Data Mining:
Concepts and Techniques

— Chapter 8 —

8.2 Mining time-series data

Jiawei Han and Micheline Kamber
Department of Computer Science
University of Illinois at Urbana-Champaign
www.cs.uiuc.edu/~hanj


THOMAS M. SIEBEL CENTER FOR COMPUTER SCIENCE

Chapter 8. Mining Stream, Time-Series, and Sequence Data

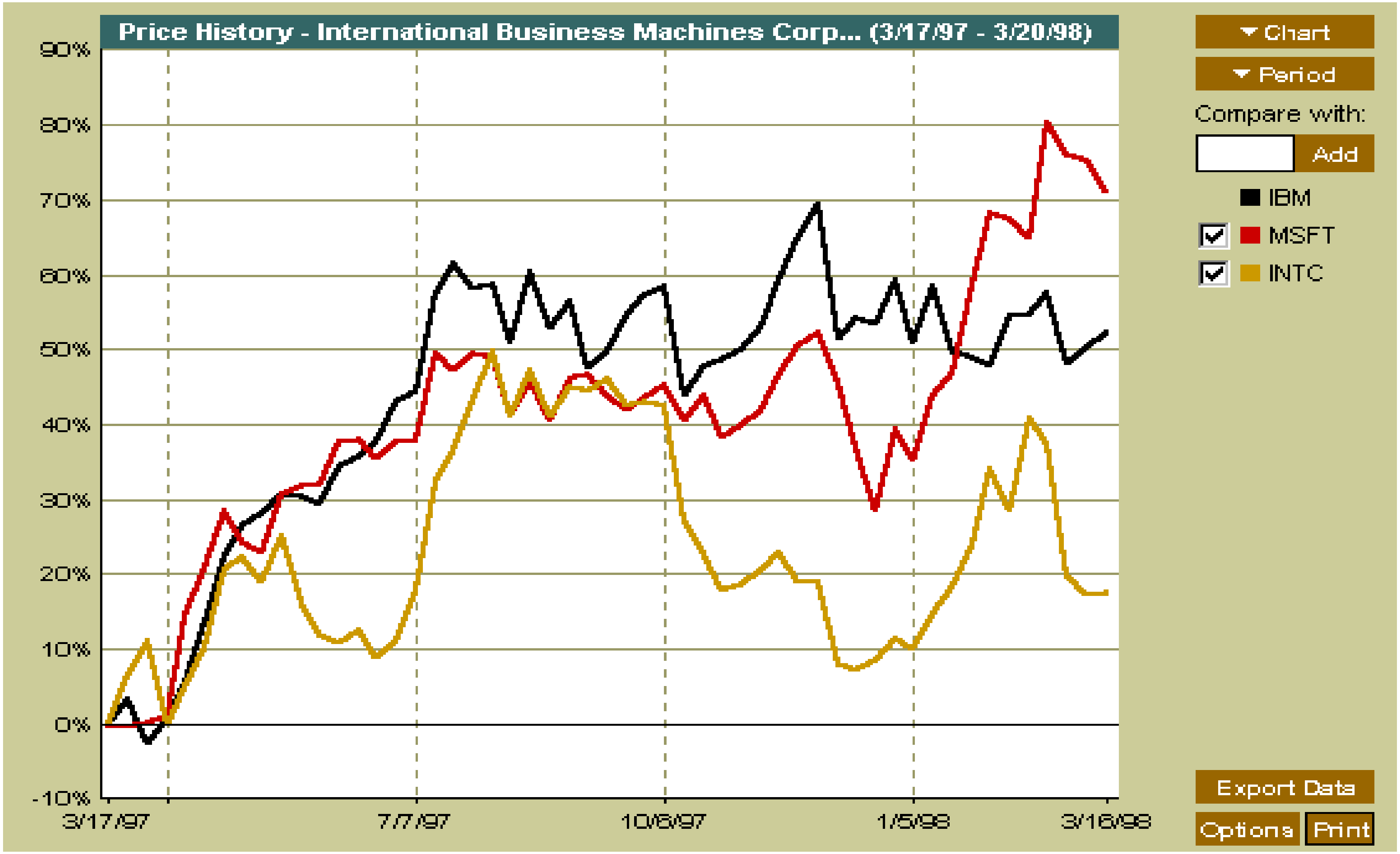
- **Mining data streams**
- **Mining time-series data**
- **Mining sequence patterns in transactional databases**
- **Mining sequence patterns in biological data**

