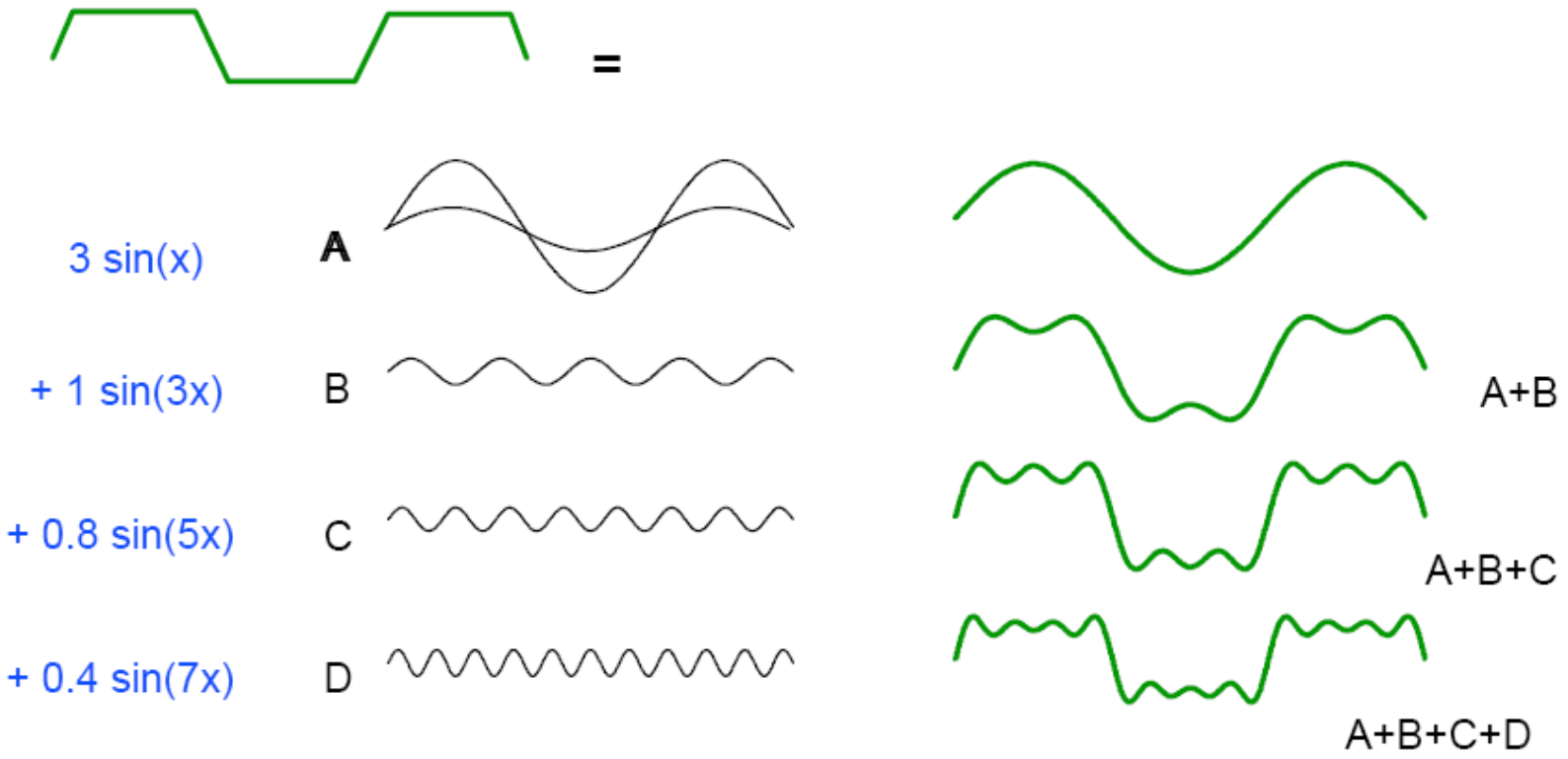
4.2 Matching-basierte Analyse

- Diskrete Fourier Transformation (DFT)
  - Idee
    - Beschreibe beliebige periodische Funktion als gewichtete Summe periodischer Grundfunktionen (Basisfunktionen) mit unterschiedlicher Frequenz
    - Basisfunktionen: sin und cos

