Speeding Up Monte Carlo Simulations for the Adaptive Sum of Powered Score Test with Importance SamplingYangqing Deng[1,3], Yinqiu He[2], Gongjun Xu[2], Wei Pan[1,4]

[1]*Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, USA*

[2]*Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA*

[3]*Department of Mathematics, University of North Texas, Denton, TX 76203, USA*

[4]*Corresponding author: panxx014@umn.edu*

SUMMARY: A central but challenging problem in genetic studies is to test for (usually weak) associations between a complex trait (e.g. a disease status) and sets of multiple genetic variants. Due to the lack of a uniformly most powerful test, data-adaptive tests, such as the adaptive sum of powered score (aSPU) test, are advantageous in maintaining high power against a wide range of alternatives. However, there is often no closed-form to accurately and analytically calculate the p-values of many adaptive tests like aSPU, thus Monte Carlo (MC) simulations are often used, which can be time-consuming to achieve a stringent significance level (e.g. 5e-8) used in GWAS. To estimate such a small p-value, we need a huge number of MC simulations (e.g. 1e+10). As an alternative, we propose using importance sampling to speed up such calculations. We develop some theory to motivate a

proposed algorithm for the aSPU test, and show that the proposed method is computationally more efficient than the standard MC simulations. Using both simulated and real data, we demonstrate the superior performance of the new method over the standard MC simulations.

Keywords: Adaptive test; aSPU; Genome-wide association studies; GWAS; SNPs.

1. Introduction

Genome-wide association studies (GWASs), such as one of the first and most influential, conducted by the Wellcome Trust Case Control Consortium (WTCCC) (2010), have successfully identified many genetic variants associated with common disease and complex traits by single SNP-based analysis, where SNP stands for single-nucleotide polymorphism. However, these results can only explain a small proportion of the heritability for most human traits. One possible reason is that many SNPs have too small effect sizes to be detected by single SNP-based analysis. Driven by the idea of aggregating small and possibly sparse signals of multiple SNPs to gain power, researchers have proposed multiple SNP-based analyses, in which the goal is to test association between a trait and multiple SNPs, which can be drawn from a gene or a pathway. Since there is no uniformly most powerful test, it is desirable to apply an adaptive test such that high power can be maintained against various alternatives (e.g. Chen et al. 2010; Lee et al. 2012; Zhang et al. 2014; Su et al. 2015; Huang et al. 2016; Su et al. 2017; Ma and Wei 2019; Yang et al. 2019), most of which require Monte Carlo methods to calculate their p-values. Pan et al. (2014) proposed such a test, called the adaptive sum of powered score (aSPU) test, based on combining a family of so-called sum of powered score (SPU) tests, which cover some existing tests, such as the Sum (or burden) test, the sum of squared score test (Pan 2009) and minP test as special

cases. In particular, compared to some existing tests, such as the popular Sequence Kernel Association Test (SKAT) (Wu et al. 2011) and the optimized SKAT (SKAT-O) (Lee et al. 2012), the aSPU test performs better in situations where a large number of non-associated variants exist (i.e. with sparse signals). The aSPU has also been extended to pathway-based analysis (Pan et al. 2015) and can simply work with GWAS summary statistics and a reference panel (Kwak and Pan 2015). An asymptotic theory for the aSPU test for "large n and large p" has been developed (Xu et al. 2016), where n and p refer to the sample size and the number of SNPs respectively. For "large n and small p", although the aSPU test statistic is a non-linear function of a multivariate normal score vector, it is unknown how to calculate its p-value analytically. Thus, Monte Carlo (MC) simulations (Lin 2005) or other resampling methods have been used so far. Given the stringent significance threshold for genome-wide testing (e.g. at p<5e-8), it requires at least 1e+9 to 1e+11 MC simulations (Yu et al. 2011) to estimate a small p-value around 5e-8, which will be time-consuming. Hence, we propose importance sampling to speed up p-value calculations for aSPU (and possibly other adaptive tests).

Importance sampling was originally proposed to reduce variations in MC simulations (Cochran 1977; Hesterberg 1995; Asmussen and Glynn, 2007). Its intuition is to sample from a proposed distribution that overweighs an important target region so that we can obtain samples more frequently from that region, while the standard Monte Carlo sampling will seldom get any from that region if the corresponding probability in that region is extremely small. The proposed distribution is often called a proposal distribution. This method has been frequently used to evaluate the extremes of Gaussian random fields and other rare-event probabilities, and its efficiency has been carefully studied (Adler et al. 2012; Liu and Xu 2014a, 2014b; Jiang, et al. 2017; He and Xu 2018; Li and Xu 2018). Since the p-value calculation involves sampling from a rejection region with possibly a small probability from a null distribution, we propose using importance sampling to speed up the SPU and aSPU tests. The

applications of importance sampling or other Markov chain Monte Carlo (MCMC) techniques for such a purpose are a common theme in the literature (e.g. Kimmel and Shamir 2006; Liang et al 2007; Shi et al. 2007; Yu et al. 2011). The recent work of Shi et al. (2018) is most closely related to ours: they also proposed an importance sampling approach called MCMC-CE (Markov Chain Monte Carlo-Cross Entropy) to more efficiently estimate small p-values. A key difference is that we consider a wider range of tests with different and possibly computationally more efficient proposal distributions. We derive some theoretical results to support the computational efficiency of our proposed method. We use simulations to demonstrate that our new method can yield good estimates of extreme p-values with much less iterations (and thus less time) than the standard MC. We apply the standard MC and the importance sampling approaches to the WTCCC data to confirm the effectiveness of the new method.

2. Methods

2.1 *Importance sampling for SPU*

The SPU tests (Pan et al. 2014) were originally proposed to improve power under each of multiple alternatives with varying association patterns between a set of variants and a trait of interest. They can be applied to either common variants or rare variants. For this paper, we focus on common variants, and will comment on rare variants in the Discussion Section. Suppose we have the (marginal) Z-statistics $\boldsymbol{Z} = (z^{(1)}, z^{(2)}, \ldots, z^{(p)})'$ for p SNPs (in a gene), and $\mathbf{R}$ is an estimate of their covariance matrix with diagonal elements equal to 1. $\mathbf{R}$ is often regarded as an estimate of the LD covariance matrix of the SNPs. Each Z-statistic is often based on the Wald or score test on each SNP from a marginal model of the trait on the SNP. For common variants, the Z-statistics are usually normally distributed (with a large sample size). Hence, without loss of generality, we assume throughout this paper that the null distribution of $\boldsymbol{Z}$ is MVN(0, $\mathbf{R}$). We also assume $\mathbf{R}$ is nonsingular (which is reasonable with large sample sizes and possible pruning of highly correlated SNPs).

To be powerful for an unknown alternative, multiple SPU tests are designed, each tailored for a type of alternative hypotheses. A power index $0 < \gamma < \infty$ is used to make a corresponding SPU test powerful for a specific alternative: *a* $\mathrm{SPU}(\gamma, \boldsymbol{Z})$ test with a smaller/larger γ is more powerful for a more sparse/dense alternative. We define

$$\mathrm{SPU}(\gamma, \boldsymbol{Z}) = T_\gamma = \begin{cases} \sum_{i=1}^{p} z^{(i)^\gamma} & (0 < \gamma < \infty) \\ \max_i |z^{(i)}| & (\gamma = \infty) \end{cases}$$

where γ is usually chosen from {1, 2, ..., 8, ∞}, which has been shown empirically to perform well in previous studies. The standard MC approach to calculating a p-value is to sample $\boldsymbol{Z}_b$ $(b = 1,2,\ldots,B)$ from its null distribution $\mathrm{MVN}(\mathbf{0}, \mathbf{R})$, then calculate

$$P_{\mathrm{SPU}(\gamma,\boldsymbol{Z})} = \frac{1}{B}\sum_{b=1}^{B} I(|\mathrm{SPU}(\gamma, \boldsymbol{Z}_b)| > |\mathrm{SPU}(\gamma, \boldsymbol{Z})|).$$

However, to accurately estimate a small p-value (e.g. p-value < 5e-8), we need a large B (e.g. > 1e+10), which can be extremely time-consuming. We propose using importance sampling to estimate a small p-value with a relatively small B.

