Biometrika (2021), **00**, 0, *pp.* 1–17
Printed in Great Britain
doi: 10.1093/biomet/asaa098

Covariate adaptive familywise error rate control for genome-wide association studies

BY HUIJUAN ZHOU

Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China
huijuan@stat.tamu.edu

XIANYANG ZHANG

Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.
zhangxiany@stat.tamu.edu

AND JUN CHEN

Division of Biomedical Statistics and Informatics, Mayo Clinic,
200 First St. SW, Rochester, Minnesota 55905, U.S.A.
Chen.Jun2@mayo.edu

SUMMARY

The familywise error rate has been widely used in genome-wide association studies. With the increasing availability of functional genomics data, it is possible to increase detection power by leveraging these genomic functional annotations. Previous efforts to accommodate covariates in multiple testing focused on false discovery rate control, while covariate-adaptive procedures controlling the familywise error rate remain underdeveloped. Here, we propose a novel covariate-adaptive procedure to control the familywise error rate that incorporates external covariates which are potentially informative of either the statistical power or the prior null probability. An efficient algorithm is developed to implement the proposed method. We prove its asymptotic validity and obtain the rate of convergence through a perturbation-type argument. Our numerical studies show that the new procedure is more powerful than competing methods and maintains robustness across different settings. We apply the proposed approach to the UK Biobank data and analyse 27 traits with 9 million single-nucleotide polymorphisms tested for associations. Seventy-five genomic annotations are used as covariates. Our approach detects more genome-wide significant loci than other methods in 21 out of the 27 traits.

Some key words: EM algorithm; External covariate; Familywise error rate; Multiple testing.

1. INTRODUCTION

Multiple testing arises when we face a large number of hypotheses and aim to discover signals while controlling specific error measures. Table 1 lists the possible outcomes when testing multiple hypotheses. The familywise error rate, FWER, and the false discovery rate, FDR, are two commonly used error measures employed in a wide range of scientific studies. The FWER is the probability of making one false discovery, while the FDR is the expected proportion of false positives. The FWER provides stringent control of Type I errors, and is preferable if the

Table 1. *Possible outcomes when testing multiple hypotheses*

	Not rejected	Rejected	Total
True nulls	U	V	m_0
True alternatives	T	S	m_1
Total	$m - R$	R	m

overall conclusion from various individual inferences is likely to be erroneous when at least one of them is, or the existence of a single false claim would cause significant loss. In contrast, the FDR control procedure provides less stringent control of Type I errors, and it generally delivers higher power at the cost of an increased number of Type I errors.

Consider the problem of simultaneously testing m hypotheses. We reject the hypotheses whose p-values are less than a cut-off t^*. For many FWER and FDR controlling procedures, the t^* that controls either one of them at level α is obtained by solving the constraint optimization problem

$$\text{maximize}_{t\in[0,1]} R(t), \qquad M(t) \leqslant \alpha, \tag{1}$$

where $R(t)$ denotes the total number of rejections given the threshold t, and $M(t)$ is a conservative estimate of the FWER or FDR. The most fundamental procedure for controlling the FWER is the Bonferroni method. It corresponds to the choice of $M(t) = mt$, which is the union bound on the FWER under the assumption that the null p-values are uniformly distributed, or superuniform, on $[0, 1]$. The classical Benjamini–Hochberg procedure for controlling the FDR can also be formulated using (1), with $M(t) = mt/R(t)$ being a conservative estimate of the FDR (Benjamini & Hochberg, 1995).

The formulation in (1) assumes that the hypotheses for different features are exchangeable. However, in many scientific applications there are informative covariates for each hypothesis that could reflect the group structure among the hypotheses or provide information on prior null probabilities. For example, in genome-wide association studies, single-nucleotide polymorphisms, SNPs, in active chromatin state are more likely to be significantly associated with the phenotype (GTEx Consortium., 2017). In a meta-analysis where samples are pooled across studies, the loci-specific sample sizes and population-level frequency can be informative for association analyses (Boca & Leek, 2018). For a fixed sample size, the power to detect significant associations is determined by the effect size, minor allele frequency, and levels of linkage disequilibrium at causal and noncausal variants (Kichaev et al., 2019). It is thus promising to incorporate these covariates to improve the detection power in genome-wide association studies.

Multiple testing procedures that leverage different types of covariates information have received considerable attention in the literature, especially for false discovery rate control. Genovese et al. (2006) pioneered multiple testing procedures with prior information using weighted p-values, and demonstrated that their weighted procedure controls the FWER and FDR while improving power. Roeder & Wasserman (2009) further explored this p-value weighting procedure by introducing an optimal weighting scheme for the FWER control. Inspired by the above works, Hu et al. (2010) developed a group Benjamini–Hochberg procedure by estimating the proportions of null hypotheses for each group separately. Bourgon et al. (2010) developed a particular weighting method called independent filtering, which first filters hypotheses by a criterion independent of the p-values and only tests hypotheses passing the filter. Ignatiadis et al. (2016) proposed independent hypothesis weighting for multiple testing with covariate information. The idea is to bin the covariate into several groups and then apply the weighted Benjamini–Hochberg procedure with piecewise constant weights. A similar idea has been used in the structure-adaptive

Benjamini–Hochberg algorithm introduced in Li & Barber (2019), where the weight assigned for each p-value is the reciprocal of the estimated null probability of the corresponding hypothesis. The null probabilities were estimated by utilizing censored p-values and structural information believed to be present among the hypotheses. Boca & Leek (2018) employed a similar approach by using the censored p-values and a regression approach to estimate null probabilities based on informative covariates. The above procedures can all be viewed to some extent as different variants of the weighted Benjamini–Hochberg or Bonferroni procedure. On the other hand, there are FDR-controlling procedures designed to find an optimal decision threshold by taking into account the p-value distribution under the alternatives, mostly based on the local FDR framework. For example, Sun et al. (2015) developed a local-FDR-based procedure to incorporate spatial information. Scott et al. (2015) and Tansey et al. (2018) proposed expectation maximization type algorithms to estimate the local FDR by taking into account covariate and spatial information, respectively. Lei & Fithian (2018) proposed the AdaPT procedure, which iteratively estimates the p-value thresholds based on a two-group mixture model using the partially masked p-values together with the covariates. Zhang & Chen (2021) proposed a more computationally efficient procedure to assign each p-value a covariate-adaptive threshold. Another related method, AdaFDR, used a mixture of the generalized linear model and Gaussian mixture for a threshold function to capture the covariate information and reflect the bump and slope structures (Zhang et al., 2019). Other relevant works include Ferkingstad et al. (2008), Zablocki et al. (2014), Dobriban et al. (2015), Wen (2016), Lei et al. (2021), Li & Barber (2017), Stephens (2017), and Xiao et al. (2017).

