Time Series Data Mining (TSDM):

A New Temporal Pattern Identification Method for Characterization and Prediction of Complex Time Series Events

Xin Feng, Ph.D.
Department of Electrical and Computer Engineering
Marquette University
Milwaukee, Wisconsin 53233, USA
(414)288-3504
xin.feng@mu.edu

November 2013

An Ancient Chinese Say …

- **You are not "grown-ups" till you reach the age of 30.**
- **You are not "free from doubts" till the age of 40.**
- **You do not know God's Will till the age of 50.**
- **The bottom line:**

Learning is a life-long process

11/4/2013 2

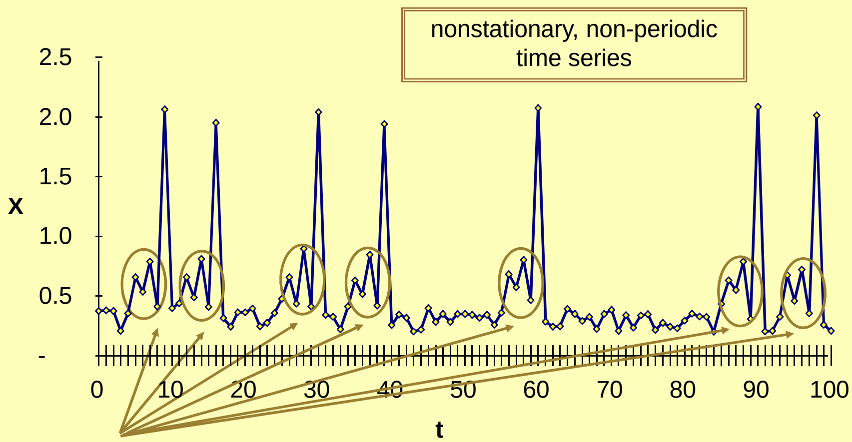
Overview of Presentation

- **Problem Statement**
  - Graphical Problem Statement
  - Time Series Analysis Literature
  - Innovative New Approach
- **Algorithm**
  - Phase Space
  - Mathematical Formulation
  - Algorithm Results
- **Applications**
  - Progression of Time Series
  - Engineering
  - Financial

Collaborators/Credits

- **Dr. Noveen Bansal, MSCS**
- **Dr. Richard Povinelli, EECE**
- **Hai Huang, Microsoft, Inc.**
- **Odilon K. Senyana, FAA**
- **Dr. Wenjing Zhang, Discover, Inc.**
- **Shaobo Wang, MU**

Graphical Problem Statement

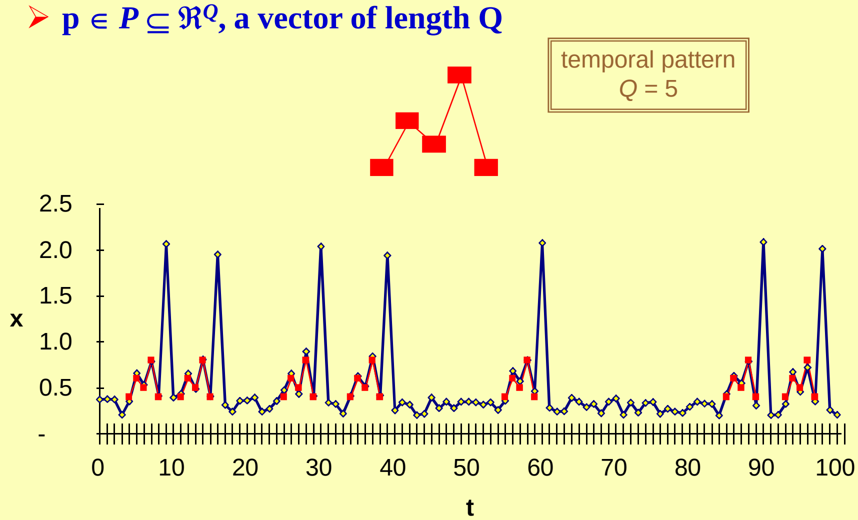


- Can we automatically characterize these sequences (temporal patterns)?
- Can we use such temporal patterns for prediction?

11/4/2013 5

Temporal Pattern

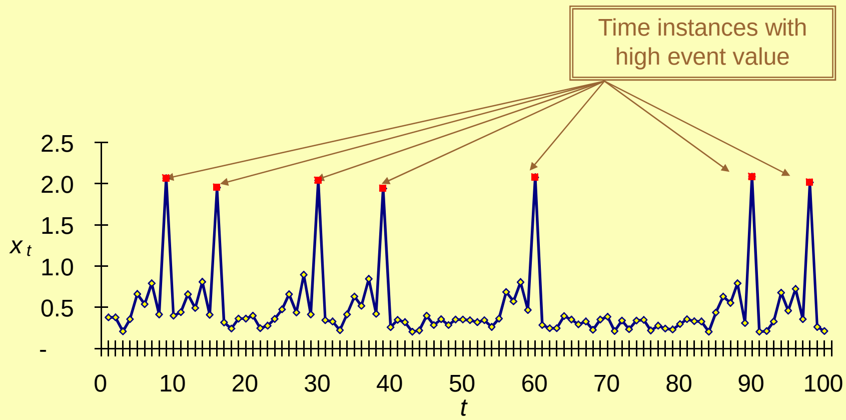
- Find temporal patterns
- $p \in P \subseteq \Re^Q$, a vector of length Q



11/4/2013 6

Events

- **Temporal patterns that characterize and predict events**
- **Chosen a priori**
  - Algorithm not restricted by event definition
  - Event definition is problem specific



Time Series Analysis Literature

- **Box-Jenkins, ARMA**
  - Pandit and Wu (1983)
  - Bowerman and O'Connell (1993)
- **Chaotic deterministic**
  - Takens (1981)
  - Sauer (1991)
  - Casdagli (1989)
  - Abarbanel (1990, 1994)
  - Ghoshray (1996)

An Innovative Approach

- **Time Series Data Mining**
  - Time Series, $X = \{x_t, t = 1, \ldots, N\}$
  - Find temporal patterns that are characteristic and predictive of an event $g(x_t)$
    $\mathbf{p} \in P \subseteq \Re^Q$
  - Nonstationary, non-periodic time series
  - Chaotic deterministic time series whose attractors are non-stationary
- **Local model, local prediction**
  - Not concerned with characterizing and predicting everywhere (every time)
  - Characterize and predict events
- **Applying Genetic Algorithm (GA) to find the "optimal" local model**
- **Exciting, new results with difficult real world time series**

11/4/2013 9

Introduction to Data Mining

- **A step in the knowledge discovery process**
- **Application of algorithms to extract *meaningful* patterns**

11/4/2013 10

Data Mining and Knowledge Discovery

- **"The nontrivial extraction of implicit, previously unknown, and potentially useful information from data"[1]**
- **Uses artificial intelligence, statistical and visualization techniques to discover and present knowledge in a form which is easily comprehensible to humans.**

[1] W. Frawley and G. Piatetsky-Shapiro and C. Matheus, Knowledge Discovery in Databases: An Overview. AI Magazine, Fall 1992, pgs 213-228.

11/4/2013 11

Data Mining and Knowledge Discovery

More recently (already obsolete):

- *“The huge amount of tracking data available from web sites as an "information gold mine.”*
  - **Business Week**
- “74% of large companies expect to mine web data to increase profits by 2002.”
  - **Forrester Research (1998)**

11/4/2013 12

Popular Data Mining Problems

- Associating – Identifies patterns or groups of items

  "men who buy red ties also often buy cigars."

- Classifying – Identifies clusters of items with common attributes

  "men who buy red ties and cigars also usually have wine at lunch and pay by credit card. "

Popular Data Mining Problems

- Sequencing – Identifies the order of events
  " men tend to buy red ties before lunch and cigars after lunch."

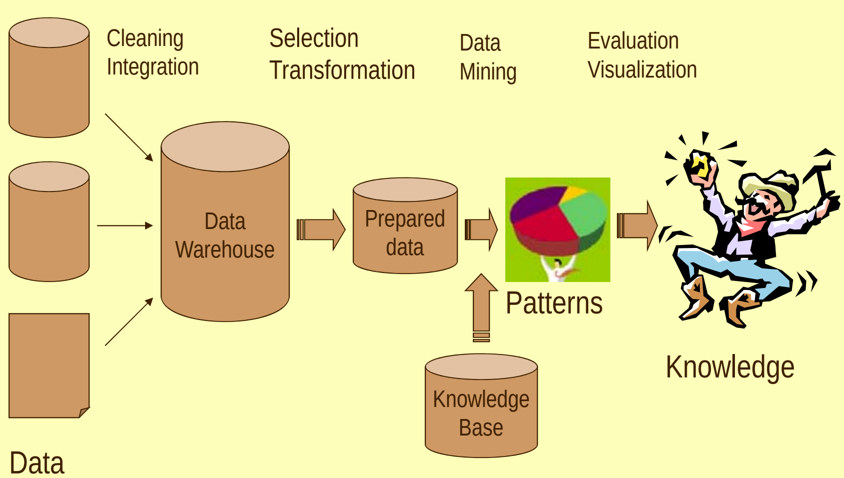
- Predictive Modeling – Identifies a likely outcome from item clusters.
  "men who buy red ties and cigars and have wine at lunch are very likely to buy a silver Benz within two years and finance their purchase by borrowing money from bank"

Gold Mining Analogy

- **Where do prospectors search for the gold?**
  - Geological formation
    - Quartz and ironstone
    - Structures such as banded iron formations
  - Data Mining
    - Define formations that point to nuggets of information
    - Define patterns that identify an information strike
- **Definition of "gold"**
  - For Gold Mining
    - Size of nuggets makes a difference
    - Mining for oil or silver is different
  - Data Mining
    - Definition of knowledge
    - Clearly define the desired nuggets of information

11/4/2013 15

Knowledge Discovery in Databases



11/4/2013 16

Features of Data Mining

- **Intend to discover knowledge which are previously unknown**
- **A Multi-disciplinary field growing out of many areas**
  - Mathematical modeling and statistics
  - Pattern recognition
  - Computational intelligence

11/4/2013 17

Data Mining Tools

- **Traditional pattern recognition algorithms:**
  - Statistics, Mathematical Modeling, Algorithms.
  - Data Visualization, Graphics
- **Intelligent computing:**
  - Artificial Neural networks, Fuzzy Logic, Genetic Algorithms/Evolutionary Computing, Expert Systems, Natural Language Processing, etc.

