Z-process 法による変化点問題の研究

西山　陽一　　数理・推論研究系　准教授

1　序

大学院時代の友人が PMDA で統計解析の仕事をしています.
　『安全部に移ってから，仕事で変化点問題を扱っていて，

1. ある特定の薬剤でのある特定の種類の副作用の発生について，医療機関からの報告を随時受け付ける;
2. 毎週の副作用発生数が強度 λ のポアソン分布に従うと仮定する；
3. ある週以降に副作用の発生数が増加している（λが増加方向に変化した）と判断されたら,「何かまずいことが生じているかもしれない」というアラートを発して医学的な検討を開始する；

というシステムを作ろうとしています.』
　『副作用の発生時刻の間隔の長さが指数分布に従うとし，ある時点からその平均が小さくなったかどうかを検定するというアプローチも考えています.』

2　パラメトリック変化点問題

測度空間 $(\mathcal{X}, \mathcal{A}, \mu)$，確率密度 $f(x,\theta)$, $\theta \in \Theta \subset \mathbb{R}^d$. 独立データ $\{X_i\}_{i=1,...,n}$ について：

　H_0：全ての $i = 1, ..., n$ について $\theta = \theta_0$.
　H_1：ある異なる θ_0, θ_1 と，ある $u_* \in (0,1)$ が存在して，$i \le u_* n$ に対しては $\theta = \theta_0$，　$i > u_* n$ に対しては $\theta = \theta_1$.

　この問題へのアプローチは，大別して

- **尤度比の方法**（通常の「尤度比統計量」に対応）
- **Z-process 法**（通常の「Rao 統計量」に対応）
- **逐次 MLE 法**（通常の「Wald 統計量」に対応）

がある.

2.1　「Z-process 法」について

$l(x,\theta) = \log f(x,\theta)$ とおいて

$$U_{n,k} = \frac{1}{n}\sum_{i=1}^{k} \dot{l}(X_i, \widehat{\theta}_n)$$
$$\widehat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \dot{l}(X, \widehat{\theta}_n)\dot{l}(X_i, \widehat{\theta}_n)^\top$$

（ただし $\widehat{\theta}_n$ は $X_1, ..., X_n$ に基づく最尤推定量）と定義して

$$\mathcal{S}_n = n \max_{1\le k\le n} U_{n,k}^\top \widehat{I}_n^{-1} U_{n,k}$$

を提案すると，帰無仮説 H_0 のもとで

$$\mathcal{S}_n \to^d \sup_{u\in[0,1]} \sum_{p=1}^{d} |B^{\circ,p}(u)|^2.$$

ただし $B^{\circ,1}, ..., B^{\circ,d}$ は独立な標準ブラウン橋.
　対立仮説 H_1 のもとでは，記号

$$I_{\theta_0}(\theta) := E_{\theta_0}\dot{l}(X,\theta)\dot{l}(X,\theta)^\top$$
$$u_* E_{\theta_0}\dot{l}(X,\theta) + (1-u_*)E_{\theta_1}\dot{l}(X,\theta) = 0 \quad \text{の解 } \theta = \theta_*$$

を導入すると，

$$2\mathcal{S}_n \ge n(v_*^\top I_*^{-1} v_* - o_P(1)) - O_P(1)$$

が成り立つ．ただし

$$I_* = u_* I_{\theta_0}(\theta_*) + (1-u_*) I_{\theta_1}(\theta_*)$$

は正定値行列であり，

$$v_* = u_*(1-u_*)(E_{\theta_0}\dot{l}(X,\theta_*) - E_{\theta_1}\dot{l}(X,\theta_*))$$

はゼロではないベクトルである．従って，検定は一致性もつ.

2.2　確率過程への拡張の一例

拡散過程モデル

$$X_t = X_0 + \int_0^t S(X_s,\theta)ds + \int_0^t \sigma(X_s)dW_s$$

を考える.

$$U_{T,t} = \frac{1}{T}\int_0^t \frac{\dot{S}(X_s,\theta)}{\sigma(X_s)^2}(dX_s - S(X_s,\theta)ds)\Big|_{\theta=\widehat{\theta}_T},$$

$$\widehat{I}_T = \frac{1}{T}\int_0^T \frac{\dot{S}(X_s,\widehat{\theta}_T)\dot{S}(X_s,\widehat{\theta}_T)^\top}{\sigma(X_s)^2}ds,$$

$$\mathcal{S}_T = T \sup_{0\le t\le T} U_{T,t}^\top \widehat{I}_T^{-1} U_{T,t}.$$

とおくと，これに対して，独立データのときと同様の結論が得られる.

3　ノンパラメトリック変化点問題

1 次元の独立データ $\{X_i\}_{i=1,...,n}$ について：

　H_0： 全ての $i = 1, ..., n$ について，同一の連続分布 F に従う.
　H_1：ある $u_* \in (0,1)$ が存在して，$i \le u_* n$ は F_0 に従い，$i > u_* n$ は F_1 に従う．ただし，ある $x \in \mathbb{R}$ に対して $F_0(x) \ne F_1(x)$.

　$\{X_i\}_{i=1,...,n}$ の順位統計量を $\{R_i\}_{i=1,...,n}$ とし，次を提案する：

$$D_n = \sqrt{n} \max_{1\le i,j\le n} \left| \frac{1}{n}\sum_{q=1}^{n} 1\{q \le i,\ R_q \le j\} - \frac{ij}{n^2} \right|.$$

　H_0 のもとで,

$$D_n \to^d \sup_{0\le s,t\le 1} |B^\circ(s,t)|,$$

ただし B° は平均ゼロの正規確率場であって共分散構造が

$$EB^\circ(s_1,t_1)B^\circ(s_2,t_2) = (s_1 \wedge s_2 - s_1 s_2)(t_1 \wedge t_2 - t_1 t_2)$$

であるもの（standard Brownian pillow と呼ばれています）.
　H_1 のもとで,

$$D_n \ge \sqrt{n}\left\{ u_*(1-u_*) \sup_x |F_0(x) - F_1(x)| - o(1) \right\} - O_P(1).$$

　重要な注意：我々の統計量は，H_0 のもとで，漸近的分布不変であるだけでなく，n を止めても分布不変である.