Time Series Discretization via MDL-based Histogram Density Estimation

Yoshitaka Kameya
Tokyo Institute of Technology

Outline

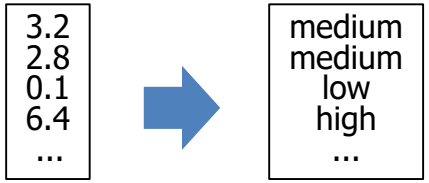
- Background: Unsupervised discretization of time series data
- Our proposal: Histogram-based discretization
- Experiments
- Conclusion/Future work

Outline

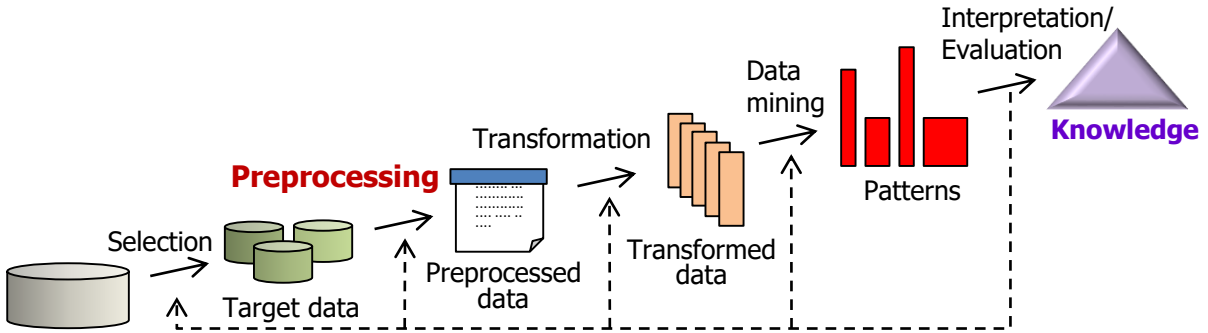
- Background: Unsupervised discretization of time series data
- Our proposal: Histogram-based discretization
- Experiments
- Conclusion/Future work

Discretization

- ... converts numeric data into symbolic data



- ... is a *preprocessing* task in knowledge discovery



[Fayyad et al. 1995]

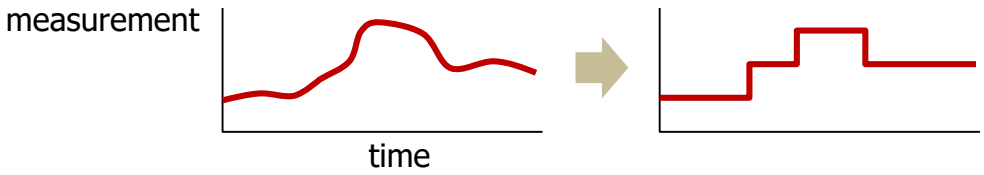
- ... may lead to noise reduction and data abstraction
  - We wish to have *interpretable* discrete levels
- ... may help *symbolic* data mining
  - Frequent pattern mining
  - Inductive logic programming

Unsupervised discretization of time series data

Common strategy:

- *Smoothing* at the time (x) axis
- *Binning* or *clustering* at the measurement (y) axis

combined sequentially or simultaneously



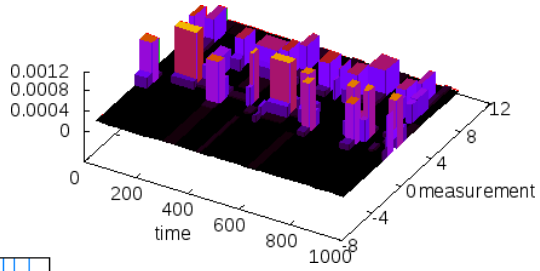
- Binning:
  - Equal width binning
  - Equal frequency binning
  - …
- Clustering:
  - Hierarchical clustering [Dimitrova et al. 05]
  - K-means
  - Gaussian mixture models [Mörchen et al. 05b]
  - …
- Smoothing:
  - Regression trees [Geurts 01]
  - Smoothing filters
    - Moving averaging
    - Savitzky-Golay filters [Mörchen et al. 05b]
  - …
- All-in-one methods:
  - SAX [Lin et al. 07]
  - Persist [Mörchen et al. 05a]
  - Continuous hidden Markov models [Mörchen et al. 05a]

Outline

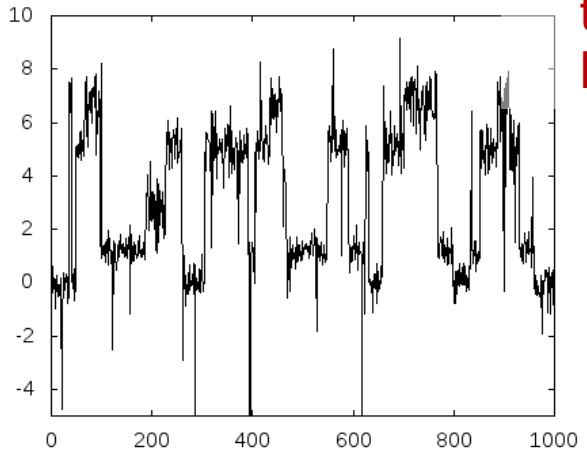
- ✓ Background: Unsupervised discretization of time series data
- Our proposal: Histogram-based discretization
- Experiments
- Conclusion/Future work

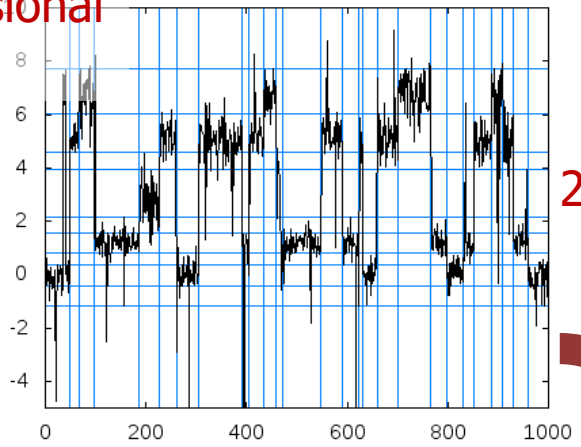
Histogram-based discretizer

- Three-step procedure:

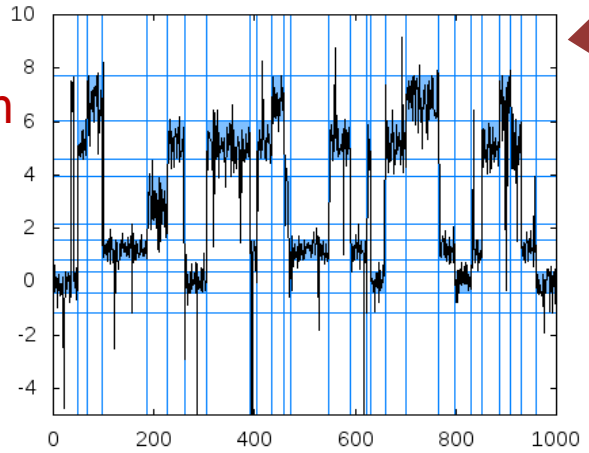


1. Build a two-dimensional histogram

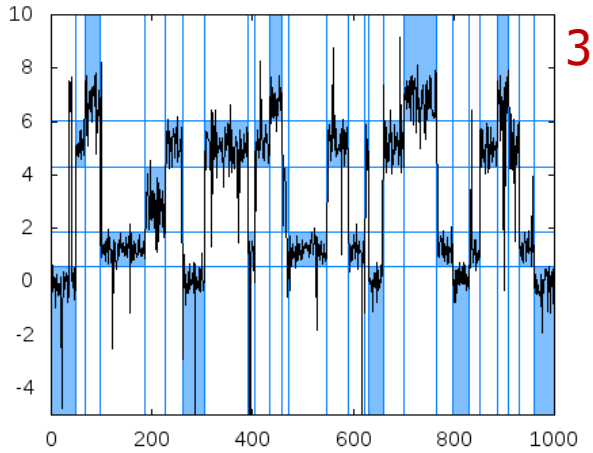




2. Choose the bin with the highest volume at each time interval

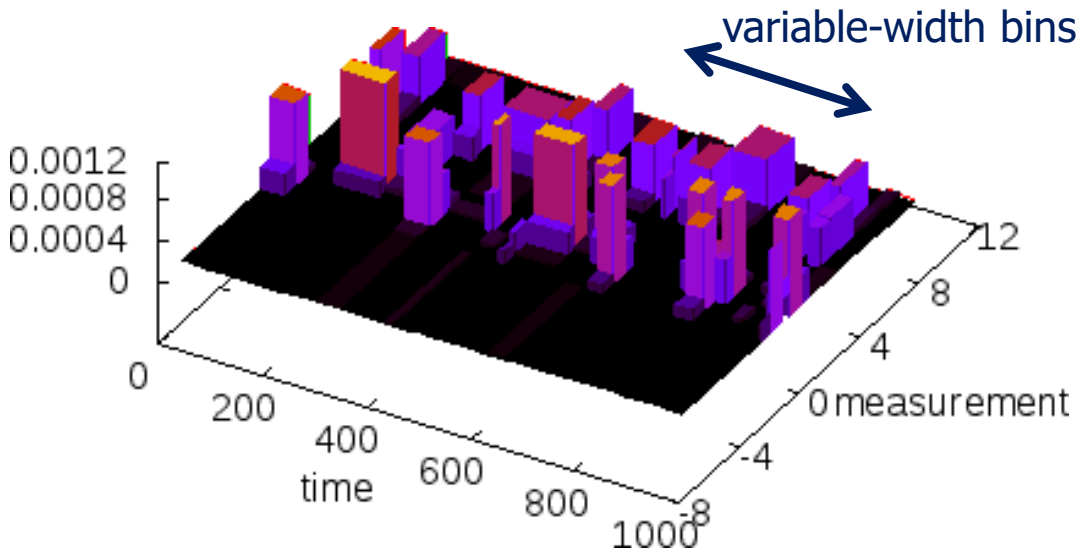


3. Suppress unchosen levels



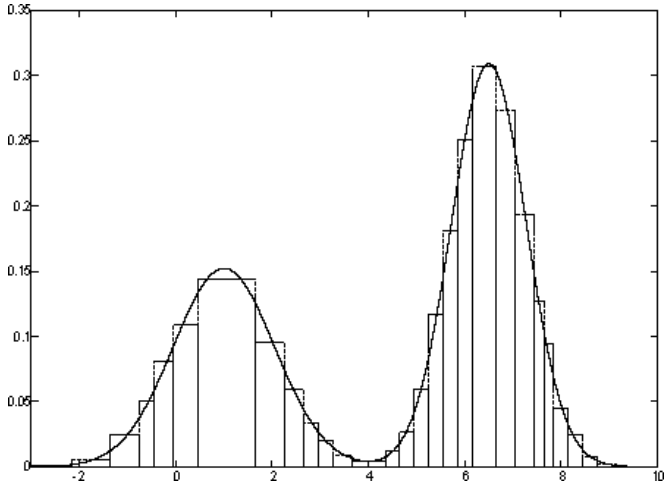
Histogram-based discretizer (cont'd)

- Advantages over previous methods:
  - More neutral ... less assumptions are required
    (discretization is just a preprocessing)
  - More intuitive ... bins are understood as propositions like "$a < X \leq b$"
  - More robust ... smoothing along with x-axis would get stronger
    (continuous time is taken into account)



Basis: the K&M method

- Efficient optimization of a 1-D histogram
  [Kontkanen and Myllymäki 07b]
  - Variable-width bins are allowed
  - # of bins and bin-widths are optimized based on an MDL score: normalized maximum likelihood (NML)
    - No hyperparameters need to be specified



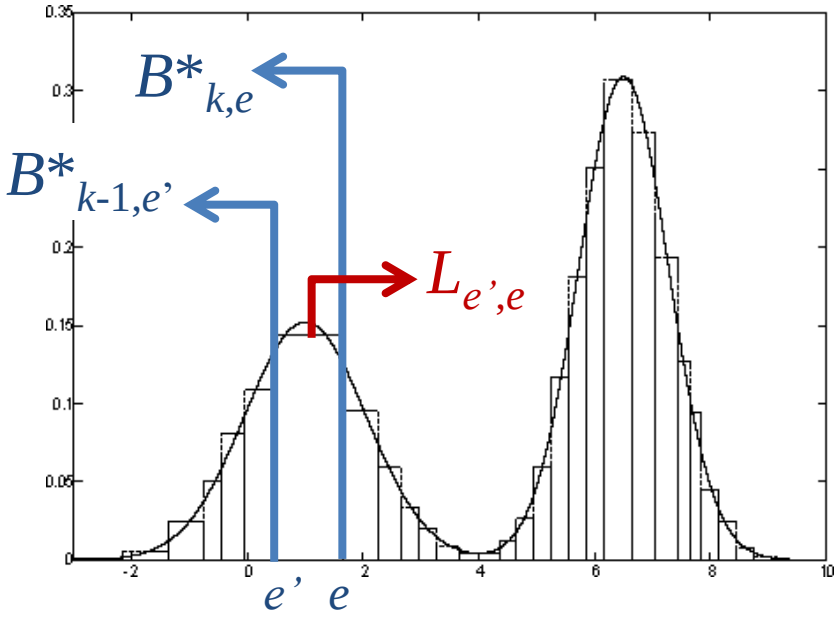
Basis: the K&M method (cont'd)

