

# パラメータの変化点検出問題に対するベイズ決定理論的アプローチ

1G00H105-2 堀井 俊佑

指導 松嶋 敏泰 教授

## 1 序論

データが発生している母集団がいくつかの水準に分かれ、その母集団の間にパラメータの変化がいくつか存在するモデルを仮定する．本研究では、観測したデータをもとに変化の位置を検出する問題を扱う．このような問題の例としては、薬品の投与量に対して、投与量を何段階かに分けて、その薬の効果の変化を解析するような問題が挙げられる．

この問題に対し、各水準間での検定を複数回行う方法として多重比較法がある [1]．多重比較法は検定の多重性を考慮し、仮説全体としての第 1 種の過誤の確率を有意水準以下に抑えつつ、検定を行う方法である．しかし、この手法では水準の数が多くなると検出力が著しく落ちるという問題点があった．

この問題点を受けて、パラメータの変化点と各水準のパラメータの組を一つの確率モデルと考え、モデル選択基準により一つのモデルを選び、変化点を検出する手法が研究されている．その中で、各モデル間の変化回数の差を損失関数として定義し、その損失をベイズ基準のもとで最小化するモデル選択方法が提案されている [2]．この手法では、データの数がある有限の時に、上記の損失に対してベイズ最適であるということが保証されるが、変化点の検出問題として、確率モデルに損失を置くことは必ずしも妥当ではないと考えられる．

本研究では、実際には無い変化点を誤って検出、もしくは本来ある変化点を見逃してしまうことに対して損失を定義し、その損失をベイズ基準のもとで最小化する変化点検出方法を提案する．更にシミュレーションにより、多重比較法・モデル選択手法と比較し、提案手法の性能評価を行う．

## 2 問題設定

$r$  個の水準を  $C_1, C_2, \dots, C_r$  とする．第  $i$  水準のデータ  $x^{n_i} = x_{i1}, \dots, x_{in_i}$  は互いに独立に  $p(x^{n_i}|\theta_i)$  に従っていると仮定する． $s_i$  を水準  $i$  と  $i+1$  の間で変化があれば 1、なければ 0 をとるインデックスとする．全データ  $x^n = (x^{n_1}, \dots, x^{n_r})$  の確率密度関数は  $\prod_{i=1}^{r-1} p(x^{n_i}, x^{n_{i+1}}|\theta_i, \theta_{i+1}, s_i)$  と表すことができる．

## 3 モデル選択を用いた変化点検出法

$s_1$  から  $s_{r-1}$  の組み合わせをモデルとし、これを  $m$  で表す．すなわち  $m = \{s_1, \dots, s_{r-1}\}$  である．モデル選択は、複数のモデルの中から、一つのモデル  $A_m$  を決定する問題として考えられる．

従来手法では損失関数を、決定するモデル  $A_m$  に対し定義し、以下のように設定する．

$$L_m(m : A_m) = \begin{cases} 0 & \text{if } A_m = m, \\ a_0 & \text{if } A_m \neq m, k_{A_m} = k_m, \\ a_1(k_m - k_{A_m}) & \text{if } k_{A_m} < k_m, \\ a_2(k_{A_m} - k_m) & \text{if } k_{A_m} > k_m. \end{cases} \quad (1)$$

ここで、 $k_m$  はモデル  $m$  における変化の回数を表す．

損失関数は確率変数の実現値  $x^n$  により左右するため、決定理論の枠組みでは真の分布による期待値をとった危険関数により評価する．

$$R_m(A_m) = \int_{X^n} L(m : A_m) p(x^n|m, \theta) dx^n. \quad (2)$$

現実には、真の  $m, \theta$  は事前には未知であり、この危険関数を最小化することはできない．そこで、以下のベイズ危険関数を定義する．

$$BR_m(A_m) = \sum_m \int_{X^n} \int_{\Theta} L(m : A_m) p(x^n|m, \theta) \times p(\theta|m) p(m) dx^n d\theta. \quad (3)$$

ベイズ危険関数を最小にする決定は、以下の条件付ベイズ危険を最小化する決定と等価である．

$$BL_m(A_m) = \sum_m L(m : A_m) p(m|x^n). \quad (4)$$

ここで、

$$p(m|x^n) = \frac{p(x^n|m) p(m)}{\sum_m p(x^n|m) p(m)}. \quad (5)$$

$$p(x^n|m) = \int_{\Theta} p(x^n|m, \theta) p(\theta|m) d\theta. \quad (6)$$

である．

ベイズ最適なモデル  $\hat{m}$  は以下で与えられる．

$$\hat{m} = \arg \min_{A_m} \sum_m L_m(A_m, m) p(m|x^n). \quad (7)$$

## 4 提案法

本来知りたいのはパラメータの変化点であり，変化の回数をモデル間の損失としたモデル選択は，この目的に対して適切であるとは言えない．

そこで，本研究では変化の有無の決定  $As_i$  に対して，以下のような損失を定義する．

$$l(s_i : As_i) = \begin{cases} 0 & \text{if } As_i = s_i, \\ b_1 & \text{if } As_i = 1, s_i = 0, \\ b_2 & \text{if } As_i = 0, s_i = 1. \end{cases} \quad (8)$$

