

# Vorlesung: Räumliche Statistik WS 04/05

basierend auf einer Mitschrift vom SS 2001  
durch Marcel Dettling und Christian Sangiorgio

Hansruedi Künsch  
Seminar für Statistik  
ETH Zürich

Dezember 2005



# Inhaltsverzeichnis



# Kapitel 1

## Einführung

Die folgenden Beispiele zeigen einige typische Fragestellungen der räumlichen Statistik und der Bildanalyse:

**Cadmiumgehalt im Oberboden** (aus einer unpublizierten Arbeit von Andreas Parritz, Institut fuer terrestrische Ökologie, ETHZ). Das Ziel ist die Interpolation des Cadmiumgehalts im Oberboden aus einem Netz von Beobachtungspunkten. Man will den Cadmiumgehalt (mit Genauigkeitsangabe) an einer Stelle vorhersagen, wo keine Messung gemacht wurde.

Methoden: *Geostatistik, Kriging*.

**Räumliche Verteilung von Krankheiten** (siehe z.B. Besag, York und Mollié, Ann. Inst. Statist. Math. 43, 1991, 1–59). Das Ziel ist, den Einfluss von erklärenden Variablen zu untersuchen. Dazu betrachtet man auch die räumliche Korrelation der Residuen, was Hinweise auf im Modell nicht erfasste Variable liefern kann.

Methoden: *Bayes-Verfahren, Markov-Felder*.

**Feldversuche** (siehe z.B. Besag und Higdon, J. Royal Statist. Soc. B 61, 1999, 691–746). Mit Hilfe der räumlichen Korrelation der Residuen wird versucht, die Bodenbeschaffenheit (latente Variable) aus den Daten zu schätzen.

Methoden: *Bayes-Verfahren, Markov-Felder*.

**Satellitenbilder (low level vision)** (siehe z.B. Geman und Geman, IEEE Trans. Pattern Anal. Machine Intell. 6, 1984, 721–741). Das Ziel ist, die “Verschmierung” (“blurring”) rückgängig zu machen. Das Rauschen im Bild soll allerdings unter Beibehaltung der wesentlichen Strukturen (Kanten, etc.) entfernt werden.

Methoden: *Bayes-Verfahren, Wavelets*.

**Objekterkennung (high level vision)** (siehe z.B. Fleuret und Geman, Intern. J. Computer Vision 41, 2001, 85–107). Das Ziel ist, in Bildern vorgängig definierte Objekte zu erkennen. Man benützt die Stochastik zur Formalisierung von Mehrdeutigkeit auf lokalem Niveau.

Methoden: *Klassifikation, Feature-Extrahierung*.

**Standorte von Bäumen** (siehe z.B. Ogata und Tanemura, J. Royal Statist. Soc. B 46, 1984, 496–518). Das Ziel ist, zu untersuchen ob die Bäume regelmässig, bzw. zufällig oder geclustert wachsen.

Methoden: *(markierte) Punktmuster*.

**Oberflächen, Texturen, 3-dimensionale Strukturen** (siehe z.B. Ohser und Mücklich, *Statistical Analysis of Microstructures in Materials Science*, Wiley, 2000). Das Ziel ist die Beschreibung, Klassifizierung und Simulation 3-dimensionaler Eigenschaften aus 2-dimensionalen Schnitten.

Methoden: *Stochastische Geometrie, Markovfelder.*

# Kapitel 2

## Gauss'sche Modelle, Geostatistik

### 2.1 Begriffe und Grundlagen

Wir betrachten hier Modelle, wo man potentiell an jedem beliebigen Punkt im  $\mathbb{R}^d$  eine Messgrösse hat. Dazu führen wir den Begriff des Zufallsfeldes ein. Danach werden wichtige Spezialfälle von Zufallsfeldern definiert und diskutiert.

**Definition 2.1** *Ein Zufallsfeld ist eine Kollektion von Zufallsvariablen  $(Z(x))_{x \in \mathbb{R}^d}$  auf einem gemeinsamen Wahrscheinlichkeitsraum  $(\Omega, \mathcal{F}, P)$ .*

Ein Zufallsfeld ist also eine Abbildung  $Z : \mathbb{R}^d \times \Omega \rightarrow \mathbb{R}$ ,  $(x, \omega) \rightarrow Z(x, \omega)$ . Für festes  $x$  ist  $Z(x, \cdot)$  eine Zufallsvariable, und für festes  $\omega$  ist  $Z(\cdot, \omega)$  eine Funktion von  $\mathbb{R}^d$  nach  $\mathbb{R}$ . Im Fall  $d = 1$  wird  $x$  meist als Zeit aufgefasst, und man spricht dann von einem Zufallsprozess.

Mathematisch ist die Konstruktion eines Zufallsfelds, d.h. die Konstruktion eines Wahrscheinlichkeitsraums  $(\Omega, \mathcal{F}, P)$  und der zugehörigen Abbildung  $Z$  eine anspruchsvolle Aufgabe. Wir gehen in dieser Vorlesung aber nicht näher darauf ein. Alles, was wir verwenden werden, sind die endlich-dimensionalen Verteilungen, d.h. die Wahrscheinlichkeiten  $P[Z(x_1) \leq a_1, \dots, Z(x_n) \leq a_n]$  für  $n \in \mathbb{N}$  und  $x_j \in \mathbb{R}^d$ .

Am einfachsten sind die Gauss'schen Zufallsfelder:

**Definition 2.2** *Ein Zufallsfeld heisst Gauss'sch, falls alle endlich-dimensionalen Verteilungen Gauss'sch sind.*

Ein Gauss'sches Zufallsfeld ist festgelegt durch den Erwartungswert  $m(x) = \mathbf{E}[Z(x)]$  und die Kovarianz  $C(x, x') = \text{Cov}(Z(x), Z(x'))$ .

Im Unterschied zur Statistik von i.i.d. Modellen beobachten wir in der räumlichen Statistik meist nur eine einzige Realisierung  $z$  von  $Z$  an endlich vielen Punkten  $x_1, \dots, x_n$ . Ohne zusätzliche Annahmen ist es daher nicht möglich, Rückschlüsse auf die zu Grunde liegende Verteilung zu ziehen. Die einfachste solche Annahme ist die **Stationarität**. Dies bedeutet, dass sich die endlich-dimensionalen Verteilungen nicht ändern, wenn man alle Messpunkte um den gleichen Vektor verschiebt:  $(Z(x_1), \dots, Z(x_n))$  und  $(Z(x_1+h), \dots, Z(x_n+h))$  sind identisch verteilt für alle  $n$  und alle  $x_j$ . Weil im Gauss'schen Fall alle endlich-dimensionalen Verteilungen durch Erwartungswert und Kovarianz bestimmt sind, vereinfacht sich diese Bedingung.

**Definition 2.3** Ein Gauss'sches Zufallsfeld heisst stationär, falls der Erwartungswert konstant ist, d.h.  $m(x) \equiv m$ , und falls die Kovarianz invariant unter Verschiebungen ist,  $C(x+h, x'+h) = C(x, x')$ . In diesem Fall gilt  $C(x, x') = C(x-x')$  und  $C(\cdot)$  heisst Autokovarianzfunktion.

Wenn das Zufallsfeld nicht Gauss'sch ist, dann heisst die Invarianz der ersten beiden Momente schwache Stationarität, während die Verschiebungsinvarianz der endlich-dimensionalen Verteilungen als strikte Stationarität bezeichnet wird.

Die folgende Eigenschaft ist schwächer als die Stationarität.

**Definition 2.4** Ein Gauss'sches Zufallsfeld heisst intrinsisch, falls der Erwartungswert konstant ist, d.h.  $m(x) \equiv m$ , und falls  $\frac{1}{2} \text{Var}(Z(x) - Z(x'))$  nur vom Abstand  $x - x'$  abhängt. In diesem Fall nennen wir  $\gamma(h) = \frac{1}{2} \text{Var}(Z(x+h) - Z(x))$  das Semivariogramm.

Bei einem Gauss'schen intrinsischen Zufallsfeld haben also  $Z(x) - Z(x')$  und  $Z(x+h) - Z(x'+h)$  die gleiche Verteilung. Die Verteilung von Zuwächsen ändert sich nicht bei Verschiebungen.

Wie die folgende Rechnung zeigt, folgt aus der Eigenschaft "stationär" die Eigenschaft "intrinsisch":

$$\begin{aligned} \text{Var}(Z(x+h) - Z(x)) &= \text{Var}(Z(x+h)) + \text{Var}(Z(x)) - 2 \text{Cov}(Z(x), Z(x+h)) \\ &= C(0) + C(0) - 2C(h) = 2(C(0) - C(h)). \end{aligned}$$

Im stationären Fall ist also insbesondere  $\gamma(h) = C(0) - C(h)$ .

Umgekehrt folgt aber aus der Eigenschaft "intrinsisch" nicht die Eigenschaft "stationär". Wir betrachten als Gegenbeispiel den Fall, wo  $d = 1$  ist und  $Z(x)$  gleich der Brown'schen Bewegung ist. In diesem Fall ist  $\mathbf{E}[Z(x)] = 0$ , aber  $\text{Var}(Z(x)) = |x|$ , und damit ist  $Z$  nicht stationär. Aus der Eigenschaft der unabhängigen Zuwächse folgt jedoch  $\text{Var}(Z(x+h) - Z(x)) = |h|$ , und damit ist  $Z$  intrinsisch. Bei Zeitreihen entspricht "intrinsisch" der Eigenschaft, dass die ersten Differenzen stationär sind.

Das nächste einfache Lemma wird im Folgenden eine wichtige Rolle spielen

**Lemma 2.1** a) Falls  $Z(x)$  stationär ist, dann gilt:

$$\text{Cov} \left( \sum_{j=1}^n \beta_j Z(x_j), \sum_{j=1}^n \beta'_j Z(x_j) \right) = \sum_{j,k=1}^n \beta_j \beta'_k C(x_k - x_j). \quad (2.1)$$

b) Falls  $Z(x)$  intrinsisch ist, dann gilt:

$$\text{Cov} \left( \sum_{j=1}^n \lambda_j Z(x_j), \sum_{j=1}^n \lambda'_j Z(x_j) \right) = - \sum_{j \neq k} \lambda_j \lambda'_k \gamma(x_k - x_j) \quad (2.2)$$

für alle  $n$ , für alle  $x_1, \dots, x_n$  und für alle  $\lambda_j, \lambda'_j$  mit  $\sum \lambda_j = \sum \lambda'_j = 0$ .

**Beweis:**

a) Rechenregel für die Kovarianz.

b) Wir benützen

$$\sum_{j=1}^n \lambda_j Z(x_j) = \sum_{j=1}^n \lambda_j (Z(x_j) - Z(x_1))$$

(und analog für die zweite Linearkombination). Daraus folgt

$$\text{Cov} \left( \sum \lambda_j Z(x_j), \sum \lambda'_j Z(x_j) \right) = \sum_{j,k} \lambda_j \lambda'_k \text{Cov} (Z(x_j) - Z(x_1), Z(x_k) - Z(x_1))$$

Mit Hilfe von

$$Z(x_j) - Z(x_k) = Z(x_j) - Z(x_1) + Z(x_1) - Z(x_k)$$

erhalten wir

$$2\gamma(x_j - x_k) = 2\gamma(x_j - x_1) + 2\gamma(x_k - x_1) - 2\text{Cov} (Z(x_j) - Z(x_1), Z(x_k) - Z(x_1))$$

und somit

$$\begin{aligned} \text{Cov} \left( \sum \lambda_j Z(x_j), \sum \lambda'_j Z(x_j) \right) &= \sum_{j,k} \lambda_j \lambda'_k (\gamma(x_j - x_1) + \gamma(x_k - x_1) - \gamma(x_j - x_k)) \\ &= 0 + 0 - \sum_{j \neq k} \lambda_j \lambda'_k \gamma(x_k - x_j). \end{aligned}$$

□

Neben der Invarianz unter Verschiebungen kann man auch noch Invarianz unter Drehungen betrachten.

**Definition 2.5** *Ein intrinsisches Gauss'sches Zufallsfeld heisst isotrop, falls  $\gamma(h) = \gamma(\|h\|)$ .*

## 2.2 Nichtparametrische Schätzung der ersten beiden Momente

Das Ziel in diesem Abschnitt ist die Schätzung des Erwartungswerts  $m$  und der Autokovarianzfunktion  $C(\cdot)$  bzw. des Semivariogramms  $\gamma(\cdot)$  aus Beobachtungen  $(z(x_1), \dots, z(x_n))$ . Zur Vereinfachung der Notation setzen wir ausserdem meist noch die Isotropie des zu Grunde liegenden Zufallsfelds voraus. Einfach und naheliegend sind nichtparametrische Schätzungen für  $m$ ,  $C(\cdot)$  und  $\gamma(\cdot)$ .

- 1) Für den Erwartungswert wird "klassisch" gemittelt, d.h.  $\hat{m} = \frac{1}{n} \sum_{i=1}^n Z(x_i)$ .
- 2) Für das Semivariogramm bestimmt man das Mittel über Paare mit ungefähr gleichen Abständen, d.h.  $\hat{\gamma}(\|h\|) = \text{Mittel von } \frac{1}{2}(Z(x_i) - Z(x_j))^2 \text{ über alle Paare } i, j \text{ für die } \|x_i - x_j\| \approx h$ .

Zur praktischen Durchführung werden meist Klassen von Paaren gebildet, die Beobachtungen mit ähnlichen Abständen enthalten (analog zu einem Histogramm). Dann wird für

jede Klasse das (gewöhnliche) Mittel gebildet. Stattdessen kann man aber auch mit einem Kernschätzer eine Glättung durchführen:

$$\hat{\gamma}(h) = \frac{\frac{1}{2} \sum_{i,j} (Z(x_i) - Z(x_j))^2 \cdot K(\|x_i - x_j\| - \|h\|)}{\sum_{i,j} K(\|x_i - x_j\| - \|h\|)}$$

Ein Nachteil dieses Vorgehens ist die grosse Variabilität der nichtparametrischen Schätzungen. Ein weiterer Nachteil ist, dass die so geschätzten Grössen für  $C(\cdot)$  und  $\gamma(\cdot)$  oft gar nicht einem möglichen Modell entsprechen. Gemäss Lemma

Um den Nachteilen der nichtparametrischen Schätzungen aus dem Weg zu gehen, passt man parametrische Modelle an mit Kovarianzfunktionen, von denen man weiss, dass sie positiv definit sind. Eine Liste solcher Modelle folgt im nächsten Abschnitt.

## 2.3 Parametrische Modelle für Kovarianz und Variogramm

Hier folgt zuerst ein Überblick über die wichtigsten parametrischen Modelle:

**“Weisses Rauschen”, “Nugget-Modell”** In diesem Modell wird vorausgesetzt, dass die Werte an verschiedenen Stellen unabhängig sind. Damit erhalten wir die Autokovarianzfunktion

$$C(h) = \begin{cases} \sigma^2, & \text{falls } h = 0 \\ 0, & \text{falls } h \neq 0 \end{cases} \quad (2.3)$$

**“Allgemeines exponentielles Modell”**

$$C(h) = \sigma^2 \cdot \exp\left(-\left(\frac{\|h\|}{\rho}\right)^\nu\right) \quad (2.4)$$

wobei  $0 < \nu \leq 2$  sein muss, damit  $C(\cdot)$  die im Lemma 1.1 bewiesenen Eigenschaften einer Autokovarianzfunktion erfüllt. Am häufigsten verwendet werden die Parameter  $\nu = 1$  bzw.  $\nu = 2$ , die unter der Bezeichnung “exponentielles Modell” bzw. “Gauss’sches Modell” bekannt sind.

**“Sphärisches Modell”**

$$C(h) = \sigma^2 \cdot \frac{|B(0,1) \cap B(h/\rho,1)|}{|B(0,1)|} \quad (2.5)$$

wo  $B(x,r)$  die Sphäre mit Zentrum  $x$  und Radius  $r$  in  $\mathbb{R}^d$  ist, und  $|\cdot|$  für das Volumen (bzw. die Fläche) dieser Sphäre steht.

**“Matérn-Modell”**

$$C(h) = \sigma^2 \cdot \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(2\sqrt{\nu}\frac{\|h\|}{\rho}\right)^\nu K_\nu\left(2\sqrt{\nu}\frac{\|h\|}{\rho}\right), \quad (2.6)$$

wo  $K_\nu(\cdot)$  die Besselfunktion und  $\Gamma(\cdot)$  die Gammafunktion ist.

**“Potenz-Modell”**

$$\gamma(h) = \sigma^2 \cdot \|h\|^\nu \quad (2.7)$$

mit  $0 < \nu < 2$ . Für  $\nu = 1$  und  $d = 1$  entspricht dieses Modell gerade der Brown’schen Bewegung.

Die unbekannt Parameter sind  $\sigma^2$ ,  $\rho$  und  $\nu$ . Dabei dient  $\sigma^2$  zur Skalierung von  $Z$  ( $Z \rightarrow \text{const} \cdot Z$ ),  $\rho$  dient zur Skalierung der Distanz ( $x \rightarrow \text{const} \cdot x$ ), und  $\nu$  ist ein Formparameter, der die Glattheit von  $C$  bzw.  $\gamma$  bei 0 regelt. Weil  $\gamma(h) = \mathbf{E} [(Z(x+h) - Z(x))^2] / 2$  und  $\gamma(0) = 0$ , steht das in direkter Beziehung mit der Glattheit des Zufallsfelds.

Insbesondere ist  $Z(x)$  stetig im quadratischen Mittel, d.h.  $Z(x') \rightarrow Z(x)$  in  $L_2$  für  $x' \rightarrow x$ , genau dann wenn  $\gamma(\cdot)$  bzw.  $C(\cdot)$  im Nullpunkt stetig sind. Beim Nugget-Modell ist also  $Z$  nicht stetig im quadratischen Mittel. Konvergenz in  $L_2$  ist schwächer als fast sichere Konvergenz, d.h. aus Stetigkeit im quadratischen Mittel folgt nicht, dass fast alle Realisationen  $Z(\cdot, \omega)$  stetig sind.

