Published in final edited form as:

Biometrika. 2013 ; 100(2): 495–502. doi:10.1093/biomet/ast001.

The optimal power puzzle: scrutiny of the monotone likelihood ratio assumption in multiple testing

Hongyuan Cao,

Department of Health Studies, University of Chicago, Chicago, Illinois 60637, U.S.A

Wenguang Sun, and

Department of Information and Operation Management, Marshall School of Business, University of Southern California, Los Angeles, California 90089, U.S.A

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

Hongyuan Cao: hycao@uchicago.edu; Wenguang Sun: wenguang@marshall.usc.edu; Michael R. Kosorok: kosorok@unc.edu

Summary

In single hypothesis testing, power is a non-decreasing function of type I error rate; hence it is desirable to test at the nominal level exactly to achieve optimal power. The puzzle lies in 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 implicitly used 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 closely examine 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 $S = S_0 \cup S_1$: $\delta = 0$ 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 , 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 $\alpha(t) = \alpha_0$, should satisfy $\alpha(t^*) = \alpha_0$. In other words, we wish to 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 $H_1, \dots, H_m$ simultaneously based on a random vector $X = (X_1, \dots, X_m)$. Let $\theta_1, \dots, \theta_m$ be independent and identically distributed Bernoulli (p) random variables, where $\theta_i = 0$ if H_i is a null and $\theta_i = 1$ otherwise. Assume

$$X_i \sim (1 - \theta_i)F_0 + \theta_i F_1, \quad (1)$$

where F_0 and F_1 are the null and non-null 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 $\delta = (\delta_1, \dots, \delta_m) \in \{0, 1\}^m$, where $\delta_i = 1$ if we reject H_i and $\delta_i = 0$ otherwise. As an example, consider a testing rule which 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 non-null 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

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

The false non-discovery rate, missed discovery rate, and average power can be used to describe the power of an false discovery rate procedure:

$$\begin{aligned} \text{FNR}(t) &= E \left[\frac{\sum_{i=1}^m (1 - \delta_i) \theta_i}{\{\sum_{i=1}^m (1 - \delta_i)\} \vee 1} \right] = \frac{p\{1 - G_1(t)\}}{p\{1 - G_1(t)\} + (1-p)(1-t)} + O(m^{-1/2}), \\ \text{MDR}(t) &= E \left\{ \frac{\sum_{i=1}^m \theta_i (1 - \delta_i)}{(\sum_{i=1}^m \theta_i) \vee 1} \right\} = 1 - G_1(t) + O(m^{-1/2}), \end{aligned} \quad (3)$$

$\text{AP}(t) = 1 - \text{MDR}(t)$. Similar to the situation in single hypothesis testing, it is often assumed in the multiple testing literature that

$$\text{FDR}(t) \text{ increases in } t \text{ and } \text{FNR}(t) \text{ decreases in } t; \text{ therefore } \text{FNR}(t) \text{ decreases in } \text{FDR}(t). \quad (4)$$

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 cutoff t^* should solve the equation

$$\frac{(1-p)t^*}{(1-p)t^* + pG_1(t^*)} = \alpha. \quad (5)$$

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 (4) always holds. This question yields a logical gap in methodological development. If (4) 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 (5) is not optimal and all attempts to achieve precise false discovery rate control must fail. Surprisingly, (4) can be violated in various important scenarios.

2. The Monotone Likelihood Ratio Condition

Consider a decision rule $\delta = (\delta_1, \dots, \delta_m)$, where $\delta_i = I(T_i < t)$. Various statistics T_i have been proposed for multiple testing in the literature, 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 is desirable to develop a general principle which guarantees that (4) is fulfilled by different T_i . To focus on the main idea, we assume for the moment that 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 $i = 1, \dots, m$. Let $g_j(t) = (d/dt)G_j(t)$ ($j = 0, 1$) be the corresponding conditional densities. The monotone likelihood ratio condition can be stated as

$$g_1(t)/g_0(t) \text{ is monotonically decreasing in } t. \quad (6)$$

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 (6) if the null p -value distribution is uniform. A significant advantage of condition (6), compared to condition (4), 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 equations (2)–(3) are referred to as the marginal false discovery rate and marginal false non discovery rate, respectively. The property of a testing rule is essentially characterized by these approximations. We mainly use these marginal measures hereinafter to simplify our discussion while still preserving the key features of the problem. The main finding is that condition (6), 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 (6) is fulfilled.

