The 2014 ACM SIGMOD/PODS Conference
June 22-27, 2014, Snowbird, Utah, USA
ACM SIGMOD/PODS 2014

AutoPlait: Automatic Mining of Co-evolving Time Sequences

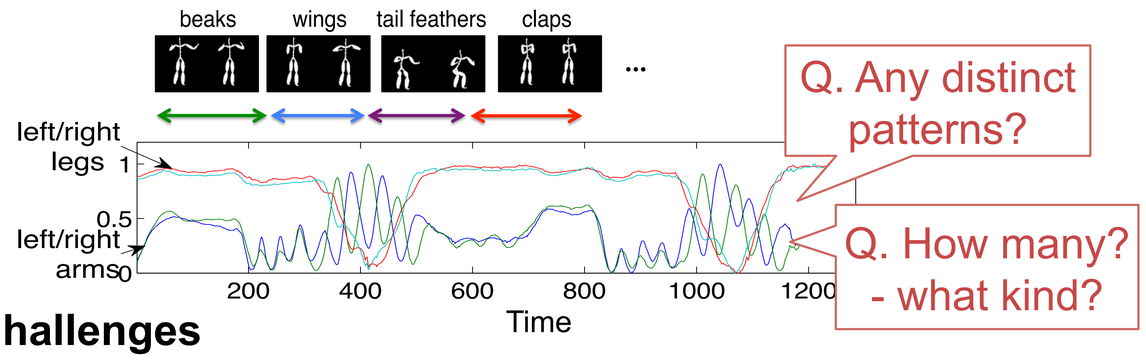
Yasuko Matsubara
Kumamoto University
yasuko@cs.kumamoto-u.ac.jp

Yasushi Sakurai
Kumamoto University
yasushi@cs.kumamoto-u.ac.jp

Christos Faloutsos
Carnegie Mellon University
christos@cs.cmu.edu

Motivation - Given: Co-evolving time-series

e.g., Mocap (leg/arm sensors) - "chicken dance"



Challenges

(1) Unknown # of patterns (2) Different durations

Q. Can we summarize it ***automatically***??

Goal: find patterns that agree with human intuition

AutoPlait: "fully-automatic" mining algorithm

Importance of "fully-automatic"

No magic numbers! … because,

Manual	- sensitive to the parameter tuning - it takes a very long time (hours, days, …)
Automatic	- no expert tuning required

Big data mining: we cannot afford human intervention!!

Problem formulation - Key concepts

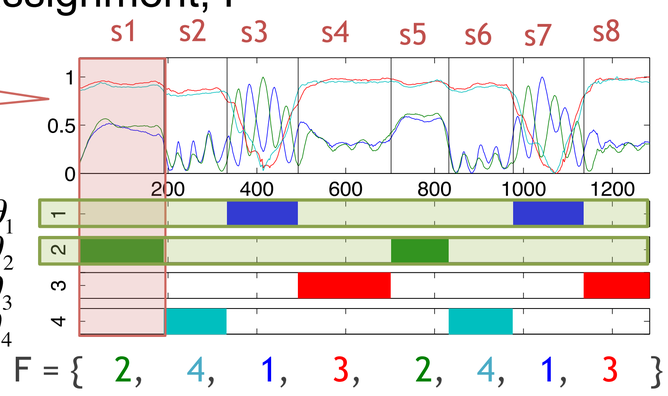
- Bundle (given) : d co-evolving sequences, $X = \{x_1,\ldots,x_n\}^{d\times n}$
- Segment : convert X -> m segments, S
- Regime : segment groups, Θ
- Segment-membership: assignment, F

θ_r : model params of regime r

Segment (m=8) $S = \{s_1,\ldots,s_m\}$

Regimes (r=4) $\Theta = \{\theta_1,\theta_2,\ldots\theta_r,\Delta_{r\times r}\}$

Segment-membership $F = \{f_1,\ldots,f_m\}$



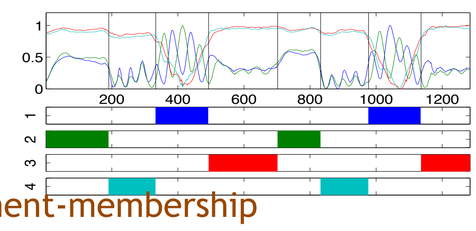
Problem definition

Given: bundle $X = \{x_1,\ldots,x_n\}$

Find: compact description C of X

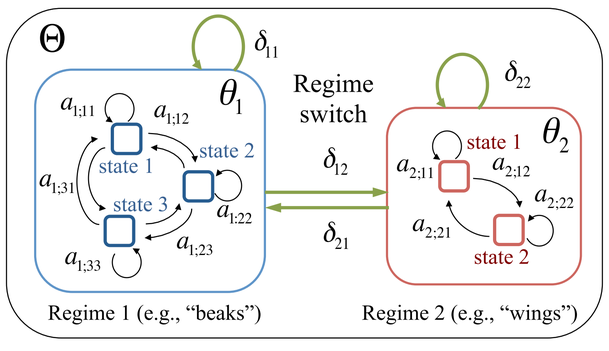
$C = \{m, r, S, \Theta, F\}$

m segments, r regimes, Segment-membership



Proposed method: AutoPlait

Main idea (1): MLCM: multi-level chain model



$\Theta = \{\theta_1,\theta_2,\ldots\theta_r,\Delta_{r\times r}\}$

r HMMs (regimes) — across-regime transition prob.