Das folgende Lemma zeigt den Zusammenhang zwischen der Differenzierbarkeit von  $Z$  und Differenzierbarkeit von  $\gamma$ .

**Lemma 2.2** *Ein intrinsisches Zufallsfeld  $Z$  ist differenzierbar im quadratischen Mittel, d.h. die Richtungsableitung in eine beliebige Richtung  $h$  existiert in  $L_2$ , genau dann wenn  $\gamma(\cdot)$  2-mal stetig differenzierbar ist. In diesem Fall gilt*

$$\text{Cov}(D_x Z(x'), D_x Z(x)) = D_x^2 \gamma(x - x').$$

**Beweis:** Falls  $(Z(x+th) - Z(x))/t$  in  $L_2$  konvergiert, dann konvergiert insbesondere das zweite Moment gegen das zweite Moment der Richtungsableitung. Also konvergiert

$$\mathbf{E} \left[ \left( \frac{Z(x+th) - Z(x)}{t} \right)^2 \right] = \frac{2\gamma(th)}{t^2}.$$

Damit ist die zweimalige Differenzierbarkeit bei Null zumindest plausibel. Für einen exakten Beweis und für die zweimalige Differenzierbarkeit an andern Stellen, braucht man die Spektraldarstellung. Wir verzichten darauf.

Für die Umkehrung der Behauptung zeigt man, dass  $t \rightarrow \frac{Z(x+th) - Z(x)}{t}$  eine Cauchyfolge in  $L_2$  ist, d.h zu jedem  $\varepsilon > 0$  existiert ein  $\delta$ , so dass

$$\begin{aligned} & \mathbf{E} \left[ \left( \frac{Z(x+th) - Z(x)}{t} - \frac{Z(x+sh) - Z(x)}{s} \right)^2 \right] \\ &= \frac{2\gamma(th)}{t^2} + \frac{2\gamma(sh)}{s^2} - 2 \frac{\gamma(th) + \gamma(sh) - \gamma((t-s)h)}{ts} \leq \varepsilon \end{aligned}$$

für alle  $t, s$  mit  $|t|, |s| \leq \delta$ . Die beiden ersten Terme konvergieren je gegen  $h^T D_x^2 \gamma(0) h$ . Für den dritten Term setzen wir  $f(t, s) = \gamma(th) - \gamma((t-s)h)$ ,  $f_1 = \partial f / \partial t$  und  $f_{12} = \partial^2 f / \partial t \partial s$ . Weil  $f_1(t, 0) = 0$ , folgt dann mit dem Zwischenwertsatz dass

$$\frac{f(t, s) - f(0, s)}{ts} = \frac{f_1(\theta_1 t, s)}{s} = f_{12}(\theta_1 t, \theta_2 s).$$

Weil  $\gamma$  zweimal stetig differenzierbar ist, konvergiert  $f_{12}(\theta_1 t, \theta_2 s)$  gegen  $f_{12}(0, 0) = h^T D_x^2 \gamma(0) h$ . Die Formel für die Kovarianz der partiellen Ableitungen  $D_x Z$  folgt mit einer einfachen Rechnung.  $\square$

### 2.3.1 Schätzung der Parameter

Die unbekannt Parameter sind  $m$  sowie  $\theta = (\sigma^2, \rho, \nu)$ . Zur Schätzung dieser Parameter gibt es mindestens zwei mögliche Verfahren

- 1) Anpassung über die nichtparametrisch geschätzte Autokovarianzfunktion  $\widehat{C}(\cdot)$ , bzw. über das Semivariogramm  $\widehat{\gamma}(\cdot)$ . Dies macht man entweder "von Auge" oder dann aber mit "kleinsten Quadraten":

$$\sum_{j=1}^k w_j \cdot \left( \widehat{C}(h_j) - C_{\theta}(h_j) \right)^2$$

wird bezüglich  $\theta$  minimiert. Die  $w_j$  sind Gewichte, z.B. die Anzahl Paare, die bei der Berechnung von  $\widehat{C}(h_j)$  verwendet werden.

- 2) Maximum-Likelihood. Weil wir ein Gauss'sches Zufallsfeld angenommen haben, ist die Likelihood basierend auf den Beobachtungen  $\mathbf{z} = (z(x_1), \dots, z(x_n))^T$  gegeben durch

$$2 \log L(\mu, \theta) = -\log \det(C_n(\theta)) - (\mathbf{z} - \mu \mathbf{1})^T C_n(\theta)^{-1} (\mathbf{z} - \mu \mathbf{1}).$$

Dabei ist  $\mathbf{1} = (1, \dots, 1)^T$  und  $C_n(\theta)_{ij} = \frac{1}{\sigma^2} C(x_i - x_j)$ . Die Berechnung und Maximierung der Likelihood-Funktion ist meist aufwändig. Für gegebenes  $(\rho, \nu)$  kann man die Maximierung bezüglich  $\mu$  und  $\sigma^2$  geschlossen durchführen. Setzt man das Ergebnis ein, so erhält man die sogenannte Profil-Likelihood, welche dann nur noch von zwei Parametern abhängt.

## 2.4 Kriging

Wir beobachten

$$\mathbf{Z} := (Z(x_1), Z(x_2), \dots, Z(x_n))^T$$

und wir wollen  $Z(x_0)$  schätzen. Neben einer *Punktvorhersage*  $\widehat{Z}(x_0)$  möchte man mindestens noch einen *Vorhersagefehler*  $\mathbf{E} \left[ (Z(x_0) - \widehat{Z}(x_0))^2 \right]^{1/2}$  angeben. Für eine vollständige Beschreibung der Unsicherheit braucht man sogar die bedingte Verteilung von  $Z(x_0)$  gegeben  $\mathbf{Z}$ .

### 2.4.1 Einfaches (simple) Kriging

Wir betrachten zunächst die lineare Vorhersage

$$\widehat{Z}(x_0) = \alpha + \boldsymbol{\beta}^T \mathbf{Z}$$

und wir wollen  $\alpha$  und  $\boldsymbol{\beta}$  so bestimmen, dass

$$\mathbf{E} \left[ (Z(x_0) - \widehat{Z}(x_0))^2 \right]$$

minimal wird. Das Feld muss weder Gauss'sch, noch stationär oder intrinsisch sein.

Die Lösung lautet (vergleiche Vorlesung Lineare Regression)

$$\begin{aligned} \boldsymbol{\beta}_0 &= C_n^{-1} \mathbf{c}_n \\ \alpha_0 &= m(x_0) - (m(x_1), \dots, m(x_n)) \boldsymbol{\beta}_0 \end{aligned}$$

mit

$$\begin{aligned} C_n &= (\text{Cov}(Z(x_i), Z(x_j)))_{1 \leq i, j \leq n} \\ \mathbf{c}_n &= (\text{Cov}(Z(x_0), Z(x_i)))_{1 \leq i \leq n} \end{aligned}$$

Es folgt also

$$\widehat{Z}(x_0) = \mathbf{E}[Z(x_0)] + \boldsymbol{\beta}_0^T (\mathbf{Z} - \mathbf{E}[\mathbf{Z}])$$

und der Vorhersagefehler ist

$$\mathbf{E} \left[ (Z(x_0) - \widehat{Z}(x_0))^2 \right] = \text{Var}(Z(x_0)) - \mathbf{c}_n^T C_n^{-1} \mathbf{c}_n$$

**Lemma 2.3** *Bei einem Gauss'schen Zufallsfeld minimiert die optimale lineare Prognose  $\widehat{Z}(x_0)$  den mittleren quadratischen Prognosefehler unter allen (auch nichtlinearen) Vorhersagen und*

$$Z(x_0) | \mathbf{Z} \sim \mathcal{N} \left( \widehat{Z}(x_0), \mathbf{E} \left[ (Z(x_0) - \widehat{Z}(x_0))^2 \right] \right).$$

**Beweis:** Für die bedingte Dichte von  $Z(x_0)$  gegeben  $\mathbf{Z}$  gilt:

$$p(z(x_0) | \mathbf{z}) = \frac{p(z(x_0), \mathbf{z})}{p(\mathbf{z})} \propto p(z(x_0), \mathbf{z}).$$

Die Proportionalität bedeutet "bis auf einen Faktor, der von  $\mathbf{z}$ , aber nicht von  $z(x_0)$  abhängen darf". Diesen Faktor erhält man dann aus der Normierung  $\int p(z(x_0) | \mathbf{z}) dz(x_0) = 1$ .

Die gemeinsame Dichte  $p(z(x_0), \mathbf{z})$  ist eine  $(n+1)$ -dimensionale Normalverteilungsdichte. Mit quadratischem Ergänzen kann man diese auf die Form bringen

$$p(z(x_0), \mathbf{z}) = \text{const}_1 \cdot \exp(-\text{const}_2 \cdot z(x_0)^2 + \text{const}_3 z(x_0)),$$

wobei  $\text{const}_1$  und  $\text{const}_3$  von  $\mathbf{z}$  abhängen. Daraus ist ersichtlich, dass  $Z(x_0)$  gegeben  $\mathbf{Z}$  normalverteilt ist.

Mit dem gleichen Argument sieht man ferner, dass der Erwartungswert dieser bedingten Verteilung linear ist in  $\mathbf{z}$  und dass deren Varianz nicht von  $\mathbf{z}$  abhängt. Der Rest des Lemmas folgt nun aus dem allgemeinen Resultat, dass die beste (nichtlineare) Vorhersage gegeben ist durch  $\mathbf{E}[Z(x_0)|\mathbf{Z}]$ .  $\square$

Die Schwierigkeit bei der Anwendung dieser Resultate ist, dass  $m(x_i)$  und  $C(x_i, x_j)$  unbekannt sind. Der einfachste Ausweg ist, diese Grössen unter der Annahme von Stationarität zu schätzen und die erhaltenen Schätzungen in die obigen Formeln einzusetzen. Unbefriedigend dabei ist, dass der Effekt des Schätzfehlers so ignoriert wird.

### 2.4.2 Gewöhnliches (ordinary) Kriging

Wir setzen jetzt voraus, dass der Erwartungswert von  $Z$  konstant ist, d.h.  $m(x) \equiv m$  und betrachten lineare Vorhersagen der Form

$$\widehat{Z}(x_0) = \boldsymbol{\lambda}^T \mathbf{Z} \quad \text{mit} \quad \sum \lambda_j = \boldsymbol{\lambda}^T \mathbf{1} = 1$$

Durch die zusätzliche Bedingung an die Koeffizienten nimmt der Vorhersagefehler natürlich zu. Die folgenden Überlegungen zeigen, dass diese Zusatzbedingung trotzdem vernünftig ist:

- Zur Bestimmung des optimalen  $\boldsymbol{\lambda}$  braucht man  $m$  nicht.
- Wenn man  $m$  erwartungstreu und linear schätzt und beim simple Kriging einsetzt, hat  $\widehat{Z}(x_0)$  diese Form: Sei  $\widehat{m} = \mathbf{w}^T \mathbf{Z}$  mit  $\mathbf{E}[\widehat{m}] = m$ . Dies ist äquivalent zu  $\mathbf{w}^T \mathbf{1} = 1$ . Einsetzen ergibt

$$\widehat{Z}(x_0) = \widehat{m} + \boldsymbol{\beta}_0^T (\mathbf{Z} - \widehat{m} \mathbf{1}) = (1 - \boldsymbol{\beta}_0^T \mathbf{1}) \mathbf{w}^T \mathbf{Z} + \boldsymbol{\beta}_0^T \mathbf{Z} = \boldsymbol{\lambda}^T \mathbf{Z}$$

mit

$$\boldsymbol{\lambda} = (1 - \boldsymbol{\beta}_0^T \mathbf{1}) \mathbf{w} + \boldsymbol{\beta}_0.$$

Dieses  $\boldsymbol{\lambda}$  erfüllt  $\boldsymbol{\lambda}^T \mathbf{1} = 1$ .

Zur Bestimmung der optimalen Koeffizienten  $\boldsymbol{\lambda}$  berechnen wir für ein intrinsisches Feld

$$\begin{aligned} \mathbf{E} \left[ (Z(x_0) - \widehat{Z}(x_0))^2 \right] &= \mathbf{E} \left[ (Z(x_0) - \boldsymbol{\lambda}^T \mathbf{Z})^2 \right] = \text{Var} (Z(x_0) - \boldsymbol{\lambda}^T \mathbf{Z}) \\ &= - \sum_{i \neq j} \lambda_i \lambda_j \gamma(x_i - x_j) + 2 \sum_i \lambda_i \gamma(x_i - x_0) \end{aligned}$$

(die letzte Gleichung folgt aus Lemma

Die Optimierung dieses Ausdrucks mit Nebenbedingung  $\boldsymbol{\lambda}^T \mathbf{1} = 1$  führt auf ein lineares Gleichungssystem. Die konkrete Form dieses Gleichungssystems ist für uns nicht von grosser Bedeutung, da wir sowieso Programmpakete benutzen werden.

Man kann ferner zeigen, dass man die Lösung des gewöhnlichen Krigings bekommt, wenn man einfaches Kriging benutzt und für  $m$  den besten linearen erwartungstreuen Schätzer

$$\hat{m} = \frac{\mathbf{1}^T C_n^{-1} \mathbf{Z}}{\mathbf{1}^T C_n^{-1} \mathbf{1}}$$

einsetzt.

### 2.4.3 Bayes'sches Kriging

In der Bayes'schen Statistik betrachtet man die unbekannt Parameter  $(m, \sigma^2, \rho, \nu)$  ebenfalls als zufällig mit einer sogenannten a priori Dichte. Die Dichte der Beobachtungen  $\mathbf{Z}$  wird dann als bedingte Dichte gegeben die Parameter interpretiert, und man berechnet mit Hilfe der Bayes-Formel die a posteriori Dichte, das heisst die Dichte von  $(m, \sigma^2, \rho, \nu)$  gegeben die Beobachtungen:

$$p((m, \sigma^2, \rho, \nu) | \mathbf{z}) = \frac{p(\mathbf{z} | m, \sigma^2, \rho, \nu) p(m, \sigma^2, \rho, \nu)}{p(\mathbf{z})} \propto p(\mathbf{z} | m, \sigma^2, \rho, \nu) p(m, \sigma^2, \rho, \nu).$$

Proportionalität bedeutet "bis auf einen Faktor, der von  $\mathbf{z}$ , aber nicht von  $m, \sigma^2, \rho, \nu$  abhängt". Diesen Faktor kann man dann durch Integration über  $m, \sigma^2, \rho, \nu$  bestimmen.

Als a priori Dichte wählt man etwas, was wenn möglich die Rechnungen vereinfacht und möglichst "nicht informativ" ist, also Dichten mit einer grossen Streuung. Meist nimmt man an, dass  $m, \sigma^2, \rho, \nu$  a priori unabhängig sind. Für  $m$  und  $\log(\sigma^2)$  verwendet man häufig die "Gleichverteilung", obwohl das natürlich keine Wahrscheinlichkeitsdichten sind. In einem gewissen Sinn sind das die am wenigsten informativen Dichten, wir gehen jedoch nicht näher darauf ein. Eine mögliche Wahl für die Dichten von  $\rho$  und  $\nu$  ist

$$p(\rho) = \frac{1}{(1 + \rho)^2}, \quad p(\nu) = \frac{1}{(1 + \nu)^2},$$

da dies wegen der Langschwänzigkeit nicht sehr informativ ist. Als a posteriori Dichte erhält man dann

$$p((m, \sigma^2, \rho, \nu) | \mathbf{z}) \propto \frac{p(\rho)p(\nu)}{(\sigma^2)^{n/2} (\det(C_n(\rho, \nu)))^{1/2}} \exp\left(-\frac{1}{2\sigma^2} (\mathbf{z} - m\mathbf{1})^T C_n^{-1} (\mathbf{z} - m\mathbf{1})\right).$$

Die a posteriori Dichte gibt einem einmal eine Angabe, welche Parameterwerte im Lichte der Daten plausibel sind. Man ist also insbesondere an den Randdichten interessiert, weil man diese grafisch darstellen kann. Durch Ausintegrieren von  $m$  und  $\sigma^2$  erhält man die gemeinsame Dichte von  $\rho$  und  $\nu$ . Die Dichte von  $\rho$ , bzw.  $\nu$  allein muss dann numerisch berechnet werden.

Die a posteriori Dichte spielt auch eine wichtige Rolle bei der Vorhersage für  $Z(x_0)$ . Gemäss dem Satz der totalen Wahrscheinlichkeit gilt

$$p(z(x_0) | \mathbf{z}) = \int \underbrace{p(z(x_0) | \mathbf{z}, m, \sigma^2, \rho, \nu)}_{\text{Normalverteilungsdichte}} p(m, \sigma^2, \rho, \nu | \mathbf{z}) dm d\sigma^2 d\rho d\nu$$

Statt geschätzte Parameter einzusetzen, wird in der Bayes'schen Statistik also gemittelt gemäss der a posteriori Verteilung.

## 2.5 Weitere Themen und Resultate

Wir geben hier noch einen stichwortartigen Überblick über weitere Themen und Resultate:

- Nugget-Effekt: falls vorhanden, dann  $\widehat{Z}(x_0) \nrightarrow Z(x_i)$  falls  $x_0 \rightarrow x_i$ . Man macht dann also eine echte Glättung.
- Lineares Modell für  $m(x)$ :

$$m(x) = m + \sum_{j=1}^p \alpha_j f_j(x)$$

Analoge Überlegungen wie vorher. Man kann lineare Vorhersagen betrachten, für die man die Koeffizienten  $\alpha_j$  nicht braucht (universelles Kriging).

