Biometrika (2011), **98**, 2, *pp.* 273–290
© 2011 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/asr003

Sample size and power analysis for sparse signal recovery in genome-wide association studies

By JICHUN XIE

Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, Pennsylvania, 19010, U.S.A.
jichun@mail.med.upenn.edu

T. TONY CAI

Department of Statistics, University of Pennsylvania, Philadelphia, Pennsylvania, 19010, U.S.A.
tcai@wharton.upenn.edu

and HONGZHE LI

Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, Pennsylvania, 19010, U.S.A.
hongzhe@upenn.edu

Summary

Genome-wide association studies have successfully identified hundreds of novel genetic variants associated with many complex human diseases. However, there is a lack of rigorous work on evaluating the statistical power for identifying these variants. In this paper, we consider sparse signal identification in genome-wide association studies and present two analytical frameworks for detailed analysis of the statistical power for detecting and identifying the disease-associated variants. We present an explicit sample size formula for achieving a given false non-discovery rate while controlling the false discovery rate based on an optimal procedure. Sparse genetic variant recovery is also considered and a boundary condition is established in terms of sparsity and signal strength for almost exact recovery of both disease-associated variants and nondisease-associated variants. A data-adaptive procedure is proposed to achieve this bound. The analytical results are illustrated with a genome-wide association study of neuroblastoma.

Some key words: False discovery rate; False non-discovery rate; High-dimensional data; Multiple testing; Oracle exact recovery.

1. Introduction

Genome-wide association studies have emerged as an important tool for discovering regions of the genome that harbour genetic variants conferring risk for complex diseases. Such studies often involve scanning hundreds of thousands of single-nucleotide polymorphism markers in the genome. Many novel genetic variants have been identified from recent genome-wide association studies, including variants for age-related macular degenerative diseases (Klein et al., 2005), breast cancer (Hunter et al., 2007) and neuroblastoma (Maris et al., 2008). The Wellcome Trust Case-Control Consortium (2007) published a study of seven diseases using 14 000 cases and 3000 shared controls and has identified several markers associated with each of these complex diseases. The success of these studies has provided solid evidence that the genome-wide association studies represent a powerful approach to the identification of genes involved in common human diseases.

Due to the genetic complexity of many common diseases, it is generally assumed that multiple genetic variants are associated with disease risk. The key question is to identify true disease-associated markers from the multitude of nondisease-associated markers. The most common approach to this is to perform a single marker score test derived from the logistic regression model and to control for multiplicity of testing using stringent criteria such as the Bonferroni correction (McCarthy et al., 2008). Similarly, the most commonly used approach for sample size analysis or power analysis for such large-scale association studies is based on detecting markers of given odds ratios using a conservative Bonferroni correction. However, this correction is often too conservative for large-scale multiple testing problems, leading to reduced power in detecting disease-associated markers. Analytical and simulation studies by Sabatti et al. (2003) have shown that the false discovery rate procedure of Benjamini & Hochberg (1995) can effectively control the false discovery rate for the dependent tests encountered in case–control association studies and increase power.

Despite the success of genome-wide association studies, questions remain as to whether current sample sizes are large enough to detect most or all of the disease markers. This is related to power and sample size analysis and is closely related to sparse signal detection and discovery. However, power analysis has not been addressed fully for simultaneous testing of hundreds of thousands of null hypotheses. Efron (2007) presented an important alternative for power analysis for large-scale multiple testing problems in the framework of local false discovery rate. Gail et al. (2008) investigated the probability of detecting the disease-associated markers in case–control genome-wide association studies in the top T largest chi-square values from the trend tests of association. The detection probability is related to false non-discovery and T is related to false discovery, although these terms are not explicitly used in that paper.

The goal of the present paper is to provide an analytical study of the power and sample size issues in genome-wide association studies. We treat the vector of the score statistics across all the markers as a sequence of Gaussian random variables. Since we expect only a small number of markers to be associated with disease, the true mean of the vector of score statistics should be very sparse, although the degree of sparsity is unknown. Our goal is to recover those sparse and relevant markers. This is deeply connected with hypothesis testing in the context of multiple comparisons and false discovery rate control. Both false discovery rate control and sparse signal recovery have been areas of intensive research in recent years. Sun & Cai (2007) showed that the large-scale multiple testing problem has an equivalent weighted classification formulation, in the sense that the optimal solution to the multiple testing problem is also the optimal decision rule for the weighted classification problem. They further proposed an optimal false discovery rate controlling procedure that minimizes the false non-discovery rate. Donoho & Jin (2004) studied the detection of sparse heterogeneous mixtures using higher criticism, focusing on testing the global null hypothesis against the alternative where only a fraction of the data comes from a normal distribution with a common nonnull mean. It is particularly important that the detectable region is identified on the amplitude or sparsity plane so that higher criticism can completely separate the two hypotheses asymptotically.

In this paper, we present analytical results on sparse signal recovery in case–control genome-wide association studies. We investigate two frameworks for detailed analysis of the statistical power for detecting and identifying the disease-associated markers. In a similar setting as in Gail et al. (2008), we first present an explicit sample-size formula for achieving a given false non-discovery rate while controlling the false discovery rate based on the optimal false discovery rate controlling procedure of Sun & Cai (2007). This provides results on how odds ratios and marker allele frequencies affect the power of identifying the disease-associated markers. We also consider sparse marker recovery, establishing the theoretical boundary for almost exact recovery of both the disease-associated markers and nondisease-associated markers. Our results further extend the amplitude or sparsity boundary of Donoho & Jin (2004) for almost exact recovery of the signals. Finally, we construct a data adaptive procedure to achieve this bound.

2. Problem setup and score statistics

We consider a case-control genome-wide association study with m markers genotyped, where the minor allele frequency for marker i is p_i. Let $G_i = 0, 1, 2$ be the number of minor alleles at marker i $(i = 1, \ldots, m)$, where $G_i \sim \text{Bin}(2, p_i)$. Assume that the total sample size is n and $n_1 = rn$ and $n_2 = (1 - r)n$ are the sample sizes for cases and controls. Let $Y = 1$ for diseased and $Y = 0$ for nondiseased individuals. Suppose that in the source population, the probability of disease is given by

$$\text{logit}\{\text{pr}(Y \mid G_i, i = 1, \ldots, m)\} = a + \sum_{i=1}^{m} G_i b_i,$$

where only a very small fraction of the b_is are nonzero. Gail et al. (2008) showed that for rare diseases or for more common diseases over a confined age range such as 10 years, for case-control population, if the markers are independent of each other, it follows that

$$\text{logit}\{\text{pr}(Y \mid G_i)\} = a_i + G_i b_i \quad (i = 1, \ldots, m), \tag{1}$$

approximately. This implies that a logistic regression model can be fitted for each single marker i separately.

Based on the results of Prentice & Pyke (1979), the maximum likelihood estimation for a cohort study applied to case-control data with model (1) yields a fully efficient estimate of b_i and a consistent variance estimate. For a given case-control study, let j be the index for individuals in the samples, Y_j be the disease status of the jth individual and G_{ij} be the genotype score for the jth individual at the ith marker. The profile score statistic to test $\text{H}_i^0 : b_i = 0$ can be written as

$$U(b_i) = \sum_j G_{ij}(Y_j - \bar{Y}),$$

where $\bar{Y} = \sum_{j=1}^{n} Y_j$. In the following, we let pr_{i0}, E_{i0}, var_{i0} denote the probability, expectation and variance calculated under the null hypothesis $\text{H}_i^0 : b_i = 0$, and pr_{i1}, E_{i1}, var_{i1} those calculated under the alternative hypothesis $\text{H}_i^1 : b_i = 1$. Under $\text{H}_i^0 : b_i = 0$, $\text{pr}_{i0}(Y_j = y \mid G_{ij} = g) = K\,\text{I}(y = 1) + (1 - K)\,\text{I}(y = 0)$, where $\text{I}(\cdot)$ is the indicator function. We have $E_{i0}\{U(b_i) \mid Y\} = 0$, $\text{var}_{i0}\{U(b_i) \mid Y\} = 2p_i(1 - p_i)r(1 - r)n$.

