Multiple Testing in Change-Point Problem with Application to Safety Signal Detection

Jie Chen

Merck Research Laboratories

E-Mail: jie_chen@merck.com

The 5th International Conference on Multiple Comparison Procedures

Vienna, Austria

July 11, 2007

Overview

Introduction (1)

- Change-point analysis concerns with the inference on the point(s) in a sequence of random process at which the distribution changes
- Applications
  - Statistical quality control – Online detection of changes in quality operations
  - Public Health – Sequentially monitoring the number of cases of a disease for potential outbreak
  - Medicine – Post-marketing surveillance, dose-finding
  - Biomedical signal processing – Online detection of biomedical signals such as Eletroencephalogram (EEG) and electrocardiogram (ECG)
  - Meteorology – Global warming
  - Finance – Detection of business cycles
  - Seismology

Introduction (2)

Classifications of change-point analysis

- Continuous versus discrete
- Retrospective (fixed sample) versus prospective (online sequential)
- Parametric versus non-parametric
- Frequentist versus Bayesian
- One change-point versus multiple change-points

Introduction (3)

Statistical methods for change-point analysis

- Likelihood ratio procedures for parametric models
- Non-parametric methods – Mann-Whitney U test, Wilcoxon rank test
- Regression-based methods (including curve fitting)
- Cumulative sum (CUSUM) methods
- Bayesian analysis and its variations
- Sequential methods
- information criterion
- Wavelet transformation

One Change-Point Problem (1)

- Let $(X_1, \ldots, X_T)$ be a sequence of independent random variables, ordered in time interval, each with density function $f(X_i|\mu_i)$ where $E(X_i) = \mu_i, i = 1, \ldots, T$
- Consider the model for one change-point in means at time interval τ
- The null hypothesis

$$H_0 : \mu_1 = \ldots = \mu_T$$

against against

$$H_1 : \mu_1 = \ldots = \mu_\tau \neq \mu_{\tau+1} = \ldots = \mu_T$$

for an unknown τ

One Change-Point Problem (2)

- Let $\mu_1 = \ldots = \mu_\tau = \mu_0'$ with known μ_0' before the change-point τ and $\mu_{\tau+1} = \ldots = \mu_T = \mu_1'$ with known μ_1' after the change-point. Then the log likelihood function is

$$\ell(\tau) = \sum_{i=1}^{\tau} \log f(X_i|\mu_0') + \sum_{i=\tau+1}^{T} \log f(X_i|\mu_1') \quad (1)$$

- The log likelihood ratio test statistic for testing H_0 against H_1 is

$$\log \ell(\tau) = \sum_{i=\tau+1}^{T} \log f(X_i|\mu_0') - \sum_{i=\tau+1}^{T} \log f(X_i|\mu_1') \quad (2)$$

One Change-Point Problem (3)

- Traditionally, the null hypothesis H_0 of no change-point against H_1 of one change-point over the T time intervals is rejected if

$$2 \sup_{\tau} \log \ell(\tau) > \chi^2_{\alpha,1} \tag{3}$$

- The MLE $\hat{\tau}$ of τ is obtained by maximizing (1)
- When μ_0' and μ_1' are unknown, the MLE's $\hat{\mu}_0'$, $\hat{\mu}_1'$ and $\hat{\tau}$ can be obtained by simultaneously maximizing (1) w.r.t. μ_0', μ_1' and τ

Multiple Change-Points (1)

- The test continues sequentially for testing H_{m-1} of $m-1$ change-points against H_m of m change-points, $m = 1, \ldots, T-1$, until an acceptance occurs
- That is, the test starts from $m = 0$ (against $m = 1$) towards $m = T - 1$

$$
\begin{aligned}
H_0 : \quad & \mu_1 = \ldots = \mu_T \\
H_1 : \quad & \mu_1 = \ldots = \mu_\tau \neq \mu_{\tau+1} = \ldots = \mu_T \\
H_2 : \quad & \mu_1 = \ldots = \mu_{\tau_1} \neq \mu_{\tau_1+1} = \\
& \qquad \ldots = \mu_{\tau_2} \neq \mu_{\tau_2+1} = \ldots = \mu_T \\
\vdots \quad & \quad \vdots \\
H_{T-1} : \quad & \mu_1 \neq \ldots \neq \mu_T
\end{aligned}
$$

