Change point estimation: another look at multiple testing problems

Hongyuan Cao

Department of Statistics, University of Missouri-Columbia

Joint work with Wei Biao Wu

September 30, 2016
WHOA-PSI, St Louis, MO
Outline

1. Introduction
2. Testing procedure
3. Asymptotic results
4. Numerical studies and real data analysis
5. Summary and discussion
### 1.1 Multiple testing

**Table:** Outcomes when testing $m$ hypotheses

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Accept</th>
<th>Reject</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null true</td>
<td>$U$</td>
<td>$V$</td>
<td>$m_0$</td>
</tr>
<tr>
<td>Alternative true</td>
<td>$F$</td>
<td>$S$</td>
<td>$m_1$</td>
</tr>
<tr>
<td>Total</td>
<td>$W$</td>
<td>$R$</td>
<td>$m$</td>
</tr>
</tbody>
</table>

- $V$: # of false discoveries; $F$: # of missed discoveries.

**Calibration criteria**

- False discovery rate: $\text{FDR} = E(FDP) = E(V/R) \leq \alpha$.
- False non-discovery rate: $\text{FNR} = E(FNP) = E(F/W)$.
- Missed discover rate: $\text{MDR} = E(F/m_1)$. 
1.2 FDR literature

Hypotheses: $H_1, \ldots, H_m$ with individual $p$-values $p_1, \ldots, p_m$.

- **Benjamini and Hochberg (BH, 1995) step-up procedure:**
  1. Rank the $p$-values: $p(1) \leq \cdots \leq p(m)$.
  2. Define the threshold $k = \max \{i : p(i) \leq \frac{i}{m} \alpha\}$.
  3. Reject all $H(i) : i \leq k$.
  4. Remark: the actual control level is $\alpha \pi_0$, where $\pi_0$ is the proportion of null hypotheses.

- **Mixture model approach (Storey, 2002, 2003):**
  1. $\theta_1, \ldots, \theta_m$ are independent and identically distributed Bernoulli random variables, where $\theta_i = 0$ if $H_i$ is null and $\theta_i = 1$ otherwise.
  2. $p_i \sim (1 - \theta_i)F_0 + \theta_i F_1$, where $F_0$ and $F_1$ are the null and non-null distributions of $p$-values.
1.3 FDR literature (Cont.)

- **Empirical processes** based methods were proposed by Storey, Taylor and Siegmund (2003) and Genovese and Wasserman (2004).
- Fan, Hall and Yao (2007) addressed the issue on the number of hypotheses that can be simultaneously tested based on asymptotic approximations.
- Cao and Kosorok (2011) extended their work to find simultaneous critical values for multiple testing based on $t$-statistics.
- Cai and Sun (2009) developed a hidden Markov model to incorporate neighbouring information.
- Fan, Han and Gu (2012) used the factor model to integrate the multiple testing dependence structure.
- Recent development can be found in Barber and Candés (2014), Dobriban et al. (2015), Sun et al. (2015), among others.
1.4 Spatial multiple testing

- Hypotheses can exhibit spatial cluster effect—a location and its adjacent neighbours fall in a similar type of region, either significant or non-significant.

- Clark and Hall (2009) investigated the clustering effect of alternative hypotheses under positive stochastic dependence.

- Siegmund, Zhang and Yakir (2011) proposed a scan statistic and treated each cluster as a testing unit.

- Zhang, Fan and Yu (2011) employed a $p$-value smoothing approach by taking local median of $p$-values in a neighbourhood and developed testing procedures based on the smoothed $p$-values.

- We related the multiple testing problems to the change-point theory.
1.5 Change-point theory

- Given a sequence of data, a change-point is a position or time at which the structure of this sequence changes.
- Change point detection in a single sequence has been extensively studied, and reviewed by Zack (1983) and Bhattacharya (1994).
- For multiple change point problem, a comprehensive review can be found in Lai and Xing (2011).
- Our work has a similar flavor as Niu and Zhang (2012), where change point is the point where the distribution changes. We allow the alternative hypotheses sequence to be of any distribution.
- Recent development on change-point analysis can be found in Killick et al. (2012), Fryzlewicz (2014) and Frick et al. (2014), among others.
2.1 Model set up

