PROSEMINAR DATENREDUKTIONSTECHNIKEN SS‘99
SELEKTIVITÄTSSCHÄTZUNG DURCH WAVELETS

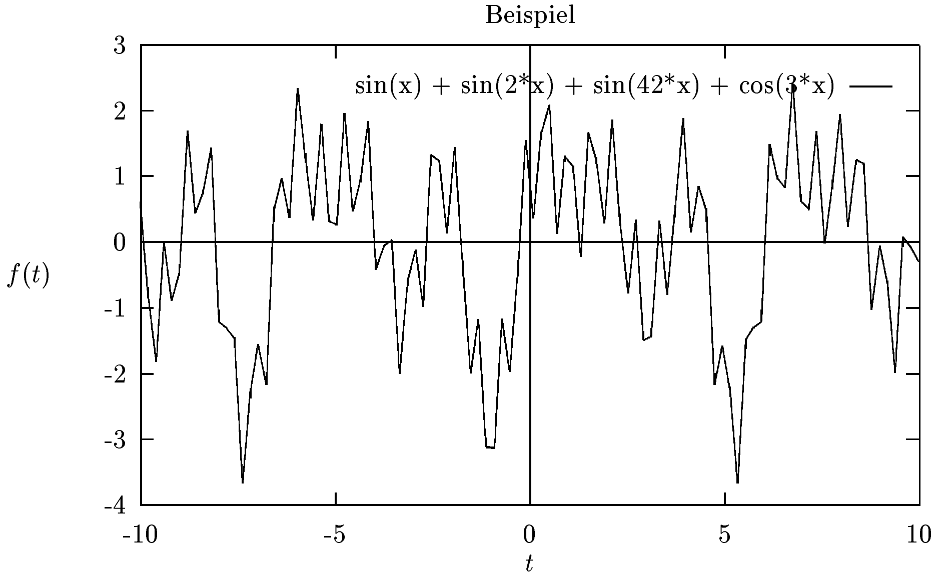
ANDREAS HILDEBRANDT
BETREUER: ARND CHRISITAN KÖNIG

ZUSAMMENFASSUNG. Um die Zeit, die eine Abfrage in einem Datenbanksystem in Anspruch nimmt, minimieren zu können, ist es u.a. nötig, eine grobe Abschätzung der sogenannten „Selektivität“ (d.h. den Anteil der Datensätze in der Datenbank, die von dieser Abfrage erfaßt würden) angeben zu können. Offensichtlich kann diese nur von Nutzen sein, wenn sie sich schnell und platzsparend berechnen läßt. In dieser Ausarbeitung wird ein moderner und interessanter Lösungsansatz - die Verwendung von Wavelets - für dieses Problem vorgestellt.

1. MATHEMATISCHE GRUNDLAGEN

1.1. **Grundlagen der Signalanalyse.** In dieser Ausarbeitung wird im folgenden immer wieder von „Signalen“ die Rede sein. Darunter verstehe ich für die Zwecke dieses Textes eine (eindimensionale) Funktion der Zeit, also $f = f(t)$, die für jedem Zeitpunkt genau einen Funktionswert zurückliefert. Eine Verallgemeinerung auf z.B. mehrdimensionale Signale ist zwar leicht möglich aber unanschaulich und würde damit den Rahmen dieses Textes sprengen.

Beispiele für solche Signale wären Wechselspannungen, Töne oder Bewegungen.



Die „Signalanalyse“ beschäftigt sich damit, so viel Informationen wie möglich aus einem Signal zu gewinnen. Dabei werden die zugrundeliegenden Daten üblicherweise auf die eine oder andere Weise transformiert, um Eigenschaften des Signals einfach ablesen zu können, die in der ursprünglichen Fassung nicht offensichtlich waren.

Ein wichtiges Werkzeug, mit dem solche Transformationen durchgeführt werden können, ist die Darstellung durch ein vollständiges Funktionensystem.

Aus der linearen Algebra kennt man die Zerlegung eines Vektors bezüglich eines vollständigen Systems $\{\vec{e}_1, \dots, \vec{e}_d\}$ von sogenannten „Basisvektoren“.

$$\vec{\xi} = \sum_{i=1}^{d} \xi_i \, \vec{e}_d$$

(Im dreidimensionalen Fall enstpricht dies z.B. den Projektionen auf die $x-$, $y-$, und $z-$Achsen).

Dieses Konzept läßt sich punktweise auf Funktionen übertragen: Ein Funktionensystem $\{\hat{f}_1, \dots, \hat{f}_n\}$ heißt *vollständig*, wenn sich jede Funktion an jedem Punkt ihres Definitionsbereiches durch dieses System darstellen läßt.