Under the alternative, if b_i is known,

$$u_{1gi} = \text{pr}_{i1}(Y_j = 1 \mid G_{ij} = g) = \frac{\exp(a_i + b_i g)}{1 + \exp(a_i + b_i g)},$$

$$u_{0gi} = \text{pr}_{i1}(Y_j = 0 \mid G_{ij} = g) = \frac{1}{1 + \exp(a_i + b_i g)},$$

where a_i can be determined by $\sum_{g=0}^{2} u_{1gi}\mathrm{pr}(G_{ij}=g)=D$, with D being the disease prevalence and $\mathrm{pr}(G_{ij}=g)=p_i^2\,\mathrm{I}(g=2)+2p_i(1-p_i)\,\mathrm{I}(g=1)+(1-p_i)^2\,\mathrm{I}(g=0)$. Then

$$w_{ygi}=\mathrm{pr}_{i1}(G_{ij}=g\mid Y_j=y)=\frac{u_{y,g,i}\mathrm{pr}(G_{ij}=g)}{\sum_{g'=0}^{2}u_{yg'i}\mathrm{pr}(G_{ij}=g')}$$

and

$$E_{i1}\{U(b_i)\mid Y\}=r(1-r)n\{E_{i1}(G_{ij}\mid Y_j=1)-E_{i1}(G_{ij}\mid Y_j=0)\},$$
$$\mathrm{var}_{i1}\{U(b_i)\mid Y\}=r(1-r)n\{(1-r)\,\mathrm{var}_{i1}(G_{ij}\mid Y_j=1)+r\,\mathrm{var}_{i1}(G_{ij}\mid Y_j=0)\},$$

where $E_{i1}(G_{ij}\mid Y_j=y)=\sum_{g=0}^{2}g w_{ygi}$, $E_{i1}(G_{ij}^2\mid Y_j=y)=\sum_{g=0}^{2}g^2 w_{ygi}$, and

$$\mathrm{var}_{i1}(G_{ij}\mid Y_j=y)=E_{i1}(G_{ij}^2\mid Y_j=y)-\{E_{i1}(G_{ij}\mid Y_j=y)\}^2.$$

Let $X_i=U(b_i)/\mathrm{var}_{i0}\{U(b_i)|Y\}^{1/2}$ be the score statistic for testing the association between the marker i and the disease. If b_i is known, then under H_i^0, $X_i\sim\mathrm{N}(0,1)$ and under H_i^1, $X_i\sim\mathrm{N}(\mu_{n,i},\sigma_i)$ asymptotically, where

$$\mu_{n,i}=n^{1/2}\mu_i\equiv n^{1/2}\left\{\frac{r(1-r)}{2p_i(1-p_i)}\right\}^{1/2}\{E_{i1}(G_{ij}\mid Y_j=1)-E_{i1}(G_{ij}\mid Y_j=0)\},\qquad(2)$$

$$\sigma_i=\left\{\frac{(1-r)\,\mathrm{var}_{i1}(G_{ij}\mid Y_j=1)+r\,\mathrm{var}_{i1}(G_{ij}\mid Y_j=0)}{2p_i(1-p_i)}\right\}^{1/2}.\qquad(3)$$

Given b_i, p_i, r and D, $\mu_{n,i}$ and σ_i can be determined. The alternative mean $\mu_{n,i}$ increases with the sample size n at the order $\surd n$. Intuitively, as the sample size increases, it is easier to distinguish the signals and zeros. So, the problem of power analysis and sample size analysis can be formulated as the sparse normal mean problem: we have score statistics X_i, $(i=1,\ldots,m)$ and only a very small proportion ϵ_m of them have nonzero means, and the goal is to identify those markers whose score statistics have nonzero means.

3. MARKER RECOVERY

3·1. *Effect size and false discovery rate control*

As in Genovese & Wasserman (2002) and Sun & Cai (2007), we use the marginal false discovery rate, defined as MFDR $=E(N_{01})/E(R)$, and marginal false non-discovery rate, defined as MFNR $=E(N_{10})/E(S)$, as our criteria for multiple testing, where R is the number of rejections, N_{01} is the number of nulls among these rejections, S is the number of non-rejections and N_{10} is the number of nonnulls among these nonrejections. Genovese & Wasserman (2002) and Sun & Cai (2007) showed that under weak conditions, the marginal false discovery rate (MFDR) and the false discovery rate (FDR), the marginal false non-discovery rate (MFNR) and the false non-discovery rate (FNR) are asymptotically the same in the sense that MFDR $=$ FDR $+\,O(m^{-1/2})$ and MFNR $=$ FNR $+\,O(m^{-1/2})$. In the following, we use MFDR and MFNR for our analytical analysis, but we use the notation, FDR and FNR.

Consider the sequence of score statistics X_i $(i=1,\ldots,m)$, as defined in the previous section, where under H_i^0, $X_i\sim\mathrm{N}(0,1)$ and under H_i^1, $X_i\sim\mathrm{N}(\mu_{n,i},\sigma_i^2)$, $\mu_{n,i}\neq 0$, $\sigma_i\geqslant 1$. Usually, $\mu_{n,i}$ and σ_i are different across different disease-associated markers due to different minor allele frequencies and different effect sizes as measured by the odds ratios; see (2) and (3). However, since most of the markers in genome-wide association studies are not rare and the observed odds ratios range from 1·2 to 1·5, we can reasonably assume that they are on a comparable scale. In addition, for the purpose of sample size calculation, one should always consider the worst case scenarios for all the markers , which leads to a conservative estimate of the required sample sizes. To simplify the notation and analysis and to obtain closed-form analytical results, we thus assume that all the markers considered have the same minor allele frequency and all the relevant markers have the same effect size. Specifically, we assume that under H_i^1, $X_i \sim \mathrm{N}(\mu, \sigma^2)$, $\sigma \geqslant 1$. Suppose that the proportion of nonnull effects is ϵ_m. Defining a sequence of binary latent variables, $\theta = (\theta_1, \ldots, \theta_m)$, the model under consideration can be stated as follows:

$$\begin{gathered} \theta_1, \ldots, \theta_m \text{ are independently and identically distributed as Bernoulli}(\epsilon_m), \\ \text{if } \theta_i = 0,\ X_i \sim \mathrm{N}(0, 1);\ \theta_i = 1,\ X_i \sim \mathrm{N}(\mu, \sigma^2). \end{gathered} \tag{4}$$

Assume that σ is known. Our goal is to find the minimum μ that allows us to identify θ with the false discovery rate asymptotically controlled at α_1 and the false non-discovery rate asymptotically controlled at α_2. We particularly consider the optimal false discovery rate controlling procedure of Sun & Cai (2007), which simultaneously controls the false discovery rate at α_1 and minimizes the false non-discovery rate asymptotically. Different from the p-value-based false discovery rate procedures, this approach considers the distribution of the test statistics and involves the following.

Algorithm 1. Calculating the optimal false discovery rate.

Step 1. Given the observation $X = (X_1, \ldots, X_m)$, estimate the nonnull proportion $\hat{\epsilon}_m$ using the method in Cai & Jin (2010). This estimator is based on Fourier transformation and the empirical characteristic function, which can be written as $\hat{\epsilon}_m = 1 - m^{-(1-\eta)} \sum_{i=1}^m \cos\{(2\eta \log m)^{1/2} X_i\}$, where $\eta \in (0, 1/2)$ is a tuning parameter, which needs to be small for the sparse case considered in this paper. Based on our simulations, $\eta = 10^{-4}$ works well.

Step 2. Use a kernel estimate $\hat{f}$ to estimate the mixture density f of the X_i,

$$\hat{f}(x) = m^{-1+\rho} \sum_{j=1}^{m} K\{m^{\rho}(x - X_j)\}, \qquad \rho \in (0, 1/2), \tag{5}$$

where K is a kernel function and ρ determines the bandwidth. We use $\rho = 1{\cdot}34\, m^{-1/5}$ as recommended by Silverman (1986).

Step 3. Compute $\hat{T}_i = (1 - \hat{\epsilon}_m)\varphi(X_i)/\hat{f}(X_i)$, where $\varphi(.)$ is the standard normal density function.

Step 4. Let

$$k = \max\left\{ i : i^{-1} \sum_{j=1}^{i} \hat{T}_{(j)} \leqslant \alpha_1 \right\},$$

then reject all $\mathrm{H}_{(i)}$ $(i = 1, \ldots, k)$. This adaptive step-up procedure considers the average posterior probability of being null and is adaptive to both the global feature ϵ_m and local feature of $\varphi(X_i)/f(X_i)$.

The procedure has a close connection with a weighted classification problem. Consider the loss function

$$L_\lambda(\theta,\delta)=\sum_{i=1}^{m}\{\lambda\,\mathrm{I}\,(\theta_i=0)\delta_i+\mathrm{I}\,(\theta_i=1)(1-\delta_i)\}. \tag{6}$$

The Bayes rule under the loss function (6) is

$$\theta_i(\lambda)=\mathrm{I}\left\{\Lambda(X_i)=\frac{(1-\epsilon_m)f^0(X_i)}{\epsilon_m f^1(X_i)}<\frac{1}{\lambda}\right\}, \tag{7}$$

where $f^0(x)=\varphi(x;0,1)$ and $f^1(x)=\varphi(x;\mu,\sigma^2)$. Sun & Cai (2007) show that the threshold λ is a decreasing function of false discovery rate, and for a given level α_1, one can determine a unique thresholding λ to ensure the false discovery rate level to be controlled under α_1. The following theorem gives the minimal signal μ required in order to obtain the pre-specified false discovery rate and false non-discovery rate using this optimal procedure.

