Biometrika Advance Access published March 26, 2013

Biometrika (2013), *pp.* 1–8
doi: 10.1093/biomet/ast001

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 Operations 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, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27514, U.S.A.
kosorok@unc.edu

Summary

In single hypothesis testing, power is a nondecreasing function of Type I error rate; hence it is desirable to test at the nominal level exactly to achieve optimal power. The optimal power puzzle arises from the fact that for multiple testing under the false discovery rate paradigm, such a monotonic relationship may not hold. In particular, exact false discovery rate control may lead to a less powerful testing procedure if a test statistic fails to fulfil the monotone likelihood ratio condition. In this article, we identify different scenarios wherein the condition fails and give caveats for conducting multiple testing in practical settings.

Some key words: False discovery rate; Heteroscedasticity; Monotone likelihood ratio; Multiple testing dependence.

1. Introduction

We study an important assumption that has been used implicitly in the multiple testing literature. In the context of false discovery rate analysis (Benjamini & Hochberg, 1995), we show that the assumption can be violated in many important settings. The goal of this article is to explicitly state the assumption to bridge the gap in conventional methodological development, rigorously investigate the legitimacy of the assumption in various settings, and give caveats for conducting multiple testing in practice.

To identify this assumption, it is helpful to first examine closely the framework of single hypothesis testing. Suppose we want to test H_0 versus H_1 based on the observed value of a continuous random variable X. A binary decision rule $\delta \in \{0, 1\}$ divides the sample space S into two regions, $S = S_0 \cup S_1$, such that $\delta = 0$ when $X \in S_0$ and $\delta = 1$ when $X \in S_1$. Let $T(\cdot)$ be a function of X, with small values indicating evidence against H_0. The critical region S_1 can be expressed as $S_1 = \{x \in S : T(x) < t\}$. Correspondingly, we have a testing rule $\delta = I\{T(X) < t\}$ where $I(\cdot)$ is an indicator function and t is the rejection threshold. Denote by F_0 and F_1 the conditional distributions of X under H_0 and H_1, respectively, and by G_0 and G_1 the conditional distributions of $T(X)$ under H_0 and H_1. The Type I and Type II error rates of δ are $\alpha(t) = \text{pr}_{H_0}\{T(X) < t\} = G_0(t)$ and $\beta(t) = \text{pr}_{H_1}\{T(X) > t\} = 1 - G_1(t)$, respectively. Since $\alpha(t)$ increases in t and $\beta(t)$ decreases in t, we conclude that $\beta(t)$ decreases in $\alpha(t)$. Therefore the optimal choice of t^* which minimizes $\beta(t)$ subject to $\alpha(t) \leqslant \alpha_0$ should satisfy $\alpha(t^*) = \alpha_0$. In other words, one should test H_0 at the nominal level exactly in order to minimize the Type II error rate.

Now suppose we want to test m hypotheses $H_1, \ldots, H_m$ simultaneously based on a random vector $X = (X_1, \ldots, X_m)$. Let $\theta_1, \ldots, \theta_m$ be independent and identically distributed Ber(p) random variables, where $\theta_i = 0$ if H_i is a null and $\theta_i = 1$ otherwise. Assume that

$$X_i \sim (1 - \theta_i)F_0 + \theta_i F_1, \tag{1}$$

where F_0 and F_1 are the null and nonnull distributions, respectively. Let $T_i(\cdot)$ be a function of X for testing H_i. The solution to a multiple testing problem can be represented by a vector of binary decisions $\delta = (\delta_1, \ldots, \delta_m) \in \{0, 1\}^m$, where $\delta_i = 1$ if we reject H_i and $\delta_i = 0$ otherwise. As an example, consider a testing rule that rejects H_i when $P_i < t$, where P_i is the p-value. Then $T_i(X) = P_i$ and we can write $\delta_i = I(P_i < t)$. Denote the conditional distribution of P_i under the alternative by G_1. The mixture distribution of P_i is $G(t) = (1 - p)t + pG_1(t)$, where p is the proportion of nonnull hypotheses. The false discovery rate is the expected proportion of false positives among all rejections. Let $x \vee y = \max(x, y)$. Genovese & Wasserman (2002) showed that the false discovery rate, as a function of the p-value threshold t, is

$$\textsc{fdr}(t) = E\left\{\frac{\sum_{i=1}^m (1 - \theta_i)\delta_i}{(\sum_{i=1}^m \delta_i) \vee 1}\right\} = \frac{(1 - p)t}{(1 - p)t + pG_1(t)} + O(m^{-1/2}). \tag{2}$$

