MURI Update 5.12

Phyllis Wan, Richard Davis

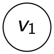
Columbia University

May 12, 2017

Part I – Inference of preferential attachment model

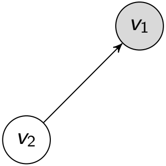
Linear preferential attachment model

- Add one edge each time t



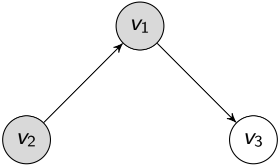
Linear preferential attachment model

- Add one edge each time t



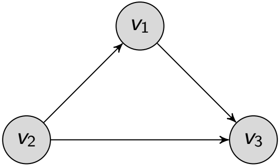
Linear preferential attachment model

- Add one edge each time t



Linear preferential attachment model

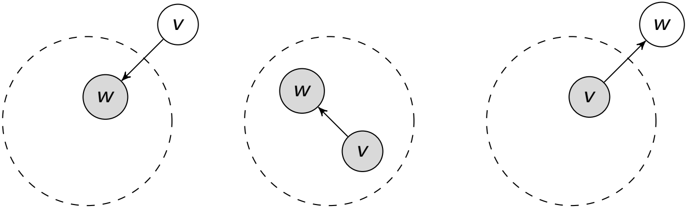
- Add one edge each time t



Linear preferential attachment model $(\alpha, \beta, \gamma, \delta_{\text{in}}, \delta_{\text{out}})$

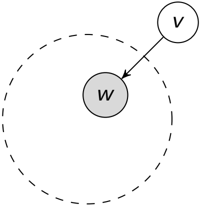
- How to add an edge?

With probability α With probability β With probability γ



Linear preferential attachment model $(\alpha, \beta, \gamma, \delta_{\text{in}}, \delta_{\text{out}})$

With probability α

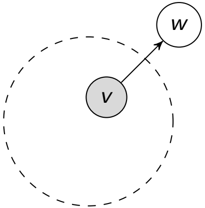


$$\Pr(\text{node } w \text{ is chosen}) \propto D_{\text{in}}(w) + \delta_{\text{in}}$$

- Small δ_{in} – popular nodes get more attention
- Very large δ_{in} – all node get (almost) equal attention

Linear preferential attachment model $(\alpha, \beta, \gamma, \delta_{\text{in}}, \delta_{\text{out}})$

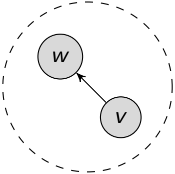
With probability γ



$\Pr(\text{node } v \text{ is chosen}) \propto D_{\text{out}}(v) + \delta_{\text{out}}$

Linear preferential attachment model $(\alpha, \beta, \gamma, \delta_{\text{in}}, \delta_{\text{out}})$

With probability β



$$\Pr(\text{nodes } w, v \text{ are chosen}) \propto (D_{\text{in}}(w) + \delta_{\text{in}})(D_{\text{out}}(v) + \delta_{\text{out}})$$

Power laws in degree distributions

- As $n \to \infty$,

$$\frac{N_i^{\text{in}}(n)}{N(n)} \xrightarrow{a.s.} p_i^{\text{in}} \sim i^{-a_1}$$

$$\frac{N^{\text{out}} j(n)}{N(n)} \xrightarrow{a.s.} p_j^{\text{out}} \sim j^{-a_2}$$

  - with

$$a_1 = 1 + \frac{1 + \delta_{\text{in}}(\alpha + \gamma)}{\alpha + \beta}, \quad a_2 = 1 + \frac{1 + \delta_{\text{out}}(\alpha + \gamma)}{\beta + \gamma}.$$

Inference on the parameters $(\alpha, \beta, \gamma, \delta_{\text{in}}, \delta_{\text{out}})$?

The kind of data likely to be available...