- Efficient optimization of a 1-D histogram
  [Kontkanen and Myllymäki 07]
  - Dynamic programming
    → $O(KE^2)$-time (K: max. # of bins, E: # of candidates for cut points)
    → ≈1.5 sec with $K = 5$, $E = 100$ (Intel Core i7 2.66GHz)

Partial histogram's score

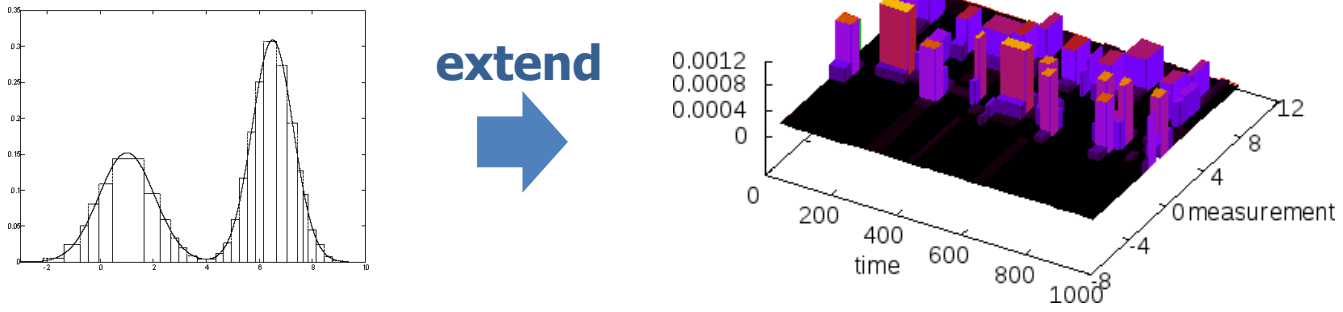
$$B^*_{k,e} = \min_{k-1 \le e' \le e-1}\{B^*_{k-1,e'} + L_{e',e}\}$$

Local contribution from e' to e



Proposed method

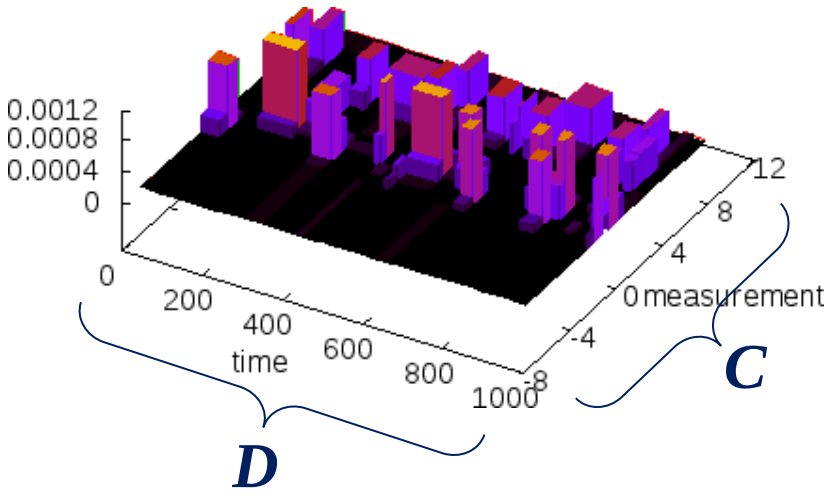
- Main task: Density estimation by a 2D histogram
- We extend the K&M method into the 2D case



- Major modifications:
  - Computation of NML in the 2D case (following [Kontokanen et al. 05])
  - Iterative optimization of the bins
    between time-axis and mesurement-axis

Proposed method: Dynamic programming

- Simultaneous finding of the cut points at both axes seems intractable



- Iterative optimization:

Start with equal-interval cut points D at the **time** axis

Run a modified K&M method for optimal cut points C at the **measurement** axis (with D fixed)

dynamic programming

Run a modified K&M method for optimal cut points D at the **time** axis (with C fixed)

dynamic programming

Local optimality is guaranteed

Proposed method: Time complexity

- Computation of NML: $O(n^2 \log \min\{K_{\max}, K'_{\max}\})$
- Iteration for finding the cut points: $O(\{E^2 + E'^2\} K_{\max} K'_{\max})$

n	# of data points
E	# of candidates for the cut points at the measurement (y) axis
E'	# of candidates for the cut points at the time (x) axis
K_{max}	max. # of bins at the measurement (y) axis
K'_{max}	max. # of bins at the time (x) axis