THEOREM 1. *Consider the model in* (4) *and the optimal oracle procedure of* Sun & Cai (2007). *In order to control the false discovery rate under α_1 and the false non-discovery rate under α_2, the minimum μ has to be no less than $(\sigma^2-1)\hat{d}$, where $(\hat{c},\hat{d})$ are the roots of the equations*

$$\begin{cases}\Phi(c-d\sigma)-\Phi(-c-d\sigma)=(1-c_1)/(c_2-c_1),\\ \Phi(-c\sigma-d)+\Phi(-c\sigma+d)=c_1(c_2-1)/(c_2-c_1),\end{cases} \tag{8}$$

where Φ is the cumulative density function of the standard normal distribution, and

$$c_1=\frac{\alpha_1\epsilon_m}{(1-\alpha_1)(1-\epsilon_m)},\quad c_2=\frac{(1-\alpha_2)\epsilon_m}{\alpha_2(1-\epsilon_m)}.$$

As discussed in § 2, in genome-wide association studies, $\mu=\sqrt{n}\mu_1$, where μ_1 as defined in (2) can be determined by the effect of the marker, log odds ratio b, the minor allele frequency p and the case ratio or control ratio r. The following corollary provides the minimum sample size needed in case-control genome-wide association studies in order to control false discovery rate and false non-discovery rate.

COROLLARY 1. *Consider a case-control genome-wide association study, with m markers, a total sample size of n, with $n_1=nr$ cases and $n_2=(1-r)n$ controls. Assume that all the markers have the same minor allele frequency of $p_i=p$ and all the relevant markers have the same log odds ratio $b_i=b$. In order to control false discovery rate and false non-discovery rate under α_1 and α_2 asymptotically, the minimum total sample size must be at least $n=\{\hat{d}(\sigma^2-1)/\mu_1\}^2$, where $\hat{d}$ is the root of* (8), *and μ_1 and σ_1 are defined in* (2) *and* (3).

Corollary 1 gives the minimum sample size required in order to identify the markers with minor allele frequency p and effective odds ratio of b with false discovery rate less than α_1 and false non-discovery rate less than α_2 using the optimal false discovery rate controlling procedure of Sun & Cai (2007). Since this procedure asymptotically controls false discovery rate and minimizes false non-discovery rate, the sample size obtained here is a lower bound. This implies that asymptotically, one cannot control both false discovery rate and false non-discovery rate using a smaller sample size with any other procedure. In practice, α_2 should be set to less than ϵ_m, which can be achieved by not rejecting any nulls.

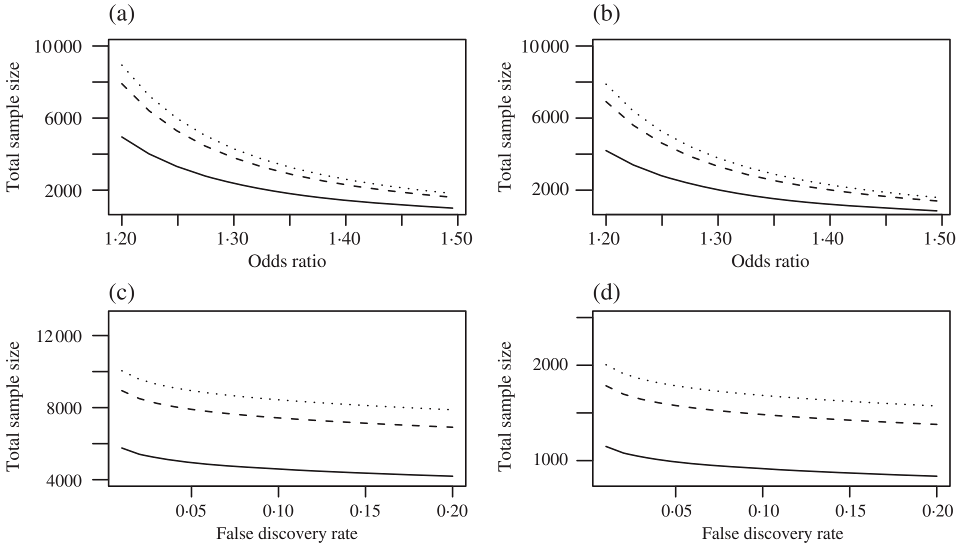

Fig. 1. Minimum sample size needed for a given false discovery rate and false non-discovery rate for a genome-wide association study with $m = 500\,000$ markers and the relevant markers proportion $\epsilon_m = 2 \times 10^{-4}$. Assume that the case proportion $r = 0{\cdot}40$ and the minor allele frequency $p = 0{\cdot}40$. (a) FDR = 0·05 and different values of odds ratio; (b) FDR = 0·20 and different values of odds ratio; (c) odds ratio of 1.20 and different values of false discovery rate; (d) odds ratio of 1.50 and different values of false discovery rate. For each plot, the solid, dash and dot lines correspond to FNR $= 10^{-4}$, 2×10^{-5} and 10^{-6}, respectively.

3·2. *Illustration*

We present the minimum sample sizes needed for different levels of false discovery rate and false non-discovery rate in Fig. 1 for $\epsilon_m = 2 \times 10^{-4}$, which corresponds to assuming 100 disease-associated markers. We assume a disease prevalence of $D = 0{\cdot}10$. As expected, the sample size needed increases as the false discovery rate level or the target false non-discovery rate level decreases. It also decreases with higher signal strength. The false non-discovery rate levels presented in the plot are very small since it cannot exceed the proportion of the relevant markers ϵ_m. When m is extremely large, ϵ_m is usually very small. In Fig. 1, $\epsilon_m = 2 \times 10^{-4}$ and false non-discovery rate we choose should be less than ϵ_m, since FNR $= \epsilon_m$ can be achieved by not rejecting any of the markers. Among the 100 relevant markers, FNR $= 10^{-4}$, 2×10^{-5} and 10^{-6} corresponds to an expected number of nondiscovered relevant markers of fewer than 50, 10 and 0·5, respectively. As a comparison, with the same sample size for FDR = 0·05 and FNR $= 10^{-4}$, 2×10^{-5} and 10^{-6}, the power of the one-sided score test using the Bonferonni correction to control the genome-wide error rate at 0·05 is 0·22, 0·64 and 0·95, respectively. We observe that the sample sizes needed strongly depend on the false non-discovery rate and the effect size of the markers, but are less dependent on false discovery rate. Similar trends were observed for $\epsilon_m = 2 \times 10^{-5}$, which corresponds to assuming 10 disease-associated markers.

We performed simulation studies to evaluate the sample size formula presented in Corollary 1. We assume that $m = 500\,000$ markers are tested with 100 being associated with disease, which corresponds to $\epsilon_m = 2 \times 10^{-4}$. We first considered the setting that all the 100 relevant markers have the same odds ratio on disease risk with a minor allele frequency of 0·40. For a prespecified FDR = 0·05 and FNR $= 10^{-4}$, we determined the sample size based on Corollary 1. We then applied the false discovery rate controlling procedure of Sun & Cai (2007) to the simulated data and examined the values of the observed false discovery rates and false non-discovery

Table 1. *Empirical false discovery rate and false non-discovery rate* (10^{-4} *unit) and their standard errors for the optimal false discovery rate procedure based on* 100 *simulations where the sample sizes are determined by Corollary* 1 *for the independent sequences with the same alternative odds ratio for* FDR $= 0{\cdot}05$ *and* FNR $= 10^{-4}$.

