SCS CMU

Civil and Environmental ENGINEERING

Outline

- Motivation
- Traditional tools
  - Similarity Search and Indexing
  - DSP (Digital Signal Processing)
  - Linear Forecasting
  - ICA
- Recent streaming tools
- Intro to water quality [Jeanne]
- Conclusions

SCS CMU

Civil and Environmental ENGINEERING

Traditional tools: ICA

SCS CMU

Civil and Environmental ENGINEERING

SVD - quality

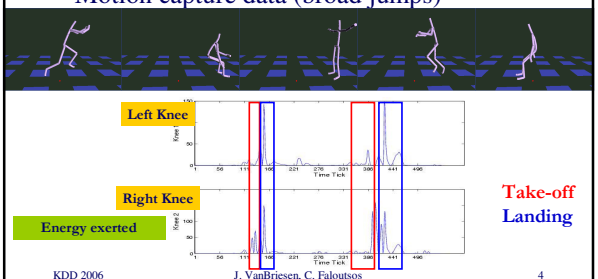
- Remember – from SVD:
- Q: can we find better 'hidden variables'?
- A: yes – with Independent Component Analysis (ICA) – see later

SCS CMU

Civil and Environmental ENGINEERING

(Q1) Find patterns in data

- Motion capture data (broad jumps)

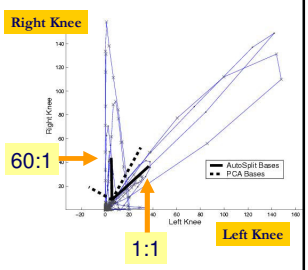


SCS CMU

Civil and Environmental ENGINEERING

(Q1) Find patterns in data

- Human would say
  - Pattern 1: along diagonal
  - Pattern 2: along vertical axis
- How to find these automatically?

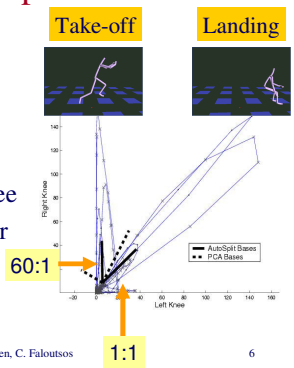


SCS CMU

Civil and Environmental ENGINEERING

Q1: Find patterns

- Patterns found
  - Landing: both knees
  - Take-off: right knee
  - Right-handed actor



SCS CMU

Civil and Environmental ENGINEERING

(Q2) Find hidden variables

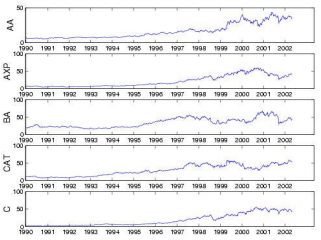
Alcoa

American Express

Boeing

Caterpillar

Citi Group



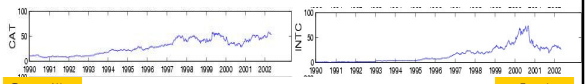
Find common hidden variables, and weights.

Dow Jones Industrial Average

SCS CMU

Civil and Environmental ENGINEERING

(Q2) Find hidden variables



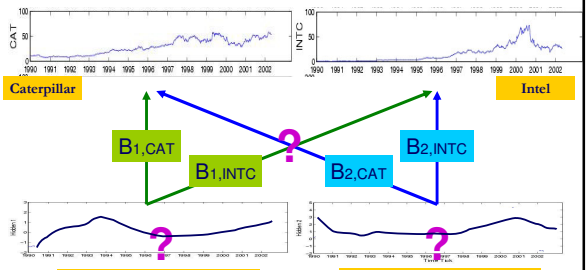
Caterpillar

Intel

SCS CMU

Civil and Environmental ENGINEERING

(Q2) Find hidden variables



Caterpillar

Intel

$B_{1,CAT}$

$B_{1,INTC}$

?

$B_{2,CAT}$

$B_{2,INTC}$

?

?

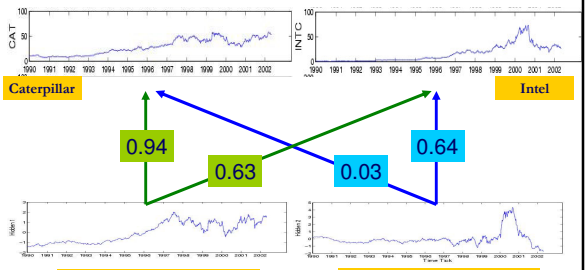
Hidden variable 1

Hidden variable 2

SCS CMU

Civil and Environmental ENGINEERING

(Q2) Find hidden variables



Caterpillar

Intel

0.94

0.63

0.03

0.64

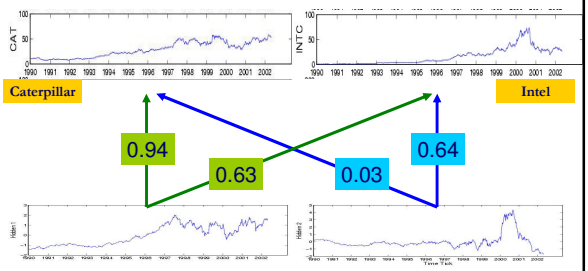
Hidden variable 1

Hidden variable 2

SCS CMU

Civil and Environmental ENGINEERING

(Q2) Find hidden variables



Caterpillar

Intel

0.94

0.63

0.03

0.64

"General trend"