Recent developments on covariate-adaptive multiple testing focus on FDR control, while methods for FWER control lag behind. Existing FWER-controlling methods can all be thought to be variants of the weighted Bonferroni method, with the weights reflecting only the prior null probabilities. It has been demonstrated clearly in the FDR literature that incorporating the distribution of p-values under the alternative leads to the optimal rejection region in theory and more power in practice; see, e.g., Efron (2010). Given the popularity of FWER control in genome-wide association studies, we introduce a new covariate-adaptive FWER-controlling procedure, which takes into account the prior null probabilities as well as the distribution of p-values under the alternative, making it distinct from the existing FWER-controlling procedures. To illustrate the idea, suppose we are given a set of p-values p_i together with the external covariates x_i. Our method is motivated by the two-group mixture model

$$p_i \mid x_i \sim \pi(x_i) f_0(\cdot) + \{1 - \pi(x_i)\} f_1(\cdot)$$

with $\pi(x_i)$ and $f_1(\cdot)$ reflecting the heterogeneity of the probabilities of being null and the distributional characteristics of signals. We construct an objective function to control a conservative estimate of FWER while maximizing the expected number of true rejections. Specifically, we formulate the following constrained optimization problem:

$$\max_{t_i} \sum_{i=1}^{m} \{1 - \pi(x_i)\} F_1(t_i), \qquad \sum_{i=1}^{m} \pi(x_i) F_0(t_i) \leqslant \alpha,$$

where F_0 and F_1 are the cumulative distribution functions of f_0 and f_1, respectively. To establish the asymptotic FWER control, and the rate of convergence, new theoretical developments are needed. Existing theoretical analysis techniques developed for FDR-controlling procedures are not applicable to the FWER-controlling procedure, since we aim to control a sum instead of a proportion. The arguments based on the Rademacher complexity in Li & Barber (2019) do not provide a meaningful bound on the FWER. Employing a perturbation-type argument, we develop

a more delicate analysis for each of the summands, which leads to a useful bound on the sum and thus the FWER.

The main contributions of the paper are twofold. First, we propose a powerful covariate-adaptive FWER-controlling procedure that can incorporate multi-dimensional covariates, and exploit the information from both the null probability and the alternative distribution. We prove asymptotic FWER control of the proposed procedure when the pairs of covariates and p-values across different hypotheses are independent, and derive the exact rate of convergence based on a novel perturbation technique. We emphasize that our proofs do not rely on the correct specification of the two-group mixture model. Second, we develop an efficient algorithm to implement the proposed method and demonstrate its usefulness in handling big datasets arising from genome-wide association studies. In the application to the genome-wide association study of about 9 million SNPs and 75 covariates, we could complete the analysis in hours.

Numerical studies show that our procedure controls the FWER in the strong sense and is more powerful than the competing methods. It maintains robustness across different settings, including scenarios of model misspecification and correlated hypotheses. Even when the covariates are not informative, our procedure is as powerful as the traditional methods.

2. METHODOLOGY

2.1. *Set-up*

Denote by $\|v\|$ the Euclidean norm of a vector v. With some abuse of notation, let $\|A\|$ be the spectral norm of a matrix A. For two symmetric matrices A and B, $A \preceq B$ means that $B - A$ is positive semidefinite. For $a, b \in \mathbb{R}$, write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. Throughout the paper, we use c to denote a positive constant which can be different from line to line.

We consider the problem of covariate-adaptive multiple testing to control the FWER. Suppose we are given m hypotheses, among which m_0 are true nulls. For each hypothesis, we observe a p-value p_i as well as a covariate x_i lying in some space $\mathcal{X} \subseteq \mathbb{R}^d$, which encodes potentially useful external information concerning the presence of a signal. Let $H_i = 0$ if the ith null hypothesis is true and $H_i = 1$ otherwise. Denote by $\mathcal{M}_0$ the set of all true null hypotheses. We transform the ith p-value based on a map $T_i : [0, 1] \to \mathbb{R}_+$ that will be estimated from the covariates and p-values. The larger $T_i(p_i)$ is, the more likely the ith hypothesis is over the alternative. The motivation for such a transformation will be discussed in the next section. In a nutshell, the optimal T_i is the likelihood ratio between the ith p-value distributions under the alternative and the null.

2.2. *Optimal rejection rule*

Let $f_0(\cdot)$ be the null p-value distribution and $f_1(\cdot)$ denote the alternative p-value distribution. Denote by $F_0(\cdot)$ and $F_1(\cdot)$ the corresponding cumulative distribution functions. Suppose we reject the ith hypothesis if $p_i \leqslant t_i$ for some cut-off t_i. Before presenting the procedure that inspires the choice of T_i, it is worth clarifying the definition of the FWER from both the frequentist and Bayesian perspectives. The key difference between these two viewpoints lies in whether we treat the indicators $\{H_i\}$ as fixed or random quantities. From the frequentist perspective, the indicators $\{H_i\}$ are deterministic and we have, by the union bound,

$$\mathrm{FWER}_{\mathrm{Freq}} = \mathbb{P}(p_i \leqslant t_i \text{ for some } i \in \mathcal{M}_0) \leqslant \sum_{i=1}^{m} \mathbb{I}(H_i = 0) F_0(t_i).$$

From a Bayesian's point of view, it is natural to posit the two-group mixture model

$$p_i \mid x_i \sim \pi(x_i) f_0(\cdot) + \{1 - \pi(x_i)\} f_1(\cdot).$$

In this case, conditional on x_i, H_i is assumed to be a Bernoulli random variable with success probability $1 - \pi(x_i)$. The Bayesian FWER can be bounded as

$$\mathrm{FWER}_{\mathrm{Bay}} = \mathbb{P}(p_i \leqslant t_i \text{ for some } i \in \mathcal{M}_0) \leqslant \sum_{i=1}^{m} \mathbb{P}(p_i \leqslant t_i, H_i = 0) = \sum_{i=1}^{m} \mathbb{E}\{\pi(x_i)\} F_0(t_i). \tag{2}$$

To motivate our procedure, it is more convenient to adopt the Bayesian viewpoint. But we emphasize that the proposed procedure indeed provides asymptotic FWER control in the usual frequentist sense, as shown in § 3.

We aim to find $\{t_i\}$ to maximize the expected number of true rejections given by

$$\mathbb{E}\left\{\sum_{i=1}^{m} \mathbb{I}(H_i = 1, p_i \leqslant t_i)\right\} = \sum_{i=1}^{m} \mathbb{E}[\{1 - \pi(x_i)\}] F_1(t_i)$$

while controlling the FWER at a desired level α. To achieve both goals, we formulate the following constraint optimization problem:

$$\max_{t_i} \sum_{i=1}^{m} \{1 - \pi(x_i)\} F_1(t_i) \;\text{ such that }\; \sum_{i=1}^{m} \pi(x_i) F_0(t_i) \leqslant \alpha, \tag{3}$$

where $\sum_{i=1}^{m} \pi(x_i) F_0(t_i)$ serves as a conservative estimate of the Bayesian FWER based on the derivations in (2). The Lagrangian for problem (3) is

$$L(t_1, \ldots, t_m; \lambda) = \sum_{i=1}^{m} \{1 - \pi(x_i)\} F_1(t_i) - \lambda \left\{\sum_{i=1}^{m} \pi(x_i) F_0(t_i) - \alpha\right\},$$

with $\lambda > 0$. Differentiating the Lagrangian with respect to t_i and setting the derivative to be zero, at the optimal value t_i^*, we obtain

$$\frac{\{1 - \pi(x_i)\} f_1(t_i^*)}{\pi(x_i) f_0(t_i^*)} = \lambda.$$

Motivated by the above observation, we set

$$T_i(p) = \frac{\{1 - \pi(x_i)\} f_1(p)}{\pi(x_i) f_0(p)}.$$

We note that $T_i(p)$ is related to the local FDR as follows:

$$\frac{1}{T_i(p) + 1} = \frac{\pi(x_i) f_0(p)}{\pi(x_i) f_0(p) + \{1 - \pi(x_i)\} f_1(p)} = \mathbb{P}(H_i = 0 \mid p, x_i).$$