Proposition 1—(Sun & Cai, 2007). Consider random mixture model (1). Let $T_i = T(X_i)$ be the test statistic and $\delta(T, t) = \{\delta_i : i = 1, \dots, m\} = \{I(T(X_i) < t) : i = 1, \dots, m\}$ the testing rule. If T_i satisfies condition (6), then (i) $\text{mFDR}(t)$ increases in t ; (ii) $\text{mFNR}(t)$ decrease in t ; and (iii) $\text{mFNR}(t)$ decrease in $\text{mFDR}(t)$. In particular, results (i), (ii) and (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 conditions (4) and (6) are violated. First consider a heteroscedastic normal mixture model

$$Z_i | \theta_i \sim (1 - \theta_i)N(0, 1) + \theta_i N(\mu, \sigma^2), \quad i = 1, \dots, m, \quad (7)$$

where $\theta_1, \dots, \theta_m$ are independent Bernoulli(p) variables. The next proposition shows that the standard approach, which thresholds the z -value or equivalently, the one-sided p -value $P_i = \Pr\{N(0, 1) > Z_i\}$, may fail to fulfill condition (6).

Theorem 1—Consider the normal mixture model (7). Define the one-sided p -value $P_i = \Pr\{N(0, 1) > Z_i\}$. Let $\delta = (\delta_i: i = 1, \dots, m)$ be a testing rule, where $\delta_i = I(P_i < t)$. Then condition (6) always holds when $\sigma = 1$ but fails when $\sigma < 1$.

The heteroscedastic model (7) can arise from applications such as sign tests. Suppose we want to test whether random variable Y_i has median 0 based on replicated observations $Y_{i1}, \dots, Y_{in}$, ($i = 1, \dots, m$). Let $q = \Pr(Y_i > 0)$. The hypotheses can be stated as $H_{0i}: q = 0.5$ versus $H_{1i}: q \neq 0.5$. 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. Next we provide a numerical example to illustrate the failure of the condition in a heteroscedastic model.

Example 1—We generate $m = 2000$ independent Bernoulli(p) variables $\theta_1, \dots, \theta_m$ with $p = 0.1$, and generate Z_i according to model (7) with $\mu = 2.5$. The one-sided p -value is obtained as $P_i = \Pr\{N(0, 1) > Z_i\}$. We vary the critical value t from 1.95 to 4 and calculate false discovery proportion $\text{FDP}(t)$. Then $\text{FDR}(t)$ is obtained by averaging the $\text{FDP}(t)$ over 2000 replications. The results are summarized in the first row of Fig. 1. We can see that when $\sigma = 1$, $\text{FDR}(t)$ decreases monotonically in t . However, when $\sigma = 0.5$, $\text{FDR}(t)$ first decreases and then increases in t . The violation of monotonicity leads to testing results that are not interpretable. For example, the right panel of the first row of Fig. 1 suggests that if we threshold at $t = 3.8$, the false discovery rate is 0.12, but if we threshold at $t = 3.0$, the false discovery rate is 0.07. In fact larger threshold does not necessarily control false discovery rate at a lower level when $\sigma < 1$. This heteroscedasticity resulted in the violation of (4) and (6).

3.2. Correlated tests

This section discusses the violation of condition (6) 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 (6) can be violated under strong dependency.