"Internet bubble"

SCS CMU

Civil and Environmental ENGINEERING

Main idea behind ICA

- Find directions that are 'independent' (instead of just orthogonal)
- = minimum mutual information
- = sparse encoding
- (for algo's etc, see citations:
  - first do PCA;
  - then find 'indep. directions')

SCS CMU

Civil and Environmental ENGINEERING

Alternate naming: BSS

- Blind source separation
- Cocktail discussion problem:
  - given N discussions and N ears
  - separate the discussions (and 'lock' on one of them)

SCS CMU

Civil and Environmental ENGINEERING

Experimental Results

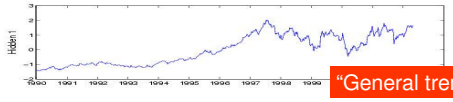
- hidden variable discovery (of stock prices)
- Scalability

SCS CMU

Civil and Environmental ENGINEERING

Companies related to h.v. 1

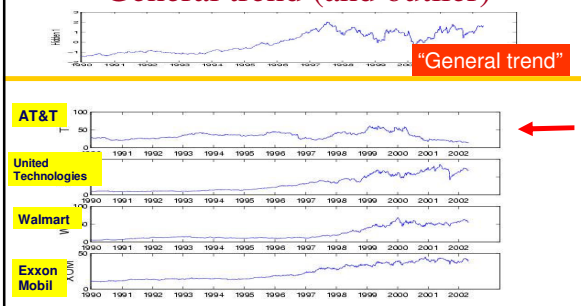
$B_{1,j}$			
Highest		Lowest	
Caterpillar	0.938512	AT&T	0.021885
Boeing	0.911120	WalMart	0.624570
MMM	0.906542	Intel	0.638010
Coca Cola	0.903858	Home Depot	0.647774
Du Pont	0.900317	Hewlett-Packard	0.658768



SCS CMU

Civil and Environmental ENGINEERING

General trend (and outlier)



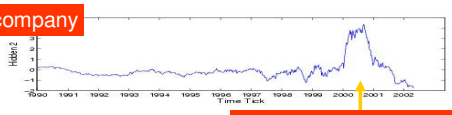
SCS CMU

Civil and Environmental ENGINEERING

Companies related to h.v. 2

$B_{2,j}$			
Highest		Lowest	
Intel	0.641102	Philip Morris	-0.194843
Hewlett-Packard	0.621159	International Paper	-0.089569
GE	0.509164	Caterpillar	0.031678
American Express	0.504871	Procter and Gamble	0.109576
Disney	0.490529	Du Pont	0.133337

Tech company



2000-2001 "Internet bubble"

SCS CMU

Civil and Environmental ENGINEERING

Experimental Results

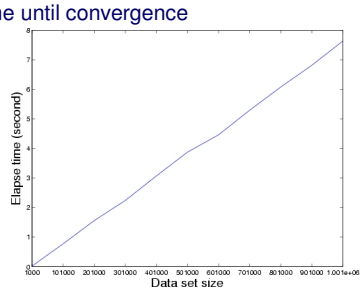
- hidden variable discovery (of stock prices)
- Scalability

SCS CMU

Civil and Environmental ENGINEERING

Scale with data set size - linear

Run time until convergence

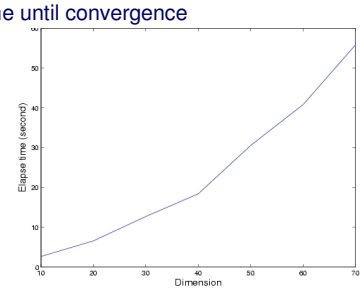


SCS CMU

Civil and Environmental ENGINEERING

Scale up with dimensionality

Run time until convergence



SCS CMU

Civil and Environmental ENGINEERING

Conclusions

- **ICA = sparse coding = Blind Source Separation**
  - Can find patterns (Q1)
  - Can find hidden variables (Q2)
  - **Better than PCA**
  - Can handle streams
  - Scalable

SCS CMU

Civil and Environmental ENGINEERING

References

- Aapo Hyvärinen, Juha Karhunen, Erkki Oja: *Independent Component Analysis,* John Wiley & Sons, 2001