- We have $m$ hypotheses $H_1, \ldots, H_m$, with corresponding p-values $p_1, \ldots, p_m$.
- Our global null hypothesis is

$$H_0 : p_1, \ldots, p_m \sim U(0,1).$$ (1)

- Global alternative hypothesis: there exist

$$1 = \tau_0 < \tau_1 < \tau_2 < \ldots < \tau_l < n = \tau_{l+1}, l < \infty$$ such that

$$p_1, \ldots, p_{\tau_1-1} \sim U(0,1), \ p_{\tau_1}, \ldots, p_{\tau_2-1} \not\sim U(0,1),$$
$$p_{\tau_2}, \ldots, p_{\tau_3-1} \sim U(0,1), \ p_{\tau_3}, \ldots, p_{\tau_4-1} \not\sim U(0,1), \ldots.$$ (2)

- $\tau_i$ can be interpreted as change-points from true to false when $i$ is odd, and from false to true if $i$ is even.
- Based on the $p$-values $p_1, \ldots, p_m$, we shall estimate the change-points $\mathcal{T} = \{\tau_i, i = 1, \ldots, l\}$. 
Define **local discrepancy**

\[
L_i = \left| \frac{1}{k} \sum_{j=i-k}^{i-1} p_j - \frac{1}{2} \right|, \quad R_i = \left| \frac{1}{k} \sum_{j=i}^{i+k-1} p_j - \frac{1}{2} \right|,
\]

where \(k \to \infty\) is a sequence of the sliding window lengths for which \(k/m \to 0\).

Clearly \(L_{i+k} = R_i\). If there are no change-points, so all \(p_i \sim \mathcal{U}(0, 1)\), we expect that \(\max_{k+1 \leq i \leq m} L_i\) is small.

**Global discrepancy**

\[
\Delta_m = \max_{k \leq i \leq m-k} L_i.
\]
2.3 Testing procedure

1. For a chosen window size $k$, calculate $L_i$ and $R_i$ in (3).
2. For a pre-specified cutoff value $\gamma > 0$, let $Q_i = 1(L_i > \gamma) + 1(R_i > \gamma)$.
3. Decompose $\{1, \ldots, m\} = W_0 \cup W_1 \cup W_2$, where $i \in W_0$ if $Q_i = 0$, $i \in W_1$ if $Q_i = 1$ and $i \in W_2$ if $Q_i = 2$.
4. Let $\mathcal{M}_1, \ldots, \mathcal{M}_\hat{l}$ be connected components of $W_1$ whose lengths are larger than $k/2$.
5. For each $\mathcal{M}_i$, let $\hat{\tau}_i = \arg\max_{j \in \mathcal{M}_i} \{\max(L_j, R_j)\}$ be estimates of $\tau_i$. 
### 3.1 Asymptotic results

**Theorem 1 (How to choose \( \gamma \)?)**

When all hypotheses are from the null, and

\[
\frac{(\log m)^3}{k} + \frac{k}{m} \to 0 \text{ as } m \to \infty. \tag{5}
\]

Let \( A_T = 2 \log T + 2^{-1} \log \log T + \log(\pi^{-1/2}) \) and \( g_m = m/k - 1 \). Then

\[
(1/12)^{-1/2} \sqrt{2k \log(g_m)} \Delta_m - A_{g_m} \Rightarrow E, \tag{6}
\]

where \( E \) has the extreme value distribution \( P(E \leq x) = \exp\{-2e^{-x}\} \).

**Remark:**

Due to the slow rate of convergence, in practice, we can use a simulation assisted method to get \( \gamma \) in numerical studies.
3.2 Assumptions

(A1) The change-points are $\tau_i = \lfloor \eta_i m \rfloor$, $i = 1, \ldots, l$, where $0 < \eta_1 < \cdots < \eta_l < 1$. In addition, there exists a constant $c > 0$ such that $\eta_i - \eta_{i-1} \geq c$, for $i = 1, \ldots, l + 1$, where $\tau_0 = 0$ and $\tau_{l+1} = 1$.

(A2) There exists a constant $0 < \rho < 1/2$ such that, under $H_{1j}$, $E(p_j) \leq \rho$.

