzigen Kinder von τ_α sind. Somit ergibt sich

$$\begin{aligned}
\boldsymbol{P}(\tau_\alpha | \cdot) &\propto \tau^{0.01-1} \mathrm{e}^{-0.01\tau_\alpha} \prod_{i=1}^{106} \tau_\alpha^{1/2} \exp\left(-\frac{1}{2}\tau_\alpha(\alpha_i - \alpha_0)^2\right) \\
&= \tau^{0.01+106/2-1} \exp\left(-\tau_\alpha\left(0.01 + \frac{1}{2}\sum_{i=1}^{106}(\alpha_i - \alpha_0)^2\right)\right) \\
&\propto \mathit{\Gamma}\left(0.01 + \frac{106}{2}, 0.01 + \frac{1}{2}\sum_{i=1}^{106}(\alpha_i - \alpha_0)^2\right).
\end{aligned}$$

Somit ist die volle bedingte Verteilung von τ_α wieder eine Gammaverteilung.

Entsprechend sind auch die vollen bedingten Verteilungen von σ_β^{-2} und σ^{-2} wieder Gammaverteilungen. Sowohl aus Normal- wie auch Gammaverteilungen läßt es sich bequem Sampeln; dafür stehen Standardalgorithmen zur Verfügung.

1.4.6 Darstellung (Monitoring) und Kontrolle der Ergebnisse

Die Werte für die unbekannten Größen, die mit Hilfe des Gibbs Samplers generiert wurden, müssen graphisch und statistisch zusammengefaßt werden um z.B. gute Konvergenz zu garantieren. Von besonderem Interesse ist die Frage, ob die 'burn in time' schon erreicht wurde. D.h. man muß prüfen, ob zwei Realisierungen des Gibbs Samplers mit verschiedenen Startzuständen 'zusammengelaufen' sind. Dies prüft man für jeden Skalar einzeln mit verschiedenen Startwerten.

Dazu läßt man den Gibbs Sampler für jede interessierende Größe ψ m mal mit verschiedenen Startwerten laufen: Seien ψ_{ij}, $j = \ldots, n$, $i = 1, \ldots, m$ die Werte, die beim Sampling in m Läufen der Länge n generiert wurden. Ähnlich wie bei der Varianzanalyse werden 'inter'-Varianzen B und 'intra'-Varianzen W berechnet:

$$\begin{aligned} B &= \frac{n}{m-1}\sum_{i=1}^{m}(\bar{\psi}_{i\cdot} - \bar{\psi}_{\cdot\cdot})^2, \;\; \bar{\psi}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n}\psi_{ij}, \; \bar{\psi}_{\cdot\cdot} = \frac{1}{m}\sum_{i=1}^{m}\bar{\psi}_i \\ W &= \frac{1}{m}\sum_{i=1}^{m}s_i^2, \;\; s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(\psi_{ij} - \bar{\psi}_{i\cdot})^2. \end{aligned}$$

Die Intervarianz B enthält einen Faktor n, weil es auf den inneren Mitteln $\bar{\psi}_{i\cdot}$ basiert welche selbst Mittel von n Werten ψ_{ij} sind. Aus diesen Varianzkomponenten werden zwei Schätzer der Varianz von Π gebildet. Zunächst wird

$$v\hat{a}r(\psi) = \frac{n-1}{n}W + \frac{1}{n}B$$

betrachtet. Dies ist unter Annahme der Stationarität, d.h. daß alle 1-dim Marginalverteilungen der Markovkette gleich Π sind, ein erwartungstreuer Schätzer der Varianz von Π. Da aber am Anfang der Simulation die einzelnen Ketten noch nicht stationär sind, überschätzt $v\hat{a}r(\psi)$ die Varianz, wenn die Startpunkte sehr verstreut gewählt werden. Auf der anderen Seite unterschätzt W die Varianz, weil in den einzelnen Läufen der Bereich von Π nur teilweise durchforstet wurde. Im Lauf der Annäherung strebt $v\hat{a}r(\psi)$ von oben und W von unten gegen die wahre Varianz. Man bildete nun

$$\sqrt{\hat{R}} = \sqrt{v\hat{a}r(\psi)/W},$$

die Gelman-Rubin-Statistik, welche gegen eins strebt. Diese Zahl berechnet man nun für mehrere interessierende Kenngrößen von Π und nimmt die Nähe zu eins als Gütemaß für die Simulation.

Im Beispiel wurden folgende Startwerte gewählt: Der erste Startwert ist nach den Daten plausibel, der zweite repräsentiert kleine Meßfehler und hohe Intergruppenvarianz und der dritte ähnliche Individuen und hohe Meßfehler. Es wurden mit diesen Startwerten 3 Runs mit 5000 Iterationen durchgeführt

Parameter	Run 1	Run 2	Run 3
α_0	5.0	20.0	-10.00
β_0	-1.0	-5.0	5.00
σ_α	2.0	20.0	0.20
σ_β	0.5	5.0	0.05
σ	1.0	0.1	10.00

Tabelle 1.1: Startwerte für verschiedene Runs des Gibbs Samplers

und für verschiedene Variablen die Mediane und 97,5 Perzentile der Gelman-Rubin Statistik geplottet; sie sollten bei eins liegen. Die Statistiken selbst hatten folgende Werte: Die Ergebnisse waren konsistent mit früheren Un-

Parameter	Schätzung	97,5 Perzentil
β_0	1.03	1.11
σ_β	1.01	1.02
σ	1.00	1.00

Tabelle 1.2: Gelman-Rubin Statistik für 3 Parameter

tersuchungen; z.B. ergab sich für den mittleren Gradienten eine Verteilung, die um minus eins zentriert war (also das vermutete Ergebnis). Anschließend wurde das Modell verfeinert, indem z.B. robustere a priori Verteilungen benutzt wurden, um die offensichtlichen Ausreißer besser zu modellieren, und Kovariablen eingebaut wurden.

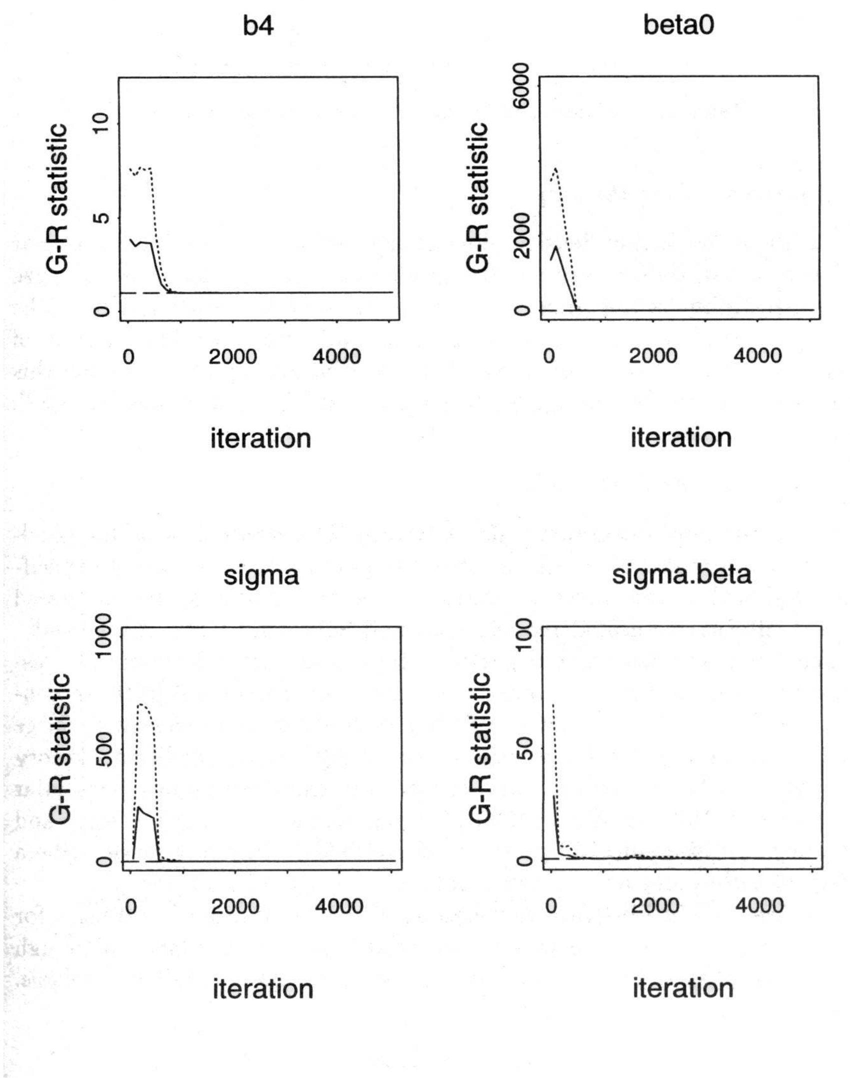


Abbildung 1.6: Gelman-Rubin Statistik für 4 Parameter und 3 Runs. Es wurden die Mediane (Linie) und 97,5 Perzentile (gestrichelte Linie) geplottet. Aus [12] mit freundlicher Erlaubnis von W.R. Gilks *et al.*

Kapitel 2

Konditionale Modelle

Wir befassen uns nun mit konditionalen Modellen; sie werden mit Hilfe von Markovketten dargestellt. Die Abhängigkeiten finden sich in den Übergangskernen wieder. Für diese brauchen wir eine flexible Klasse von Verteilungen, die wir zunächst einführen. Als Literatur für Exponentialfamilien und das verallgemeinerte lineare Modell verwenden wir [13], für konditionale Modelle und einige Beispiele [10].

2.1 Die Modellierung

Wir führen Exponentialfamilien, das verallgemeinerte lineare Modell und konditionale Modelle ein.

2.1.1 Exponentialfamilien

Wir werden uns im folgenden mit Verallgemeinerungen des klassischen linearen (Regressions-)Modelles der Statistik befassen. Bei letzterem beruht die (einfache und elegante) Theorie wesentlich auf der Normalverteilungsannahme. Für die Verallgemeinerung benötigen wir eine flexiblere Klasse von Verteilungen. Sie muß insbesondere stetige, diskrete und sogar kategerielle bzw. nominale Skalen zulassen.

Seien $b(\vartheta), c(\vartheta), a(y), t(y)$ reelle meßbare Funktionen. Eine Verteilung gehört zur *Exponentialfamilie* zu $c(\vartheta)$ und $t(y)$, wenn sie eine Dichte der folgenden Gestalt hat:

$$f(y, \vartheta) = \exp\left(c(\vartheta)t(y) + a(y) - b(\vartheta)\right)$$

(im Fall diskreter Verteilungen ist f als Zähldichte aufzufassen). Falls $t(y) = y$, so liegt die *kanonische Form* vor. $c(\vartheta)$ ist der *natürliche Parameter* der Verteilung.

Beispiel 2.1.1 (Die Gaußverteilung) hat die wohlbekannte Dichte

$$\tilde{f}(y;\mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right).$$

Setze $\vartheta = \frac{\mu}{\sigma^2}$. Wird σ als bekannt vorausgesetzt, so liegt eine einparametrische Familie vor und wir können rechnen

$$\begin{aligned} f\left(y,(\vartheta\sigma^2,\sigma^2)\right) &= \exp\left(\underbrace{\vartheta}_{c(\vartheta)}\cdot\underbrace{y}_{t(y)} - \underbrace{\frac{y^2}{2\sigma^2}}_{a(y)} - \underbrace{\left(\frac{\vartheta^2\sigma^2}{2} + \ln\sqrt{2\pi\sigma^2}\right)}_{b(\vartheta)}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\mu y}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) = \tilde{f}(y,\mu,\sigma^2). \end{aligned}$$

Konstanten im Exponenten können wir natürlich nach belieben in a oder b aufnehmen.

Beispiel 2.1.2 (Die Gammaverteilung) hat schon fast die Exponentialgestalt:

$$\tilde{f}\Big(y,(\alpha,\beta)\Big) = \frac{1}{\Gamma(\beta)}\alpha^\beta y^{\beta-1}e^{-\alpha y}, \quad y > 0.$$

Setzt man $\vartheta = -\alpha$ und nimmt bekanntes β an, so ergibt sich

$$\begin{aligned} f(y,\vartheta) &= \exp\left(\vartheta y + (\beta-1)\ln\left[y + \ln(-\vartheta)^\beta/\Gamma(\beta)\right]\right) \\ &= \frac{1}{\Gamma(\beta)}\alpha^\beta y^{\beta-1}e^{-\alpha y} = \tilde{f}(y,(\alpha,\beta)). \end{aligned}$$

Beispiel 2.1.3 (Die Binomialverteilung) hat die Zähldichte

$$\tilde{f}\left(y,(n,p)\right) = \binom{n}{y} p^y(1-p)^{n-y}.$$

Wir setzen n als bekannt voraus und

$$\vartheta = \ln\frac{p}{1-p} = \operatorname{logit} p.$$

$$
\begin{aligned}
f(y,\vartheta) &= \exp\left(\vartheta y + \underbrace{\ln\binom{n}{y}}_{a(y)} - \underbrace{n\ln(1+e^{\vartheta})}_{b(\vartheta)}\right)\\
&= \left(\frac{p}{1-p}\right)^{y}\binom{n}{y}\underbrace{\frac{1}{(1+e^{\vartheta})^{n}}}_{\frac{1-p}{1-p}+\frac{p}{1-p}=\frac{1}{1-p}} = \binom{n}{y}p^{y}(1-p)^{n-y}.
\end{aligned}
$$

Im Bernoullifall $n = 1$ reduziert sich die Formel auf

$$f(y,\vartheta) = \exp\big(\vartheta y - \ln(1+e^{\vartheta})\big) = \frac{e^{\vartheta y}}{1+e^{\vartheta}};$$

insbesondere gelten

$$f(1,\vartheta) = \frac{e^{\vartheta}}{1+e^{\vartheta}} = \frac{1}{1+e^{-\vartheta}}; \quad f(0,\vartheta) = \frac{1}{1+e^{\vartheta}},$$

was uns an das Ising Modell erinnert.

Beispiel 2.1.4 (Die Poissonverteilung) Hier hat man

$$\tilde{f}(y,\lambda) = e^{-\lambda}\cdot\frac{\lambda^{y}}{y!} = \exp(y\cdot\underbrace{\ln\lambda}_{\vartheta} - \underbrace{\ln y!}_{a(y)} - \lambda) = \exp(y\vartheta - \ln y! - e^{\vartheta}).$$

Nützlich und erleuchtend ist der folgende Satz 2.1.6. Wir schicken eine Definition und ein Lemma voraus. Als *natürlichen Parameterraum* Θ einer Exponentialfamilie definiert man die Menge der ϑ mit

$$\int \exp(c(\vartheta)t(y) + a(y))\,dy < \infty.$$

Er ist für die Wahl eines $c(\vartheta)$ aus den obigen Beispielen ein Intervall mit nichtleerem Innerem. Die Funktion f ist eine Dichte, hat deshalb Integral eins und auf Θ gilt somit

$$\int \exp(c(\vartheta)t(y) + a(y))\,dy = \exp(b(\vartheta). \tag{2.1}$$

Lemma 2.1.5 *Die Zufallsvariable Y habe die Dichte*

$$f(y,\vartheta) = \exp\Big(\vartheta \cdot t(y) + a(y) - b(\vartheta)\Big)$$

und Φ erfülle

$$\boldsymbol{E}_\vartheta(|\Phi(Y)|) < \infty$$

auf Θ. Dann ist die Funktion

$$\vartheta \mapsto \boldsymbol{E}_\vartheta(|\Phi(Y)|)$$

auf Θ^o beliebig oft differenzierbar. Insbesondere ist b auf Θ^o unendlich oft differenzierbar.

Beweis. Die erste Aussage findet sich in jeder Analysisvorlesung unter dem Stichwort 'Vertauschung von Differentiation und Integration', die Exponentialfunktion ist unendlich oft differenzierbar in ϑ und der letzte Teil gilt wegen (2.1). □

Der Satz liefert Erwartungswert und Varianz von Y (bzw. $t(Y)$).

Theorem 2.1.6 *Seien die Voraussetzungen des Lemmas erfüllt. Dann gelten für $\vartheta \in \Theta^o$ die Identitäten*

$$\boldsymbol{E}_\vartheta\Big(t(y)\Big) = \frac{db(\vartheta)}{d\vartheta}, \quad \boldsymbol{V}_\vartheta\Big(t(y)\Big) = \frac{d^2b(\vartheta)}{d\vartheta^2}.$$

Beweis. Es ist

$$\begin{aligned}
\boldsymbol{E}_\vartheta\Big(t(y)\Big) &= \exp\big(-b(\vartheta)\big)\int t(y)\exp\Big(\vartheta t(y) + a(y)\Big)\,dy \\
&= \exp\Big(-b(\vartheta)\Big)\int \frac{d}{d\varepsilon}\exp\Big(\vartheta t(y) + a(y)\Big)\,dy \\
&= \exp\Big(-b(\vartheta)\Big)\frac{d}{d\varepsilon}\int \exp\Big(\vartheta t(y) + a(y)\Big)\,dy \\
&= \exp\Big(-b(\vartheta)\Big)\frac{d}{d\vartheta}\,\exp\Big(b(\vartheta)\Big) = b'(\vartheta).
\end{aligned}$$

Genauso gilt

$$\begin{aligned}
\boldsymbol{E}_\vartheta\Big(t(y)^2\Big) &= e^{-b(\vartheta)}\int t(y)^2\exp\Big(\vartheta t(y) + a(y)\Big)\,dy \\
&= e^{-b(\vartheta)}\int \frac{d^2}{\partial\vartheta^2}\exp\Big(\vartheta t(y) + ay\Big)\,dy \\
&= e^{-b(\vartheta)}\frac{d^2}{d\vartheta^2}\exp(e^{b(\vartheta)}) = b''(\vartheta) + b'(\vartheta)^2
\end{aligned}$$

Die Behauptung folgt nun aus der Verschiebungsformel und der ersten Identität. □

In mehreren Dimensionen gilt die selbe Formel, wobei die Multiplikation als Skalarmultiplikation zu lesen ist; jetzt muß natürlich auf Dimensionsverträglichkeit geachtet werden.

$$\vartheta \in \mathbb{R}^d, \quad y \in \mathbb{R}^n, \quad t(y) \in \mathbb{R}^d, \quad b(\vartheta) \in \mathbb{R}, \quad a(y) \in \mathbb{R},$$

also

$$f(y, \vartheta) = \exp\left(\vartheta' t(y) + a(y) - b(\vartheta)\right).$$

Auch hier gilt

$$\exp\Big(b(\vartheta)\Big) = \int \exp\Big(\vartheta' t(y) + a(y)\Big)\, dy.$$

Für die Momente gilt entsprechend zu Satz 2.1.6:

$$\begin{aligned} \boldsymbol{E}_\vartheta\Big(t(Y)\Big) &= \frac{d}{d\vartheta} b(\vartheta) = \nabla b(\vartheta), \\ \boldsymbol{V}_\vartheta(t(Y)) &= \frac{d^2}{d\vartheta^2}\Big(b(\vartheta)\Big) (= \text{ Hessematrix von } b), \end{aligned}$$

wobei wir die Varianz einer Zufallsvariablen ξ mit $\boldsymbol{V}(\xi)$ bezeichnen[1].

Beispiel 2.1.7 Für die Gaußverteilung errechnet man (σ^2 bekannt):

$$\vartheta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^T, \quad t(y) = (y, y^2)^T, \quad a = 0,$$
$$b(\vartheta) = \frac{\mu^2}{2\sigma^2} + \ln\sqrt{2\pi\sigma^2};$$

der Ausdruck a ist in das t und ϑ gerutscht, weil σ^2 als Parameter aufgefaßt wird.

Beispiel 2.1.8 (Multinomialverteilung) Seien

$$p = (p_1, \ldots, p_m)T, \quad p_i > 0, \sum_{i=1}^m p_i < 1,$$
$$y = (y_1, \ldots, y_m)T, \quad y_i \in \mathbb{N}_0, \sum_{i=1}^m y_i \leq n$$

[1] Analog bezeichnet $\boldsymbol{V}(\xi)$ die Kovarianzmatrix eines Zufallsvektors $\xi = (\xi_1, \ldots, \xi - d)$

und

$$c(y) = \frac{n!}{y_1! \cdots y_m!(n - \sum_{j=1}^m y_j)!}.$$

Die Multinomialverteilung $M_m(n,p)$ ist dann gegeben durch

$$\begin{aligned} & f((y_1, \ldots y_m); (n,p)) \\ = \; & c(y) \prod_{i=1}^m p_i^{y_i} \cdot (1 - \sum_{i=1}^m p_i)^{n - \sum_{i=1}^m y_i} \\ = \; & \exp\left(\sum_{i=1}^m \left(\ln \frac{p_i}{1 - \sum_{j=1}^m p_j} \right) y_i + \ln c(y) - (- n \ln(1 - \sum_{i=1}^n p_i)) \right). \end{aligned}$$

Wir wollen wieder ein einfaches (Skalar-)Produkt im natürlichen Parameter erreichen und setzen daher

$$\vartheta_i = \ln \frac{p_i}{1 - \sum_{j=1}^n p_j},$$

so daß gilt

$$M_m(n,p)(y) = \exp\left(\vartheta' y + \ln c(y) - \left(n \ln(1 + \sum_{i=1}^m e^{\vartheta_i}) \right) \right).$$

2.1.2 Das verallgemeinerte lineare Modell (GLM)

Das klassische lineare Modell für ungruppierte normale Zielvariablen und deterministische Kovariablen hat die Gestalt

$$y_i = z_i' \beta + \eta_i, \quad i = 1, \ldots, n, \tag{2.2}$$

wobei der Designvektor z_i eine Funktion der Kovariablen x_i und β der Vektor der unbekannten Parameter ist. Die Kriteriumsvariable y_i setzt sich also linear aus einer Erwartungswertfunktion $z_i'\beta$ und einer Fehlervariable η_i zusammen; wesentlich sind

(1) die Fehlervariable η_i ist normalverteilt,

(2) die Erwartungswertfunktion $z_i'\beta$ ist linear in den unbekannten Modellparametern β.

In vielen wichtigen Fällen sind aber (1) und/oder (2) verletzt, z.B.

(1)
- Die Kriteriumsvariable ist nominal oder ordinal skaliert.
- Die Kriteriumsvariable ist zwar intervall-skaliert, aber nicht in eine normalverteilte transformierbar.

(2) Bei vielen biologischen Problemen ist die Erwartungswertfunktion exponentiell oder logistisch in den Modellparametern, oder - wie bei Kontingenztafelanalysen - logarithmisch.

Deshalb wird das Modell erweitert:

(1) Die Annahme von Normalverteilungen wird zu Exponentialfamilie abgeschwächt.

(2) Der Erwartungswert ist (erst) nach einer monotonen Transformation durch eine *Linkfunktion* linear in den Modellparametern.

Was ist der Preis für diese Verallgemeinerung?

- Es gibt keine direkten Verfahren zur Parameterschätzung (hier kommen den MCMC-Methoden ähnliche Verfahren ins Spiel).
- Es gibt i.a. keine exakten, sondern nur noch asymptotische Testverfahren.
- Die Analyseverfahren sind (noch?) nicht so gut ausgebaut wie im linearen Modell.

Wir definieren nun das verallgemeinerte lineare Modell. Dazu benötigen wir zwei wesentliche Zutaten:

(1) *Verteilungsannahmen*
Gegeben die Kovariablen x_i, sind die y_i unabhängig; sie gehören zu einer Exponentialfamilie mit (bedingter) Erwartung $\boldsymbol{E}(y_i|x_i) = \mu_i$. Möglicherweise kommt noch ein Störparameter τ^2 ins Spiel, der nicht von i abhängt:

$$f(z, \vartheta_i) = \exp\left(\frac{1}{\tau^2}(z\vartheta_i + a(z,\tau) - b(\vartheta_i))\right). \tag{2.3}$$

Wir setzen $b'' > 0$ auf Θ voraus.

(2) *Strukturelle Annahme*
Die Erwartung μ_i steht zum linearen Prädiktor $z_i'\beta$ in der Beziehung

$$g(\mu_i) = z_i'\beta \quad \text{bzw.} \quad \begin{pmatrix} g(\mu_1) \\ \vdots \\ g(\mu_n) \end{pmatrix} = z\beta. \tag{2.4}$$

Die Funktionen bedeuten:

– g ist die sogenannte *Linkfunktion*; sie muß invertierbar sein mit $g^{-1} = h$; d.h.

$$\mu_i = h(z_i'\beta).$$

– h heißt *Responsefunktion* .

Bemerkung 2.1.9 (a) Setzen wir in einem GLM mit unabhängigen Zufallsvariablen Y_i, deren Erwartungswerte die μ_i sind, in formaler Analogie zum linearen Modell

$$Y = \mu + \varepsilon, \tag{2.5}$$

mit dem Vektor von Zufallsvariablen $\varepsilon' = (\varepsilon_1, \ldots \varepsilon_n)$, so haben die ε_i den Erwartungswert 0. Ihre Dichte ist

$$f_{\varepsilon_i}(z) = f\Big(z + \mu_i, \vartheta_i\Big) = \exp\left(\frac{1}{\tau^2}\left(z\vartheta_i + a(z + \vartheta_i, \tau) - (b(\vartheta_i) - \mu_i\vartheta_i)\right)\right).$$

Die ε_i bleiben zwar unabhängig, sie sind aber *nicht* identisch verteilt. Wir sollten also nicht mit (2.5) arbeiten, sondern direkt mit (2.3) und (2.4).
(b) Es gelten

$$\begin{aligned} \mu_i &= \boldsymbol{E}(Y_i) = b'(\vartheta_i) \\ \sigma_i^2 &= \boldsymbol{V}(Y_i) = \tau^2 b''(\vartheta_i). \end{aligned}$$

Formal gilt für die Linkfunktion

$$g(\mu_i) = g(b'(\vartheta_i)),$$

d.h. g muß auf $b'(\Theta)$ definiert sein.