- Intrinsische Modelle höherer Ordnung. Statt für die Differenzen  $Z(x+h) - Z(x)$  verlangt man Verschiebungsinvarianz nur für Zuwächse höherer Ordnung. Ein Beispiel einer Differenz zweiter Ordnung ist  $Z(x+h) + Z(x-h) - 2Z(x)$ .
- Block-Kriging: Vorhersage von  $\int_A Z(x)dx$  durch  $\int_A \widehat{Z}(x)dx$ .
- Nichtlineare Vorhersagen:
  - Transformieren (z. Bsp. mit Hilfe der log Funktion) führt auf log-normales Kriging.
  - Disjunctive Kriging. Hier betrachtet man Vorhersagen der Form  $\widehat{Z}(x) = \sum g_i(Z(x_i))$ .
- Anisotropie: Man kann ein anisotropes Variogramm schätzen, indem man die quadrierten Differenzen  $(Z(x_i) - Z(x_j))^2$  zweidimensional als Funktion von  $x_i - x_j$  glättet. Aus einem isotropen parametrischen Variogramm kann man mit  $h \rightarrow \gamma(\|Ah\|)$  ein gültiges anisotropes Modell erhalten (entspricht einer Verzerrung des Raumes). Eine andere anisotrope Klasse bilden die Produktvariogramme  $h \rightarrow \prod_{i=1}^d \gamma_i(h_i)$ , wobei die  $\gamma_i$  beliebige Variogramme für  $d = 1$  sind.

# Kapitel 3

## Markovmodelle auf einem Gitter

Die Spezifikation nicht-Gauss'scher Modelle  $(Z(x))_{x \in \mathbb{R}^d}$  mit kontinuierlichem Parameter  $x$  ist schwierig. Einfacher ist es,  $(Z(x))_{x \in L}$  auf einem endlichen Gitter  $L \subset \mathbb{R}^d$  zu definieren, zum Beispiel  $L = \{1, 2, \dots, N\}^d$ . Dazu genügt es, die Wahrscheinlichkeit bzw. die Dichte von  $(Z(x))_{x \in L}$  anzugeben.

Gesucht sind Modelle mit einfacher Struktur.

### 3.1 Zeitliche und räumliche Markoveigenschaft

Sei zuerst  $L = \{1, 2, \dots, N\}$ .  $t \in L$  kann als Zeit interpretiert werden.

**Definition 3.1**  $(Z_t)_{t \in L}$  heisst *zeitdiskreter Markovprozess falls*

$$\mathcal{L}(Z_t | (Z_s, s < t)) = \mathcal{L}(Z_t | Z_{t-1}) \quad t \in \{3, 4, \dots, N\}$$

Unter allen vorangegangenen Werten ist also nur der letzte relevant für die Prognose des momentanen Wertes.

Bei einem Markovprozess kann man  $\mathcal{L}(Z_t | Z_{t-1})$  beliebig vorgeben, zum Beispiel mit einer Übergangsdichte  $p_t(z_t | z_{t-1})$  (im stetigen Fall) oder einer Übergangsmatrix (im diskreten Fall). Zusammen mit  $\mathcal{L}(Z_1)$  ist dann  $\mathcal{L}((Z_t)_{t \in L})$  festgelegt. Die gemeinsame Dichte sieht dann so aus:

$$p((z_1, \dots, z_N)) = p_1(z_1) \prod_{t=2}^N p_t(z_t | z_{t-1}). \quad (3.1)$$

Man multipliziert also einfach die Übergangsdichten miteinander.

Für  $L \subset \mathbb{R}^d, d > 1$ , existiert jedoch keine natürliche Ordnung. Daher lässt sich obige Definition nicht so einfach verallgemeinern; es ist nicht klar, was mit den vorangegangenen Werten gemeint ist. Im Raum betrachtet man stattdessen die bedingte Verteilung von  $Z(x)$  gegeben *alle* andern Werte. Diese bedingte Verteilung soll nur von denjenigen  $Z(x')$  abhängen, für die  $x'$  ein Nachbar von  $x$  ist. Im Fall  $L = \{1, 2, \dots, N\}^d$  sind die Nachbarn von  $x$  alle Punkte im Abstand 1. Wir können jedoch ohne zusätzliche Komplikationen beliebige Nachbarschaftsrelationen  $x \sim x'$  betrachten. Das einzige, was man braucht, sind die Symmetrie  $x \sim x' \Leftrightarrow x' \sim x$  und  $x \approx x$ .

**Definition 3.2**  $(Z_x)_{x \in L}$  heisst ein Markov-Feld bezüglich der Nachbarschaftsrelation  $\sim$ , falls

$$\mathcal{L}(Z_x | Z_{x'}, x' \neq x) = \mathcal{L}(Z_x | Z_{x'}, x' \sim x).$$

Für  $L = \{1, 2, \dots, N\}$  folgt die räumliche Markoveigenschaft aus der zeitlichen.

**Lemma 3.1** Ein Markovprozess ist auch ein Markov-Feld bezüglich der Relation  $s \sim t \Leftrightarrow |s - t| = 1$ .

**Beweis:** Unter Verwendung von (

Die Umkehrung gilt, falls alle  $p(z)$  strikt positiv sind, siehe Abschnitt

Wie sieht nun  $p((z_x)_{x \in L})$  aus für ein Markov-Feld? Kann man auch  $p(z_x | z_{x'}, x \sim x')$  beliebig wählen und dann eine analoge Formel wie (

## 3.2 Die Gibbsdarstellung für Markovfelder

### 3.2.1 Der Satz von Hammersley-Clifford

Wenn wir jeweils zwei benachbarte Punkte von  $L$  durch eine Kante verbinden, erhalten wir einen (ungerichteten) Graphen. Zunächst führen wir einen Begriff aus der Graphentheorie ein.

**Definition 3.3**  $C \subset L$  heisst Clique falls  $x \in C, x' \in C$  impliziert  $x = x'$  oder  $x \sim x'$ , d.h.  $C$  ist ein vollständiger Teilgraph.

Wir brauchen noch etwas zusätzliche Notation. Der Wertebereich von  $Z_x$  sei  $S$ , wobei  $S$  diskret oder ein Intervall sein kann. Wir bezeichnen dann die möglichen Konfigurationen des Feldes mit  $z := (z_x)_{x \in L} \in S^L$  und die Restriktion einer Konfiguration auf  $C \subset L$  als  $z_C := (z_x)_{x \in C}$ . Wenn  $A \cap B = \emptyset$  dann kann man zwei Konfigurationen auf  $A$ , bzw.  $B$  zusammensetzen:  $(z_A z'_B)_x = z_x$  falls  $x \in A$  und  $(z_A z'_B)_x = z'_x$  falls  $x \in B$ .

Schliesslich bezeichnen wir mit  $p(z)$  die Wahrscheinlichkeitsfunktion (im diskreten Fall) bzw. die gemeinsame Dichte (im stetigen Fall) des Zufallsfelds.

**Satz 3.1 (Hammersley-Clifford ~ 1970)** Wenn  $p(z) > 0$  für alle  $z$  und  $p(z)$  stetig ist, dann ist das zugehörige Zufallsfeld ein Markov-Feld genau dann wenn Funktionen  $g_C : S^C \rightarrow \mathbb{R}^+$  existieren, so dass

$$p(z) = \prod_{C \text{ Cliquen}} g_C(z_C).$$

**Beweis:** Der Beweis von “ $\Leftarrow$ ” geht analog zum Beweis des Lemmas

Für den Beweis von “ $\Rightarrow$ ” fixieren wir eine beliebige Konfiguration  $w \in S^L$  und definieren

$$\begin{aligned} \Psi_A(z_A) &= -\log(p(z_A w_{L \setminus A})) \quad A \subset L \\ \Phi_A(z_A) &= \sum_{B \subset A} (-1)^{|A|-|B|} \Psi_B(z_B). \end{aligned}$$

Der Übergang von  $(\Psi_A)$  zu  $(\Phi_A)$  heisst die *Moebius transformation*. Die Umkehrung dieser Transformation lautet (siehe z.B. S. Lauritzen, Graphical Models (1966), Lemma A.2):

$$\Psi_A(z_A) = \sum_{B \subset A} \Phi_B(z_B).$$

Also gilt

$$p(z) = \exp(-\Psi_L(z_L)) = \exp\left(-\sum_{B \subset L} \Phi_B(z_B)\right) = \prod_{B \subset L} g_B(z_B)$$

mit  $g_B(z_B) = \exp(-\Phi_B(z_B))$

Es bleibt zu zeigen: Falls  $B$  keine Clique ist, dann ist  $g_B \equiv 1$ , bzw.  $\Phi_B \equiv 0$ . Sei also  $B$  mit  $x \in B$ ,  $x' \in B$ ,  $x \neq x'$ ,  $x \approx x'$ . Für  $A \subset B \setminus \{x, x'\}$  definieren wir  $A_1 = A \cup x$ ,  $A_2 = A \cup x'$  und  $A_3 = A \cup \{x, x'\}$ . Dann gilt

$$\begin{aligned} \Phi_B(z_B) &= \sum_{A \subset B} (-1)^{|B|-|A|} \Psi_A(z_A) \\ &= \sum_{A \subset B \setminus \{x, x'\}} (-1)^{|B|-|A|} (\Psi_A(z_A) - \Psi_{A_1}(z_{A_1}) - \Psi_{A_2}(z_{A_2}) + \Psi_{A_3}(z_{A_3})) \\ &= \sum_{A \subset B \setminus \{x, x'\}} (-1)^{|B|-|A|} \log \left( \frac{p(z_{A_1} w_{L \setminus A_1}) p(z_{A_2} w_{L \setminus A_2})}{p(z_A w_{L \setminus A}) p(z_{A_3} w_{L \setminus A_3})} \right) \\ &= \sum_{A \subset B \setminus \{x, x'\}} (-1)^{|B|-|A|} \log \left( \frac{p(z_x | z_A w_{L \setminus A_1}) p(w_x | z_{A_2} w_{L \setminus A_3})}{p(w_x | z_A w_{L \setminus A_1}) p(z_x | z_{A_2} w_{L \setminus A_3})} \right). \end{aligned}$$

Jeder Term in der Summe ist gleich Null, weil  $x \approx x'$ . □

Einige Bemerkungen und Folgerungen zu diesem wichtigen Satz:

- $\Phi_\emptyset = \Psi_\emptyset = -\log(p(w))$  ist eine Konstante. Es folgt:

$$p(z) \propto \prod_{C \text{ Clique, } C \neq \emptyset} g_C(z_C).$$

Die Normierungskonstante lässt sich aber im Allgemeinen nicht berechnen. Im diskreten Fall hat man eine Summe mit sehr vielen Termen, und im stetigen Fall ein hochdimensionales Integral.

- Modelle dieser Form heissen *Gibbsverteilungen*. Sie spielen in der statistischen Mechanik eine wichtige Rolle. Dort sind  $\Phi_B$  und  $\Psi_B$  umgekehrt proportional zur absoluten Temperatur  $T$ ,  $T\Phi_B$  ist das *Potential* der Konfiguration  $z_B$  und  $T\Psi_B$  ist die Energie von  $z_B$ . Es gilt also

$$p(z) \propto \exp\left(-\frac{1}{T} \text{Energie}\right).$$

Wir nehmen die Temperatur  $T$  meist als eins an und verwenden auch für  $(\Phi_C)$  die Bezeichnung Potential.

- Die Darstellung von  $p(z)$  als  $\prod g_C(z_C)$  ist nicht eindeutig (und damit natürlich das Potential  $(\Phi_C)$  ebenfalls nicht). Das im Beweis konstruierte Potential hat die Eigenschaft  $\Phi_C(z_C) = 0$  falls  $z_x = w_x$  für ein  $x \in C$ . Mit dieser Zusatzbedingung ist  $(\Phi_C)$  eindeutig.

- Wie bei Markovprozessen lässt sich also die gemeinsame Dichte als ein Produkt von Faktoren schreiben, die jeweils nur wenige Variablen enthalten. Die Faktoren sind aber *nicht* die bedingten Dichten von jeweils einer Variablen gegeben der Rest:

$$\prod_{x \in L} p(z_x \mid z_{x'}, x' \sim x) \neq p(z),$$

vergleiche (

### 3.2.2 Beispiele

#### Binäre Markovfelder

Als erstes betrachten wir den Wertebereich  $S = \{0, 1\}$ , es handelt sich also um den binären Fall mit nur zwei Zuständen (“weiss”, bzw. “schwarz”). Weiter wählen wir  $w_x = 0$  für alle  $x$  (“alles weiss”) als Bezugsconfiguration und wir betrachten ein normiertes Potential  $\Phi_C(z_C) = 0$  falls  $z_x = 0$  für ein  $x$  in  $C$ . Dann gibt es Konstanten  $J_C$  derart dass

$$\Phi_C(z_C) = J_C \cdot \prod_{x \in C} z_x.$$

Damit ist also

$$p(z) \propto \exp \left( - \sum_C J_C \cdot \prod_{x \in C} z_x \right).$$

Falls  $J_C < 0$  ist, so besteht eine erhöhte Tendenz, dass alle Punkte in  $C$  schwarz sind.

Die bedingten Verteilungen von  $Z(x)$  gegeben den Rest sehen dann wie folgt aus

$$P [z_x = 1 \mid z_{x'}, x' \sim x] = \frac{\exp \left( - \sum_{\text{Cliques } C; x \in C} J_C \prod_{x' \in C, x' \neq x} z_{x'} \right)}{1 + \exp \left( - \sum_{\text{Cliques } C; x \in C} J_C \prod_{x' \in C, x' \neq x} z_{x'} \right)}$$

Den einfachsten Spezialfall erhalten wir, wenn wir  $J_C = 0$  setzen, falls  $|C| > 2$ , d.h. es handelt sich um Paarpotentiale. Im Fall des kubischen Gitters mit  $2d$  nächsten Nachbarn bestehen alle Cliques nur aus einem oder zwei benachbarten Punkten, das heisst wir haben dann immer ein Paarpotential. Wir vereinfachen die Notation, indem wir  $J_x$  anstelle von  $J_{\{x\}}$  schreiben und  $J_{xy} = J_{yx} = \frac{1}{2} J_{\{x,y\}}$  setzen. Es gilt dann

$$p(z) \propto \exp \left( - \sum_{x \in L} J_x z_x - \sum_{x,y; x \sim y} J_{xy} z_x z_y \right).$$

Wir formen die zweite Summe in der Exponentialfunktion um. Weil  $z_x^2 = z_x$ , folgt

$$\sum_{x,y; x \sim y} J_{xy} z_x z_y = -\frac{1}{2} \sum_{x,y; x \sim y} J_{xy} (z_x - z_y)^2 + \sum_{x \in L} z_x \sum_{y \sim x} J_{xy},$$

und wir erhalten schliesslich

$$p(z) \propto \exp \left( - \sum_{x \in L} z_x (J_x + \sum_{y, y \sim x} J_{xy}) + \frac{1}{2} \sum_{x \sim y} J_{xy} (z_x - z_y)^2 \right).$$

Falls  $J_x = -\sum_{y, y \sim x} J_{xy}$ , dann besteht also eine Symmetrie bezüglich der Vertauschung von 0 und 1.  $J_{xy} < 0$  bedeutet eine Tendenz benachbarter Punkte gleich zu sein ("Anziehung"), während  $J_{xy} > 0$  einer Abstossung gleicher Werte entspricht. Wenn  $J_{xy}$  konstant gleich  $J$  ist hat man das sogenannte Ising-Modell. Dann wird in der zweiten Summe gezählt, wieviele Paare von benachbarten Punkten mit verschiedenen Werten es gibt. Bei nächsten Nachbarn ist dies einfach die Länge der Grenze zwischen den beiden Zuständen "schwarz" und "weiss".

Beim kubischen Gitter mit  $2d$  nächsten Nachbarn hat man für ein festes  $x$  wegen der Gibbsdarstellung  $2d+1$  freie Parameter für die Wahrscheinlichkeiten  $P[z_x = 1 | z_{x'}, x' \sim x]$ . Die Anzahl möglicher Randbedingungen ( $z_{x'}, x' \sim x$ ) ist jedoch  $2^{2d} > 1 + 2d$ . Man kann darum  $P[z_x = 1 | z_{x'}, x' \sim x]$  im Unterschied zu zeitlichen Markovketten nicht beliebig wählen. Wegen der Gibbsdarstellung bestehen ausserdem Beziehungen zwischen diesen bedingten Wahrscheinlichkeiten für verschiedene  $x$ .

Oft erhält man aber interessantere Modelle, wenn man nicht nur Paarpotentiale betrachtet. Tjelmeland und Besag (1998) arbeiten mit einem hexagonalen Gitter und lassen "zweitnächste" Nachbarn zu. Es entstehen dann Cliques der Grösse  $\leq 7$ . Anstatt ein normiertes Potential zu nehmen, kann man  $\Phi_C = 0$  setzen für  $|C| < 7$ . Mit Translations- und Rotationsinvarianz ergeben sich schliesslich 26 verschiedene Konfigurationen als Argumente für  $\Phi_C$ , d.h. man hat 26 freie Parameter.

### Endliche oder abzählbare Wertebereiche

Sei  $S = \{0, 1, 2, \dots, k\}$ . Wenn diese  $k+1$  Werte ungeordnete Kategorien, z.B. Farben, bezeichnen, dann ist ein naheliegende Modell gegeben durch

$$p(z) \propto \exp \left( - \sum_x \Phi_x(z_x) - \sum_{y \sim x} J_{xy} 1_{[z_x \neq z_y]} \right).$$

Wenn  $k = 1$  ist, so entspricht es dem Paarpotentialmodell von vorher.

Wenn die Werte von  $S$  ordinal sind, z.B.  $(k+1)$  Grauwerte, ist dieses Modell nicht sinnvoll. Naheliegender ist dann ein Modell, wo

$$\mathcal{L}(Z_x | Z_{x'}, x' \sim x) = \text{Bin}(k, \pi(z_{\partial x})),$$

d.h. man hat eine Binomialverteilung für  $Z_x$  mit einer Erfolgswahrscheinlichkeit, die von den benachbarten Werten abhängt. Wann sind diese bedingten Wahrscheinlichkeiten von der Gibbs'schen Form? Es gilt

$$\begin{aligned} p(z_x | z_{x'}, x' \sim x) &= \binom{k}{z_x} \pi(z_{\partial x})^{z_x} (1 - \pi(z_{\partial x}))^{k - z_x} \\ &= \binom{k}{z_x} (1 - \pi(z_{\partial x}))^k \exp \left( z_x \log \left( \frac{\pi(z_{\partial x})}{1 - \pi(z_{\partial x})} \right) \right) \end{aligned}$$

Dies ist genau dann von der Gibbs'schen Form, falls

$$\log \left( \frac{\pi(z_{\partial x})}{1 - \pi(z_{\partial x})} \right) = -J_x - \sum_{y \sim x} J_{xy} z_y,$$

und wir erhalten die folgenden Potentiale

$$\begin{aligned}\Phi_x(z_x) &= J_x z_x - \log \binom{k}{z_x} \\ \Phi_{xy}(z_x, z_y) &= J_{xy} z_x z_y.\end{aligned}$$

Das ist analog zum Modell der logistischen Regression, wobei die gleiche Variable  $Z$  als Ziel- und erklärende Grösse dient. Man spricht daher von einem “autologischen” Modell.