They all use "mechanical" computing power:

- Database Systems; Client-Server; Internet/WWW; Software/programming; etc.

11/4/2013 18

Chaotic Deterministic Time Series

- **No universal definition**
- **Characterized by the following criteria**
  - Sensitivity to initial conditions
  - Positive Lyapunov exponent
  - Broadband Fourier spectra
  - Finite, possibly fractional attractors

11/4/2013 19

Attractors

- **Given a manifold M and a map f: $M \to M$, an invariant set S is defined as follows**

$$S = \left\{ x_0 : x_0 \in S,\ f^{\,n}\left(x_0\right) \in S, \forall n \right\} \subset M$$

- **A positively invariant set requires that n ≥ 0.**
- **A set $A \subset M$ is an attracting set if A is a closed invariant set and there exists a neighborhood U of A such that U is a positively invariant set and**

$$f^{\,n}\left(x\right) \to A \quad \forall \quad x \in U$$

- **An attractor is defined as an attracting set that contains a dense orbit.**

N. B. Tufillaro, T. Abbott, and J. Reilly, An Experimental Approach to Nonlinear Dynamics and Chaos. Redwood City, CA: Addison-Wesley, 1992.

11/4/2013 20

Gold Mining Analogy

- **Where do prospectors search for the gold?**
  - Geological formations
    - Quartz and ironstone
    - Structures such as banded iron formations
  - Data Mining
    - Define formations that point to nuggets of information (events)
    - Define temporal patterns that identify an information strike.
- **Definition of gold**
    - Size of nuggets makes a difference in how the mining is approached.
    - Mining for oil or silver is different
  - Data Mining
    - Definition of knowledge
    - Clearly define the desired nuggets of information (events)
    - Define event function $g(x_t)$

11/4/2013 21

Two ideas arise from the analogy

- **Patterns (Temporal Patterns)**
  - Patterns that guide and direct to the nuggets of information need to be understood and identified
  - Temporal patterns and time series embedding
- **Gold (Events)**
  - Nuggets of information require clear definition
  - Time series events
- **Relationship between a temporal pattern and event**
  - Find "temporal patterns" that indicate a high event value
  - Example
    - A sequence of charge flow values in the brain that precede a physical response
    - A series of voltage values that precede the release of a droplet of metal from a welder
    - A sequence of stock prices that precede a rapid increase in the stock price

11/4/2013 22

Presentation Status

- **Problem Statement**
  - Graphical Problem Statement
  - Time Series Analysis Literature
  - Innovative New Approach
- **Algorithm**
  - Phase Space
  - Mathematical Formulation
  - Algorithm Results
- **Applications**
  - Progression of Time Series
  - Engineering
  - Financial

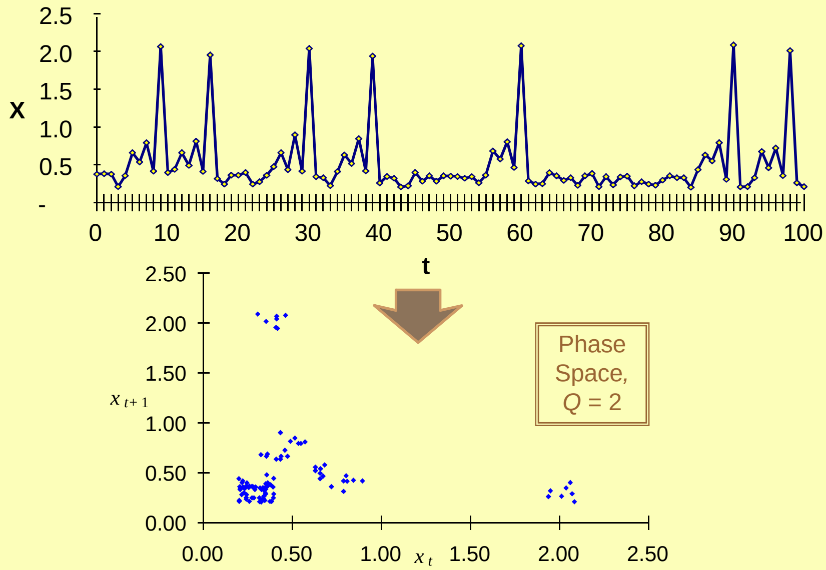
11/4/2013 23

Algorithm

- **Search for the optimal temporal pattern that yields the maximal "event" values**
- **Choose the support, *Q*, of the temporal pattern p**
- **Define the event function, *g***
- **Embed the time series into a phase space**
  - Define a metric to allow comparison between the temporal pattern and embedded time series
- **Find the "optimal" pattern cluster**
  - Pattern neighborhood that contains the maximum average "eventness"
  - Genetic Algorithm

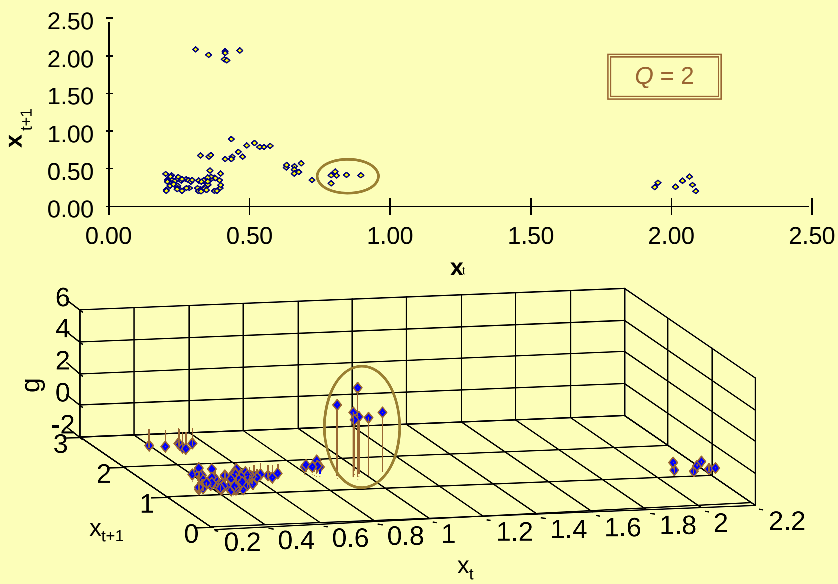
11/4/2013 24

Graphical Mapping



11/4/2013 25

Phase Space with Event Values



11/4/2013 26

Mathematical Formulation I

- **Non-stationary training time series**
  - *N* is the length of the series

$$X = \{x_t, t = 1, \ldots, N\}$$

- **Testing time series**

$$Y = \{x_t, t = R, \ldots, S\}, \quad R > N$$

- **Define Event Function**
  - Inputs must chosen carefully
  - Relationship to temporal pattern is important
  - Example *g*
    - The percentage change in the time series between times $t+Q$ and $t+Q+1$
    - Q is the length of the temporal pattern

$$g(x_t) = \frac{x_{t+Q} - x_{t+Q-1}}{x_{t+Q-1}}$$

11/4/2013 27

Mathematical Formulation II

- **Define $P \subseteq \Re^Q$**
  - A Q dimensional real space
  - Phase space
- **Embed X into P, likewise for Y create $\mathbf{y}_t$**
  - Subsets, $\mathbf{x}_t$, of X of cardinality Q
  - Chosen by any consistent rule

$$\mathbf{x}_t^T = \left(x_t, x_{t+\tau_1}, \ldots, x_{t+\tau_{Q-1}}\right), \qquad t = 1, \ldots, N - \tau_{Q-1}$$

$$\tau_1 < \tau_2 < \ldots < \tau_{Q-1}$$

- **Define metric d for P**
  - $d(\mathbf{p}, \mathbf{x}_t)$
  - Compare the embedded time series and the temporal pattern
  - $\mathbf{p}$ - temporal pattern, a vector of length Q
  - $\mathbf{x}_t$ - embedded time series