- The full history of the network
- A snapshot of the network

MLE estimation for parameters $(\alpha, \beta, \delta_{\text{in}}, \delta_{\text{out}})$

- If the full history of the network with n edges is observed

$$
\begin{aligned}
L(\alpha, \beta, &\delta_{\text{in}}, \delta_{\text{out}}) \\
&= \prod_{t=1}^{n} \left(\alpha \frac{D_{\text{in}}^{(t-1)}(v_t^{(2)}) + \delta_{\text{in}}}{t - 1 + \delta_{\text{in}} N(t-1)} \right)^{\mathbf{1}_{\{J_t = 1\}}} \\
&\quad \times \prod_{t=1}^{n} \left(\beta \frac{D_{\text{in}}^{(t-1)}(v_t^{(2)}) + \delta_{\text{in}}}{t - 1 + \delta_{\text{in}} N(t-1)} \frac{D_{\text{out}}^{(t-1)}(v_t^{(1)}) + \delta_{\text{out}}}{t - 1 + \delta_{\text{out}} N(t-1)} \right)^{\mathbf{1}_{\{J_t = 2\}}} \\
&\quad \times \prod_{t=1}^{n} \left((1 - \alpha - \beta) \frac{D_{\text{out}}^{(t-1)}(v_t^{(1)}) + \delta_{\text{out}}}{t - 1 + \delta_{\text{out}} N(t-1)} \right)^{\mathbf{1}_{\{J_t = 3\}}}
\end{aligned}
$$

Theorem

The MLE estimator

$$\hat{\boldsymbol{\theta}} := (\hat{\alpha}, \hat{\beta}, \hat{\delta}_{\text{in}}, \hat{\delta}_{\text{out}}) = \arg\max L(\alpha, \beta, \delta_{\text{in}}, \delta_{\text{out}})$$

is unique, strongly consistent, and asymptotically normal, i.e.

$$\hat{\boldsymbol{\theta}} \stackrel{a.s.}{\to} (\alpha, \beta, \delta_{\text{in}}, \delta_{\text{out}}) =: \boldsymbol{\theta},$$

and

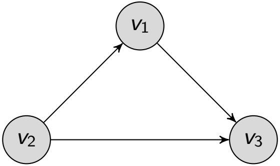
$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \stackrel{d}{\to} N(\mathbf{0}, I^{-1}(\boldsymbol{\theta})),$$

where $I(\boldsymbol{\theta})$ is the asymptotic Fisher information matrix for the parameters.

One-snapshot estimation for parameters $(\alpha, \beta, \delta_{\text{in}}, \delta_{\text{out}})$

- If only a snapshot of the network with n edges is observed,

$$G(n) = (V(n), E(n))$$



One-snapshot estimation for parameters $(\alpha, \beta, \delta_{\text{in}}, \delta_{\text{out}})$

- approximate the score functions using law of large numbers

$$\frac{1}{n}\frac{\partial \log L}{\partial \delta_{\text{in}}} \approx \sum_{i=0}^{\infty} \frac{N^{\text{in}}_{>i}(n)/n}{i+\delta_{\text{in}}} - \frac{1-\alpha-\beta}{\delta_{\text{in}}} - \frac{1-\beta}{1+\delta_{\text{in}}(1-\beta)}(\alpha+\beta) = 0,$$

$$\frac{1}{n}\frac{\partial \log L}{\partial \delta_{\text{out}}} \approx \sum_{j=0}^{\infty} \frac{N^{\text{out}}_{>j}(n)/n}{j+\delta_{\text{out}}} - \frac{\alpha}{\delta_{\text{out}}} - \frac{1-\beta}{1+\delta_{\text{in}}(1-\beta)}(1-\alpha) = 0.$$

One-snapshot estimation for parameters $(\alpha, \beta, \delta_{\text{in}}, \delta_{\text{out}})$

- approximate the score functions using law of large numbers

$$\frac{1}{n}\frac{\partial \log L}{\partial \delta_{\text{in}}} \approx \sum_{i=0}^{\infty} \frac{N^{\text{in}}_{>i}(n)/n}{i+\delta_{\text{in}}} - \frac{1-\alpha-\beta}{\delta_{\text{in}}} - \frac{1-\beta}{1+\delta_{\text{in}}(1-\beta)}(\alpha+\beta) = 0,$$

$$\frac{1}{n}\frac{\partial \log L}{\partial \delta_{\text{out}}} \approx \sum_{j=0}^{\infty} \frac{N^{\text{out}}_{>j}(n)/n}{j+\delta_{\text{out}}} - \frac{\alpha}{\delta_{\text{out}}} - \frac{1-\beta}{1+\delta_{\text{in}}(1-\beta)}(1-\alpha) = 0.$$

