1 Likelihood restriction

This section explicates main restrictions of the likelihood function. They are essential for the proposed algorithm of change point detection. Further assume that log-likelihood function $L(\theta) = L(Y, \theta)$, $Y = (Y_1, \ldots, Y_n)$, has rather precise approximation by its quadratic part in local region $\Theta_0(r)$ of θ^*, $\Theta_0(r) \subseteq I\!R^p$, where

$$\theta^* = \operatorname*{argmax}_\theta I\!E L(\theta), \quad \widehat{\theta} = \operatorname*{argmax}_\theta L(\theta)$$

and $\Theta_0(r) = \{\|D(\theta - \theta^*)\| < r\}$. [**?**] provides required conditions for justified quadratic approximation and parameter concentration in the local region. Approximation error involves the next variables for its estimation:

$$\alpha(\theta, \theta_0) = L(\theta) - L(\theta_0) - (\theta - \theta_0)^T \nabla L(\theta_0) + \frac{1}{2}\|D(\theta - \theta_0)\|^2,$$

$$\chi(\theta, \theta_0) = D^{-1}\nabla\alpha(\theta, \theta_0) = D^{-1}(\nabla L(\theta) - \nabla L(\theta_0)) + D(\theta - \theta_0).$$

Let in region $\Theta_0(r)$ with probability $1 - e^{-x}$:

$$\frac{|\alpha(\theta, \theta^*)|}{\|D(\theta - \theta^*)\|} \le \diamondsuit(r, x), \quad \|\chi(\theta, \theta^*)\| \le \diamondsuit(r, x), \tag{A}$$

where $\diamondsuit(r, x) = (\delta(r) + 6v_0 z_H(x)\omega)r$,

$$D^2(\theta) = -\nabla^2 I\!E L(\theta), \quad D = D(\theta^*), \tag{D}$$

$$\|D^{-1}D^2(\theta)D^{-1} - I_p\| \le \delta(r), \tag{dD}$$

$$\forall\lambda \le g,\ \gamma_1\gamma_2 \in I\!R^p: \quad \log I\!E \exp\left\{\frac{\lambda}{\omega}\frac{\gamma_1^T\nabla^2\zeta(\theta)\gamma_2}{\|D\gamma_2\|\|D\gamma_2\|}\right\} \le \frac{\nu_0^2\lambda^2}{2}, \tag{ED2}$$

$$z_H(x) = \sqrt{H} + \sqrt{2x} + \frac{g^{-2}x + 1}{g}H, \quad H = 6p,$$

$$\omega = \omega(n) \sim \frac{1}{\sqrt{n}}, \quad \delta(r) \sim \frac{r}{\sqrt{n}}, \quad r \sim \sqrt{p}.$$

Condition (dD) ensures quadratic approximation of $I\!E L(\theta)$ and (ED2) ensures linear approximation of centered likelihood $\zeta(\theta) = L(\theta) - I\!E L(\theta)$.

2 LRT Statistic

Provide simple non-strict explanation of what kind of distribution the main statistic T_h is supposed to have. Review T_h as

$$T_h = L(\widehat{\theta}) - L(\widehat{\theta}_{H_0}), \quad L(\theta) = L_1(\theta_1) + L_2(\theta_2), \quad L_1 = L(Y[1:h]),\ L_2 = L(Y[h:2h]),$$

where $\widehat{\theta}_{H_0}$ is argmax of L under condition $H_0 : \theta_1 = \theta_2$. Then under quadratic approximation assumption T_h could be presented in Tailor equation with point $\widehat{\theta}$:

$$T_h \approx \frac{1}{2}\|D(\widehat{\theta} - \widehat{\theta}_{H_0})\|^2.$$

If $\widehat{\theta}$ and $\widehat{\theta}_{H_0}$ tend to be Normal and H_0 is true then their difference are close to a centered Normal variable. If H_0 is false – the Normal variable will have mean that is equal to $\theta^* - \theta^*_{H_0}$.

The next two theorems includes more formal properties of T_h statistic.

Theorem 1. Assume condition (L*) and quadratic Laplace approximation (A) of L_1 and L_2 are fulfilled with probability $1 - 2e^{-x}$, additionally with probability $1 - 2e^{-x}$

$$\|\xi_i\| \le z(x), \quad z^2(x) = \max_i p_{B_i} + 6\lambda_{B_i} x,$$

$$B_i = D_i^{-1} \operatorname{Var}(\nabla L_i(\theta^*)) D_i^{-1}, \quad p_B = \operatorname{tr}(B), \quad \lambda_B = \lambda_{\max}(B). \tag{B}$$

Then in the local region with probability $1 - 8e^{-x}$

$$2T_h = \|\xi_{12} + \theta^*_{12}\|^2 + O(\{r + z(x)\}\diamondsuit(r, x)),$$

where

$$\xi_{12} = \Sigma(D_2^{-1}\xi_2 - D_1^{-1}\xi_1), \quad \theta^*_{12} = \Sigma(\theta^*_2 - \theta^*_1).$$

Remark. In increasing sample size $n \to \infty$ the stochastic component tends to Normal distribution:

$$\xi_{12} \to \mathcal{N}(0, B_1 + B_2).$$

Remark. For the condition $\widehat{\theta} \in \Theta_1(r) \cap \Theta_2(r)$ the restriction of the parameter variability θ^* is required

$$\|D(\theta^*_1 - \theta^*_2)\| \le r. \tag{L*}$$

Theorem 2. Assume condition (L*) and quadratic Laplace approximation (A) with probability $1 - 2e^{-x}$ are fulfilled. Then with probability $1 - 4e^{-x}$ in the local region $\Theta_1(r) \cap \Theta_2(r)$ took place

$$\left| \sqrt{2T_h} - \|\xi_{12} + \theta^*_{12}\| \right| \le 10\diamondsuit(r, x).$$

where ξ_{12} and θ^*_{12} are defined in theorem 1.