11/4/2013 28

Property 1

- **Find a temporal pattern p $\in P$ and a threshold $\delta \in \Re$ that have the following two properties**
- **Property 1**
  - Given the following definitions

$$M_{train} = \{t : d(\mathbf{p}, \mathbf{x}_t) \le \delta\}, \qquad t = 1, \ldots, N - \tau_{Q-1}$$

$$\mu_{M_{train}} = \frac{1}{c(M_{train})} \sum_{t \in M_{train}} g(x_t)$$

$$\mu_X = \frac{1}{N - Q + 1} \sum_{t=1}^{N-Q+1} g(x_t)$$

  - That $\mu_{M_{train}} > \mu_X$
  - and the set $\{g(x_t) : t \in M_{train}\}$

  is statistically different from the set $\{g(x_t) : t = 1, \ldots, N - Q + 1\}$

11/4/2013 29

Property 2

- **Property 2**
  - Given the following definitions

$$M_{test} = \{t : d(\mathbf{p}, \mathbf{y}_t) \le \delta\}, \qquad t = R, \ldots, S - \tau_{Q-1}$$

$$\mu_{M_{test}} = \frac{1}{c(M_{test})} \sum_{t \in M_{test}} g(x_t)$$

$$\mu_Y = \frac{1}{S - Q - R + 2} \sum_{t=R}^{S-Q+1} g(x_t)$$

  - That $\mu_{M_{test}} > \mu_Y$
  - and that the set $\{g(x_t) : t \in M_{test}\}$
  - is statistically different from the set $\{g(x_t) : t = R, \ldots, S - Q + 1\}$

11/4/2013 30

Mathematical Formulation III

- **Define threshold δ**
  - In terms of normalized threshold δ_d
  - Use the mean and standard deviation statistics of $d(\mathbf{p}, \mathbf{x}_t)$

$$\mu_d = \frac{1}{N - \tau_{Q-1}} \sum_{t=1}^{N-\tau_{Q-1}} d(\mathbf{p}, \mathbf{x}_t)$$

$$\sigma_d^2 = \frac{1}{N - \tau_{Q-1}} \sum_{t=1}^{N-\tau_{Q-1}} \left(d(\mathbf{p}, \mathbf{x}_t) - \mu_d\right)^2$$

$$\delta = \mu_d + \delta_d \sigma_d$$

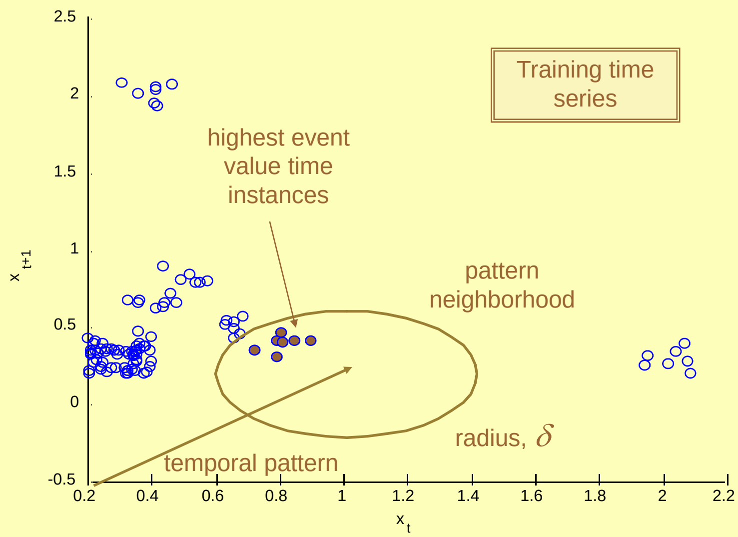
- **Optimization formulation**
  - Constraint forces cluster to be greater than 1 in size

$$\underset{\mathbf{p},\delta}{\text{maximize}} \quad f(\mathbf{p}, \delta, X, g) \equiv \mu_{M_{train}} = \frac{1}{c(M)} \sum_{t \in M_{train}} g(x_t)$$