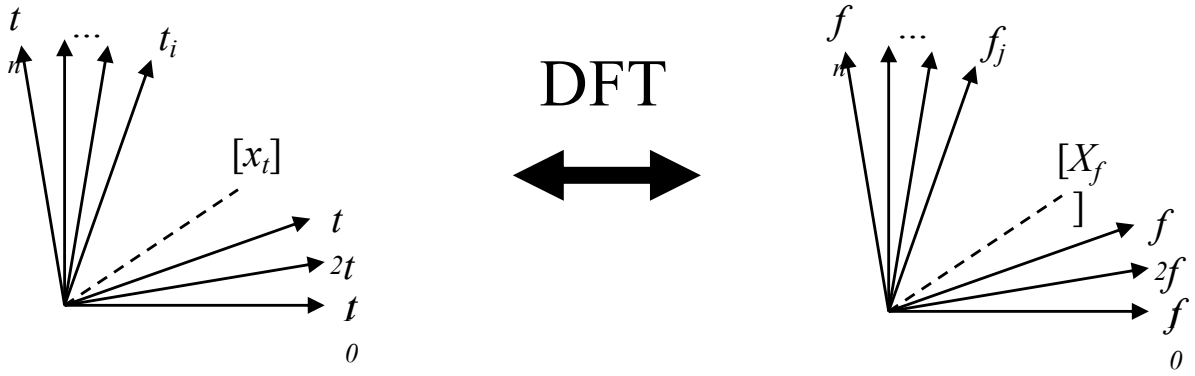


4.2 Matching-basierte Analyse

- DFT
  - Fouriers Theorem:

    Jede beliebige periodische Funktion lässt sich darstellen als Summe von Kosinus- und Sinus-Funktionen unterschiedlicher Frequenzen.

  - Transformation verändert eine Funktion nicht, sondern stellt sie nur anders dar
  - Transformation ist umkehrbar => inverse DFT
  - Analogie: Basiswechsel in der Vektorrechnung



  - Was ist diese andere „Basis“?

4.2 Matching-basierte Analyse

- Formal
  - Gegeben sei eine Zeitreihe der Länge n: $x = [x_t]$, $t = 0, \ldots, n-1$
  - Die DFT von x ist eine Sequenz $X = [X_f]$ von n komplexen Zahlen für die Frequenzen $f = 0, \ldots, n-1$ mit

$$X_f = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x_t \cdot e^{\frac{-j2\pi ft}{n}} =$$

$$\underbrace{\frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x_t \cos\left(\frac{2\pi ft}{n}\right)}_{\text{Realteil}} - j \cdot \underbrace{\frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x_t \sin\left(\frac{2\pi ft}{n}\right)}_{\text{Imaginärteil}}$$

wobei j die imaginäre Einheit bezeichnet, d.h. $j^2 = -1$.

  - Der Realteil gibt den Anteil der Kosinus- und der Imaginärteil den Anteil der Sinusfunktionen in der jeweiligen Frequenz f an.

4.2 Matching-basierte Analyse

- Durch die inverse DFT wird das ursprüngliche Signal *x* wieder hergestellt:

$$x_t = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} X_f \cdot e^{\frac{j 2\pi f t}{n}}$$

$t = 0, \ldots, n-1$ (t: Zeitpunkte)

$[x_t] \leftrightarrow [X_f]$ bezeichnet ein Fourier-Paar,
d.h. $\text{DFT}([x_t]) = [X_f]$ und $\text{DFT}^{-1}([X_f]) = [x_t]$.

- Die DFT ist eine lineare Abbildung, d.h. mit $[x_t] \leftrightarrow [X_f]$ und $[y_t] \leftrightarrow [Y_f]$ gilt auch:
  - $[x_t + y_t] \leftrightarrow [X_f + Y_f]$ und
  - $[ax_t] \leftrightarrow [aX_f]$ für ein Skalar $a \in IR$
- Energie einer Sequenz
  - Die Energie *E(c)* von *c* ist das Quadrat der Amplitude: $E(c) = |c|^2$.
  - Die Energie *E(x)* einer Sequenz *x* ist die Summe aller Energien über die Sequenz:

$$E(x) = \|x\|^2 = \sum_{t=0}^{n-1} |x_t|^2$$

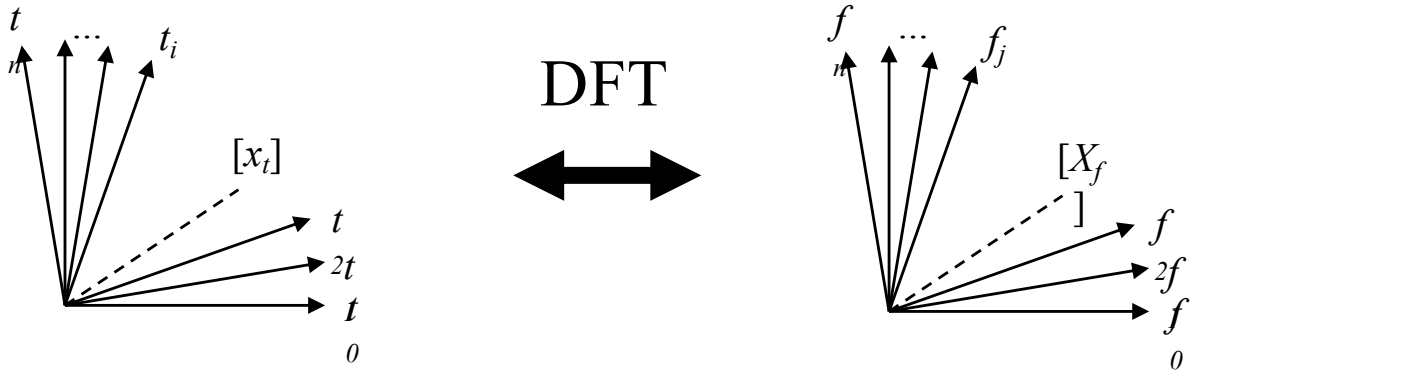
4.2 Matching-basierte Analyse

- Satz von Parseval

  Die Energie eines Signals im Zeitbereich ist gleich der Energie im Frequenzbereich.
  Formal: Sei *X* die DFT von *x*, dann gilt:

$$\sum_{t=0}^{n-1} |x_t|^2 = \sum_{t=0}^{n-1} |X_f|^2$$

- Folge aus Parsevals Satz und der Linearität der DFT: Die euklidische Distanz zweier Signale x und y stimmt im Zeit- und im Frequenzbereich überein: $|| x - y ||^2 = || X - Y ||^2$



„Zeitbreich (-raum)“ „Frequenzbereich (-raum)“

4.2 Matching-basierte Analyse

- Grundidee der Anfragebearbeitung:
  - Als Ähnlichkeitsfunktion für Sequenzen wird die euklidische Distanz verwendet:

$$dist(x,y) = \|x - y\| = \sqrt{\sum_{t=0}^{n-1} |x_t - y_t|^2}$$

  - Der Satz von Parseval ermöglicht nun, die Distanzen im Frequenz- statt im Zeitbereich zu berechnen: $dist(x,y) = dist(X,Y)$
- Kürzen der Sequenzen für die Indexierung
  - In der Praxis haben die tiefsten Frequenzen die größte Bedeutung.
  - Die ersten Frequenz-Koeffizienten enthalten also die wichtigste Information.
  - Für den Aufbau eines Index werden die transformierten Sequenzen gekürzt, d.h. von $[X_f]$, $f = 0, 1, \ldots, n-1$ werden nur die ersten c Koeffizienten $[X_f < c]$, $c < n$, indexiert.
  - Im Index kann dann eine untere Schranke der echten Distanz berechnet werden:

$$dist_c(x,y) = \sqrt{\sum_{f=0}^{c-1} |x_f - y_f|^2} \leq \sqrt{\sum_{f=0}^{n-1} |x_f - y_f|^2} = dist(x,y)$$

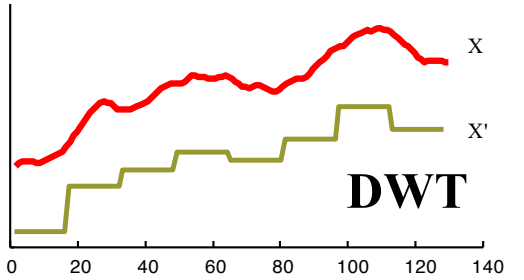
  - Filter-Refinement: Filterschritt auf gekürzten Zeitreihen (mit Indexunterstützung), Refinement auf kompletten Zeitreihen

