Probabilistische Aspekte der kantenerhaltenden Glättung

Gerhard Winkler [1]

GSF - Forschungszentrum für Umwelt und Gesundheit GmbH
IBB - Institut für Biomathematik und Biometrie

Postfach 1129, D-85758 Oberschleißheim

Dezember 2000

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

Vorwort

Gegenstand dieses technischen Berichtes ist die Unterdrückung von Rauschen auf ein- oder mehrdimensionalen Signalen. Mehrdimensionale Signale werden auch kurz 'Bilder' genannt. Die Rauschunterdrückung bei eindimensionalen Signalen oder bei Bildern ist ein grundlegendes und klassisches Problem in Signal- und Bildanalyse. Es wurde u.a. eine umfangreiche Theorie der linearen Filterung entwickelt, beruhend z.B. auf Transformationsmethoden wie Fourier- oder Waveletanalyse. Das Ziel ist, das Signal zu glätten, d.h. durch das Rauschen verursachte Rauhigkeit zu entfernen. Das Hauptproblem ist, in solchen Bereichen *nicht zu glätten*, in denen die Daten im Vergleich zur Rauschintensität hohe Sprünge aufweisen, d.h. auf Brüche oder Ränder hindeuten. Lineare Filter sind bekanntlich hierfür ungeeignet: sie transformieren das Signal lediglich - üblicherweise unter Informationsverlust. Nur nichtlineare Verfahren können 'Entscheidungen' - etwa über das Vorliegen einer Kante - fällen, .

Ein klassisches Beispiel eines nichtlinearen Filters ist der Medianfilter, der in der elementaren Statistik als gegen Ausreißer robuster Ersatz für das empirische Mittel empfohlen wird. In der Tat spielen Intensitäten auf einer Seite eines Sprunges bzgl. der Glättung auf der anderen Seite die Rolle von Ausreißern. Allgemeiner formuliert, sind also robuste Verfahren von Interesse. Praktisch alle vorgestellten robusten Verfahren entsprechen einer nichtlinearen Filterung im weitesten Sinne.

Während die Theorie der linearen Filterung bis in letzte Details entwickelt ist und zum Standard gehört, ist schon im elementaren Fall des Medianfilters weit weniger allgemein bekannt. Im folgenden wird versucht, zunächst am Beispiel des Medianfilters zu demonstrieren, wie eine deterministische und wie eine statistische Analyse nichtlinearer Filter aussehen könnte. Anschließend wird ein vollständig modellbasierter - bayesianischer - Ansatz vorgestellt, für dessen Verständnis die im elementaren Fall gewonnenen Erkennt-

nisse hilfreich sind. Schließlich werden neuere nichtlineare -'bayesnahe' Filter und Schätzer - wie etwa 'truncated means', σ-Filter, nichtlineare Gaußfilter, Kaskaden solcher Filter, lokale M-Glätter und lokal adaptive statistische Verfahren vorgestellt und verglichen.

Die Beschäftigung mit solchen Verfahren ist wesentlicher Bestandteil der wissenschaftlichen Arbeit der Arbeitsgruppe 2 des Institutes für Biomathematik und Biometrie an der GSF. Ich bedanke mich bei allen Kollegen der Arbeitsgruppe 2 des IBB, die mir sämtlich sehr geholfen haben. Besonders danke ich V. Aurich, Düsseldorf, der sein gesamtes Bildmaterial zur Verfügung stellte. F. Friedrich, Heidelberg, hat Simulationen zu Kantenmodellen, Mustern und Algorithmen für uns gerechnet.

Gerhard Winkler München im Dezember 2000

Inhaltsverzeichnis

Kapitel 1

Einleitung: Informationsgehalt von Kanten

Kantenerhaltende Glättung ist eng mit Kantenfindung verwandt. Diese kann als eigene Disziplin der klassischen Bildverarbeitung aufgefaßt werden und hat sich mittlerweile zu einer Art 'Kunsthandwerk' entwickelt. Trotzdem wurden in den letzten Jahren neue - nichtlineare - Methoden gefunden, die zu einer Verbesserung der kantenerhaltenden Glättung beitragen sollten. Diese Entwicklung war nicht zuletzt deshalb möglich, weil sich immer mehr Mathematiker und Statistiker mit Signal- und Bildverarbeitung beschäftigen. Insbesondere flossen neue, z.B. semi- oder nichtparametrische Methoden der modernen Statistik ein.

Obwohl wir uns mit kantenerhaltender Glättung und nicht mit Kantenfindung beschäftigen wollen, sollen einige Bemerkungen über letztere die Bedeutung von Kanten illustrieren. In der Tat sind kantenerhaltende Glättung, Kantenfindung und Segmentierung eng verwandt. Nach Glättung können Kanten als Intensitätssprünge detektiert werden, die Kanten andererseits unterteilen ein Bild in Gebiete mit glatten Intensitätsverläufen. Dies entspricht einer Segmentierung des Bildes nach dem Kriterium 'glatte Teilbereiche'.

Die Bedeutung von Kanten kann beispielhaft an ihrem Informationsgehalt illustriert werden. Die Arbeiten U. Daub, [10] (1995) und V. Aurich und U. Daub, [2] (1996) von V. Aurich *et al.* befassen sich mit der Untersuchung dieses Aspektes.

Daß Kanten erhebliche Information tragen können, wird schon am Beispiel von Zeichnungen oder Schrift klar. Es stellt sich die Frage, wie die der Informationsgehalt präzisiert und gemessen werden kann. Diese Fragestel-

lung wurde in [10] und [2] präzisiert und untersucht. Die Idee der Autoren kann wie folgt zusammengefaßt werden:

Aus einem Grauwertebild werden mit Hilfe eines geeigneten Kantenfinders die Kanten detektiert. In den Bildbeispielen Abb. 1.1, 1.2, 1.3 und 1.4 wurde dafür Aurichs Filterkette, vgl. Abschnitt 4.3 verwendet. Natürlich ist der Bedarf an Speicherplatz für diese Kanten erheblich geringer als für das ganze ursprüngliche Intensitätsmuster. Um einen besseren optischen Eindruck zu gewährleisten, wird ein Undersampling der Intensitäten durchgeführt, d.h. die Grauwerte werden auf einem sehr groben Raster abgetastet und gespeichert. Auch dies erfordert nur geringen Speicherplatz. Um aus diesen Daten eine Version des ursprünglichen Bildes zu rekonstruieren, werden die ausgedünnten Intensitäten auf das ursprüngliche - feine - Gitter interpoliert. Um den Kontrast zu erhalten, wird nur auf solchen Pixeln interpoliert, die von einem Pixel des Teilgitters aus 'gesehen' werden, d.h. die vom Pixel des Teilgitters aus gesehen nicht jenseits einer Kante liegen.

Die Bewertung der Qualität dieser Rekonstruktion schließlich ist subjektiv: Das Undersampling der Intensitäten erfolgt mit einer Rate, bei der ein menschlicher Betrachter die Rekonstruktion für eine gute Version des ursprünglichen Bildes hält. Die erreichte Kompressionsrate dient dann als Maß für den Informationsgehalt der Kanten (zusammen mit dem Undersampling). Die Methode liefert außerdem ausgezeichnete Rekonstruktionen verrauschter Bilder.

Daß Kanten erhebliche Information tragen, wird eindrucksvoll durch Beispiele belegt. Wir demonstrieren dies an Bildern aus [10] mit Kompressionsraten unter 4 %. In den Abbildungen 1.1, 1.2 und 1.3 wurden natürliche Szenen bearbeitet. Abbildung 1.4 stellt die Ergebnisse dieses Kompressionsverfahrens mit dem derzeitigen Standardverfahren JPEG komprimierten Bildern gegenüber. JPEG beruht auf einer Cosinustransformation. In den Bildunterschriften benutzen wir die Abkürzung PSC (piecewise smooth compression) für die AURICH-DAUB Methode. Wir sehen, daß die auf kantenerhaltender Glättung beruhende Methode qualitativ völlig andere Ergebnisse als JPEG liefert. Sie eignen sich sehr gut für anschließende Analysen, wie z.B. Schrifterkennung. Abb. 1.5 zeigt die Ergebnisse der Kompression mit dem Verfahren EPIC, der Waveletkompression, und YF, einer fraktalen Kompression, beide mit ähnlicher Kompressionsrate wie in Abb. 1.4.

Abbildung 1.1: Natürliche Szene (oben) mit PSC komprimiert (unten), Kompressionsrate 3.87% (aus [10])

Abbildung 1.2: Natürliche Szene (oben) mit PSC komprimiert (unten), Kompressionsrate 4,64 % (aus [10])

Abbildung 1.3: Natürliche Szene (oben) mit PSC komprimiert (unten), Kompressionsrate 2.94% (aus [10])

Abbildung 1.4: Ziffern auf einem Chip mit PSC komprimiert, Rate kleiner als 5 %. Oben: Originalbild; mitte: PSC Kompression; unten: Kompression mit JPEG mit vergleichbarer Rate (aus [10])

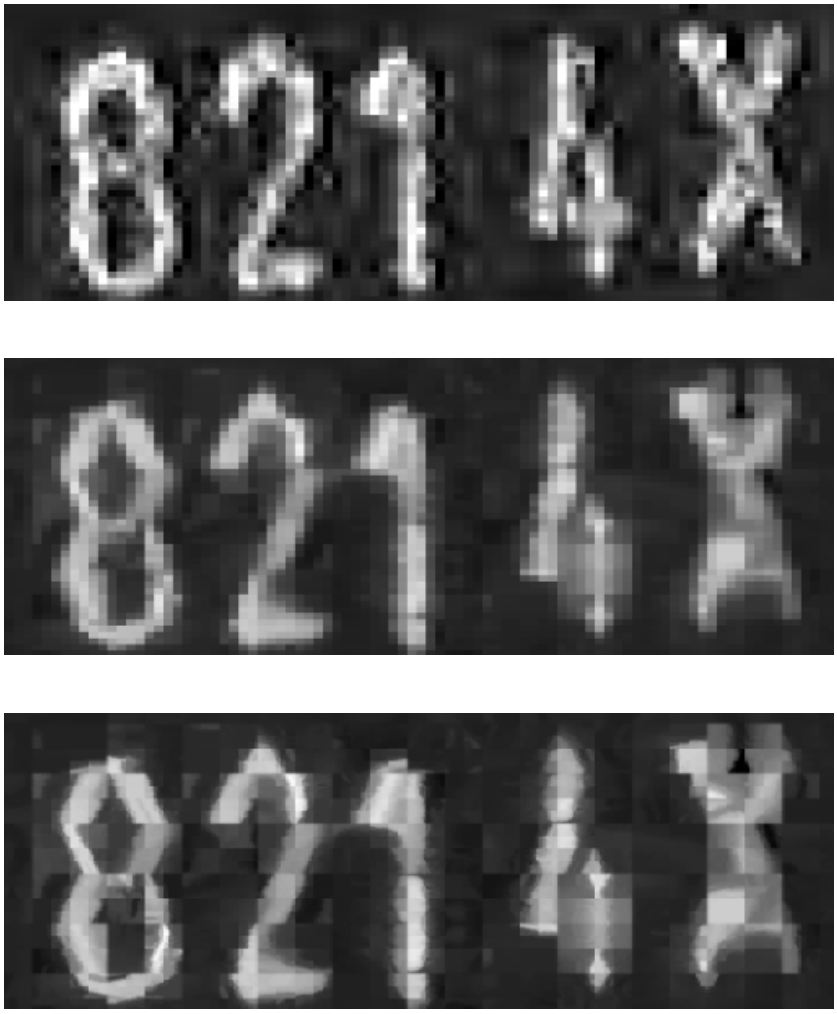


Abbildung 1.5: Ziffern aus Abb. 1.4 komprimiert, Rate kleiner als 5 %. Komprimiert mit: oben: EPIC; mitte: YF; unten: YF, vierfach vergrößerter Ausschnitt (aus [10])

Kapitel 2

Analyse nichtlinearer Filter am Beispiel des Medianfilters

Glättungsfilter verschmieren das Signal, beseitigen Rauhigkeiten nicht vollständig, zerstören scharfe Kanten. Gerade letztere sind wesentliche primitive Bildelemente. Sogenannte Rangordnungsfilter können hier teilweise Abhilfe schaffen. Der Name kommt aus der Statistik - wir kommen später darauf zurück. Das einfachste Beispiel ist der *Medianfilter*. An diesem Beispiel wollen wir illustrieren, wie eine sorgfältige Analyse eines (nichtlinearen) Filters aussehen könnte (und sollte). Wir führen - soweit möglich - eine deterministische und eine statistische Analyse durch. 'Analyse' wird in diesem Text im strikt mathematischen bzw. statistischen Sinne verstanden.

2.1 Deterministische Analyse

Wir beginnen mit der deterministischen Analyse des Medianfilters. Dieser induziert eine *nichtlineare* Abbildung zwischen Signalen. Deshalb steht z.B. die Fourieranalyse nicht zur Verfügung. Im folgenden beschränken wir uns auf das Studium der Fixpunkte, insbesondere in der Hoffnung, daß wiederholte Anwendung des Medianfilters approximativ Fixpunkte liefert bzw. daß sich typische Eigenschaften der Fixpunkte in den Ergebnissen wiederfinden lassen. Die folgenden Abschnitte sind im wesentlichen eine Ausarbeitung von Teilen in [29]. Wir betrachten nur eindimensionale Signale, da schon in zwei Dimensionen alleine die Definition geeigneter Monotonieeigenschaften sehr schwierig ist. Einen Eindruck davon vermittelt [29], 6.3.

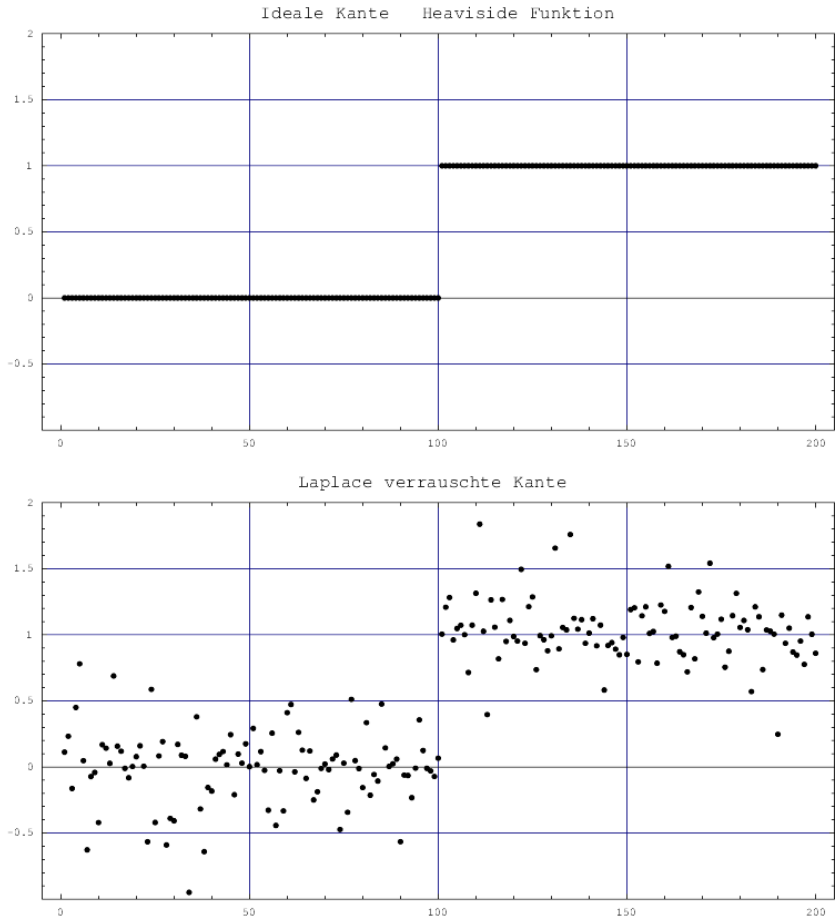


Abbildung 2.1: Heavisidekante mit Sprunghöhe 1 mit additivem laplaceverteiltem Rauschen mit Erwartungswert 0 und Streuung $0,2$.

2.1.1 Einführung

Wir beschränken uns auf den Fall einer Dimension, d.h. wir wählen die Menge $\mathbb{Z}$ der ganzen Zahlen als Indexmenge. Ein *Bild* oder *Signal* ist also durch eine Doppelfolge $x = (x_i)_{i\in\mathbb{Z}}$ reeller Zahlen gegeben. Der *Medianfilter* oder *gleitende Median* bildet jede Doppelfolge $x = (x_i)$ auf eine Doppelfolge $\mathbb{M}_{2k+1}$ ab gemäß der Vorschrift

$$M_{2k+1}(x)_n = \mathbb{M}(x_{n-k}, \ldots, x_n, \ldots, x_{n+k}).$$

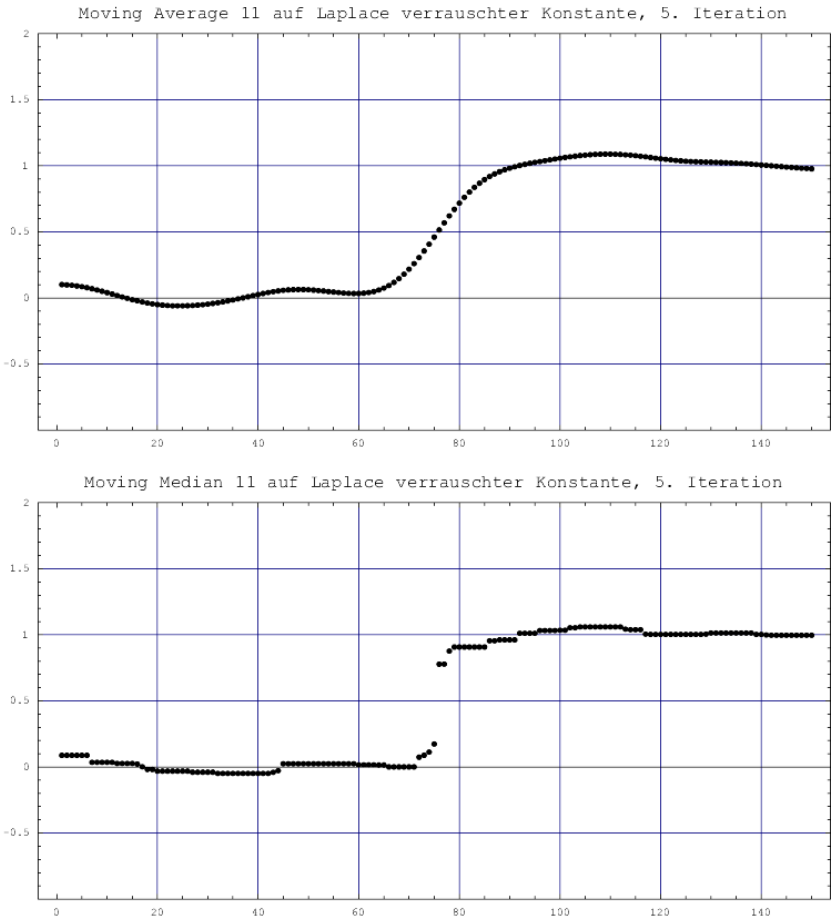


Abbildung 2.2: Verrauschte Heavisidekante aus Abb. 2.1 nach fünfter Iteration des Moving Average mit Maskenbreite 11 und fünfter Iteration des Moving Median mit Maskenbreite 11

Dabei ist $k \in \mathbb{N}_0$, der Menge der natürlichen Zahlen $\mathbb{N}_0 = \{0, 1, 2, 3, \ldots\}$, und $2k+1$ die (hier ungerade) *Fenster*-, oder *Maskenbreite*. Die Menge $\{n-k, \ldots, n+k\}$ von Indizes wird auch als *Fenster* oder *Filtermaske* bezeichnet. $\mathbb{M}\{\cdots\}$ ist das Element $x_{(n)}$ der *geordneten Stichprobe*

$$(x_{(j)})_{j=n-k}^{n+k} = (x_{(n-k)}, \cdots, x_{(n-1)}, x_{(n)}, x_{(n+1)}, \cdots, x_{(n+k)})$$

definiert durch

$$\{x_{(j)} : n-k \le j \le n+k\} = \{x_j : n-k \le j \le n+k\},$$
$$x_{(n-k)} \le \cdots \le x_{(n-1)} \le x_{(n)} \le x_{(n+1)} \le \cdots \le x_{(n+k)}).$$

Die x_i werden also der Größe nach geordnet und das in der Reihenfolge mittlere Element ist der *Median*. In der Praxis wendet man den Filter meist

Abbildung 2.3: Eindimensionale Indexmenge $\mathbb{Z}$: Allgemeine Pixel sind durch Kreise symbolisiert, das aktuelle Pixel durch den großen schwarzen Punkt und die Elemente der Filtermaske des Medianfilters $\mathbb{M}$ durch kleine schwarze Punkte.

Abbildung 2.4: Pixelgitter, Pixel und Filtermasken in $\mathbb{Z}^2$

nicht nur einmal, sondern mehrmals hintereinander an. Das Ergebnis einer solchen Iteration ist $\mathbb{M}^n(x) = \mathbb{M} \circ \cdots \circ \mathbb{M}(x)$.

Beispiel (a) Wir geben ein Beispiel für eine Stichprobe, die zugehörige geordnete Stichprobe und den Median:

$$z = \underbrace{(2,2,3,-1,0)}_{\text{ungeordnet}} \longrightarrow \underbrace{(-1,0,2,2,3)}_{\text{geordnet}} \longrightarrow \underbrace{2}_{\mathbb{M}(z)}.$$

(b) Der Median ist 'robust' gegen Ausreißer:

$$z = (2, 1000, 3, -1, 0) \longrightarrow \mathbb{M}(z) = 2.$$

Mittelwerte sind *nicht* robust gegen Ausreißer:

$$\bar{z} = \frac{1}{5} \cdot 6 = \frac{6}{5} \sim 1; \quad \bar{y} = \frac{1004}{5} \sim 201.$$

Es gibt recht wenige Eigenschaften, die man dem Medianfilter direkt ansehen kann. Offensichtlich sind:
(1) Gilt $x_{-k} \leq \cdots \leq x_0 \leq \cdots \leq x_k$, so $\mathbb{M}(x_{-k}, \ldots, x_0, \ldots, x_k) = x_0$;
(2) Ist $g : \mathbb{R} \to \mathbb{R}$ monoton, so $\mathbb{M}(g(x_1), \cdots, g(x_n)) = g(\mathbb{M}(x_1, \cdots, x_n))$.

Unsere erste Annäherung an solche Filter M_{2k+1} ist das Studium ihrer *Fixpunkte* $M_{2k+1}x = x$. Hat ein x erwünschte Gestalt - in unserem Fall etwa glatte oder stufenförmige (ideal) - sollte es vom Filter nicht zerstört werden, also (ungefähr) ein Fixpunkt sein. Unerwünschte Gestalt sollte beseitigt werden, also bei Fixpunkten nicht vorkommen. Als nächstes wären die Einzugsbereiche zu studieren; Signale sollten bei Iteration des Filters gegen (Fix-) Punkte mit erwünschten Eigenschaften streben. Ersteres ist schon schwer genug - wir beschränken uns zunächst darauf.

2.1.2 Lokal monotone Fixpunkte

Aus der Eigenschaft (1) folgt:

Beispiel 2.1.1 Monotone Signale sind Fixpunkte des gleitenden Medians. Somit sind auch ideale Kanten, repräsentiert durch *Heaviside-Funktionen*[1] $\chi_{[a,\infty)}$ Fixpunkte.

Wegen der endlichen Maskengröße sollte Monotonie des Signals über genügend lange Segmente hinreichend für die Fixpunkteigenschaft sei. Wir befassen uns deshalb im folgenden mit Fixpunkten des Medianfilters, welche lokale Monotonieeigenschaften haben.

Definition 2.1.2 *x ist* lokal monoton der Länge *m, wenn jedes Segment der Länge m monoton ist. Die Menge der lokal monotonen Doppelfolgen $x = (x_n)_{n\in\mathbb{Z}}$ werde mit $LOMO(m)$ bezeichnet.*

Eine lokal monotone Folge muß nicht global monoton sein, da die Trends wechseln können.

[1] Heavyside-Funktionen auf $\mathbb{Z}$ sind von der Gestalt $\chi_{[a,\infty)}$, $a \in \mathbb{Z}$, wobei χ_A die Indikatorfunktion der Menge A ist, d.h. $\chi_A(i) = 1$ falls $i \in A$ und $\chi_A(i) = 0$ andernfalls.

Definition 2.1.3 *Ein* Trendwechsel *bei p und q mit $p < q$, liege vor, wenn $x_p < x_{p+1}$ und $x_{q-1} > x_q$ oder wenn $x_p > x_{p+1}$ und $x_{q-1} < x_q$.*

Bei lokal monotonen Folgen dürfen Stellen mit Trendumkehr nicht zu nahe beieinander liegen.

Lemma 2.1.4 *Hat x einen Trendwechsel bei p und q und gilt $x \in LOMO(m)$, so ist x konstant auf einem Segment der Länge $m-1$ in $[p+1, q-1]$.*

Beweis Es liege ein Trendwechsel bei $p < q$ vor. Wir nehmen $x_p < x_{p+1}$ und $x_{q-1} > x_q$ an. Sei u der größte Index $u < q$ mit $x_u < x_{u+1}$ und sei v der kleinste Index $v > u$ mit $x_{v-1} > x_v$. Insbesondere liegt ein Trendwechsel zwischen u und v vor. Wegen $x \in LOMO(m)$ gilt:

$$x_u < x_{u+1} \leq \cdots \leq x_{u+m-1}.$$

Wegen der Maximalität von u gilt:

$$x_u < x_{u+1} = \cdots = x_{u+m-1}$$

Genau so gilt

$$x_{v-m+1} = \cdots = x_{v-1} > x_v.$$

Jedenfalls ist x konstant auf einem Segment der Länge mindestens $m-1$ zwischen u und v und somit auch zwischen p und q. □

Satz 2.1.5 *Jedes $x \in LOMO(m)$ ist ein M_{2k+1} - Fixpunkt für alle $k \in \mathbb{N}$ mit $k \leq m-2$.*

Beispiel 2.1.6 Die periodische Fortsetzung von $-1, -1, 1, 1$ ist $LOMO(3)$ und in Übereinstimmung mit der Aussage des Satzes Fixpunkt von M_3. Sie ist *nicht* Fixpunkt von M_5 und M_7. Deren Anwendung bewirkt Spiegelung des Signales an der 0. Andererseits ist sie Fixpunkt von M_9. Wir werden diesen Aspekt später genauer untersuchen.

Beweis Sei $M_{2k+1} = M$. Wir berechnen $(Mx)_n$. Dazu betrachten wir das Segment V der Gestalt

$$\overbrace{x_{n-m+2}, \cdots, x_{n-1}}^{m-2}, \overbrace{x_n}^{1}, \overbrace{x_{n+1}, \cdots, x_{n+m-2}}^{m-2}$$

der Länge $l = 2(m-2)+1 = 2m-3 (\geq 2k+1)$. Es enthält also alle Elemente innerhalb der Filtermaske um n. Ist V monoton, so ist $M(x)_n = x_n$. Ist V nicht monoton, so enthält es nach Lemma 2.1.4 ein konstantes Stück der Länge $m-1$. Es ist $2m-2 > l$ und deshalb enthält dieses Segment das Element x_n. Die Elemente dieses konstanten Segmente haben also alle den Wert x_n. Weil $k+1 \leq m-1$ ist, enthält die Maske mindestens $k+1$ Elemente, die gleich x_n sind, und der Median ist daher ebenfalls gleich x_n. □

Insbesondere sind $LOMO(k+2)$-Folgen Fixpunkte von M_{2l+1} für $1 \leq l \leq k$. Mehr über $LOMO(k+2)$ sagt:

Satz 2.1.7 *Sei x ein Fixpunkt von M_{2k+1} und enthalte ein monotones Segment der Länge $k+1$. Dann gilt $x \in LOMO(m)$, $m-2 = k$.*

Also sind insbesondere Fixpunkte von M_{2k+1} aus $LOMO(k+1)$ schon in $LOMO(k+2)$. Der Beweis dieses Satzes beruht auf dem wichtigen Lemma 2.1.10 unten. Dessen Beweis ist etwas umfangreicher und deshalb demonstrieren wir zuerst die Bedeutung von Satz 2.1.7. Wir beweisen die Umkehrung von Satz 2.1.5.

Satz 2.1.8 *Ist x Fixpunkt von M_{2k+1} für alle $k \leq m-2$, so gilt $x \in LOMO(m)$.*

Bemerkung 2.1.9 Beispiel 2.1.6 zeigt, daß die Bedingung 'für alle $k \leq m-2$ nicht überflüssig ist.

Beweis von Satz 2.1.8 Wir führen eine vollständige Induktion nach $p = m-2$ durch.
Die Aussage ist trivialerweise richtig für $p = 0$, d.h.$m = 2$, da jede Doppelfolge in $LOMO(2)$ ist. Für $p = 1$, bzw $m = 3$, ist $M_{2k+1} = M_3$ und da stets $x \in LOMO(2)$ ist ein Fixpunkt x von M_3 in $LOMO(3)$ wegen Satz 2.1.7. Der Fall $p(= m-2) = 1$ ist damit fertig und somit der Induktionsanfang.
Sei der Satz für p gültig und x Fixpunkt von M_{2k+1} für $k = 1, \cdots, p$. Dann ist nach Induktionsvoraussetzung $x \in LOMO(p+2)$, enthält also monotone Stücke der Länge $p+2$ und ist nach Satz 2.1.7 aus $LOMO(p+3) = LOMO(m+1)$. □

Lemma 2.1.10 *Seien $m < n$ und*

$$x_m < x_i < x_n \quad \textit{für} \quad m < i < n. \tag{2.1}$$

Seien ferner

$$\mathbb{M}(x_{m-p},\ldots,x_{m+p}) = x_m,\ \ \mathbb{M}(x_{n-q},\ldots,x_{n+q}) = x_n, \tag{2.2}$$

wobei

$$m-p \le n-q,\ \ m+p \le n+q. \tag{2.3}$$

Dann ist

$$\Big(\{x_{m-p},\ldots,x_m\}\bigcup\{x_n,\ldots,x_{n+q}\}\Big)\bigcap(x_m,x_n) = \emptyset. \tag{2.4}$$

Überdies gilt:

$$\begin{aligned} x_j &\le x_m,\ \ j \le \min\{m, n-q\} \\ x_j &\ge x_n,\ \ j \ge \max\{n, m+p\}. \end{aligned}$$

Beweis Wegen (2.2) gibt es in $(x_{m-p},\ldots,x_{m+p})$ (ohne x_m) mindestens p Größen x_l mit $x_l \le x_m$. Wegen (2.1) kommen dafür Größen x_l mit $m < l < n$ nicht in Betracht. Also liegen diese p Größen in

$$A = \{x_{m-p},\ldots,x_{m-1}\}\bigcup\{x_{n+1},\ldots,x_{n+q}\}.$$

Genau so liegen in A mindestens q Größen $x_r \ge x_n$. Weil aber A nur $p+q$ Größen enthält, liegen in A *genau* p Größen $x_l \le x_m$ und *genau* q Größen $x_r \ge x_n$. Damit ist (2.4) verifiziert.
Nun liegen ja genau p Größen mit $x_l \le x_m$ (ohne x_m) in A. In der kleineren Menge

$$\{x_{m-p},\ldots,x_{m+p}\}\backslash\{x_{m+1},\ldots,x_{n-1}\}$$

liegen ebenfalls mindestens p solcher Größen. Analoges gilt für die $x_r \ge x_n$. Somit muß für $j < n-q$ gelten, daß $x_j \le x_n$, denn sonst würde das Element in $[n-q, n+q]$ fehlen. Damit ist der Beweis vollständig. □

Beweis von Satz 2.1.7 . Habe x ein monotones Segment $(x_{n-k},\ldots,x_n)$ der Länge $k+1$. Dieses ergänzt sich zu einem monotonen Segment der Länge $k+2$. Wir unterscheiden zwei Fälle:
(1) Wenn $x_{n-k} = x_n$, so gilt $x_{n-k} = \cdots = x_n$ wegen der Monotonie. Dann sind

$$(x_{n-k-1},\ldots,x_n),\ \ (x_{n-k},\ldots,x_n,x_{n+1})$$

jedenfalls monoton der Länge $k+2$.
(2) Sei $x_{n-k} \neq x_n$; gelte o.E. $x_{n-k} < x_n$ (insbesondere ist das Stück isoton). Wenn $x_n \leq x_{n+1}$ ist, so hat man ein monotones Segment der Länge $k+2$. Sei nun $x_n > x_{n+1}$. Wegen der Fixpunkteigenschaft ist Lemma 2.1.10 anwendbar (mit $m \approx n$ und $n \approx n+1$ und vertauschten Ungleichungen). Damit folgt $x_{n-k} \geq x_n$ im Widerspruch zu $x_{n-k} < x_n$. Dieser Fall kommt also nicht vor und es ist nur $x_{n+1} \geq x_n$ möglich. Damit hat man wieder ein monotones Segment der Länge $k+2$. Genauso ist $(x_{n-k-1}, \ldots, x_n)$ monoton und von der Länge $k+2$. Mit diesen beiden Stücken wiederhole man das Argument und fahre so fort. Damit folgt $x \in LOMO(k+2)$. □

Wir fassen zusammen.

Satz 2.1.11 *Für ein Signal $x = (x_n)_{n\in\mathbb{Z}}$ und $p \in \mathbb{N}$ sind äquivalent:*
(a) x ist Fixpunkt aller $M_{2k+1}, 1 \leq k \leq p$.
(b) $x \in LOMO(p+2)$.

Beweis Man verbinde Satz 2.1.5 und Satz 2.1.8. □

2.1.3 Nichtmonotone Fixpunkte

Wir haben nun die Fixpunkte mit Monotonieeigenschaften charakterisiert. Sie sind überall lokal monoton. Leider haben gleitende Mediane weitere Fixpunkte. Diese sollten wegen Satz 2.1.7 nirgends monoton sein (vgl. Beispiel 2.1.6).

Definition 2.1.12 *Ein Signal ist* nirgends $LOMO(m)$*, wenn es kein monotones Segment der Länge m enthält.*

Beispiel 2.1.13 *Die Doppelfolge x gegeben durch $\ldots, 1, -1, 1, -1, 1, -1, \ldots$ ist $LOMO(2)$ und nirgends $LOMO(m)$ für $m > 2$. Sie ist Fixpunkt von M_1, M_5, M_9, ...; allgemein ist sie Fixpunkt von M_{2k+1} für $k = 2j$, $j \in \mathbb{N}_0$. Für $k = 2j+1$, $k \in \mathbb{N}$, alterniert die Folge $(M^n_{2k+1}x)_{n\geq 0}$.*

Also gibt es Fixpunkte, die nirgends $LOMO(k)$ sind. Solche Fixpunkte sind allerdings starken Restriktionen unterworfen.

Satz 2.1.14 *Ist x Fixpunkt von M_{2k+1} und nirgends LOMO(k+1), dann nimmt x genau zwei Werte an.*

Bevor wir zum Beweis übergehen, diskutieren wir diese Aussage.

Definition 2.1.15 *Die Fixpunkte aus Satz 2.1.11 seien vom* Typ I *und die Fixpunkte aus Satz 2.1.14 seien vom* Typ II.

Für M_3 ist $k = 1$ und $k + 1 = 2$. Da jede Folge $LOMO(2)$ ist, hat M_3 keine Fixpunkte vom Typ II. Nach Satz ist ja auch 2.1.7 jeder Fixpunkt von M_3 vom Typ LOMO(3).

Spezielle Fixpunkte beider Typen lassen sich leicht konstruieren.

Beispiel 2.1.16 Sei $k \geq 1$ und

$$\ldots; a_0, a_1, \ldots, a_k; -a_0, -a_1, \ldots, -a_k; \ldots$$

periodisch erzeugt aus $a_0, \ldots, a_k = \pm 1$, d.h. insbesondere mit Periode $2k+2$. Wir bezeichnen die Folge mit x. Natürlich ist x ein Fixpunkt von M_{2k+1}, denn es liegen nach Konstruktion in $[n-k, n+k]\backslash\{n\}$ genau so viele positive wie negative Elemente. Das Element x_n gibt dann den Ausschlag in seine eigene Richtung. (Man veranschauliche sich dies mit Hilfe der folgenden Zeile, indem man den gekennzeichneten Bereich sukzessive um 1 verschiebt.)

$$-a_0, \underbrace{-a_1, -a_2, -a_3, \{a_0\}, a_1, a_2, a_3}, -a_0, -a_1, -a_2, -a_3, a_0$$

1. Fall: $(a_0, \ldots, a_k)$ ist monoton. Dann ist x LOMO(k+2) nach Satz 2.1.7 und somit vom Typ I.
2. Fall: $(a_0, \ldots, a_k)$ ist nicht monoton. Dann ist x nirgends $LOMO(k+1)$, denn sonst gäbe es ein monotones Stück der Länge $k+1$ und wegen der Fixpunkteigenschaft wäre $x \in LOMO(k+2)$ im Gegensatz zur Annahme. Deshalb ist x vom Typ II.
Typ II Fixpunkte von M_5 und M_7 sind von dieser Art. Für M_9 gilt dies nicht, denn der Fixpunkt aus Beispiel 2.1.6 hat Periode 4, aber nicht Periode 10 wie in der Konstruktion.

Beweis von Satz 2.1.14 Da x nicht konstant ist, sei o.E.
(I) $x_0 < x_1$
Nach dem Lemma 2.1.10 wissen wir:
(II) $x_i \leq x_0$ oder $x_i \geq x_1$ für $-k \leq i \leq k+1$.
(III) $x_{-k} \leq x_0 < x_1 \leq x_{k+1}$.

Wir zeigen:
(i) es gibt $-k+1 \leq p \leq -1$ mit $x_p \geq x_1$ und
(ii) es gibt $2 < q \leq k$ mit $x_q \leq x_0$.
Dies ist also das Pendant zu (III). Wir verifizieren jetzt (ii). Nach dem Lemma gibt es keine Elemente x_l mit $x_0 < x_l < x_1$. Deshalb genügt es zu zeigen, daß es ein $q \in \{2, \ldots, k\}$ gibt mit $x_q < x_1$. Um einen Widerspruch herbeizuführen, nehmen wir an, daß

$$x_i \geq x_1, \text{ für alle } i \in \{2, \ldots, k\}.$$

Weil $x_1 > x_0$ gilt dann $x_l \geq x_0$ für $k+1$ Elemente in $[0, k]$. Wegen der Medianeigenschaft von x_0 muß dann gelten, daß

$$x_l \leq x_0, \ -k \leq l \leq -1. \tag{2.5}$$

Aus der Annahme und (III) folgt das Analogon

$$x_l \geq x_1 > x_0 \text{ für alle } 2 \leq l \leq k+1.$$

Weil x kein monotones Stück der Länge $k+1$ enthält, gibt es $j \in [1, k]$ mit $x_j > x_{j+1}$. Das Lemma, angewandt auf $m = j+1$, $n = j$, besagt, daß $x_{j-k} \geq x_j (\geq x_0)$ ist. Weil $j - k \leq 0$ widerspricht dies (2.5). Damit ist (ii) verifiziert und aus Symmetriegründen auch (i).
(b) Wir zeigen schließlich, daß x nur die Werte x_0 und x_1 annimmt. Es genügt, dies für die Indizes in $R = \{2, \ldots, k+1\}$ nachzuweisen. Wir wissen, daß es $j \in \{2, \ldots, k\}$ gibt mit $x_j \leq x_0$.
Wir nehmen zuerst an, daß $x_p < x_0$ für ein $p \in R$. Dazu wählen wir jeweils $q \in R$ mit $x_q \geq x_1$ (z.B. x_1 selbst). Wir können p und q sogar so wählen, daß

$$x_j \in (\min\{x_p, x_q\}, \max\{x_p, x_q\}) \text{ für alle } j \in (\min\{p, q\}, \max\{p, q\}).$$

Dann ist $x_p < x_0 < x_q$. Weil sowohl x_p als auch x_q die Medianeigenschaft haben, widerspricht dies dem ersten Teil des Lemmas 2.1.10, da

$$0 \in [\min\{p, q\} - k, \min\{p, q\}].$$

Falls es $p \in R$ gibt mit $x_p > x_1$ argumentieren wir *mutatis mutandis* genauso. Damit ist der Satz ist bewiesen. □

Bemerkung 2.1.17 *Offen* scheint z.B. die Frage, ob alle Typ II Fixpunkte periodisch sind. In Simulationen werden die oben gezeigten Eigenschaften jedenfalls sichtbar.

Die im Beispiel betrachteten Typ II Fixpunkte sind sehr regelmäßig aufgebaut und treten bei zufällig gestörten Signalen i.a. kaum auf.

2.1.4 Gemischte Filter

Man kann überlegen, wie man die - unerwünschten - Typ II Fixpunkte langer Filtermasken vermeidet, z.B. durch Aufsetzen kurzer Glätter und anschließendes 'reroughing'. Ein weiterer Ausweg ist die Konstruktion kombinierter Filter. Einige Aussagen seien unbewiesen berichtet. Das erste Beispiel zeigt schon, wie kompliziert die Situation ist.

Satz 2.1.18 *Seien* $a_0 \neq 1$, $a_k \geq 0$ *und* $\sum_{k=0}^{n} a_k = 1$. *Ferner sei*

$$T = \sum_{k=0}^{n} a_k \mathbb{M}_3^k.$$

Dann gelten:
(a) Ist $a_k \neq 0$ *für ein ungerades* k, *dann hat* T *nur* $LOMO(3)$ *Folgen als Fixpunkte.*
(b) Ist $a_k = 0$ *für alle ungeraden* k, *dann hat* T *sowohl* $LOMO(3)$ *als auch alternierende Fixpunkte.*
(c) Fixpunkte von T^m, $m \geq 2$ *sind entweder* $LOMO(3)$ *oder alternierend. Alternierende Fixpunkte treten hierbei nur auf, falls*

- m *gerade und* $a_k = 0$ *für alle geraden* k, *oder*
- $a_k = 0$ *für alle ungeraden* k.

Das Symbol I stehe für die Identität.

Satz 2.1.19 *Für*

$$T = \alpha I + (1 - \alpha) M_3, \;\; 0 < \alpha < 1,$$

gilt

$$T^m x \longrightarrow z \in LOMO(3).$$

Von diesem Typ von Konvergenzaussagen hätte man gerne mehr. Gewisse Typ II Fixpunkte lassen sich durch Kombination von Medianen verhindern. So war z.B. $\ldots, -1, 1, -1, 1, \ldots$ kein Fixpunkt von M_3, jedoch Fixpunkt von M_5.

Satz 2.1.20 *Für*

$$T = \alpha M_3 + (1 - \alpha) M_5, \;\; 0 < \alpha < 1,$$

gilt

$$Tx = x \Longleftrightarrow x \in LOMO(4).$$

Der eigentlich interessante Fall der Dimension 2 ist höchst kompliziert, schon die Definition der lokalen Monotonie stößt auf Schwierigkeiten. Für isolierte Resultate vgl. man [29].

2.2 Stochastische Analyse des Medianfilters

Üblicherweise werden Signale u.a. durch zufällige Einflüsse gestört. Dies kommt einmal daher, daß ein genuin stochastisches Rauschen vorliegt, wie z.B. ein Elekronenrauschen in Röhren, oder daß nur eine stochastische Modellierung der (Zer-) Störung sinnvoll ist wie beim Ausfall ganzer Zeilen oder Spalten bei CCD-Chips. Wir deuten nun an einfachsten Beispielen an, wie eine stochastische Analyse des – wir betonen nichtlinearen – Medianfilters aussehen könnte. Im wesentlichen sind die folgenden Abschnitte eine Ausarbeitung des ersten Teils von [19].

Wir werden additives und nicht additives Rauschen betrachten. Bei additivem Rauschen nimmt man an, daß ein ideales Signal $(x_i)_{i\in S}$ additiv gestört wird, d.h. man hat das Modell

$$\xi_i = x_i + \eta_i,$$

wobei (η_i) die Rausch(zufalls)variablen und (ξ_i) die (nunmehr zufälligen) Beobachtungen sind. Sind die η_i i.i.d mit $\mathbb{E}(\eta_i) = 0$, so nennt man $\eta = (\eta_i)$ oft *weißes Rauschen*.

Im allgemeinen ist x unbekannt und man versucht es aus den Beobachtungen heraus wiederherzustellen, im gegenwärtigen Fall durch Anwendung des Medianfilters. Uns interessiert zunächst: Gibt es Situationen, wo der empirische Median den empirischen Mittelwert schlägt? Bei welchen Signalen ist das der Fall und warum erhält er Kanten?

Wir müssen uns auf einfachste Signale x beschränken, konkret auf konstante und stufenförmige Signale. Als Einstieg betrachten wir die Wirkung des Medianfilters auf das einfachste verrauschte Bild, nämlich eine verrauschte (konstante) *Grauwertfläche*:

$$\xi_i = \mathbf{m} + \eta_i, \;\; i \text{ 'Pixel' (aus } \mathbb{Z}, \mathbb{Z}^2, \ldots)$$

Es ist sinnvoll, zunächst einfache Aussagen über den Median für i.i.d. Zufallsvariablen zusammenzustellen. Um die grobe Richtung unserer Überlegungen anzudeuten, kommentieren wir vorher das gleitende Mittel, mit dem wir den Median vergleichen wollen.