$$\text{subject to} \quad c(M_{train}) > \beta N, \quad 0 < \beta \le 1.$$

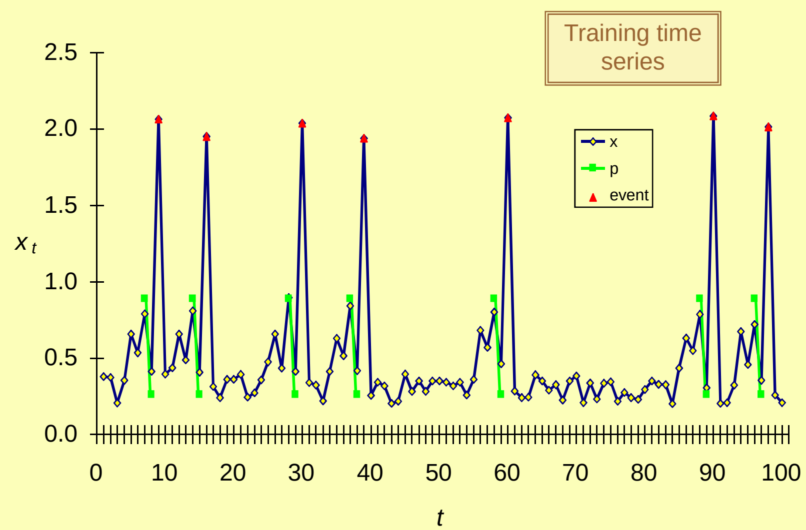
11/4/2013 31

Algorithm Results



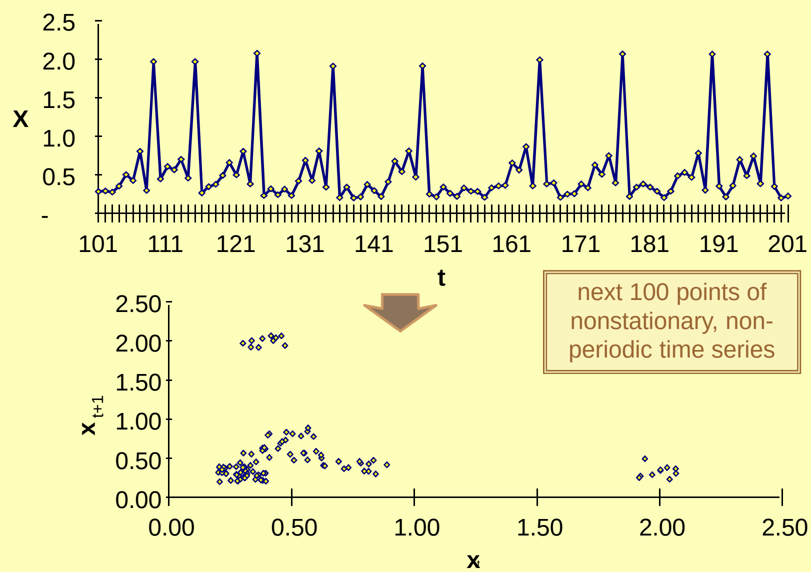
11/4/2013 32

Time Series View of Results



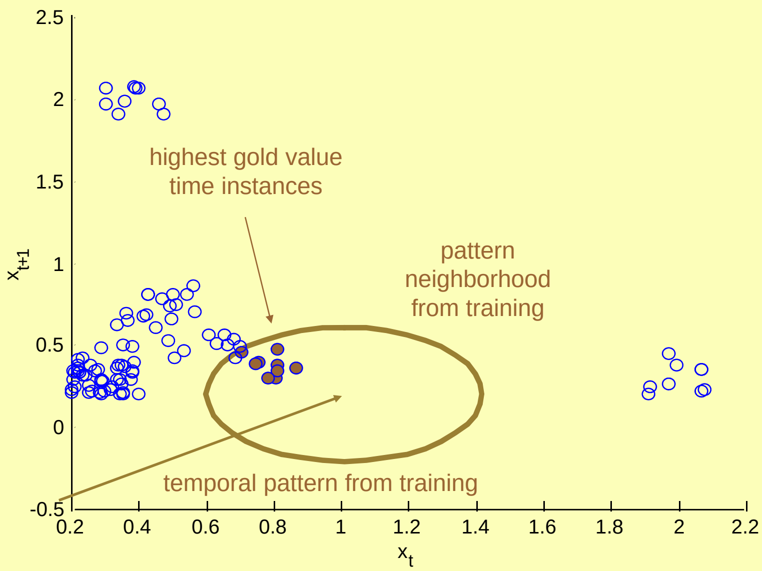
11/4/2013 33

Test Series Phase Space



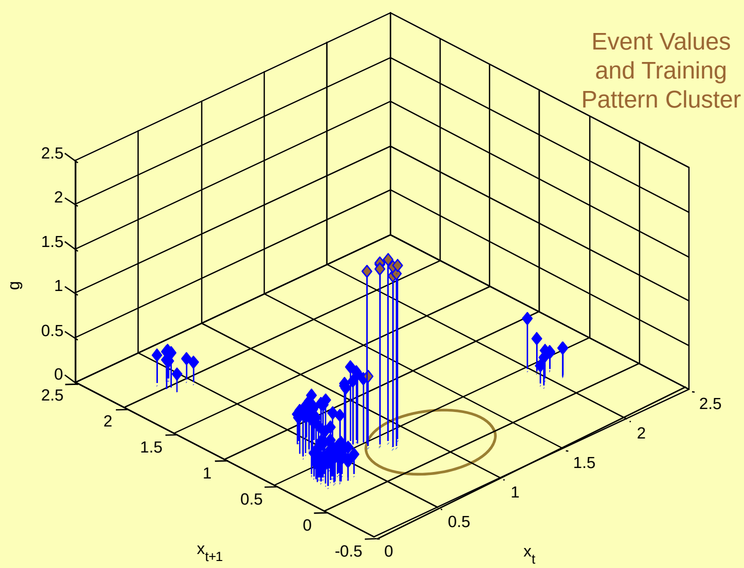
11/4/2013 34

Test Series Results



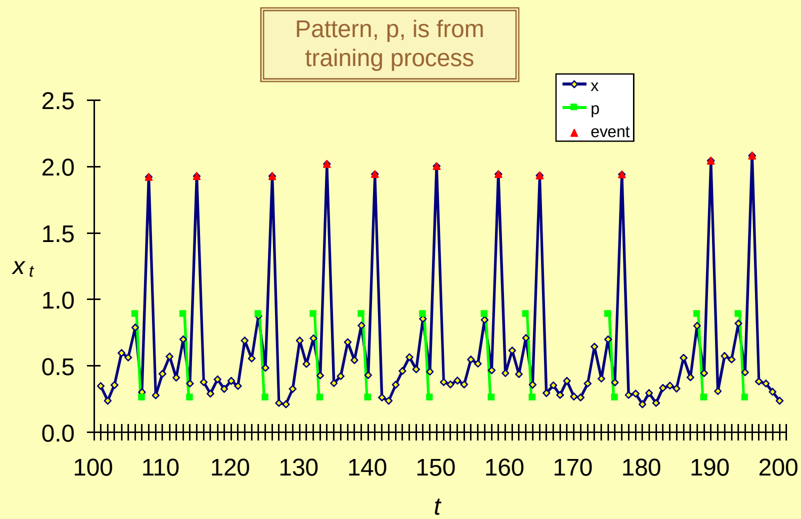
11/4/2013 35

Testing Phase Space



11/4/2013 36

Time Series View of Testing



11/4/2013 37

Presentation Status

- **Problem Statement**
  - Graphical Problem Statement
  - Time Series Analysis Literature
  - Innovative New Approach
- **Algorithm**
  - Phase Space
  - Mathematical Formulation
  - Algorithm Results
- **Applications**
  - Progression of Time Series
  - Engineering
  - Financial
- **Proposed Work**

11/4/2013 38

Applications

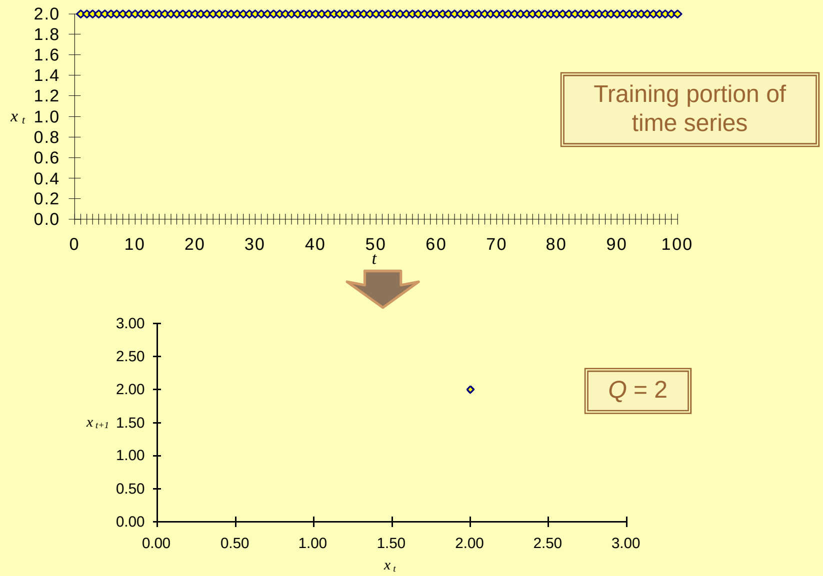
- **Set of progressively more complex time series**
  - Show how each is embedded into the phase space
  - Show results of algorithm
  - Constant
  - Sinusoidal
  - Chirp
  - Random noise
- **Engineering Application**
  - Welding Time Series
- **Financial Application**
  - Stock Open Price

Progression of Time Series

- **Choose the support of the pattern, *Q* = *2*, to allow graphical presentation**
- **Define the gold function**
  - The *t*+*Qth* value in the time series
  - Q is the length of the pattern