3 Multiplier Bootstrap

Likelihood function in bootstrap case is a zipped sum with i.i.d weights $(u_1, \ldots, u_n)$:

$$L^\flat(\theta) = \sum_{i=1}^n u_i l_i(\theta).$$

Each weight element has $\operatorname{Var}_\flat u_i = 1$ and $I\!E_\flat u_i = 1$, that results in ($\{l_i(\theta)\}$ assumed to be independent)

$$I\!E_\flat L^\flat(\theta) = L(\theta), \quad \operatorname{Var}_\flat \nabla L^\flat(\theta) = \sum_{i=1}^n \nabla l_i(\theta) \nabla l_i(\theta)^T \approx \operatorname{Var} \nabla L(\theta) + \sum_{i=1}^n I\!E \nabla l_i(\theta) I\!E \nabla l_i(\theta)^T.$$

Variable $\zeta = \sum_i \zeta_i$ denotes stochastic part of the likelihood ($\zeta_i(\theta) = l_i(\theta) - I\!E l_i(\theta)$).

Lemma (Deviations of empirical process norm)**.** Given exponential moment restriction with $\|\gamma_1\| = \|\gamma_2\| = 1$

$$\log I\!E \exp\left\{ \frac{\lambda}{\omega} \gamma_1 D^{-1} \nabla^2 \zeta_i(\theta) D^{-1} \gamma_2 \right\} \le \frac{\lambda^2 \nu_i^2}{2}, \quad \sum_i \nu_i^2 = \nu_0^2, \tag{EDi}$$

in local region $\Theta_0(r)$ of radius r with probability $1 - e^{-x}$ the next statement is fulfilled for all θ, θ_0

$$\|D_0^{-1}(\nabla\zeta(\theta) - \nabla\zeta(\theta_0))\| \le 12\nu_0 z(x)\omega(n) r.$$

Paper [**?**] contains proof for this lemma.

Function $\alpha^\flat(\theta,\theta_0)$ determines quadratic approximation error for the weighted likelihood function.

$$\alpha^\flat(\theta,\theta_0) = L^\flat(\theta) - L^\flat(\theta_0) - (\theta-\theta_0)^T\nabla L^\flat(\theta_0) + \frac{1}{2}\|D(\theta-\theta_0)\|^2$$

Mean and deviation of the approximation error are

$$\|D^{-1}\nabla I\!\!E_\flat\alpha^\flat(\theta,\theta_0)\| = \|D^{-1}\nabla\alpha(\theta,\theta_0)\| \le 2\diamondsuit(r,x).$$

$$S^\flat(\theta,\theta_0) = D^{-1}\{\nabla\alpha^\flat(\theta,\theta_0) - I\!\!E_\flat\nabla\alpha^\flat(\theta,\theta_0)\} = \sum_{i=1}^n D^{-1}\{\nabla l_i(\theta) - \nabla l_i(\theta_0)\}(u_i-1). \tag{Sb}$$

Centered function $\overset{o}{S}_\flat = S^\flat(\theta,\theta_0) - I\!\!E_Y S^\flat(\theta,\theta_0)$ with probability $1-e^{-x}$ fulfills

$$\|\overset{o}{S}_\flat\| \le 12\nu_0 z(x)\omega r\sqrt{\sum_{i=1}^n(\nu_i^2/\nu_0^2)(u_i-1)^2}.$$

Let u_i has restricted exponential moment:

$$\log I\!\!E_\flat e^{(u_i-1)^2} \le 1, \tag{Eu}$$

then with probability $1-e^{-t}$

$$\sqrt{\sum_{i=1}^n(\nu_i^2/\nu_0^2)(u_i-1)^2} \le \sqrt{1+t}$$

Consequently

$$\|\overset{o}{S}_\flat\| \le 12\nu_0 z(x)\omega r\sqrt{1+x}.$$

One could get restriction for $I\!\!E S(\theta_2,\theta_1)$ from Hoeffding inequality 8 restricting its components norms:

$$\|D^{-1}\nabla^2 I\!\!E l_i(\theta)D^{-1}\| \le \frac{C_i(r)}{n}, \tag{dDi}$$

$$\|I\!\!E S^\flat(\theta,\theta_0)\| \le \frac{1+\sqrt{2x}}{2n}C(r)rc_u,$$

where $C(r) = \sqrt{\sum_i C_i^2(r)}$, $c_u = \max(u_i-1)$.

Theorem 3 (Boot, Wilks)**.** Under conditions (EDi), (dDi), (A) and restrictions for weights (Eu), in local r-region it holds with probability $1-4e^{-x}$

$$|\alpha^\flat(\theta,\theta_0)| \le \|D(\theta-\theta_0)\|\left(2\diamondsuit(r,x) + 12\nu_0 z(x)\omega r\sqrt{1+x} + \frac{1+\sqrt{2x}}{2n}C(r)rc_u\right) = \tag{Ab}$$

$$= \|D(\theta-\theta_0)\|\,2\diamondsuit^\flat(r,x).$$

Fisher theorem could be proved from condition

$$\chi^\flat(\theta,\theta_0) = D^{-1}(\nabla L^\flat(\theta) - \nabla L^\flat(\theta_0)) + D(\theta-\theta_0),$$

$$\chi^\flat(\theta,\theta_0) = D^{-1}\nabla\alpha^\flat(\theta,\theta_0).$$

With analog steps to 3 theorem proof one get an upper bound for $\chi^\flat(\theta,\theta_0)$.

Theorem 4 (Boot, Fisher)**.** Under conditions (EDi), (dDi), (A) and restrictions for weights (Eu), in local r-region it holds with probability $1-4e^{-x}$

$$\|\chi^\flat(\widehat{\theta}^\flat,\widehat{\theta})\| = \|D(\widehat{\theta}^\flat-\widehat{\theta}) - D^{-1}\nabla L^\flat(\widehat{\theta})\| \le 2\diamondsuit^\flat(r,x),$$

where $\widehat{\theta}^\flat$, $\widehat{\theta}$ – MLE of weighted and non-weighted likelihood functions.

