Change Point Detection in Streams

Ran Wolff, Univ. of Haifa, Israel

Joint work with Murad Badarna

- CPD Theory
- A New Algorithm
- Initial results

Changes in Data Streams

- Stream mining requires learning a little from every sample and then just patience
- Changes make life more interesting
  - Stock market
  - Industrial process control
- Cost of late detection can be high
  - Winner takes all scenarios
  - Production pipelines

Some Statistical CPD Theory

- Necessary **focus**:
  - A priori definition of interesting changes
  - We focus on change of mean
- Necessary **tradeoff** between contradictive objectives:
  - CPD theory: $ARL-TP \geqslant \log\left(ARL-FP\right)/magnitude$
- Algorithm is **meaningful** if: ARL-TP < ARL-FP

Optimal CPD

- Given the current prefix $x_1, x_2, \ldots, x_n$
- Best parametrized estimate before and after each sample

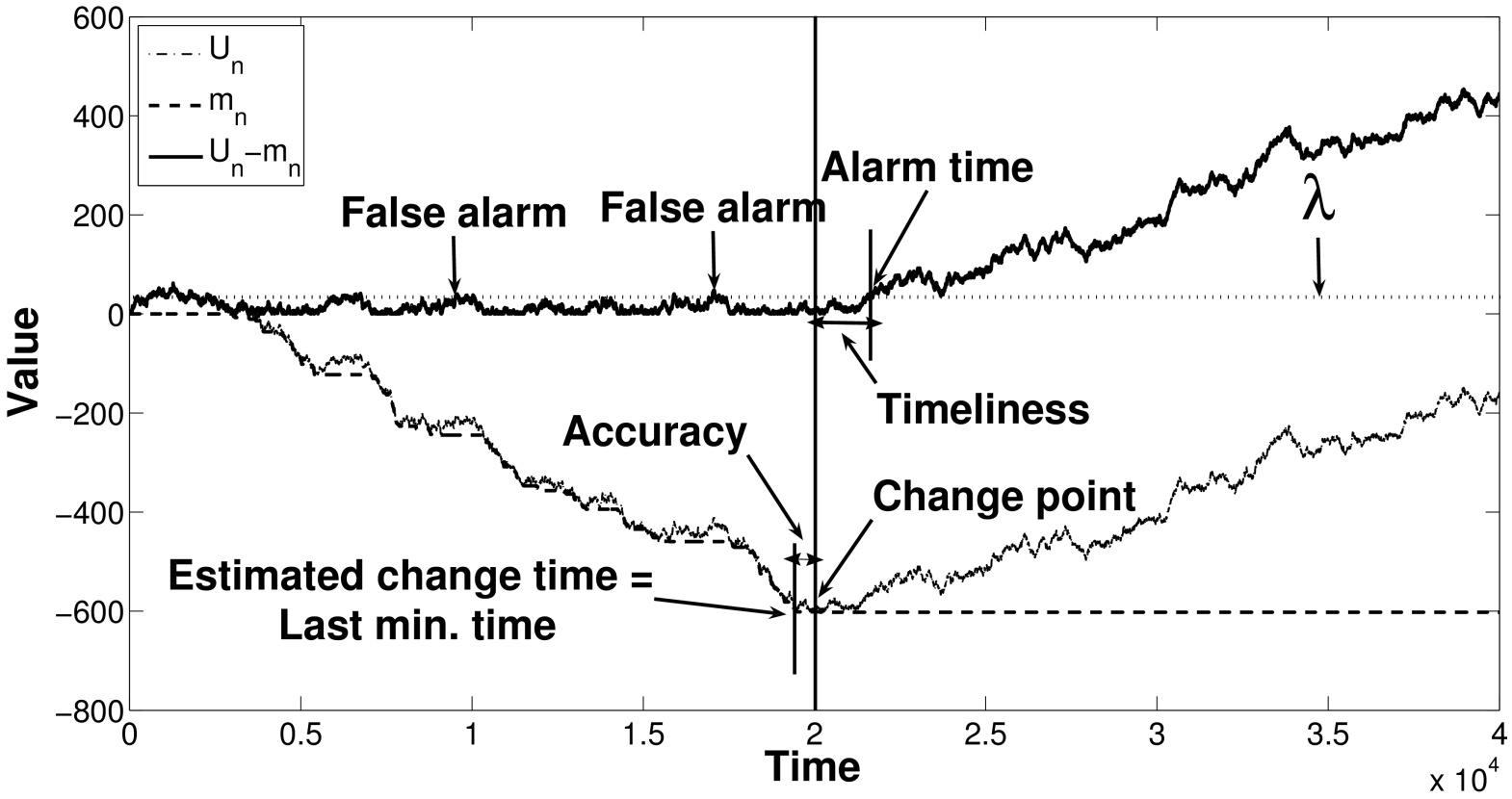
$$X_i^{pre}, X_i^{post}$$

- Compute the maximal joint probability

$$cp = arg\,max_i \;\; Pr\left(x_1, \ldots, x_i \sim X_i^{pre}, x_{i+1}, \ldots, x_n \sim X_i^{post}\right)$$