Hierfür wird man im allgemeinen allerdings unendlich viele Basisfunktionen benötigen. Daraus (sehr viele sehr kleine Summanden) ergibt sich auch ein Übergang von einer Summe zum Integral. Also:

$$f(x) = \int_{-\infty}^{\infty} a(\omega) \, \hat{f}(\omega, x) d\omega$$

1.2. **Die klassische Fouriertransformation.** Das wahrscheinlich wichtigste und bekannteste vollständige Funktionensystem bilden die sinus- und cosinus-Funktionen verschiedener Frequenzen: $\{\sin(\omega x), \cos(\omega x)\}$. Der von diesem System aufgespannte Raum heißt „Frequenz-“ oder „Fourierraum“.

Mit Hilfe der Eulerschen Formel $e^{i\phi} = \cos(\phi) + i \sin(\phi)$ ergibt sich nach längerer Rechnung dann für den Übergang vom Zeit- in den Fourierraum die folgende Formel (*„Fouriertransformation“*):

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) \, e^{i\omega t} \, dt$$

Es wird also die ursprüngliche Funktion mit sinus- und cosinus-Funktionen (Eulersche Formel!) multipliziert und das Produkt über alle Frequenzen integriert. Der Vorfaktor $\frac{1}{\sqrt{2\pi}}$ dient schließlich noch der Normierung.

Die erhaltene Funktion $\hat{f}(\omega)$ läßt sich folgendermaßen interpretieren: Zu jeder Frequenz ω gibt $\hat{f}$ die Amplitude an, die ein sinus bzw. cosinus dieser Frequenz haben müßte, damit man, nachdem man alle sinus- und cosinus-Funktionen mit den dazugehörigen Amplituden multipliziert und aufeinanderaddiert hat, wieder das ursprüngliche Signal erhält.

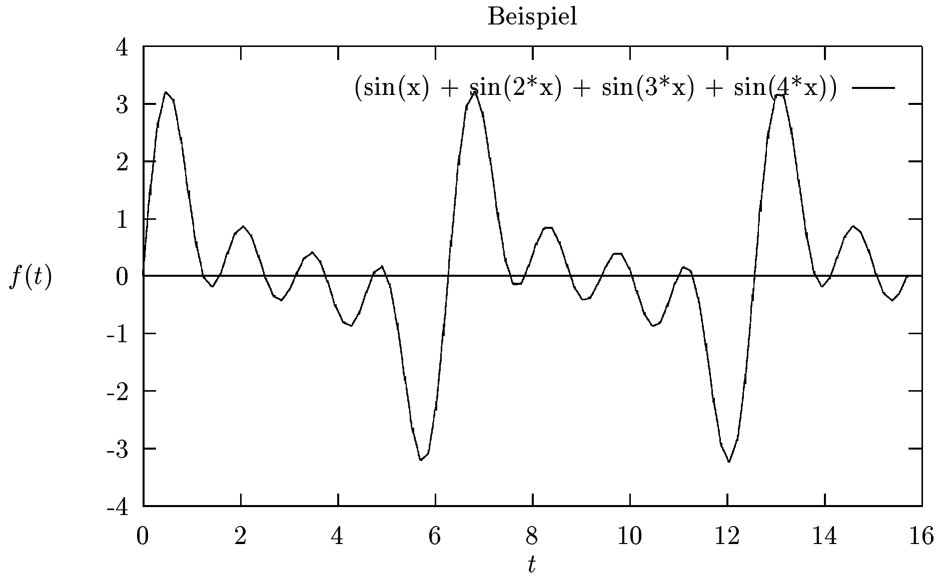
Anschaulich gesprochen sagt $\hat{f}$ aus, welche Frequenzen in dem Signal „vorkommen“ und wie „wichtig“ sie in diesem sind.

Durch die Fouriertransformation hat man also Informationen über das Frequenzverhalten des Signales gewonnen.

Auch die Rücktransformation - die sogenannte inverse Fouriertransformation - ist verlustlos möglich:

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(\omega) \, e^{-i\omega t} \, d\omega$$

Ein Beispiel: Die Fouriertransformierte der folgenden Funktion:



besteht aus vier sehr hohen, sehr dünnen Spitzen („peaks") bei den Frequenzen $\omega = 1, 2, 3$ und 4.

Um zu verdeutlichen, *warum* man überhaupt Informationen über das Frequenzverhalten gewinnen möchte, noch ein weiteres Beispiel: Ein wichtiges Mittel der Hirnforschung ist die Aufnahme von EEG's, also Hirnströmen, die auf irgendeine Art und Weise durch die nervliche Aktivität in verschiedenen Regionen des Hirns zustandekommt. Man kennt nun von verschiedenen Regionen typische Frequenzen, d.h. bei Aktivität verursachen sie Wechselspannungen oder -ströme mit einer bestimmten Frequenz. Führt man nun eine Fouriertransformation des gemessenen EEG's durch, so kann man den auftretenden Frequenzen Hirnregionen zuordnen. Die „Wichtigkeit" der Frequenz im Signal entspricht dabei der Aktivität der entsprechenden Hirnregion.

Leider gibt es bei der Fouriertransformation noch ein Problem: Sie sagt nichts darüber aus, *wann* die Frequenzen in einem Signal vorkommen. Sie geht sozusagen davon aus, daß sich der Anteil der verschiedenen Frequenzen am Signal zeitlich gesehen nicht ändert (solche Signale nennt man *stationär*).