	Independent sequence with the same alternative odds ratios		Short-ranged dependent sequence with different alternative odds ratios	
Odds ratio	Empirical FDR	Empirical FNR	Empirical FDR	Empirical FNR
1·20	0·050 (0·028)	1·04 (0·10)	0·051 (0·030)	1·04 (0·10)
1·23	0·048 (0·028)	1·03 (0·12)	0·051 (0·032)	1·04 (0·11)
1·26	0·045 (0·030)	1·04 (0·11)	0·055 (0·033)	1·04 (0·11)
1·29	0·046 (0·030)	1·15 (0·13)	0·053 (0·031)	1·03 (0·11)
1·33	0·053 (0·033)	1·04 (0·11)	0·054 (0·032)	1·05 (0·10)
1·36	0·047 (0·029)	1·01 (0·13)	0·049 (0·029)	1·00 (0·12)
1·39	0·051 (0·030)	1·02 (0·11)	0·051 (0·033)	1·01 (0·11)
1·43	0·046 (0·033)	1·05 (0·12)	0·044 (0·029)	1·04 (0·11)
1·46	0·046 (0·027)	1·04 (0·10)	0·046 (0·025)	1·02 (0·11)
1·50	0·051 (0·027)	1·03 (0·13)	0·053 (0·033)	1·03 (0·12)

FDR, false discovery rate; FNR, false non-discovery rate.

rates. Table 1 shows the empirical false discovery rates and false non-discovery rates over 100 simulations for different values of the odds ratios. The empirical false discovery rates and false non-discovery rates are very close to the pre-specified values, indicating that the sample size formula can indeed result in the specified false discovery rate and false non-discovery rate.

Since our power and sample size calculation is derived under the assumption of common effect sizes for all associated markers and independent test statistics, we also evaluated our results under the assumption of short-range dependency and unequal effect sizes. To approximately mimic the dependency structure among markers due to linkage disequilibrium, we assumed a block-diagonal covariance structure for the score statistics with exchangeable correlation of 0·20 for block sizes of 50. We simulated marker effects with a mean odds ratio ranging from 1·20 to 1·50. For a given mean odds ratio, we obtained the sample size and simulated the corresponding marker effects by generating $\mu_{n,i}$ from $N(4{\cdot}40, 1{\cdot}00)$. We then applied the false discovery rate controlling procedure and calculated the empirical false discovery rates and false non-discovery rates for different mean odds ratios. The results are presented in Table 1, indicating that the false discovery rate controlling procedure can indeed control false discovery rates and obtain the pre-specified false non-discovery rates.

4. A PROCEDURE FOR SPARSE MARKER RECOVERY

4·1. *Oracle sparse marker discovery*

Besides controlling false discovery rate and false non-discovery rate, an important alternative in practice is to control the numbers of false discoveries and false non-discoveries. One limitation with the use of false discovery rate and false non-discovery rate in power calculations is that they do not provide a very clear idea about roughly how many markers are falsely identified and how many are missed until the analysis is conducted. In genome-wide association studies, the goal is to make the number of misidentified markers, including both the false discoveries and false non-discoveries, converge to zero when the number of the markers m and the sample sizes are sufficiently large. In this section, we investigate the conditions on the parameters $\mu = \mu_m$ and ϵ_m in (4) that can lead to almost exact recovery of the null and nonnull markers. This in turn provides a useful formula for the minimum sample size required for an almost exact recovery of disease-associated markers.

Consider (4) and a decision rule δ_i $(i = 1, \ldots, m)$. We define false discoveries as the expected number of null scores that are misclassified to the alternative, and false non-discoveries as the expected number of nonnull scores that are misclassified to null,

$$\text{FD} = E\left(\sum_{i=1}^{m} \mathrm{I}(\theta_i = 0)\delta_i\right), \quad \text{FN} = E\left(\sum_{i=1}^{m} \mathrm{I}(\theta_i = 1)(1 - \delta_i)\right).$$

Define the loss function as

$$L(\theta, \delta) = \sum_i \mathrm{I}(\theta_i = 0)\delta_i + \mathrm{I}(\theta_i = 1)(1 - \delta_i) = \sum_i \mathrm{I}(\theta_i \neq \delta_i),$$

which is simply the Hamming distance between the true θ and its estimate δ. Note that $E\{L(\theta, \delta)\} = \text{FD} + \text{FN}$.

Under (4), the Bayes oracle rule (7) with $\lambda = 1$, defined as

$$\delta_{\mathrm{OR},i} = \mathrm{I}\left\{\Lambda_{\mathrm{OR}}(X_i) = \frac{(1 - \epsilon_m) f^0(X_i)}{\epsilon_m f^1(X_i)} < 1\right\}, \tag{9}$$

minimizes the risk $E\{L(\theta, \delta)\}$. We consider the condition on μ_m and ϵ_m such that this risk converges to zero as m goes to infinity, where the nonnull markers are assumed to be sparse. We calibrate ϵ_m and μ_m as follows. Define

$$\epsilon_m = m^{-\beta}, \quad \beta \in (1/2, 1), \tag{10}$$

which measures the sparsity of the disease-associated markers, and

$$\mu_m = (2\tau \log m)^{1/2}, \tag{11}$$

which measures the strength of the disease-associated markers. Here we assume that the disease markers are sparse. We first study the relationship between τ and β so that both false discoveries and false non-discoveries converge to zero as m goes to infinity.

THEOREM 2. *Consider* (4) *and reparameterize* ϵ_m *and* $\mu = \mu_m$ *in terms of* β *and* τ *as in* (10) *and* (11). *For any* $\gamma \geqslant 0$, *under the condition*

$$\tau \geqslant \{(1 + \gamma)^{1/2} + \sigma(1 + \gamma - \beta)^{1/2}\}^2, \tag{12}$$

the Bayes rule (9) *guarantees that both false discoveries and false non-discoveries converge to zero with convergence rate* $1/\{m^{\gamma}(\log m)^{1/2}\}$.

Theorem 2 shows the relationship between the strength of the signal and the convergence rate of the expected false discoveries and false non-discoveries. As expected, a stronger nonzero signal leads to faster convergence of the expected number of false identifications to zero and easier separation of the alternative from the null.

In (12), when $\gamma = 0$, the convergence rate of false discoveries and false non-discoveries is the slowest, $1/(\log m)^{1/2}$. Letting $\sigma \rightarrow 1$ yields the phase diagram in Fig. 2, which is an extension to the phase diagram in Donoho & Jin (2004). Three lines separate the $\tau - \beta$ plane into four

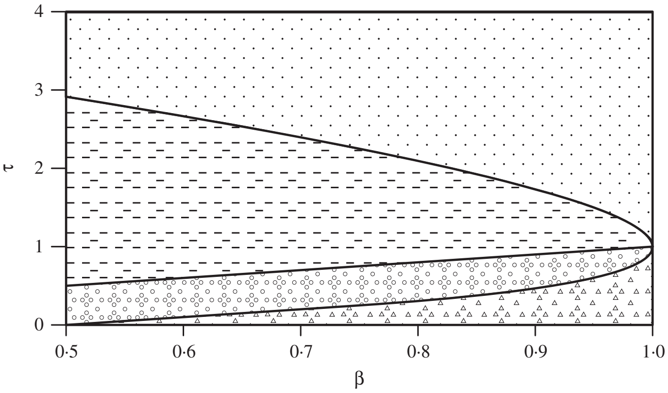

Fig. 2. Four regions of the $\tau - \beta$ plane, where $\epsilon_m = m^{-\beta}$ represents the percentage of the true signals and $\mu_m = (2\tau \log m)^{1/2}$ measures the strength of the signals: almost exact recovery (dots); estimable (dashes); detectable (circles); undetectable (triangles).

regions. The detection boundary

$$\tau = \begin{cases} \beta - 1/2, & (1/2 < \beta \leqslant 3/4), \\ \{1 - (1-\beta)^{1/2}\}^2, & (3/4 < \beta < 1), \end{cases}$$

separates the detectable region and the undetectable region, when μ exceeds the detection boundary, the null hypothesis $\mathrm{H}_0 : X_i \sim N(0, 1)$, $(i = 1, \ldots, m)$ and the alternative $\mathrm{H}_1^{(m)} : X_i \sim (1 - \epsilon_m) N(0, 1) + \epsilon_m N(\mu, 1)$, $(i = 1, \ldots, m)$ separate asymptotically. Above the estimation boundary $\tau = \beta$, it is possible not only to detect the presence of nonzero means, but also to estimate these means. These two regions were identified by Donoho & Jin (2004). Based on Theorem 2, we can obtain the almost exact recovery boundary

$$\tau = \{1 + (1-\beta)^{1/2}\}^2,$$

which provides the region that we can recover the whole sequence of θ with a very high probability converging to one when $m \to \infty$, since

$$\mathrm{pr}\left\{ \sum_i \mathrm{I}\,(\theta_i \neq \delta_i) \right\} \leqslant E\big(\|\theta - \delta\|_2^2\big) \asymp 1/(\log m)^{1/2},$$

where $\asymp$ represents asymptotic equivalence. In other words, in this region, we can almost fully classify the θ_i into the nulls and the non-nulls.

In large-scale genetic association studies, $\mu = \surd n \mu_1$. Based on Theorem 2, we have the following corollary.