In the following discussions, we suppose f_0 is the uniform distribution on $[0,1]$ and f_1 is strictly decreasing, which is a common assumption in the literature, e.g., Sun & Cai (2007) and Cao et al. (2013). As T_i is strictly decreasing in this case, we may reduce our attention to the rejection rule $p_i \leqslant t_i^*$ as $T_i(p_i) \geqslant T_i(t_i^*) := \tau^*$. The cut-off can then be expressed as

$$t_i^* = f_1^{-1}\left\{\frac{\pi(x_i)\tau^*}{1-\pi(x_i)}\right\},$$

where f_1^{-1} denotes the inversion of f_1. The expected number of true rejections and the conservative estimate of the Bayesian FWER in (2) are both monotonically decreasing in τ. Therefore, the solution to (3) satisfies

$$\tau^* = \min\left[\tau > 0 : \sum_{i=1}^{m} \pi(x_i) f_1^{-1}\left\{\frac{\pi(x_i)\tau}{1-\pi(x_i)}\right\} \leqslant \alpha\right]. \tag{4}$$

In practice, both π and f_1 are unknown and need to be replaced by estimates from the data. We provide detailed discussions about estimating the unknowns in the next subsection.

2.3. *A feasible procedure*

We describe a feasible procedure based on suitable estimates of π and f_1. To avoid overfitting and facilitate the theoretical analysis, we adopt the idea of censoring p-values as in Storey (2002), Boca & Leek (2018) and Li & Barber (2019). Under the two-group mixture model, for a prespecified $0 < \gamma < 1$, we have

$$\mathbb{I}(p_i > \gamma) \mid x_i \sim \pi(x_i)\mathrm{Ber}(1-\gamma) + \{1-\pi(x_i)\}\mathrm{Ber}\{1-F_1(\gamma)\},$$

where $\mathrm{Ber}(1-\gamma)$ denotes the Bernoulli distribution with success probability $1-\gamma$. We model f_1 using the beta distribution $f_1(p) = kp^{k-1}$ for $0 < k < 1$, as it provides reasonably good approximation to a wide range of alternative distributions, as demonstrated in Zhang & Chen (2021). Here we treat k as fixed and will discuss the choice of data-driven k in § 2.4.

Before presenting our method, it is worth clarifying the rationale behind our procedure. Notice that $\pi(x_i)$ appears both inside and outside the function f_1^{-1} in (4). To achieve asymptotic FWER control, we need a conservative estimate for $\pi(x_i)$ outside the function f_1^{-1}, while we require the one inside f_1^{-1} to depend on the covariates to reflect the heterogeneity among signals while retaining a certain form of stability; for more details see § 3. The reason will become clear by inspecting the proof of Proposition 1. We first observe that

$$\mathbb{E}\left\{\frac{\mathbb{I}(p_i > \gamma)}{1-\gamma}\middle| x_i\right\} = \pi(x_i) + \{1-\pi(x_i)\}\frac{1-F_1(\gamma)}{1-\gamma} \geqslant \pi(x_i).$$

Therefore, we suggest replacing the $\pi(x_i)$ outside f_1^{-1} with $\mathbb{I}(p_i > \gamma)/(1-\gamma)$. To estimate $\pi(x_i)$ inside f_1^{-1}, we consider the logistic model

$$\log\left\{\frac{\pi(x_i)}{1-\pi(x_i)}\right\} = x_i^{\mathrm{T}}\beta.$$

The quasi-loglikelihood function is then

$$L_m(\beta) = \sum_{i=1}^{m} \log\left[\pi(x_i)(1-\gamma)^{y_i}\gamma^{1-y_i} + \{1-\pi(x_i)\}(1-\gamma^k)^{y_i}\gamma^{k(1-y_i)}\right],$$

where $\pi(x_i) = (1+\mathrm{e}^{-x_i^{\mathrm{T}}\beta})^{-1}$ and $y_i = \mathbb{I}(p_i > \gamma)$. Define the corresponding quasi maximum likelihood estimator as

$$\hat{\beta} = \arg\max_{\beta\in\mathcal{B}} L_m(\beta), \tag{5}$$

where $\mathcal{B}$ is some compact subset of $\mathbb{R}^d$. Let $\hat{\pi}(x_i) = \{\tilde{\pi}(x_i) \vee \varepsilon_1\} \wedge \varepsilon_2$, where $\tilde{\pi}(x_i) = (1+\mathrm{e}^{-x_i^{\mathrm{T}}\hat{\beta}})^{-1}$ and $0 < \varepsilon_1 < \varepsilon_2 < 1$. We have used winsorization to prevent $\hat{\pi}(x_i)$ being too close to zero and one. Further denote

$$\hat{\tau} = \min\left[\tau \geqslant \varepsilon : \sum_{i=1}^{m} \frac{\mathbb{I}(p_i > \gamma)}{1-\gamma} f_1^{-1}\left\{\frac{\hat{\pi}(x_i)\tau}{1-\hat{\pi}(x_i)}\right\} \leqslant \alpha\right]$$

for some $\varepsilon > 0$. It is straightforward to show that $\hat{\tau} = \tilde{\tau} \vee \varepsilon$ with

$$\tilde{\tau} = k\left[\sum_{i=1}^{m} \frac{\mathbb{I}(p_i > \gamma)}{\alpha(1-\gamma)}\left\{\frac{1-\hat{\pi}(x_i)}{\hat{\pi}(x_i)}\right\}^{1/(1-k)}\right]^{1-k}.$$

Finally, we set

$$\hat{t}_i = \left[\frac{\{1-\hat{\pi}(x_i)\}k}{\hat{\pi}(x_i)\hat{\tau}}\right]^{1/(1-k)},$$

and reject the ith hypothesis if $p_i \leqslant \hat{t}_i \wedge \gamma$.

Remark 1 (Connection to the weighted Bonferroni procedure). Suppose $\varepsilon_1 = \varepsilon = 0$ and $\varepsilon_2 = 1$. Then we have

$$\hat{t}_i = \left[\frac{\{1-\tilde{\pi}(x_i)\}k}{\tilde{\pi}(x_i)\tilde{\tau}}\right]^{1/(1-k)} = \alpha w_i,$$

where

$$w_i = \exp\left\{-\frac{x_i^{\mathrm{T}}\hat{\beta}}{1-k}\right\}\left[\sum_{i=1}^{m} \frac{\mathbb{I}(p_i > \gamma)}{1-\gamma}\exp\left\{-\frac{x_i^{\mathrm{T}}\hat{\beta}}{1-k}\right\}\right]^{-1}.$$

We reject the ith hypothesis if

$$p_i \leqslant \alpha w_i \wedge \gamma.$$

In this sense, our procedure can be viewed as a particular type of weighted Bonferroni procedure. However, different from existing methods, our weight incorporates the information regarding the alternative p-value distribution, which often leads to more rejections and thus higher power, as observed in our numerical studies.

2.4. *EM algorithm*

Algorithm 1 provides the details of our iterative algorithm to solve problem (5).

Algorithm 1. EM algorithm for problem (5).