Wir stellen nun den *Zusammenhang zwischen μ und ϑ* her. Es galt

$$g(\mu_i) = z_i'\beta, \quad h(z_i'\beta) = \mu_i.$$

Nun ist aber

$$\mu_i = b'(\vartheta_i).$$

Wir nehmen an, daß b' invertierbar ist mit Inverser

$$\psi = b'^{-1}.$$

Dann kann man die Gleichung nach ϑ_i auflösen und erhält

$$\vartheta_i = \psi(\mu_i) = \psi\Big(h(z_i'\beta)\Big).$$

Die Linkfunktion g heißt *natürlich*, wenn

$$g = \psi$$

d.h., wenn gilt

$$\vartheta_i = \psi(h(z_i'\beta)) = z_i'\beta.$$

Merkregel: Natürliche Link- bzw. Responsefunktionen sind gerade diejenigen, welche zu

$$\vartheta_i = z_i'\beta \tag{2.6}$$

führen.

Durch die Verknüpfung von ϑ und β wird auch σ^2 eine Funktion von $z_i'\beta$ nämlich

$$\sigma_i^2 = \tau^2 b''(\vartheta_i) = \tau^2 b''\left(\psi(h(z_i'\beta))\right).$$

Auch andere wichtige Funktionen, wie die Likelihood können ähnlich ausgedrückt werden.

Beispiel 2.1.10 (Das lineare Modell als GLM) Wir brauchen

$$Y_i = z_i'\beta + \varepsilon_i, \tag{2.7}$$

wobei ε_i i.i.d. $\mathcal{N}(0, \sigma^2)$. Wir haben die Normalverteilung schon als Mitglied einer Exponentialfamilie entlarvt mit $b(\vartheta_i) = \frac{\vartheta_i^2}{2} + c_i$. Es ist $\mu_i = b'(\vartheta_i) = (\frac{\vartheta_i^2}{2} + c_i)' = \vartheta_i$, d.h. $\psi = Id$. Also gilt im natürlichen Fall $g = h = Id$ und $\mu_i = \vartheta_i = z_i'\beta$. Wir erhalten das Modell (2.7) also mit natürlicher Linkfunktion.

Beispiel 2.1.11 (Dichotome kategorielle Regression) Seien $Y_1, \ldots, Y_n$ Zufallsvariablen mit Werten $0, 1$ und

$$\boldsymbol{P}(Y_i = 1) = \pi_i = \boldsymbol{E}_{\pi_i}(Y_i).$$

Die π_i seien unbekannt. Die Daten seien

$$x_i' = (x_{0i}, \ldots, x_{pi}).$$

Mit einer beliebigen Linkfunktion g gilt:

$$g(\pi_i) = x_i'\beta.$$

Für $B(1, \pi_i)$ haben wir die Zähldichte

$$f(y, \pi_i) = \exp\left(y \ln \frac{\pi_i}{1 - \pi_i} + \ln(1 - \pi_i)\right)$$

mit natürlichen Parametern

$$\vartheta_i = \ln \frac{\pi_i}{1 - \pi_i} = \operatorname{logit} \pi_i.$$

Entscheiden wir uns für die natürliche Linkfunktion, so ist

$$g(\pi_i) = \psi(\pi_i) = \psi(\mu_i) = \vartheta_i = \ln \frac{\pi_i}{1 - \pi_i} = g(\pi_i).$$

Als Umkehrfunktion von ψ gilt dann für $h = b'$, daß

$$b'(\vartheta_i) = h(\vartheta_i) = \frac{e^{\vartheta_i}}{1 + e^{\vartheta_i}}.$$

Für b ergibt sich

$$b(\vartheta_i) = \ln\left(1 + \mathrm{e}_i^{\vartheta}\right) + c.$$

Rechnung ergibt $c = 0$. Damit haben wir alle Funktionen und Parameter im natürlichen Fall bestimmt. Das Modell heißt *binäres (dichotomes) logistisches Regressionsmodell*. Wegen $g(\pi_i) = g(\mu_i) = x_i'\beta$ hat man

$$x_i'\beta = g(\pi_i) = \ln \frac{\pi_i}{1 - \pi_i};$$

$$\pi_i = h(x_i'\beta) = \frac{1}{1 + e^{-x_i'\beta}}. \tag{2.8}$$

Wir haben also in Entsprechung zur Erwartungswertformel $\boldsymbol{E}_\beta(Y_i) = x_i'\beta$ die einfache Formel

$$\operatorname{logit} \pi_i = x_{i0}\beta_0 + \cdots x_{ip}\beta_p.$$

Beispiel 2.1.12 (Probitanalyse) Hier wählt man $h = \phi$, $g = \phi^{-1}$, wobei ϕ die Verteilungsfunktion der Standardnormalverteilung ist. Man beachte, daß sowohl die Responsefunktion h aus dem logistischen Regressionsmodell (vgl. (2.8) als auch $h = \phi$ aus dem Probitmodell ähnliche sigmoide Funktionen sind (vgl. Abb. 2.1.).

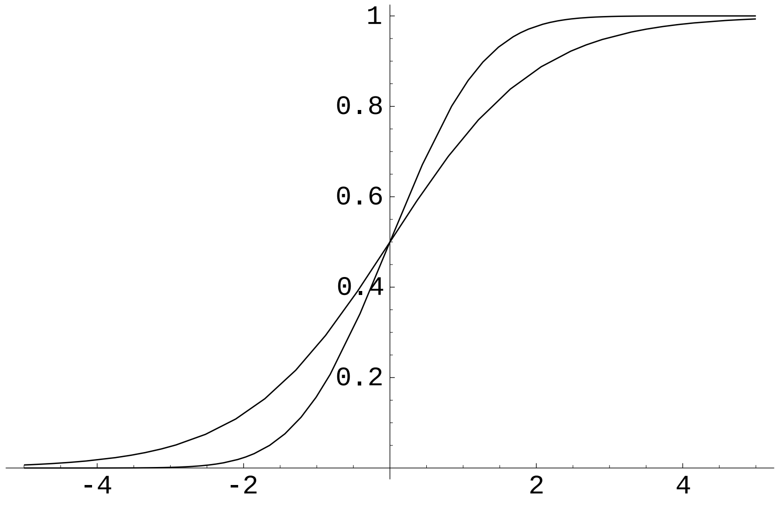


Abbildung 2.1: Graph der Verteilungsfunktion von $\mathcal{N}(0,1)$ und von $S(x) = 1/(1+\exp(-x))$ (im Ursprung flacher).

2.1.3 Zeitreihen und Longitudinaldaten

Zeitreihen sind wiederholte Beobachtungen einer Responsevariablen Y, die von eigentlichem Interesse ist und möglicherweise einem Vektor x von Kovariablen zu Zeitpunkten $t = 1, \ldots, T$. Man erhält also eine Zeitreihe

$$(y_1, x_1), \ldots, (y_T, x_T).$$

Bei *longitudinalen* oder *Paneldaten* an Einheiten $i = 1, \ldots, n$ sind die y_t Vektoren (y_{it}). Man kann sie als (Zeit-)Schnitt durch mehrere eindimensionale Zeitreihen auffassen. Modelle für (nichtgaußische) Daten sind relativ neu ($\geq$ 1970); solche sind z.B. binäre und multikategorielle Daten wie

es regnet	ja / nein
es regnet	nicht / wenig/ stark
Polioindex	Fallzahl im Monat.

2.1.4 Konditionale Modelle

Bedingte Verteilungen, gegeben die Vergangenheit zusammen mit Kovariablen ersetzen die unabhängigen Variablen.

Wir behandeln zunächst *verallgemeinerte autoregressive Modelle*. Bezeichne

$$\mathcal{H}_t = \{y_{t-1}, \ldots, y_1, x_t, x_{t-1}, \ldots, x_1\}$$

die *t-Vergangenheit*[2]. Diese Modelle sind charakterisiert durch:

(1) die *bedingten* Dichten

$$f(y_t|\mathcal{H}_t),\ t = 1, 2, \ldots, T$$

gehören zu einer Exponentialfamilie.

(2) die *bedingten* Erwartungen

$$\mu_t = \boldsymbol{E}(y_t|\mathcal{H}_t)$$

sind von der Gestalt

$$\mu_t = h(z_t'\beta),$$

wobei z_t nur von $\mathcal{H}_t$ abhängt. Für die bedingte Varianz ergibt sich (vgl. (2.1.6)):

$$\sigma_t^2 = \tau^2 b''(\vartheta),$$

wobei τ^2 und b'' zum Modell gehören.

Häufig bedingt man nicht durch alle $y_{t-1}, \ldots, y_1$ sondern nur durch $y_{t-1}, \ldots, y_{t-l}$. Dann erhält man ein *verallgemeinertes lineares Modell der Ordnung l*. Im Fall $l = 1$ reduziert sich dies auf eine gewöhnliche Markovkette.

Beispiel 2.1.13 (Binäre Zeitreihen) Sei (y_t), $y_t \in \{0, 1\}$. Dann ist die bedingte Verteilung von y_t gegeben $\mathcal{H}_t$ bestimmt durch

$$\pi_t = \boldsymbol{P}(y_t = 1|\mathcal{H}_t) = \boldsymbol{E}(y_t|\mathcal{H}_t).$$

Ohne Kovariablen ist das *reine autoregressive Modell der Ordnung l* gegeben durch

$$\pi_t = h(\beta_0 + \beta_1 y_{t-1} + \ldots + \beta_l y_{t-l}) = h(z_t'\beta),$$

[2]eigentlich die σ-Algebra, die von den vergangenen Variablen und Kovariablen erzeugt wird, sowie der gegenwärtigen Kovariablen; solange wir mit endlichem Zustandsraum arbeiten, können wir jedoch alles elementar niederschreiben.

wobei

$$\beta = (\beta_0, \ldots, \beta_l)', \;\; z_t' = (1, y_{t-1}, \ldots, y_{t-l}).$$

Die bedingte Varianz ist

$$\sigma_t^2 = \pi_t(1 - \pi_t);$$

Bei natürlicher Linkfunktion $g = \psi$ ist $h = b'$ und somit $\tau^2 = 1$. Man kann Kovariable aufnehmen, etwa

$$\pi_t = h(\beta_0 + \beta_1 y_{t-1} + \beta_l y_{t-l} + x_t'\gamma) = h(z'\beta), \; t > l,$$

mit

$$z_t' = (1, y_t, \ldots, y_{t-l}, x_t'), \; \beta' = (\beta_0, \beta_1, \ldots, \beta_l, \gamma).$$

Wegen des Einflusses der Kovariablen wird die Markovkette der Ordnung l *inhomogen.*

Im logistischen oder Logit-Modell wäre z.B.

$$h(\mu) = \frac{e^\mu}{1 + e^\mu}$$

(bei Parameter $\mu = b'(\vartheta)$).

Ein Fall, den wir genau anschauen werden, ist

$$\pi_t(b, a) = \boldsymbol{P}(y_t = a | y_{t-1} = b, x_t), \quad a, b \in \{0, 1\}.$$

Wieder induziert die Kovariable x_t eine zeitliche Inhomogenität. Bei Longitudinaldaten hätte man

$$\pi_{i,t}(b, a) = \boldsymbol{P}(y_{it} = a | y_{i,t-1} = b, x_{it}),$$

wobei i der Index für die Individuen ist. Im Falle der logistischen Regression hat man

$$\begin{array}{rcl} y_{it} & = & B(1, \pi_{it}) \\ \mathrm{logit}\pi_{it} & = & \beta_0 + \beta_1 y_{t-1} + \gamma x_t. \end{array}$$

Es können auch Interaktionen zwischen vergangenen Beobachtungen und den Kovariablen eingebaut werden, wie z.B. in

$$\mathrm{logit}\, \pi_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + \beta_3 y_{t-1} x_t.$$

Dies ist äquivalent zur folgenden parametrischen Form der Übergangswahrscheinlichkeit eines (homogenen) Markovprozesses

$$\pi_t(i,1) = P(y_t = 1 | y_{t-1} = i, x_t) = h(\alpha_{0i} + \alpha_{1i}x_t).$$

Dies drückt explizit die Abhängigkeit von vergangenem Zustand aus. Die Äquivalenz ergibt sich mit

$$\alpha_{00} = \beta_0,\ \alpha_{0,1} = \beta_0 + \beta_1,\ \alpha_{10} = \beta_2, \alpha_{11} = \beta_2 + \beta_3.$$

Zur Probe setzen wir ein; für $y_{t-1} = 0$ ergibt sich

$$\begin{aligned} \alpha_{00} + \alpha_{10}x_t &= \beta_0 + \beta_2 x_t \\ &= \beta_0 + \beta_1 \cdot 0 + \beta_2 x_t + \beta_3 \cdot 0 \cdot x_t \cdot x_t); \end{aligned}$$

für $y_{t-1} = 1$ erhalten wir

$$\alpha_{01} + \alpha_{11}x_t = \beta_0 + \beta_1 \cdot 1 + \beta_2 x_t + \beta_3 \cdot 1 \cdot x_t,$$

also die obige Identität. Ähnliches kann auch für Abhängigkeiten höherer Ordnung durchgeführt werden.

Bevor wir uns der statistischen Inferenz konditionaler Modelle zuwenden, betrachten wir das GLM. Einiges davon läßt sich übertragen.

2.2 Statistische Inferenz

Die statistische Inferenz für konditionale Modelle ist oft eine Übertragung der Methoden aus dem gewöhnlichen GLM auf den konditionalen Fall, wobei dort die (Marginal-) Verteilungen der i.i.d. Variablen durch bedingte Verteilungen ersetzt werden.

Da der Index jetzt nicht mehr die Nummer von austauschbaren i.i.d. Variablen, sondern die Zeit bedeutet schreiben wir jetzt t statt i.

2.2.1 Inferenz im GLM

Zunächst schauen wir uns die Loglikelihoodfunktion an. Für unabhängige Beobachtungen $Y_1, \ldots, Y_T$ nehmen wir eine Verteilung aus einer Exponentialfamilie an, also eine Dichte

$$f(y_t, \vartheta_t) = \exp\Big(\frac{1}{\tau^2}(y_t\vartheta_t + a(y_t,\tau) - b(\vartheta_t))\Big). \tag{2.9}$$

Theorem 2.2.1 *Gegeben sei die Beobachtung $y = (y_1, \ldots, y_T)'$. In Abhängigkeit von β ist die Loglikelihoodfunktion gegeben durch*

$$l(\beta) = \sum_{t=1}^{T} \frac{1}{\tau^2} \Big(y_t \psi(h(\eta_t)) + a(y_t, \tau) - b(\psi(h(\eta_t))) \Big)$$

wobei $\eta_t = x_t'\beta$ als Funktion von β aufgefaßt wird.

Beweis. Produktbildung in (2.9) und Logarithmieren ergibt

$$\sum_{t=1}^{T} \frac{1}{\tau^2} \Big(y_t \vartheta_t + a(y_t, \tau) - b(\vartheta_t) \Big).$$

Es waren g die Linkfunktion und $h = g^{-1}$ die Responsefunktion, ferner $\psi = b'^{-1}$. Wir hatten

$$\vartheta_t = \psi(\mu_t) = \psi(h(\eta_t)),$$

wobei $\eta_t = x_t'\beta$. Dies wird einfach eingesetzt und die Formel ist verifiziert. □

ML-Schätzer maximieren die Loglikelihoodfunktion. Der erste Ansatz dazu ist,

$$U(\beta) = \nabla_\beta l(\beta) = 0$$

zu setzen. Der Gradient $U(\beta)$ heißt *Scorefunktion*. Dann wäre zu prüfen, ob die Lösung wirklich ein Maximum liefert und nicht ein lokales Minimum oder einen Sattelpunkt. Wir schreiben

$$l^{(t)}(\vartheta_t) = Y_t \vartheta_t + a(Y_t, \tau) - b(\vartheta_t),\ \vartheta_t = \psi(\mu_t(\eta_t))(\beta).$$

Mit der Kettenregel hat man

$$\frac{\partial l^{(t)}}{\partial \beta_j} = \frac{\partial l^{(t)}}{\partial \vartheta_t} \cdot \frac{\partial \vartheta_t}{\partial \mu_t} \cdot \frac{\partial \mu_t}{\partial \eta_t} \cdot \frac{\partial \eta_t}{\partial \beta_j}. \tag{2.10}$$

Für die einzelnen partiellen Ableitungen bekommt man

$$\begin{aligned}
\frac{\partial l^{(t)}}{\partial \vartheta_t} &= Y_t - b'(\vartheta_t) = Y_t - \mu_t \\
\frac{\partial \mu_t}{\partial \vartheta_t} &= b''(\vartheta_t) = \frac{1}{\tau^2}\sigma_t^2. \quad \text{Umkehrfunktion !} \\
\frac{\partial \eta_t}{\partial \beta_j} &= x_{tj}.
\end{aligned}$$

Es fehlt noch

$$\frac{d\mu_t}{d\eta_t} = \left.\frac{dh(\eta)}{d\eta}\right|_{\eta=\eta_t=x_t'\beta} = \left.\left(\frac{dg(\mu)}{d\mu}\right)^{-1}\right|_{\mu=\mu_t=h(x_t'\beta)}$$

Wir setzen in (2.10) ein und erhalten

$$\frac{\partial l^{(t)}}{\partial \beta_j} = (Y_t - \mu_t) \cdot \frac{\tau^2}{\sigma_t^2} \cdot \frac{d\mu_t}{d\eta_t} \cdot x_{tj}$$

und somit

$$U_j(\beta) = \sum_{t=1}^{T} x_{tj}(Y_t - \mu_t(\beta)) \left(\frac{d\mu_t}{d\eta_t}\right) \cdot \frac{1}{\sigma_t^2(\beta).}$$

Wichtig ist auch die Matrix

$$I(\beta) = (I_{jk}(\beta)), \quad I_{jk} = \boldsymbol{E}_\beta\Big(U_j(\beta)U_k(\beta)\Big),$$

die $p \times p$ *Fisher-Informationsmatrix*. Weil die Y_t unabhängig sind, gilt

$$\boldsymbol{E}_\beta(Y_t - \mu_t)(Y_s - \mu_s) = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases}$$

und deshalb

$$\begin{aligned}
&I_{jk}(\beta) \\
&\quad = \boldsymbol{E}_\beta\Big(U_j(\beta)U_k(\beta)\Big) \\
&\quad = \boldsymbol{E}\left(\sum_t x_{tj}\Big(Y_t - \mu_t(\beta)\Big)\left(\frac{d\mu_t}{d\eta_t}\right)\frac{1}{\sigma_t^2(\beta)} \sum_t x_{tk}\Big(Y_t - \mu_t(\beta)\Big)\left(\frac{d\mu_t}{d\eta_t}\right)\frac{1}{\sigma_t^2(\beta)}\right) \\
&\quad = \sum_t x_{tj}x_{tk}\left(\frac{d\mu_t}{d\eta_t}\right)^2 \boldsymbol{E}_\beta\Big((Y_t - \mu_t)^2\Big)\frac{1}{\sigma_t^4(\beta)} \\
&\quad = \sum_t x_{tj} \cdot x_{tk}\left(\frac{d\mu_t}{d\eta_t}\right)\frac{1}{\sigma_t^2}.
\end{aligned}$$

Um das in Matrixschreibweise zu übersetzen definieren wir

$$\begin{aligned}
X &= (x_1, \ldots, x_T), \quad \frac{d\mu}{d\eta} = \operatorname{Diag}\left(\frac{d\mu_t}{d\eta_t}\right) \\
V(\beta) &= \operatorname{Diag}\Big(\sigma_t^2(\beta)\Big) = \tau^2\Big(b''(\psi(\mu_t))\Big)
\end{aligned}$$

und fassen zusammen:

Proposition 2.2.2 *Im GLM (mit i.i.d. Zufallsvariablen $Y_1, \ldots, Y_T$) gelten:*

$$\begin{aligned} U_j(\beta) &= \sum_{t=1}^{n} x_{tj}(Y_t - \mu_t(\beta)) \left(\frac{d\mu_t}{d\eta_t}\right) \cdot \frac{1}{\sigma_t^2(\beta)} \\ I_{jk}(\beta) &= \sum_{t=1}^{n} x_{tj} \cdot x_{tk} \left(\frac{d\mu_t}{d\eta_t}\right)^2 \frac{1}{\sigma_t^2}, \end{aligned}$$

bzw.

$$\begin{aligned} U(\beta) &= X'V^{-1}(\beta) \left(\frac{d\mu}{d\eta}\right) \Big(Y - \mu(\beta)\Big), \\ I(\beta) &= X'V^{-1}(\beta) \left(\frac{\mu_t}{d\eta_t}\right)^2 X. \end{aligned}$$

Neben dem Vektor $U(\beta)$ der ersten Ableitungen der Loglikelihoodfunktion ist auch die $p \times p$-Matrix

$$W(\beta) = (W_{jk}((\beta)) = \frac{d^2 l(\beta)}{d\beta d\beta'}$$

der zweiten Ableitungen von Interesse. Dazu führen wir - mit der Schreibweise $\sigma_t^2 = \sigma^2(\eta_t)$ und mit $u = \psi \circ h$ - die Abkürzung

$$v_t(\beta) = \frac{d}{d\eta}\left(\frac{1}{\sigma^2(\eta)} \cdot \frac{dh(\eta)}{d\eta}\right)\Big|_{\eta=\eta_t} = \frac{1}{\tau^2}\frac{d^2}{d\eta^2}u(\eta)\Big|_{\eta=\eta_t}, \quad (\eta_t = x_t'\beta)$$

ein, sowie die $n \times n$-Matrix

$$R(\beta) = \mathrm{Diag}\Big(v(\beta)\Big).$$

Proposition 2.2.3 *Für ein GLM gilt*

$$W(\beta) = X'R(\beta)\mathrm{Diag}\Big(Y - \mu(\beta)\Big)X - I(\beta),$$

Beweis. Unter Benutzung der Formel für U_j gilt mit $v_i \equiv v_i(\beta)$ für das Element $W_{j,k}$:

$$\begin{aligned} W_{jk}(\beta) &= \frac{\partial}{\partial\beta_k} U_j(\beta) \\ &= \sum_t x_{tj}(Y_t - \mu_t) v_t \cdot \frac{\partial\eta_t}{\partial\beta_k} - \sum_t x_{tj} \left(\frac{d\mu_t}{d\eta_t}\right)^2 \frac{\partial\eta_t}{\partial\beta_k} \cdot \frac{1}{\sigma_t^2} \\ &= \sum_t x_{tj} x_{tk}(Y_t - \mu_t) v_t - \sum_t x_{tj} x_{tk} \left(\frac{d\mu_t}{d\eta_t}\right)^2 \cdot \frac{1}{\sigma_t^2}, \end{aligned}$$

woraus mit der Formel für I_{jk} die Behauptung folgt. □

Für die natürliche Linkfunktion

$$g = \psi = (b')^{-1}$$

sieht alles etwas freundlicher aus. Man erinnere sich: es gilt

$$\vartheta_t = \psi\Big(h\underbrace{(x_t'\beta)}_{\eta_t}\Big) = \eta_t.$$

Bei natürlichen Linkfunktionen fallen die Parameter ϑ_t und η_t zusammen, und wir haben für den Parameter ϑ der Exponentialfamilie ein lineares' Modell vorliegen:

$$g(\mu_t) = \vartheta_t = \eta_t = x_t'\beta,$$

Für die berechneten Größen erhalten wir im Fall natürlicher Linkfunktion vereinfachte Ausdrücke:

Theorem 2.2.4 *Für ein GLM mit natürlicher Linkfunktion gilt*

$$\begin{aligned}
l(\beta) &= \sum_{t=1}^{n} \frac{1}{\tau^2}(Y_t\eta_t + a(Y_t,\tau) - b(\eta_t)), \quad \eta_t = x_t'\beta \\
U_j(\beta) &= \sum_{t=1}^{n} \frac{1}{\tau^2} x_{tj}\Big(Y_t - \mu_t(\beta)\Big), \quad U(\beta) = \frac{1}{\tau^2}X'\Big(Y - \mu(\beta)\Big) \\
I_{tj}(\beta) &= \sum_{t=1}^{n} \frac{1}{\tau^4} x_{tj}x_{tk}\sigma_t^2(\beta), \quad I(\beta) = \frac{1}{\tau^4}X'V(\beta)X \\
W_{tk}(\beta) &= -I_{tk}(\beta), \quad W(\beta) = -I(\beta)
\end{aligned}$$

Beweis. Setze $\vartheta_t = \psi\Big(h(\eta_t)\Big) = \eta_t$ in die allgemeine Form ein. Wegen

$$\frac{d\mu_t}{d\eta_t} = b''(\vartheta_t) = \frac{1}{\tau^2}\sigma_t^2$$

ist

$$\left(\frac{d\mu}{d\eta}\right) = \frac{1}{\tau^2}V(\beta)$$

einzusetzen, und wegen $u = \psi \circ h = Id$ gilt $R = 0$. □

Beispiel 2.2.5 *Bei der binären Regression hatten wir*

$$g(\mu_t) = g(\pi_t) = \text{logit}(\pi_t) = \ln \frac{\pi_t}{1-\pi_t} (= \vartheta_t) = \eta_t = z_t'\beta.$$

Also gilt

$$\begin{aligned} l(\beta) &= \sum_{t=1}^{n} \Big(Y_t \eta_t - \ln(1 + e^{\eta_t}) \Big) \\ U(\beta) &= X'(Y - \pi); \quad \pi_t = \frac{1}{1 + e^{-\eta_t}} \\ I(\beta) &= X'\text{Diag}\Big(\pi_t(1-\pi_t)\Big) X \\ W(\beta) &= -I(\beta). \end{aligned}$$

Setzt man voraus, daß X vollen Rang hat und $0 < \pi_t < 1$, so ist die Lösung von $U(\beta) = 0$ (falls existent) eindeutig. Zur Berechnung der Lösung gibt es mehrere Algorithmen, z.B. das Newton-Verfahren (vgl. Abschnitt 2.2.2).

