Message-passing sequential detection of multiple change points in networks

Long Nguyen, Arash Amini — *University of Michigan*
Ram Rajagopal — *Stanford University*

ISIT, Boston, July 2012

Sequential change-point detection

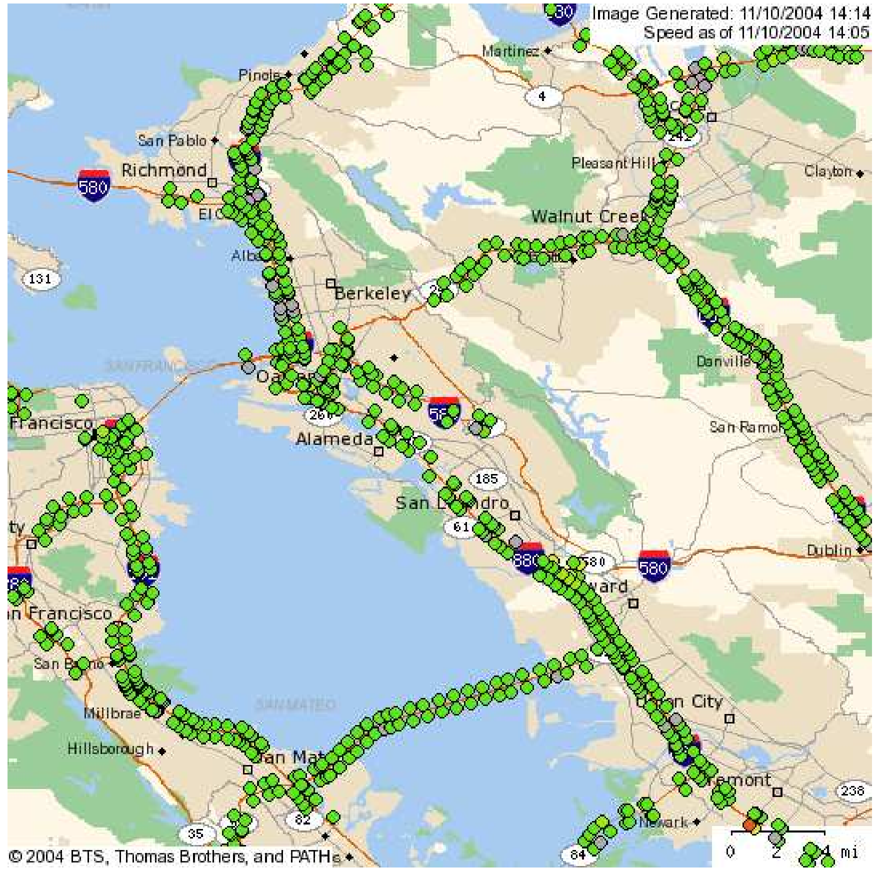
- Quickest detection of change in distribution of a sequence of data
  - data collected sequentially over time
  - tradeoff between false alarm rate and detection delay time
  - extensions to decentralized network with a fusion center
  - classical setting involves only one single change point variable

Sequential change-point detection

- Quickest detection of change in distribution of a sequence of data
  - data collected sequentially over time
  - tradeoff between false alarm rate and detection delay time
  - extensions to decentralized network with a fusion center
  - classical setting involves only one single change point variable

- We study problems requiring detection of multiple change points in multiple sequences across network sites
  - multiple change points are statistically dependent
  - need to borrow information across network sites
  - no fusion center – needs message-passing type algorithm
  - new elements of modeling and asymptotic theory

Example – Simultaneous traffic monitoring



Problem: detecting in real-time potential hotspots in a traffic network

- data are sequences of measurements of traffic volume at multiple sites
- sequential change point detection for each site

Sequential detection for single change point

- network site j collects sequence of data X_j^n for $n = 1, 2, \ldots$
- time $\lambda_j \in \mathbb{N}$ is change point variable for site j
- data are i.i.d. according to density g before the change point; and i.i.d. according to f after

Sequential detection for single change point

- network site j collects sequence of data X_j^n for $n = 1, 2, \ldots$
- time $\lambda_j \in \mathbb{N}$ is change point variable for site j
- data are i.i.d. according to density g before the change point; and i.i.d. according to f after
- a sequential change point detection procedure is a stopping time τ_j, i.e., $\{\tau_j \leq n\} \sim \sigma(X_j^{[n]})$

Sequential detection for single change point

- network site j collects sequence of data X_j^n for $n = 1, 2, \ldots$
- time $\lambda_j \in \mathbb{N}$ is change point variable for site j
- data are i.i.d. according to density g before the change point; and i.i.d. according to f after
- a sequential change point detection procedure is a stopping time τ_j, i.e., $\{\tau_j \leq n\} \sim \sigma(X_j^{[n]})$
- Neyman-Pearson criterion:
  - constraint on false alarm error

$$PFA(\tau_j) = P(\tau_j < \lambda_j) \leq \alpha \text{ for some small } \alpha$$