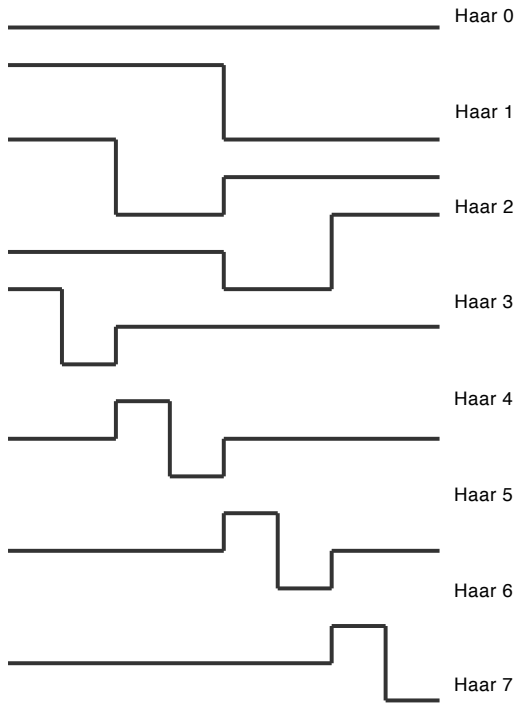
4.2 Matching-basierte Analyse

- Diskussion
  - Vorteile
    - (Sehr) gute Komprimierung natürlicher Signale
    - Effizient zu berechnen (O(n log n))
    - Kann zeitliche Verschiebungs-invariante Anfragen unterstützen
  - Nachteile
    - Zeitreihen müssen gleiche Länge haben
    - Keine Unterstützung gewichteter Distanzfunktionen

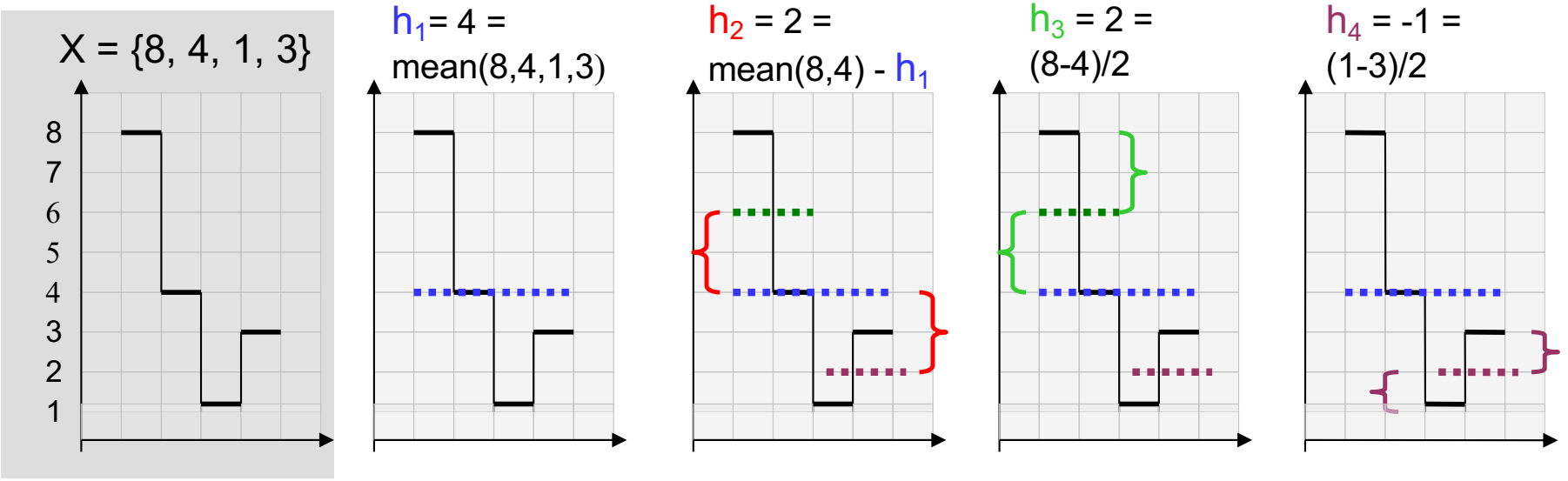
4.2 Matching-basierte Analyse

- Diskrete Wavelet Transformation (DWT)
  - Idee:
    - Repräsentiere Zeitreihe als Linearkombination von Wavelet-Basisfunktionen
    - Speichere nur die ersten Kooefizienten
    - Meist werden Haar-Wavelets als Basisfunktion gewählt (leicht zu implementieren)