Suppose $f(\boldsymbol{Z}_b)$ is the density function of the desired/true distribution of $\boldsymbol{Z}_b$ under the null. In this case, $f(\boldsymbol{Z}_b)$ is the density function of $\mathrm{MVN}(\mathbf{0}, \mathbf{R})$. We want to use a different distribution, called the importance distribution, to sample $\boldsymbol{Z}_b$, and the density function of this importance distribution is $g(\boldsymbol{Z}_b)$. $T(\boldsymbol{Z}_b)$ is a function of $\boldsymbol{Z}_b$. It can be any test statistic based on $\boldsymbol{Z}_b$ (e.g. the SPU test statistic with any power index). Based on the importance sampling theory, we want to choose an appropriate density function $g(\cdot)$ such that the ratio $f(\boldsymbol{Z}_b)/g(\boldsymbol{Z}_b)$ is well-defined and then

$$\mathrm{E}_f(T(\boldsymbol{Z}_b)) = \mathrm{E}_g(\frac{f(\boldsymbol{Z}_b)}{g(\boldsymbol{Z}_b)} T(\boldsymbol{Z}_b)),$$

where E_f and E_g denote the expectations over the density $f(\cdot)$ and $g(\cdot)$ respectively. This equation suggests that if we sample $\boldsymbol{Z}_b$ under g, then $[\sum_{b=1}^{B} w_b T(\boldsymbol{Z}_b)]/B$ is an unbiased estimate of $E(T(\boldsymbol{Z}_b))$, where $w_b = f(\boldsymbol{Z}_b)/g(\boldsymbol{Z}_b)$. Particularly, we consider

$$T(\boldsymbol{Z}_b) = I(|\mathrm{SPU}(\gamma, \boldsymbol{Z}_b)| > |\mathrm{SPU}(\gamma, \boldsymbol{Z})|).$$

$\boldsymbol{Z}$ is the observed statistic, treated as fixed for now. As a result, $\frac{1}{B}\sum_{b=1}^{B} w_b I(|\mathrm{SPU}(\gamma, \boldsymbol{Z}_b)| > |\mathrm{SPU}(\gamma, \boldsymbol{Z})|)$ is an unbiased estimate of $P_f(|\mathrm{SPU}(\gamma, \boldsymbol{Z}_b)| > |\mathrm{SPU}(\gamma, \boldsymbol{Z})|) = E_f(I(|\mathrm{SPU}(\gamma, \boldsymbol{Z}_b)| > |\mathrm{SPU}(\gamma, \boldsymbol{Z})|))$, which is the p-value of $\mathrm{SPU}(\gamma, \boldsymbol{Z})$. Here P_f represents the probability over the density $f(\cdot)$.

Now the only concern is how to choose a proper g to sample $\boldsymbol{Z}_b$. Denote the observed $\mathrm{SPU}(\gamma, \boldsymbol{Z})$ by t_γ. Based on the idea that g should allow $\boldsymbol{Z}_b$ to get to extreme values (i.e. comparable to the observed $\boldsymbol{Z}$) much more easily, we propose the following procedure for different γ's. We will discuss the theoretical efficiency of the proposed importance sampling procedure in the following section, which will also explain why we use different algorithms for $\gamma \leq 2$ and $\gamma > 2$.

1. for $\gamma = 1$ or 2: Sample $\mathbf{Z}_b$ from $\mathrm{MVN}(\boldsymbol{\mu}^*, \mathbf{R})$ or $\mathrm{MVN}(-\boldsymbol{\mu}^*, \mathbf{R})$ with the same probabilities, where $\boldsymbol{\mu}^* = (\frac{t_1}{p}, \frac{t_1}{p} \ldots \frac{t_1}{p})'$ if $\gamma = 1$; $\boldsymbol{\mu}^* = (\sqrt{\frac{t_2}{p}}, \sqrt{\frac{t_2}{p}} \ldots \sqrt{\frac{t_2}{p}})'$ if $\gamma = 2$ $(b = 1, 2 \ldots B)$. Calculate $\mathrm{SPU}(\gamma, \mathbf{Z}_b)$ and the p-value: $P_{\mathrm{SPU}(\gamma, \boldsymbol{Z})} = \frac{1}{B}\sum_{b=1}^{B} w_b I(|\mathrm{SPU}(\gamma, \boldsymbol{Z}_b)| > |\mathrm{SPU}(\gamma, \boldsymbol{Z})|)$, where

$$\frac{1}{w_b} = \frac{1}{2}\left(\frac{e^{-\frac{1}{2}(\boldsymbol{Z}_b - \boldsymbol{\mu}^*)' \boldsymbol{R}^{-1} (\boldsymbol{Z}_b - \boldsymbol{\mu}^*)}}{e^{-\frac{1}{2}\boldsymbol{Z}_b' \boldsymbol{R}^{-1} \boldsymbol{Z}_b}} + \frac{e^{-\frac{1}{2}(\boldsymbol{Z}_b + \boldsymbol{\mu}^*)' \boldsymbol{R}^{-1} (\boldsymbol{Z}_b + \boldsymbol{\mu}^*)}}{e^{-\frac{1}{2}\boldsymbol{Z}_b' \boldsymbol{R}^{-1} \boldsymbol{Z}_b}}\right).$$

2. $\gamma > 2$: For each b, simulate τ from $\mathrm{Unif}(1, 2 \ldots p)$ only once. Sample ${Z_b}^{(\tau)}$ from $N({t_\gamma}^*, 1)$ or $N(-{t_\gamma}^*, 1)$ with the same probabilities. If γ is finite, ${t_\gamma}^* = {t_\gamma}^{1/\gamma}$. If γ is infinite, ${t_\gamma}^* = t_\gamma$. Conditioning on ${Z_b}^{(\tau)}$, generate $({Z_b}^{(1)}, \ldots, {Z_b}^{(\tau-1)}, {Z_b}^{(\tau+1)}, \ldots, {Z_b}^{(p)})'$ from $\mathrm{MVN}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ with

$\boldsymbol{\mu}^* = {Z_b}^{(\tau)}\boldsymbol{\Sigma}_{12}$ and $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{12}'$. $\boldsymbol{\Sigma}_{11}$ is $\mathbf{R}$ without its τth column and row, and $\boldsymbol{\Sigma}_{12} = (R_{\tau 1} \ldots R_{\tau,\tau-1}\, R_{\tau,\tau+1} \ldots R_{\tau p})'$. Calculate $\mathrm{SPU}(\gamma, \boldsymbol{Z}_b)$ and p-value: $P_{\mathrm{SPU}(\gamma,\boldsymbol{Z})} = \frac{1}{B}\sum_{b=1}^{B} w_b I(|\mathrm{SPU}(\gamma, \boldsymbol{Z}_b)| > |\mathrm{SPU}(\gamma, \boldsymbol{Z})|)$, where w_b is defined by

$$\frac{1}{w_b} = \frac{1}{2p}\sum_{i=1}^{p} \frac{e^{-\frac{1}{2}({Z_b}^{(i)} - {t_\gamma}^*)^2}}{e^{-\frac{1}{2}({Z_b}^{(i)})^2}} + \frac{1}{2p}\sum_{i=1}^{p} \frac{e^{-\frac{1}{2}({Z_b}^{(i)} + {t_\gamma}^*)^2}}{e^{-\frac{1}{2}({Z_b}^{(i)})^2}}.$$

Note that for $\gamma = 1$ or 2, $\mathrm{SPU}(\gamma, \boldsymbol{\mu}^*) = t_\gamma$, and for $\gamma > 2$, $\mathrm{SPU}\big(\gamma, ({t_\gamma}^*, 0, \ldots, 0)'\big) = \ldots = \mathrm{SPU}\big(\gamma, (0, \ldots, {t_\gamma}^*)'\big) = t_\gamma$. This means that the SPU test statistics resulting from the shifted mean vectors are comparable to the observed ones, ensuring that we will be able to sample many extreme statistics but not too many. More rigorous justification for our proposed choice of the proposal distributions (i.e. the normal distributions with the shifted mean vectors) is provided by the two Propositions and following discussions in Section 2.3. More information on deriving w_b is also provided in Web Appendix A (page 7).

In addition to efficiently estimate the rare-event p-values, according to the importance sampling theory (Asmussen and Glynn, 2007), for each γ, an estimator of the variance of $P_{\mathrm{SPU}(\gamma,\boldsymbol{Z})}$ is $Var(P_{\mathrm{SPU}(\gamma,\boldsymbol{Z})}) = \frac{1}{B^2}\sum_{b=1}^{B}\big[w_b I(|\mathrm{SPU}(\gamma, \boldsymbol{Z}_b)| > |\mathrm{SPU}(\gamma, \boldsymbol{Z})|) - P_{\mathrm{SPU}(\gamma,\boldsymbol{Z})}\big]^2$. With the above variance estimator, we can obtain that the standard error of our importance sampling estimator $P_{\mathrm{SPU}(\gamma,\boldsymbol{Z})}$ based on B replications is $\sqrt{Var(P_{\mathrm{SPU}(\gamma,\boldsymbol{Z})})}$.

