Biometrika (2011), **94**, 2, *pp*. 1–6

Advance Access publication on 31 July 2008
Printed in Great Britain

Supplementary material for the optimal power puzzle: scrutiny of the monotone likelihood ratio assumption in multiple testing

BY HONGYUAN CAO
Department of Health Studies, University of Chicago, Chicago, Illinois 60637, U.S.A.
hycao@uchicago.edu

WENGUANG SUN
Department of Information and Operation Management, Marshall School of Business, University of Southern California, Los Angeles, California 90089, U.S.A.
wenguans@marshall.usc.edu

AND MICHAEL R. KOSOROK
Department of Biostatistics and Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27514, U.S.A.
kosorok@unc.edu

SUMMARY

This Supplementary Material provides simulation studies on grouped hypothesis testing and marginal false discovery rate analysis, a revisit to Example 1 and details on the proofs of Theorem 1, Theorem 2 and Theorem 3.

1. SIMULATION STUDIES

1·1. *Grouped hypothesis testing*

In many large-scale studies, the data are collected from various sources and the test statistics may exhibit different characteristics. For example, in a brain imaging study considered by Schwartzman et al. (2008) for comparing dyslexic versus normal children, it was found that the z-values from the front and back halves of the brain centered at different means: the estimated null distributions of z-values for the front and back halves of the brain are $N(0.06, 1.09^2)$ and $N(-0.29, 1.01^2)$, respectively. In the adequate yearly progress study of California high schools by Rogosa (2003) for comparing academic performance of socioeconomically advantaged versus disadvantaged students, the z-value distributions vary significantly according to school sizes. The problem can be formulated as testing groups of hypotheses. It was argued by Efron (2008) that it can be problematic to combine all tests together without taking into account the grouping information. The issue was further studied by Ferkingstad et al. (2008), Cai & Sun (2009), Hu et al. (2010) and Peña et al. (2011), among others. We tackle the problem from a different angle, with the specific goal to show that the monotone likelihood ratio condition can be violated when grouping information is ignored. Consider the following example.

Suppose $m_1 = 2000$ hypotheses come from Group 1 and $m_2 = 1000$ hypotheses come from Group 2. Let $m = m_1 + m_2 = 3000$. The test statistics in the first group $Z_1, \ldots, Z_{2000}$ fol-

low a two-component normal mixture model $F^1(t) = 0.9N(1, 0.5^2) + 0.1N(-2, 0.5^2)$, and the test statistics in the second group $Z_{2001}, \ldots, Z_{3000}$ follow distribution $F^2(t) = 0.8N(0, 1) + 0.2N(2, 1)$. In both cases, the first component in the mixture is the null distribution and the second component is the non-null distribution. Our simulation setting is motivated by a brain imaging study considered in Schwartzman et al. (2008), where the null distributions of the two groups have different means and variances. The non-null proportions and alternative distributions are also set to be different across groups. It is easy to check that the monotone likelihood ratio condition holds separately in both groups. However, we will show that the monotone likelihood ratio condition may fail if we combine the two groups into a single group without adjustment. Specifically, consider the following two multiple testing strategies.

The first strategy, referred to as the pooled analysis, is to ignore the grouping information and pool all tests together. Under this framework, the hypotheses will be ranked according to absolute deviation from the sample median of $Z_1, \ldots, Z_{3000}$. In the simulation, we plot the false discovery rate levels as functions of the critical values. The results are summarized in the left panel of Figure 1. We can see that the false discovery rate level decreases first and then increases with the critical values, indicating that the monotone likelihood ratio condition is violated.

The second strategy, suggested by Efron (2008) and referred to as the separate analysis, is to utilize the grouping information and analyze the data separately. For example, let $\tilde{\mu}^k$ be the sample median of the test statistics in group k, $k = 1, 2$. We can compute group-specific p-values $P_i = 2\Phi(-|X_i - \tilde{\mu}^1|/0.5)$ if X_i comes from group 1 and $P_i = 2\Phi(-|X_i - \tilde{\mu}^2|)$ if X_i comes from group 2. Then we rank the hypotheses according to their group-specific p-values. In the simulation, we vary the critical values and plot the false discovery rate levels on the right panel of Figure 1. We can see that the false discovery rate level is now monotonically decreasing in the critical values when we test different groups separately.

