Change-Point Detection in a Sequence of Bags-of-Data

An Extension of Anomaly Analysis

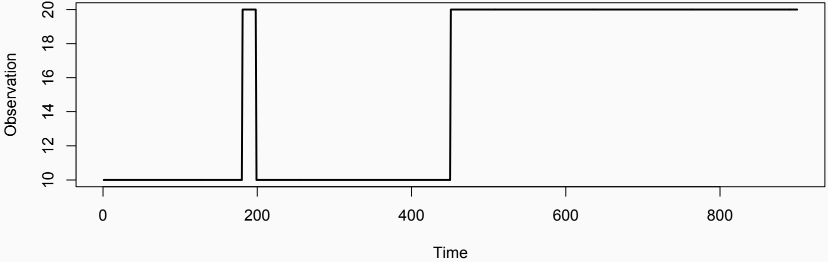
Noboru Murata

June 19, 2023

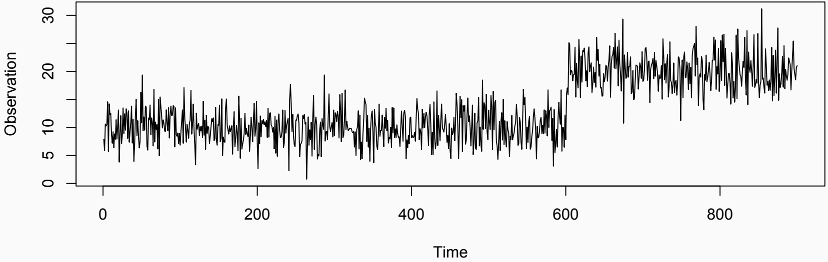
https://noboru-murata.github.io/

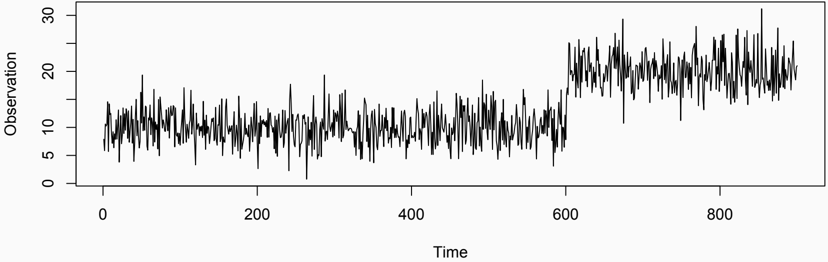
- Introduction
  - motivated examples
  - target problem
- Problem Formulation
  - change-point in bags-of-data
  - metric of bags-of-data
  - two sample problem for bags-of-data
- Numerical Examples
  - enron corpus analysis
- Conclusion

Introduction



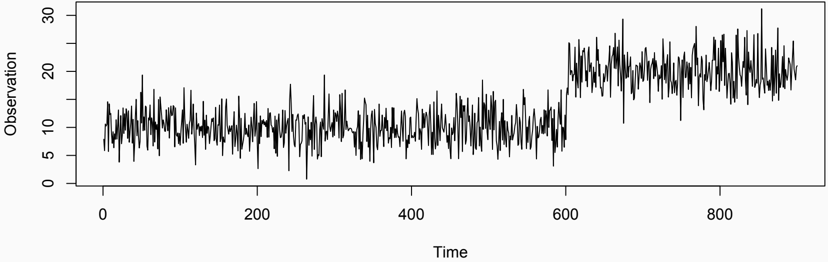
- objective
  - anomaly detection
    — find an outlier of time series
  - change-point detection
    — find a drastic change of time series


Observation
0
5
10
20
30
0
200
400
600
800
Time



- generating mechanism

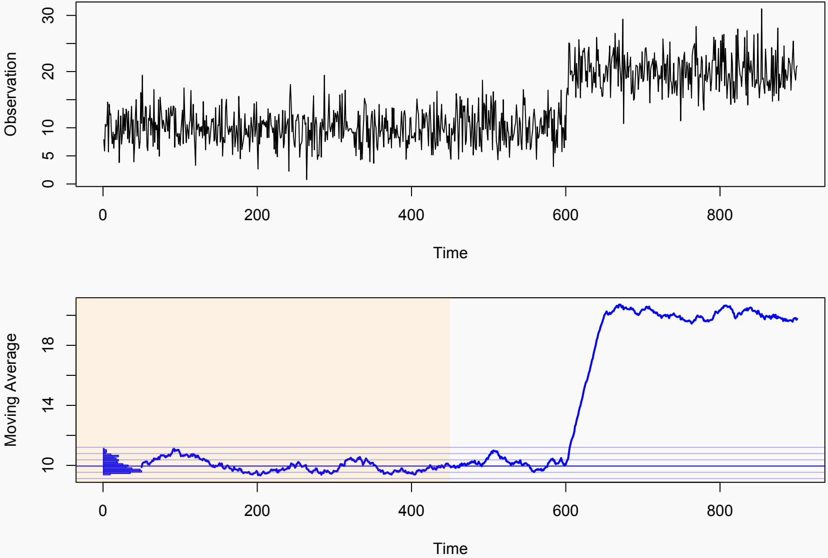
$$X_t = \begin{cases} c_0 + \varepsilon_t, & t < t_0, \\ c_1 + \varepsilon_t, & t \geq t_0, \end{cases} \qquad \varepsilon_t \sim P$$