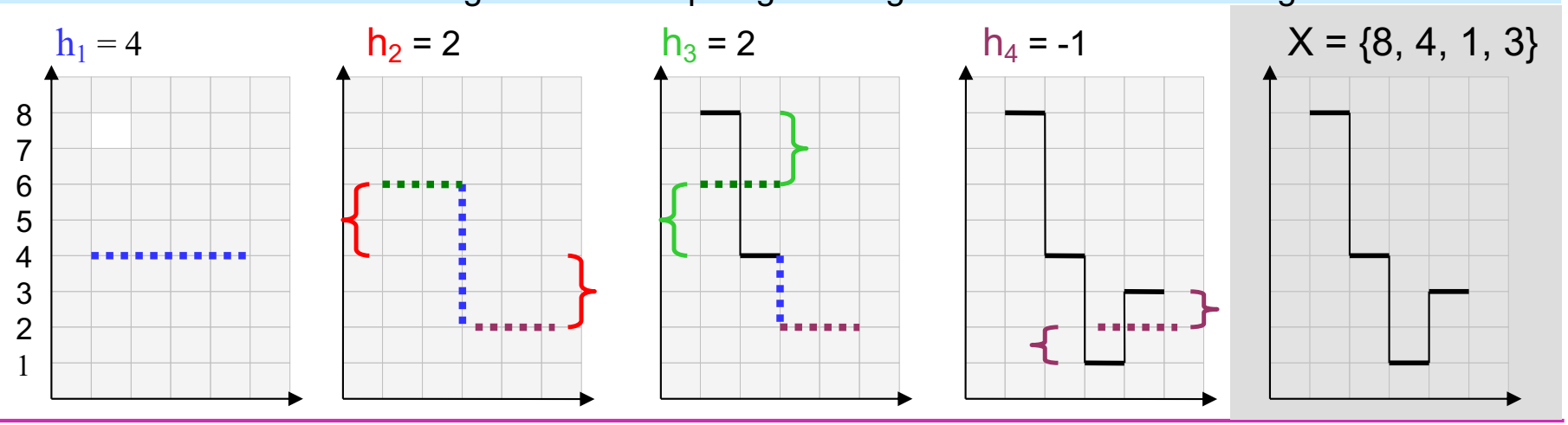




4.2 Matching-basierte Analyse



Schrittweise Transformation der Zeitreihe X = {8, 4, 1, 3} in die Haar Wavelet Darstellung H = [4, 2 , 2,-1]
Aus der Haar Wavelet Darstellung kann das ursprüngliche Signal verlustfrei wieder hergestellt werden.

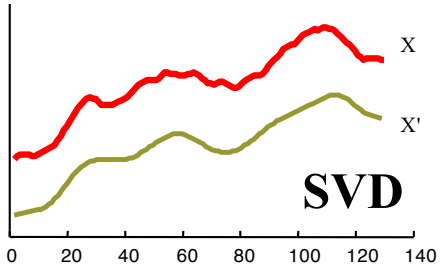


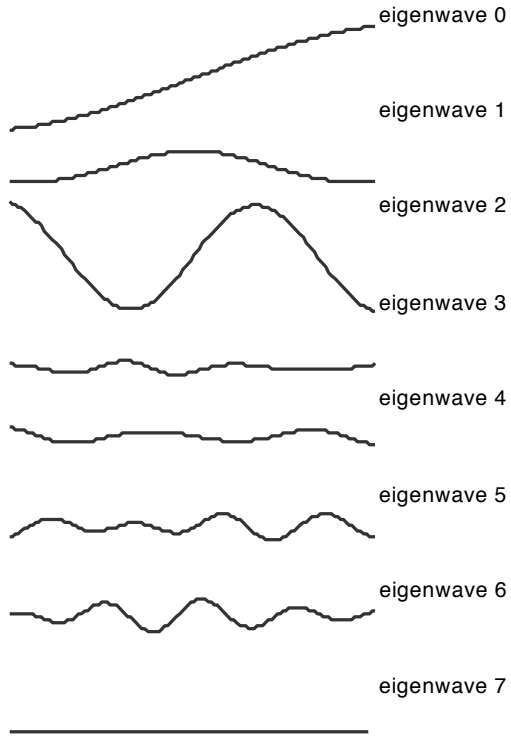
4.2 Matching-basierte Analyse

- Diskussion
  - Vorteile
    - Gute Kompression v.a. bei stationären Signalen
    - Kompression in linearer Laufzeit
    - Einfache Unterstützung nicht-Euklidischer Distanzmaße (z.B. DTW)
  - Nachteile
    - Länge der originalen Zeitreihen muss 2er-Potenz sein
    - Länge der reduzierten Zeitreihen sollte 2er-Potenz sein
    - Keine Unterstützung gewichteter Distanzmaße