The false nondiscovery rate, missed discovery rate, and average power can be used to describe the power of a false discovery rate procedure:

$$\textsc{fnr}(t) = E\left[\frac{\sum_{i=1}^m (1 - \delta_i)\theta_i}{\{\sum_{i=1}^m (1 - \delta_i)\} \vee 1}\right] = \frac{p\{1 - G_1(t)\}}{p\{1 - G_1(t)\} + (1 - p)(1 - t)} + O(m^{-1/2}), \tag{3}$$

$$\textsc{mdr}(t) = E\left\{\frac{\sum_{i=1}^m \theta_i(1 - \delta_i)}{(\sum_{i=1}^m \theta_i) \vee 1}\right\} = 1 - G_1(t) + O(m^{-1/2}),$$

and $\textsc{ap}(t) = 1 - \textsc{mdr}(t)$. Similar to the situation in single hypothesis testing, it is often assumed in the multiple testing literature that the following holds.

Assumption 1. The function $\textsc{fdr}(t)$ increases in t and $\textsc{fnr}(t)$ decreases in t; therefore $\textsc{fnr}(t)$ decreases in $\textsc{fdr}(t)$.

Consequently, to achieve the optimal power, we should control the false discovery rate at the nominal level α exactly; that is, the optimal p-value cut-off t^* should solve the equation

$$\frac{(1 - p)t^*}{(1 - p)t^* + pG_1(t^*)} = \alpha. \tag{4}$$

In Genovese & Wasserman (2002), the testing rule $\delta_i = I(P_i < t^*)$ is referred to as the oracle false discovery rate procedure. In the literature, considerable effort has been devoted to the development of data-driven methods aiming to mimic the oracle for precise false discovery rate control (Benjamini & Hochberg, 2000; Genovese & Wasserman, 2004; Benjamini et al., 2006). The tacit assumption is that the closer a test gets to the upper bound α, the more powerful the test is. However, a fundamental question is whether Assumption 1 always holds. This question reveals a logical gap in the methodological development. If Assumption 1 is not true, then a false discovery rate procedure at level $\alpha^* < \alpha$ can be more powerful than a procedure at level α. Consequently, the oracle procedure (4) would not be optimal, and all attempts to achieve precise false discovery rate control must fail. Surprisingly, Assumption 1 can be violated in several important scenarios.

2. The monotone likelihood ratio condition

Consider a decision rule $\delta = (\delta_1, \ldots, \delta_m)$, where $\delta_i = I(T_i < t)$. Various statistics T_i have been proposed in the literature for multiple testing, including the local false discovery rate (Efron et al., 2001), the weighted p-value (Genovese et al., 2006), the local index of significance (Sun & Cai, 2009) and t-statistics (Cao & Kosorok, 2011). Therefore, it would be desirable to develop a general principle which guarantees that Assumption 1 is satisfied by different T_i. To focus on the main idea, we assume for the moment that the T_i are identically distributed with $G_0(t) = \text{pr}(T_i < t \mid \theta_i = 0)$ and $G_1(t) = \text{pr}(T_i < t \mid \theta_i = 1)$ for $i = 1, \ldots, m$. Let $g_j(t) = (\text{d}/\text{d}t)\, G_j(t)$ $(j = 0, 1)$ be the corresponding conditional densities. The monotone likelihood ratio condition can be stated as

$$g_1(t)/g_0(t) \text{ is monotonically decreasing in } t. \tag{5}$$

It is commonly assumed that $G_1(t)$, the p-value distribution under the alternative, is a concave function. Such an assumption has been made in Storey (2002), Genovese & Wasserman (2002, 2004) and Kosorok & Ma (2007), among others. This concavity assumption is a special case of condition (5) if the null p-value distribution is uniform. A significant advantage of (5), compared to Assumption 1, is that it can be roughly checked in practice. For a p-value testing procedure, we can first estimate the mixture density by $\hat{g}_P(t)$; then $\hat{g}_P(t)$ would be decreasing in t if the monotone likelihood ratio condition holds.