全ての水準間における変化の有無を決定することに対する累積損失を以下で定義する．

$$L_s(s^r : As^r) = \sum_i l(s_i : As_i). \quad (9)$$

ここで， $s^r = (s_i)_{i=1}^{r-1}$  である．

(9) 式に対するベイズ危険関数は以下で与えられる．

$$BR(As^r) = \sum_i \sum_{s_i} \int_{X^n} \int_{\Theta} l(s_i : As_i) p(x^{n_i}, x^{n_{i+1}} | s_i, \theta_i, \theta_{i+1}) \times p(\theta_i, \theta_{i+1} | x^{n_i}, x^{n_{i+1}}) p(s_i) dx^{n_i} dx^{n_{i+1}} d\theta_i d\theta_{i+1}. \quad (10)$$

このとき，ベイズ最適な決定は， $s^r$  の各  $s_i$  に対して，以下のような決定となる．

$$\hat{s}_i = \arg \min_{As_i} \sum_{s_i} L(s_i : As_i) p(s_i | x^{n_i}, x^{n_{i+1}}). \quad (11)$$

## 5 シミュレーションによる評価

### 5.1 シミュレーション目的

#### 実験．1

(9) 式の累積損失に関し，多重比較法（テューキーの方法）・モデル選択・提案手法に関して比較し，提案手法が他の手法に比べ，平均的に累積損失を小さくしているかどうかを検証する．

#### 実験．2

(9) 式以外の評価基準で提案手法の性能を評価するため，全ての変化を見逃すことなく検出できる確率である，検出力で比較を行う．

### 5.2 シミュレーション条件

#### 実験．1

$r = 4, 5, 6$  の場合で，各水準のデータの数はいずれも 10 とした．分布としては正規分布を仮定し， $p(s_i)$  は全て 0.5 とした．各水準の平均値，分散はいずれも 0 から 1 の間の一様乱数に従って発生させた．モデル選択と提案手法における損失はそれぞれ  $a_0 = a_1 = a_2 = 1$ ， $b_1 = b_2 = 1$  とし，実験回数は 100000 回とした．

#### 実験．2

実験 1 と同様の環境で実験を行った．ただし，モデル

選択・提案手法ともに，多重比較法のように第 1 種の過誤の確率を 5%以下に抑えるという明確な基準が無いため，事前実験で，この値を 5%以下に抑えられる損失の設定を探し，その損失のもとで実験を行った．

### 5.3 シミュレーション結果

表 1: 実験 1(平均累積損失の比較)

|         | 多重比較   | モデル選択  | 提案手法   |
|---------|--------|--------|--------|
| 4 水準の場合 | 1.0273 | 0.9373 | 0.9216 |
| 5 水準の場合 | 1.4291 | 1.2649 | 1.2357 |
| 6 水準の場合 | 1.8028 | 1.5809 | 1.5191 |

表 2: 実験 2(検出力の比較)

|         | 多重比較   | モデル選択  | 提案手法   |
|---------|--------|--------|--------|
| 4 水準の場合 | 0.3362 | 0.3634 | 0.3752 |
| 5 水準の場合 | 0.2299 | 0.2523 | 0.2629 |
| 6 水準の場合 | 0.1787 | 0.1819 | 0.1959 |

## 6 考察

表 1 より，提案手法は全ての水準数において平均損失が最小となっていることが確認できる．

検出力に関しても，モデル選択よりも提案手法のほうが優れている．同様の事前分布を仮定しているにも関わらず，検出力が上がるということは，パラメータの変化点検出という問題に対する損失の設定法として，従来のモデル選択よりも，提案手法による損失の設定法が妥当であることを表していると考えられる．

## 7 結論

変化点の検出問題として，変化の有無に関して損失を設定し，ベイズ最適となる決定方法を提案した．またシミュレーションにより，モデル選択を行った場合より検出力が高くなることを示した．より一般的な問題を考えると，水準間のパラメータに自然な順序が仮定できる場合がある．例えば，薬効問題において，投与量を増やすことにより効果は増えることはあっても減ることは無いという状況である．このような場合には，ベイズ最適な決定が解析的には出来なくなる．この問題に関しては今後の課題としたい．

## 参考文献

- [1] Y.Hochberg and A.C.Tamhane(1987),“ Multiple Comparison Procedures, ”John Wiley & Sons.
- [2] M.Gotoh, T.Matsushima, and S.Hirasawa, “ Statistical Model Selection based on Bayes Decision Theory and Its Application to Change Detection Problems, ”International Journal of Production Economics, Vol.60-61,1999,pp629-638.