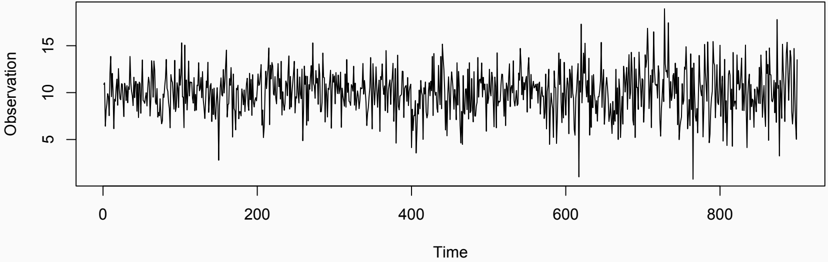


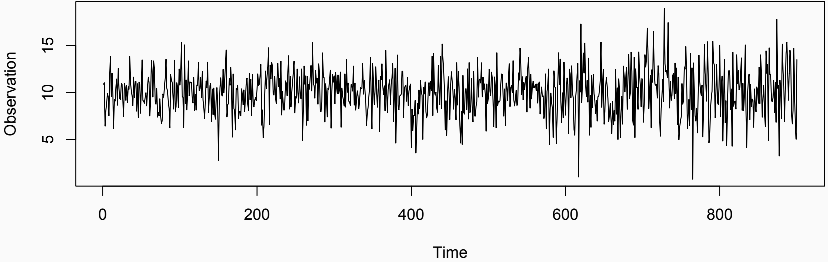
- summary statistics

$$\bar{X}_t = \frac{1}{\tau} \sum_{i=0}^{\tau-1} X_{t-i}$$

estimates of mean values (moving average)

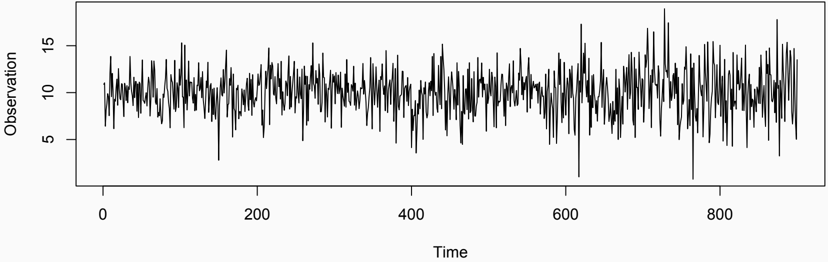

Observation
0
5
10
20
30
0
200
400
600
800
Time
Moving Average
10
14
18
0
200
400
600
800
Time


Observation
15
10
5
0
200
400
600
800
Time



- generating mechanism

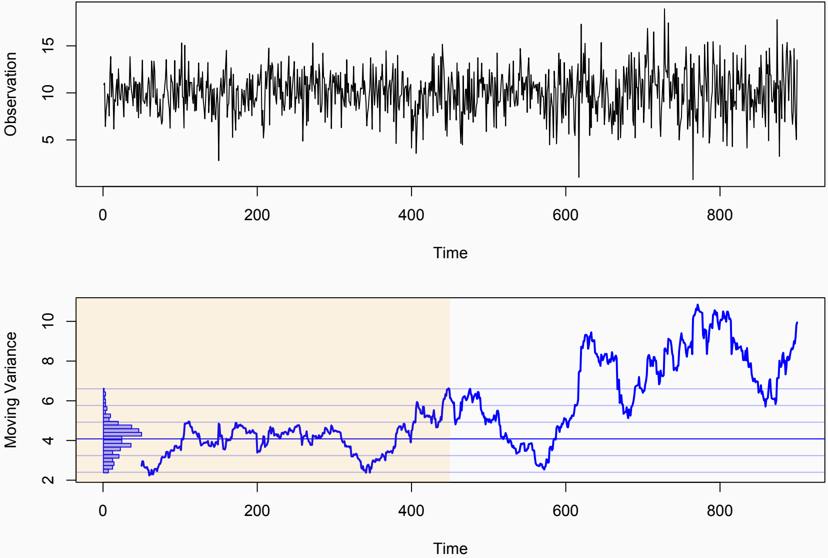
$$X_t = \begin{cases} c_0 + \varepsilon_t, & t < t_0, \quad \varepsilon_t \sim P \\ c_0 + \xi_t, & t \geq t_0, \quad \xi_t \sim Q \end{cases}$$

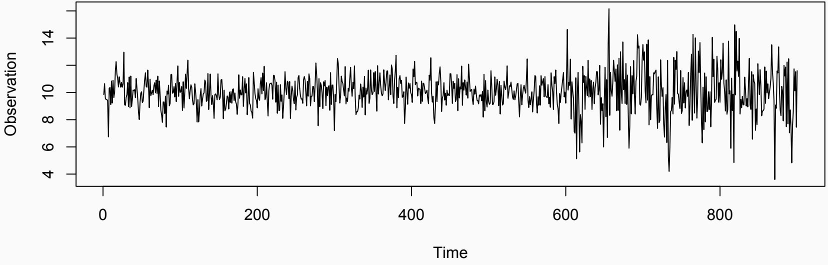


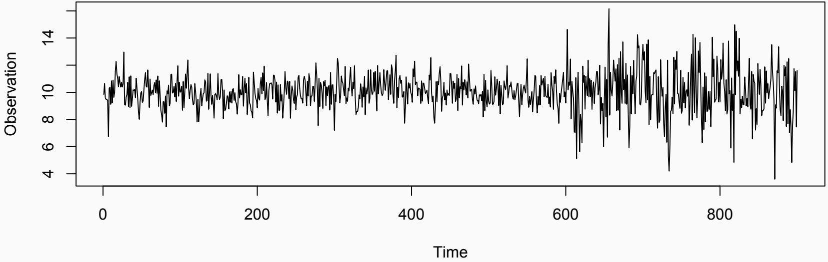
- summary statistics:

$$V_t = \frac{1}{\tau'} \sum_{i=0}^{\tau'-1} (X_{t-i} - \bar{X}_t)^2$$

estimates of variances

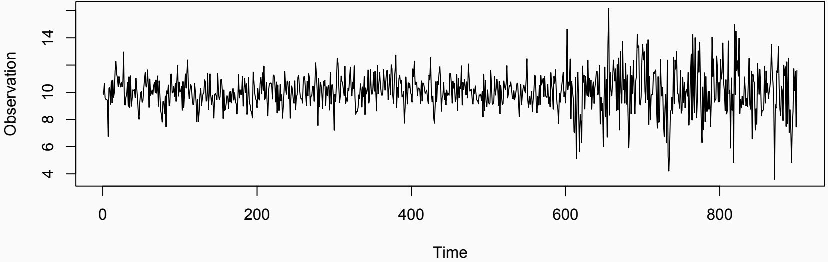

Observation
15
10
5
0
200
400
600
800
Time
Moving Variance
10
8
6
4
2
0
200
400
600
800
Time