$\theta_i = \{\pi, A, B\}$ } Single HMM parameters

Main idea (2): Model description cost

of segments / regimes

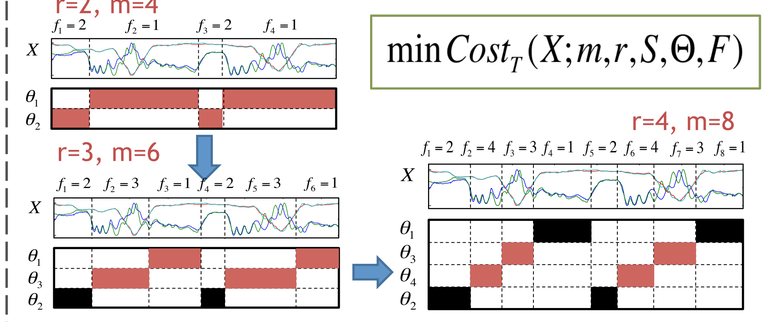
$$Cost_T(\boldsymbol{X};\mathcal{C}) = Cost_T(\boldsymbol{X};m,r,\mathcal{S},\boldsymbol{\Theta},\mathcal{F})$$
$$= \log^*(n) + \log^*(d) + \log^*(m) + \log^*(r) + m\log(r)$$
$$+ \sum_{i=1}^{m-1} \log^*|s_i| + Cost_M(\boldsymbol{\Theta}) + Cost_C(\boldsymbol{X}|\boldsymbol{\Theta}) \quad (6)$$

duration / dimensions — segment member-ship F — segment lengths — Model description cost of Θ — Coding cost of X given Θ

AutoPlait (outer-loop algorithm)

Split regimes r=2,3,…, as long as **cost** keeps decreasing

- Find appropriate # of regimes



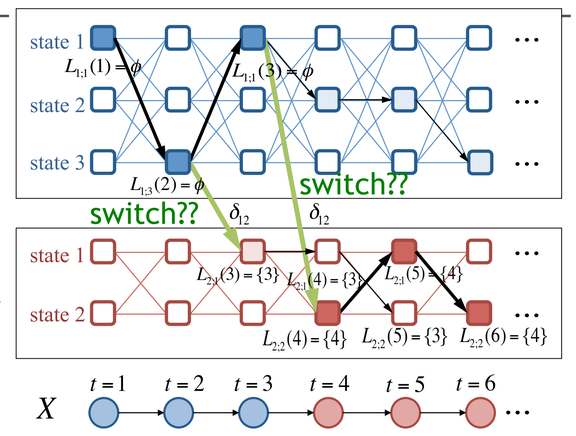
CutPointSearch (inner-most loop)

Given: X, regimes $\Theta = \{\theta_1,\theta_2,\Delta\}$

Find: cut-points of segs: S_1, S_2

$\{S_1,S_2\} = \underset{S_1,S_2}{\operatorname{argmax}} P(X \mid S_1,S_2,\Theta)$

DP algorithm to compute likelihood: $P(X \mid \Theta)$

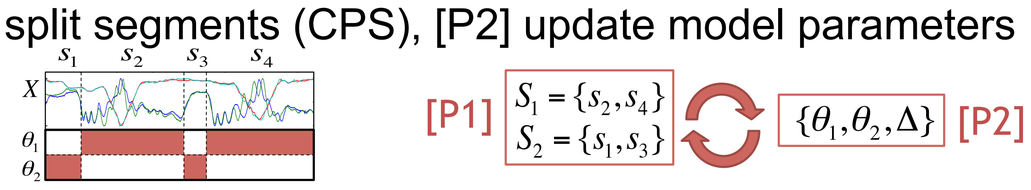


RegimeSplit (inner-loop algorithm) Given: X,

Find (1) two segment sets: S_1, S_2 (2) two regimes: $\Theta = \{\theta_1,\theta_2,\Delta\}$