- moment matching

$$\frac{N^{\text{in}}_0(n)}{n} \to p^{\text{in}}_0 = \frac{\alpha}{(1+\frac{\alpha+\beta}{1+\delta_{\text{in}}(1-\beta)}\delta_{\text{in}})},$$

$$\frac{N^{\text{out}}_0(n)}{n} \to p^{\text{in}}_0 = \frac{1-\alpha-\beta}{(1+\frac{1-\alpha}{1+\delta_{\text{out}}(1-\beta)}\delta_{\text{out}})}.$$

- Solve the above equations to obtain $\tilde{\boldsymbol{\theta}} = (\tilde{\alpha}, \tilde{\beta}, \tilde{\delta}_{\text{in}}, \tilde{\delta}_{\text{out}})$

Theorem: strong consistency of the one-snapshot estimation

The estimator $\tilde{\boldsymbol{\theta}} := (\tilde{\alpha}, \tilde{\beta}, \tilde{\delta}_{\text{in}}, \tilde{\delta}_{\text{out}})$ is unique and strongly consistent, i.e.

$$\tilde{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}.$$

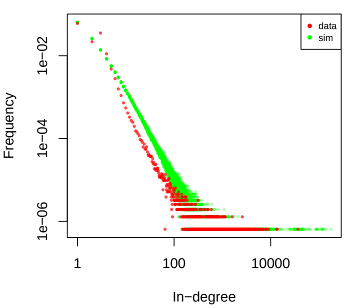
- Asymptotic normality is implied from simulations but yet to be proved (Tiandong?)
- From the simulations, $\tilde{\boldsymbol{\theta}}$ can achieve $\sim 20\%$ efficiency compared to MLE, since less information is provided.

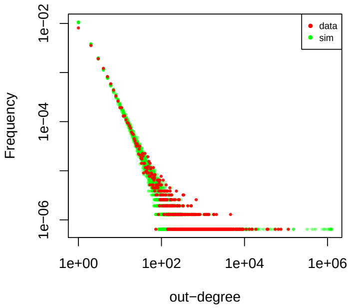
Dutch Wiki talk network

- A $\rightarrow$ B: user A writing a message on the talk page of user B
- 225,749 nodes (users) and 1,554,699 edges (messages)
- full history available

Dutch Wiki talk network

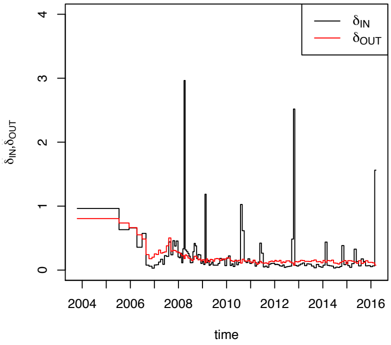
- Goodness-of-fit of degree distributions





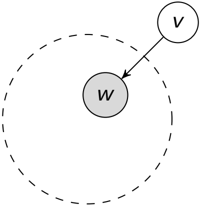
Dutch Wiki talk network

- Separate the network into ∼ 150 time intervals
- Fit the model in each interval



Linear preferential attachment model $(\alpha, \beta, \gamma, \delta_{\text{in}}, \delta_{\text{out}})$

With probability α

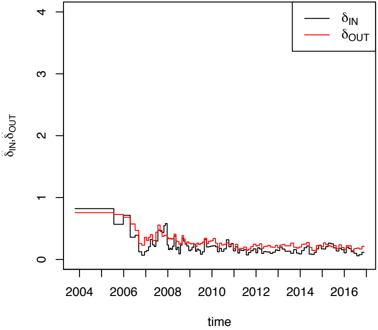


$$\Pr(\text{node } w \text{ is chosen}) \propto D_{\text{in}}(w) + \delta_{\text{in}}$$

- Small δ_{in} – popular nodes get more attention
- Very large δ_{in} – all node get (almost) equal attention

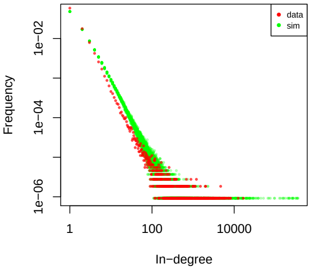
Dutch Wiki talk network

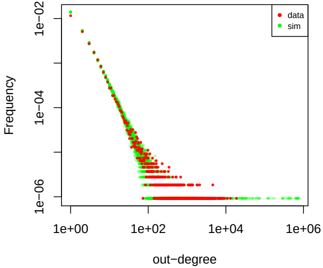
- Discover the existence of accounts with broadcasting feature
- Remove 37 administrative accounts