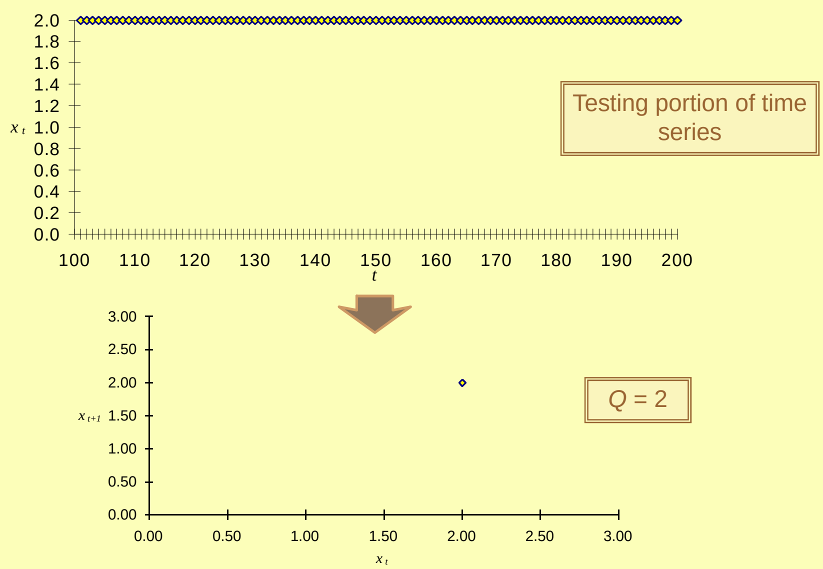
$$g\left(x_t\right) = B^{-(Q+1)} x_t$$

Constant Train Phase Space



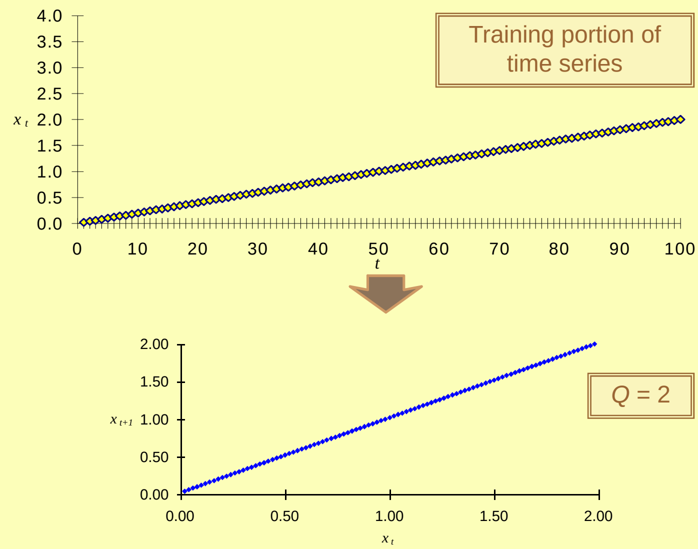
11/4/2013 41

Constant Test Phase Space



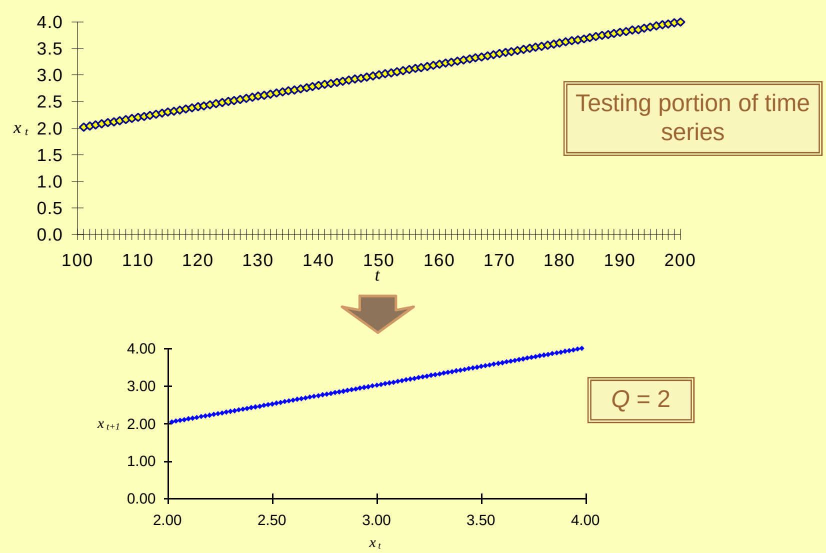
11/4/2013 42

Linearly Increasing Train



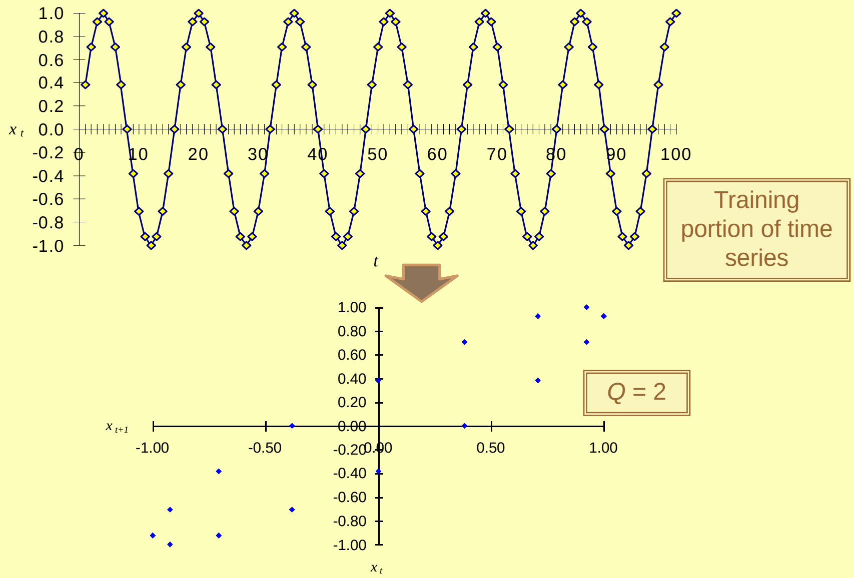
11/4/2013 43

Linearly Increasing Test



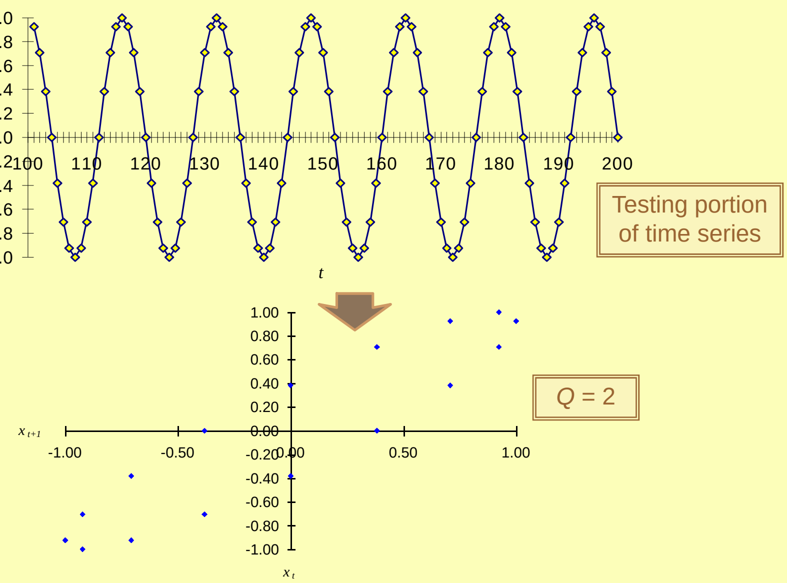
11/4/2013 44

Sinusoidal Train Phase Space



11/4/2013 45

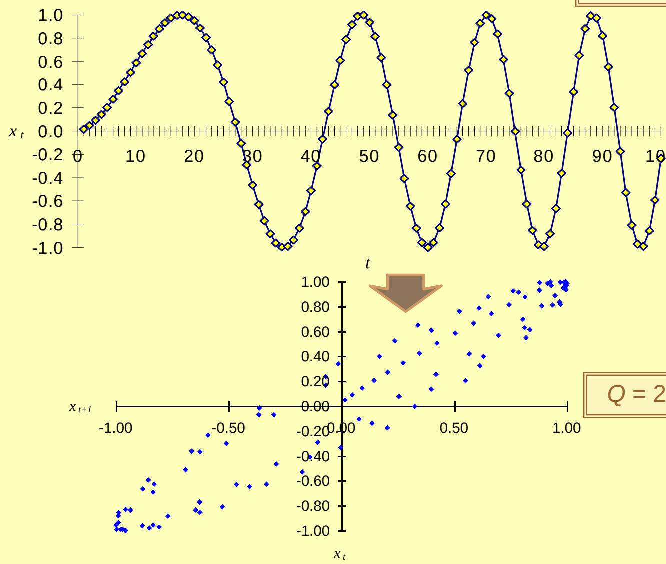
Sinusoidal Test Phase Space



11/4/2013 46

Chirp Train Phase Space

Training portion of time series

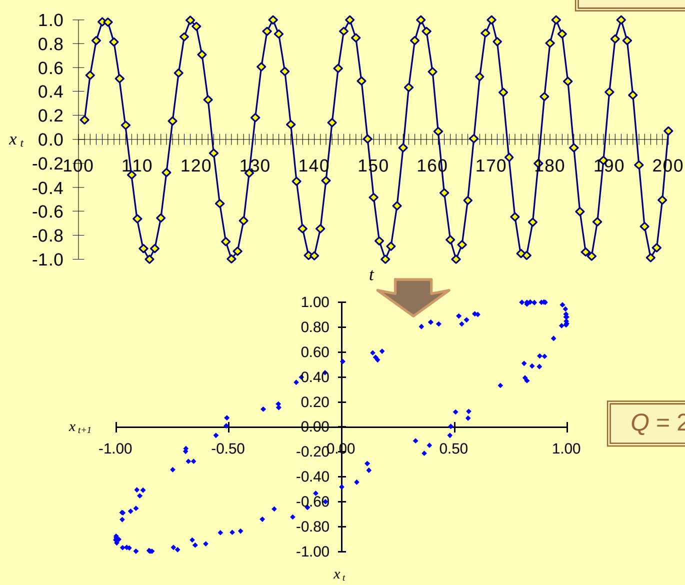


Q = 2

11/4/2013 47

Chirp Test Phase Space

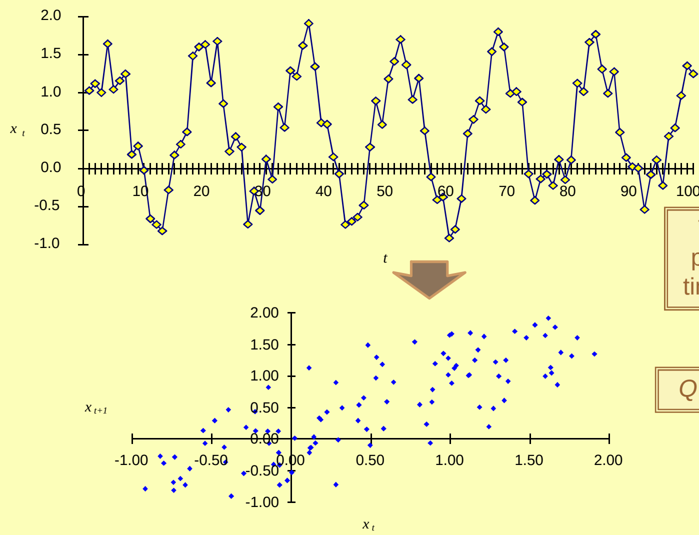
Testing portion of time series



Q = 2

11/4/2013 48

Noisy Sinusoidal Train Phase Space

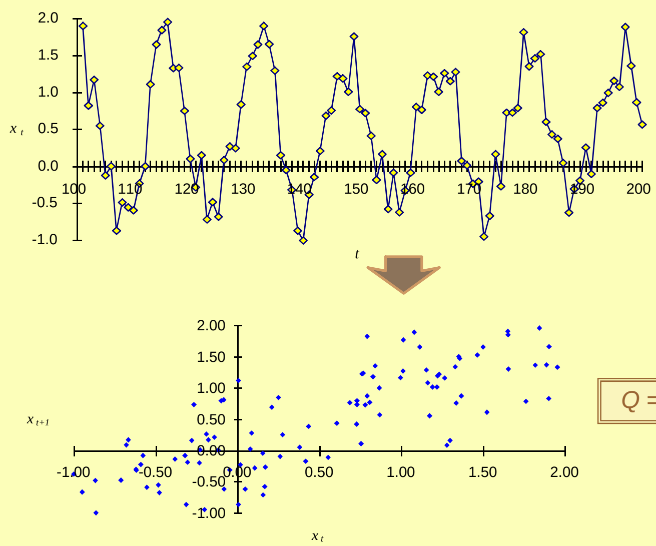


