M-Statistic for Kernel Change-Point Detection

Shuang Li, Yao Xie, Hanjun Dai, Le Song

Georgia Institute of Technology

Motivation

Design a nonparametric change-point detection method (M-statistic) in both the offline and online setting.

- **Distribution Free:** kernel approaches are distribution free as they provide consistent results over larger classes of data distributions.
- **Efficient to compute:** split data to compute the offline M-statistics in a novel structured way; update M-statistics recursively in the online setting.
- **Analytical way to obtain threshold:** accurately characterize the tail probability of the M-statistics under null hypothesis in both offline and online setting.
- **Powerful:** numerically demonstrate that our algorithm is more powerful and more robust compared to conventional parametric approaches (e.g., Hoteling's T^2).

MMD

Assume there are two sets with n observations from a domain $\mathcal{X}$, where $X = \{x_1, x_2, \ldots, x_n\}$ are drawn iid from distribution P, and $Y = \{y_1, y_2, \ldots, y_n\}$ are drawn iid from distribution Q.

- The maximum mean discrepancy (MMD) is defined as [1]

$$\mathrm{MMD}_0[\mathcal{F}, P, Q] := \sup_{f\in\mathcal{F}} \{\mathbb{E}_x[f(x)] - \mathbb{E}_y[f(y)]\}.$$

- An unbiased estimate of MMD_0^2 can be obtained using U-statistic

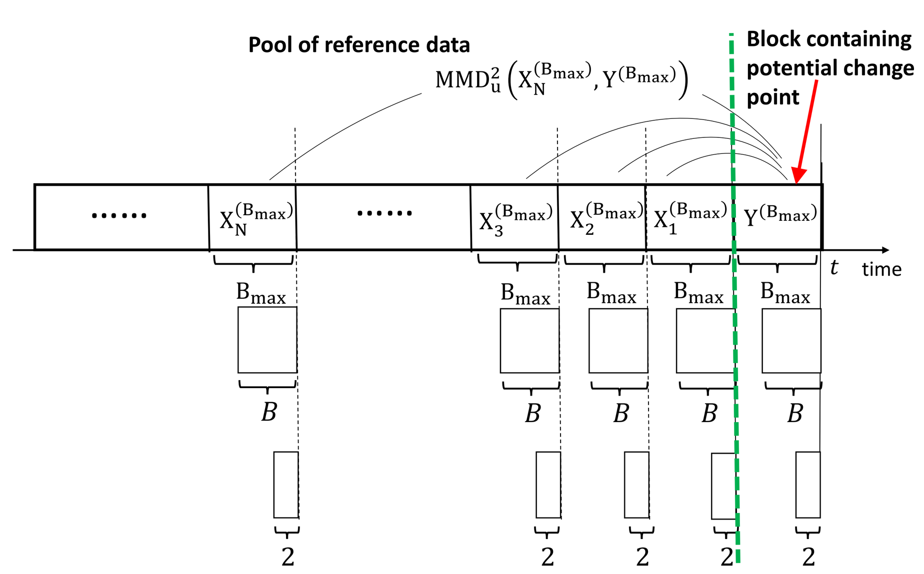
$$\mathrm{MMD}_u^2[\mathcal{F}, X, Y] = \frac{1}{n(n-1)} \sum_{i,j=1, i\neq j}^{n} h(x_i, x_j, y_i, y_j),$$

where $h(\cdot)$ is the kernel of the U-statistic defined as

$$h(x_i, x_j, y_i, y_j) = k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i).$$

- Intuitively, the empirical test statistic MMD_u^2 is expected to be small (close to zero) if $P = Q$, and large if P and Q are far apart.

Offline M-statistic



(a): offline

Figure: Illustration of (a) offline case: data are split into blocks of size $B_{\max}$, indexed backwards from time t, and we consider blocks of size B, $B = 2, \ldots, B_{\max}$.