Input: $\{x_i, y_i\}_{i=1}^m, \gamma, k$; initializer: $\beta^{(0)}$
Output: $\hat{\beta}$
Notation: $b_{0i} = (1-\gamma)^{y_i}\gamma^{1-y_i}$; $b_{1i} = (1-\gamma^k)^{y_i}\gamma^{k(1-y_i)}$; tol: tolerance level
Iteration:
E step:
$Q_i^{(t)} = \mathbb{E}\{\mathbb{I}(H_i = 0) \mid y_i, x_i, \beta^{(t)}\} = \pi_i^{(t)} b_{0i} / \{\pi_i^{(t)} b_{0i} + (1-\pi_i^{(t)}) b_{1i}\}$,
where $\pi_i^{(t)} = (1 + \mathrm{e}^{-x_i^{\mathrm{T}}\beta^{(t)}})^{-1}$
M step:
$\beta^{(t+1)} = \arg\max_{\beta\in\mathcal{B}} \sum_{i=1}^m \{Q_i^{(t)} \log(\pi_i) + (1-Q_i^{(t)}) \log(1-\pi_i)\}$,
where $\pi_i = (1 + \mathrm{e}^{-x_i^{\mathrm{T}}\beta})^{-1}$
Until: $|L_m(\beta^{(t+1)}) - L_m(\beta^{(t)})| / |L_m(\beta^{(t)})| < \text{tol}$
Return: $\beta^{(t+1)}$ after a sufficient number of iterations

The theory in § 3 shows that our procedure controls the FWER asymptotically for any fixed k. However, a suitable choice of k that produces a beta-distribution closer to the true f_1, especially on the small-p-value region, will improve the statistical power. In practice, an EM algorithm can be used to estimate the k and β jointly. To be precise, we define the quasi-loglikelihood function

$$L_m(\beta, k) = \sum_{i=1}^m \log\left[\pi(x_i)(1-\gamma)^{y_i}\gamma^{1-y_i} + \{1-\pi(x_i)\}(1-\gamma^k)^{y_i}\gamma^{k(1-y_i)}\right].$$

Then we estimate (β, k) jointly by the quasi maximum likelihood estimator defined as

$$(\hat{\beta}, \hat{k}) = \arg\max_{\beta\in\mathcal{B}, k\in(0,1)} L_m(\beta, k). \tag{6}$$

We summarize the algorithm for solving problem (6) in Algorithm 2.

Algorithm 2. EM algorithm for problem (6).

Input: $\{x_i, y_i\}_{i=1}^m, \gamma$; initializer: $\beta^{(0)}, k^{(0)}$
Output: $\hat{\beta}, \hat{k}$
Notation: $b_{0i} = (1-\gamma)^{y_i}\gamma^{1-y_i}$; tol: tolerance level
Iteration:
E step:
$Q_i^{(t)} = \mathbb{E}\{\mathbb{I}(H_i = 0) \mid y_i, x_i, \beta^{(t)}, k^{(t)}\} = \pi_i^{(t)} b_{0i} / \{\pi_i^{(t)} b_{0i} + (1-\pi_i^{(t)}) b_{1i}^{(t)}\}$,
where $\pi_i^{(t)} = (1 + \mathrm{e}^{-x_i^{\mathrm{T}}\beta^{(t)}})^{-1}$, $b_{1i}^{(t)} = (1-\gamma^{k^{(t)}})^{y_i}\gamma^{k^{(t)}(1-y_i)}$
M step:
$\beta^{(t+1)} = \arg\max_{\beta\in\mathcal{B}} \sum_{i=1}^m \{Q_i^{(t)} \log(\pi_i) + (1-Q_i^{(t)}) \log(1-\pi_i)\}$,
where $\pi_i = (1 + \mathrm{e}^{-x_i^{\mathrm{T}}\beta})^{-1}$
$k^{(t+1)} = \arg\max_{k\in(0,1)} \sum_{i=1}^m (1-Q_i^{(t)}) \{y_i \log(1-\gamma^k) + k(1-y_i)\log(\gamma)\}$

Until: $|L_m(\beta^{(t+1)}, k^{(t+1)}) - L_m(\beta^{(t)}, k^{(t)})|/|L_m(\beta^{(t)}, k^{(t)})| < \text{tol}$
Return: $\beta^{(t+1)}, k^{(t+1)}$ after a sufficient number of iterations

3. Asymptotic familywise error rate control

In this section we prove asymptotic FWER control for the procedure proposed in § 2.3. Throughout this section we shall adopt the frequentist viewpoint, i.e., we view the indicators $\{H_i\}$ as a deterministic sequence.

Let $p_{j\to a} = (p_1, \ldots, p_{j-1}, a, p_{j+1}, \ldots, p_m)^{\mathrm{T}} \in \mathbb{R}^m$ for $a = 0, 1$. We define $\hat{\beta}(p_{j\to a})$ and $\hat{t}_i(p_{j\to a})$ by setting the jth p-value to be equal to a when estimating the corresponding quantities. We make the following assumption.

Assumption 1. Denote by F_{0i} the cumulative distribution function for p_i with $H_i = 0$. Suppose that $\{p_i\}_{i\in\mathcal{M}_0}$ are superuniform, i.e., $F_{0i}(t) \leqslant t$ for all $t \in [0, 1]$ and $i \in \mathcal{M}_0$.

Assumption 1 is standard in the literature, see, e.g., Benjamini & Yekutieli (2001).

Proposition 1. *If $\{p_i\} \in \mathcal{M}_0$ are mutually independent and are independent with the nonnull p-values, then under Assumption 1 we have*

$$\mathrm{FWER} \leqslant J_m + \alpha \leqslant c(J_{m,1} + J_{m,2}) + \alpha,$$

where

$$J_m = \sum_{j=1}^{m} \mathbb{E}\left\{\left|\hat{t}_j(p_{j\to 0}) - \hat{t}_j(p_{j\to 1})\right|\right\},$$

$$J_{m,1} = \sum_{j=1}^{m} \mathbb{E}\left[\frac{|x_j^{\mathrm{T}}\{\hat{\beta}(p_{j\to 0}) - \hat{\beta}(p_{j\to 1})\}|}{\left\{c\alpha^{-1}\sum_{i\neq j}\mathbb{I}(p_i > \gamma)\right\} \vee \varepsilon^{1/(1-k)}}\right],$$

$$J_{m,2} = \sum_{j=1}^{m} \mathbb{E}\left(\frac{\alpha^{-1}\sum_{i\neq j}\mathbb{I}(p_i > \gamma)|x_i^{\mathrm{T}}\{\hat{\beta}(p_{j\to 0}) - \hat{\beta}(p_{j\to 1})\}| + \alpha^{-1}}{\left[\left\{c\alpha^{-1}\sum_{i\neq j}\mathbb{I}(p_i > \gamma)\right\} \vee \varepsilon^{1/(1-k)}\right]^2}\right),$$

and ε is defined as in § 2.3.

The above proposition shows that the validity of the asymptotic FWER control relies on the stability of $\hat{t}_j$, i.e., the smallness of $|\hat{t}_j(p_{j\to 0}) - \hat{t}_j(p_{j\to 1})|$, which in turn depends on $\|\hat{\beta}(p_{j\to 0}) - \hat{\beta}(p_{j\to 1})\|$. Set $z_i = (x_i, y_i)$, where $y_i = \mathbb{I}\{p_i > \gamma\}$. Define

$$l(\beta; z_i) = \log\left\{\frac{1}{1 + \mathrm{e}^{-x_i^{\mathrm{T}}\beta}}(1-\gamma)^{y_i}\gamma^{1-y_i} + \frac{\mathrm{e}^{-x_i^{\mathrm{T}}\beta}}{1 + \mathrm{e}^{-x_i^{\mathrm{T}}\beta}}(1-\gamma^k)^{y_i}\gamma^{k(1-y_i)}\right\}$$

and $\mathbb{P}_m l(\beta) = m^{-1}\sum_{i=1}^{m} l(\beta; z_i)$. To ensure $\|\hat{\beta}(p_{j\to 0}) - \hat{\beta}(p_{j\to 1})\|$ is small, we impose the following assumptions.