Bemerkung 2.2.1 (empirisches Mittel) Das ideale Bild sei konstant gleich $\mathbf{m}$. Bei additivem Rauschen läuft Restauration auf die Schätzung von $\mathbf{m} = \mathbb{E}(\xi_i)$ hinaus. Für das *empirische Mittel*

$$\bar{\xi} = \frac{1}{n} \sum_{i=1}^{n} \xi_i$$

gilt $\mathbb{E}(\bar{\xi}) = \mathbf{m}$, d.h. es ist *erwartungstreu* (*unbiased*). Außerdem ist $\bar{\xi}$ der *kleinste-Quadrate Schätzer* , d.h.

$$\bar{\xi} = \text{argmin}_a \sum_{i=1}^{n} (\xi_i - a)^2.$$

Sind die ξ_i gaußisch, so ist $\bar{\xi}$ auch der *Maximum-Likelihoodschätzer* , denn er maximiert die gemeinsame Dichte der ξ_i, gegeben durch

$$\frac{1}{\sqrt{2\pi\sigma^2}^n} \cdot \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\xi_i - a)^2 \right).$$

In der linearen Theorie der Statistik zeigt man überdies: $\bar{\xi}$ ist ein BLUE (best linear unbiased estimator). Dabei bedeuten:

- *linear*: Die Abbildung
$$(\xi_1, \ldots, \xi_n) \longmapsto \bar{\xi}$$
ist linear

- *unbiased*: $\bar{\xi}$ ist ein erwartungstreuer Schätzer für den Erwartungswert:
$$\mathbb{E}(\bar{\xi}) = \mathbb{E}(\xi_i) = \mathbf{m}$$

- *best*: Unter allen linearen erwartungstreuen Schätzern für den Erwartungswert $\mathbf{m}$ hat das empirische Mittel die *kleinste Varianz*.

Sind die ξ_i Gaußisch, so ist $\bar{\xi}$ sogar unter *allen* erwartungstreuen Schätzern der beste.

2.2.1 Elementare Theorie

Wir betrachten ein endliches Fenster und bezeichnen die durch die Pixel indizierten Zufallsvariablen mit

$$\xi_1, \ldots, \xi_n.$$

Für $1 \leq s \leq n$ sei die *s-te Ordnungsstatistik* definiert durch

$$\xi_{(s)} : \omega \to \xi_{(s)}(\omega).$$

Die folgenden Überlegungen sind gerechtfertigt, weil die Ordnungsstatistiken $\xi_{(s)}$ Zufallsvariablen, d.h. meßbare Funktionen sind:

Lemma 2.2.2 *Für Zufallsvariablen $\xi_1, \ldots, \xi_n$ ist jede Ordnungsstatistik $\xi_{(s)}$ meßbar.*

Beweis Zu jedem $\omega \in \Omega$ gibt es eine Permutation $\pi = \pi_\omega$ von $J = \{1, \ldots, n\}$ mit

$$\xi_{\pi(1)}(\omega) \leq \cdots \leq \xi_{\pi(n)}(\omega) = \xi_{(1)}(\omega) \leq \cdots \leq \xi_{(n)}(\omega).$$

Für jede Permutation π von J ist die Menge

$$\Omega_\pi = \{\xi_{\pi(1)} \leq \cdots \leq \xi_{\pi(n)}\}$$

meßbar. Durch Bildung aller endlichen Schnitte der Ω_π erhält man eine endliche meßbare Partition $\{P\}$ von Ω. Auf jedem P gilt $\xi_{\pi(1)} \leq \cdots \leq \xi_{\pi(n)}$ für mindestens eine Permutation π und deshalb ist $\xi_{(s)} = \xi_{\pi(s)}$ auf P und somit meßbar auf P. Also ist jede Ordnungsstatistik $\xi_{(s)}$ meßbar auf ganz Ω wie behauptet. □

Im Gegensatz zu den ξ_i sind die $\xi_{(s)}$ weder unabhängig noch identisch verteilt.

Beispiel Sei $\Omega = \{0, 1\}^2$ mit der Gleichverteilung versehen und ξ_i, $i = 1, 2$, die Projektion auf die i-te Komponente. Dann ist $\mathbb{P}(\xi_{(1)} = 0, \xi_{(2)} = 0) = \mathbb{P}((0, 0)) = 1/4$. Andererseits sind $\mathbb{P}(\xi_{(1)} = 0) = \mathbb{P}((0, 0), (0, 1), (1, 0)) = 3/4$ und $\mathbb{P}(\xi_{(2)} = 0) = \mathbb{P}((0, 0)) = 1/4$. Deshalb sind die Zufallsvariablen $\xi_{(i)}$, $i = 1, 2$, weder identisch verteilt noch unabhängig.

Der Median ist eine spezielle Ordnungsstatistik.

Definition 2.2.3 *Sei der Stichprobenumfang n ungerade. Dann heißt die $(n+1)/2$-te Ordnungsstatistik $\mathbb{M}(\xi_1,\ldots,\xi_n)) = \xi_{((n+1)/2)}$* empirischer Median.

Für Stichproben mit geradem Umfang ist die Definition komplizierter und der Median wird zu einer mengenwertigen Abbildung. Zur Berechnung von Momenten und anderen Kenngrößen benötigt man die Verteilung des empirischen Medians.

Satz 2.2.4 *Seien $\xi_1,\ldots,\xi_n$ i.i.d. mit Dichte f und Verteilungsfunktion F. Sei $\xi_{(s)}$ die s-te Ordnungsstatistik. Dann gilt für $G_s(x) = \mathbb{P}(\xi_{(s)} \le x)$, daß*

$$\begin{aligned} G_s(x) &= \int_{-\infty}^{x} n\binom{n-1}{s-1} F(y)^{s-1}(1-F(y))^{n-s} f(y)\,dy \\ &= \frac{n!}{(s-1)!(n-s)!}\int_0^{F(x)} y^{s-1}(1-y)^{n-s}dy. \end{aligned}$$

Beweis Für $z \in \mathbb{R}$ sei

$$P_\delta = \mathbb{P}(\xi_1,\ldots,\xi_{s-1} \le \xi_s \le \xi_{s+1},\ldots,\xi_n; z < \xi_s < z+\delta).$$

Nun kann man $\xi_{(s)}$ auf n Arten auswählen und die untere Gruppe auf $\binom{n-1}{s-1}$ Arten. Also ist

$$P(z < \xi_{(s)} < z+\delta) = n\binom{n-1}{s-1} P_\delta.$$

Um P_δ zu berechnen, schreiben wir untere und obere Schranken auf:

$$\begin{aligned} P_\delta &\le \mathbb{P}\left(\xi_1,\ldots,\xi_{s-1} \le z+\delta, \xi_{s+1},\ldots\xi_n \ge z; z < \xi_s < z+\delta\right) \\ &= F(z+\delta)^{s-1}\left(1-F(z)\right)^{n-s}\int_z^{z+\delta} f(y)dy = P_\delta^o \\ P_\delta &\ge \mathbb{P}(\xi_1,\ldots,\xi_{s-1} \le z \le \xi_s \le z+\delta \le \xi_{s+1},\ldots,\xi_n) \\ &= F(z)^{s-1}\left(1-F(z+\delta)\right)^{n-s}\int_z^{z+\delta} f(y)\,dy = P_\delta^u. \end{aligned}$$

Es gilt für $\delta \to 0$, daß

$$\begin{aligned} \frac{1}{\delta}P_\delta^o &\longrightarrow F(z)^{s-1}\left(1-F(z)\right)^{n-s} f(z). \\ \frac{1}{\delta}P_\delta^u &\longrightarrow F(z)^{s-1}\left(1-F(z)\right)^{n-s} f(z). \end{aligned}$$

Für die Dichte g_s von $\xi_{(s)}$ folgt daraus

$$g_s(z) = \lim_{\delta \to 0} \frac{1}{\delta} \cdot P(z < \xi_{(s)} < z + \delta) = n \binom{n-1}{s-1} F(z)^{s-1} (1 - F(z))^{n-s} f(z).$$

Damit ist die erste Identität verifiziert. Die zweite ist ein Spezialfall einer Aussage, die von allgemeinem Interesse ist. Deshalb formulieren wir sie separat. Sie beruht im wesentlichen auf dem Integraltransformationssatz für Bildmaße[2]. □

Lemma 2.2.5 *Sei μ eine Verteilung (auf der reellen Achse) mit stetiger Verteilungsfunktion F. Dann gelten für meßbare Mengen A und bzgl. μ integrierbare Funktionen h die Identitäten*

$$\begin{aligned} \mu \circ F^{-1}(A) &= \lambda(A) \\ \int_{-\infty}^{x} h \circ F \, d\mu &= \int_0^{F(x)} h \, d\lambda, \end{aligned}$$

wobei λ das Lebesguemaß[3] *bezeichnet.*

Beweis von Lemma 2.2.5 Sei $a \in [0,1]$. Dann gilt wegen der Stetigkeit und der Isotonie von F, daß

$$\mu \circ F^{-1}[0,a] = \mu\left(-\infty, \max\{y : F(y) \leq a\}\right) = F(\max\{y : F(y) \leq a\}) = a,$$

[2]Seien $(\Omega, \mathcal{F})$ und $(\Omega', \mathcal{F}')$ Meßräume und $\varphi : \Omega \to \Omega'$ eine meßbare Abbildung. Das *Bildmaß* eines Wahrscheinlichkeitsmaßes μ auf $\mathcal{F}$ unter φ ist das Wahrscheinlichkeitsmaß $\mu \circ \varphi^{-1}$ auf $\mathcal{F}'$ gegeben durch

$$\mu \circ \varphi^{-1}(F') = \mu(\varphi^{-1}[F']), \;\; F' \in \mathcal{F}'.$$

Der Integraltransformationssatz verallgemeinert diese Identität: Sei $h : \Omega' \to \mathbb{R}$ meßbar. Dann ist $h \circ \varphi$ bezüglich μ integrierbar genau dann, wenn h bezüglich $\mu \circ \varphi^{-1}$integrierbar ist und in diesem Falle gilt für jedes $A \in \mathcal{F}'$, daß

$$\int_A h \, d\mu \circ \varphi^{-1} = \int_{\varphi^{-1}[A]} h \circ \varphi \, d\mu.$$

[3]Das Lebesguemaß λ ist das eindeutig bestimmte Maß auf der (von den Intervallen erzeugten) Borel-σ-Algebra von $\mathbb{R}$ mit $\lambda((a,b]) = b - a$ für alle $a < b$.

d.h $\mu \circ F^{-1}$ ist die Gleichverteilung auf $[0,1]$ und die erste Identität ist bewiesen. Für den Nachweis der zweiten Identität setzen wir im Integraltransformationssatz (Fußnote (2)) $\varphi = F$ (F ist isoton und somit meßbar). Dies ergibt zusammen mit der ersten Identität

$$\int_0^{F(x)} h\,dx = \int_{[0,F(x)]} h\,d\mu \circ F^{-1} = \int_{F^{-1}[[0,F(x)]]} h \circ F\,d\mu.$$

Wegen der Stetigkeit (und der Isotonie) von F ist

$$F^{-1}[[0,F(x)]] = (-\infty, \max\{y : F(y) \leq F(x)\}] = (-\infty, \max\{y : F(y) = F(x)\}].$$

Damit ist die zweite Identität und damit das Lemma bewiesen. □

Bemerkung Die simple erste Gleichung ist die Grundlage für die *Inversionsmethode* zur Generierung von Pseudozufallszahlen mit Verteilung μ: Für auf $[0,1)$ gleichverteiltes u ist $F^{-1}(u)$ verteilt gemäß μ, wobei F die Verteilungsfunktion von μ ist (ist F nicht invertierbar gilt das mit einer kleinen Modifikation, vgl. Satz 2.2.6). Für invertierbares F beruht das auf den trivialen Äquivalenzen

$$a < F^{-1}(u) \leq b \Longleftrightarrow F(a) < F \circ F^{-1}(u) \leq F(b) \Longleftrightarrow F(a) < u < F(b).$$

Allgemeiner gilt:

Satz 2.2.6 *Seien η eine reelle Zufallsvariable mit Verteilungsfunktion $F(t) = \mathbb{P}(\eta \leq t)$ und*

$$F^-(u) = \min\{t : F(t) \geq u\}$$

die verallgemeinerte Inverse von F. Dann hat $\xi = F^-(U)$ mit einer in $[0,1)$ gleichverteilten Zufallsvariablen U die Verteilungsfunktion F.

Korollar 2.2.7 *Seien F eine invertierbare Verteilungsfunktion und die Zufallsvariable U in $[0,1)$ gleichverteilt. Dann hat die Zufallsvariable $\xi = F^{-1}(U)$ die Verteilung F.*

Beweis von Satz 2.2.6 Da F rechtsstetig ist existieren die Minima in der Definition von F^-. Der Supergraph von F^- stimmt mit dem Subgraphen von F überein, d.h.

$$\{(u,t) : F^-(u) \leq t\} = \{(u,t) : u \leq F(t)\}.$$

In der Tat gilt für ein Element (u,t) der linken Menge, daß $F^-(u) \leq t$. Deshalb gilt

$$F(t) \geq F\left(F^-(u)\right) = F\left(\min\{t : F(t) \geq u\}\right) \geq u$$

und (u,t) ist in der rechten Menge enthalten. Sei umgekehrt $u \leq F(t)$. Da F^- wächst, gilt

$$F^-(u) \leq F^-(F(t)) = \min\{r : F(r) \geq F(t)\} = t,$$

wieder wegen der Rechtsstetigkeit. Somit gilt

$$\mathbb{P}(X \leq t) = \mathbb{P}\left(F^-(U) \leq t\right) = \mathbb{P}\left(U \leq F(t)\right) = F(t)$$

und der Beweis ist vollständig. □

Als einfaches Beispiel betrachten wir die Exponentialverteilung mit Verteilungsfunktion $F(t) = 1 - e^{-\alpha t}$, $\alpha > 0$. Die Inverse ist $F^{-1}(u) = -\frac{1}{\alpha}\ln(1-u)$. Nach dem Korollar ist $\eta = -\frac{1}{\alpha}\ln(1-U)$ exponentialverteilt und weil $1-U$ die gleiche Verteilung wie U hat, ist auch $\xi = -\frac{1}{\alpha}\ln U$ exponentialverteilt.
Ein Beispiel für die allgemeinere Aussage im Satz ist die Erzeugung diskreter Zufallsvariablen. Nimmt die Zufallsvariable ξ genau die Werte $w_1, \ldots, w_n$ mit den Wahrscheinlichkeiten $p_1, \ldots, p_n$, $p_i > 0$, an, so partitioniert man das Einheitsintervall $[0,1)$ in Intervalle I_j der Länge p_j und nimmt w_j als Realisierung, wenn U in I_j fällt. Dies entspricht genau der Bildung der verallgemeinerten Inversen (Skizze!).

Beweis von Satz 2.2.4, 2. Teil Wir wenden das Lemma auf die zur Verteilungsfunktion F gehörige Verteilung μ und $\varphi = F$ an. Dies ist zulässig, da F eine Dichte besitzt und somit stetig ist. Mit $c = n\binom{n-1}{s-1}$ folgt, daß

$$G_s(x) = \int_\infty^x c \cdot F(y)^{s-1}(1-F(y))^{n-s}\, d\mu(y) = \int_0^{F(x)} c \cdot y^{s-1}(1-y)^{n-s}\, dy.$$

Damit ist Satz 2.2.4 vollständig bewiesen. □

Korollar 2.2.8 *Seien $\xi_1, \ldots, \xi_n$ i.i.d. Zufallsvariablen. Dann hat der empirische Median $\mathbb{M}(\xi_1, \ldots, \xi_n)$ die Dichte*

$$g(x) = n\binom{n-1}{(n-1)/2} F^{(n-1)/2}(x)\left(1-F(x)\right)^{(n-1)/2} f(x). \tag{2.6}$$

Beweis Für den Median ist $s = (n+1)/2$, $s-1 = (n-1)/2$ und $n-s = (n-1)/2$. □

Wir ziehen eine einfache aber wichtige Folgerung: In Spezialfällen ist der empirische Median erwartungstreu.

Korollar 2.2.9 *Die Zufallsvariablen ξ_i, $1 \leq i \leq n$, seien i.i.d mit endlichem Erwartungswert $\mathbb{E}(\xi_i) = \mathbf{m}$ und um $\mathbf{m}$ symmetrischer Dichte f. Dann gilt für ungeraden Stichprobenumfang n, daß*

$$\mathbb{E}(\mathbb{M}(\xi_1, \ldots, \xi_n)) = \mathbb{E}(\xi_i) = \mathbf{m}.$$

Beweis Sei o.E. $\mathbf{m} = 0$. Der Faktor vor f in der Dichte g aus (2.6) ist beschränkt. Also impliziert die Existenz von

$$\mathbb{E}(\xi_i) = \int x f(x)\, dx$$

die Existenz von $\int xg(x)\, dx$. Des weiteren ist f symmetrisch und deshalb

$$F(-x) = 1 - F(x).$$

Somit gilt

$$F(-x)(1 - F(-x)) = (1 - F(x))F(x).$$

Potenzieren der beiden Seiten erhält die Gleichheit. Deshalb ist

$$F(x)^r(1 - F(x))^r f(x)$$

und somit ist g symmetrisch. Daraus folgt

$$\mathbb{E}(\mathbb{M}(\xi_1, \ldots, \xi_n)) = 0 = \mathbf{m}$$

wie behauptet. im Falle $\mathbf{m} \neq 0$ ersetze man die Zufallsvariablen ξ_i durch die Zufallsvariablen $\xi_i - \mathbf{m}$. □

Das empirische Mittel bei i.i.d. Zufallsvariablen ist eng mit dem Erwartungswert – dem theoretischen Mittel verknüpft. Genauso entspricht der empirische Median dem theoretischen Median. Der *theoretische Median* einer Zufallsvariablen ξ ist eine Zahl $\tilde{\mathbb{M}}$ mit

$$\mathbb{P}(\xi \leq \tilde{\mathbb{M}}) \geq \frac{1}{2}, \;\; \mathbb{P}(\xi \geq \tilde{\mathbb{M}}) \geq \frac{1}{2}.$$

Dadurch ist der theoretische Median natürlich nicht eindeutig bestimmt (man müßte ihn eigentlich als mengenwertige Abbildung definieren). Offensichtlich gelten:

$$\mathbb{P}\{\xi > \tilde{\mathbb{M}}\} \leq \frac{1}{2}, \;\; \mathbb{P}\{\xi < \tilde{\mathbb{M}}\} \leq \frac{1}{2} \tag{2.7}$$

Die Definition vereinfacht sich unter Stetigkeit:

Lemma 2.2.10 *Ist die Verteilungsfunktion F der Zufallsvariablen ξ stetig, so gilt*

$$\mathbb{P}(\xi \leq \tilde{\mathbb{M}}) = \frac{1}{2} = \mathbb{P}(\xi \geq \tilde{\mathbb{M}})$$

Beweis Dies ist klar. □

Der theoretische Median minimiert den erwarteten absoluten Abstand. Dies ist das Analogon zur *Mean Square* Eigenschaft des Erwartungswertes.

Satz 2.2.11 *Für eine integrierbare Zufallsvariable ξ gilt*

$$\tilde{\mathbb{M}} = \operatorname{argmin}_a \mathbb{E}|\xi - a|.$$

Beweis Sei $a \geq \tilde{\mathbb{M}}$. Falls $\tilde{\mathbb{M}} < \xi < a$, so gilt

$$|a-\xi| - |\tilde{\mathbb{M}} - \xi| = (a-\xi) - (\xi - \tilde{\mathbb{M}}) = a + \tilde{\mathbb{M}} - 2\xi = -(a - \tilde{\mathbb{M}}) + 2(a-\xi).$$

Unter Beachtung von (2.7) rechnet man

$$\begin{aligned}
& \mathbb{E}(|\xi - a|) - \mathbb{E}(|\xi - \tilde{\mathbb{M}}|) = \int |\xi - a| \, d\mathbb{P} - \int |\xi - \tilde{\mathbb{M}}| \, d\,\mathbb{P} \\
= \; & \int_{\xi \leq \tilde{\mathbb{M}}} (a - \tilde{\mathbb{M}}) \, d\,\mathbb{P} - \underbrace{\int_{\xi \geq a} (a - \tilde{\mathbb{M}}) \, d\,\mathbb{P} - \int_{\tilde{\mathbb{M}} < \xi < a} (a - \tilde{\mathbb{M}}) \, d\,\mathbb{P}}_{\xi > \tilde{\mathbb{M}}} \\
& \qquad + 2 \cdot \int_{\tilde{\mathbb{M}} < \xi < a} (a - \xi) \, d\,\mathbb{P} \\
= \; & \int_{\xi \leq \tilde{\mathbb{M}}} (a - \tilde{\mathbb{M}}) \, d\,\mathbb{P} - \int_{\xi > \tilde{\mathbb{M}}} (a - \tilde{\mathbb{M}}) \, d\,\mathbb{P} + 2 \cdot \int_{\tilde{\mathbb{M}} < \xi < a} (a - \tilde{\mathbb{M}}) \, d\,\mathbb{P} \\
= \; & (a - \tilde{\mathbb{M}})(\mathbb{P}(\xi \leq \tilde{\mathbb{M}}) - \mathbb{P}(\xi > \tilde{\mathbb{M}}) + 2 \int_{\tilde{\mathbb{M}} < \xi < a} (a - \tilde{\mathbb{M}}) \, d\,\mathbb{P} \\
\geq \; & (a - \tilde{\mathbb{M}})(\frac{1}{2} - \frac{1}{2}) + 2 \int_{\tilde{\mathbb{M}} < \xi < a} (a - \tilde{\mathbb{M}}) \, d\,\mathbb{P} \geq 0.
\end{aligned}$$

Der Fall $a < \tilde{\mathbb{M}}$ geht ähnlich. □

Bemerkung 2.2.12 *Seien $x_1, \ldots, x_n \in \mathbb{R}$ gegeben. Sei ξ eine Zufallsvariable mit $\mathbb{P}(\xi = x_i) = 1/n$. Dann hat ξ die diskrete Verteilung $\mu = (1/n)\sum_{i=1}^n \varepsilon_{x_i}$. Es gilt*

$$\frac{1}{n}\sum_{i=1}^{n} |x_i - a| = \mathbb{E}_\mu(|\xi - a|).$$

Ebenso ist $\mathbb{M}(x_1, \ldots, x_n)$ der theoretische Median zu μ. Der Satz liefert also:

$$\mathbb{M}(x_1, \ldots, x_n) = \operatorname{argmin}_a \; a \longmapsto \sum_{i=1}^{n} |x_i - a|$$

Sind die Beobachtungen ξ_i bzw. die Störungen η_i nicht symmetrisch verteilt, so liegen dennoch theoretischer Median und der Erwartungswert bei kleiner theoretischer Streuung eng beisammen. Dies ergibt sich aus der folgenden (nicht allzu bekannten) Abschätzung[4].

Satz 2.2.13 *Für jede integrierbare Zufallsvariable ξ mit Erwartungswert $\mathbf{m}$ und Streuung σ gilt*

$$|\tilde{\mathbb{M}} - \mathbf{m}| \leq \sigma.$$

Beweis Wegen der Jensenschen Ungleichung gilt

$$|\tilde{\mathbb{M}} - \mathbf{m}| = |\mathbb{E}(\tilde{\mathbb{M}} - \xi)| \leq \mathbb{E}(|\tilde{\mathbb{M}} - \xi|) \leq \mathbb{E}(|\mathbf{m} - \xi|) \leq \sqrt{\mathbb{E}(|\mathbf{m} - \xi|^2)} = \sigma.$$

Die zweite Ungleichung beruht auf (2.2.11), d.h. $\tilde{\mathbb{M}} = \operatorname{argmin}_a \mathbb{E}(|\xi - a|)$, und damit ist die Ungleichung verifiziert. □

2.2.2 Verrauschte Grauwertflächen

Verrauschte Grauwertflächen entsprechen dem Fall von i.i.d. Zufallsvariablen. Sogar in diesem einfachen Fall ist die Berechnung selbst der niedrigen Momente des empirischen Medians schwierig. Für Zufallsvariablen, die in $[0, 1]$ gleichverteilt sind, gelingt die exakte Berechnung.

[4] Persönliche Mitteilung H.v. Weizsäcker, Kaiserslautern, 3.1.1998

Beispiel 2.2.14 (Gleichverteiltes Rauschen) Die Zufallsvariablen $\xi_i, 1 \leq i \leq n$, seien unabhängig und in $[0,1]$ gleichverteilt. Dann hat $\xi_{(s)}$ nach Korollar 2.2.8 die Dichte

$$g_s(x) = n\binom{n-1}{s-1}x^{s-1}(1-x)^{n-s}, 0 \leq x \leq 1,$$

und $g_s(x) = 0$ sonst. Erwartungswert und Varianz haben die Gestalt

$$\mathbb{E}(\xi_{(s)}) = \frac{s}{n+1}, \ \mathbb{V}(\xi_{(s)}) = \frac{s(n-s+1)}{(n+1)^2(n+2)}. \tag{2.8}$$

g_s ist die Dichte einer *Beta-Verteilung* $\beta(u,v)$ mit $u = s$ und $v = n-s+1$. Wir erinnern daran, daß die Dichte der allgemeinen Beta-Verteilung $\beta(u,v)$ zu den Parametern u und v gegeben ist durch

$$\frac{1}{B(u,v)}t^{u-1}(1-t)^{v-1},$$

wobei $B(u,v)$ die *Beta-Funktion*[5] zu den Parametern u und v ist.

Beweis Wir berechnen zunächst den Normierungsfaktor $B(u,v)$. Wir haben es mit ganzzahligen Parametern zu tun. Es gilt allgemein, daß

$$B(u,k) = \frac{(k-1)!}{u(u+1)\dots(u+k-1)}, \ k = 1,2,\dots$$

Für unseren Spezialfall $u = s$ und $k = n-s+1$ ergibt das

$$B(s,n-s+1) = \frac{(n-s)!}{s(s+1)\dots(s+n-s+1-1)} = \frac{(n-s)!(s-1)!}{n(n-1)!} = \frac{1}{n\binom{n-1}{s-1}},$$

[5]Die Betafunktion (Eulersches Integral erster Gattung) ist formal gegeben durch

$$B(u,v) = \int_0^1 t^{u-1}(1-t)^{v-1}dt.$$

Dieses konvergiert für $u,v > 0$ (und divergiert für $u \leq 0$ oder $v \leq 0$). Ferner gelten

$$B(u,v) = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)}; \quad \text{und:} \quad \text{falls } u,v = 1,2,\dots \text{ so } \Gamma(u+1) = u!$$

mit der Gamma-Funktion Γ, der Verallgemeinerung der Fakultät.

Damit haben wir den Normierungsfaktor der interessierenden Betaverteilung. Der Erwartungswert ist

$$\begin{aligned}\mathbb{E}(\xi_{(s)}) &= \frac{1}{B(s, n-s+1)} \int_0^1 x x^{s-1}(1-x)^{n-s}\, dx \\ &= \frac{1}{B(s, n-s+1)} \int_0^1 x^{(s+1)-1}(1-x)^{n-s}\, dx.\end{aligned}$$

Deshalb berechnen wir noch

$$B(s+1, n-s+1) = \frac{(n-s)!}{(s+1)\dots(n+1)} = \frac{(n-s)!s!}{(n+1)!};$$

Wir multiplizieren mit der Normierung und erhalten:

$$\mathbb{E}(\xi_{(s)}) = \frac{n(n-1)!s!(n-s)!}{(s-1)!(n-s)!(n+1)!} = \frac{s}{n+1}$$

(für den Spezialfall des Medians ergibt sich natürlich wegen der Symmetrie der Gleichverteilung um 1/2 sofort $\mathbb{E}(\xi_{(n+1)/2}) = 1/2$ aus Satz 2.2.9). Für die Varianz berechnen wir analog das zweite Moment des Integrals:

$$B(s+2, n-s+1) = \frac{(n-s)!}{(s+2)\dots(n+2)} = \frac{(n-s)!(s+1)!}{(n+2)!}$$

Multiplikation mit dem Vorfaktor ergibt:

$$\frac{n(n-1)!(s+1)!(n-s)!}{(s-1)!(n-s)!(n+2)!} = \frac{s(s+1)}{(n+1)\cdot(n+2)} = \mathbb{E}(\xi_{(s)}^2)$$

Die Verschiebungsformel gibt die Varianz:

$$\begin{aligned}\mathbb{V}(\xi_{(s)}) &= \mathbb{E}(\xi_{(s)}^2) - \mathbb{E}(\xi_{(s)})^2 \\ &= \frac{s(s+1)}{(n+1)(n+2)} - \frac{s^2}{(n+1)^2} \\ &= \frac{s}{n+1}\left(\frac{(s+1)(n+1) - s(n+2)}{(n+1)(n+2)}\right) \\ &= \frac{s}{(n+1)^2(n+2)}(sn + s + n + 1 - sn - 2s) \\ &= \frac{s(n-s+1)}{(n+1)^2(n+2)}\end{aligned}$$

Damit sind die Formeln verifiziert. □

Für den Median ist $s = (n+1)/2$, d.h.

$$\begin{aligned}\mathbb{E}(\mathbb{M}(\xi_1,\ldots,\xi_n)) &= \frac{1}{2};\\ \mathbb{V}(\mathbb{M}(\xi_1,\ldots,\xi_n)) &= \frac{1}{4}\frac{(n+1)(n+1)}{(n+1)^2(n+2)} = \frac{1}{4(n+2)}.\end{aligned}$$

Es ist

$$\begin{aligned}\mathbb{E}(\xi_i) &= \frac{1}{2};\\ \mathbb{V}(\xi_i) &= \int_0^1 x^2dx - \frac{1}{4} = \frac{1}{3}x^3\Big|_0^1 - \frac{1}{4} = \frac{1}{3} - \frac{1}{4} = \frac{1}{12},\end{aligned}$$

und daher:

$$\mathbb{V}(\mathbb{M}(\xi_1,\ldots,\xi_n)) = \frac{3}{n+2}\cdot\mathbb{V}(\xi_i).$$

Wir vergleichen mit

$$\mathbb{V}(\bar{\xi}) = \frac{1}{n}\cdot\mathbb{V}(\xi_i)$$

und sehen, daß der Median hier schlecht abschneidet. Für ein 5×5 Fenster mit $n = 25$ z.B. ergibt sich mit $\sigma^2 = \mathbb{V}(\xi_i)$:

$$\mathbb{V}(\bar{\xi}) = \frac{\sigma^2}{25};\ \ \mathbb{V}(\mathbb{M}(\xi_1,\ldots,\xi_n)) = \frac{\sigma^2}{9}.$$

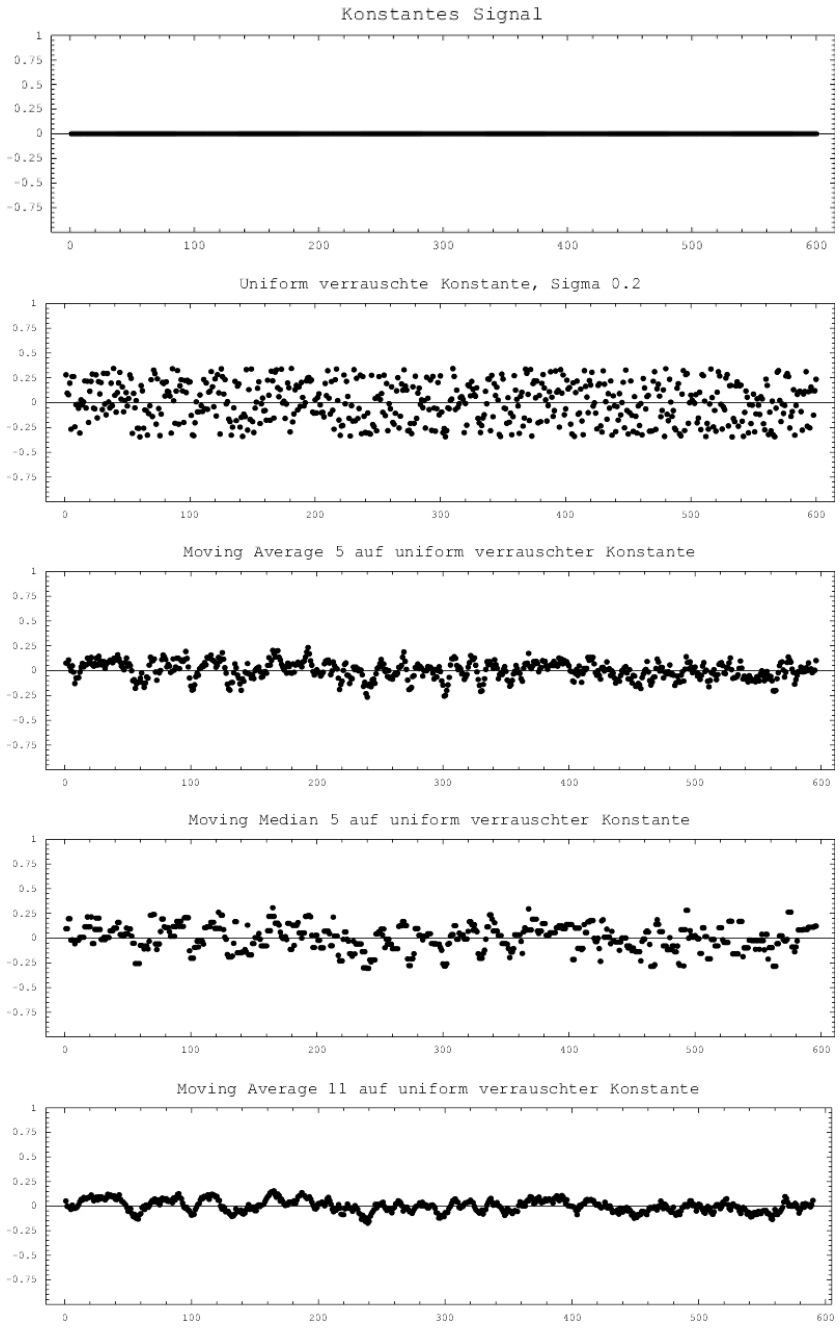


Abbildung 2.5: Konstantes Signal, mit additivem zentriertem weißem auf $[-\sqrt{3}/5, \sqrt{3}/5]$ gleichverteiltem Rauschen; die Streuung ist 0,2. Gefiltert mit: Moving Average bzw. Moving Median der Maskenbreite 5 und Moving Average der Maskenbreite 11.

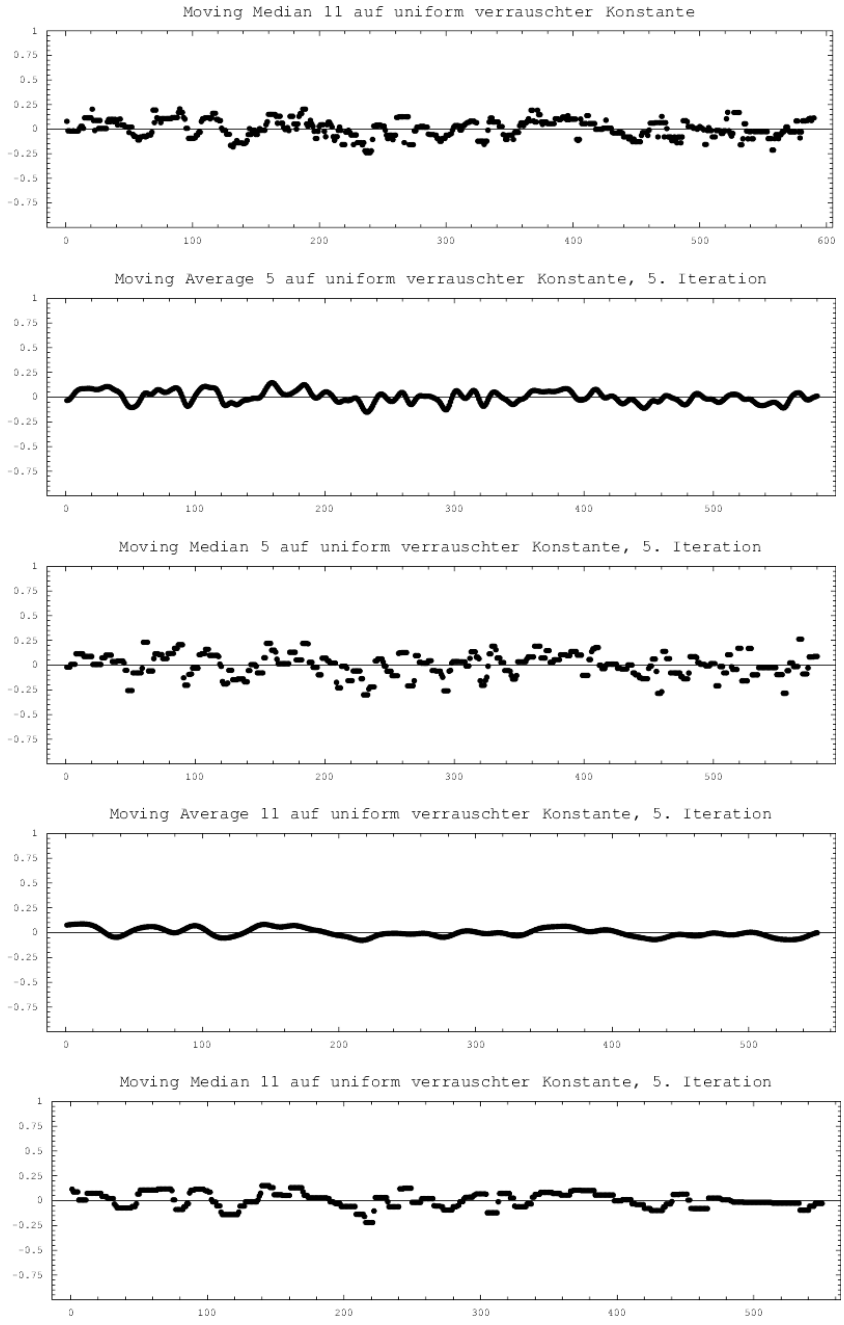


Abbildung 2.6: Uniform verrauschtes konstantes Signal aus Abb. 2.5 gefiltert mit Moving Median der Maskenbreite 11 sowie nach 5 maliger iterierter Anwendung dieser Filter.

Der tiefere Grund dafür ist, daß *auf den Tails der Dichte der Gleichverteilung kein Gewicht liegt.* Die Wirkung von gleitenden Medianen und iterierten gleitenden Medianen auf uniform verrauschte Konstanten ist in den Abbildungen 2.5 und 2.6 illustriert.

Selbst in Standardfällen wie gaußischem weißem Rauschen kann die Varianz des Medianes nach [19], Example 5.2, nur noch durch numerische Integration bestimmt werden.

Bemerkung 2.2.15 *Allgemein* gilt (siehe [19], (5.9)): Der empirische Median $\mathbb{M}((\xi_1,\ldots,\xi_n)$ ist für große n approximativ $\mathcal{N}(\tilde{\mathbb{M}},\sigma_n^2)$ - also gaußisch - verteilt mit Varianz

$$\mathbb{V}(\mathbb{M}(\xi_1,\ldots,\xi_n)) \approx \frac{1}{4nf^2(\tilde{\mathbb{M}})} = \sigma_n^2,$$

wobei $\tilde{\mathbb{M}}$ der theoretische Median ist. Für kleine n ersetzt man n durch $n+b$, mit $b = 1/(4f^2(\tilde{\mathbb{M}})\sigma^2) - 1$:

$$\mathbb{V}(\mathbb{M}(\xi_1,\ldots,\xi_n)) \approx \frac{1}{4(n+1/(4f^2(\tilde{\mathbb{M}})\sigma^2)-1)f^2(\tilde{\mathbb{M}})}. \tag{2.9}$$

Diese Formel ist für $n = 1$ exakt, d.h. der Bruch hat den Wert σ^2.

Wir betrachten nun Verteilungen mit zunehmenden Gewicht auf den Tails.

Beispiel 2.2.16 (Gaußisches weißes Rauschen) Sind die ξ_i gaußisch verteilt gemäß $\mathcal{N}(\mu,\sigma^2)$, so ist $f^2(\tilde{\mathbb{M}}) = 2\pi\sigma^2$ und (2.9) wird zu:

$$\mathbb{V}(\mathbb{M}(\xi_1,\ldots,\xi_n)) \approx \frac{\sigma^2}{n+\frac{\pi}{2}-1}\cdot\frac{\pi}{2},\ \ n = 1,3,5,\ldots.$$

Die Approximation ist für ungerades n sehr gut. Wir stellen dies

$$\mathbb{V}(\bar{\xi}) = \frac{\sigma^2}{n}$$

gegenüber: Es ist

$$\frac{\mathbb{V}(\bar{\xi})}{\mathbb{V}(\mathbb{M})} = \frac{2}{\pi}\frac{n+\pi/2-1}{n} = (2/\pi)(1+(\pi-2)/(2n)).$$

Der Term $(\pi-2)/(2n)$ hat schon für eine 3×3-Maske einen Wert kleiner $0,57/9 < 0,063$. Der Vorfaktor ist etwa 0,64 mit Kehrwert 1,57.

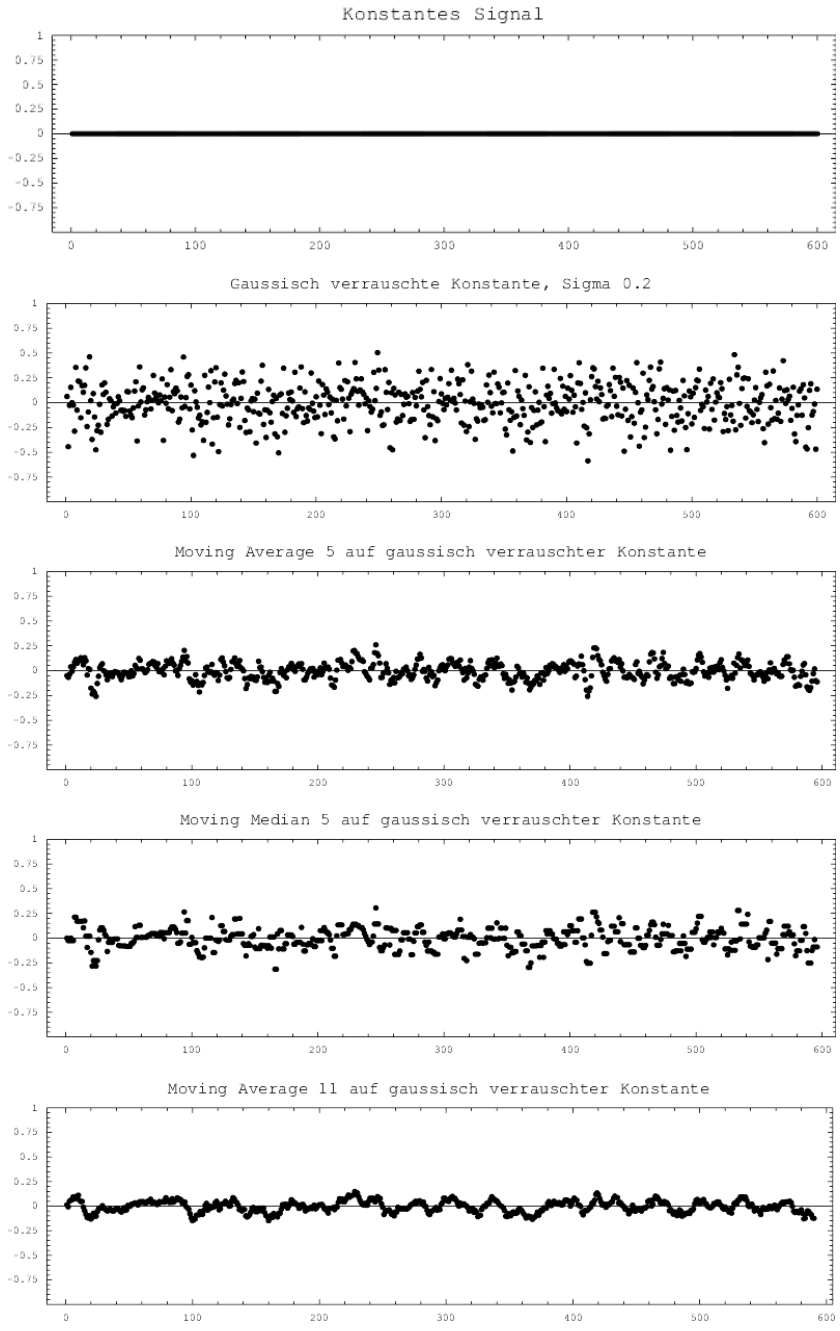


Abbildung 2.7: Konstantes Signal, mit additivem zentriertem weißem gaußischem Rauschen mit Streuung 0,2. Gefiltert mit: Moving Average bzw. Moving Median der Maskenbreite 5 und Moving Average der Maskenbreite 11.

