Detection of an anomalous path in a noisy network

Shirshendu Chatterjee,
City College & Graduate Center, CUNY

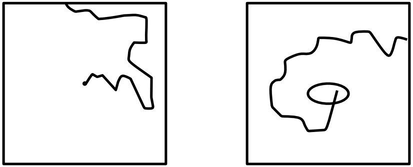
Joint Work with **Ofer Zeitouni**

CUNY Probability Seminar, Feb 2018

Background

- The topic fits into broad framework of nonparametric detection.
- Based on measurements about variables related through a graphical model, detect whether there is a sequence of connected nodes which exhibit a "peculiar behavior".
- Consider a graph-indexed process: a model problem for detecting whether or not there is a chain of connected nodes in a given network which exhibit an "unusual behavior".
  - Trace the existence of a polluter in a network of streams
  - Detecting atypical gene behavior in a given gene network
- Based on one realization of this process,
  - can one reliably detect if there is a chain of nodes (hidden in the background noise) that stand out?
  - How subtle a difference can one detect?

Our problem set up



Suppose G_n is a $n \times n$ two dimensional graph with node set V_n and path set P_n. Paths in P_n are nonintersecting and each has order n many nodes. Suppose each node $\mathbf{v}$ has a r.v. $X_{\mathbf{v}}$ attached to it.

Observable: $(X_{\mathbf{v}}, \mathbf{v} \in V_n)$.

- **Null hypothesis H_0:** The random variables $\{X_{\mathbf{v}} : \mathbf{v} \in V_n\}$ are *i.i.d.* with common distribution $N(0,1)$.
- **Alternate (signal) hypothesis $H_{1,n}$:** it is a composite hypothesis $\cup_{\pi \in P(G_n)} H_{1,\pi}$, where, under $H_{1,\pi}$, the random variables $\{X_{\mathbf{v}} : \mathbf{v} \in V_n\}$ are independent with

$$X_{\mathbf{v}} \stackrel{d}{=} \begin{cases} N(\mu_n, 1) & \text{if } \mathbf{v} \in \pi \\ N(0,1) & \text{otherwise} \end{cases} \quad \text{for some } \mu_n > 0.$$

Main Question

Let $\mathbb{P}_0$ and $\mathbb{P}_\pi$ be the probability distribution of $(X_\mathbf{v}, \mathbf{v} \in V_n)$ under the hypothesis H_0 and $H_{1,\pi}$.

Question. For which values of μ_n, the probability distributions $\mathbb{P}_0$ and $\cup_{\pi \in P_n} \mathbb{P}_\pi$ are distinguishable?

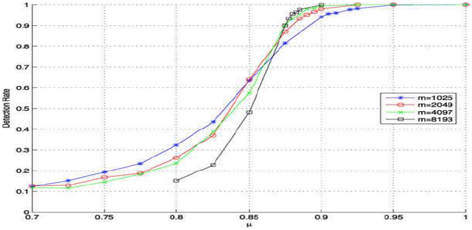
More precisely, consider the following.

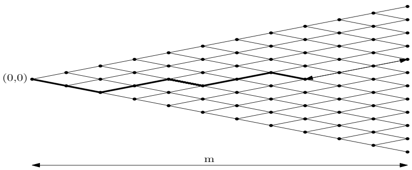
- A test T_n is a $\{0,1\}$-valued function of $(X_\mathbf{v}, \mathbf{v} \in V_n)$
- $\{T_n = 1\} \leftrightarrow$ "Accept signal hypothesis", and $\{T_n = 0\} \leftrightarrow$ "Accept null hypothesis"
- Minimax risk: For any test T_n, $\gamma(T_n) := \mathbb{P}(TypeI) + \sup_{\pi \in P_n} \mathbb{P}(TypeII) = \mathbb{P}_0(T_n = 1) + \sup_{\pi \in P_n} \mathbb{P}_\pi(T_n = 0)$.
- A test T_n is
    - asymptotically powerful if $\gamma(T_n) \to 0$.
    - asymptotically powerless if $\gamma(T_n)$ is close to 1.

Question. What values of μ_n (signal per anomalous node) can be detected reliably?

GLRT

The GLRT would reject H_0 for large values of $M_n := \max_{\pi \in P_n} \sum_{\mathbf{v} \in \pi} X_{\mathbf{v}}$. Its performance is not optimal.





Consider the two dimensional graph with vertex set

$$V_n := \{(i,j) : 0 \leq i \leq n-1, |j| \leq i, \text{ and } i,j \text{ have same parity}\},$$

and path set P_n consisting of directed paths starting at the origin.

Theorem (Castro, Candes, Helgason, Zeitouni (2008))

If $\mu_n \sqrt{\log(n)} \to \infty$, then there is an asymptotically powerful test. If $\mu_n \log(n)\sqrt{\log\log(n)} \to 0$, then all tests are asymptotically powerless.