- Select an alert threshold to satisfy ARL-FP requirements
- Cost per sample is O(N)

Efficient On-Line Solutions



- Fails when data is noisy, otherwise – optimal ARL

ProTO Algorithm

- Break with current dichotomy
  - All candidates or best candidate
- Manage candidate CP smartly and efficiently
  - Keep candidates which may become best
- Two parts:
  - A test statistics
  - Upper & lower bounds

ProTO-T: Test Statistics

- Student's T - a standard two sample mean test

$$\frac{\hat{R}-\hat{S}}{\sqrt{\frac{v_R}{|R|}+\frac{v_S}{|S|}}}$$

- Applied to the i^{th} head & tail, R_i and S_i

$$T_i(n)=\frac{\hat{R}_i-\hat{S}_i}{\sqrt{\frac{v_{R_i}}{i}+\frac{v_{S_i}}{n-i}}}$$

  - Incrementally computable in O(1) per candidate

ProTO-T: Upper&Lower Bounds

- Convergence:

$$T_i(n) = \frac{\hat{R}_i - \hat{S}_i}{\sqrt{\frac{v_{R_i}}{i} + \frac{v_{S_i}}{n-i}}} \overset{n\to\infty}{\to} \left(\hat{R}_i - \mu_{post}\right)\sqrt{\frac{i}{v_{R_i}}}$$

  - Standard conf. interval:

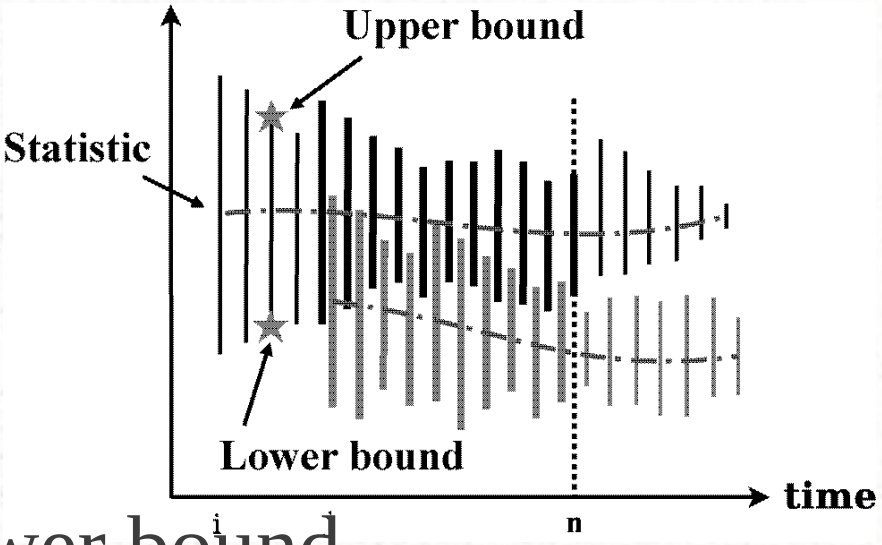
$$\mu_{post}(n-i) = \hat{S}_i \pm t^*_{1-\alpha/2}\frac{sd}{n-i}$$

  - w.p. $1-\alpha$

$$T_i(\infty) \in \left(\hat{R}_i - \hat{S}_i \pm t^*_{1-\alpha/2}\frac{sd}{n-i}\right)\sqrt{\frac{i}{v_{R_i}}}$$