Note that the accuracy of the $\mathbf{R}$ estimate will generally affect the result of SPU tests and aSPU test (which will be introduced next), no matter whether we use importance sampling or the standard approach. The main focus of this paper is to compare our new method with the standard approach given the same $\mathbf{R}$ estimate. More information on estimating $\mathbf{R}$ can be found elsewhere (Wen and Stephens 2010; Kwak and Pan 2016).

2.2 Importance sampling for aSPU

For the SPU test, smaller γ's work better when the signals are dense (i.e. many SNPs in the region have nonzero effects) while larger γ's work better when the signals are sparse (i.e. only a few SNPs have nonzero effects). The main idea of the aSPU test is to combine the SPU tests with different γ's, so that the test will perform well in various situations. Suppose we choose γ from $\Gamma = \{\gamma_1, \gamma_2 \ldots \gamma_r\}$. The aSPU test statistic, aSPU$(\boldsymbol{Z})$, is simply the minimum of $P_{\mathrm{SPU}(\gamma,Z)}$'s. Applying the importance sampling method for each SPU test separately can give us aSPU$(\boldsymbol{Z})$, but that does not directly give us the p-value of this test statistic. We propose to use a mixture sampling procedure that can calculate not only aSPU$(\boldsymbol{Z})$ but also its p-value.

First, select weight q_i for each γ_i $(i = 1, 2, \ldots, r)$. We set $q_i = \frac{1}{r}$ for all i. For each b, randomly choose γ according to the weights. Follow the procedure for γ to obtain $\boldsymbol{Z}_b$. For each $\boldsymbol{Z}_b$, we can calculate the SPU test statistics SPU(γ_i, Z_b) and the weight w_b:

$$\frac{1}{w_b} = \sum_{i=1}^{r} \frac{q_i}{w_b^{(\gamma_i)}},$$

$$\frac{1}{w_b^{(\gamma_i)}} = \begin{cases} \frac{1}{2}\left(\frac{e^{-\frac{1}{2}(\boldsymbol{Z}_b - \boldsymbol{\mu}^*)' \boldsymbol{R}^{-1} (\boldsymbol{Z}_b - \boldsymbol{\mu}^*)}}{e^{-\frac{1}{2}\boldsymbol{Z}_b' \boldsymbol{R}^{-1} \boldsymbol{Z}_b}} + \frac{e^{-\frac{1}{2}(\boldsymbol{Z}_b + \boldsymbol{\mu}^*)' \boldsymbol{R}^{-1} (\boldsymbol{Z}_b + \boldsymbol{\mu}^*)}}{e^{-\frac{1}{2}\boldsymbol{Z}_b' \boldsymbol{R}^{-1} \boldsymbol{Z}_b}} \right) & \text{if } \gamma_i \le 2 \\ \frac{1}{2p} \sum_{j=1}^{p} \frac{e^{-\frac{1}{2}(Z_b^{(j)} - t_{\gamma_i}{}^*)^2}}{e^{-\frac{1}{2}(Z_b^{(j)})^2}} + \frac{1}{2p} \sum_{j=1}^{p} \frac{e^{-\frac{1}{2}(Z_b^{(j)} + t_{\gamma_i}{}^*)^2}}{e^{-\frac{1}{2}(Z_b^{(j)})^2}}, \quad t_{\gamma_i}{}^* = t_{\gamma_i}{}^{1/\gamma} \text{ if } 2 < \gamma_i < \infty \\ \frac{1}{2p} \sum_{j=1}^{p} \frac{e^{-\frac{1}{2}(Z_b^{(j)} - t_{\gamma_i}{}^*)^2}}{e^{-\frac{1}{2}(Z_b^{(j)})^2}} + \frac{1}{2p} \sum_{j=1}^{p} \frac{e^{-\frac{1}{2}(Z_b^{(j)} + t_{\gamma_i}{}^*)^2}}{e^{-\frac{1}{2}(Z_b^{(j)})^2}}, \quad t_{\gamma_i}{}^* = t_{\gamma_i} \text{ if } \gamma_i = \infty \end{cases}$$

where $\boldsymbol{\mu}^* = \left(\sqrt[\gamma_i]{\frac{t_{\gamma_i}}{p}}, \sqrt[\gamma_i]{\frac{t_{\gamma_i}}{p}} \ldots \sqrt[\gamma_i]{\frac{t_{\gamma_i}}{p}} \right)'$ for $\gamma_i \le 2$. Note that for each b, we only simulate γ once and use it to generate Z_b. Then we calculate $w_b = 1/(\sum_{i=1}^{r} q_i / w_b^{(\gamma_i)})$, using all γ_i's we consider (e.g. 1, 2, 4, 8, ∞). This is like simulating Z_b from a mixture of distributions.

Next, we calculate the p-values for the SPU tests and the aSPU test statistics:

$$P_{\mathrm{SPU}(\gamma_i,\boldsymbol{Z})} = \frac{1}{B}\sum_{b=1}^{B} w_b I(|\mathrm{SPU}(\gamma_i,\boldsymbol{Z}_b)| > |\mathrm{SPU}(\gamma_i,\boldsymbol{Z})|),$$

$$P_{\mathrm{SPU}(\gamma_i,\boldsymbol{Z}_b)} = \frac{1}{B-1}\sum_{b\prime=1,\ldots,B;\, b\prime\neq b} w_{b\prime} I(|\mathrm{SPU}(\gamma_i,\boldsymbol{Z}_{b\prime})| > |\mathrm{SPU}(\gamma_i,\boldsymbol{Z}_b)|),$$

$$\mathrm{aSPU}(\boldsymbol{Z}) = \min_i\left(P_{\mathrm{SPU}(\gamma_i,\boldsymbol{Z})}\right), \quad \mathrm{aSPU}(\boldsymbol{Z}_b) = \min_i\left(P_{\mathrm{SPU}(\gamma_i,\boldsymbol{Z}_b)}\right).$$

Finally, we can obtain the p-value of the aSPU test: $P_{\mathrm{aSPU}(\boldsymbol{Z})} = \frac{1}{B}\sum_{b=1}^{B} w_b I(\mathrm{aSPU}(\boldsymbol{Z}_b) < \mathrm{aSPU}(\boldsymbol{Z}))$, and its variance estimate $\frac{1}{B^2}\sum_{b=1}^{B}\left[w_b I(\mathrm{aSPU}(\boldsymbol{Z}_b) < \mathrm{aSPU}(\boldsymbol{Z})) - P_{\mathrm{aSPU}(\boldsymbol{Z})}\right]^2$.

Note that a good choice of q_i, the weight of each γ_i, may make the new approach more efficient. We propose an adaptive sampling approach that updates the weights four times during the process. Start from $q_i = \frac{1}{r}$ to get the first $B/5$ samples. Then we calculate the p-values of the observed SPU test statistics only based on the weights and test statistics of these samples. If γ_k has the smallest p-value, we double its weight, which means now $q_i = \frac{2}{r+1}$ for $i = k$, and $q_i = \frac{1}{r+1}$ for $i \neq k$. Generate another $B/5$ samples with the new weights and update the weights again in the same way (double the weight of γ_i that gives the smallest p-value), using the $B/5$ new samples. Continue this process to get B samples and combine them to finish the testing procedure. For convenience, we call the importance sampling methods with fixed weights and updated weights IMP and IMP2 respectively. Running IMP or IMP2 once will give us the p-values of aSPU and the corresponding SPU tests, and our experience suggests that obtaining the p-values of SPU from mixed samples is not worse, and sometimes better, than using the approach described in the previous subsection. Hence, we will only use IMP and IMP2 in the following. Note that based on the importance sampling theory, IMP and IMP2 provide unbiased estimates. Their computational efficiency (compared to the standard approach) will be shown in the simulation studies.

2.3 Theoretical properties

We have already shown that the estimates generated by importance sampling are unbiased. In this subsection, we discuss the efficiency of the new method. Let $L_\gamma(x)$ denote an estimator of the rare-event probability $\alpha_\gamma(x) = \mathrm{P}(|\mathrm{SPU}(\gamma, \mathbf{Z})| > |x|)$, which goes to 0 as $|x| \to \infty$. We consider $\alpha_\gamma(x) \in (0,1)$. To estimate $\alpha_\gamma(x)$, which is like computing the p-value of an observed x, we simulate B i.i.d. samples of $L_\gamma(x)$, $\{L_\gamma^{(b)}(x): b = 1, \ldots, B\}$ and obtain the average estimator $\bar{L}_\gamma(x) = B^{-1}\sum_{b=1}^{B} L_\gamma^{(b)}(x)$.