Training portion of time series

Q = 2

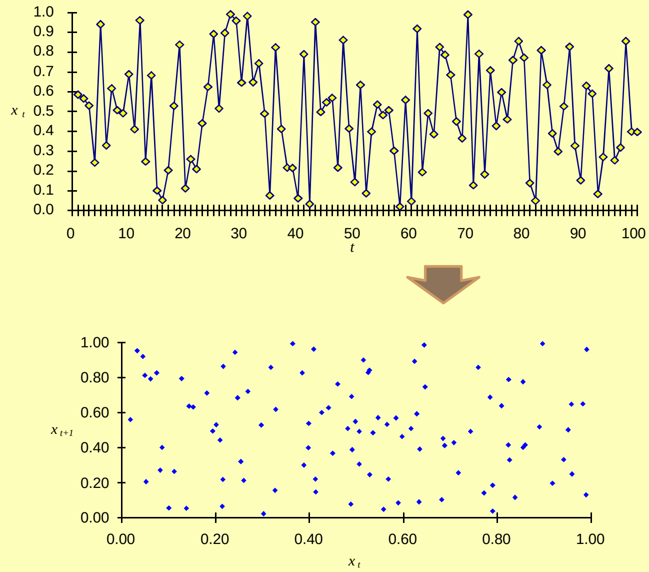
Noisy Sinusoidal Test Phase Space

Testing portion of time series



Q = 2

Noise Train Phase Space

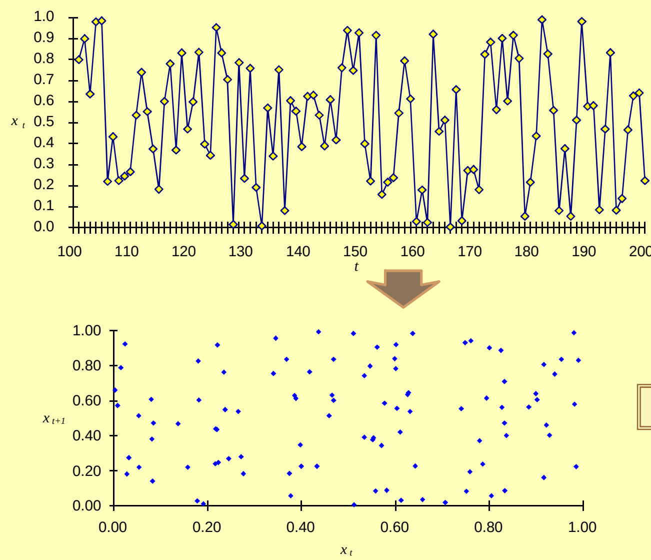


Uniform Density Random Variable

Q = 2

11/4/2013 51

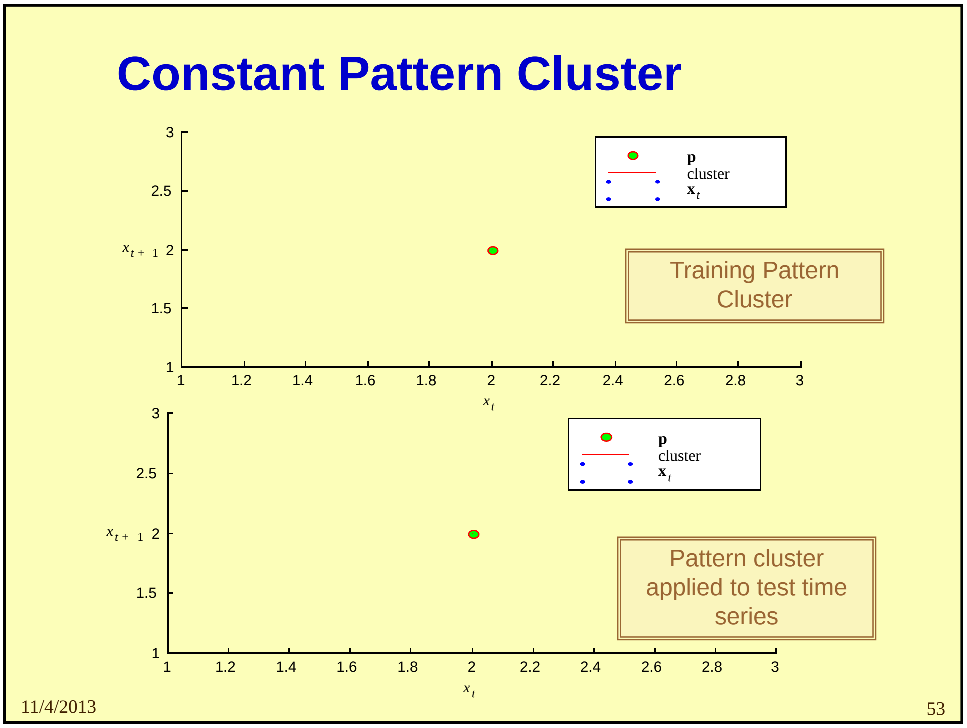
Noise Test Phase Space

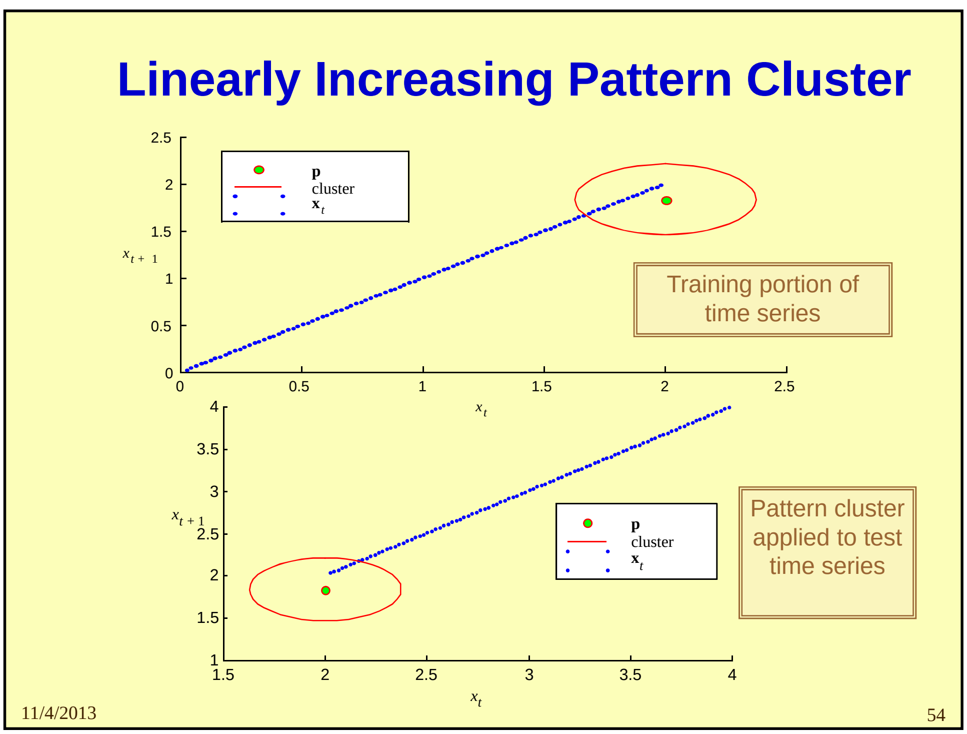


Testing portion of time series

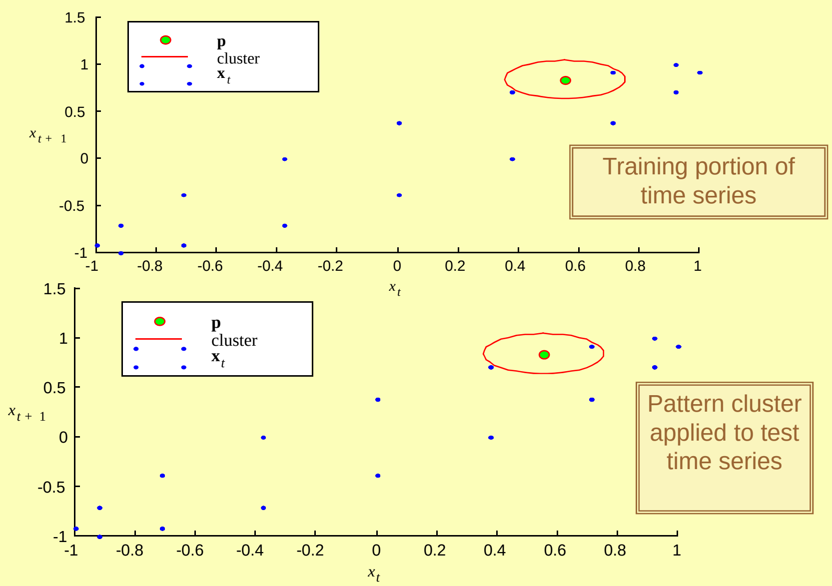
Q = 2

11/4/2013 52


Constant Pattern Cluster
p
cluster
x_t
Training Pattern Cluster
x_{t+1}
x_t
p
cluster
x_t
Pattern cluster applied to test time series
x_{t+1}
x_t
11/4/2013
53

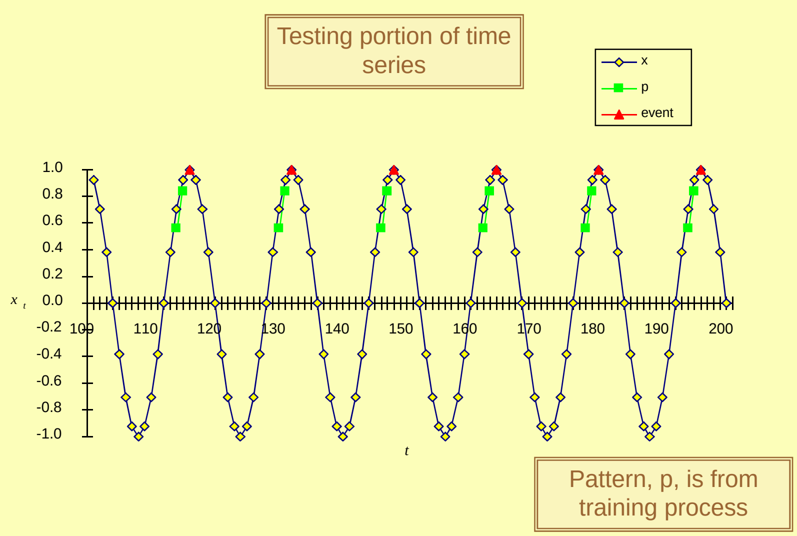

Linearly Increasing Pattern Cluster
p
cluster
x_t
Training portion of time series
x_{t+1}
x_t
p
cluster
x_t
Pattern cluster applied to test time series
x_{t+1}
x_t
11/4/2013
54