Schliesslich betrachten wir noch den abzählbaren Fall  $S = \{0, 1, 2, \dots\}$ . Man denke dabei an Zählraten, z.B. die Anzahl Krankheitsfälle pro Ort. Wir suchen ein Modell, wo

$$\mathcal{L}(Z_x | Z_{x'}, x' \sim x) = \text{Pois}(\lambda(z_{\partial x})).$$

Aus dieser Forderung ergibt sich

$$p(z_x | z_{x'}, x' \sim x) = \exp(-\lambda(z_{\partial x})) \frac{\lambda(z_{\partial x})^{z_x}}{z_x!} = \exp(-\lambda(z_{\partial x}) + z_x \log \lambda(z_{\partial x}) - \log(z_x!)).$$

Damit dies von der Gibbs’schen Form ist, muss

$$\log(\lambda(z_{\partial x})) = -J_x - \sum_{y \sim x} J_{xy} z_y$$

sein. Dabei tritt eine Komplikation auf, denn die Normierung der Gibbsverteilung

$$\sum_z \exp \left( - \sum_x (\log(z_x!) + J_x z_x) - \sum_{y \sim x} J_{xy} z_x z_y \right)$$

muss endlich sein. Dies ist erfüllt, falls alle  $J_{xy} \geq 0$ , denn dann ist jeder Summand kleiner oder gleich

$$\prod_x \frac{\exp(-J_x z_x)}{z_x!}.$$

Sobald ein  $J_{xy} < 0$  ist, dann ist die Normierung unendlich. Um dies zu sehen, beschränkt man sich auf Konfigurationen, wo  $z_x = z_y = k$  ist und  $z_{x'} = 0$  für alle andern Orte. Dann gilt

$$\sum_{k=0}^{\infty} \frac{1}{(k!)^2} \exp(-(J_x + J_y)k - J_{xy}k^2) = \infty$$

wegen der Stirling-Formel. Weil  $J_{xy} \geq 0$  bedeutet, dass grosse Werte bei den Nachbarn von  $x$  kleine Werte bei  $x$  begünstigen, ist das Modell nur beschränkt brauchbar.

### Gauss’sche Markovfelder

Wir können die gemeinsame Dichte schreiben als

$$\begin{aligned}p(z) &= \frac{1}{Z} \exp \left( -\frac{1}{2} (z - \mu)^T \Sigma^{-1} (z - \mu) \right) \\ &= \frac{1}{Z'} \exp \left( -\frac{1}{2} z^T \Sigma^{-1} z + z^T \Sigma^{-1} \mu \right)\end{aligned}$$

wobei  $Z$  und  $Z'$  Normierungskonstanten sind.

Dies ist von der Gibbs'schen Form wenn man die Nachbarschaftsrelation definiert durch  $x \sim x' \Leftrightarrow (\Sigma_{xx'}^{-1}) \neq 0$ . Ausserdem hat man stets ein Paarpotential:  $\Phi_C \equiv 0$  falls  $|C| > 2$ . Für die Markoveigenschaft geht man also von  $\Sigma^{-1}$  aus.

Die bedingte Dichte von  $Z(x)$  gegeben der Rest ist

$$\begin{aligned} p(z_x | z_{\partial x}) &\propto \exp \left( (\Sigma^{-1}\mu)_x z_x - \frac{1}{2} (\Sigma^{-1})_{xx} z_x^2 - z_x \sum_{x' \sim x} (\Sigma^{-1})_{xx'} z_{x'} \right) \\ &\propto \exp \left( -\frac{1}{2} (\Sigma^{-1})_{xx} \left( z_x - \frac{(\Sigma^{-1}\mu)_x}{(\Sigma^{-1})_{xx}} + \frac{\sum_{x' \sim x} (\Sigma^{-1})_{xx'} z_{x'}}{(\Sigma^{-1})_{xx}} \right)^2 \right), \end{aligned}$$

also wieder eine Normalverteilungsdichte. Insbesondere sind die

$$-\frac{(\Sigma^{-1})_{xx'}}{(\Sigma^{-1})_{xx}}$$

die Interpolationsgewichte (Kriging-Gewichte), und  $1/(\Sigma^{-1})_{xx}$  ist die Interpolationsvarianz (Kriging-Varianz).

### 3.2.3 Ergänzungen und Kommentare

- Markovfelder auf einem unendlichen Gitter, zum Beispiel  $\mathbb{Z}^d$ . Die Menge aller Konfigurationen ist dann überabzählbar, und die Formel

$$p(z) \propto \exp \left( - \sum_C \Phi_C(z_C) \right)$$

macht keinen Sinn mehr. Die Formeln für die bedingten Verteilungen von  $Z_x$  gegeben der Rest,

$$p(z_x | z_{x'}, x' \neq x) \propto \exp \left( - \sum_{C; x \in C} \Phi_C(z_C) \right)$$

hingegen macht weiterhin Sinn (zumindest wenn jedes  $x$  nur endlich viele Nachbarn hat). Man definiert daher eine Gibbsverteilung auf dem unendlichen Gitter durch die Forderung, dass alle bedingten Verteilungen von  $Z_x$  gegeben der Rest von dieser Form sein müssen. Dann muss man natürlich zeigen, dass solche Gibbsverteilungen existieren. Dies kann man mit einem Grenzwert (“thermodynamischer Limes”) tun.

Sei  $S$  kompakt und  $L_n = \{-n, -n+1, \dots, n\}^d \nearrow \mathbb{Z}^d$  für  $n \rightarrow \infty$ . Ferner sei  $p_n$  ein Gibbsfeld auf  $L_n$  mit einem (translationsinvarianten) Potential. Dann existiert  $\lim_{n \rightarrow \infty} p_n$  im Sinne der schwachen Konvergenz (wenn man zu einer Teilfolge übergeht). Unter Umständen hängt der Limes aber davon ab, wie man  $p_n$  am Rand von  $L_n$  definiert. Damit ist eine Gibbsverteilung auf einem unendlichen Gitter im Allgemeinen nicht eindeutig. Man spricht von *Phasenübergang*.

- Für  $L = \{1, 2, \dots, n\}$  impliziert die räumliche Markov-Eigenschaft die zeitliche Markov-Eigenschaft, sofern  $p$  von der Gibbs'schen Form ist.

**Beweis:**

$$p(z_k | z_{k-1}, \dots, z_1) = \frac{p(z_1, \dots, z_k)}{p(z_1, \dots, z_{k-1})} = \frac{\sum_{z_{k+1}, \dots, z_n} p(z)}{\sum_{z_k, \dots, z_n} p(z)}.$$

Wir verwenden die Gibbsdarstellung. Die Cliques bestehen entweder aus einem oder aus zwei benachbarten Punkten. Da Terme  $g_C$  mit  $C \subset \{1, 2, \dots, k-1\}$  wegfallen, kriegen wir

$$p(z_k | z_{k-1}, \dots, z_1) = \frac{\sum_{z_{k+1}, \dots, z_n} \prod_{j=k}^n g_j(z_j) g_{j-1, j}(z_{j-1}, z_j)}{\sum_{z_k, \dots, z_n} \text{gleiche Terme wie im Zähler}}.$$

Offensichtlich kommen  $z_{k-2}, \dots, z_1$  nicht vor. □

**Bemerkung:**

Man kann diese Summen effizient berechnen mit Hilfe der Rekursion  $n \rightarrow n-1 \rightarrow \dots \rightarrow 1$  gegeben durch:

$$\begin{aligned} h_n(z_n) &\equiv 1 \\ h_{k-1}(z_{k-1}) &:= \sum_{z_k} g_{k-1, k}(z_{k-1}, z_k) g_k(z_k) h_k(z_k). \end{aligned}$$

Dann ist nämlich

$$p(z_k | z_{k-1}, \dots, z_1) = \frac{h_k(z_k) g_k(z_k) g_{k-1, k}(z_{k-1}, z_k)}{h_{k-1}(z_{k-1})}.$$

- Im Allgemeinen kann man die Randverteilung  $Z_x$  nicht explizit berechnen. Im diskreten Fall gilt

$$p(z_x) = \sum_{z_{L \setminus x}} p(z) = \frac{\sum_{z_{L \setminus x}} \exp(-\sum_C \Phi_C(z_C))}{\sum_z \exp(-\sum_C \Phi_C(z_C))},$$

und man hat sowohl im Nenner als auch im Zähler zu viele Terme, über die summiert werden muss.

- Verfeinerung oder Vergrößerung des Gitters macht Probleme: Markov-Eigenschaft geht dabei im Allgemeinen verloren. Das sieht man am einfachsten im Gauss'schen Fall: Wenn man bei  $\Sigma$  Zeilen und Spalten streicht (oder hinzufügt), dann ändern sich alle Elemente von  $\Sigma^{-1}$ .

### 3.2.4 Schätzung der Potentiale

Die Gibbsdarstellung ergibt die richtige Parametrisierung der bedingten Verteilungen von  $Z_x$  gegeben den Rest. Wir betrachten nun das Problem, wie man die Potentiale aus einer Realisierung von  $Z = (Z_x)$  schätzen kann. Wir nehmen an, dass die Potentiale bis auf einen endlich dimensional Parameter  $\theta$  bekannt sind:  $\Phi_C(z_C) = \Phi_C(z_C, \theta)$ . Üblicherweise nimmt man dazu an, dass das Potential translationsinvariant ist.

Die Maximum-Likelihood-Schätzung ist im Allgemeinen nicht direkt anwendbar. Im diskreten Fall gilt

$$p_\theta(z) = \frac{\exp(-\sum_C \Phi_C(z_C, \theta))}{\sum_{z'} \exp(-\sum_C \Phi_C(z'_C, \theta))},$$

Der Nenner hängt ebenfalls von  $\theta$  ab, lässt sich aber nicht explizit berechnen.

Es gibt Verfahren, die die Likelihoodfunktion mit Hilfe von Simulationen approximieren und dann diese approximative Likelihood maximieren. Wir gehen hier jedoch nicht darauf ein, sondern betrachten ein einfacheres Verfahren, welches die sogenannte Pseudo-Likelihood maximiert. Diese ist definiert als

$$L_p(\theta) = \prod_{x \in L} p_\theta(z_x | z_{\partial x}) = \prod_{x \in L} \frac{\exp(-\sum_{x \in C} \Phi_C(z_C, \theta))}{N(z_{\partial x}, \theta)}.$$

Der entscheidende Punkt ist, dass man hier die Normierung  $N(z_{\partial x}, \theta)$  in vielen Fällen berechnen kann (man summiert nur über  $|S|$  Terme).

Warum ist die Pseudo-Likelihood-Methode sinnvoll? Diese Methode gehört zur Klasse der M-Schätzer, und wir besprechen hier kurz die Heuristik für solche Schätzer.

**Exkurs: M-Schätzer** Wir betrachten eine (im Allgemeinen hoch-dimensionale) Zufallsvariable  $Z$ , deren Verteilung von einem Parameter  $\theta$  abhängt. Dann sind M-Schätzer definiert durch

$$T(z) = \arg \max_{\theta} \rho(z, \theta)$$

wobei  $\rho$  eine zunächst beliebige ‘Kontrastfunktion’ ist. Damit dies einen vernünftigen Schätzer definiert, müssen zwei Bedingungen erfüllt sein:

– Wenn  $Z \sim p_{\theta_0}(z)$ , dann

$$\mathbf{E}_{\theta_0} [\rho(Z, \theta)] \leq \mathbf{E}_{\theta_0} [\rho(Z, \theta_0)]$$

d.h. die gemittelte Funktion  $\rho$  hat das Maximum beim wahren Parameter.

–  $\rho(z, \theta)$  ist mit grosser Wahrscheinlichkeit nahe bei  $\mathbf{E}_{\theta_0} [\rho(Z, \theta)]$ .

In unserem Fall ist

$$\rho(z, \theta) = \sum_{x \in L} \log(p_\theta(z_x | z_{\partial x})).$$

Um die erste Bedingung zu verifizieren, berechnen wir den Erwartungswert von einem Summanden in  $\rho(Z, \theta)$ :

$$\begin{aligned} & \mathbf{E}_{\theta_0} [\log(p_\theta(Z_x | Z_{\partial x}))] \\ &= \mathbf{E}_{\theta_0} [\mathbf{E}_{\theta_0} [\log(p_\theta(Z_x | Z_{\partial x})) | Z_{\partial x}]] = \mathbf{E}_{\theta_0} \left[ \int p_{\theta_0}(z_x | Z_{\partial x}) \log(p_\theta(z_x | Z_{\partial x})) dz_x \right] \\ &= \mathbf{E}_{\theta_0} \left[ \int p_{\theta_0}(z_x | Z_{\partial x}) \log(p_{\theta_0}(z_x | Z_{\partial x})) dz_x + \int p_{\theta_0}(z_x | Z_{\partial x}) \log \left( \frac{p_\theta(z_x | Z_{\partial x})}{p_{\theta_0}(z_x | Z_{\partial x})} \right) dz_x \right]. \end{aligned}$$

Das zweite Integral ist stets negativ, weil

$$\log \left( \frac{p_\theta(z_x | z_{\partial x})}{p_{\theta_0}(z_x | z_{\partial x})} \right) \leq \frac{p_\theta(z_x | z_{\partial x})}{p_{\theta_0}(z_x | z_{\partial x})} - 1.$$

Für die zweite Bedingung braucht man ein Gesetz der Grossen Zahlen. Wir gehen nicht weiter darauf ein.

### 3.3 Anwendungen

#### 3.3.1 Klassifikation und Synthese von Texturen

Bei der Synthese von Texturen sucht man einen Algorithmus zur Erzeugung künstlicher Texturen, die sich visuell nicht wesentlich von einer vorgegebenen Textur unterscheiden. Eine Möglichkeit dafür ist die Simulation von einem Markovfeld, bei dem das Potential aus der vorgegebenen Textur geschätzt wurde. Die Schätzung der Potentiale kann mit der Maximierung der Pseudo-Likelihood erfolgen. Die Simulation von Markovmodellen besprechen wir in dieser Vorlesung nur am Rande, da dies in der Vorlesung “Stochastische Simulation” ausführlicher geschieht. Alle Verfahren sind iterativ, d.h. man erzeugt nicht eine Realisierung auf einen Schlag, sondern man beginnt mit irgendeiner Konfiguration und modifiziert dann diese iterativ so, dass sich im Limes eine Realisierung gemäss dem Markovmodell ergibt. Der einfachste Algorithmus ist der Gibbs-Sampler: Dabei wird jeweils die Konfiguration nur an einem Punkt  $x$  modifiziert gemäss der richtigen bedingten Verteilung  $p(z_x | z_{x'}, x' \sim x)$ .

Es ist jedoch ziemlich schwierig, mit diesem Vorgehen gute Resultate zu erhalten. Die Bilder in *Cross & Jain, IEEE Pattern Anal. Machine Intell. 5, 1983* sind vermutlich fast “zu gut”, weil der Simulationsalgorithmus nicht oft genug iteriert wurde.

Die Klassifikation von Texturen ist einfacher: Zur Verfügung steht ein Trainingsdatensatz von Texturen, die in endlich viele mögliche Klassen eingeteilt sind. Damit soll ein Algorithmus konstruiert werden, der eine neue Textur automatisch einer dieser Klassen zuordnen kann. Anstatt direkt das ganze Bild als Input eines Klassifikationsverfahrens zu benutzen, ist es meist besser, zunächst gewisse “Features” aus dem Bild zu extrahieren (Dimensionsreduktion), die man dann als Input nimmt. Die geschätzten Potentiale eines Markovfelds ergeben oft sehr trennscharfe Features.

Diese Idee wurde von *Baron et al., Technometrics 43, 2001* verwendet. In dieser Arbeit geht es um die Qualitätskontrolle bei der Fabrikation von Halbleitern. Dabei werden Chips auf eine dünne Silikonplatte (Wafer) aufgebracht. Man möchte die räumliche Anordnung von defekten Chips auf einem Wafer benutzen, um fehlerhafte Wafers einem von acht möglichen Ursachen im Produktionsprozess zuordnen zu können. Dazu werden funktionierende/defekte Chips als 0/1 kodiert und an das so entstandene Bild wird ein Markovmodell mit 8 nächsten Nachbarn und 10 unbekanntenen Werten des Potentials angepasst. Dies 10 Werte dienen dann als Input eines neuronalen Netzes zur Klassifikation.