Dutch Wiki talk network

- Re-fit the linear preferential attachment model





Part II – Threshold selection for multivariate heavy-tailed data

Multivariate regular variation

Let $(R, \boldsymbol{\Theta}) = (\|\mathbf{X}\|, \frac{\mathbf{X}}{\|\mathbf{X}\|})$, $\mathbf{X} \sim \mathbf{RV}(\alpha)$ if and only if $R \sim RV(\alpha)$ and

$$P(\boldsymbol{\Theta} \in \cdot | R > r) \to S(\cdot), \quad r \to \infty.$$

Given $(R_1, \boldsymbol{\Theta}_1), \ldots, (R_n, \boldsymbol{\Theta}_n)$, how to model $S(\cdot)$?

- Take $\{\boldsymbol{\Theta}_i\}$ for r_0 large

How to choose r_0?

- Choose r_0 s.t. when conditional on $R > r_0$, $\boldsymbol{\Theta}$ and R are approximately independent

How to measure the dependence between $\boldsymbol{\Theta}$ and R?

- Distance covariance

A simulated example

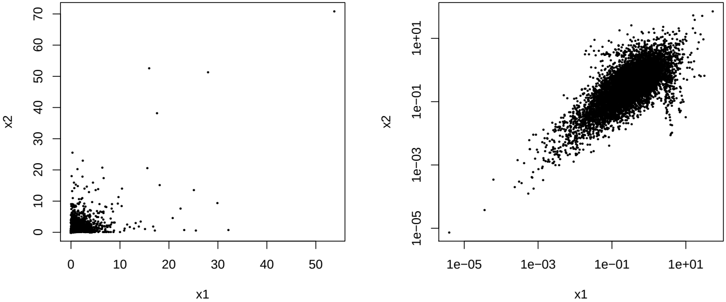


Figure: Simulated $\mathbf{X}$ where R is independent of Θ only when conditioning on $R > r_{0.9}$.

A simulated example (cont.)

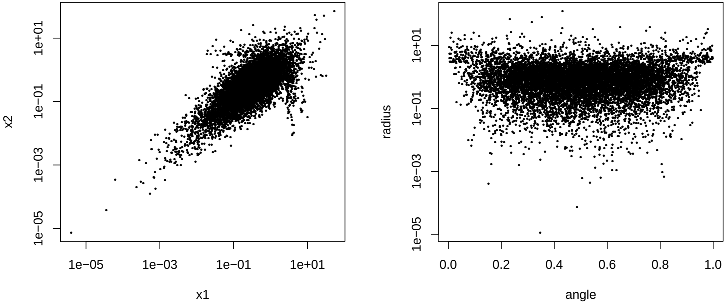


Figure: Simulated $\mathbf{X}$ where R is independent of Θ only when conditioning on $R > r_{0.9}$.

A simulated example (cont.)

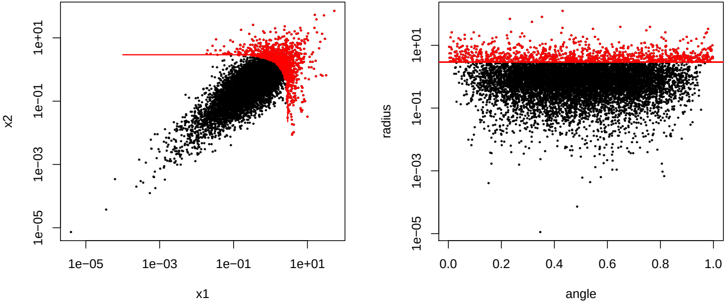


Figure: Simulated $\mathbf{X}$ where R is independent of Θ only when conditioning on $R > r_{0.9}$.

A simulated example (cont.)

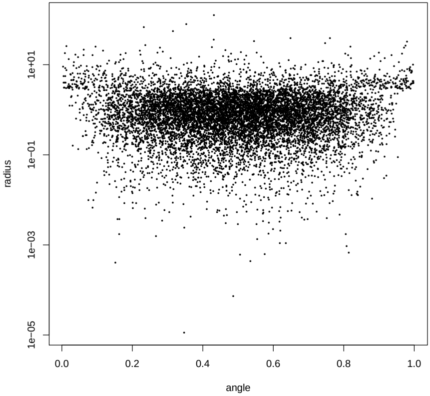


Figure: Simulated $\mathbf{X}$ where R is independent of Θ only when conditioning on $R > r_{0.9}$.