Sinusoidal Pattern Cluster



11/4/2013 55

Time Series View of Sinusoidal



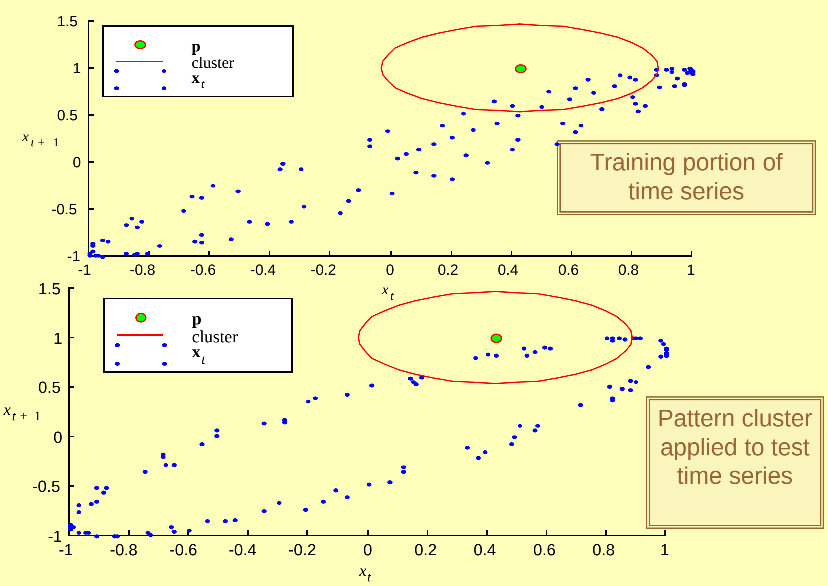
11/4/2013 56

Sinusoidal Statistical Tests

- **Runs Test**
  - H_0: There is no difference between the matched time series and the whole time series.
  - H_A: There is significant difference between the matched time series and the whole time series.
  - Using a 1% probability of Type I error ($\alpha = 0.01$).
  - $\alpha = 1.87e\text{-}018$ which means the null hypothesis can easily be rejected.
- **Difference of two independent means**
  - Although the two populations are probably dependent, this can be ignored because it makes the statistics more conservative, i.e., it will tend to overestimate the Type I error.
  - H_0: $\mu_M - \mu_{g(X)} = 0$.
  - H_A: $\mu_M - \mu_{g(X)} > 0$.
  - Using a 1% probability of Type I error ($\alpha = 0.01$).
  - $\alpha = 5.179741e\text{-}043$ shows that the null hypothesis can be rejected.

11/4/2013 57

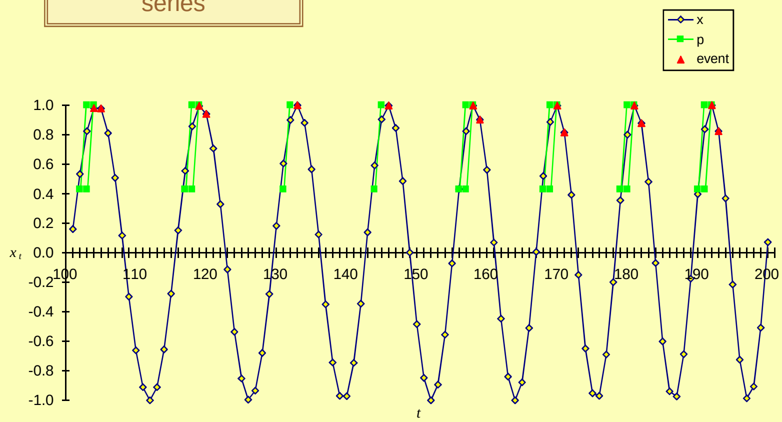
Chirp Pattern Cluster



11/4/2013 58

Time Series View of Chirp

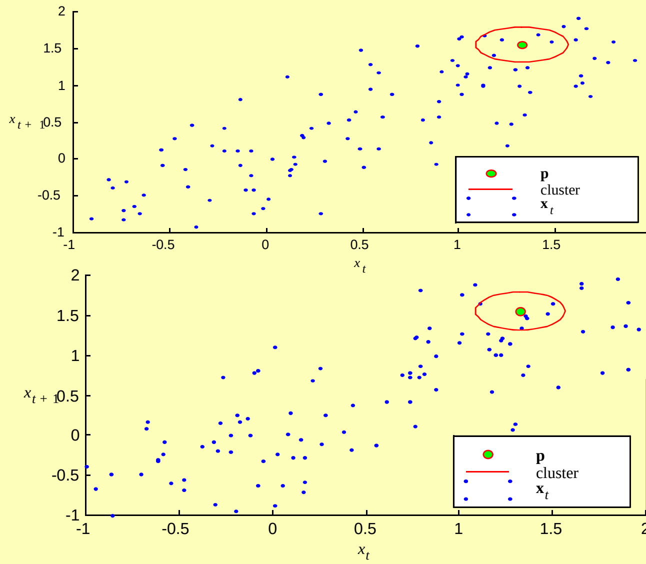
Testing portion of time series



Pattern, p, is from training process

11/4/2013 59

Noisy Sinusoidal Pattern Cluster



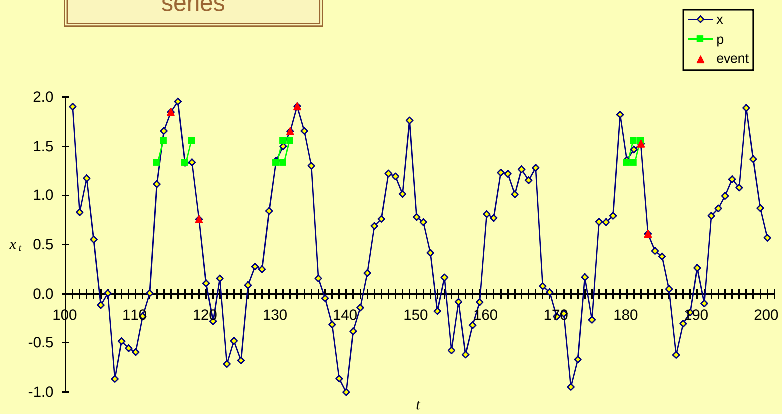
Training portion of time series

Pattern cluster applied to test time series

11/4/2013 60

Time Series View of Noisy Sinusoidal

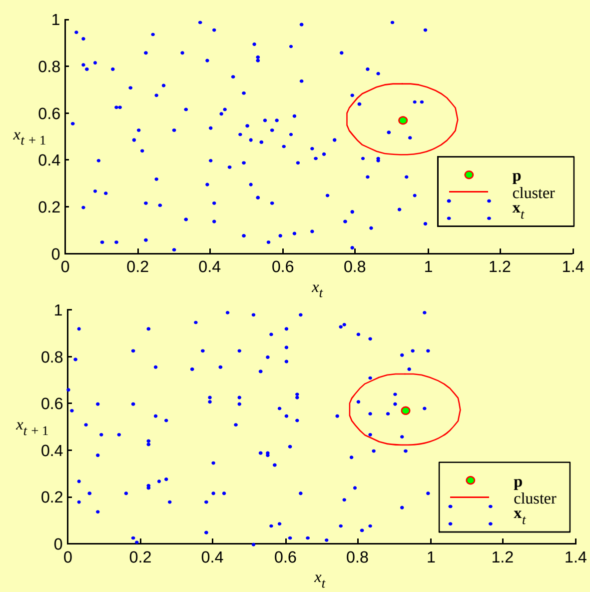
Testing portion of time series



Pattern, p, is from training process

11/4/2013 61

Noise Pattern Cluster



Training portion of time series

Pattern cluster applied to test time series

11/4/2013 62

Welding Application

- **Welding Process**
  - Two pieces of metal joined into one by making a joint between them
  - Arcing current is created between welder and metal to be joined
  - Wire is pushed out of welder
  - Tip of wire melts, forming a droplet of metal that elongates (sticks out) until it releases
  - Goal: Predict when droplet will release
    - Can't be done by traditional methods

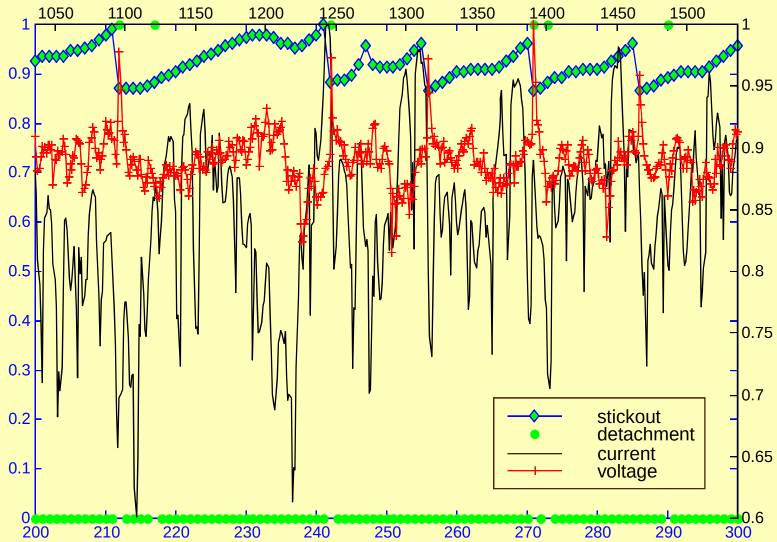
11/4/2013 63

Welding Data Set

- **Goal: Predict when a droplet will release**
- **Four time series**
  - Release (event), 1kHz sampling rate (~5000 data points)
  - Stickout, 1kHz sampling rate (~5000 data points)
  - Current, 5kHz sampling rate (~35,000 data points)
  - Voltage, 5kHz sampling rate (~35,000 data points)
  - Data not originally synchronized
- **First Pass**
  - Used release time series as event function
  - Used stickout as time series
    - "not too reliable"
  - Used a set of phase spaces (*Q*) from dimension 1 to 16
  - Generated 16 temporal patterns