  - minimum detection delay

$$\mathbb{E}[(\tau_j - \lambda_j) | \tau_j \geq \lambda_j].$$

Optimal rule for single change point detection

- taking a Bayesian approach, λ_j is endowed with a prior
- under some conditions, optimal sequential rule obtained by thresholding the posterior of λ_j: (Shiryaev, 1978)

$$\boxed{\tau_j = \inf\{n : \Lambda_n \geq 1 - \alpha\},}$$

where

$$\Lambda_n = \mathbb{P}(\lambda_j \leq n | X_j^{[n]}).$$

- well-established asymptotic properties (Tartakovsky & Veeravalli, 2006):
  - false alarm:
  $$PFA(\tau_j) \leq \alpha.$$
  - detection delay:
  $$D(\tau_j) = \frac{|\log \alpha|}{I_j + d}\Big(1 + o(1)\Big) \text{ as } \alpha \to 0.$$
  - here $I_j = KL(f_j || g_j)$, the Kullback-Leibler information, constant d depends on the prior

Extensions to network setting.

- survey paper by Tsitsiklis (1993)
- decentralized sequential detection: Veeravalli, Basar and Poor (1993), Mei (2008), Nguyen, Wainwright and Jordan (2008)
- sequential change diagnosis: Dayanik, Goulding and Poor (2008)
- multiple sequence change point detection: Xie and Siegmund (2010)
- sequential detection of a markov process: Raghavan and Veeravalli (2010)
- ...

Talk outline

- statistical formulation for sequential detection of *multiple* change points in a network setting
  - probabilistic graphical models
  - extension of sequential analysis to multiple change point variables

- sequential and "real-time" message-passing detection algorithms
  - decision procedures with limited data and computation

- asymptotic theory characterizing detection delay and algorithm convergence
  - roles of graphical models in asymptotic analysis

Graphical models for multiple change points

- m network sites labeled by $U = \{1, \ldots, m\}$
- given a graph $G = (U, E)$ that specifies the the connections among $u \in U$
- each site j experiences a change at time $\lambda_j \in \mathbb{N}$
  - λ_j is endowed with (independent) prior distribution π_j

Graphical models for multiple change points

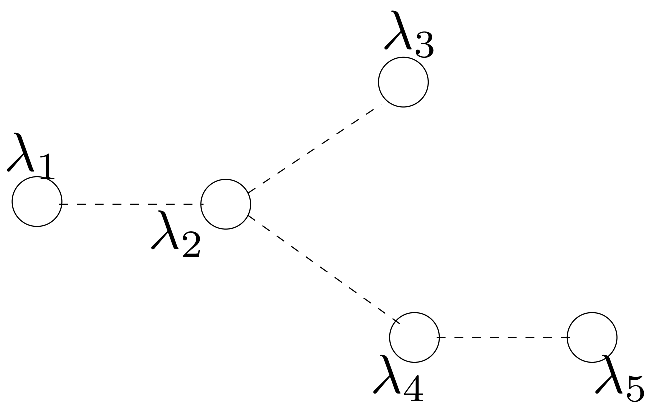
- m network sites labeled by $U = \{1, \ldots, m\}$
- given a graph $G = (U, E)$ that specifies the the connections among $u \in U$
- each site j experiences a change at time $\lambda_j \in \mathbb{N}$
  - λ_j is endowed with (independent) prior distribution π_j
- there may be private data sequence $(X_j^n)_{n \geq 1}$ for site j
  - private data sequence changes its distribution after λ_j

Graphical models for multiple change points

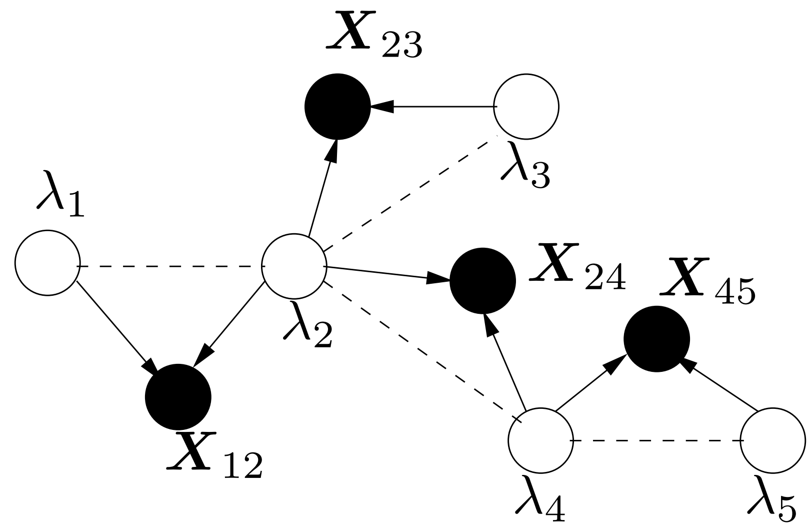
- m network sites labeled by $U = \{1, \ldots, m\}$
- given a graph $G = (U, E)$ that specifies the the connections among $u \in U$
- each site j experiences a change at time $\lambda_j \in \mathbb{N}$
  - λ_j is endowed with (independent) prior distribution π_j
- there may be private data sequence $(X_j^n)_{n \geq 1}$ for site j
  - private data sequence changes its distribution after λ_j
- there is shared data sequence $(X_{ij}^n)_{n \geq 1}$ for each edge $e = (i, j)$ connecting *neighboring pair* of sites j and i:

$$
\begin{aligned}
X_{ij}^n &\stackrel{iid}{\sim} g_{ij}(\cdot), \quad \text{for } n < \lambda_{ij} := \min(\lambda_i, \lambda_j) \\
&\stackrel{iid}{\sim} f_{ij}(\cdot), \quad \text{for } n \geq \lambda_{ij} = \min(\lambda_i, \lambda_j)
\end{aligned}
$$

Graphical model of change points and data sequences



(a) Topology of sensor network



(b) Graphical model of random variables

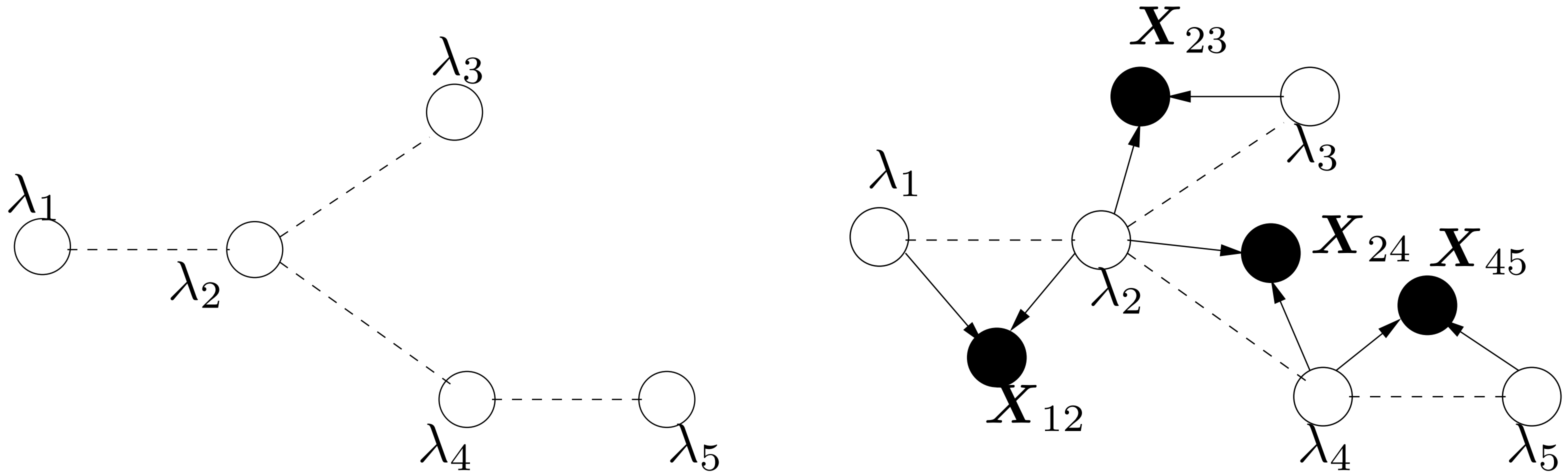
Joint distribution of change points and observed network data at time n:

$$\boxed{P(\lambda_*, \boldsymbol{X}_*^n) = \prod_{j \in V} \pi_j(\lambda_j) \prod_{j \in V} P(\boldsymbol{X}_j^n | \lambda_j) \prod_{(ij) \in E} P(\boldsymbol{X}_{ij}^n | \lambda_i \wedge \lambda_j)}$$

Star notations: $\lambda_* := (\lambda_1, \ldots, \lambda_m)$, $\boldsymbol{X}_*^n = (X_1^n, \ldots, X_m^n)$.

- Change point variables are statistically dependent a posteriori!

Min-functional of change points



Let S be a subset of network sites. Define the earliest change point among any sites in S:

$$\lambda_S := \min_{u \in S} \lambda_j.$$

Question: what is the optimal stopping rule τ_S for estimating λ_S?

$$\tau_S \sim \sigma(X_*^{[n]}).$$

A natural rule is by thresholding the posterior probability:

$$\boxed{\tau_S = \inf\{n : \mathbb{P}(\min_{u \in S} \lambda_j \leq n | X_*^{[n]}) \geq 1 - \alpha\},}$$

for small $\alpha > 0$.