Beispiel 2.2.6 (Zähldaten) Wir betrachten zunächst die Exponentialform der Poissonverteilung

$$\tilde{f}(y, \lambda_t) = e^{-\lambda_t} \frac{\lambda_t^y}{y!} = \exp\left(y \ln \lambda_t - \ln y! - \lambda_t\right).$$

Der natürliche Parameter ist $\vartheta_t = \ln \lambda_t$, d.h. wir bekommen die Exponentialform

$$f(y, \vartheta_t) = \exp\left(y\vartheta_t - \ln y! - e^{\vartheta_t}\right).$$

Also ist $b(\vartheta) = e^{\vartheta}$; $b'(\vartheta) = e^{\vartheta}$; $\psi(\mu) = \ln(\mu)$ und deshalb für die natürliche Link- und Responsefunktion

$$\eta_t = z_t'\beta = g(\mu_t) = \ln(\mu_t); \quad \mu_t = h(z_t'\beta) = e^{z_t'\beta}.$$

Wegen $\mu_t = \lambda_t$ bekommen wir mit den Designvariablen $(1, y_{t-1}, \ldots, y_{t-l}, x_t')$ und den Modellparametern $\beta = (\beta_0, \beta_1, \ldots, \beta_l, \gamma')$ das log-lineare Poisson-Modell

$$\lambda_t = \boldsymbol{E}(y_t | H_t) = h(z_t'\beta) = \exp(\beta_0 + \beta_1 y_{t-1} + \ldots + \beta_l y_{t-l} + x_t'\gamma),$$

bzw.

$$\ln \lambda_t = \beta_0 + \beta_1 y_{l-1} + \ldots + \beta_l y_{t-l} + x_t'\gamma.$$

Für $\gamma = 0$ ist das ein reines autoregressives Modell. Es ist

$$\begin{aligned} l_t(\beta) &= y_t z_t'\beta - e^{z_t'\beta}, \;\; l(\beta) = \sum_t l_t(\beta) \\ U_j(\beta) &= \sum_t y_t z_{tj} - z_{tj} e^{z_t'\beta} \\ &= \sum_t z_{tj}(y_t - \underbrace{e^{z_t'\beta}}_{\mu_t = \lambda_t}). \end{aligned}$$

(in Übereinstimmung mit dem allgemeinen Ausdruck).

2.2.2 Inferenz in konditionalen Modellen

Die gemeinsame Dichte von $Y_1, \ldots, Y_T$ bei β läßt sich bei deterministischen Kovariablen faktorisieren[3]:

$$f_\beta(y_1, \ldots, y_T) = \prod_{t=1}^{T} f_\beta(y_t | y_{t-1}, \ldots, y_1).$$

Sind die Kovariablen stochastisch, so hat man

$$f_\beta(y_1, \ldots, y_T; x_1, \ldots, x_T) = \prod_{t=1}^{T} f_\beta(y_t | \mathcal{H}_t) \prod_{t=1}^{T} f_\beta(x_t | \mathcal{C}_t),$$

wobei

$$\mathcal{H}_t = (y_{t-1}, \ldots, y_1; x_t, \ldots, x_1); \quad \mathcal{C}_t = (y_{t-1}, \ldots, y_1; x_{t-1}, \ldots, x_1).$$

Unter der Annahme, daß der x-Term nicht von β abhängt, kann man sich bei der Parameterschätzung auf den ersten Term beschränken. In jedem Fall hat man

$$l(\beta) = \sum_{t=1}^{T} l_t(\beta), \quad l_t(\beta) = \ln f_\beta(y_t | \mathcal{H}_t),$$

[3]Dies entspricht der elementaren Formel für Ereignisse $A_1, \ldots, A_T$ mit $\boldsymbol{P}(A_1 \cap \ldots \cap A_T) > 0$, nämlich:

$$P(A_1 \cap \ldots \cap A_T) = P(A_1) P(A_2 | A_1) \cdots P(A_T | A_1 \cap \ldots \cap A_{T-1}).$$

wobei die bedingten Dichten durch das autoregressive Modell gegeben sind.

Für Scorefunktion etc. gelten die Formeln wie bei den GLM, jedoch mit bedingten Größen; so gilt z.B.

$$U(\beta) = \sum_{t=1}^{T} X_t' V_t^{-1} \left(\frac{d\mu}{\partial \eta} \right) (y_t - \mu_t(\beta)),$$

wobei $\mu_t(\beta) = h(X_t'(\beta))$ der bedingte Erwartungswert ist, $V_t = \mathrm{cov}(y_t|\mathcal{H}_t)$ und $\frac{d\mu}{d\eta}$ ausgewertet in $X_t\beta$ sind. Für univariaten Response ist $X_t = x_t'$ und $V_t(\beta) = \mathrm{Diag}(\sigma_t^2)$ die bedingte Varianz.

Merkregel. Die für die konditionale Maximum-Likelihood-Schätzung relevanten Größen sind formal identisch mit denen für i.i.d. Variablen, wenn die Zielvariablen $y_1, \ldots, y_{t-1}$ als zusätzliche Kovariablen aufgefaßt werden.

Somit können übliche Prozeduren für ML-Schätzung bei geeigneter Wahl der Parameter auch im konditionalen Modell eingesetzt werden.

Dies besagt noch lange nichts über die theoretische Fundierung der ML-Methode. Für das GLM im i.i.d. Fall existieren zumindest asymptotische Aussagen über die Konvergenz der Schätzer von der Art $\hat{\beta} \overset{n \to \infty}{\sim} \mathcal{N}(\beta, G^{-1})(\tilde{\beta})$. (mit wachsender Stichprobe) gegen normalverteilte Zufallsvariablen (siehe Pruscha (1996) VII. 3, [13]). Darauf können theoretisch untermauerte Tests und Konfidenzintervalle aufgebaut werden. Solche Resultate sind für das konditionale Modell nur in Einzelfällen verfügbar; vgl. Pruscha, Diskussion Paper und Fahrmeir-Tutz (1994), S. 195, [10]. Wir sind hier im Bereich der aktuellen Forschung.

Annahmen unter denen asymptotische Normalität gilt, sind natürlich solche, die das Modell nahe am i.i.d. Fall halten, wie (asymptotische) Stationarität und Ergodizität, die uns ja auch im 1. Kapital Grenzwertsätze erlaubt haben.

Zur Berechnung der Schätzer kann z.B. das *Newton-Verfahren* verwendet werden: Sei $G \subset \mathbb{R}^P$ offen und die Zielfunktion

$$l: \mathbb{R}^P \longrightarrow \mathbb{R}, (\beta_1, \ldots, \beta_P) \longrightarrow l(\beta)$$

mindestens zweimal stetig differenzierbar. Mit $W(\beta)$ bezeichnen wir die Hessematrix der gemischten zweiten partiellen Ableitungen von l im Punkte β. Wir fordern, daß W nichtsingulär im Punkte β ist. Das Iterationsverfahren ist gegeben durch die Vorschrift:

Sei $\beta^{(0)}$ der Startwert; ist $\beta^{(n)}$ gegeben, so approximieren wir l durch eine quadratische Funktion

$$l^{(n)}(\beta) = l(\beta^{(n)}) + \nabla l(\beta^{(n)})(\beta - \beta^{(n)}) + \frac{1}{2}(\beta - \beta^{(n)})' W(\beta^{(n)})(\beta - \beta^{(n)}).$$

und bestimmen $\beta^{(n+1)}$ als Wurzel der Gleichung

$$\nabla l^{(n)}(\beta) = 0.$$

Ist $W(\beta^{(n)})$ nichtsingulär, so ist $\beta^{(n+1)}$ eindeutig bestimmt. Das sieht man so:

$$\begin{aligned}
&\frac{\partial}{\partial \beta_i} l^{(n)}(\beta) \\
&= \underbrace{\frac{\partial}{\partial \beta_i} l^{(n)}(\beta^{(n)})}_{=0} + \nabla l^{(n)}(\beta^{(n)}) \frac{\partial}{\partial \beta_i}(\beta - \beta^{(n)}) + \frac{1}{2}\frac{\partial}{\partial \beta_i} W(\beta^{(n)}) \\
&= \frac{\partial}{\partial \beta_i} \sum_{j=1} (\beta_i - \beta^{(n)})^2 \nabla l^{(n)}(\beta^{(n)}) + W(\beta^{(n)})(\beta^{n+1} - \beta^{(n)}) = 0
\end{aligned}$$

d.h.

$$\beta^{(n+1)} = \beta^{(n)} - W(\beta^{(n)})^{-1} \nabla l(\beta^{(n)}). \tag{2.11}$$

Die Rekursionsformel (2.11) nennt man das Newton-Verfahren. Um $\beta^{(n+1)}$ zu berechnen, ist es nicht notwendig, die Matrix W zu invertieren(Inversion ist ein instabiles Verfahren). Es genügt, das Gleichungssystem

$$W(\beta^{(n)})\beta = W(\beta^{(n)})\beta^{(n)} - \nabla l^{(n)}(\beta^{(n)})$$

zu lösen.

2.2.3 Polio Inzidenz in den USA

Gemessen wurde die monatliche Polio Inzidenz der USA von 1970-1983[4]. Dies ergibt eine Zeitreihe von Zähldaten. Traditionelle Methoden erscheinen anhand der Daten wenig geeignet. Die Daten suggerieren saisonale Einflüsse

[4]Inzidenz ist die Anzahl des (Neu-)Auftretens einer Krankheit in einem Einheitszeitraum, z.B. pro Tag, Woche, Monat, Jahr.

und ein eventuelles Abnehmen bei der Polioinzidenz. Es bietet sich ein loglineares Poisson-Modell mit Trend und saisonalen Komponenten als Kovariablen an. Außerdem hängen die Beobachtungen sicher voneinander ab, d.h. konditionale Modelle könnten eingesetzt werden.

In [10], Beispiel 6.1, S. 199, wird folgende Vorgehensweise beschrieben. Es wird eine Regression von y_t auf einen linearen Trend, sowie auf Cosinus- und Sinuspaare für die jährlichen und halbjährlichen saisonalen Komponenten angesetzt. Ferner werden vergangene Zählungen durch Konditionierung berücksichtigt. Also wird von einem bedingten log-linearen Poissonmodell ausgegangen:

$$\lambda_t = \boldsymbol{E}(y_t|y_{t-1},\ldots,y_{t-l}) = h(z_t'\beta) = \exp(\alpha + \beta t \cdot 10^{-3} + z_t'\delta + \sum_{j=1}^{l} \gamma_j y_{t-j})$$

wobei der Term z_t Komponenten der Form

$$\cos(2\pi t/12),\ \sin(2\pi t/12),\ \cos(2\pi t/6),\ \sin(2\pi t/6)$$

enthält und den saisonalen Teil des Modelles repräsentiert.

Für das autoregressive Modell der Ordnung $l = 5$ wurden über die ML-Methode die Parameter $\alpha, \beta, \delta_1, \ldots, \delta_4, \gamma_1, \ldots, \gamma_5$ geschätzt (siehe Tab. 2.1). Ein abnehmender Trend wird durch das negative Vorzeichen von $\hat{\beta}$ nahegelegt.

2.2.4 Die indonesische Kindergesundheitsuntersuchung (ICHS)

In Westjava wurden 3000 Kinder vierteljährlich untersucht, um den Einfluß und Ursachen von Vitamin-A-Mangel zu studieren. Bestimmt wurden das Vorliegen von respiratorischen Darminfektionen (RD) und Xerophtalmie, einer Folge am Auge des Vitaminmangels. Die Infektionen sind eine der Hauptursachen für Kindermortalität in der dritten Welt. Einen Zusammenhang nachzuweisen ist wichtig, weil der Vitamin-A-Mangel (X) mit wenigen Pfennigen pro Kind und Jahr gebessert werden kann (unsere Quelle ist Diggle, Liang, Zeger: Analysis of Longitudinal Data (1994), [7]).

Die gesamten vorkommenden empirischen Übergangshäufigkeiten von RD (alle Übergänge) sind in der Tabelle 2.2 zusammengefaßt.

Insbesondere hat der Übergang $1 \to 1$ mit 13,5 % fast die doppelte Häufigkeit

Term	MLE	p-Wert
1	0.160	0.523
$t \cdot 10^{-3}$	-3.332	0.095
$\cos(2\pi t/12)$	-0.217	0.116
$\sin(2\pi t/12)$	-0.462	0.002
$\cos(2\pi t/6)$	0.128	0.354
$\sin(2\pi t/6)$	-0.372	0.008
y_{t-1}	0.085	0.022
y_{t-2}	0.040	0.445
y_{t-3}	-0.040	0.501
y_{t-4}	0.029	0.532
y_{t-5}	0.081	0.059

Tabelle 2.1: Parameterschätzung für das Polio Inzidenz Model aus Abschnitt 2.2.3 (nach [10], S. 199)

		Y_{it}		
		0	1	
Y_{it-1}	0	721 0.923	60 0.077	781 1.0
	1	64 0.865	10 0.135	74 1.0

Tabelle 2.2: Empirische Übergangshäufigkeiten für RD

wie der Übergang $0 \rightarrow 1$ mit 7,7 %.

Von Interesse ist der Einfluß von (X) auf (RD). Für die Untersuchungen Nr. 2-7 ergibt sich die Tabelle 2.3. (RD) kommt bei Kindern mit (X) $0,119/0,08 = 1,49$ mal so häufig vor als bei solchen ohne (X).

Wir beziehen nun die Abhängigkeit von einer vorhergehenden Infektion mit ein, indem wir nach $Y_{it-1} = 0$ und $Y_{it-1} = 1$ trennen (vgl. Tab. 2.4):

		Y_{it}		
		0	1	
X_{it}	0	7.48 0.942	65 0.08	813 1.0
	1	37 0.881	5 0.119	42 (1.0)
				855

Tabelle 2.3: Einfluß von (X) auf (RD)

		Y_{it}		
		0	1	
X_{it}	0	688 0.925	56 0.075	744 1.0
	1	33 0.892	4 0.108	37 1.0
$Y_{it-1} = 0$				781

		Y_{it}		
		0	1	
X_{it}	0	60 0.87	9 0.13	69 1.0
	1	4 0.8	1 0.2	5 1.0
$Y_{it-1} = 0$				74

Tabelle 2.4: Einfluß der Vergangenheit

Unter den vorher nicht infizierten Kindern ist die Häufigkeit einer jetzigen Infektion bei (X) $1,44 = 0/0,075$ höher als ohne (X). Bei den vorher infizierten ergibt sich $1,54 = 0,2/0,13$ also ein ähnliches Verhältnis.

Solche Überlegungen legen ein Modell

$$\operatorname{logit} \pi_{it} = x'_{it}\beta + \alpha y_{it-1}$$

nahe. Im einfachsten Fall geht nur (X) ein:

$$\operatorname{logit} \pi_{it} = \beta_0 + \underbrace{X_{it}}_{=0,1} \beta_1,$$

wobei x_{it} univariat ist.

Das Modell sollte Tab. 2.3 reproduzieren. Die Häufigkeit von (RD) unter $(X) = 0$ ist 0,8 %. Im Modell wäre das

$$e^{\beta_0}/(1+e^{\beta_0}) = e^{-2,44}/(1+e^{-2,44}) = 8\%$$

womit der Eintrag in 2.3 übereinstimmt.

	Model				
Variable	1	2	3	4	5
Intercept	-2.44	-2.51	-2.51	-2.85	-2.81
Current Xerophthalmia (1=yes;0=no)	0.44	0.40	0.42	0.79	0.78
Age-36 (months)				-0.024	-0.023
Season (1=2nd qtr; 0=other)				1.23	1.11
Y_{ij-1}		0.61	0.62	0.82	0.62
Y_{ij-1} by Xerophthalmia		0.11		-0.11	
Y_{ij-1} by Age				0.00063	
Y_{ij-1} by Season				-1.24	

Tabelle 2.5: Modelle und ihre Parameter. Aus [7]

Eine weitere Größe ist die *Logoddsratio*[5]. In der Studie sei

$$p_1 = P(Y = 1|X = 1),\ p_2 = (Y = 1|X = 0).$$

Dann ist

$$R = \operatorname{logit} \pi_{it}(x = 1) - \operatorname{logit} \pi_{it}(X = 0) = \beta_1$$

Die Maximumlikelihoodschätzung ergibt $\beta_1 = 0,44$ (vgl. Tab. 2.5, Modell 1, 2. Zeile). In weiteren Modellen werden Kovariablen wie Alter, $Y_{t-1} - (X)$-Interaktion usw. hinzugefügt. Weiteres findet man in [7], S. 197ff.

[5]Seien Y und Y_2 zwei binomialverteilte Zufallsvariablen mit Erfolgswahrscheinlichkeiten P_1 und p_2. Dann ist die *Logoddsratio* definiert als

$$R = \ln \left(\frac{\frac{p_1}{(1-p_1)}}{\frac{p_2}{(1-p_2)}} \right) = \ln \frac{p_1}{(1 - p_2)} - \ln \frac{p_2}{(1 - p_2)}. = \operatorname{logit} p_1 - \operatorname{logit} p_2.$$

Kapitel 3

Dynamische Modelle und der Kalman-Filter

Konditionale und dynamische Modelle sind entgegengesetzte Ansätze um Abhängigkeiten der Beobachtungen einer Zeitreihe $y_1, \ldots, y_T$ oder von Longitudinaldaten $y_{i1}, \ldots, y_{iT}$ zu modellieren. Bei konditionalen Modellen modelliert man den Erwartungswert von Y_t gegeben die Vergangenheit direkt als Funktion von $Y_{t-1}, \ldots, Y_1$. Ohne Kovariablen läßt sich ein konditionales Modell 1. Ordnung graphisch wie folgt darstellen:

$$Y_{t-2} \longrightarrow Y_{t-1} \longrightarrow Y_t \longrightarrow Y_{t+1}.$$

Im Gegensatz dazu wird bei dynamischen Modellen angenommen, daß die Y_t *bedingt unabhängig* sind, gegeben latente, unbeobachtete Zustände x_t. Diese entwickeln sich nach eigener Dynamik:

$$\begin{array}{ccccccc} X_{t-2} & \longrightarrow & X_{t-1} & \longrightarrow & X_t & \longrightarrow & X_{t+1} \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ Y_{t-2} & \longrightarrow & Y_{t-1} & \longrightarrow & Y_t & \longrightarrow & Y_{t+1} \end{array}$$

Hier liegt also ein einfaches *hierarchisches Modell* vor, mit der Hierarchieebene I der beobachteten Y_t und der Ebene II der latenten X_t. Diese werden als die wirklichen Zustände eines Systems aufgefaßt, deren Beobachtungen Y_t verfälscht sind. Deshalb heißen dynamische Modelle auch *Zustandsraummodelle* (*State-Space Models*), oder auch *Hidden-Markov-Modelle*.

Generell haben dynamische Modelle folgende Eigenschaften:

1. Sie sind schwieriger zu schätzen als konditionale Modelle, deren Likelihood sich faktorisieren läßt.

2. Sie sind als Modellklasse flexibler.
    - Sie ermöglichen die Modellierung langer zeitlicher Abhängigkeiten.
    - Sie können oft durch substanzwissenschaftliche Überlegungen motiviert werden.

Überdies gibt es eine Verbindung zu nonparametrischen Glättungsverfahren; sie sind 'nonparametrisch', weil die Anzahl der unbekannten Modellparameter $x_1, \ldots, x_T$ mit T wächst. Bei konditionalen Modellen ist sie hingegen fest.

Historisch hat sich folgende Einteilung herausgebildet:

x_t	y_t	Bezeichnung
normal	normal	Zustandsraummodell
normal	diskret	dynamisches Modell
diskret	stetig/diskret	Hidden-Markov-Modell

3.1 Zustandsraummodelle

Wir betrachten zunächst den 'klassischen' Fall. Gegeben sei eine Zeitreihe $y_1, \ldots, y_T$ (von Vektoren). Die *Zustände* oder unbeobachteten Parameter $x_1, \ldots, x_T$ seien ebenfalls Vektoren. Letztere entwickeln sich nach einer Dynamik, die durch die *Systemgleichung*

$$x_t = A_{t-1}x_{t-1} + w_t,\ w_t \sim \mathcal{N}(0, W_t), x_0 \sim \mathcal{N}(\mu_0, S)$$

gegeben ist, wobei A_t eine quadratische Matrix ist. Es liegt also eine durch Rauschen gestörte lineare Dynamik vor. Die Beobachtungen folgen linearen Modellen

$$y_t = C_t x_t + v_t,\ v_t \sim \mathcal{N}(0, V_t).$$

Alle Größen A_t, W_t, C_t, V_t und S werden (zunächst) als bekannt vorausgesetzt; v_t und w_t, $t = 1, \ldots, T$ sind alle voneinander unabhängig.

Bemerkung 3.1.1 Oft sind A_t usw. nicht zeitabhängig, ebenso wie die Kovarianzmatrizen (Dispersion) V_t und W_t. In Anwendungen sind μ_0, S und W_t häufig unbekannt und müssen mitgeschätzt werden. Dies ist nicht einfach.

3.1.1 Einfache Trendmodelle

Wir geben einige einfache Beispiele an.

Beispiel 3.1.2 Die einfachste Form, einen zeitlichen Trend zu modellieren ist das *Steady State Modell*

$$\begin{aligned} x_t &= x_{t-1} + w_t \\ y_t &= x_t + v_t, \ \ t = 1, \ldots, T. \end{aligned}$$

Die Systemgleichung beschreibt also eine Irrfahrt erster Ordnung, d.h. die ersten Differenzen sind unabhängig.

Beispiel 3.1.3 Eine Verallgemeinerung ist das *lokal-lineare Trendmodell*

$$\begin{aligned} x_t &= \begin{pmatrix} \tau_t \\ \lambda_t \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}}_{A} \begin{pmatrix} \tau_{t-1} \\ \lambda_{t-1} \end{pmatrix} + v_t, \, V = \begin{pmatrix} \sigma_\tau^2 & 0 \\ 0 & \sigma_\lambda^2 \end{pmatrix}. \\ y_t &= (1,0) x_t + w_t. \end{aligned}$$

Wir schreiben das aus:

$$\begin{aligned} \tau_t &= \tau_{t-1} + \lambda_{t-1} + v_{t1}, \\ \lambda_t &= \lambda_{t-1} + v_{t2}. \end{aligned}$$

Die zweite Komponente repräsentiert einen latenten Steigungsparameter, der einer Irrfahrt folgt. Für die Beobachtung gilt

$$y_t = \tau_t + w_t$$

Mit $(\lambda_t) \equiv 0$ ergibt sich das Steady-State Modell.

Beispiel 3.1.4 Ferner lassen sich Markovketten höherer Ordnung unter das Modell subsummieren. Wir betrachten als Beispiel die Irrfahrt 2. Ordnung

$$\tau_t = 2\tau_{t-1} - \tau_{t-2} + \tilde{v}_t.$$

Wir können sie schreiben als

$$\begin{aligned} \begin{pmatrix} \tau_t \\ \tau_{t-1} \end{pmatrix} &= \begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \tau_{t-1} \\ \tau_{t-2} \end{pmatrix} + \begin{pmatrix} \tilde{v}_t \\ 0 \end{pmatrix} \\ y_t &= (1,0) \begin{pmatrix} \tau_t \\ \tau_{t-1} \end{pmatrix}. \end{aligned}$$

Man nennt das auch 'Markovisierung'.

Bemerkung 3.1.5 Die Modellierung kommt aus den Ingenieurwissenschaften. Dort betrachtet man die allgemeineren Gleichungen

$$\begin{aligned} \alpha_{t+1} &= A_t\alpha_t + B_tu_t + w_{t+1}, \\ \gamma_t &= C_t\alpha_t + D_tu_t + v_t; \end{aligned}$$

wobei γ beobachtet wird und α der innere Zustand des Systems ist. Dieser entwickelt sich gemäß einer linearen Dynamik, wobei u_t ein bekannter deterministischer Input ist (bzw. eine Kontrollgröße). Solche Systeme können in ein rein deterministisches und ein rein stochastisches lineares System zerlegt werden:

$$\begin{array}{rclrcl} z_{t+1} &=& A_tz_t + B_tu_t, & x_{t+1} &=& A_tx_t + w_{t+1} \\ s_t &=& C_tz_t + D_tu_t, & y_t &=& C_tx_t + v_t \end{array} .$$

mit

$$\gamma_t = s_t + y_t,\ \alpha_t = z_t + x_t.$$

Die Lösung für z_t ist wohlbekannt in der Theorie linearer Systeme und optimaler Kontrolle. Somit kann man sich auf die optimale Schätzung von x_t im stochastischen System beschränken.

3.1.2 Statistische Inferenz im Zustandsraummodell

Im Modell haben wir es mit der Verteilung der x_t zu tun, implizit gegeben durch die Systemgleichung und von $y_t|x_t$, in der Beobachtungsgleichung. Es liegt also ein Bayessches Schätzproblem für $x_t|y_1, \ldots, y_t$ vor. Die optimale Lösung ist

$$x_{t|t} = \boldsymbol{E}(x_t|y_1, \ldots, y_t),$$

möglicherweise noch mit Kovariablen. Da alle Größen gaußisch sind, ist auch die a posteriori Verteilung gaußisch und es ist der negative Exponent zu minimieren. Die zugehörige Präzisionsmatrix ist nicht von vornherein bekannt und muß zusätzlich berechnet werden. Dies ist schon wegen der Dimension der Probleme schwierig. Zudem möchte man Echtzeitschätzungen haben, muß also einen sehr schnellen Algorithmus haben. Die Lösung wird durch den Kalmanfilter gegeben, den wir im Anschluß herleiten werden.

Beispiel 3.1.6 Für das einfache Steady-State Modell können wir die Größen noch berechnen. Es war

$$x_t = x_{t-1} + w_t (\in \mathbb{R})$$

$$y_t = x_t + v_t (\in \mathbb{R})$$

Für die a posteriori Verteilung bei festem y gilt:

$$\begin{aligned}
f(x|y) &\propto f(x) \cdot f(y|x) = \prod_{t=2}^{T} f(x_t|x_{t-1}) \cdot \prod_{t=1}^{T} f(y_t|x_t) \\
&= \exp\left(-\frac{1}{2} \cdot \frac{1}{W} \sum_{t=2}^{T} (x_t - x_{t-1})^2\right) \cdot \exp\left(-\frac{1}{2} \cdot \frac{1}{V} \sum_{t=1}^{T} (y_t - x_t)^2\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{1}{W} x' K x + \frac{1}{V}(y-x)'(y-x)\right)\right)
\end{aligned}$$

mit (Probe!)

$$K = \begin{pmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & 0 \\ & -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & & \\ 0 & & -1 & 2 & -1 & \\ & & & -1 & 1 & \end{pmatrix}$$

Wir wissen, daß dies eine gaußische Dichte ist, d.h. der Exponent läßt sich als vollständiges Quadrat

$$(x - c(y))' D^{-1} (x - c(y))$$

schreiben (wie das genau geht, behandeln wir später).