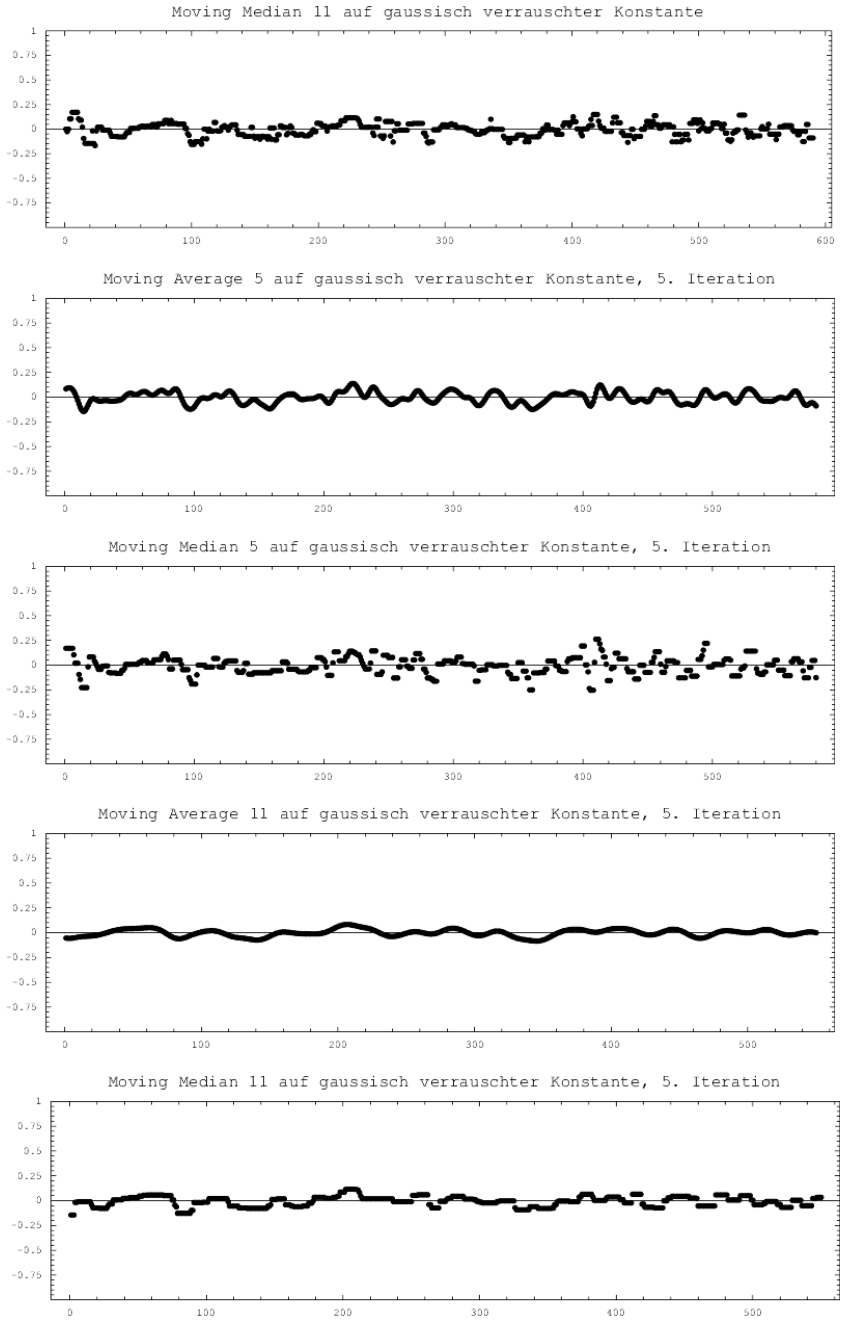


Abbildung 2.8: Verrauschtes konstantes Signal aus Abb. 2.7 gefiltert mit Moving Median der Maskenbreite 11 und nach 5 maliger iterierter Anwendung der Filter.

Also ist $\mathbb{V}(\mathbb{M}(\xi_1,\ldots,\xi_n))$ etwa 57% größer als $\mathbb{V}(\bar{\xi})$ (bei gleicher Fenstergröße). Für die Streuung sind es noch etwa 25 %. Man kann das durch größere Fenster ausgleichen, d.h. n um 57% vergrößern. *Wir merken an, daß die Gaußverteilung* wenig *Gewicht auf den Tails hat.* Die Abbildungen 2.7 und 2.8 illustrieren den diskutierten Effekt.

Beispiel 2.2.17 (Laplace-verteiltes Rauschen) Die ξ_i seien *doppelt exponential verteilt* (oder *Laplace-verteilt*) mit der Dichte

$$f(x) = \frac{1}{\sqrt{2}\sigma} \exp\left(-\frac{\sqrt{2}}{\sigma}|x-\mathbf{m}|\right).$$

Da die Verteilung symmetrisch ist, sind Erwartungswert und theoretischer Median gleich $\mathbf{m}$. Um die Varianz zu berechnen, erinnern wir uns an die *Exponentialverteilung* zum Parameter $\lambda > 0$ mit der Dichte

$$h_\lambda(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{falls } x \geq 0 \\ 0 & \text{sonst} \end{cases}.$$

Wir wissen, daß sie den Erwartungswert λ^{-1} und die Varianz λ^{-2} hat. Daraus schließen wir mit Hilfe der Verschiebungsformel für eine exponentialverteilte Zufallsvariable η, daß

$$\mathbb{E}(\eta^2) = \mathbb{V}(\eta) + \mathbb{E}(\eta)^2 = 2/\lambda^2.$$

Daraus folgt

$$\int_0^\infty x^2 \exp(-\lambda x)\, dx = 2/\lambda^3.$$

Nun gilt

$$\int_0^\infty \exp(-\lambda x)\, dx = \lambda^{-1} \exp(-\lambda x)\big|_\infty^0 = \lambda^{-1}, \quad \int_{-\infty}^\infty \exp(-\lambda x)\, dx = 2/\lambda.$$

Das ist der Normierungsfaktor der Laplace-Verteilung mit $\lambda = \sqrt{2}/\sigma$. Wir fassen dies zusammen:

$$\frac{\lambda}{2}\int_{-\infty}^\infty x^2 \exp(-\lambda x)\, dx = 2 \cdot \frac{\lambda}{2}\int_0^\infty x^2 \exp(-\lambda x)\, dx = \frac{2}{\lambda} = 2 \cdot \frac{\lambda}{2} \cdot \frac{2}{\lambda^3} = \frac{2}{\lambda^2}.$$

Mit $\lambda = \sqrt{2}/\sigma$ ergibt sich also für die Varianz gerade σ^2, was die obige Schreibweise der Dichte der Laplaceverteilung rechtfertigt. Mit der Approximation (2.9) erhält man:

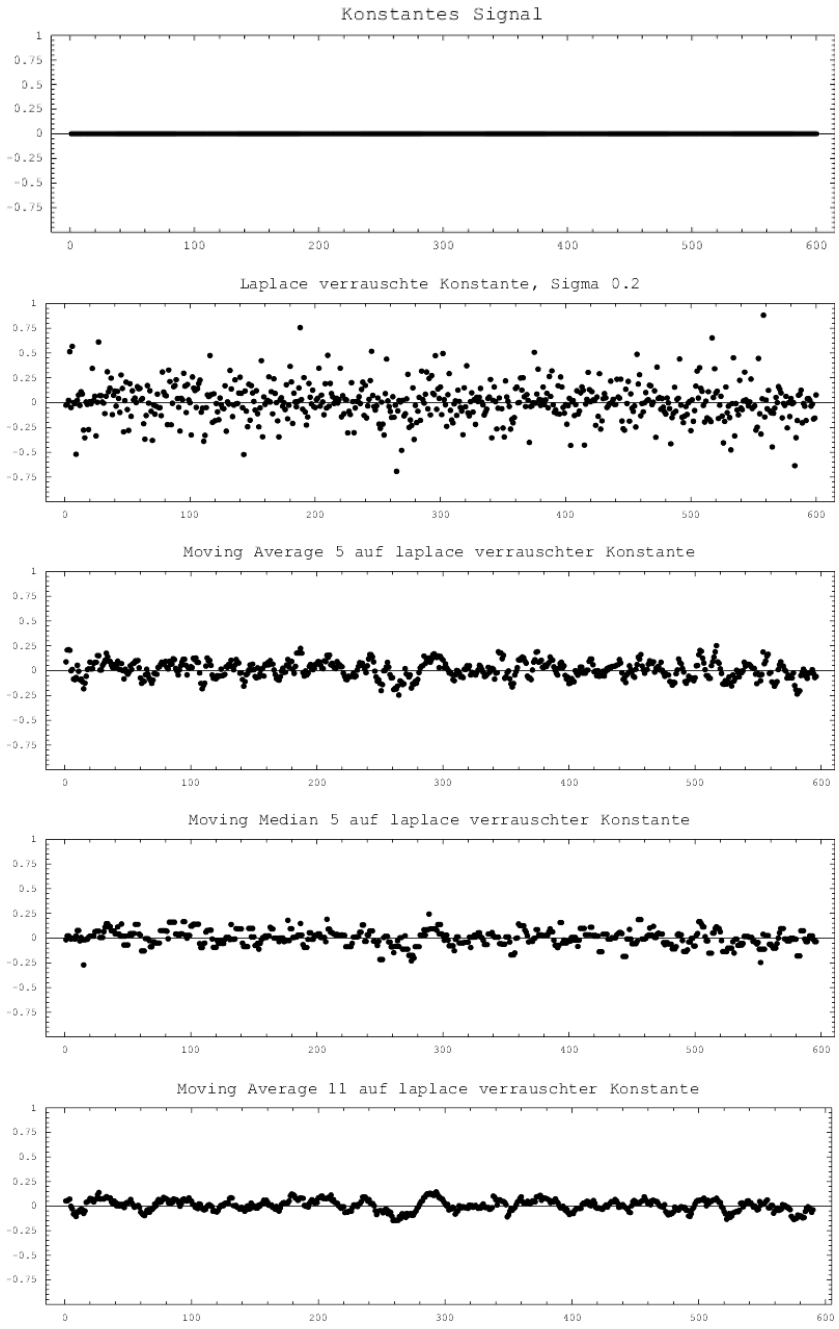


Abbildung 2.9: Konstantes Signal, mit additivem zentriertem weißem Laplaceschem Rauschen mit Streuung 0,2. Gefiltert mit: Moving Average bzw. Moving Median der Maskenbreite 5 und Moving Average der Maskenbreite 11.

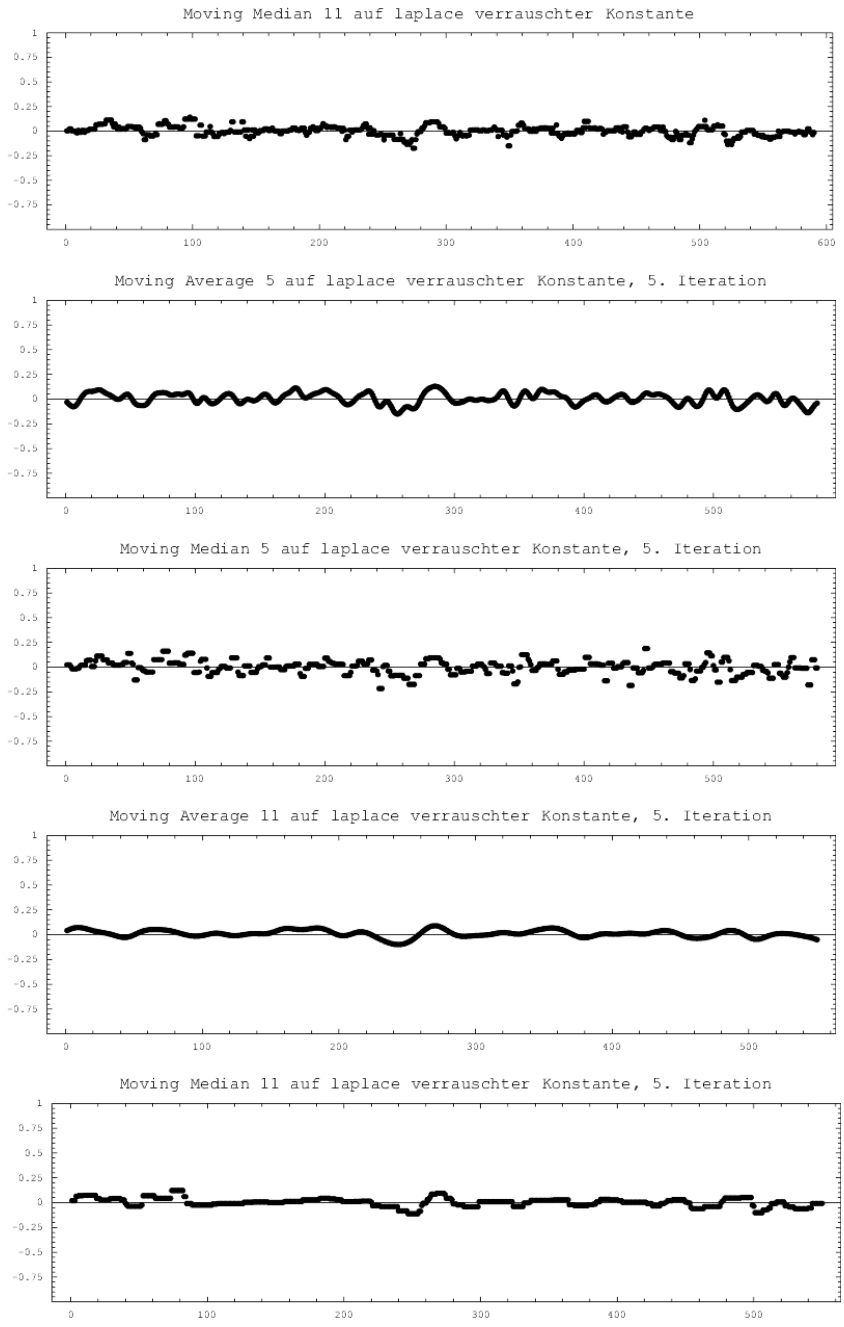


Abbildung 2.10: Signale aus Abb. 2.9 Gefiltert mit Moving Median der Maskenbreite 11 nach 5 maliger iterierter Anwendung von Moving Average und Moving Median der Maskenbreite 5 und 11

$$\mathbb{V}(\mathbb{M}(\xi_1,\ldots,\xi_n)) \approx \frac{2\sigma^2}{4\left(n+\frac{2\sigma^2}{4\sigma^2}-1\right)} = \frac{\sigma^2}{2(n-1/2)}.$$

Dies ist etwa halb so groß wie die Varianz σ^2/n des empirischen Mittels $\bar{\xi}$. Der Median schneidet hier also deutlich besser ab, als das empirische Mittel. *Wir merken an, daß die doppelte Exponentialverteilung relativ* viel *Gewicht auf den Tails hat.*

Bemerkung 2.2.18 Es wurde argumentiert, daß der empirische Median robuster als das empirische Mittel ist. Dazu wurde deren Wirkung auf die Gleich-, Gauß- und Laplaceverteilung untersucht. In Abbildung 2.11 sind Dichten von Gleich-, Gauß- und Laplaceverteilung mit Standardabweichung 0,2 dargestellt[6]. Offensichtlich nimmt die Masse innerhalb des dargestellten Bereiches von der Gleich- über die Gauß- zur Laplaceverteilung hin ab und dementsprechend die Masse auf den Tails zu. Dies wird noch deutlicher in Abbildung 2.12 illustriert, die einen Auschnitt der beiden Dichten 'weiter draußen' zeigt.

Wir zeigen, daß der empirische Median für die (*heavytailed*) doppelte Exponentialverteilung der Maximumlikelihoodschätzer ist (wie dies das empirische Mittel für die Gaußverteilung war). Wir formulieren den Kern dieser Beobachtung gesondert.

Satz 2.2.19 *Sei $x_1,\ldots,x_{2k+1}$ eine Stichprobe. Dann gilt für den empirischen Median $\mathbb{M}$, daß*

$$m = \mathbb{M} = \operatorname{argmin}_a \sum_i |x_i - a|. \tag{2.10}$$

Beweis Dies ist ein Spezialfall von Satz 2.2.11, siehe auch Bemerkung 2.2.12. Wir geben noch einen netten direkten Beweis an: Sei o.E. die Stichprobe geordnet und (aus Gründen der Notation) geeignet numeriert:

$$x_{-k} \le \cdots \le x_0 = \mathbb{M} \le \cdots \le x_k$$

Sei a eine beliebige Zahl. Dann gelten

$$\begin{aligned} |x_{-i}-x_0| + |x_i - x_0| = |x_i - x_{-i}| &\le |x_i - a| + |x_{-i} - a|, \ i = 1,\ldots,k, \\ |x_0 - x_0| &\le |x_0 - a|. \end{aligned}$$

[6] Zentriertes gleichverteiltes Rauschen auf $[-\sqrt{3}/5, \sqrt{3}/5]$ hat Standardabweichung 0,2.

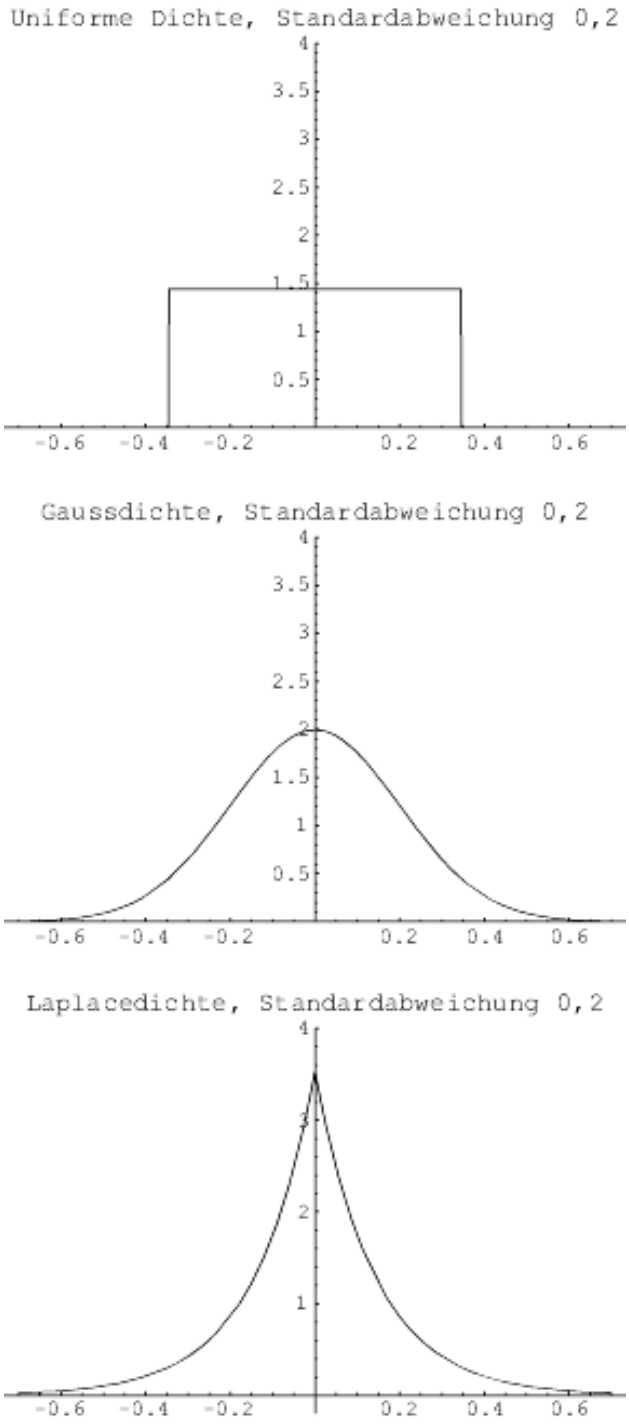


Abbildung 2.11: Dichten der Gleich-, Gauß- und Laplaceverteilung mit Standardabweichung 0,2. Für diese Standardabweichung ist die Gleichverteilung auf dem Intervall $[-\sqrt{3}/5, \sqrt{3}/5]$ konzentriert.

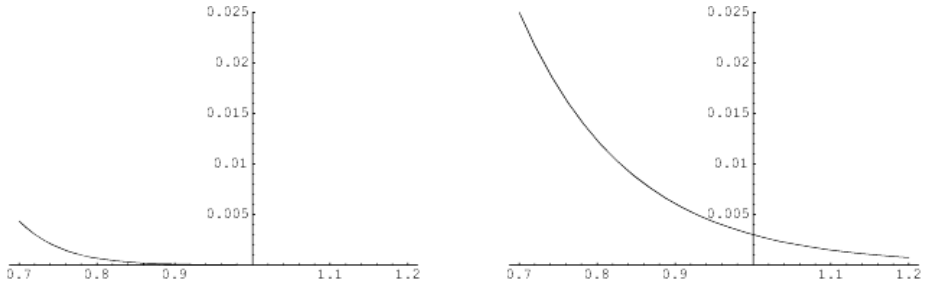


Abbildung 2.12: Ausschnitte der Tails für die Gauß- und die Laplaceverteilung der Standardabweichung 0,2

Wegen $x_0 = \mathbb{M}$ ergibt Summation

$$\sum_{i=-k}^{k} |x_i - \mathbb{M}| \leq \sum_{i=-k}^{k} |x_i - a|.$$

Damit ist (2.10) verifiziert. □

Wir folgern

Satz 2.2.20 *Für die doppelte Exponentialverteilung ist der empirische Median der Maximumlikelihoodschätzer für Erwartungswert und theoretischen Median.*

Beweis Da die Verteilung symmetrisch um μ ist, sind Erwartungswert und theoretischer Median gleich μ und der Satz ist mit (2.10) bewiesen. □

Satz 2.2.13 hat eine weitere interessante Konsequenz. Sie besagt, daß bei kleinem Rausch-Signalverhältnis, d.h. wenn das (additive, weiße) Rauschen kleine Streuung hat, sich die Wirkung von empirischem Mittel und empirischem Median nur wenig unterscheiden.

Proposition 2.2.21 *Für Werte $x_1, \ldots, x_n$ gilt*

$$|\mathbb{M} - \bar{x}| \leq \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2} \ (:= s'_n).$$

Die Größe s'_n ist *eine* Form der empirischen Streuung. In der Statistik wird sie allerdings nicht benutzt, weil sie bei quadratintegrierbaren i.i.d. Zufallsvariablen nicht erwartungstreu für die theoretische Streuung

$$\sigma(\xi_i) = \sqrt{\mathbb{E}((\xi_i - \mu)^2)}$$

ist. Erwartungstreu für σ ist vielmehr die empirische Größe

$$s_n = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (\xi_i - \bar{\xi})^2}.$$

Beweis von Proposition 2.2.21 Wir wenden Satz 2.2.13 auf die Zufallsvariable ξ an, welche die Werte $x_1, \ldots, x_n$ jeweils mit Wahrscheinlichkeit $1/n$ annimmt. Für ξ gilt

$$\mathbb{E}(\xi) = \sum_{i=1}^{n} x_i \cdot \frac{1}{n} = \bar{x}, \;\; \mathbb{V}(\xi) = \sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \frac{1}{n} = (s'_n)^2.$$

Also ergibt sich die behauptete Ungleichung. □

Bei i.i.d. Zufallsvariablen ist bei größerem Stichprobenumfang n die theoretische Streuung nahe bei s_n, was die vor dem Satz aufgestellte Behauptung untermauert.

Im Mittel, d.h. im stochastischen Kontext, gilt:

Proposition 2.2.22 *Für integrierbare i.i.d. Zufallsvariablen $\xi_1, \ldots, \xi_n$ mit Streuung σ gilt*

$$\mathbb{E}(|\mathbb{M}(\xi_1, \ldots, \xi_n) - \bar{\xi}|) \leq \sqrt{\frac{n-1}{n}} \cdot \sigma.$$

Beweis Wir ersetzen in Lemma 2.2.21 die Größen x_i durch die Realisierungen $\xi_i(\omega)$ der Zufallsvariablen $\xi_i : \Omega \to \mathbb{R}$. Integration bzw. Bildung des Erwartungswertes liefert

$$\mathbb{E}(|\mathbb{M} - \bar{\xi_i}| \leq \mathbb{E}\left(\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (\xi_i - \bar{\xi})^2}\right) = \sqrt{\frac{n-1}{n}} \cdot \sigma;$$

dafür wurde die Erwartungstreue von s_n benutzt. Damit ist der Beweis vollständig. □

Obige Betrachtungen gelten allgemein für die Pixel eines Bildes mit einer konstanten, die Filtermaske enthaltenden Umgebung. Sie legen nahe, daß der Medianfilter seine Stärke bei Verteilungen, die 'heavy tailed' sind, beweist.

Wir betrachten nun zwei Beispiele mit impulsartigem nichtadditivem Rauschen. Insbesondere verschwindet der Erwartungswert der Rauschvariablen nicht. Sei wieder $(x_s)_{s\in S}$ das Idealbild, $\xi = (\xi_s)_{s\in S}$ repräsentiere die (zufälligen) Daten. Im ersten Fall führen alle Fehler zum selben Grauwert.

Beispiel 2.2.23 (Impulsrauschen, fester Wert) Sei S eine Menge von Zeitpunkten, Pixeln oder Voxeln. Das Signal (x_s) sei durch Impuls- oder Kanalrauschen mit festem Fehler gestört, d.h. man beobachtet i.i.d.Zufallsvariablen $\xi_s, s \in S$ mit

$$\xi_s = \begin{cases} d & \text{mit Wahrscheinlichkeit} \quad p \\ x_s & \text{mit Wahrscheinlichkeit} \quad q = 1-p. \end{cases}$$

Wir betrachten wieder den einfachsten Fall des konstanten Signals $x_s \equiv c$ mit $c \neq d$. Sei nun A mit ungeradem $n = |A|$ eine Filtermaske um das aktuell zu bearbeitende Pixel. $\mathbb{M}(\xi_t : t \in A)$ bezeichne wieder den empirischen Median. Die Rekonstruktion ist korrekt im Pixel s, d.h. $\mathbb{M}(\xi_t : t \in A) = x_s$, genau dann, wenn die Anzahl der Fehler in A höchstens gleich $(n-1)/2 = v$ ist. Die Wahrscheinlichkeit dafür ist

$$\mathbb{P}(\mathbb{M}(\xi_t : t \in A) = x_s) = \sum_{k=0}^{v} \binom{n}{k} p^k q^{n-k} \equiv Q(n,p). \tag{2.11}$$

Die Tabelle 2.1 enthält Werte für $1 - Q(n,p)$; diese zeigen, daß für z.B. $p \leq 0,3$ die Rauschunterdrückung auch für kleine Masken recht hoch ist.

Das Kanalrauschen erzeuge nun zufällige Fehler.

Beispiel 2.2.24 (Impulsrauschen, gleichverteilt) Es sei wieder $x = (x_s)$ das Signal. Die Rauschvariablen $\eta_s, s \in S$, seien i.i.d. und gleichverteilt in $[0, d\,]$. Beobachtet werden Zufallsvariablen ξ_s gegeben durch

$$\xi_s = \begin{cases} \eta_s & \text{mit Wahrscheinlichkeit} \quad p \\ x_s & \text{mit Wahrscheinlichkeit} \quad 1-p. \end{cases}$$

Wir betrachten wieder nur ein konstantes Signal, d.h. wir nehmen ohne Einschränkung an, daß $x_s \equiv 0$.

Fehlerrate p	$1-Q(n,p)$ für Fenstergröße n			
	3	5	9	25
0.01	0.00030	0.0000099	0.0000000	0
0.05	0.00725	0.0016000	0.0000330	0.0000000
0.10	0.02800	0.0086000	0.0008900	0.0000002
0.15	0.06080	0.0266000	0.0056300	0.0000170
0.20	0.10400	0.0580000	0.0196000	0.0003700
0.30	0.21600	0.1630000	0.0990000	0.0170000
0.40	0.35200	0.3170000	0.2670000	0.1540000
0.50	0.50000	0.5000000	0.5000000	0.5000000

Tabelle 2.1: 'Mißklassifikationsraten' $1-Q(n,p)$ für das Impulsrauschen in Beispiel 2.2.23

Sei wieder A eine Filtermaske um das aktuelle Pixel mit ungerader Kardinalität $|A|$. Wir indizieren die Zufallsvariablen in A und schreiben $\xi_1,\ldots,\xi_n$ etc. Die Wahrscheinlichkeit korrekter Rekonstruktion ist wie oben in (2.11). Wir interessieren uns für die Rauschunterdrückung. Es gilt:

Satz 2.2.25 *Im Modell Beispiel 2.2.24 gilt mit $u=(n+1)/2$ für den Erwartungswert der Rekonstruktion, daß*

$$\mathbb{E}(\mathbb{M}(\xi_1,\ldots,\xi_n)) = d\cdot\sum_{k=u}^{n}\frac{k-(n-1)/2}{k+1}\binom{n}{k}p^k q^{n-k},$$

Unter der Bedingung, daß die Rekonstruktion falsch ist, gilt

$$\mathbb{E}(\mathbb{M}(\xi_1,\ldots,\xi_n)) \mid \mathbb{M}(\xi_1,\ldots,\xi_n)\neq x_s) = \mathbb{E}(\mathbb{M}(\xi_1,\ldots,\xi_n))(1-Q(n,p)).$$

Beweis Die zufällige Anzahl κ der Fehler innerhalb der Filtermaske A um das Pixel s ist binomialverteilt. Unter der Bedingung $\{\kappa=k\}$, daß innerhalb der Maske genau k Fehler auftreten, gilt ohne Einschränkung

$$(\xi_1,\ldots,\xi_n)\sim((\underbrace{0,\ldots,0}_{n-k\text{ mal}},\eta_1,\ldots,\eta_k)),$$

wobei für Zufallsvektoren ξ und η das Symbol $\xi\sim\eta$ bedeutet, daß ξ wie η verteilt ist. Offensichtlich nimmt der Median den Wert 0 an, wenn weniger

als $(n-1)/2$ Fehler vorliegen. Sind es mehr, so ist der Median der r-t kleinste Wert unter den $\eta_1, \ldots, \eta_k$ mit $r = (n+1)/2-(n-k) = k-(n-1)/2$, also die r-te Ordnungsstatistik $\eta_{(k,r)}$ zum Stichprobenumfang k. Zusammenfassend gilt also

$$\mathbb{M}(0, \ldots, 0, u_1, \ldots, u_k) = \begin{cases} 0 & \text{falls } k \leq (n-1)/2 \\ \eta_{(k,r)} & \text{falls } k \geq (n+1)/2 \end{cases}, .$$

Wir zerlegen den Erwartungswert nach den Bedingungen $\{\kappa = k\}$. Weil

$$\mathbb{E}(\mathbb{M}|\kappa = k) = 0, \;\; \kappa < (n+1)/2,$$

ergibt dies

$$\begin{aligned} \mathbb{E}(\mathbb{M}) &= \sum_{k=0}^{n} \mathbb{E}(\mathbb{M}|\kappa = k)\mathbb{P}(\kappa = k) \\ &= \sum_{k=(n+1)/2}^{n} \mathbb{E}(\mathbb{M}|\kappa = k)\mathbb{P}(\kappa = k) \qquad (2.12) \\ &= \sum_{k=(n+1)/2}^{n} \mathbb{E}(\eta_{(k,r)})\mathbb{P}(\kappa = k). \end{aligned}$$

Gemäß (2.8) gilt

$$\mathbb{E}(\eta_{(s)}) = d \cdot \frac{s}{k+1}.$$

und deshalb

$$\mathbb{E}(\eta_{(k,k-(n-1)/2,k)}) = d \cdot \frac{2k-n+1}{2(k+1)}.$$

Schließlich ergibt sich durch Einsetzen in (2.12) für den Erwartungswert

$$\mathbb{E}(\mathbb{M}) = d \cdot \sum_{k=(n+1)/2}^{n} \frac{k-(n-1)/2}{k+1} \binom{n}{k} p^k q^{n-k}.$$

Zur Berechnung des bedingten Erwartungswertes setze man in (2.12) statt $\mathbb{P}(\kappa = k)$ die bedingte Verteilung

$$\mathbb{P}(\kappa = k|\text{Fehler in } s) = \frac{\mathbb{P}(\kappa = k, \mathbb{M} \neq x_s)}{\mathbb{P}(\mathbb{M} \neq x_s)}$$

ein. Da die Zufallsvariablen η_i den Wert x_s fast sicher nicht annehmen, gilt

$$\mathbb{P}(\kappa = k, \mathbb{M} \neq x_s) = \mathbb{P}(\kappa = k), \;\; k \geq (n+1)/2.$$

Damit folgt

$$\mathbb{P}(\kappa = k|\text{Fehler in } s) = \frac{\mathbb{P}(\kappa = k)}{\mathbb{P}(\mathbb{M} \neq x_s)} = \frac{\mathbb{P}(\kappa = k)}{1 - Q(n,p)}$$

und in Analogie zu (2.12)

$$\begin{aligned}\mathbb{E}(\mathbb{M}|\text{Fehler in } s) &= \sum_{k=(n+1)/2}^{n} \mathbb{E}(\eta_{(r)})\mathbb{P}(\kappa = k|\text{Fehler in } s) \\ &= (1 - Q(n,p)) \sum_{k=(n+1)/2}^{n} \mathbb{E}(\eta_{(r)})\mathbb{P}(\kappa = k) \\ &= (1 - Q(n,p)) \cdot \mathbb{E}(\mathbb{M}).\end{aligned}$$

Damit ist auch die Formel für den bedingten Erwartungswert verifiziert. □

Beispiel 2.2.26 Seien in Beispiel 2.2.24 etwa $n = 9$ und $p = 0,2$. Dann sollte der Anteil der Falschrekonstruktionen von $0,2$ auf etwa $1 - Q(9,0,2) = 0,0196$ gedrückt werden. Der Erwartungswert sollte von $\mathbb{E} = 0,5 \cdot d$ auf $0,187 \cdot d$ sinken.

2.2.3 Die verrauschte Kante

Die stochastische Analyse von verrauschten Kanten ist natürlich schwieriger als die der verrauschten Graufläche. Bei letzterer hatte man es ja mit dem klassischen Fall von i.i.d. Zufallsvariablen zu tun. Selbst bei unabhängigen Rauschvariablen kommen bei Kanten mehrere Mittelwerte ins Spiel.

Ein Beispiel ist die ideale eindimensionale Kante gegeben durch

$$x_i = h \cdot \chi_{[1,\infty)}, \;\; i \in \mathbb{Z}, \;\; h \geq 0.$$

Bei additivem Rauschen (η_i) hat man das Modell

$$\xi_i = x_i + \eta_i, \;\; i \in \mathbb{Z}.$$

Sind die η_i identisch verteilt wie η, gilt $\xi_i \sim \eta$ für $i \leq 0$ und $\xi_i \sim \eta + h$ für $i \geq 1$.

Für den Median spielt es nur eine Rolle, wie viele Variablen ξ_i innerhalb der Maske gemäß der einen bzw. anderen Verteilung verteilt sind. Die Geometrie einer Kante - etwa in $\mathbb{Z}^2$ - beeinflußt den Median nur über die Lage des

aktuellen Pixels und die Form der Filtermaske. Deshalb reduzieren sich die folgenden Betrachtungen auf Zufallsvariablen ξ_i, die zwar meist unabhängig sind, jedoch eine von zwei verschiedenen Verteilungen haben.

Beispiel 2.2.27 Bei additivem gaußischem weißem Rauschen sind die Beobachtungen verteilt gemäß

$$\xi_i \sim \mathcal{N}(0,\ \sigma^2),\ \ i \leq 0,\ \ \xi_i \sim \mathcal{N}(h, \sigma^2)),\ \ i \geq 1.$$

In der Nähe der Kante kommen Variablen beider Verteilungstypen in der Maske vor.

Ein erstes Qualitätskriterium für die Güte einer Rekonstruktion sind die ersten Momente. Sie hängen nun davon ab wieviele Variablen des jeweiligen Verteilungstyps in der Filtermaske enthalten sind.

Beispiel 2.2.28 Bei gaußischem weißem Rauschen lassen sich Erwartungswert und Varianz des empirischen Mittels $\bar{\xi}$ innerhalb einer Filtermaske sofort angeben. Seien n die Maskenlänge und p die Zahl der $\mathcal{N}(h, \sigma^2)$-verteilten Variablen in der Maske (und deshalb $n - p$ die Zahl der $\mathcal{N}(0, \sigma^2)$-verteilten Variablen in der Maske). Dann gilt:

$$\bar{\xi_p} \sim \mathcal{N}(hp/n, \sigma^2/n).$$

Beweis der Behauptung in 2.2.28 Wegen der Unabhängigkeit ist die Varianz klar. Für den Erwartungswert gilt natürlich

$$\mathbb{E}(\bar{\xi_p}) = \frac{1}{n}(p \cdot h + 0 \cdot (n - p)) = (ph)/n$$

wie behauptet. □

Zur Berechnung von Momenten und anderen Kenngrößen des Medians benötigen wir die Dichte. Diese ist natürlich nicht einfacher als im homogenen Fall zu berechnen.

Satz 2.2.29 *Seien n unabhängige Zufallsvariablen $\xi_1, \ldots, \xi_n$ gegeben. Für $p < n$ haben die Variablen $\xi_1, \ldots, \xi_p$ die Dichte f_1 und die Verteilungsfunktion F_1; die Variablen $\xi_{p+1}, \ldots, \xi_n$ haben die Dichte f_2 und die Verteilungsfunktion F_2. Dann hat die s-te Ordungsstatistik $\eta = \xi_{(s)}$ die Dichte*

$$g = g_1 + g_2,$$

wobei

$$g_1 = \sum_j p\binom{p-1}{j}\binom{n-p}{s-j-1} f_1 F_1^j F_2^{s-j-1}(1-F_1)^{p-j-1}(1-F_2)^{n-p-s+j+1}$$

$$g_2 = \sum_j (n-p)\binom{p}{j}\binom{n-p-1}{s-j-1} f_2 F_1^j F_2^{s-j-1}(1-F_1)^{p-j}(1-F_2)^{n-p-s+j}$$

Die Summation erstreckt sich dabei über alle Binomialkoeffizienten $\binom{u}{v}$ *mit* $u \geq v \geq 0$.

Wir skizzieren lediglich die Beweisidee. Die rigorose Umsetzung erfolgt genau wie in Satz 2.2.29.

Beweisskizze Wir betrachten die infinitesimale Menge

$$\Delta = [x, x + dx]$$

0 und die disjunkten infinitesimalen Ereignisse

$$\begin{aligned} A_1 &= \{\text{genau ein } \xi_1, \ldots, \xi_p \text{ in } \Delta\} \\ A_2 &= \{\text{genau ein } \xi_{p+1}, \ldots, \xi_n \text{ in } \Delta\}, \end{aligned}$$

sowie die paarweise disjunkten Ereignisse

$$B_j = \{\text{genau } j \text{ der } \xi_1, \ldots, \xi_p \text{ in } (-\infty, x),\ j = 0, 1, \ldots, p\}.$$

Dann gilt offensichtlich

$$\mathbb{P}(\eta \in \Delta) = \sum_{i,j} \mathbb{P}(\eta \in \Delta, A_i \cap B_j).$$

Wir betrachten ein Ereignis in der Summe mit $i = 1$, also für jedes $j = 0, 1, \ldots, p$ die Wahrscheinlichkeit

$$\mathbb{P}(\text{genau ein } \xi_1, \ldots, \xi_p \text{ in } \Delta,\ \text{genau } j \text{ der } \xi_1, \ldots, \xi_p \text{ in } (-\infty, x)).$$

Der entsprechende Term in der Summe von g_1 kommt so zustande:

- Man wählt ein Element gemäß f_1; das geht auf p Weisen, d.h. man bekommt den Faktor $p \cdot f_1(x)$.

- Eine feste Auswahl von j Elementen der ersten Gruppe ist links von x mit Wahrscheinlichkeit $F(x)^j$.
- Die übrigen Elemente der ersten Gruppe liegen rechts von $x + dx$ mit Wahrscheinlichkeit $(1 + F_1(x - dx))^{p-j-1}$.
- Das geht auf $\binom{p-1}{j}$ Weisen, was den entsprechenden Multiplikator

  $$p\binom{p-1}{j} \cdot f_1(x)F(x)^j(1 - F_1(x + dx))^{p-j-1}$$

  ergibt.
- Nun sind $j + 1$ Elemente verbraucht. Also müssen genau $s - j - 1$ Element der zweiten Gruppe links von x liegen; die Wahrscheinlichkeit ist genau mit der obigen Argumentation

  $$\binom{n-p}{s-j-1} F_2(x)^{s-j-1}(1 - F_2(x + dx))^{n-p-s+j+1}.$$

  Multiplikation der einzelnen Faktoren und Summation über j ergibt nun g_1.

Den zweiten Term g_2 erhält man ebenso. □

Mit Hilfe der Dichte lassen sich nun Momente des Medians numerisch berechnen.

Beispiel 2.2.30 Die Zufallsvariablen $\xi_1, \ldots, \xi_n$ seien unabhängig und normalverteilt mit Streuung σ. Der Erwartungswert von $\xi_1, \ldots, \xi_q$ sei h, der von $\xi_{q+1}, \ldots, \xi_n$ verschwinde. Diese Variablen repräsentieren in geeigneter Anordung einen Signalsprung. Tabelle 2.13 gibt einen Überblick über Erwartungswert und Streuung des Medians (mit entsprechender Maskenlänge $2k + 1 = n$ für verschiedene Sprunghöhen h, p und n. In allen Fällen ist $\sigma = 1$. Abbildung 2.13 veranschaulicht das Verhältnis der Erwartungswerte für einen Moving Average und einen Moving Median Filter mit Maskengröße $n = 3$, Sprunghöhe $h = 5$ und $\sigma = 1$.

Als weiteres Maß, um die Kantenerhaltung zu quantifizieren, bietet sich die erwartete quadratische Abweichung von x über die Kante hinweg an:

$$\mathcal{E} = \sum_{i=-p}^{p+1} \mathbb{E}((\xi_i - x_i)^2).$$

			h				
n	p		1.0	2.0	3.0	4.0	≥ 5.0
3	1	$\mathbb{E}(\mathbb{M})$	0.305	0.486	0.549	0.562	0.564
		$\sigma(\mathbb{M})$	0.697	0.760	0.806	0.822	0.826
9	3	$\mathbb{E}(\mathbb{M})$	0.318	0.540	0.626	0.641	0.642
		$\sigma(\mathbb{M})$	0.424	0.471	0.513	0.527	0.529
5	1	$\mathbb{E}(\mathbb{M})$	0.179	0.270	0.294	0.297	0.297
		$\sigma(\mathbb{M})$	0.551	0.580	0.596	0.600	0.600
5	2	$\mathbb{E}(\mathbb{M})$	0.386	0.676	0.808	0.841	0.846
		$\sigma(\mathbb{M})$	0.560	0.631	0.705	0740	0.748
25	5	$\mathbb{E}(\mathbb{M})$	0.184	0.286	0.312	0.315	0.315
		$\sigma(\mathbb{M})$	0.256	0.271	0.280	0.282	0.282
25	10	$\mathbb{E}(\mathbb{M})$	0.391	0.719	0.900	0.944	0.948
		$\sigma(\mathbb{M})$	0.260	0.295	0.346	0.371	0.375

Tabelle 2.2: Momente des verrauschten Sprunges aus Beispiel 2.2.30

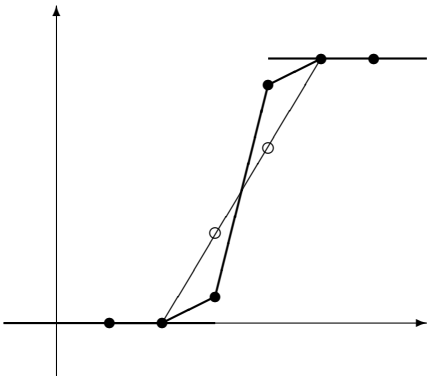


Abbildung 2.13: Erwartungswerte des Moving Average (∘) und Moving Median (•) in Beispiel 2.2.30 für $h = 5$, $n = 3$ und $\sigma = 1$.

Numerische Rechnung zeigt, daß diese Größe im Bereich $h < 2$ $(= 2\sigma)$ für das Mittel (erwartungsgemäß) etwas besser als für den Median liegt. Mit wachsender Sprunghöhe wird dagegen der Median erheblich besser (vgl. Plot in [19], Seite 175).

Weitere Untersuchungen findet man in [19] und [18][7].

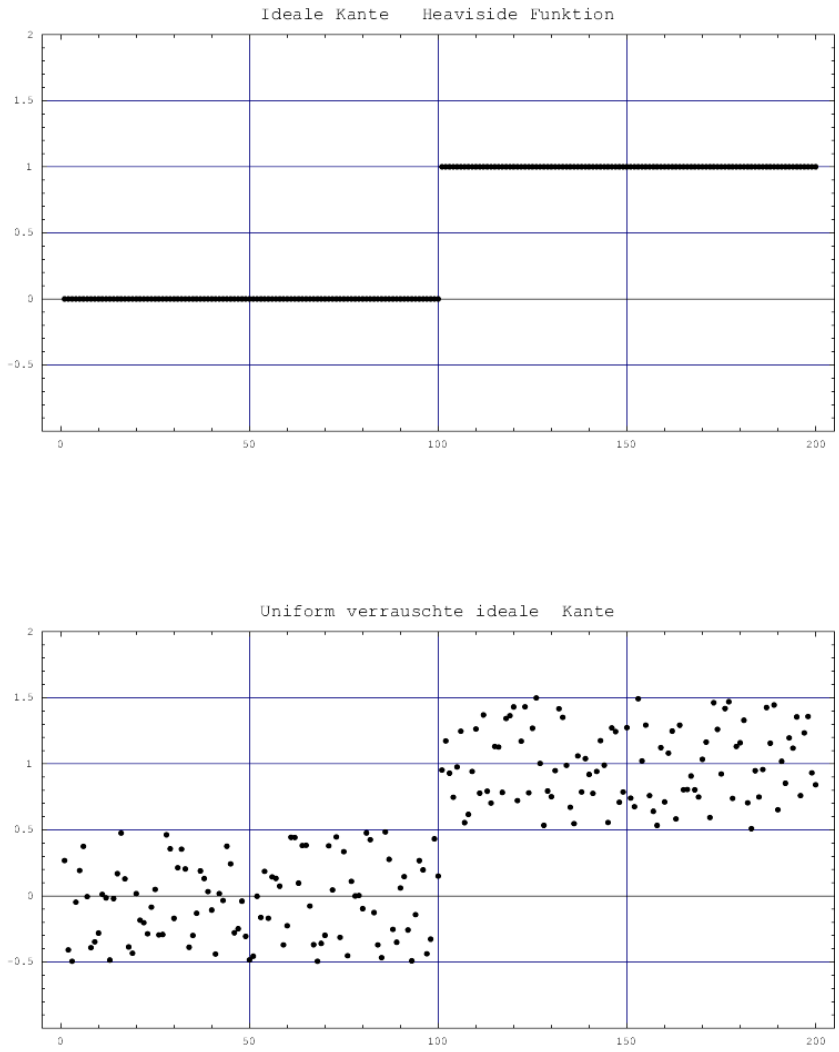


Abbildung 2.14: Heavisidekante der Höhe 1, uniform verrauscht mit Erwartungswert 0 und Standardabweichung 0,2

[7] Letzteres Papier liegt dem Autor nicht vor.