Observation
4
6
8
10
14
0
200
400
600
800
Time



- generating mechanism

$$X_t = aX_{t-1} + bX_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim \begin{cases} P, & t < t_0, \\ Q, & t \geq t_0 \end{cases}$$

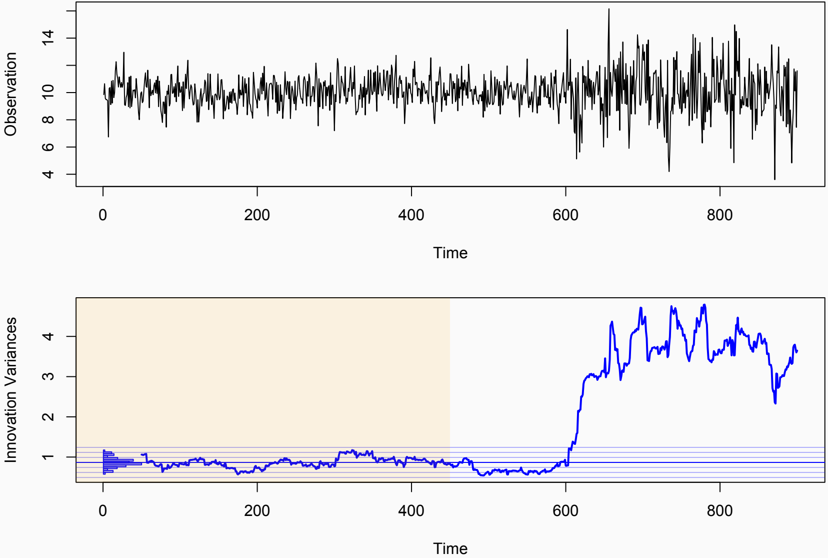


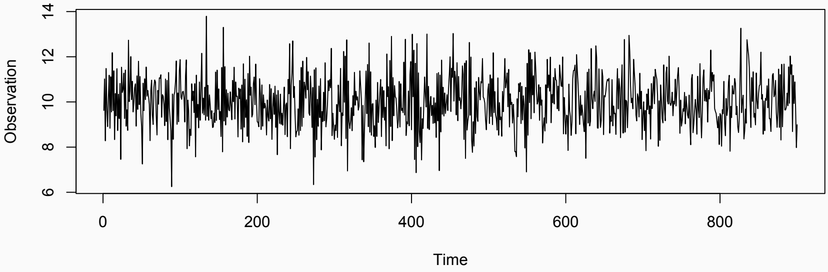
- summary statistics

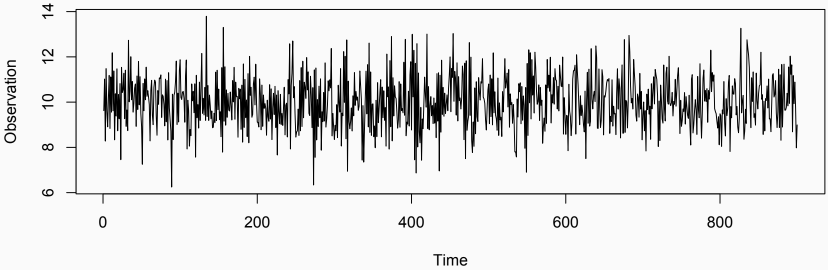
$$Var(\hat{\varepsilon}_t) \quad \text{(estimated from } X_t, X_{t-1}, \ldots)$$

estimates of innovation variances

$$\hat{\varepsilon}_t = X_t - \hat{X}_t = X_t - (\hat{a}X_{t-1} + \hat{b}X_{t-2})$$

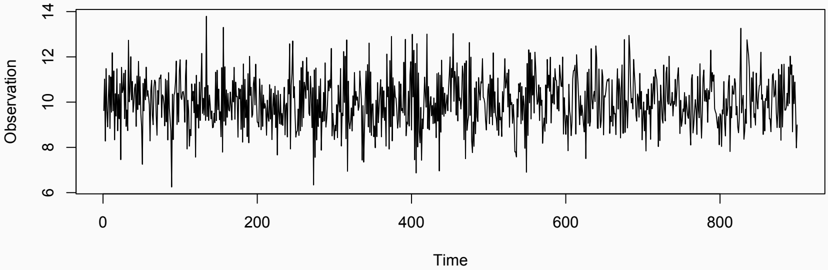

Observation
4
6
8
10
14
0
200
400
600
800
Time
Innovation Variances
1
2
3
4
0
200
400
600
800
Time


Observation
Time
6
8
10
12
14
0
200
400
600
800



- generating mechanism

$$X_t = \begin{cases} a_0 X_{t-1} + b_0 X_{t-2} + \varepsilon_t, & t < t_0, \\ a_1 X_{t-1} + b_1 X_{t-2} + \varepsilon_t, & t \geq t_0, \end{cases} \quad \varepsilon_t \sim P$$