A simulated example (cont.)

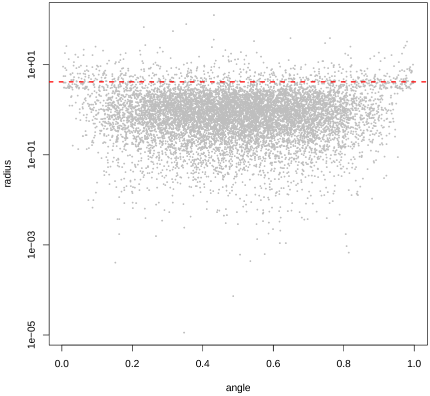


Figure: Simulated **X** where R is independent of Θ only when conditional on $R > r_{0.9}$.

A simulated example (cont.)

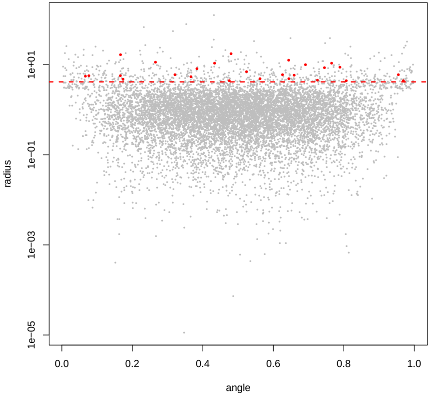


Figure: Simulated **X** where R is independent of Θ only when conditional on $R > r_{0.9}$.

A simulated example (cont.)

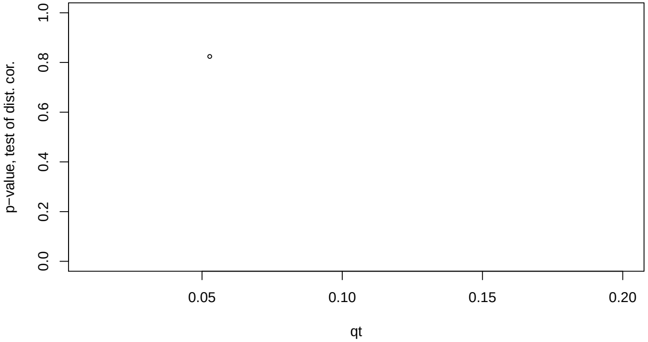


Figure: P-value of test of independence of truncated R and Θ vs. the truncation level. The true independence level is 0.1.

A simulated example (cont.)

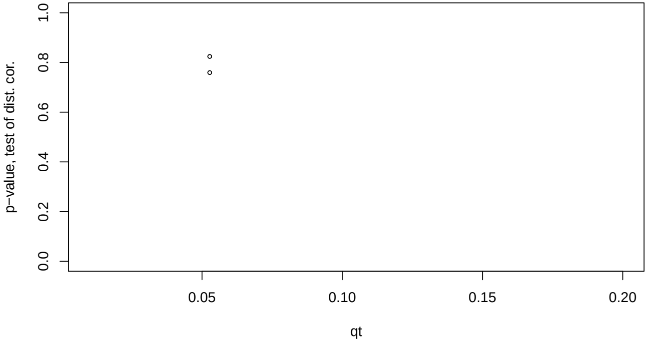


Figure: P-value of test of independence of truncated R and Θ vs. the truncation level. The true independence level is 0.1.

A simulated example (cont.)

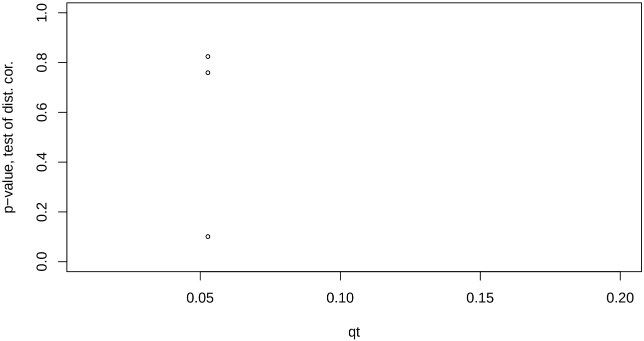


Figure: P-value of test of independence of truncated R and Θ vs. the truncation level. The true independence level is 0.1.