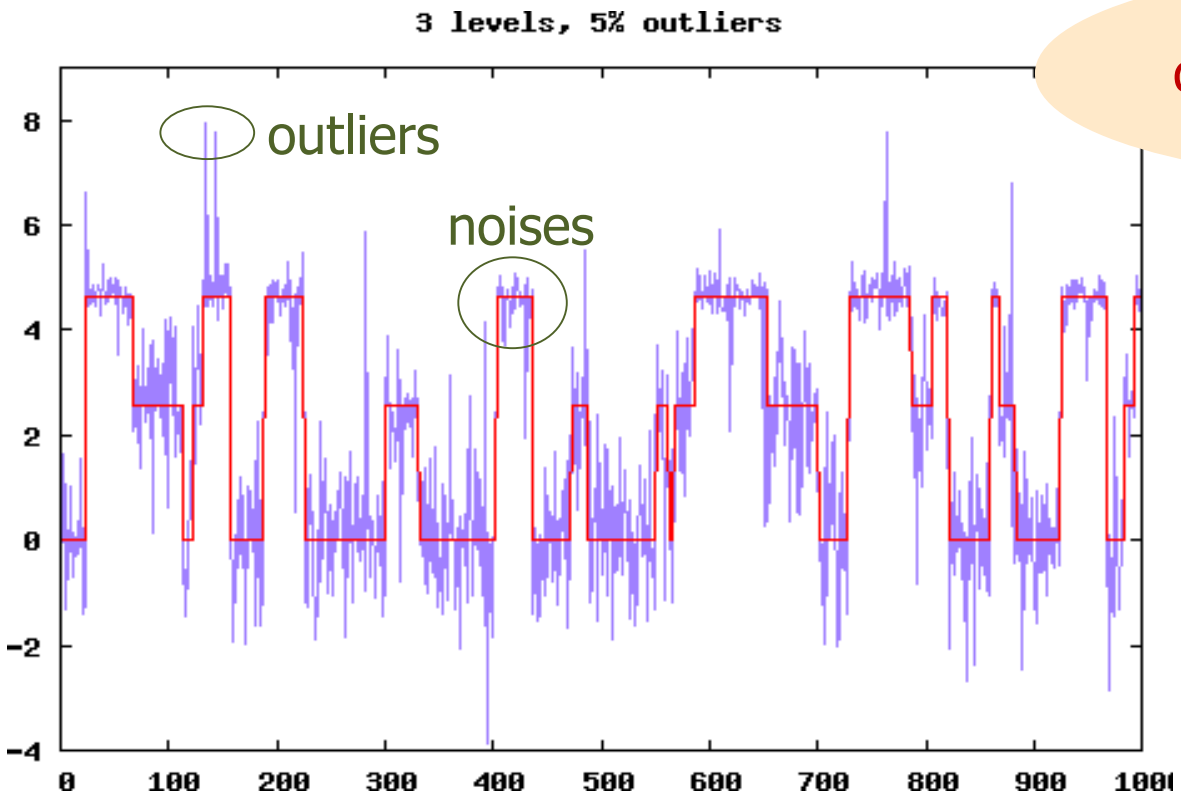
→ Control parameters
(determined by the trade-off between time and quality)

Outline

- ✓ Background: Unsupervised discretization of time series data
- ✓ Our proposal: Histogram-based discretization
- Experiments
- Conclusion/Future work

Experiment 1: enduring-state dataset

- Originally introduced in [Mörchen et al. 05]
- Comparison on the predictive performance among the discretizers

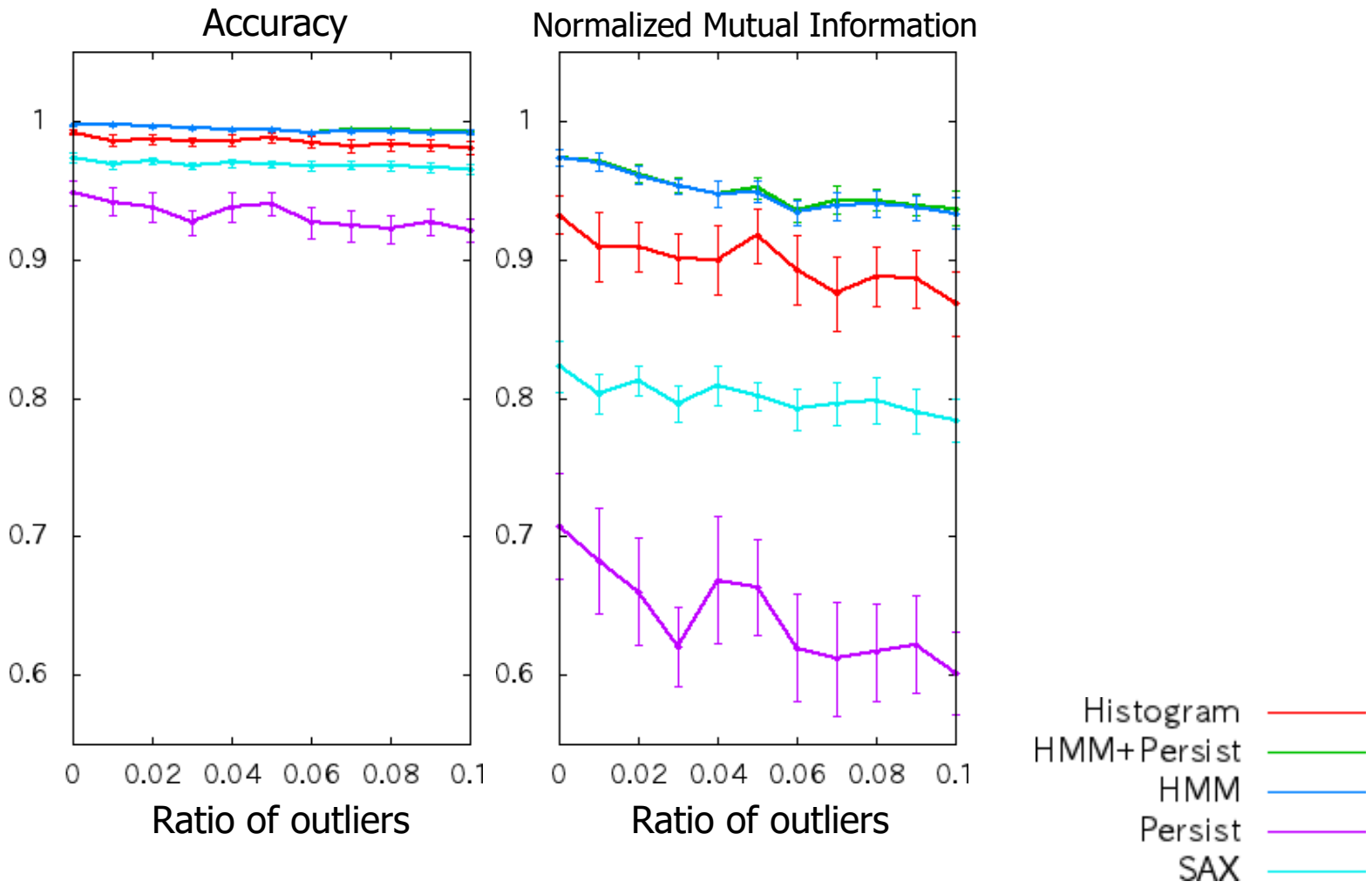


How well do the discretizers recover the answers?

