Notes on "Change point localization in dependent dynamic nonparametric random dot product graphs" by Padilla, Yu, and Priebe

7/28/20

The goal of this paper is to detect change points in a sequence of dynamic networks. Specifically, given a sequence of adjacency matrices $\{A(t)\}_{t=1}^T \subset \{0,1\}^{n\times n}$ which come from distributions $\{\mathcal{L}_t\}_{t=1}^T$, where for all $t = 1,\dots,T$, $\mathcal{L}_{t-1} \neq \mathcal{L}_t$ if and only if $t \in \{\eta_1,\dots,\eta_K\}$, for the sequence of change points $\{\eta_k\}_{k=1}^K \subset \{2,\dots,T\}$ with $1 = \eta_0 < \eta_1 < \cdots < \eta_K \leq T < \eta_{K+1} = T+1$, the authors want to estimate the $\{\eta_k\}_{k=1}^K$ accurately.

The authors consider time-varying RDPGs according to the following model:

Model 1. $\{A(t)\}_{t=1}^T \subset \{0,1\}^{n\times n}$ are a sequence of adjacency matrices of RDPGs satisfying the following:

1. For all t,
$$\mathbb{P}\{A(t)|X(t)\} = \prod_{1\leq i<j\leq n} (X_i(t)^\top X_j(t))^{A_{ij}(t)}(1 - X_i(t)^\top X_j(t))^{1-A_{ij}(t)},$$
where $X(t) = (X_1(t),\dots,X_n(t))^\top \in \mathbb{R}^{n\times d}$ satisfies the following:
There exists a sequence $1 = \eta_0 < \eta_1 < \cdots < \eta_k \leq T, \eta_{K+1} = T+1$ of change points. For $k \in \{0,\dots,K\}$,
$$X_i(\eta_k) \in \mathbb{R}^d \overset{\text{ind}}{\sim} F_{\eta_k}, i = 1,\dots,n,$$
and for $t \in \{\eta_k + 1,\dots,\eta_{k+1} - 1\}$,
$$X_i(t) \begin{cases} = X_i(t-1) & \text{with probability } \rho \\ \overset{\text{ind}}{\sim} F_{\eta_k} & \text{with probability } 1-\rho \end{cases},$$
for inner product distributions F_t. For all t, $P_t := X(t)X(t)^\top$.

2. The minimal spacing between two consecutive change points satisfies
$$\min_{k=1,\dots,K+1}\{\eta_k - \eta_{k-1}\} = \Delta > 0$$

3. For each $k \in \{0,\dots,K\}$ and for any $X, Y \overset{i.i.d.}{\sim} F_{\eta_k}$, denote
$$G_{\eta_k}(z) = \mathbb{P}\left\{X^\top Y \leq z\right\}, z \in [0,1].$$
The magnitudes of the changes in the data generating distribution are such that
$$\min_{k=1,\dots,K} \sup_{z\in[0,1]} |G_{\eta_k}(z) - G_{\eta_{k-1}}(z)| = \min_{k=1,\dots,K+1} \kappa_k = \kappa > 0.$$

4. For every $k \in \{0, \ldots, K\}$ and $i \in \{1, \ldots, n\}$,

$$\mathbb{E}\left\{X_i(\eta_k)X_i(\eta_k)^\top\right\} = \Sigma_k \in \mathbb{R}^{d\times d},$$

where Σ_k has eigenvalues $\mu_1^k \geq \cdots \geq \mu_d^k > 0$, with $\{\mu_\ell^k : k = 0, \ldots, K, \ell = 1, \ldots, d\}$ all being universal constants.

So

1. the latent positions at a change point are obtained via random draw from F_{η_k}. Between change points, the latent positions either stay the same, or are a new draw from F_{η_k}. At change points, this distribution F_{η_k} may also change.
2. We can't have multiple change points on top of each other, and the minimum distance between change points is $\Delta > 0$.
3. The changes between $F_{\eta_{k-1}}$ and F_{η_k} are sufficiently large in the sense that the distributions G_{η_k} of the inner product arising from two i.i.d. draws of a given F_{η_k}, always differ in Kolmogorov-Smirnov distance by at least κ.
4. The covariance matrices for all F_{η_k} are positive definite.
5. As time passes, the correlation between latent positions is exponentially decaying, and if a change point has occurred between the two time points, they are independent.