ProTO-T: Algorithm

Consider two candidates



- Find the maximal lower bound

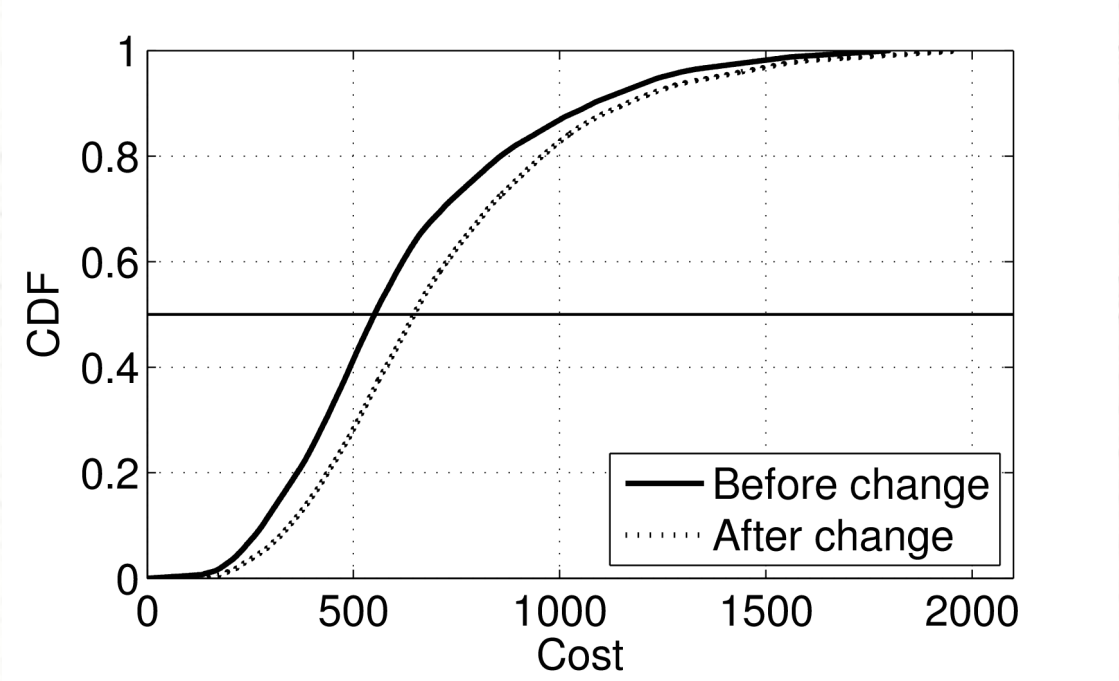
$$T_{max}^{low}(n) = \left(\hat{R}_{max} - \hat{S}_{max} - t^{*}_{1-\alpha/2} \frac{sd}{n-max} \right) \sqrt{\frac{max}{v_{R_{max}}}}$$

- Discard any candidate below it

$$T_{j}^{high}(n) = \left(\hat{R}_{j} - \hat{S}_{j} + t^{*}_{1-\alpha/2} \frac{sd}{n-j} \right) \sqrt{\frac{j}{v_{R_j}}} < T_{max}(n)$$

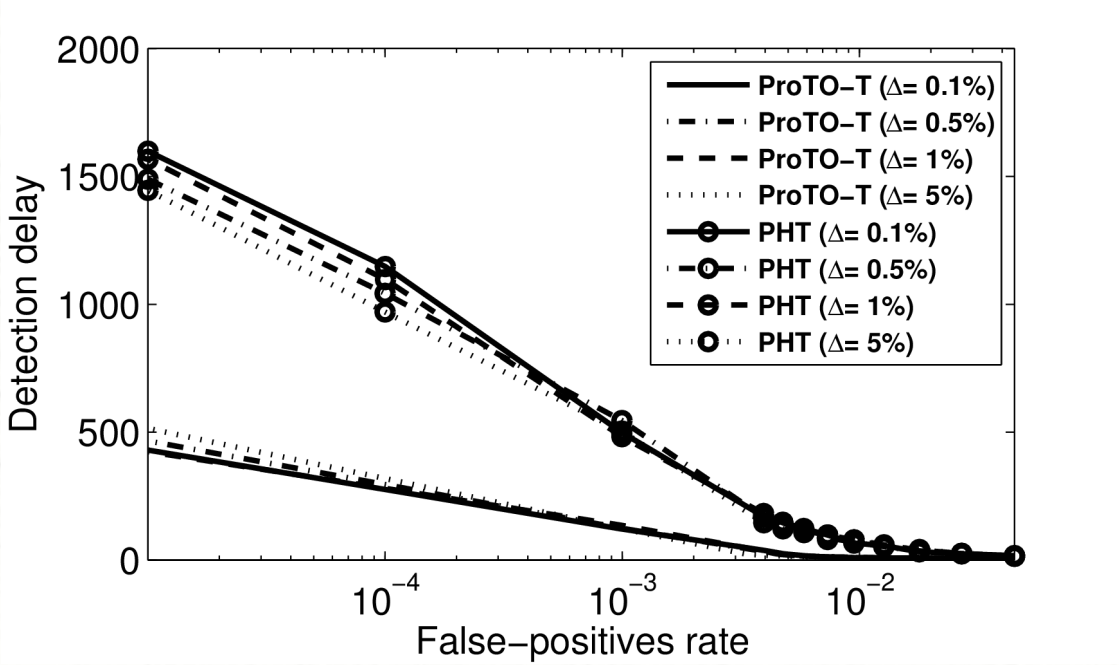
Results

- Constant expectancy for # of candidates



Results

- Supperior detection trade-off



Open Questions

- Applications
- Real data
- Multi dimentional data
- Relation to CPD theory

Thank you!