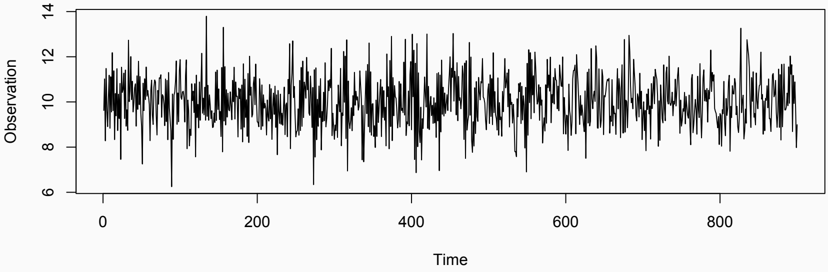


- summary statistics

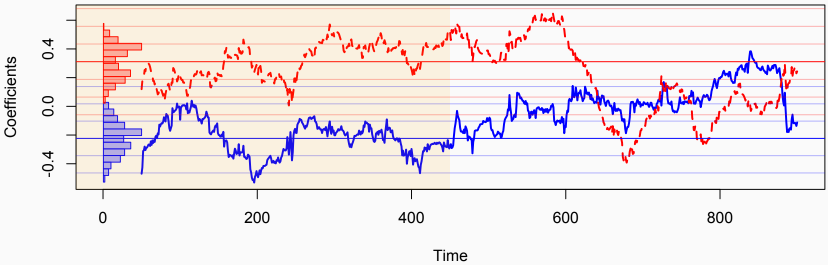
$$\hat{a}_t, \hat{b}_t \quad (\text{estimated from } X_t, X_{t-1}, \ldots)$$

estimates of coefficients

note: multi-dimensional problem



Observation
14
12
10
8
6
0
200
400
600
800
Time



Coefficients
0.4
0.0
-0.4
0
200
400
600
800
Time

Problem

find time points at which the generating mechanism of time series suddenly changes

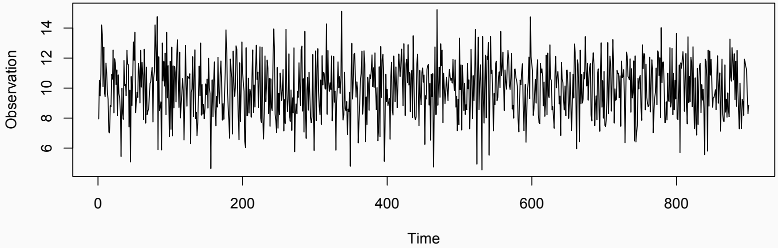
- applications
  - intrusion detection in computer networks
  - irregular-motion detection in vision systems
  - signal segmentation in data stream
  - fraud detection in cellular systems
  - fault detection in engineering systems
  - etc.

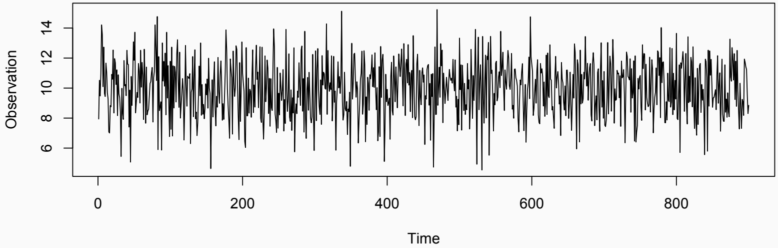
- framework
  - datum at time t: X_t
    a random variable (stochastic process)
    fixed length data vectors are considered
  - objective
    examine whether $X_t, X_{t+1}, \ldots$ differ from $X_{t-1}, X_{t-2}, \ldots$
    (or whether % X_t can be predicted from $X_{t-1}, X_{t-2}, \ldots$)
  - typical approach: define change-point scores, e.g.

$$\mathrm{score}(X_t) = -\log \Pr(X_t | X_{t-1}, X_{t-2}, \ldots)$$

summary statistics are used for specifying probability models

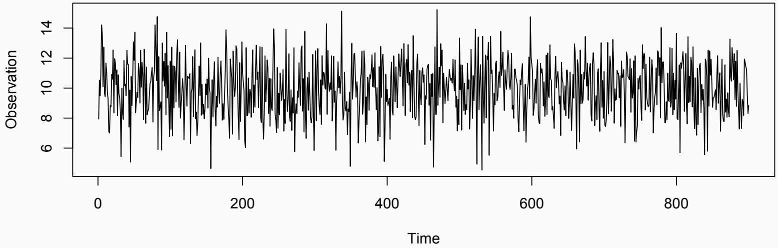
- representative algorithms
  - Singular Spectrum Analysis
    (Moskvinaa & Zhigljavskya, 2003)
  - ChangeFinder
    (Takeuchi & Yamanishi, 2006)
  - Kullback-Leibler Importance Estimation Procedure
    (Sugiyama et al. 2007)
- differences of these approaches
  - generative models of time series
  - computational costs
  - scalability of data size
  - sensitivity to change of regularity


Observation
6
8
10
12
14
0
200
400
600
800
Time



- generating mechanism

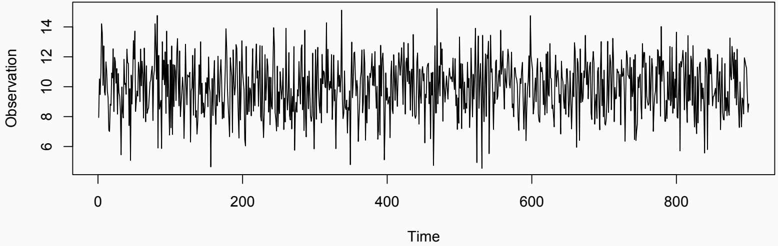
$$X_t = \begin{cases} c_0 + \varepsilon_t, & t < t_0, \quad \varepsilon_t \sim P \\ c_0 + \xi_t, & t \geq t_0, \quad \xi_t \sim Q \end{cases}$$