4 Uniform bands

Quadratic approximation of $\alpha(\theta_0,\theta)$ is correct with high probability for each window position t, but it takes place uniformly for all the windows with higher probability than independent events because of correlations. Corresponding probability of the uniform approximation could be obtained by grouping window positions into high correlation regions. After split the whole data range of size T into such regions of size r_t the uniform approximations are to be true for all the regions with probability $1-e^{-x_t}$, $x_t = x - \log(T/r_t)$. The first grouping approach uses bootstrap related equations.

$$L(\theta,t) = \sum_{i=1}^{n} l_i(\theta) u_i(t), \quad 1 \le t \le n = h + 2r_t,$$

where $u_i(t)$ denotes kernel function that selects elements around coordinate t with distance less than $h/2$:

$$u_i(t) = \mathbb{1}\{|t-i| \le h/2\}, \quad \mathbb{E}_t u_i(t) = \frac{h}{n}.$$

Weights $u_i(t)$ have probabilistic nature since they are dependently generated by sampling position $t \in [1,n]$ uniformly. The expectation comes from all window positions of circle closed data ($Y_{n+i} = Y_i$).

$$\alpha(t,\theta,\theta_0) = L(\theta,t) - L(\theta_0,t) - (\theta-\theta_0)\nabla L(\theta_0,t) + \frac{1}{2}\|D_h(\theta-\theta_0)\|^2$$

Precise quadratic approximation requires small value of $\nabla\alpha(t,\theta,\theta_0)$ for all From (A) and weights mean follows

$$\|\mathbb{E}_t D_h^{-1} \nabla\alpha(t)\| \le \sqrt{\frac{h}{n}} \cdot 2\diamondsuit(r(\theta^*),x,n) \sim \sqrt{\frac{h}{n}}\sqrt{\frac{p}{n}}.$$

Defining variable similar to (Sb) one obtains its mean and deviation bounds in separate steps.

$$S(t) = D_h^{-1}\{\nabla\alpha(t) - \mathbb{E}_t\nabla\alpha(t)\}, \quad \|S(t) - \mathbb{E}_Y S(t)\| \le \sqrt{1-\frac{h}{n}} \cdot 12\nu_0 z(x,p)\omega(h) \max_i r(\theta_i^*).$$

Here $r(\theta_i^*)$ denotes max distance from argmax $\mathbb{E}l_i(\theta)$ to (θ,θ_0). From asymptotic assumptions for $\omega(h) \sim 1/\sqrt{h}$ and $z(x) \sim \sqrt{p}$ one get

$$\|S(t) - \mathbb{E}_Y S(t)\| \sim \sqrt{\frac{p}{h}}\overline{r}.$$

The mean part requires additional condition (dDi) for divergence of the likelihood components

$$\|\mathbb{E}_Y S(t)\| \le \max_i \frac{C(r_i)}{h} r_i \sum_{j=1}^{n} |u_j - \mathbb{E}_t u_j| = 2\max_i C(r_i) r_i \left(1 - \frac{h}{n}\right),$$

where

$$\sum_{j=1}^{n} |u_j - \mathbb{E}_t u_j| = (n-h)\left|0 - \frac{h}{n}\right| + h\left|1 - \frac{h}{n}\right| = 2h\left(1-\frac{h}{n}\right).$$

Finally, the uniform estimation error for region n is

$$12\nu_0 z(x,p) w(h) \max_i r(\theta_i^*) + 2\max_i C(r_i) r_i \left(1 - \frac{h}{n}\right).$$

From which with $C(r_i) \sim r_i$ one get $1/\sqrt{h} \sim (1-h/n)$, consequently $n - h \sim \sqrt{h}$.

The second grouping approach treats window position t as a part of parameter θ. Define variable y as difference of stochastic likelihood part gradients:

$$y(\theta, t) = D_h^{-1} \left(\nabla\zeta(\theta, t) - \nabla\zeta(\theta_0, t_0) \right).$$

Define new parameters u, τ aiming for restriction $\|u\| \le r_0$, $\|\tau\| \le r_0$:

$$u = D_h(\theta - \theta_0), \quad \tau = \frac{r_0}{r_t}(t - t_0), \quad |t - t_0| \le r_t.$$

Estimation for r_t comes from equating parts of y gradients from each parameter.

$$\nabla_u y = D_h^{-1} \nabla^2 \zeta(\theta, t) D_h^{-1}, \quad \nabla_\tau y = \frac{r_t}{r_0} D_h^{-1} \nabla_t \nabla\zeta(\theta, t)$$

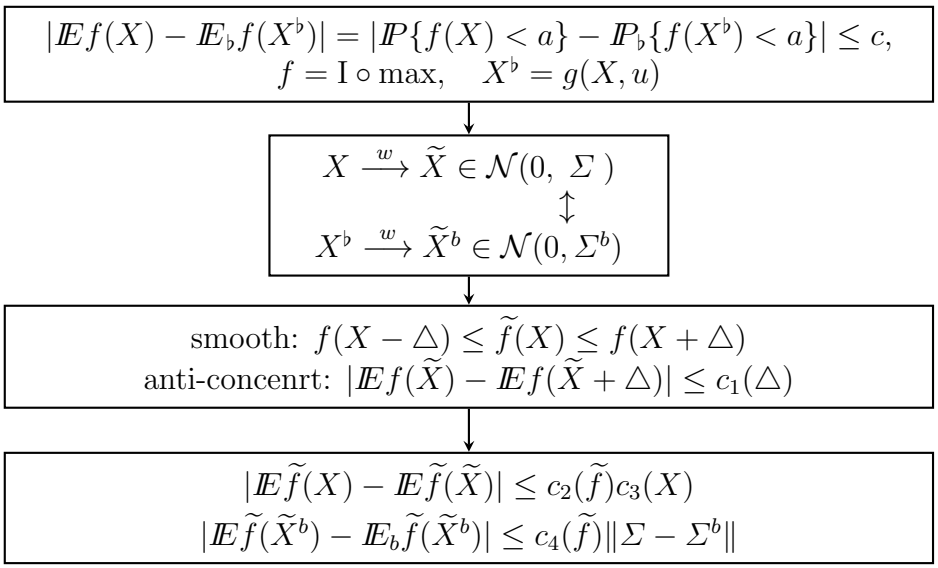
$$\|D_h^{-1} \nabla_t \nabla\zeta(\theta, t)\| \le 2\|D_h^{-1} \nabla l_i(\theta_i^*, t)\| + \|D_h^{-1} \nabla^2 \zeta_i(\cdot, t) D_h^{-1}\| r_0 \sim \sqrt{\frac{p}{h}}.$$

$$\|D_h^{-1} \nabla^2 \zeta(\theta, t) D_h^{-1}\| \sim \omega(h) z(x) \sim \sqrt{\frac{p}{h}}.$$

Finally, this approach results in $n - h = r_t \sim r_0$.