COROLLARY 2. *Consider the same model as in Corollary* 1. *If the sample size*

$$n \geqslant \left[\frac{\{2(1+\gamma)\log m\}^{1/2} + \sigma\{2(1+\gamma)\log m + 2\log \epsilon_m\}^{1/2}}{\mu_1} \right]^2,$$

then both false discoveries and false non-discoveries of the oracle rule converge to zero at rate $1/\{m^\gamma (\log m)^{1/2}\}$.

4·2. *An adaptive sparse recovery procedure*

Theorem 2 shows that the convergence rate $1/\{m^{\gamma}(\log m)^{1/2}\}$ can be achieved using the oracle rule (9). However, this rule involves unknown parameters, ϵ_m, $f^1(x)$ and $f^0(x)$, which need to be estimated. Estimation errors can therefore affect the convergence rate. We instead propose an adaptive estimation procedure for model (4) that can achieve the same convergence rate for false discoveries and false non-discoveries under a slightly stronger condition than that in Theorem 2.

The Bayes oracle rule given in (9) can be rewritten as

$$\delta_{\mathrm{OR},i} = \mathrm{I}\{(1-\epsilon_m)f^0(X_i) < \epsilon_m f^1(X_i)\} = \mathrm{I}\left\{\frac{f(X_i)}{f^0(X_i)} > 2(1-\epsilon_m)\right\}, \qquad (13)$$

where $f(x) = (1-\epsilon_m)f^0(x) + \epsilon_m f^1(x)$ is the mixture density of the X_i. The density f is typically unknown, but is easily estimable. Let $\hat{f}(x)$ be a kernel density estimate based on the observations $X_1, X_2, \ldots, X_m$, defined by (5). For $i = 1, \ldots, m$, if $|X_i| > \{2(1+\gamma)\log m\}^{1/2}$, then reject the null $\mathrm{H}_i^0 : \theta_i = 0$; otherwise, calculate $\hat{S}(X_i) = \hat{f}(X_i)/\varphi(X_i)$ and reject the null if and only if $\hat{S}(X_i) > 2$. The threshold value 2 is based on (13) since the proportion ϵ_m is vanishingly small in our setting.

The next theorem shows that this adaptive procedure achieves the same convergence rate as the optimal Bayesian rule under a slightly stronger condition.

THEOREM 3. *Consider* (4) *and reparameterize* $\epsilon_m = n^{-\beta}$ *and* $\mu = \mu_m = (2\tau\log m)^{1/2}$ *as in* (10) *and* (11). *Suppose that the Gaussian kernel density estimate* $\hat{f}$, *defined in* (5), *with* $\rho = 1{\cdot}34 \times m^{-1/5}$ *is used in the adaptive procedure. For any* $\gamma \geqslant 0$, *if*

$$\tau > \{(1+\gamma)^{1/2} + \sigma(1+\gamma-\beta)^{1/2}\}^2,$$

then the adaptive procedure guarantees that both false discoveries and false non-discoveries converge to zero with convergence rate $1/\{m^{\gamma}(\log m)^{1/2}\}$.

4·3. *Illustration of marker recovery*

The sample sizes needed to obtain an almost exact recovery of the true disease-associated markers are presented in Fig. 3 for $\epsilon_m = 2 \times 10^{-4}$ and $\epsilon_m = 2 \times 10^{-5}$, which correspond to 100 and 10 disease-associated markers. We assume a disease prevalence of $D = 0{\cdot}10$. We observe that a larger γ implies a faster rate of convergence, and therefore also larger sample size required to the achieve the rate. Generally speaking, the sample sizes in Fig. 3 are much larger than those in Fig. 1, simply because almost exact recovery of the disease-associated markers is a much stronger goal than controlling the false discovery rate and false non-discovery rate at given levels. In addition, for markers with the same effect size, it is easier to achieve full recovery of the relevant markers for the sparser cases, see of Fig. 3(b).

We performed simulation studies to investigate how sample size affects the convergence rate of false discoveries and false non-discoveries. We considered the same setting as above, where the number of markers $m = 500\,000$, case proportion $r = 0{\cdot}40$, minor allele frequency $p = 0{\cdot}40$ for all the markers and the relevant markers proportion is $\epsilon_m = 2 \times 10^{-4}$. For a given odds ratio and a given γ parameter, we obtained the minimum sample size n for almost exact recovery and evaluated the convergence rate of the oracle and the adaptive procedures under different sample sizes. Table 2 shows the average number of FD+FN over 100 simulations. We observe a clear decrease of FD+FN as we increase the sample sizes for both the oracle and the adaptive procedures. For the adaptive procedure, when the sample size is larger than the minimum sample

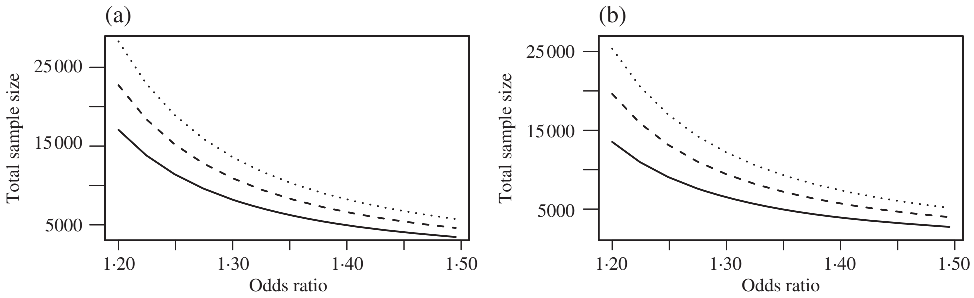

Fig. 3. Minimum sample size required for almost exact recovery of the disease-associated markers for a case-control genome-wide association study with the number of markers $m = 500\,000$, case proportion $r = 0{\cdot}40$, minor allele frequency $p = 0{\cdot}40$ for all the markers. (a) The relevant markers proportion is $\epsilon_m = 2 \times 10^{-4}$; (b) The relevant markers proportion is $\epsilon_m = 2 \times 10^{-5}$. For each plot, the solid, dash and dot lines correspond to γ values of 0·00, 0·20 and 0·40.

Table 2. *The performances of the oracle and adaptive false discovery and false non-discovery controlling procedures for varying sample sizes, where n is the theoretical sample size. For a given γ, the average number of* FD+FN *and its standard error over* 100 *simulations are shown for each sample size and odds ratio combination.*

			$\gamma = 0$			$\gamma = 0{\cdot}2$	
OR		$0{\cdot}8n$	n	$1{\cdot}2n$	$0{\cdot}8n$	n	$1{\cdot}2n$
	O	0·77 (0·89)	0·14 (0·35)	0·02 (0·14)	0·08 (0·27)	0·00 (0·00)	0·00 (0·00)
1·20	A	3·08 (3·11)	3·10 (3·23)	2·51 (3·04)	2·57 (3·02)	3·12 (3·24)	2·59 (3·12)
	O	0·57 (0·79)	0·13 (0·37)	0·01 (0·10)	0·03 (0·17)	0·00 (0·00)	0·00 (0·00)
1·30	A	3·93 (4·22)	2·61 (2·97)	2·92 (3·36)	3·41 (3·99)	2·66 (3·00)	2·92 (3·34)
	O	0·50 (0·72)	0·1 (0·30)	0·04 (0·20)	0·02 (0·14)	0·00 (0·00)	0·01 (0·10)
1·40	A	3·31 (3·30)	2·57 (3·11)	2·89 (3·45)	2·97 (3·39)	2·55 (3·11)	2·84 (3·42)

OR, odds ratio; O, oracle; A, adaptive.

size, the decrease in FD+FN is small since the minimum sample size has already achieved almost exact recovery.