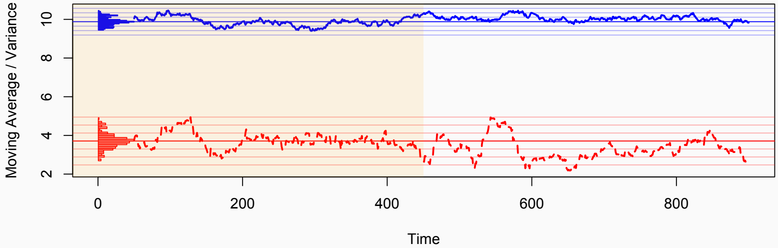


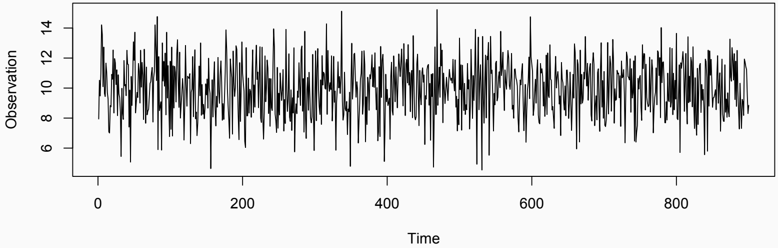
- summary statistics

$$\bar{X}_t = \frac{1}{\tau} \sum_{i=0}^{\tau-1} X_{t-i} \qquad \text{(moving average),}$$

$$V_t = \frac{1}{\tau'} \sum_{i=0}^{\tau'-1} (X_{t-i} - \bar{X}_t)^2 \qquad \text{(volatility)}$$


Observation
6
8
10
12
14
0
200
400
600
800
Time

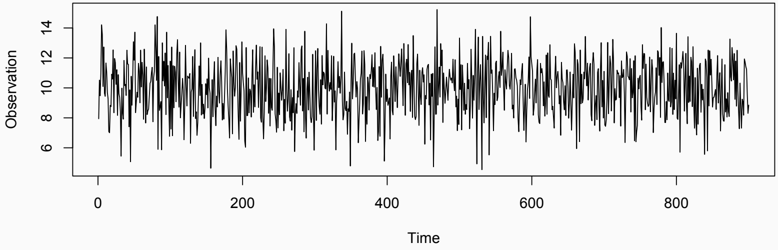

Moving Average / Variance
2
4
6
8
10
0
200
400
600
800
Time

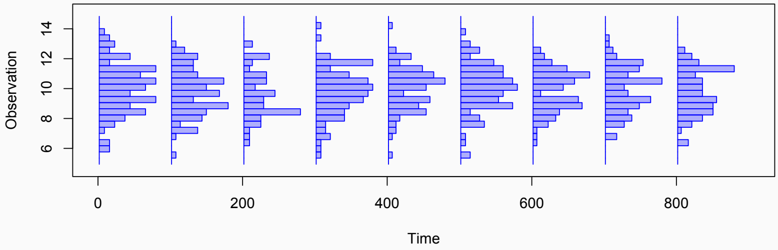


- summary statistics

$$\hat{P}_t = (\text{density estimates of } X_t, X_{t-1}, \ldots)$$

i.e. histogram, kernel density estimate, etc.

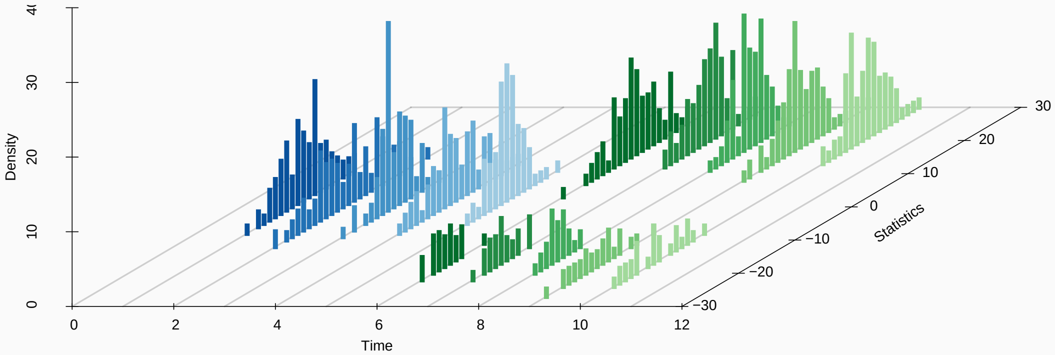

Observation
14
12
10
8
6
0
200
400
600
800
Time


Observation
14
12
10
8
6
0
200
400
600
800
Time

Problem Formulation

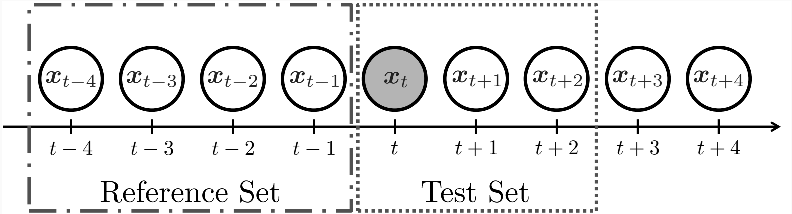
- framework
  - datum at time t: $B_t = \{X_i; i = 1, \ldots, n_t\}$
    a set of random variables, i.e. a bag of data
    size of bag can be different in time
  - objective:
    examine whether $B_t, B_{t+1}, \ldots$ differ from $B_{t-1}, B_{t-2}, \ldots$

    in statistical setup:
    examine whether $\Pr(B_t)$ is predictable from $\Pr(B_{t-1}), \Pr(B_{t-2}), \ldots$