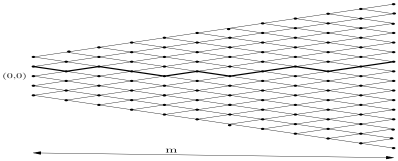
The statistic separating the two hypothesis

- Consider a linear statistic $\sum_{(i,j)\in V_n} w_{i,j} X_{(i,j)}$, where $(w_{i,j})_{(i,j)\in V_n}$ are weights.
- Choose the weights which maximizes the signal-noise ratio.
- Optimal weight $w_{i,j}$ turns out to be proportional to $1/(i+1)$.
- So the (weighted average) linear statistic separating the two hypothesis is

$$WAS := \sum_{(i,j)\in V_n} \frac{X_{(i,j)}}{i+1},$$

- $\mathbb{E}_0(WAS) = 0$, $\mathbb{E}_\pi(WAS) = \mu_n \log(n)$, $Var_0(WAS) = Var_\pi(WAS) = \log(n)$, so the signal-noise ratio is $\mu_n \sqrt{\log(n)}$.
- So if $\mu_n \sqrt{\log(n)} \to \infty$, then WAS separates the two hypothesis.

Unknown initial location and directed paths



Consider the two dimensional graph with vertex set

$$V_n := \{(i,j) : 0 \le i \le n-1, |j| \le an+i, \text{ and } i,j \text{ have same parity}\},$$

and path set P_n consisting of directed paths starting at the left hyperplane.

Theorem (C. and Zeitouni (2017))

If $\mu_n\sqrt{\log(n)} \ge C$ for some large constant C, then there is an asymptotically powerful test.
If $\mu_n \log(n)\sqrt{\log\log(n)} \to 0$, then all tests are asymptotically powerless.

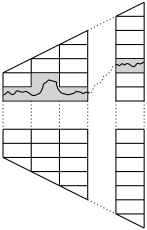
The statistic separating the two hypothesis

- The *WAS* loses its power in this case.
- We develop a polynomial statistic (PS), which is a polynomial in $(X_{(i,j)}, (i,j) \in V_n)$.
- The *PS* is obtained inductively using a quadratic statistic as the building black.
- Call $(i,j) \leftrightsquigarrow (i',j')$ if a path hitting (i,j) can visit (i',j') and vice versa.
- Define the quadratic forms

$$Q_n := \sum_{(i,j),(i',j') \in V_n, (i,j) \leftrightsquigarrow (i',j')} \frac{1}{|i-i'|} X_{(i,j)} X_{(i',j')}.$$

- $\mathbb{E}_0(Q_n) = 0$, $\mathbb{E}_\pi(Q_n) = \mu_n^2 n \log(n)$, $Var_0(Q_n) = n^2 \log(n)$ and $Var_\pi(Q_n) = (1+o(1))n^2 \log(n)$, so the signal-noise ratio is $\mu_n^2 \sqrt{\log(n)}$.
- So if $\mu_n \gg (\log(n))^{-1/4} \to \infty$, then Q_n separates the two hypothesis.

Renormalization Argument



We partition the graph into disjoint squares and half-squares having side length $\sqrt{n}$, and consider

- the coarse grained graph, where each square (and half-square) represents a node,
- the coarse grained path on the above graph.

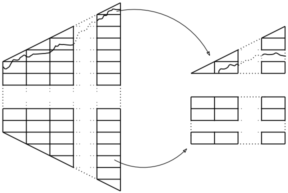
The random variable attached to a coarse grained node is the normalized quadratic form $Q_{\sqrt{n}}$ computed on the associated square (or half-square).

Renormalization Argument (continued)

Thus, we get a renormalized version of the original problem, where n is replaced by $\sqrt{n}$ and μ_n is replaced by $\mu_n^2\sqrt{\log(\sqrt{n})} = c\mu_n^2\sqrt{\log(n)}$ (the signal noise ratio for $Q_{\sqrt{n}}$).

If we compute a similar quadratic form of the quadratic forms corresponding to the squares (and thus use a polynomial of degree 4), then the new signal-noise ration becomes

$$\asymp \left(\mu_n^2\sqrt{\log(n)}\right)^2\sqrt{\log(n)} = \mu_n^4\log^{3/2}(n) \to \infty \text{ if } \mu_n \gg (\log(n))^{-3/8}.$$



Repeating the renormalization argument 'few' times, we get the upper bound for detection threshold to be $C/\sqrt{\log(n)}$.