A simulated example (cont.)

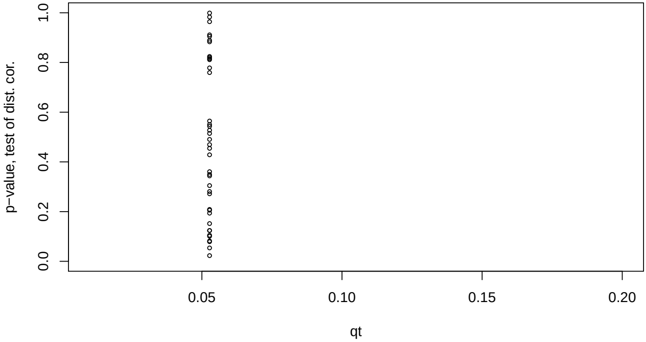


Figure: P-value of test of independence of truncated R and Θ vs. the truncation level. The true independence level is 0.1.

A simulated example (cont.)

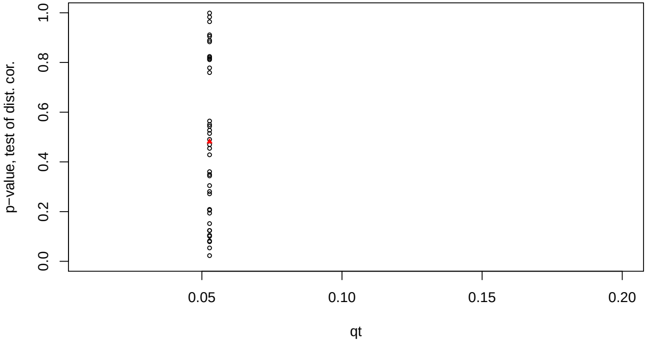


Figure: P-value of test of independence of truncated R and Θ vs. the truncation level. The true independence level is 0.1.

A simulated example (cont.)

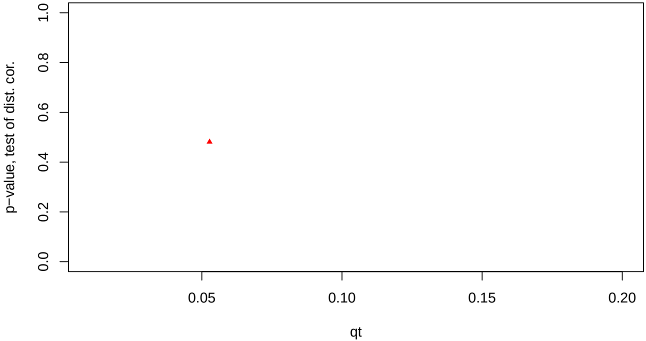


Figure: Mean p-value of test of independence of truncated R and Θ vs. the truncation level. The true independence level is 0.1.

A simulated example (cont.)

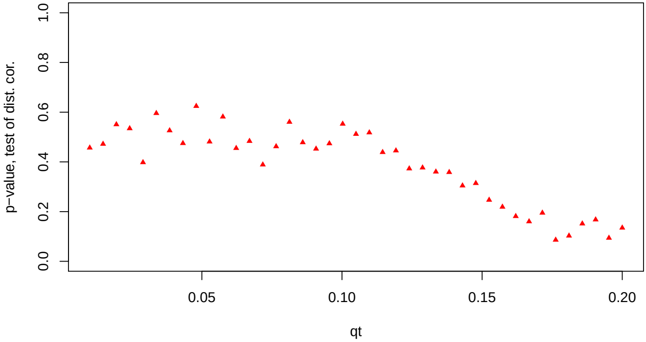


Figure: Mean p-value of test of independence of truncated R and Θ vs. the truncation level. The true independence level is 0.1.

A simulated example (cont.)

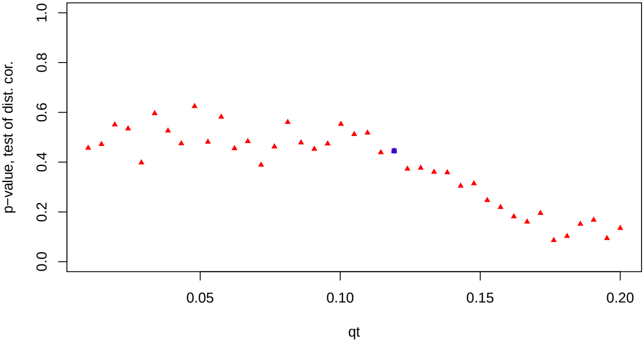


Figure: Mean p-value of test of independence of truncated R and Θ vs. the truncation level. The true independence level is 0.1.