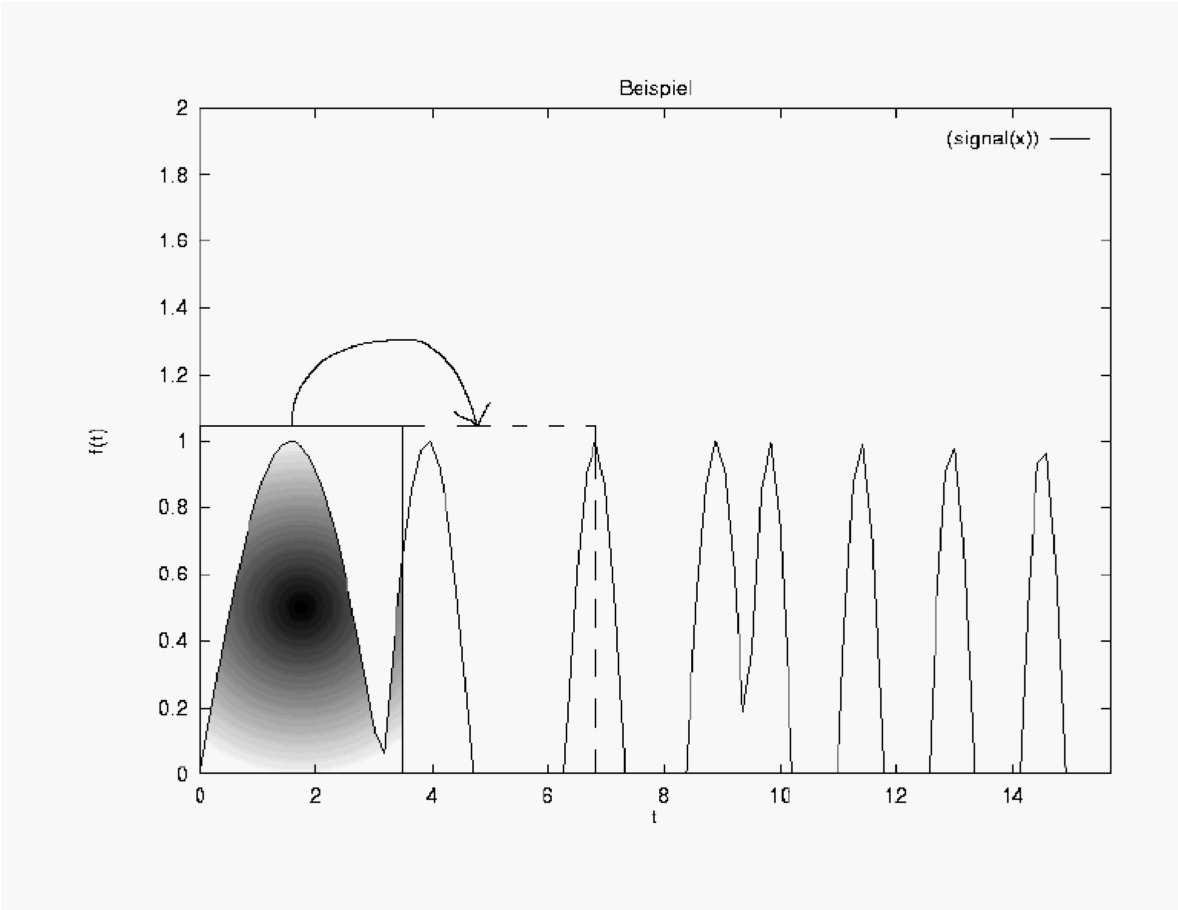
Im obigen Beispiel würde dies bedeuten, daß man zwar wüßte *welche* Hirnregionen aktiv sind, aber nicht *wann*.

Aufgrund dieser Tatsache können Signale, die sich nicht im geringsten ähnlich sehen, die gleiche Fouriertransformierte haben.

1.3. **Die gefensterte Fouriertransformation (*STFT*).** Um dieses Problem zu umgehen, wurde 1946 die sogenannte *gefensterte Fouriertransformation* entwickelt.

Hintergrund dieser Transformation ist der folgende Ansatz: Wenn die Funktion schon nicht zeitlich gesehen konstantes Frequenzverhalten zeigt, so wird sie sich doch sicherlich in so kleine „Funktionsbruchstücke" zerlegen lassen, daß in diesen stationäres Verhalten herrscht. Auf diese Stücke läßt sich dann die klassische Fouriertransformation anwenden.

Realisiert wird diese Zerlegung durch die Multiplikation mit einer sogenannten „Fensterfunktion", die nur ein kleines Stück der Funktion übrig läßt. Dieses Stück wird fouriertransformiert und die Fensterfunktion um ein ein kleines Stück auf der ursprünglichen Funktion verschoben. Auf diese Art und Weise fährt man fort, bis man die gesamte Funktion transformiert hat.