The dominant terms on the right-hand sides of (2) and (3) are referred to as the marginal false discovery rate and marginal false nondiscovery rate, denoted by mfdr(t) and mfnr(t), respectively. The property of a testing rule is essentially characterized by these approximations. Hereafter we shall mainly use these marginal measures, to simplify our discussion while still preserving the key features of the problem. The main finding is that condition (5), although not affecting the validity of a multiple testing procedure, plays an important role in optimality analysis. The next proposition shows that exact false discovery rate control leads to the most powerful test when condition (5) is fulfilled.

Proposition 1 (Sun & Cai, 2007). *Consider the random mixture model* (1)*. Let $T_i = T(X_i)$ be the test statistic and $\delta(T, t) = \{\delta_i : i = 1, \ldots, m\} = \{I(T(X_i) < t) : i = 1, \ldots, m\}$ the testing rule. If T_i satisfies condition* (5)*, then:* (i) mfdr(t) *increases in t;* (ii) mfnr(t) *decreases in t;* (iii) mfnr(t) *decreases in* mfdr(t)*. In particular, assertions* (i)–(iii) *hold when $T(X_i) = P_i$ and the p-value distribution function under the alternative is concave.*

As pointed out by a reviewer, the monotonicity relationship is derived only for single-step thresholding procedures $\delta(T, t)$. The results in Genovese & Wasserman (2002) indicate that, in a random mixture model, a broad class of stagewise testing procedures have asymptotically equivalent versions in the family of single-step thresholding procedures. Therefore our result remains relevant when stagewise procedures such as the step-up procedure of Benjamini & Hochberg (1995) are considered.

3. Violation of the monotone likelihood ratio condition

3·1. *Heteroscedastic models*

This section explores several important situations where Assumption 1 and condition (5) are violated. First, consider a heteroscedastic normal mixture model

$$Z_i \mid \theta_i \sim (1 - \theta_i) N(0, 1) + \theta_i N(\mu, \sigma^2) \quad (i = 1, \ldots, m), \tag{6}$$

where $\theta_1, \ldots, \theta_m$ are independent Ber(p) variables. The next theorem shows that the standard approach, which thresholds the z-value or, equivalently, the one-sided p-value $P_i = \text{pr}\{N(0, 1) > Z_i\}$, may fail to satisfy condition (5).

Theorem 1. *Consider the normal mixture model* (6)*. Define the one-sided p-value by $P_i = \text{pr}\{N(0, 1) > Z_i\}$. Let $\delta = (\delta_i : i = 1, \ldots, m)$ be a testing rule, where $\delta_i = I(P_i < t)$. Then condition* (5) *always holds when $\sigma \geqslant 1$ but fails when $\sigma < 1$.*