One way to look at whether $\bar{L}_\gamma(x)$ is efficient is to consider the relative error $|\bar{L}_\gamma(x) - \alpha_\gamma(x)|/\alpha_\gamma(x)$. We want to control it so that for some prescribed $\eta, \delta > 0$,

$$\mathrm{P}\{|\bar{L}_\gamma(x) - \alpha_\gamma(x)|/\alpha_\gamma(x) > \eta\} < \delta. \quad (1)$$

The standard Monte Carlo method generates samples from $\mathrm{MVN}(\mathbf{0}, \mathbf{R})$ and uses $L_\gamma(x) = I(|\mathrm{SPU}(\gamma, \mathbf{Z}_b)| > |x|)$. As a result, $\mathrm{E}(\bar{L}_\gamma(x)) = \mathrm{E}(L_\gamma(x)) = \alpha_\gamma(x)$. By Markov's inequality,

$$\mathrm{P}\{|\bar{L}_\gamma(x) - \alpha_\gamma(x)|/\alpha_\gamma(x) > \eta\} \le \mathrm{Var}\{\bar{L}_\gamma(x)\}/\{\alpha_\gamma^2(x)\eta^2\} = \mathrm{Var}\{L_\gamma(x)\}/\{B\alpha_\gamma^2(x)\eta^2\}. \quad (2)$$

To achieve the relative error control, we want $\mathrm{var}\{L_\gamma(x)\}/\{B\alpha_\gamma^2(x)\eta^2\} < \delta$. As $\mathrm{var}\{L_\gamma(x)\} = \alpha_\gamma(x)(1 - \alpha_\gamma(x))$ for the standard Monte Carlo estimator, it needs $B > \eta^{-2}\delta^{-1}\alpha_\gamma^{-1}(x)(1 - \alpha_\gamma(x))$. As $\alpha_\gamma(x) \to 0$, the standard Monte Carlo method becomes inefficient and even infeasible because $\eta^{-2}\delta^{-1}\alpha_\gamma^{-1}(x)(1 - \alpha_\gamma(x)) \to \infty$.

A more efficient estimator is the logarithmic efficient estimator (Asmussen and Glynn, 2007). An unbiased estimator $L_\gamma(x)$ of $\alpha_\gamma(x)$ is called logarithmic efficient if

$$\limsup_{|x|\to\infty} \frac{\mathrm{Var}\{L_\gamma(x)\}}{\alpha_\gamma^{2-\varepsilon}(x)} = 0, \tag{3}$$

for any $\varepsilon > 0$. When $L_\gamma(x)$ is logarithmic efficient, inequality (2) also holds by Markov's inequality, and therefore (3) implies that we only need $B = O\{\eta^{-2}\delta^{-1}\alpha_\gamma^{-\varepsilon}(x)\}$ i.i.d. replicates of $L_\gamma(x)$, for any $\eta > 0$. Compared with the standard MC, the logarithmic efficient estimators substantially reduce the computational cost, especially for small $\alpha_\gamma(x)$.

To construct a logarithmic efficient estimator, we use the importance sampling method. We construct the change of measure estimator (Section V.1 of Asmussen and Glynn, 2007) $L_\gamma(x) = w_b I(|\mathrm{SPU}(\gamma, \mathbf{Z}_b)| > |x|)$, where $w_b = f_\gamma(\mathbf{Z}_b)/g_\gamma(\mathbf{Z}_b)$. To have an insight of how the choice of $g_\gamma(\mathbf{Z}_b)$ could satisfy requirement (3), we first notice that (3) is equivalent to

$$\liminf_{|x|\to\infty} \frac{\log[\mathrm{Var}\{L_\gamma(x)\}]}{\log \alpha_\gamma^2(x)} \geq 1.$$

In addition, since $\mathrm{E}\{L_\gamma^2(x)\} \geq \mathrm{Var}\{L_\gamma(x)\}$ and $\mathrm{E}\{L_\gamma^2(x)\} \geq [\mathrm{E}\{L_\gamma(x)\}]^2 = \alpha_\gamma^2(x)$, (3) is also equivalent to

$$\lim_{|x|\to\infty} \frac{\log \mathrm{E}\{L_\gamma^2(x)\}}{\log \alpha_\gamma^2(x)} = 1, \tag{4}$$

the proof of which is provided in the supplementary materials. Then for the change of measure estimator $L_\gamma(x)$, if we choose $g_\gamma(\mathbf{Z}_b) = f_\gamma(\mathbf{Z}_b)$, $L_\gamma(x)$ reduces to the standard MC estimator $I(|\mathrm{SPU}(\gamma, \mathbf{Z}_b)| > |x|)$, and the left hand side of (4) then equals 1/2, which is smaller than 1, the right hand side. On the other hand, consider $G_x(\cdot)$ to be the conditional probability measure given $|\mathrm{SPU}(\gamma, \mathbf{Z}_b)| > |x|$, and has the density $f_\gamma^{G_x}(\mathbf{Z}_b) = \alpha_\gamma^{-1}(x) f_\gamma(\mathbf{Z}_b)$; then if we choose $g_\gamma(\mathbf{Z}_b) = f_\gamma^{G_x}(\mathbf{Z}_b)$, the left hand side of (4) is exactly 1. Note that this change of measure is of no practical use since $L_\gamma(x)$ depends on the unknown $\alpha_\gamma(x)$. But if we can find a change of measure $g_\gamma(\mathbf{Z}_b)$ that is a

good approximation of $f_\gamma^{G_x}(\boldsymbol{Z}_b)$, the conditional probability measure given $|\mathrm{SPU}(\gamma, \mathbf{Z}_b)| > |x|$, we would expect (4) to hold and the corresponding estimator $L_\gamma(x)$ to be efficient. In other words, the logarithmic efficiency criterion requires that the new density $g_\gamma(\mathbf{Z}_b)$ is a good approximation of the conditional distribution of interest. We then show that the new density $g_\gamma(\mathbf{Z}_b)$ given by the proposed importance sampling could approximate the conditional probability measure $f_\gamma^{G_x}(\boldsymbol{Z}_b)$ well. We first present the theoretical result (Proposition 1) showing that the proposed sampling method is logarithmic efficient for $\gamma = 1$ and $\gamma = \infty$.

Proposition 1 *When* $|x| \to \infty$, $L_\gamma(x) = w_b I(|\mathrm{SPU}(\gamma, \boldsymbol{Z}_b)| > |x|)$ *is logarithmic efficient for* $\gamma = 1, \infty$.

Note that when proving Proposition 1 for $\gamma = 1$, instead of choosing $\boldsymbol{\mu}^* = (t_1, t_1 \dots t_1)'/p$ as previously described, we use $\boldsymbol{\mu}^* = \mathbf{R1}a_0$, where $a_0 = |t_1|/(\mathbf{1}'\mathbf{R1})$ and $\mathbf{1}$ is a p-dimensional vector with all elements equal to 1. If $\mathbf{R}$ is an identity matrix, then $\boldsymbol{\mu}^* = \mathbf{R1}a_0 = (t_1, t_1 \dots t_1)'/p$. From our experience, using $(t_1, t_1 \dots t_1)'/p$ or $\mathbf{R1}a_0$ does not make a huge difference. Hence, we will still use the former in this paper, while the function in our package allows both ways.

We then discuss the sampling procedures for $\gamma = 2$ and $2 < \gamma < \infty$, which will be shown to be similar to the cases of $\gamma = 1$ and $\gamma = \infty$ respectively. When $\gamma = 2$, $\mathrm{SPU}(2, \mathbf{Z}) = \sum_{i=1}^{p} (z^{(i)})^2$, and we know $(z^{(i)})^2 \sim \chi_1^2$, which is a light tail distribution, i.e., it decays at an exponential rate or faster. Note that when $\gamma = 1$, $\mathrm{SPU}(1, \mathbf{Z}) = \sum_{i=1}^{p} z^{(i)}$ and $z^{(i)} \sim \mathrm{N}(0,1)$ also has light tail. We then know that the two cases when $\gamma = 1$ and $\gamma = 2$ are similar in the sense that they simulate the tail probabilities of a sum of light tail distributions. Following the discussion on rare-event simulations with light tail distributions in Chapter V1.2 in Asmussen and Glynn (2007), this motivates us to use a similar sampling procedure when $\gamma = 1, 2$. Then analogous to the case when $\gamma = 1$, we expect that when $\gamma = 2$, the change of measure $g_2(\boldsymbol{Z}_b)$ is also a good approximation to the conditional probability measure $f_2^{G_x}(\boldsymbol{Z}_b)$. We

provide an illustrative example for $\gamma = 2$ when $p = 1$ in Figure 1; the figure appears in color in the electronic version of this article, and any mention of color refers to that version. The targeted conditional distribution is $f_1(z) = P(z \mid z^2 > 4^2)$ and the proposed bimodal mixture distribution is f_3. For comparison, we also consider a single normal f_2 with a large variance to approximate f_1. It can be seen that the bimodal mixture distribution f_3 provides a better approximation of f_1 than f_2 does. For $2 < \gamma < \infty$, we show below that $(z^{(i)})^\gamma$'s follow heavy tail distributions.