Der Kasten im obigen Bild soll für die Fensterfunktion stehen (in der realen Anwendung wird man aus bestimmten Gründen wohl eher eine Gauß'sche Glockenkurve als ein Rechteck verwenden).

Aber auch bei der *STFT* gibt es ein Problem: Wie findet man die optimale Fensterfunktion für gegebene Daten? Denn es gilt: Wählt man sein Fenster sehr eng, so kann man den Zeitpunkt, zu dem die Frequenzen im Signal auftreten zwar sehr stark eingrenzen. Aber niedrige Frequenzen fallen aus dem Fenster heraus, werden also nicht mehr registriert. Bei einem sehr breiten Fenster „sieht" man zwar auch niedrige Frequenzen, man kann aber den Zeitpunkt ihres Auftretens nicht mehr genau eingrenzen.

Zusammengefasst:

	Frequenzdarstellung	zeitliche Darstellung
enges Fenster	schlecht	gut
weites Fenster	gut	schlecht

Man kann sogar zeigen, daß es keine ideale Fensterfunktion geben kann. Sie muß immer den vorliegenden Daten angepaßt sein. Dies bezeichnet man als *„Auflösungsproblem"*.

Jedesmal, wenn man Informationen über das zeitliche Verhalten gewinnt, opfert man Informationen über das Frequenzverhalten und umgekehrt. Genauer formuliert: man kennt nie gleichzeitig das genaue Frequenz- und Zeitverhalten einer Funktion in einem ihrer Punkte. Dies ist die sogenannte *Heisenbergsche Unschärferelation*.

1.4. **Die kontinuierliche Wavelettransformation (KWT).** Wavelets wurden entwickelt, um dieses Problem zu umgehen. Die Unschärferelation gilt natürlich weiterhin, läßt sich aber durch das folgende Vorgehen „austricksen“: Da man ja weiß, daß es keine optimale Fensterfunktion geben kann, probiert man einfach verschiedene aus.

Genauer: Man beginnt mit einer grundlegenden Basisfunktion („Mutterwavelet“). Diese wird dann, wie zuvor das Fenster, über die Funktion geschoben und dabei „auf sie angewendet“. Danach wird sie „gedehnt“ oder „gestaucht“ und in der neuen Form wieder über das Signal geschoben usw.

Man benötigt bei diesem Vorgehen auch keine Fouriertransformation mehr: Die unterschiedliche Ausdehnung des Wavelets durch Stauchen/Dehnen sorgt schon von selbst für die Trennung der Frequenzen. Dies kann man sich einfach veranschaulichen: In einem engen Fenster haben nur hohe Frequenzen „Platz“, niedrige dagegen fallen heraus. In einem weiten Fenster hingegen werden zu hohe Frequenzen nicht mehr „wahrgenommen“, denn über einen großen Bereich betrachtet werden sich hohe Frequenzen nur noch als undeutliches Rauschen bemerkbar machen (man „zoomt“ sozusagen aus dem Signal heraus - dabei werden hohe Frequenzen immer undeutlicher erscheinen).

Durch diese Art der Transformation kann man also ein Signal gleichzeitig in verschiedenen Auflösungen oder Skalen betrachten.

Welche Funktionen kommen nun als Wavelets in Frage?

Mathematisch gesehen ist jede quadratintegrable Funktion ψ mit der Eigenschaft

$$0 < c_\psi := 2\pi \int_R \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty$$

ein Wavelet. Dies nennt sich „Zulässigkeitsbedingung“.

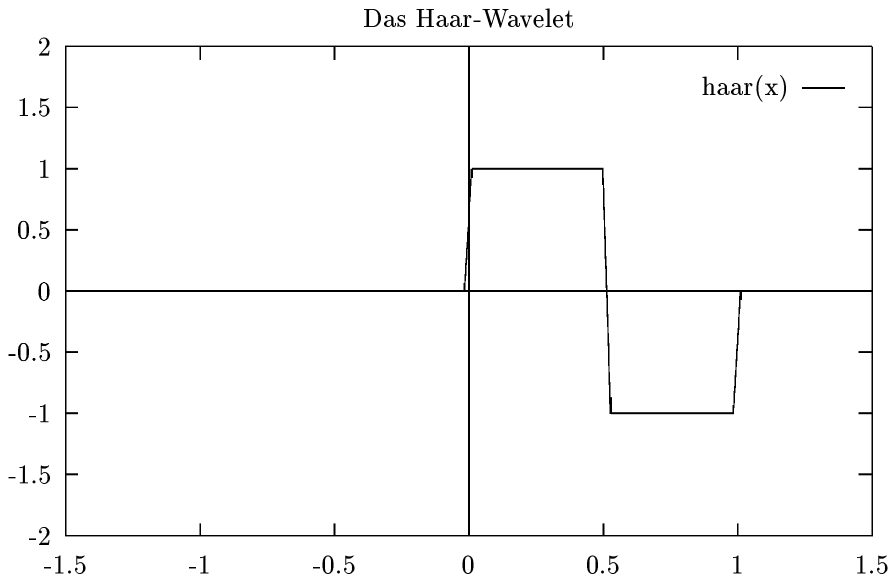
Dabei ist $\hat{\psi}(\omega)$ die Fouriertransformierte von ψ.