Time-Series and Sequential Pattern Mining

- Regression and trend analysis—A statistical approach
- Similarity search in time-series analysis
- Sequential Pattern Mining
- Markov Chain
- Hidden Markov Model

Mining Time-Series Data

- Time-series database
  - Consists of sequences of values or events changing with time
  - Data is recorded at regular intervals
  - Characteristic time-series components
    - Trend, cycle, seasonal, irregular
- Applications
  - Financial: stock price, inflation
  - Industry: power consumption
  - Scientific: experiment results
  - Meteorological: precipitation



- A time series can be illustrated as a time-series graph which describes a point moving with the passage of time

Categories of Time-Series Movements

- Categories of Time-Series Movements
  - Long-term or trend movements (trend curve): general direction in which a time series is moving over a long interval of time
  - Cyclic movements or cycle variations: long term oscillations about a trend line or curve
    - e.g., business cycles, may or may not be periodic
  - Seasonal movements or seasonal variations
    - i.e, almost identical patterns that a time series appears to follow during corresponding months of successive years.
  - Irregular or random movements
- Time series analysis: decomposition of a time series into these four basic movements
  - Additive Modal: $TS = T + C + S + I$
  - Multiplicative Modal: $TS = T \times C \times S \times I$

Estimation of Trend Curve

- The freehand method
  - Fit the curve by looking at the graph
  - Costly and barely reliable for large-scaled data mining
- The least-square method
  - Find the curve minimizing the sum of the squares of the deviation of points on the curve from the corresponding data points
- The moving-average method

Moving Average

- Moving average of order n

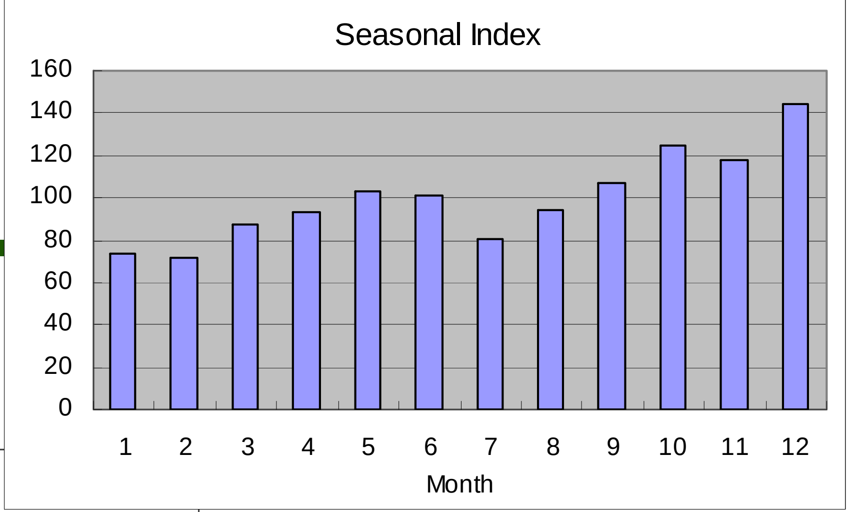
$$\frac{y_1 + y_2 + \cdots + y_n}{n}, \frac{y_2 + y_3 + \cdots + y_{n+1}}{n}, \frac{y_3 + y_4 + \cdots + y_{n+2}}{n}, \cdots$$