- SAX
- Persist
- HMMs
- HMMs + Persist
  [Kameya et al. 2010]
- Histogram
  (our proposal)

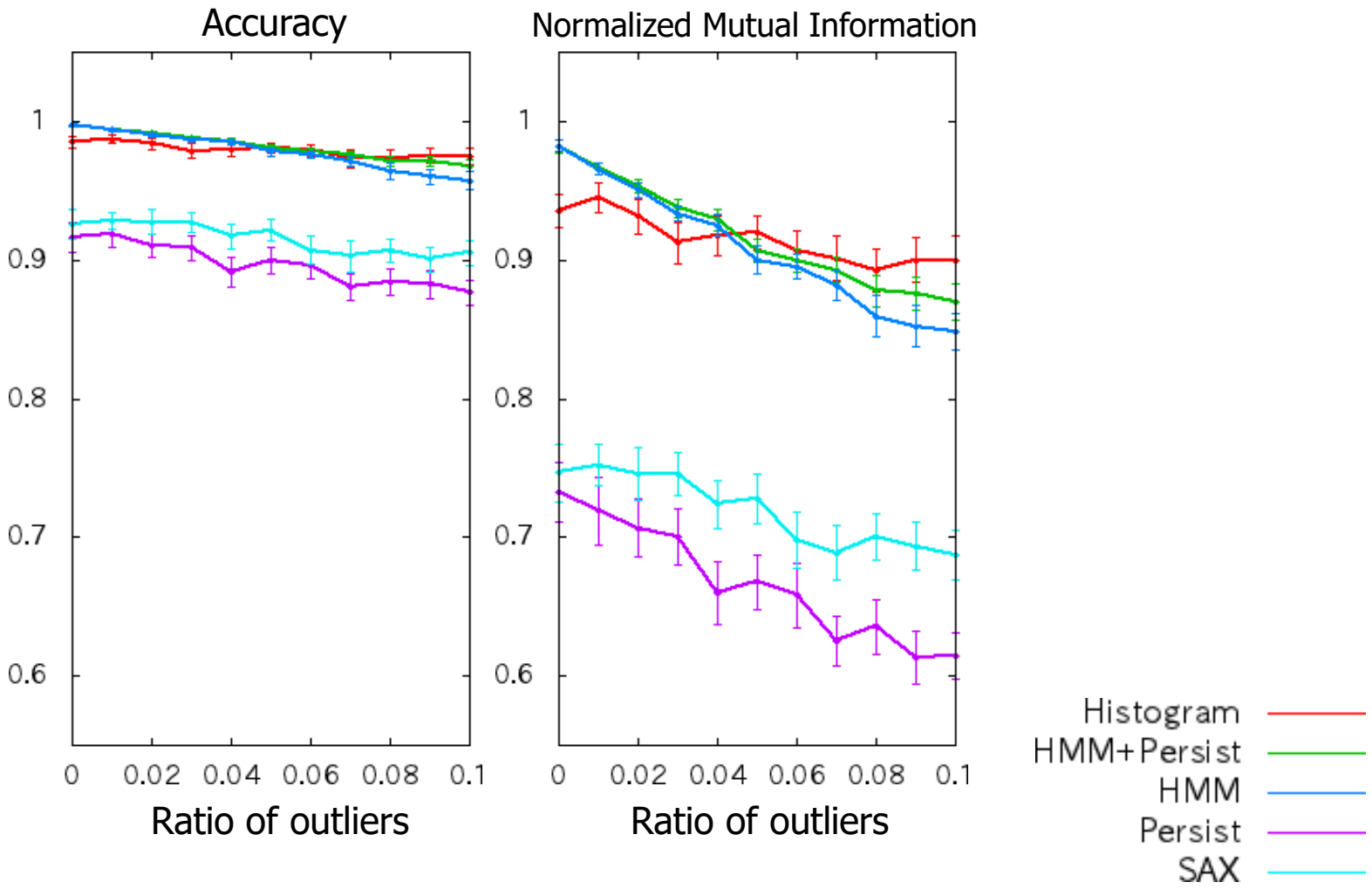
Experiment 1: enduring-state dataset (cont'd)

- # of discrete levels = **2**



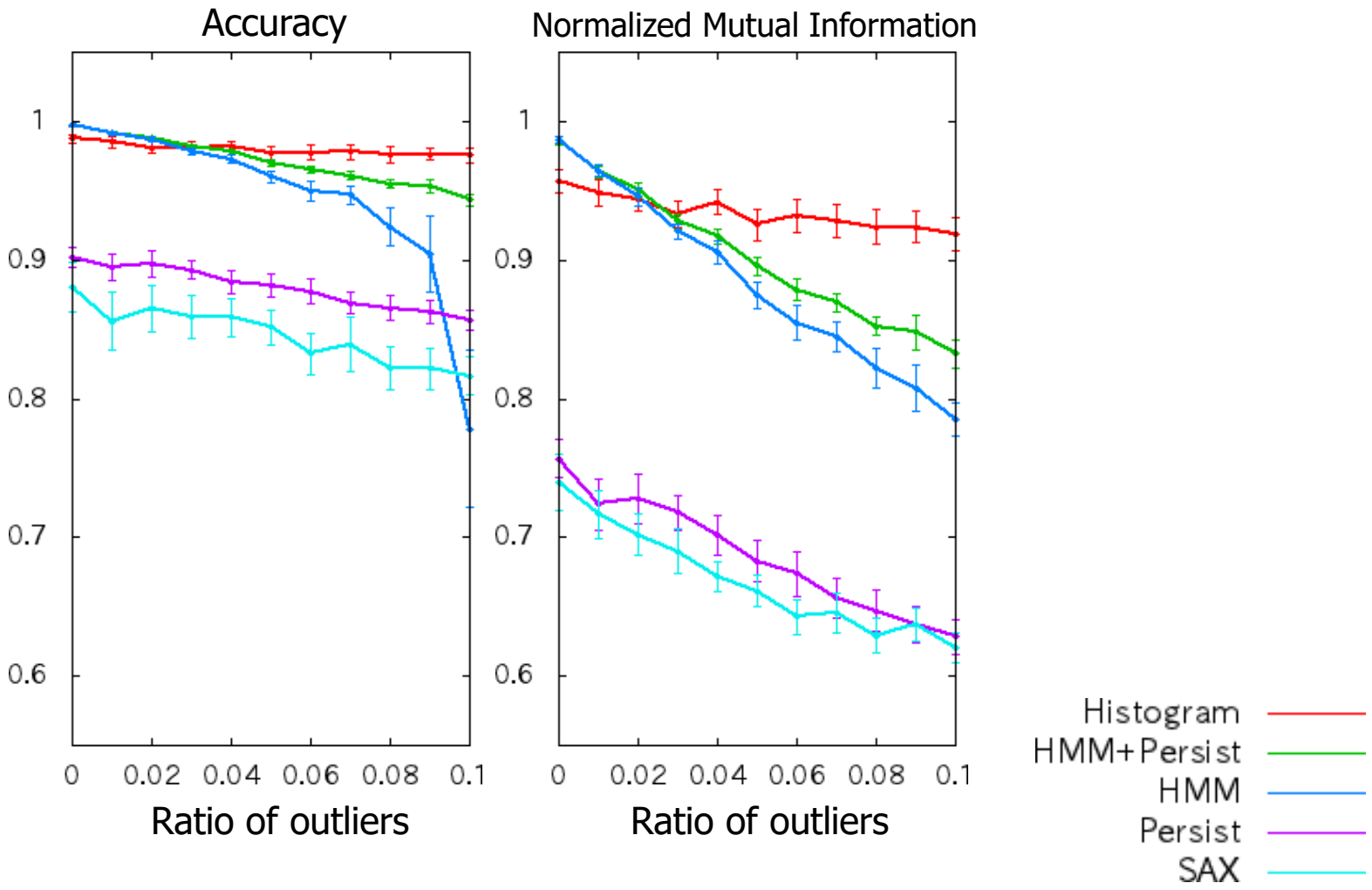
Experiment 1: enduring-state dataset (cont'd)