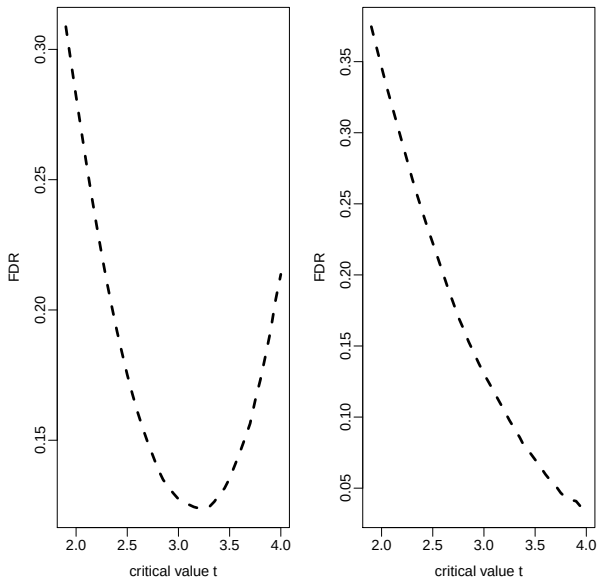

Fig. 1. Left panel represents pooled analysis and right panel represents separate analysis

Our numerical results show that even when the monotone likelihood ratio condition holds in separate groups, the condition can be violated in a pooled analysis. Unfortunately, a pooled

analysis is often what is done in practice. The issue can be resolved if a separate analysis is conducted.

1·2. *Marginal false discovery rate analysis*

We showed that under the normal mixture model, false discovery rate and marginal false discovery rate are asymptotically equivalent if test statistics are independent. The situation is quite different under dependence. In a simulation study with the same set up as in Example 2, we vary the critical value t from 1.95 to 4 and calculate the false discovery rate and marginal false discovery rate, respectively. The results are summarized in Figure 2.

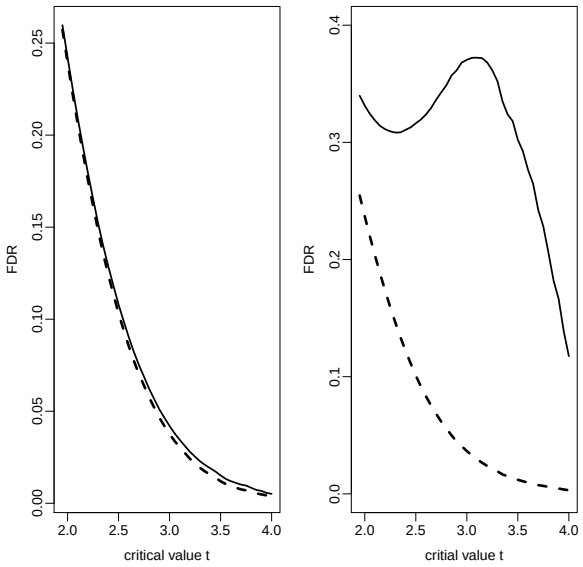

Fig. 2. Left panel is for weak correlation and right panel is for strong correlation. Solid line represents FDR and dotted line represents mFDR

From the plot, we can see that false discovery rate and marginal false discovery rate can be very different especially when strong correlation exists. It seems to be clear that dependence has a big impact on the variability of the number of rejections R, and large variability in R would further result in big discrepancies between false discovery rate and marginal false discovery rate. In addition, if the number of rejections R is small or the cut-off is large, the relative variability would be further increased.