Proposition 2 *For* $\gamma = \infty$,

$$\lim_{|t_\infty| \to \infty} \frac{\mathrm{P}\{\mathrm{SPU}(\infty, \mathbf{Z}) > |t_\infty|\}}{\mathrm{P}\{|z^{(i)}| > |t_\infty|\}} = p.$$

For even γ, $2 < \gamma < \infty$,

$$\liminf_{|t_\gamma| \to \infty} \frac{\mathrm{P}\{\mathrm{SPU}(\gamma, \mathbf{Z}) > |t_\gamma|\}}{\mathrm{P}\left\{\left(z^{(i)}\right)^\gamma > |t_\gamma|\right\}} \geq p; \tag{5}$$

when we further assume $\mathbf{R} = \mathbf{I}_p$,

$$\lim_{|t_\gamma| \to \infty} \frac{\mathrm{P}\{\mathrm{SPU}(\gamma, \mathbf{Z}) > |t_\gamma|\}}{\mathrm{P}\left\{\left(z^{(i)}\right)^\gamma > |t_\gamma|\right\}} = p. \tag{6}$$

Note that (6) suggests that marginally $(z^{(i)})^\gamma$'s with $2 < \gamma < \infty$ satisfies the definition of "subexponential" distributions (see, e.g., Teugels, 1975), which is a type of heavy tail distribution. Specifically, the proof of Proposition 2 shows that $\mathrm{P}\{(z^{(i)})^\gamma > x\}$ decays at rate $e^{-\frac{1}{2}x^{2/\gamma}}$, which is slower than the exponential rate as $\gamma > 2$. In addition, when $z^{(i)}$'s are jointly independent, Proposition 2 shows that $\mathrm{P}\{\mathrm{SPU}(\infty, \mathbf{Z}) > |t_\infty|\} \sim p\mathrm{P}\{|z^{(i)}| > |t_\infty|\}$ and $\mathrm{P}\{\mathrm{SPU}(\gamma, \mathbf{Z}) > |t_\gamma|\} \sim p\mathrm{P}\{(\mathrm{z}^{(i)})^\gamma > |t_\gamma|\}$ for $2 < \gamma < \infty$. Therefore,

$$\mathrm{P}\{\mathrm{SPU}(\gamma,\mathbf{Z}) > |t_\gamma|\} \sim \mathrm{P}\left\{\mathrm{SPU}(\infty,\mathbf{Z}) > |t_\gamma|^{1/\gamma}\right\} = \mathrm{P}\left\{\max_{1\le i\le p}(z^{(i)})^\gamma > |t_\gamma|\right\}.$$

This implies that the conditional probability

$$\mathrm{P}\{\max_{1\le i\le p}(z^{(i)})^\gamma > |t_\gamma| \mid \mathrm{SPU}(\gamma,\mathbf{Z}) > |t_\gamma|\} = \frac{\mathrm{P}\{\max_{1\le i\le p}(z^{(i)})^\gamma > |t_\gamma|,\ \mathrm{SPU}(\gamma,\mathbf{Z}) > |t_\gamma|\}}{\mathrm{P}\{\mathrm{SPU}(\gamma,\mathbf{Z}) > |t_\gamma|\}} \to 1, \quad (7)$$

as $\mathrm{P}\{\max_{1\le i\le p}(z^{(i)})^\gamma > |t_\gamma|, \mathrm{SPU}(\gamma,\mathbf{Z}) > |t_\gamma|\} = \mathrm{P}\{\max_{1\le i\le p}(z^{(i)})^\gamma > |t_\gamma|\}$. For a general correlation $\mathbf{R}$, following similar analysis, we have the conditional probability (7) ≤ 1, and under certain technical weak-dependence assumptions, it can be further shown that (7) is close to 1 (Geluk and Tang, 2009). The conditional probability (8) approaching 1 suggests that conditioning on the event $\mathrm{SPU}(\gamma,\mathbf{Z}) > |t_\gamma|$, the probability of one single $z^{(i)}$ to become large tends to be big. Intuitively, our proposed sampling procedures for $2 < \gamma < \infty$ and $\gamma = \infty$ mimic this tail behavior of one single $z^{(i)}$ being large. Specifically, we randomly sample one index τ and shift the mean of $z^{(\tau)}$ to $\pm t_\gamma^*$ so that the probability of $(z^{(\tau)})^\gamma$ being large is big. Then conditioning on this "large" one, we sample the remaining $z^{(i)}$'s. Following the discussions on the rare-event simulation of the summation of heavy tail distributions in Chapter VI.3 in Asmussen and Glynn (2007), we also expect that the change of measure $g_\gamma(\mathbf{Z}_b)$ could approximate the conditional probability measure $f_\gamma^{G_x}(\mathbf{Z}_b)$ well, similarly to the case when $\gamma = \infty$.

As for aSPU, we also expect the proposed sampling procedure to be efficient because it is based on a combination of the procedures for SPU. Proofs of the propositions can be found in Web Appendix A.

2.4 Comparison with Other Methods

We would like to mention that there are other approaches that can potentially improve upon standard MC, but most of them cannot be directly applied to the aSPU test. For example, the saddle-point

approximation (Daniels 1954) cannot be easily applied because it requires the moment generating function of the target statistic. The moment generating function of the aSPU statistic does not have a simple closed-form, as the aSPU statistic is a complicated combination of the high order moments of multivariate Gaussian random vectors. The MCMC-CE (Markov Chain Monte Carlo-Cross Entropy) method proposed by Shi et al. (2018) seems comparable to our method since it also incorporates the importance sampling technique and can be used to test SNP sets. Nevertheless, this approach cannot be directly applied to aSPU since the aSPU test statistic cannot be written as an explicit function of some multivariate normal variables. In fact, MCMC-CE cannot be easily applied to SPU(γ) with $\gamma > 2$ either. More discussions and some simulation results for comparing IMP with MCMC-CE are provided in the supplementary materials.

3. Simulation Studies

We did some simulations to compare the numerical performance and speed of the importance sampling and the standard MC (denoted STD) with GWAS summary statistics. First, we looked at the Blood Cell Consortium (BCX) GWAS data (Chami et al. 2016) on hematocrit (HCT) with 808 subjects of European ancestry, focusing on chromosome 22. We used 381 subjects of European ancestry from the 1000 Genomes Project data (The 1000 Genomes Project Consortium 2015) as our reference panel. Then we selected $p = 20$ SNPs that were present in both datasets with minor allele frequencies > 0.05. For each subject i, the trait Y_i was simulated by a linear model $Y_i = \sum_{j=1}^{p} X_{ij}\beta_{ij} + e_i$, where e_i followed a standard normal distribution, and X_{ij} was the genotype of SNP j for subject i. We assumed that $\beta_{ij} = \theta > 0$ for $j = 1, \ldots, k$, and $\beta_{ij} = 0$ for $j = k+1, \ldots, q$, which means the first k SNPs were causal with effect size θ, and the rest of the SNPs had no effect on the trait. The data were generated once, and then analyzed by different methods multiple times. We also applied the standard MC-based

aSPU test with at least 1e+7 iterations to give the approximate "true" p-values of SPU and aSPU, as well as their confidence intervals.

As shown in Table 1, the standard MC (STD) failed to give good approximations to the p-values when B was not large enough, while IMP and IMP2 were always able to provide better estimates. In most cases, IMP was very efficient since its estimates were close to the results of STD with at least 1e+8 iterations and had relatively small standard deviations. IMP2 performed close to, but not significantly better than, IMP. In terms of speed, also shown in Table 1, even with the same number of iterations, the functions of IMP and IMP2 implemented in Rcpp were much faster than the standard function for SPUs. We performed the analyses on a single laptop. For the IMP function, 1e+5 iterations took about one second and 1e+6 iterations took about 10 seconds. In practice, 1e+5 is usually large enough for the importance sampling methods, while the standard MC needs a huge B (e.g. 1e+8 or 1e+9) to work well, which means using IMP and IMP2 may save a large amount of computing time.