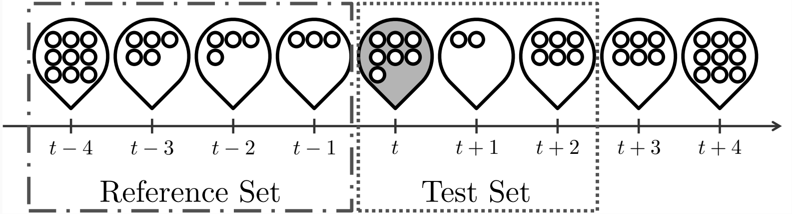


detect a change of distributions behind bags

- standard problem setting



- our problem setting



- graph-structured examples: sender-receiver scenario
  - internet incident detection
    (relation between source and destination hosts)
  - Enron email dataset
    (relation between mail senders and receivers)
  - market trading analysis
    (relation between buyers and sellers)
- other examples: multi-variate data
  - multi-sensor plant data
    (colinearlity analysis of non-stationary data)
  - follow-up surveys
    (random missing)

- parametric model

$$B_t = \{X_i\} \sim P_{\theta_t}$$

reduce to the change-point detection problem of $\{\theta_t\}$

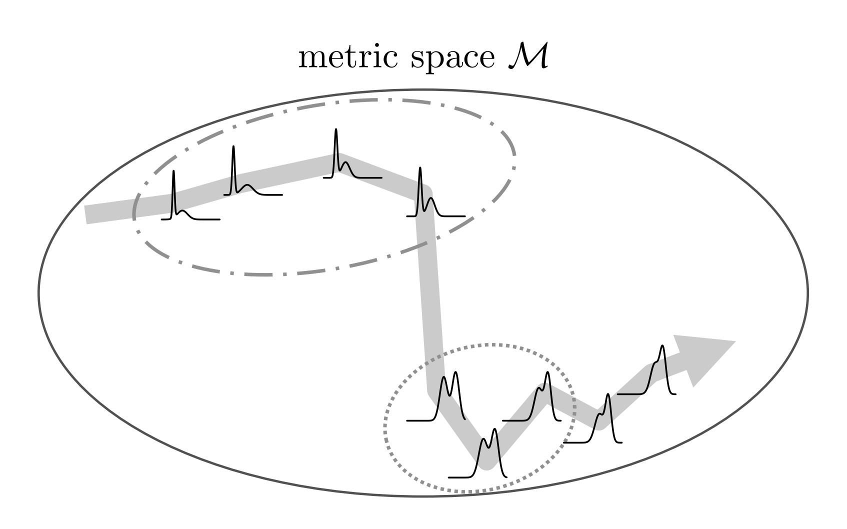
- non-parametric model

$$B_t = \{X_i\} \sim P_{B_t} \quad \text{(histogram, Parzen window, etc)}$$

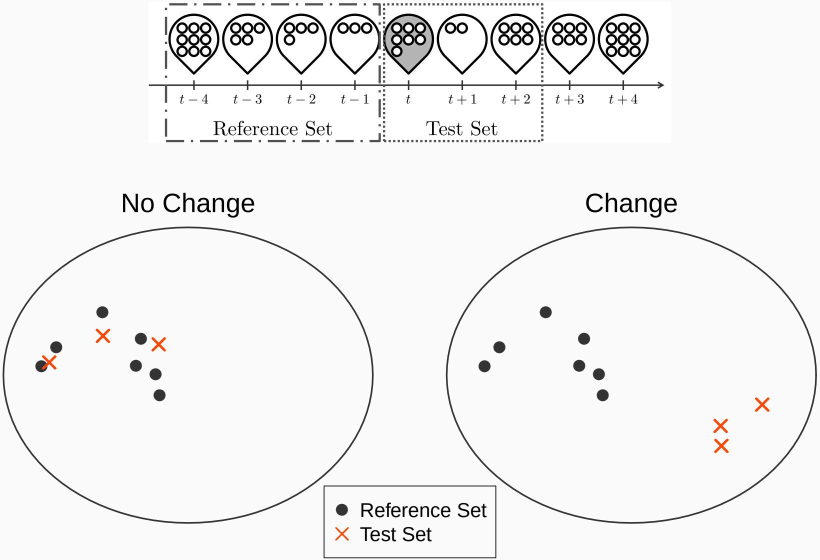
deal with probability distributions $\{P_{B_t}\}$

- non-parametric model: weighted data sets (histograms)
  - flexible for modeling various distributions
  - scalable for large sparse graphs

- twofold procedure for detection
  - embed each P_{B_t} in an appropriate metric space
  - examine whether fluctuation of $\{P_{B_t}\}$ is anomalous or not



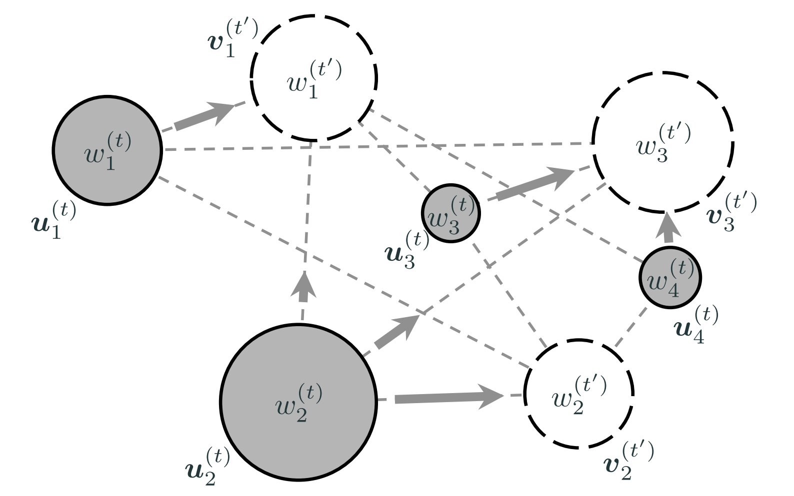
detect a significant change by following a path of bags