Multiple Change-Points (2)

- Binary segmentation (Vostrikova 1981)
  - Test for no change point H_0 against one change-point H_1
  - If H_0 is rejected, test the two subsequences before and after the change-point identified in the above step separately for a change
  - Repeat the process until no change-points are found in any of the subsequences
  - The collection of change-points identified from the above steps are $\{\hat{\tau}_1, \ldots, \hat{\tau}_k\}$ and the estimated number of change-points is then k

Multiple Change-Points (3)

A hierarchy of sub-hypothesis tests (Hogg 1961)

- Let Ω denote the total parameter space and Ω^* a subspace of Ω.
- It is desired to test $H_0' : \theta \in \Omega^*$ against $H_1' : \theta \in \Omega - \Omega^*$.
- Suppose there are certain intermediate hypotheses. Let Ω_i be a subset of Ω_{i-1}, $i = 1, \ldots, t-1$, such that

$$\Omega = \Omega_0 \supset \Omega_1 \supset \ldots \supset \Omega_{t-1} = \Omega^*$$

  where each Ω_i corresponds to an intermediate hypothesis

- Testing H_0' against H_1' can be carried out by iteratively testing the following hypotheses:

$$H_0^i : \theta \in \Omega_i \text{ versus } H_1^i : \theta \in \Omega_{i-1} - \Omega_i,$$

  $i = 1, \ldots, t-1$

Multiple Change-Points (4)

A hierarchy of sub-hypothesis tests (cont'd)

- To test H_0' against H_1', we first test

$$H_0^1 : \theta \in \Omega_1 \text{ against } H_1^1 : \theta \in \Omega_0 - \Omega_1$$

- If H_0^1 is accepted, we then test

$$H_0^2 : \theta \in \Omega_2 \text{ against } H_1^2 : \theta \in \Omega_1 - \Omega_2$$

- In general, if H_0^{i-1} is accepted, we continue to test

$$H_0^i : \theta \in \Omega_i \text{ against } H_1^i : \theta \in \Omega_{i-1} - \Omega_i$$

- H_0' is rejected if any one of $H_0^1, \ldots, H_0^{t-1}$ is rejected
- H_0' is accepted if and only if all of $H_0^1, \ldots, H_0^{t-1}$ are accepted

Multiple Testing in Change-Point Analysis (1)

- "If you torture the data long enough, it will confess anything you want" – Nobel Laureate Ronald Coase
- Generalized likelihood ratio (GLR) test
  - Let $\lambda_i = L(\hat{\Omega}_i)/L(\hat{\Omega}_{i-1})$ be the likelihood ratio for testing H_0^i against H_1^i, $i = 1, \ldots, t-1$
  - The GLR for H_0' against H_1' is given by

$$\lambda = \frac{L(\hat{\Omega}_{t-1})}{L(\hat{\Omega}_0)} = \prod_{i=1}^{t-1} \frac{L(\hat{\Omega}_i)}{L(\hat{\Omega}_{i-1})} = \prod_{i=1}^{t-1} \lambda_i \qquad (4)$$

  - The λ_i's are mutually stochastically independent test statistics
  - Significance level for each test $\alpha_i = 1 - (1-\alpha)^{1/(t-1)}$, where α is the family-wise type I error rate

Multiple Testing in Change-Point Analysis (2)

Closure principle

- Suppose $T = 4$, then there exist at most $m = 3$ change-points. Reformulation of the H_i's

$$\begin{aligned} H_0 :& \quad \{\mu_1 = \mu_2 = \mu_3 = \mu_4\} \\ H_1 :& \quad \{\mu_1 = \mu_2 = \mu_3\}, \{\mu_2 = \mu_3 = \mu_4\}, \\ & \quad \{\mu_1 = \mu_2\} \cap \{\mu_3 = \mu_4\} \\ H_2 :& \quad \{\mu_1 = \mu_2\}, \{\mu_2 = \mu_3\}, \{\mu_3 = \mu_4\} \\ H_3 :& \quad \{\mu_1\}, \{\mu_2\}, \{\mu_3\}, \{\mu_4\} \end{aligned}$$