5. GENOME-WIDE ASSOCIATION STUDY OF NEUROBLASTOMA

Neuroblastoma is a paediatric cancer of the developing sympathetic nervous system and is the most common form of solid tumour outside the central nervous system. It is a complex disease, with rare familial forms occurring due to mutations in paired-like homeobox 2b or anaplastic lymphoma kinase genes (Mosse et al., 2008), and several common variations being enriched in sporadic neuroblastoma cases (Maris et al., 2008). The latter genetic associations were discovered in a genome-wide association study of sporadic cases, compared with children without cancer, conducted at The Children's Hospital of Philadelphia. After initial quality controls on samples and marker genotypes, our discoveries dataset contained 1627 neuroblastoma case subjects of European ancestry, each of which contained 479 804 markers. To correct the potential effects of population structure, 2575 matching control subjects of European ancestry were selected based on their low identity-by-state estimates with case subjects. Analysis of this dataset has led to identification of several markers associated with neuroblastoma, including three markers on 6p22 containing the predicted genes FLJ22536 and FLJ44180 with allelic odds ratio of about 1·4 (Maris et al., 2008), and markers in BARD1 genes on chromosome 2 (Capasso et al., 2009) with odds ratio of about 1·68 in high-risk cases. The question that remains to be answered

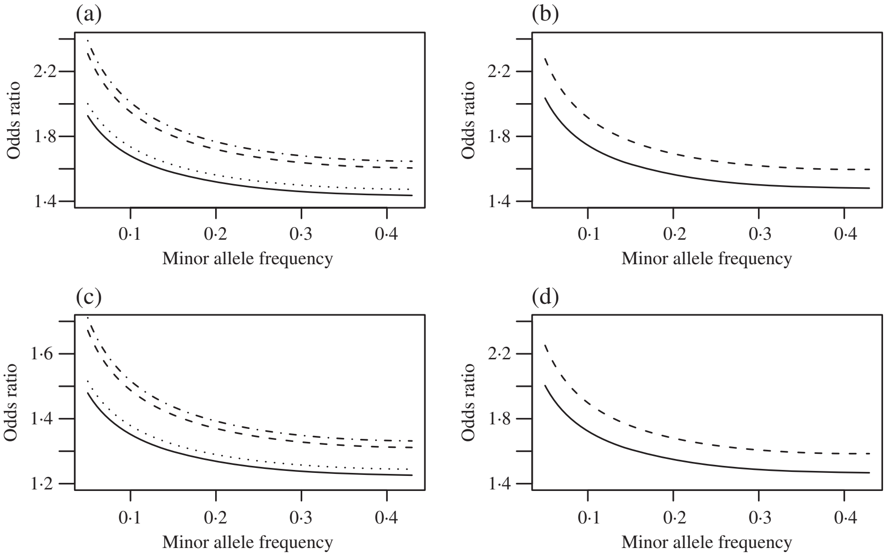

Fig. 4. Power analysis for neuroblastoma genome-wide association study. Plots (a) and (b) assume that $\epsilon = 2 \times 10^{-4}$, which corresponds to 98 relevant markers and plots (c) and (d) assume that $\epsilon = 10^{-4}$, which corresponds to 49 relevant markers. (a) and (c): detectable odds ratios for markers with different minor allele frequencies. For plot (a), the solid, dot, dash and dot-dash lines correspond to ($\text{FDR} = 5\%$, $\text{FNR} = 10^{-4}$), ($\text{FDR} = 1\%$, $\text{FNR} = 10^{-4}$), ($\text{FDR} = 5\%$, $\text{FNR} = 10^{-5}$), and ($\text{FDR} = 1\%$ and $\text{FNR} = 10^{-5}$), respectively. For plot (c), the solid, dot, dash and dot-dash lines correspond to ($\text{FDR} = 5\%$, $\text{FNR} = 5 \times 10^{-5}$), ($\text{FDR} = 1\%$, $\text{FNR} = 5 \times 10^{-5}$), ($\text{FDR} = 5\%$, $\text{FNR} = 5 \times 10^{-6}$), and ($\text{FDR} = 1\%$ and $\text{FNR} = 5 \times 10^{-6}$), respectively. (b) and (d): detectable odds ratios for markers with different minor allele frequencies for almost exact recovery with different convergence rates determined by $\gamma = 0{\cdot}10$ (solid) and $0{\cdot}40$ (dash).

is whether the current sample size is large enough to identify all the neuroblastoma-associated markers.

We estimated $\epsilon_m = 2 \times 10^{-4}$ using the procedure of Cai and Jin (2010) with tuning parameter $\eta = 10^{-4}$. A small value of η was chosen to reflect that we expect a very small number of disease-associated markers. With this estimate, we assume that there are 96 markers associated with neuroblastoma. For the given sample size and number of markers to be tested, Fig. 4(a) shows the minimum detectable odds ratios for markers with different minor allele frequencies for various false discovery rate and false non-discovery rate levels. The false non-discovery rates of 10^{-4}, 5×10^{-5} and 10^{-5} correspond to an expected number of nondiscovered markers of fewer than 48, 24 and 4·8, respectively. This plot indicates that with the current sample size, it is possible to recover about half of the disease-associated markers with an odds ratio around 1·4 and a minor allele frequency of around 20%.

Figure 4(b) shows the minimum detectable odds ratios for markers with different minor allele frequencies for almost exact recovery at different rates of convergence as specified by the parameter γ, indicating that the current sample size is not large enough to obtain almost exact recovery of all the disease-associated markers with odds ratio around 1·4.

Figure 4(c) and (d) shows similar plots if we assume that $\epsilon_m = 10^{-4}$, which implies that there are 48 associated markers. This plot indicates that with the current sample size, it is possible to recover most of the disease-associated markers with odds ratio around 1·4 and minor allele frequency around 20% for false discovery rates of 1 or 5%. However, the sample size is still not large enough to obtain almost exact recovery of all disease-associated markers with odds ratio around 1·4.

6. Discussion

Our work complements and differs from that of Zaykin & Zhivotovsky (2005) and Gail et al. (2008), who based their analyses on the individual marker P-values, aiming to evaluate how likely the markers selected based on the P-value ranking are disease-associated. Although both papers considered the issues of false positives and false negatives, neither links these to false discovery and false non-discovery rates in a formal way. Our analysis is mainly based on the distribution of the score statistics derived from the logistic regressions for case-control data and is set in the framework of multiple testing and sparse signal recovery. In the setting of thousands of hypotheses, it is natural to consider the power and sample size in terms of false discovery and false non-discovery rates, which correspond to Type I and Type II errors in single hypothesis testing. In addition, we have derived an explicit sample size formula under the assumption of fixed allele frequency and fixed common marker effect.

In presenting the results for genome-wide association studies, we made several simplifications. First, we assumed that the score statistics were independent across all the m markers. This obviously does not hold due to linkage disequilibrium among the markers. However, we expect such dependency to be short-range and our recent unpublished results show that the optimal false discovery rate controlling procedure of Sun & Cai (2007) is still valid and remains optimal. The results and the sample size formulae remain valid for such short-range dependent score statistics. Zaykin & Zhivotovsky (2005) and Gail et al. (2008) showed that correlations of P-values within linkage disequilibrium blocks of markers or among such blocks have little effect on selection or detection probabilities because such correlations do not extend beyond a small portion of the genome. It is likely, therefore, that our results were also little affected by such correlations. This was further verified in our simulation studies shown in Table 1. Second, to simplify the presentation and the derivation and to obtain a clean sample size formula as we presented in Corollaries 1 and 2, we assumed that all the markers had the same minor allele frequency and the effect sizes of the relevant markers were the same. One can only employ simulation for power analysis if the minor allele frequencies and the marker effects are all different. In fact, the analytical results that were used to validate their simulations in Gail et al. (2008) were also derived under the assumption of fixed allele frequency and fixed common effect. As in most of the power calculations, in practice, one should consider different scenarios in terms of minor allele frequencies and odds ratios.

Acknowledgement

This research was supported by grants from the National Institutes of Health and the National Science Foundation. We would like to thank the reviewers and associate editor for helpful comments.

Appendix

We present the proofs of Theorem 1, Theorem 2 and Theorem 3. We use the symbol $\asymp$ to represent asymptotic equivalence. If $a \asymp b$, then $a = O(b)$ and $b = O(a)$.

Proof of Theorem 1. When $\delta_i = 1$, we have

$$\frac{(1-\epsilon_m)\,\varphi(X_i; 0, 1)}{\epsilon_m \varphi(X_i; \mu, \sigma^2)} < \frac{1}{\lambda},$$

which is equivalent to $(X_i - \mu)/\sigma \in (-\infty, -c - d\sigma) \cup (c - d\sigma, +\infty)$, where

$$c = \left[\frac{2}{\sigma^2 - 1}\log\left\{\frac{\lambda\sigma(1-\epsilon_m)}{\epsilon_m}\right\} + \frac{\mu^2}{(\sigma^2-1)^2}\right]^{1/2}, \quad d = \frac{\mu}{\sigma^2 - 1}. \tag{A1}$$