We looked at another scenario with 3 significant SNPs and much smaller p-values. As shown in Table 2, IMP provided very good estimates for most tests with 1e+5 (and sometimes even 1e+4) iterations, which agreed well with the results from STD with 1e+9 iterations. Again, STD failed to obtain valid results in most cases, and IMP2 performed similarly to IMP.

To further examine how the computation time changes with different numbers of SNPs and different numbers of iterations, we carried out more simulation studies with $p = 20, 50, 100, 150$ and $B = 10^3, 10^4, 10^5, 10^6$. Since these studies were only focused on computation time, to keep things simple, we simulated Z-scores $(Z_1, Z_2, \ldots, Z_p)$ from a multivariate normal distribution with mean 0 and $Cov(Z_{j_1}, Z_{j_2}) = (0.5)^{|j_1 - j_2|}$. We applied different methods to these Z-scores 10 times and plotted the mean computation time in Figure 2. We only looked at aSPU since the SPU tests share the same computation time as a part of the aSPU function. As expected, all methods took longer time as p

increased, while IMP and IMP2 (implemented with Rcpp) were already able to save time compared to the standard MC (STD) (implemented in the aSPU package) with the same and not too small number of iterations (e.g. 1e+5 or 1e+6), which was consistent with Table 1. Since in reality IMP and IMP2 require much less iterations than STD (e.g. 1e+6 vs. 1e+9), the actual time saved may be even more. For $p = 150$, STD with 1e+9 iterations is expected to take more than 60 hours (since it already takes about 4 minutes to do 1e+6 iterations), while IMP with 1e+6 iterations only takes about 2 minutes. These results were based on a single CPU with R version 3.5.3. In practice, multiple cores can be used to shorten the computation time, but usually we need to test thousands of SNP sets instead of just one, so IMP and IMP2 will still have an obvious advantage.

In addition to shown in Tables 1-2, we examined how our new methods performed in some more extreme scenarios with even smaller p-values. The results showed that IMP was able to yield good estimates with 1e+6 iterations, which could potentially cut down the computation time from more than 20 days (for STD) to less than 30 seconds. Detailed information is provided in the supplementary materials.

4. Applications to the WTCCC Data

To further demonstrate the advantage of IMP and IMP2 over the standard MC approach with real data, we applied the methods to the 4572 genes with complete data in the WTCCC (Wellcome Trust Case Control Consortium 2010) GWAS of Crohn's disease, which had 1748 cases (i.e. patients with Crohn's disease) and 2938 controls after preprocessing. The number of SNPs in each gene varied; in total there were 66849 SNPs. One aim was to identify which genes were associated with Crohn's disease by applying the powerful aSPU test. Our goal was to show that in such an analysis, for non-extreme p-values IMP, IMP2 and standard MC would give similar results, while for extreme p-values, IMP and IMP2 could yield valid estimates with much fewer iterations than the standard MC. Before applying

different methods, we obtained the Z-scores from marginal logistic regression for each SNP based on the individual level case-control data, and then used the whole control group as a reference panel to estimate **R**. For the importance sampling approach, we first tried 1e+3 iterations. If the p-value of one gene turned out to be less than 0.01, we would use 1e+5 iterations to get a more accurate p-value for this gene. Then we compared the QQ plots of IMP, IMP2 and the standard MC (as implemented in aSPUs). The standard MC (STD) used 1e+6 iterations. As Figure 3 shows, the results of IMP, IMP2 and the standard MC were very similar for almost all (large) p-values, except that IMP and IMP2 were able to give three extremely small p-values (e.g. < 1e-10) using no more than 1e+5 iterations, while the smallest (non-zero) p-value that the standard MC could give was 1e-6 using 1e+6 iterations. The bottom panels of Figure 3 directly compare the p-values of the methods to confirm that IMP and IMP2 perform closely to STD when the p-value is not extreme. To see whether IMP and IMP2 could control type I errors, we did the same analysis with a control-control design. We randomly chose half of the 4686 subjects to be the control group and the rest to be the "case" group. Then we used the new "disease status" to calculate the Z-scores and applied different methods with the same reference panel as done before. As shown in Figure 4, without extremely small p-values, all three methods gave similar results, and were able to control the type I error rate satisfactorily. These two studies (case-control and control-control) showed IMP and IMP2 had no problem estimating non-extreme p-values (the results were almost the same as the standard MC) while being able to calculate extreme p-values with much fewer iterations.

5. Discussion

We have presented an importance sampling approach, IMP, to speed up the p-value calculation for the adaptive aSPU test. The main idea is to design a suitable proposal distribution to facilitate more frequent sampling of more extreme values (from the tails of the null distribution), and then calculate the

p-values using the samples and their weights. Our derived theoretical results show that the new method achieves high efficiency. In addition, we have also proposed another more sophisticated implementation of importance sampling, called IMP2, which updates the weights during the sampling process. Although we conjectured that IMP2 might perform better than IMP, the simulation study and real data application showed their similar performance, both estimating small p-values better and much faster than the standard Monte Carlo simulation. To estimate a p-value around 5e-8, the standard MC approach usually needs at least 1e+9 iterations (taking about 20+ hours), while IMP can do it efficiently with 1e+5 iterations (about 1 second).

In the future, we may modify the scheme of updating the weights to make IMP2 more efficient. More importantly, it is possible and worthwhile to extend our proposed importance sampling approach to other tests that require Monte Carlo simulations with normal variates (or with other known distribution) to calculate their p-values. Besides, we have been focusing on common variants in this paper. A main challenge for applying our new method to rare variants is that their Z statistics can no longer be well approximated by a normal (or other known) distribution, and thus we cannot directly apply our proposed importance sampling technique with a multivariate normal distribution. More investigation is warranted.

Acknowledgements

In our opinion, the first two authors should be treated as the co-first-authors. We thank the editors and reviewers for many insightful and helpful comments. This research was supported by NIH grants R01GM113250, R01GM126002, R01HL105397, R01HL116720, R21AG057038 and R01AG065636; by NSF grants DMS 1711226, DMS 1712717, SES 1659328 and SES-1846747; and by the MSI at the University of Minnesota. This study makes use of data generated by the WTCCC; a full list of the

investigators who contributed to the generation of the data is available from www.wtccc.org.uk; funding for the project was provided by the Wellcome Trust under award 076113 and 085475.

Data Availability Statement

The Blood Cell Consortium (BCX) GWAS data concerning red blood cells (Chami et al. 2016) is publicly available at http://www.mhi-humangenetics.org/resources. The 1000 Genomes Project data (The 1000 Genomes Project Consortium 2015) is available at http://www.internationalgenome.org/data. The Wellcome Trust Case Control Consortium (WTCCC) data (Wellcome Trust Case Control Consortium 2010) is available upon request at https://www.wtccc.org.uk.

References

Adler R. J., Blanchet J. H. and Liu J. (2012). Efficient Monte Carlo for high excursions of Gaussian random fields. *The Annals of Applied Probability*, 22(3), 1167-1214.

Asimit A.V., Furman E., Tang Q., Vernic R. (2011). Asymptotics for risk capital alloca- tions based on conditional tail expectation. *Insurance: Mathematics and Economics, 49,* 310–324.

Asmussen S. and Glynn P. W. (2007). Stochastic simulation: algorithms and analysis, Volume 57. Springer Science & Business Media.

Chami N., Chen M. H., Slater A. J., Eicher J. D., Evangelou E., Tajuddin S. M., Love-Gregory L., Kacprowski T., Schick U. M., Nomura A. *and others.* (2016). Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits. *Am. J. Hum Genet*, *99*, 8-21.

Chen L. S., Hutter C. M., Potter J. D., Liu Y., Prentice R. L., et al., (2010). Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *Am. J. Hum. Genet. 86*, 860–871.

Cochran W. G. (1977). Sampling Techniques (3rd Ed). John Wiley & Sons, New York.

He Y. and Xu G. (2018). Estimating tail probabilities of the ratio of the largest eigenvalue to the trace of a Wishart matrix. *Journal of Multivariate Analysis*, 166, 320–334.

Daniels, H. E. (1954). Saddlepoint Approximations in Statistics. *The Annals of Mathematical Statistics*, 25 (4): 631–650, doi:10.1214/aoms/1177728652

Geluk, J. and Tang, Q. (2009). Asymptotic tail probabilities of sums of dependent subexponential random variables. *Journal of Theoretical Probability*, 22(4):871.

Hesterberg T. (1995). Weighted Average Importance Sampling and Defensive Mixture Distributions. *Technometrics, 37*(2), 185-194.

Huang J., Wang K., Wei P., Liu X., Liu X., Tan K., Boerwinkle E., Potash J.B., Han S. (2016). FLAGS: A Flexible and Adaptive Association Test for Gene Sets Using Summary Statistics. *Genetics, 202*, 919-929.