- This forms the closure of the family by taking all possible intersections
- The closed family resembles the one that is formed for all pair-wise comparisons, but much smaller; it consists of hypotheses of homogeneity of successive means and their intersections

Multiple Testing in Change-Point Analysis (3)

Closed test procedure for dose-response (Marcus, et al, 1976; Begun and Gabriel, 1981; Rom, et al, 1994) – Let k denote the cardinality of a subset homogeneity hypothesis H_K

- Reject a subset homogeneity hypothesis H_K at level $\alpha_k = \alpha k/t$
- Retain H_K at level α
- Otherwise, if H_K is rejected at level α but not at level α_k, then H_K is rejected if every hypothesis H_R that concerns means in the complement of K is rejected at level α_r

Multiple Testing in Change-Point Analysis (4)

Proposed modification of the above closed test procedure

- Reject H_0^i
  - If Ω_i is not contained in any accepted set, and
  - If H_0^i is rejected at level $\alpha_i = 1 - (1-\alpha)^{(t-1-i)/(t-1)}$
- Retain H_0^i
  - If Ω_i is contained in another accepted set, or
  - If H_0^i is not rejected at level $\alpha_i = 1 - (1-\alpha)^{(t-1-i)/(t-1)}$

Two-Sequence Change-Point Problem

- Suppose that there are two independent sequences of random processes X_{i1} and X_{i2}, with $X_{ij} \sim f(\mu_{ij})$, $i = 1, \ldots, T$ and $j = 1, 2$
- The question of interest is whether there are an abrupt change in the ratios of the two random variables across the time period
- Let $\gamma_i = \mu_{i1}/\mu_{i2}$, $i = 1, \ldots, T$. Then this is equivalent to simultaneously testing the null hypotheses

$$H_0 : \gamma_1 = \ldots = \gamma_T$$

against H_1 : there is at least one change point

An Example (1)

Pharmacovigilance and Post-Marketing Safety Surveillance

- An integrated part of biopharmaceuticcal development
- Activities involve in detection, assessment, understanding and prevention of adverse effects and any drug-related problems
- Pharmacovigilance plan (PvP)
  - As part of Marketing Authorization Application, the PvP must be prepared in compliance with regulatory request on potential safety impact of product modification
  - The PvP describes routine pharmacovigilance practice, as well as special action plan including Post-marketing Safety Surveillance Analysis (PSSA)
  - The PSSA will evaluate all potential safety signals, with special attention to proportional change of a particular adverse event (system) over time