5 LRT Critical values calibration

Scheme of $I\!P$ measure and variable $f(X)$ (f is non random) approximation by corresponded bootstrap (empirical) measure $I\!P_\flat$ and variable $f(X^\flat)$ could be done in three steps: approximate $X, X^\flat$ by Normal distributions, approximate $f(\cdot)$ by smooth function, present measures distance as product of f properties and its arguments distribution parameters distance.



Weighted LRT statistic $T_h^\flat$ from multiplier bootstrap has analog of theorem 2 due to 4 and 3.

Theorem 5. Let (**??**) and (**??**) are true. Then with probability $1 - 16e^{-x}$ in local region $\Theta_1(r) \cap \Theta_2(r)$

$$\left| \sqrt{2T_h^\flat(\theta_1^\flat, \theta_2^\flat)} - \|\xi_{12}^\flat + \widehat{\theta}_{12}\| \right| \le 10 \diamondsuit^\flat(r, x),$$

where

$$\xi_{12}^\flat = \Sigma(D_2^{-1}\xi_2^\flat - D_1^{-1}\xi_1^\flat), \quad \widehat{\theta}_{12} = \Sigma(\widehat{\theta}_2 - \widehat{\theta}_1), \quad \theta^\flat = \operatorname*{argmax}_\theta L^\flat(\theta).$$

Remark. Generator of $\xi_{12}^\flat$ use statistic $T_h^\flat = \sqrt{2T_h(\theta_1^\flat, \theta_1^\flat + \widehat{\theta}_{12})}$, for which 5 theorem is also applicable with $\widehat{\theta}_{12} = 0$,

$$\left|T_h^\flat - \|\xi_{12}^\flat\|\right| \le 10 \diamondsuit^\flat(r, x).$$

Precision of distribution estimation $\xi_{12} + m$ using bootstrap (it's analog $\xi_{12}^\flat + m^\flat$, m, $m^\flat$ non random) cold be obtained by normal approximation for both variables. The sequence of comparison is

1. $\xi_{12}^\flat(\widehat{\theta}_1, \widehat{\theta}_2) \approx \xi_{12}^\flat(\theta_1^*, \theta_2^*)$;
2. $\|\xi_{12}^\flat(\theta_1^*, \theta_2^*) + m^\flat\| \approx \|\widetilde{\xi}^\flat\|$, $\widetilde{\xi}^\flat \sim \mathcal{N}(m^\flat, \Sigma^\flat)$;
3. $\|\xi_{12}(\theta_1^*, \theta_2^*) + m\| \approx \|\widetilde{\xi}\|$, $\widetilde{\xi} \sim \mathcal{N}(m, \Sigma)$;
4. $\|\widetilde{\xi}\| \approx \|\widetilde{\xi}^\flat\|$.

1) Compare $\xi_{12}^\flat(\widehat{\theta}_1, \widehat{\theta}_2)$ and $\xi_{12}^\flat(\theta_1^*, \theta_2^*)$. Due to

$$\xi_i^\flat(\theta) = D^{-1} \sum_{i=1}^{n} \nabla l_i(\theta)(u_i - 1),$$

and estimation for (Sb), the next inequality holds with probability $1 - 3e^x$, $i = \{1, 2\}$

$$\|\xi_i^\flat(\theta_2) - \xi_i^\flat(\theta_1)\| \le \left(12\nu_0 z(x)\omega r\sqrt{1+x} + \frac{1+\sqrt{2x}}{2n} C(r) r c_u\right) = \diamondsuit_\xi^\flat(r, x).$$

By definition

$$\xi_{12}^\flat = A \begin{pmatrix} \xi_1^\flat \\ \xi_2^\flat \end{pmatrix}, \quad A = \Sigma\left(-D_1^{-1}, D_2^{-1}\right),$$

which leads to statement

Lemma.

$$\|\xi_{12}^\flat(\widehat{\theta}_1, \widehat{\theta}_2) - \xi_{12}^\flat(\theta_1^*, \theta_2^*)\| \le 2\diamondsuit_{\xi,F}^\flat(r, x),$$

where $F = A^T A$, $H_2 = H_2(F) + 4p$,

$$\diamondsuit_{\xi,F}^\flat(r, x) = \left(12\nu_0 z_F(x)\omega r\sqrt{1+x} + \frac{1+\sqrt{2x}}{2n} C(r) r c_u\right),$$

$$z_F(x) = \ldots.$$

2),3) This step is weak approximation for $\|\xi_{12}(\theta_1^*, \theta_2^*) + m\| \approx \|\widetilde{\xi}\|$, $\widetilde{\xi} \sim \mathcal{N}(m, \Sigma)$. Namely, one need to compare functions of these norms (pointwise argument comparison for each $\|\xi_{12}(\theta_1^*, \theta_2^*) + m\|(t)$ has worse precision for its aggregations):

$$\max_\tau \sum_{t \in \tau} P(t)\|\xi_{12}(\theta_1^*, \theta_2^*) + m\|(t). \tag{1}$$

For non-limiting simplification assume that $m^\flat = m = 0$. Norm of a vector ξ could be presented as $\max_{\|\gamma\|=1} \gamma^T \xi$, and max as smooth max function:

$$h_\beta(x) = \beta^{-1} \log\left(\sum_i e^{\beta x_i}\right), \quad x = \{x_i\}.$$

After smoothing the considered expression (1) is

$$h_\beta\left(\sum_{t\in\tau} P(t)\gamma_t^T\xi_{12}(t)\right) = h_\beta\left(\sum_{t\in\tau} P(t)\gamma_t^T\sum_{i=1}^n w_i(t)\varepsilon_i\right) = h_\beta\left(\sum_{i=1}^n c(\tau)^T\varepsilon_i\right),$$

where $w_i(t)$ are kernel weights for window position t, $P(t)$ pattern values, ε_i – components of ξ_{12}. Function h_β corresponds to max over iterable arguments (τ, γ_τ).