Assumption 2. Suppose $z_i \in \mathbb{R}^{d+1}$ are independent and possibly nonidentically distributed.

Assumption 2 is not uncommon in the multiple testing literature, see, e.g., Ignatiadis et al. (2016). We suspect that the results still hold when z_i is a sequence of weakly dependent variables, although a rigorous proof is left for future investigation.

Assumption 3. There exists a continuous function of β, denoted by $\mathcal{L}(\beta)$, such that

$$\lim_{m\to+\infty} \sup_{\beta\in\mathcal{B}} |\mathbb{E}\{\mathbb{P}_m l(\beta)\} - \mathcal{L}(\beta)| = 0.$$

Assumption 4. Suppose $\mathcal{L}(\beta)$ has a unique global maximizer β^* over the compact space $\mathcal{B}$.

Assumption 4 is needed in our perturbation argument. If the maximizer is not unique, there seems to be no guarantee that the difference between $\hat{\beta}(p_{j\to 0})$ and $\hat{\beta}(p_{j\to 1})$ will be small.

PROPOSITION 2. *Suppose Assumptions 2–4 are satisfied, and further assume that* $\sup_{1\leqslant i\leqslant m} \mathbb{E}\left(\|x_i\|^8\right) < \infty$. *Then we have*

$$\hat{\beta}(p_{j\to 0}) - \hat{\beta}(p_{j\to 1}) = (S_j^* + \Delta_j)^{-1}(U_j^* + \Pi_j),$$

where S_j^* *and* U_j^* *are the leading terms such that* $S_j^* = -\sum_{i\neq j} \nabla^2 l(\beta^*; z_i)$ *and* $\sup_{1\leqslant j\leqslant m} \|U_j^*\| = O_{\mathbb{P}}(1)$, *and* Δ_j *and* Π_j *are the remainder terms satisfying*

$$\sup_{1\leqslant j\leqslant m} \|\Delta_j\| = o_{\mathbb{P}}(m) \text{ and } \sup_{1\leqslant j\leqslant m} \|\Pi_j\| = o_{\mathbb{P}}(1).$$

Given Propositions 1 and 2, we have the following theorem of asymptotic FWER control.

THEOREM 1. *Suppose the following conditions are satisfied:*

(i) *Assumptions 1–4 hold;*
(ii) *for some* $q \geqslant 2$ *and* $\epsilon > 0$, *we have* $\sup_{1\leqslant i\leqslant m} \mathbb{E}\left(\|x_i\|^{4q+\epsilon}\right) < \infty$;
(iii) *we have* $\sup_{\beta\in\mathcal{B}} |\mathbb{E}\{\mathbb{P}_m l(\beta)\} - \mathcal{L}(\beta)| = O(m^{-1/2})$;
(iv) *the function* $\mathcal{L}(\beta)$ *is twice continuously differentiable;*
(v) *the global maximizer* β^* *is not on the boundary of* $\mathcal{B}$;
(vi) *for some* $c > 0$, *we have* $\nabla^2\mathcal{L}(\beta^*) \preceq -cI$, *where* I *denotes the identity matrix;*
(vii) *for large enough* m *and some* $c > 0$, *we have* $\mathbb{E}\left\{\nabla^2\mathbb{P}_m l(\beta^*)\right\} \preceq -cI$; *and*
(viii) *the number of true null hypotheses* m_0 *satisfies* $\liminf m_0/m > 0$.

Then

$$\text{FWER} \leqslant J_m + \alpha = \begin{cases} o(\alpha m^{\frac{1-q}{4}}) + \alpha, & \text{if } 2 \leqslant q \leqslant 2+\sqrt{5}, \\ O(\alpha m^{\frac{-q}{1+q}}) + \alpha, & \text{if } q > 2+\sqrt{5}. \end{cases}$$

Theorem 1 derives the bound and its exact order on the FWER. Interestingly, the order of the bound depends crucially on the tail behaviour of the covariates, and it shows an interesting phase transition depending on the value of q. We briefly explain this result: from Proposition 1, we can see that the FWER is upper bounded by an expression of the form $\alpha + \sum_{i=1}^m r_i$ with $r_i \geqslant 0$. Our argument optimizes the summation $\sum_{i=1}^m r_i$ in the upper bound. Depending on the

value of q, the dominant term in this summation will change, which eventually leads to different convergence rates. When the covariates have exponential tails, the rate of convergence can be as close to m^{-1} as possible. The details of the proof are provided in the Supplementary Material. As we discussed earlier, Assumptions 1–4 enable us to show that the upper bound on the FWER relies on the smallness of $\|\hat{\beta}(p_{j\to 0}) - \hat{\beta}(p_{j\to 1})\|$ and to get the expression for $\hat{\beta}(p_{j\to 0}) - \hat{\beta}(p_{j\to 1})$. As our goal is to quantify the exact rate of convergence of the FWER upper bound to the nominal level α, we further need to quantify the exact difference between $\hat{\beta}(p_{j\to 0})$ and $\hat{\beta}(p_{j\to 1})$. Through conditions (iii)–(v) and the strong-concavity condition (vi), we obtain the concentration inequality for $\|\hat{\beta}(p_{j\to a}) - \beta^*\|$. Conditions (ii) and (vii) are used for controlling the inverse $(S_j^* + \Delta_j)^{-1}$ in the expression of $\hat{\beta}(p_{j\to 0}) - \hat{\beta}(p_{j\to 1})$. Condition (viii) requires the number of true null hypotheses to be at least some positive proportion of all hypotheses, which is fairly mild. We give one toy example where all the conditions are satisfied.

Example 1. Suppose all hypotheses are true nulls and (x_i, p_i) are independent and identically distributed with x_i being one-dimensional, $x_i \perp\!\!\!\perp p_i$ and $p_i \sim \mathrm{Un}([0,1])$, i.e., the uniform distribution on $[0,1]$. Then $\mathcal{L}(\beta) = \mathbb{E}\{\mathbb{P}_m l(\beta)\} = \mathbb{E}\{l(\beta; z_1)\}$. If x_i follows a distribution symmetric about zero and $\mathbb{P}(x_i \neq 0) > 0$, it can be shown that $\mathcal{L}'(\beta) = 0$ if $\beta = 0$, $\mathcal{L}'(\beta) > 0$ if $\beta < 0$, $\mathcal{L}'(\beta) < 0$ if $\beta > 0$ and $\mathcal{L}'(\beta) = -\mathcal{L}'(-\beta)$. Thus, $\beta^* = 0$ is the unique maximizer. We can further prove that $\mathcal{L}''(0) \leqslant -c$ as long as $\mathbb{E}(x_i^2) > c'$ for some $c' > 0$. Other conditions are naturally satisfied. When x_i follows a nonsymmetric distribution, we also illustrate its obedience to these conditions. One mandatory requirement for the distribution of x_i is that $\mathbb{P}(x_i > 0) > 0$ and $\mathbb{P}(x_i < 0) > 0$. In practice, we could always achieve this by shifting the covariate via subtracting the median, or by standardizing the covariate. For more details see the Supplementary Material.