Lower bound for detection threshold

- Suppose that the anomalous path π follow a prior distribution Π on the set of paths.
- Let L_n be the likelihood ratio: $L_n(\mathbf{X}) := \frac{d\mathbb{P}_\Pi}{d\mathbb{P}_0}(\mathbf{X})$. Consider the test $\mathbf{1}_{\{L_n \geq 1\}}$.
- It is well known that

$$\inf_{\text{all tests } T_n} \gamma(T_n) = \gamma(\mathbf{1}_{\{L_n \geq 1\}}) = \mathbb{P}_0(L_n \geq 1) + \mathbb{P}_\Pi(L_n < 1) =: \gamma_n^*.$$

- Clearly $\mathbb{E}_0(L_n) = 1$. A standard calculation shows

$$\gamma_n^* = 1 - \frac{1}{2}\mathbb{E}_0|L_n - 1|.$$

- Need to find condition on μ_n such that $\mathbb{E}_0|L_n - 1| \to 0$.
- Instead, we use Caichy-Schwartz inequality to have

$$\gamma_n^* \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}_0[(L_n - 1)^2]}$$

and find conditions on μ_n so that $\mathbb{E}_0[(L_n - 1)^2] \to 0$.

Connection with number of intersection of two walks

- Recall that $L_n = \frac{d\mathbb{P}_\Pi}{d\mathbb{P}_0}(\mathbf{X}) = \mathbb{E}_\Pi \exp(\sum_{\mathbf{v}\in\Pi} \mu_n X_{\mathbf{v}} - \frac{1}{2}\mu_n^2)$.
- $\mathbb{E}_0[(L_n - 1)^2] = \mathbb{E}_0(L_m^2) - 1$ and

$$\mathbb{E}_0(L_n^2) = \mathbb{E}_0\mathbb{E}_{\Pi_1\times\Pi_2} \exp\left(\sum_{\mathbf{v}\in\Pi_1} \mu_n X_{\mathbf{v}} - \frac{1}{2}\mu_n^2 + \sum_{\mathbf{v}\in\Pi_2} \mu_n X_{\mathbf{v}} - \frac{1}{2}\mu_n^2\right)$$

$$= \mathbb{E}_{\Pi_1\times\Pi_2}\left[\prod_{\mathbf{v}\in\Pi_1\cap\Pi_2} \mathbb{E}_0 e^{2\mu_n X_{\mathbf{v}} - \mu_n^2} \prod_{\mathbf{v}\in\Pi_1\triangle\Pi_2} \mathbb{E}_0 e^{\mu_n X_{\mathbf{v}} - \frac{1}{2}\mu_n^2}\right]$$

$$= \mathbb{E}_{\Pi_1\times\Pi_2} e^{\mu_n^2 N_n},$$

where N_n is the number of intersections between two independent walks having prior Π.

Predictability Profile

- The strategy is to construct a prior on the family of paths with a low predictability profile, that is, a process whose location in the future is hard to predict from its current state and history.
- **Predictability profile** (Benjamini, Pemantle and Peres, 1998) of a stochastic process $(S_t)_{t\geq 0}$ is

$$PRE_S(k) := \sup_{x, history} \mathbb{P}(S_{t+k} = x \mid S_0, S_1, \ldots, S_t), \ k \in \mathbb{N}.$$

Theorem (Haggstrom & Mossel (1998) improving Benjamini et al)

If $(f_k)_{k\geq 1}$ is a decreasing and positive sequence such that $\sum_{k\geq 1} \frac{f_k}{k} < \infty$, then there exists a nearest-neighbor walk $(S_t)_{t\geq 0}$ starting at $S_0 = 0$ satisfying $PRE_S(k) \leq \frac{C}{kf_k}$.

Hoffman (1998) proved if $(f_k)_{k\geq 1}$ is a decreasing positive sequence with $\sum_{k\geq 1} \frac{f_k}{k} = \infty$, then the above predictability profile is impossible to achieve.

Predictability Profile and Number of Intersections

- The number of intersections between two independent nearest-neighbor walks drawn from a prior with low predictability profile is "small", namely it has exponential tails.

Lemma

For a walk $(S_t)_{0 \le t \le n-1}$ with finite number of steps, if

$$\sum_{1 \le k < n/B} PRE_S(kB) \le \theta < 1 \text{ for some } B,$$

then for any (possibly deterministic) sequence $(v_t)_{0 \le t \le n-1}$,

$$\mathbb{P}(|\mathbf{S} \cap \mathbf{v}| > k) \le B \cdot \theta^{k/B}.$$

Using this estimate and some work,

$\mathbb{E}_{\Pi_1 \times \Pi_2} e^{\mu_n^2 N_n} \lessapprox e^{\mu_n^2 \log^2(n) \log\log(n)} \to 1$ if $\mu_n \ll [\log(n)\sqrt{\log\log(n)}]^{-1}$.

This will imply $\liminf \gamma_n^* \ge 1$. So reliable detection is impossible.

Comments

- Our algorithm also applies to the ‘known initial location’ case.
- Our algorithm also applies to the case where the paths are not necessarily directed.
- The detection threshold vanishes as n grows.
- *Other distribution.* Similar results are available for other distributions from exponential family.
- *Other graph.* The detection threshold is sensitive to the underlying graph. The threshold is nonvanishing when the graph has a tree structure or has high dimension.

Future directions

The anomalous subset may be evolving with time.

- It may be associated to the unfolding of a stochastic process.
- It may be associated to or controlled by an underlying "particle system" running on the network.

Thank you