2. A REVISIT TO EXAMPLE 1

Now we apply Theorem 3 in heteroscedastic model (7). First estimate p using the method in Jin & Cai (2007) and f using a kernel density estimator. The null density $f_0(x)$ is the standard normal. Let $\hat{p}$ and $\hat{f}$ denote the estimates and define $\widehat{\text{Lfdr}}(X_i)$ to be the plug-in statistic. The false discovery rate is calculated for varying number of rejections. We choose $m = 2,000$, $p = 0.1$, $\mu = 2.5$ and $\sigma = 0.5$, and then apply both p-value and local false discovery rate based testing procedures for 2,000 simulated data sets; the results are summarized in Figure 3. The false discovery rate of the p-value method first decreases and then increases in the p-value cut-off. Therefore the monotone likelihood ratio condition is violated by P_i. Consequently, a smaller

p-value cutoff may correspond to a higher false discovery rate. In contrast, the false discovery rate of the local false discovery rate method increases monotonically in the local false discovery rate cutoff. This is consistent with our theoretical prediction. It is clear that the confusing situation caused by p-value methods is avoided by using the local false discovery rate which yields an increasing curve.

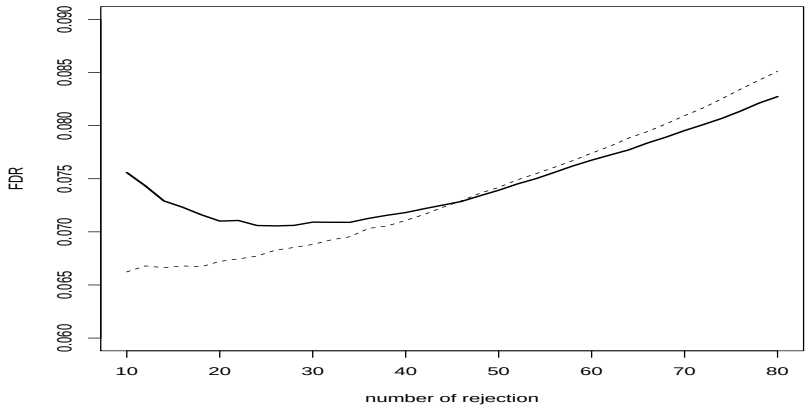

Fig. 3. Solid line represents p-value method and dotted line represents local false discovery rate method

3. PROOF OF THEOREMS

3·1. *Proof of Theorem 1*

Denote by Φ and ϕ the cdf and pdf of a standard normal deviate, respectively. Observe that

$$G_P^1(t) = pr_{H_i=1}(p_i < t) = \Phi\left\{\frac{\Phi^{-1}(t)+\mu}{\sigma}\right\},$$

and the conditional pdf of the p-value is

$$\begin{aligned} g_P^1(t) &= \frac{1}{\sigma}\phi\left\{\frac{\Phi^{-1}(t)+\mu}{\sigma}\right\}/\phi\{\Phi^{-1}(t)\} \\ &= \begin{cases} (1/\sigma)\exp\left[-\frac{1-\sigma^2}{2\sigma^2}\left\{\Phi^{-1}(t)+\frac{\mu}{1-\sigma^2}\right\}^2+\frac{\mu^2}{2(1-\sigma^2)}\right] & \text{if } \sigma<1 \\ (1/\sigma)\exp\left[\frac{\sigma^2-1}{2\sigma^2}\left\{\Phi^{-1}(t)-\frac{\mu}{\sigma^2-1}\right\}^2-\frac{\mu^2}{2(\sigma^2-1)}\right] & \text{if } \sigma>1 \\ \exp\left\{-\Phi^{-1}(t)\mu-\frac{1}{2}\mu^2\right\} & \text{if } \sigma=1 \end{cases}. \end{aligned}$$

The critical region for inference is the interval $t \in (0,\eta)$, where η is usually very small. In order to guarantee that $G_P^1(t)$ is concave, $g_P^1(t)$ should be decreasing in t. It is easy to see that $g_P^1(t)$ is a decreasing function for $t \in (0,\eta)$ when $\sigma \geq 1$. However, $g_P^1(t)$ is increasing in t for $t < t_0 = \Phi\left\{-\mu/(1-\sigma^2)\right\}$ and decreasing in t for $t \geq t_0$ when $\sigma < 1$ ($\Phi^{-1}(t) < \Phi^{-1}(\eta) < \Phi^{-1}(1/2) = 0$). Therefore (i) is straightforward. To see (ii), write

$$mFDR = \frac{(1-p)t}{(1-p)t + pG_p^1(t)} = \frac{1}{1+\frac{p}{1-p}\frac{G_P^1(t)}{t}}.$$