Remark:
- (A1) requires that change-points are well-separated and the number of change-points is finite.
- (A2) makes it possible to distinguish alternative hypotheses from the null hypotheses.
3.3 Accurate detection

**Theorem 2**

Assume (A1)-(A2) and $\gamma + \rho < 1/2$, where $\gamma \asymp (k^{-1} \log m)^{1/2}$. Then we have

$$\mathbb{P} \left\{ \hat{l} = l, \max_{i \leq \hat{l}} |\hat{\tau}_i - \tau_i| \leq 2k\gamma (1/2 - \rho)^{-1/2} \right\} \geq 1 - 4k^{-1}me^{-3k\gamma^2/(4+4\gamma)}.$$ 

**Remark:**

- Theorem 2 asserts the **uniform closeness** of our estimated change-points to the real ones.
- Let $k = \lambda \log m$, where $\lambda$ is a constant, $\max_{i \leq \hat{l}} |\hat{\tau}_i - \tau_i| = O_p(\log m)$. 
3.4 Comparison with Higher Criticism

- Data are modelled as $m$ i.i.d. observations from one of two possible situations:

\[ H_0 : X_i \sim N(0, 1), \quad 1 \leq i \leq m, \]
\[ H_1^{(m)} : X_i \sim (1 - \epsilon)N(0, 1) + \epsilon N(\mu, 1), \quad 1 \leq i \leq m. \]

- Under $H_1^{(m)}$, a fraction $\epsilon = m^{-\beta}, \beta \in (1/2, 1)$ of the data come from $N(\mu, 1)$.

- Signal strength $\mu = \mu_m$ has to be at least $c \sqrt{\log m}$.

- If we know signals are clustered, for example, letting $X_i \sim N(\mu_i, 1)$ with $\mu_i = \mu$ if $i \leq \tau = \lceil m(1 - \epsilon) \rceil$ and $\mu_i = 0$ if $i > \tau$. Choose $k \asymp \lceil (\log m)^2 \rceil$. Then $H_0$ can be rejected under the weaker condition $\sqrt{\log m/k} = o(\mu_m)$. 
4.1 Simulation set up

- The number of change-points is 5. $z$-value at each hypothesis follows a normal distribution with mean exhibited in the table below and variance 1.
- We change the signal strength at the first 2.5% loci of $m$ tests.
- $p$-values are calculated based on standard normal distribution.
- Total number of testing is $m = 20,000$ and $m = 100,000$.
- Each simulation example is repeated 1,000 times.

<table>
<thead>
<tr>
<th>Signal</th>
<th>Segment</th>
<th>2.5</th>
<th>2.5</th>
<th>30</th>
<th>2.5</th>
<th>30</th>
<th>2.5</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>-1.5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1.5 and 1 alternating</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.2 Implementation of our method

- Use $k = \lfloor m^{1/2} \rfloor$ and $\lfloor (\log m)^2 \rfloor$.
- We obtain $10^4$ independent realizations of $B_m = 12^{1/2}\max_{k+1 \leq i \leq m} L_i$. This is computed based on $m$ independent $U(0, 1)$ random variables.
- Choose critical value $\gamma$ to be 95% sample quantile of empirical distribution of $B_m$.
- FDR, FNR and MDR can be obtained.
- At the same level of FDR, we compare our method with the spatial smoothing method of Zhang, Fan and Yu (2011).
4.3 p-value smoothing

Zhang, Fan and Yu (2011)

- Original null $p$-values are uniformly distributed and the median filtered null $p$-values will approximately be symmetrically distributed about 0.5.
- The distribution of smoothed $p$-values under the null hypothesis can be estimated by

$$
\hat{G}^*(t) = \begin{cases} 
\frac{\sum_{i=1}^{m} I\{p_i^* \geq 1-t\}}{2 \sum_{i=1}^{m} I(p_i^*>0.5)+\sum_{i=1}^{m} I(p_i^*=0.5)} & \text{if } 0 \leq t \leq 0.5 \\
1 - \frac{\sum_{i=1}^{m} I\{p_i^* \geq t\}}{2 \sum_{i=1}^{m} I(p_i^*>0.5)+\sum_{i=1}^{m} I(p_i^*=0.5)} & \text{if } 0.5 < t \leq 1,
\end{cases}
$$

where $p_i^*$ is the median of the $p$-values in the $k^*$th neighbourhood of $i$th hypothesis.
The estimated FDR is

\[ \hat{FDR}_{ZFY}(t) = \frac{W^*(\lambda)\hat{G}^*(t)}{\{R^*(t) \vee 1\}\{1 - \hat{G}^*(\lambda)\}}, \]

\(\lambda\) is a tuning parameter, \(W^*(\lambda) = \sum_{i=1}^{m} I\{p^*_i > \lambda\}\) and \(R^*(t) = \sum_{i=1}^{m} I\{p^*_i \leq t\}\).