#### 3.3.2 Bildrekonstruktion

Sei  $z = (z_x)_{x \in L}$  das “richtige”, ungestörte Bild, das wir aber nicht beobachten können. Wir sehen nämlich nur  $y = (y_x)_{x \in L}$ , ein gestörtes, degradiertes Bild, das verrauscht und verschmiert wurde. Der Zusammenhang zwischen  $z$  und  $y$  ist gegeben durch

$$y_x = \sum_{x'} \gamma(x, x') z_{x'} + \varepsilon_x,$$

wo  $\gamma(\cdot)$  die sogenannte *point-spread-Funktion* und  $\varepsilon_x$  ein *iid-Fehler* ist. Meist geht man davon aus, dass  $\gamma(x, x') = \gamma(x - x')$  und  $\varepsilon_x \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ , wobei  $\gamma(\cdot)$  und  $\sigma_\varepsilon^2$

bekannt sind. Wenn wir mit  $N(x)$  die Menge aller Punkte  $x'$  bezeichnen, für die  $\gamma(x, x') \neq 0$  ist, dann gilt

$$y | z \sim \prod_{x \in L} q(y_x | z_{N(x)}),$$

mit

$$q(y_x | z_{N(x)}) = \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} \exp\left(-\frac{1}{2\sigma_\varepsilon^2} \left(y_x - \sum_{x'} \gamma(x, x') z_{x'}\right)^2\right).$$

Statt additivem Gauss'schen Rauschen können wir allgemeinere bedingte Verteilungen  $q$  betrachten, z.B. für binäre Beobachtungen

$$\frac{\mathbb{P}[Y_x = 1 | z]}{\mathbb{P}[Y_x = 0 | z]} = \exp\left(\alpha + \sum_{x'} \gamma(x, x') z_{x'}\right)$$

oder für Anzahlen

$$y_x | z \sim \text{Pois}\left(\exp\left(\alpha + \sum_{x'} \gamma(x, x') z_{x'}\right)\right).$$

Ohne die dazwischengeschaltete Exponentialfunktion tritt der letzte Fall auf bei SPECT (single photon emission computerized tomography) : Dabei ist  $z_x$  die unbekannte Isotopenkonzentration an der (diskretisierten) Stelle  $x$  und  $\gamma(x, x')$  ist proportional zur Wahrscheinlichkeit, dass ein an der Stelle  $x$  emittiertes Photon an der Stelle  $x'$  des Detektors ankommt. Das gleiche Modell tritt auch bei der Modellierung der räumlichen Verteilung von Krankheiten auf: Dort ist  $z_x$  die unbekannte Rate im Distrikt  $x$ ,  $\alpha = \alpha_x$  ist die logarithmierte Grösse der Risikopopulation im Distrikt  $x$  und  $\gamma(x, x') = 0$  für  $x \neq x'$ .

Die Rekonstruktion von  $z$  aus  $y$  ist (wie die meisten Aufgaben in der Statistik) ein sogenanntes *inverses Problem*. Wir können die wahren, aber unbekanntenen Bildwerte  $(z_x)_{x \in L}$  als unbekannte Parameter betrachten. Die Maximum-Likelihood-Methode

$$\hat{z} = \arg \max_z \sum_x \log q(y_x | z_{N(x)})$$

liefert aber meist keine brauchbare Schätzung, weil die Anzahl zu schätzender Parameter gleich der Anzahl Beobachtungen ist. Dies sieht man sehr leicht im Fall, wo  $y_x = z_x + \varepsilon_x$ . Dann ist

$$\log q(y_x | z_{N(x)}) = -\frac{1}{2\sigma_\varepsilon^2} \sum_x (y_x - z_x)^2 + \text{const.},$$

und damit ist der Maximum Likelihood Schätzer  $\hat{z}_x = y_x$ . Die Schätzung des Originalbildes ist einfach gleich dem beobachteten, verrauschten Bild. Eine alternative Möglichkeit wäre die Verwendung des Steinschätzers  $\hat{z}_x = \lambda y_x + (1 - \lambda)\bar{y}$ . Weil hier nur mit dem globalen Mittelwert korrigiert wird, ist dieses Vorgehen aber ebenfalls ungenügend.

Die sinnvollste Ansatz zur Bildrekonstruktion ist das Ausnützen von Vorinformationen der Art "benachbarte Werte sind ähnlich". Dies entspricht einem Bayes-Ansatz, d.h. wir arbeiten mit einer a-priori-Verteilung  $z \sim \pi(z)$ , welche Konfigurationen  $z$

bevorzugt, bei denen benachbarte Werte ähnlich sind. Die Bayes-Formel ergibt dann die a-posteriori-Verteilung

$$p(z | y) \propto \pi(z) \prod_{x \in L} q(y_x | z_{N(x)}).$$

Als Schätzung von  $z$  wird entweder der Modus dieser a posteriori Verteilung (abgekürzt *MAP*) verwendet,

$$\hat{z} = \arg \max_z p(z | y) = \arg \max_z \log(p(z | y)) = \arg \max_z (\log \pi(z) + \sum_{x \in L} \log q(y_x | z_{N(x)})),$$

oder der Erwartungswert von  $z$  bezüglich der a posteriori Verteilung:  $\hat{z} = (\hat{z}_x; x \in L)$  mit

$$\hat{z}_x = \mathbf{E}[Z_x | y] = \int \dots \int z_x p(z | y) \prod_{x' \in L} dz_{x'}.$$

Der MAP-Schätzer unterscheidet sich vom Maximum Likelihood Schätzer durch den zusätzlichen Term  $\log \pi(z)$  in der zu maximierenden Funktion. Dies kann als zusätzlicher Regularisierungsterm aufgefasst werden: Konfigurationen mit tiefer a priori Wahrscheinlichkeit werden bestraft.

Um das Vorgehen in die Praxis umzusetzen, müssen wir den Mittelwert oder den Modus der a-posteriori-Verteilung berechnen können und eine geeignete a-priori-Verteilung  $\pi(\cdot)$  wählen.

### Markovfelder als a priori Verteilungen

Wir betrachten als a priori Verteilungen spezielle Gibbs-Verteilungen

$$\pi(z) \propto \exp \left( -\beta_1 \sum_x \phi(z_x) - \beta_2 \sum_{x \sim x'} \rho \left( \frac{z_x - z_{x'}}{\delta} \right) \right)$$

mit einer geeigneten Nachbarschaftsrelation  $x \sim x'$ . Die Funktion  $\rho$  ist gerade mit  $\rho(0) = 0$  und  $\rho(z) > 0$  für  $z \neq 0$ . Für  $\beta_2 > 0$  bevorzugt diese a priori Verteilung also lokal konstante Konfigurationen. Wenn der Wertebereich  $S \subset \mathbb{R}$  gross ist, ist die Form von  $\rho$  für die Eigenschaften von  $\hat{z}$  ziemlich wichtig. Man kann mindestens die folgenden Fälle unterscheiden:

$$\begin{aligned} \rho(z) &= \frac{z^2}{2}, \\ \rho(z) &= \begin{cases} \frac{z^2}{2} & \text{falls } |z| \leq 1 \\ |z| - \frac{1}{2} & \text{falls } |z| \geq 1 \end{cases} \\ \rho(z) &= \frac{z^2}{1 + z^2}. \end{aligned}$$

Der zweite Fall gehört zum Typ “ $\rho$  konvex mit linearer Asymptote” und enthält im Grenzfall  $\delta \rightarrow 0$  den Fall  $\rho(z) = |z|$ . Der dritte Fall ist konvex in der Mitte und beschränkt. Im Extremfall erhält man  $\rho(z) = 1_{\{z \neq 0\}}$ . Der Unterschied liegt darin, wie stark die zugehörige a priori Dichte Sprünge, die von Kanten im Bild herrühren, durch niedrige Wahrscheinlichkeit bestraft. Bei quadratischem  $\rho$  ist diese Bestrafung am stärksten und nimmt dann sukzessive ab.

Die Funktion  $\phi$  erlaubt es noch, a priori zwischen den verschiedenen möglichen Werten zu differenzieren. Meist wird  $\phi \equiv 0$  gesetzt, so dass alle möglichen Werte a priori gleich wahrscheinlich sind. Wenn  $S$  unbeschränkt ist, ist dann aber  $\pi(z)$  keine Verteilung mehr (die Normierung ist unendlich).

Der schwierigste Punkt bei der Wahl der a priori Verteilung ist die Festlegung der Parameter  $\beta_1$ ,  $\beta_2$  und  $\delta$ . Wir diskutieren die Wahl dieser Parameter unten noch etwas genauer.

Die a priori Verteilung ist also ein Markovfeld. Um zu entscheiden, ob die a posteriori Verteilung auch wieder die Markoveigenschaft hat, berechnen wir die bedingten Dichten

$$p(z_x | z_{L \setminus \{x\}}, y) \propto \pi(z) \prod_{x' \in L} q(y'_x | z_{N(x')}) \propto \prod_{C, x \in C} g_C(z_C) \prod_{x', x \in N(x')} q(y'_x | z_{N(x')}).$$

Auf der rechten Seite kommt daher  $z_{x''}$  nur vor, wenn entweder  $x'' \sim x$  oder wenn ein  $x'$  existiert, so dass  $x \in N(x')$  und  $x'' \in N(x')$ . Wenn die Verschmierung nur Punkte in der Nachbarschaft involviert, dann ist also die a posteriori Verteilung wieder ein Markovfeld, allerdings "verdoppelt" sich die Nachbarschaft. Falls es keine Verschmierung gibt, so sind die Nachbarschaften der a priori und der a posteriori Verteilung identisch. Dies wird die Berechnung von  $\hat{z}$  etwas vereinfachen.

### Berechnung von $\hat{z}$

Wir betrachten zunächst den MAP

$$\hat{z} = \arg \max_z \left( \log \pi(z) + \sum_x \log q(y_x | z_{N(x)}) \right).$$

Wenn die a priori-Verteilung und die  $q(y_x | z_{N(x)})$  Gauss'sch sind, dann kann man  $\hat{z}$  geschlossen berechnen:

$$\hat{z} = \arg \min_z \left( \beta_1 z^T z + \beta_2 (Dz)^T Dz + \frac{1}{2\sigma_\varepsilon^2} (y - \Gamma z)^T (y - \Gamma z) \right).$$

Auf der rechten Seite steht eine quadratische Form in  $z$ , und die Bestimmung von  $\arg \min$  führt auf ein lineares Gleichungssystem. Die Berechnung sind zwar einfach, aber die Resultate sind im allgemein nicht befriedigend: Innerhalb von homogenen Regionen wird zuwenig geglättet oder die Kanten werden verschmiert oder beides.

In den meisten andern Fällen muss man iterative Verfahren zur Berechnung von  $\hat{z}$  verwenden. Ein einfacher Algorithmus ist die iterative Maximierung bezüglich einer Komponente, wobei alle anderen Komponenten festgehalten werden.

$$\hat{z}_x^{neu} = \arg \max_{z_x} p(z_x \hat{z}_{L \setminus \{x\}}^{alt} | y) = \arg \max_{z_x} p(z_x | \hat{z}_{L \setminus \{x\}}^{alt}, y).$$

Wie wir oben gesehen haben, hängt die zu maximierende Funktion dann nur von wenigen Werten  $z_{x'}^{alt}$  ab, was die Berechnungen vereinfacht. Die Nachteile dieses Verfahrens sind, dass häufig viele Iterationen notwendig sind (wenn die Funktion einen langen schmalen Grat hat, der nicht parallel zu einer Koordinatenachse läuft), und dass man in einem lokalen Maximum steckenbleiben kann.

Nebenmaxima können nicht auftreten, wenn  $S$  ein Intervall ist und wenn die zu maximierende Funktion konkav ist. In allen Beispielen, die wir angeschaut haben, ist  $\sum_x \log q(y_x | z_{N(x)})$  konkav in  $z$ . Die Berechnung von MAP vereinfacht sich daher, wenn man auch  $\log \pi(z)$  konkav wählt. Insbesondere kann es von Vorteil sein, die Funktion  $\rho$  von vorher konvex zu wählen. Konkavität der Zielfunktion allein genügt aber nicht, damit die iterative Maximierung bezüglich einer Komponente gegen das eindeutige Maximum konvergiert: Man kann leicht Gegenbeispiele mit zwei Punkten und  $\rho(z) = |z|$  konstruieren. Für die Konvergenz braucht man noch Differenzierbarkeit.

Dass in der Regel viele lokale Maxima existieren, sieht man an folgendem Beispiel.

**Beispiel 3.1** Seien  $(z_x)$  und  $(y_x)$  binär (wir betrachten also schwarz/weiss Bilder). Weiter soll gelten:

- $y_x = z_x$  mit Wahrscheinlichkeit  $p$  und  $y_x = 1 - z_x$  mit Wahrscheinlichkeit  $1 - p$ .  
Dann ist

$$\log p(y_x | z_x) = \log(1 - p) + 1_{[y_x = z_x]} \log(p/(1 - p)).$$

Für verschiedene  $x$ , sind die  $y_x$  bedingt unabhängig.

- $\pi(z)$  ist ein Markovfeld:

$$\pi(z) \propto \exp\left(-\beta \sum_{x \sim x'} 1_{[z_x = z_{x'}]}\right)$$

Dann ist die a posteriori Verteilung

$$\log p(z | y) = \gamma \sum_x 1_{[z_x = y_x]} - \beta \sum_{x \sim x'} 1_{[z_x = z_{x'}]} + \text{const.},$$

wobei  $\gamma = \log(p/(1-p))$ . Für  $\beta > 0$  und  $\gamma > 0$  betrachten wir das folgende beobachtete Bild  $y$ :

Dann ist  $z = y$  ein lokales Maximum, aber die folgende Konfiguration  $z$  hat höhere a posteriori Wahrscheinlichkeit, falls  $3\gamma < 2\beta$  ist

Um aus lokalen Maxima herauszukommen, verwendet man randomisierte Algorithmen. Die einfachste Form eines solchen Algorithmus' ist die folgende. Man hält iterativ alle Werte ausser an einem Punkt  $x$  fest, und man wählt einen Kandidaten  $\zeta$  für  $z_x$  zufällig. Dann setzt man

$$\hat{z}_x^{\text{neu}} = \begin{cases} \zeta & \text{mit Wahrscheinlichkeit} & \min\left(1, \left(\frac{p(\zeta, z_{L \setminus \{x\}}^{\text{alt}} | y)}{p(\hat{z}_x^{\text{alt}} | y)}\right)^{1/T}\right) \\ \hat{z}_x^{\text{alt}} & \text{mit Wahrscheinlichkeit} & 1 - \min\left(1, \left(\frac{p(\zeta, z_{L \setminus \{x\}}^{\text{alt}} | y)}{p(\hat{z}_x^{\text{alt}} | y)}\right)^{1/T}\right) \end{cases}$$

Wenn der Kandidat die a posteriori Wahrscheinlichkeit vergrössert, nimmt man ihn auf jeden Fall, sonst nur mit einer gewissen Wahrscheinlichkeit, die davon abhängt,

wieviel kleiner die a posteriori Wahrscheinlichkeit ist.  $T = \infty$  entspricht einer reinen Zufallssuche und  $T = 0$  einem iterativen Maximieren wie vorher. Man sucht ein Kompromiss:  $T \rightarrow 0$  langsam im Iterationsprozess (simuliertes annealing).

Die Berechnung des a posteriori Erwartungswert wird mit Simulation durchgeführt. Um zu simulieren, kann man gleich vorgehen wie bei simuliertem annealing, aber mit  $T = 1$  fest. Für den Erwartungswert wird über die verschiedenen  $\hat{z}_x^{(j)}$  gemittelt:

$$\hat{z}_x = \frac{1}{R} \sum_{j=1}^R \hat{z}_x^{(j)}.$$

### 3.3.3 Schätzung der Parameter

Wenn es unbekannte Parameter in  $p(y|z)$  bzw.  $\pi(z)$  gibt, müssen diese geschätzt, bzw. festgelegt werden. Dies kann auf mehrere Arten geschehen:

- Durch Kalibrierungsexperimente: Man kann z.B. ein typisches, unverraushtes Bild nehmen, Rauschen dazu addieren und dann testen, welche a priori Verteilungen zu guten Rekonstruktionen führen.
- Aus dem beobachteten  $y$  allein, z.B. mit Kreuzvalidierung, Maximierung der marginalen Likelihood oder mit hierarchischen Bayes-Verfahren.

Die Verfahren, die rein auf den beobachteten Daten basieren, sind allerdings alle schwierig zu berechnen. Zur Vereinfachung nehmen wir an, dass nur die a priori Dichte  $\pi_\theta(z)$  von einem unbekanntem Parameter  $\theta$  abhängt, während die bedingte Dichte  $p(y|z)$  bekannt ist (was in unseren Anwendungen mindestens approximativ richtig ist).

Die marginale Likelihood ist definiert als

$$p_\theta(y) = \int p(y | z) \pi_\theta(z) dz,$$

und man kann versuchen,  $\theta$  so zu bestimmen, dass diese marginale Likelihood maximal wird. Für  $\pi_\theta(z) \propto \exp(-\theta U(z))$ , ist

$$p_\theta(y) = \frac{\int p(y | z) \exp(-\theta U(z)) dz}{\int \exp(-\theta U(z)) dz}.$$

Man muss also den Quotienten von zwei Normierungskonstanten bestimmen, was höchstens approximativ möglich ist.

Bei der Kreuzvalidierung maximiert man nicht die marginale Likelihood, sondern analog zur Pseudo-Likelihood das Produkt

$$\prod_{x \in L} p_\theta(y_x | y_{x'}, x' \neq x).$$

Wenn wir wie zuvor annehmen, dass  $p(y | z) = \prod_x q(y_x | z)$ , dann gilt

$$p_\theta(y_x | y_{x'}, x' \neq x) = \int q(y_x | z) p_\theta(z | y_{x'}, x' \neq x) dz = \frac{\int q(y_x | z) \prod_{x' \neq x} q(y_{x'} | z) \pi_\theta(z) dz}{\int \prod_{x' \neq x} q(y_{x'} | z) \pi_\theta(z) dz}.$$

Dies kann man wieder auffassen als den Quotienten von zwei Normierungskonstanten. Man kann auch schreiben

$$p_\theta(y_x | y_{x'}, x' \neq x) = \frac{1}{\int \frac{1}{q(y_x|z)} \pi_\theta(z | y) dz}.$$

Wenn man gemäss der a posteriori Verteilung  $\pi_\theta(z | y)$  simuliert, um den a posteriori Erwartungswert zu berechnen, dann kann man damit gleichzeitig auch noch das Kreuzvalidierungskriterium approximieren. Eine andere mögliche Approximation lautet

$$p_\theta(y_x | y_{x'}, x' \neq x) \approx q(y_x | \hat{z}^{(-x)}),$$

wobei

$$\hat{z}^{(-x)} = \arg \max p_\theta(z | y_{x'}, x' \neq x) = \arg \max_{x' \neq x} \left( \sum \log(q(y_{x'} | z)) + \log(\pi_\theta(z)) \right)$$

der MAP-Schätzer ohne die Beobachtung  $y_x$  ist. Dies kann man näherungsweise berechnen, indem ausgehend vom MAP iterativ eine Komponente von  $z$  maximiert und die andern Komponenten festhält.