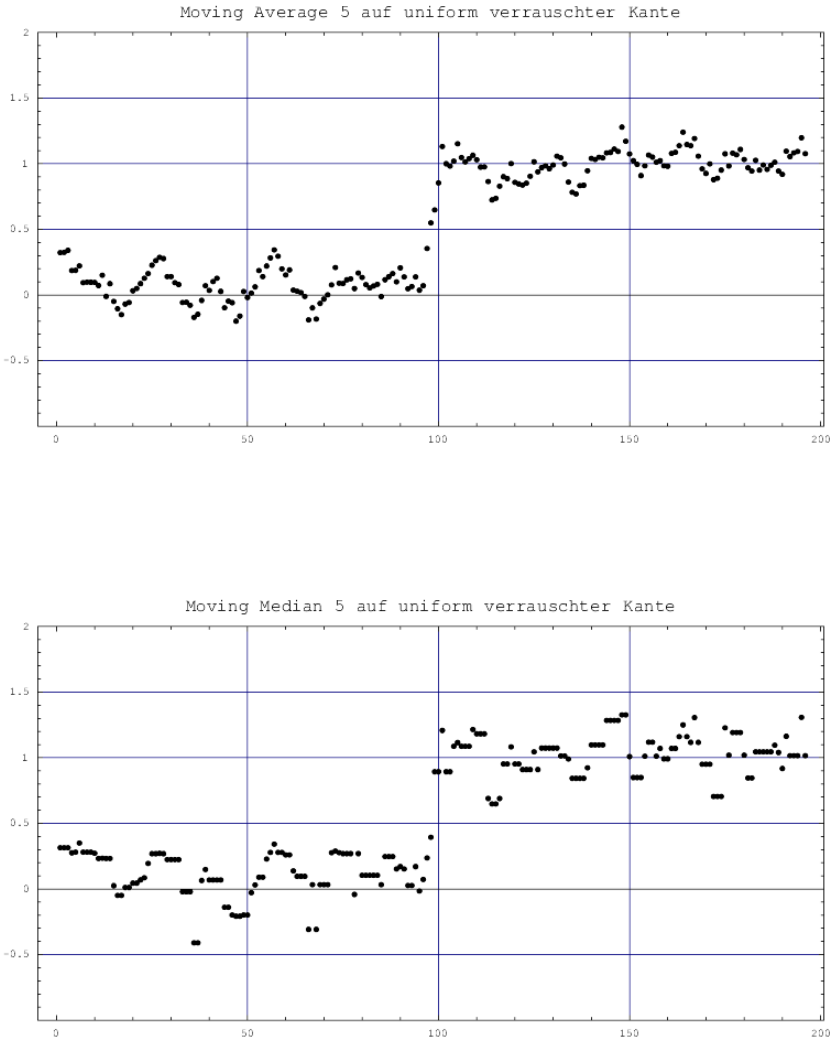


Abbildung 2.15: Signal aus Abb. 2.14 gefiltert mit Moving Average und Median der Länge 5. Standardabweichung des Rauschens 0.2

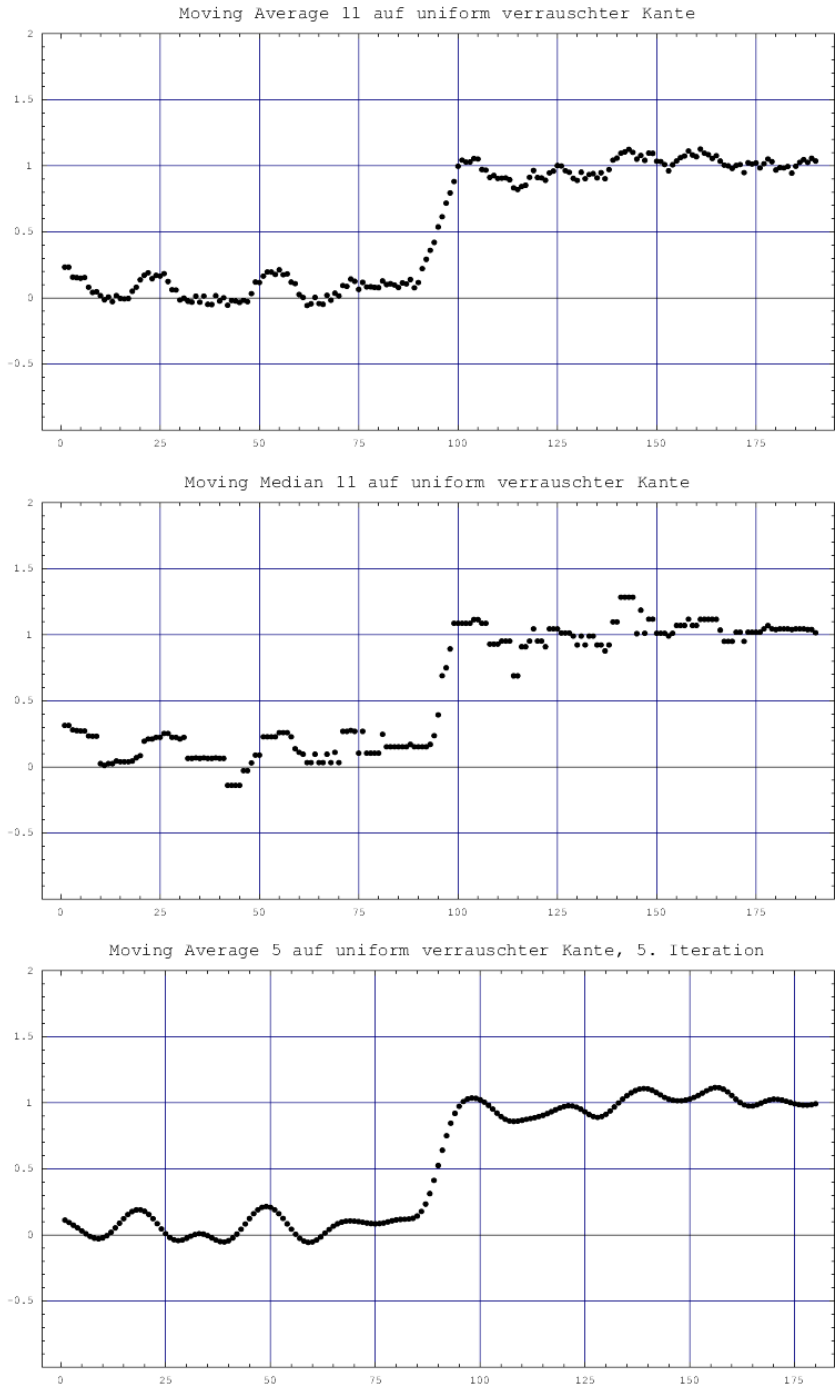


Abbildung 2.16: Uniform verrauschte Kante gefiltert mit: Moving Average bzw. Moving Median der Maskenbreite 11. 5. Iteration des Moving Average der Breite 5

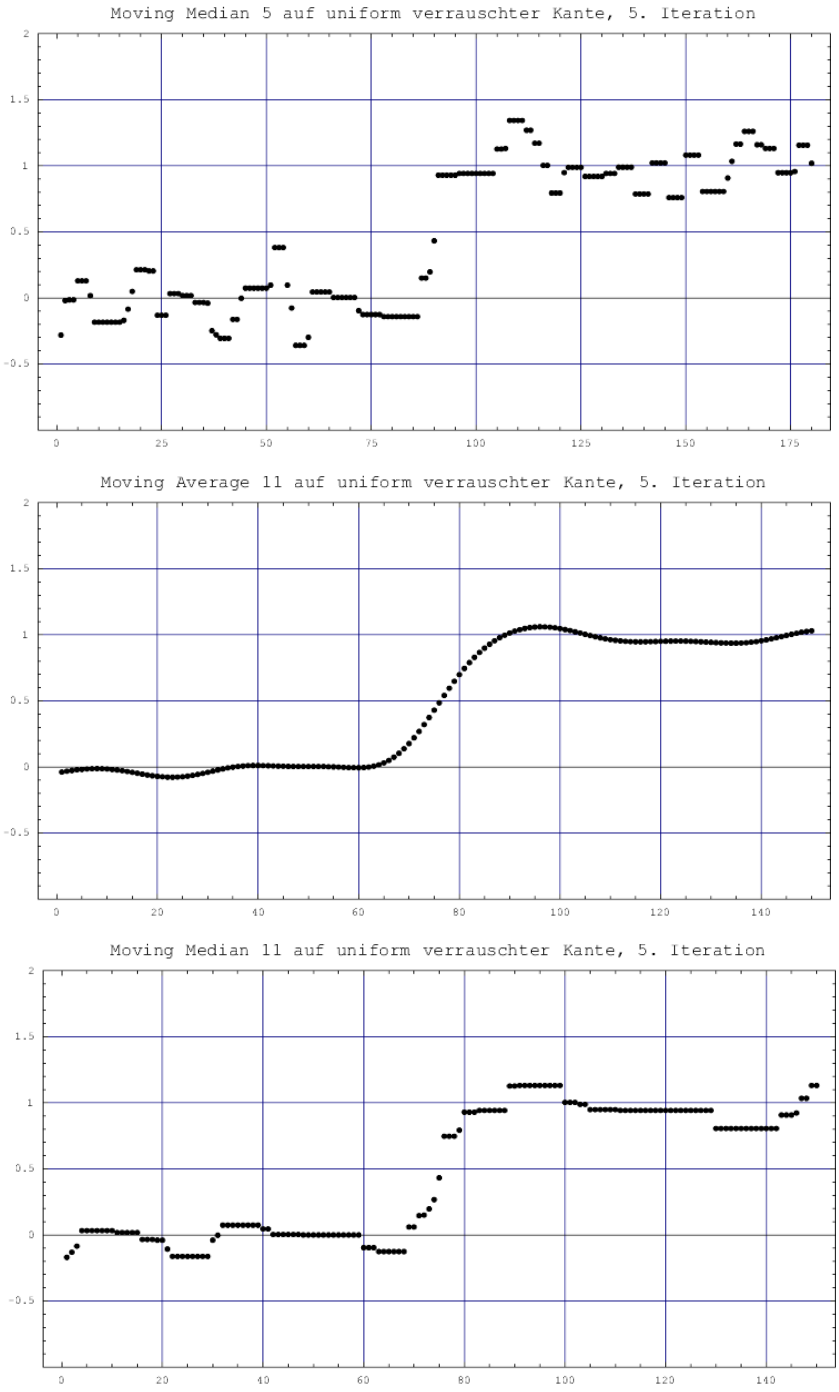


Abbildung 2.17: Uniform verrauschte Kante nach 5 maliger iterierter Anwendung von Moving Average der Breiten 5 und 11, sowie des Moving Median der Breite 11.

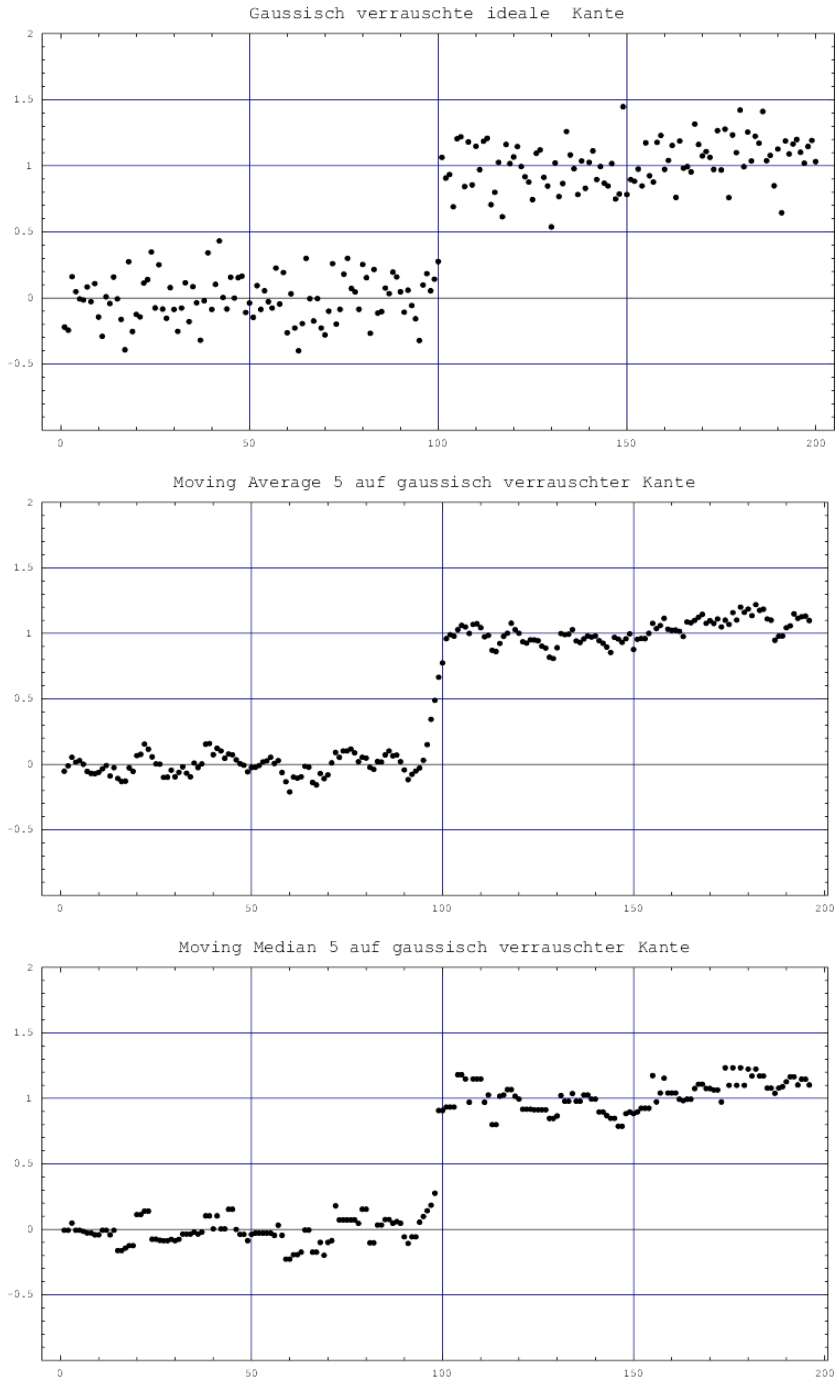


Abbildung 2.18: Gaußisch verrauschte Kante, gefiltert mit Moving Average und Median der Länge 5. Standardabweichung des Rauschens 0.2

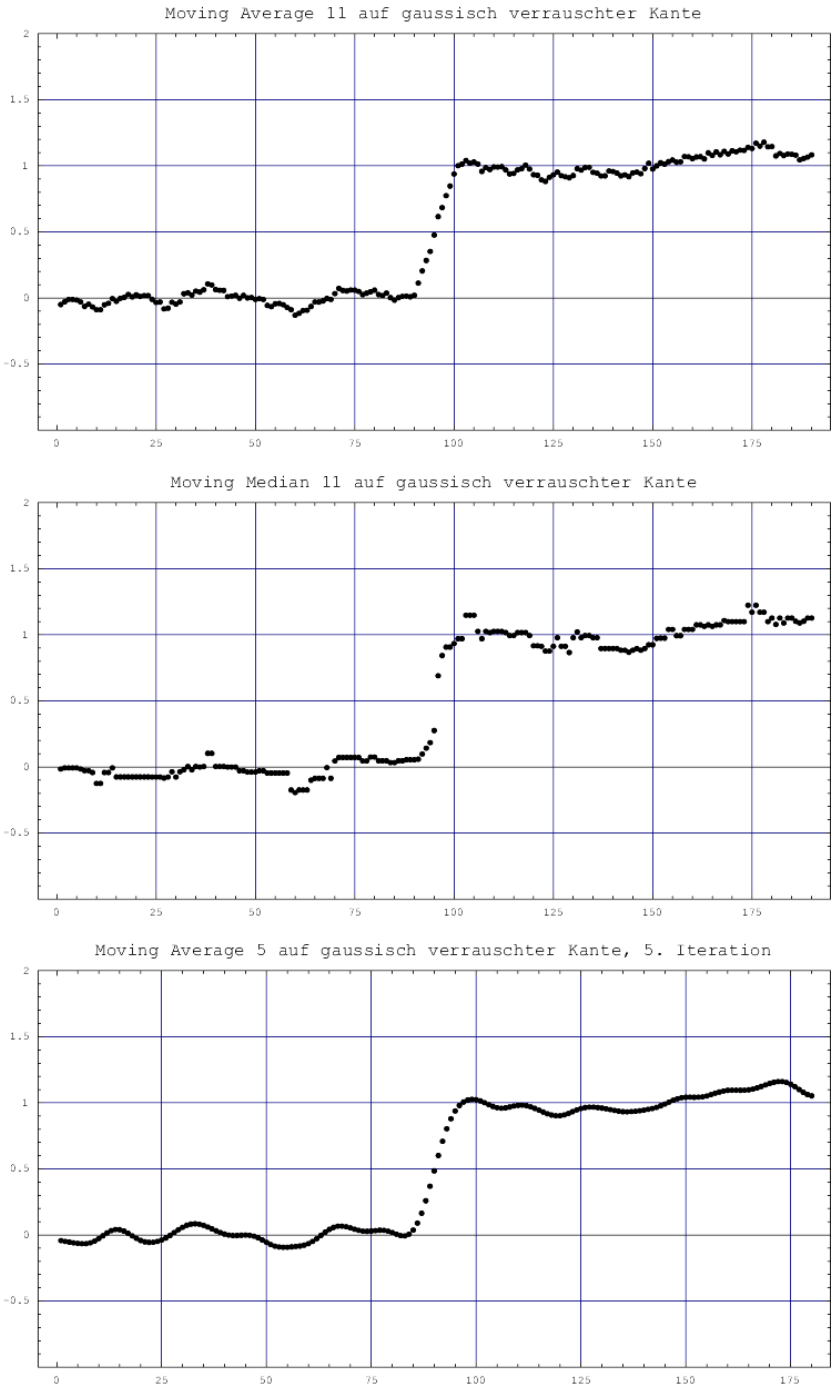


Abbildung 2.19: Gaußisch verrauschte Kante gefiltert mit: Moving Average bzw. Moving Median der Maskenbreite 11. 5. Iteration des Moving Average der Breite 5

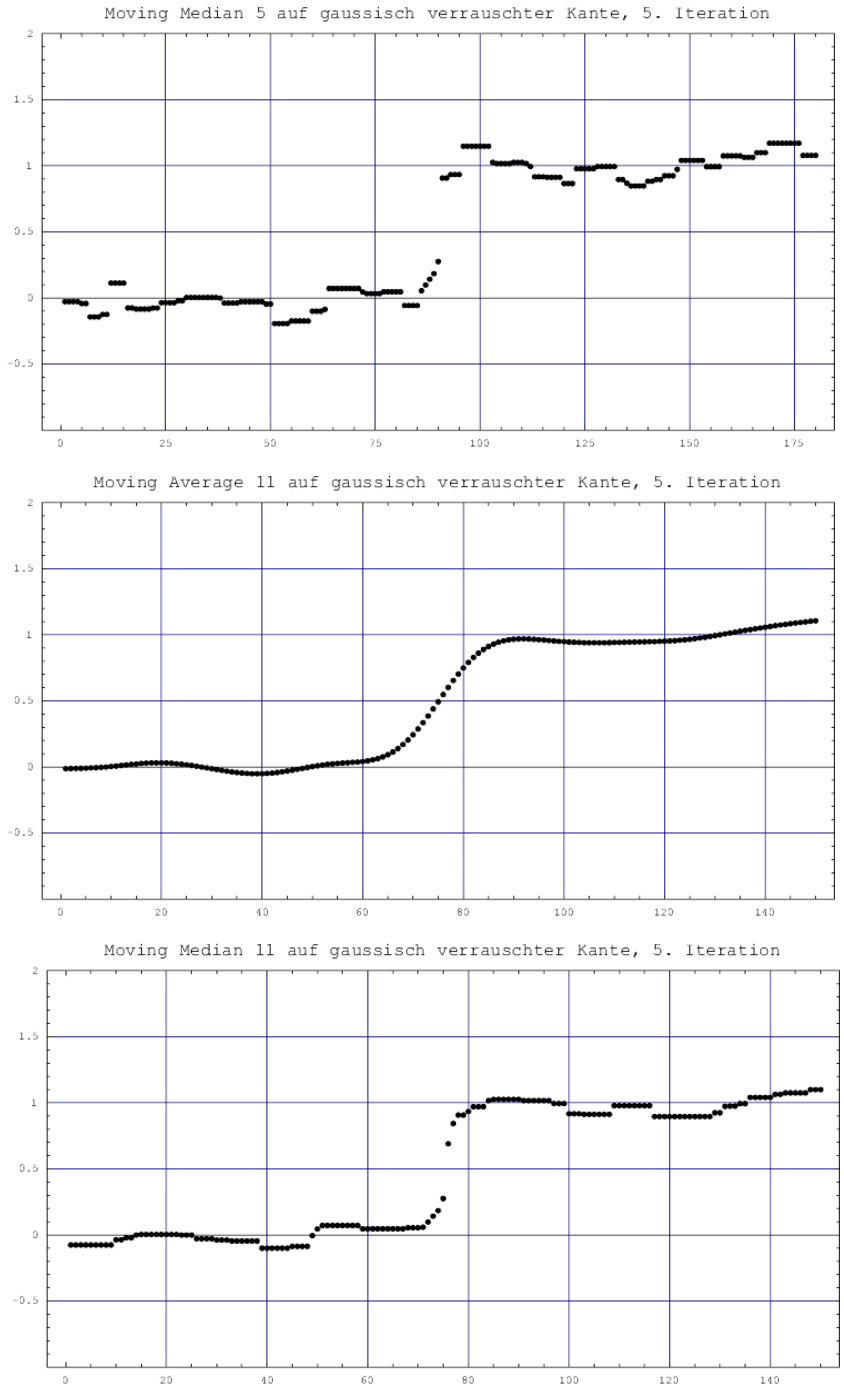


Abbildung 2.20: Gaußisch verrauschte Kante nach 5 maliger iterierter Anwendung von Moving Average der Breiten 5 und 11, sowie des Moving Median der Breite 11.

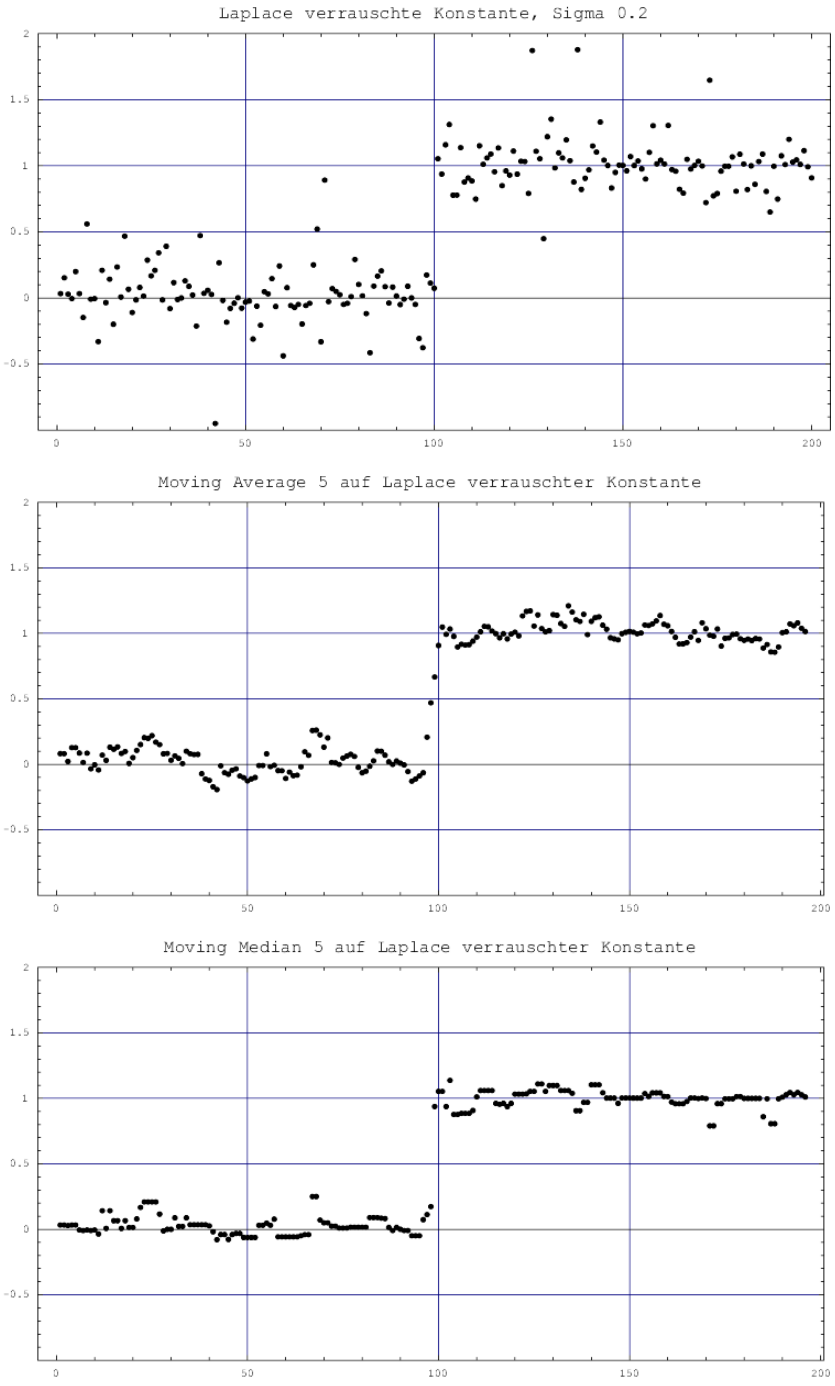


Abbildung 2.21: Laplace verrauschte Kante, gefiltert mit Moving Average und Median der Länge 5. Standardabweichung des Rauschens 0.2

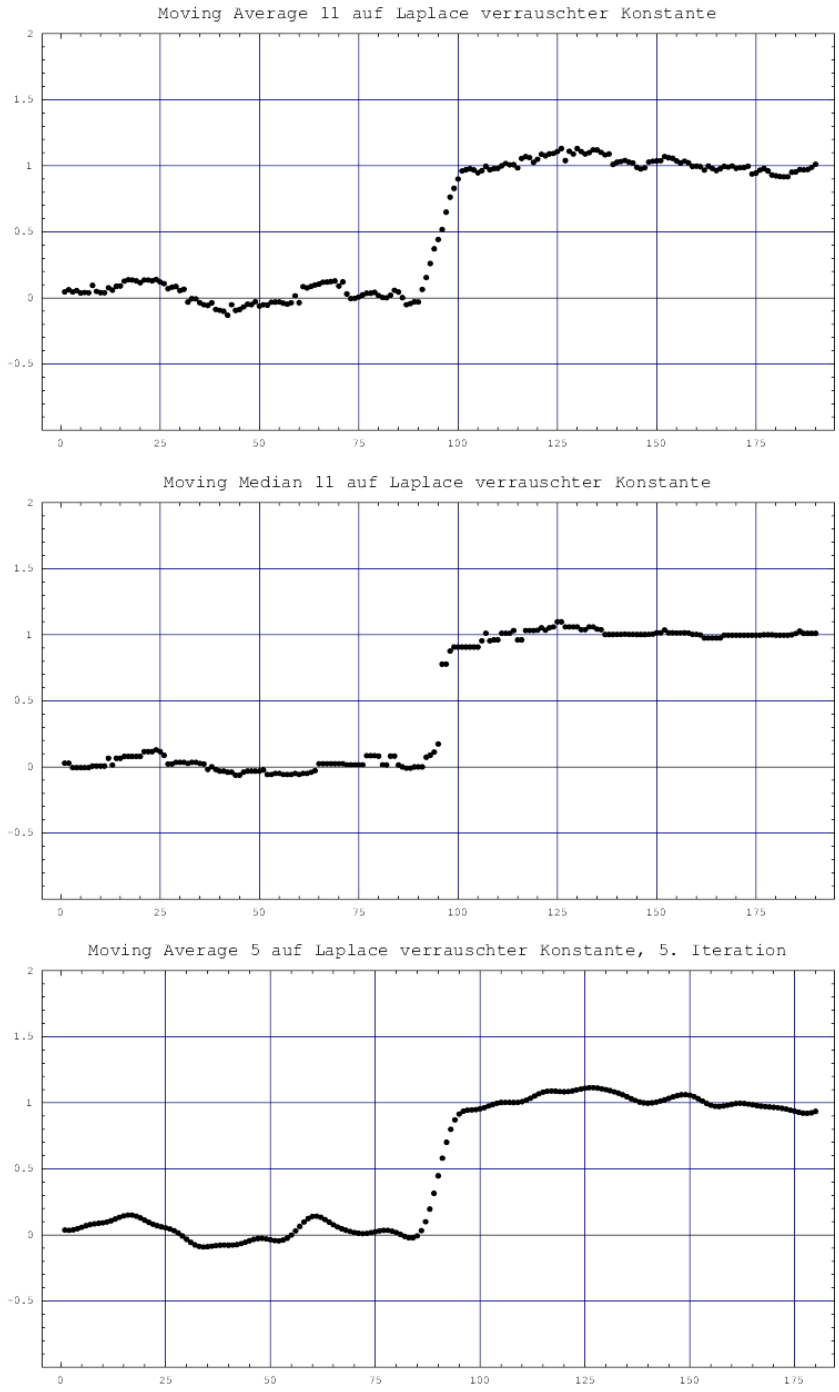


Abbildung 2.22: Laplace verrauschte Kante gefiltert mit: Moving Average bzw. Moving Median der Maskenbreite 11. 5. Iteration des Moving Average der Breite 5

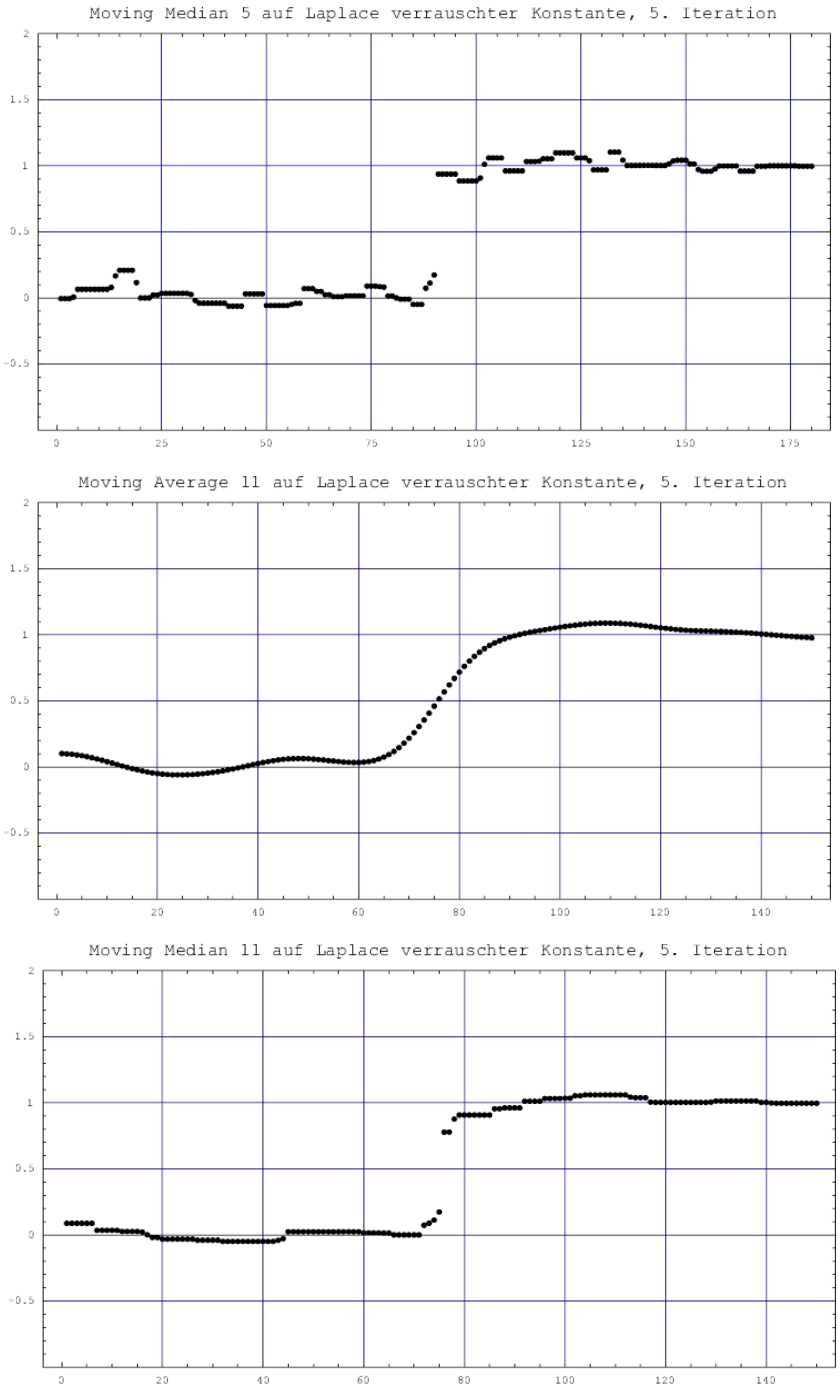


Abbildung 2.23: Laplace verrauschte Kante nach 5 maliger iterierter Anwendung von Moving Average der Breiten 5 und 11, sowie des Moving Median der Breite 11.

Kapitel 3

Bayessche Modelle

Wieder geht es um die Rekonstruktion eines im allgemeinen unbekannten Idealbildes x aus einem Datensatz y. Dabei stellt man sich vor, daß die Daten eine gestörte Version von x darstellen. Die Zuordnung eines x zu y stellt ein *inverses Problem* dar, das praktisch nie eindeutig lösbar ist. Um die Lösungsmenge einzuschränken, stellt man Zusatzbedingungen, die ein gewisses Vorwissen über oder Erwartungen an die Rekonstruktion widerspiegeln. Eine Möglichkeit dies zu tun, ist die Angabe einer Gütefunktion für x. Diese wird so aufgebaut, daß Bildern x mit unerwünschten Eigenschaften eine geringe Güte und solchen mit erwünschten Eigenschaften eine hohe Güte zugeordnet wird. Unter allen Bildern entscheidet man sich dann z.B. für dasjenige als Rekonstruktion, welches einerseits zu den Daten paßt und andererseits die beste Güte hat.

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

Umfassende Darstellungen der bayesschen Bildanalyse findet man in den Monographien von B. Chalmond (2000), [8], X. Guyon (1995), [14], und G. Winkler (1995), [32] ([33] für die Volksrepublik China).

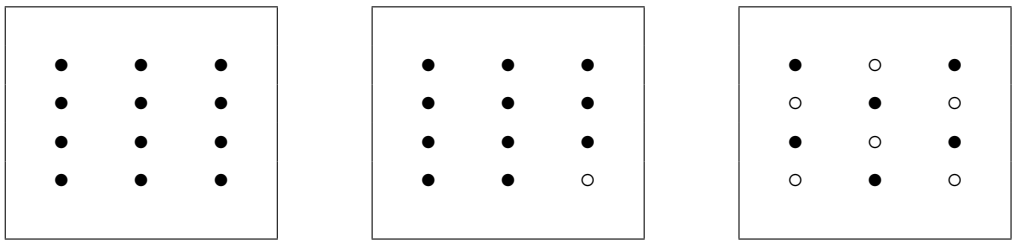
3.1 Einleitung: Bewertung von Bildern

Im nächsten Unterabschnitt verschaffen wir uns anhand von Binärbildern eine grobe Vorstellung davon, was die Bewertung von Bildern bedeuten könnte. Im zweiten Teil diskutieren wir einen speziellen Ansatz zur kantenerhalten-

den Glättung etwas genauer.

3.1.1 Qualität und Datentreue

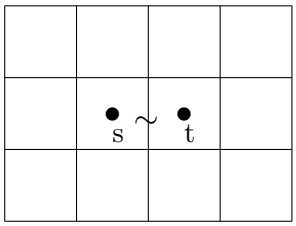
Um eine Vorstellung von der Wahl einer Gütefunktion zu bekommen, betrachten wir ein einfaches Beispiel. Gegeben seien die 'Bilder' in Abbildung 3.1 Wir wollen die Muster nach dem Kriterium 'Glattheit' bewerten: Glat-



erstklassig ordentlich katastrophal

Abbildung 3.1: Glatte und rauhe Muster

te Muster sollen gute (=niedrige) Noten und rauhe Muster schlechte Noten bekommen. Verbal könnten wir von links nach rechts die Noten 'erstklassig', 'ordentlich' und 'katastrophal' vergeben. Wir suchen aber ein quantitatives Gütemaß. Denkbar wäre z.B. die Anzahl von benachbarten Pixeln (in horizontaler und vertikaler Richtung) mit verschiedenen Farben. Dies gäbe die Noten 0, 2 und 17 statt erstklassig, ordentlich und katastrophal. Um dies



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

Abbildung 3.2: Pixelgitter, Pixel und die elementare 'nächste Nachbarschaftsbeziehung', d.h. nächste Nachbarn eines Pixels sind die geometrisch nächsten Pixel in östlicher, nördlicher, westlicher und südlicher Richtung

formal hinschreiben zu können, bezeichnen wir die Pixel mit $s, t, \ldots$ und

schreiben $s \sim t$, wenn s und t benachbart sind (dies ist in Abb. 3.2 veranschaulicht).
Die Farbe in Pixel s sei x_s, $x = (x_s)$ sei das ganze Muster. Dann entsteht die obige Bewertung als Auswertung der Funktion

$$G(x) = \sum_{s \sim t} \chi_{\{\{a,b\}:a \neq b\}}(\{x_s, x_t\}).$$

Aus historischen Gründen und aus Gründen der Konvention wählen wir eine äquivalente Beschreibung. Wir setzen $x_s = \pm 1$, und

$$G(x) = -\sum_{s \sim t} x_s x_t. \tag{3.1}$$

Die Terme $-x_s x_t$ nehmen dann für gleiche Paare den Wert -1 und für ungleiche Paare den Wert 1 an. Die Muster in Abb. 3.1 bekommen dann die Bewertungen -17, -13 und +17. Offensichtlich gehen die beiden Modell durch die Beziehung $-x_s x_t = 2\chi_{\{a,b\}:a \neq b\}}(\{x_s, x_t\}) - 1$ auseinander hervor. Das Modell in (3.1) kommt aus der statistischen Physik und ist unter dem Namen *Ising Modell* bekannt. Es wird weiter unten ausführlicher diskutiert werden.

Der natürliche Rahmen, um Datenquellen, Vorinformation und Regularitätsbedingungen zu verknüpfen ist das *Bayessche Paradigma*[1], auf das wir in bälde zurückkommen werden. In seinem Rahmen können sinnvolle Schätzer für x formuliert und begründet werden. Im Anschluß daran muß die Berechnung dieser Schätzer diskutiert werden.

Um hiervon eine Idee davon zu geben, führen wir das Beispiel fort. Sei unser Bewertungsmaßstab für die Qualität eines Bildes durch das grobe Kriterium (3.1) gegeben. Wir nehmen nun an, daß wir ein Bild $y = (y_s)$ beobachten. Dieses sei aus irgendeinem - uns nicht bekannten - $\tilde{x} = (\tilde{x}_s)$ hervorgegangen. Bei diesem Vorgang wurden die Werte $\tilde{x}_s$ möglicherweise verfälscht, indem z.B. ein (zufälliges) Rauschen $\eta = (\eta_s)$ addiert wurde:

$$y_s = x_s + \eta_s.$$

Wir suchen nun eine Approximation $x^* = (x_s^*)$ von $\tilde{x}$. Wenn wir nichts über $\tilde{x}$ wissen, ist dieses Unterfangen hoffnungslos. Nehmen wir andererseits an, daß wir wenigstens wissen, daß das Bild 'glatt' und das Rauschen moderat ist. Dann sollte einerseits die Approximation $\tilde{x}$ unter der Bewertung G in (3.1)

[1] *Paradigma*: Beispielhafte Vorgehensweise

gut abschneiden und sich andererseits nicht all zu sehr von der Beobachtung y unterscheiden. Die Abweichung eines Musters x von y können wir durch irgend ein Abstandsmaß messen, z.B. durch die quadratische Abweichung

$$D_2(x,y) = \sum_s (x_s - y_s)^2 \Big(=: \sum_s d_2(x_s, y_s) \Big)$$

oder die absolute Abweichung

$$D_1(x,y) = \sum_s |x_s - y_s| \Big(=: \sum_s d_1(x_s, y_s) \Big).$$

Es ergeben sich

$$d_2(a,b) = \begin{cases} 0 & \text{falls } a = b \\ 4 & \text{falls } a \neq b \end{cases}, \qquad d_1(a,b) = \begin{cases} 0 & \text{falls } a = b \\ 2 & \text{falls } a \neq b \end{cases},$$

$$-ab + 1 = \begin{cases} 0 & \text{falls } a = b \\ 2 & \text{falls } a \neq b \end{cases}, \quad a, b = \pm 1.$$

Jedenfalls halten wir fest, daß

$$D_i(x,a) = -c \cdot \sum_s x_s y_s + d, \;\; c > 0. \tag{3.2}$$

ist. Die konstante Verschiebung der Bewertungsskala um d spielt inhaltlich keine Rolle, die Konstante c steuert die Stärke der Datentreue. Also können wir uns im gegenwärtigen Falle von Binärbildern und additivem pixelweisem Rauschen auf einen Datenterm der Gestalt (3.2) beschränken.

Die Bildqualität bewerten wir also mit Hilfe von (3.1). Wir kombinieren sie nun mit der Datentreue in der Funktion

$$H(x) = G(x) + D(x,y),$$

und erhalten

$$H(x) = -\left(\sum_{s \sim t} x_s x_t + c \cdot \sum_s x_s y_s \right). \tag{3.3}$$

Auf diese Weise haben wir nun die Forderungen nach Regularität und Datentreue in einer Zielfunktion kombiniert; in diesem Sinne wäre ein x^* optimal, welches H minimiert. Über die Konstante c wird das Gewicht von Datentreue gegenüber Regularität gesteuert. Im nächsten Abschnitt verallgemeinern wir dieses Modell auf mehr als zwei Intensitätsstufen.

Beispiel 3.1.1 Wir illustrieren die Methode am Beispiel der Restauration einer verrauschten Strichzeichnung (cf. Abb. 3.3). Die Restauration ist eine Minimalstelle der Funktion

$$H(x) = \Big(\underbrace{\underbrace{-\beta \sum_{s,t} x_s x_t}_{\text{Ising Modell}} + \underbrace{h \sum_{s} x_s}_{\text{äußeres Feld}}}_{\text{Isingmodell mit äußerem Feld}} \Big) + \underbrace{\underbrace{\frac{1}{2} \ln\left(\frac{p}{1-p}\right)}_{c \text{ in (3.2)}} \underbrace{\sum_{s} x_s y_s}_{\text{Summe in (3.2)}}}_{\text{Datenterm}} \quad , \quad (3.4)$$

wobei β die Stärke der Glättung steuert, p das Rauschen charakterisiert (mehr dazu später) und y eine gegebene Beobachtung bezeichnet. Die Summe im zusätzlichen Term zählt im wesentlichen die Pixel einer ausgezeichneten Farbe - etwa die schwarzen. Mit dem Parameter h läßt sich die Tendenz zu dieser Farbe steuern. Der Datenterm ist von der oben abgeleiteten Gestalt, wobei sich der multiplikative Faktor $\frac{1}{2} \ln(p/(1-p))$ über das bayessche Paradigma aus der Gestalt des Rauschens ergibt (vgl (3.2.7). Wir gehen in Abschnitt 3.2 näher darauf ein.

Modelle vom Typ in (3.3) treten unter verschiedenen Namen und in verschiedenem Zusammenhang auf. In der Numerik wird der erste Term manchmal als *Regularisierung* bezeichnet (meist für stetige Funktionen) und in der Statistik erscheinen sie als häufig als *penalized likelihood*. Wir werden sie in den Rahmen der bayesschen Statistik einbetten, wo sie als Exponenten von a posteriori Verteilungen bei Exponentialverteilungen auftreten.

3.1.2 Ein Beispiel: kantenerhaltende Glättung

Wir sehen uns ein Beispiel an, bei dem zunächst (scheinbar) die Stochastik immer noch keine Rolle spielt. Abbildung 3.2 veranschaulicht eine elementare Situation und faßt einige Bezeichnungen zusammen.