Following property $h_\beta(x)$, $x \in \mathbb{R}^M$ characterise the error of smooth max approximation

$$\max_i(x_i) \le h_\beta(x) \le \max_i(x_i) + \frac{\log(M)}{\beta}.$$

Dimension includes $\gamma_t \in G_\varepsilon$, $t \in \{1, \ldots, n\}$ and τ, which results in $M < n|G_\varepsilon|$. The next lemma apply indicator for both parts of the above inequality.

Lemma. For function g_δ which is smooth indicator (g_δ grows from 0 to 1 inside interval of size δ) took place following inequality with $\delta = \beta^{-1}\log(M)$

$$g_\delta h_\beta(x-\delta) \le \mathbb{1}\left[\max_{0\le i\le M} x_i > 0\right] \le g_\delta h_\beta(x+\delta).$$

Consequently, under condition $|\mathbb{E}g_\delta h_\beta(x) - \mathbb{E}g_\delta h_\beta(\widetilde{x})| \le C(\beta, M)\mu_n$,

$$\left|\mathbb{P}\left(\max_{0\le i\le M} x_i > 0\right) - \mathbb{P}\left(\max_{0\le i\le M} \widetilde{x}_i > \pm 2\delta\right)\right| \le C(\delta, M)\mu_n. \tag{2}$$

The shift of length $\pm 2\delta$ could be moved outside applying anti-concentration lemma for distribution density $\max_i \widetilde{x}_i$.

Lemma. Let $x \in \mathcal{N}(m, \Sigma) \in \mathbb{R}^M$, $\sigma_1 \le \sqrt{\Sigma_{ii}} \le \sigma_2$, $a_M = \max_i(\widetilde{x}_i - m_i)/\sqrt{\Sigma_{ii}}$, then $\forall c$

$$\mathbb{P}(|\max_i \widetilde{x}_i - c| \le \varepsilon) \le C_{\text{ak}}(M, \Sigma),$$

$$C_{\text{ak}}(M,\Sigma) = \frac{4\varepsilon}{\sigma_1}\left(\frac{\sigma_2}{\sigma_1}a_M + \left(\frac{\sigma_2}{\sigma_1} - 1\right)\sqrt{2\log\left(\frac{\sigma_1}{\varepsilon}\right)} + 2 - \frac{\sigma_1}{\sigma_2}\right) \approx 4\varepsilon\frac{\sigma_2}{\sigma_1^2}\sqrt{2\log\left(\frac{\sigma_1 M}{\varepsilon}\right)}.$$

Combining this lemma and 2 one get result for measure difference of max of vector x and normal vector $\widetilde{x}$.

$$\left|\mathbb{P}\left(\max_{0\le i\le M} x_i > 0\right) - \mathbb{P}\left(\max_{0\le i\le M} \widetilde{x}_i > 0\right)\right| \le C(\delta, M)\mu_n + \delta C_{\text{ak}}(M, \Sigma). \tag{3}$$

The next required step is involved upper bound of obtaining

$$|\mathbb{E}g_\delta h_\beta(x) - \mathbb{E}g_\delta h_\beta(\widetilde{x})| \le C(\beta, M)\mu_n.$$

Lemma. Function $f = g_\delta h_\beta$, where g_δ has bounded third order derivations, $\delta = \beta^{-1}\log(M)$, has representation in Tailor form

$$\left|f(x+d) - f(x) - d^T f'(x) - d^T f''(x)d/2\right| \le C(\delta, M)\|d\|_\infty^3,$$

where

$$C(\delta, M) = \frac{1}{6\delta^3}(|g'''| + \log(M)|g''| + \log^2(M)|g'|).$$

Basing on this lemma use that

$$x_{\tau,\gamma} = \sum_{i=1}^n c_i(\tau)^T\varepsilon_i,$$

where $\mathrm{Var}\sum_i c_i(\tau)^T\varepsilon_i = \sum_i c_i(\tau)^T V_i^2 c_i(\tau)$.

The final statement for approximation of ε_i by normal variables $\widetilde{\varepsilon}$ with the same mean and variance is

$$\left| I\!\!E g_\delta h_\beta \left(\sum_{i=1}^n c(\tau)^T \varepsilon_i \right) - I\!\!E g_\delta h_\beta \left(\sum_{i=1}^n c(\tau)^T \widetilde{\varepsilon}_i \right) \right| \le C(\delta, M)\mu_n, \tag{4}$$

where

$$\mu_n = \max_\tau \sum_{i=1}^n (\|\varepsilon_i\|_\infty^3 + C\log^{3/2}(M)\sigma_i^3)\|c_i(\tau)\|^3,$$

$$\|c_i(\tau)\| \le \sum_{t\in\tau} P(t) w_i(t) \|\gamma_t^T\| \sim 1\!\!1(\tau - 2h \le i \le \tau + 2h), \quad \mu_n \sim \frac{1}{\sqrt{h}}.$$

2'),3') Consider the other type of smoothing for (1):

$$h_\beta \left(\sum_{i\in\tau} P(i) \left\| \xi_{12} \right\| (i) \right).$$

One have to estimate derivative

$$(g_\delta h_\beta\{\xi_{12} + td\})'''_t = g'_\delta h'''_\beta + 3g''_\delta h'_\beta h''_\beta + g'''_\delta (h'_\beta)^3.$$

Assume that approximately

$$\|\nabla_P h_\beta(P_1, \dots, P_n)\|_1 \le 1, \quad \left\|\nabla_P^2 h_\beta(P_1, \dots, P_n)\right\|_1 \le \beta, \quad \left\|\nabla_P^3 h_\beta(P_1, \dots, P_n)\right\|_1 \le \beta^2,$$

$$|g'_\delta|_\infty \le \frac{1}{\delta}, \quad |g''_\delta|_\infty \le \frac{1}{\delta^2}, \quad |g'''_\delta|_\infty \le \frac{1}{\delta^3}.$$

Also the estimation for $\max_\tau P'_\tau(t)$ is required, where $P_\tau(t) = \sum_{i\in\tau} P(i) \left\| \xi_{12} + td \right\| (i)$:

$$\max_\tau P'_\tau(t) \le \|d\| \max_\tau \sum_{i\in\tau} P(i).$$

Consider pattern weights with sum equal to 1, which leads to

$$(g_\delta h_\beta\{\xi_{12} + td\})'''_t \le \left(\frac{\beta^2}{\delta} + 3\frac{\beta}{\delta^2} + \frac{1}{\delta^3} \right) \|d\|^3.$$

4) Compare the same functions $I\!\!E g_\delta h_\beta$ from two normal vectors $X = \widetilde{\xi}$ and $Y = \widetilde{\xi}^\flat$ with different mean and variance.