- # of discrete levels = **3**



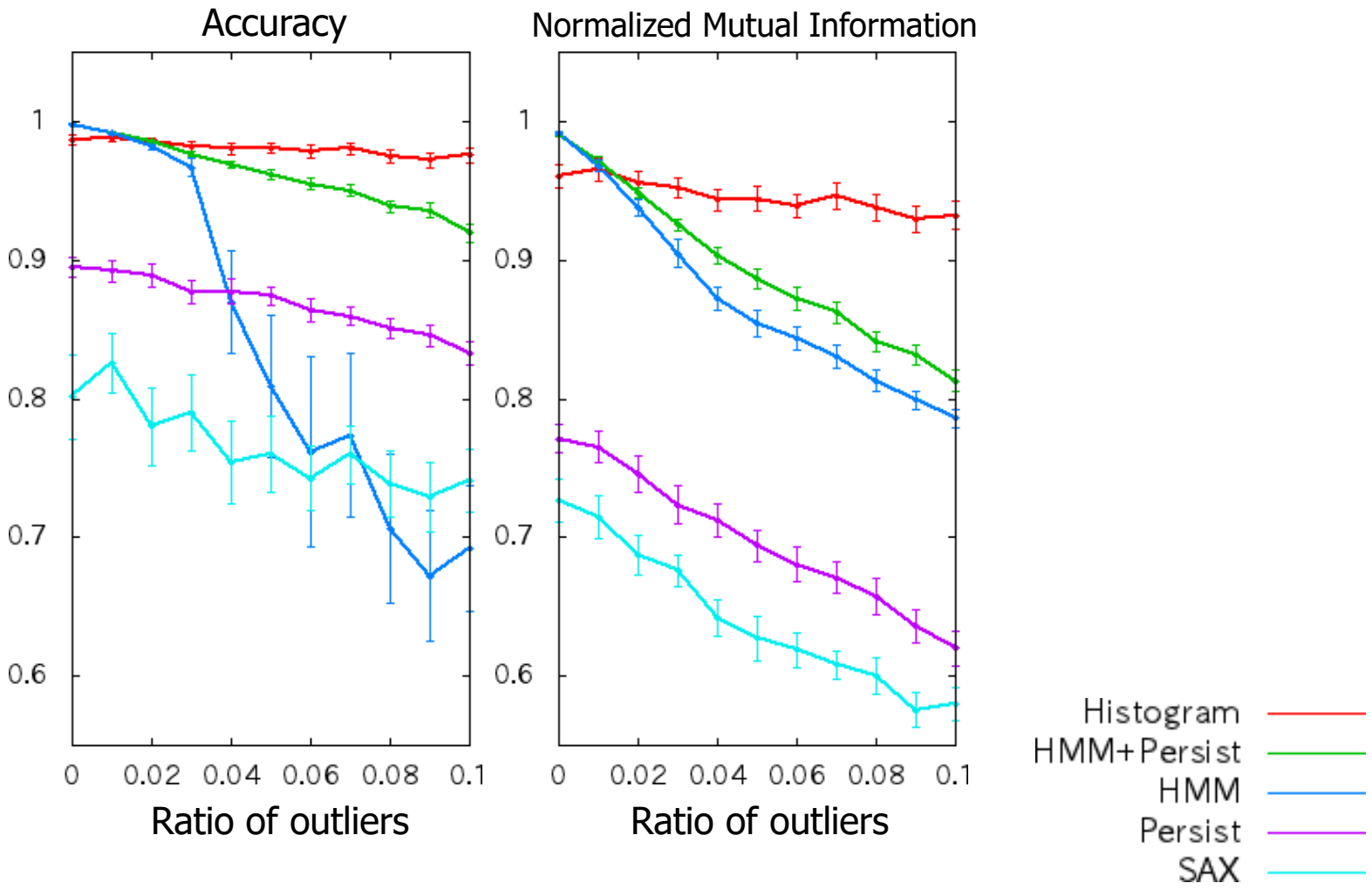
Experiment 1: enduring-state dataset (cont'd)