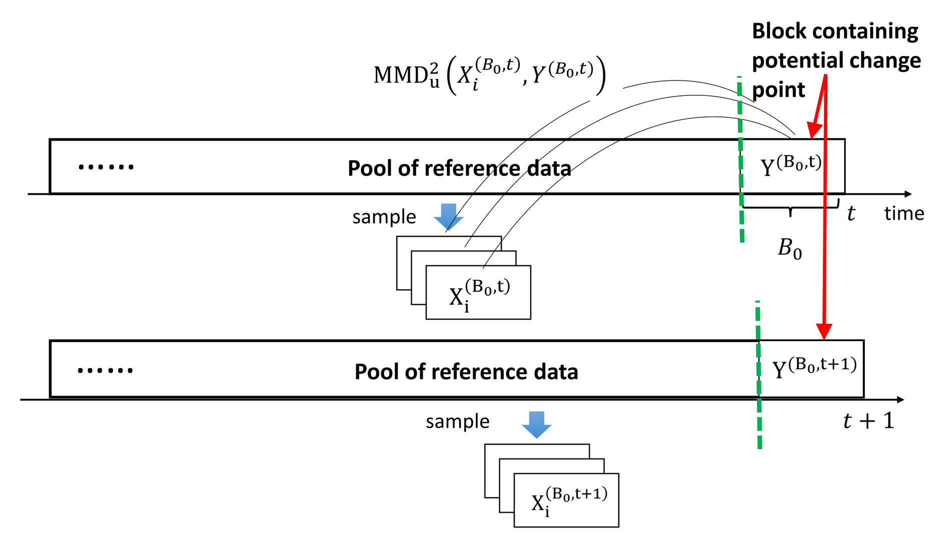
- Search for a location B $(2 \le B \le B_{\max})$ for a change-point.

$$Z_B := \frac{1}{N}\sum_{i=1}^{N} \mathrm{MMD}_u^2(X_i^{(B)}, Y^{(B)}) = \frac{1}{NB(B-1)} \sum_{i=1}^{N} \sum_{j,l=1, j\neq l}^{B} h(X_{i,j}^{(B)}, X_{i,l}^{(B)}, Y_j^{(B)}, Y_l^{(B)}).$$

- Detect a change-point whenever the M-statistic exceeds the threshold $b > 0$:

$$M := \max_{B\in\{2,3,\ldots,B_{\max}\}} Z_B/\sqrt{\mathrm{Var}[Z_B]} > b. \quad \{\text{offline change-point detection}\}$$

Online M-statistic



(b): sequential

Figure: Illustration of (b) online case. Assuming we have large amount of reference or background data that follows the null distribution.

- Fixed block size B_0.

$$Z_{B_0,t} := \frac{1}{N}\sum_{i=1}^{N} \mathrm{MMD}_u^2(X_i^{(B_0,t)}, Y^{(B_0,t)}).$$

- Detect a change-point whenever the normalized B-statistic exceeds a pre-determined threshold $b > 0$:

$$T = \inf\{t : Z_{B_0,t}/\sqrt{\mathrm{Var}[Z_{B_0}]} > b\}. \quad \{\text{online change-point detection}\}$$

- Recursive update:

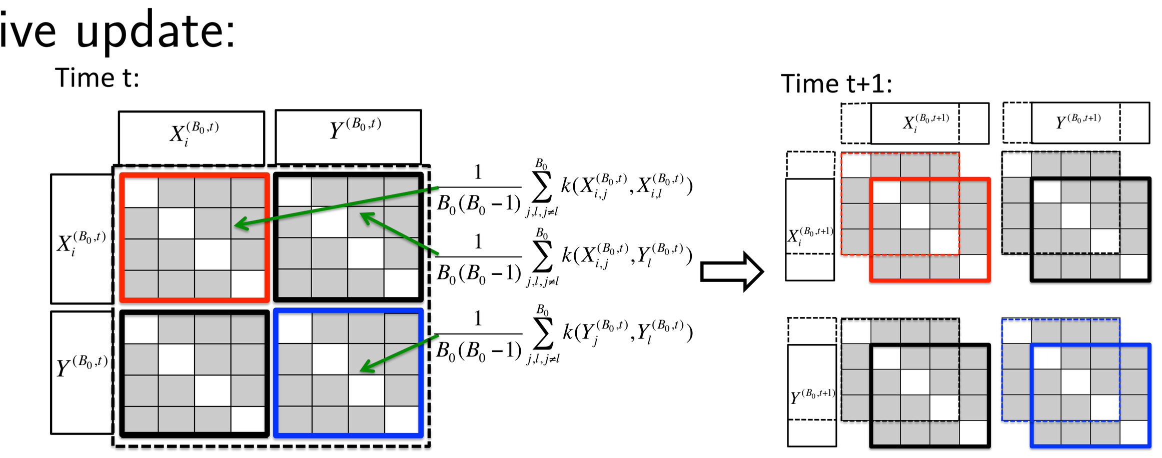
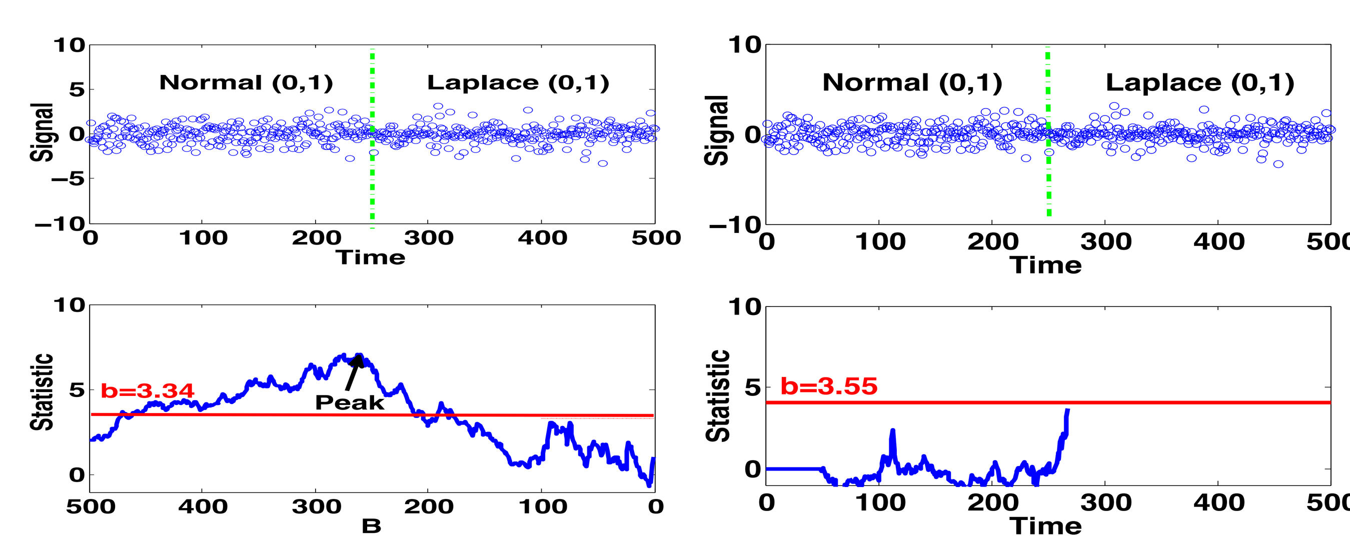


Figure: Recursively update the Gram matrix when calculating the online M-statistics.

Examples of M-Statistics



(a): Offline, $\tau = 250$ (b): Online, $\tau = 250$

Tail Probability Approximation under H_0

- **Theorem 1. Significance Level Approximation.** When $b \to \infty$ and $b/\sqrt{B_{\max}} \to c$ for some constant c, the significant level of the offline M-statistic is given by

$$\mathbb{P}^{\infty}\left\{\max_{B\in\{2,3,\ldots,B_{\max}\}} \frac{Z_B}{\sqrt{\mathrm{Var}[Z_B]}} > b\right\} = b^2 e^{-\frac{1}{2}b^2} \cdot \sum_{B=2}^{B_{\max}} \frac{(2B-1)}{2\sqrt{2\pi}B(B-1)} \nu\left(b\sqrt{\frac{2B-1}{B(B-1)}}\right) + o(1),$$