SCS CMU

Civil and Environmental ENGINEERING

References

- Jia-Yu Pan, Hiroyuki Kitagawa, Christos Faloutsos and Masafumi Hamamoto, *AutoSplit: Fast and Scalable Discovery of Hidden Variables in Stream and Multimedia Databases* PAKDD 2004, Sydney, Australia

SCS CMU

Civil and Environmental ENGINEERING

Streaming tools

SCS CMU

Outline

- Motivation
- Traditional tools (DWT, AR, etc)
- Recent streaming tools
  - single-sequence mining (AWSOM)
  - lag correlations (BRAIDS)
  - hidden variable detection (SPIRIT)
- Intro to water quality [Jeanne]
- Conclusions

SCS CMU

Motivation: AR

- In AR, we need to choose a window of *w*
- How to choose it?

SCS CMU

Problem#2: Forecast

- Solution: try to express

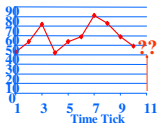
  x_t

  as a linear function of the past: x_{t-2}, x_{t-2}, …,

  (up to a window of *w*)

Formally:

$$x_t \approx a_1 x_{t-1} + \ldots + a_w x_{t-w} + noise$$



SCS CMU

How to choose '*w*'?

- goal: capture arbitrary periodicities
- with NO human intervention
- on a semi-infinite stream

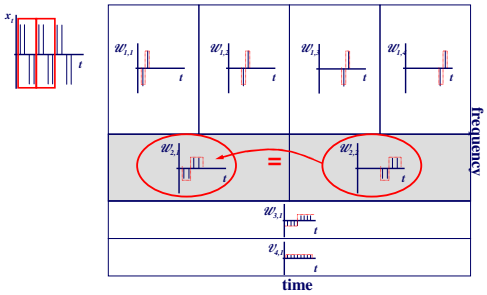
SCS CMU

Answer:

- MULTI-RESOLUTION:
- 'AWSOM' (Arbitrary Window Stream fOrecasting Method) [Papadimitriou+, vldb2003]
- idea: do AR on each wavelet level
- in detail:

SCS CMU

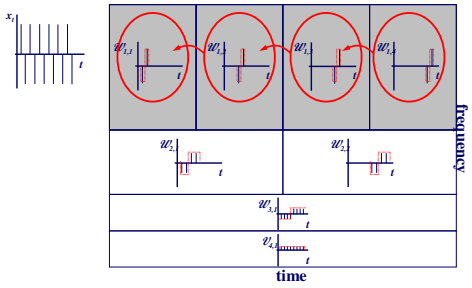
AWSOM



SCS CMU

Civil and Environmental ENGINEERING

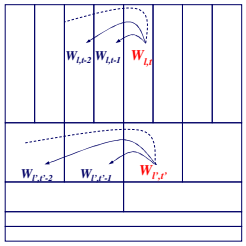
AWSOM



SCS CMU

Civil and Environmental ENGINEERING

AWSOM - idea



$W_{l,t} = \beta_{l,1} W_{l,t-1} + \beta_{l,2} W_{l,t-2} + \ldots$

$W_{l',t'} = \beta_{l',1} W_{l',t'-1} + \beta_{l',2} W_{l',t'-2} + \ldots$

SCS CMU

Civil and Environmental ENGINEERING

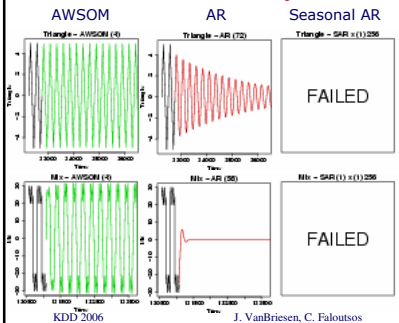
More details…

- Update of wavelet coefficients (incremental)
- Update of linear models (incremental; RLS)
- Feature selection (single-pass)
  - Not all correlations are significant
  - Throw away the insignificant ones ("noise")

SCS CMU

Civil and Environmental ENGINEERING

Results - Synthetic data

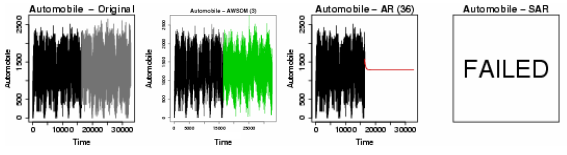


- Triangle pulse
- Mix (sine + square)
- AR captures wrong trend (or none)
- Seasonal AR estimation fails

SCS CMU

Civil and Environmental ENGINEERING

Results - Real data

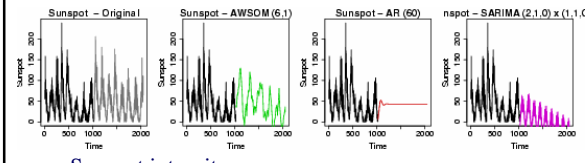


- Automobile traffic
  - Daily periodicity
  - Bursty "noise" at smaller scales
- AR fails to capture any trend
- Seasonal AR estimation fails

SCS CMU

Civil and Environmental ENGINEERING

Results - real data



- Sunspot intensity
  - Slightly time-varying "period"
- AR captures wrong trend
- Seasonal ARIMA
  - wrong downward trend, despite help by human!