4.2 Matching-basierte Analyse

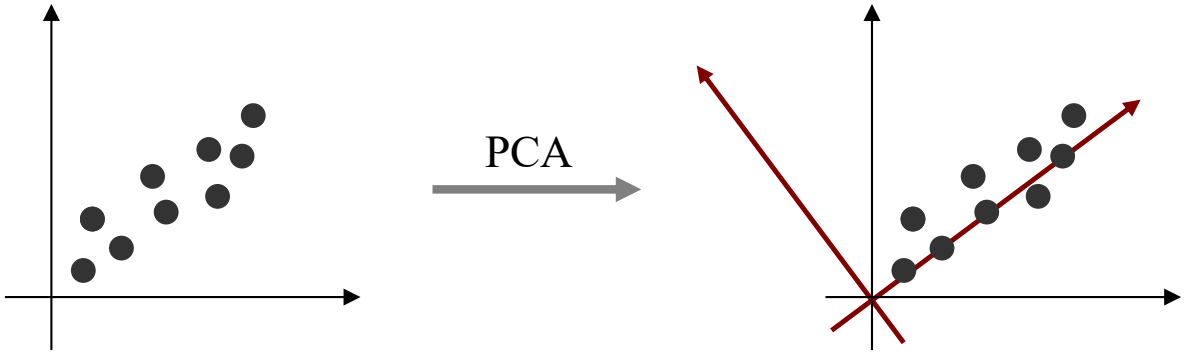
- Singular Value Decomposition (SVD)
  - Idee
    - Repräsentiere Zeitreihen als Linearkombination von Eigenwellen (Eigen Waves)
    - Speichere nur die ersten (wichtigsten) Koeffizienten
  - Vergleich SVD mit DFT und DWT
    - Ähnlich: Linearkombination von Funktionen, die die Form der Zeitreihen modellieren
    - Unterschied: SVD ist abhängig von den Daten (DFT: sin/cos, DWT: konstante Linien unterschiedlicher Amplituden)
  - SVD ist in der Textverarbeitung und dem Information Retrieval auch unter dem Acronym „Latent Semantic Indexing“





4.2 Matching-basierte Analyse

- Bestimmung der Eigenwellen
  - Zeitreihen der Länge *n* = *n*-dimensionale Punkte
  - PCA dieser Punkte
  - Rotation der Zeitachsen auf die Hauptachsen aller Zeitreihen
  - Erste Achse entlang der Richtung mit maximaler Varianz
  - Zweite Achse entlang der Richtung mit maximaler Varianz orthogonal zur ersten Achse
  - …



Alle Zeitreihenobjekte
als Punkte repräsentiert

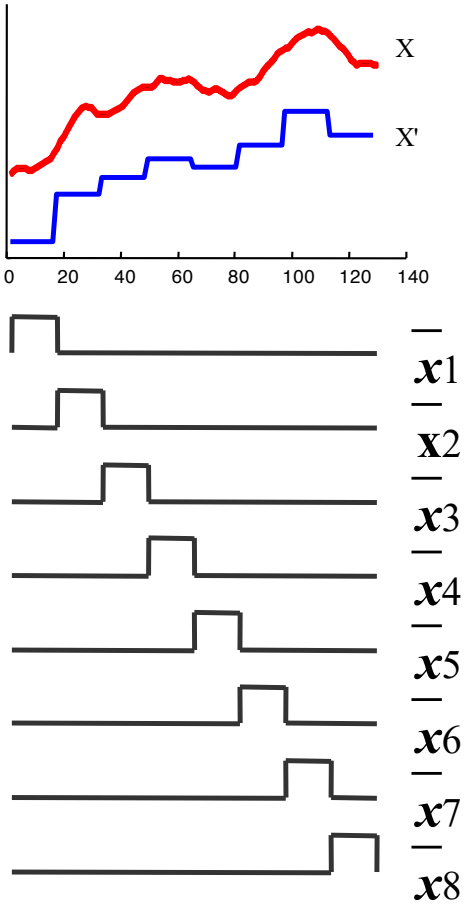
  - Transformiere Zeitreihen in das neue Koordinatensystem (gegeben durch Hauptachsen) und speichere nur die ersten k Werte, da diese Dimensionen die meiste Varianz der Zeitreihen repräsentieren
  - SVD minimiert den quadratischen Fehler, der durch das Weglassen von (*n*-*k*) Achsen gemacht wird