where the special function

$$\nu(u) \approx \frac{(2/u)(\Phi(u/2) - 0.5)}{(u/2)\Phi(u/2) + \phi(u/2)},$$

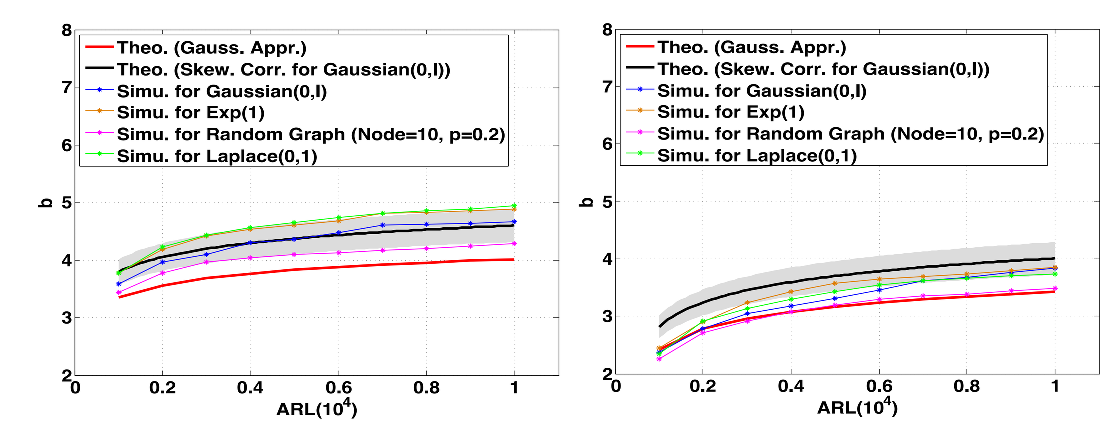
ϕ is the pdf and $\Phi(x)$ is the cdf of the standard normal distribution, respectively.

- Comparison of thresholds, determined by simulation, Theorem 1, and Skewness Correction(SC).

α	$B_{\max} = 10$			$B_{\max} = 20$			$B_{\max} = 50$		
	b (sim)	b (th3)	b (SC)	b (sim)	b (th3)	b (SC)	b (sim)	b (th3)	b (SC)
0.10	2.29	**2.40**	2.65 (0.10)	2.47	**2.60**	2.90 (0.12)	2.70	**2.80**	3.14 (0.17)
0.05	2.72	**2.72**	3.02 (0.12)	2.88	**2.90**	3.25 (0.14)	3.15	**3.08**	3.46 (0.19)
0.01	3.74	3.30	**3.71** (0.16)	3.68	3.46	**3.87** (0.16)	4.08	3.62	**4.02** (0.19)

- **Theorem 2. Average Run Length (ARL) Approximation.** When $b \to \infty$ and $b/\sqrt{B_0} \to c'$ for some constant c', the average run length (ARL) of the stopping time T is given by

$$\mathbb{E}^{\infty}[T] = \frac{e^{b^2/2}}{b^2} \cdot \left\{\frac{(2B_0-1)}{\sqrt{2\pi}B_0(B_0-1)} \cdot \nu\left(b\sqrt{\frac{2(2B_0-1)}{B_0(B_0-1)}}\right)\right\}^{-1} + o(1).$$



(a): $B_0 = 10$ (b): $B_0 = 200$

Figure: In online case, for a range of ARL values, comparison b obtained from simulation, from Theorem 2, and from skewness correction under various null distributions.

Power Analysis

Table: Power, offline, thresholds for all methods are calibrated so that $\alpha = 0.05$.

	Case 1	Case 2	Case 3	Case 4	Case 5
M-statistic	**0.71**	**1.00**	**0.26**	**1.00**	**0.44**
Hotelling's T^2	0.18	0.88	0.07	0.87	0.03
GLR	0.03	0.05	0.07	0.12	0.04

Note: Case 1 to 4, change from $\mathcal{N}(0, I)$ to $\mathcal{N}(0.1, I)$, $\mathcal{N}(0.2, I)$, $\mathcal{N}(0, \Sigma), \mathcal{N}(0.2, \Sigma), respectively.$
Case 5, change from $\mathcal{N}(0, 1)$ to Laplace(0,1).

[1] A Kernel Two-Sample Test. *(Gretton,A., et.al. JMLR 2012)*