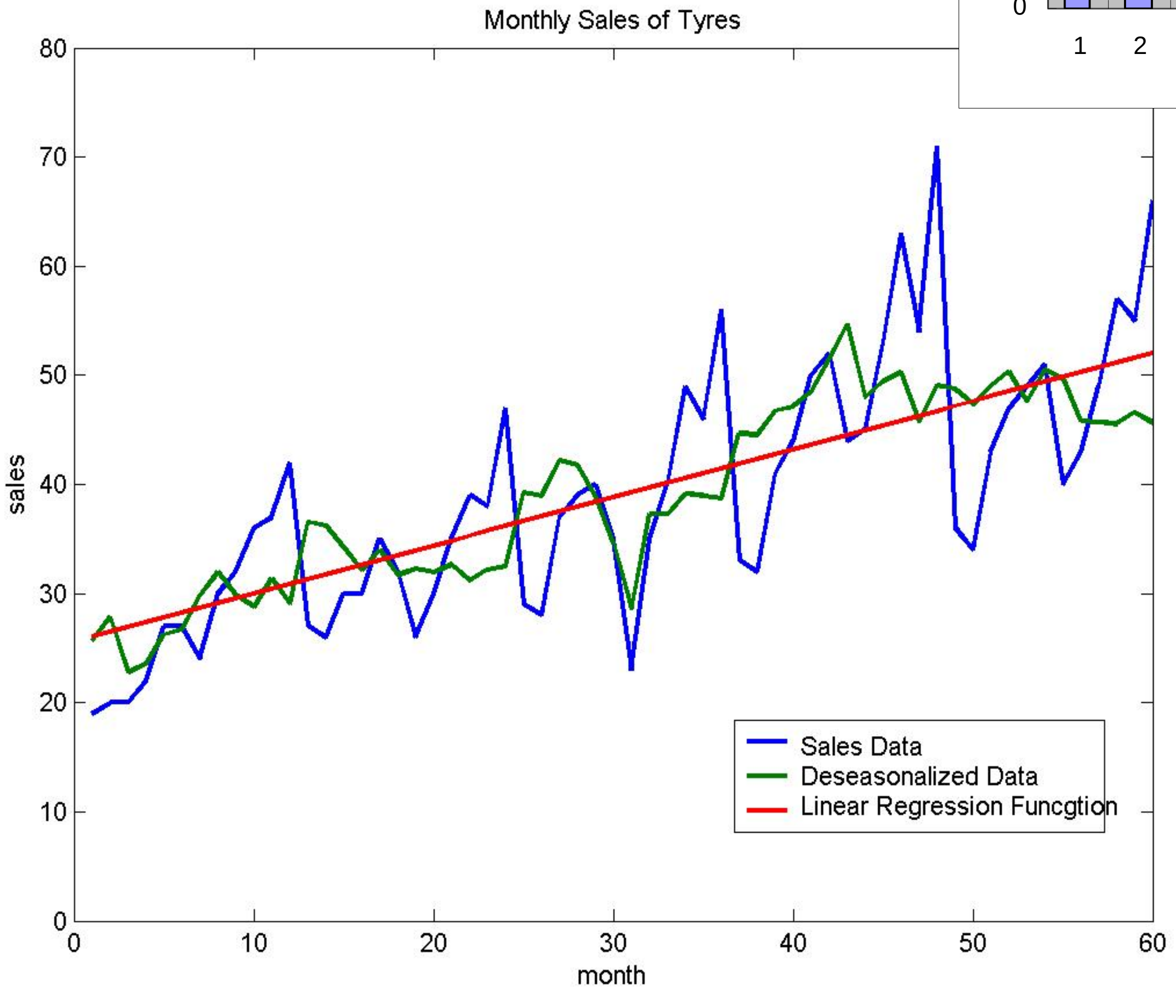
  - Smoothes the data
  - Eliminates cyclic, seasonal and irregular movements
  - Loses the data at the beginning or end of a series
  - Sensitive to outliers (can be reduced by weighted moving average)

Trend Discovery in Time-Series (1): Estimation of Seasonal Variations

- Seasonal index
  - Set of numbers showing the relative values of a variable during the months of the year
  - E.g., if the sales during October, November, and December are 80%, 120%, and 140% of the average monthly sales for the whole year, respectively, then 80, 120, and 140 are seasonal index numbers for these months
- Deseasonalized data
  - Data adjusted for seasonal variations for better trend and cyclic analysis
  - Divide the original monthly data by the seasonal index numbers for the corresponding months

Seasonal Index





Raw data from http://www.bbk.ac.uk/manop/man/docs/QII_2_2003%20Time%20series.pdf

Trend Discovery in Time-Series (2)

- Estimation of cyclic variations
  - If (approximate) periodicity of cycles occurs, cyclic index can be constructed in much the same manner as seasonal indexes
- Estimation of irregular variations
  - By adjusting the data for trend, seasonal and cyclic variations
- With the systematic analysis of the trend, cyclic, seasonal, and irregular components, it is possible to make long- or short-term predictions with reasonable quality

Time-Series & Sequential Pattern Mining

- Regression and trend analysis—A statistical approach
- Similarity search in time-series analysis
- Sequential Pattern Mining
- Markov Chain
- Hidden Markov Model

Similarity Search in Time-Series Analysis

- Normal database query finds exact match
- Similarity search finds data sequences that differ only slightly from the given query sequence
- Two categories of similarity queries
  - Whole matching: find a sequence that is similar to the query sequence
  - Subsequence matching: find all pairs of similar sequences
- Typical Applications
  - Financial market
  - Market basket data analysis
  - Scientific databases
  - Medical diagnosis