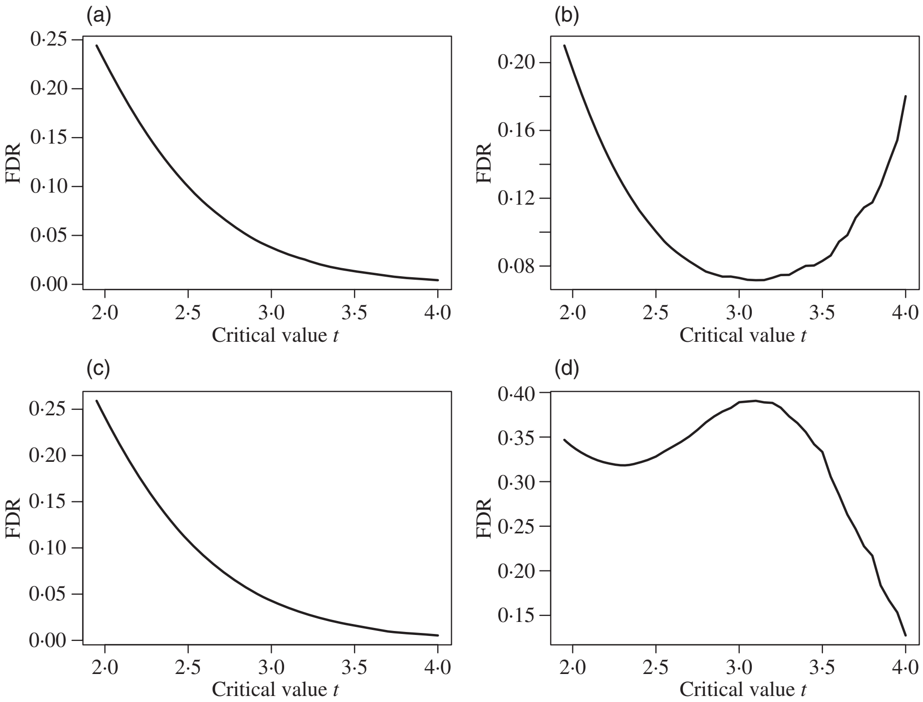

Fig. 1. Computed false discovery rate plotted as a function of the critical value. The top row corresponds to heteroscedastic models (Example 1) with: (a) $\sigma = 1$; (b) $\sigma = 0{\cdot}5$. The bottom row corresponds to correlated tests (Example 2) with: (c) weak correlation; (d) strong correlation.

The heteroscedastic model (6) can arise from applications such as sign tests. Suppose we want to test whether the random variable Y_i has median 0 based on replicated observations $Y_{i1}, \ldots, Y_{in}$ $(i = 1, \ldots, m)$. Let $q = \text{pr}(Y_i > 0)$. The hypotheses can be stated as $H_{0i} : q = 0{\cdot}5$ versus $H_{1i} : q \neq 0{\cdot}5$. The test statistic is $Z_i = n^{-1/2} \sum_{j=1}^{n} \text{sign}(Y_{ij}) = n^{-1/2} \sum_{j=1}^{n} \{2I(Y_{ij} > 0) - 1\}$. We have $E(Z_i) = 2q - 1$, $\text{var}(Z_i) = \sigma_q^2 = 4q(1-q)$, $Z_i \sim N(0, 1)$ under H_{0i}, and $Z_i \sim N(2q - 1, \sigma_q^2)$ under H_{1i} with $\sigma_q^2 < 1$. Therefore the sign test gives rise to a heteroscedastic model asymptotically. Below we provide a numerical example to illustrate the failure of the monotonicity condition in a heteroscedastic model.

Example 1. We generate $m = 2000$ independent $\text{Ber}(p)$ variables $\theta_1, \ldots, \theta_m$ with $p = 0{\cdot}1$, and generate Z_i according to model (6) with $\mu = 2{\cdot}5$. The one-sided p-value is obtained as $P_i = \text{pr}\{N(0, 1) > Z_i\}$. We vary the critical value t from 1·95 to 4 and calculate the false discovery proportion $\textsc{fdp}(t)$. Then $\textsc{fdr}(t)$ is obtained by averaging $\textsc{fdp}(t)$ over 2000 replications. The results are summarized in Fig. 1(a)–(b). We can see that when $\sigma = 1$, $\textsc{fdr}(t)$ decreases monotonically in t. However, when $\sigma = 0{\cdot}5$, $\textsc{fdr}(t)$ first decreases and then increases in t. The violation of monotonicity leads to testing results that are not interpretable. For example, Fig. 1(b) suggests that if we threshold at $t = 3{\cdot}8$, the false discovery rate is 0·12, but if we threshold at $t = 3{\cdot}0$, the false discovery rate is 0·07. In fact, a larger threshold does not necessarily control the false discovery rate at a lower level when $\sigma < 1$. This heteroscedasticity resulted in the violation of Assumption 1 and condition (5).

3·2. *Correlated tests*

This section discusses the violation of condition (5) under dependency. An additional example on multiple testing with groups is discussed in the Supplementary Material. The dependency issue has attracted much attention in the multiple testing literature (Benjamini & Yekutieli, 2001; Efron, 2007; Wu, 2008; Sun & Cai, 2009). The next example shows that condition (5) can be violated under strong dependency.