4. Numerical studies

4.1. *Simulation set-ups*

We conduct comprehensive simulations to evaluate the finite-sample performance of the proposed method and compare it to competing methods. For genome-scale multiple testing, the numbers of hypotheses could range from thousands to millions. For demonstration purposes, we start with $m = 10\,000$ hypotheses. To study the impact of signal density and strength, we simulate three levels of signal density, sparse, medium and dense signals, and six levels of signal strength, from very weak to very strong. To demonstrate the power improvement by using external covariates, we simulate covariates of varying informativeness, noninformative, moderately informative and strongly informative. For simplicity, we simulate one covariate $x_i \sim N(0,1)$ for $i = 1, \ldots, m$. Given x_i, we denote $\pi(x_i)$ by π_i and let

$$\pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}, \qquad \eta_i = \eta_0 + k_d x_i,$$

where η_0 and k_d determine the baseline signal density and the informativeness of the covariate, respectively. We set $\eta_0 = 3.5$, 2.5 and 1.5, which achieves a signal density of around 3%, 8% and 18%, respectively, at the baseline, i.e., no covariate effect, representing sparse, medium and dense signals. Here, k_d is set to be 0, 1 and 1.5, representing a noninformative, moderately informative and strongly informative covariate. Based on π_i, the underlying truth H_i is simulated from $H_i \sim \mathrm{Ber}(1 - \pi_i)$. Finally, we simulate independent z-scores using $z_i \sim N(k_s H_i, 1)$, where k_s controls the signal strength, the effect size, and we use six values equally spaced on [2, 2.8]

and label them $\{1, 2, \ldots, 6\}$. The one-sided formula $1 - \Phi(z_i)$ converts z-scores into p-values. The p-values together with x_i are used as the input for the proposed method.

In addition to the basic setting, denoted S0, we investigate other settings to study the robustness of the proposed method.

Set-up S1 (Additional f_1 distribution). Instead of simulating normal z-scores under f_1, we simulate z-scores from a noncentral gamma distribution with the shape parameter 2. The scale/noncentrality parameters of the noncentral gamma distribution are chosen to match the variance and mean of the normal distribution under S0.

Set-up S2 (Correlated hypotheses). We further investigate the effect of dependency among hypotheses by simulating correlated multivariate normal z-scores. Four correlation structures, including two block correlation structures and two AR(1) correlation structures, are investigated. For the block correlation structure we divide the 10 000 hypotheses into 500 equal-sized blocks. Within each block, we simulate equal positive correlations ($\rho = 0.5$) (S2.1). On top of S2.1, we divide the block into 2×2 sub-blocks, and simulate negative correlations, $\rho = -0.5$, between the two sub-blocks (S2.2). For the AR(1) structure, we investigate both $\rho = 0.75^{|i-j|}$ (S2.3) and $\rho = (-0.75)^{|i-j|}$ (S2.4).

4.2. *Competing methods*

We compared the proposed covariate-adaptive FWER-controlling procedure, denoted by CAMT.fwer, to IHW-Bonferroni, weighted Bonferroni and Holm's step-down methods (Holm, 1979). The covariate-adaptive FWER-controlling procedure, implemented using the CAMT.fwer function in the `R` package `CAMT` (R Development Core Team, 2021), used the model $\log[\pi(x_i)/\{1 - \pi(x_i)\}] = x_i^{\mathrm{T}}\beta$, set $f_1(p) = kp^{k-1}$, and estimated β and k jointly using Algorithm 2. The weighted Bonferroni method rejected the ith hypothesis if $p_i < \alpha/(m\pi_i)$, where the π_i were estimated from CAMT.fwer. The IHW-Bonferroni method was implemented using the `R` package `IHW`, and Holm's step-down method using the `holm` function from the `R` package `mutoss`. We also implemented an oracle procedure based on the proposed optimal rejection rule, where the π_i and f_1 were the true null probabilities and alternative density that generated the data.

Storey et al. (2004) proposed the bootstrap method to estimate the overall null probability π, which is implemented in the `R` package `qvalue`. The method uses censored p-values $\mathbb{I}\{p_i > \lambda\}$ with $\lambda = 0.05, 0.1, \ldots, 0.95$ to obtain the corresponding estimates of the null probability, π_λ, and returns the best $\pi_{\hat{\lambda}}$. We set $\gamma = \hat{\lambda}$. We evaluated the performance based on the FWER control, probability of making at least one false positive, and power, true positive rate, with a target FWER level of 5%. Results were averaged over 1000 simulation runs. In addition, we investigated the FWER control across different target levels, $\alpha = 0.01, 0.05, 0.1, 0.15, 0.2$, for cases where there are no signals and under Set-up S0 with moderate signal density, $\eta_0 = 2.5$, signal strength, $k_s = 2.4$, and covariate informativeness, $k_d = 1$.

4.3. *Simulation results*

We present the simulation results of Set-up S0 in Fig. 1, and Set-ups S1 and S2 in the Supplementary Material along with the FWER control across different target levels. All methods control the FWER around the 5% target level, see Fig. 1(a). We additionally draw the 95% confidence intervals of the proposed method CAMT.fwer, and observe that almost all the intervals cover the 5% target level, see the dashed line in Fig. 1(a), which suggests adequate FWER control of CAMT.fwer under finite samples. In terms of power, Fig. 1(b) shows that generally the five competing methods from the best to the worst are oracle, CAMT.fwer, IHW-Bonferroni and weighted Bonferroni, the performance of these two methods depends on the cases, and Holm's

step-down methods. The oracle procedure represents the performance upper bound and dominates other methods.

We now study the impact of the external prior information, signal density and strength; see Fig. 1(b). First, the power increases with the signal strength for all methods as expected. Second, as the prior informativeness increases, the performance difference between methods widens. CAMT.fwer is close to the oracle procedure: it is as powerful as other methods when the prior is not informative, and is substantially more powerful when the prior is highly informative. Both the IHW-Bonferroni and weighted Bonferroni methods improve over Holm's step-down method when the prior is informative. Third, the proposed method maintains high power across different signal densities. In contrast, the IHW-Bonferroni method performs better than the weighted Bonferroni method when the signal is sparse, and performs worse when the signal is dense.

In the Supplementary Material figures show the weak and strong FWER control of the competing methods across different target levels. All the methods including CAMT.fwer control the FWER at the target level. We compare the power across different target levels at moderate signal density, signal strength and prior informativeness. CAMT.fwer remains more powerful than other methods. In fact, as the target level increases, the power difference becomes larger.

We next study the robustness of the proposed method under Set-ups S1 and S2. The general trend remains similar to Set-up S0, indicating that CAMT.fwer is robust to different f_1 distributions and various correlation structures. Interestingly, as we generate z-scores from noncentral gamma distributions for the alternative in Set-up S1, the power of CAMT.fwer is even closer to that of the oracle procedure, indicating that the beta distribution can model the alternative p-value distribution very accurately in this case.

5. Application to UK Biobank data

To demonstrate the use of the proposed procedure in real-world applications, we applied CAMT.fwer to UK Biobank data (Kichaev et al., 2019). We downloaded the data, which includes p-values and functional annotations, from `https://data.broadinstitute.org/alkesgroup/UKBB/` and `https://data.broadinstitute.org/alkesgroup/FINDOR/`. The genome-wide association p-values for 9 million SNPs and 27 traits were calculated using BOLT-LMM (Loh et al., 2018) based on 459K samples. The annotation data consists of 75 coding, conserved, regulatory and linkage-disequilibrium-related annotations that have previously been shown to be enriched for disease heritability (Kichaev et al., 2019). We compared our method with IHW-Bonferroni, weighted Bonferroni and Holm's step-down methods. For the IHW-Bonferroni method, as it can only deal with one-dimensional covariates, we chose the covariate that had the maximum Spearman correlation with the p-values out of the 75 covariates for the 27 traits separately. For the weighted Bonferroni method, we rejected the ith hypothesis if $p_i < \alpha/(m\pi_i)$, where the π_i were estimated from CAMT.fwer. The details of the use of CAMT.fwer are given below.