At FDR level \(\alpha\), threshold \(\hat{t}\) is chosen as the largest \(t\) that satisfies \(\hat{FDR}_{ZFY}(\hat{t}) \leq \alpha\).

We set the tuning parameter \(\lambda = 0.1\) and size of neighbourhood \(k^* = k^{1/2}\).
### 4.5 Comparison with Zhang, Fan and Yu (2011)

<table>
<thead>
<tr>
<th>m</th>
<th>k</th>
<th>µ</th>
<th>Our procedure</th>
<th>ZFY procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>FDR</td>
<td>FNR</td>
</tr>
<tr>
<td>20000</td>
<td>141</td>
<td>0.1</td>
<td>0.0014</td>
<td>0.0404</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.8</td>
<td>0.0012</td>
<td>0.0275</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.8</td>
<td>0.0010</td>
<td>0.0200</td>
</tr>
<tr>
<td>98</td>
<td></td>
<td>0.1</td>
<td>0.0008</td>
<td>0.0460</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.8</td>
<td>0.0008</td>
<td>0.0427</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.8</td>
<td>0.0006</td>
<td>0.0239</td>
</tr>
<tr>
<td>100000</td>
<td>316</td>
<td>0.1</td>
<td>0.0002</td>
<td>0.0297</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.8</td>
<td>0.0002</td>
<td>0.0059</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.8</td>
<td>0.0002</td>
<td>0.0057</td>
</tr>
<tr>
<td>132</td>
<td></td>
<td>0.1</td>
<td>0.0002</td>
<td>0.0325</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.8</td>
<td>0.0018</td>
<td>0.0300</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.8</td>
<td>0.0001</td>
<td>0.0069</td>
</tr>
</tbody>
</table>
4.6 Application to a genome wide association study

- We study the association between single nucleotide polymorphisms (SNPs) on chromosome 8 and HDL cholesterol.
- There are $m = 353,488$ SNPs and we obtain $p$-values through marginal regression.
- $p$-values show substantial serial dependence and we estimate the long run variance using the batched mean estimate (Brockwell and Davis, 2009)

$$
\hat{\sigma}^2(l_m) = \frac{l_m}{m - l_m + 1} \sum_{j=1}^{m-l_m+1} (l_m^{-1} \sum_{i=j}^{j+l_m-1} p_i - \bar{\rho}_m)^2,
$$

where $\bar{\rho}_m = \sum_{i=1}^{m} p_i / m$, and $l_m$ is the window size satisfying $l_m \to \infty$ and $l_m/m \to 0$. We use $l_m = 28$, which is between $m^{1/3}$ and $m^{1/4}$ and obtain $\hat{\sigma}^2 = 0.8736$. 
4.7 Analysis results

- We use $k = \lfloor (\log m)^2 \rfloor = 163$.
- Critical value $\gamma$ is the multiplication of $\hat{\sigma}$ and 95% empirical quantile of $B_m$, which is 0.3422.
- The alternative hypothesis blocks are $\{42845, \ldots, 43045\}$, $\{83839, \ldots, 84430\}$, $\{282143, \ldots, 282344\}$ and $\{297063, \ldots, 297225\}$.
- The total number of rejections is 1158, which is roughly 0.3% of the total number of tests.
- The last two clusters belong to the same gene set.
- Our result agrees well with others in the literature: the closest genes to these regions are PPP1R3B (Teslovich et al., 2010), INTS10 (Wilke, 2011) and TRPS1 (Teslovich et al., 2010).
Figure: Identified gene sets. Top panel: z-values; bottom panel: p-values.
5 Concluding remarks

- We propose a change-point detection algorithm to take into account spatial information for large scale multiple testing.
- A new test statistic is proposed and its asymptotic distribution is investigated.
- It is of interest to use the original data instead of $p$-values and derive corresponding results.
Thank You!
Questions?