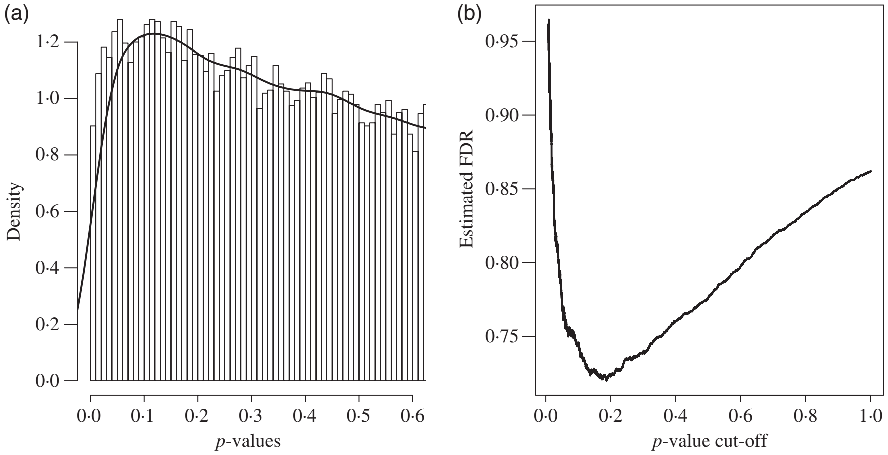

Fig. 2. Results from a DNA methylation study dataset: (a) histogram and density of p-values; (b) estimated false discovery rate plotted against p-value cut-off.

Example 2. Suppose that we observe $X = (X_1, \ldots, X_m)$ from the model

$$X = \mu + \epsilon$$

and want to identify nonzero elements in $\mu = (\mu_1, \ldots, \mu_m)$. In many important applications, such as imaging analysis and signal processing, it is commonly believed that the null cases are independent but the nonnull cases are clustered (Logan et al., 2008). We consider such a setting. In our simulation, the total number of tests is $m = 2000$ and the proportion of nonnull hypotheses is $p = 0{\cdot}1$. Let $m_0 = m(1 - p)$. Without loss of generality, we assume that the first m_0 elements, $X^0 = (X_1, \ldots, X_{m_0})$, are null cases and the remaining $m - m_0$ elements, $X^1 = (X_{m_0+1}, \ldots, X_m)$, are nonnull cases. Under the null, $X_1, \ldots, X_{m_0}$ are independent $N(0, 1)$ observations. Under the alternative, X^1 follows a multivariate normal distribution with mean $\mu^1 = \mu 1_{m-m_0}$ and equicorrelated covariance matrix $\Sigma = (1 - \rho)I + \rho J$, where 1_{m-m_0} is a vector of ones, I is the identity matrix and J is a square matrix of ones.

We vary the critical value t from 1·95 to 4 and calculate the false discovery rate by averaging over 2000 replications. The results are summarized in Fig. 1(c)–(d). Figure 1(c) shows the weakly correlated case, where $\mu = 2{\cdot}5$ and $\rho = 0{\cdot}1$; Fig. 1(d) shows the strongly correlated case, where $\mu = 2{\cdot}5$ and $\rho = 0{\cdot}9$. We can see that under weak correlation, the false discovery rate is monotonically decreasing in the threshold. In contrast, under strong correlation, Assumption 1 is violated because the false discovery rate first decreases, then increases, and finally decreases with the critical value t.

Inspired by a reviewer's comment, we investigated the relationship between the marginal false discovery rate and the false discovery rate under dependency. The two error measures can be very different when the tests are highly correlated. We present the results related to the false discovery rate here, since it is more commonly used; see the Supplementary Material for results on the marginal false discovery rate.

3·3. *A real-data example*

In this subsection we present an example from a DNA methylation study. The study was conducted by Teschendorff et al. (2010) to investigate the mechanisms of diabetic nephropathy, which often develops in patients with chronic diabetes. The dataset contains 96 cases and 98 controls on 25 880 markers. We are interested in identifying markers at which the proportions of methylation are different between the cases and the controls. A two-sample t-statistic is calculated for each gene, and the t-statistics are then converted to p-values.

Figure 2(a) shows the histogram of p-values overlaid with the density estimate $\hat{g}(t)$. The mixture distribution is $G(t) = (1 - p)t + pG_1(t)$. Condition (5) implies that $G_1(t)$ is concave. Hence a roughly decreasing pattern is expected for $\hat{g}(t)$ should the monotone likelihood ratio condition hold. However, we can see that $\hat{g}(t)$ first increases and then decreases, indicating that condition (5) is violated. A direct consequence is that the false discovery rate is not a monotone function of the p-value cut-off, which makes the search for the optimal threshold impossible. To see this, we apply the q-value false discovery rate approach (Storey, 2002) to estimate the nonnull proportion as $\hat{p} = 0{\cdot}49$. The false discovery rate for a given cut-off t can be approximately estimated as $\widehat{\text{FDR}}(t) = (1-\hat{p})t/\{m^{-1}\sum_i I(P_i < t)\}$. Figure 2(b) plots the false discovery rate estimates against a grid of p-value cut-offs; the graph first decreases and then increases. The pattern is very counter-intuitive, and, moreover, the results are uninterpretable since a larger p-value may correspond to a smaller false discovery rate level in the range between 0 and 0·20. We suspect that in this dataset the p-value ranking is inappropriate. In other words, small p-values do not necessarily indicate strong evidence against the null. This example shows that multiple testing results should be interpreted with caution. In particular, further investigation is required into possible effects of the normality assumption, heteroscedasticity, grouping and dependence among tests.