For $t \in (0, t_0)$, we have $\frac{d}{dt}(G_P^1(t)/t) = (g_P^1(t)t - G_P^1(t))/t^2 = [g_P^1(t) - g_P^1(t^*)]/t > 0$, where $0 < t^* < t < t_0$ by the mean value theorem and using the fact that $G_p^1(0) = 0$. So $G_P^1(t)/t$ is an increasing function. Since $h(x) = 1/(1 + ax)$ is a monotone decreasing function (for $a > 0$), we have that marginal false discovery rate is decreasing in t when $0 < t < t_0$. So (6) fails in this scenario. □

3·2. *Proof of Theorem 2*

For (i), it suffices to show that $\frac{d\text{mFDR}(t)}{dt} \geq 0, \forall 0 < t < 1$. Note that

$$\text{mFDR} = \frac{\sum_{i=1}^m (1-p_i)G_{i0}(t)}{\sum_{i=1}^m (1-p_i)G_{i0}(t) + p_i G_{i1}(t)}.$$

We have

$$\begin{aligned}
&\frac{d\text{mFDR}(t)}{dt} \\
&= \frac{[\sum_{i=1}^m (1-p_i)g_{i0}(t)][\sum_{i=1}^m (1-p_i)G_{i0}(t) + p_i G_{i1}(t)] - [\sum_{i=1}^m (1-p_i)G_{i0}(t)][\sum_{i=1}^m (1-p_i)g_{i0}(t) + p_i g_{i1}(t)]}{[\sum_{i=1}^m (1-p_i)G_{i0}(t) + p_i G_{i1}(t)]^2} \\
&= \frac{[\sum_{i=1}^m p_i G_{i1}(t)][\sum_{i=1}^m (1-p_i)g_{i0}(t)] - [\sum_{i=1}^m p_i g_{i1}(t)][\sum_{i=1}^m (1-p_i)G_{i0}(t)]}{[\sum_{i=1}^m (1-p_i)G_{i0}(t) + p_i G_{i1}(t)]^2} \\
&= \frac{[\sum_{i=1}^m p_i \int_0^t g_{i1}(x)dx][\sum_{i=1}^m (1-p_i)g_{i0}(t)] - [\sum_{i=1}^m p_i g_{i1}(t)][\sum_{i=1}^m (1-p_i)\int_0^t g_{i0}(x)dx]}{[\sum_{i=1}^m (1-p_i)G_{i0}(t) + p_i G_{i1}(t)]^2} \\
&= \frac{[\int_0^t \frac{\sum_{i=1}^m p_i g_{i1}(x)}{\sum_{i=1}^m (1-p_i)g_{i0}(x)} \sum_{i=1}^m (1-p_i)g_{i0}(x)]dx[\sum_{i=1}^m (1-p_i)g_{i0}(t)] - [\sum_{i=1}^m p_i g_{i1}(t)][\sum_{i=1}^m (1-p_i)\int_0^t g_{i0}(x)dx]}{[\sum_{i=1}^m (1-p_i)G_{i0}(t) + p_i G_{i1}(t)]^2} \\
&\geq \frac{[\sum_{i=1}^m p_i g_{i1}(t)][\sum_{i=1}^m \int_0^t (1-p_i)g_{i0}(x)dx] - [\sum_{i=1}^m p_i g_{i1}(t)][\sum_{i=1}^m (1-p_i)\int_0^t g_{i0}(x)dx]}{[\sum_{i=1}^m (1-p_i)G_{i0}(t) + p_i G_{i1}(t)]^2} = 0.
\end{aligned}$$