Aus dieser Bedingung folgt u.a. direkt, daß der zeitliche Mittelwert jedes Wavelets verschwindet:

$$0 = \hat{\psi}(0) = \frac{1}{\sqrt{2\pi}} \int_R \psi(t) dt$$

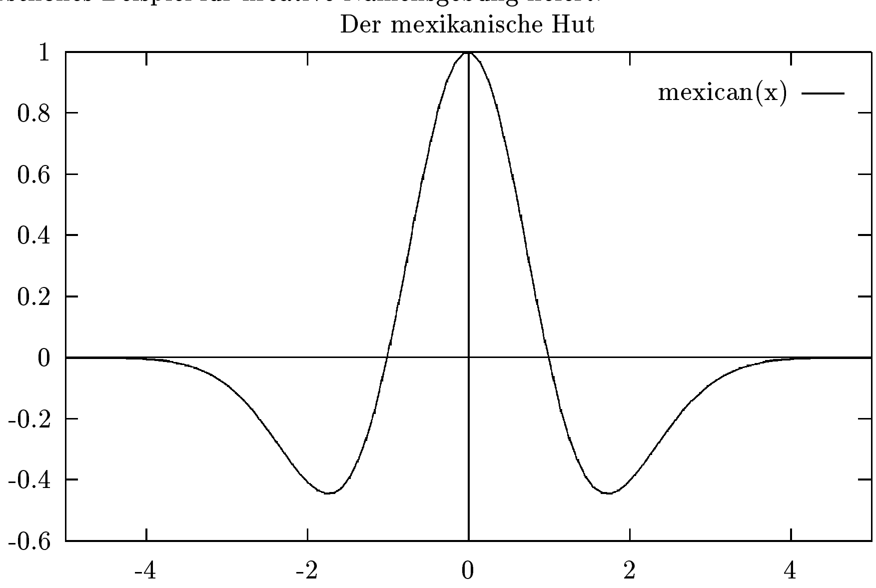
Diese Beziehung war es auch, die dieser Klasse von Funktionen ihren Namen eingebracht hat: *Wavelet* bedeutet zu deutsch in etwa *kleine Welle*, denn das Verschwinden des Mittels entspricht einer Schwingung. Wieso die Wellen ausgerechnet „klein“ sein sollen, ist allerdings nicht so eindeutig zu beantworten.

Das einfachste Wavelet ist das sogenannte *Haar-Wavelet:*



Dieses Wavelet hat die sehr angenehme Eigenschaft, daß es sehr einfach zu beschreiben, zu berechnen und anzuwenden ist. Leider hat es aber andere negative Eigenschaften (insbesondere ist es nicht stetig differenzierbar), die dazu führen, daß es in realen Anwendungen kaum angewandt wird. Es handelt sich also sozusagen um ein „Übungswavelet".

In der realen Anwendung schon wesentlich wichtiger, dafür aber auch erheblich schwieriger zu berechnen ist der sogenannte *Mexikanische Hut*, der gleichzeitig ein wunderschönes Beispiel für kreative Namensgebung liefert:



Was genau versteht man nun unter der „Anwendung" eines Wavelets?

Mathematisch gesehen ist die Wavelettransformierte $L_\psi f(\tau, \chi)$ zum Wavelet ψ folgendermaßen definiert:

$$L_\psi f(\tau,\chi) := \frac{(2\pi \int_R \frac{|\hat{\psi(\omega)}|^2}{|\omega|} d\omega)^{-\frac{1}{2}}}{\sqrt{\tau}} \int_R f(t)\psi(\frac{t-\chi}{\tau})dt$$

Offenbar greift dies einen Bereich um $t = \chi$ aus der Funktion heraus, τ bestimmt dessen Ausdehnung ($lim_{\tau \to 0}$ „zoomt" in die Funktion hinein).

$L_\psi f(\tau, \chi)$ bildet (da ja f eindimensional vorausgesetzt war) eine Fläche. Aus dieser läßt sich zu jeder „Zeit" χ ablesen, wie „wichtig" eine bestimmte Frequenz ($\propto \frac{1}{\tau}$) gerade im Signal ist.

Auch die Rücktransformation ist möglich:

$$f(t) = \frac{1}{\sqrt{c_\psi}} \int_R \int_R L_\psi f(\tau, \chi) \frac{1}{\sqrt{|\tau|}} \psi(\frac{t - \chi}{\tau}) \frac{d\tau \, d\chi}{\tau^2}$$

Doch auch die Wavelettransformation bereitet Probleme:

1. Berechenbarkeit. Die KWT erfordert mehrere Integrationen. Außerdem sind τ, χ kontinuierliche Variablen. Also ist die Auswertung an unendlich vielen Stellen nötig. Dies ist offensichtlich in der realen Anwendung nicht durchführbar, es muß also auf irgendeine Art und Weise diskretisiert oder gesampelt werden.
2. Nicht auf diskrete Werte anwendbar. Bisher waren wir von kontinuierlichen Funktionen ausgegangen. Für die reale Anwendung ist aber eine Anpassung an diskrete Werte (z.B. Messwerte) nötig, denn in den seltensten Fällen wird man wirklich kontinuierliche Signale bearbeiten.