SCS CMU

Skip

Complexity

- Model update
  Space: $O(lgN + mk^2) \approx O(lgN)$
  Time: $O(k^2) \approx O(1)$
- Where
  - *N:* number of points (so far)
  - *k:* number of regression coefficients; fixed
  - *m:*number of linear models; $O(lgN)$

SCS CMU

References

- Spiros Papadimitriou, Anthony Brockwell and Christos Faloutsos *Adaptive, Hands-Off Stream Mining* VLDB 2003, Berlin, Germany, Sept. 2003

SCS CMU

Outline

- Motivation
- Traditional tools (DWT, AR, etc)
- Recent streaming tools
  - single-sequence mining (AWSOM)
  - lag correlations (BRAIDS)
  - hidden variable detection (SPIRIT)
- Intro to water quality [Jeanne]
- Conclusions

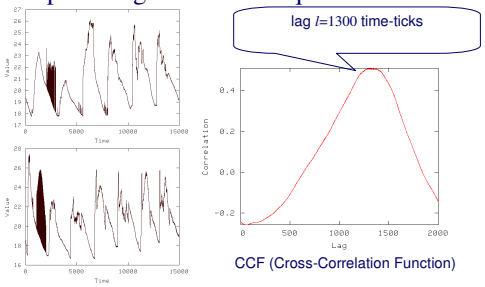
SCS CMU

Streaming tools: lag correlations

SCS CMU

Lag Correlations

- Example of lag-correlated sequences



CCF (Cross-Correlation Function)

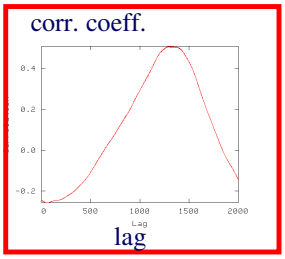
SCS CMU

Lag Correlations

- Example of lag-correlated sequences

CCF (Cross-Correlation Function)

how to compute it
•quickly
•cheaply
•incrementally



SCS CMU

Civil and Environmental ENGINEERING

Challenging Problems

- Problem definitions
  - P1: For given two co-evolving sequences *X* and *Y*, determine
    - Whether there is a lag correlation
    - If yes, what is the lag length *l*
  - P2: For given *k* numerical sequences, $X_1,\ldots,X_k$, report
    - Which pairs have a lag correlation
    - The corresponding lag for each pair

SCS CMU

Civil and Environmental ENGINEERING

Our solution

- Ideal characteristics:
  - 'Any-time' processing, and fast
  - Nimble (space < O(N))
  - Accurate

SCS CMU

Civil and Environmental ENGINEERING

Challenge

- How to compute the corr. coeff. of *X1*, *X2*
- at lag, say, *l*=1million
- without keeping 1million timeticks info?

SCS CMU

Civil and Environmental ENGINEERING

Main ideas

1. corr. coeff. *cc(X1, X2, lag=0)* can be computed incrementally
2. geometric probing: consider only lags 1, 2, 4, ..., 2**i
3. do smoothing: averages for disjoint windows of length 1, 2, ..., 2**i

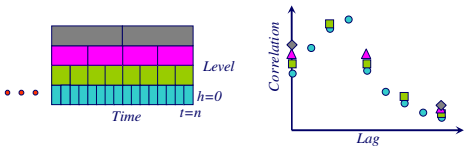
(MULTI-RESOLUTION approach)

SCS CMU

Civil and Environmental ENGINEERING

Main Ideas

- Geometric lag probing

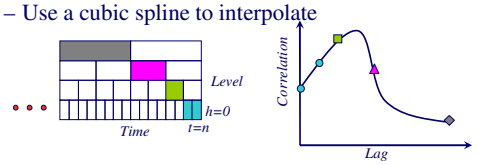


SCS CMU

Civil and Environmental ENGINEERING

Main Ideas

- Geometric lag probing
  - Use colored windows
  - Keep track of only a geometric progression of the lag values: $l=\{0,1,2,4,8,\ldots,2^h,\ldots\}$
  - Use a cubic spline to interpolate



SCS CMU

Civil and Environmental ENGINEERING

Experimental results

- Setup
  - Datasets:

    Sines, *SpikeTrains*, *Humidity*, *Light*, *Temperature*,

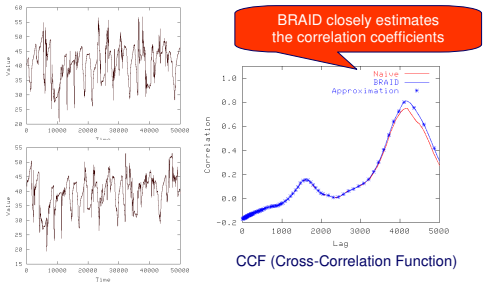
    Kursk, *Sunspots*
  - Enhanced BRAID, b=16
- Evaluation: error; scalability

SCS CMU

Civil and Environmental ENGINEERING

Detecting Lag Correlations (3)

- *Humidity*



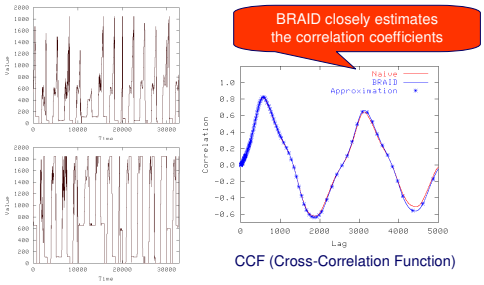
CCF (Cross-Correlation Function)

SCS CMU

Civil and Environmental ENGINEERING

Detecting Lag Correlations (4)

- *Light*



CCF (Cross-Correlation Function)

SCS CMU

Civil and Environmental ENGINEERING

Error

Datasets	Lag correlation		Estimation error (%)
	Naive	BRAID	
Sines	716	716	0.000
SpikeTrains	2841	2830	0.387
Humidity	3842	3855	0.338
Light	567	570	0.529
Kursk	1463	1472	0.615
Sunspots	1156	1168	1.038

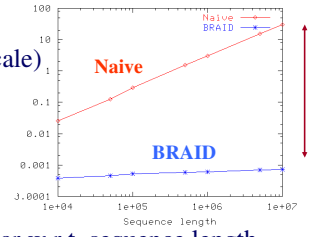
- Largest relative error is about 1%

SCS CMU

Civil and Environmental ENGINEERING

Performance

wall-clock time (log scale)



several orders of magnitude

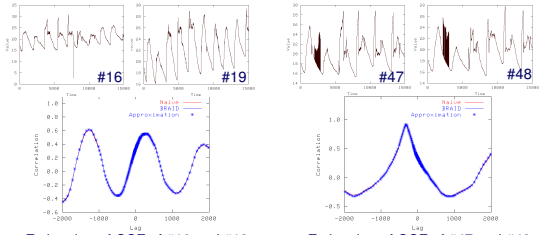
- ~linear w.r.t. sequence length
- Up to 40,000 times faster

SCS CMU

Civil and Environmental ENGINEERING

Group Lag Correlations

- Two correlated pairs from 55 *Temperature* sequences
- Each sensor is located in a different place



Estimation of CCF of #16 and #19 Estimation of CCF of #47 and #48

SCS CMU

Civil and Environmental ENGINEERING

Conclusions

Proposed method (BRAID) detects lag correlations:

- It is 'any-time'
- Nimble: ***O*(*log n*)** space, ***O*(1)** time
- Fast: **40,000x** faster than naive
- Accurate: <=1% relative error

SCS CMU

Civil and Environmental ENGINEERING

References

- Yasushi Sakurai, Spiros Papadimitriou and Christos Faloutsos *BRAID: Stream Mining through Group Lag Correlations* SIGMOD 2005, Baltimore, MD

SCS CMU

Civil and Environmental ENGINEERING

Streaming tools: hidden variables

SCS CMU

Civil and Environmental ENGINEERING

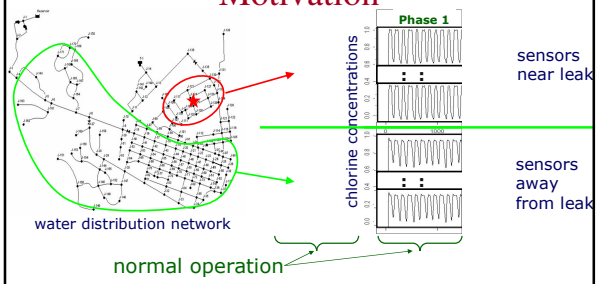
Outline

- Motivation
- Traditional tools (DWT, AR, etc)
- Recent streaming tools
  - single-sequence mining (AWSOM)
  - lag correlations (BRAIDS)
  - hidden variable detection (SPIRIT)
- Intro to water quality [Jeanne]
- Conclusions