Als Bayesianer würde man für die Schätzung von  $\theta$  eine weitere a priori Dichte  $p_0(\theta)$  für  $\theta$  annehmen. Man spricht dann von einem hierarchischen Modell: Auf der untersten Stufe ist das beobachtete Bild  $y$ , auf der mittleren Stufe das ungestörte Bild  $z$  und zuoberst der Parameter  $\theta$ . Die gemeinsame Dichte von  $(\theta, z, y)$  ist

$$p(\theta, z, y) = p_0(\theta) \pi_\theta(z) p(y|z),$$

und die gemeinsame a posteriori Dichte von  $\theta$  und  $z$  gegeben  $y$  ist

$$p(\theta, z | y) \propto p_0(\theta) \pi_\theta(z) p(y|z).$$

Diese a posteriori Dichte möchte man nun als Bayesianer simultan bezüglich  $\theta$  und  $z$  maximieren, bzw. man möchte den a posteriori Erwartungswert

$$\hat{z}_x = \mathbf{E} [z_x | y] = \int_{S^L} \int_{\Theta} z_x p(\theta, z | y) d\theta dz = \int_{\Theta} \int_{S^L} z_x p(z | \theta, y) dz p(\theta | y) d\theta$$

berechnen. Diese Grössen sind aber im Allgemeinen noch schwieriger zu berechnen, weil jetzt zum hochdimensionalen  $z$  noch eine zusätzliche Variable auftritt.

Im folgenden wichtigen Spezialfall kann man die Normierungskonstante angeben: Wenn  $S = \mathbb{R}$  und  $U(z)$  homogen ist vom Grad  $p$ , d.h.  $U(cz) = c^p U(z)$ , dann folgt mit einer Substitution, dass die Normierung  $N(\theta)$  der Gibbs-Verteilung

$$\pi_\theta(z) = N(\theta)^{-1} \exp(-\theta U(z))$$

die Form hat

$$N(\theta) = \theta^{|L|/p} N(1).$$

Ein wichtiges Beispiel ist

$$U(z) = \sum_{x \sim x'} |z_x - z_{x'}|^p.$$

Allerdings führt das auf eine uneigentliche Verteilung: Weil  $U(z)$  konstant ist in der Richtung  $(1, 1, \dots, 1)$ , gilt  $N(\theta) = \infty$  für alle  $\theta$ . Wenn man  $\pi_\theta$  als Dichte auf dem Unterraum der Zuwächse von  $(Z_x)$  auffasst, dann erhält man mit dem gleichen Argument

$$N(\theta) = \theta^{(|L|-1)/p} N(1).$$

Für solche a priori Dichten kann man damit die gemeinsame a posteriori Dichte  $p(\theta, z|y)$  maximieren, bzw. gemäss dieser Dichte simulieren.



# Kapitel 4

## Punktmuster und zufällige Mengen

### 4.1 Grundlegende Definitionen für Punktmuster

**Definition 4.1** Ein Punktmuster ist eine zufällige, lokal endliche Teilmenge  $Z = \{x_1, x_2, \dots\}$  des  $\mathbb{R}^d$  mit  $x_i \neq x_j$  für alle  $i \neq j$ . "Lokal endlich" heisst, dass es in jedem beschränkten  $B$  nur endlich viele Punkte  $x_i$  gibt.

Die Anzahl Punkte in  $B$ ,  $B \subset \mathbb{R}^d$ , bezeichnen wir mit

$$N(B) = \sum_{i=1}^{\infty} 1_B(x_i) \in \{0, 1, 2, \dots, \infty\}.$$

Für eine strenge Behandlung der Punktmuster braucht man Masstheorie. Hier werden wir einen intuitiven Ansatz benutzen. Unter der Verteilung von  $Z$  verstehen wir die Verteilung von  $(N(B_1), \dots, N(B_k))$  für alle  $k$  und alle beschränkten  $B_1, \dots, B_k$  in  $\mathbb{R}^d$ .

**Definition 4.2** – Ein Punktmuster  $Z$  heisst stationär, falls  $(N(B_1), \dots, N(B_k))$  und  $(N(B_1 + x), \dots, N(B_k + x))$  die gleiche Verteilung haben für alle  $k$ , alle  $B_1, \dots, B_k$  und alle  $x$  in  $\mathbb{R}^d$ .

– Ein Punktmuster  $Z$  heisst isotrop, falls  $(N(B_1), \dots, N(B_k))$  und  $(N(R(B_1)), \dots, N(R(B_k)))$  die gleiche Verteilung haben für alle  $k$ , alle  $B_1, \dots, B_k$  und alle Drehungen  $R$  um Null.

Wie bei Gauss'schen Modellen dient die Stationarität und Isotropie als Rechtfertigung für Mittelung zur Schätzung.

**Definition 4.3** Die Intensität einer Menge  $B$  in  $\mathbb{R}^d$  ist definiert durch

$$\Lambda(B) = \mathbf{E}[N(B)] \in [0, \infty]$$

Wenn  $\Lambda(B) < \infty$  ist für alle beschränkten Mengen  $B$ , dann sagen wir, der Punktprozess habe endliche Intensität.  $\Lambda$  ist ein Mass auf der  $\sigma$ -Algebra der Borelmengen.

Die Additivität ist klar, und die  $\sigma$ -Additivität

$$\Lambda \left( \bigcup_{i=1}^{\infty} B_i \right) = \sum_{i=1}^{\infty} \Lambda(B_i) \quad \text{für } B_i \cap B_j = \emptyset, \forall i \neq j.$$

folgt aus monotoner Konvergenz, weil

$$0 \leq N \left( \bigcup_{i=1}^n B_i \right) \nearrow N \left( \bigcup_{i=1}^{\infty} B_i \right)$$

eine monoton wachsende Folge ist und damit die Erwartungswerte konvergieren.

Im stationären Fall gilt

$$\Lambda(B+x) = \Lambda(B).$$

Daraus folgt aus einem Satz der Masstheorie, dass  $\Lambda$  ein Vielfaches  $\lambda (\in [0, \infty])$  des Lebesguemasses ist. Wir nennen  $\lambda$  ebenfalls die Intensität.

Intuitiv nimmt man an, dass für  $\Lambda(B) \rightarrow 0$  gilt

- $P[N(B) = 1] \approx \Lambda(B)$ ,
- $P[N(B) = 0] \approx 1 - \Lambda(B)$ ,
- $P[N(B) \geq 2] = o(\Lambda(B))$ .

Die Bedingungen dafür, dass dies korrekt ist, sind relativ kompliziert. Sie sind erfüllt im stationären Fall falls  $0 < \lambda < \infty$ .

## 4.2 Das Poisson-Punktmuster

Das Poisson-Punktmuster ist das einfachste Modell, bei dem “alle Punkte unabhängig” sind:

**Definition 4.4 Poisson Punktmuster:** Sei  $\Lambda$  ein beliebiges Mass auf  $\mathbb{R}^d$  mit  $\Lambda(\{x\}) = 0$  für alle  $x$ .  $Z$  heisst ein Poisson Punktmuster mit Intensität  $\Lambda$  falls

- $N(B) \sim \text{Pois}(\Lambda(B))$  für alle  $B$ .
- $(N(B_1), N(B_2), \dots, N(B_k))$  unabhängig, falls  $B_i \cap B_j = \emptyset$  für alle  $i \neq j$ .

**Satz 4.1** Ein Poisson Punktmuster existiert, falls  $\Lambda(B) < \infty$  für alle beschränkten Mengen  $B$  ist.

**Beweis:** Es handelt sich um einen konstruktiven Beweis, man kann so ein Poisson-Punktmuster simulieren. Wir überdecken  $\mathbb{R}^d$  mit abzählbar vielen paarweise disjunkten und beschränkten Mengen  $A_i$  und konstruieren  $Z \cap A_i$ , unabhängig für  $i = 1, 2, \dots$ , gemäss folgendem Rezept:

- $N_i := N(A_i) \sim \text{Pois}(\Lambda(A_i))$ ,
- gegeben  $N_i = n$ , wähle  $n$  Punkte  $x_1, x_2, \dots, x_n$  i.i.d mit

$$P[x_i \in B] = \frac{\Lambda(A_i \cap B)}{\Lambda(A_i)}$$

Nachweis der Definition des Poisson Punktmusters:

– Zuerst sei  $B \subset A_i$ :

$$\begin{aligned}
 \mathbb{P}[N(B) = k] &= \sum_{n=k}^{\infty} \mathbb{P}[N(B) = k | N(A_i) = n] \cdot \mathbb{P}[N(A_i) = n] \\
 &= \sum_{n=k}^{\infty} \binom{n}{k} \left( \frac{\Lambda(B)}{\Lambda(A_i)} \right)^k \left( 1 - \frac{\Lambda(B)}{\Lambda(A_i)} \right)^{n-k} \cdot \exp(-\Lambda(A_i)) \frac{\Lambda(A_i)^n}{n!} \\
 &= \exp(-\Lambda(A_i)) \frac{\Lambda(B)^k}{k!} \sum_{n=k}^{\infty} \frac{1}{(n-k)!} (\Lambda(A_i) - \Lambda(B))^{n-k} \\
 &= \exp(-\Lambda(A_i)) \frac{\Lambda(B)^k}{k!} \exp(\Lambda(A_i) - \Lambda(B)) \\
 &= \exp(-\Lambda(B)) \frac{\Lambda(B)^k}{k!}
 \end{aligned}$$

Für ein allgemeines  $B$  benutzen wir

$$N(B) = \sum_i N(B \cap A_i).$$

Da  $N(B \cap A_i) \sim \text{Pois}(\Lambda(B \cap A_i))$  ist, und die einzelnen Summanden nach Konstruktion unabhängig sind, folgt, dass die Summe  $N(B)$  wieder Poisson verteilt ist.

– Sei  $B_1 \subset A_i$ ,  $B_2 \subset A_i$  mit  $B_1 \cap B_2 = \emptyset$ . Mit  $B_3 = A_i \setminus (B_1 \cup B_2)$  folgt

$$\begin{aligned}
 &\mathbb{P}[N(B_1) = k_1, N(B_2) = k_2] \\
 &= \sum_{n=k_1+k_2}^{\infty} \mathbb{P}[N(B_1) = k_1, N(B_2) = k_2 | N(A_i) = n] \cdot \mathbb{P}[N(A_i) = n] \\
 &= \sum_{n=k_1+k_2}^{\infty} \frac{n!}{k_1! k_2! (n - k_1 - k_2)!} \left( \frac{\Lambda(B_1)}{\Lambda(A_i)} \right)^{k_1} \left( \frac{\Lambda(B_2)}{\Lambda(A_i)} \right)^{k_2} \left( \frac{\Lambda(B_3)}{\Lambda(A_i)} \right)^{n-k_1-k_2} \\
 &\quad \cdot \exp(-\Lambda(A_i)) \frac{\Lambda(A_i)^n}{n!} \\
 &= \exp(-\Lambda(A_i)) \frac{\Lambda(B_1)^{k_1}}{k_1!} \frac{\Lambda(B_2)^{k_2}}{k_2!} \sum_{n=k_1+k_2}^{\infty} \frac{1}{(n - k_1 - k_2)!} \Lambda(B_3)^{n-k_1-k_2} \\
 &= \exp(-\Lambda(B_1) - \Lambda(B_2)) \frac{\Lambda(B_1)^{k_1}}{k_1!} \frac{\Lambda(B_2)^{k_2}}{k_2!} = \mathbb{P}[N(B_1) = k_1] \mathbb{P}[N(B_2) = k_2].
 \end{aligned}$$

Der Rest des Beweises geht analog wie oben.

□

### 4.3 Zweite Momente und andere Kennzahlen

**Definition 4.5** Das zweite Momentenmass  $\Lambda^{(2)}$  ist definiert als

$$\Lambda^{(2)}(B_1 \times B_2) = \mathbf{E}[N(B_1)N(B_2)].$$

$\Lambda^{(2)}$  ist ein Mass auf  $\mathbb{R}^{2d}$ , nämlich die Intensität des Produktmusters  $((x_i, x_j) \in \mathbb{R}^{2d} \mid i = 1, 2, \dots; j = 1, 2, \dots)$ .

Insbesondere gilt

$$\text{Cov}(N(B_1), N(B_2)) = \Lambda^{(2)}(B_1 \times B_2) - \Lambda(B_1)\Lambda(B_2)$$

**Beispiel 4.1** Berechnung von  $\Lambda^{(2)}$  beim Poisson Punktmuster.

$$\begin{aligned} \text{Cov}(N(B_1), N(B_2)) &= \text{Cov}(N(B_1 \cap B_2) + N(B_1 \cap B_2^c), N(B_1 \cap B_2) + N(B_1^c \cap B_2)) \\ &= \text{Var}(N(B_1 \cap B_2)) + 0 + 0 + 0 = \Lambda(B_1 \cap B_2) \end{aligned}$$

(die 3 Terme sind null wegen der zweiten Eigenschaft des Poisson-Punktmusters).  
Es folgt:

$$\Lambda^{(2)}(B_1 \times B_2) = \Lambda(B_1 \cap B_2) + \Lambda(B_1)\Lambda(B_2)$$

Nicht nur im Poisson-Fall, sondern für alle Punktmuster hat  $\Lambda^{(2)}$  stets den Anteil  $\Lambda$  auf der Diagonalen. Insbesondere hat  $\Lambda^{(2)}$  nie eine Dichte bezüglich des Lebesgue-masses auf  $\mathbb{R}^{2d}$ . Für die meisten Modelle hat man aber ausserhalb der Diagonalen eine Dichte:

$$\Lambda^{(2)}(B_1 \times B_2) = \Lambda(B_1 \cap B_2) + \int_{B_1} \int_{B_2} \lambda_2(x, x') \, dx dx'.$$

Beim Poisson Punktmuster mit  $\Lambda(dx) = \lambda(x)dx$  ist insbesondere

$$\lambda_2(x, x') = \lambda(x)\lambda(x').$$

Die Dichte  $\lambda_2$  hat die folgende Interpretation für  $x \neq x'$ :

$$\lambda_2(x, x') \, dx dx' = \mathbf{P}[N(dx) = 1, N(dx') = 1].$$

Mit  $\lambda(x)dx = \mathbf{P}[N(dx) = 1]$  folgt damit auch

$$\frac{\lambda_2(x, x')dx'}{\lambda(x)} = \mathbf{P}[N(dx') = 1 \mid x \in Z],$$

sowie

$$\int_B \frac{\lambda_2(x, x')dx'}{\lambda(x)} = \mathbf{E}[N(B) \mid x \in Z] - 1_B(x)$$

Wenn  $Z$  stationär ist, so ist  $\lambda(x) \equiv \lambda$  und

$$\lambda_2(x, x') = \lambda_2(x - x'),$$

insbesondere also

$$\mathbf{E}[N(B) \mid x \in Z] = \mathbf{E}[N(B - x) \mid 0 \in Z].$$

Für stationäre Punktmuster definieren wir das reduzierte zweite Momentenmass  $K$  wie folgt:

$$K(B) = \frac{\mathbf{E}[N(B) \mid 0 \in Z] - 1_B(0)}{\lambda} = \frac{\int_B \lambda_2(x)dx}{\lambda^2}.$$

Für das stationäre Poisson-Punktmuster ist  $K(B)$  gleich dem Lebesguemass von  $B$ . Wenn der Punktprozess zusätzlich noch isotrop ist, dann gilt

$$\lambda_2(x, x') = \lambda_2(\|x - x'\|),$$

und das reduzierte zweite Momentenmass ist festgelegt durch die sogenannte  $K$ -Funktion

$$K(r) = \frac{\mathbf{E}[N(\{x; 0 < \|x\| < r\}) \mid 0 \in Z]}{\lambda} = \frac{\nu_d}{\lambda^2} \int_0^r s^{d-1} \lambda_2(s) ds,$$

wobei  $\nu_d$  die Oberfläche der Einheitskugel im  $\mathbb{R}^d$  bezeichnet.

Die zweiten Momente und insbesondere die  $K$ -Funktion geben eine Beschreibung der Abhängigkeit in einem Punktmuster. Es ist jedoch wichtig zu bemerken, dass so nur ein Teil der Abhängigkeit erfasst wird. Es gibt Fälle, wo zwei Punktmuster trotz identischen ersten und zweiten Momenten visuell unterschiedliche Realisierungen haben, vgl. Baddeley und Silverman, *Biometrics* 40, 1984.

Aus diesem Grund zieht man auch andere Kenngrößen in Betracht. Die beiden wichtigsten sind die “empty space function”  $F$  und die “nearest neighbor function”  $G$ .  $F$  ist die Verteilungsfunktion des Abstandes von einer festen Stelle  $x \in \mathbb{R}^d$  zum nächsten Punkt des Punktmusters  $Z$ , d.h.  $F(r)$  ist gleich der Wahrscheinlichkeit, dass in der Kugel mit Radius  $r$  und Zentrum  $x$  mindestens ein Punkt des Punktmusters liegt:

$$F(r) = \mathbf{P}[N(\{x' \mid \|x' - x\| \leq r\}) > 0].$$

Für ein stationäres Punktmuster hängt  $F$  nicht von  $x$  ab.

Die Funktion  $G$  ist die Verteilungsfunktion des Abstands von einem  $x_i \in Z$  zum nächsten Punkt des Punktmusters  $Z$ , d.h.  $G(r)$  ist die Wahrscheinlichkeit, dass in der Kugel mit Radius  $r$  und Zentrum  $x_i \in Z$  noch ein weiterer Punkt liegt (im Zentrum der Kugel befindet sich ja schon ein Punkt):

$$G(r) = \mathbf{P}[N(\{x' \mid \|x' - x\| \leq r\}) > 1 \mid x \in Z].$$

Die Form von  $G$  und  $F$  geben zusätzliche Hinweise, wie die Punkte verteilt sind. Meist vergleicht man  $F$  und  $G$  mit dem, was bei einem homogenen Poisson-Muster auftritt. Dort sind die beiden Funktionen gleich, es gilt

$$F(r) = 1 - \exp(-\lambda |\{x \mid \|x\| \leq r\}|) = G(r).$$

## 4.4 Modelle mit Abhängigkeit

### 4.4.1 Cox-Punktmuster

Dies ist ein zweistufiges Modell: Auf der ersten Stufe haben wir ein nichtnegatives Zufallsfeld  $\xi(x)$ ,  $x \in \mathbb{R}^d$ . Auf der zweiten Stufe haben wir – bedingt auf  $\xi$  – ein inhomogenes Poisson-Punktmuster  $Z$  mit Intensität  $\Xi(B) = \int_B \xi(x) dx$ . Dabei wird  $\xi$  nicht beobachtet, nur  $Z$ . So kann das Punktmuster  $Z$  zum Beispiel den Standorten von Pflanzen entsprechen, und  $\xi$  ist dann eine zufällig variierende Umweltbedingung, die den Pflanzenwuchs fördert bzw. hemmt.