Since

$$m\text{FDR} = \frac{\sum_{i=1}^m \text{pr}(\delta_i = 1 \mid \theta_i = 0)(1-\epsilon_m)}{\sum_{i=1}^m \{\text{pr}(\delta_i = 1 \mid \theta_i = 0)(1-\epsilon_m) + \text{pr}(\delta_i = 1 \mid \theta_i = 1)\epsilon_m\}} \leqslant \alpha_1,$$

we have

$$\Phi(-c\sigma - d) + \Phi(-c\sigma + d) \leqslant c_1\{\Phi(-c - d\sigma) + \Phi(-c + d\sigma)\}, \tag{A2}$$

where $c_1 = \alpha_1\epsilon_m/\{(1-\alpha_1)(1-\epsilon_m)\}$. Similarly, since

$$m\text{FNR} = \frac{\sum_{i=1}^m \text{pr}(\delta_i = 0 \mid \theta_i = 1)\epsilon_m}{\sum_{i=1}^m \{\text{pr}(\delta_i = 0 \mid \theta_i = 0)(1-\epsilon_m) + \text{pr}(\delta_i = 0 \mid \theta_i = 1)\epsilon_m\}} \leqslant \alpha_2,$$

we have

$$\Phi(c\sigma - d) - \Phi(-c\sigma - d) \geqslant c_2\{\Phi(c - d\sigma) - \Phi(-c - d\sigma)\}, \tag{A3}$$

where $c_2 = (1-\alpha_2)\epsilon_m/\{\alpha_2(1-\epsilon_m)\}$. Setting (A2) and (A3) as equalities and simplifying, we obtain (8). Let $(\hat{c}, \hat{d})$ be the solution of this equation, then (A1) leads to $\hat{\mu} = \hat{d}(\sigma^2 - 1)$. □

Proof of Theorem 2. Define c and d as in (A1). Since

$$\text{FD} = m\text{pr}(\delta_i = 1 \mid \theta_i = 0)(1-\epsilon_m) = m(1-\epsilon_m)\{\Phi(-c\sigma - d) + \Phi(-c\sigma + d)\},$$

in order for false discoveries to converge to zero, $-c\sigma + d$ should be negative and small enough. In addition, since

$$\begin{aligned} -c\sigma + d &= -\sigma\left[\frac{2}{\sigma^2-1}\log\left\{\sigma(1-m^{-\beta})\right\} + \frac{2\{\beta(\sigma^2-1)+\tau\}}{(\sigma^2-1)^2}\log m\right]^{1/2} + \frac{(2\tau\log)^{1/2}}{\sigma^2-1} \\ &\asymp \frac{\surd 2}{\sigma^2-1}\left[-\sigma\left\{\beta(\sigma^2-1)+\tau\right\}^{1/2} + \tau^{1/2}\right](\log m)^{1/2}, \end{aligned}$$

as $m \to \infty$, $-c\sigma + d < 0$. Because $\Phi(-c\sigma - d) \leqslant \Phi(-c\sigma + d)$, we have

$$\text{FD} \asymp \frac{m}{c\sigma - d}\varphi(c\sigma - d) \asymp \frac{1}{(\log m)^{1/2}} m^{1-C_1},$$

where

$$C_1 = \frac{[\sigma\{\beta(\sigma^2-1)+\tau\}^{1/2} - \tau^{1/2}]^2}{(\sigma^2-1)^2}.$$

In order for false discoveries to converge to zero with the rate $1/\{m^\gamma(\log m)^{1/2}\}$, we require $1 - C_1 \leqslant -\gamma$, which leads to $\tau \geqslant \{(1+\gamma)^{1/2} + \sigma(1+\gamma-\beta)^{1/2}\}^2$. Similarly,

$$\text{FN} = m\text{pr}(\delta_i = 0 \mid \theta_i = 1)\epsilon_m = m\epsilon_m\{\Phi(c - d\sigma) + \Phi(-c - d\sigma)\},$$

and

$$\begin{aligned} c - d\sigma &= \left[\frac{2}{\sigma^2-1}\log\left\{\sigma(1-m^{-\beta})\right\} + \frac{2\{\beta(\sigma^2-1)+\tau\}}{(\sigma^2-1)^2}\log m\right]^{1/2} - \frac{\sigma(2\tau\log m)^{1/2}}{\sigma^2-1} \\ &\asymp \frac{\surd 2}{\sigma^2-1}\left[\{\beta\left(\sigma^2-1\right)+\tau\}^{1/2} - \sigma\tau^{1/2}\right](\log m)^{1/2}. \end{aligned}$$

As $m \to \infty$, in order for $c - d\sigma < 0$, we require $\{\beta(\sigma^2 - 1) + \tau\}^{1/2} < \sigma/\tau^{1/2}$, which implies $\tau > \beta$. Then,

$$\text{FN} \asymp \frac{m}{d\sigma - c}\varphi(d\sigma - c) \asymp \frac{1}{(\log m)^{1/2}} m^{1-\beta-C_2},$$

where

$$C_2 = \frac{[\sigma(\tau)^{1/2} - \{\beta(\sigma^2 - 1) + \tau\}^{1/2}]^2}{(\sigma^2 - 1)^2}.$$

It is easy to show that

$$1 - \beta - C_2 \leqslant -\gamma, \quad \tau > \beta,$$

is equivalent to $\tau \geqslant \{(1+\gamma)^{1/2} + \sigma(1 + \gamma - \beta)^{1/2}\}^2$. □

Proof of Theorem 3. Define $S(x) = f(x)/\varphi(X_i)$, $\tilde{S}(x) = S(x)/(1 - \epsilon_m)$. We have

$$\begin{aligned}\text{pr}\{\hat{S}(X_i) > 2 \mid \theta_i = 0\} \leqslant\; & \text{pr}[|X_i| > \{2(1+\gamma)\log m\}^{1/2} \mid \theta_i = 0] \\ & + \text{pr}[|X_i| \leqslant \{2(1+\gamma)\log m\}^{1/2}, \tilde{S}(x) > 3/2 \mid \theta_i = 0] \\ & + \text{pr}[|X_i| \leqslant \{2(1+\gamma)\log m\}^{1/2}, |\hat{S}(X_i) - \tilde{S}(X_i)| > 1/2 \mid \theta_i = 0] \\ =\; & (A) + (B) + (C),\end{aligned}$$

say. We control (A), (B) and (C), respectively. First, we have

$$(A) = 2\Phi[-\{2(1+\gamma)\log m\}^{1/2}] \leqslant Cm^{-1-\gamma}/(\log m)^{1/2}.$$

Using the same technique as the proof of Theorem 2, if $\tau \geqslant \{(1+\gamma)^{1/2} + \sigma(1+\gamma-\beta)^{1/2}\}^2$, then

$$(B) \leqslant \text{pr}\{\tilde{S}(X_i) > 3/2 \mid \theta_i = 0\} = \text{pr}\{\Lambda_{OR}(X_i) < 2 \mid \theta_i = 0\} \leqslant Cm^{-1-\gamma}/(\log m)^{1/2}.$$

Consider the set $\mathcal{X}_i = \{X_i : |X_i| \leqslant \{2(1+\gamma)\log m\}^{1/2}\}$, for all $X_i \in \mathcal{X}_i$,

$$\text{pr}\{|\hat{S}(X_i) - \tilde{S}(X_i)| > 1/2 \mid X_i\} \leqslant \text{pr}\{|m\hat{S}(X_i) - mS(X_i)| + |mS(X_i) - m\tilde{S}(X_i)| > m/2 \mid X_i\},$$

and

$$|mS(X_i) - m\tilde{S}(X_i)| = \frac{m^{1-\beta} f(X_i)}{(1-\epsilon_m)\varphi(X_i)}.$$

Since the alternative density can be modelled as a Gaussian location-scale mixture, we have

$$f(X_i) = (1 - \epsilon_m)\varphi(X_i) + \sum_{l=1}^{L} \frac{1}{\sigma_l}\varphi\left(\frac{X_i - \mu_l}{\sigma_l}\right)\epsilon_{m,l},$$

where $\epsilon_{m,l}$ is the mixture proportion, representing the probability that X_j is from $N(\mu_l, \sigma_l)$). Under the condition

$$\mu_l \geqslant \tau\{(1+\gamma)^{1/2} + \sigma_l(1+\gamma-\beta')^{1/2}\}^2$$

and $\sigma_l > 1$ for all $l = 1, \ldots, L$, we have $f(X_i)/\varphi(X_i) < m^{\beta'}$, for all X_i. Then,

$$|mS(X_i) - m\tilde{S}(X_i)| \leqslant 2m^{1-(\beta-\beta')}, \beta > \beta'.$$

Therefore, for m sufficiently large,

$$\text{pr}\{|\hat{S}(X_i) - \tilde{S}(X_i)| > 1/2 \mid X_i \in \mathcal{X}_i\} \leqslant \text{pr}\{|m\hat{S}(X_i) - mS(X_i)| > m/3 \mid X_i\}.$$

Let $h = m^{-\rho}$, we have

$$E\{m\hat{S}(X_i) \mid X_i\} - mS(X_i) = m\int \frac{K(y)}{\varphi(X_i)} f(X_i - yh)\,\mathrm{d}y - m\frac{f(X_i)}{\varphi(X_i)}.$$