- # of discrete levels = **4**



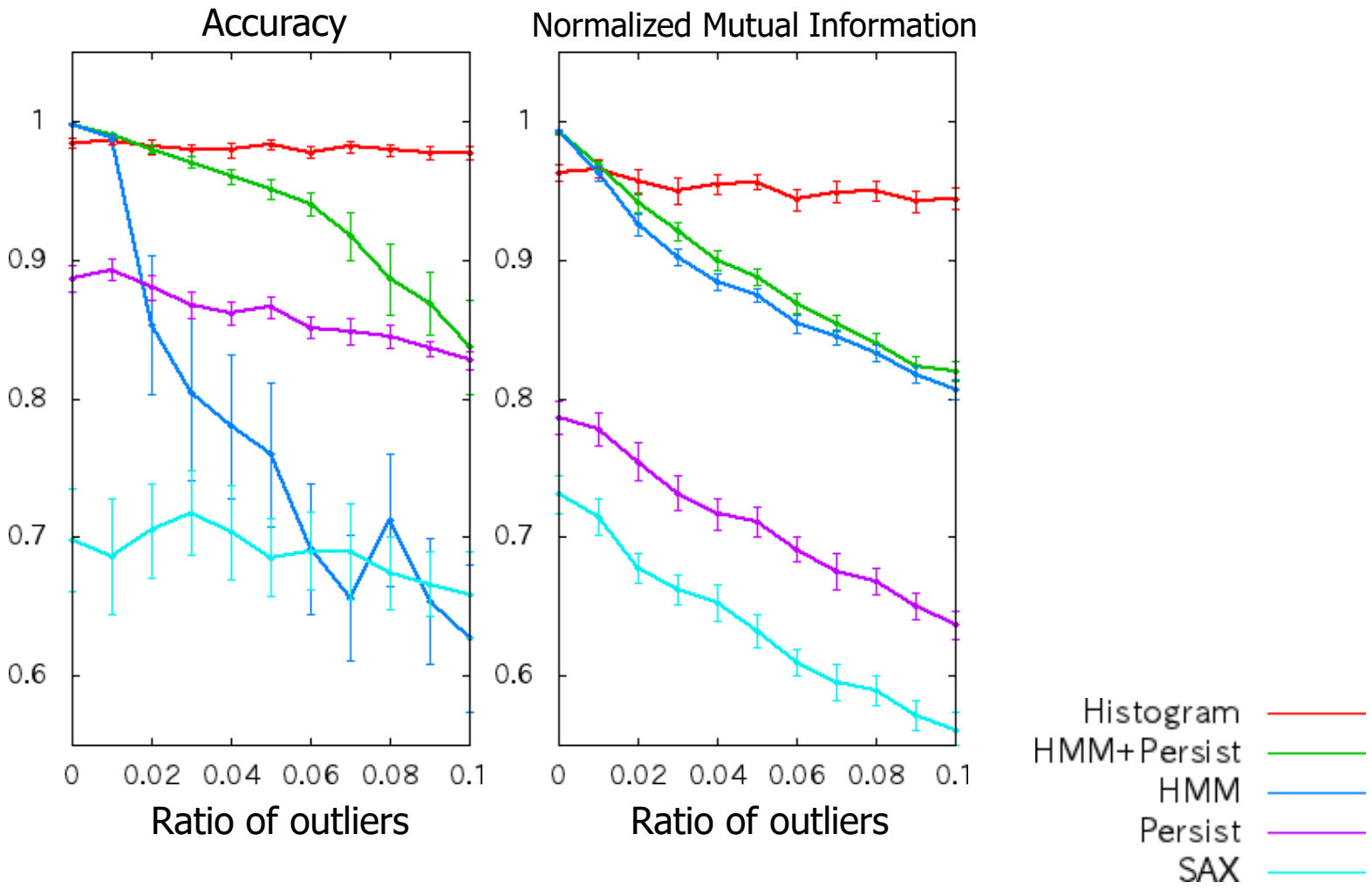
Experiment 1: enduring-state dataset (cont'd)

- # of discrete levels = **5**



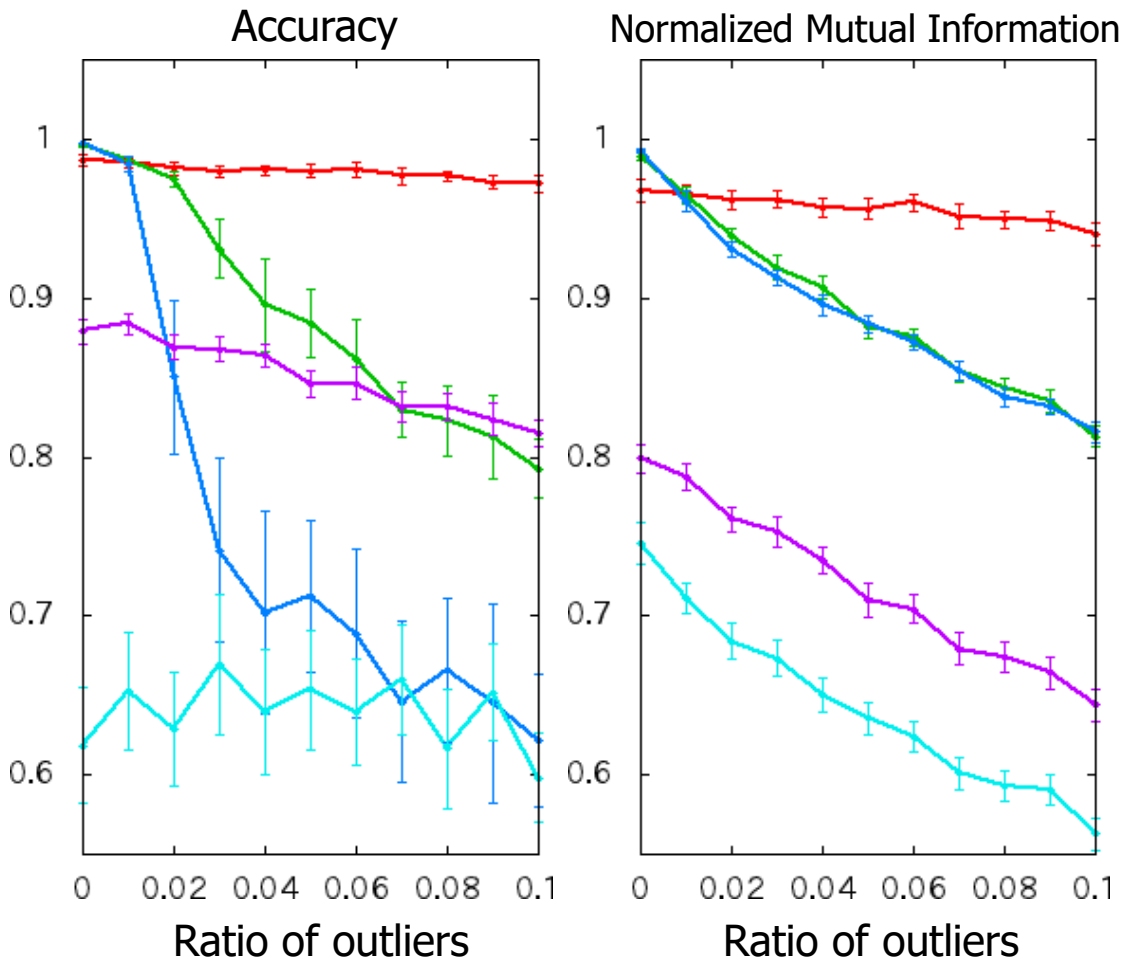
Experiment 1: enduring-state dataset (cont'd)