Example 2—Suppose we observe $X = (X_1, \dots, X_m)$ from the model

$$X = \mu + \varepsilon, \quad (8)$$

and want to identify non-zero elements in $\mu = (\mu_1, \dots, \mu_m)$. In many important applications such as imaging analysis and signal processing, it is commonly believed that the null cases are independent but the non-null 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 non-null hypotheses is $p = 0.1$. Let $m_0 = m(1 - p)$. Without loss of generality, we assume that the first m_0 elements $X^0 = (X_1, \dots, X_{m_0})$ are null cases and the remaining $m - m_0$ elements $X^1 = (X_{m_0+1}, \dots, X_m)$ are non-null cases. Under the null, $X_1, \dots, X_{m_0}$ are independent observations from $N(0, 1)$. Under the alternative, X^1 follows a multivariate normal distribution with mean $\mu^1 = \mu 1_{m-m_0}$ and equi-correlated variance 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 the second row of Figure 1. The left and right panels consider the weakly correlated case where $\mu = 2.5$ and $\rho = 0.1$ and the strongly correlated case where $\mu = 2.5$ and $\rho = 0.9$. We can see that under weak correlation, the false discovery rate is monotonically decreasing in the threshold. In contrast, under strong correlation, condition (4) is violated because the false discovery rate first decreases and then increases and finally decreases in the critical value t .

Inspired by a comment from a reviewer, we investigated the relationship between the marginal false discovery rate and 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 more results on the marginal false discovery rate.

3.3. A real data example

Next 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 data set contains 96 cases and 98 controls on 25880 markers. We are interested in identifying markers at which the proportions of methylation are different between cases and controls. A two sample t -statistic is calculated for each gene and the t -statistics are then converted to p -values.

The left panel of Figure 2 contains 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 (6) 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 (6) is violated. A direct consequence is that the false discovery rate is not a monotone function of the p -value cutoff, which makes the search for optimal threshold impossible. To see this, we apply the q -value false discovery rate approach (Storey, 2002) to estimate the non-null proportion as $p \hat{=} 0.49$. The false discovery rate for a given cutoff t can be approximately estimated as

$\widehat{\text{FDR}}(t) = (1 - \hat{p})t / \{m^{-1} \sum_i I(P_i < t)\}$. The right panel of Figure 2 plots the false discovery rate estimates against a grid of p -value cutoffs; it 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 data set 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 the multiple testing results should be interpreted with caution. In particular, further investigation is required for possible effects of the normality assumption, heteroscedasticity, grouping and dependence among tests.

4. Generalized Monotone Ratio Condition

Let $T = (T_1, \dots, T_m)$ be the test statistics and $\theta = (\theta_1, \dots, \theta_m)$ be Bernoulli(p_i) variables with $p_i = \text{pr}(\theta_i = 1)$, $i = 1, \dots, m$. Suppose that $T_i | \theta_i \sim (1 - \theta_i)G_{i0} + \theta_i G_{i1}$. Condition (6) requires all G_{i0} 's (and G_{i1} 's) to be identical. Now we generalize condition (6) by allowing 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

$$\frac{\sum_{i=1}^m p_i g_{i1}(t)}{\sum_{i=1}^m (1-p_i) g_{i0}(t)} \text{ is monotonically decreasing in } t. \quad (9)$$

The next theorem generalizes Proposition 1.

Theorem 2—Consider a decision rule of the form $\delta = \{\delta_i: i = 1, \dots, m\} = \{I(T_i < t): i = 1, \dots, m\}$. If T_i satisfies (9), then (i) mFDR(t) increases in t ; (ii) mFNR(t) decreases in t ; and (iii) mFNR(t) decreases in mFDR(t).

Next we propose a class of test statistics which always satisfy the generalized condition (9). Let $\theta_i \sim \text{Bernoulli}(p_i)$. Suppose we observe $X = (X_1, \dots, X_m)$ from the following model

$$X = \mu + \varepsilon, \quad (10)$$

where $\mu_i | \theta_i \sim (1 - \theta_i)f_{i0}(\mu) + \theta_i f_{i1}(\mu)$ and $E(\varepsilon) = 0$. The use of $f_{i0}(\mu)$ and $f_{i1}(\mu)$ allows the null and non-null 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 (10) which always obey (9).