This proves (i). (ii) can be similarly proved by noting that $\text{mFNR} = \frac{\sum_{i=1}^m p_i(1-G_{i1}(t))}{\sum_{i=1}^m [p_i(1-G_{i1}(t)) + (1-p_i)(1-G_{i0}(t))]}$, and we get (iii) by combining (i) and (ii). □

3·3. *Proof of Theorem 3*

Let λ be the penalty for a false positive versus a false negative. We first consider a weighted classification problem with loss function $L(\vec{\theta}, \vec{\delta}) = m^{-1} \sum_i \{\lambda(1-\theta_i)\delta_i + \theta_i(1-\delta_i)\}$. Let $\vec{X} = (X_1, \ldots, X_m)$ and $\vec{s} = (s_1, \ldots, s_m)$. The posterior risk is $E_{\vec{\theta}|\vec{X},\vec{s}}\{L_\lambda(\vec{\theta}, \vec{\delta})\} = \frac{1}{m}\sum_i E_{\theta_i|\vec{X},\vec{s}}\{\lambda(1-\theta_i)\delta_i + \theta_i(1-\delta_i)\} = \frac{1}{m}\sum_{i=1}^m \lambda\delta_i T_{OR}^i + (1-\delta_i)(1-T_{OR}^i)$. Therefore the classification risk is minimized by $\delta_i = I(\lambda T_{OR}^i < 1 - T_{OR}^i) = I(T_{OR}^i < (1+\lambda)^{-1})$, for $i = 1, \ldots, m$. Let $G_{i0}(t), G_{1i}(t), g_{i0}(t)$ and $g_{i1}(t)$ be defined as before (with respect to $\vec{T}_{OR}$). The goal is to show that (9) holds. For the weighted classification problem, take $\lambda = 1/t$. Let $t^* > 0$. Suppose $\vec{\delta}(\vec{T}_{OR}, t^*) = I(\vec{T}_{OR} < t^*\vec{1})$ is used for classification. The risk is $R = (mt)^{-1}\sum_i (1-p(s_i))G_{i0}(t^*) + m^{-1}\sum_i p(s_i) - m^{-1}\sum_i p(s_i)G_{i1}(t^*)$. The optimal cutoff t^* that minimizes this risk satisfies

$$\frac{\sum_i p(s_i)g_{i1}(t^*)}{\sum_i (1-p(s_i))g_{i0}(t^*)} = \frac{1}{t}.$$

Meanwhile, note that the optimal cutoff t^* is given by $t^* = 1/(1+\lambda) = t/(1+t)$. Hence

$$\frac{\sum_i p(s_i) g_{i1}(t^*)}{\sum_i (1-p(s_i)) g_{i0}(t^*)} = \frac{1-t^*}{t^*}.$$

By definition $0 < t^* < 1$. Thus $(1-t^*)/t^*$ is decreasing in t^* and the result follows. □.

REFERENCES

CAI, T. T. & SUN, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Amer. Statist. Assoc.* 488 1467–1481.

EFRON, B. (2008). Simultaneous inference: When should hypothesis testing problems be combined? *Annals of Applied Statistics* 2 197–223.

FERKINGSTAD, E., FRIGESSI, A., RUE, H., THORLEIFSSON, G. & KONG, A. (2008). Unsupervised empirical Bayesian multiple testing with external covariates. *Ann. Appl. Stat.* 2 714–735.

HU, J., ZHAO, H. & ZHOU, H. (2010). False discovery rate control with groups. *J. Amer. Statist. Assoc.* 105 1215–1227.

JIN, J. & CAI, T. T. (2007). Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* **102** 495–506.

PEÑA, E. A., HABIGER, J. D. & WU, W. (2011). Power-enhanced multiple decision functions controlling family-wise error and false discovery rates. *Ann. Statist.* 39 556–583.

ROGOSA, D. (2003). Accuracy of api index and school base report elements: 2003 academic performance index, california department of education.

SCHWARTZMAN, A., DOUGHERTY, R. F. & TAYLOR, J. E. (2008). False discovery rate analysis of brain diffusion direction maps. *Annals of Applied Statistics* 2 153–175.

[*Received November* 2011. *Revised November* 2012]