Lemma. For X, Y independent normal vectors with m_X, m_Y, Σ_X, Σ_Y,

$$\triangle m = m_2 - m_1, \quad \triangle\Sigma = \Sigma_2 - \Sigma_1.$$

it holds

$$|I\!\!E g(\delta^{-1}h(X)) - I\!\!E g(\delta^{-1}h(Y))| \le \left(\frac{\beta\|g'\|_\infty}{\delta} + \frac{\|g''\|_\infty}{2\delta^2} \right) \|\triangle\Sigma\|_{\mathrm{op}} + \frac{\|g'\|_\infty}{2\delta} \|\triangle m\|_{\mathrm{op}}.$$

Following Section deals with estimation of $\|\triangle\Sigma\|_{\mathrm{op}}$.

6 Matrix deviations

6.1. Matrix inequality in change point statistic

A sequence of noise variables produced by different window positions.

$$\xi(t) = D_h^{-1}(\nabla L(\theta^*) - \nabla I\!E L(\theta^*)) = D_h^{-1}\sum_{i=t}^{t+h}\nabla l_i(\theta^*),$$

$$\xi^b(t) = D_h^{-1}(\nabla L(\widehat{\theta}) - \nabla I\!E_b L(\widehat{\theta})) = D_h^{-1}\sum_{i=t}^{t+h}\nabla l_i(\widehat{\theta})(u_i - 1).$$

Unite all variables involved into LRT statistic.

$$X = \xi_{12}(t)_{t=1}^n, \quad X^b = \xi_{12}^b(t)_{t=1}^n, \quad \xi_{12}(t) = \frac{1}{\sqrt{2}}(\xi(t+1) - \xi(t))$$

Consider difference in variance between X and X^b in case of regression model:

$$Y = \Psi_{[n\times p]}^T\theta + \varepsilon, \quad \varepsilon \in \mathcal{N}(0, \Sigma), \quad \Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)$$

Variables X and X^b cold be presented as linear transformation of a standart Normal vectors ε and ε^b correspondingly

$$\nabla l(Y_i) = \Psi_i\varepsilon_i$$

$$\nabla l^b(Y_i)(u_i - 1) = \Psi_i(Y - \Pi Y)_i(u_i - 1)\Psi_i\varepsilon_i^b, \quad \Pi = \Psi^T(\Psi\Psi^T)^{-1}\Psi$$

$$X = AHD_h^{-1}D(\Psi)\varepsilon = V\varepsilon, \quad X^b = AHD_h^{-1}D(\Psi)\varepsilon^b = V\varepsilon^b.$$

$$D(\Psi) = \operatorname{diag}(\Psi_1, \ldots, \Psi_n),$$

$$H_{np,np} = \begin{pmatrix} I_p & I_p & \cdots & I_p & 0 & 0\cdots & 0 \\ 0 & I_p & \cdots & I_p & I_p & 0\cdots & 0 \\ 0 & 0 & \cdots & I_p & I_p & I_p\cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & & \\ 0 & 0 & \cdots\cdots\cdots & 0 & 0 & & \end{pmatrix} \quad A_{np,np} = \begin{pmatrix} I_p & -I_p & \cdots & 0 & \\ 0 & I_p & -I_p & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0\cdots & 0 & I_p & -I_p & \end{pmatrix}$$

We are to compare S and $S^\flat$ by

$$\|S\|\|S^{-1/2}(S^\flat - S)S^{-1/2}\|$$

$$S = \operatorname{Var}(X) = V\Sigma V^T, \quad S^\flat = \operatorname{Var}_\flat(X^\flat) = V\operatorname{diag}((Y - \Pi Y)(Y - \Pi Y)^T)V^T$$

Involving variable $\mathcal{U}$ the goal estimation $(S^\flat - S)$ becomes

$$\mathcal{U} = S^{-1/2}V\Sigma^{1/2}, \quad \mathcal{U}\mathcal{U}^\top = I_{pn}$$

$$S^{-1/2}(S^\flat - S)S^{-1/2} = \mathcal{U}\Sigma^{-1/2}\operatorname{diag}((Y - \Pi Y)(Y - \Pi Y)^T)\Sigma^{-1/2}\mathcal{U}^\top - I_{np}$$

Variable $Y - \Pi Y$ has decomposition to bias and noise part

$$\Sigma^{-1/2}(Y - \Pi Y) = \Sigma^{-1/2}(f - \Pi f) + \Sigma^{-1/2}(\varepsilon - \Pi\varepsilon) = B + \widetilde{\varepsilon}$$

$$S^{-1/2}(S^\flat - S)S^{-1/2} = \\ = \mathcal{U}\operatorname{diag}(BB^\top)\mathcal{U}^\top + 2\mathcal{U}\operatorname{diag}(\widetilde{\varepsilon}B^\top)\mathcal{U}^\top + \\ + \mathcal{U}\operatorname{diag}(\widetilde{\varepsilon}\widetilde{\varepsilon}^\top - I\!\!E\widetilde{\varepsilon}\widetilde{\varepsilon}^\top)\mathcal{U}^\top + \mathcal{U}\operatorname{diag}(I\!\!E\widetilde{\varepsilon}\widetilde{\varepsilon}^\top)\mathcal{U}^\top - I_{np}.$$

The last component in this equation is a part of the non-random component, which becomes high in very large model, while B is small in large model.