Jiang T., Leder K. and Xu G. (2017). Rare-event analysis for extremal eigenvalues of white Wishart matrices. *The Annals of Statistics*, *45(4)*, 1609-1637.

Kimmel G., Shamir R. (2006). A fast method for computing high-significance disease association in large population-based studies. *Am J Hum Genet, 79*, 481-492.

Kwak I. Y. and Pan W. (2015). Adaptive gene- and pathway-trait association testing with GWAS summary statistics. *Bioinformatic*, *32*(8), 1178-84.

Lee S., Emond M. J., Bamshad M. J., Barnes K. C., Rieder M. J., Nickerson D. A., NHLBI GO Exome Sequencing Project-ESP Lung Project Team, Christiani D. C., Wurfel M. M., Lin X. (2012). Optimal unified approach for rare variant association testing with application to small sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet. 91*, 224–237.

Li X. and Xu G. (2018). Uniformly efficient simulation for extremes of Gaussian random fields. *Journal of Applied Probability, 55*(1), 157-178. doi:10.1017/jpr.2018.11

Liang F., Liu C., Carroll R.J. (2007). Stochastic approximation in Monte Carlo computation. *Journal of the American Statistical Association, 102*, 305–320.

Lin D.Y. (2005) An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics, 21*, 781–787.

Liu J. and Xu G. (2014a). On the Conditional Distributions and the Efficient Simulations of Exponential Integrals of Gaussian Random Fields. *Annals of Applied Probability* 24(4), 1691–1738.

Liu J. and Xu G. (2014b). Efficient Simulations for the Exponential Integrals of Hölder Continuous Gaussian Random Fields. *ACM Trans on Modeling Comp Sim*, 24(2), 9:1–9:24.

Ma Y., Wei P. (2019). FunSPU: A versatile and adaptive multiple functional annotation-based association test of whole-genome sequencing data. *PLoS Genet, 15(4)*, e1008081.

Pan W. (2009). Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic epidemiology*, *33*(6), 497-507.

Pan W., Kim J., Zhang Y., Shen X. and Wei P. (2014). A powerful and adaptive association test for rare variants. *Genetics, 197*(4), 1081-95.

Pan W., Kwak I. Y. and Wei, P. (2015). A Powerful Pathway-Based Adaptive Test for Genetic Association with Common or Rare Variants. *American journal of human genetics*, *97*(1), 86-98.

Shi J., Siegmund D., Yakir B. (2007). Importance sampling for estimating p values in linkage analysis. *Journal of the American Statistical Association, 102*, 929–937.

Shi Y., Wang M., Shi W., Lee J.-H., Kang H. and Jiang H. (2019). Accurate and efficient estimation of small P-values with the cross-entropy method: applications in genomic data analysis, Bioinformatics, 35, 2441–2448.

Su Y.R., Gauderman W.J., Berhane K., Lewinger J.P. (2015). Adaptive Set-Based Methods for Association Testing. *Genet Epi, 40*, 113-122.

Su Y.-R., Di C.-Z., Hsu L., Genetics and Epidemiology of Colorectal Cancer Consortium. (2017). A unified powerful set-based test for sequencing data analysis of GxE interactions. *Biostatistics, 18*,119–131.

Teugels J. L. (1975). The class of subexponential distributions. *The Annals of Probability*, 1000–1011.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation, *Nature,* 526, 68-74.

Wellcome Trust Case Control Consortium (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, *464*(7289), 713-20.

Wen X. and Stephens M. (2010). Using linear predictors to impute allele frequencies from summary or pooled genotype data. *Ann. Appl. Stat.*, 4, 1158-1182.

Wu M. C., Lee S., Cai T., Li Y., Boehnke M. and Lin X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet. 89*, 82–93.

Xu G., Lin L., Wei P., and Pan W. (2016). An adaptive two-sample test for high-dimensional means. *Biometrika*, *103*(3), 609-624.

Yang T., Chen H., Tang H., Li D., Wei P. (2019). A powerful and data-adaptive test for rare-variant–based gene-environment interaction analysis. *Statistics in Med, 38*, 1230–1244.

Yu K., Liang F., Ciampa J., Chatterjee N. (2011). Efficient p-value evaluation for resampling-based tests. *Biostatistics, 12*, 582-593.

Zhang H., Shi J., Liang F., Wheeler W., Stolzenberg-Solomon R., Yu K. (2014). A fast multilocus test with adaptive SNP selection for large-scale genetic-association studies. *European Journal of Human Genetics, 22*, 696–702.

Supporting Information

Web Appendices, Tables, and Figures referenced in Sections 2.3, 2.4 and 3 are available with this paper at the Biometrics website on Wiley Online Library. The new methods implemented in an R package will be available at https://github.com/yangq001/IMP.

Table 1 Mean p-values (SD) [computation time in seconds] of different methods on one dataset using 12 runs. $k = 10,\ \theta = 0.26,\ \Gamma = (1, 2, 4, 8, \infty)$. Average computation times for the SPU tests are the same as those for aSPU.

	SPU(1) (1.5e-6, 2.1e-6)*			
	$B = 10^3$	$B = 10^4$	$B = 10^5$	$B = 10^6$
STD	0 (0)	0 (0)	0 (0)	1.2e-6 (1.0e-6)
IMP	2.3e-6 (1.3e-6)	1.8e-6 (4.0e-7)	1.7e-6 (8.0e-8)	1.7e-6 (3.0e-8)
IMP2	1.9e-6 (7.0e-7)	1.7e-6 (2.1e-7)	1.7e-6 (3.7e-8)	1.7e-6 (2.7e-8)
	SPU(2) (1.6e-4, 1.7e-4)*			
	$B = 10^3$	$B = 10^4$	$B = 10^5$	$B = 10^6$
STD	2.5e-4 (6.2e-4)	1.6e-4 (1.4e-4)	1.7e-4 (3.7e-5)	1.7e-4 (8.4e-6)
IMP	1.7e-4 (1.1e-4)	1.5e-4 (3.3e-5)	1.7e-4 (1.5e-5)	1.6e-4 (5.3e-6)
IMP2	9.9e-5 (7.7e-5)	1.4e-4 (9.0e-5)	1.6e-4 (2.0e-5)	1.6e-4 (6.8e-6)
	SPU(4) (6.5e-4, 6.6e-4)*			
	$B = 10^3$	$B = 10^4$	$B = 10^5$	$B = 10^6$
STD	5.8e-4 (6.7e-4)	6.3e-4 (2.6e-4)	6.3e-4 (8.6e-5)	6.6e-4 (3.1e-5)

IMP	6.9e-4 (1.4e-4)	6.5e-4 (2.7e-5)	6.6e-4 (1.2e-5)	6.6e-4 (4.0e-6)
IMP2	6.4e-4 (1.2e-4)	6.5e-4 (4.9e-5)	6.6e-4 (1.2e-5)	6.5e-4 (4.6e-6)
	SPU(∞) (2.5e-2, 2.5e-2)*			
	$B = 10^3$	$B = 10^4$	$B = 10^5$	$B = 10^6$
STD	2.4e-2 (4.3e-3)	2.4e-2 (1.4e-3)	2.5e-2 (6.5e-4)	2.5e-2 (1.3e-4)
IMP	2.6e-2 (3.0e-3)	2.5e-2 (3.3e-4)	2.5e-2 (2.1e-4)	2.5e-2 (5.5e-5)
IMP2	2.6e-2 (2.8e-3)	2.5e-2 (9.0e-4)	2.5e-2 (2.9e-4)	2.5e-2 (9.8e-5)
	aSPU (5.8e-6, 6.8e-6)*			
	$B = 10^3$	$B = 10^4$	$B = 10^5$	$B = 10^6$
STD	1.0e-3 (0)	1.0e-4 (0)	1.0e-5 (0)	5.1e-6 (3.4e-6)
	[0.068]	[0.57]	[5.9]	[48]
IMP	9.0e-6 (4.1e-6)	6.6e-6 (1.7e-6)	6.0e-6 (3.0e-7)	6.0e-6 (9.3e-8)
	[0.024]	[0.12]	[1.1]	[11]
IMP2	1.3e-5 (6.9e-6)	7.1e-6 (2.0e-6)	5.9e-6 (2.0e-7)	5.9e-6 (8.9e-8)
	[0.024]	[0.099]	[0.72]	[7.9]

* 95% confidence intervals of p-values based on STD with 10^8 iterations.

Table 2 Mean p-values (SD) of different methods on one dataset using 12 runs. $k = 3$, $\theta = 0.41$, $\Gamma = (1, 2, 4, 8, \infty)$. Average computation times are similar to Table 1.