A natural rule is by thresholding the posterior probability:

$$\boxed{\tau_S = \inf\{n : \mathbb{P}(\min_{u \in S} \lambda_j \leq n | X_*^{[n]}) \geq 1 - \alpha\},}$$

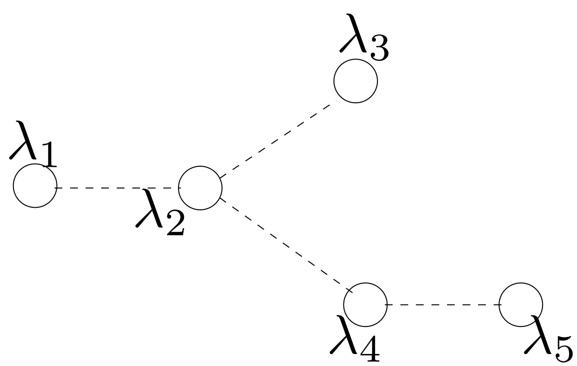
for small $\alpha > 0$.

This rule is sub-optimal (unlike the single change point case, which is optimal under some conditions on the prior).

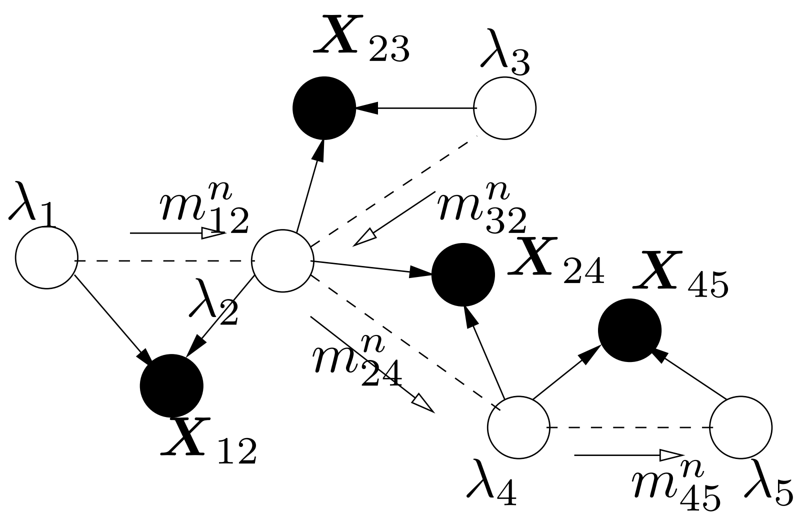
But it will be shown to be asymptotically optimal and computationally tractable.

Message-passing distributed computation via sum-product algorithm:
the issue to compute posterior probabilities, assuming that data and statistical messages can be only be passed through the graphical structure:

$$P(\lambda_S \leq n | \boldsymbol{X}_*^{[n]}) \geq 1 - \alpha\}.$$



(a) Topology of sensor network



(b) Message-passing in network

Simple to implement via an adaptation of the sum-product algorithm

Computational complexity. When G is a tree, the computational complexity of the message passing algorithm at each time step n is $O((|V| + |E|)n)$, but linearity in n is not desirable.

Mean-field approximation.