Let $\operatorname{Var}(\widetilde{\varepsilon}) = R$ and $\|(S^{-1/2}V)_i\| \le \delta$, then from the consequence for theorem 13 with probability $1 - e^x$

$$\|\mathcal{U}\operatorname{diag}(\widetilde{\varepsilon}\widetilde{\varepsilon}^\top - I\!\!E\widetilde{\varepsilon}\widetilde{\varepsilon}^\top)\mathcal{U}^\top\|_{\mathrm{op}} \le \lambda_{\max}(R)(2\delta\sqrt{\mathtt{x}_{np}} + 2\delta^2\mathtt{x}_{np}),$$

where $\mathtt{x}_{np} = x + \log(np)$,

$$\|\mathcal{U}\operatorname{diag}(\widetilde{\varepsilon}B^\top)\mathcal{U}^\top\|_{\mathrm{op}} \le \sqrt{\lambda_{\max}(R)}\delta^2\|B\|\sqrt{2\mathtt{x}}.$$

The other two components are restricted by means of

Lemma.

$$\|\mathcal{U}A\mathcal{U}^\top\|_{\mathrm{op}} \le \min(\|A\|_{\mathrm{op}}, \delta^2\operatorname{tr}\{A\})$$

Proof.

$$\left\|\mathcal{U}A\mathcal{U}^\top\right\|_{\mathrm{op}} = \sup_{\|\gamma\|=1}\gamma^\top\mathcal{U}A\mathcal{U}^\top\gamma \le \sup_{\|\gamma\|=1,\max\gamma\le\delta}\gamma^\top A\gamma \le \min(\|A\|_{\mathrm{op}}, \delta^2\operatorname{tr}\{A\})$$

□

Applying the lemma one get

$$\left\|\mathcal{U}\operatorname{diag}(BB^\top)\mathcal{U}^\top\right\|_{\mathrm{op}} \le \min(\|B\|_{\mathrm{op}}^2, \delta^2\|B\|^2)$$

and

$$\left\|\mathcal{U}\operatorname{diag}(I\!\!E\widetilde{\varepsilon}\widetilde{\varepsilon}^\top)\mathcal{U}^\top - I_{np}\right\|_{\mathrm{op}} = \left\|\mathcal{U}(I_n - \Pi)^2 - I_n\mathcal{U}^\top\right\|_{\mathrm{op}} = \left\|\mathcal{U}\Pi\mathcal{U}^\top\right\|_{\mathrm{op}} \le 1.$$

Finally the upper bound for variance difference is

$$\lambda_{\max}(R)(2\delta\sqrt{\mathtt{x}_{np}} + 2\delta^2\mathtt{x}_{np}) + \sqrt{\lambda_{\max}(R)}\delta^2\|B\|\sqrt{2\mathtt{x}} + \min(\|B\|_{\mathrm{op}}^2, \delta^2\|B\|^2) + 1.$$

The last step is bound estimation for $\|S\|_{\mathrm{op}}$ which is

$$\|A\|_{\mathrm{op}}^2\|H\|_{\mathrm{op}}^2\left\|D_h^{-1}\right\|_{\mathrm{op}}^2\left\|D(\Psi)D(\Psi)^\top\right\|_{\mathrm{op}}$$

$$\left\|D(\Psi)D(\Psi)^\top\right\|_{\mathrm{op}} \le \max_i\left\|\Psi_i\Psi_i^\top\right\|_{\mathrm{op}}$$

Lemma. Spectrum of an upper triangular matrix A is its diagonal.

Proof. $\operatorname{Res}(A)$ set is equal to $\operatorname{Res}(\operatorname{diag}(A))$, because of $\operatorname{rank}(A - \lambda I) = \operatorname{rank}(\operatorname{diag}(A) - \lambda I)$. □

Correspondingly, $\|A\|_{\mathrm{op}}^2 = \|H\|_{\mathrm{op}}^2 = 1$. Bound for $\left\|D_h^{-1}\right\|_{\mathrm{op}}^2$ is $O(h^{-1})$.

6.2. Generalization to i.i.d model

Retreat regression model with an i.i.d model. In this case comparing variables would be

$$X = AHD_h^{-1}\begin{pmatrix}\nabla l_1(\theta_1^*)\\ \vdots \\ \nabla l_n(\theta_n^*)\end{pmatrix} = AHD_h^{-1}\nabla, \quad X = AHD_h^{-1}\begin{pmatrix}\nabla l_1(\widehat{\theta}_1)(u_1-1)\\ \vdots \\ \nabla l_n(\widehat{\theta}_n)(u_n-1)\end{pmatrix} = AHD_h^{-1}\nabla^\flat.$$

Here instead of noise variance one get block-diagonal one

$$\Sigma = \begin{pmatrix}\operatorname{Var}\nabla l_1(\theta_1^*) & \cdots & 0\\ \ddots & \ddots & \\ 0 & \cdots & \operatorname{Var}\nabla l_n(\theta_n^*)\end{pmatrix}$$

Analogically divide $\widehat{\nabla} = \nabla(\widehat{\theta})$ into mean and stochastic parts

$$\Sigma^{-1/2}\widehat{\nabla} = \Sigma^{-1/2}I\!\!E\widehat{\nabla} + \Sigma^{-1/2}(\widehat{\nabla} - I\!\!E\widehat{\nabla}) = B + \zeta.$$

Normalized variance difference term is

$$\begin{aligned}S^{-1/2}(S^\flat - S)S^{-1/2} = \\ &= \mathcal{U}\operatorname{blockDiag}(BB^\top)\mathcal{U}^\top + 2\mathcal{U}\operatorname{blockDiag}(\zeta B^\top)\mathcal{U}^\top + \\ &\quad + \mathcal{U}\operatorname{blockDiag}(\zeta\zeta^T - I\!\!E\zeta\zeta^T)\mathcal{U}^\top + \mathcal{U}\operatorname{blockDiag}(I\!\!E\zeta\zeta^T)\mathcal{U}^\top - I_{np}.\end{aligned}$$

Where blockDiag determines block diagonal matrix with block shape $[p\times p]$, corresponded to each component in sequence $\nabla l_1, \ldots, \nabla l_n$. Start consideration with term $\mathcal{U}\operatorname{blockDiag}(\zeta\zeta^T - I\!\!E\zeta\zeta^T)\mathcal{U}^\top$.