Appropriate initial values of (β, k) are important for the algorithm to reach convergence in fewer iterations and reduce the computation time significantly. To achieve this end, we estimate those initial values based on small p-values, so the initial beta distribution fits the small-p-value region more accurately. Let π^{s} be the estimate of the proportion of the true null hypotheses based on Storey's procedure. We define the small p-values as the first $m(1-\pi^{\mathrm{s}})$ smallest p-values and let u be the maximum value of those small p-values. We have that

$$f(p \mid p < u) = \frac{\pi + (1-\pi)kp^{k-1}}{\pi u + (1-\pi)u^k}$$

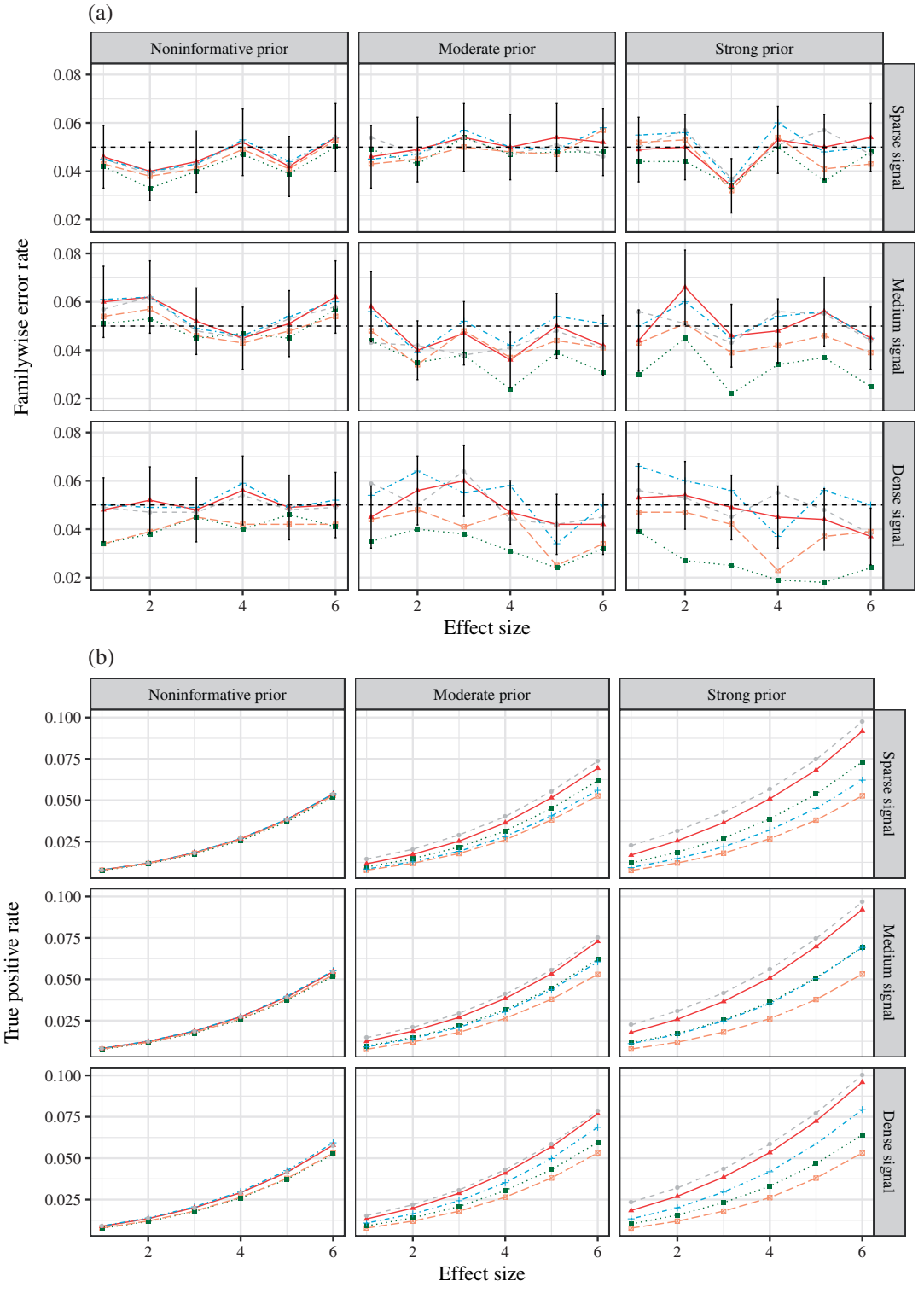

Fig. 1. Performance comparison under Set-up S0. (a) Familywise error rates; and (b) true positive rates were averaged over 1000 simulation runs. The dashed grey, solid red, dotted green, dot-dashed blue and long-dashed orange lines represent the oracle, CAMT.fwer, IHW-Bonferroni, weighted Bonferroni and Holm's step-down methods, respectively. The error bars in (a) represent the 95% confidence intervals of the method CAMT.fwer, and the dashed horizontal line indicates the target FWER level of 0.05.

Table 2. *Significant loci detected at the FWER level of* 0.05. Improve = (CAMT.fwer − Holm)/Holm × 100%. *The numbers with subscript* ∗ *are the maximum numbers of rejections among the four competing methods for the corresponding traits*

	Holm	IHW	Weighted Bonferroni	CAMT.fwer	Improve
Balding Type I	836_*	836_*	836_*	833	−0.4%
BMI	1287	1287	1347	1364_*	6.0%
Heel T Score	2104	2104	2144	2146_*	2.0%
Height	3463	3460	3555_*	3550	2.5%
Waist-hip Ratio	909	909	937	952_*	4.7%
Eosinophil Count	1750	1750	1817_*	1797	2.7%
Mean Corpular Hemoglobin	1913	1913	1953_*	1925	0.6%
Red Blood Cell Count	1570	1570	1609	1633_*	4.0%
Red Blood Cell Distribution Width	1470	1470	1493_*	1470	0.0%
White Blood Cell Count	1393	1393	1430	1462_*	5.0%
Auto Immune Traits	179	179	180_*	138	−22.9%
Cardiovascular Diseases	512	512	529	540_*	5.5%
Eczema	423	423	426	431_*	1.9%
Hypothyroidism	373	373	377	424_*	13.7%
Respiratory and Ear-nose-throat Diseases	228	228	231	236_*	3.5%
Type 2 Diabetes	156	156	158	160_*	2.6%
Age at Menarche	634	634	648	652_*	2.8%
Age at Menopause	200	200	201	203_*	1.5%
FEV1-FVC Ratio	1537	1537	1575	1599_*	4.0%
Forced Vital Capacity (FVC)	867	867	924	947_*	9.2%
Hair Color	1606	1606	1616	1629_*	1.4%
Morning Person	204	204	217	229_*	12.3%
Neuroticism	176	115	189	198_*	12.5%
Smoking Status	221	159	232	254_*	14.9%
Sunburn Occasion	232	232	232	237_*	2.2%
Systolic Blood Pressure	1108	1108	1148	1157_*	4.4%
Years of Education	383	383	416	447_*	16.7%

Holm, Holm's procedure; IHW, independent hypothesis weighting; CAMT.fwer, our proposed method.

is the conditional density of the mixture model $f(p) = \pi + (1-\pi)kp^{k-1}$ given that the value is less than u. We estimate π and k by maximizing the conditional loglikelihood function,

$$(\tilde{\pi}, \tilde{k}) = \arg \max_{\pi \in (0,1), k \in (0,1)} \sum_{i: p_i < u} \log \left\{ \pi + (1-\pi) k p_i^{k-1} \right\} - n \log\{\pi u + (1-\pi) u^k\},$$

where n is the number of p-values that are smaller than u. Let $\tilde{\beta} = (\log\{\tilde{\pi}/(1-\tilde{\pi})\}, 0)^{\mathrm{T}}$. Then we set $(\tilde{\beta}, \tilde{k})$ as the initializer in Algorithm 2.