The main test statistic employed is the CUSUM statistic, which takes $\hat{X}(t) = U_A(t)\Lambda_A(t)^{1/2}$, and forms

$$\hat{Y}^t_{i,n/2+i} := \hat{X}_i(t)^\top \hat{X}_{n/2+i}(t),$$

where $\hat{X}_i(t)^\top$ is the ith row of $\hat{X}(t)$, for all $i = 1, \ldots, n/2$ (assume n is even). Then for $(s, t, e), 0 \leq s < t < e \leq T$, $z \in \mathbb{R}$, the CUSUM statistic is

$$D^t_{s,e} = \sup_{z\in[0,1]} |D^t_{s,e}(z)|,$$

where for $\alpha_{s,e}(t) = [(e-t)/(t-s)]^{1/2}$, $D^t_{s,e}(z)$ is defined as

$$D^t_{s,e}(z) = \sqrt{\frac{2}{n(e-s)}}\left|\alpha_{s,e}(t)\sum_{k=s+1}^{t}\sum_{i=1}^{n/2}\chi_{\leq z}(\hat{Y}^k_{i,n/2+i}) - \alpha_{s,e}(t)^{-1}\sum_{k=t+1}^{e}\sum_{i=1}^{n/2}\chi_{\leq z}(\hat{Y}^k_{i,n/2+i})\right|.$$

In essence, the first sum in $D^t_{s,e}(z)$ approximates $G_t(z)$, while the second sum approximates $G_{t+1}(z)$, and $D^t_{s,e}$ estimates the KS distance between these. So long as there are no change points between s and t, as s decreases, the first estimate should improve. Similarly, so long as there are no change points between $t+1$ and e, the second estimate should improve. This explains assumptions 2 and 3 in the model: We need the gap between change points to be large enough that we can get a decent estimate of the KS distance between G_t and G_{t+1}, which requires a number of steps before and after t without a change point; and we need the true KS distance at change points to be large enough that we can detect true changes even with our noisy estimate of this distance.

Algorithm 2 chops the full time sequence $1, \ldots, T$ into subintervals (α_m, β_m), identifies the most likely change points in each subinterval such that $D^t_{\alpha_m,\beta_m}$ exceeds some threshold τ, chooses the most likely change point b_{m^*} among these, and adds it to the set of estimated change points. It then repeats the procedure on the two smaller time sequences $1, \ldots, b_{m^*}$, and $b_{m^*}+1, \ldots, T$, and continues until all smaller time sequences are either too short or do not have any value of $D^t_{s_m,e_m}$ which exceeds the threshold, where s_m and e_m are endpoints chosen by intersecting the intervals (α_m, β_m) with the new time sequence (s, e).

Theory

The first few lemmas show that when $G \neq \tilde{G}$, $F \neq \tilde{F}$, which should be clear since the distributions G are obtained from the distributions F by taking the inner product of two independent draws from F. When G and $\tilde{G}$ differ enough in KS distance, this further implies that the distributions of $A \sim \text{RDPG}(F, n)$ and $\tilde{A} \sim \text{RDPG}(\tilde{F}, n)$ are different. This makes sense, since the distributions of $X_i^\top X_j$ and $\tilde{X}_i^\top \tilde{X}_j$ should be fairly different at this point, and these are the means of A_{ij} and $\tilde{A}_{ij}$, respectively. Finally, this is equivalent to the first $n-1$ moments of F and $\tilde{F}$ not being identical, which can be shown by writing the probability mass function for $A|X$ in terms of these moments of X.

The next assumption essentially says that $\kappa\sqrt{\Delta}$ grows sufficiently quickly as $T \to \infty$. Recall that κ tells us about the minimum KS distance between distributions before and after a change point, and Δ gives the minimum number of time steps between change points, so we should expect that increasing these will improve the level of signal. Also, the assumption makes clear that $\rho < 1$ is important for the method to proceed, since they require more than one sample from F_{η_k}, and these only appear between change points when $\rho < 1$.

Under this assumption, the authors prove the following Theorem:

Theorem 1. *Suppose d in Algorithm 2 is the true dimension, and $\tau \in [c_1, c_2]$, a certain interval depending on $\kappa, \Delta, n, \rho, T$, and d. Then if the subintervals in Algorithm 2 satisfy*

$$\max_{m=1,\ldots,M} |\alpha_m - \beta_m| \leq C\Delta,$$

the estimated change points $\{\hat{\eta}_k\}_{k=1}^{\hat{K}}$ satisfy

$$\mathbb{P}\left\{\hat{K} = K, |\hat{\eta}_k - \eta_k| \leq C\frac{T\log(n)}{\kappa_k^2 n(1-\rho)}\ \forall k\right\} \geq 1 - \delta(n, T),$$

where δ is decaying at least polynomially quickly in n and T.

They also discuss some possible extensions and run several simulations, the latter of which reveal that in the presence of correlation of the graphs at the time steps between change points, their approach improves on existing methods.