Lemma. Let $\{U_i\}$ be submatrices of matrix $\mathcal{U}$ corresponded to each block, such that $\mathcal{U} = (U_1, \ldots, U_n)$. Assume that

$$\left\|U_i^T U_i\right\|_{\mathrm{op}} \le \delta^2.$$

Than

$$\left\|(U_i^T U_i)^{-1}\log I\!\!E\exp\{(U_i^T U_i)^{1/2}(\zeta_i\zeta_i^T - I\!\!E\zeta_i\zeta_i^T)(U_i^T U_i)^{1/2}\}\right\|_{\mathrm{op}} \le \frac{\nu_l^2}{2}\delta^2.$$

Proof. Denote $\zeta_i\zeta_i^T - I\!\!E\zeta_i\zeta_i^T$ as A and $(U_i^T U_i)^{1/2}$ as U. With condition $I\!\!EA = 0$:

$$\left\|\log I\!\!Ee^{UAU}\right\|_{\mathrm{op}} = \left\|\left(\frac{1}{2}UI\!\!E[AU^2A]U + O((UA)^3)\right)\right\|_{\mathrm{op}}.$$

With condition $\|UA\|_{\mathrm{op}} < 1/2$ (with high probability):

$$\left\|U^{-2}\log I\!\!Ee^{UAU}\right\|_{\mathrm{op}} \le \left\|U^2\right\|_{\mathrm{op}}\left\|I\!\!EA^2\right\|_{\mathrm{op}}. \tag{5}$$

$$\left\|I\!\!E(\zeta_i\zeta_i^T - I\!\!E\zeta_i\zeta_i^T)^2\right\|_{\mathrm{op}} = \left\|I\!\!E\|\zeta_i\|^2\zeta_i\zeta_i^T - (I\!\!E\zeta_i\zeta_i^T)^2\right\|_{\mathrm{op}},$$

where

$$\|\zeta_i\|^2 \sim p, \quad \left\|I\!\!E\zeta_i\zeta_i^T\right\|_{\mathrm{op}} = \left\|\Sigma_i^{-1/2}\Sigma_i(\widehat{\theta}_i)\Sigma_i^{-1/2} - I_p + I_p\right\|_{\mathrm{op}} \le 1 + \delta_\Sigma(r).$$

Finally, $\nu_l^2 \sim 2p(1+\delta_\Sigma(r))$. □

From proof for theorem 13

$$I\!\!P\left(\left\|\mathcal{U}\operatorname{diag}(\zeta\zeta^T - I\!\!E\zeta\zeta^T)\mathcal{U}^\top\right\|_{\mathrm{op}} > t\right) \le 2np\inf_\theta \exp\left\{-\theta t + \frac{\delta^2\theta^2\nu_l^2}{2}\right\},$$

Subsequently, with $t = \sqrt{2x_{np}}\delta\nu_l$, $x_{np} = x + \log(2np)$:

$$I\!P\left(\left\|\mathcal{U}\operatorname{diag}(\zeta\zeta^T - I\!E\zeta\zeta^T)\mathcal{U}^\top\right\|_{\mathrm{op}} > t\right) \le e^x.$$

For the other term $\mathcal{U}\,\mathrm{blockDiag}(\zeta B^\top)\mathcal{U}^\top$ by means of 5 one get with $\|\zeta_i\|^2 \sim p$

$$\left\|U^{-2}\log I\!E\exp\{U\zeta_i B_i^T U\}\right\|_{\mathrm{op}} \le p\left\|B_i\right\|^2\delta^2.$$

Correspondingly, from proof for theorem 13

$$I\!P\left(\left\|\mathcal{U}\operatorname{diag}(\zeta B^T)\mathcal{U}^\top\right\|_{\mathrm{op}} > t\right) \le 2np\inf_\theta \exp\left\{-\theta t + p\max_i\left\|B_i\right\|^2\delta^2\right\},$$

Finally, with $t = 2\sqrt{x_{np}p}\delta\max_i\|B_i\|$, $x_{np} = x + \log(2np)$:

$$I\!P\left(\left\|\mathcal{U}\operatorname{diag}(\zeta B^T)\mathcal{U}^\top\right\|_{\mathrm{op}} > t\right) \le e^x.$$

7 Max-norm Gaussian

Comparison of $\max_t\langle P,(\|X_1\|,\dots,\|X_n\|)\rangle(t)$ and $\max_t\langle P,(\|X_1^b\|,\dots,\|X_n^b\|)\rangle(t)$ could be done by Slepian inequality and replacing norm with max-product function $(\max_t P\max_{\gamma_1\dots\gamma_n}\Gamma X)$, where

$$\Gamma = \operatorname{diag}(\gamma_1^T\dots\gamma_n^T),\quad P = \operatorname{diag}(P_h^T\dots P_h^T)$$

Lemma.

$$\max_{\gamma\in G_\varepsilon}\gamma^T S \le \|S\| \le \frac{1}{1-\varepsilon}\max_{\gamma\in G_\varepsilon}\gamma^T S,$$

where $|G_\varepsilon| \le (2/\varepsilon)^{p-1}$, $S(t)\in I\!R^p$, $1\le t\le M$ By anti-concentration lemma one could obtain inequality for measures

$$\left|I\!P(\max_t\|S(t)\| < q) - I\!P(\max_t\max_{\gamma\in G_\varepsilon}\gamma^T S(t) < q)\right| \le 8\sqrt{p}\varepsilon q\frac{\sigma_2}{\sigma_1^2}\sqrt{\log\left(M\frac{2}{\varepsilon}\right)}$$

Since max_{γ_i} functions are independent the next theorem take place with $|Gh_\varepsilon| \le h(2/\varepsilon)^{p-1}$

Theorem 6.

$$\left|I\!P(\max_t\langle P,(\|X_1\|,\dots,\|X_n\|)\rangle(t) < q) - I\!P(\max_t\max_{\gamma_1\dots\gamma_n} P\Gamma X < q)\right| \le 8\sqrt{p}\varepsilon q\frac{\sigma_2}{\sigma_1^2}\sqrt{\log\left(\frac{2nh}{\varepsilon}\right)}$$

Theorem 7.

$$I\!P(\max_t\max_{\gamma_1\dots\gamma_n} P\Gamma X < q) - I\!P(\max_t\max_{\gamma_1\dots\gamma_n} P\Gamma X^b < q) \le$$

$$\le \left(\frac{\beta\|g'\|_\infty}{\delta} + \frac{\|g''\|_\infty}{2\delta^2}\right)\|\triangle P\Gamma\Sigma\Gamma^T P^T\|_{\mathrm{op}} + 8\sqrt{p}\delta\frac{\sigma_2}{\sigma_1^2}\sqrt{\log\left(\frac{2nh}{\varepsilon}\right)}$$

8 Appendix