Jedenfalls liegt ein Optimierungsproblem in T Dimensionen inklusive Inversion von $T \times T$-Matrizen vor. Inversion ist instabil und der Aufwand steigt mit T^3. Der nun zu entwickelnde Kalman-Filter arbeitet mit $o(T)$. Nur die letzte Beobachtung wird in diesem rekursiven Verfahren jeweils verarbeitet.

3.2 Optimalitätskriterien und Bayessche Inferenz

Bei dynamischen Modellen haben wir ein zufälliges Design

$$\begin{aligned}
x_{t+1} &= A_t x_t + w_{t+1} \\
y_t &= C_t x_t + v_t,
\end{aligned}$$

d.h. in das lineare Regressionsmodell gehen die zufälligen Größen x_t ein, deren Verteilung durch die Systemgleichung bestimmt ist. Wir machen *keine* Voraussetzung über die Verteilung. Geschätzt werden sollen die x_t aus Beobachtungen der y_t. Um Schätzer zu bestimmen, gibt es verschiedene Kriterien, darunter die Methode der kleinsten Quadrate und diverse Bayesschätzer. Wir wollen deren Zusammenhang durchleuchten und so besser verstehen, was hinter den später eingeführten Algorithmen steckt. Die Argumentation folgt [4].

3.2.1 Kleinste-Quadrate Schätzer

Bezeichne $x(y)$ eine Funktion von y und $x^*(y)$ stehe für die Schätzer. Ein kleinstes Quadrate-Kriterium ist:

$$x^*(y) \text{ minimiere } \boldsymbol{E}_{x,y}(||x - x(y)||^2) = \boldsymbol{E}_{x,y}((x - x(y))'(x - x(y)),$$

wobei der Erwartungswert bezüglich der gemeinsamen Verteilung von x und y gebildet wird. Eine Minimalstelle $x^*(\cdot)$ heißt *kleinster Quadrate-Schätzer (MMS-Schätzer)*. Natürlich wird dies impliziert, falls

$$\boldsymbol{E}_x(||x - x^*(y)||^2|y) = \min_z \boldsymbol{E}(||x - z||^2|y) \text{ für jedes } y.$$

Proposition 3.2.1 *Der MMS-Schätzer ist der bedingte Erwartungswert*

$$x^*(y) = \boldsymbol{E}(x|y).$$

Beweis. Sei z gegeben. Dann ist

$$\boldsymbol{E}_x(||x - z||^2|y) = \boldsymbol{E}(||x||^2|y) - 2z'\boldsymbol{E}(x|y) + ||z||^2, \tag{3.1}$$

Die Funktion

$$z \longrightarrow ||z||^2 - 2z'a =: f(z)$$

ist quadratisch mit Minimum in

$$\frac{\partial f(z)}{\partial z_i} = 2z_i - 2a_i = 0$$

d.h. die Lösung von (3.1) ist $z^* = \boldsymbol{E}(x|y)$. □

3.2.2 Lineare Kleinste-Quadrate Schätzer

Die Funktion $y \longrightarrow \boldsymbol{E}(x|y)$ kann sehr kompliziert sein. Wir schränken deshalb die Klasse der Schätzer auf die einfachsten ein, nämlich *lineare Schätzer* der Form

$$x(y) = Ay + b,$$

d.h. wir suchen $\hat{A}$, $\hat{b}$, welche

$$h : (A, b) \longrightarrow \boldsymbol{E}_{x,y}(||x - (Ay + b)||^2)$$

minimieren (über alle dimensionsverträglichen Matrizen A und Vektoren b).

Beispiel 3.2.2 (Der gaußsche Fall) Eine durchsichtige Diskussion ist im gaußischen Fall möglich. Sei

$$z = \begin{pmatrix} x \\ y \end{pmatrix}$$

ein Gaußischer Vektor mit

$$\boldsymbol{E}(z) = \begin{pmatrix} \boldsymbol{E}(x) \\ \boldsymbol{E}(y) \end{pmatrix} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} = \mu$$

und Kovarianzmatrix

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}. \tag{3.2}$$

Für die Dichte gilt

$$f(z) \propto \exp\left(-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)\right).$$

Die Verteilung von y hat Dichte

$$f(y) \propto \exp\left(-\frac{1}{2}(y-\mu_y)'\Sigma_{yy}^{-1}(y-\mu_y)\right),$$

so daß gilt

$$f(x|y) \propto \exp\left(-\frac{1}{2}\Big\{(z-\mu)'\Sigma^{-1}(z-\mu) - (y-\mu_y)'\Sigma_{yy}^{-1}(y-\mu_y)\Big\}\right). \tag{3.3}$$

Wir wissen, daß dies eine gaußische Dichte ist. Diese können wir mit einer quadratischen Form im Exponenten schreiben mit einem linearen Schätzer als Erwartungswert und D als Kovarianzmatrix der (gaußischen) a posteriori Verteilung, d.h.

$$\cdots = (x - (Ay + b))' D^{-1} (x - (Ay + b)), \tag{3.4}$$

Wir brauchen ein Lemma aus der linearen Algebra.

Lemma 3.2.3 *Sei*

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

Dann gilt

$$M^{-1} = \begin{pmatrix} Q & -QBD^{-1} \\ -D^{-1}CQ & D^{-1} + D^{-1}CQBD^{-1} \end{pmatrix}. \tag{3.5}$$

mit $\quad Q = (A - BD^{-1}C)^{-1}$, *falls D und $A - BD^{-1}C$ invertierbar sind.*

Beweis. Verifiziere $MM^{-1} = I$ durch Nachrechnen. □

Damit kann man die Größen in (3.4) bestimmen.

Proposition 3.2.4 . *Für die a posteriori Verteilung von $x|y$ gilt (3.4) mit*

$$A = \Sigma_{xy}\Sigma_{yy}^{-1},\ b = \mu_x - \Sigma_{xy}\Sigma_{yy}^{-1}\mu_y,\ D = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}. \tag{3.6}$$

Beweis. Man setzt die Größen in (3.4) ein und formt (3.3) und (3.4) um, wobei man für Σ^{-1} die Formel (3.5) benutzt. □

Damit haben wir gezeigt:

Theorem 3.2.5 *Im gaußschen Fall gilt für den a posteriori Erwartungwert*

$$\boldsymbol{E}(x|y) = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y). \tag{3.7}$$

Insbesondere ist dies der a posteriori Erwartungswert, Modus, Median und Maximumlikelihoodschätzer.

Beweis. Dies ist $Ay + b$ mit A, b aus (3.6). □

Wir folgern

Korollar 3.2.6 *Sind x und y gemeinsam gaußisch, so ist der MMS-Schätzer $\boldsymbol{E}(x|y)$ sogar ein linearer MMS-Schätzer.*

Beweis. Wir haben den MMS-Schätzer als $\boldsymbol{E}(x|y)$ entlarvt und dies im gaußischen Fall explizit berechnet. □

Wir charakterisieren nun im *allgemeinen* Fall den *linearen* MMS-Schätzer. Wir benutzen die eingeführten Bezeichnungen sinngemäß.

Proposition 3.2.7 *Seien x und y nicht notwendig gaußische Zufallsvariablen. Die nötigen ersten und zweiten Momente (der gemeinsamen Verteilung) mögen existieren; die Kovarianzmatrix $\sum_{yy}$ sei invertierbar. Dann ist der lineare MMS-Schätzer durch*

$$\hat{x}(y) = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y) \tag{3.8}$$

gegeben. Die Fehlerkovarianzmatrix ist

$$\boldsymbol{E}_{x,y}\Big((x - \hat{x}(y))'(x - \hat{x}(y))\Big) = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}.$$

Bevor wir dies beweisen, stellen wir fest: *Im gaußischen Fall ist der MMS-Schätzer der a posteriori Schätzer für alle wichtigen Skalenparameter von x und die Fehlerkovarianzmatrix ist gerade die Kovarianzmatrix der a posteriori Verteilung von $x|y$.*

Beweis. Der lineare MMS-Schätzer ist gegeben durch

$$\hat{x}(y) = \hat{A}y + \hat{b},$$

wobei $\hat{A}$ und $\hat{b}$ die Funktion

$$h : (A, b) \longrightarrow \boldsymbol{E}_{x,y}(||x - Ay - b||^2)$$

über A und b minimieren. Wie üblich setzen wir die Ableitung nach A und b gleich 0, d.h.

$$\begin{aligned} 0 &= \frac{\partial h}{\partial A}(\hat{A}, \hat{b}) = 2\boldsymbol{E}_{x,y}(y(\hat{b} + \hat{A}y - x)') \\ 0 &= \frac{\partial h}{\partial b}(\hat{A}, \hat{b}) = 2\boldsymbol{E}_{x,y}(\hat{b} + \hat{A}y - x). \end{aligned}$$

Die zweite Gleichung liefert zusammen mit Einsetzen in die erste

$$\begin{aligned} &\hat{b} = \mu_x - \hat{A}\mu_y \\ &\boldsymbol{E}_{xy}(y(\hat{A}(y - \mu_y) - (x - \mu_x))') = 0. \end{aligned}$$

Trivialerweise gilt:

$$\boldsymbol{E}_{xy}(-\mu_y(\hat{A}(y-\mu_y)-(x-\mu_x))') = -\mu_y\boldsymbol{E}_{xy}((\hat{A}(y-\mu_y)-(x-\mu_x))') = 0.$$

Wir addieren die letzten beiden Gleichungen und erhalten

$$\boldsymbol{E}_{xy}((y-\mu_y)(\hat{A}(y-\mu_y)-(x-\mu_x))') = 0,$$

was sich schreiben läßt als

$$\Sigma_{yy}\hat{A}' - \Sigma_{yx} = 0,$$

woraus folgt

$$\hat{A} = \Sigma'_{yx}\Sigma_{yy}^{-1} = \Sigma_{xy}\Sigma_{yy}^{-1}.$$

Damit erhalten wir die behauptete Form des Schätzers, nämlich

$$\hat{x}(y) = \hat{A}y + \hat{b} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y).$$

Die Fehlerkovarianzmatrix bekommt man durch Einsetzen und Ausrechnen.
□

Der MMS-Schätzer ist erwartungstreu:

Korollar 3.2.8 *Es gilt*

$$\mu_x = \boldsymbol{E}_x(x) = \boldsymbol{E}_y(\hat{x}(y)).$$

Beweis. Man bilde in (3.8) den Erwartungswert bezüglich y. □

Korollar 3.2.9 *Der Schätzfehler $x - \hat{x}(y)$ ist unkorreliert mit y und $\hat{x}(y)$:*

$$\boldsymbol{E}_{xy}(y(x-\hat{x}(y))') = 0,\ \boldsymbol{E}_{xy}(\hat{x}(y)(x-\hat{x}(y))') = 0.$$

Beweis. Man setze $\hat{x} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y)$ ein, bilde die Erwartungswerte und beachte Korollar 3.2.8. □

Dies ist das *orthogonale Projektionsprinzip*.

Bemerkung 3.2.10 Unkorreliertheit von y und $x - \hat{x}(y)$ bedeutet

$$\boldsymbol{E}_{xy}((y - \mu_y)(x - \hat{x}(y) - \boldsymbol{E}_{xy}(x - \hat{x}(y)))') = 0.$$

Nun verschwindet der innere Erwartungswert wegen der Erwartungstreue und aus demselben Grund ist

$$\boldsymbol{E}_{xy}(\mu_y(x - \hat{x}(y))') = \mu_y \boldsymbol{E}((x - \hat{x}(y))') = 0.$$

Deshalb ist Unkorreliertheit äquivalent zur Gleichung im Lemma. Für die zweite Gleichung rechnet man genau so.

Bemerkung 3.2.11 Die erste Beziehung bedeutet[1]

$$y \perp x - \hat{x}(y) \in \mathcal{L}^2.$$

Dies bedeutet nach Definition, daß $\hat{x}(\cdot)$ die orthogonale Projektion von x auf den von den Zufallsvariablen $1, y_1, \ldots, y_s$ aufgespannten linearen Teilraum $\mathcal{H}$ von $\mathcal{L}^2$ ist. Nach dem Orthogonalitätsprinzip ist das äquivalent mit

$$||x - \hat{x}(\cdot)||^2 = \min_{z \in \mathcal{H}} ||x - z||^2.$$

Nun läßt sich jedes $z \in \mathcal{H}$ als Linearkombination von $1, y_1, \ldots, y_s$ schreiben. Wir wenden das einzeln auf jede Komponente von $\hat{x}(\cdot)$ an und erhalten die Form

$$\hat{x}(y) = \hat{A}(y) + b.$$

Der beste *lineare* MMS-Schätzer von x ist also die orthogonale Projektion von x auf den von $1, y_1, \ldots, y_s$ erzeugten Teilraum von $\mathcal{L}^2$.

Wir wollen nun nicht x selbst, sondern eine lineare Funktion von x schätzen:

Korollar 3.2.12 *Sei C eine geeignet dimensionierte Matrix und $z = Cx$. Der lineare MMS-Schätzer $\hat{z}(y)$ für z ist $\hat{z} = C\hat{x}(y)$ mit Fehlerkovarianz*

$$\boldsymbol{E}_{zy}((z - \hat{z}(y))(z - \hat{z}(y))') = C\boldsymbol{E}_{xy}((x - \hat{x}(y))(x - \hat{x}(y))')C'.$$

---

[1] $\mathcal{L}^2$ ist der lineare Raum der quadratintegrierbaren Funktionen, d.h. derjenigen meßbaren Funktionen f auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \boldsymbol{P})$ mit $\int f^2(\omega)\, d\boldsymbol{P}(\omega) < \infty$.

Beweis. Dies ist eine einfache Rechnung. Natürlich gelten:

$$\begin{aligned} \boldsymbol{E}_z(z) &= C\mu_x \\ \Sigma_{zz} &= C\Sigma_{xx}C', \\ \Sigma_{zy} &= \boldsymbol{E}_{xy}((z-\mu_z)(y-\mu_y)') = C\Sigma_{xy} \\ \Sigma_{yz} &= \Sigma'_{zy} = \Sigma_{yx}C'. \end{aligned}$$

Nach Proposition 3.2.7 ist

$$\begin{aligned} \hat{z}(y) &= \mu_z + \Sigma_{zy}\Sigma_{yy}^{-1}(y-\mu_y) \\ &= C\mu_x + C\Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y) = C\hat{x}(y). \end{aligned}$$

Genauso gilt

$$\begin{aligned} \boldsymbol{E}_{xy}((z-\hat{z}(y))(z-\hat{z}(y))') &= \Sigma_{zz} - \Sigma_{zy}\Sigma_{yy}^{-1}\Sigma_{yz} \\ &= C(\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx})C' \\ &= C\boldsymbol{E}((x-\hat{x}(y))(x-\hat{x}(y))')C'. \end{aligned}$$

□

Korollar 3.2.13 *Sei die Zufallsvariable z unkorreliert zu y. Dann ist die lineare MMS-Schätzung $\hat{x}(y,z)$ (gegeben (y,z)) von der Gestalt*

$$\hat{x}(y,z) = \hat{x}(y) + \hat{x}(z) - \mu_x,$$

wobei $\hat{x}(y)$ und $\hat{x}(z)$ die linearen MMS-Schätzer von x gegeben y bzw. z sind. Die Fehlerkovarianzmatix ist

$$\boldsymbol{V}(x-\hat{x}(y,z)) = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}$$

Beweis. Seien

$$w = \begin{pmatrix} y \\ z \end{pmatrix}, \quad \mu_w = \begin{pmatrix} \mu_y \\ \mu_z \end{pmatrix}.$$

Nach Proposition 3.2.7 gilt

$$\hat{x}(w) = \mu_x + \Sigma_{xw}\Sigma_{ww}^{-1}(w-\mu_w).$$

Nun ist

$$\Sigma_{xw} = \left(\Sigma_{xy}, \Sigma_{xz}\right)$$

und weil y und z unkorreliert sind, ergibt sich

$$\Sigma_{ww} = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zy} & \Sigma_{zz} \end{pmatrix} = \begin{pmatrix} \Sigma_{yy} & 0 \\ 0 & \Sigma_{zz} \end{pmatrix}.$$

Wir setzen ein und erhalten

$$\begin{aligned} \hat{x}(y,z) &= \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y) + \Sigma_{xz}\Sigma_{zz}^{-1}(z-\mu_z) \\ &= \hat{x}(y) + \hat{x}(z) - \mu_x. \end{aligned}$$

Analog berechnet man die Fehlerkovarianzmatrix. □

Korollar 3.2.14 *Es seien nun im vorigen Korollar 3.2.13 y und z* nicht *notwendig unkorreliert. Dann gelten die Formeln*

$$\hat{x}(y,z) = \hat{x}(y) + \hat{x}(z - \hat{z}(y)) - \mu_x$$

und

$$\begin{aligned} &\boldsymbol{E}_{xyz}((x-\hat{x}(y,z))(x-\hat{x}(y,z))') \\ &= \boldsymbol{E}_{xy}((x-\hat{x}(y))(x-\hat{x}(y))') - \boldsymbol{E}_{xyz}((x-\mu_x)(z-\hat{z}(y))') \\ &\cdot\ (\boldsymbol{E}_{yz}((z-\hat{z}(y))(z-\hat{z}(y))'))^{-1}\boldsymbol{E}_{xyz}((z-\hat{z}(y))(x-\mu_x)'). \end{aligned}$$

Beweis. Nach Korollar 3.2.9 sind y und $(z - \hat{z}(y)$ unkorreliert. Aus Lemma 3.2.13 bekommt man deshalb $\hat{x}(y, z - \hat{z}(y))$ zusammen mit der Fehlerkovarianz. Schließlich überzeugt man sich von

$$\hat{x}(y,z) = \hat{x}(y, z - \hat{z}(y)).$$

Dies sieht man am einfachsten so: Es ist $\hat{x}(y,z)$ die orthogonale Projektion von x auf $\mathcal{H} = \mathrm{span}\{1, y, z\}$. Ferner ist $\hat{z}(y)$ die orthogonale Projektion von z auf $\mathrm{span}\{1, y\}$, also $z = \hat{z}(y) + (z - \hat{z}(y))$. Offensichtlich erzeugen aber 1, y und $z - \hat{z}(y)$ denselben Raum wie 1, y und z. □

3.3 Kalmanfilter und -glätter

Wir wenden uns wieder dem ursprünglichen Ziel zu, die aktuellen Systemzustände aus den vergangenen Systemzuständen und Beobachtungen, sowie der aktuellen Beobachtung optimal zu schätzen. Wieder lehnen wir uns an D.P. BERTSEKAS (1976), [4] an. Eine ausführliche - wenn auch sehr technische - Diskussion enthält C.K. CHUI und G. CHEN (1987), [6].

3.3.1 Der Kalmanfilter

Es seien

$$x_{t+1} = A_t x_t + w_t, \quad t = 0, 1, \ldots, T-1,$$

die *Systemgleichung*, $x_t \in \mathbb{R}^n$ der nicht beobachtbare Systemzustand, $w_t \in \mathbb{R}^n$ ein Systemrauschen, A_t deterministisch und bekannt, und

$$z_t = C_t x_t + v_t, \quad t = 0, 1, \ldots, T,$$

die *Beobachtungsgleichung* mit den Beobachtungen $z_t \in \mathbb{R}^s$ und der Störung $v_t \in \mathbb{R}^s$.

Annahmen 3.3.1 Die Zufallsgrößen

$$x_0, w_0, w_1, \ldots, w_{T-1}, v_0, \ldots, v_T$$

seien unabhängig mit vorgegebenen Verteilungen und *zentriert*. Wir verwenden die Bezeichnungen

$$S = \boldsymbol{E}(x_0 x_0'), \;\; M_t = \boldsymbol{E}(w_t w_t'), \;\; N_t = \boldsymbol{E}(v_t v_t')$$

und setzen voraus, daß alle N_t positiv definit sind. Ferner schreiben wir

$$Z_t = (z_0, \ldots, z_t).$$

Für die Schätzer und deren Fehlerkovarianzmatrizen schreiben wir

$$\begin{aligned} \hat{x}_{t|t-1} &= \hat{x}_t(Z_{t-1}), \;\; \Sigma_{t|t-1} = \boldsymbol{V}(x_t - \hat{x}_t(Z_{t-1})) && (3.9) \\ \hat{x}_{t|t} &= \hat{x}_t(Z_t), \;\; \Sigma_{t|t} = \boldsymbol{V}(x_t - \hat{x}_t(Z_t)). && (3.10) \end{aligned}$$

Die Schätzer sollen rekursiv berechnet werden; deshalb nehmen wir an, daß die Größen (3.9) schon berechnet sind. Wir wollen die Größen in (3.10) berechnen, wobei nur die Informationen (3.9) *und* z_t benutzt werden sollen; für den nächsten Schritt müssen wir dann noch die Größen (3.9) mit inkrementierter Zeit berechnen. Nach Korollar 3.2.14 gilt

$$\hat{x}_t(Z_t) = \hat{x}_t(Z_{t-1}) + \hat{x}_t(z_t - \hat{z}_t(Z_{t-1})) - \boldsymbol{E}(x_t).$$

Wegen der Anfangsbedingung $\boldsymbol{E}(x_0) = 0$ und der Systemgleichung ist $\boldsymbol{E}(x_t) = 0$. Als nächstes berechnen wir

$$\hat{x}_t(z_t - \hat{z}_t(Z_{t-1})) = \mu_{x_t} + \Sigma_{x_t, z_t - \hat{z}_t(Z_{t-1})} \Sigma^{-1}_{z_t - \hat{z}_t(Z_{t-1}), z_t - \hat{z}_t(Z_{t-1})} (z_t - \hat{z}_t(Z_{t-1})). \tag{3.11}$$

Wegen der Linearität der Schätzer ist

$$\hat{z}_t(Z_{t-1}) = (\widehat{C_t x_t})(Z_{t-1}) + \hat{v}_t(Z_{t-1}) = C_t \hat{x}_t(Z_{t-1}),$$

denn $\hat{v}_t(Z_{t-1}) = 0$, weil es die orthogonale Projektion von v_t auf $\operatorname{span}\{Z_{t-1}\}$ ist, worauf v_t wegen der Unkorreliertheit senkrecht steht. Der Rest folgt aus Korollar 3.2.12. Ebenfalls aus der Unkorreliertheit folgt

$$\begin{aligned} \boldsymbol{V}(z_t - \hat{z}_t(Z_{t-1})) &= \boldsymbol{V}(C_t(x_t - \hat{x}_t(Z_{t-1})) + v_t) \\ = \quad & C_t \Sigma_{t|t-1} C_t' + N_t. \end{aligned}$$

Ähnlich berechnen wir

$$\begin{aligned} & \boldsymbol{V}(x_t, z_t - \hat{z}_t(Z_{t-1})) = \boldsymbol{E}(x_t \cdot (z_t - \hat{z}_t(Z_{t-1}))') \\ = \quad & \boldsymbol{E}(x_t(x_t - \hat{x}_t(Z_{t-1}))' C_t') + \boldsymbol{E}(x_t v_t) \\ = \quad & \boldsymbol{E}((x_t - \hat{x}_t(Z_{t-1}))(x_t - \hat{x}_t(Z_{t-1}))') C_t' \\ + \quad & \boldsymbol{E}(\hat{x}_t(Z_{t-1})(x_t - \hat{x}_t(Z_{t-1}))' C_t'). \end{aligned}$$

Weil der Schätzer unkorreliert zum Schätzfehler ist, verschwindet der letzte Term. Damit haben wir gezeigt:

$$\boldsymbol{V}(x_t, z_t - \hat{z}_t(Z_{t-1})) = \Sigma_{t|t-1} C_t'.$$

Wir setzen die Ergebnisse in (3.11) ein und erhalten

$$\hat{x}_t(z_t - \hat{z}_t(Z_{t-1})) = \Sigma_{t|t-1} C_t' \left(C_t \Sigma_{t|t-1} C_t' + N_t \right)^{-1} (z_t - C_t \hat{x}_t(Z_{t-1})).$$

Zusammenfassend ergibt sich

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + \Sigma_{t|t-1} C_t' \left(C_t \Sigma_{t|t-1} C_t' + N_t \right)^{-1} (z_t - C\hat{x}_t(Z_{t-1}));$$

weiter folgt aus der Systemgleichung

$$\hat{x}_{t+1|t} = A_t \hat{x}_{t|t}.$$

Ähnlich berechnet man

$$\Sigma_{t+1|t} = A_t \Sigma_{t|t} A_t' + M_t.$$

Ähnlich wie $x_{t|t}$ berechnet man

$$\Sigma_{t|t} = \Sigma_{t|t-1} - \Sigma_{t|t-1} C_t' \left(C_t \Sigma_{t|t-1} C_t' + N_t \right)^{-1} C_t \Sigma_{t|t-1}.$$

Zusammen mit dem Anfangsschritt

$$\hat{x}_0 = 0, \;\; \Sigma_{0|-1} = S$$

ergibt das den Kalmanfilter.

Wir können den Algorithmus so aufschreiben: Sei

$$K_t = \Sigma_{t|t-1} C_t' (C_t \Sigma_{t|t-1} C_t' + N_t)^{-1}$$

der *Kalman Gain*. Dann berechnen wir:

$$\begin{aligned}
x_{0|0} &= x_0, \Sigma_{0|0} = S && (3.12) \\
& \quad \text{Für } k = 1, \ldots, N: \\
x_{t|t-1} &= A_{t-1} x_{t-1|t-1} && (3.13) \\
\Sigma_{t|t-1} &= A_{t-1} \Sigma_{t-1|t-1} A_{t-1}' + M_{t-1} && (3.14) \\
x_{t|t} &= x_{t|t-1} + K_t (z_t - C_t x_{t|t-1}) && (3.15) \\
\Sigma_{t|t} &= \Sigma_{t|t-1} - K_t C_t \Sigma_{t|t-1}. && (3.16)
\end{aligned}$$

In der ersten Zeile wird der Algorithmus initialisiert. In der zweiten wird aus der vorherigen Schätzung eine Vorhersage des jetzigen Zustandes gemacht, indem die Systemgleichung angewendet wird. Letztlich wird die aktuelle Beobachtung z_t benutzt, um diese Vorhersage durch den Einschrittvorhersagefehler optimal gewichtet durch den Kalman Gain zu korrigieren. Nun hat man die neue Zustandsschätzung.