Further, we take $K(y) = \varphi(y)$ or any other symmetric kernel with a smaller tail than φ when $|y| > \{2(1+\gamma)\log m\}^{1/2}$. Using the fact that $\int K(y)\,\mathrm{d}y = 1$ and $\int yK(y)\,\mathrm{d}y = 0$,

$$E\{m\hat{S}(X_i)\mid X_i\} - mS(X_i) = m\sum_{l=2}(-1)^l h^l\left\{\int y^l K(y)\,\mathrm{d}y\right\}\frac{f^{(l)}(X_i)}{\varphi(X_i)}.$$

Let $H^l(x)$ be the lth Hermite polynomial. For $X_i \in \mathcal{X}_i$,

$$|(-1)^l f^{(l)}(X_i)/\varphi(X_i)| \leqslant (1-\epsilon_m)|H^l(X_i)| + \epsilon_m\sum_{k=1}^{L}\frac{1}{\sigma_k^{l+1}}\left|H^l\left(\frac{X_i-\mu_k}{\sigma_k}\right)\right| \leqslant C(\log m)^{l/2}.$$

Therefore, for m sufficiently large,

$$\begin{aligned}\text{pr}\{|\hat{S}(X_i) - \tilde{S}(X_i)| > 1/2 \mid X_i\} &\leqslant \text{pr}[|m\hat{S}(X_i) - mE\{S(X_i)\mid X_i\}| \\ &\quad + |E\{m\hat{S}(X_i)\mid X_i\} - mS(X_i)| > m/3 \mid X_i] \\ &\leqslant \text{pr}[|m\hat{S}(X_i) - mE\{S(X_i)\mid X_i\}| + Cm^{1-2\rho}\log m > m/3 \mid X_i] \\ &\leqslant \text{pr}[|m\hat{S}(X_i) - mE\{S(X_i)\mid X_i\}| > m/4 \mid X_i].\end{aligned}$$

Let $X_k^l = (X_k, \ldots, X_l), l \geqslant k$. For example, $X_1^m = (X_1, \ldots, X_m) = X$. Given $X_i \in \mathcal{X}_i$, define

$$g(X) = \frac{1}{h}\sum_{j=1}^{m}\frac{K\left(\frac{X_i-X_j}{h}\right)}{\varphi(X_i)} = \hat{S}(X_i),$$

then

$$\left|\frac{K\left(\frac{X_i-X_k}{h}\right)}{h\varphi(X_i)}\right| \leqslant \frac{1}{h}\exp\left[-\frac{1-h^2}{2h^2}\left\{X_i - \frac{h}{1-h^2}X_k\right\}^2 - \frac{1-2h^2}{2h^2(1-h^2)}X_k^2\right] \leqslant \frac{1}{h}.$$

For $k \neq i$,

$$|g(X_1^k, X_k, X_{k+1}^m) - g(X_1^k, \tilde{x}_k, X_{k+1}^m)| = \left|\frac{K\left(\frac{X_i-X_k}{h}\right) - K\left(\frac{X_i-\tilde{x}_k}{h}\right)}{h\varphi(X_i)}\right| \leqslant \frac{2}{h}.$$

By McDiarmid's inequality (McDiarmid, 1989),

$$\begin{aligned}\text{pr}\{|\hat{S}(X_i) - \tilde{S}(X_i)| > 1/2 \mid X_i \text{ with } X_i \in \mathcal{X}_i\} &\leqslant \text{pr}[|m\hat{S}(X_i) - mE\{S(X_i)\mid X_i\}| > m/4 \mid X_i \text{ with } X_i \in \mathcal{X}_i] \\ &\leqslant 2\exp(-m^{1-2\rho}/8).\end{aligned}$$

Therefore,

$$(C) \leqslant \int_{X_i\in\mathcal{X}_i} 2\exp(-m^{1-2\rho}/8)\varphi(X_i)\,\mathrm{d}X_i \leqslant 2\exp(-m^{1-2\rho}/8).$$

Thus,

$$\text{FD} = m(1-\epsilon_m)\text{pr}\{\hat{S}(X_i) > 2 \mid \theta_i = 0\} \leqslant Cm^{-\gamma}/(\log m)^{1/2}.$$

Similarly,

$$\begin{aligned}\text{FN} &= m\epsilon_m\text{pr}\{\hat{S}(X_i) \leqslant 2 \mid \theta_i = 1\} \\ &= m^{1-\beta}\text{pr}\{\hat{S}(X_i) \leqslant 2, |X_i| \leqslant \{2(1+\gamma)\log m\}^{1/2} \mid \theta_i = 1\} \\ &\leqslant m^{1-\beta}[\text{pr}\{S(X_i) \leqslant 5/2 \mid \theta_i = 1\} \\ &\quad + \text{pr}\{|\hat{S}(X_i) - \tilde{S}(X_i)| > 1/2, |X_i| \leqslant [2(1+\gamma)\log m]^{1/2} \mid \theta_i = 1\}] \\ &= m^{1-\beta}\{(D) + (E)\}.\end{aligned}$$

Using the same technique as the proof of Theorem 2, if

$$\tau \geqslant \{(1+\gamma)^{1/2} + \sigma(1+\gamma-\beta)^{1/2}\}^2,$$

then $(D) \leqslant Cm^{-1+\beta-\gamma}/(\log m)^{1/2}$. Finally,

$$(E) \leqslant \int_{X_i \in \mathcal{X}_i} \text{pr}\{|\hat{S}(X_i) - \tilde{S}(X_i)| > 1/2 \mid X_i\} f^1(X_i)\,\mathrm{d}X_i \leqslant 2\exp(-m^{1-2\rho}/8).$$

Thus,

$$\text{FD} = m(1-\epsilon_m) = m\{(D)+(E)\} \leqslant Cm^{-\gamma}/(\log m)^{1/2}. \qquad \square$$

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.

CAI, T. & JIN, J. (2010). Optimal rates of convergence for estimating the null and proportion of non-null effects in large-scale multiple testing. *The Ann. Statist.* **38**, 100–45.

CAPASSO, M., HOU, C., ASGHARZADEH, S., et al. (2009). Common variations in the bard1 tumor suppressor gene influence susceptibility to high-risk neuroblastoma. *Nature Genet.* **41**, 718–23.

DONOHO, D. & JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32**, 962–94.

EFRON, B. (2007). Size, power and false discovery rates. *Ann. Statist.* **35**, 1351–77.

GAIL, M., PFEIFFER, R., WHEELER, W. & PEE, D. (2008). Probability of detecting disease-associated single nucleotide polymorphisms in case-control genome-wide association studies. *Biostatistics* **9**, 201–15.

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

HUNTER, D., KRAFT, P., JACOBS, K., et al. (2007). A genome-wide association study identifies alleles in fgfr2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genet.* **39**, 870–4.

KLEIN, R., ZEIS, C., CHEW, E., TSAI, J., SACKLER, R., HAYNES, C., HENNING, A., SANGIOVANNI, J., MANE, S., MAYNE, S., BRACKEN, M., FERRIS, F., OTT, J., BARNSTABLE, C. & HOH, J. (2005). Complement factor h polymorphism in age-related macular degeneration. *Science* **3085**, 385–9.

MARIS, J., YAEL, P., BRADFIELD J., HOU, C., MONNI, S., SCOTT, R., ASGHARZADEH, S., ATTIVEH, E., DISKIN, S., LAUDENSLAGER, M., et al. (2008). A genome-wide association study identifies a susceptibility locus to clinically aggressive neuroblastoma at 6p22. *New Engl. J. Med.* **358**, 2585–93.

MCCARTHY, M., ABECASIS, G., CARDON, L., GOLDSTEIN, D., LITTLE, J., IOANNIDIS, J. & HIRSCHHORN, J. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9**, 356–69.

MCDIARMID, C. (1989). On the method of bounded differences. *Surveys in Combinatorics* **141**, 148–88.

MOSSE, Y., LAUDENSLAGER, M., LONGO, L., et al. (2008). Identification of alk as a major familial neuroblastoma predisposition gene. *Nature* **455**, 930–35.

PRENTICE, R. & PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–11.

SABATTI, C., SERVICE, S. & FREIMER, N. (2003). False discovery rates in linkage and association linkage genome screens for complex disorders. *Genetics* **164**, 829–33.

SILVERMAN, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

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

WELCOME TRUST CASE-CONTROL CONSORTIUM (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–78.

ZAYKIN, D. & ZHIVOTOVSKY, L. (2005). Ranks of genuine associations in whole-genome scans. *Genetics* **171**, 813–23.

[*Received October* 2010. *Revised January* 2011]