4.2 Matching-basierte Analyse

- Vorteile
  - Optimale Dimensionsreduktion durch Minimierung des kleinsten quadratischen Fehlers
  - Daten-abhängig
- Nachteile
  - Sehr aufwendig: $O(n^3)$
  - Nur für einfache Euklidische Distanz (keine gewichtete oder nicht-Euklidische Distanzfunktion verwendbar)
  - Schlecht für dynamische Daten: bei Einfügen/Löschen von Zeitreihen muss die gesamte SVD neu berechnet werden, da sich die Eigenwellen ändern könnten

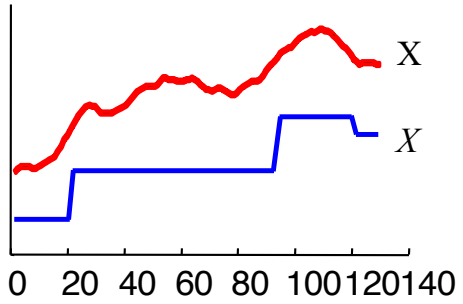
4.2 Matching-basierte Analyse

- Piecewise Aggregate Approximation (PAA)
  - Idee
    - Repräsentiere Zeitreihen als Sequenz von Box-Basisfunktionen
    - Jede Box hat dieselbe Länge
    - Je länger die Boxen, desto niedriger die resultierende Approximation
  - Vorteile
    - Schnell und einfach zu berechnen
    - Unterstützt alle Arten von Distanzfunktionen
    - Unterstützt Zeitreihen verschiedener Länge

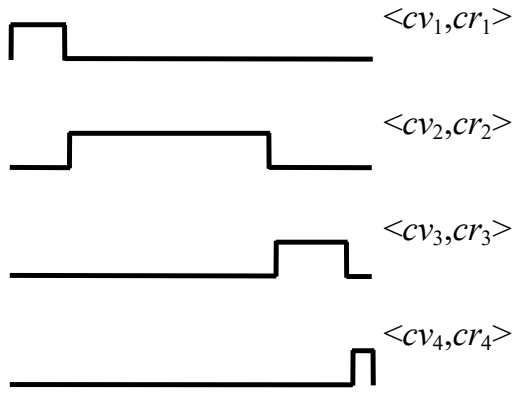


4.2 Matching-basierte Analyse

- Erweiterung: Adaptive Piecewise Constant Approximation (APCA)
  - Motivation
    - Viele Zeitreihen haben Bereiche mit geringer Detailmenge (wenige Informationen) und Bereiche mit hoher Detailmenge
    - PAA approximiert jeden Bereich mit derselben Detailtiefe



  - Idee
    - Verwende Basisboxen mit unterschiedlicher Länge
    - Speichere jedes Segment mit 2 Werten (vorher 1)



  - Vorteile
    - Schnell und einfach zu berechnen
    - Unterstützt alle Arten von Distanzfunktionen
    - Unterstützt Zeitreihen verschiedener Länge
  - Nachteil
    - Relativ komplexe Implementierung

4.2 Matching-basierte Analyse

- Piecewise Linear Approximation (PLA)
  - Idee
    - Repräsentiere Zeitreihen als Sequenz von Linien-Segmenten
      S= (Länge,linke Höhe, rechte Höhe)
    - Zwei aufeinanderfolgender Segmente können verbunden sein (müssen aber nicht), dann genügt
      S= (Länge,linke Höhe)
  - Berechnungskomplexität
    - Optimale Lösung benötigt $O(n^2N)$ => für viele Anwendungen zu langsam
    - Lineare Laufzeit durch Verwendung von Heuristiken möglich.
  - Vorteile
    - Geeignet für „natürliche“ Signale
    - Schnell und einfach zu berechnen (nicht optimal)
    - Unterstützt alle Arten von Distanzfunktionen (insb. gewichtete Distanzfunktionen)
  - Nachteil:
    - Keine Indexstruktur für PLA bekannt (aber: schneller sequentieller Scan möglich)