Für konkrete Konstruktionen kann man z.B.  $\log \xi$  als ein Gauss'sches Zufallsfeld wählen mit Erwartungswert  $m(x)$  und Kovarianzfunktion  $C(x, x')$ , oder man setzt

$$\xi(x) = \sum_i k(x, s_i),$$

wobei  $\{s_1, s_2, \dots\}$  ein stationäres Poisson-Punktmuster auf  $\mathbb{R}^d$  ist und  $k$  eine Funktion  $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+$  mit  $\int k(x, y) dy < \infty$  für alle  $x$ .

Eigenschaften des Cox-Musters:

- 1)  $Z$  ist stationär, falls  $\xi$  stationär ist.
- 2) Die Intensität ist  $\lambda(x) = \mathbf{E}[\xi(x)]$ , denn mit Fubini gilt

$$\Lambda(B) = \mathbf{E}[N(B)] = \mathbf{E}[\mathbf{E}[N(B) | \xi]] = \mathbf{E}\left[\int_B \xi(x) dx\right] = \int_B \mathbf{E}[\xi(x)] dx$$

- 3) Für das zweite Moment gilt  $\lambda_2(x, x') = \mathbf{E}[\xi(x)\xi(x')] = \text{Cov}(\xi(x), \xi(x')) + \lambda(x)\lambda(x')$ . Zur Begründung beachte man

$$\begin{aligned} \Lambda^{(2)}(B_1 \times B_2) &= \mathbf{E}[\mathbf{E}[N(B_1)N(B_2) | \xi]] \\ &= \mathbf{E}\left[\int_{B_1 \cap B_2} \xi(x) dx + \int_{B_1} \int_{B_2} \xi(x)\xi(x') dx dx'\right] \\ &= \int_{B_1 \cap B_2} \lambda(x) dx + \int_{B_1} \int_{B_2} \mathbf{E}[\xi(x)\xi(x')] dx dx' \end{aligned}$$

Wenn man nur eine Realisierung eines Cox-Musters beobachtet hat, dann ist das nicht unterscheidbar von einem inhomogenen Poisson-Muster. Bei einem inhomogenen Poisson-Muster kann man die unbekannte Intensität auch nichtparametrisch mit einem Kern-Schätzer schätzen:

$$\hat{\xi}(x) = \sum_{x_i \in Z} h^{-d} k\left(\frac{x - x_i}{h}\right).$$

Mit der Annahme, dass  $\xi$  zufällig ist und einem einfachen parametrischen Modell genügt, kann man unter Umständen die Intensität besser schätzen (vor allem an den Rändern).

#### 4.4.2 Cluster-Punktmuster

Diese Modelle, die auch Neyman-Scott Punktmuster heissen, sind wie folgt definiert. Die Zentren  $\{x'_1, x'_2, x'_3, \dots\}$  der Cluster bilden ein stationäres Poisson-Punktmuster mit  $\Lambda(dx) = \lambda_0 \cdot dx$ . Weiter seien die Cluster-Grössen  $M_1, M_2, M_3, \dots$  (mit Werten in  $\mathbb{N}_0$ ) iid, unabhängig von den Cluster-Zentren. Schliesslich sind die relativen Positionen  $D_i$  der Punkte eines Clusters relativ zu ihrem Zentrum iid mit der Dichte  $f$  auf  $\mathbb{R}^d$ . Das heisst, das Punktmuster besteht aus den Punkten

$$x'_1 + D_1, x'_1 + D_2, \dots, x'_1 + D_{M_1}, x'_2 + D_{M_1+1}, \dots, x'_2 + D_{M_1+M_2}, x'_3 + D_{M_1+M_2+1}, \dots$$

Im Beispiel der Standorte von Pflanzen fasst man die Cluster-Zentren als die Positionen der Eltern auf, die  $M_i$ 's als die Anzahl Nachkommen und die  $D_i$ 's als die Distanzen zwischen Nachkommen und Eltern.

Obwohl die Cluster-Modelle auf einer ganz anderen heuristischen Vorstellung beruhen als die Cox-Modelle, sind die beiden Klassen nicht disjunkt: Wenn die  $M_i$ 's nämlich Poisson-verteilt sind (mit Parameter  $\lambda$ ), dann hat das zugehörige Cluster-Punktmuster die gleiche Verteilung wie ein Cox-Punktmuster mit der zufälligen Intensität

$$\xi(x) = \lambda \sum_i f(x - x'_i)$$

wobei  $f$  die Dichte der  $D_i$  bezeichnet. Das sieht man wie folgt ein: Gegeben  $x'_i$  bilden die Punkte des  $i$ -ten Clusters ein inhomogenes Poisson-Punktmuster mit Intensität  $f(x - x'_i)$  (vgl. den Beweis von Satz

Eigenschaften eines Cluster-Punktmusters  $Z$ :

- 1)  $Z$  ist stationär.
- 2) Die Intensität ist  $\lambda = \lambda_0 \mathbf{E}[M_i]$ . Begründung:

$$\begin{aligned} \lambda(x)dx &= \mathbf{P}[N(dx) = 1] = \mathbf{E}[\mathbf{P}[N(dx) = 1 \mid x'_1, M_1, x'_2, M_2, \dots]] \\ &= \mathbf{E}\left[\sum_i M_i f(x - x'_i)\right] dx. \end{aligned}$$

Die Behauptung folgt jetzt, weil die  $M_i$ 's und die  $x'_i$ 's unabhängig sind und weil für eine beliebige messbare positive Funktion  $h$  auf  $\mathbb{R}^d$  und für ein Punktmuster  $Z$  mit Intensität  $\Lambda$  gilt

$$\mathbf{E}\left[\sum_{x_i \in Z} h(x_i)\right] = \int h(x)\Lambda(dx).$$

Dies ist per Definition richtig für Indikatorfunktionen  $h$ , und wegen der Linearität auch für einfache Funktionen. Durch Betrachtung monotoner Limiten kann man es schliesslich auf alle nichtnegativen Funktionen ausdehnen.

- 3)  $\lambda_2(x, y) = \lambda^2 + \lambda_0 \mathbf{E}[M_i(M_i - 1)] \cdot \int f(x - u)f(y - u)du$ . Die Begründung geht analog:

$$\begin{aligned} &\mathbf{P}[N(dx) = 1, N(dy) = 1 \mid x'_1, M_1, x'_2, M_2, \dots] \\ &= \sum_i M_i(M_i - 1) f(x - x'_i)f(y - x'_i)dx dy + \sum_{i \neq j} M_i M_j f(x - x'_i)f(y - x'_j)dx dy \end{aligned}$$

(die erste Summe enthält die Fälle, wo die beiden Punkte an den Stellen  $x$ , bzw.  $y$  aus dem gleichen Cluster kommen, die zweite Summe die Fälle, wo sie aus verschiedenen Clustern kommen).

### 4.4.3 Inhibitionsmodelle

Das Ziel bei den Inhibitionsmodellen ist es, eine "hard-core-Distanz"  $r_0$  einzuhalten: Zwei Punkte des Musters sollen mindestens den Abstand  $r_0$  haben. Es gibt verschiedene Möglichkeiten, dies zu erreichen.

Eine Möglichkeit ist, die Punkte sequentiell zu platzieren unter Beachtung des Minimalabstands  $r_0$ . Wir wählen also  $x_1$  mit Verteilung  $f(x)dx$ , dann  $x_2 \mid x_1$  mit

Verteilung  $c_1 \cdot f(x) \cdot 1_{[\|x-x_1\|>r_0]} dx$ , dann  $x_3 \mid x_2, x_1$  mit Verteilung  $c_2 \cdot f(x) \cdot 1_{[\|x-x_1\|>r_0, \|x-x_2\|>r_0]} dx$ , usw.. Das entstehende Punktmuster ist natürlich nicht stationär.

Eine andere Möglichkeit benutzt Elimination. Man geht aus von einem Poisson-Punktmuster  $\{x'_1, x'_2, x'_3, \dots\}$  mit Parameter  $\lambda_0$ . Dann wird von zwei Punkten, deren Abstand kleiner als  $r_0$  ist, einer eliminiert. Damit dieses Vorgehen eindeutige Resultate produziert, braucht es genaue Vorschriften, in welcher Reihenfolge die Punkte eliminiert werden. Wir können zum Beispiel alle Punkte  $x'_i$  mit iid uniformen Zufallszahlen  $U_i$  versehen und dann bei jedem Konflikt den Punkt mit kleinerem  $U_i$  markieren. Am Schluss werden dann alle markierten  $x'_i$  eliminiert.

Unter Umständen werden dabei mehr Punkte eliminiert, als unbedingt nötig wäre, wie man im folgenden Beispiel sieht: Es seien  $\|x'_1 - x'_2\| < r_0$  und  $\|x'_2 - x'_3\| < r_0$ , während alle andern Abstände, an denen  $x'_1$ ,  $x'_2$  oder  $x'_3$  beteiligt ist, grösser als  $r_0$  sein sollen. Dann folgt:

- Falls nur  $x'_2$  markiert wird, bleiben  $x'_1$  und  $x'_3$  (die Wahrscheinlichkeit ist  $1/3$ ).
- Falls  $x'_1$  und  $x'_3$  markiert werden, bleibt nur  $x'_2$  (die Wahrscheinlichkeit ist  $1/3$ ).
- Falls  $x'_2$  und  $x'_3$  markiert werden, bleibt nur  $x'_1$  (die Wahrscheinlichkeit ist  $1/6$ ).
- Falls  $x'_1$  und  $x'_2$  markiert werden, bleibt nur  $x'_3$  (die Wahrscheinlichkeit ist  $1/6$ ).

Die ersten und zweiten Momente bei diesem Eliminationsmodell lassen sich berechnen. Die Intensität  $\lambda$  ist

$$\lambda = \frac{1 - \exp(-\lambda_0 V_0)}{V_0}$$

mit

$$V_0 = |\{\|x\| \leq r_0\}| = \omega_d r_0^d$$

( $\omega_d$  ist das Volumen der Einheitskugel im  $\mathbb{R}^d$ ). Begründung:

$$\lambda dx = \text{P}[\text{Punkt in } dx] = \lambda_0 dx \text{P}[\text{Punkt überlebt}].$$

Weil die Punkte  $x'_i$  mit  $U_i > u$  ein Poissonmuster mit Intensität  $(1-u)\lambda_0$  bilden, folgt

$$\begin{aligned} \text{P}[\text{Punkt } x'_i \text{ überlebt}] &= \text{P}[\text{Kein anderer Punkt } x'_j \text{ mit grösserem } U \text{ in } \{x \mid \|x - x'_i\| \leq r_0\}] \\ &= \int_0^1 \exp(-(1-u)\lambda_0 V_0) du = \frac{1 - \exp(-\lambda_0 V_0)}{\lambda_0 V_0}. \end{aligned}$$

Im Grenzfall  $\lambda_0 \rightarrow \infty$  konvergiert  $\lambda$  gegen  $\frac{1}{V_0}$ . Mit dem entstehenden Muster erhält man eine nicht-überlappende Packung von Kugeln mit dem Radius  $r_0/2$  im  $\mathbb{R}^d$ . Diese Packung ist aber nicht sehr dicht, der überdeckte Anteil ist nämlich

$$\frac{N(B)\omega_d(r_0/2)^d}{|B|} \sim \frac{\omega_d(r_0/2)^d}{\omega_d r_0^d} = 2^{-d}.$$

Zum Vergleich: beim kubischen Gitter ist der überdeckte Anteil  $\omega_d 2^{-d}$ , für  $d = 2$  ergibt das hexagonale Gitter die dichteste Kugelpackung mit einem Überdeckungsanteil von  $\omega_2/(2\sqrt{3}) = 0.2887 \omega_2$ , und für  $d = 3$  hat das flächenzentrierte kubische Gitter den maximalen Überdeckungsanteil von  $\omega_3/(4\sqrt{2}) = 0.1768 \omega_3$ .

Analog erhält man das zweite Moment (siehe z.B. Stoyan et al., p. 164): Für  $\|x - x'\| \leq r_0$  ist  $\lambda_2(x, x') = 0$ , und für  $\|x - x'\| \geq r_0$  gilt

$$\lambda_2(x, x') = \frac{2(2V_0 - \gamma_0)(1 - \exp(-\lambda_0 V_0)) - 2V_0(1 - \exp(-\lambda_0(2V_0 - \gamma_0)))}{V_0(V_0 - \gamma_0)(2V_0 - \gamma_0)}.$$

Dabei ist  $\gamma_0$  das Volumen des Durchschnitts zweier Kugeln mit Radien  $r_0$  und Zentren  $x$ , bzw.  $x'$ :

$$\gamma_0 = \gamma_0(x, x') = |\{x'' \mid \|x'' - x\| \leq r_0\} \cap \{x'' \mid \|x'' - x'\| \leq r_0\}|$$

( $\gamma_0$  ist gleich 0 falls  $\|x - x'\| \geq 2r_0$  ist).

#### 4.4.4 Gibbs-Modelle

Gibbs-Modelle sind gegeben durch die Wahrscheinlichkeiten, dass an der Stelle  $x$  ein Punkt ist, gegeben das Muster auf  $\mathbb{R}^d \setminus \{x\}$ :

$$\lambda(x \mid z) = \frac{\mathbb{P}[N(dx) = 1 \mid Z \cap (\mathbb{R}^d \setminus \{x\}) = z]}{dx}.$$

Wie im Fall von Gibbs-Modellen auf einem Gitter kann man sich fragen, was für eine Form die bedingten Intensitäten  $\lambda(x \mid z)$  haben müssen, damit ein zugehöriges Punktmuster existiert, und wie dieses Punktmuster dann aussieht.

Wir betrachten Gibbs-Modelle nur auf einer beschränkten Menge  $B \subset \mathbb{R}^d$ . Dann besteht  $Z$  nur aus endlich vielen Punkten. Die Gibbs'sche Form der bedingten Intensitäten lautet dann

$$\lambda(x \mid x_1, \dots, x_n) = \exp \left( -\phi_1(x) - \sum_{k=1}^n \sum_{i_1 < \dots < i_k \leq n} \phi_{k+1}(x, x_{i_1}, \dots, x_{i_k}) \right).$$

Die Potentiale  $\phi_k$  sind dabei symmetrische Funktionen der  $k$  Argumente. Die zugehörige Gibbs-Verteilung wird dann definiert, indem man die Radon-Nikodym Dichte  $f$  bezüglich des Poisson-Punktmusters mit Intensität eins angibt, nämlich

$$f(\{x_1, \dots, x_n\}) = f(\emptyset) \exp \left( -\sum_{k=1}^n \sum_{i_1 < \dots < i_k \leq n} \phi_k(x_{i_1}, \dots, x_{i_k}) \right).$$

Die Konstante  $f(\emptyset)$  ergibt sich dann aus der Normierung

$$\sum_{n=0}^{\infty} \frac{\exp(-|B|)}{n!} \int_{B^n} f(\{x_1, \dots, x_n\}) dx_1 \dots dx_n = 1.$$

Man braucht aber Bedingungen an  $f$ , damit die linke Seite endlich ist.

Das einfachste Beispiel ist  $\phi_k(x) \equiv 0$  für alle  $k > 1$ . Dann ist das zugehörige Gibbs-Punktmuster einfach ein Poisson-Punktmuster mit Intensität  $\exp(-\phi_1(x))$ . Das einfachste nichttriviale Beispiel ist das sogenannte Strauss-Punktmuster. Dort ist  $\phi_1(x) \equiv a$ ,  $\phi_2(x, y) = b1_{\{\|x-y\| \leq r\}}$  und  $\phi_k(x) \equiv 0$  für  $k > 2$ . Für  $b > 0$  hat man Abstossung: Je mehr Punkte im Abstand  $\leq r$  von  $x$  schon vorhanden sind, desto kleiner ist die Wahrscheinlichkeit für einen weiteren Punkt an der Stelle  $x$ . Für

$b = +\infty$  erhält man ein Inhibitionsmodell. Für  $b < 0$  hat man umgekehrt Anziehung. Allerdings gibt es im Fall  $b < 0$  gar kein zugehöriges Gibbs-Modell, weil die Normierung unendlich ist. Man behilft sich dann damit, dass man das Modell bedingt auf  $N(B) = n$  betrachtet: Dann muss man nur die gemeinsame Dichte  $f_n$  der  $n$  Punkte angeben, welche für das Strauss-Modell proportional ist zu

$$\exp \left( -b \sum_{i < j} 1_{\{\|x_i - x_j\| \leq r\}} \right).$$

Eine andere Lösung ist, dass man

$$\phi_2(x, y) = \begin{cases} b_1 > 0 & \|x - y\| < r_0 \\ b_2 < 0 & r_0 \leq \|x - y\| < r \\ 0 & \|x - y\| \geq r \end{cases}$$

setzt, d.h. in einer kleinen Distanz eine Abstossung einführt.

## 4.5 Statistik für Punktmuster

Wir beobachten ein stationäres und isotropes Punktmuster  $Z$  auf einem beschränkten Beobachtungsfenster  $W: \{x_1, x_2, \dots, x_{N(W)}\}$ . Zusätzlich gibt es Punkte ausserhalb  $W$ , die nicht beobachtet sind.