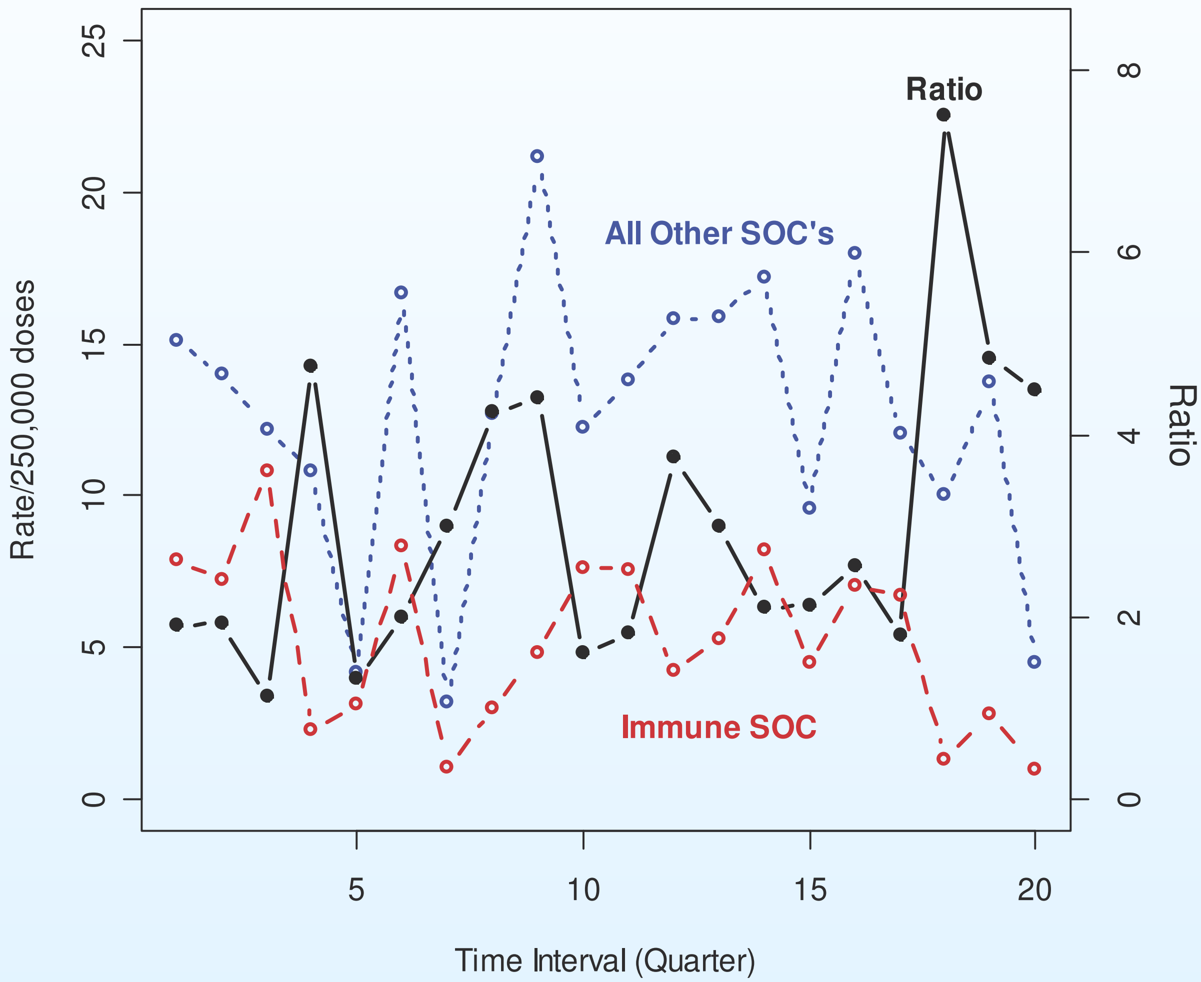
An Example (2)

PSSA

- Uses spontaneous reporting databases (AERS, VAERS, company's spontaneous reporting databases, etc.) and other epidemiological studies
- Data mining tools – Empirical Bayes, neural network, etc.
- Proportional change over time – useful to detect the impact of drug modification on the reporting of a particular (body system) adverse experience. For example,
  - Name change
  - Combination of two or more independent vaccines
  - New technology $\Rightarrow$ manufacturing process change
  - $\cdots$

An Example (3)

Figure 1: AE reporting rates for autoimmune SOC and other SOC's



An Example (4)

Estimates of the number of change-points:

- All computations are carried out using compiled R functions and SAS® macros
- No multiplicity adjustment
  - Each tested at $\alpha = 0.05$
  - 4 change-points: 3, 4, 9, and 17
- Šidák inequality
  - Each tested at $\alpha^* = 1 - (1 - 0.05)^{1/19} = 0.0027$
  - 1 change-point: 17
- Proposed method
  - Each λ_i is tested at $0.05 \times (19 - i)/19$
  - 2 change-points: 3 and 17

Summary

- Exact method for computing rejection probabilities
- Simulations
- Constant rate within interval
- multivariate change-point problem
- Sequential or online change-point problem
- Continuous-time estimate of change-point
- Bayesian attempt
- Potential biases in post-marketing spontaneous data reporting