SCS CMU

Civil and Environmental ENGINEERING

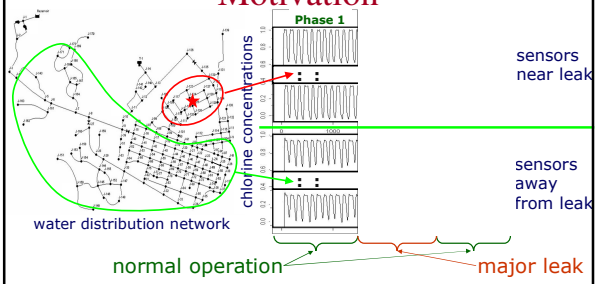
Motivation



SCS CMU

Civil and Environmental ENGINEERING

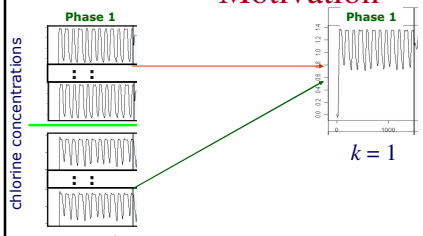
Motivation



SCS CMU

Civil and Environmental ENGINEERING

Motivation



actual measurements (*n* streams)

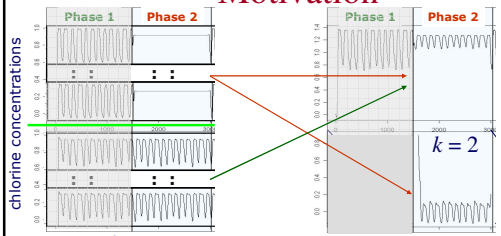
k hidden variable(s)

spot: "hidden (latent) variables"

SCS CMU

Civil and Environmental ENGINEERING

Motivation



actual measurements (*n* streams)

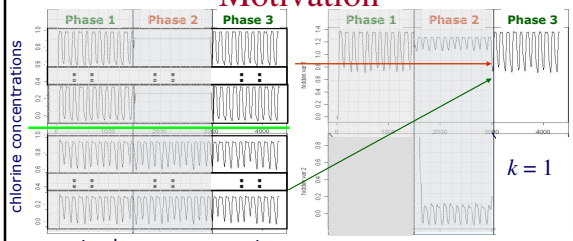
k hidden variable(s)

spot: "hidden (latent) variables"

SCS CMU

Civil and Environmental ENGINEERING

Motivation



actual measurements (*n* streams)

k hidden variable(s)

spot: "hidden (latent) variables"

SCS CMU

Civil and Environmental ENGINEERING

Goals

- Discover "hidden" (latent) variables for:
  - Summarization of main trends for users
  - Efficient forecasting, spotting outliers/anomalies
- Incremental, real-time computation
- Limited memory requirements
- No special parameters to tune

SCS CMU

Civil and Environmental ENGINEERING

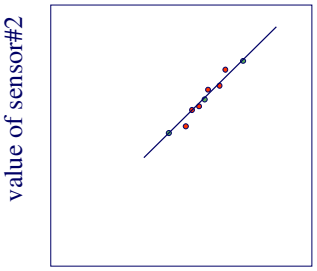
Stream correlations

- **Step 1:** How to capture correlations?
- Step 2: How to do it incrementally, when we have a very large number of points?
- Step 3: How to dynamically adjust the number of hidden variables?

SCS CMU

Civil and Environmental ENGINEERING

1. How to capture correlations



A: SVD

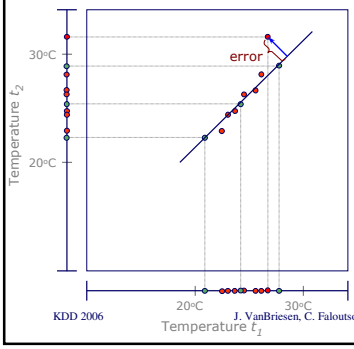
SCS CMU

Stream correlations

- Step 1: How to capture correlations?
- **Step 2:** How to do it incrementally, when we have a very large number of points?
- Step 3: How to dynamically adjust the number of hidden variables?

SCS CMU

2. Incremental update

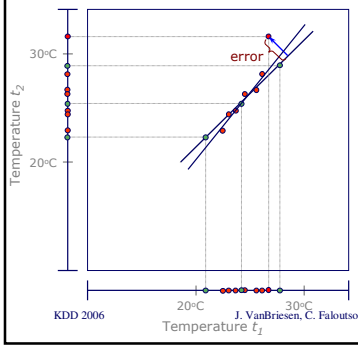


For each new point
- Project onto current line
- Estimate error

New value

SCS CMU

2. Incremental update

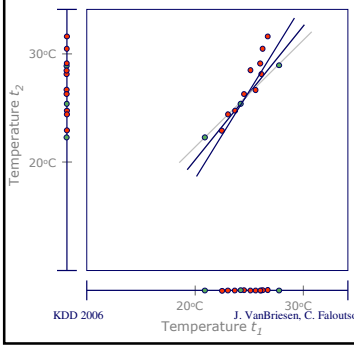


For each new point
- Project onto current line
- Estimate error
- Rotate line
- O(n) time