3.3.2 Prädiktion und Glättung

Allgemeiner interessiert man sich für die optimale Schätzung von x_t, gegeben gewisse der Beobachtungen $z_1, \ldots, z_T$. Dies nennt man

- Filtern für $t = T$
- Prädiktion für $t > T$
- Glättung für $t < T$

Das Filtern wurde besprochen, ebenso wie die Prädiktion (im Spezialfall $t = T + 1$). Bei der Glättung berechnet man $x_{t|T} = \hat{x}_t(Z_T)$ für $t \leq T$. Der Glättungsalgorithmus hat die Gestalt:

$$\begin{aligned} x_{t-1|T} &= x_{t-1|t-1} + B_t(x_{t|T} - x_{t|t-1}) \\ \Sigma_{t-1|T} &= \Sigma_{t-1|t-1} + B_t(\Sigma_{t|T} - \Sigma_{t|t-1})B_t' \\ B_t &= \Sigma_{t-1|t-1} A_{t-1}' \Sigma_{t|t-1}^{-1}. \end{aligned}$$

In jedem Schritt erhält man die Glättungsschätzung aus der Filterschätzung durch Addition der gewichteten Differenz zwischen der letzten Glättungsschätzung und der Prädiktion. Für Details verweisen wir auf die Literatur, z.B. [10].

3.4 Anhang

Als Referenz fassen wir die benötigten Fakten über orthogonale Projektionen und bedingte Erwartungen zusammen.

3.4.1 Das Orthogonalitätsprinzip

Gegeben sei ein Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \boldsymbol{P})$. Es sei

$$\mathcal{L}^2(\boldsymbol{P}) = \left\{ X : \int X^2 \, d\boldsymbol{P} < \infty \right\}$$

der Raum der Zufallsvariablen mit existierendem zweiten Moment, wobei X und Y in $\mathcal{L}^2(\boldsymbol{P})$ identifiziert werden, falls $\boldsymbol{P}\{X \neq Y\} = 0$. Mit den üblichen Operationen ist $\mathcal{L}^2(\boldsymbol{P})$ ein reeller linearer Raum. Durch

$$\langle X, Y \rangle = \int XY \, d\boldsymbol{P}, \quad \|X\|^2 = \langle X, X \rangle,$$

sind ein Skalarprodukt und eine Norm definiert. Damit ist $\mathcal{L}^2(\boldsymbol{P})$ vollständig.

Bemerkung 3.4.1 *(a) Jedes X mit zweitem Moment hat auch ein erstes Moment, denn für geeignetes c ist*

$$|X| \leq X^2 + c.$$

(b) Es gibt $\boldsymbol{V}(X, Y) = \langle X - \boldsymbol{E}(X), Y - \boldsymbol{E}(Y)\rangle$; für zentrierte Zufallsvariablen ist also das Skalarprodukt gerade die Kovarianz. Das macht die Bedeutung des $\mathcal{L}^2(\boldsymbol{P})$ in der Theorie insbesondere stationärer Prozesse und bei Zeitreihen aus.

Man sagt, X sei *orthogonal* zu Y, wenn

$$X \perp Y :\Longleftrightarrow \langle X, Y\rangle = 0$$

und Teilmengen $\mathcal{M}$ und $\mathcal{N}$ sind orthogonal wenn

$$X \perp Y \text{ für alle } X \in \mathcal{M}, Y \in \mathcal{N}.$$

Sei nun $\mathcal{H}$ ein abgeschlossener linearer Teilraum von $\mathcal{L}^2(\boldsymbol{P})$. Dann gibt es zu jedem $X \in \mathcal{L}^2(\boldsymbol{P})$ eine eindeutige Zerlegung

$$X = p + (X - p), p \in \mathcal{H}, X - p \perp \mathcal{H}$$

und $p = P_{\mathcal{H}}X$ heißt *orthogonale Projektion* von X auf $\mathcal{H}$. Die Abbildung $X \mapsto P_{\mathcal{H}}X$ ist eine stetige lineare Abbildung, der Projektionsoperator.

Bemerkung 3.4.2 *Es ist $p = P_{\mathcal{H}}X$ genau dann, wenn*

$$\langle X, Y\rangle = \langle p, Y\rangle \text{ für jedes } Y \in \mathcal{H}. \tag{3.17}$$

Für uns entscheidend ist

Theorem 3.4.3 (Das Orthogonalitätsprinzip) *Seien $\mathcal{H}$ ein abgeschlossener linearer Teilraum von $\mathcal{L}^2(\boldsymbol{P})$ und $X \in \mathcal{L}^2(\boldsymbol{P})$. Dann gilt*

$$p = P_{\mathcal{H}}X \Longleftrightarrow \|X - p\|^2 = \min_{Y \in \mathcal{H}} \|X - Y\|^2.$$

Beweis. Seien $p = P_{\mathcal{H}}X$ und $Y = p + Z \in \mathcal{H}$. Dann gilt

$$\|X-Y\|^2 = \|(X-p)-Z\|^2 = \|X-p\|^2 - 2\underbrace{\langle X - p, Z\rangle}_{=0} + \|Z\|^2 = \|X-p\|^2 + \|Z\|^2.$$

Dies ist minimal genau dann, wenn $Z = 0$, d.h. $Y = p$. □

3.4.2 Beste MMS-Schätzer und bedingte Erwartung

Sei $\mathcal{F}$ eine Unter-σ-Algebra von $\mathcal{A}$. Es sei ferner

$$\mathcal{H} = \{Y \in \mathcal{L}^2(\boldsymbol{P}) : Y \ \ \mathcal{F} - \text{meßbar}\}.$$

Sei ferner $X \in \mathcal{L}^2(\boldsymbol{P})$. Dann heißt

$$\boldsymbol{E}(X|\mathcal{F}) = P_{\mathcal{H}}X$$

die *bedingte Erwartung* von X geben $\mathcal{F}$. Wegen (3.17) sind äquivalent:

$$\begin{array}{ll} & \boldsymbol{E}(X|\mathcal{F}) \text{ bedingte Erwartung} \\ \Longleftrightarrow & \langle Y, X\rangle = \langle Y, \boldsymbol{E}(X|\mathcal{F})\rangle \text{ für alle } Y \in \mathcal{H} \\ \Longleftrightarrow & \displaystyle\int YX\, d\boldsymbol{P} = \int Y\boldsymbol{E}(X|\mathcal{F})\, d\boldsymbol{P} \text{ für alle } Y \in \mathcal{H} \\ \Longleftrightarrow & \displaystyle\int_F X\, d\boldsymbol{P} = \int_F \boldsymbol{E}(X|\mathcal{F})\, d\boldsymbol{P} \text{ für alle } F \in \mathcal{F}. \end{array}$$

Eine Richtung der letzten Äquivalenz ist trivial, die andere folgt aus Monotoniesätzen.

Bemerkung 3.4.4 In der Wahrscheinlichkeitstheorie definiert man die bedingte Erwartung üblicherweise auf dem Raum aller Zufallsvariablen mit existierendem *ersten* Moment durch die letzte Gleichung. Aus unserer Definition erhält man das ebenso, da man jede Zufallsvariable mit erstem Moment durch solche mit zweitem Moment approximieren kann.

Sei nun $\mathcal{Y}$ eine Familie von Zufallsvariablen mit zweitem Moment. Dann definiert man die bedingte Erwartung durch

$$\boldsymbol{E}(X|\mathcal{Y}) := \boldsymbol{E}(X|\mathcal{F}),$$

wobei $\mathcal{F}$ die von $\mathcal{Y}$ erzeugte σ-Algebra ist.

Sie hat die Darstellung

$$\mathcal{F} = \sigma\{\{Y \leq a\} : a \in \mathbb{R}, Y \in \mathcal{Y}\}.$$

$\mathcal{F}$ enthält also die Ereignisse, für die man aufgrund der Beobachtung von Y entscheiden kann, ob sie eingetreten sind oder nicht. Die bedingte Erwartung ist meßbar bezüglich dieses $\mathcal{F}$, beinhaltet also nicht mehr Information als die

Beobachtung. Es kann als eine Version von X gedeutet werden, die sich bei Mittelung von X (also Vergröberung) über Ereignisse aus $\mathcal{F}$ genauso verhält wie X selbst. Man sieht leicht

$$\mathcal{H} = \{\varphi(\mathcal{Y}_a) : \mathcal{Y}_a \subset \mathcal{Y} \text{ abzählbar}, \varphi : \mathbb{R}^{\mathsf{N}} \longrightarrow \mathbb{N} \text{ mit } \varphi(\mathcal{Y}_a) \in \mathcal{L}^2(\boldsymbol{P})\}.$$

Nach Definition ist $\boldsymbol{E}(X|\mathcal{Y}) \in \mathcal{H}$. Somit kann die bedingte Erwartung allerdings eine sehr komplizierte Funktion der Beobachtungen sein.

In Proposition 3.2.1 wurde gezeigt, daß die bedingten Verteilung der optimale MMS–Schätzer ist. Jetzt haben wir ihn in anschaulicher Weise als Projektion auf den Raum der Zufallsvariablen mit zweitem Moment entlarvt, die sich als Funktion der Beobachtungen schreiben lassen.

Der beste *lineare* MMS-Schätzer ist eine *lineare* Funktion von $Y_1, \ldots, Y_n$ und der konstanten Funktion $\mathbf{1}$, hat also die Gestalt[2]

$$\hat{X}(Y) = a'Y + b.$$

Also suchen wir $\hat{X}(Y) \in \text{span}\{\mathbf{1}, Y_1, \ldots Y_n\}$ mit minimalem Abstand zu X. *Dieses* ist aber nach dem Orthogonalitätsprinzip die orthogonale Projektion von X auf

$$\text{span}\{Y_1, \ldots Y_n\}.$$

[2]um Irritationen vorzubeugen, bemerken wir, daß wir uns gegenwärtig in der univariaten Situation befinden. Für die Vektoren im Text gilt das hier Gesagte komponentenweise.

Kapitel 4

Verweildauer- und Ereignisanalyse

Verweildauer- (Survival-) und Ereignisanalyse bilden einen Grundpfeiler der Statistik. Sie finden Anwendung in Versicherungsmathematik, Demographie, Epidemiologie, Medizin, Zuverlässigkeitsanalyse, Mikrosoziologie und vielen anderen Gebieten. In dieser Theorie werden individuelle Lebensabläufe als Pfade eines stochastischen Prozesses aufgefaßt, der sich zwischen Zuständen eines diskreten Raumes bewegt. Die Zustände dieses Prozesses repräsentieren verschiedene Zustände der Individuen, einer Unfallversicherung, einer technischen Komponente usw., während die Übergänge zwischen den Zuständen Ereignisse von Interesse repräsentieren, wie Tod des Versicherungsnehmers, Defekt einer technischen Komponente usw. Die Häufigkeit, mit der solche interessierenden Ereignisse auftreten, werden durch Raten oder Intensitäten repräsentiert. Somit enthalten statistische Modelle der Ereignisanalyse typischerweise Beschreibungen, wie diese Intensitäten von der Zeit oder von individuellen Eigenschaften und äußeren beobachteten Ereignissen abhängen.

In der einfachsten Situation werden nur zwei Zustände - tot und lebendig, funktioniert und funktioniert nicht, usw. - modelliert. Die Theorie dazu wird oft Sterbetafelanalyse (im Englischen life-table analysis) genannt. In diesem Fall ist die Intensität des Ereignisses 'Tod' einfach die Hazardrate der Überlebenszeitfunktion.

Die kumulierte Hazardrate zu einer gewissen Zeit ist das bis dahin aufgesammelt Risiko - eine theoretische Größe. Empirisch spiegelt sich dies in der Zahl der bisher beobachteten Ereignisse wider, die durch einen Zählprozeß beschrieben wird. Inferenz über die theoretische Größe 'kumuliertes

Risiko' fordert also ein Modell, welches die empirischen Zähldaten als Summe des kumulierten Risikos und eines Störterms darstellt. Solche Zerlegungen sind Gegenstand der Doob-Meyer Zerlegung aus der stochastischen Analysis. Vor allem deshalb baut die moderne Ereignisanalyse auf stochastischer Analysis auf und erreicht so eine sehr hohe Stufe der Theorie, im Gegensatz zur herkömmlichen Sterbetafelmethode, die in elementarer Weise mit empirischen Häufigkeiten arbeitet. Wichtig dabei ist, daß wichtige Prozesse dabei Zählprozesse sind und somit prinzipiell unstetig, während viele Bücher über stochastische Analysis stetige Prozesse - wie Diffusionsprozesse in den Mittelpunkt stellen. Dieser Aspekt wird in [16] besonders berücksichtigt.

Ein weiterer spezieller Aspekt dieses Gebietes ist, daß man selten die vollständige Lebensgeschichte erfassen kann. Dieses Phänomen erfaßt man durch 'Zensierung'. Es kann z.B. sein, daß Individuen vor Ende der Studie ausscheiden, obwohl das interessierende Ereignis (z.B. 'Tod') noch nicht eingetreten ist. Formal läuft dies auf das 'Stoppen' des zugrundeliegenden Prozesses hinaus - er wird zum Zeitpunkt der Zensur eingefroren. Dies ist eine Standardtechnik in der Theorie der stochastischen Prozesse.

Zusammenfassend stellen wir fest, daß die moderne Ereignisanalyse untrennbar mit modernsten mathematischen Gebieten wie der Theorie stochastischer Prozesse und stochastischer Analysis verknüpft ist.

Um die Problemstellung anzudeuten, skizzieren wir ein Beispiel aus (dem für dieses Gebiet grundlegenden) Buch P.K. Andersen, Ø. Borgan, R.D. Gill und N. Keiding (1993), 767 Seiten: *Statistical Models Based on Counting Processes*, [2], Example I.3.1.

Beispiel 4.0.5 (Überleben mit malignem Melanom)
Unter Beobachtung standen Patienten mit malignem Melanom (Hautkrebs charakterisiert durch dunkle Hautstellen) nach einer Radikaloperation. Der Tumor wurde zusammen mit der Haut im Umkreis von 2,5 cm vollständig entfernt. Die Patienten wurden an der Universitätsklinik von Odense, Dänemark, zwischen 1962 und 1977 beobachtet. Dies war eine historisch prospektive klinische Studie mit dem Ziel, den Einfluß von Risikofaktoren auf das Überleben zu studieren. Die Zeitskala war *nicht* die Kalenderzeit, sondern die Zeit ab der Operation. Es wurden Risikofaktoren wie Geschlecht und Alter sowie klinische Merkmale wie Tumordurchmesser und -ort oder histologische Befunde wie Tumordicke, Wachstum, Typ und Krebsartigkeit aufgenommen. Die Überlebenszeit ist nur für die Patienten bekannt, die vor Ende 1977 starben. Der Rest ist 'zensiert'. Manche Patienten starben zwar im Beobach-

tungszeitraum, jedoch nicht am Melanom.

In diesem Beispiel gibt es zwei Anlässe für zensierte Beobachtungen: Patienten konnten aus anderen Ursachen (und nicht am Melanom) sterben oder das Ende der Studie überleben.

Wichtige Quellen für dieses Kapitel sind der Übersichtsartikel 'Counting Process Models for Life History Data: A Review' von P. Kragh und Ø. Borgan, 1985, [1], das schon erwähnte Standardwerk [2] und ein unveröffentlichtes Vorlesungsskript von H. Pruscha, [14].

Als Referenz für die Grundlagen von Punkt- und Zählprozessen wird P. Brémaud: Point Processes and Queues. Martingale Dynamics (1981), [5], empfohlen, die grundlegende Theorie der stochastischen Analysis, insbesondere für nicht stetige Prozesse, wie hier benötigt, findet man in H.v. Weizsäcker und G. Winkler: Stochastic Integrals, [16]. Dort wird auch auf die hinderlichen 'üblichen Bedingungen' vieler anderer Bücher über stochastische Analysis verzichtet.

4.1 Herkömmlicher Zugang

Bevor wir zur Theorie kommen, stellen wir einige heuristische Überlegungen zu den Grundbegriffen herkömmlicher Art an. Außerdem skizzieren wir die (wohl jahrhunderte alte) Sterbetafelmethode.

4.1.1 Survivalfunktion und Hazardrate

Wir führen einige wichtige statistische Kenngrößen der Verweildauer im Ein-Episodenfall ein. Es wird von einer homogenen Population ausgegangen. Die Verweildauer, d.h. die Dauer der Episode, wird durch eine Zufallsvariable mit Werten in $[0,\infty]$ repräsentiert. Ihre Verteilungsfunktion sei F; diese sei absolutstetig mit Dichte f. Die *Survivalfunktion*

$$\boldsymbol{S}(t) = \boldsymbol{P}(T \geq t) = 1 - \boldsymbol{F}(t)$$

gibt die Wahrscheinlichkeit an, daß ein Individuum bis zum Zeitpunkt t noch lebt; die allgemeine Interpretation ist, daß bis zur Zeit t noch kein interessierendes Ereignis aufgetreten ist.

Die *Hazardrate* wird herkömmlich definiert durch

$$\lambda(t) = \lim_{h\to 0} \frac{1}{h} \boldsymbol{P}(t \leq T \leq t+h | T \geq t), \tag{4.1}$$

falls dieser Grenzwert existiert.

Sie gibt (idealisiert) die Wahrscheinlichkeit an, daß das interessierende Ereignis im Zeitintervall $[s, t]$ eintritt, falls dieses überhaupt erreicht wird, wobei 'Rate' meint, daß diese Größe auf die Länge $t - s$ dieses Intervalles bezogen wird und die Länge des Intervalles infinitesimal klein ist! Sie ist natürlich keine bedingte Wahrscheinlichkeit mehr, sie kann größer als eins sein. (Am ehesten kann man diese Raten noch mit z.B. Geschwindigkeiten als 'Fortbewegungsraten' interpretieren). Andere Bezeichnungen sind Intensitäts- oder Risikofunktion, Übergangsrate oder Mortalitätsrate. Die Hazardrate informiert über den weiteren Verlauf, falls ein Individuum den Zeitpunkt t überlebt. Das Integral

$$\Lambda(t) = \int_0^t \lambda(s)\, ds$$

ist die *kumulierte Hazardrate*, die das bis t angehäufte Risiko einsammelt. Sie steht zu λ in derselben Beziehung wie $\boldsymbol{F}$ zu f. Aus der Definition folgt heuristisch

$$\begin{aligned}\lambda(t) &= \lim_{h\to 0} \frac{1}{h} \cdot \boldsymbol{P}(t \leq T \leq t+h | T \geq t) = \lim_{h\to 0} \boldsymbol{P}\frac{1}{h}\frac{(t \leq T \leq t+h)}{\boldsymbol{P}(T \geq t)} \\ &= \frac{1}{\boldsymbol{P}(T \geq t)} \lim_{h\to 0} \frac{1}{h} \boldsymbol{P}(t \leq T \leq t+h) = \frac{f(t)}{\boldsymbol{S}(t)},\end{aligned}$$

kurz

$$\lambda(t) = \frac{f(t)}{\boldsymbol{S}(t)} = \frac{f(t)}{1 - \boldsymbol{F}(t)}.$$

Umgekehrt rechnet man:

$$\begin{aligned}\int_0^t \lambda(s) ds &= -\int_0^t \frac{-f(s)}{1 - \boldsymbol{F}(s)} ds = -\ln(1 - \boldsymbol{F}(s))|_0^t \\ &= -\ln(1 - \boldsymbol{F}(t)) = -\ln \boldsymbol{S}(t).\end{aligned}$$

Daraus folgt

$$\boldsymbol{S}(t) = e^{-\int_0^t \lambda(s)\, ds},$$

d.h. die Survivalfunktion fällt exponentiell mit der kumulierten Hazardrate. Aus diesen Beziehungen ergibt sich, daß jede der drei Größen $f(t), \lambda(t)$ und $\boldsymbol{S}(t)$ die beiden anderen bestimmt.

Beispiel 4.1.1 (Die Exponentialverteilung) Sei die Hazardrate $\lambda(t) = \lambda,\ \lambda > 0$ konstant. Dann bekommen wir

$$\boldsymbol{S}(t) = e^{-\lambda t}, \boldsymbol{F}(t) = 1 - e^{-\lambda t}, f(t) = \lambda e^{-\lambda t},$$

also die Exponentialverteilung. Die erwartete Verweildauer ist $\boldsymbol{E}(T) = \lambda^{-1}$, sie sinkt mit höherem 'Risiko' λ. Die Varianz $\boldsymbol{V}(T) = \lambda^{-2}$ wächst mit zunehmender erwarteter Verweildauer.

4.1.2 Die Sterbetafel-Methode

Sie ist eine der einfachsten Methoden zur Analyse von Verweildauern. Methodisch paßt sie eigentlich nicht in unsere (martingalbasierte) Theorie. Aufgrund ihrer praktischen Bedeutung sei sie dennoch skizziert. Wir folgen [9], Kapitel 7.2.1.

Die Methode wurde früher vor allem von Demographen und Versicherungsmathematikern angewendet (Beiträge, Gesundheitssystem). Die *Kohorten-Sterbetafel (Generations-, Populations-Sterbetafel)* enthält Überlebenszeiten in einer Kohorte, d.h. von Personen, die in einem festen Zeitraum geboren wurden. Solche Längsschnittanalysen können erst abgeschlossen werden, wenn alle Kohortenmitglieder gestorben sind. Dies mindert natürlich ihren praktischen Wert, z.B. wegen der mangelnden Aktualität (was kürzlich in Deutschland politisch relevant war). Deshalb werden meist *Perioden-* oder *Querschnitts-Sterbetafeln* erstellt. Hier werden Personen aus der momentan lebenden Bevölkerung über wenige Jahre beobachtet. Neben wirklichen Sterbefällen sind Dauer von Krankenhausaufenthalten, von Arbeitslosigkeit etc. interessierende Zielgrößen.

Kovariablen werden nicht eingebracht. Ihr Einfluß kann erst nachträglich durch Gruppenvergleiche studiert werden. Man arbeitet in diskreter Zeit. Seien also

$$I_k = [a_{k-1}, a_k),\ a_0 = 0,\ a_{q+1} = \infty,\ k = 1, \dots, q$$

aufeinanderfolgende Zeitintervalle. Die diskreten Hazardraten sind

$$\lambda_k = \boldsymbol{P}(T \in [a_{k-1}, a_k) | T \geq a_{k-1}).$$

Mit

$$p_k = \boldsymbol{P}(T \geq a_k | T \geq a_{k-1}),\ \ P_k = \boldsymbol{P}(T \geq a_k)$$

ist $\lambda_k = 1 - p_k$. Weiter gilt $P_k = p_k \cdot P_{k-1}$ und somit $P_k = p_k \cdot \ldots \cdot p_1$.
Folgende Daten werden erhoben:

N: die Gesamtzahl der Beobachtungen zu Beginn der Studie,

d_k: Anzahl der Sterbefälle während I_k

w_k: Anzahl der Zensierungen in I_k, also der Fälle, für die 'Tod' in I_k nicht festgestellt werden konnte und das Erreichen von I_{k+1} unbekannt war.

Die Größe n_k der Risikogruppe in I_k ist

$$n_1 = N, \;\; n_k = n_{k-1} - d_{k-1} - w_{k-1}.$$

Ohne Zensierung in I_k ist

$$\hat{\lambda}_k = d_k / n_k$$

ein Schätzer für die Hazardrate. Bei Zensierung erfolgt üblicherweise die Korrektur

$$\hat{\lambda}_k^C = \frac{d_k}{n_k - w_k/2},$$

was, falls nicht detailliert begründet, eine gewisse Willkür beinhaltet. Mit $\hat{p_k} = 1 - \hat{\lambda}_k$ folgt

$$\hat{P}_k = \hat{p}_1 \cdot \ldots \cdot \hat{p}_k.$$

als Schätzung für die Überlebenszeitwahrscheinlichkeiten zu den Zeiten a_k. $\hat{P}_k$ als Schätzung der Survivalfunktion $\boldsymbol{S}(a_k)$ heißt *kumulative Überlebensrate*. Für Überlebenszeiten für die Intervallmitten $m_k = (a_k + a_{k-1})/2$ schätzt man

$$\hat{P}(T \geq m_k) = \frac{\hat{P}_k + \hat{P}_{k-1}}{2} = \frac{\hat{P}_{k-1}(1 + \hat{p}_k)}{2}.$$

Außerdem schätzt man

$$\hat{P}(a_{k-1} \leq T < a_k) = \hat{P}_{k-1} - \hat{P}_k,$$

und die Dichte

$$\hat{f}_k = \frac{\hat{P}_{k-1} - \hat{P}_k}{l_k} = \frac{\hat{P}_{k-1}\hat{\lambda}_k}{l_k}, \;\; l_k = a_k - a_{k-1}.$$

und die 'mittlere Hazardrate'

$$\hat{\lambda}(m_k) = \frac{\hat{f}_k}{\hat{P}(T \geq m_k)} = \frac{2\hat{\lambda}_k}{1 + \hat{p}_k}.$$

Beispiel 4.1.2 (Sterbetafel 'Dauer von Arbeitslosigkeit') Aus den ersten sechs Wellen des Sozioökonomischen Panels wurden für den Zeitraum 1983 bis 1988 N = 1669 Arbeitslosigkeitsdauern ermittelt. Die Dauern sind in Monaten gemessen. Die 'Sterbetafel'-Schätzungen sind in Tabelle 4.1 wiedergegeben.

k	$[a_{k-1}, a_k]$ in Monaten	n_k	w_k	$n_k - w_k/2$	d_k	$\hat{\lambda}_k$	$\hat{P}_k$	$\hat{f}_k$	$\hat{\lambda}(m_k)$
1	[0,1)	1669	0	1669.0	0	0	1.0000	0	0
2	[1,2)	1669	131	1603.5	197	.1229	.8771	.1229	.1309
3	[2,3)	1341	7	1337.5	178	.1331	.7604	.1167	.1426
4	[3,4)	1156	12	1150.0	159	.1383	.6553	.1051	.1485
5	[4,5)	985	3	983.5	89	.0905	.5960	.0593	.0948
6	[5,6)	893	9	888.5	86	.0968	.5383	.0577	.1017
7	[6,7)	798	6	795.0	81	.1019	.4834	.0548	.1074
8	[7,8)	711	4	709.0	53	.0748	.4473	.0361	.0777
9	[8,9)	654	5	651.5	58	.0890	.4075	.0398	.0932
10	[9,10)	591	7	587.5	45	.0766	.3763	.0312	.0796
$\vdots$									
33	[32,33)	43	0	43.0	1	.0233	.0470	.0011	.0235
34	[33,34)	42	2	41.0	2	.0488	.0447	.0023	.0500
35	[34,35)	38	0	38.0	3	.0789	.0412	.0035	.0822
36	[35,36)	35	0	35.0	0	0	.0412	0	0
37	[36,∞)	35	0	35.0	35	1.0000	0		

Tabelle 4.1: 'Sterbetafel' zur Dauer von Arbeitslosigkeit der ersten sechs Wellen des Sozioökonomischen Panels. Auszug aus [9], Tab. 2.1, Seite 320

4.2 Submartingale

In diesem Abschnitt wird die benötigte Theorie stochastischer Prozesse angerissen. Von besonderer Bedeutung sind die Submartingale, deren wichtiger Spezialfall die Zählprozesse darstellen. Im Zentrum des Interesses steht die Zerlegung von Submartingalen in wachsende Prozesse - die gewisse Trends repräsentiernen - und Martingale als Modelle für zufällige Schwankungen.