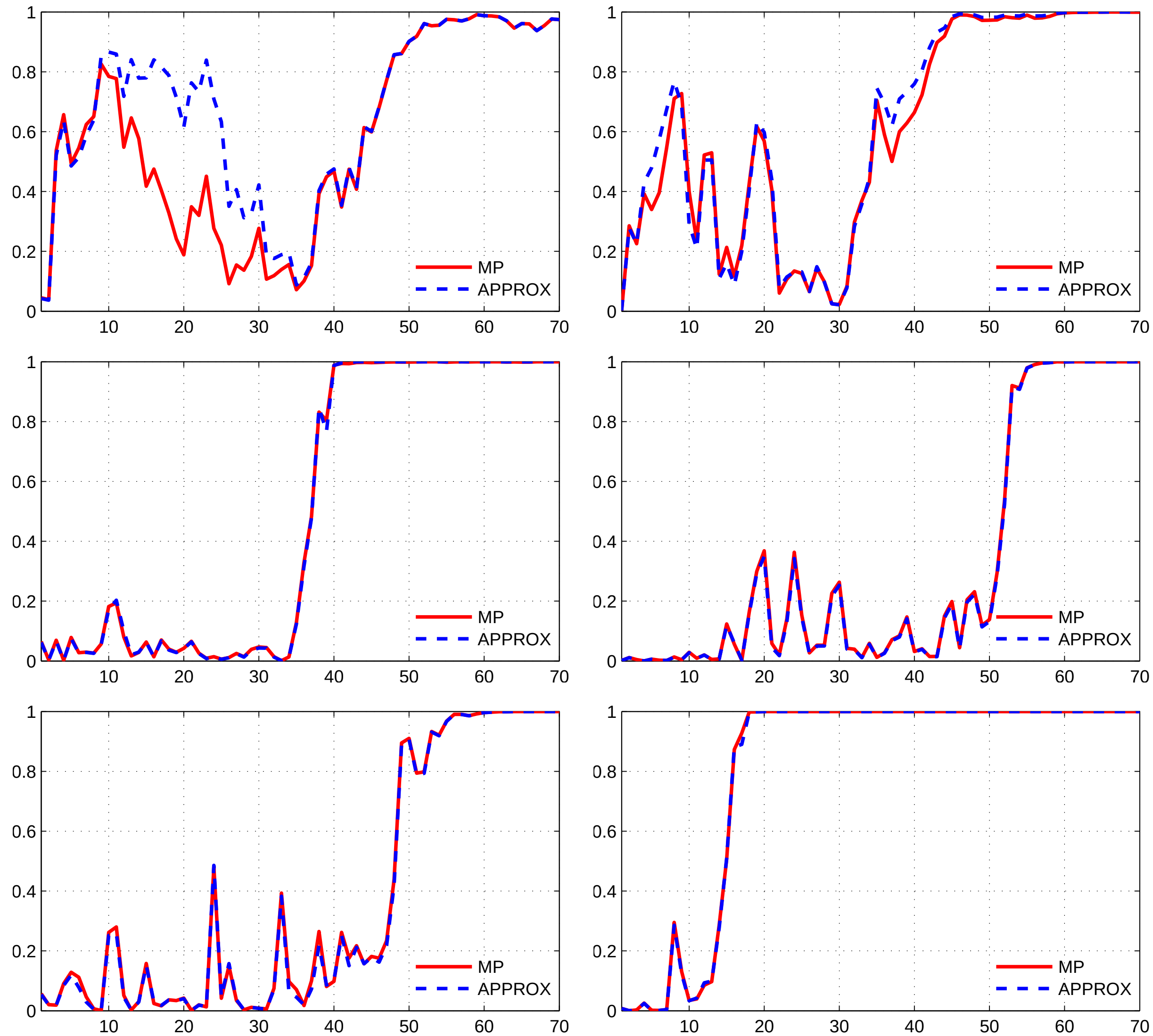
- Define latent binary variables $Z_j^n = \mathbb{I}(\lambda_j \le n)$.
- Compute $P(Z_*^n | \boldsymbol{X}_*^{[n]})$ in terms of $P(Z_*^n | \boldsymbol{X}_*^{[n-1]})$ by Bayes rule.
- Decoupling approximation: As n gets large, due to concentration, the variables Z_j^n become decoupled across the graph. So, approximate:

$$\boxed{\tilde{P}(Z_*^n | \boldsymbol{X}_*^{[n-1]}) \approx \prod_{j \in V} P(Z_j^n | \boldsymbol{X}_*^{[n-1]})}$$

- In effect, we have avoided marginalization over time at every time step, resulting in O(1) computational complexity in n.

Theorem 1. Both exact message-passing algorithm and mean-field approximation algorithm construct a Markov sequence of posterior probabilities that obey a contraction map. This entails that both sequences converge to 1 almost surely.

Approximation of posterior paths, $n \mapsto P(\lambda_j \le n | X_*^{[n]})$.



Main Theorem (optimal delay theorem).

Assume that

(a) The change points λ_j are endowed with independent geometric priors.

(b) The likelihood ratio functions are bounded from above.

Then the proposed stopping rule τ_S satisfies:

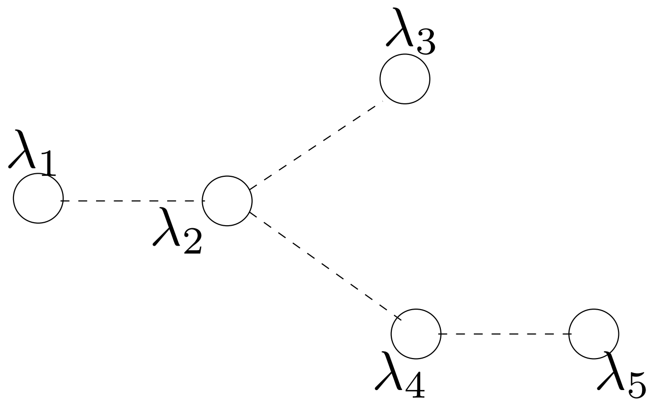
(i) False alarm rate: $\mathbb{P}(\tau_S \leq \lambda_S) \leq \alpha$.