Beispiel 3.1.2 Wieder sei die Güte eines Bildes sei seine Glattheit. Eine Möglichkeit diese bei Binärbildern zu quantifizieren ist die Gütefunktion (3.1). Für mehrere diskrete oder für reelle Werte bietet sich die Funktion

$$G(x) = \beta \sum_{s \sim t} (x_s - x_t)^2, \;\; \beta > 0$$

an. Sie ist offensichtlich nichtnegativ und verschwindet genau für konstante Muster x. Jedes Paar $s \sim t$ mit hoher Grauwertdifferenz verschlechtert die

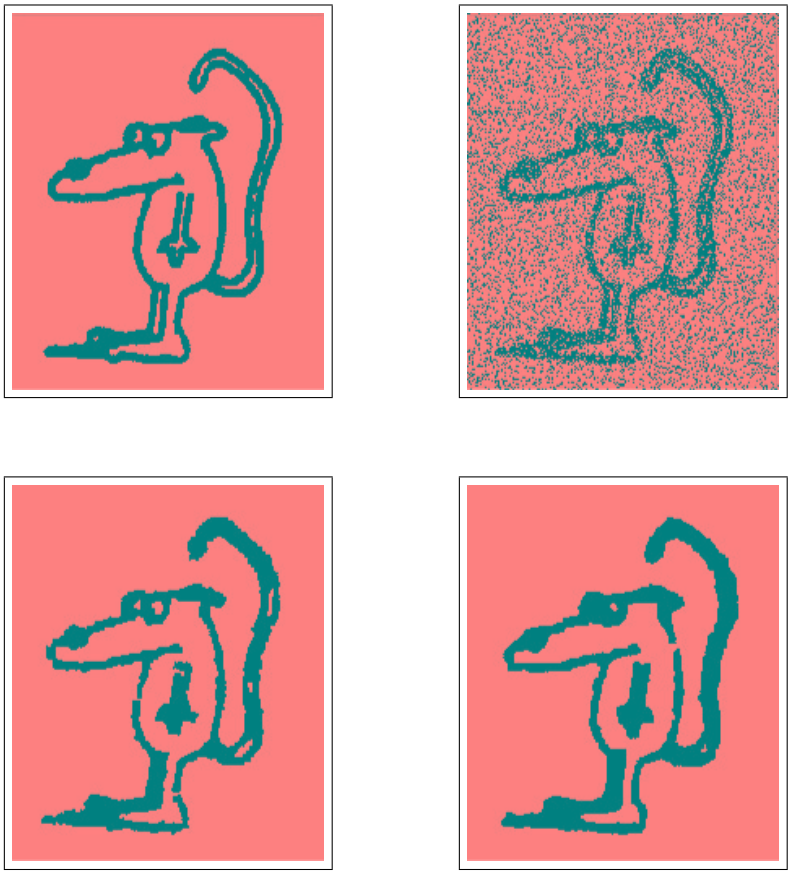


Abbildung 3.3: Comic-Figur Ratbert. L.o.: Original, r.o.: verrauscht, l.u.: Minimalstelle von (3.4) für die Parameter $\beta = 0,99$, $h = -0,087$ und $p = 0,2$, y ist das Bild rechts oben, r.u.: Minimalstelle von (3.4) für $\beta = 1$ und $h = 0$. Bild: F. Friedrich, Heidelberg

Bilanz beträchtlich. Wir beobachten, daß G der Funktion $a \mapsto \sum_s (x_s - a)^2$ entspricht, die durch $\bar{x}$ minimiert wird. G wird nun ergänzt durch einen Term, welcher die Datentreue mißt, etwa

$$D(x, y) = \sum_s (y_s - x_s)^2$$

was zusammen die Funktion

$$H(x|y) = G(x) + D(x, y) = \beta \sum_{s \sim t} (x_s - x_t)^2 + \sum_s (y_s - x_s)^2$$

ergibt. Für gute x müssen sich nun Glattheit und Datentreue die Waage halten. Die Bedeutung beider Kriterien kann über den Parameter β gesteuert werden. Das Resultat $\operatorname{argmin}_x H(x|y)$ wird jedenfalls eine glatte Approximation der Daten sein, gegebenenfalls mit verschmierten Kanten. Rückblickend stellen wir fest, daß wir eigentlich nicht 'Glattheit', sondern 'Konstantheit' als Gütekriterium gewählt haben.

A. Blake und A. Zisserman (1984), [5], schlagen einen physikalisch motivierten Ausweg vor. Das Bild x wird als elastische Membran mit Affinität zu den Daten y_s gedacht. In Ruhestellung ist diese möglichst kurz, also konstant. Der von einer Streckung verursachte Energiezuwachs wird durch das Quadrat der ersten Ableitung ausgedrückt. Die diskrete Richtungsableitung zwischen s und einem benachbarten t ist $(x_s - x_t)/h$ ist der Streckungsfaktor der Membran. Im vorliegenden Fall ist $h = 1$ und somit ist die Membran zwischen s und t um $|x_t - x_s|$ länger als im Ruhezustand. Der entsprechende Term ist also $(x_s - x_t)^2$ (wie gehabt). Die Membran darf nun brechen, wenn die Daten dies erzwingen. Jeder Bruch führt zu Energiekosten α. Andererseits eliminiert er die Streckungsenergie zwischen den Pixeln. Dazu führt man Schaltervariablen b_{st} ein, welche die Glättung aus- und das Brechen einschalten: Für benachbarte Pixel s, t gemäß der Skizze definieren wir die *Mikrokante* $s \sim t$ als das Element des dualen Gitters zwischen s und t. Für die Mikrokanten definieren wir die Schaltervariablen

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

Die resultierende Gütefunktion ist

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

wobei die Konstanten aus der Physik kommen. Wir diskutieren: Ist $d = \lambda^2(x_s - x_t)^2 > \alpha$, so lohnt es sich, die Strafe α für b_{st} zu bezahlen, d.h. der Glättungsterm wird ausgeschaltet. Ob dieser Fall sich lohnt, hängt wiederum vom Datenterm $(y_s - x_s)^2$ ab. Wird umgekehrt von diesem Term ein kleines d favorisiert, so lohnt sich der Bruch nicht und d wirkt wie im letzten Beispiel. Wir werden an dieser Stelle noch nicht auf die Minimierung von H eingehen. Der erste Schritt dorthin ist jedoch von eigenständigem Interesse.

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

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

die Funktion φ ist dabei gegeben durch

$$\varphi(\Delta) = \min\{\lambda^2\Delta^2, \alpha\}. \tag{3.5}$$

Um x^* zu berechnen genügt es also, eine Minimierung nur in der Komponente

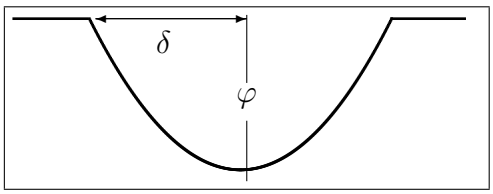


Abbildung 3.4: Die 'Topffunktion'

x durchzuführen. Die zweite Komponente b^* der Minimalstelle (x^*, b^*) ist dann eindeutig rekonstruierbar:

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

Bemerkung 3.1.3 Im Zusammenhang mit Schätzern hatten wir Maße für die Abweichungen der Schätzer von den Daten betrachtet. Sie waren gegeben durch Funktionen des Typs

$$a \longmapsto \sum_i \psi(x_i - a),$$

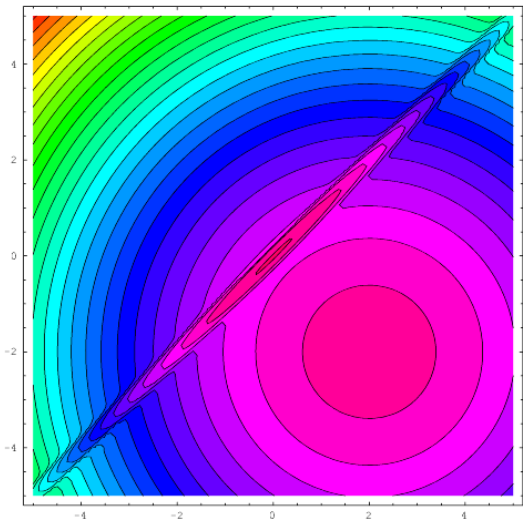


Abbildung 3.5: Höhenlinien von $(x_s) \longmapsto \sum_{s\sim t} \varphi(x_s - x_t) + \sum_s (y_s - x_s)^2$ für $S = \{1, 2\}$ und φ aus (3.5) in den Koordinaten x_1 und x_2. Der Datenvektor (y_1, y_2) bildet den Mittelpunkt der konzentrischen Kreise, d.h. die Basis der nach oben geöffneten Parabel. Sie wird überlagert durch den von φ verursachten 'Graben' entlang der Diagonale.

sogenannten *Verlustfunktionen*. Schätzer wurden als Minimalstellen von Verlustfunktion charakterisiert. Für $\psi(u) = u^2$ wurde die Verlustfunktion durch das empirische Mittel $\bar{x}$ minimiert, für $\psi(u) = |u|$ durch den empirischen Median (vgl. Bemerkungen 2.2.1 und 2.2.12). Der empirische Median erwies sich als robuster als das empirische Mittel, was dadurch erklärt wurde, daß die Betragsfunktion weniger steil ansteigt als das Quadrat. Nun sind wir noch radikaler und setzen obige Topffunktion φ für ψ ein; φ ist für große $|u|$ sogar konstant. Davon erwarten wir uns noch stärkere Robustheit mit den entsprechenden Konsequenzen für die kantenerhaltende Glättung. Die entsprechende Verlustfunktion wurde von F.R. Hampel, [15], vorgeschlagen.

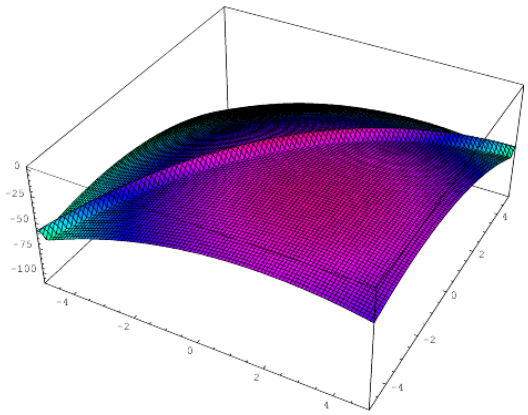


Abbildung 3.6: Der Graph der Funktion $(x_s) \mapsto \sum_{s\sim t} \varphi(x_s - x_t) + \sum_s (y_s - x_s)^2$ aus Abb. 3.5 von unten gesehen.

3.2 Das Bayessche Paradigma

Im obigen Beispiel wurden Regularitätsforderungen durch eine Gütefunktion H ausgedrückt und durch einen Term ergänzt, der Datentreue fordert. Dabei wurde die spezielle Natur des Rauschens nicht diskutiert.

Wir beschreiben nun, wie Regularitätsforderungen und das Wissen um die statistischen Eigenschaften des Rauschens im Rahmen der bayesschen Statistik verknüpft werden können.

3.2.1 Einführung

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

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

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

eine gemeinsame Verteilung auf $\boldsymbol{X} \times \boldsymbol{Y}$ definiert. Offensichtlich sind $\Pi(y|x) =$

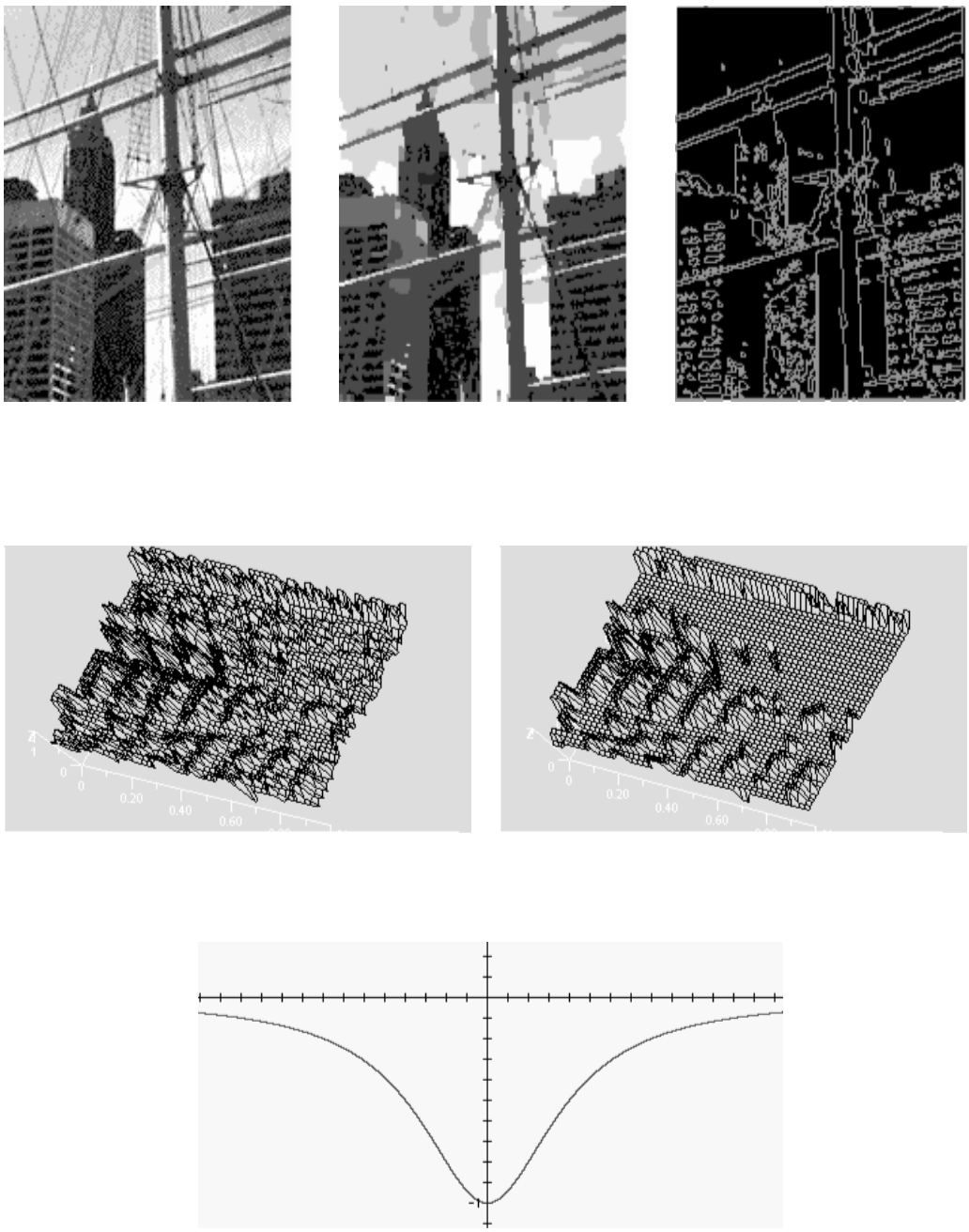


Abbildung 3.7: Minimierung von $(x_s) \longmapsto \sum_{s\sim t} \varphi(x_s - x_t) + \sum_s (y_s - x_s)^2$. Obere Reihe: links: die Daten y; mitte: eine Minimalstelle x^*; rechts: die zugehörigen Kanten b^*. Untere Reihe: Surfaceplots, links: der rechten unteren Ecke von y; rechts: der rechten unteren Ecke von x^*; außerdem: die in diesem Beispiel verwendete Funktion $\varphi(u) = -1/(1 + (u/\delta)^2)$. Bild: F.Friedrich, Heidelberg

$P(x, y)$ und

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

Interpretieren wir x als Parameter (in der Statistik meist mit ϑ bezeichnet) und y als Beobachtung, so erlaubt diese Formel Rückschlüsse von der Beobachtung auf den zugrundeliegenden Parameter. $\Pi|\mathbf{X}$ modelliert ein Vorwissen über die Parameter oder (weiche) Regularitätsforderungen; nach Beobachtung von y wird Π zu $\Pi(\cdot|y)$ modifiziert. Deshalb heißt Π die *a priori* und $\Pi(\cdot|y)$ die *a posteriori Verteilung*.

Diese Art der Verknüpfung von Vorwissen und Daten ist weit flexibler als die im oben geschilderten Zugang. Zum ersten können beliebige Formen des Rauschens modelliert werden und zum zweiten steht die gesamte Schätztheorie (angewandt auf die a posteriori Verteilung) zur Verfügung.

Für Verteilungen mit Dichten ist der Zugang analog.

Beispiel 3.2.1 Sei die a priori Verteilung gegeben durch

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

Das Rauschen sei additiv, weiß, gaußisch, also

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

Dann gilt

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

Verteilungen der Art $\Pi(x) \propto \exp(-H(x))$ nennt man *Gibbssche Verteilungen*. Die Funktion H heißt dann *Energiefunktion*. Die Normierungskonstante $Z = \sum_z \exp(-H(z))$ heißt *Zustandssumme*. Da Gibbsverteilungen strikt positiv sind, sind sie nicht völlig allgemein. Andererseits läßt sich *jede* strikt positive Verteilung mit $H(x) = -\ln \Pi(x) - \ln(Z)$, wobei Z eine positive Zahl ist, als Gibbsmaß schreiben. Es ist nämlich $\exp(-H) = \Pi \cdot Z$; Z ist dann notwendig die Zustandssumme.

Analog gilt für die a posteriori Verteilung

Proposition 3.2.2 *Ist $P(x, dy)$ eine beliebige strikt positive (Zähl-) Dichte $f_x(y)\,dy$ und $\Pi(x) \propto \exp(-H(x))$ ein Gibbsmaß, so gilt*

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

Beweis Man multipliziere aus. □

3.2.2 A priori Verteilungen: Beispiele

Wir ergänzen die obigen Beispiele durch einige klassische Modelle. Alle sind von der Gibbschen Gestalt

$$\Pi(x) \propto \exp(-H(x)), \Pi(dx) \propto \exp(-H(x))dx,$$

je nachdem, ob es sich um diskrete Verteilungen oder Verteilungen mit Dichten handelt. Somit genügt es, die jeweiligen Energiefunktionen H anzugeben.

Wir beginnen mit Binärbildern mit zweiwertigen x_s. Meist gilt also $X = \{-1, 1\}^s$ oder $X = \{0, 1\}^s$.

Beispiel 3.2.3 (Das Ising Modell) Das *Ising Modell* Energiefunktion

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

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

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

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

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

die Zahl der ungleichen Nachbarn, also die Konturlänge. Konfigurationen mit niedrigem $H(x)$ haben also kurze Konturen, ähnlich wie im Modell von Blake & Zisserman. Wir hatten erwähnt, daß dies eine gewisse Regularität der Konturen erzwingt. Der Parameter β steuert den Grad der Regularität. An diesem Beispiel sieht man ein typisches Phänomen dieser Modelle: Neben den Minima gibt es eine Vielzahl lokaler Minima, die sich im Wert nur wenig von den Minima unterscheiden - in diesem diskreten Fall nennen wir x ein

lokales Minimum, wenn die Änderung in einem beliebigen Pixel - d.h. einer Koordinate von $X = \{-1, 1\}^s$ - den Wert von H nicht verschlechtert. So haben in einem $n \times n$ Gitter S alle Bilder x, welche in zwei gleichfarbige Teile, getrennt durch eine senkrechte oder waagerechte Gerade, denselben Wert

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

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

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

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

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

Das *Ising Modell mit äußerem Feld* ist gegeben durch

$$H(x) = -\beta \sum_{s \sim t} x_s x_t + \gamma \sum_s x_s \, .$$

Hier werden die Werte zusätzlich - je nach Vorzeichen von γ - in eine Richtung getrieben.
Das Modell stammt aus der Physik, wo es die Gestalt

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

hat. J ist eine Materialkonstante, T die absolute Temperatur, k die Boltzmannkonstante, m hängt wieder vom Material ab und B quantifiziert ein äußeres magnetische Feld. $J > 0$ treibt die x_s in dieselbe, $J < 0$ in entgegengesetzte Richtung. Interpretiert man die x_s als Richtung von Spins - oder Elementarmagneten - in einem Kristallgitter S, so entspricht $J > 0$ einem Ferromagneten und $J < 0$ einem Antiferromagneten.

Das Modell ist nach (dem Deutschen) E. Ising benannt, der es in seiner Doktorarbeit 1925 bei W. Lenz untersucht hat (und teilweise falsche Schlüsse von Dimension 1 auf höhere Dimenion zog; dadurch wurde die statistische Physik um Jahre zurückgeworfen).

Beispiel 3.2.4 (Das Potts-Modell) Bei endlichem Grauwertevorrat G ist eine weitere natürliche Verallgemeinerung des homogenen Isingmodells das *Potts-Modell*

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

wieder mit

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

Für spätere Zwecke bemerken wir: Ersetzen wir dieses φ durch

$$\psi = 2 \cdot \varphi + 1, \text{ d.h. } \psi(0) = -1, \ \psi(u) = 1 \text{ sonst},$$

so entsteht für Parameter $\beta/2$ ein äquivalentes Modell (d.h. die Energie ändert sich nur um eine Konstante und somit wird das selbe Gibbsmaß induziert).

Beispiel 3.2.5 (Allgemeine Binärmodelle) 'Glattheit' kann man jetzt verallgemeinern: Im Ising-Modell bedeutet es ja die Nähe zu den zwei konstanten Mustern. Diese können wir durch beliebige Muster (und ihre Inverse) ersetzen, indem wir die Vorzeichen der Terme $x_s x_t$ geeignet wählen:

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

Noch allgemeiner kann man die Stärke der Ähnlichkeit lokal steuern durch

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

oder noch allgemeiner

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

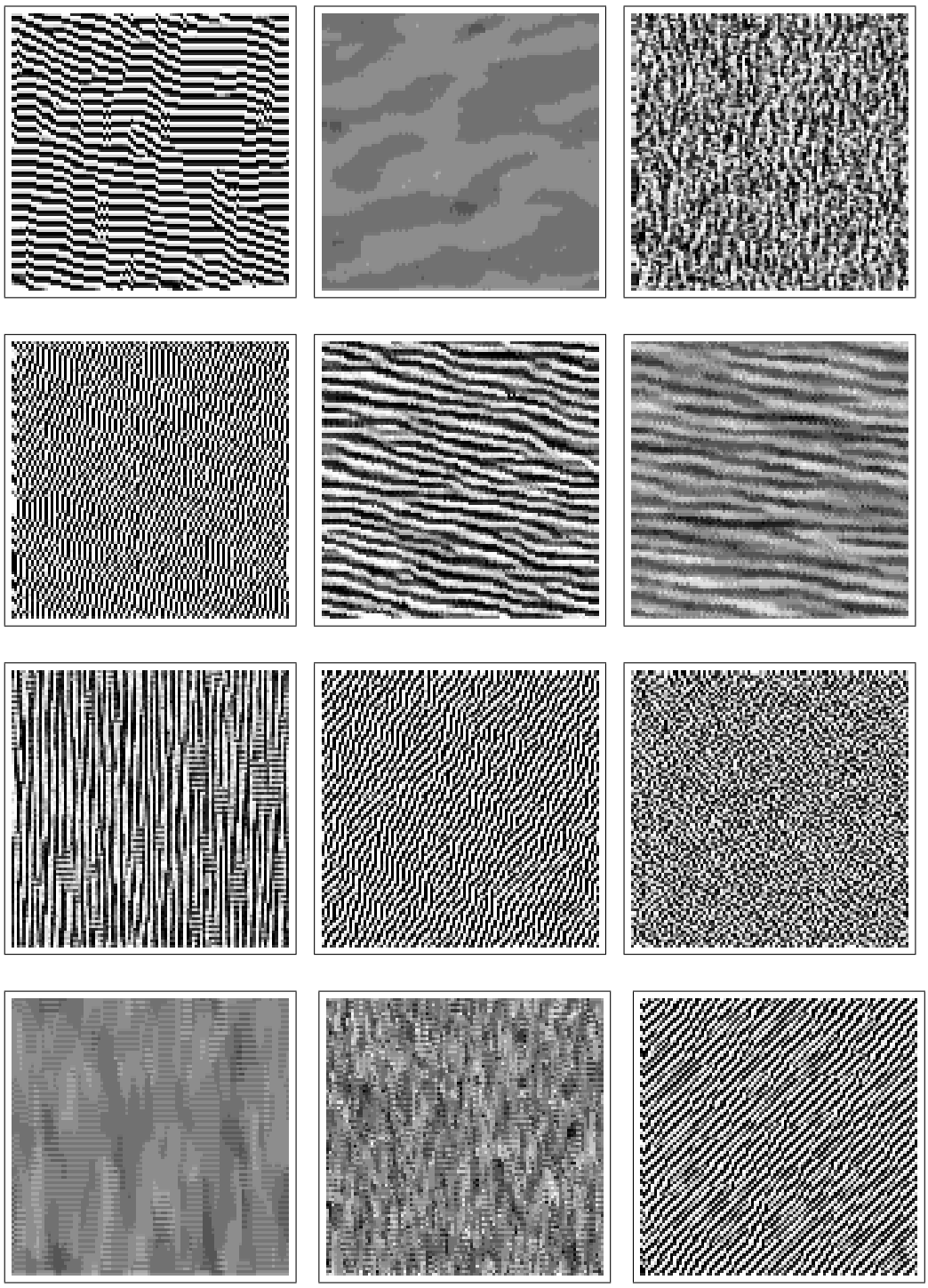


Abbildung 3.8: Verschiedene Muster durch den Gibbsampler aus verallgemeinerten Texturmodellen erzeugt (vgl. Bsp. 3.2.5). Bild: F. Friedrich, Heidelberg

Durch solche Modelle lassen sich beliebige binäre Muster charakterisieren. Die Modelle lassen sich auch in mannigfacher Weise auf mehr als zwei Intensitätsstufen verallgemeinern (vgl. z.B. G. WINKLER, (1995), [32]. Da wir uns in diesem Text nicht mit Texturanalyse beschäftigen, begnügen wir uns, die Syntese von Texturen anhand einiger Beispiele zu illustrieren, vgl. Abb. 3.8.

Wir schließen diesen Abschnitt mit einem Modell, welches D. und S. Geman 1984 unabhängig von Blake & Zisserman entwickelten, das aber auf ähnlichen Ideen beruht.

Beispiel 3.2.6 (Das Modell von Geman und Geman) D. und S. GEMAN (1984), [12], schlugen ein viel beachtetes bayessches Modell zur kantenerhaltenden Glättung vor. Die Grundidee ist die gleiche wie bei BLAKE & ZISSERMAN, deren Modell unabhängig zur selben Zeit entwickelt wurde. Es gehen wieder Intensitätsvariablen x_s und Kantenvariablen b_{st} ein. Die a priori Energie setzt sich aus zwei Teilen zusammen:

$$H(x,b) = H_1(x,b) + H_2(b). \tag{3.6}$$

Der erste Term ist wieder für An- bzw. Abschalten der Glättung zuständig:

$$H_1(x,b) = \vartheta \sum_{s\sim t} \varphi(x_s - x_t)(1 - b_{st}), \;\; \vartheta > 0.$$

Wir beginnen mit einer besonders einfachen Funktion φ, nämlich

$$\varphi(0) = -1 \quad \text{and} \quad \varphi(u) = 1 \quad \text{sonst},$$

die wir schon im Zusammenhang mit dem Potts Modell kennengelernt haben. Die Terme in der Summe nehmen die Werte -1, 0 oder 1 gemäß folgender Tabelle an:

	Kontrast	
Kante	nein	ja
an	0	0
aus	-1	1

Dies Autoren halten diese Funktion bei wenigen Graustufen (etwa weniger als 15) für sinnvoll. Für größere Grauwertebereiche schlagen sie Funktionen vom Typ des früheren φ vor, z. B.

$$\varphi(u) = 1 - \frac{2}{1 + (u/\delta)^2}. \tag{3.7}$$

Der Term

$$H_2(b) = -\gamma W(b), \ \gamma > 0, \tag{3.8}$$

dient als Organisationsterm für glatte Kantenverläufe. Die Funktion W zählt ausgezeichnete lokale Kantenkonfigurationen (vgl. Abbildung 3.9)) und gewichtet sie je nach dem Grad ihrer Regularität. Glatte Flächen enthalten keine Kanten und deshalb bekommen leere Kantenkonfigurationen große Gewichte w_0. Entsprechend bekommen glatte Kantenstücke hohe Gewichte $w_1 < w_0$; Knicke und T-förmige Stücke werden mit $w_3 \leq w_2 \leq w_1$ schlechter behandelt und blinde Enden oder Kreuze werden durch $w_4 < w_3$ bestraft. Die lokalen Kantenkonfigurationen sind in Abbildung 3.9dargestellt. Man könnte noch einen Term für Zusammenhang einfügen; das Modell würde dann wesentlich komplizierter.

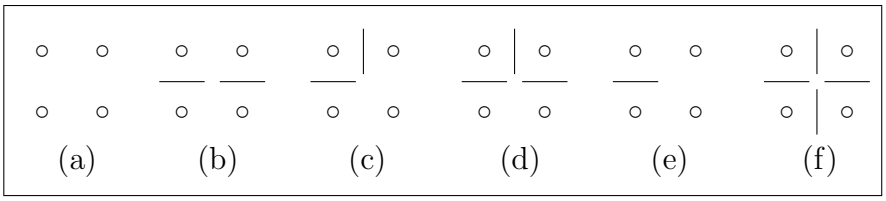


Abbildung 3.9: Ausgewählte lokale Konfigurationen aktiver Mikrokanten

Die a priori Energie ist nun spezifiziert. Sie wird schließlich durch den Datenterm ergänzt, der z.B. bei weißem gaußischem Rauschen wie bei Blake&Zisserman aussieht.
Das Modell von A. Blake & A. Zisserman war

$$H(x,b) = \lambda^2 \sum_{s\sim t} (x_s - x_t)^2 (1 - b_{st}) + \alpha \sum_{s\sim t} b_{st} + \sum_s (x_s - y_s)^2. \tag{3.9}$$

Es kann als der Spezialfall von (3.6) für $\gamma = 0$ in (3.8) und mit

$$\varphi(u) = u^2 - \alpha$$

aufgefaßt werden: Es gilt für $n \in \mathbb{N}$, daß

$$(x_s - x_t)^2 (1 - b_{st}) + \alpha b_{st} = \big((x_s - x_t)^2 - \alpha\big)(1 - b_{st}) + \alpha;$$

Also unterscheidet sich (3.9) von (3.6) nur durch die additive Konstante $|\{s \sim t\}|\alpha$, die das entsprechende Gibbsmaß jedoch nicht ändert.

Abb. 3.10 zeigt ein hübsches (und arbeitsintensives) Beispiel, wie mit solchen Modellen etwa Kanten extrahiert werden können. Das Hauptproblem besteht in der Bestimmung der Parameter, die noch nicht abschließend geklärt ist. Die Arbeit wurde von Felix Friedrich, Heidelberg, getan.

Abbildung 3.10: Links: Original, rechts: Kantenextraktion mit einem Kantenmodell vom Typ Beispiel 3.2.6. Bild: F. Friedrich, Heidelberg

Zunächst ist allerdings nicht klar, wie etwa Minima der Energiefunktion in den betrachteten Modellen berechnet werden sollen. Dies wird uns später noch beschäftigen.

Die diskutierten Ideen finden auch in vordergründig völlig anderem Zusammenhang Anwendung; z.B. in der Bewegungsanalyse, in der Tomographie usw. Weitere Beispiele findet man in Winkler (1995), [32].

3.2.3 Übergangswahrscheinlichkeiten

Wir kennen schon den Fall additiven (gaußischen) Rauschens. Impulsrauschen haben wir ebenfalls kennengelernt - es ist nicht additiv. Diese Eigenschaft hat auch das folgende Beispiel.

Beispiel 3.2.7 (Kanalrauschen) Wir betrachten Binärbilder x mit $x_s = \pm 1$. In den Pixeln werden die Zustände unabhängig geflippt:

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

Dann gilt

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

und

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

Wegen

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

ergibt sich

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

3.2.4 A posteriori Schätzer

Unser Ziel ist, mit Hilfe der a posteriori Verteilung möglichst gute Rekonstruktionen x aus den Daten y abzuleiten. Hier liegt also ein statistisches Schätzproblem vor.

Wie wir schon gesehen haben, stehen Schätzer in engem Zusammenhang mit assoziierten Verlustfunktionen. Eine *Verlustfunktion* $L : \boldsymbol{X} \times \boldsymbol{X} \to [0, \infty)$ erfüllt $L(x,x) = 0$.

Im bayesschen Kontext sind die Parameter x durch die a priori Verteilung gewichtet. Deshalb definiert man das *Bayesrisiko* eines Schätzers $\hat{X}(x)$ für x als

$$R(\hat{X}) = \sum_{x,y} L(x, \hat{X}(y)) \Pi(x,y)\,,$$

den erwarteten Verlust. Ein Schätzer der R minimiert, heißt *Bayesschätzer* zu L.

Beispiel 3.2.8 Sei

$$L(x,\hat{x}) = \begin{cases} 0 \text{ falls } x = \hat{x} \\ 1 \text{ sonst} \end{cases}.$$

Es gilt

$$\begin{aligned} \hat{R} &= \sum_{x,y \neq \hat{X}(y)} \Pi(x,y) = \Pi((x,y) : x \neq \hat{X}(y), y \in \boldsymbol{Y}) \\ &= 1 - \Pi((x,y) : x = \hat{X}(Y)) = 1 - \Pi((\hat{X}(y),y) : y \in \boldsymbol{Y}). \end{aligned}$$

Es ist

$$\Pi((\hat{X}(y),y) : y \in \boldsymbol{Y}) = \sum_y \Pi(\hat{X}(y)|y)\, \Pi(y).$$

Dies wird maximal für

$$\hat{X}(y) = \text{argmax}_x\, \Pi(x|y),$$

den *maximum a posteriori Schätzer* (*MAP*) Schätzer (welcher nicht unbedingt eindeutig ist). Die lokale Version des *MAP* Schätzers wird ebenfalls benutzt.

Beispiel 3.2.9 Die Zahl

$$d(x,\hat{x}) = |\{s \in S : x_s \neq \hat{x}_s\}|$$

heißt *Hamming Abstand*. Sei

$$L(x,\hat{x}) = |S|^{-1} d(x,\hat{x}).$$

Eine Rechnung analog zum letzten Beispiel zeigt, daß der zugeordnete Bayesschätzer gegeben ist durch:

$$(\hat{X}(y))_s = \text{argmax}_{x_s} \Pi(x_s|y), \;\; s \in S$$

Als letztes Beispiel betrachten wir kleinste Quadrate.

Beispiel 3.2.10 Für die Verlustfunktion

$$L(x,\hat{x}) = \sum_s (x_s - \hat{x}_s)^2$$

ist

$$\hat{X}(y) = \mathbb{E}_{\Pi(\cdot|y)} = \sum_x \Pi(x|y)$$

der Bayesschätzer. Er heißt (a posteriori) kleinster Quadrate Schätzer ('minimum mean square estimator' - *MMS*E).

3.3 Sampling und Annealing

Wir befassen uns nun mit der Berechnung oder Approximation der genannten Bayesschätzer, wobei wir uns im wesentlichen auf die *MAP*- und die *MMS*-Schätzung beschränken. Bei gegebenen Daten y sind also eine Minimalstelle x der a posteriori Energie $H(x)$ bzw. der Erwartungswert der a posteriori Verteilung $\Pi(\cdot|y)$ zu bestimmen. Da letzterer bei gegebenem y eine Gibbsverteilung ist, unterdrücken wir das Symbol y und stellen die Aufgabe: Sei eine Energiefunktion

$$H : \boldsymbol{X} \longrightarrow \mathbb{R}, \;\; x \longmapsto H(x)$$

gegeben.

1. Finde eine Minimalstelle x^* von H.
2. Bestimme den Erwartungswert der Gibbsverteilung Π zu H.

Aufgrund der hohen Zahl der Komponenten von x stehen analytische Hilfsmittel zur Bestimmung von x^* im kontinuierlichen Fall nur in Spezialfällen zur Verfügung. Sind die Komponenten von $\boldsymbol{X}$ diskret, so ist die Mächtigkeit von $\boldsymbol{X}$ leicht in der Größenordnung $256^{256\times 256}$; für allgemeine H sind keine diskreten Optimierungsmethoden bekannt. Die Erwartungswerte für (2) können ebenfalls nicht direkt berechnet werden, da die Zustandssumme Z über $|\boldsymbol{X}|$ Terme gebildet wird. Einen möglichen Ausweg bilden stochastische Verfahren, die auf bedingten Verteilungen beruhen. Diese sind nämlich für eine große Klasse praktisch relevanter Energiefunktionen handhabbar.

3.3.1 Bedingte Verteilungen

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

Beispiel 3.3.1 (Potentiale) Sei $\boldsymbol{X} = \Pi_{x\in S}\boldsymbol{X}_s$ und H von der Gestalt

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

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

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

von $\Pi \propto \exp(-H(x))$, daß

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

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

Beispiel 3.3.2 (Ising Modell) Im Ising Modell

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

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

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

d.h.

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

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

3.3.2 Der Kontraktionskoeffizient

In diesem Abschnitt bezeichnen $\boldsymbol{X}$ eine endliche Menge, P eine *Übergangswahrscheinlichkeit* oder einen *Markovkern* auf $\boldsymbol{X}$ und μ, ν, ϱ usw. Verteilungen auf $\boldsymbol{X}$, d.h. P erfüllt $P(x,y) \geq 0$ und $\sum_y P(x,y) = 1$, $x \in \boldsymbol{X}$. Für jedes $x \in \boldsymbol{X}$ ist also $P(x,\cdot)$ eine Wahrscheinlichkeitsverteilung auf $\boldsymbol{X}$.

Um die Qualität der Konvergenz zu messen, benötigen wir eine Metrik. $\mathcal{P}$ bezeichne die Menge aller Wahrscheinlichkeitsmaße auf $\boldsymbol{X}$. Für $\mu, \nu \in \mathcal{P}$ sei

$$\|\mu - \nu\| = \sum_x |\mu(x) - \nu(x)|$$

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

Bemerkung 3.3.3 Es gilt

$$\|\mu - \nu\| = 2\sum_x (\mu(x) - \nu(x))^+ = 2\left(1 - \sum_x \min\{\mu(x), \nu(x)\}\right).$$

Beweis der Behauptung in Bemerkung 3.3.3 Elementare Rechnung liefert:

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

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

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

und somit die erste Identität. Die zweite folgt aus:

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

□

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

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

Bemerkung 3.3.4 Es gilt offensichtlich

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

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

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

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

Die beiden Ungleichungen im folgenden Lemma sind fundamental; ihr Nachweis ist elementar und deshalb seien sie unbewiesen aus WINKLER (1995) zitiert:

Lemma 3.3.5 *Auf $\mathcal{P}$ gelten stets*

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

Schließlich benötigen wir noch eine einfache Abschätzung.

Lemma 3.3.6 *Für jeden Markovkern auf einem endlichen Raum $\boldsymbol{X}$gilt*

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

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

Beweis Mit Bemerkung 3.3.3 folgt

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

für Verteilungen μ and ν. Also gilt

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

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

3.3.3 Homogene Markovketten

Eine *homogene Markovkette* ist gegeben durch eine Startverteilung ν und einen Markovkern P. Die n-te Marginalverteilung ist

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

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

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

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

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

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

Lemma 3.3.7 *Für jeden strikt positiven Markovkern Q ist $c(Q) < 1$.*

Beweis Dies folgt unmittelbar aus Lemma 3.3.6. □

Daraus folgt schon der Konvergenzsatz.

Satz 3.3.8 *Sei P primitiv mit invarianter Verteilung μ. Dann gilt*

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

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

Beweis Es gilt für große n und $k(n) = \max\{k \geq 1 : k \cdot \tau \leq n\}$, daß

$$\begin{aligned} & \|\nu P^n - \mu\| = \|\nu P^n - \mu P^n\| \leq \|\nu - \mu\| c(P^n) \\ \leq \;& 2c(P^{k(n)\cdot\tau})c(P^{n-k(n)\cdot\tau}) \leq 2(c(P)^\tau)^{k(n)} \longrightarrow 0, n \to \infty, \end{aligned} \tag{3.10}$$

wodurch der Satz bewiesen ist. □

Satz 3.3.9 *Jeder primitive Markovkern besitzt genau eine invariante Verteilung. Diese ist strikt positiv.*

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

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

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

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

Schließlich liefert Normalisierung von ν eine invariante strikt positive Verteilung μ. Die Eindeutigkeit folgt schon aus Satz 3.3.8.
Sei P schließlich primitiv mit $P^\tau > 0$. Wieder hat P den Eigenwert 1 und es gibt $\nu \in \mathbb{R}^{\boldsymbol{X}} \backslash \{0\}$, so daß $\nu P = \nu$ und somit auch $\nu P^\tau = \nu$. Da $P^\tau > 0$, hat ν die Gestalt $a\mu$, $a \in \mathbb{R}$, für die eindeutig bestimmte invariante Verteilung μ von P^τ. Somit ist μ auch die eindeutig bestimmte invariante Verteilung von P und sie ist überdies strikt positiv. Ist $\tilde{\nu}$ eine weitere invariante Verteilung von P, so ist $\tilde{\nu}$ auch P^τ-invariant und deshalb gleich ν. □

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

Satz 3.3.10 (Schwaches Gesetz der großen Zahlen) *Sei $\boldsymbol{X}$eine endliche Menge und sei P ein Markovkern auf $\boldsymbol{X}$ mit invarianter Verteilung μ and $c(P) < 1$. Dann gilt für jede Startverteilung ν und jede Funktion f auf $\boldsymbol{X}$, daß*

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

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

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

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

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

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

$$\begin{aligned}
&\mathbb{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} f(\xi_i) - \mathbb{E}(f)\right)^2\right) = \mathbb{E}\left(\left(\sum_x f(x) g_x\right)^2\right) \\
\leq\ &\mathbb{E}\left(\left(\sqrt{\sum_x f(x)^2}\sqrt{\sum_x g_x^2}\right)^2\right) = \|f\|_2^2 \sum_x \mathbb{E}(g_x^2). \qquad (3.11)
\end{aligned}$$

Man berechnet

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

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

$$\begin{aligned}
&\sum_x |\nu P^i(x)\varepsilon_x P^k(x) - \mu(x)^2| \\
\leq\ &\sum_x |\nu P^i(x)\varepsilon_x P^k(x) - \mu P^i(x)\varepsilon_x P^k(x)| + |\mu(x)\varepsilon_x P^k(y) - \mu(x)\mu P^k(x)|
\end{aligned}$$

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

Daraus folgt

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

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

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

Mit Hilfe von (3.10) können das zweite und dritte Mittel abgeschätzt werden:

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

Also gilt

$$\sum_x \mathbb{E}\left(g_x^2\right) \leq \frac{13}{n(1-c(P))}.$$

Zusammen mit (3.11) ist damit der erste Teil der Behauptung bewiesen. Der zweite Teil folgt daraus mit der Markovschen Ungleichung. □

3.3.4 Gibbs und Metropolis Sampler

Unser Ziel war, Stichproben gemäß einem Gibbsmaß Π zu ziehen (*sampling*). Im Hinblick auf die Grenzwertsätze für Markovketten ist also eine homogene

Markovkette mit invarianter Verteilung Π zu konstruieren, die rechentechnisch handhabbar ist. *Eine* Möglichkeit der Konstruktion basiert auf den bedingten Verteilungen zu Π; der entsprechende Algorithmus heißt *Gibbs Sampler*. Er wird später in Konkurenz zu leichter simulierbaren Ketten – wie dem Metropolis Sampler – stehen. Die Theorie des Gibbs Samplers ist jedoch besonders elegant und durchsichtig.

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

$$\Pi_I(x,y) = \begin{cases} Z_I^{-1} \exp(-H(y_I x_{S\setminus I})) & \text{falls } y_{S\setminus I} = x_{S\setminus I} \\ 0 & \text{sonst} \end{cases} \tag{3.12}$$
$$Z_I = \sum_{z_I} \exp(-H(z_I x_{S\setminus I})).$$

Diese Markovkerne nennt man *lokale Charakteristiken* von Π. Sampeln aus $\Pi_I(x,\cdot)$ ändert x höchstens auf I. Die Gibbs Verteilung Π ist invariant für Π_I. Dies folgt aus einer stärkeren, aber leichter nachzuweisenden Eigenschaft.

Lemma 3.3.11 *Die Gibbs Verteilung Π und ihre lokalen Charakteristiken Π_I erfüllen die* detailed balance *Gleichung, d.h. für alle $x, y \in \boldsymbol{X}$ und $I \subset S$ gilt*

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

Allgemeiner erfüllen die Verteilung μ und der Markovkern P die *Detailed Balance Gleichung*, wenn

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

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

Beweis von Lemma 3.3.11 Entweder beide Seiten der Identität verschwinden oder es gilt $y_{S\setminus I} = x_{S\setminus I}$. Da $x = x_I y_{S\setminus I}$ und $y = y_I x_{S\setminus I}$, gilt weiter

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

Daraus folgt die detailed balance Gleichung. □

Stationarität folgt daraus leicht.

Satz 3.3.12 *Falls μ und P die detailed balance Gleichung erfüllen, so ist μ invariant für P. Insbesondere sind Gibbs Verteilungen invariant bezüglich ihrer lokalen Charakteristiken.*

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

Eine Abzählung $S = \{s_1, \ldots, s_\sigma\}$, $\sigma = |S|$, von S heißt *Besuchsschema*. Zur Vereinfachung schreiben wir $S = \{1, \ldots, \sigma\}$. Der Markovkern, welcher den Übergang während eines Sweeps beschreibt, ist definiert durch

$$P(x, y) = \Pi_{\{1\}} \ldots \Pi_{\{\sigma\}}(x, y). \tag{3.13}$$

Man beachte, daß die Verknüpfungen in (3.13) Matrixmultiplikationen sind. Die homogene Markovkette zum Kern P entspricht dem folgenden Algorithmus:

- Ziehe Startkonfiguration x gemäß Startverteilung ν (z.B. $\nu = \delta_x$).
- Update x durch die neue, zufällige Komponente y_1 gezogen gemäß $\Pi_{\{1\}}(x, \cdot x_{S\setminus\{1\}})$. Dies ergibt eine neue Konfiguration $y = y_1 x_{S\setminus\{1\}}$, die dann in der 2. Komponente upgedated wird usw. Nach dem Schritt Nummer σ ist ein *Sweep* beendet.
- Führe nun (viele) weitere Sweeps durch.