1.5. **Multi-Skalen Analyse, diskrete und schnelle Wavelettransformation.** Diese Probleme lassen sich glücklicherweise beide auf einen Schlag beseitigen. Um dies zu bewerkstelligen, verwendet man die sogenannte Multi-Skalen Analyse, angewandt auf Wavelets.

Die Multi-Skalen Analyse zerlegt - anschaulich betrachtet - ein Signal in einen „glatten" und einen „rauhen" oder „detaillierten" Anteil: $f = g + d$. Dabei kann man den glatten Anteil mit niedrigen Frequenzen, den rauhen mit hohen Frequenzen assoziieren (dies kann man sich z.B. anhand eines Sinus verdeutlichen: Ein Sinus mit sehr geringer Frequenz wirkt glatt, ein sehr hochfrequentiger hingegen springt immer schärfer hin und her). Man speichert nun d und zerlegt g noch einmal wie oben.

Die Aufspaltung erfolgt hierbei durch die Anwendung von Wavelets.

Da g „glatter" ist als f, muß man auch nur noch weniger Stützpunkte speichern als für f selbst, d.h. es genügt, weniger Werte zu speichern, zwischen denen dann interpoliert werden kann. Formal folgt dies aus dem Nyquist-Theorem (die Samplingrate eines Signales muß immer mindestens doppelt so hoch sein wie die höchste im Signal vorkommende Frequenz).

Man führt nun die Aufspaltung so lange fort, bis nur noch ein einziger Stützpunkt benötigt wird. Zusammen mit den gespeicherten d's bezeichnet man diese als die „Wavelettransformierte" des diskreten Signales.

Wie läuft nun die Anwendung der Wavelets auf die diskreten Werte im einzelnen ab?

Dazu wird das Wavelet auf je zwei benachbarte Punkte angewandt und dann auf die nächsten beiden Punkte verschoben. Jeweils zwei Punkte werden dabei zu einem Wert „verschmolzen". Die neuen Werte liegen jetzt natürlich weiter auseinander als die Ausgangswerte (es sind ja nur noch halb so viele). Also muß das Wavelet im nächsten Schritt erst gedehnt werden. Mit diesem gedehnten Wavelet wiederholt man dann die Anwendung auf die eben berechneten Punkte.

Man kann zeigen, daß bei diskreten Daten die Anwendung des Haar-Wavelets auf die Punkte x_i, x_{i+1} für den „glatten" Anteil

$$\bar{x} := \frac{x_i + x_{i+1}}{2}$$

also genau deren Mittelwert ergibt. Die Details entsprechen dem Abstand der beiden Punkte:

$$2\,|\bar{x} - x_i| = 2\,|\bar{x} - x_{i+1}|$$

Man wendet das Haar-Wavelet nun auf die $\bar{x}_i$ an und iteriert dies, bis nur noch ein Mittelwert übrig bleibt.

Auf diese Art ergeben sich für das Haar-Wavelet die folgenden Transformationsformeln:

$$\begin{aligned}
S(0) &= \hat{S}(0) - \frac{1}{2}\hat{S}(1) - \frac{1}{2}\hat{S}(2) - \frac{1}{2}\hat{S}(4) \\
S(1) &= \hat{S}(0) - \frac{1}{2}\hat{S}(1) - \frac{1}{2}\hat{S}(2) + \frac{1}{2}\hat{S}(4) \\
&\vdots \\
S(7) &= \hat{S}(0) + \frac{1}{2}\hat{S}(1) + \frac{1}{2}\hat{S}(3) + \frac{1}{2}\hat{S}(7)
\end{aligned}$$

2. Anwendung von Wavelets zur Selektivitätsschätzung

Viele Algorithmen, die auf Wavelets beruhen, machen sich eine bestimmte wichtige Eigenschaft der Wavelettransformierten zunutze: In der Transformierten von real vorkommenden Signalen sind üblicherweise viele Detailkoeffizienten sehr klein, d.h. ihre Vernachlässigung wird das Ergebnis nicht sehr verfälschen.

Diese Eigenschaft nutzen auch Matias, Wang und Vitter in ihrem Paper, um ein Verfahren zur effizienten Durchführung von Selektivitätsschätzungen zu entwickeln.

Zugrunde liegt folgender Algorithmus:

- Vorverarbeitung: berechne die zum Attribut X gehörige erweiterte kumulative Datenverteilung T^{C^+}. [1]. Diese enthalte n Tupel.
- Bilde die Wavelettransformierte von T^{C^+}. Diese besteht aus n Koeffizienten.
- Wähle aus diesen n Werten nur die $m < n$ wichtigsten Werte aus. Die restlichen $n - m$ Werte werden ignoriert.