4. GENERALIZED MONOTONE RATIO CONDITION

Let $T = (T_1, \ldots, T_m)$ be the test statistics, and let $\theta = (\theta_1, \ldots, \theta_m)$ be Ber(p_i) variables with $p_i = \text{pr}(\theta_i = 1)$ for $i = 1, \ldots, m$. Suppose that $T_i \mid \theta_i \sim (1-\theta_i)G_{i0} + \theta_i G_{i1}$. Condition (5) requires all of the G_{i0}, and the G_{i1}, to be identical. Now we generalize condition (5) by allowing the G_{i0} and G_{i1} to vary across i so that we can handle a wider class of test statistics, such as weighted p-values (Genovese et al., 2006) and the local index of significance (Sun & Cai, 2009). Let g_{i0} and g_{i1} be the corresponding densities. Define the following generalized monotone ratio condition:

$$\frac{\sum_{i=1}^m p_i g_{i1}(t)}{\sum_{i=1}^m (1-p_i) g_{i0}(t)} \text{ is monotonically decreasing in } t. \tag{7}$$

The next theorem generalizes Proposition 1.

THEOREM 2. *Consider a decision rule of the form* $\delta = \{\delta_i : i = 1, \ldots, m\} = \{I(T_i < t) : i = 1, \ldots, m\}$. *If* T_i *satisfies* (7), *then:* (i) $\text{MFDR}(t)$ *increases in* t*;* (ii) $\text{MFNR}(t)$ *decreases in* t*;* (iii) $\text{MFNR}(t)$ *decreases in* $\text{MFDR}(t)$.

Next, we propose a class of test statistics which always satisfy the generalized condition (7). Let $\theta_i \sim \text{Ber}(p_i)$. Suppose that we observe $X = (X_1, \ldots, X_m)$ from the model

$$X = \mu + \epsilon, \tag{8}$$

where $\mu_i \mid \theta_i \sim (1-\theta_i) f_{i0}(\mu) + \theta_i f_{i1}(\mu)$ and $E(\epsilon) = 0$. Use of the notation $f_{i0}(\mu)$ and $f_{i1}(\mu)$ allows the null and nonnull distributions to vary with i. We also assume that θ and ϵ follow some multivariate distribution with arbitrary covariance matrices Σ_θ and Σ_ϵ, respectively. The next theorem derives a class of test statistics for model (8) which always obey (7).

THEOREM 3. *Consider model* (8). *Denote by* Θ *the collection of all model parameters* p_i, f_{i0}, f_{i1}, Σ_θ *and* Σ_ϵ. *Suppose an oracle knows* Θ. *Let* $T_{\text{OR}}^i = \text{pr}_\Theta(\theta_i = 0 \mid X)$ *be the oracle test statistic and let* $T_{\text{OR}} = \{T_{\text{OR}}^i : i = 1, \ldots, m\}$. *Then* T_{OR} *satisfies condition* (7).

The oracle statistic involves unknown parameters which require accurate estimation in practice. In situations where Θ and T_{OR} can be estimated well, Theorem 3 can be directly applied to avoid the failure of condition (7). For example, suppose that $X_1, \ldots, X_m$ form a random sample from the mixture density $f(x) = (1-p) f_0(x) + p f_1(x)$. Then condition (7) reduces to condition (5) and T_{OR}^i reduces to the local false discovery rate $\text{LFDR}(X_i) = (1-p) f_0(X_i)/f(X_i)$, which, by Theorem 3, obeys (5). Similarly, test statistics which obey (7) can be derived in, for example, hidden Markov models and the multigroup model considered by Efron (2008) and Cai & Sun (2009). In the Supplementary Material we revisit Example 1 to demonstrate an important application of Theorem 3. Theorems 2 and 3 together provide a useful framework for choosing proper test statistics in practice. However, the scope of our result is limited, since strong distributional assumptions are needed and the estimation of unknown Θ can be very challenging. By revealing the interesting connection between estimation and testing in problems arising from model (8), we show that much research is still needed towards a more general estimation and testing theory in large-scale simultaneous inference.