Due to the linkage disequilibrium between SNPs, after getting the rejected SNPs we used PLINK's linkage-disequilibrium-based clumping algorithm with a 5 Mb window and an r^2 threshold of 0.01 to form clumps of SNPs. The British population in the 1000 genomes data (1000 Genomes Project Consortium, 2015) was used to calculate the linkage disequilibrium. The rejected SNPs belonging to the same clump count for only one significant locus. The numbers of significant loci at the 5% FWER level detected by the four competing methods are presented in Table 2. We present the numbers of rejections before clumping in the Supplementary Material.

Our proposed method CAMT.fwer detected more loci than other methods in 21 out of the 27 traits. Averaged across traits, our approach attained a 4.20% increase in significant loci detected compared with Holm's method.

6. DISCUSSION

To conclude, we point out a few future research directions. First, in the two-group mixture model, we assume that the success probabilities $\pi(x_i)$ vary with x_i while f_1 is independent of x_i. This assumption is reasonable in some applications, but it can be restrictive when the covariates also affect the effect sizes. It is thus of interest to develop a procedure by allowing f_1 to be dependent on x_i in such scenarios. Second, modelling f_1 and π using nonparametric procedures would give us the flexibility to capture more complicated signal patterns. Finally, extending the method to accommodate more general structural information, such as the phylogenetic tree structure (Xiao et al., 2017), is an interesting direction.

ACKNOWLEDGEMENT

Zhou and Zhang acknowledge support from the National Science Foundation. Zhou was also partially supported by the China Scholarship Council. Chen acknowledges support from the Mayo Clinic Center for Individualized Medicine. We thank Dr. Kejun He for help with the proof of Lemma S4, and the associate editor and two reviewers for their insightful comments, which substantially improved the paper. Zhou is also affiliated with Texas A & M University. Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health (R21HG011662). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

SUPPLEMENTARY MATERIAL

Supplementary Material available at *Biometrika* online includes further discussion on Example 1, the proofs for all theoretical results, additional simulation results and the numbers of rejections before clumping. The proposed method is implemented in the `R` package `CAMT` available at `https://github.com/jchen1981/CAMT`.

REFERENCES

1000 GENOMES PROJECT CONSORTIUM (2015). A global reference for human genetic variation. *Nature* **526**, 68–74.

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

BOCA, S. M. & LEEK, J. T. (2018). A direct approach to estimating false discovery rates conditional on covariates. *PeerJ* **6**, e6035.

BOURGON, R., GENTLEMAN, R. & HUBER, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proc. Nat. Acad. Sci.* **107**, 9546–51.

CAO, H, SUN, W. & KOSOROK, M. R. (2013). The optimal power puzzle: scrutiny of the monotone likelihood ratio assumption in multiple testing. *Biometrika* **100**, 495–502.

DOBRIBAN, E., FORTNEY, K., KIM, S. K. & OWEN, A. B. (2015). Optimal multiple testing under a Gaussian prior on the effect sizes. *Biometrika* **102**, 753–66.

EFRON, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge: Cambridge University Press.

FERKINGSTAD, E., FRIGESSI, A., RUE, H., THORLEIFSSON, G. & KONG, A. (2008). Unsupervised empirical Bayesian multiple testing with external covariates. *Ann. Appl. Statist.* **2**, 714–35.

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

GTEx Consortium., Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., Jo, B., Mohammadi, P., Park, Y.-S., Parsana, P. et al. (2017), Genetic effects on gene expression across human tissues. *Nature* **550**, 204–13.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65–70.

Hu, J. X., Zhao, H. & Zhou, H. H. (2010). False discovery rate control with groups. *J. Am. Statist. Assoc.* **105**, 1215–27.

Ignatiadis, N., Klaus, B., Zaugg, J. B. & Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Meth.* **13**, 577–80.

Kichaev, G., Bhatia, G., Loh, P., Gazal, S., Burch, K., Freund, M. K., Schoech, A., Pasaniuc, B. & Price, A. L. (2019). Leveraging polygenic functional enrichment to improve GWAS power. *Am. J. Hum. Genet.* **104**, 65–75.

Lei, L. & Fithian, W. (2018). AdaPT: an interactive procedure for multiple testing with side information. *J. R. Statist. Soc.* B **80**, 649–79.

Lei, L., Ramdas, A. & Fithian, W. (2021). A general interactive framework for FDR control under structural constraints. *Biometrika* **108**, 253–67.

Li, A. & Barber, R. F. (2017). Accumulation tests for FDR control in ordered hypothesis testing. *J. Am. Statist. Assoc.* **112**, 837–49.

Li, A. & Barber, R. F. (2019). Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm. *J. R. Statist. Soc.* B **81**, 45–74.

Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. (2018). Mixed-model association for Biobank-scale datasets. *Nature Genet.* **50**, 906–8.

R Development Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: 280 R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Roeder, K. & Wasserman, L. (2009). Genome-wide significance levels and weighted hypothesis testing. *Statist. Sci.* **24**, 398–413.

Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P. & Kass, R. E. (2015). False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *J. Am. Statist. Assoc.* **110**, 459–71.

Stephens, M. (2017). False discovery rates: a new deal. *Biostatistics*, **18**, 275–94.

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

Storey, J. D., Taylor, J. E. & Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc.* B **66**, 187–205.

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., Reich, B. J., Cai, T. T., Guindani, M. & Schwartzman, A. (2015). False discovery control in large-scale spatial multiple testing. *J. R. Statist. Soc.* B **77**, 59–83.

Tansey, W., Koyejo, O., Poldrack, R. A. & Scott, J. G. (2018). False discovery rate smoothing. *J. Am. Statist. Assoc.* **113**, 1156–71.

Wen, X. (2016). Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control. *Ann. Appl. Statist.* **10**, 1619–38.

Xiao, J., Cao, H. & Chen, J. (2017). False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing. *Bioinformatics* **33**, 2873–81.

Zablocki, R. W., Schork, A. J., Levine, R. A., Andreassen, O. A., Dale, A. M. & Thompson, W. K. (2014). Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics* **30**, 2098–104.

Zhang, M. J., Xia, F. & Zou, J. (2019). Fast and covariate-adaptive method amplifies detection power in large-scale multiple hypothesis testing. *Nature Commun.* **10**, 3433.

Zhang, X. & Chen, J. (2021). Covariate adaptive false discovery rate control with applications to omics-wide multiple testing. *J. Am. Statist. Assoc.*, to appear, DOI:10.1080/01621459.2020.1783273.

[*Received on* 17 *February* 2020. *Editorial decision on* 2 *November* 2020]