Wir wollen nun folgende Fragestellungen untersuchen:

1. Wie findet man nichtparametrische Schätzungen für die Kenngrößen  $\lambda$ ,  $K$ ,  $\lambda_2$ ,  $G$ ,  $F$ ?
2. Wie schätzt man Parameter in einem parametrischen Modell? (z.B. bei einem Cox Muster mit lognormaler stationärer Intensität  $\xi$ , wenn man ein parametrisches Modell für die Kovarianzfunktion von  $\log(\xi)$  verwendet).

### 4.5.1 Nichtparametrische Schätzung von Kenngrößen

Wegen der Stationarität ist

$$\hat{\lambda} = \frac{N(W)}{|W|}$$

die naheliegende Schätzung der Intensität. Sie ist stets erwartungstreu, aber ihre Varianz hängt von zweiten Moment ab:

$$\begin{aligned} \text{Var}(\hat{\lambda}) &= \frac{1}{|W|^2} \text{Var}(N(W)) = \frac{1}{|W|^2} \left( \mathbf{E}[N(W)^2] - \mathbf{E}[N(W)]^2 \right) \\ &= \frac{1}{|W|^2} (\Lambda^{(2)}(W \times W) - \lambda^2 |W|^2) \\ &= \frac{1}{|W|^2} \left( \lambda |W| + \int_W \int_W \lambda_2(\|x - x'\|) dx dx' - \lambda^2 |W|^2 \right) \\ &= \frac{\lambda}{|W|} + \frac{1}{|W|^2} \int_W \int_W (\lambda_2(\|x - x'\|) - \lambda^2) dx dx'. \end{aligned}$$

Es gilt

$$\left| \int_W \int_W (\lambda_2(\|x - x'\|) - \lambda^2) dx dx' \right| \leq |W| \int_{\mathbb{R}^d} |\lambda_2(\|x\|) - \lambda^2| dx.$$

Die Varianz ist also umgekehrt proportional zur Grösse des Beobachtungsfensters, wenn  $\lambda_2(\|x\|) - \lambda^2$  integrierbar ist.

Die Schätzung der andern Kenngrössen ist schwieriger. Auf Grund der Definition von  $K$ ,  $\lambda_2$ ,  $G$  und  $F$  sind die folgenden Schätzer naheliegend ( $B(x, r)$  bezeichnet die Kugel mit Mittelpunkt  $x$  und Radius  $r$ ):

$$\begin{aligned} \widehat{K}(r) &= \frac{1}{\widehat{\lambda}N(W)} \sum_{i=1}^{N(W)} (N(B(x_i, r) \cap W) - 1) = \frac{1}{\widehat{\lambda}N(W)} \sum_{i=1}^{N(W)} \sum_{j \neq i} 1_{[\|x_i - x_j\| \leq r]} \\ \widehat{G}(r) &= \frac{1}{N(W)} \sum_{i=1}^{N(W)} 1_{[N(B(x_i, r) \cap W) > 1]} \\ \widehat{F}(r) &= \frac{1}{|W|} |\{x \in W \mid N(B(x, r) \cap W) > 0\}| = \frac{|\cup_i B(x_i, r) \cap W|}{|W|} \\ \widehat{\lambda}_2(r) &= \frac{1}{N(W)^2} \sum_{i \neq j} \frac{1}{h^d} k\left(\frac{\|x_i - x_j\| - r}{h}\right). \end{aligned}$$

Der Schätzer  $\widehat{\lambda}_2(r)$  ist einfach ein Kernschätzer für eine rotationsinvariante Dichte auf  $\mathbb{R}^d$ .

Alle obigen Schätzer unterschätzen aber die wahren Grössen zum Teil massiv wegen der nicht beobachteten Punkte ausserhalb  $W$ . Eine mögliche Verbesserung ist die Reduktion des Beobachtungsfensters: Anstelle von  $W$  verwendet man das reduzierte Fenster

$$W_{\ominus r} = \{x \in W \mid B(x, r) \subset W\}.$$

Dann ist

$$\widehat{K}(r) = \frac{1}{\widehat{\lambda}N(W_{\ominus r})} \sum_{i=1}^{N(W_{\ominus r})} \sum_{j \neq i} 1_{[\|x_i - x_j\| \leq r]} = \frac{1}{\widehat{\lambda}N(W_{\ominus r})} \sum_{i=1}^{N(W_{\ominus r})} (N(B(x_i, r)) - 1)$$

und analog für die anderen Grössen.

Dies bedeutet jedoch, dass man die vorliegende Information nicht voll ausnutzt. Ferner sind  $\widehat{F}$  und  $\widehat{G}$  nicht mehr immer monoton. Es gibt bessere Methoden.

Für  $K$  gibt es eine einfache Gewichtung zur Biaskorrektur:

$$\widehat{K}(r) = \frac{1}{\widehat{\lambda}N(W)} \sum_{i=1}^{N(W)} \sum_{j \neq i} 1_{[\|x_i - x_j\| \leq r]} w_{ij}$$

mit

$$w_{ij} = \frac{2\pi \cdot \|x_j - x_i\|}{\text{Länge von } (\{x \mid \|x - x_i\| = \|x_j - x_i\|\} \cap W)},$$

d.h. man gewichtet mit dem Inversen des Anteils des Kreisumfangs, der in  $W$  enthalten ist. Paare in der Nähe des Randes erhalten so ein grösseres Gewicht, das im Mittel die nicht-beobachteten Punkte kompensiert.

Die so erhaltene Schätzung ist bis auf den Nenner erwartungstreu:

**Satz 4.2** *Es gilt:*

$$\mathbf{E} \left[ \widehat{\lambda}^2 \widehat{K}(r) \right] = \lambda^2 K(r)$$

**Beweis:** Weggelassen □

Für die Schätzung von  $G$  benützt man eine Analogie zur Überlebenszeitanalyse. Wir definieren

$$T_i = \min_j \|x_i - x_j\|.$$

Dieser Abstand von  $x_i$  zu demjenigen Punkt des Musters, der am nächsten liegt, kann nicht immer beobachtet werden, weil der nächste Punkt des Musters auch ausserhalb des Beobachtungsfensters  $W$  liegen kann. Wenn wir noch den kleinsten Abstand vom Punkt  $x_i$  zum Rand des Beobachtungsfensters  $W$ ,

$$C_i = \min_{x \notin W} (\|x - x_i\|),$$

eingeführen, dann beobachten wir nur  $Y_i = \min(T_i, C_i)$  zusammen mit dem Indikator  $\Delta_i = 1_{[T_i \leq C_i]}$ , der angibt, ob  $T_i$  beobachtet wurde.

In der Überlebenszeitanalyse kommen alle obigen Grössen auch vor, haben aber eine etwas andere Bedeutung:  $T_i$  ist dort die Lebenszeit des Individuums  $i$ ,  $C_i$  ist die Zeit bis zum Abbruch der Studie, bzw. bis zum vorzeitigen Ausscheiden des Patienten  $i$ . Auch hier ist  $Y_i = \min(T_i, C_i)$  die beobachtete Grösse und  $\Delta_i = 1_{[T_i \leq C_i]}$  die Indikatorfunktion, die die Zensierung anzeigt. In der Überlebenszeitanalyse gibt es einen Standardschätzer, um die Verteilungsfunktion  $F(\cdot)$  von  $T$  auf Grund der zensierten Daten  $(Y_i, \Delta_i)$  zu schätzen, den sogenannten Kaplan-Meier Schätzer. Dieser lässt sich direkt anwenden zur Schätzung von  $G$ .

### Der Kaplan-Meier-Schätzer

Für eine absolut stetige Verteilung  $F$  auf  $[0, \infty)$  mit Dichte  $f$  ist die Sterberate, bzw. Hazardfunktion  $\lambda$  definiert als

$$\lambda(t) = \lim_{h \rightarrow 0} \frac{\mathbf{P}[t \leq T \leq t+h \mid T \geq t]}{h} = \lim_{h \rightarrow 0} \frac{\mathbf{P}[t \leq T \leq t+h]}{h \mathbf{P}[T \geq t]} = \frac{f(t)}{1-F(t)} = -\frac{d}{dt} \log(1-F(t)).$$

Der letzte Ausdruck zeigt, wie man  $F$  aus der Sterberate berechnen kann:

$$F(t) = 1 - \exp\left(-\int_0^t \lambda(s) ds\right).$$

Falls  $F$  diskret ist und die Werte  $0 \leq t_1 < t_2 < \dots$  annimmt, so ist die Hazardfunktion  $\lambda(\cdot)$  ebenfalls diskret:

$$\lambda(t_i) = \frac{\mathbf{P}[T = t_i]}{\mathbf{P}[T \geq t_i]} = \frac{F(t_i) - F(t_i-)}{1 - F(t_i-)},$$

und es gilt dann

$$F(t) = 1 - \prod_{t_i \leq t} (1 - \lambda(t_i)).$$

**Beweis:** Es gilt

$$1 - F(t) = P[T > t] = P[T > t_k]$$

falls  $t_k \leq t < t_{k+1}$ . Wir bilden dann das Produkt

$$1 - F(t) = \frac{P[T > t_k]}{P[T > t_{k-1}]} \cdot \frac{P[T > t_{k-1}]}{P[T > t_{k-2}]} \cdot \dots \cdot \frac{P[T > t_1]}{1}.$$

Daraus folgt die Behauptung, denn es gilt

$$\frac{P[T > t_k]}{P[T > t_{k-1}]} = \frac{P[T \geq t_k] - P[T = t_k]}{P[T \geq t_k]} = 1 - \lambda(t_k).$$

□

Die Kaplan-Meier-Schätzung ergibt eine diskrete Verteilung, die auf der Menge der beobachteten, unzensierten  $T_i$ 's konzentriert ist. Sie schätzt die Hazardfunktion an diesen Stellen einfach mit der naheliegenden relativen Häufigkeit

$$\hat{\lambda}(t_i) = \frac{|\{j | Y_j = t_i, \Delta_j = 1\}|}{|\{j | Y_j \geq t_i\}|}$$

und verwendet dann obige Produktformel.

#### 4.5.2 Schätzung in parametrischen Modellen

Die Maximum-Likelihood-Schätzung ist im Allgemeinen schwierig zu berechnen, da sie für die hier besprochenen Modelle typischerweise ein sehr kompliziertes Integral enthält (im stetigen Fall), bzw. eine komplizierte Summe (im diskreten Fall). Einfacher ist die Momentenschätzung:

$$\hat{\theta} = \arg \min_{\theta} \int_{r_1}^{r_2} \left| \hat{K}(r)^\alpha - K_\theta(r)^\alpha \right| w(r) dr,$$

wobei man den Bereich  $[r_1, r_2]$  und den Exponenten  $\alpha$  im Prinzip frei wählen kann. Oft verwendet man  $\alpha = 1/2$ , weil dann die Varianz von  $\hat{K}(r)$  für alle  $r$  etwa gleich ist. Diese Methode bedeutet einfach, dass die nichtparametrische Schätzung  $\hat{K}$  möglichst gut mit der Funktion  $K$  im angepassten Modell übereinstimmen soll. Oft führt man die Minimierung noch unter der Nebenbedingung durch, dass die Intensitäten  $\hat{\lambda}$  und  $\lambda_{\hat{\theta}}$  gleich sein sollen.

## 4.6 Zufällige Mengen

Sei  $Z$  eine zufällige, abgeschlossene Teilmenge des  $\mathbb{R}^d$ . Deren Konstruktion ist masstheoretisch anspruchsvoll, wir gehen jedoch nicht darauf ein. Stationarität einer zufälligen Menge bedeutet, dass  $Z$  und  $Z + x$  die gleichen Verteilungen haben sollen. Wie bei den Punktmustern kann man zufällige Mengen charakterisieren durch Kennzahlen. Die folgenden Definitionen und Formeln gelten für den stationären Fall.

- Der Volumenanteil  $p$  beschreibt, wie gross der Anteil “weiss/schwarz” im Bild ist:

$$p = P[x \in Z] = \frac{\mathbf{E}[|Z \cap B|]}{|B|}.$$

– Die Kovarianzfunktion ist

$$C(h) = P[\{x \in Z\} \cap \{x + h \in Z\}].$$

– Kontakt-Verteilungen  $H_B(\cdot)$  messen, wie stark die beiden Phasen “ineinander verzahnt” sind,

$$H_B(r) = 1 - \frac{P[Z \cap rB = \emptyset]}{1 - p},$$

wo  $B$  eine Testmenge, wie z.B. ein Kreis, Kreissegment, Polygon, etc. ist. Diese Definition wird klarer, wenn wir die Zufallsvariable

$$R = \inf\{s \mid Z \cap sB \neq \emptyset\},$$

betrachten. Da  $R = 0$  bedeutet “ $0 \in Z$ ”, hat  $R$  im Allgemeinen eine Punktmasse bei 0. Wegen

$$P[R > r \mid R > 0] = \frac{P[Z \cap rB = \emptyset]}{1 - p} = 1 - H_B(r)$$

ist  $H_B$  also die bedingte Verteilungsfunktion von  $R$ , gegeben dass  $R > 0$ .

#### 4.6.1 Modelle für zufällige Mengen

Am einfachsten zu behandeln sind Boole’sche Modelle

$$Z = \bigcup_{i=1}^{\infty} \{Z_i + x_i\},$$

wo  $\{x_1, x_2, x_3, \dots\}$  ein homogenes Poisson-Punktmuster mit Parameter  $\lambda$  ist und  $Z_1, Z_2, \dots$  eine iid Folge von zufälligen, kompakten Mengen, unabhängig vom Punktmuster. Man kann z.B. Kugeln mit zufälligem Radius nehmen:  $Z_i = B(0, R_i)$  mit  $R_i$  iid, oder Segmente von zufälliger Länge und Richtung. Für beliebiges, kompaktes  $K$  gilt (ohne Beweis)

$$P[Z \cap K = \emptyset] = \exp(-\lambda \cdot \mathbf{E}[|Z_i \oplus \check{K}|]),$$

wo  $\check{K} = \{x \mid -x \in K\}$  der Spiegelung von  $K$  in 0 entspricht, und  $A \oplus B = \{x + x' \mid x \in A, x' \in B\}$ . Wählt man  $K = \{0\}$ , so folgt daraus, dass der Volumenanteil gleich

$$1 - \exp(-\lambda \cdot \mathbf{E}[|Z_i|])$$

ist. Analog kann man auch die Kovarianzfunktion berechnen.

Leider sind die Boole’schen Modelle häufig nicht realistisch, denn die wahren zufälligen Mengen sind keine Vereinigungen von Kreisen oder sonstigen “einfachen” Körpern. Eine Möglichkeit, realistischere Modelle zu erhalten, ist darum ein “Verschmieren” von Boole’schen Modellen mit den Öffnungs- bzw. Abschluss-Operationen der mathematischen Morphologie.



## Kapitel 5

# Hinweise zur Literatur

Das Buch Cressie (1993) versucht, die ganze räumliche Statistik abzudecken. Entsprechend ist es 900 Seiten dick, und trotzdem hilft es oft nicht weiter, wenn man etwas genau wissen will. Die 70 Seiten Literatur sind aber auf jeden Fall nützlich. Das Buch Ripley (1981) ist relativ leicht zu lesen, aber das Gebiet hat sich in den vergangenen Jahren natürlich sehr stark weiterentwickelt. Das Buch Ripley (1988) behandelt 5 weit auseinander liegende Themen auf unterschiedlichem Niveau. Zwei der Themen betreffen Bildanalyse.

Für die Geostatistik ist Chilès and Delfiner (1999) das Standardwerk. Es ist sehr ausführlich, aber angenehm zu lesen. Daneben möchte ich noch das Buch von Stein (1999) erwähnen, das mehr theoretische Aspekte behandelt.

Über Markovfelder und Anwendungen in der Bildanalyse gibt es inzwischen auch sehr viel Literatur. Guyon (1995) behandelt Gauss- und Markovfelder nebeneinander, auf einem etwas technischen und trockenen Niveau. Winkler (2003) ist eine gut lesbare Einführung vom Standpunkt eines Mathematikers/Statistikers. Li (1995) zeigt den Standpunkt eines Ingenieurs.

Für Punktmuster und zufällige Mengen ist Stoyan, Kendall and Mecke (1995) das Standardwerk. Ohser and Mücklich (2000) befasst sich spezieller mit zufälligen Mengen und enthält ebenfalls viel nützliche Information. Eine Einführung in Punktmuster ist Diggle (2003), welches ursprünglich 1983 erschienen ist und dann nach 20 Jahren überarbeitet und neu aufgelegt wurde. Das Buch Møller and Waagepetersen (2004) setzt den Schwerpunkt auf die statistische Analyse von Punktmustern mit Hilfe von Simulationen und geht daher in vielen Punkten weiter als die Vorlesung.



# Literaturverzeichnis

- Chilès, J.-P. and Delfiner, P. (1999). *Geostatistics. Modeling Spatial Uncertainty*, Wiley Series in Probability and Statistics, Wiley, N.Y.
- Cressie, N. A. C. (1993). *Statistics for Spatial Data*, Wiley Series in Probability and Mathematical Statistics, 2nd edn, Wiley, N. Y.
- Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*, 2nd edn, Arnold, London.
- Guyon, X. (1995). *Random Fields on a Network. Modeling, Statistics, and Applications*, Springer, New York.
- Li, S. Z. (1995). *Markov Random Field Modeling in Computer Vision*, Springer, Tokyo.
- Møller, J. and Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes*, Monographs on Statistics and Applied Probability 100, Chapman & Hall.
- Ohser, J. and Mücklich, F. (2000). *Statistical Analysis of Microstructures in Materials Science*, Statistics in Practice, Wiley.
- Ripley, B. D. (1981). *Spatial Statistics*, Wiley, N. Y.
- Ripley, B. D. (1988). *Statistical Inference for Spatial Processes*, Cambridge University Press, Cambridge.
- Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, New York.
- Stoyan, D., Kendall, W. S. and Mecke, J. (1995). *Stochastic Geometry and its Applications*, Wiley Series in Probability and Statistics, 2nd edn, Wiley.
- Winkler, G. (2003). *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods : A Mathematical Introduction*, Applications of Mathematics, 2nd edn, Springer, New York.