4.2.1 Filtrationen und Stoppzeiten

Sei wie üblich ein Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \boldsymbol{P})$ gegeben. Sei $I \subset \mathbb{R}_+$ eine Menge von Zeitpunkten. Eine *Filtration* ist eine aufsteigende Familie von Unter-σ-Algebren $\mathcal{F}_t$, $t \in I$, von $\mathcal{F}$, d.h.

$$\mathcal{F}_s \subset \mathcal{F}_t, \quad s \leq t.$$

Fast immer wird I ein Intervall $[0, a]$ oder $[0, \infty]$ sein.

Beispiel 4.2.1 (a) Ist $(X_t)_{t \in I}$ ein stochastischer Prozeß, so ist durch

$$\mathcal{F}_t^X = \sigma(X_s : s \leq t)$$

die *natürliche Filtration* gegeben.
(b) Ist $(Y_t)_{t \in I}$ ein weiterer stochastischer Prozeß, so definiert

$$\mathcal{F}_t = \sigma(X_s, Y_s : s \leq t)$$

eine Filtration, die im allgemeinen weder für (X_t) noch (Y_t) die natürliche ist. Es gelten nur:

$$\mathcal{F}_t^X \subset \mathcal{F}_t, \;\; \mathcal{F}_t^Y \subset \mathcal{F}_t.$$

Wir werden im folgenden einer Reihe von Filtrationen begegnen, die nicht unbedingt von Prozessen erzeugt sind. Jedenfalls repräsentiert die Filtration den mit der Zeit wachsenden Wissensstand; bei natürlichen Filtrationen beruht er auf der laufenden Beobachtung eines Prozesses.

Allgemeiner heißt (X_t) bzgl. $(\mathcal{F}_t)$ *adaptiert*, wenn für jedes $t \in I$ die Variable X_t bezüglich $\mathcal{F}_t$ meßbar ist. Der gesamte Wissensstand der Filtration wird in

$$\mathcal{F}_\infty = \sigma(\bigcup_{t \in I} \mathcal{F}_t)$$

zusammengefaßt. Insbesondere ist

$$(\mathcal{F}_t)_{t \in I \cup \{\infty\}}$$

eine Filtration. Weitere Filtrationen sind von Interesse z.B.

$$\mathcal{F}_t^+ = \bigcap_{s > t} \mathcal{F}_s;$$

die σ-Algebra $\mathcal{F}_t^+$ erlaubt einen infinitesimalen Blick in die Zukunft. Die σ-Algebra

$$\mathcal{F}_{t-} = \sigma\left(\bigcup_{s<t} \mathcal{F}_s\right) \tag{4.2}$$

beinhaltet die Ereignisse bis unmittelbar vor t.

Deterministische Zeitpunkte $t \in I$ reichen nicht aus, um wichtige Aussagen über die Prozesse zu formulieren. Zum Beispiel ist der Zeitpunkt des Ruins eines Spielers, der Todeszeitpunkt oder allgemeiner des Eintretens gewisser Ereignisse von den Individuen in der Population und somit vom Zufall abhängig. Also müssen wir zufällige Zeitpunkte $T(\omega)$ betrachten. Von besonderem Interesse sind solche, von denen wir in jedem Zeitpunkt $t \in I$ entscheiden können, ob sie eingetreten sind oder nicht. Dies schlägt sich in der Forderung.

$$T : \Omega \longrightarrow [0, \infty], \ \ \{T \le t\} \in \mathcal{F}_t, \ \ t \in I,$$

nieder. Zufällige Zeiten mit dieser Eigenschaft heißen *Stoppzeiten*.

Beispiel 4.2.2 Sei (X_t) ein Prozeß mit stetigen Pfaden auf $I = \mathbb{R}_+$. Dann ist für jede *abgeschlossene* Menge G die Eintrittszeit

$$T^G(\omega) = \inf\{t \ge 0 : X_t(\omega) \in G\}$$

eine Stoppzeit (Beweis z.B. in [16], Abschnitt 2.3). Somit sind die zufälligen Zeiten des Über- oder Unterschreitens gewisser Schranken Stoppzeiten. Bei fast allen Zählprozessen sind die Auftrittszeiten der Ereignisse Stoppzeiten.

Umgekehrt können Prozesse aus Stoppzeiten konstruiert werden.

Beispiel 4.2.3 (Der Ein-Sprung oder Ein-Episoden-Zählprozeß) Sei $T : \Omega \longrightarrow [0, \infty]$ eine zufällige Zeit. Dann ist der Indikatorprozeß

$$N_t(\omega) = \mathbf{1}_{\{T(\omega) \le t\}}(\omega), \ \ t \ge 0,$$

ein stochastischer Prozeß [1]. Es ist

$$N_t(\omega) = \begin{cases} 0 & \text{falls}\, t < T(\omega) \\ 1 & \text{falls}\, t \ge T(\omega) \end{cases}$$

[1] mit rechtsstetigen Pfaden, die überdies linksseitige Limiten besitzen. Diese Eigenschaft wird oft cadlag genannt (continué à droite, limites à gauche).

(N_t) ist $(\mathcal{F}_t)$-adaptiert genau dann, wenn T eine Stoppzeit ist, denn es ist

$$\{T \leq t\} = \{\omega : T(\omega) \leq t\} = \{\omega : N_t(\omega) = 1\}.$$

Diese Menge ist $\mathcal{F}_t$-meßbar genau dann, wenn (N_t) adaptiert ist. Dies ist das einfachste Beispiel eines Zählprozesses.

Bemerkung 4.2.4 Stoppzeiten müssen als Funktionen $T : \Omega \longrightarrow [0, \infty]$ nicht notwendig $\mathcal{F}$-meßbar sein.

Zu jeder Stoppzeit T gehört die σ-Algebra

$$\mathcal{F}_T = \{F \in \mathcal{F} : F \cap \{T \leq t\} \in \mathcal{F}_t,\ t \in I\}$$

der *T-Vergangenheit*. Das sind die Ereignisse, auf denen wir zu jeder Zeit t entscheiden können, ob das zu $\{T \leq t\}$ gehörige Ereignis zur Zeit t eingetreten ist oder nicht.

4.2.2 Bedingte Erwartungen und (Sub-) Martingale

Eine wichtige Klasse von Prozessen sind Martingale, welche z.B. faire Spielabläufe beschreiben. Bevor sie eingeführt werden, muß noch etwas über bedingte Erwartungen gesagt werden. Im Abschnitt 3.4 wurde für Zufallsvariablen X mit zweitem Moment, d.h. in $\mathcal{L}^2(\boldsymbol{P})$, die bedingte Erwartung $\boldsymbol{E}(X|\mathcal{G})$ bezüglich einer Unter-σ-Algebra $\mathcal{G}$ von $\mathcal{F}$ als orthogonale Projektion auf den Raum

$$\mathcal{H} = \{Z \in \mathcal{L}^2(\boldsymbol{P}) : Z \,\text{ist bzgl.}\, \mathcal{G}\,\text{meßbar}\}$$

eingeführt. Anschließend wurde die Glättungseigenschaft

$$\int_G X\, dP = \int_G \boldsymbol{E}(X|\mathcal{F})\, d\boldsymbol{P},\ G \in \mathcal{G}, \tag{4.3}$$

nachgewiesen. Bedingte Erwartungen für Zufallsvariablen aus dem Raum

$$\mathcal{L}^1(\boldsymbol{P}) = \{X \text{ ist Zufallsvariable mit } \boldsymbol{E}(|X|) < \infty\}$$

der Zufallsvariablen mit endlichem *ersten* Moment sind durch die Identität (4.3) definiert.

Bemerkung 4.2.5 Der Existenzbeweis ist (zumindest von der Idee her) einfach: Zunächst gilt

$$\mathcal{L}^2(\boldsymbol{P}) \subset \mathcal{L}^1(\boldsymbol{P}),$$

was in Bemerkung 3.4.1 gezeigt wurde. Ferner kann man jede Zufallsvariable X mit existierendem ersten Moment durch solche mit existierendem zweiten Moment in $\mathcal{L}^1(\boldsymbol{P})$ approximieren, z.B. durch

$$X^{(n)} = \max\{\min\{X, n\}, -n\},$$

und so die bedingte Erwartung auf von $\mathcal{L}^2(\boldsymbol{P})$ auf $\mathcal{L}^1(\boldsymbol{P})$ ausdehnen.

Folgende Eigenschaften der bedingten Erwartung sind nützlich. Die meisten lassen sich aus den Eigenschaften orthogonaler Projektionen in $\mathcal{L}^2(\boldsymbol{P})$ ablesen und anschließend auf $\mathcal{L}^1(\boldsymbol{P})$ übertragen.

Eigenschaften 4.2.6 *(a) Ist $X = Y$ f.s., so $\boldsymbol{E}(X|\mathcal{G}) = \boldsymbol{E}(Y|\mathcal{G})$ f.s.*

(b) Die Abbildung $X \mapsto \boldsymbol{E}(X|\mathcal{G})$ ist linear.

(c) Ist X bezüglich $\mathcal{G}$ meßbar, so ist

$$\boldsymbol{E}(X|\mathcal{G}) = X \ \ f.s.$$

(d) Ist X von $\mathcal{G}$ unabhängig, so

$$\boldsymbol{E}(X|\mathcal{G}) = \boldsymbol{E}(X).$$

(g) Ist $\mathcal{G}_1 \subset \mathcal{G}_2$, so

$$\boldsymbol{E}(\boldsymbol{E}(X|\mathcal{G}_2)|\mathcal{G}_1) = \boldsymbol{E}(X|\mathcal{G}_1).$$

Eine grundlegende Klasse von Prozessen für die modernen ‘Life-History-Analysis’ sind Martingale: Ein adaptierter stochastischer Prozeß $(M_t)_{t\in I}$ mit $M_t \in \mathcal{L}^1(\boldsymbol{P})$ heißt *Martingal*, wenn für $s \leq t$

$$\boldsymbol{E}(M_t|\mathcal{F}_s) = M_s \ \ \boldsymbol{P} - f.s.$$

Falls $M_t \in \mathcal{L}^2(\boldsymbol{P})$, $t \in I$, so heißt M_t ein $\mathcal{L}^2$-*Martingal*.

Bemerkung 4.2.7 Jedenfalls für $\mathcal{L}^2$-Martingale haben wir bereits eine sehr intuitive Interpretation. Sei $(\mathcal{F}_t^M)_{t\in I}$ die natürliche σ-Algebra für einen Prozeß $(M_t)_{t\in I}$. Dann ist

$$\boldsymbol{E}(M_t|\mathcal{F}_s) = \boldsymbol{E}(M_t|M_u : u \leq s).$$

Nach Proposition 3.2.1 ist dies aber der (beste) MMS-Schätzer für M_t gegeben die gesamte Geschichte vor (dem gegenwärtigen) Zeitpunkt s. Ist (M_t) ein Martingal, so ist also die beste MMS-Schätzung für den zukünftigen Zustand M_t gegeben unser gesamtes Wissen bis zum Zeitpunkt $s \leq t$ der gegenwärtige Zustand M_s. Repräsentiert (M_t) beispielsweise unser Kapital bei einem Spiel, so wäre dieses 'fair' zu nennen, da wir unseren Zugewinn aufgrund des beobachteten Verlaufes weder ins Positive noch ins Negative bewegen können. In $\mathcal{L}^1(\boldsymbol{P})$ ist die Interpretation nicht ganz so einleuchtend; es bleibt uns aber immer noch die Glättungsinterpretation, aus der definierenden Identität (4.3).

In der Survivalanalyse werden Martingale als zufälliges Rauschen interpretiert, da sie ohne deterministischen Trend sind.

Beispiel 4.2.8 (Der Brownsche Prozeß) Eine der Definitionen des *Brownschen Prozesses* [2] $(B_t)_{t\geq 0}$ ist:

(i) für $0 \leq t_0 \leq \cdots \leq t_n$ sind die Zuwächse

$$B_{t_1} - B_{t_0}, \ldots, B_{t_n} - B_{t_{n-1}}$$

unabhängig.

(ii) $B_0 = 0$ und $\boldsymbol{V}(B_t - B_s) = t - s$

(iii) $B_t - B_s \sim \mathcal{N}(0, t - s)$.

[2] Die unregelmäßige Bewegung von Pollen in Wasser (deren Ursache später durch die Molekularbewegung erklärt wurde) wurde 1828 vom Botaniker Robert Brown im Mikroskop beobachtet. Der theoretische Durchbruch erfolgte 1905 durch Albert Einstein in seiner Arbeit über die Erklärung der Bewegung kleiner Partikel in Flüssigkeiten aus der Molekularen Theorie der Wärme [8] (für Arbeiten aus diesem Gebiet bekam er den Nobelpreis und nicht für die Relativitätstheorie). Die mathematische Fundierung erfolgte durch Norbert Wiener. Wesentliche Aspekte der Brownschen Bewegung wurden aber schon 1900 vom französischen Mathematiker Louis Bachelier in seiner Dissertation mit dem Titel 'Théorie de la spéculation' herausgearbeitet.

(iv) Die Pfade $t \mapsto B_t(\omega)$ sind fast sicher stetig.

Jedenfalls gilt: $B_t - B_s$ ist unabhängig von allen $B_u = B_u - B_0$, $u \leq s$, also ist auch $B_t - B_s$ unabhängig von $\mathcal{F}_s^B$. Prozesse mit dieser Eigenschaft nennt man *Prozesse mit unabhängigen Zuwächsen*. Man kann diese Definition leicht erweitern, indem man fordert: für alle $s < t < u$ ist $B_u - B_t$ unabhängig von $\mathcal{F}_s$.

Dieses Beispiel kann wesentlich ausgeweitet werden:

Beispiel 4.2.9 (Prozesse mit unabhängigen Zuwächsen) Sie liefern eine Vielfalt von Martingalen: Ein stochastischer Prozeß $(X_t)_{t\in I}$ hat *unabhängige Zuwächse* bzgl. der Filtration $(\mathcal{F}_t)_{t\in I}$, wenn gilt

$$X_t - X_s \text{ ist unabhängig von } \mathcal{F}_s,\ s \leq t.$$

Das letzte Beispiel zeigt, daß die Brownsche Bewegung unabhängige Zuwächse hat.

Theorem 4.2.10 *Jeder Prozeß in $\mathcal{L}^1(\boldsymbol{P})$ mit unabhängigen mittelwertfreien Zuwächsen ist ein Martingal.*

Beweis. Für $s < t$ gilt

$$\boldsymbol{E}(X_t|\mathcal{F}_s) = \boldsymbol{E}(X_s|\mathcal{F}_s) + \boldsymbol{E}(X_t - X_s|\mathcal{F}_s).$$

Wegen der Unabhängigkeit der Zuwächse und Eigenschaft 4.2.6(d) ist der letzter Term gleich

$$\boldsymbol{E}(X_t - X_s) = 0 \ \ f.s.$$

Wegen 4.2.6(c) ist $\boldsymbol{E}(X_s|\mathcal{F}_s) = X_s$ f.s. d.h.

$$\boldsymbol{E}(X_t|\mathcal{F}_s) = X_s \ \ f.s.$$

Damit ist der Satz bewiesen.

Beispiel 4.2.11 Da die Brownsche Bewegung orthogonale Zuwächse hat, ist sie ein Martingal.

In Verallgemeinerung der Martingale definiert man Prozesse mit monotonem Trend. Ein $(\mathcal{F}_t)$-adaptierter Prozeß (X_t) in $\mathcal{L}^1(\boldsymbol{P})$ heißt *Submartingal*, wenn

$$M_s \leq \boldsymbol{E}(M_t|\mathcal{F}_s) \ f.s., \ s \leq t$$

und *Supermartingal*, wenn

$$M_s \geq \boldsymbol{E}(M_t|\mathcal{F}_s) \ f.s., \ s \leq t.$$

Beispiel 4.2.12 Der *Poisson Prozeß* $(N_t)_{t\geq 0}$ teilt die beiden ersten Eigenschaften mit der Brownschen Bewegung; man kann ihn so definieren:

(i) (N_t) hat unabhängige Zuwächse.

(ii) $N_0 \equiv 0$ und $N_t - N_s$ hat Varianz $t - s$ für $s \leq t$.

(iii) Die Zuwächse $N_t - N_s$ sind poissonverteilt mit Mittel $t - s$.

Da die Erwartungswerte $\boldsymbol{E}(N_t) = t$ wachsen, ist der Poisson Prozeß offensichtlich kein Martingal bzgl. seiner natürlichen Filtration $(\mathcal{F}_t^N)_{t\geq 0}$. Wir machen ihn mittelwertfrei durch

$$M_t = N_t - t.$$

(M_t) ist $(\mathcal{F}_t^N)$-adaptiert, hat unabhängige Zuwächse und es ist $\boldsymbol{E}(M_t) = 0$ für jedes t. Also ist (M_t) ein Martingal, das sogenannte *Poissonmartingal*. Es ist für $s \leq t$

$$\boldsymbol{E}(N_t|\mathcal{F}_s) = \boldsymbol{E}(N_t - t|\mathcal{F}_s) + t = (N_s - s) + t = N_s + (t - s),$$

d.h. (N_t) ist ein Submartingal. Wichtig ist, daß sich (N_t) (eindeutig wie gleich gezeigt wird) zerlegen läßt gemäß:

$$N_t = M_t + t, \tag{4.4}$$

wobei (M_t) ein Martingal und t ein völlig deterministischer 'Prozeß' ist. Solche *'Doob-Meyer'-Zerlegungen* sind der Schlüssel zum Studium von Zählprozessen.

4.2.3 Vorhersehbarkeit und Doob-Meyer Zerlegung

Zerlegungen eines Zählprozesses - wie speziell des Poissonprozesses in (4.4) - in einen deterministischen, 'vorhersehbaren' Trend und ein als Rauschen zu deutendes Martingal ist zentral in der Theorie der Zählprozesse. Wir machen uns das am Beispiel diskreter Zeit klar.

Beispiel 4.2.13 (Vorhersehbarkeit in diskreter Zeit) Für diskrete Zeit sei $I = \{0, 1, 2, \ldots\}$; ferner sei $(\mathcal{F}_n)_{n \geq 0}$ eine Filtration. Ein stochastischer Prozeß (X_n) ist natürlich '$(\mathcal{F}_n)$-vorhersehbar' , wenn die Information, die X_{n+1} liefert, schon zur Zeit n bekannt ist, d.h. jedes X_{n+1} bezüglich $\mathcal{F}_n$ meßbar ist. Sei nun (X_n) ein $(\mathcal{F}_n)$-Submartingal, d.h.

$$X_n \leq \boldsymbol{E}(X_{n+1} | \mathcal{F}_n) \;\; f.s. \text{ für jedes } n \geq 0.$$

Intuitiv gesprochen, hat (X_n) einen Trend nach oben (jedenfalls im Mittel), denn es gilt

$$\boldsymbol{E}(X_n) \leq \boldsymbol{E}(X_{n+1}), \;\; n \geq 0.$$

Wir wollen nun (X_n) zerlegen in einen 'Rauschanteil', d.h. hier ein Martingal (M_n), und einen wachsenden Trend (A_n), *welcher vorhersehbar sein soll.* Wir setzen $M_0 = X_0$ und $A_0 = 0$ und sodann rekursiv

$$\begin{aligned} A_n &= A_{n-1} + (\boldsymbol{E}(X_n | \mathcal{F}_{n-1}) - X_{n-1}), \\ M_n &= X_n - A_n. \end{aligned}$$

Jedes A_n ist $\mathcal{F}_{n-1}$-meßbar; deshalb ist (A_n) vorhersehbar. Weil (X_n) ein Submartingal ist, ist die Differenz in der Klammer nichtnegativ und somit (A_n) wachsend. Schließlich gilt $f.s.$:

$$\begin{aligned} & \boldsymbol{E}(M_n - M_{n-1} | \mathcal{F}_{n-1}) = \boldsymbol{E}(X_n - X_{n-1} - (A_n - A_{n-1}) | \mathcal{F}_{n-1}) \\ = \;\; & \boldsymbol{E}(X_n | \mathcal{F}_{n-1}) - X_{n-1} - (A_n - A_{n-1}) = 0 \end{aligned}$$

nach Definition von (A_n). Dies bedeutet

$$\boldsymbol{E}(M_n | \mathcal{F}_{n-1}) = M_{n-1}, \; n \geq 1, \;\; f.s.,$$

d.h. (M_n) ist ein Martingal.

Dieses Programm müssen wir nun auf kontinuierliche Zeit übertragen. Da wir das Hauptresultat nur zitieren werden, müssen wir uns über den Begriff 'Vorhersehbarkeit' klar werden. Eine gewisse Anschauung liefert folgender Satz (v. Weizsäcker, Winkler (1990), [16], Prop. 6.12.)

Theorem 4.2.14 *Für eine σ-Algebra $\mathcal{P}$ auf $(0,\infty)\times\Omega$ sind äquivalent: $\mathcal{P}$ wird erzeugt durch jede der folgenden Klassen von Mengen oder Prozessen:*

(a) den stochastischen Intervallen

$$(0,T] = \{(t,\omega) : 0 < t \leq T(\omega)\},$$

T Stoppzeit,

(b) den vorhersehbaren Rechtecken $(s,t]\times F,\ s<t,\ F\in\mathcal{F}_s$,

(c) den Abbildungen $(t,\omega)\longmapsto X_t(\omega)$ für alle stetige adaptierte Prozesse $(X_t)_{t>0}$,

(d) den Abbildungen $(t,\omega)\longmapsto X_t(\omega)$ für alle linksstetigen adaptierte Prozesse.

Dabei ist '(links-)stetig' pfadweise zu verstehen. Zur Interpretation betrachten wir eines der stochastischen Intervalle $(0,T]$ aus (a). Im Gegensatz zu 4.2.3 ist der Indikatorprozeß

$$X_t(\omega) = \mathbf{1}_{\{T<t\}}(\omega)$$

linksstetig und konstant $=0$ für $t\leq T(\omega)$ und konstant gleich 1 für $t>T(\omega)$. Jeder Wert $X_t(\omega)$ ist bekannt, wenn man $X_s(\omega)$ in einem beliebig kleinen Intervall $(t-\varepsilon,t)$ (rechts offen!) kennt. Daraus kann man $X_t(\omega)$ *vorhersagen*. Ähnliches gilt für allgemeinere (links-)stetige Prozesse.

Ein stochastischer Prozeß $(X_t)_{t\geq 0}$ heißt *vorhersehbar*, wenn X_0 bezüglich $\mathcal{F}_0$ und die Abbildung

$$(t,\omega)\longrightarrow X_t(\omega)$$

auf $(0,\infty)\times\Omega$ bezüglich $\mathcal{P}$ meßbar ist.

Bemerkung 4.2.15 Folgende Beobachtung veranschaulicht das Konzept: Ist (X_t) vorhersehbar und adaptiert, so ist jedes X_t bezüglich

$$\mathcal{F}_{t-} = \sigma(\bigcup_{s<t}\mathcal{F}_s).$$

meßbar, d.h. (X_t) ist $(\mathcal{F}_{t-})$-adaptiert. Für linksstetige adaptierte Prozesse ist das klar, da

$$\{X_t\leq a\} = \bigcap_{r>0,r\in\mathbb{Q}}\ \bigcup_{p>0,p\in\mathbb{Q}}\ \bigcap_{t-p<q<t,q\in\mathbb{Q}}\{X_q\leq a+r\}\in\mathcal{F}_{t-}.$$

Damit gilt die Aussage insbesondere für Indikatorprozesse zu vorhersehbaren Rechtecken, d.h.

$$X_r(\omega) = \mathbf{1}_{(u,v]\times F}(r,\omega),\ u < v,\ F \in \mathcal{F}_u.$$

Da sich jeder vorhersehbare Prozeß durch Linearkombinationen solcher Prozesse approximieren läßt, gilt die Aussage allgemein.

Wir schreiben den entscheidenden Satz jetzt auf. Unbekannte *termini technici* erläutern wir später.

Theorem 4.2.16 *Sei (X_t) ein rechtsstetiger adaptierter Prozeß. Betrachte die Aussagen*

(a) (X_t) ist ein (gleichgradig integrierbares) Submartingal.

(b) (X_t) besitzt eine Zerlegung

$$X_t = M_t + A_t$$

mit einem Martingal (M_t) und einem wachsenden vorhersehbaren rechtsstetigen Prozeß (A_t).

Dann wird (b) von (a) impliziert. Ergänzt man in (b): (M_t) ist gleichgradig integrierbar und (A_t) ist lokal $\mathcal{L}^2$-beschränkt, so sind (a) und (b) äquivalent und die Zerlegung ist f.s. eindeutig.

Der Prozeß (A_t) heißt *Kompensator* von (X_t). Für Beweise siehe [16], Theorem 6.6.1. Die technischen Einzelheiten lassen wir beiseite. Wichtig ist, daß diese Zerlegung auf Zählprozesse anwendbar ist, da sie in Verallgemeinerung von Beispiel 4.2.12 Submartingale sind:

Lemma 4.2.17 *Jeder adaptierte Prozeß in $\mathcal{L}^1(\boldsymbol{P})$ mit wachsenden Pfaden ist ein Submartingal.*

Beweis. Für $s < t$ gilt

$$\boldsymbol{E}(X_t|\mathcal{F}_s) \geq \boldsymbol{E}(X_s|\mathcal{F}_s) = X_s \ \ f.s.$$

Dies ist die Submartingaleigenschaft. □

In vielen Fällen der Praxis ist die Integrierbarkeitsbedingung in Satz 4.2.16 trivialerweise erfüllt. Für den Ein-Episodenprozeß $(X_t) = (N_t)$ z.B. ist ja stets $0 \leq N_t \leq 1$. Ähnliches gilt für Zählprozesse für Studien über n Individuen, wo $0 \leq N_t \leq n$ ist. In vielen anderen Fällen kann man die Bedingung durch den Trick der Lokalisierung erzwingen.