- # of discrete levels = **6**

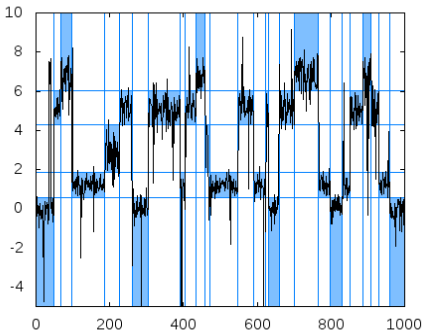


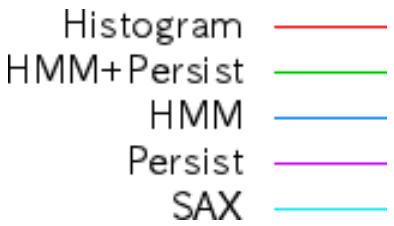
Experiment 1: enduring-state dataset (cont'd)

- # of discrete levels = **7**



Histogram-based discretizer works quite robustly than existing discretizers due to capturing the global behavior and strong smoothing

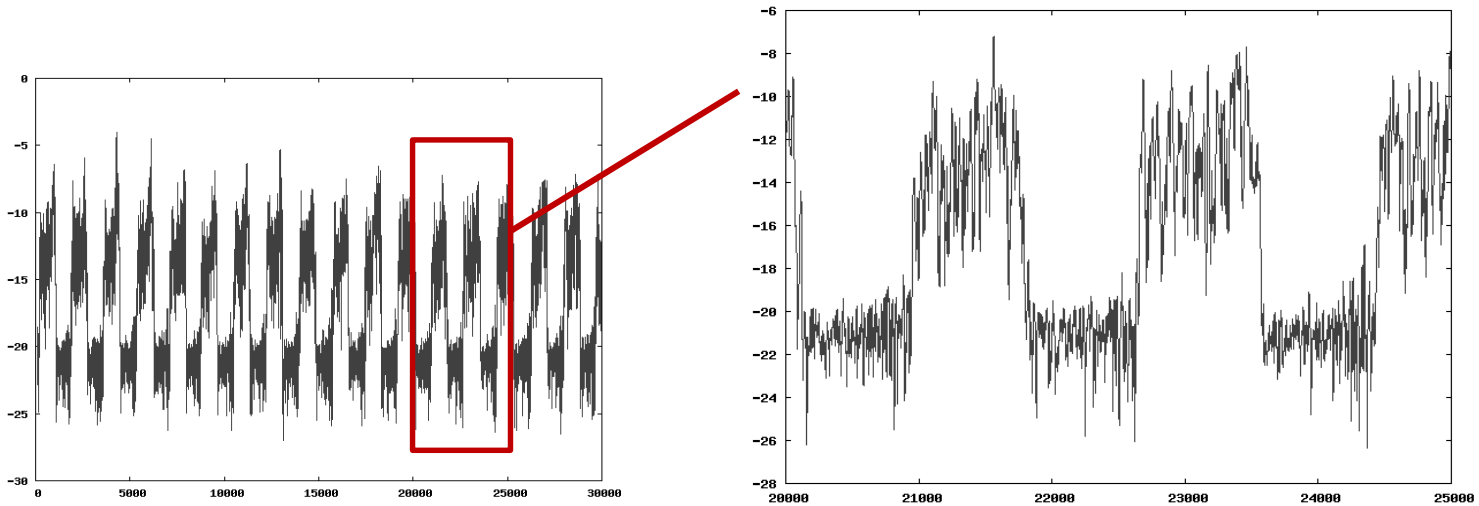




http://en.wikipedia.org/wiki/File:Patinaje_Euro2004.jpg

Experiment 2: Background

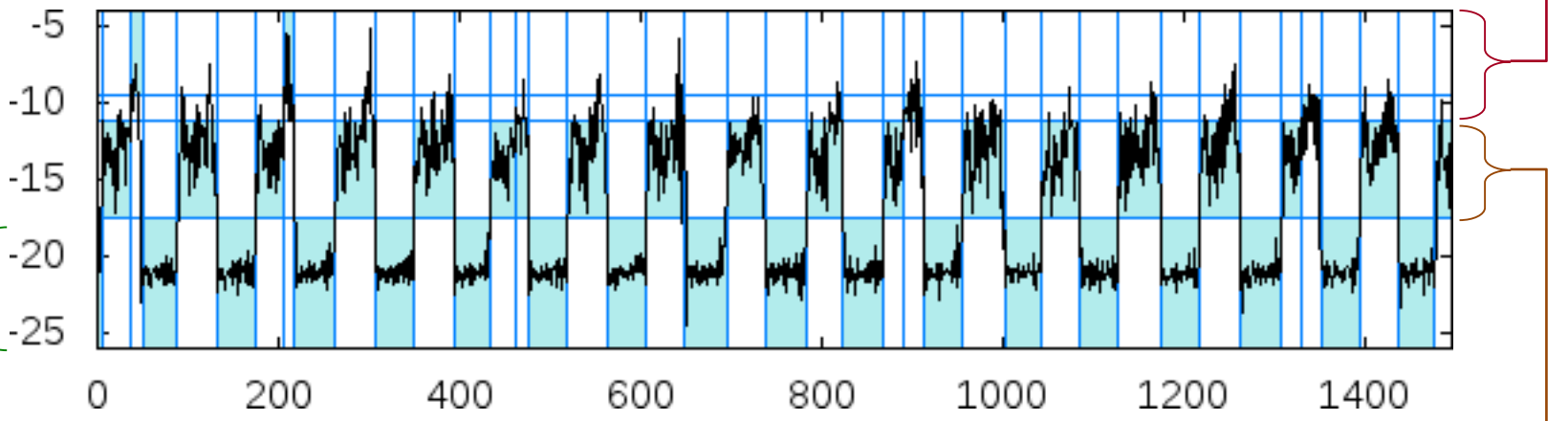
- Also used in [Mörchen et al. 05a]
- Data on muscle activation of a professional inline speed skater
  - Nearly 30,000 points recorded in log-scale
  - Time series is compressed by PAA (piece-wise approximate aggregation) [Lin et al. 07]



Experiment 2: Result

- A plausible # of discrete levels is *automatically* estimated with NML
- Cyclic behavior is clearly uncovered

Last kick to the ground to move forward



Muscle is not used

Gliding phase
(muscle is used to keep stability)

Outline

- ✓ Background: Unsupervised discretization of time series data
- ✓ Our proposal: Histogram-based discretization
- ✓ Experiments
- Conclusion/Future work

Conclusion

- Histogram-based discretizer for time series data
  - Based on the K&M method for finding optimal 1D histograms
  - More neutral, more intuitive and more robust
  - Polynomial-time complexity

Future work

- Handling long trends
- Applying pattern mining to discretized time series
- Histogram-based discretization of supervised tabular data