$t-4$ $t-3$ $t-2$ $t-1$ t $t+1$ $t+2$ $t+3$ $t+4$
Reference Set
Test Set
No Change
Change
Reference Set
Test Set

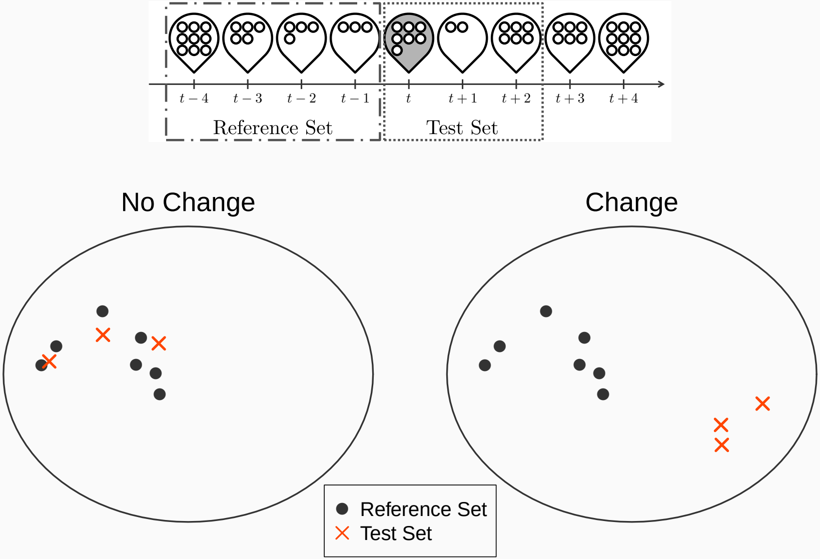
- distance between distributions P and Q:
  - the least amount of work needed to match two distributions, i.e. a kind of edit distance
  - proposed as a perceptually natural dissimilarity measure in computer vision
  - efficiently calculated by linear programming
  - mathematically equivalent to Wasserstein/Mallows distance

$$D(P, Q) = \inf_R \mathbb{E}_{(X,Y\sim R)}[d(X, Y)], \ (d \text{ can be any distance})$$

$$\text{where } P(X) = \int R(X, dy), \ \text{and } Q(Y) = \int R(dx, Y)$$



histogram: $\{(\text{bin}, \text{freq})\}$; $P = \{(\boldsymbol{u}, w)\}$, $Q = \{(\boldsymbol{v}, w')\}$


$t-4$
$t-3$
$t-2$
$t-1$
t
$t+1$
$t+2$
$t+3$
$t+4$
Reference Set
Test Set
No Change
Change
Reference Set
Test Set

Problem

given i.i.d. observations $\{x_i; i = 1, \ldots, m\} \sim P$ and $\{y_j; j = 1, \ldots, n\} \sim Q$, examine whether $P \neq Q$

- possible criteria
    - empirical mean (moment matching)
    - KL divergence with parametric models
    - KL divergence without models

- distance-based entropy estimators
  - bags with weights: $\mathfrak{D} = \{(B_i, w_i); i = 1, \ldots, n\}$
  - information content

$$I(B;\mathfrak{D}) = c + d \sum_{B_i \in \mathfrak{D}} w_i \log D(B_i, B) \qquad (c, d: \ \text{const.})$$

  - cross-entropy

$$H(\mathfrak{D}, \mathfrak{D}') = c + d \sum_{B_i \in \mathfrak{D}, B'_j \in \mathfrak{D}'} w_i w'_j \log D(B_i, B'_j)$$

  - auto-entropy

$$H(\mathfrak{D}) = c + d \sum_{B_i, B_j \in \mathfrak{D}, B_j \neq B_i} \frac{w_i w_j}{1 - w_i} \log D(B_i, B_j)$$

- reference and test datasets

$$\begin{aligned} \mathfrak{D}_t^{\text{ref}} &= \{(B_i, w_i); i = t-1, t-2, \dots\} && \text{(past bags)} \\ \mathfrak{D}_t^{\text{test}} &= \{(B_i, w_i); i = t, t+1, \dots\} && \text{(future bags)} \end{aligned}$$

  where weights are used as discounting factors
- likelihood ratio (f: density)

$$\text{score}_t = \log \frac{f_{\text{test}}(B_t)}{f_{\text{ref}}(B_t)} = I(B_t; \mathfrak{D}_t^{\text{ref}}) - I(B_t; \mathfrak{D}_t^{\text{test}})$$

- symmetric Kullback-Leibler divergence

$$\text{score}_t = \frac{2H(\mathfrak{D}_t^{\text{ref}}, \mathfrak{D}_t^{\text{test}}) - H(\mathfrak{D}_t^{\text{ref}}) - H(\mathfrak{D}_t^{\text{test}})}{2}$$

- Bayesian bootstrap: Bayesian analogue of the bootstrap

  instead of resampling from an empirical distribution, weighted samples are used where weights are sampled from the Dirichlet distribution

$$(N_1, \dots, N_k) \sim \mathrm{Mult}(n; \rho_1, \dots, \rho_k) \qquad \textit{(resampling)}$$