Data Transformation

- Many techniques for signal analysis require the data to be in the frequency domain
- Usually data-independent transformations are used
  - The transformation matrix is determined a priori
    - discrete Fourier transform (DFT)
    - discrete wavelet transform (DWT)
- The distance between two signals in the time domain is the same as their Euclidean distance in the frequency domain

Discrete Fourier Transform

from $\vec{x} = [x_t], t = 0, \ldots, n-1$ to $\vec{X} = [X_f], f = 0, \ldots, n-1$:

$$X_f = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x_t \exp(-j2\pi ft/n), \; f = 0, 1, \ldots, n-1$$

- DFT does a good job of concentrating energy in the first few coefficients
- If we keep only first a few coefficients in DFT, we can compute the lower bounds of the actual distance
- Feature extraction: keep the first few coefficients (F-index) as representative of the sequence

DFT (continued)

- Parseval's Theorem

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

- The Euclidean distance between two signals in the time domain is the same as their distance in the frequency domain
- Keep the first few (say, 3) coefficients underestimates the distance and there will be no false dismissals!

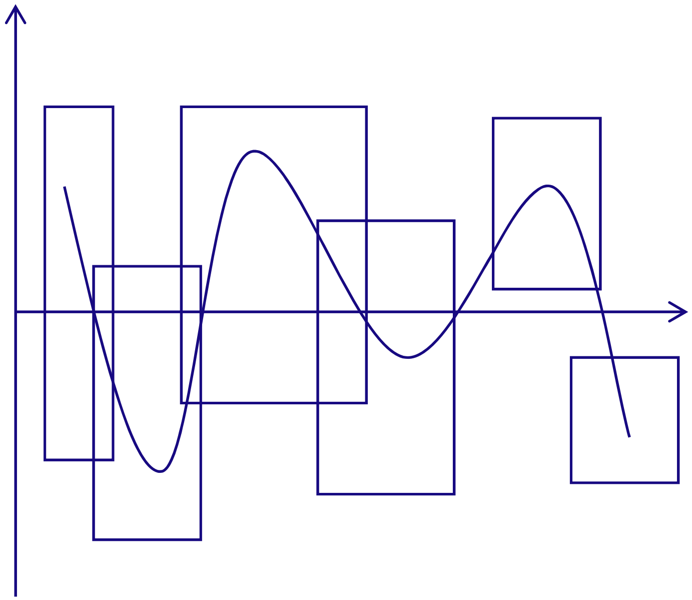
$$\sum_{t=0}^{n} | S[t] - Q[t] |^2 \le \varepsilon \Rightarrow \sum_{f=0}^{3} | F(S)[f] - F(Q)[f] |^2 \le \varepsilon$$