11/4/2013 64

Welding Time Series



Welding Initial Results

- **Training Set**
  - Runs $\alpha = 0$
  - Means $\alpha = 3.02 \times 10^{-49}$
  - true positives: 125 (5.6%), false positives: 99 (4.4%)
  - true negatives: 2005 (89.3%), false negatives: 17 (0.8%)
  - 94.8% accuracy
- **Testing Set**
  - Runs $\alpha = 0$
  - Means $\alpha = 1.34 \times 10^{-51}$
  - true positives: 97 (3.5%), false positives: 52 (1.9%)
  - true negatives: 2470 (89.1%), false negatives: 55 (2.0%)
  - 95.1% accuracy

Statistical Tests

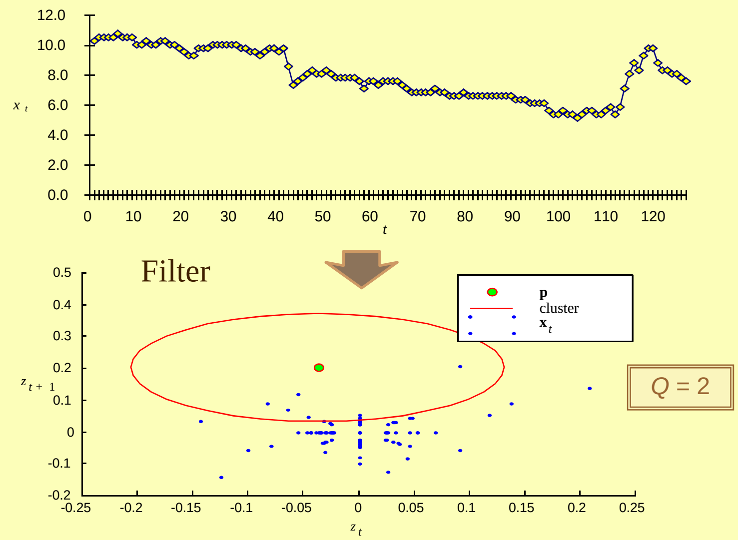
- **Runs Test**
  - H_0: There is no difference between the matched time series and the whole time series.
  - H_A: There is significant difference between the matched time series and the whole time series.
  - Using a 1% probability of Type I error ($\alpha = 0.01$).
- **Difference of two independent means**
  - Although the two populations are probably dependent, this can be ignored because it makes the statistics more conservative, i.e., it will tend to overestimate the Type I error.
  - H_0: $\mu_M - \mu_{g(X)} = 0$.
  - H_A: $\mu_M - \mu_{g(X)} > 0$.
  - Using a 1% probability of Type I error ($\alpha = 0.01$).

ICN Time Series

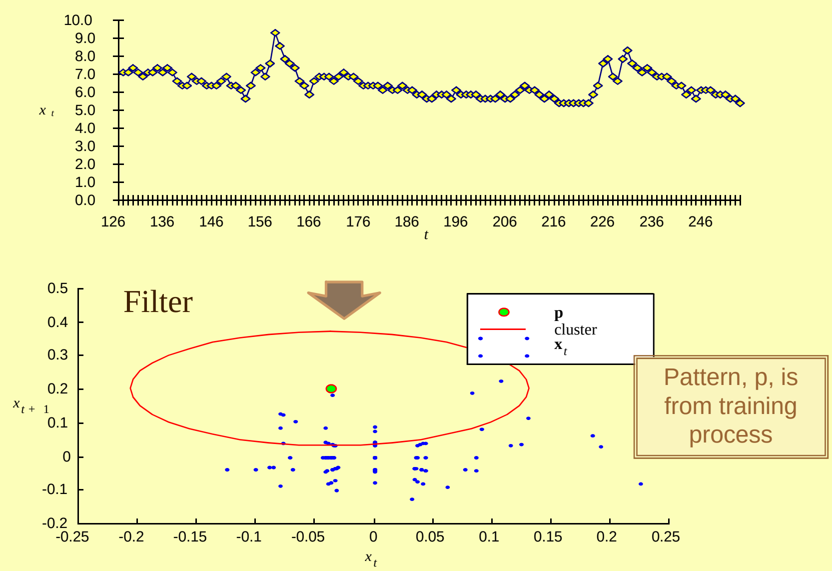
- **ICN is a pharmaceutical company whose stock trades on the NYSE**
- **Training Time Series**
  - Daily open price for 1st 126 trading days of 1990
- **Testing Time Series**
  - Daily open price for the 2nd 127 trading days of 1990
- **Stock prices tend to grow exponentially**
  - A filter is applied to the time series

$$Z = \left\{ z_t = \frac{(1-B)}{B} x_t \right\}$$

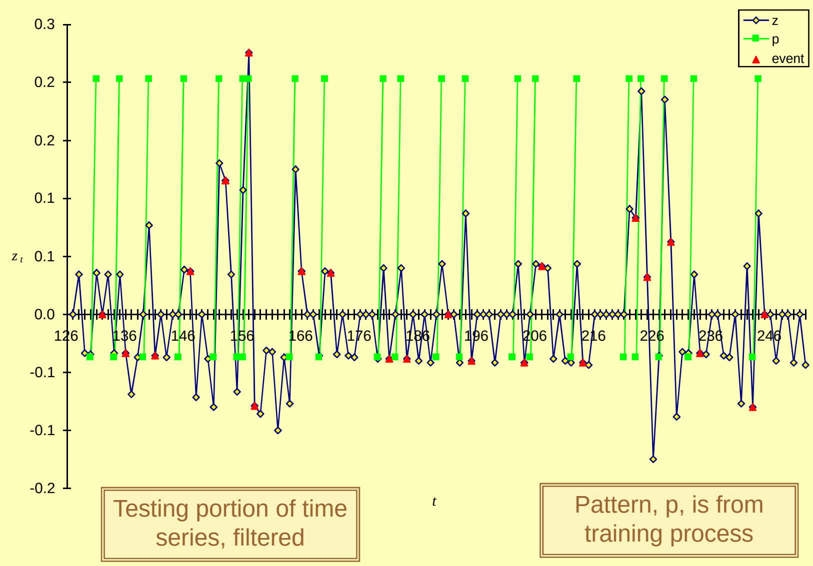
ICN Train Phase Space



ICN Test Phase Space



Time Series View of ICN



11/4/2013 71

ICN Initial Results

- **Training Set**
  - Using Buy and Hold Strategy: -26.2% (125 days)
  - Using Temporal Patterns: 51.0% (8 days)
  - Runs $\alpha = 2.21 \times 10^{-3}$
  - Means $\alpha = 2.42 \times 10^{-2}$
- **Testing Set**
  - Using Buy and Hold Strategy: -24.1% (126 days)
  - Using Temporal Patterns: 17.3% (22 days)
  - Runs $\alpha = 3.63 \times 10^{-3}$
  - Means $\alpha = 2.54 \times 10^{-1}$

11/4/2013 72

Presentation Status

- Problem Statement
  - Graphical Problem Statement
  - Time Series Analysis Literature
  - Innovative New Approach
- Algorithm
  - Phase Space
  - Mathematical Formulation
  - Algorithm Results
- Applications
  - Progression of Time Series
  - Engineering
  - Financial

11/4/2013 73

Thank You

- **References**

  - Povinelli, R. J., and Feng, X. (1998). "Temporal Pattern Identification of Time Series Data using Pattern Wavelets and Genetic Algorithms" Proceedings, 8th International Conference on Artificial Neural Networks in Engineering (ANNIE'99), Vol. 8, pp.691-696, November 1998.
  - Povinelli and Xin Feng. (1999). "Data Mining of Multiple Non-stationary Time Series" Proceedings, 9th International Conference on Artificial Neural Networks in Engineering (ANNIE'99), Vol. 9, pp. 511-516, November 1999.
  - Richard J. Povinelli and Xin Feng. "Improving Genetic Algorithms Performance By Hashing Fitness Values" Proceedings, 9th International Conference on Artificial Neural Networks in Engineering (ANNIE'99), Vol. 9, pp. 399-404, November 1999.
  - Richard J. Povinelli and Xin Feng. "Characterization And Prediction of Welding Droplet Release Using Time Series Data Mining" Proceedings, 10th International Conference on Artificial Neural Networks in Engineering (ANNIE2000), Vol. 9, pp. 857-862, November, 2000
  - Richard J. Povinelli and Xin Feng, "A New Temporal Pattern Identification Method for Characterization and Prediction of Complex Time Series Events," IEEE Transactions On Knowledge And Data Engineering, Vol. 15, No. 2, March/April 2003.

11/4/2013 74

xin.feng@mu.edu

THANK YOU!

Questions???