Dieser Algorithmus wird durch folgende Aussage gerechtfertigt:

Satz 3.3.13 *Für jedes $x \in \boldsymbol{X}$ gilt*

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

gleichmäßig in allen Startverteilungen ν.

Beweis Die Gibbsverteilung $\mu = \Pi$ ist invariant bzgl. ihrer lokalen Charakteristiken nach Satz 3.3.12 und deshalb auch bzgl. deren Komposition P. Außerdem ist $P(x, y)$ strikt positiv, da in jedem $s \in S$ die Wahrscheinlichkeit, ein y_s zu sampeln strikt positiv ist. Deshalb ist der Satz ein Spezialfall von Satz 3.3.8. □

Man kann die Pixel auch in zufälliger Reihenfolge besuchen: Sei G eine Verteilung auf S. Man ersetze die lokalen Charakteristiken (3.12) in (3.13) durch Kerne

$$\tilde{\Pi}(x,y) = \begin{cases} G(s)\Pi_{\{s\}}(x,y) & \text{wenn } y_{S\setminus\{s\}} = x_{S\setminus\{s\}} \text{ für ein } s \in S \\ 0 & \text{sonst} \end{cases} \tag{3.14}$$

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

Satz 3.3.14 *Sei G eine strikt positive Verteilung auf S. Dann gilt*

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

für alle $x \in \boldsymbol{X}$.

Beweis Π und $\tilde{P}$ erfüllen die detailed balance Gleichung und somit ist Π invariant für $\tilde{P}$. Weiter folgt aus der strikten Positivität von G, daß $\tilde{P}$ strikt positiv ist und die Konvergenz ergibt sich nach Satz 3.3.8. □

Beispiel 3.3.15 Wir illustrieren, wie sich die Muster unter dem Gibbssampler entwickeln und wie sie von den Parametern abhängen. Dazu betrachten wir das Ising Modell mit äußerem Feld, gegeben durch die Energiefunktion

$$H(x) = -\beta \sum_{s,t} x_s x_t + h \sum_s x_s, \;\; x_s = \pm 1. \tag{3.15}$$

Abb. 3.11 zeigt für verschiedene Parameter 'Schnappschüsse' aus dem Gibbssampler, d.h. die Zustände nach ausgewählten Schritten. $x_s = 1$ bedeutet dabei, daß die Farbe von Pixel s Schwarz ist. Kleines β, d.h. hohe Temperatur, führt zu klein strukturierten Mustern, hohes β, d.h. niedrige Temperatur, zu gleichmäßigen Flächen. Hohes h bestraft schwarze Pixel und drängt das Muster in Richtung Weiß.

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

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

ist die *Oszillation* von H in s und

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

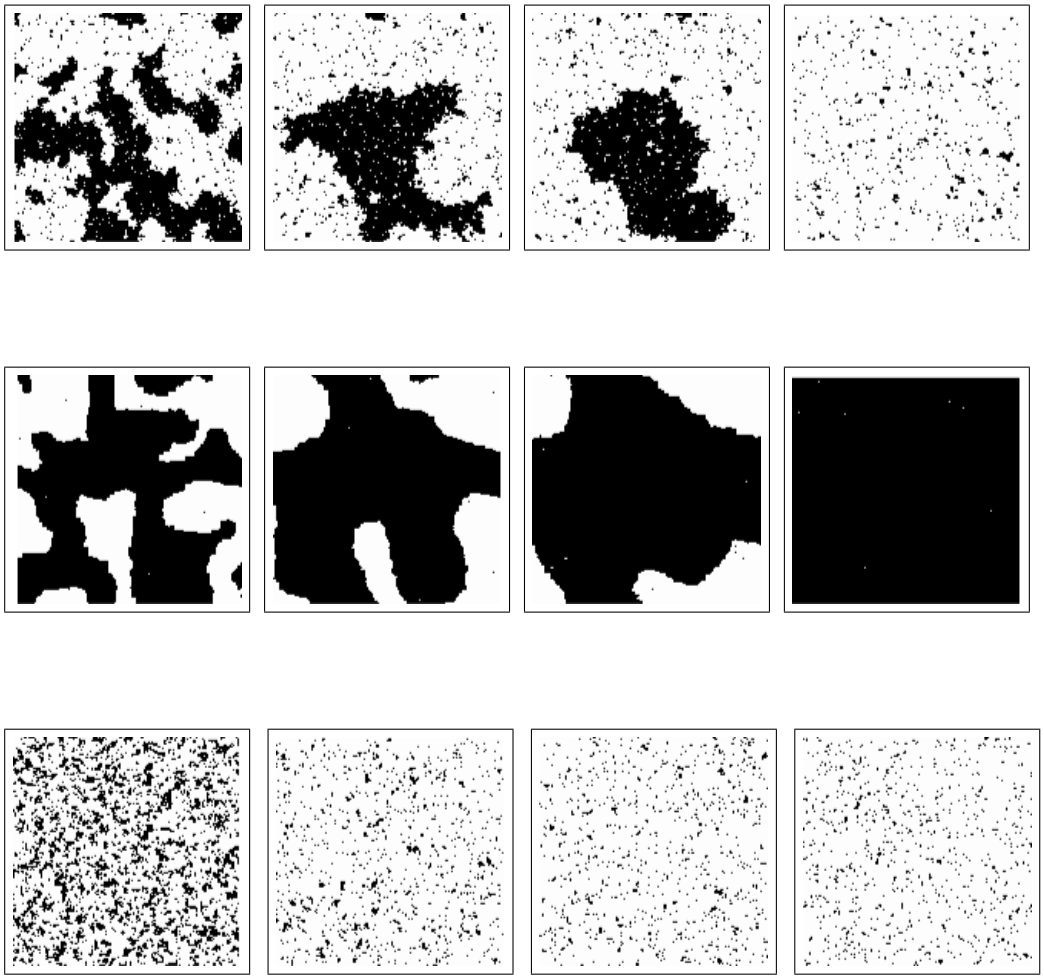


Abbildung 3.11: Schnappschüsse aus dem Sampler für das Ising Modell (mit äußerem Feld) mit Energiefunktion (3.15). 1. Reihe: Parameter $\beta = 0.5$, $h = 0$, $100, 500, 1000, 5000$ Sweeps. 2. Reihe: Parameter $\beta = 1$, $h = 0$, $100, 500, 1000, 5000$ Sweeps. 3. Reihe: Parameter $\beta = 0,4$, $h = 0,2$, $2, 5, 10, 100$ Sweeps. Bild: F. Friedrich, Heidelberg

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

Satz 3.3.16 *Sei (ξ_i) durch (3.13) oder (3.14) induziert. Dann gilt für jede Funktion f auf $\boldsymbol{X}$, daß*

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

in L^2 und in Wahrscheinlichkeit. Für jedes $\varepsilon > 0$ gilt

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

wobei $c = 13\|f\|^2$ für (3.13) und $c = 13\|f\|^2 \min_S G(s)^{-\sigma}$ für (3.14).

Beweis Der Markovkern P in (3.13) ist strikt positiv und deshalb folgt die L^2-Konvergenz aus Satz 3.3.10. Für das Gesetz der großen Zahlen müssen wir den Kontraktionskoeffizienten abschätzen.
Für $x \in \boldsymbol{X}$ sei z_s eine Minimalstelle in s, d.h.

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

Dann gilt

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

und somit

$$\min_{x,y} P(x,y) \geq \prod_{s=1}^{\sigma} \big(|\boldsymbol{X}_s|^{-1} e^{-\sigma_s}\big) = |\boldsymbol{X}|^{-1}\mathrm{e}^{-\Delta\sigma}.$$

Die Abschätzung in Lemma 3.3.6 liefert

$$c(P) \leq 1 - |\boldsymbol{X}| \cdot \min_{x,y} P(x,y) \leq 1 - \mathrm{e}^{-\Delta\sigma}. \tag{3.16}$$

Damit folgt das Gesetz der großen Zahlen im Fall (3.13). Der Beweis für (3.14) unterscheidet sich nur in offensichtlichen Details. □

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

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

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

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

folgt aus dem Satz:

Proposition 3.3.17 *Unter den Voraussetzungen von Satz 3.3.16 gilt* $A_{x,n} \longrightarrow \Pi(x)$ *in Wahrscheinlichkeit.*

Insbesondere besucht der Gibbs Sampler jeden Zustand unendlich oft.

Die Konvergenzsätze für den Gibbs Sampler lassen sich in besonders eleganter Weise herleiten. Deshalb wurde er für den theoretischen Teil benutzt. In der Praxis erweisen sich oft andere Verfahren als vorteilhaft oder leichter zu programmieren. Wir skizzieren nun die vielleicht bekannteste Alternative, den *Metropolis Sampler*. Sein Ursprung wird üblicherweise auf die Arbeit (M. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller und E. Teller (1953)), [22], zurückgeführt. Sei H eine Energiefunktion (möglicherweise ersetzt durch die parametrisierte Version βH) und sei x die Konfiguration, die gerade aktualisiert werden soll. Das Updating erfolgt in zwei Schritten:

1. Der Vorschlag.

   Eine neue Konfiguration y wird erzeugt durch Sampeln aus einer Verteilung $G(x, \cdot)$ auf $\boldsymbol{X}$.

2. Der Annahmeschritt.

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

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

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

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

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

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

Beispiel 3.3.18 Bei Bildern ist ein natürlicher Vorschlag, ein Pixel rein zufällig zu wählen und dort einen zufälligen Grauwert einzusetzen. Genauer heißt das

$$G(x,y) = \begin{cases} \frac{1}{\sigma(N-1)} & \text{if } x_s \neq y_s \text{ für genau ein } s \in S \\ 0 & \text{sonst} \end{cases} \tag{3.17}$$

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

Er hat explizit die Gestalt

$$\pi(x,y) = \begin{cases} G(x,y)\exp(-(H(y)-H(x))^+) & \text{if } x \neq y \\ 1 - \sum_{z \in \boldsymbol{X}\setminus\{x\}} \pi(x,z) & \text{if } x = y \end{cases} \tag{3.18}$$

Definition 3.3.19 *Eine stochastische Matrix $(G(x,y)_{x,y\in X}$ heißt* irreduzibel*, wenn es zu jedem x und y in $\boldsymbol{X}$ eine Folge $x = x_0, x_1, \ldots, x_{n(x,y)-1}, x_{n(x,y)} = y$ gibt mit $G(x_{i-1}, x_i) > 0$ für alle $1 \leq i \leq n(x,y)$.*

Bemerkung 3.3.20 Primitive stochastische Matrizen sind irreduzibel. Irreduzibilität erzwingt *nicht* die Existenz eines n, das für alle Paare gemeinsam obige Kettenbedingung erfüllt (wie bei primitiven Matrizen). Existiert ein solcher Index, so ist die Matrix irreduzibel und *aperiodisch*.

Der entsprechende Konvergenzsatz lautet

Satz 3.3.21 *Sei $\boldsymbol{X}$ eine endliche Menge, H eine nichtkonstante Funktion auf $\boldsymbol{X}$ und Π die Gibbs Verteilung zu H. Die Vorschlagsmatrix G sei symmetrisch und irreduzibel. Dann gilt:*

(a) Für jedes $x \in \boldsymbol{X}$ und jede Startverteilung ν auf $\boldsymbol{X}$ gilt

$$\nu\pi^n(x) \longmapsto \Pi(x) \quad as \quad n \to \infty.$$

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

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

in L^2 und in Wahrscheinlichkeit.

3.3.5 Inhomogene Markovketten

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

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

Wir beweisen wieder einen Konvergenzsatz für die Marginalverteilungen.

Lemma 3.3.22 *Seien μ_n, $n \geq 1$, Verteilungen auf $\boldsymbol{X}$, so daß $\sum_n \|\mu_{n+1} - \mu_n\| < \infty$. Dann gibt es eine Verteilung μ_∞, so daß $\mu_n \to \mu_\infty$ (in $\|\cdot\|$), wenn $n \to \infty$.*

Da $\boldsymbol{X}$endlich ist, fallen punktweise und L^1-Normkonvergenz bzgl. $\|\cdot\|$ zusammen.

Beweis Für $m < n$ gilt

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

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

Satz 3.3.23 *Seien P_n, $n \geq 1$, Markovkerne und jedes P_n besitze eine invariante Verteilung μ_n. Ferner seien die folgenden Bedingungen erfüllt:*

$$\sum_n \|\mu_n - \mu_{n+1}\| < \infty, \tag{3.19}$$

$$\lim_{n \to \infty} c(P_i \ldots P_n) = 0 \quad \textit{für jedes} \quad i \geq 1. \tag{3.20}$$

Dann existiert eine Verteilung μ_∞ für die

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

gleichmäßig in allen Startverteilungen ν gilt.

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

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

Außerdem gilt

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

Zum Beweis: der erste Term in der Summe ist

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

der letzte ist gleich

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

und die Kombination von Paaren aufeinanderfolgender Terme liefert

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

Damit gilt

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

Daraus folgt

$$\|\mu_\infty P_i \dots P_{i+k} - \mu_\infty\| \leq 2 \cdot \sup_{n \geq i} \|\mu_\infty - \mu_n\| + \sum_{n \geq i} \|\mu_n - \mu_{n+1}\|. \tag{3.21}$$

Verwendet wurden dabei Lemma 3.3.5 und die Beschränktheit des Kontraktionskoeffizienten durch 1. Mit Bedingung (3.19) und weil μ_∞ existiert, wird

für großes i der Ausdruck auf der rechten Seite klein. Für $2 \leq i \leq n$ fahren wir fort mit

$$\begin{aligned} & \|\nu P_i \dots P_n - \mu_\infty\| \\ = \; & \|(\nu P_i \dots P_n - \mu_\infty P_i \dots P_n) + \mu_\infty P_i \dots P_n - \mu_\infty\| \qquad (3.22) \\ \leq \; & 2 \cdot c(P_i \dots P_n) + \|\mu_\infty P_i \dots P_n - \mu_\infty\| \end{aligned}$$

Für große n wird der erste Term wegen (3.20) klein. Damit ist der Beweis vollständig. □

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

Lemma 3.3.24 *Für Verteilungen μ_n, $n \geq 1$ gilt Bedingung (3.19) falls jede der Folgen $(\mu_n(x))_{n\geq 1}$ schließlich fällt oder wächst.*

Beweis Nach Lemma 3.3.4 gilt

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

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

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

für alle großen N. Deshalb ist die Doppelsumme endlich und Bedingung (3.19) ist erfüllt. Damit ist der Beweis vollständig. □

Lemma 3.3.25 *Bedingung (3.20) ist erfüllt, falls*

$$\prod_{k\geq i} c(P_k) = 0 \quad \textit{für jedes} \quad i \geq 1. \qquad (3.23)$$

oder falls

$$c(P_n) > 0 \quad \textit{für jedes } n \textit{ und} \quad \prod_{k\geq 1} c(P_k) = 0. \qquad (3.24)$$

Beweis Bedingung (3.23) impliziert (3.20) wegen der zweiten Regel in Lemma 3.3.5, und (3.24) impliziert offensichtlich (3.23). □

3.3.6 Simulated Annealing

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

Sei H eine Funktion auf $\boldsymbol{X}$. Die Funktion βH hat für große β dieselben Minima wie H, jedoch sind diese tiefer.

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

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

Sei M die Menge der globalen Minimalstellen von H.

Proposition 3.3.26 *Sei Π eine Gibbs Verteilung mit Energiefunktion H. Dann gilt*

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

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

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

Beweis Sei m der minimale Wert von H. Dann gilt

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

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

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

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

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

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

Bemerkung 3.3.27 Mit $\beta \to 0$ konvergiert die Gibbs Verteilung Π^β gegen die Gleichverteilung auf $\boldsymbol{X}$, denn in

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

strebt jeder Exponent gegen 1.

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

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

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

Satz 3.3.28 *Sei $(\beta(n))_{n\geq 1}$ ein Abkühlschema, welches gegen Unendlich wächst, so daß schließlich*

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

Dann gilt

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

gleichmäßig in allen Startverteilungen ν.

Der Satz stammt von S. und D. Geman (1984). Wir schicken dem Beweis ein Lemma voraus.

Lemma 3.3.29 *Für die reellen Folgen (a_n) und (b_n) gelte $0 \leq a_n \leq b_n \leq 1$. Dann folgt aus $\sum_n a_n = \infty$, daß $\prod_n (1 - b_n) = 0$.*

Beweis Aus der bekannten Ungleichung

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

folgt

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

Da die Summen divergieren, gilt

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

was zu

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

äquivalent ist. □

Beweis des Satzes Wir müssen $\beta(n)$ so wählen, daß die Voraussetzungen von Satz 3.3.23 mit obigen P_n und $\mu_n = \Pi^{\beta(n)}$ erfüllt sind. Daraus und mit Satz 3.3.26 folgt dann die Behauptung.
Das Gibbsfeld μ_n ist invariant bezüglich der Kerne P_n nach Satz 3.3.12. Da $(\beta(n))$ wächst, sind die Folgen $(\mu_n(x))$, $x \in \boldsymbol{X}$ schließlich monoton wegen Proposition 3.3.26 und deshalb gilt (3.19) wegen Lemma 3.3.24. Wie in (3.16) gilt

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

Wir leiten nun die hinreichende Bedingung (3.23) d.h. $\prod_{k \geq i} c(P_k) = 0$ für alle i her. Nach Lemma 3.3.29 gilt dies, falls $\exp(-\beta(n)\Delta\sigma) \geq a_n$ für $a_n \in [0, 1]$ mit divergenter unendlicher Reihe. Eine naürliche Wahl ist $a_n = n^{-1}$ und damit ist

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

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

Das logarithmische Abkühlschema ist willkürlich gewählt, da die Bedingung

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

entscheidend ist. Oft benutzt man stückweise konstante Schemata.

Der Satz gilt auch für zufällige Besuchsschemata in (3.14). Dann setzt man

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

Es ist

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

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

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

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

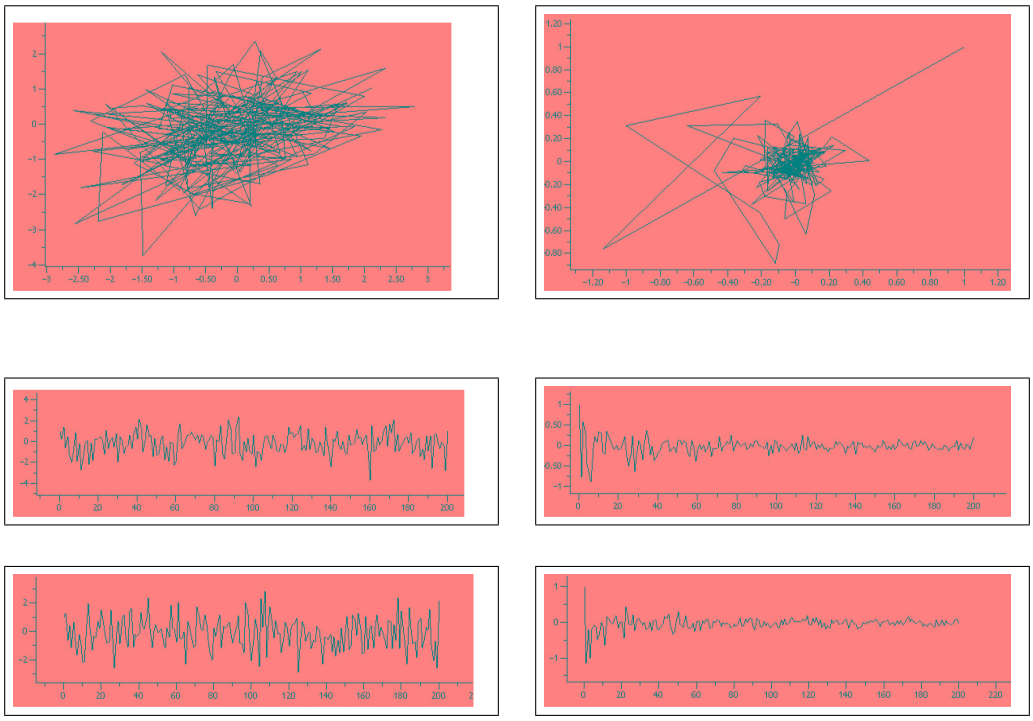


Abbildung 3.12: Linke Spalte: Ein Pfad des Samplers für (3.25), darunter der Verlauf der x_1- und x_2-Komponente. Rechte Spalte: Analog für Simulated Annealing. Bild: F.Friedrich, Heidelberg

Beispiel 3.3.30 Sampling und Annealing können auch auf kontinuierlichen Räumen durchgeführt werden, etwa für gewisse Verteilungen mit Dichten. An

die Stelle der bedingten diskreten Verteilungen, aus denen gesampelt wird, treten dann eben bedingte Dichten. Das folgende einfache Beispiel illustriert das Verhalten von Pfaden der entsprechenden Markovketten. Bei endlichem Zustandsraum erhält man qualitativ ähnliche Effekte.

Wir betrachten auf $\mathbb{R}^2$ die Gaußverteilung mit negativem Exponenten

$$H(x) = (x_1 - x_2/5)^2 + (x_2 - x_1/5)^2, \;\; f(x) \propto \exp(-H(x)). \tag{3.25}$$

Wir generieren mit Hilfe des Gibbssamplers und des gibbsschen Annealings Pfade des assoziierten Markovprozesses. Abb. 3.12 zeigt in der linken Spalte einen Pfad des Gibbssamplers in der Ebene sowie den Verlauf der x_1- bzw. x_2-Komponente. In der rechten Spalte ist analog ein Pfad des Annealing dargestellt. Man sieht deutlich, wie der Sampler die Gaußverteilung 'exploriert', während sich der Bereich des Gibbssamplers immer mehr um die Minimalstelle $x^* = (0, 0)$ konzentriert.

3.4 Exakter MAP für das Pottsmodell

In diesem Abschnitt stellen wir für ein spezielles Modell einen exakten Algorithmus zur Berechnung von *MAP* Schätzern vor. Als a priori Verteilung wählen wir das nach dem Ising Modell einfachste Modell, nämlich das Potts Modell aus Beispiel 3.2.4. Es beruht auf der a priori Energiefunktion

$$H(x) = \gamma' \sum_{s \sim t} \chi_{\{(a,b):a \neq b\}}(x_s, x_t) = \gamma' |s \sim t : x_s \neq x_t|,$$

wobei γ' ein - üblicherweise nichtnegativer - Hyperparameter ist. Die a posteriori Verteilung hat also die Energiefunktion

$$\gamma' |s \sim t : x_s \neq x_t| - \ln P(x, y).$$

Wählen wir für die Störung des idealen Signales ein additives weißes gaußisches Rauschen, d.h.

$$y_s = x_s + \eta_s$$

mit i.i.d. zentrierten gaußischen Variablen η_s mit Varianz σ^2, so ergibt sich eine a posteriori Verteilung mit Energiefunktion

$$\gamma' |s \sim t : x_s \neq x_t| + \frac{1}{2\sigma^2} \sum_s (y_s - x_s)^2$$

(bis auf eine additive Konstante). Für laplace-verteiltes weißes Rauschen mit Dichte

$$f(x) = \frac{1}{\sqrt{2}\sigma} \exp\left(-\frac{\sqrt{2}}{\sigma}|x|\right).$$

bekommt man

$$\gamma'|s \sim t : x_s \neq x_t| + \frac{\sqrt{2}}{\sigma} \sum_s |y_s - x_s|,$$

ebenfalls bis auf eine additive Konstante. Da der *MAP* nicht von der Multiplikation mit einer positiven Konstante abhängt, können wir (bei bekanntem σ) äquivalent die Funktion

$$\gamma|s \sim t : x_s \neq x_t| + \sum_s (y_s - x_s)^2 \tag{3.26}$$

oder

$$\gamma|s \sim t : x_s \neq x_t| + \sum_s |y_s - x_s|. \tag{3.27}$$

minimieren, wobei γ dem Wert $\gamma' \cdot 2 \cdot \sigma^2$ bzw. $\gamma'\sigma/\sqrt{2}$ entspricht.

Bemerkung 3.4.1 Man kann zeigen, daß es im Falle (3.26) bis auf eine Lebesguesche Nullmenge $L \subset \mathbb{R}^{|S|}$ für alle $y \in \mathbb{R}^{|S|} \backslash L$ eine eindeutige Minimalstelle $x^*(y)$ gibt (vgl. [20]). Daraus folgt insbesondere, daß die Minimalstelle von (3.26) fast sicher in y eindeutig ist. Diese Aussage läßt sich auf allgemeineres Rauschen ausdehnen ([20]).

3.4.1 Der Algorithmus

Wir beschränken uns auf eine Dimension und betrachten folgende Situation:

1. $S = \{1, \ldots, N\}$ ist ein eindimensionales Gitter mit nächsten Nachbarn.

2. Das Rauschen des Gesamtsignals ist exponential verteilt mit der Energiefunktion

$$D(x, y) = \sum_{s \in S} \rho(x_s - y_s).$$

Die für den *MAP* zu minimierende Funktion hat dann in Verallgemeinerung von (3.26) und (3.27) die Gestalt

$$H(x) = \gamma \cdot |\{s \sim t : x_s \neq x_t\}| + \sum_{s \in S} \rho(x_s - y_s).$$

Wir beschreiben die (idealen) Signale in einer dieser Situation angepaßten Weise:

Ein Signal x ist eindeutig beschrieben durch eine Partition $\mathcal{P} = \{I\}$ von S in (diskrete) Intervalle I und die konstanten Werte μ_I auf diesen Intervallen. Umgekehrt kann jedes Signal x aus dem Paar $(\mathcal{P}, \mu_{\mathcal{P}})$, bestehend aus einer Partition $\mathcal{P}$ und einer Familie $\mu_{\mathcal{P}} = (\mu_I)_{I \in \mathcal{P}}$ der Werte μ_I auf den Intervallen $I \in \mathcal{P}$ rekonstruiert werden. In dieser Notation nimmt die Funktion H die Gestalt

$$\begin{aligned} & H(x) = H((\mathcal{P}, \mu_{\mathcal{P}})) = H_{\mathcal{P}(x)}(\mu_{\mathcal{P}(x)}) \\ = \quad & \gamma \cdot (|\mathcal{P}(x)| - 1) + \sum_{I \in \mathcal{P}(x)} \sum_{s \in I} \rho(y_s - \mu_I), \end{aligned} \tag{3.28}$$

an, wobei $|\mathcal{P}|$ die Anzahl der Intervalle I der Partition $\mathcal{P}$ bezeichnet, und μ_I jeweils der konstante Wert von x auf dem Intervall I ist. Sei der *Hyperparameter* γ nun fixiert.

Ehe wir versuchen, die Energiefunktion zu minimieren, halten wir folgende Beobachtung fest: Sobald die Partition $\mathcal{P}$ gegeben ist - und somit der erste Term in (3.28) fest liegt - reduziert sich das Problem auf die Minimierung des zweiten Terms. Dieser wird dann minimal, wenn jede einzelne Summe

$$\sum_{s \in I} \rho(y_s - \mu_I)$$

minimal ist; sei die Minimalstelle jeweils mit μ_I^* bezeichnet. Somit können wir für diejenigen speziellen Funktionen ρ, für die wir optimale Schätzer μ_I^* kennen, diese einsetzen. Einfache Beispiele sind:

- Für $\rho(u) = |u|$ ist der Wert μ_I^* gerade der Median der y_s, $s \in I$;
- für $\rho(u) = u^2$ ist μ_I^* das empirische Mittel der y_s, $s \in I$.

Somit können wir das Optimierungsproblem vereinfachen: Wir minimieren $H_{\mathcal{P}}(\mu_{\mathcal{P}}^*)$ über alle Partitionen $\mathcal{P}$. Wir definieren nun eine Intervallenergiefunktion H_I durch

$$H_I(\mu_I^*) = \gamma + \sum_{s \in I} \rho(y_s - \mu_I^*).$$

Sei $\mathcal{P} = \{I_1, \ldots, I_k\}$ eine Partition von S. Die Summe aller Intervallenergiefunktionen H_I ist nicht $H_{\mathcal{P}}$, sondern

$$\begin{aligned}
\sum_{I_l \in \mathcal{P}} H_{I_l}(\mu^*_{I_l}) &= \gamma + \sum_{s \in I_1} \rho(y_s - \mu^*_{I_1}) + \gamma + \sum_{s \in I_2} \rho(y_s - \mu^*_{I_2}) \\
&\quad + \ldots + \gamma + \sum_{s \in I_k} \rho(y_s - \mu^*_{I_k}) \\
&= \gamma \cdot \underbrace{k}_{=|\mathcal{P}|} + \sum_{I \in \mathcal{P}} \sum_{s \in I} \rho(y_s - \mu^*_I) \\
&= \underbrace{\gamma \cdot (|\mathcal{P}| - 1) + \sum_{I \in \mathcal{P}} \sum_{s \in I} \rho(y_s - \mu^*_I)}_{=H_{\mathcal{P}}(\mu^*_{\mathcal{P}})} + \gamma \\
&= H_{\mathcal{P}}(\mu^*_{\mathcal{P}}) + \gamma
\end{aligned}$$

Schreiben wir H für die Partition $\mathcal{P}$ auf, so bekommen wir die Beziehung

$$\begin{aligned}
H_{\{I_1,\ldots,I_k\}}(\mu_{\{I_1,\ldots,I_k\}}) &= -\gamma + \sum_{l=1}^{k} H_{I_l}(\mu_{I_l}) \\
&= -\gamma + \sum_{l=1}^{k-1} H_{I_l}(\mu_{I_l}) + H_{I_k}(\mu_{I_k}) \\
&= H_{\{I_1,\ldots,I_{k-1}\}}(\mu_{\{I_1,\ldots,I_{k-1}\}}) + H_{I_k}(\mu_{I_k}).
\end{aligned}$$

Wir definieren für jedes $1 \leq n \leq N$ die Funktion

$$B(n) = \min_{\mathcal{P}(n)} H_{\mathcal{P}(n)}(\mu^*_{\mathcal{P}(n)}), \tag{3.29}$$

wobei sich die Minimierung über alle Partitionen $\mathcal{P} = \mathcal{P}(n)$ von $\{1, \ldots, n\}$ erstreckt. Mit obiger Rechnung erhalten wir

$$B(n) = \min_{\mathcal{P}(n)} H_{\mathcal{P}(n)}(\mu^*_{\mathcal{P}(n)}) = \min_{0 \leq r \leq n-1} [B(r) + \min_{\mu \in \mathbb{R}} H_{[r+1,n]}(\mu)]$$

für alle $1 \leq n \leq N$. Man beachte, daß $B(N)$ das Minimum der Energiefunktion $H(x)$ ist. B ist eine *Bellmann Funktion*. Das Optimierungsproblem wird mit Hilfe *dynamischer Programmierung* gelöst.

Der entsprechende Algorithmus kann wie folgt dargestellt werden :

1. Berechne für alle Intervalle $[r, s]$ für $1 \leq r < s \leq N$ die Minimierer

$$\mu^*_{[r,s]} = \arg\min_{\mu \in \mathbb{R}} H_{[r,s]}(\mu)$$

und $H_{[r,s]}(\mu^*_{[r,s]})$. Beachte, daß $H_{[r,r]}(\mu^*_{[r,r]}) = \gamma$.

2. Setze $B(0) = -\gamma$ und $B(1) = 0$.

3. Bestimme rekursiv $B(n)$ für alle $1 < n \leq N$ und speichere mindestens ein $r^*_n \in \{0, \ldots, n-1\}$, in welchem das Minimum angenommen wird.

4. Konstruiere rekursiv vom Ende her eine Partition von S:

Die Minimalstelle r^*_N von $B(N)$ wird der letzte Punkt des vorletzten Intervalles. Das letzte Intervall ist damit $[r^*_N + 1, N]$. Suche nun die Minimalstelle von $B(r^*_N), \ldots$, um die nächsten Unterteilungspunkte zu bekommen. Diese Konstruktion ergibt eine Partition $\{I_1, \ldots, I_k\}$ von {1,. . . ,N } und x^* ist gegeben durch $x^*_s = \mu^*_I$ für $s \in I$.

Dieser Algorithmus liefert in Abhängigkeit von γ stückweise konstante Approximationen der Daten. Im Limes $\gamma \to 0$ werden die Daten zurückgeliefert, für $\gamma \to \infty$ bekommt man das konstante (mittlere) Signal. So kontrolliert γ die Glattheit von x^*. Deshalb liegt es nahe, das optimale x^* für *jeden Wert* von γ zu berechnen.

Hierzu berechnen wir ähnlich zur Intervallenergiefunktion H_I, eine Intervallfehlerfunktion D_I durch

$$D_I(\mu) = H_I(\mu) - \gamma$$

und die Partitionsfehlerfunktion $D_\mathcal{P}$ für eine Partition $\mathcal{P}$ durch

$$D_P(\mu_\mathcal{P}) = H_\mathcal{P}(\mu_\mathcal{P}) - \gamma \cdot (|\mathcal{P}| - 1) = \sum_{I \in \mathcal{P}} \sum_{s \in I} \rho(y_s - \mu_I).$$

Ähnlich wie für $H_{\{I_1, \ldots I_k\}}$ bekommt man

$$D_{\{I_1,\ldots,I_k\}}(\mu_{\{I_1,\ldots,I_k\}}) = D_{\{I_1,\ldots,I_{k-1}\}}(\mu_{\{I_1,\ldots,I_{k-1}\}}) + D_{I_k}(\mu_{I_k}).$$

Das Minimierungsproblem kann umgeschrieben werden in

$$\min_x H(x) = \min_\mathcal{P}[D_\mathcal{P}(\mu^*_\mathcal{P}) + \gamma \cdot (|\mathcal{P}| - 1)] = \min_{1 \leq k \leq N}[\min_{|\mathcal{P}|=k}(D_\mathcal{P}(\mu^*_\mathcal{P})) + \gamma \cdot (k-1)],$$

wobei $\mathcal{P}$ eine Partition von $\{1,\dots,N\}$ ist. Definiere nun

$$\tilde{B}(k,n) = \min_{|\mathcal{P}|=k}(D_P(\mu^*_{\mathcal{P}})),$$

wobei sich die Minimierung über alle Partitionen $\mathcal{P}$ von $\{1,\dots,n\}$ erstreckt. Aus den obigen Berechnungen ergibt sich die Formel

$$\tilde{B}(k,n) = \min_{1\le r\le n-1}[\tilde{B}(k-1,r) + \min_{\mu\in\mathbb{R}} D_{[r+1,n]}(\mu)]$$

für $N \ge n \ge k > 1$.

Wir wollten $H(x)$ für jeden Wert von γ über alle x minimieren. Sei nun

$$h(\gamma) = \min_x H(x) = \min_{1\le k\le N}[\tilde{B}(k,N) + \gamma\cdot(k-1)].$$

Wir beschreiben nun den Algorithmus:

1. Berechne für alle Intervalle $[r,s]$ for $1 \le r < s \le N$ die Minimierer

   $$\mu^*_{[r,s]} = \arg\min_{\mu\in\mathbb{R}} D_{[r,s]}(\mu)$$

   und $D_{[r,s]}(\mu^*_{[r,s]})$. Beachte, daß $D_{[r,r]}(\mu^*_{[r,r]}) = 0$.

2. Setze $\tilde{B}(1,n) = D_{[1,n]}(\mu^*_{[1,n]})$.

3. Bestimme $\tilde{B}(k,n)$ für alle $1 < k \le n \le N$ und speichere mindestens ein $r^*_{k,n}$ für welches das Minimum angenommen wird.

4. Konstruiere rekursiv Partitionen $\{I^k_1,\dots,I^k_k\}$ von $S = \{1,\dots,N\}$ aus den Minimierern von $\tilde{B}(k,N)$, $\tilde{B}(k-1,r^*_{k,N}),\dots$ und $x^{*,k}$ ist bestimmt durch $x^{*,k}_s = \mu^*_{I^k_l}$ für $s \in I^k_l$.

5. Konstruiere die stückweise lineare Funktion $h(\gamma)$.

Wir betrachten nun den letzten Schritt dieses Algorithmus, nämlich:

$$h(\gamma) = \min_x H(x) = \min_{1\le k\le N}[\tilde{B}(k,N) + \gamma\cdot(k-1)].$$

h ist stückweise linear. Sei I das (offene) Intervall zwischen zwei aufeinander folgenden Knotenpunkten. Auf I ist h linear, wobei die Steigung die Anzahl

der Sprünge in der minimierenden Folge $x^*(\gamma)$, $\gamma \in I$, angibt. Die Knotenpunkte werden auf folgende Weise bestimmt: Auf dem Intervall $[0, \gamma_1)$ mit

$$\gamma_1 = \min_{1 \le k \le N-1} \frac{\tilde{B}(N,N) - \tilde{B}(k,N)}{k-N}\,. \tag{3.30}$$

ist h linear mit Steigung $(N-1)$. Sei k_1 der Wert von k in (3.30), an welchem das entsprechende Minimum angenommen wird. Dann ist h linear mit Steigung $(k_1 - 1)$ bis zum nächsten Knoten

$$\gamma_2 = \min_{1 \le k \le k_1 - 1} \frac{\tilde{B}(k_1,N) - \tilde{B}(k,N)}{k-k_1}\,.$$

Der nächste Knoten wird analog bestimmt. Somit kann die Minimierung von H für alle Werte von γ auf die Bestimmung der endlich vielen Knoten von h zurückgeführt werden.

Wieder stehen sich die Forderungen nach Datentreue und Glattheit - die sich in der Zahl der Sprünge widerspiegelt - gegenüber. Für $\gamma \to 0$ bekommt man die Daten als Lösung mit maximalem $k = N$, da in diesem Fall der zweite Term überwiegt. Andererseits ergibt sich für $\gamma \to \infty$ ein optimales Signal x^*, welches konstant ist, also $1 - 1 = 0$ Sprünge hat.

Bemerkung 3.4.2 Die Wahl des 'richtigen' Hyperparameters γ ist ein entscheidendes und schwieriges Problem. Es gehört zum Problemkreis *Modellwahl*. In der Vergangenheit wurden einige Kriterien vorgeschlagen, z.B. das AIC-Kriterium von H. Akaike, [1], oder eine bayesianische Variante von G. Schwarz, [27]. Es ist aber zweifelhaft, ob diese in der vorliegenden Situation sinnvolle Ergebnisse liefern. Dies ist Gegenstand laufender Untersuchungen, z.B. in [20]. Die Darstellung über *alle* Hyperparameter ist ein erster Schritt.

Im Falle der linearen Regression wurden exakte AIC-Kriterien in [7] hergeleitet.

Wir schließen diesen Abschnitt mit einem Hinweis auf ein verwandtes Modell.

Bemerkung 3.4.3 Ein Prior, der weniger robust ist als der Potts Prior, aber eine gewisse Popularität genießt, ist der L^1-Prior. Er wurde in den letzten Jahren von einer Reihe von Autoren studiert. Wir beschränken uns auf den Fall einer Dimension. Wir gehen von einem Regressionsmodell der Gestalt

$$Y_s = m(s) + \eta_s$$

aus. Dabei sei m eine glatte Funktion auf dem Einheitsintervall $[0,1]$, die $s \in [0,1]$ seien die *Designpunkte*, d.h. die Zeiten oder Orte, an denen das verrauschte Signal gemessen wird. Wir nehmen gaußisches weißes Rauschen an. Ziel ist die Minimierung einer *penalized Likelihood* der Gestalt

$$L(m) = \lambda \mathrm{TV}(m) + \sum_s (y_s - m(s))^2, \;\; \lambda > 0, \tag{3.31}$$

wobei $\mathrm{TV}(m)$ die Totalvariation von m bezeichnet. Ist zum Beispiel m differenzierbar, so ist

$$\mathrm{TV}(m) = \int_0^1 |m'(t)|\, dt.$$

In der diskreten Situation kann die diskrete Ableitung immer gebildet werden. Die Funktion $L(m)$ geht dann über in den Ausdruck

$$L_d(m) = \lambda \sum_{s\tilde{t}} |m(s) - m(t)| + \sum_s (y_s - m(s))^2, \;\; \lambda > 0.$$

Dies kann als Energiefunktion einer a posteriori Verteilung mit a priori Energie

$$G(m) = \lambda \sum_{s \sim t} |m(s) - m(t)|$$

interpretiert werden. In [21] werden die Minimalstellen von (3.31) exakt bestimmt. Sie erweisen sich als Output des *Taut-String*-Algorithmus, der sich anschaulich wie folgt beschreiben läßt: Betrachte die kumulierten Daten

$$Z(s) = \sum_{t:t\le s} Y_t.$$

Betrachte ferner den Schlauch zwischen

$$Z^o = Z + \lambda, \;\; Z^u = Z - \lambda.$$

Fädle nun einen Faden durch den Schlauch und ziehe ihn stramm; er hat nun die Form einer stückweise linearen stetigen Funktion, eines *Taut-String* (wobei über die Ränder noch zu reden wäre). E. Mammen und S. van de Geer zeigen in [21], daß durch Differenzieren dieses '*Taut-String*' Minimalstellen der penalized Likelihood (3.31) entstehen. Es sind dies stückweise konstante Funktionen mit Sprüngen an den Knoten des *Taut-String*.

P.L. Davies und A. Kovac entwickeln in [11] eine lokal-adaptive Variante. Sie basiert darauf, den Schlauch räumlich variabel so lange zusammenzuziehen, d.h. λ ortsabhängig zu verringern, bis gewisse Multiresolutionskoeffizienten, ähnlich den Waveletkoeffizienten, der Residuen $Y_s - \hat{m}(s)$ des entsprechenden Taut-String $\hat{m}$ hinreichend klein werden. Dies wird dann als Hinweis darauf interpretiert, daß es sich bei den Residuen um Rauschen handelt und deshalb $\hat{m}$ eine vernünftige Schätzung von m ist.

P.L. Davies und A. Kovac argumentieren ferner, daß $\hat{m}$ ein asymptotisch modentreuer Schätzer ist, d.h. Anzahl und Lokalisation der Moden von m richtig wiedergibt (vgl. [11], Theorem 3.2. Dies ist im Hinblick auf das folgende Beispiel spannend, bei dem wir das Potts Modell ausschließlich zur Bestimmung der Zahl der Moden einsetzen. Ein systematischer Vergleich wird in [20] durchgeführt.

3.4.2 Beispiel: Gehirndaten aus der funktionellen Magnetresonanztomographie

Die Daten in Tabelle 3.1 stammen aus der funktionellen Magnetresonanztomographie (fMRT, im englischen auch fMRI). Bei dieser Technik werden mit Hilfe von Magnetfeldern Änderungen der paramagnetischen Eigenschaften des Blutes erfaßt. Diese treten insbesondere in Arealen des menschlichen Gehirnes auf, in denen erhöhte Gehirnaktivität stattfindet. Bei Aktivierung wird dort mehr Sauerstoff benötigt als in Ruhe und deshalb angefordert. Paradoxerweise erhöht sich die Sauerstoffkonzentration bei Aktivität trotz erhöhtem Verbrauch. Im konkreten Fall wurde die Testperson einem visuellen äußerem Reiz ausgesetzt: Ein schachbrettartiges Muster wird periodisch auf einem Bildschirm für 30 Sekunden gezeigt, bzw. abgeschaltet. Dies entspricht schematisch einem Signal, das aus einer Folge von kastenförmigen Teilstücken besteht, genannt *Boxcar-Signal*. Der ungefähre Verlauf eines solchen Signales ist in Abb. 3.13 dargestellt. Man erwartet, daß Bereiche des visuellen Kortex, darauf mit erhöhter Sauerstoffkonzentration 'antworten'. In einem vermutlich aktivierten Bereich wurden Daten gemessen, wie sie z.B. in Tabelle 3.1 als Punkte dargestellt sind; die Daten sind bereits grob von Delay, d.h. Zeitverzögerung, und dem linearen Trend bereinigt. Das Ziel ist festzustellen, ob ein Bereich des Gehirnes auf den Stimulus 'antwortet' oder nicht. Dazu versucht man, das Reizsignal 'Boxcar' im Response wiederzufinden. 'Gebiet' sind, von der Technik diktiert, gegenwärtig *Voxel* von der Größe

$3 \times 3 \times 5$ mm^3 (Voxel sind die dreidimensionale Entsprechung der zweidimensionalen Pixel). Der Response wird allerdings eine gestörte Version des

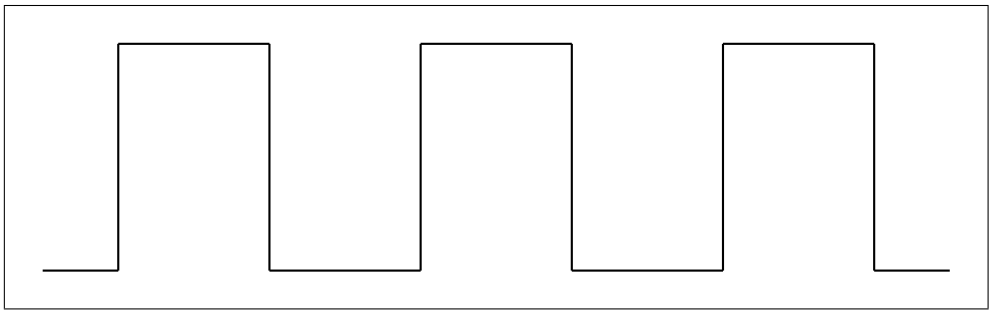
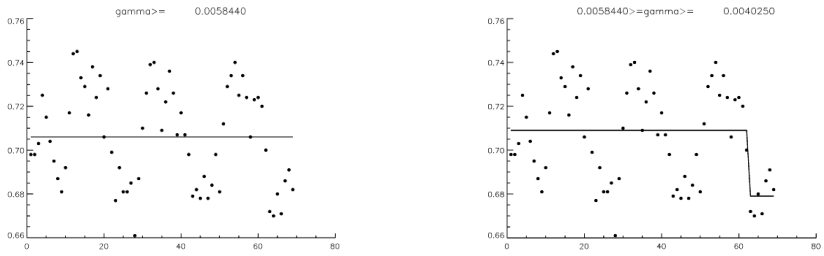
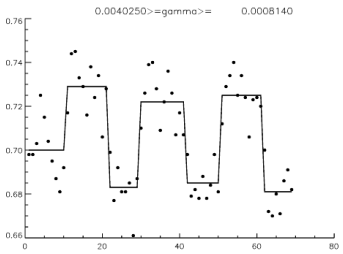


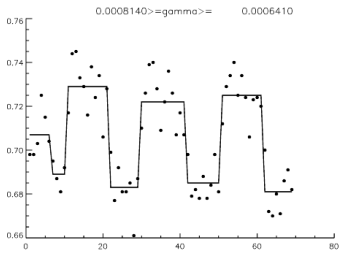
Abbildung 3.13: Idealisierte Gestalt eines Boxcarsignales .