Multidimensional Indexing in Time-Series

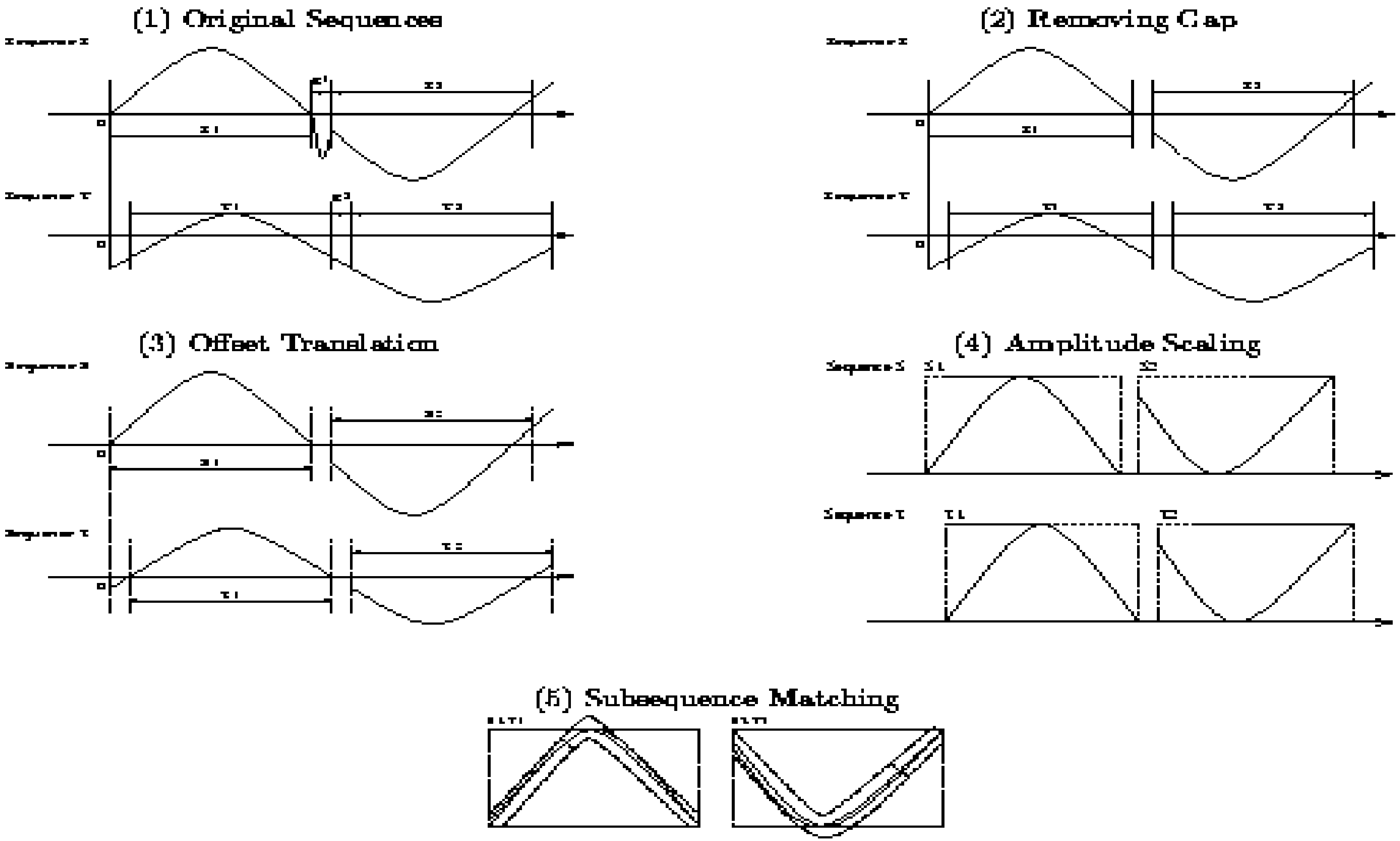
- Multidimensional index construction
  - Constructed for efficient accessing using the first few Fourier coefficients
- Similarity search
  - Use the index to retrieve the sequences that are at most a certain small distance away from the query sequence
  - Perform post-processing by computing the actual distance between sequences in the time domain and discard any false matches

Subsequence Matching

- Break each sequence into a set of pieces of window with length *w*
- Extract the features of the subsequence inside the window
- Map each sequence to a "trail" in the feature space
- Divide the trail of each sequence into "subtrails" and represent each of them with minimum bounding rectangle
- Use a multi-piece assembly algorithm to search for longer sequence matches



Analysis of Similar Time Series



Enhanced Similarity Search Methods

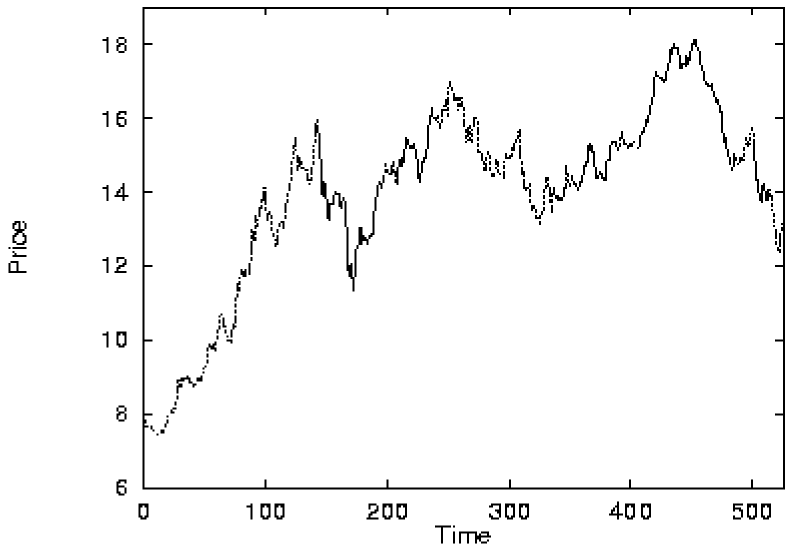
- Allow for gaps within a sequence or differences in offsets or amplitudes
- Normalize sequences with amplitude scaling and offset translation
- Two subsequences are considered similar if one lies within an envelope of ε width around the other, ignoring outliers
- Two sequences are said to be similar if they have enough non-overlapping time-ordered pairs of similar subsequences
- Parameters specified by a user or expert: sliding window size, width of an envelope for similarity, maximum gap, and matching fraction