New value

SCS CMU

2. Incremental update



For each new point
- Project onto current line
- Estimate error
- Rotate line
- repeat

SCS CMU

Skip

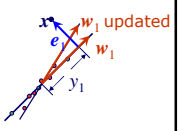
2. Incremental update

Given number of hidden variables k

- Assuming k is **known**
- We know how to update the slope

For each new point $\boldsymbol{x}$ and for $i = 1, \ldots, k$:

• y_i	:= $\boldsymbol{w}_i^T \boldsymbol{x}$	(proj. onto $\boldsymbol{w}_i$)
• d_i	← $\lambda d_i + y_i^2$	(energy ∝ i-th eigenval.)
• $\boldsymbol{e}_i$	:= $\boldsymbol{x} - y_i \boldsymbol{w}_i$	(error)
• $\boldsymbol{w}_i$	← $\boldsymbol{w}_i + (1/d_i)\, y_i \boldsymbol{e}_i$	(update estimate)
• $\boldsymbol{x}$	← $\boldsymbol{x} - y_i \boldsymbol{w}_i$	(repeat with remainder)



SCS CMU

Stream correlations

- Step 1: How to capture correlations?
- Step 2: How to do it incrementally, when we have a very large number of points?
- **Step 3:** How to dynamically adjust k, the number of hidden variables?

SCS CMU

Estimating *k*

- Decide on two thresholds
  - t_{low} (say, 0.95) and t_{high} (say, 0.98)
- introduce/drop hidden variables, to keep the reconstruction error within the bounds

SCS CMU

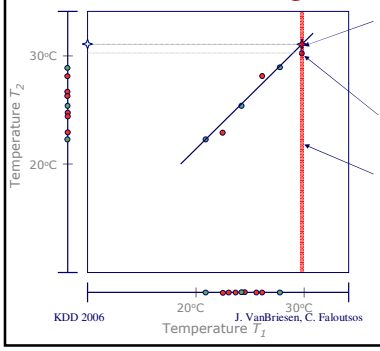
mini-overview

- Main ideas
- Discussion
- Experiments

SCS CMU

Skip

Missing values



best guess (given correlations: intersection)

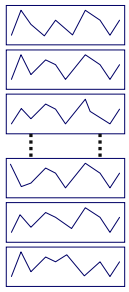
true values (pair)

all possible value pairs (given **only** t_1)

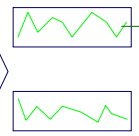
SCS CMU

Skip

Forecasting



hidden variables



k hidden vars

n streams

k* << *n and already capture correlations

Only *k* simple models

Efficiency & robustness

SCS CMU

Time/space requirements

O(*nk*) space (total) and time (per tuple), i.e.,

- Independent of # points
- Linear w.r.t. # streams (*n*)
- Linear w.r.t. # hidden variables (*k*)

In fact,

- **real time:** demo at www.warsteiner.db.cs.cmu.edu/demo/intemon.jsp

SCS CMU

mini-overview

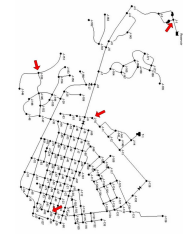
- Main ideas
- Discussion
- Experiments

SCS CMU

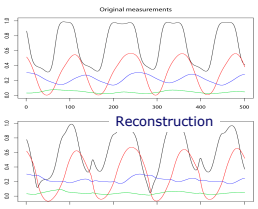
Civil and Environmental ENGINEERING

Experiments

Chlorine concentration



[CMU Civil Engineering]



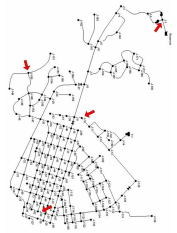
166 streams
2 hidden variables (~4% error)

SCS CMU

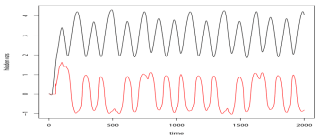
Civil and Environmental ENGINEERING

Experiments

Chlorine concentration



[CMU Civil Engineering]



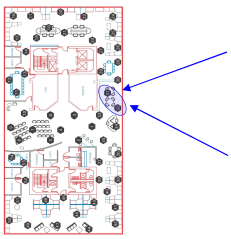
hidden variables

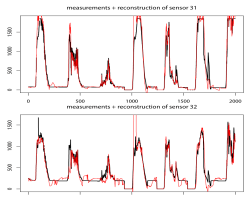
- Periodic
- with phase shift

SCS CMU

Civil and Environmental ENGINEERING

Experiments - light





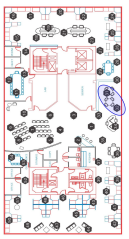
measurement
reconstruction

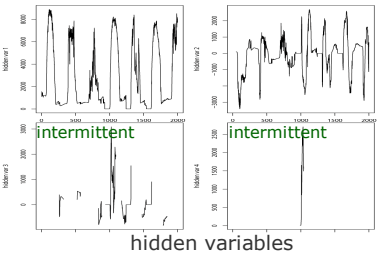
54 sensors
2-4 hidden variables (~6% error)