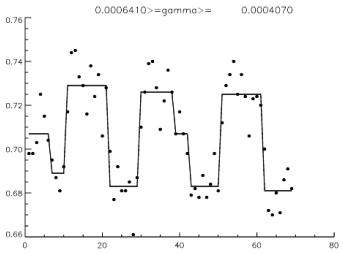
Stimulus sein. Somit scheint uns sinnvoll, nur sehr spartanische Features des Stimulus im Response zu suchen. Im vorliegenden Fall ist die Zahl der *Moden* oder der lokalen Extrema ein Kandidat. Deshalb wurden *MAP*-Schätzungen mit einer Potts a priori Verteilung, basierend auf dem dargestellten exakten Algorithmus durchgeführt. Das Rauschen wurde vereinfacht als gaußisches weißes Rauschen modelliert. Die Ergebnisse für verschiedene Werte des Hyperparameters γ sind in Tabelle 3.1 dargestellt.

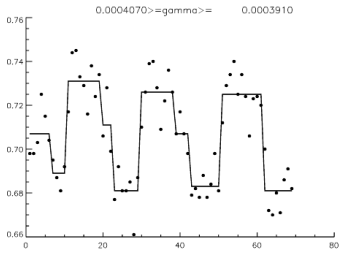
Tabelle 3.1: Exakte maximum a posteriori Schätzer im Potts Modell für die Boxcar-Daten aus der funktionellen Magnetresonanztomographie: Verschiedene Hyperparameter γ in aufsteigender Reihenfolge (aus [20]).

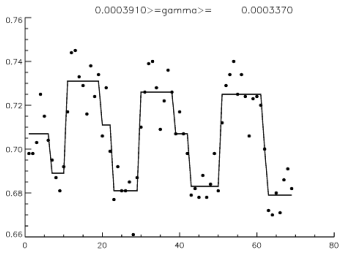


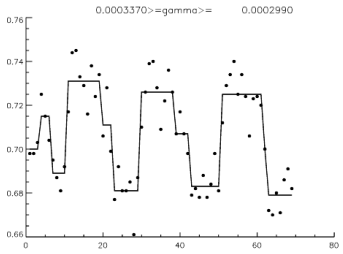

0.0040250>=gamma>= 0.0008140

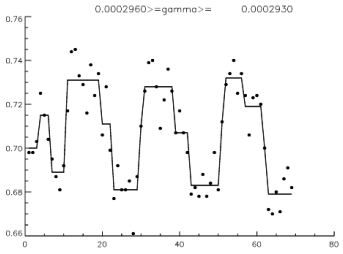

0.0008140>=gamma>= 0.0006410

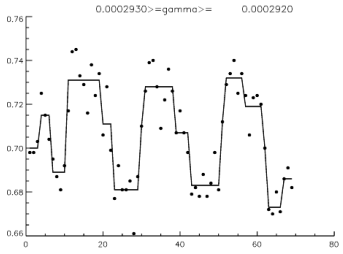

0.0006410>=gamma>= 0.0004070

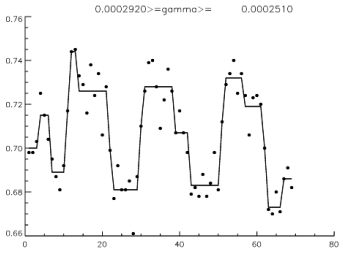

0.0004070>=gamma>= 0.0003910

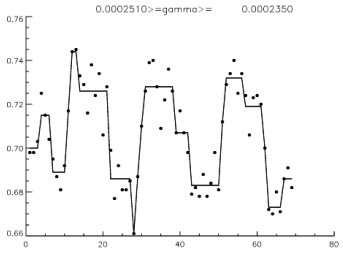

0.0003910>=gamma>= 0.0003370

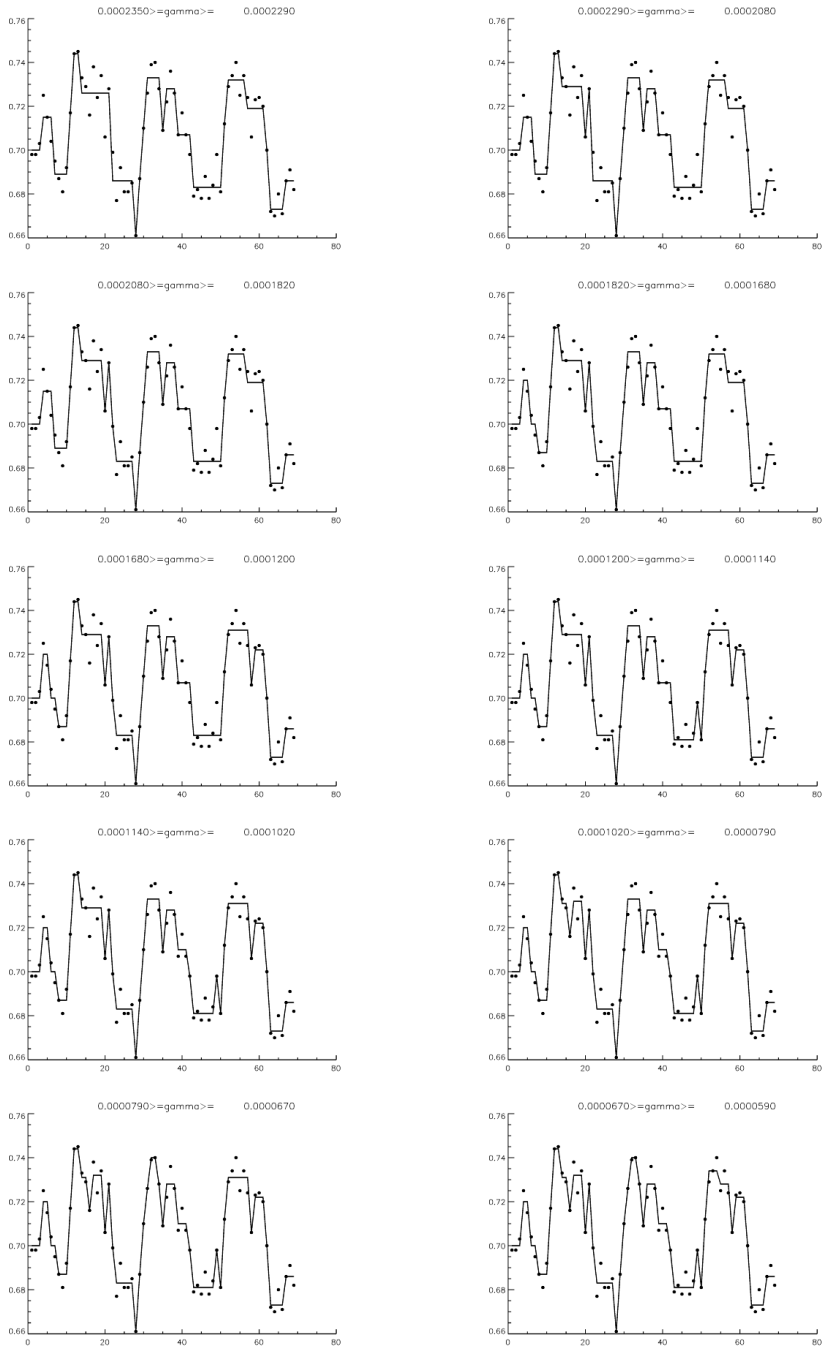

0.0003370>=gamma>= 0.0002990

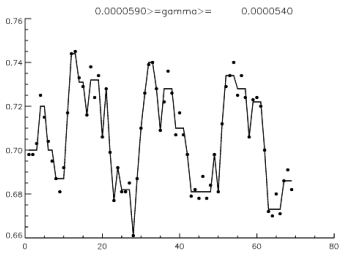

0.0002960>=gamma>= 0.0002930

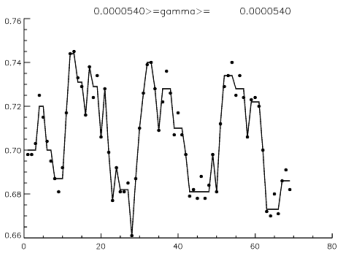

0.0002930>=gamma>= 0.0002920

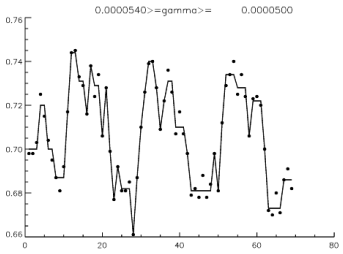

0.0002920>=gamma>= 0.0002510

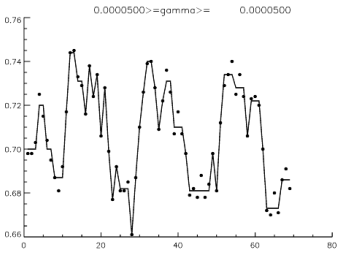

0.0002510>=gamma>= 0.0002350

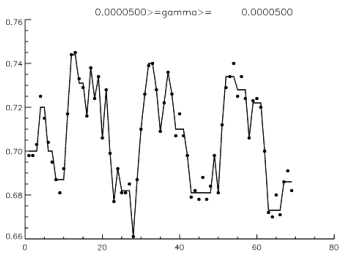

0.0002350>=gamma>= 0.0002290
0.0002290>=gamma>= 0.0002080
0.0002080>=gamma>= 0.0001820
0.0001820>=gamma>= 0.0001680
0.0001680>=gamma>= 0.0001200
0.0001200>=gamma>= 0.0001140
0.0001140>=gamma>= 0.0001020
0.0001020>=gamma>= 0.0000790
0.0000790>=gamma>= 0.0000670
0.0000670>=gamma>= 0.0000590

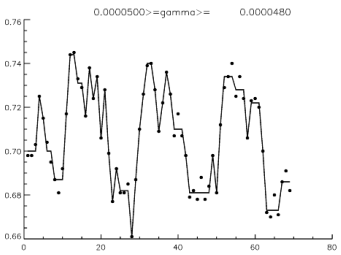

0.0000590>=gamma>= 0.0000540

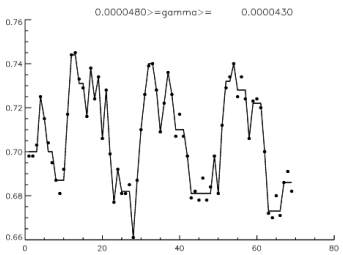

0.0000540>=gamma>= 0.0000540

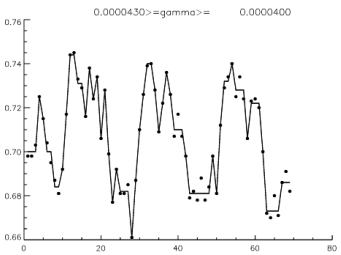

0.0000540>=gamma>= 0.0000500

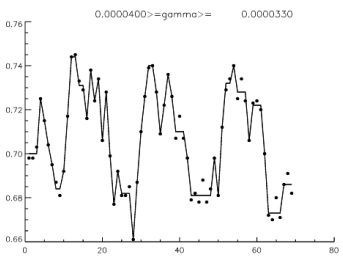

0.0000500>=gamma>= 0.0000500

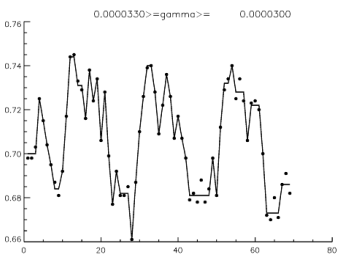

0.0000500>=gamma>= 0.0000500

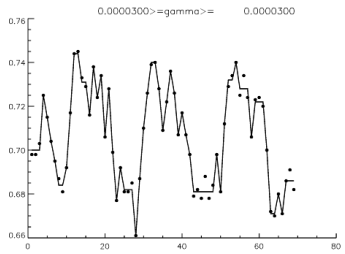

0.0000500>=gamma>= 0.0000480

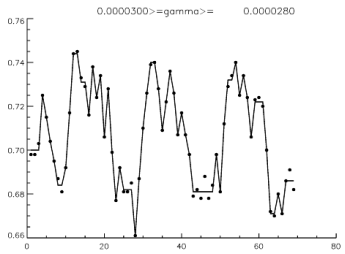

0.0000480>=gamma>= 0.0000430

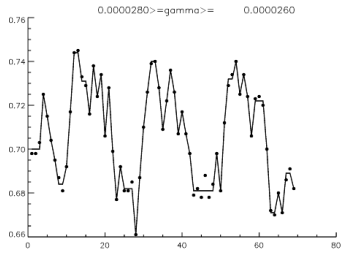

0.0000430>=gamma>= 0.0000400

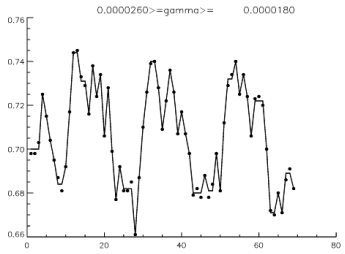

0.0000400>=gamma>= 0.0000330

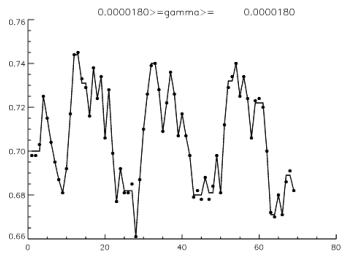

0.0000330>=gamma>= 0.0000300

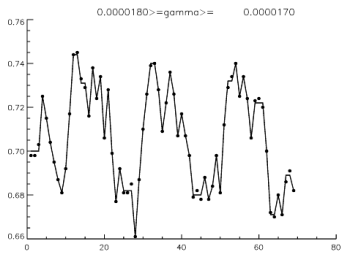

0.0000300>=gamma>= 0.0000300

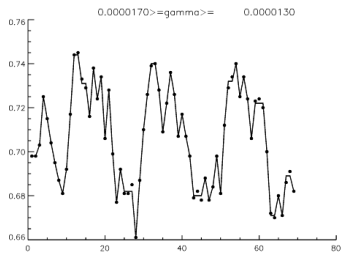

0.0000300>=gamma>= 0.0000280

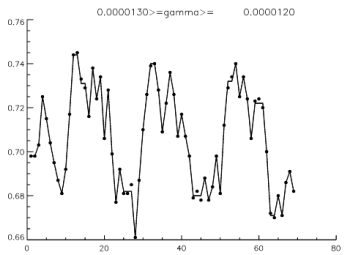

0.0000280>=gamma>= 0.0000260

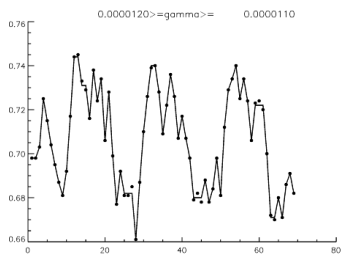

0.0000260>=gamma>= 0.0000180

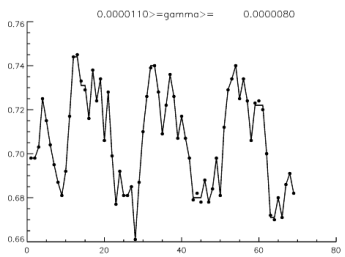

0.0000180>=gamma>= 0.0000180

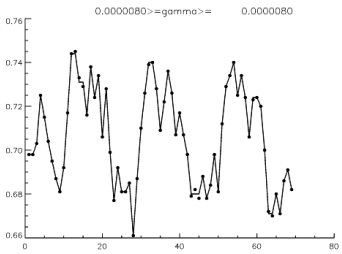

0.0000180>=gamma>= 0.0000170

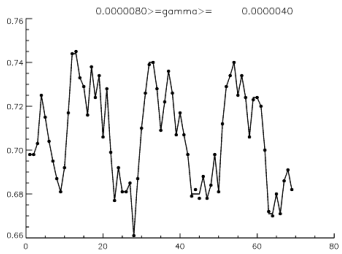

0.0000170>=gamma>= 0.0000130

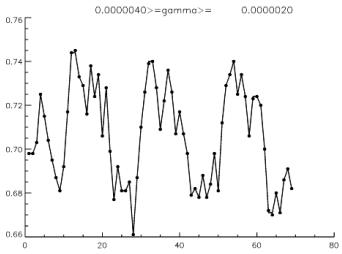

0.0000130>=gamma>= 0.0000120

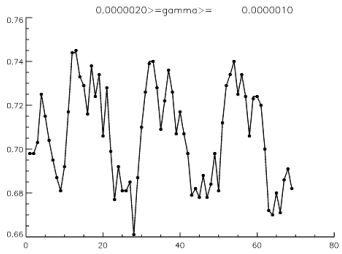

0.0000120>=gamma>= 0.0000110

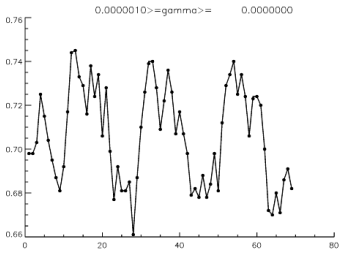

0.0000110>=gamma>= 0.0000080

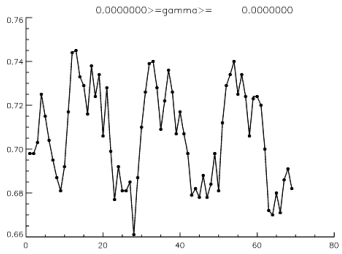

0.0000080>=gamma>= 0.0000080


0.0000080>=gamma>= 0.0000040


0.0000040>=gamma>= 0.0000020


0.0000020>=gamma>= 0.0000010


0.0000010>=gamma>= 0.0000000


0.0000000>=gamma>= 0.0000000

Kapitel 4

Eine Auswahl robuster Glätter

In diesem Kapitel stellen wir einige weitere kantenerhaltende Glätter vor, die z.B. in [34] und [35] diskutiert wurden. Allen ist gemeinsam, daß die für die Kantenerhaltung nötige Robustheit durch die Verwendung einer topfförmigen Funktion φ - wie sie im Bayesmodell auftaucht - wesentlich ist. An diesen Beispielen sollen die Gemeinsamkeiten von Filtern, d.h. Verfahren aus dem Ingenieurwesen, und Schätzern aus der Statistik herausgestellt werden. Es wird sich zeigen, daß diese stärker ist als es zunächst scheinen mag. Außerdem soll noch einmal der Aspekt der Robustheit und deren Verbindung zu Verlustfunktionen eingehender diskutiert werden.

Wir haben bereits betrachtet:

- die Bayesmethode mit a posteriori Maximumlikelihoodschätzung;

Wir diskutieren im folgenden:

- einen der Kerndichteschätzung verwandten Zugang, der auf lokaler M-Schätzung basiert;
- einen nichtlinearen Gaußfilter;
- einen Kette nichtlinearer Gaußfilter;
- ein lokal-adaptives statistisches Verfahren.

4.1 Lokale M-Glätter

Lokale M-Schätzer, wie wir sie jetzt diskutieren, wurden in C.K. Chu, I. Glad, F. Godtliebsen and J.S. Marron (1998), [9], eingeführt.

4.1.1 Globale and Lokale *M*-Schätzer

Wir diskutieren noch einmal die Rolle von topfförmigen Funktionen φ. Seien die Zufallsvariablen $Y_i, 1 \leq i \leq n$, i.i.d. gaußverteilt gemäß $\mathcal{N}(\mathbf{m}, \sigma^2)$, wobei die Varianz bekannt und der Erwartungswert unbekannt seien. Dann maximiert der Maximumlikelihoodschätzer ϑ^* für $\mathbf{m}$ die Likelihoodfunktion

$$\vartheta \longmapsto (2\pi\sigma^2)^{-n/2} \cdot \exp\left(-\frac{1}{2\sigma^2}\sum_i (Y_i - \vartheta)^2\right)$$

oder, dazu äquivalent, minimiert die Funktion

$$\vartheta \longmapsto \sum_i (Y_i - \vartheta)^2. \tag{4.1}$$

Durch Bestimmung der Nullstelle der ersten Ableitung sieht man sofort, daß das empirische Mittel $\bar{Y} = (\sum_i Y_i)/n$ das Problem löst; wir haben gesehen, daß es sogar ein BLUE ist, d.h. ein bester linearer unverfälschter Schätzer. Wir haben ferner argumentiert, daß das empirische Mittel extrem anfällig für Ausreißer ist. Wir können dies auch so ausdrücken: Es ist empfindlich gegen eine Kontamination der zugrundeliegenden (in diesem Falle gaußischen) Verteilung $\mathcal{N}(\mathbf{m}, \sigma^2)$, besonders wenn diese das Gewicht der Tails erhöht. In der Bildanalyse ist dies jedoch die generische Situation: Nehmen wir an, daß das 'ideale' Bild die Intensität $\mathbf{m}$ auf einer Seite einer Kante und Intensität $\mathbf{m}' \neq \mathbf{m}$ auf der anderen Seite. Nehmen wir weiter an, daß die Y_i verrauschte Intensitäten in einem gleitenden Fenster $B(s)$ um Pixel s sind. Wenn das Fenster sich von einer zur anderen Seite bewegt, gelangen mehr und mehr Pixel der $\mathbf{m}$-Seite in das Fenster und es enthält Beobachtungen sowohl aus $\mathcal{N}(\mathbf{m}, \sigma^2)$ als auch aus $\mathcal{N}(\mathbf{m}', \sigma^2)$. Ist die Mehrzahl der Pixel auf der $\mathbf{m}$-Seite, so möchten wir die Pixel der $\mathbf{m}'$-Seite aus dem Glättungsprozeß heraus halten. Eine andere Formulierung ist, i.i.d. Beobachtungen Y_i mit Verteilung

$$Y_i \sim (1-\alpha)\mathcal{N}(\mu, \sigma^2) + \alpha\mathcal{N}(\nu, \sigma^2)$$

zu betrachten, wobei $\alpha > 0$ der Anteil der '$\mathbf{m}'$-Pixel' ist. Diese Mischung von Gaußverteilungen hat möglicherweise stärkere Tails als die 'reine' Verteilung $\mathcal{N}(\mathbf{m}, \sigma^2)$. Dies ist genau die Situation, in der die robuste Statistik ins Spiel kommt.

Die üblichen robusten Schätzer für Lageparameter sind die sogenannten *M-Schätzer*. Die Quadratfunktion in (4.1) wird durch Funktionen φ ersetzt,

die nach außen langsamer wachsen, d.h. M-Schätzer sind als Minimalstellen ϑ^* der *M-Funktion*

$$\Phi : \vartheta \longmapsto \sum_i \varphi(Y_i - \vartheta) \tag{4.2}$$

definiert. Die Funktion φ gewichtet Ausreißer schwächer, was im Bildanalysekontext bedeuted,daß Kanten nicht zu stark überglätted werden. Typische Beispiele für Funktionen φ sind

- Gaußsche Quadrate $\varphi(u) = u^2$. (Regression wird weitgehend im Kontext von L^2-Räumen studiert). Die Lösung ist das empirische Mittel $\vartheta^* = \bar{Y}$

- Der Betrag der Abweichung $\varphi(u) = |u|$. Die L^2-Theorie wird durch die (schwierigere) L^1-Theorie ersetzt, vgl. P. Bloomfield and W.L. Steiger (1983), [6]. In diesem Falle ist ϑ^* der empirische Median. Er ist relativ robust, jedoch glättet er in glatten Gebieten schwächer als das (L^2) empirische Mittel, falls die Verteilung schwache Tails hat.

- P. Huber, [17], schlägt die Funktion φ vor, die innerhalb einer Kugel mit der Quadratfunktion übereinstimmt, konvex ist, und so schwach wie möglich außerhalb der Kugel wächst. Sie hat notwendig die Gestalt

  $$\varphi(u) = \chi_{|u|\leq 1} u^2 + \chi_{|u|>1}(2|u| - 1).$$

  Die zugehörige Gibbsverteilung heißt *Least Informative Distribution*. Dies ist eine Kombination des L^2- und des L^1-Falles. Um eine funktionalanalytische Theorie aufzubauen, müssen L^p-Räume durch Orlizs Räume ersetzt werden. Kanten werden nicht perfekt erhalten.

- F.R. Hampel, [15], schlägt Funktionen φ mit Ableitungen, die außen gegen 0 gehen, vor. Ein typisches Beispiel ist die so oft erwähnte topfförmige Funktion φ in (3.4). Funktionen dieses Typs sind am geeignetsten zur Kantenfindung.

M-Schätzer sind als stationäre Punkte der *M-Funktion* Φ in (4.2) definiert oder als Wurzeln der entsprechenden *Scorefunction* Φ' (cf. [15]). Diese Definition ist etwas verwirrend, weil Minima oder wenigstens lokale Minima von Φ wünschenswert sind. Überdies ist nicht klar, welches der eventuell zahlreichen lokalen Minima zu wählen ist. Dies führte übrigens zu einer beachtlichen

Kontroverse innerhalb der Statistik. Für die Bildanalyse ist dies eher angenehm und eine Chance. Es zeigt sich, daß lokale Minima nahe bei den Daten bessere Restaurationen liefern, als globale Minima. Dies werden wir wir im Abschnitt 4.1.2 und 4.1.3 besprechen.

4.1.2 Lokale M-Glätter

In einem Pixel s soll die Intensität abhängig von den Intensitäten $y_t, t \in B(s)$, in einem Fenster $B(s)$ um s neu bestimmt werden. Um Glättung über Kanten hinweg zu vermeiden, greifen wir auf die oben skizzierte *M-Schätzung* zurück und minimieren (4.2) entsprechend die Funktion

$$\Phi_s : \vartheta \longmapsto \sum_{t \in B(s)} \varphi(Y_t - \vartheta), \tag{4.3}$$

wobei $(Y_i)_{i=1}^n = (Y_t)_{t \in B(s)}$ und φ von dem oben diskutierten robusten Typ ist. Natürlich sind die zufälligen Intensitäten Y_t nicht mehr i.i.d. Dieser Schätzer wird über Plateaus mit glatten Rändern gut glätten. Andererseits wird er feine Details wie Spikes zerstören.

Um dies zu vermeiden, schlagen C.K. Chu, I. Glad, F. Godtliebsen and J.S. Marron (1996 - 98), ([9]), eine lokale Version vor. Die Funktion Φ_s hat im allgemeinen zahlreiche lokale Minima. Diese Tatsache nutzen die Autoren aus (sie galt lange Zeit als gravierender Nachteil der M-Schätzung).

Startend von der aktuellen Intensität Y_s im Pixel s wird als Schätzung ϑ^* das am nächsten bei Y_s liegende lokale Minimum von Φ_s gewählt; genauer ist ϑ^* die nächste lokale Minimumstelle von Φ_s links von Y_s, falls die Ableitung von Φ_s in Y_s positiv ist und die nächste rechts von Y_s sonst. Als Funktion φ wird eine negative Gaußfunktion φ_σ eingesetzt (vgl. Abb. 4.3, mittlere Zeile, linke Spalte) gegeben durch

$$g(u) = \exp(-u^2/2), \qquad \varphi_\sigma = -g(u/\sigma)/\sigma, \qquad \sigma > 0. \tag{4.4}$$

Schließlich werden die 'harten' Fenster $B(s)$ durch 'weiche' ersetzt, d.h. es wird eine glockenförmige Gewichtsfunktion φ_τ eingeführt, die weit entfernte Pixel nur schwach gewichtet. Insgesamt wird also die Funktion (4.3) durch

$$\Phi_s : \vartheta \longmapsto \sum_t \varphi_\sigma(Y_t - \vartheta) v_\tau(s - t) \tag{4.5}$$

ersetzt.

Dieser lokale *M-Glätter* erhält Kanten und Spikes; überdies 'glättet er über Schluchten hinweg', vgl. Fig. 4.1, obere Reihe. Das linke Bild zeigt eine verrauschte Treppenfunktion mit ihrer Restauration. Das rechte zeigt die Höhenlinien der Funktion $(s, \vartheta) \mapsto \Phi_s(\vartheta)$. Für jedes s auf der Abszisse started der Algorithmus in y_s und geht zum nächsten lokalen Minimum entlang der y-Richtung.

Diese Idee ist eng mit der *MAP*-Schätzung mit robuster a priori Verteilung verwandt. Der Hauptunterschied besteht in der lokalen anstelle der globalen Optimierung. Die Autoren führen auch eine asymptotische Analyse von Bias und Varianz durch (was umfangreiche Rechnung erfordert).

4.1.3 *W*-Schätzer

Wir identifizieren nun den lokalen M-Glätter mit einer lokalen Version des klassischen W-Schätzers aus der robusten Statistik. Zu diesem Zweck führen wir einige Begriffe aus der robusten Statistik ein.

M-Schätzer werden häufig als stationäre Punkte der M-Funktion Φ in (4.2) oder als Nullstellen ihrer Ableitung oder *Scorefunktion* Φ' definiert. Eine solche Definition schließt lokale Minima, Sattelpunkte und sogar lokale Maxima ein. Eigentlich sind nur Minima oder wenigstens lokale Minima von Φ von Interesse. Wir skizzieren nun, wie in der Praxis damit umgegangen wird. Vorher beschreiben wir den klassischen Rahmen.

Seien Zufallsvariablen $Y_1, \ldots, Y_n$ gegeben. Ein *W-Schätzer* ϑ^* ist dann nach ([15]) definiert durch

$$\vartheta^* = \frac{1}{\sum_i w(Y_i - \vartheta^*)} \sum_i w(Y_i - \vartheta^*) Y_i. \tag{4.6}$$

wobei die Gewichtsfunktion w positiv oder wenigsten nichtnegativ ist, so daß die Summe im Nenner nicht verschwindet. Üblicherweise werden unimodale symmetrische Funktionen mit Zentrum in 0 eingesetzt, z.B. gaußische Dichten oder Aproximationen davon mit kompaktem Träger. Für konstante Gewichte w definiert die Formel (4.6) einen gewöhnlichen linearen Filter, d.h. ein gewichtetes Mittel. Neu ist, daß die Gewichte von der Schätzung ϑ^* selbst abhängen.

Wegen der Normalisierung ist die rechte Seite eine konvexe Kombination der Daten und es gilt

$$(4.6) \Longleftrightarrow \sum_i w(Y_i - \vartheta^*) \cdot (Y_i - \vartheta^*) = 0 \Longleftrightarrow \sum_i \psi(Y_i - \vartheta^*) = 0, \tag{4.7}$$

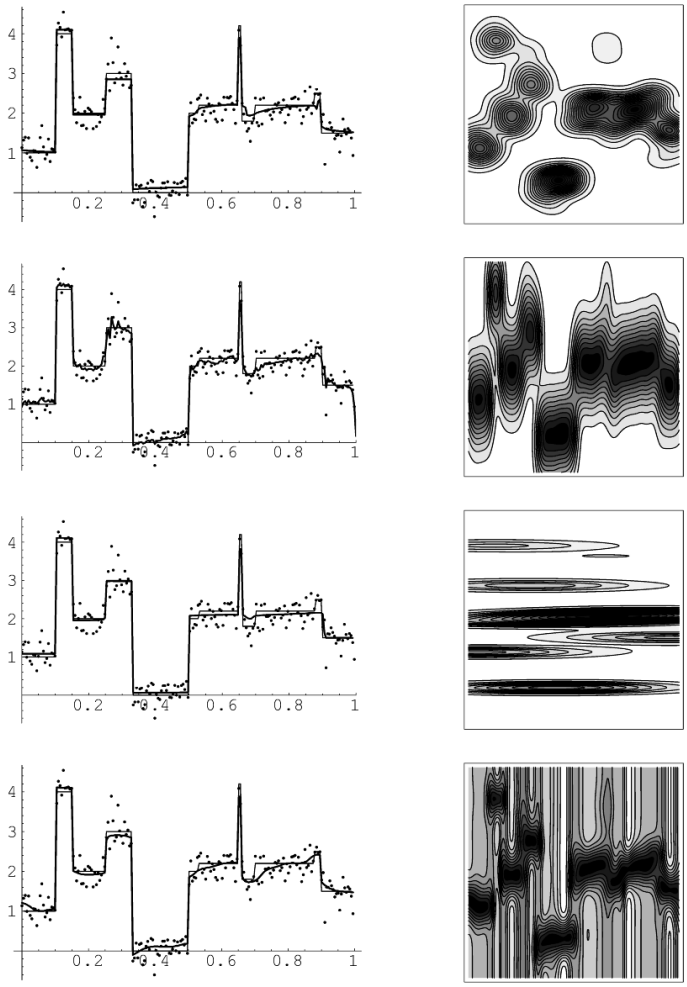


Abbildung 4.1: Linke Spalte: Eine Treppenfunktion (durchgezogene Linie), gaußisch verrauscht (Punkte). Die dicken Linien zeigen die Restaurationen mit (von oben nach unten): dem lokalen M-Glätter, dem nichtlinearen Gaußfilter, der Kette nichtlinearer Gaußfilters und einem radialen Basisfunktionsnetz, das hier nicht besprochen wird. Die rechte Spalte enthält Konturlinien der Funktionen Φ_s, (4.3), in der s-y-Ebene (erste und zweite Zeile).

wobei

$$\psi(u) = u \cdot w(u).$$

Sei $\psi = \varphi'$. Dann werden (4.6) und (4.7) impliziert von

$$\vartheta^* \text{ minimizes locally } \Phi : \vartheta \longmapsto \sum_i \varphi(Y_i - \vartheta). \tag{4.8}$$

Lösungen von (4.8) sind die lokalen M-Schätzer aus Abschnitt4.1.1 und (4.6) charakterisiert die stationären Punkte von Φ. Eine Standardmethode, um spezielle Lösungen von (4.6) zu berechnen ist der iterative Algorithmus, der durch

$$\vartheta_{k+1} = \frac{1}{\sum_i w(Y_i - \vartheta_k)} \sum_i w(Y_i - \vartheta_k) Y_i \tag{4.9}$$

gegeben ist. Da die lineare Kombination der Y_i auf der rechten Seite sogar eine Konvexkombination ist, kann die Vorschrift umformuliert werden zu

$$\begin{aligned} \vartheta_{k+1} &= \vartheta_k + \gamma_k \sum_i w(Y_i - \vartheta_k)(Y_i - \vartheta_k) \\ &= \vartheta_k + \gamma_k \sum_i \psi(Y_i - \vartheta_k) \\ &= \vartheta_k - \gamma_k \sum_i \varphi(Y_i - \cdot)'(\vartheta_k). \end{aligned} \tag{4.10}$$

Dabei nehmen wir an, daß daß γ_k die Summe der Koeffizienten zu 1 normiert. Wir lesen ab, daß obige Rekursion nichts anderes als ein klassischer (steilster) Gradientenabstieg ist. Startet man in $\vartheta_0 = y_j$, so konvergiert die Folge $(\vartheta_k)_k$ (hoffentlich) gegen das lokale Minimum der M-function Φ aus (4.8), welches talwärts am nächsten bei y_j liegt. Mit Hilfe dieses Algorithmus haben wir die Definition des W-Schätzers dahingehend präzisiert, daß er keine (isolierten) lokalen Maxima mehr liefern kann (außer er startet in einem solchen).

4.1.4 Zusammenfassung

Der W-Schätzer liefert jeweils angewendet auf Intensitätsdaten Y_t in einem Fenster $B(s)$ um ein Pixel s liefert genau die lokale M-Schätzung aus Abschnitt 4.1.2. Dabei wurde angenommen, daß der Gradientenabstieg gegen geeignete Werte konvergiert. Der einzige Unterschied ist, daß in unserer Formulierung der Optimierungsalgorithmus festgelegt wurde, während man

in der Formulierung in Abschnitt 4.1.2 beliebige Optimierungsalgorithmen eingesetzt werden können. (Diese Äquivalenz wurde unabhängig von D.G. Simpson, X. He and Y.-T. Liu, [28], und dem Autor beobachtet).

4.2 Nichtlineare Filter

In diesem Abschnitt führen wir nichtlineare Filter, insbesondere gaußische ein. Wir zeigen die enge Verwandtschaft zu einer Variante der sogenannten W-Schätzer auf und enthüllen so eine enge Beziehung zum lokalen M-Glaätter. Gleichzeitig werden die Rechnungen in Abschnitt 4.1.3 benutzt, um die Beziehung zwischen Topffunktionen φ in den M-Funktionen und den Gewichten nichtlinearer Filter aufzuzeigen.

4.2.1 *w*-Schätzer

w-Schätzer sind als das Ergebnis ϑ_1 des ersten Iterationsschrittes in (4.9) oder (4.10) definiert, cf. [15]:

$$\vartheta^* = \vartheta_1 = \frac{1}{\sum_i w^\sigma(Y_i - \vartheta_0)} \sum_i w^\sigma(Y_i - \vartheta_0)\vartheta_0. \tag{4.11}$$

Vor dem Hintergrund der obigen Diskussion, bewegt sich dieser Schätzer also von ϑ_0 aus ein Stück talwärts in Richtung des nächsten lokalen Minimums von Φ, d.h. also der lokalen M-Schätzung. Somit liegt er zwischen dem lokalen M-Schätzer und dem Startwert ϑ_0.

4.2.2 Der nichtlineare Gaußfilter

Um den w-Schätzer auf Signale oder Bilder zu übertragen, wenden wir ihn wieder lokal an. Um den gefilterten Wert der Intensität in einem Pixel s zu bestimmen, ersetzen wir die Zufallsvariablen Y_i im w-Schätzer durch zufällige Intensitäten Y_t in einem Fenster $B(s)$ um s. Der Startwert $\vartheta_0 = Y_s$ führt zu dem Filter

$$(\mathcal{F}Y)_s = \vartheta^* = \vartheta_1 = \frac{1}{\sum_t w^\sigma(Y_t - Y_s)} \sum_t w^\sigma(Y_t - Y_s)Y_t. \tag{4.12}$$

Dies ist bereits eine einfache Form des *σ-Filters*. Es ist üblich, das harte Fenster $B(s)$ wieder durch Gewichte zu ersetzen, die räumliche Entfernung

penalisieren. Dazu führt man eine zweite Gewichtsfunktion v ein - üblicherweise glockenförmig - und definiert:

$$(\mathcal{F}Y)_s = \vartheta^* = \frac{1}{\sum_t w^\sigma(Y_t - Y_s)v^\tau(t-s)} \sum_t w^\sigma(Y_t - Y_s)v^\tau(t-s)Y_t. \quad (4.13)$$

Die Parameter σ and τ messen die Bandweite von w and v. Sind z.B. w und v gaußische Dichten, so sind σ und τ die jeweiligen Standardabweichungen. (4.13) ist die allgemeine Form eines σ-*Filters*.

Wie schon erwähnt, unternimmt der Filter, startend in y_s einen Schritt talwärts in Richtung des nächsten lokalen Minimums von Φ_s, erreicht es jedoch im allgemeinen nicht. Somit liegt das Resultat zwischen den Daten und dem Ergebnis des lokalen M-Glätters. Dies erklärt, daß er glättet und Kanten weitgehend erhält. Andererseits erbt er eine gewisse Rauhheit von den Daten. Abb. 4.2 zeigt die Daten als Punkte, das Ergebnis der lokalen M-Glättung als glatte Line und das Resultat des Filters als zackige Linie dazwischen. Abb.

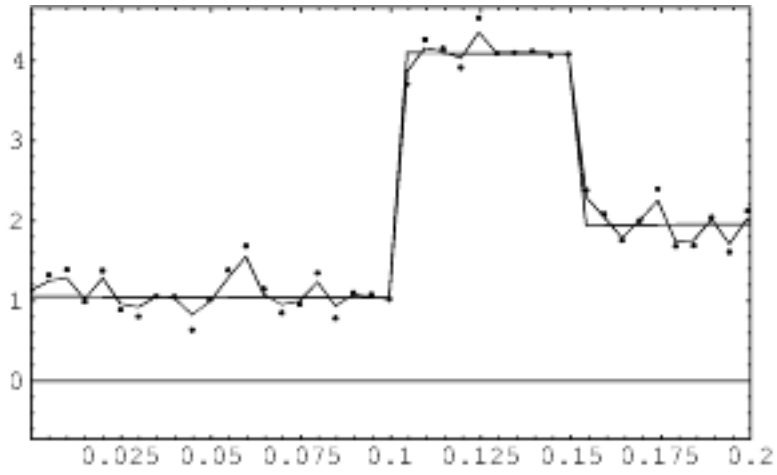


Abbildung 4.2: x-Achse: räumliche Skala; y-Achse: Intensitätsskala. Der σ-Filter (zackige Linie) bewegt sich ein Stück von den Daten y_s (Punkte) weg in Richtung auf das nächste lokale Minimum von Φ_s (glatte Linie)

4.1 zeigt die Anwendung auf die Standardtreppenfunktion. Die Höhenlinien sind dieselben wie für die M-Funktion mit den selben Gewichten w.

Beispiel 4.2.1 In Abschnitt 4.1.3 wurde eine Beziehung zwischen den topfförmigen Funktionen φ als Grundelemente der M-Funktionen, ihren Scorefunktionen ψ und gewissen Gewichten w aufgedeckt. Im selben Abschnitt und in Abschnitt 4.2.1 wurde klar, daß dies die Gewichte eines nichtlinearen Filters sind. Daraus ergibt sich eine Beziehung zwischen lokalen M-Glättern und dazugehörigen Filtern. Diese ist für verschiedene Typen von Funktionen φ in Abbildung 4.3 illustriert. Abgeschnittene Quadrate φ als M-Funktionen z.B. korrespondieren zu 'Truncated Means' als nichtlineare Filtergewichte w; negative Gaußdichten korrespondieren mit gaußischen Gewichten usw.

Sind w^σ und v^τ gaußisch, so wird (4.13) häufig *nichtlinearer Gaußfilter* genannt. Er wird z.B. in [13] und [31] untersucht.

Beispiel 4.2.2 Wir illustrieren die Wirkung des nichtlinearen Gaußfilters an einem Beispiel aus der biochemischen Pflanzenpathologie, [16]. Die Daten stammen aus einem Experiment, in dem Stress von Pflanzen mit Hilfe von photo-stimulierter Radioluminiszenz untersucht wurde. Abb. 4.4 zeigt verrauschte und geglättete Versionen eines 'Bildes' und einer einzelnen Zeile.

4.3 Ketten nichtlinearer Gaußfilter

Meist ist die Glättung durch eine einzige Anwendung des σ-Filters nicht ausreichend. In einer Reihe von Arbeiten studierten V. Aurich *et al* (1994 – 98), [3], [4], [23], [31], *Ketten* nichtlinearer Gaußfilter [1]. Ihre Wirkung ähnelt der des lokalen M-Glätters, sie arbeitet jedoch sehr schnell. Obenstehende Beobachtungen über den σ-Filter zeigen, daß eine enge Beziehung zum lokalen M-Glätter besteht.

Die Filterkette ist eine iterative Methode, deren einzelne Schritte gegeben sind durch

$$(\mathcal{F}Y)_s = \frac{1}{\sum_t w^\sigma(Y_t - Y_s)v^\tau(t-s)} \sum_t w^\sigma(Y_t - Y_s)v^\tau(t-s)Y_t$$

wobei $\sigma > 0$ und $\tau > 0$ Skalenparameter sind, z.B. Standardabweichungen falls v und w Gaußdichten sind. Während der Iteration werden σ und τ variiert. Formal ist die Filterkette definiert durch

$$\mathcal{F}^{\sigma_n,\tau_n} \circ \cdots \circ \mathcal{F}^{\sigma_1,\tau_1} Y. \tag{4.14}$$

[1] Source code under Aurich's web site

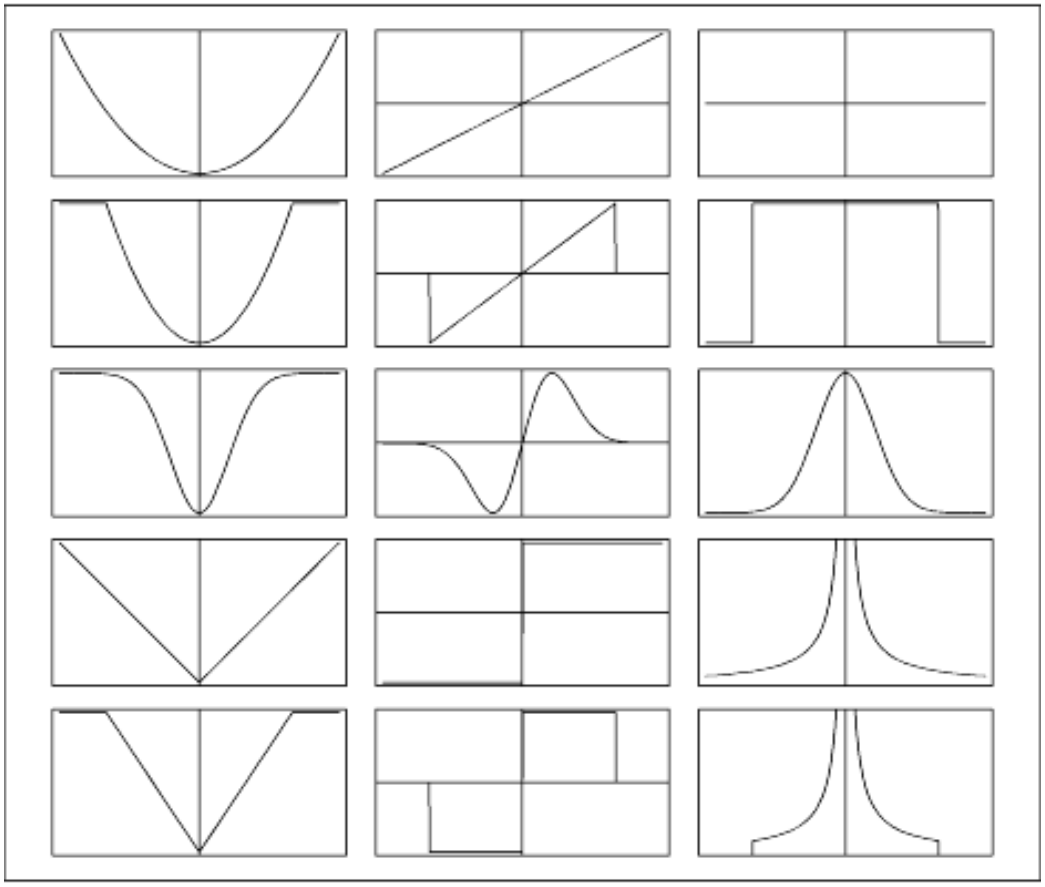


Abbildung 4.3: Einige Beispiele von Funktionen φ (linke Spalte), ihren Ableitungen ψ (mittlere Spalte) und den dazugehörigen Filtergewichten w (rechte Spalte). Von oben nach unten: φ die Quadratfunktion, die abgeschnittene Quadratfunktion, die negative Gaußfunktion, die Betragsfunktion und die abgeschnittene Betragsfunktion.