Definition 4.2.18 *Eine Folge (T_n) von Stoppzeiten heißt* lokalisierend*, wenn*

$$\boldsymbol{P}(T_n \nearrow \infty) = 1.$$

Ein Prozeß (X_t) heißt *lokal beschränkt*, wenn für eine lokalisierende Folge (T_n) gilt $(X_t)|(0, T_n] \leq c_n$ mit einer Konstanten c_n (genauer heißt das

$$X_t(\omega) \leq c_n \text{ auf der Menge } \{(t, \omega) : 0 < t \leq T_n(\omega)\}).$$

Ähnlich definiert man 'lokal integrierbar', 'lokal quadrat-integrierbar' usw. Dies kann man etwas anschaulicher formulieren: Sei (X_t) ein adaptierter Prozeß und T eine Stoppzeit. Dann definiert man den *gestoppten Prozeß* $(X_{t \wedge T})$ durch

$$X_{t \wedge T}(\omega) = \begin{cases} X_t(\omega) & \text{falls } \; t \leq T(\omega) \\ X_{T(\omega)}(\omega) & \text{falls } \; t > T(\omega) \end{cases} .$$

Nach $T(\omega)$ behält ein Pfad seinen Wert $X_{T(\omega)}(\omega)$ zur Zeit $T(\omega)$ konstant bei; vorher verhält er sich wie $X_t(\omega)$. Die Eigenschaft der lokalen Beschränktheit ist also äquivalent damit, daß jeder Prozeß $(X_{t \wedge T_n})$ (gleichmäßig) beschränkt ist.

Beispiel 4.2.19 Sei τ_n die zufällige Zeit des n-ten Sprunges des Poissonprozesses. Dann ist τ_n eine lokalisierende Folge. Ferner gilt

$$N_{\tau_n} = n.$$

und da der Poissonprozeß wächst, gilt sogar

$$N_{t \wedge \tau_n} \leq n$$

Also ist der Poissonprozeß lokal beschränkt. Insbesondere kann man die Doob-Meyer Zerlegung auf jeden Prozeß $(N_{t \wedge \tau_n})$ getrennt anwenden. Nimmt man sie wieder zusammen, so ergibt sich die eindeutige Zerlegung des Poissonprozesses in ein lokales Martingal und einen wachsenden vorhersehbaren Prozeß. Damit ist die Zerlegung (4.4) eindeutig.

Daß man sich durch das Stoppen eines Prozesses keine neuen Schwierigkeiten einhandelt, verhindert folgender Satz:

Theorem 4.2.20 *Ist (X_t) adaptiert, so auch $(X_{t_\wedge T})$ für eine Stoppzeit T.*

4.3 Modellierung durch Zähl- und Punktprozesse

Wir führen nun in die Modellierung im Rahmen der Zählprozesse ein.

4.3.1 Punkt- und Zählprozesse

Im gegenwärtigen Zusammenhang sind Punkt- und Zählprozesse zwei Seiten derselben Medaille. Wir führen beide ein und stellen den Zusammenhang her. Die Prozesse sollen die Zeitpunkte des Auftretens gewisser disjunkter diskreter Ereignisses registrieren.

Zugrunde liegen ein Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \boldsymbol{P})$ und eine Filtration $(\mathcal{F}_t)_{t\in I}$, wobei I ein Intervall $[0, b)$ oder $[0, b]$, meist mit $b = \infty$ ist. Wir betrachten gleich den multivariaten Fall; dazu sei E die abzählbare Menge von *Ereignistypen*[3]. Ein *markierter Punktprozeß* ist gegeben durch eine Doppelfolge $(\tau_n, \xi_n)_{n\geq 1}$ von Variablen

$$\tau_n : \Omega \longrightarrow [0, \infty], \;\; \xi_n : \Omega \longrightarrow E,$$

wobei gilt:

(i) τ_n ist $(\mathcal{F}_t)$- Stoppzeit,

(ii) ξ_n ist $\mathcal{F}_{\tau_n}$-meßbar,

(iii) Es ist

$$\begin{aligned} 0 < \tau_n < \tau_{n+1} \quad &\text{auf} \quad \{\tau_n < \infty\}, \\ \tau_n = \tau_{n+1} = \cdots \quad &\text{auf} \quad \{\tau_n = \infty\}. \end{aligned}$$

Man setzt noch $\tau_0 = 0$, und $\sigma_n = \tau_{n+1} - \tau_n$, falls $\tau_n < \infty$. Die Stoppzeit τ_n ist die *Auftrittszeit (occurence time)* des n-ten Ereignisses, unabhängig von seinem Typ, und ξ_n gibt den Typ des n-ten Ereignisses an (Mark, mark). Die Zeit σ_n ist die n-te *Zwischenauftrittszeit (Wartezeit, inter-occurence-time).* Sei schließlich

$$\tau_\infty(\omega) = \lim_{n\to\infty} \tau_n(\omega).$$

[3]Wir verwenden das Wort Ereignis also in einer zweiten Bedeutung.

Der Prozeß heißt *explodierend* für ω, wenn $\tau_\infty(\omega) < \infty$ und *auslöschend* für ω, wenn $\tau_n(\omega) = \infty$ für ein $n \geq 1$.

Aus dem markierten Punktprozeß läßt sich leicht der zugeordnete *Zählprozeß* konstruieren. Für jedes $i \in E$ und $t \in I$ sei

$$N_{it} = \sum_{n \geq 1} \mathbf{1}_{\{\tau_n \leq t\}} \mathbf{1}_{\{\xi_n = i\}}$$

die Anzahl der bis einschließlich der Zeit t eingetretenen Ereignisse vom Typ i. Es gilt

(a) $N_{it} \in \{0, 1, 2, \ldots\} \cup \{\infty\}$,

(b) $(N_{it})_{t \in I}$ ist $(\mathcal{F}_t)$-adaptiert.

Falls $|E| = 1$, so bekommt man einen *univariaten Zählprozeß.*
Beweis. Teil (a) ist klar. Für Teil (b) genügt es, zu zeigen daß

$$\{\tau_n \leq t\} \cap \{\xi_n = i\} \in \mathcal{F}_t,\ t \in I.$$

Nun ist $\{\xi_n = i\} \in \mathcal{F}_{\tau_n}$ und somit der fragliche Durchschnitt in $\mathcal{F}_t$ nach der Definition der τ_n-Vergangenheit.

Der Gesamtzahl der Ereignisse bis t ist

$$N_t = \sum_{i \in E} N_{it} = \sum_{n \geq 0} \mathbf{1}_{\{\tau_n \leq t\}},$$

ein ebenfalls adaptierter Prozeß. Man ergänzt durch

$$N_{i\infty}(\omega) = \lim_{t \uparrow \infty} N_{it}(\omega),\ N_\infty(\omega) = \lim_{t \uparrow \infty} N_t(\omega).$$

Es ist

$$N_\infty(\omega) < \infty \iff \text{Auslöschung auf}\, \omega.$$

Annahmen 4.3.1 *Der Prozeß (N_t) explodiere fast sicher nicht, d.h.*

$$\boldsymbol{P}(\omega : \tau_n(\omega) \nearrow \infty) = 1.$$

Unter dieser Voraussetzung ist (N_t) *lokal beschränkt*; es gilt sogar

$$\boldsymbol{P}(0 \leq N_{t \wedge \tau_n} \leq n) = 1$$

(mit $\tau_n \nearrow \infty$ f.s.). Deshalb folgt aus der Doob-Meyer Zerlegung 4.2.16 der grundlegende Satz:

Theorem 4.3.2 *Zu jedem multivariaten Zählprozeß* (N_{it}) *gibt es einen* Kompensator Prozeß (Λ_{it}) *mit den Eigenschaften*

(i) Es ist $\Lambda_{i0} = 0$, *die Pfade* $t \longmapsto \Lambda_{it}(\omega)$ *sind wachsend, rechtsstetig und adaptiert, der Prozeß* (Λ_{it}) *ist vorhersehbar*

(ii) Der Prozeß

$$M_{it} = N_{it} - \Lambda_{it}$$

ist ein (lokales) Martingal. Die Prozesse sind f.s. eindeutig bestimmt.

Bemerkung 4.3.3 Gilt

$$\boldsymbol{E}(N_{it}) < \infty \text{ für jedes } t \in I,$$

so ist jeder Prozeß (M_{it}) ein Martingal und es gilt somit stets

$$\boldsymbol{E}(\Lambda_{it}) = \boldsymbol{E}(N_{it}).$$

Es gilt sogar für jede Stoppzeit T, daß

$$\boldsymbol{E}(\Lambda_{iT}) = \boldsymbol{E}(N_{iT}).$$

Die Interpretation von (Λ_t) als systematischer Anteil oder 'Trend' wird weiter gestützt durch ein starkes Gesetz der großen Zahlen.

Theorem 4.3.4 *Es gilt*

$$\frac{N_{it}}{\Lambda_{it}} \longrightarrow 1,\ t \to \infty \ \textit{auf} \left\{ \lim_{t \uparrow \infty} \Lambda_{it} = \infty \right\}.$$

4.3.2 Der Intensitätsprozeß

Sei $(N_{it}, \mathcal{F}_t, \boldsymbol{P})$ ein multivariater Zählprozeß mit Kompensator (Λ_{it}). Um den Zusammenhang mit dem klassischen Zugang aus der Einleitung herzustellen, nehmen wir im Folgenden stets an:

Annahmen 4.3.5 Es gibt einen nichtnegativen vorhersehbaren Prozeß (λ_{it}) mit

$$\Lambda_{it} = \int_0^t \lambda_{is}\, ds.$$

Der Prozeß (λ_{it}) heißt *Intensitätsprozeß* von (N_{it}). Wir stellen nun die Beziehung zu den Intensitätsprozessen aus der 'Vormartingalzeit' her. Zunächst definieren wir

$$\begin{aligned} \gamma(t) &= \lim_{h\downarrow 0} \frac{1}{h} \boldsymbol{E}(N_{t+h} - N_t | \mathcal{F}_t) \\ \gamma^*(t) &= \lim_{h\downarrow 0} \frac{1}{h} \boldsymbol{P}(N_{t+h} - N_t = 1 | \mathcal{F}_t), \end{aligned}$$

wo auch immer diese Limiten existieren. Bevor wir die entscheidende Aussage formulieren führen wir noch einen neuen Begriff ein: Eine Filtration heißt *rechtsstetig*, wenn

$$\mathcal{F}_t = \mathcal{F}_t^+ = \bigcap_{s>t} \mathcal{F}_s.$$

Theorem 4.3.6 *Sei (N_t) ein beliebiger Zählprozeß, welcher zur rechtsstetigen Filtration $(\mathcal{F}_t)$ adaptiert ist. Ferner besitze (N_t) einen Intensitätsprozeß (λ_t) mit folgenden Eigenschaften*
(i) (λ_t) hat linksstetige Pfade mit rechtsseitigen Limiten.
(ii) Es gibt eine positive Zufallsvariable g mit $\boldsymbol{E}(g) < \infty$ und

$$\lambda_t \leq g.$$

Dann existieren die Limiten $\gamma(t)$ und $\gamma^*(t)$ und es gilt:

$$\lambda_{t+} = \gamma(t) = \gamma^*(t).$$

Beweis. Es ist $\boldsymbol{E}(N_t) = \boldsymbol{E}(\Lambda_t) < \infty$ und deshalb ist

$$M_t = N_t - \int_0^t \lambda_s \, ds$$

ein $(\mathcal{F}_t)$-Martingal. So ist

$$\boldsymbol{E}(M_{t+h} - M_t | \mathcal{F}_t) = 0$$

und

$$\frac{1}{h} \boldsymbol{E}(N_{t+h} - N_t | \mathcal{F}_t) = \frac{1}{h} \boldsymbol{E}\left(\int_t^{t+h} \lambda_s \, ds \,\middle|\, \mathcal{F}_t \right).$$

Weil

$$\frac{1}{h} \int_t^{t+h} \lambda_s \, ds \leq g, \quad \boldsymbol{E}(g) < \infty,$$

ist der Satz von der majorisierten Konvergenz anwendbar und liefert

$$\gamma(t) = \boldsymbol{E}\left(\lim_{h\to 0}\frac{1}{h}\int_t^{t+h}\lambda_s\,ds\,\middle|\,\mathcal{F}_s\right) = \boldsymbol{E}(\lambda_{t+}|\mathcal{F}_t) = \lambda_{t+},$$

da die Filtration rechtsstetig ist.

4.4 (Non-) parametrische Modelle

Wir stellen nun non- und semiparametrische Modelle vor.

4.4.1 Das multiplikative Intensitätsmodell

Sei ein multivariater Zählprozeß $(N_{it})_{i\in E,t\geq 0}$ zur Filtration $(\mathcal{F}_t)$ gegeben. Er besitze einen Intensitätsprozeß (λ_{it}).

Definition 4.4.1 *Der Prozeß* (N_{it}) *heißt* multiplikatives Intensitätsmodell *, falls*

$$\lambda_{it} = Y_{it}\cdot\alpha_i(t), \quad i\in E,$$

wobei α_{it} *eine nichtnegative linksstetige Funktion mit*

$$\int_0^t \alpha_i(s)\,ds < \infty,\ \ t\geq 0,$$

ist und (Y_{it}) *ein nichtnegativer* $(\mathcal{F}_t)$*-adaptierter Prozeß mit f.s. linksstetigen Pfaden.*

Wir setzen stets $\boldsymbol{E}(N_{it}) < \infty$ voraus (was $\tau_\infty = \infty$ f.s. impliziert). Die Linksstetigkeit der Pfade von (Y_{it}) kann zur Vorhersehbarkeit abgeschwächt werden. Die Prozesse (N_{it}) und (Y_{it}) werden als beobachtbar interpretiert. Zusätzlich kommen nun Kovariablen oder Regressoren ins Spiel. Sei also

$$X_{it} = (X_{i_1 t},\ldots,X_{i_p t}),\ \ i\in E, p>0,$$

ein adaptierter Prozeß mit linksstetigen Pfaden, der Prozeß der p-dimensionalen *Kovariablen*. Diese sollen nun parametrisch in obige α_i eingehen. Ziel ist die Schätzung der Parameter.

Beispiel 4.4.2 (Der Ein-Episoden-Fall) Wir betrachten wieder eine Stoppzeit $T > 0$ mit Dichte f und Verteilungsfunktion $\boldsymbol{F}$ und den zugehörigen Zählprozeß

$$N_t = \mathbf{1}_{\{T \leq t\}}, \ \ t \geq 0.$$

Nach Satz 4.3.6 gilt

$$\lambda_{t+} = \gamma_t^* = \lim_{h \downarrow 0} \frac{1}{h} \boldsymbol{P}(N_{t+h} - N_t = 1 | \mathcal{F}_t).$$

Wir zeigen unten

$$\begin{aligned} \boldsymbol{P}(N_{t+h} - N_t = 1 | \mathcal{F}_t) &= \mathbf{1}_{\{t<T\}} \, \boldsymbol{P}(t < T \leq t+h | T > t) \qquad (4.5) \\ &= \mathbf{1}_{\{t<T\}} \, \frac{\boldsymbol{F}(t+h) - \boldsymbol{F}(t)}{1 - \boldsymbol{F}(t)}. \end{aligned}$$

Daraus folgt dann für $\mathcal{F}_t = \mathcal{F}_t^N$ das klassische Resultat

$$\lambda_{t+} = \mathbf{1}_{\{t<T\}} \, \frac{f(t+)}{\boldsymbol{S}(t)}.$$

Wir sehen auch, daß ein multiplikatives Regressionsmodell vorliegt; setzen wir nämlich voraus, daß f stetig ist, so ergibt sich

$$\lambda_t = \mathbf{1}_{\{t \leq T\}} \frac{f(t)}{\boldsymbol{S}(t)} =: Y_t \cdot \eta(t)$$

mit $\eta(t) = f(t)/\boldsymbol{S}(t)$. Es bleibt die Indentität (4.5) nachzuweisen. Die linke Seite ist

$$\boldsymbol{E}(\mathbf{1}_{\{N_{t+h} - N_t = 1\}} | \mathcal{F}_t).$$

Bezeichnen wir die rechte Seite mit g, so wäre

$$\int_F \mathbf{1}_{\{N_{t+h} - N_t = 1\}} \, d\boldsymbol{P} = \int_F g \, d\boldsymbol{P} = R, \ \ F \in \mathcal{F}_t, \qquad (4.6)$$

nachzuweisen, d.h.

$$\begin{aligned} R &= \boldsymbol{P}(F \cap \{t < T\}) \, \frac{\boldsymbol{P}(t < T \leq t+h, \, t < T)}{\boldsymbol{P}(t < T)} \qquad (4.7) \\ &= \boldsymbol{P}(F \cap \{N_{t+h} - N_t = 1\}) =: L. \qquad (4.8) \end{aligned}$$

Es genügt, den Nachweis für einen Erzeuger von $\mathcal{F}_t$ zu führen, d.h. Mengen vom Typ $\{N_s \le a\}$ bzw. $\{N_s > a\}$, $s \le t$. Da die N_s nur die Werte 0 und 1 annehmen genügt z.B.

$$F = \left\{N_s < \frac{1}{2}\right\} = \{s < T\}.$$

Nun gilt:

$$\begin{aligned}
\boldsymbol{P}(\{s<T\} \cap \{t<T\}) &= \boldsymbol{P}(t<T) \\
\boldsymbol{P}(\{s<T\} \cap \{N_{t+h} - N_t\} = 1) &= \boldsymbol{P}(\{s<T\} \cap \{t<T\} \cap \{t+h \ge T\}) \\
&= \boldsymbol{P}(\{t<T\} \cap \{t+h \ge T\}) \\
&= \boldsymbol{P}(t < T \le t+h).
\end{aligned}$$

Also ist

$$L = \boldsymbol{P}(t < T \le t+h), \quad R = \frac{\boldsymbol{P}(t < T \le t+h)}{\boldsymbol{P}(T>t)} \boldsymbol{P}(T>t)$$

und die Behauptung ist bewiesen.

Abschließend betrachten wir noch den wachsenden Prozeß

$$\Lambda_t = \int_0^t \lambda(s)\, ds = \int_0^t \mathbf{1}_{\{s \le T\}} \frac{f(s)}{\boldsymbol{S}(s)}\, ds \; = \int_0^{t \wedge T} \frac{f(s)}{\boldsymbol{S}(s)}\, ds.$$

Im Falle der Exponentialverteilung wäre das z.B.

$$\Lambda_t = \int_0^{t \wedge T} \frac{\lambda e^{-\lambda s}}{e^{-\lambda s}}\, ds = \lambda \cdot t \wedge T.$$

Wir kommen nun zu zensierten Daten. Hier wird der Fall erfaßt, daß die Individuen nur bis zu einer (zufälligen) Zeit unter Beobachtung stehen, die nichts mit den interessierenden Zeiten zu tun hat. Sie können zum Beispiel vor dem Tod die Studie verlassen oder ähnliches. Wir erläutern das am Ein-Episoden-Fall. Sei dazu eine zufällige Zeit T gegeben.

Beispiel 4.4.3 Sei $U > 0$ eine zweite zufällige Zeit, die unabhängig von T ist (z.B. das Ausscheiden aus der Studie). Zu U gehört (äquivalent) eine Indikatorvariable.

$$D = \left\{ \begin{array}{lll} 1 & \text{falls} \quad T \le U & \text{(keine Zensur)} \\ 0 & \text{falls} \quad T > U & \text{(Zensur)} \end{array} \right\} = \mathbf{1}_{\{T \le U\}}$$

Aus der Interpretation heraus ist es sinnvoll, nur D bzw. $T \wedge U$ als beobachtbar anzunehmen. Es gilt:

$$\{T \wedge U \leq t, D = 1\} = \{T \leq t, D = 1\} = \{T \leq (t \wedge U)\} =: Z \tag{4.9}$$

Man definiert den zugehörigen zensierten Zählprozeß als

$$N_t^c = \mathbf{1}_Z = \mathbf{1}_{\{T \leq t,\, D=1\}}.$$

Bemerkung 4.4.4 Zur zufälligen Zeit T gehörte im Beispiel 4.4.2 der Indikator- oder Zählprozeß

$$N_t(\omega) = \left\{ \begin{array}{lll} 0 & \text{falls} & t < T(\omega) \\ 1 & \text{falls} & t \geq T(\omega) \end{array} \right\} = \mathbf{1}_{\{T \leq t\}}(\omega)$$

(der rechtsstetig war). Analog gehört zur Zensierung U der *Zensierungsprozeß*

$$C_t(\omega) = \left\{ \begin{array}{lll} 1 & \text{falls} & t \leq U(\omega) \\ 0 & \text{falls} & t > U(\omega) \end{array} \right\} = \mathbf{1}_{\{t \leq U\}}(\omega).$$

Dieser Prozeß ist linksstetig und somit vorhersehbar, sofern er adaptiert ist. In dieser Schreibweise ist

$$N_t^c = \int_0^t C_s \, dN_s,$$

wobei (N_t) der unzensierte Prozeß aus Beispiel 4.4.2 ist. Diese Definition hält Verallgemeinerungen stand. Weiter ist $D = C_T$ der gestoppte Zensierungsprozeß.

Ist nun $(\mathcal{F}_t^N)$ die natürliche Filtration von (N_t), so muß natürlich (C_t) nicht dazu adaptiert sein; der Zählprozeß (N_t^c) muß ebenso wenig $(\mathcal{F}_t^N)$-adaptiert sein. Will man also die allgemeinen Resultate wie die Doob-Meyer Zerlegung und die Folgerungen daraus anwenden, so muß man (N_t^c), also auch (C_t) adaptiert machen und zur Filtration

$$(\mathcal{F}_t) = \sigma(\mathcal{F}_t^N, \mathcal{F}_t^C)$$

übergehen. Das begründet erneut, weshalb wir die Theorie auf den allgemeinen Begriff der Filtration aufbauen mußten.

Das Ziel ist nun, den Intensitätsprozeß λ^c zu berechnen. Nach Beispiel 4.4.2 ist die Intensität des ursprünglichen, jetzt nicht mehr beobachtbaren Zählprozesses N gegeben durch

$$\lambda(t) = \mathbf{1}_{\{t \leq T\}} \frac{f(t)}{\boldsymbol{S}(t)} =: \mathbf{1}_{\{t \leq T\}} \cdot \alpha(t),$$

wobei $\alpha(t)$ die Hazardrate ist.

Wegen (4.9) ist $N_t^c = N_{t\wedge U}$ und hat deshalb die Intensität

$$\lambda^c(t) = \mathbf{1}_{\{t\le T\wedge U\}} \cdot \alpha(t).$$

Insbesondere liegt ein multiplikatives Intensitätsmodell

$$\lambda^c(t) = Y_t^c \cdot \alpha(t)$$

vor, wobei Y_t^c die Anzahl der zur Zeit t unter Risiko stehenden Individuen angibt (hier lediglich 0 oder 1).

4.4.2 Multiplikatives und Coxsches Regressionsmodell

Eine erste Spezialisierung ist das *multiplikative Regressionsmodell* . Es hat die Gestalt

$$\lambda_{it}(\beta) = Y_{it} \cdot \gamma(t) \cdot \eta(\beta' X_{it}), \tag{4.10}$$

wobei β der p-dimensionale Vektor der Kovariablen ist. Neben (N_{it}) und (Y_{it}) sei auch (X_{it}) beobachtbar. Typischerweise ist (Y_{it}) ein Indikator, der angibt, ob das Objekt i kurz vor der Zeit t beobachtbar ist oder nicht. Die Funktion η soll bekannt sein. Sie wird als positiv vorausgesetzt. Von γ nimmt man $\int_0^t \gamma(s)\,ds < \infty$ an. Zu schätzen sind

(i) die Modellparameter β,

(ii) die *baseline hazard intensity* γ.

Es liegt also ein *semiparametrisches Regressionsmodell* vor. Eine wichtige weitere Spezialisierung enthält die folgende Definition.

Definition 4.4.5 *Falls $\eta(\cdot) = \exp(\cdot)$, so heißt (4.10)* Coxsches Regressionsmodell*.*

4.5 Statistische Inferenz

Die Modellierung des Coxschen Regressionsmodelles beinhaltet parametrische und nonparametrische Elemente. Letztere sind dem multiplikativen Intensitätsmodell immanent, erstere werden durch das Coxsche Modell ins Spiel gebracht. Wir beginnen mit dem nonparamtrischen Teil.

4.5.1 Nelson-Aalen- und Kaplan-Meier-Schätzer

Wir gehen wieder von einem multiplikativen Regressionsmodell

$$\lambda_i(t) = Y_i(t) \cdot \alpha_i(t)$$

aus. Wir erinnern: α_i ist eine nichtnegative deterministische Funktion und $(Y_i(t))$ ist vorhersehbar. Jetzt geht es um die Schätzung von α_i bzw. von $A_i(t) = \int_0^t \alpha_i(s)\,ds$.

Wir nehmen an, daß

$$A_i(t) = \int_0^t \alpha_i(s)\,ds < \infty, \;\; i \in E, \;\; t \in I.$$

Normalerweise schätzt man zunächst letzteres Integral und verwendet dann Glättungsverfahren, um Schätzer für α selbst zu erhalten. Wir gehen dabei heuristisch vor. Die Doob-Meyer Zerlegung hat die Gestalt

$$N_i(t) = M_i(t) + \int_0^t Y_i(s)\alpha_i(s)\,ds$$

oder

$$N_i(t+h) - N_i(t) = \Big(M_i(t+h) - M_i(t)\Big) + \int_t^{t+h} Y_i(s)\alpha_i(s)\,ds.$$

Dies legt die intuitiv einleuchtende Schreibweise

$$dN_i(t) = dM_i(t) + Y_i(t)\alpha_i(t)\,dt$$

nahe (welche sich mit Hilfe der stochastischen Analysis rechtfertigen läßt, vgl. [16]).