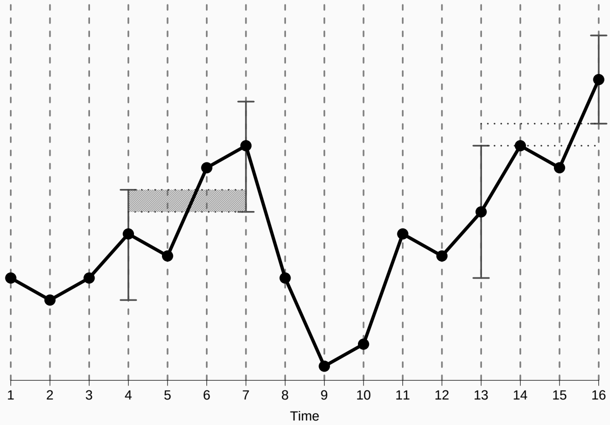
$$(W_1, \dots, W_k) \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_k) \qquad \textit{(reweighting)}$$

- if we let $\alpha_i = n\rho_i$:

$$\mathbb{E}[N_i] = \mathbb{E}[W_i] = \rho_i$$

$$\mathrm{Var}[N_i] = \mathrm{Var}[W_i] \cdot \frac{n+1}{n} = \frac{\rho_i(1-\rho_i)}{n}$$

- confidence interval with Baysian bootstrap on weights of bags
  - regular: intervals intersect each other
  - anomalous: otherwise

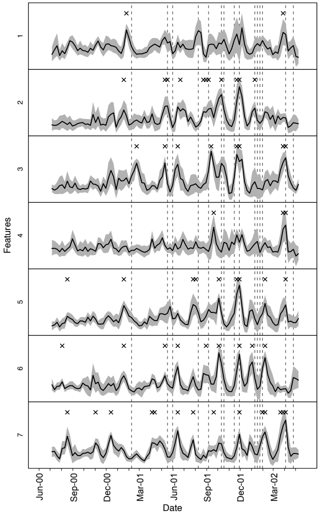


Numerical Examples

Enron Email Dataset (Cohen, 2009)

email transmission data from about 150 users, mostly senior management of Enron

- duration: 2000/6 – 2002/5 (accounting scandal: 2001)
- time window size of bags: 1 week
- size of reference datasets: 5 weeks
- size of test datasets: 3 weeks
- statistics in bags: 7 stats of bipartite graphs
    - degree of sender / receiver
    - 2nd order degree of sender-sender / receiver-receiver
    - number of messages from sender / to receiver
    - number of messages between sender and receiver
- confidence interval: 0.95



Date	Proposed	GS	Event
February 12, 2001	X	X	Jeff Skilling becomes chief executive of Enron.
May 19, 2001	X		Congress begins implementing President Bush's energy plan into legislation.
June 5, 2001	X	X	Rove divests his stocks in energy.
August 14, 2001	X	X	Skilling resigns abruptly citing personal reasons. Kenneth Lay returns to CEO.
September 11, 2001	X		Four terrorist attacks launched by al-Qaeda.
October 16, 2001	X		Enron reports a $618 million loss and a $1.2 billion reduction in shareholder equity.
October 19, 2001	X		Securities and Exchange Commission launches inquiry into Enron finances.
November 19, 2001	X	X	Enron restates its third-quarter earnings and says a $690 million debt is due Nov. 27.
November 29, 2001	X	X	Dynegy deal collapses.
December 2, 2001	X		Enron files for bankruptcy, the biggest in US history, and lays off 4,000 employees.
January 9, 2002	X	X	The justice department opens a criminal investigation of Enron.
January 17, 2002			Enron fires Andersen blaming the auditor for destoying Enron documents.
January 23, 2002		X	Kenneth Lay resigns as chairman and chief executive of Enron.
January 30, 2002	X	X	Enron names Stephen F. Cooper new CEO.
February 4, 2002	X	X	Kenneth Lay resigns from the board.
April 9, 2002	X		David Duncan, Andersen's former top Enron auditor, pleads guilty to obstruction.
April 24, 2002		X	House passes accounting reform package.

Conclusion

we consider

- change-point detection for sequence of bags of data
- a statistically appropriate distance between bags-of-data
- change-point scores based on entropy estimators
- confidence intervals with Bayesian bootstrap

possible extension would be

- on-line detection with stable entropy estimators
- on-line adaptive thresholding

Hino, Hideitsu and Noboru Murata (Oct. 2013). "Information estimators for weighted observations." In: *Neural Networks* 46, pp. 260–275. DOI: 10.1016/j.neunet.2013.06.005.

Koshijima, Kensuke, Hideitsu Hino, and Noboru Murata (Oct. 1, 2015). "Change-Point Detection in a Sequence of Bags-of-Data." In: *IEEE Transactions on Knowledge and Data Engineering* 27.10, pp. 2632–2644. DOI: 10.1109/TKDE.2015.2426693.

Moskvina, Valentina and Anatoly Zhigljavsky (2003). "An Algorithm Based on Singular Spectrum Analysis for Change-Point Detection." In: *Communications in Statistics - Simulation and Computation* 32 (2), pp. 319–352. DOI: 10.1081/SAC-120017494.

Sugiyama, Masashi et al. (2008). "Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation." In: *Advances in Neural Information Processing Systems*. Neural Information Processing Systems (Vancouver, B.C., Canada, Dec. 3–8, 2007). Ed. by John C. Platt et al. Vol. 20. Neural Information Processing Systems Foundation. Curran Associates, Inc.

- Sun, Jimeng et al. (Aug. 2007). **"GraphScope: parameter-free mining of large time-evolving graphs."** In: *Proceedings of KDD'07*. the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (San Jose, CA, USA, Aug. 12–15, 2007). Ed. by Pavel Berkhin, Rich Caruana, and Xindong Wu. SIGKDD: The community for data mining, data science and analytics. New York, NY, USA: Association for Computing Machinery, pp. 687–696. DOI: 10.1145/1281192.1281266.
- Takeuchi, Jun-ichi and Kenji Yamanishi (Apr. 2006). **"A unifying framework for detecting outliers and change points from time series."** In: *IEEE Transactions on Knowledge and Data Engineering*, pp. 482–492. DOI: 10.1109/TKDE.2006.1599387.