How to choose the change point?

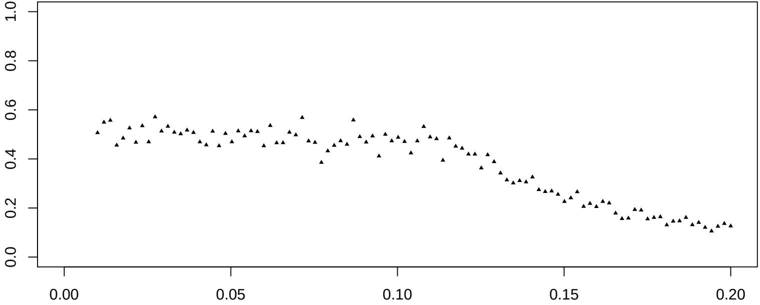


Figure: Mean p-value of test of independence of truncated R and Θ vs. the truncation level. The true independence level is 0.1.

Idea 1: CUSUM algorithm (Page (1954))

For each k,

- Consider $\sum_{i=j}^{k}(0.5 - pv_i)$ for $j = 1, \ldots, k$
- Calculate
$$H = \max_j \left[\sum_{i=j}^{k}(0.5 - pv_i) \right]$$
- Detect change point when H exceed a certain threshold h_0

Idea 1: CUSUM algorithm

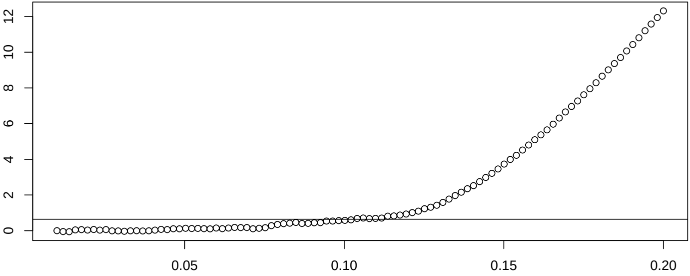


Figure: CUSUM plot for the previous p-value path.

Idea 1: CUSUM algorithm

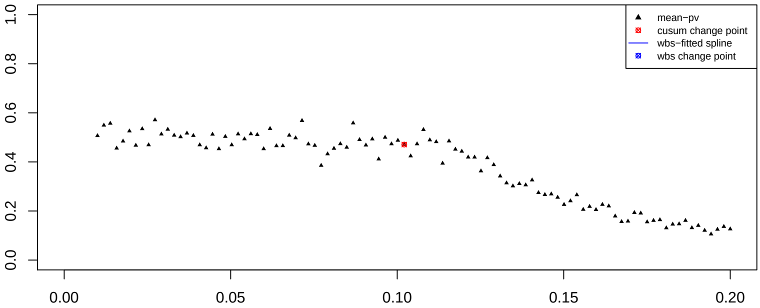


Figure: Mean p-value of test of independence of truncated R and Θ vs. the truncation level. The true independence level is 0.1.

Idea 2: Wild binary segmentation (Fryzlewicz (2014))

- Based on the CUSUM idea
- Fit a piecewise constant spline to the data

Idea 2: Wild binary segmentation

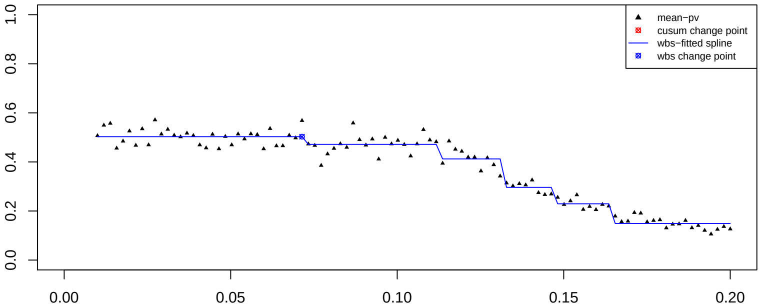


Figure: Mean p-value of test of independence of truncated R and Θ vs. the truncation level. The true independence level is 0.1.