Steps for Performing a Similarity Search

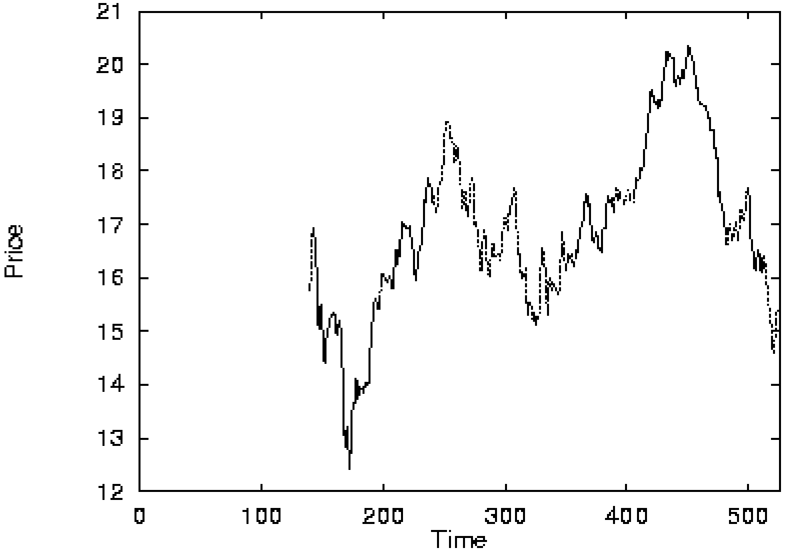
- Atomic matching
  - Find all pairs of gap-free windows of a small length that are similar
- Window stitching
  - Stitch similar windows to form pairs of large similar subsequences allowing gaps between atomic matches
- Subsequence Ordering
  - Linearly order the subsequence matches to determine whether enough similar pieces exist

Similar Time Series Analysis

VanEck International Fund



Fidelity Selective Precious Metal and Mineral Fund



Two similar mutual funds in the different fund group

Query Languages for Time Sequences

- Time-sequence query language
  - Should be able to specify sophisticated queries like

  Find all of the sequences that are similar to some sequence in class *A*, but not similar to any sequence in class *B*

  - Should be able to support various kinds of queries: range queries, all-pair queries, and nearest neighbor queries
- Shape definition language
  - Allows users to define and query the overall shape of time sequences
  - Uses human readable series of sequence transitions or macros
  - Ignores the specific details
    - E.g., the pattern up, Up, UP can be used to describe increasing degrees of rising slopes
    - Macros: spike, valley, etc.

References on Time-Series & Similarity Search

- R. Agrawal, C. Faloutsos, and A. Swami. Efficient similarity search in sequence databases. FODO'93 (Foundations of Data Organization and Algorithms).
- R. Agrawal, K.-I. Lin, H.S. Sawhney, and K. Shim. Fast similarity search in the presence of noise, scaling, and translation in time-series databases. VLDB'95.
- R. Agrawal, G. Psaila, E. L. Wimmers, and M. Zait. Querying shapes of histories. VLDB'95.
- C. Chatfield. The Analysis of Time Series: An Introduction, 3rd ed. Chapman & Hall, 1984.
- C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time-series databases. SIGMOD'94.
- D. Rafiei and A. Mendelzon. Similarity-based queries for time series data. SIGMOD'97.
- Y. Moon, K. Whang, W. Loh. Duality Based Subsequence Matching in Time-Series Databases, ICDE'02
- B.-K. Yi, H. V. Jagadish, and C. Faloutsos. Efficient retrieval of similar time sequences under time warping. ICDE'98.
- B.-K. Yi, N. Sidiropoulos, T. Johnson, H. V. Jagadish, C. Faloutsos, and A. Biliris. Online data mining for co-evolving time sequences. ICDE'00.
- Dennis Shasha and Yunyue Zhu. **High Performance Discovery in Time Series: Techniques and Case Studies**, SPRINGER, 2004