Hat man dieses Verfahren durchgeführt, so läßt sich die Selektivität einer Range Query (d.h. die Anfrage, welcher Anteil der Datensätze in der Relation bezüglich X einen Wert $a \leq X \leq b$ benutzt), folgendermaßen berechnen: Rekonstruiere aus der (durch die Datenreduktion modifizierten) Wavelettransformierten die Näherung der ursprünglichen kumulativen Frequenzen von $a-1$ und b. Diese seien c_{a-1} und c_b. Dann entspricht offenbar c_{a-1} der Anzahl von Datensätzen mit einem Wert kleiner a in R und c_b der Anzahl derer mit einem Wert kleiner oder gleich b. Zieht man nun c_{a-1} von c_b ab und dividiert das Ergebnis durch die Gesamtzahl der Datensätze, so erhält man die Selektivität der Range Query. [2]

Nun bleibt nur noch die Frage zu klären, wie die Auswahl der wichtigsten Koeffizienten abläuft.

Dazu schlagen Matias, Vitter und Wang folgendes vor:

---

[1]Diese ist folgendermaßen definiert: zu dem Wert v_i sei c_i definiert als die Anzahl aller Datensätze in der Relation R, die bezüglich des Attributes X einen Wert kleiner oder gleich v_i besitzen. Dann nennt man T^C die *kumulative Datenverteilung* Weist man nun noch jedem undefinierten Wert v_j die kumulative Frequenz $c_j = c_{j-1}$ des Vorgängers zu, so erhält man die erweiterte kumulative Datenverteilung T^{C^+}.

[2]Dieses Ergebnis ist natürlich nur eine Näherung, denn der oben angegebene Algorithmus ignoriert ja die „unwichtigeren" Koeffizienten der Wavelettransformierten, was einem Informationsverlust gleichkommt.

- Wähle zuerst ein geeignetes Fehlermaß. Das ein Koeffizient „unwichtig" ist, ist äquivalent zu der Forderung, daß seine Entfernung keinen großen Fehler im Ergebnis liefert. Dieser Fehler läßt sich auf verschiedene Weise berechnen, so z.B. durch den zur 2-Norm gemittelten Fehler:

$$||e_i||_2 := \sqrt{\frac{1}{Q} \sum_{1 \leq i \leq Q} e_j^2}$$

  Dabei sind die e_j die absoluten Fehler, die für den Koeffizienten S_j der ursprünglichen Datenverteilung bei der Elimination des Koeffizienten $\hat{S}_i$ aus der Wavelettransformierten entstehen.
- Wende irgendeinen Algorithmus an, der die m Koeffizienten $\hat{S}_i$ bestimmt, bei deren Entfernung der Fehler $||e_i||$ am geringsten ist. So könnte man z.B. mit den $2m$ größten Koeffizienten beginnen und dann m von ihnen „gierig" entfernen, d.h. in m Durchläufen jeweils denjenigen, dessen Entfernung den geringsten Fehler verursacht.

Die verschiedenen Koeffizienten der Wavelettransformierten hängen auf komplizierte Art und Weise zusammen. Entfernt man einen von ihnen, muß daher für einige andere Koeffizienten der Fehler, den ihre Entfernung nun verursachen würde, neu berechnet werden. Um effizient bestimmen zu können, für welche Koeffizienten dies der Fall ist, wendet man einen sogenannten Fehlerbaum an: Man trägt in einer Baumstruktur die Zusammenhänge der verschiedenen Koeffizienten untereinander auf und schreibt zu Beginn an die $2m$ größten von ihnen den Fehler, den ihre Entfernung verursachen würde. Nun entfernt man denjenigen mit dem geringsten assoziierten Fehlerwert und berechnet für die Koeffizienten, die im von ihm aufgespannten Teilbaum liegen und für diejenigen auf dem Pfad zur Wurzel den Fehler neu.

Beispiel:

Die zugrundeliegende Verteilung sei

$$[0, 0, 1, 2, 4, 4, 6, 10]$$

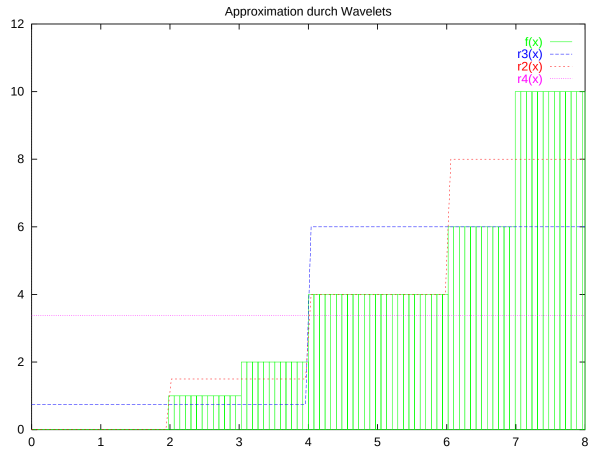
Es ergibt sich:

Auflösung	Mittelwerte	Details
8	[0, 0, 1, 2, 4, 4, 6, 10]	
4	[0, 1.5, 4, 8]	[0, 1, 0, 4]
2	[0.75, 6]	[1.5, 4]
1	[3.375]	[5.25]

Dann ist die Wavelettransformierte dieser Verteilung:

$$[3.375, 5.25, 1.5, 4, 0, 1, 0, 4]$$

Das folgende Schaubild stellt die unterschiedlichen oben berechneten Auflösungen graphisch dar.