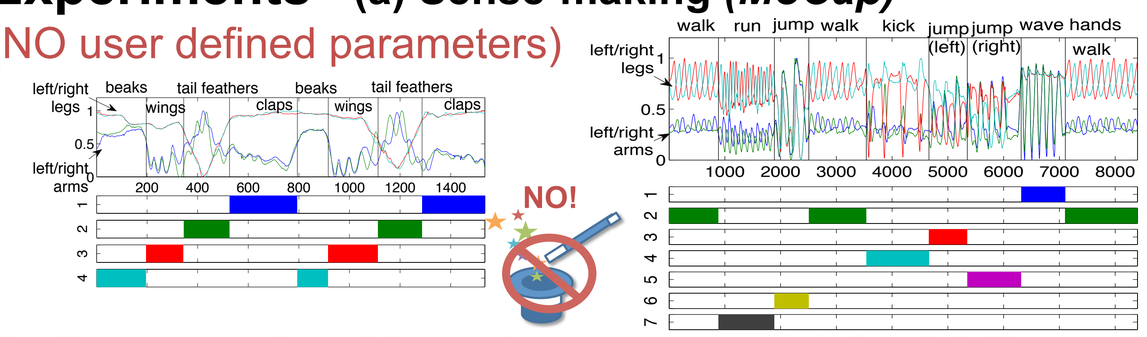
Two-phase iterative approach

[P1] split segments (CPS), [P2] update model parameters

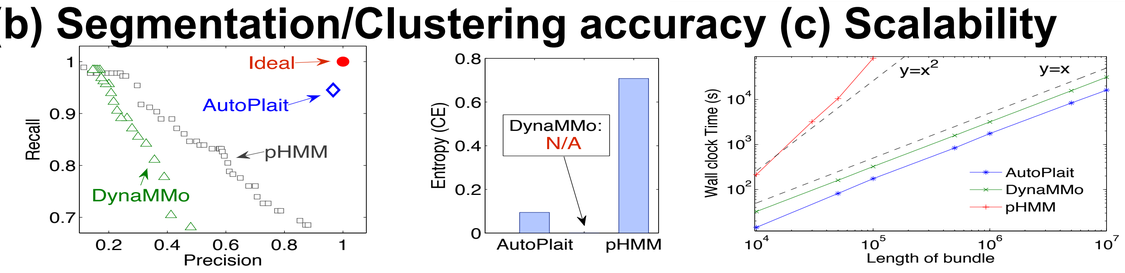


Experiments - (a) Sense-making *(MoCap)*

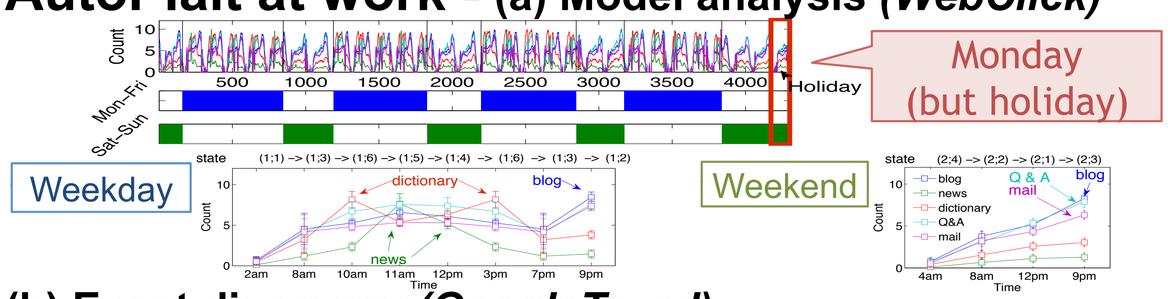
(NO user defined parameters)



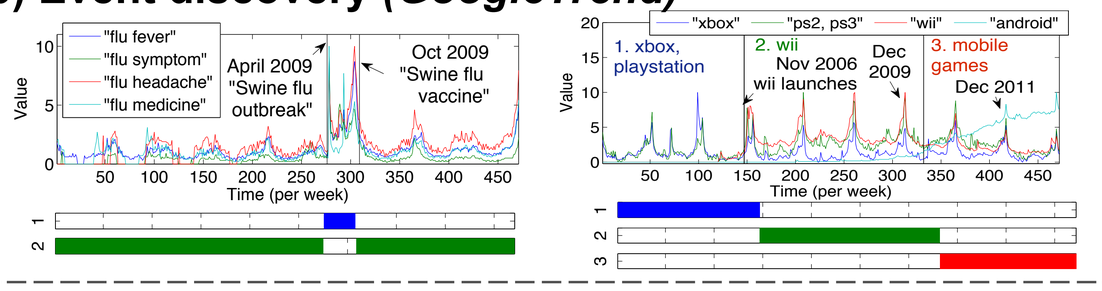
(b) Segmentation/Clustering accuracy (c) Scalability



AutoPlait at work - (a) Model analysis *(WebClick)*



(b) Event discovery *(GoogleTrend)*



Conclusions - AutoPlait has following advantages:

- Effective & Sense-making: it provides reasonable regimes
- Fully-automatic: it needs no magic numbers
- Scalable: it scales linearly with the duration n

Code: http://www.cs.kumamoto-u.ac.jp/~yasuko/software.html