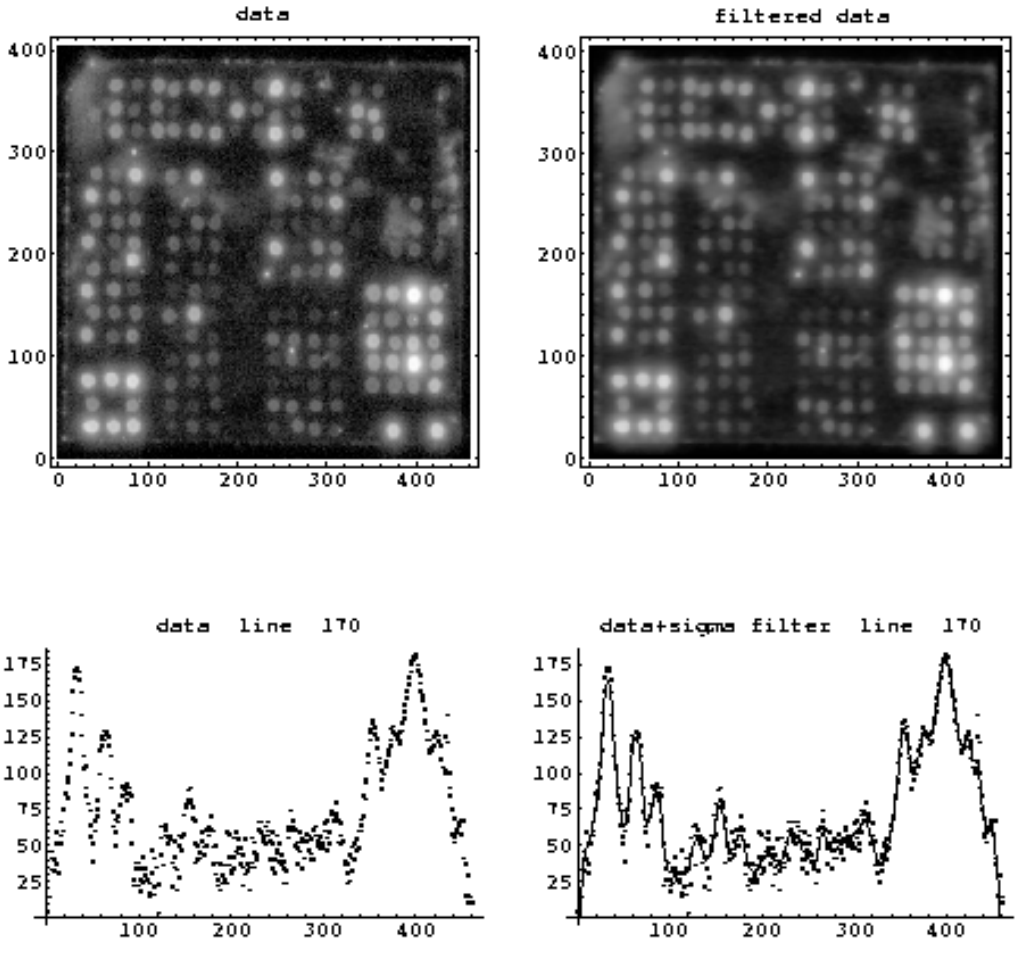


Abbildung 4.4: Daten und Restauration; Intensitätsprofil von Zeile 170

Der erste Filterschritt reduziert hauptsächlich den Kontrast in feinen Details: Es wird ein sehr kleiner räumlicher Skalenparameter τ_1 (entsprechend einem sehr kleinen Fenster) verwendet, um grobe Strukturen nicht zu verwischen. Andererseits sollte dieses Fenster von der Größe feiner Strukturen sein. Der Parameter auf der Intensitätsskala wird breit gewählt; im wesentlichen bestimmt er den maximalen Kontrast in der Feinstruktur, der schließlich eliminiert werden soll. Die folgenden Schritte bewirken weitere Reduktion des Rauschens; gleichzeitig können sie Kanten gröberer Strukturen schärfen, die in vorhergehenden Schritten verwischt wurden. Um dies zu erreichen, werden die räumlichen Skalenparameter τ vergrößert, während die Parameter σ für die Intensitätsskala vermindert werden.

In einem zusätzlichen letzten Schritt können die so erhaltenen - letzten - Gewichte noch einmal auf die Ursprungsdaten angewendet werden, um eine befriedigende Rekonstruktion zu erreichen, die weniger die Gestalt einer Treppenfunktion hat. Entrauschen und Kantenschärfung ist in Abb. 4.5 illustriert. Die Größe der Skalenparameter ist durch Fenster in der Raum-Intensitätsebene symbolisiert.

Diese Filterkette liegt der Kantenfindung in den Bildern in Abb. 1.1, 1.2, 1.3 und 1.4 in der Einleitung 1 zugrunde.

Abb. 4.6 zeigt Glättung und Kantenschärfung über einen einzelnen Sprung. Die Anzahl n von Iterationen ist klein: die Wahl $n = 4$ gibt in den meisten Anwendungen gute Resultate. die optimale Wahl der Skalenparameter wird in [23] diskutiert; sie hängt von der Dimension der Daten ab. In einer Dimension und $n = 4$ wird $\sigma_k = \sigma_{k-1}/2$ and $\tau_k = 4\tau_{k-1}$ vorgeschlagen.

Der Algorithmus ist sehr schnell und robust. In vielen Fällen liefert er ausgezeichnete Resultate. *Die rigorose mathematische Analyse ist schwierig und nur in Ansätzen durchgeführt.* Der Grund dafür ist derselbe wie für viele andere iterative Verfahren: Die Data werden in jedem Schritt nichtlinear transformiert und deshalb gehen statistische Eigenschaften wie Unabhängigkeit oder gute Eigenschaften der Verteilungen sofort verloren. Diese Probleme haben wir schon bei der Analyse des Medianfilters erkannt.

Beispiel 4.3.1 Der Algorithmus wurde auf einen Stapel von Bildern, d.h. einen Film angewandt. Die Daten sind also Elemente von $\mathbb{R}^2 \times \mathbb{R}_+$. Abb. 4.7 zeigt ein Frame mit einem einzelnen dunklen Fleck vor Hintergrund. Im Film bewegt sich der Fleck über die Fläche. Abb.4.8 zeigt zwei dunkle Flecken vor einem Hintergrund, von denen im Laufe der Zeit sich ein Fleck diagonal über die Fläche bewegt, während der andere pulsiert. Aus einem Schnappschuß

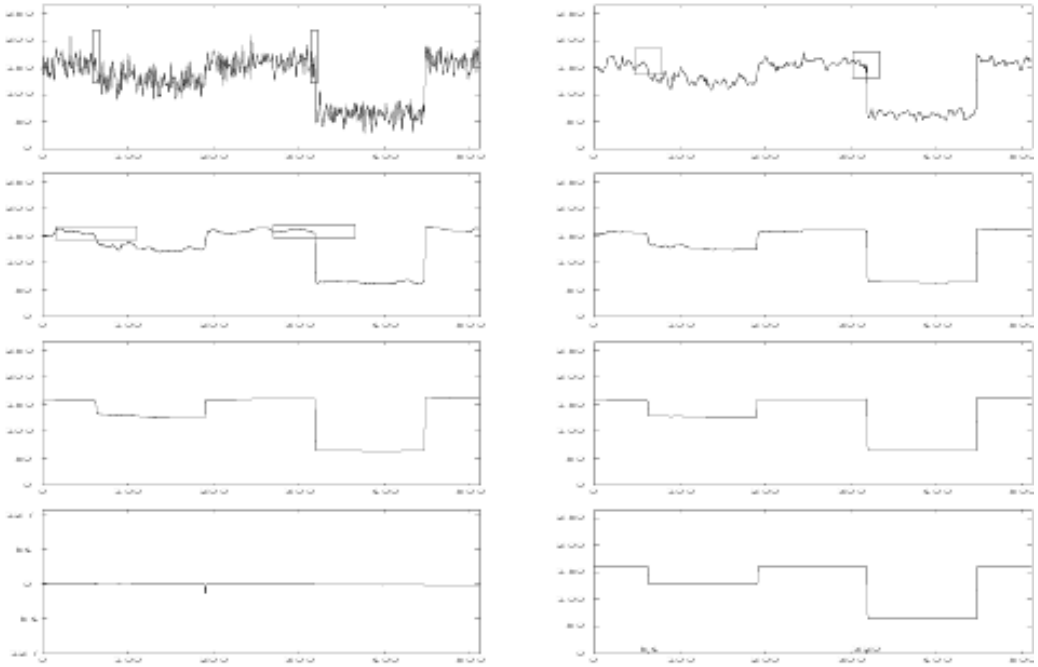


Abbildung 4.5: Verrauschtes Signal gefiltert mit sukzessiven Schritten der nichtlinearen gaußischen Filterkette. Von oben links nach unten rechts: Die Daten und Skalenparameter des ersten Schrittes; Ergebnis des ersten Schrittes; Ergebnis des zweiten Schrittes; Ergebnis des dritten Schrittes; Ergebnis des modifizierten vierten Schrittes; Ergebnis des modifizierten fünften Schrittes; Differenz zum Originalbild; Originalbild (aus [23])

oder Frame sind die Objekte in den verrauschten Bildern vom menschlichen Auge so gut wie nicht zu erkennen. Die Filterkette wird jedoch dreidimensional verwendet. So kann sie Information aus der Vergangenheit nutzen, um die einzelnen Frames zu restaurieren. Anschließend werden die Objekte detektiert und markiert. Abb. 4.9 gibt einen dreidimensionalen Eindruck von Original und Restauration.

Bemerkung 4.3.2 Bekanntlich entsprechen lineare Gaußfilter den Lösungen der Wärmeleitungsgleichung

$$\frac{\partial u(t,x)}{\partial t} = a^2 \triangle u(t,x).$$

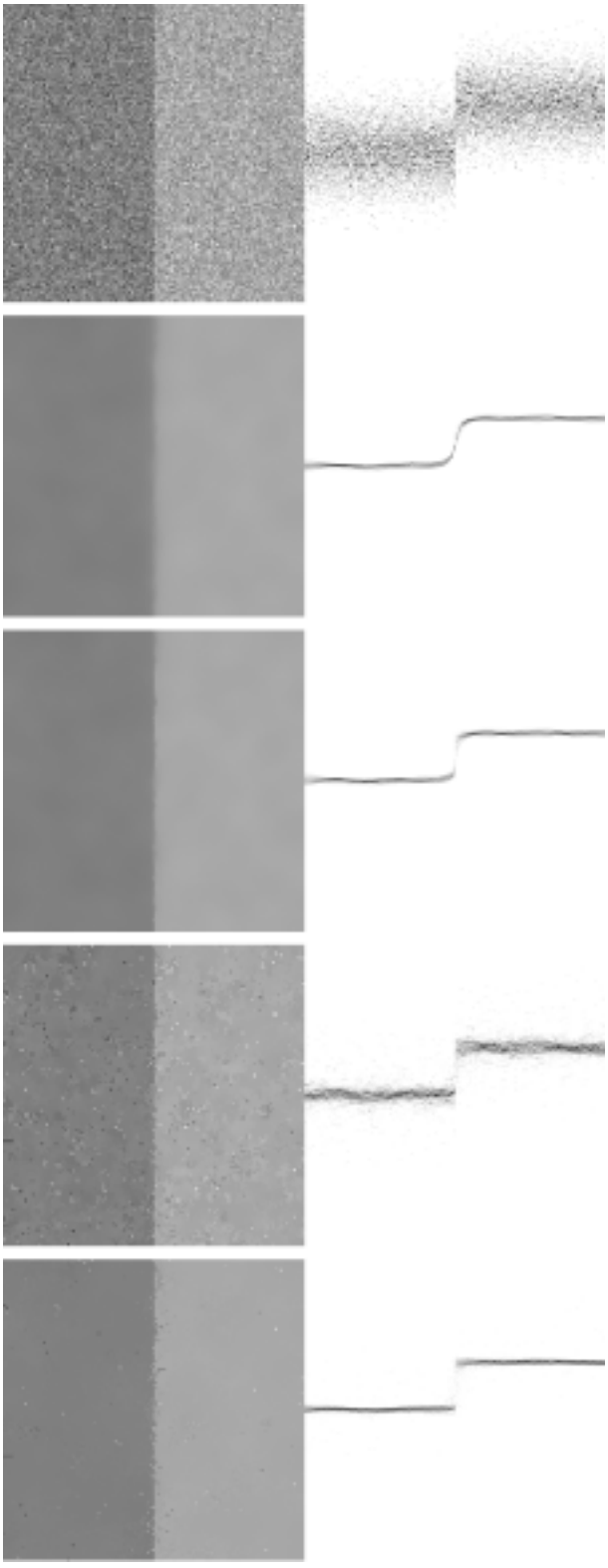


Abbildung 4.6: Glättung und Kantenschärfung über einen einzelnen Sprung durch die Aurichsche Kette nichtlinearer Gaußfilter (aus [31]).

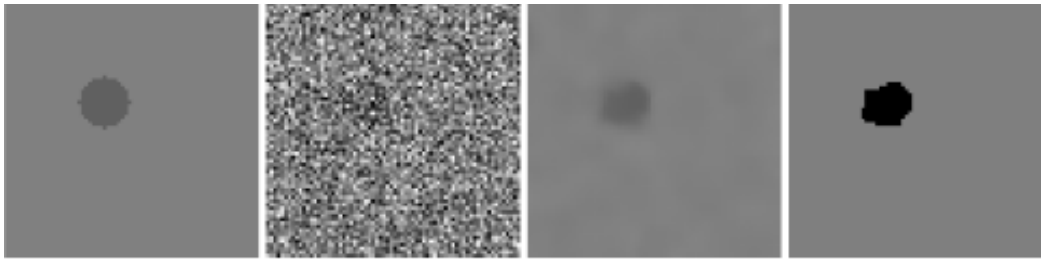


Abbildung 4.7: Eine Momentaufnahme aus dem Film 'Laufender Fleck', verrauscht, restauriert und markiert, Bild: V. Aurich, Düsseldorf

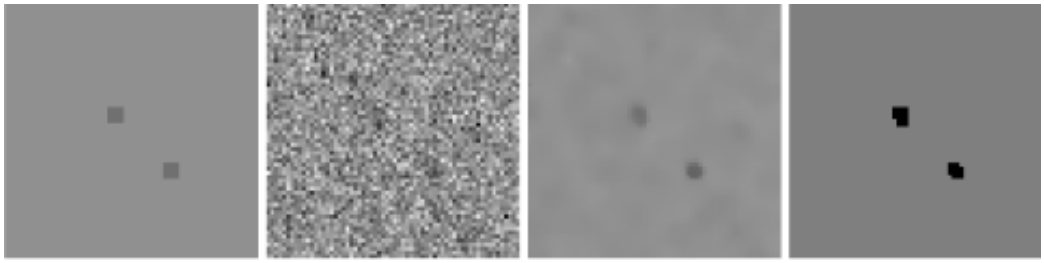


Abbildung 4.8: Eine Momentaufnahme aus dem Film 'Laufender und oszillierender Fleck', verrauscht, restauriert und markiert, Bild: V. Aurich, Düsseldorf

Analog sollte eine solche Entsprechung zwischen nichtlinearen Gaußfiltern und der sogenannten anisotropen Diffusion

$$\frac{\partial u(t,x)}{\partial t} = a^2 \text{div } D(\text{grad} u(t,x)) \text{ grad } u(t,x)$$

existieren, die zu einer anisotropen Wärmeleitungsgleichung gehört, existieren. Diese zu etablieren ist Gegenstand aktueller Forschung. Die Rolle der anisotropen Diffusion in der Bildanalyse wird in Weickert (1998), [30], diskutiert.

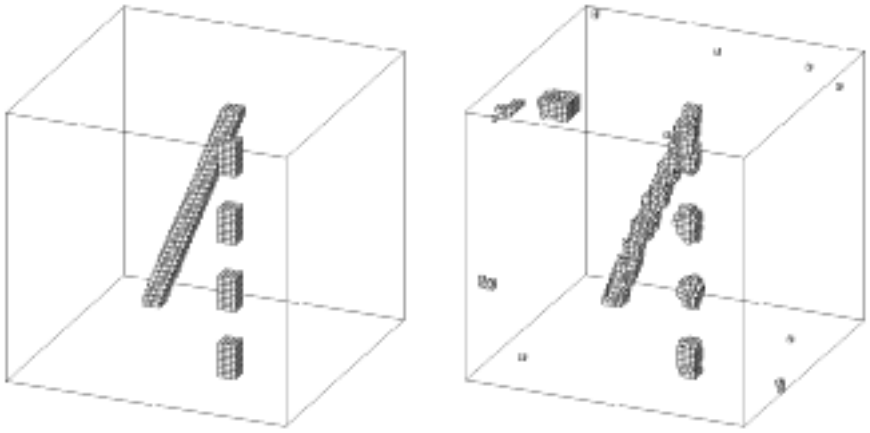


Abbildung 4.9: Dreidimensionale Darstellung des Filmes aus Abb. 4.8, Bild V. Aurich, Düsseldorf

4.4 Adaptive-Gewichte-Glättung

Diese Methode wird in einer Reihe von Arbeiten von J. Polzehl und V.G. Spokoiny (1998), [26], [25], [24] vorgeschlagen. Wir beschränken uns auf die Darstellung der in [26] und [25] vorgestellten Methode. Diese iterative Methode ist formal ähnlich wie die Aurichsche Kette von nichtlinearen Gaußfiltern (4.14). Der wesentliche Unterschied ist, daß die Größen, welche den Parametern σ_k and τ_k in (4.14) entsprechen, lokal aus den Daten geschätzt werden. Dies liefert eine *adaptive* Methode. Überdies werden die Filtergewichte direkt aus den Daten geschätzt und nicht aus transformierten Daten wie in (4.14).

4.4.1 Der Algorithmus

Zunächst werden einige Parameter fixiert: For jeden Designpunkt[2] s wird eine aufsteigende Folge von Fenstern U_s^k um s, welches $n_s^{(k)}$ Pixel enthält, gewählt. es wird vorausgesetzt, daß konsistente Schätzungen V_s der unbekannten Varianzen der Variablen Y_s vorliegen. Überdies seien Parameters $\lambda > 0$ und

[2]Designpunkte entsprechen in unserer Sprache den Stellen oder Pixeln; das Wort Design kommt aus der Theorie der Regression

$\eta > 0$ gewählt. Der Algorithmus wird mit $k = 0$ und

$$Y_s^{(0)} = \frac{1}{n_s^{(0)}} \sum_{j \in U_s^0} Y_t, \;\; V_s^{(0)} = \frac{1}{{n_s^{(0)}}^2} \sum_{j \in U_s^0} V_s$$

initialisiert. In den folgenden Schritten für $k > 0$ und $s \in S$ werden Filtergewichte gemäß

$$w^{(k)}(s;t) = g\left(\frac{Y_s^{(k-1)} - Y_t^{(k-1)}}{\lambda V_s^{(k-1)}}\right)$$

für alle Punkte $t \in U_s^k$ berechnet. Der Kern g ist glockenförmig, ähnlich wie (4.4). Im Gegensatz zu den bisher betrachteten Beispielen von Filtern sind die Gewichte hier nicht symmetrisch. Neue Schätzungen von Y werden berechnet gemäß

$$Y_s^{(k)} = \frac{1}{\sum_{t \in U_s^k} w^{(k)}(s;t)} \sum_{t \in U_s^k} w^{(k)}(s;t) Y_t,$$

und neue Schätzungen von V gemäß

$$V_s^{(k)} = \frac{1}{\left(\sum_{t \in U_s^k} w^{(k)}(s;t)\right)^2} \sum_{t \in U_s^k} w^{(k)}(s;t)^2 V_s.$$

Das Resultat wird in der folgenden Weise kontrolliert: Sei

$$\mathcal{K} = \{0, 1, 2, 4, \ldots, 2^l, \ldots\}.$$

Für jedes $l \in \mathcal{K}$, $l < k$, wird geprüft, ob

$$|Y_s^{(k)} - Y_s^{(l)}| > \eta \sqrt{V_s^{(l)}}.$$

Ist diese Ungleichung für ein solches l erfüllt, dann werden die bisherigen Schätzungen verworfen und das vorhergehende $Y_s^{(k-1)}$ und $V_s^{(k-1)}$ behalten. Der Algorithmus terminiert, falls entweder k eine vorher gewählte Schranke k_* übersteigt oder wenn $Y^{(k)} = Y^{(k-1)}$. Wir bezeichnen mit $(\hat{Y}_s)_{s \in S}$ die schließlich resultierende Schätzung.

Der Kontrollschritt verhindert, daß der Algorithmus früher entdeckte Unstetigkeiten verliert. Er ist notwendig, weil in jedem Schritt die Originaldaten verarbeitet werden. Er hat allerdings einige praktische Nachteile. Die

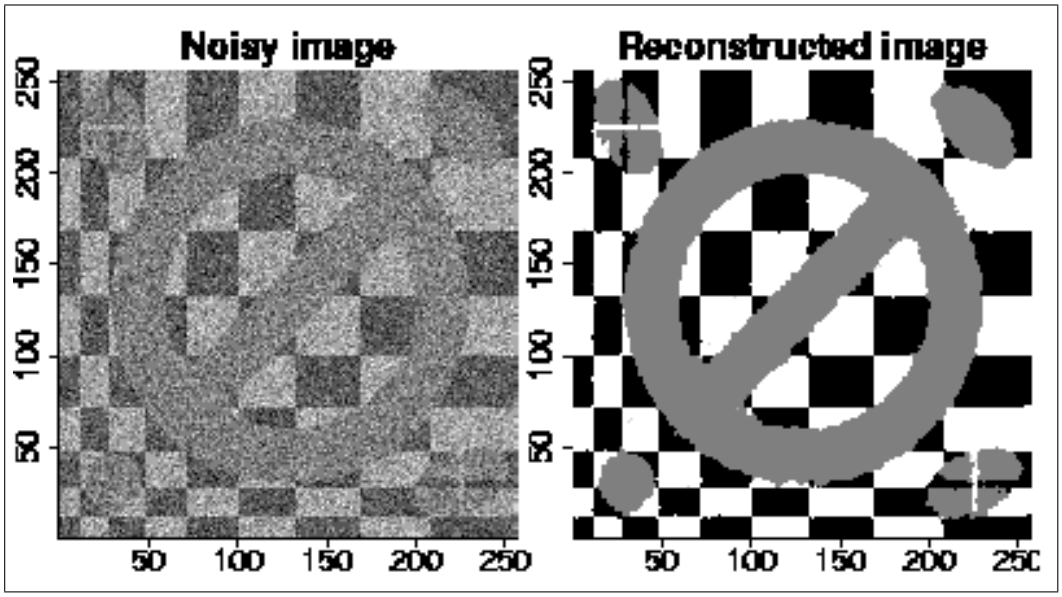


Abbildung 4.10: Grauwerte im Original 0, 0,5 und 1; Standardabweichung des Rauschens 0,4. Aus [25], J.R. Statist. Soc. Ser.B, **62**:2, 2000, 335–354, mit freundlicher Erlaubnis von J. Polzehl, Berlin, V. Spokoiny, Würzburg.

Wahl von $\mathcal{K} = \{0, 1, 2, 4, \ldots, 2^l, \ldots\}$ ist mehr oder weniger pragmatisch und nicht theoretisch untermauert. Eigentlich währe $\mathcal{K} = \{0, 1, 2, 3, 4, 5, \ldots\}$ die natürliche Wahl. Jedoch sogar im ersten Fall ist der Kontrollschritt für kleinere k zeitaufwendig.

Die Resultate dieser Methode werden allgemein als sehr gut eingeschätzt.

Beispiel 4.4.1 Die Abbildungen 4.10, 4.11 und 4.12 illustrieren die hohe Qualität dieses lokal adaptiven Glätters.

Abbbildung 4.10 basiert auf dem Testbild aus J. Polzehl und V. Spokoiny (2000), [25]. Die Grauwerte im Original sind 0, 0,5 und 1; die Standardabweichung des Rauschens ist 0,4. Für die Rekonstruktion des Bildes mittels adaptiver Gewichte Glättung war der maximale Umgebungsradius 80 Pixel, die maximale Umgebungsgröße $n^{(k_*)}$ betrug etwa 20000 Pixel.

Abbildung 4.11 zeigt Original SAR Bilder (*S*ynthetic *A*perture *R*adar) mit log-transformiertem C-Band und HH-Polarisation, aufgenommen von E. Attema vom European Space Research and Technology Centre, Noordwijk,

Niederlande. Es handelt sich um ein Gebiet nahe Thetford Forest, England. Die Daten sind unter

`ftp://peipa.essex.ac.uk/ipa/pix/books/glasbey-horgan/`

erhältlich. Die adaptive-Gewichte-Rekonstruktion hatte eine maximale Umgebungsgröße $n^{(k_*)}$ von 149 Pixeln. Es werden das verrauschte Bild und die Rekonstruktion zusammen mit den Residuen gezeigt. Die Abbildung entstammt der Arbeit [25], J.R. Statist. Soc. Ser. B, **62**:2, 2000, 335–354, von J. Polzehl, Berlin, und V. Spokoiny, Würzburg.

Abbildung 4.12 zeigt den mittleren Teil eines MRT-Bildes (*M*agnet *R*esonanz *T*omographie), die adaptive-Gewichte-Rekonstruktion und den Logarithmus der Varianzreduktion in jedem Pixel. Die Varianzreduktion in jedem Pixel ist umgekehrt proportional zur effektiven Anzahl der zur Schätzung herangezogenen Punkte. Die Bilder stammen von F. Godtliebsen, University of Tromsø.

4.4.2 Die Parameter

Die Wahl der Parameter ist vom empirischen Standpunkt aus kritisch. Ähnlich wie bei Aurich wachsen die Größen der Nachbarschaften $U_s^{(k)}$ exponentiell, typisch wie $n_s^{(k)} \propto 2^k$. Es werden z.B. die 2^k Punkte, die s in der euklidischen Distanz am nächsten liegen oder solche in wachsenden Kugeln gewählt. Die Wahl von $n_s^{(0)}$ ist ebenfalls von Bedeutung. Sei

$$c = \min\{|a-b| : a, b \text{ mögliche Intensitäten}\}$$

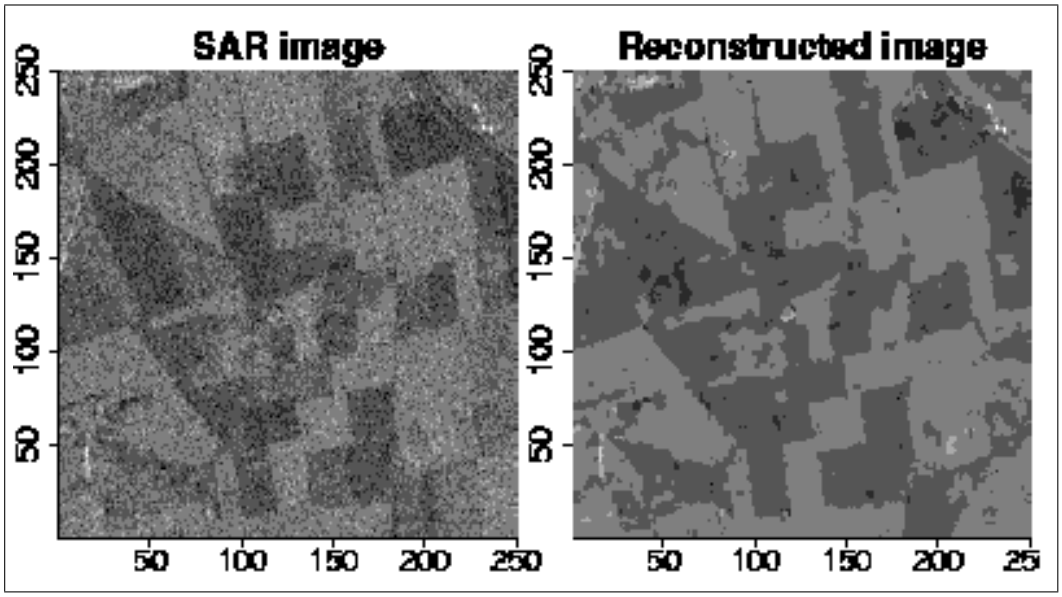
der *Bildkontrast*. Für großes Signal zu Rausch-Verhältnis $c/\sigma > 2$ empfehlen die Autoren $n_s^{(0)} = 1$; für $1 \leq c/\sigma \leq 2$, scheint ein Wert $n_s^{(0)} = 5$ und für $c/\sigma < 1$ ein $n_s^{(0)} \geq 9$ praktikabel.

Ferner empfehlen die Autoren Werte $3 \leq \eta \leq 4$ und $2.5 \leq \lambda \leq 3$. Der Parameter λ kontrolliert den Fehler erster Art, d.h. für großes λ wird die Wahrscheinlichkeit, artifizielle Sprünge zu detektieren, reduziert.

4.4.3 Rigorose Resultate

Die Autoren beweisen einige rigorose Resultate, cf. [25]. Der Kern g ist dabei rechteckig, d.h. von der Gestalt

$$g = \chi_{\{y \leq 1\}}.$$



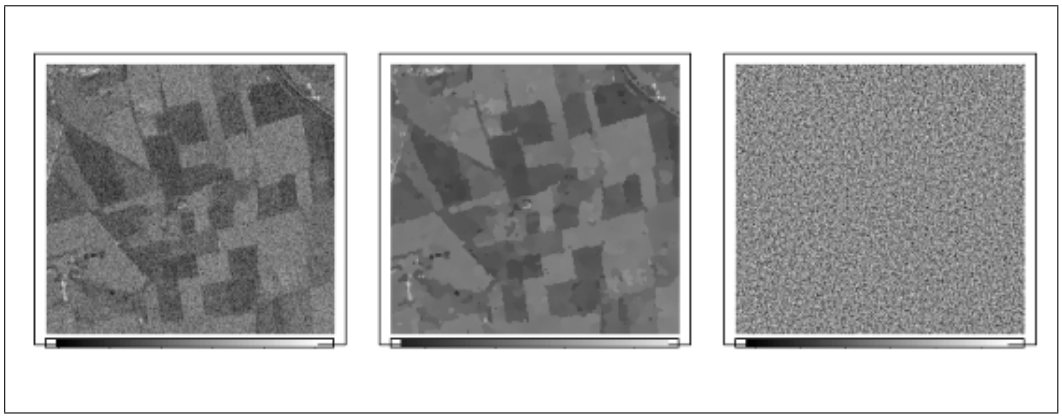


Abbildung 4.11: SAR-Bilder aufgenommen von E. Attema vom European Space Research and Technology Centre, Noordwijk, Niederlande. Es handelt sich um ein Gebiet nahe Thetford Forest, England. Daten sind unter `ftp://peipa.essex.ac.uk/ipa/pix/books/glasbey-horgan/` erhältlich. Verrauschtes Original, Rekonstruktion und Residuen. Aus [25], J.R. Statist. Soc. Ser.B, **62**:2, 2000, 335–354, mit freundlicher Erlaubnis von J. Polzehl, Berlin, und V. Spokoiny, Würzburg.

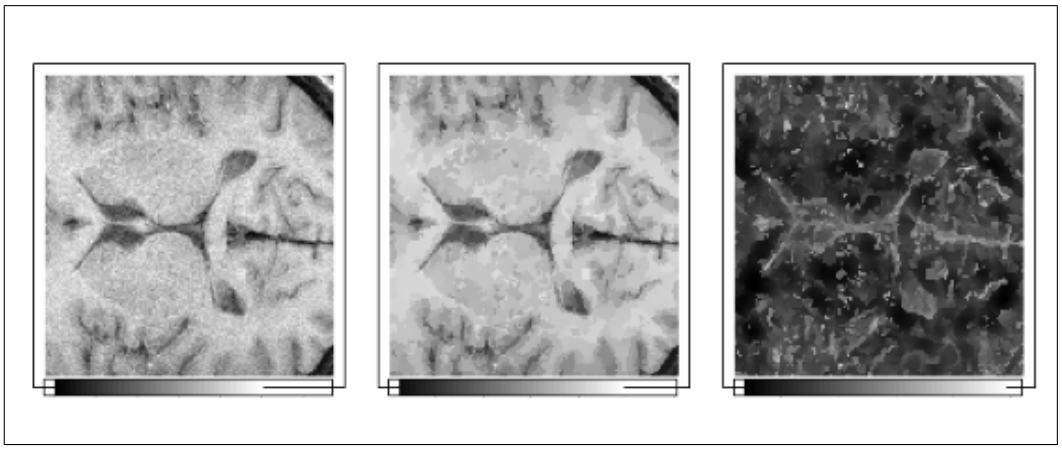


Abbildung 4.12: Mittlerer Teil eines MRT-Bildes, die adaptive-Gewichte-Rekonstruktion und der Logarithmus der Varianzreduktion in jedem Pixel. Mit freundlicher Erlaubnis von F. GODTLIEBSEN, University of Tromsø.

Für diesen Kern sind alle nichtverschwindenden Gewichte gleich.

Wie bei unserer Behandlung des Medianfilters betrifft das erste Resultat die verrauschte Fläche. Die Rauschvariablen Y_s sind i.i.d. gaußisch mit Erwartungswert a; das Modell hat also die Gestalt

$$Y_s = a + \eta_s, s \in S. \tag{4.15}$$

Es wird gezeigt, daß mit hoher Wahrscheinlichkeit die Schätzung $(\hat{Y}_s)$ eine Konstante ist und daß die Abweichungen $\hat{Y}_s - a$ von der Ordnung $n^{-1/2}$ sind.

Satz 4.4.2 *Sei in (4.15) die Größe a eine reelle Zahl und seien $\eta_s, s \in S$, unabhängige und identisch verteilte gaußische Zufallsvariablen mit Erwartungswert 0. Seien ferner*

$$g = \chi_{\{y \le 1\}}, \;\; n_s^{(k)} = n^{(k)}$$

für alle

$$s \in S, n_s^{(k_*)} = |S|,$$

und C eine reelle Konstante mit

$$n^{(1)} + \cdots + n^{(k_*)} \le C \cdot |S|$$

und

$$\lambda^2 \geq (2+\delta)\log|S|$$

mit $\delta > 0$. *Dann gilt*

$$\begin{aligned} & \mathbb{P}\left(w^{(k)}(s;t) = 0 \text{ für ein } k \leq k_*, s \neq t\right) \\ \leq\ & n^2 C/2 \exp(-\lambda^2/4) + k_* n \log_2(2k_*) \exp(-\eta^2/2). \end{aligned}$$

Die Schranke auf der rechten Seite von (4.16) ist klein, falls $\lambda^2 \geq (8+\delta)\log n$ und $\eta^2 \geq (2+\delta)\log n$ mit einer Konstante δ. Da im letzten Schritt (mit dem Index k_*) alle Pixel vom Fenster $U_s^{k_*}$ überdeckt werden, bedeutet das, daß mit hoher Wahrscheinlichkeit alle Schätzwerte $\hat{Y}_s$ mit dem Mittel aller beobachteten Daten Y_s übereinstimmen.

Die nächste Aussage betrifft stückweise konstante Regression, d.h. den Fall von Kanten. Wir nehmen der Einfachkeit halber das Modell

$$Y_s = (a \cdot \chi_A(s) + b \cdot \chi_B(s)) + \eta_s, \tag{4.16}$$

an, wobei A und B eine Partition von S in zwei disjunkte Gebiete bilden. (η_s) sei wieder ein weißes (zentriertes) gaußisches Rauschen, d.h. $Y_s \sim \mathcal{N}(0, \sigma^2)$. Das nächste Resultat sichert, daß zwischen Pixeln, im Inneren von A bzw. B, keine Mittelung stattfindet wenn der Kontrast hoch genug ist.

Wir fahren mit der Notation von Theorem 4.4.2 fort. Der Kontrast ist einfach $c = |a-b|$. Das Innere von $A \subset S$ ist die Menge der Pixel s für die A eine Umgebung enthält:

$$A^\circ = \{s \in S : U_s^0 \subset A\}.$$

Satz 4.4.3 *Sei* Y *durch (4.16) gegeben. Dann gilt*

$$w^{(k)}(s;t) = 0 \text{ für alle } s \in A^\circ, t \in B^\circ, k \leq k_*,$$

mit Wahrscheinlichkeit größer oder gleich

$$1 - \frac{1}{2} \cdot C \cdot |S|^2 \cdot \exp\left(-\frac{1}{4}\left(\frac{\sqrt{n^{(0)}}}{\sigma} \cdot |a-b| - 2\eta\right)^2\right).$$

Sind η^2 von der Größenordnung $(2+\delta)\log n$ und

$$\sigma^{-1}\sqrt{n^{(0)}}|a-b| > 4\eta$$

dann ist die Wahrscheinlichkeit im Satz durch $|S|^2 \exp(-\eta^2)$ beschränkt, was für große $|S|$ klein ist. Dadurch ist die Mißklassifikationsrate im Inneren der Gebiete klein. Fehler treten vor allem in der Nähe der Ränder auf. Dies gibt natürlich bei sehr unregelmäßigen Rändern Probleme.

Die Beweise sind einfach; sie beruhen auf Standardwissen über Gaußverteilungen. Der Grund für diese Einfachheit ist, daß in jeden Schritt die Originaldaten Y_s eingehen und somit die einfachen Eigenschaften des Rauschens erhalten bleiben. Dies steht im strikten Gegensatz zu Aurich's Filterkette, bei der die Daten sukzessive transformiert und so die angenehmen Eigenschaften des Rauschens (schon im ersten Filterschritt) zerstört werden. Andererseits bedingt dies den Kontrollschritt. Er spielt die eigentliche Rolle in den Beweisen. Es zeigt sich, daß er auch für das praktische Funktionieren des Verfahrens wesentlich ist.

Noch ein mal: Hier liegt der entscheidende Unterschied zu Aurichs Filterkette, da sie das Ergebnis eine Schrittes im nächsten weiter transformiert. Sie arbeitet zwar in der Praxis gut, ist aber aus den oben genannten Gründen einer rigorosen theoretischen Behandlung nur schwer zugänglich.

Literaturverzeichnis

[1] H. Akaike. A new look at the statistical model idenfication. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] V. Aurich and U. Daub. Bilddatenkompression mit geplanten verlusten und hoher rate. In *Mustererkennung 1996, Proceedings of the DAGM*, pages 138–146, 1996.

[3] V. Aurich, E. Mühlhaus, and S. Grundmann. Kantenerhaltende Glättung von Volumendaten bei sehr geringem Signal-Rausch-Verhältnis. In *Zweiter Aachener Workshop über Bildverarbeitung in der Medizin*, 1998.

[4] V. Aurich and J. Weule. Non-linear gaussian filters performing edge preserving diffusion. In *Proceed. 17. DAGM-Symposium, Bielefeld*, pages 538–545. Springer, 1995.

[5] A. Blake and A. Zisserman. *Visual Reconstruction*. The MIT Press Series in Artificial Intelligence. MIT Press, Massachusetts, USA, 1987.

[6] P. Bloomfield and W.L. Steiger. *Least Absolute Deviations. Theory, Applications, and Algorithms*, volume 6 of *Progress in probability and Statistics*. Birkhäuser, Boston and Basel and Stuttgart, 1983.

[7] J.E. Cavanaugh. Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters*, 33:201–208, 1997.

[8] B. Chalmond. *Eléments de modélisation pour l'analyse d'images*, volume 33 of *Mathématique & Applications*. Springer Verlag, Paris, Berlin, 2000.

[9] C.K. Chu, I. Glad, F. Godtliebsen, and J.S. Marron. Edge-preserving smoothers for image processing. *JASA*, 93(442):526–541, 1998.

[10] U. Daub. *Wieviel Information enthalten Helligheitskanten? Rekonstruktion von Bildern aus Bildmerkmalen*. PhD thesis, Heinrich-Heine-Universität Düsseldorf, 1995.

[11] P.L. Davies and A. Kovac. Local extremes, runs, strings and multresolution. *The Annals of Statistics*, 2000. To appear.

[12] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEE Trans. PAMI*, 6:721–741, 1984.

[13] F. Godtliebsen, E. Spjøtvoll, and J.S. Marron. A nonlinear Gaussian filter applied to images with discontinuities. *J. Nonparametr. Statist.*, 8:21–43, 1997.

[14] X. Guyon. *Random Fields on a Network. Modelling, Statistics, and Applications*. Probability and its Applications. Springer Verlag, New York, Berlin, Heidelberg, 1995.

[15] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust Statistics*. Wiley Series in Probability and Mathematical Statistics. Wiley & Sons, New York, 1986.

[16] B. Heidenreich, 1998. personal communication.

[17] P.J. Huber. *Robust Statistics*. Wiley Series in Probability and Mathematical Statistics. Wiley & Sons, New York, 1981.

[18] B.I. Justusson. Order statistics on stationary processes, with applications to moving medians. Technical Report TRITA-MAT-1, Royal Institute of Technology, Stockholm, 1979.

[19] B.I. Justusson. Median filtering: Statistical properties. In *Two-Dimensional Digital Signal Processing II. Transforms and Median Filters*, volume 43 of *Topics in Applied Physics*, chapter 5, pages 161–196. Springer Verlag, Berlin and Heidelberg and New York, 1981.

[20] A. Kempe. *Statistische Verfahren in der Bildanalyse*. PhD thesis, Institut für Biomathematik und Biometrie an der Gesellschaft für Umwelt und Gesundheit, München-Neuherberg, 2002.

[21] E. Mammen and S. van de Geer. Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413, 1997.

[22] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[23] E. Mühlhaus. *Die sprungerhaltende Glättung verrauschter, harmonischer Schwingungen*. PhD thesis, Heinrich-Heine-Universität Düsseldorf, 1998.

[24] J. Polzehl and V.G. Spokoiny. Image Denoising: Pointwise Adaptive Approach. Discussion Paper 38, Humbold-Universität zu Berlin, Sonderforschungsbereich 373: Quantifikation und Simulation ökonomischer Prozesse, Berlin, 1998.

[25] J. Polzehl and V.G. Spokoiny. Adaptive weights smoothing with applications to image restoration. *Journal of the Royal Statistical Society*, Ser. B, **62**(2):335–354, 2000.

[26] J. Polzehl and V.G. Spokoiny. Functional and dynamic magnet resonance imaging using vector adaptive weights smoothing. Preprint, October 2000.

[27] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[28] D.G. Simpson, X. He, and Y.-T. Liu. Comment on: C.K. Chu, I. Glad, F. Godtliebsen and J.S. Marron: Edge-preserving smoothers for image processing. *JASA*, 93(442):544–548, 1998.

[29] S.G. Tyan. Median filtering: Deterministic properties. In *Two-Dimensional Digital Signal Processing II. Transforms and Median Filters*, volume 43 of *Topics in Applied Physics*, chapter 6, pages 197–218. Springer Verlag, Berlin and Heidelberg and New York, 1981.

[30] J. Weickert. *Anisotropic Diffusion in Image Processing*. B.G. Teubner, Stuttgart, 1998.

[31] J. Weule. *Iteration nichtlinearer Gauß-Filter in der Bildverarbeitung*. PhD thesis, Heinrich-Heine-Universität Düsseldorf, 1994.

[32] G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*, volume 27 of *Applications of Mathematics*. Springer Verlag, Berlin, Heidelberg, New York, 1995.

[33] G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*. Beijing World Publishing Corporation, Beijing Peoples Republic of China, 1999. Reprint of 'Image Analysis, Random Fields and Dynamic Monte Carlo Methods', Springer Verlag, 1995.

[34] G. Winkler, V. Aurich, K. Hahn, A. Martin, and K. Rodenacker. Noise reduction in images: Some recent edge-preserving methods. *Pattern Recognition and Image Analysis. Advances in Mathematical Theory and Applications*, 9(4):749–766, 1999.

[35] G. Winkler and V. Liebscher. Smoothers for discontinuous signals. *J. Nonpar. Statist.*, page 25, 2001. to appear.

Abbildungsverzeichnis

Tabellenverzeichnis

Index