Theorem 3—Consider model (10). Denote by Θ the collection of all model parameters $p_i, f_{i0}, f_{i1}, \Sigma_\theta$ and Σ_ε . Suppose an oracle knows Θ . Let $T_{OR}^i = \text{pr}_{\Theta}(\theta_i = 0 | X)$ be the oracle test statistic and $T_{OR} = \{T_{OR}^i: i = 1, \dots, m\}$. Then T_{OR} satisfies condition (9).

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 (9). For example, suppose $X_1, \dots, X_m$ are a random sample from mixture density $f(x) = (1 - p)f_0(x) + pf_1(x)$. Then condition (9) reduces to

condition (6) 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 (6). Similarly, test statistics which obey (9) can be derived, for example, in hidden Markov models and the multi-group 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 (10), 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 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 (4) should be scrutinized in practice and optimal thresholds in multiple testing need to be carefully interpreted.

The failure of the monotonicity condition can be resulted 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 the failure of condition. However, our theory is far from solving the problem completely. Instead, the main goal is to demonstrate why one should be very careful on unknown model aspects and distributional issues in analyzing complex data sets 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 we should always proceed with caution.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the editor, an associate editor and two referees for helpful suggestions that streamlined this paper. We thank Michael Wu for providing us the DNA methylation data. This research was supported by grants from the U.S. National Science Foundation and U.S. National Institute of Health.

References

- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc B*. 1995; 57:289–300.
- Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*. 2000; 25:60–83.
- Benjamini Y, Krieger AM, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*. 2006; 93:491–507.
- Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Statist*. 2001; 29:1165–1188.
- Cai TT, Sun W. Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J Amer Statist Assoc*. 2009; 488:1467–1481.
- Cao H, Kosorok MR. Simultaneous critical values for t -tests in very high dimensions. *Bernoulli*. 2011; 17:347–394. [PubMed: 21572564]
- Efron B. Correlation and large-scale simultaneous significance testing. *J Amer Statist Assoc*. 2007; 102:93–103.
- Efron B. Simultaneous inference: When should hypothesis testing problems be combined? *Annals of Applied Statistics*. 2008; 2:197–223.

- Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. *J Amer Statist Assoc.* 2001; 96:1151–1160.
- Genovese C, Wasserman L. Operating characteristics and extensions of the false discovery rate procedure. *J R Stat Soc B.* 2002; 64:499–517.
- Genovese C, Wasserman L. A stochastic process approach to false discovery control. *Ann Statist.* 2004; 32:1035–1061.
- Genovese CR, Roeder K, Wasserman L. False discovery control with p-value weighting. *Biometrika.* 2006; 93:509–524.
- Kosorok M, Ma S. Marginal asymptotics for the “large p , small n ” paradigm: with application to microarray data. *Ann Statist.* 2007; 35:1456–1486.
- Logan B, Geliazkova M, Rowe D. An evaluation of spatial thresholding techniques in fmri analysis. *Human Brain Mapping.* 2008; 29:1379–1389. [PubMed: 18064589]
- Romano JP, Wolf M. Control of generalized error rates in multiple testing. *Ann Statist.* 2007; 35:1378–1408.
- Storey JD. A direct approach to false discovery rates. *J R Stat Soc B.* 2002; 64:479–498.
- Sun W, Cai TT. Oracle and adaptive compound decision rules for false discovery rate control. *J Amer Statist Assoc.* 2007; 102:901–912.
- Sun W, Cai TT. Large-scale multiple testing under dependence. *J R Stat Soc B.* 2009; 71:393–424.
- Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, Campan M, Nourshmehr H, Bell CG, Maxwell AP, Savage DA, Mueller-Holzner E, Marth C, Kocjan G, Gayther SA, Jones A, Beck S, Wagner W, Laird PW, Jacobs IJ, Widschwendter M. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Research.* 2010:440–446. [PubMed: 20219944]
- Wu WB. On false discovery control under dependence. *Ann Statist.* 2008; 36:364–380.

Fig. 1.

The first row corresponds to heteroscedastic models with $\sigma = 1$ (left) and $\sigma = 0.5$ (right); The second row corresponds to correlated tests with weak correlation (left) and strong correlation (right)

Fig. 2.
The left plot is the histogram and density of p-values; the right plot is the estimated false discovery rate