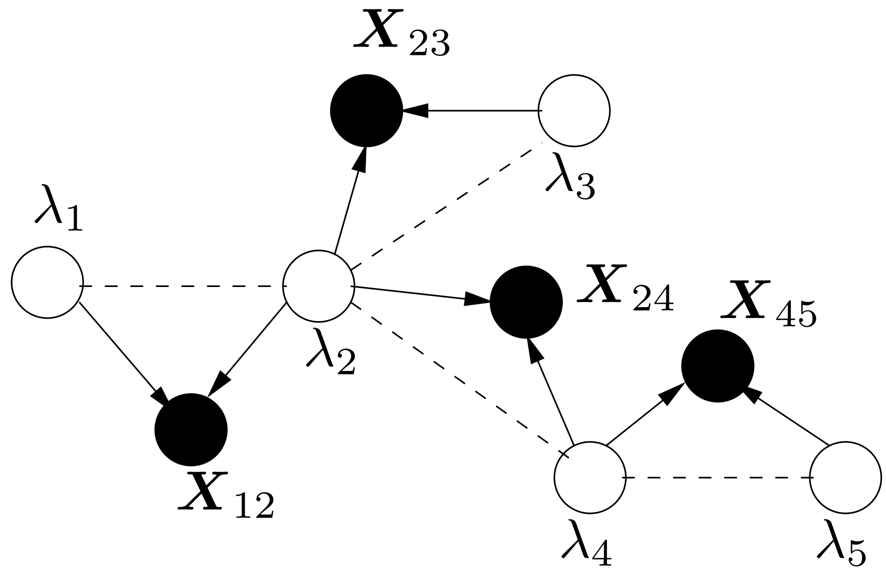
(ii) The expected delay is asymptotically optimal, and takes the form:

$$\boxed{\mathbb{E}[(\tau_S - \min_{u \in S} \lambda_j) | \tau_S \geq \min_{u \in S} \lambda_j] = \frac{|\log \alpha|(1 + o(1))}{d + \underbrace{\sum_{j \in S} I_j + \sum_{(ij) \in E \cap S} I_{ij}}_{I_{\lambda_S}}}.}$$

Here, $I_j = \int f_i \log(f_j/g_j)$, and $I_{ij} = \int f_{ij} \log(f_{ij}/g_{ij})$.

Graph-based Kullback-Leibler information ...





If $S = \{1\}$, then $I_{\lambda_S} = I_1$

If $S = \{1, 2\}$, then $I_{\lambda_S} = I_1 + I_2 + I_{12}$

If $S = \{1, 2, 3\}$, then $I_{\lambda_S} = I_1 + I_2 + I_3 + I_{12} + I_{23}$.

Concentration inequalities for marginal LRs

For $\phi = \min_{u \in S} \lambda_j$, define marginal likelihood ratio

$$\boxed{D_\phi^{k,n} := D_\phi^k(\mathbf{X}_*^n) := \frac{\mathbb{P}_\phi^k(\mathbf{X}_*^n)}{\mathbb{P}_\phi^\infty(\mathbf{X}_*^n)},}$$

where $\mathbb{P}_\phi^k$ denotes $\mathbb{P}(\cdot|\phi = k)$.

Define conditional prior probability $\pi_\phi^k(m_*) := \mathbb{P}(\lambda_* = m_* \mid \phi = k)$.

By a general result of Tartakovski & Veeravalli (2006), if

$$\mathbb{P}_\phi^k\Big[\frac{1}{N}\max_{1 \le n \le N} \log D_\phi^k(\mathbf{X}_*^{k+n}) \ge (1+\varepsilon) I_\phi\Big] \xrightarrow{N\to\infty} 0 \tag{1}$$

for all (small) $\varepsilon > 0$ and all $k \in \mathbb{N}$, then the "lower bound" follows, $\inf_{\widetilde{\tau} \in \Delta_\phi(\alpha)} \mathbb{E}\big[\widetilde{\tau} - \phi \mid \widetilde{\tau} \ge \phi\big] \ge \frac{|\log \alpha|}{q_\phi + I_\phi}\big(1 + o(1)\big)$.

Furthermore, let

$$T_\varepsilon^k := \sup\Big\{n \in \mathbb{N} : \ \frac{1}{n}\log D_\phi^k(\mathbf{X}_*^{k+n-1}) < I_\phi - \varepsilon\Big\}.$$

By Tartakovski-Veeravalli (2006), if one has

$$\mathbb{E}\, T_\varepsilon^\phi := \sum_{k=1}^{\infty} \mathbb{P}(\phi = k)\, \mathbb{E}_\phi^k(T_\varepsilon^k) < \infty, \tag{2}$$

for all (small) $\varepsilon > 0$, then the "upper bound" follows, that is, $\mathbb{E}[\tau_{\mathcal{S}} - \phi \mid \tau_{\mathcal{S}} \geq \phi] \leq \frac{|\log \alpha|}{q_\phi + I_\phi}(1 + o(1))$.