Der Martingalanteil wurde als Rauschen interpretiert. Wie (M_{it}) kann auch (M_{it}/Y_{it}) als Rauschen angesehen werden: Weil (Y_{it}) vorhersehbar und nach Bemerkung 4.2.15 bezüglich $\mathcal{F}_{t-}$ adaptiert ist, gilt:

$$\boldsymbol{E}(dM(t)/Y_t|\mathcal{F}_{t-}) = \boldsymbol{E}(dM(t)|\mathcal{F}_{t-})/Y_t = 0,$$

d.h. (M_{it}/Y_{it}) ist ein Martingal. Somit haben wir die Zerlegung

$$\frac{dN_i(t)}{Y_i(t)} = \underbrace{\frac{dM_i(t)}{Y_i(t)}}_{\text{Rauschen}} + \alpha_i(t)\,dt$$

und ein natürlicher Schätzer für $A_i(t) = \int_0^t \alpha_i(s)\, ds$ ist gegeben durch

$$\int_0^t Y_i^{-1}(s)\, dN_i(s).$$

Schließlich wird der Fall $Y_i(s) = 0$ durch die Definitionen

$$J_i(t) = \mathbf{1}_{\{Y_i(t)>0\}}, \quad \hat{A}_i(t) = \int_0^t \frac{J_i(s)}{Y_i(s)}\, dN_i(s)$$

korrekt berücksichtigt, wobei die Konvention $0/0 := 0$ getroffen wird.

Da jeder Pfad eine simple Sprungfunktion mit Sprüngen zu den Stoppzeiten $\tau_{i1} < \tau_{i2} < \cdots$ ist, kann man das Integral durch eine Summe ausdrücken und erhält den Schätzer

$$\hat{A}_i(t) = \sum_{\{j:\tau_{ij}\leq t\}} J_i(\tau_{ij}) Y_i(\tau_{ij})^{-1}.$$

Die Pfade $t \mapsto \hat{A}_i(t)$ sind wachsende, rechtsstetige Treppenfunktionen mit dem Zuwachs $Y_i(\tau_{ij})^{-1}$ im Sprung von (N_{it}) zur Zeit τ_{ij}. $\hat{A}_i(t)$ wird *Nelson-Aalen Schätzer* genannt.

Aus dem Nelson-Aalen-Schätzer für α im multiplikativen Intensitätsmodell läßt sich über die Beziehung

$$\lambda(t) = \frac{f(t)}{\boldsymbol{S}(t)}, \quad \boldsymbol{S}(t) = 1 - \boldsymbol{F}(t),$$

ein Schätzer für die Verteilungsfunktion $\boldsymbol{F}$ bzw. für die Survivalfunktion $\boldsymbol{S}$ der zufälligen Zeit T ableiten, der *Kaplan-Meier Schätzer* . Für stetige Dichte f ist nämlich

$$\ln(\boldsymbol{S}(t))' = -\lambda(t),$$

d.h.

$$\int_0^t \lambda(s)\, ds = -\ln(\boldsymbol{S}(t)),$$

und deshalb

$$\boldsymbol{S}(t) = \exp\left(-\int_0^t \lambda(s) ds\right) = \exp(-\Lambda_t).$$

Da der Nelson-Aalen Schätzer die kumulative Hazardrate schätzt, ist

$$\boldsymbol{S}^*(t) = \exp(-\hat{A}_t)$$

ein Schätzer für $\boldsymbol{S}$, beziehungsweise $1 - \boldsymbol{S}^*$ ein Schätzer für $\boldsymbol{F}$. Der Kaplan-Meier-Schätzer $\hat{\boldsymbol{S}}$ ergibt sich durch die Approximation

$$e^{-u} \sim 1 - u,$$

für kleine u, d.h.

$$\exp(-\hat{A}_t) = \prod_{k:\tau_k \leq t} \exp\Big(- J(\tau_k)/Y(\tau_k)\Big) \sim \prod_{k:\tau_k \leq t} \left(1 - \frac{J(\tau_k)}{Y(\tau_k)}\right).$$

Zusammenfassend definiert man den Kaplan-Meier-Schätzer als

$$\hat{\boldsymbol{S}}_t = \prod_{k=\tau_k \leq t} \left(1 - \frac{J(\tau_k)}{Y(\tau_k)}\right) = \bar{\prod_{s \leq t}} \left(1 - \frac{\Delta N_s}{Y_s}\right),$$

wobei $\bar{\prod}_{s \leq t}$ eine akzeptiert symbolische Schreibweise ist mit $\Delta N_t = N_t - N_{t-}$.

Mehr über die Theorie erfährt man in [2].

4.5.2 Schätzung im multiplikativen Regressionsmodell

Das multiplikative Regressionsmodell hat die Gestalt

$$\lambda_{it}(\beta) = Y_{it}\gamma(t)\alpha(\beta' X_{it}).$$

Zu schätzen sind also die Funktion γ und der Parametervektor β.

Gegeben sei nun eine Realisation $N_{is}, Y_{is}, X_{is},\ \ s \leq t$, der beobachtbaren Prozesse. Wir motivieren heuristisch die Gestalt einer 'partial Likelihoodfunktion' zur (parametrischen) Schätzung von β. Auf den nichtparametrischen Teil gehen wir später ein. Zunächst betrachten wir die bedingte Verteilung des Typs zum n-ten Sprung:

$$\boldsymbol{P}_{n,t}(i) = \boldsymbol{P}(\xi_n = i | \tau_1, \xi_1, \ldots, \tau_{n-1}, \xi_{n-1}, \tau_n = t).$$

Nach Brémaud, S. 33 gilt

$$\boldsymbol{P}_{n,t}(i) = \frac{\lambda_{it}}{\lambda_t},\ \ \lambda_t = \sum_{i \in E} \lambda_{it}.$$

Die Idee des Beweises wird an folgender heuristischen Überlegung klar: Bezeichne $\mathcal{B}_{n-1}$ die Bedingung $\tau_1, \xi_1, \ldots, \tau_{n-1}, \xi_{n-1}$. Dann ist

$$\begin{aligned}\boldsymbol{P}_{n,t}(i) &= \boldsymbol{P}(\xi_n = i | \mathcal{B}_{n-1}, \tau_n = t) \\ &= \frac{\boldsymbol{P}(dN_{it} = 1 | \mathcal{B}_{n-1})}{\boldsymbol{P}(dN_t = 1 | \mathcal{B}_{n-1})} = \frac{\lambda_{ti}}{\lambda_t}\end{aligned}$$

nach Theorem 4.3.6. Ein rigoroser Beweis findet sich in [5], I.4.T12.

Seien $N_t = n$, $\tau_{n+1} = t$, und $\tau_1, \ldots, \tau_n$ gegeben. Dann ist $\boldsymbol{P}_{n,t}$ vom Typ

$$\boldsymbol{P}_{n+1} = \boldsymbol{P}(\xi_{n+1}|\xi_1, \ldots, \xi_n)$$

und man hat die Faktorisierung

$$\begin{aligned}
& \boldsymbol{P}(\xi_1 = i_1, \ldots, \xi_{N_t} = i_{N_t}) \\
= \; & \boldsymbol{P}(\xi_1 = i_1)\boldsymbol{P}(\xi_2 = i_2|\xi_1 = i_1) \cdot \ldots \cdot \boldsymbol{P}(\xi_{N_t} = i_{N_t}|\xi_1 = i_1, \ldots, \xi_{N_t-1} = i_{N_t-1}) \\
= \; & \prod_{k=1}^{N_t} \boldsymbol{P}_{k,\tau_k}(i_k).
\end{aligned}$$

Als Funktion von β wäre dies die Likelihoodfunktion L_t; wir setzen

$$L_t = \prod_{k=1}^{N_t} \boldsymbol{P}_{k,\tau_k}(\xi_k) = \prod_{k=1}^{N_t} \frac{\lambda_{\xi_k,\tau_k}}{\lambda_{\tau_k}}.$$

Für das multiplikative Regressionsmodell

$$\lambda_{it}(\beta) = Y_{it}\gamma(t)\alpha(\beta' X_{it})$$

ergibt sich

$$\begin{aligned}
L_t(\beta) &= \prod_{k=1}^{N_t} Y_{\xi_k,\tau_k} \cdot \frac{\alpha(\beta' X_{\xi_k,\tau_k})}{\sum_j Y_{j,\tau_k}\alpha(\beta' X_{j\tau_k})} \\
&= \prod_{k=1}^{N_t} Y_{\xi_k,\tau_k} \cdot \frac{\alpha(\beta' X_{\xi_k,\tau_k})}{S_{\tau_k}(\beta)},
\end{aligned}$$

wobei

$$S_t(\beta) = \sum_{i \in E} Y_{it}\alpha(\beta' X_{it}).$$

Für die log-Likelihood $l_t(\beta) = \ln L_t(\beta)$ bekommt man dann

$$\begin{aligned}
l_t(\beta) &= \sum_{k=1}^{N_t} \ln \alpha(\beta' X_{\xi_k,\tau_k}) - \ln S_{\tau_k}(\beta) + R(Y) \\
&= \sum_{i \in E} (\int_0^t \ln \alpha(\beta' X_{is})\, dN_{is}) - \int_0^t \ln S_s(\beta)\, dN_s + R(Y),
\end{aligned}$$

wobei R weder von β noch von γ abhängt.

Speziell für das Coxsche Regressionsmodell ergibt sich (bei Weglassen irrelevanter Terme)

$$\begin{aligned} l_t(\beta) &= \beta' \sum_i \int_0^t X_{i,s}\, dN_{is} - \int_0^t \ln S_s(\beta)\, dN_s \\ &= \beta' \sum_{k=1}^{N_t} X_{\xi_k,\tau_k} - \sum_{k=1}^{N_t} \log S_{\tau_k}(\beta), \end{aligned}$$

wobei

$$S_t(\beta) = \sum_{i \in E} Y_{it} \exp(\beta' X_{it}).$$

Wie üblich bildet man den Scorevektor $U_t(\beta) = dl_t(\beta)/d\beta$ und löst $U_t(\beta) = 0$. Ferner ist die Hessematrix der gemischten zweiten partiellen Ableitungen von l_t von Interesse.

Wir betrachten zunächst

$$S_t(\beta) = \sum_i Y_{it} e^{\beta' X_{it}}$$

und berechnen

$$\frac{\partial}{\partial \beta_j} S_t(\beta) = \sum_i Y_{it} e^{\beta' X_{it}} X_{it}^{(j)};$$

wobei $X_{it}^{(j)}$ die j-te Komponente von X_{it} bezeichnet. Weiter gilt

$$\frac{\partial^2}{\partial \beta_k \partial \beta_j} S_t(\beta) = \sum_i Y_{it} e^{\beta' X_{it}} X_{it}^{(j)} X_{it}^{(k)}.$$

Damit ergibt sich

$$\begin{aligned} \frac{\partial}{\partial \beta_j} \ln S_t(\beta) &= \frac{\frac{\partial}{\partial \beta_j} S_t(\beta)}{S_t(\beta)} = \frac{\sum_i Y_{it} \exp(\beta' X_{it}) X_{it}^{(j)}}{\sum_i Y_{it} \exp(\beta' X_{it})} \\ \frac{\partial}{\partial \beta_k \partial \beta_j} \ln S_t(\beta) &= \frac{\frac{\partial^2}{\partial \beta_k \partial \beta_j} S_t(\beta) \cdot S_t(\beta) - \frac{\partial}{\partial \beta_j} S_t(\beta) \frac{\partial}{\partial \beta_k} S_t(\beta)}{S_t(\beta)^2} \\ &= \frac{\frac{\partial^2}{\partial \beta_k \partial \beta_j} S_t(\beta)}{S_t(\beta)} - \frac{\frac{\partial}{\partial \beta_j} S_t(\beta) \frac{\partial}{\partial \beta_k} S_t(\beta)}{S_t(\beta)^2}. \end{aligned}$$

Führt man formal auf E die diskrete Verteilung $p = (p_1, \ldots, p_n)$ mit

$$p_q = Y_{qt} \exp(\beta' X_{qt}) \Big/ \sum_i Y_{it} \exp(\beta' X_{it})$$

ein und Erwartungswert $\boldsymbol{E}_p$ sowie Varianz $\boldsymbol{V}_p$, so bekommt der Scorevektor U die Gestalt

$$\begin{aligned} U_t(\beta) &= \sum_i \int_0^t X_{is}\, dN_{it} - \int_0^t \frac{\sum_i Y_{it} \exp(\beta' X_{is}) X_{is}}{S_s(\beta)}\, dN_s \\ &= \sum_i \int_0^t X_{is}\, dN_{is} - \int_0^t \boldsymbol{E}_p(X_s)\, dN_s. \end{aligned}$$

Für die Hessematrix ergibt sich

$$\begin{aligned} W_t(\beta) &= -\int_0^t \Big(\frac{\partial}{\partial \beta_k \partial \beta_j} \ln S_s(\beta) \Big)_{kj} dN_s \\ &= -\int_0^t \Big\{ \boldsymbol{E}_p(X_s X_s') - \boldsymbol{E}_p(X_s) \boldsymbol{E}_p(X_s') \Big\}\, dN_s \\ &=: \int_0^t \boldsymbol{V}_p(X_s)\, dN_s. \end{aligned}$$

Formal ist dies das Negative einer Kovarianz. Somit ist W_t negativ semidefinit und l konkav. Die Größen $\boldsymbol{E}_p(X_s)$ und $\boldsymbol{V}_p(X_s)$ sind Erwartungswert und Kovarianz der Kovariablen, wenn die Individuen i mit Wahrscheinlichkeiten

$$p_i \propto Y_i \exp(\beta' X_i),$$

gewählt werden, also proportional zu ihrer Intensität.

4.5.3 Schätzung der Baseline-Hazard-Rate.

Wir schätzen zunächst die kumulative Baseline-Rate $\mathit{\Gamma}(t) = \int_0^t \gamma(s)\, ds$. Genau wie den Nelson-Aalen Schätzer im allgemeinen multiplikativen Intensitätsmodell leitet man einen Nelson-Aalen Schätzer für $\mathit{\Gamma}$ (bei gegebenem β) her; es ergibt sich

$$\begin{aligned} \hat{\mathit{\Gamma}}_t(\beta) &= \int_0^t \frac{J_s(s)}{S_s(\beta)} dN_s, \\ \text{wobei } J(s) &= \mathbf{1}_{\{Y_s > 0\}},\ Y_s = \sum_i Y_{is}. \end{aligned}$$

Setzen wir den Maximumlikelihoodschätzer $\hat{\beta}$, d.h. eine Lösung von $U(\beta) = 0$ ein, so erhalten wir den *Breslow Schätzer*.

$$\hat{\Gamma}(\hat{\beta}) = \int_0^t \frac{J(s)}{S_s(\hat{\beta})} \, dN_s$$

für die kumulierte Baseline Intensität Γ.

Eigentlich ist man an γ, also der 'Ableitung' von Γ interessiert. Für stetiges γ hat man

$$\begin{aligned} \gamma(t) &= \lim_{h \to 0} \frac{1}{h}(\Gamma(t + h/2) - \Gamma(t - h/2)) \\ &\sim (\Gamma(t + h/2) - \Gamma(t - h/2))/h \\ &= \int \frac{1}{h} \mathbf{1}_{(t-h/2,\, t+h/2)}(u) \, d\Gamma(u) \end{aligned}$$

Definiert man die Kernfunktion

$$K = \mathbf{1}_{[-1/2,\, 1/2]}$$

so ergibt sich

$$\gamma(t) \sim \int \frac{1}{h} K\left(\frac{t-s}{h}\right) \, d\Gamma(s).$$

Der entsprechende Schätzer wird zu

$$\hat{\gamma}(t) = \int \frac{1}{h} \left(K\left(\frac{t-s}{h}\right)\right) d\, \hat{\Gamma}_s(\hat{\beta}).$$

Da $\hat{\Gamma}$ Sprungfunktion ist, γ im allgemeinen aber glatt, folgt man einer Idee aus der Kerndichteschätzung und wählt glatte Kerne, z.B. symmetrische glockenförmige Dichten, die außerhalb $[-1, 1]$ verschwinden. Zusammenfassend bekommt man

$$\hat{\gamma}(t) = \int \frac{1}{h} K\left(\frac{t-s}{h}\right) \, d\hat{\Gamma}_s(\hat{\beta}) = \frac{1}{h} \sum_k K\left(\frac{t-\tau_k}{h}\right) \frac{1}{S_{\tau_k}(\hat{\beta})}.$$

Dabei erstreckt sich die Summation nur über Zeitpunkte in $[t - h, t + h]$, da wir kompakten Träger für K angenommen haben. Neben dem uniformen Kern$K = \mathbf{1}_{[-1,1]}/2$ sind der *Epanechnikov Kern*

$$K_E(u) = \frac{3}{4}(1 - u^2), \;\; -1 \leq u \leq 1,$$

oder der *Biweigth Kern*

$$K_B(u) = \frac{15}{16}(1-u^2)^2,\ -1 \le u \le 1$$

standard.

Für die statistischen Eigenschaften all dieser Schätzer verweisen wir nochmals und nachdrücklich auf [2], z.B. Kapitel 4.

Bemerkung 4.5.1 Diese Art der Glättung ist z.B. aus der Analysis wohlbekannt, wo man z.B. die Faltung mit sehr glatten Kernen benutzt, um nicht differenzierbare Funktionen differenzierbar zu machen. Es sei hier einem naheliegenden Trugschluß vorgebeugt: Faßt man K als Dichte auf, so wächst die Fähigkeit zu glätten *nicht* mit der Varianz von K. So sind die Varianzen des uniformen, Epanechnikov und Biweight Kernes gleich $1/3, 1/5$ und $1/7$. Abb. IV 2.1 in [2] zeigt jedoch, daß der Grad der Glättung wächst.

Beispiel 4.5.2 Seien $T_1, \ldots, T_n$ unabhängige positive Zufallsvariablen mit cádlàg-Dichten f_i. Seien

$$N_{it} = \mathbf{1}_{\{T_i \le t\}}$$

die zugehörigen (Ein-Episoden) Zählprozesse, zusammengefaßt im multivariaten Zählprozeß (N_{it}). Nach Beispiel 4.4.2 ist

$$\lambda_{it} = \mathbf{1}_{\{t \le T_i\}} \alpha_i(t)$$

mit $\alpha_i(t) = f_i(t)/\boldsymbol{S}_i(t)$. Dies paßt sich in das Coxsche Regressionsmodell ein, indem man definiert

$$Y_{it} = \mathbf{1}_{\{t \le T_i\}} (= 1 - N_{it-})$$

und

$$\alpha_i(t) = \gamma(t) \exp(\beta' X_{it}).$$

Den Zusammenhang mit der ursprünglichen Definition eines multivariaten Zählprozesses stellt man so her: Seien

$$\tau_1, \ldots, \tau_n = T_{(1)}, \ldots, T_{(n)}$$

die zu $T_1, \ldots, T_n$ gehörigen geordneten Größen, welche Stoppzeiten sind falls alle T_n Stoppzeiten waren. Die Marken sind dann gegeben durch

$$\xi_k = i \ \text{ auf } \ \{\tau_k = T_i\}.$$

Zu der neuen Zeitskala mit Marken gehören die Regressoren

$$X_{(k)} = X_{\xi_k \tau_k},$$

explizit gegeben durch

$$X_{(k)}(\omega) = X_{iT_i}(\omega) \text{ falls } \tau_k(\omega) = T_i(\omega).$$

Dann ist

$$S_{\tau_k}(\beta) = \sum_i \mathbf{1}_{\{\tau_k \le T_i\}} \exp(\beta' X_{i\tau_k}) = \sum_{i \in R_{(k)}} \exp(\beta' X_{i\tau_k}),$$

wobei

$$R_{(k)} = \{i : T_i \ge T_{(k)}\}$$

Die partial Likelihood ist dann gegeben durch

$$L_t(\beta) = \prod_{k=1}^{N_t} \frac{\exp(\beta' X_{(k)})}{\sum_{i \in R_{(k)}} \exp(\beta' X_{\tau_k})}$$

bzw. die partial Loglikelihood durch

$$l_t(\beta) = \beta' \sum_{k=1}^{N_t} X_{(k)} - \sum_{k=1}^{N_t} \ln S_{\tau_k}(\beta).$$

Der Nelson-Aalen Schätzer für die kumulative baseline Hazardfunktion ist

$$\hat{\Gamma}_t(\hat{\beta}) = \sum_{k=1}^{N_t} S_{\tau_k}(\hat{\beta})^{-1} = \sum_{k=1}^{N_t} \Big(\sum_{i \in R_{(k)}} \exp(\hat{\beta}' X_{iT_{(k)}}) \Big).$$

Die Ausdehnung auf zensierte Daten erfolgt analog zum Übergang von Beispiel 4.4.2 zu Beispiel 4.4.3.

Beispiel 4.5.3 Man startet mit dem in Beispiel 4.4.2 festgelegten Rahmen. Hinzu kommen die zufälligen Zeiten $U_1, \ldots, U_n$ der Zensur und daraus abgeleitet die Indikatorvariablen

$$D_i = \mathbf{1}_{\{T_i \le U_i\}};$$

beobachtbar sind also nur die Zeiten $T_i \wedge U_i$ und die D_i. Der zugeordnete beobachtbare multivariate Zählprozeß ist gegeben durch

$$N_{it} = \mathbf{1}_{\{T_i \le t, D_i = 1\}} = \mathbf{1}_{\{T_i \le t \wedge U_i\}}.$$

Aus Beispiel 4.4.2 lesen wir den Intensitätsprozeß als

$$\lambda_{it} = \mathbf{1}_{\{t \le T_i \wedge U_i\}} \alpha_i(t) =: Y_{it} \alpha_i(t)$$

ab. Über die Coxsche Parametrisierung geht dies in

$$\alpha_{it} = \eta(t) \exp(\beta' X_{i,\tau_k})$$

über. An die Stelle der Ordnungsstatistiken $T_{(1)} \le \cdots \le T_{(n)}$ der $T_1, \ldots, T_N$ tritt nun die Ordnungsstatistik

$$T_{(l_1)} < \ldots < T_{(l_m)},\ m \le n,$$

der T_j mit $D_j = 1$, der *beobachteten* geordneten Zeiten. Seien

$$R_{(k)} = \{i : T_i \wedge U_i \ge T_{(l_k)}\}$$

die *Risikomenge* zur Zeit $T_{(l_k)}$ und

$$S_{\tau_k}(\beta) = \sum_i \mathbf{1}_{\{\tau_k \le T_i \wedge U_i\}} \exp(\beta' X_{i\tau_k}) = \sum_{i \in R_{(k)}} \exp(\beta' X_{iT_{(l_k)}}),$$

$\tau_k = T_{(l_k)}$. Damit bekommt man analog wie vorher

$$L_t(\beta) = \prod_{i=1}^{N_t} \frac{\exp(\beta' X_{(l_k)})}{\sum_{i \in R_{(k)}} \exp(\beta' X_{iT_{(l_k)}})},$$

wobei $X_{(l_k)} = X_{jT_j}$ falls $\tau_k = T_{(l_k)} = T_j$, d.h. $X_{(l_k)} = X_{\xi_k \tau_k}$. Die log-Likelihood schreibt sich

$$l_t(\beta) = \beta' \sum_{k=1}^{N_t} X_{(l_k)} - \sum_{k=1}^{N_t} \ln S_{\tau_k}(\beta).$$

Literaturverzeichnis

[1] P.K. Andersen and Ø. Borgan. Counting process models for life history data: A review. *Scand. J. Statist.*, 12:97–158, 1985. with discussion.

[2] P.K. Andersen, Ø. Borgan, R.D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes.* Springer Series in Statistics. Springer-Verlag, New York, Berlin, Heidelberg ..., 1993.

[3] H. Bauer. *Wahrscheinlichkeitstheorie, 4. Auflage.* Walter de Gruyter, Berlin, New York, 1991.

[4] D.P. Bertsekas. *Dynamic Programming and Stochastic Control*, volume 125 of *Mathematics in Science and Engineering.* Academic Press, New York, San Francisco, London, 1976.

[5] Brémaud. *Point Processes and Queues. Martingale Dynamics.* Springer Series in Statistics. Springer-Verlag, New York, Heidelberg, Berlin, 1981.

[6] C.K. Chui and G. Chen. *Kalman Filtering with Real-Time Applications*, volume 17 of *Springer Series in Information Sciences.* Springer-Verlag, Berlin, Heidelberg, New York etc., 1987.

[7] P.J. Diggle, K.-Y. Liang, and S.L. Zeger. *Analysis of Longitundinal Data*, volume 13 of *Oxford Statistical Science Series.* Clarendon Press, Oxford, 1994.

[8] A. Einstein. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Ann. Physik*, 17:549–560, 1905.

[9] L. Fahrmeir, A. Hamerle, and G.Tutz, editors. *Multivariate statistische Verfahren.* Walter de Gruyter, Berlin, New York, 1996. 2., überarbeitete Auflage.

[10] L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer Series in Statistics. Springer-Verlag, New York, Berlin, 1994.

[11] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEE Trans. PAMI*, 6:721–741, 1984.

[12] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter (eds.). *Markov Chain Monte Carlo in Practice.* Interdisciplinary Statistics. Chapman & Hall, London, Weinheim, New York, Tokyo, Melbourne, Madras, 1996.

[13] H. Pruscha. *Angewandte Methoden der Mathematischen Statistik.* Teubner Skripten zur mathematischen Statistik. B.G. Teubner, Stuttgart, second edition, 1996.

[14] H. Pruscha. Punkt- und Zählprozesse. Unveröffentlichtes Vorlesungsskript, 1998.

[15] D.J. Spiegelhalter, N.G. Best, W.R. Gilks, and H. Inskip. Hepatitis b: a case study in MCMC methods. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 21–43. Chapman & Hall, London, Weinheim, New York, Tokyo, Melbourne, Madras, 1996.

[16] H. von Weizsäcker and G. Winkler. *Stochastic Integration.* Vieweg Advanced Lectures in Mathematics. Friedrich Vieweg & Sohn, Braunschweig, Wiesbaden, 1990.

[17] G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*, volume 27 of *Applications of Mathematics.* Springer Verlag, Berlin, Heidelberg, New York, 1995.

Abbildungsverzeichnis

Tabellenverzeichnis

Index