SCS CMU

Civil and Environmental ENGINEERING

Experiments - light





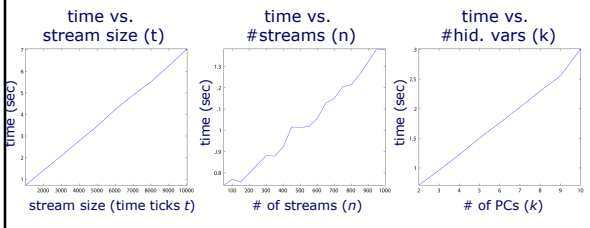
hidden variables

- 1 & 2: main trend (as before)
- 3 & 4: potential anomalies and outliers

SCS CMU

Civil and Environmental ENGINEERING

Wall-clock times



constant time per tuple and per stream

SCS CMU

Civil and Environmental ENGINEERING

Contributions

- ✓ Discover hidden variables for
  - Summarization of main trends for users
  - Efficient forecasting, spotting outliers/anomalies
- ✓ Incremental, real time computation
- ✓ With limited memory
- ✓ No special parameters to tune

SCS CMU

Civil and Environmental ENGINEERING

Streaming tools - References

- Spiros Papadimitriou, Jimeng Sun and Christos Faloutsos *Streaming Pattern Discovery in Multiple Time-Series* VLDB 2005, Trondheim, Norway.
- Sun, J., Papadimitriou, S., Faloutsos, C. *Distributed Pattern Discovery in Multiple Streams*, In the Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Singapore, Apr 9-12, 2006

SCS CMU

Civil and Environmental ENGINEERING

Outline

- Motivation
- Traditional tools (DWT, AR, etc)
- Recent streaming tools
  - single-sequence mining (AWSOM)
  - lag correlations (BRAIDS)
  - hidden variable detection (SPIRIT)
- Intro to water quality [Jeanne]
- Conclusions

SCS CMU

Civil and Environmental ENGINEERING

BREAK -> Jeanne

SCS CMU

Civil and Environmental ENGINEERING

Overall conclusions - Tools

- Similarity search: **Euclidean**/time-warping; **feature extraction** and **SAMs**
- Signal processing: **DWT** is a powerful tool
- Linear Forecasting: **AR** (Box-Jenkins) methodology

SCS CMU

Civil and Environmental ENGINEERING

Overall conclusions – Streaming tools

Challenge: '*you get to see it only once*'

Powerful ideas:

- *Multi-resolution*: wavelets, AWSOM, BRAID
- *incremental* updates (Recursive Least Squares, SPIRIT)

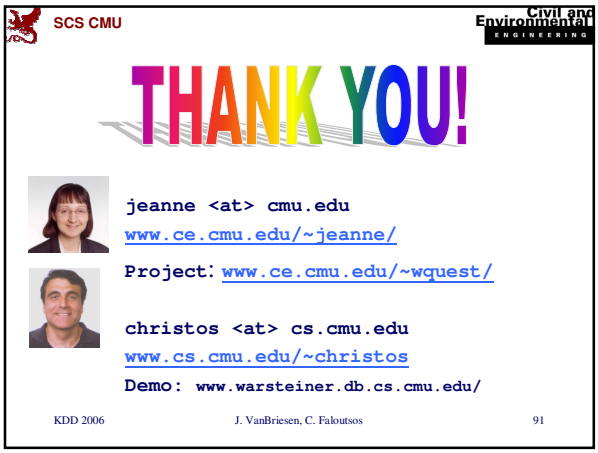
SCS CMU

Civil and Environmental ENGINEERING

Research directions

For (water) sensors:

1. **Sensor placement**
2. **Surrogate** issue: we measure chlorine; how much pathogen? (``inverse problem'')
3. **Gray-box** data mining: we know (some) of the physics/chemistry differential equations – integrate them with the measurements
   - ditto for spatio-temporal correlations


SCS CMU
Civil and Environmental
ENGINEERING
THANK YOU!
jeanne <at> cmu.edu
www.ce.cmu.edu/~jeanne/
Project: www.ce.cmu.edu/~wquest/
christos <at> cs.cmu.edu
www.cs.cmu.edu/~christos
Demo: www.warsteiner.db.cs.cmu.edu/
KDD 2006
J. VanBriesen, C. Faloutsos
91