5. Discussion

The monotone likelihood ratio condition plays an important role in optimal thresholding theory for false discovery rate analysis; it guarantees that precise false discovery rate control leads to the most powerful test. We provide important scenarios where this seemingly reasonable assumption is violated and discuss the consequence of violation using both simulated and real data. Although our discussion primarily considers the false discovery rate, we expect that similar issues exist for other important error measures in multiple testing (Romano & Wolf, 2007). We argue that the tacit assumption, Assumption 1, should be scrutinized in practice and that optimal thresholds in multiple testing need to be carefully interpreted.

The failure of the monotonicity condition can result from improper model assumptions such as homoscedasticity and normality of the distributions, as well as independence and homogeneity among the tests. We discussed a possible framework for choosing test statistics to avoid failure of the condition. However, our theory is far from solving the problem completely. Instead, the main goal is to demonstrate why one should be very careful with regard to unknown model aspects and distributional issues in analysing complex datasets from modern scientific applications, which commonly consist of a large number of variables with a small sample size. Our investigation reveals that, in addition to the existing list of concerns, the seemingly reasonable monotonicity assumption can be violated unexpectedly. Hence precise inference in the large p, small n paradigm is very difficult, and one should always proceed with caution.

Acknowledgement

We thank the editor, an associate editor and two referees for helpful suggestions that helped to streamline this paper. We thank Michael Wu for providing us with the DNA methylation data. This research was supported by the U.S. National Science Foundation and the U.S. National Institutes of Health. M. R. Kosorok is also affiliated with the Department of Statistics and Operations Research at the University of North Carolina.

Supplementary material

Supplementary material available at *Biometrika* online includes proofs of all the theorems, simulation studies on grouped hypothesis testing and marginal false discovery rate analysis, and a revisit to Example 1.

References

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289–300.

Benjamini, Y. & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Statist.* **25**, 60–83.

Benjamini, Y., Krieger, A. M. & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**, 491–507.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–88.

Cai, T. T. & Sun, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Am. Statist. Assoc.* **488**, 1467–81.

Cao, H. & Kosorok, M. R. (2011). Simultaneous critical values for t-tests in very high dimensions. *Bernoulli* **17**, 347–94.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Am. Statist. Assoc.* **102**, 93–103.

EFRON, B. (2008). Simultaneous inference: When should hypothesis testing problems be combined? *Ann. Appl. Statist.* **2**, 197–223.

EFRON, B., TIBSHIRANI, R., STOREY, J. D. & TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Am. Statist. Assoc.* **96**, 1151–60.

GENOVESE, C. & WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Statist. Soc.* B **64**, 499–517.

GENOVESE, C. & WASSERMAN, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32**, 1035–61.

GENOVESE, C. R., ROEDER, K. & WASSERMAN, L. (2006). False discovery control with p-value weighting. *Biometrika* **93**, 509–24.

KOSOROK, M. & MA, S. (2007). Marginal asymptotics for the "large p, small n" paradigm: With application to microarray data. *Ann. Statist.* **35**, 1456–86.

LOGAN, B., GELIAZKOVA, M. & ROWE, D. (2008). An evaluation of spatial thresholding techniques in fMRI analysis. *Hum. Brain Map.* **29**, 1379–89.

ROMANO, J. P. & WOLF, M. (2007). Control of generalized error rates in multiple testing. *Ann. Statist.* **35**, 1378–408.

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Statist. Soc.* B **64**, 479–98.

SUN, W. & CAI, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Statist. Assoc.* **102**, 901–12.

SUN, W. & CAI, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Statist. Soc.* B **71**, 393–424.

TESCHENDORFF, A. E., MENON, U., GENTRY-MAHARAJ, A., RAMUS, S. J., WEISENBERGER, D. J., SHEN, H., CAMPAN, M., NOURSHMEHR, H., BELL, C. G., MAXWELL, A. P., SAVAGE, D. A., MUELLER-HOLZNER, E., MARTH, C., KOCJAN, G., GAYTHER, S. A., JONES, A., BECK, S., WAGNER, W., LAIRD, P. W., JACOBS, I. J. & WIDSCHWENDTER, M. (2010). Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.* **20**, 440–6.

WU, W. B. (2008). On false discovery control under dependence. *Ann. Statist.* **36**, 364–80.

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