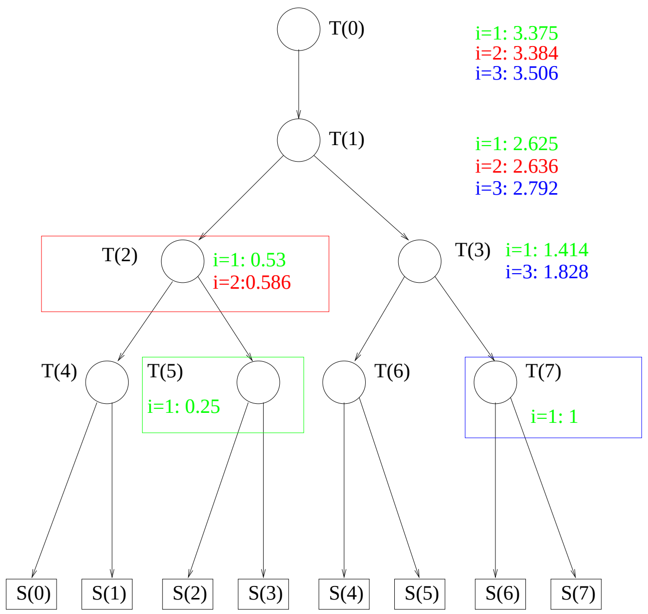


Wendet man das obige Verfahren nun mit $m = 3$ auf die obige Verteilung an, so ergibt sich als neue Wavelettransformierte:

$$[3.375, 5.25, 0, 4, 0, 0, 0, 0]$$

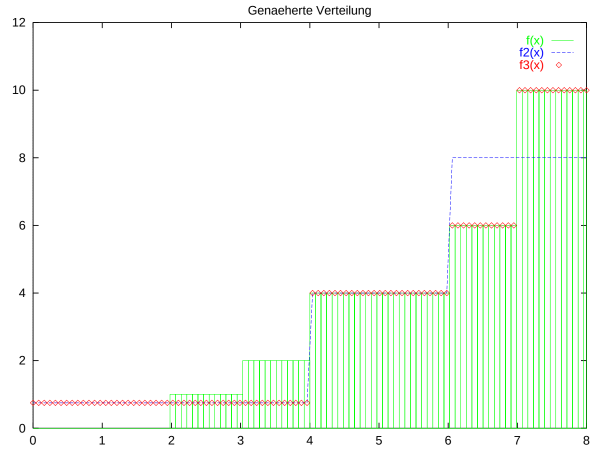
Im ersten Schritt wurde dabei $\hat{S}(5)$ aus der Transformierten entfernt (im folgenden Bild angedeutet durch den grünen Kasten), da die Gesamtfehlerzunahme mit 0.25 für diesen Koeffizienten minimal wird. Der von $\hat{S}(5)$ aufgespannte Teilbaum enthält keine Knoten, die für die Entfernung in Frage kämen. Daher muß nur für die Koeffizienten auf dem Pfad zur Wurzel - also für $\hat{S}(2), \hat{S}(1)$ und $\hat{S}(0)$ - der entstehende Fehler aktualisiert werden. Die restlichen beiden Schritte laufen analog.

Der dabei aufgebaute Fehlerbaum hat folgendes Aussehen:



(T(0), ..., T(7) seien die Elemente der Wavelettransformierten)

Rechnet man diese Näherung wieder zurück, um die ursprüngliche Datenverteilung zu erhalten, so ergibt sich das im folgenden Plot dargestellte Ergebnis (die blaue Linie entspricht einem Wert von $m = 3$, die rote $m = 4$):



3. Zusammenfassung

Wavelets stellen eine moderne und interessante Entwicklung im Bereich der Datenreduktion dar.

Experimente von Matias, Vitter und Wang zeigen, daß der oben dargestellte Algorithmus wesentlich exaktere Ergebnisse liefert als z.B. bekannte Histogrammtechniken (wie z.B. MaxDiff(V, A)). Besonders groß scheint der Vorsprung vor Random Sampling-Verfahren zu sein.

Außerdem besitzen Wavelets neben den hier erwähnten noch viele andere positive Eigenschaften. So lassen sie sich z.B. leicht auf mehrdimensionale Probleme verallgemeinern. Es existieren auch schon einige Algorithmen zur Selektivitätsschätzung, die auf dieser Eigenschaft aufbauen.

Es ist daher anzunehmen, daß sich Wavelets auch auf dem Gebiet der Datenreduktionstechniken weiterhin wachsender Beliebtheit erfreuen werden.