Both conditions (1) and (2) can be deduced from an elaborate form of concentration inequality for the marginal likelihood ratio.

Key concentration lemma. Denote by $\mathbb{P}_{\lambda_*}^{m_*}$ the conditional probability $\mathbb{P}(\cdot|\lambda_* = m_*)$. Assume that for all $m_* \in \mathbb{N}^d$ in the support of $\pi_\phi^k(\cdot)$,

$$\boxed{\mathbb{P}_{\lambda_*}^{m_*}\left\{\left|\frac{1}{n}\log D_\phi^k(\mathbf{X}_*^n) - I_\phi\right| > \varepsilon\right\} \leq q(n)\exp(-c_1 n\varepsilon^2)} \tag{3}$$

for all $n \in \mathbb{N}$ and $\varepsilon \in (0, \varepsilon_0)$ such that $\boxed{n \geq \frac{1}{\varepsilon^2}p^2(m_*, k)}$, where

- $p(\cdot)$ and $q(\cdot)$ are *polynomials with nonnegative coefficients*,
- both $\mathbb{P}(\phi = \cdot)$ and $\mathbb{P}(\lambda_j|\phi = k)$ have *finite polynomial moments*.

Then the optimal delay Theorem holds.

Probabilistic calculus of ϵ-equivalence

Definition. Consider two sequences $\{a_n\}$ and $\{b_n\}$ of random variables, where $a_n = a_n(k)$ and $b_n = b_n(k)$ could depend on a common parameter $k \in \mathbb{N}$. The two sequences are called "asymptotically ε-equivalent" as $n \to \infty$, under $\{\mathbb{P}_{\lambda_*}^{m_*} : m_* \in \operatorname{supp}(\pi_\phi^k)\}$, and denoted

$$a_n \overset{\varepsilon}{\asymp} b_n,$$

if there exist polynomials $p(\cdot)$ and $q(\cdot)$ (with constant nonnegative coefficients), and $\varepsilon_0 > 0$, such that for all $m_* \in \operatorname{supp}(\pi_\phi^k)$, we have

$$\mathbb{P}_{\lambda_*}^{m_*}(|a_n - b_n| \leq \varepsilon) \geq 1 - q(n)e^{-c_1 n \varepsilon^2}$$

for all $n \in \mathbb{N}$ and $\varepsilon \in (0, \varepsilon_0)$ satisfying $\sqrt{n}\varepsilon \geq p(m_*, k)$.

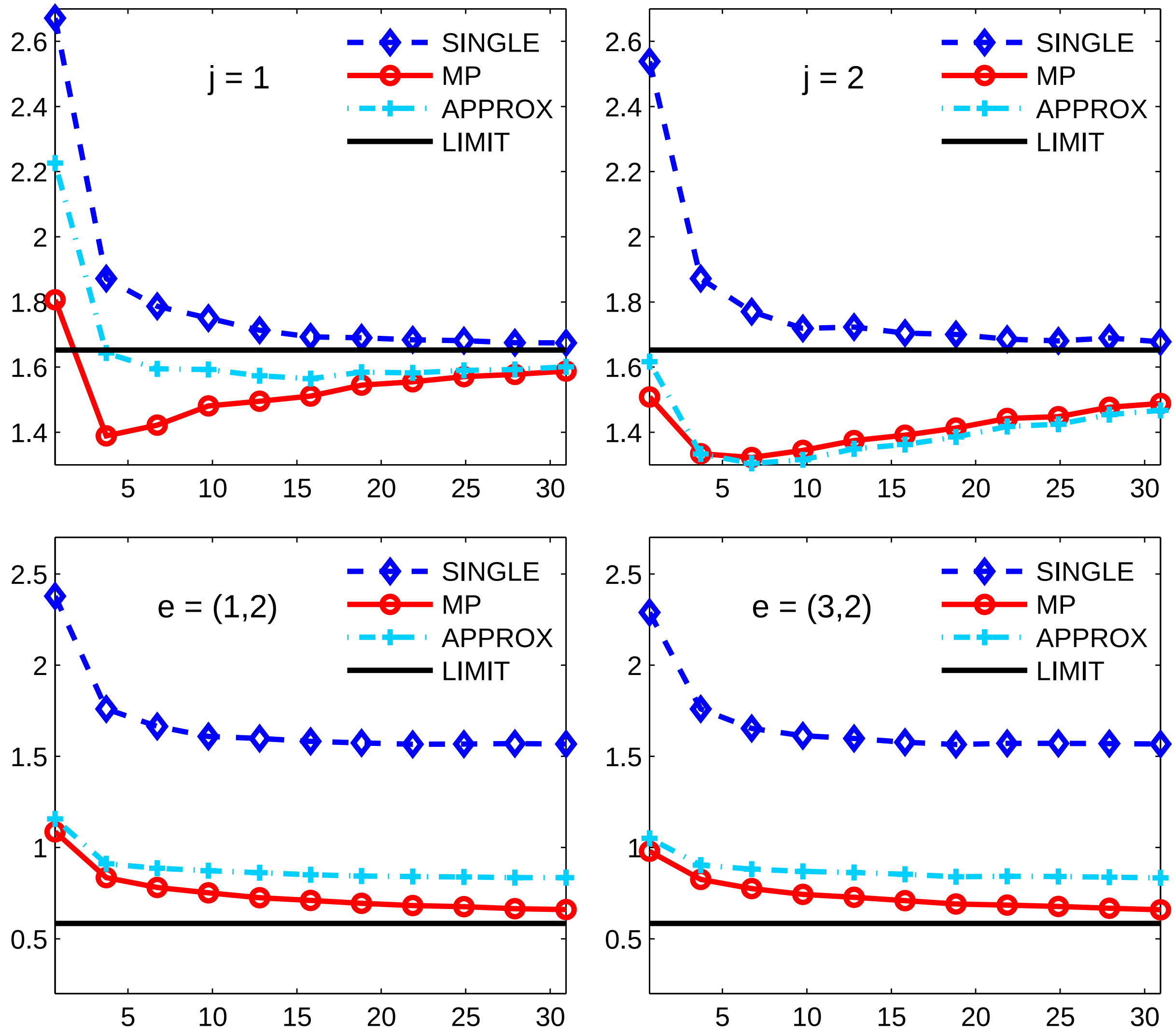
By union bound and algebraic manipulation, we obtain the following rules:

1. $a_n \overset{\varepsilon}{\asymp} b_n$ implies $a_n \overset{C\varepsilon}{\asymp} b_n$ for $C > 0$ and $\alpha a_n \overset{\varepsilon}{\asymp} \alpha b_n$ for $\alpha \in \mathbb{R}$.
2. $a_n \overset{\varepsilon}{\asymp} b_n$ and $b_n \overset{\varepsilon}{\asymp} c_n$ implies $a_n \overset{\varepsilon}{\asymp} c_n$. (Transitivity)
3. $a_n \overset{\varepsilon}{\asymp} b_n$ and $c_n \overset{\varepsilon}{\asymp} d_n$ implies $a_n \pm c_n \overset{\varepsilon}{\asymp} b_n \pm d_n$.
4. $a_n \overset{\varepsilon}{\asymp} b_n$ implies $\max\{a_n, c_n\} \overset{\varepsilon}{\asymp} \max\{b_n, c_n\}$.
5. $a_n \overset{\varepsilon}{\asymp} b_n$, $c_n \overset{\varepsilon}{\asymp} 1$ and $\{b_n\}$ bounded implies $a_n|c_n| \overset{\varepsilon}{\asymp} b_n$.
6. $a_n \overset{\varepsilon}{\asymp} a > 0$ and $b_n \overset{\varepsilon}{\preccurlyeq} -b < 0$ implies $\max\{a_n, b_n\} \overset{\varepsilon}{\asymp} a$.
7. "log–sum-max" inequality for positive sequences $\{a_n\}$ and $\{b_n\}$:

$$n^{-1} \log(a_n + b_n) \overset{\varepsilon}{\asymp} \max\{n^{-1} \log a_n, n^{-1} \log b_n\}.$$

Based on this calculus we can deduce the ϵ-equivalence of the marginal likelihood ratio from the ϵ-equivalence of the likelihood ratios defined on individual sites and edges of neighboring sites.

Plots of the slope $\frac{1}{|\log \alpha|}\mathbb{E}[\tau_S - \phi_S | \tau_S \geq \phi_S]$ for star network of (1,2,3,4) centering at 2



Summary

- decentralized sequential detection of multiple change points
  - model, algorithm and asymptotic theory needed to go beyond single change point setting

- new statistical formulation drawing from:
  - classical sequential analysis
  - probabilistic graphical models (Bayes nets)

- introduced a "message-passing" sequential detection algorithm, exploiting the benefit of "network information"

- asymptotic theory for analyzing false alarm rates and detection delay

- for more detail, see
  - A. Amini and X. Nguyen. *Sequential detection of multiple change points: A graphical models approach*. Technical report, Department of Statistics, Univ of Michigan, 2012.
  - See also: R. Rajagopal, X. Nguyen, S.C. Ergen and P. Varaiya. *Simultaneous sequential detection of multiple interacting faults*. http://arxiv.org/abs/1012.1258