	SPU(1) (1.2e-1, 1.2e-1)*			
	$B = 10^3$	$B = 10^4$	$B = 10^5$	$B = 10^6$
STD	1.2e-1 (1.2e-2)	1.2e-1 (2.7e-3)	1.2e-1 (1.2e-3)	1.2e-1 (4.6e-4)
IMP	1.2e-1 (1.7e-2)	1.1e-1 (5.2e-3)	1.2e-1 (1.6e-3)	1.2e-1 (5.0e-4)
IMP2	1.2e-1 (2.0e-2)	1.2e-1 (6.4e-3)	1.2e-1 (2.3e-3)	1.2e-1 (8.1e-4)
	SPU(2) (2.1e-6, 2.3e-6)*			
	$B = 10^3$	$B = 10^4$	$B = 10^5$	$B = 10^6$
STD	0 (0)	0 (0)	1.7e-6 (3.9e-6)	1.9e-6 (1.1e-6)
IMP	5.1e-7 (4.7e-7)	1.0e-6 (4.9e-7)	1.3e-6 (4.7e-7)	2.5e-6 (8.8e-7)
IMP2	2.7e-6 (6.1e-6)	1.1e-6 (6.0e-7)	1.7e-6 (6.8e-7)	2.4e-6 (1.0e-6)
	SPU(4) (4.7e-8, 7.9e-8)*			
	$B = 10^3$	$B = 10^4$	$B = 10^5$	$B = 10^6$
STD	0 (0)	0 (0)	0 (0)	0 (0)
IMP	5.8e-8 (1.7e-8)	6.1e-8 (7.4e-9)	6.2e-8 (2.7e-9)	6.1e-8 (8.5e-10)
IMP2	6.1e-8 (2.1e-8)	6.3e-8 (9.0e-9)	6.0e-8 (1.6e-9)	6.1e-8 (6.5e-10)

SPU(∞) (3.1e-9, 1.5e-8)*				
	$B = 10^3$	$B = 10^4$	$B = 10^5$	$B = 10^6$
STD	0 (0)	0 (0)	0 (0)	0 (0)
IMP	6.5e-9 (7.4e-10)	6.7e-9 (2.4e-10)	6.7e-9 (6.0e-11)	6.7e-9 (2.1e-11)
IMP2	6.7e-9 (9.4e-10)	6.7e-9 (2.5e-10)	6.7e-9 (6.9e-11)	6.7e-9 (2.2e-11)
aSPU (2.3e-8, 4.7e-8)*				
	$B = 10^3$	$B = 10^4$	$B = 10^5$	$B = 10^6$
STD	1.0e-3 (0)	1.0e-4 (0)	1.0e-5 (0)	1.0e-6 (0)
IMP	4.5e-6 (8.5e-6)	6.6e-8 (8.6e-8)	2.7e-8 (3.2e-9)	2.5e-8 (6.1e-10)
IMP2	1.1e-5 (2.3e-5)	6.0e-8 (4.1e-8)	2.9e-8 (6.9e-9)	2.5e-8 (1.3e-9)

* 95% confidence intervals of p-values based on STD with 10^9 iterations.

Figure 1 Density plots of three distributions. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

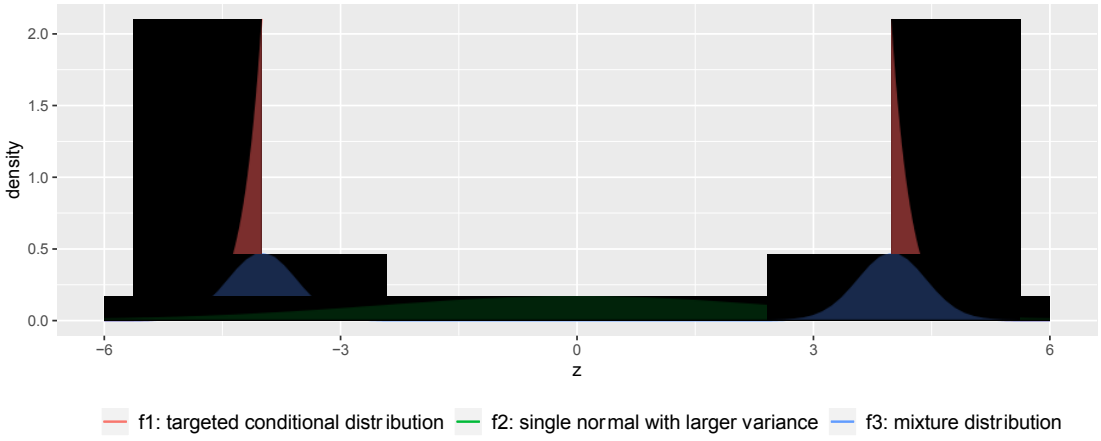


Figure 2 Computation time (in seconds) of different methods with different numbers of SNPs (p) (averaged over 10 runs). This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

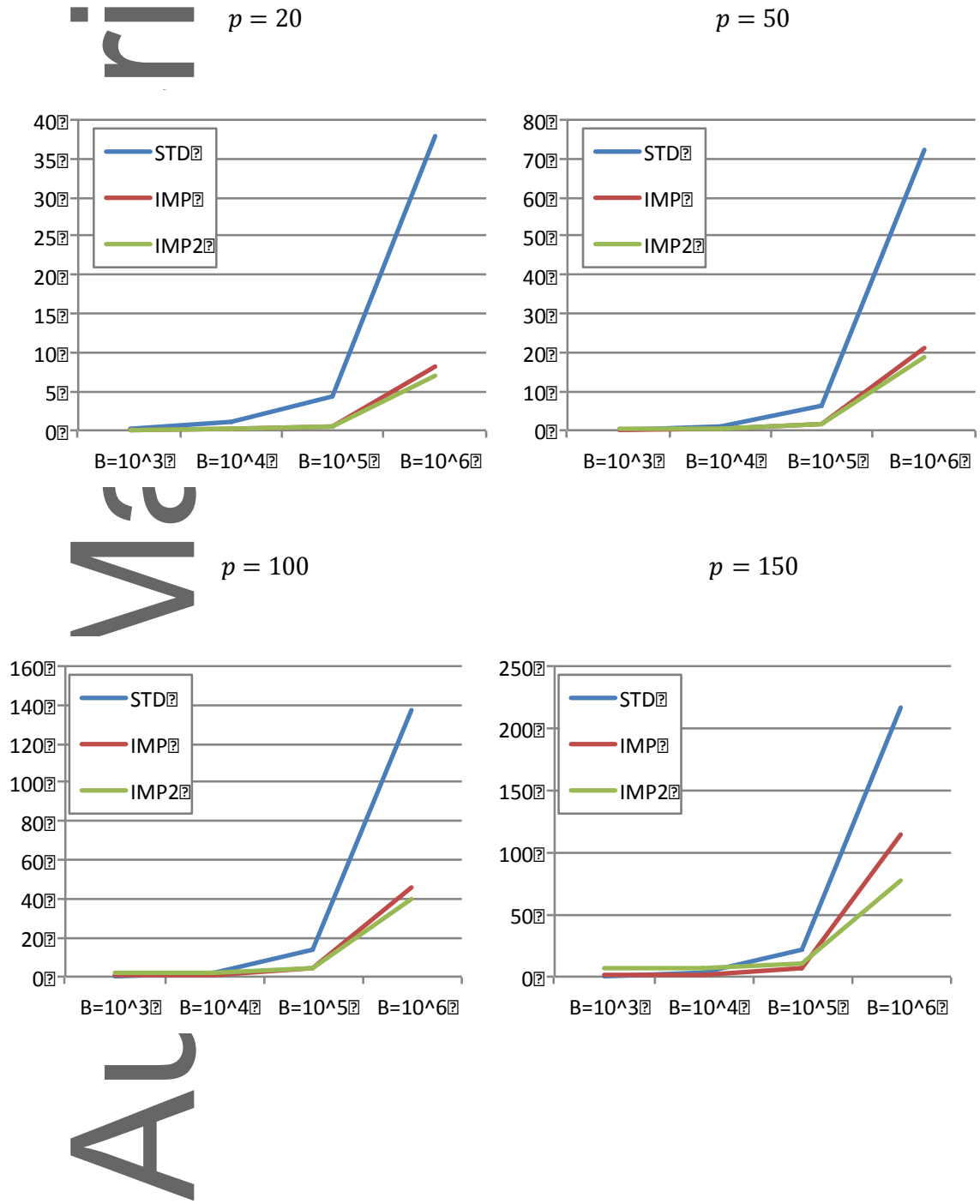
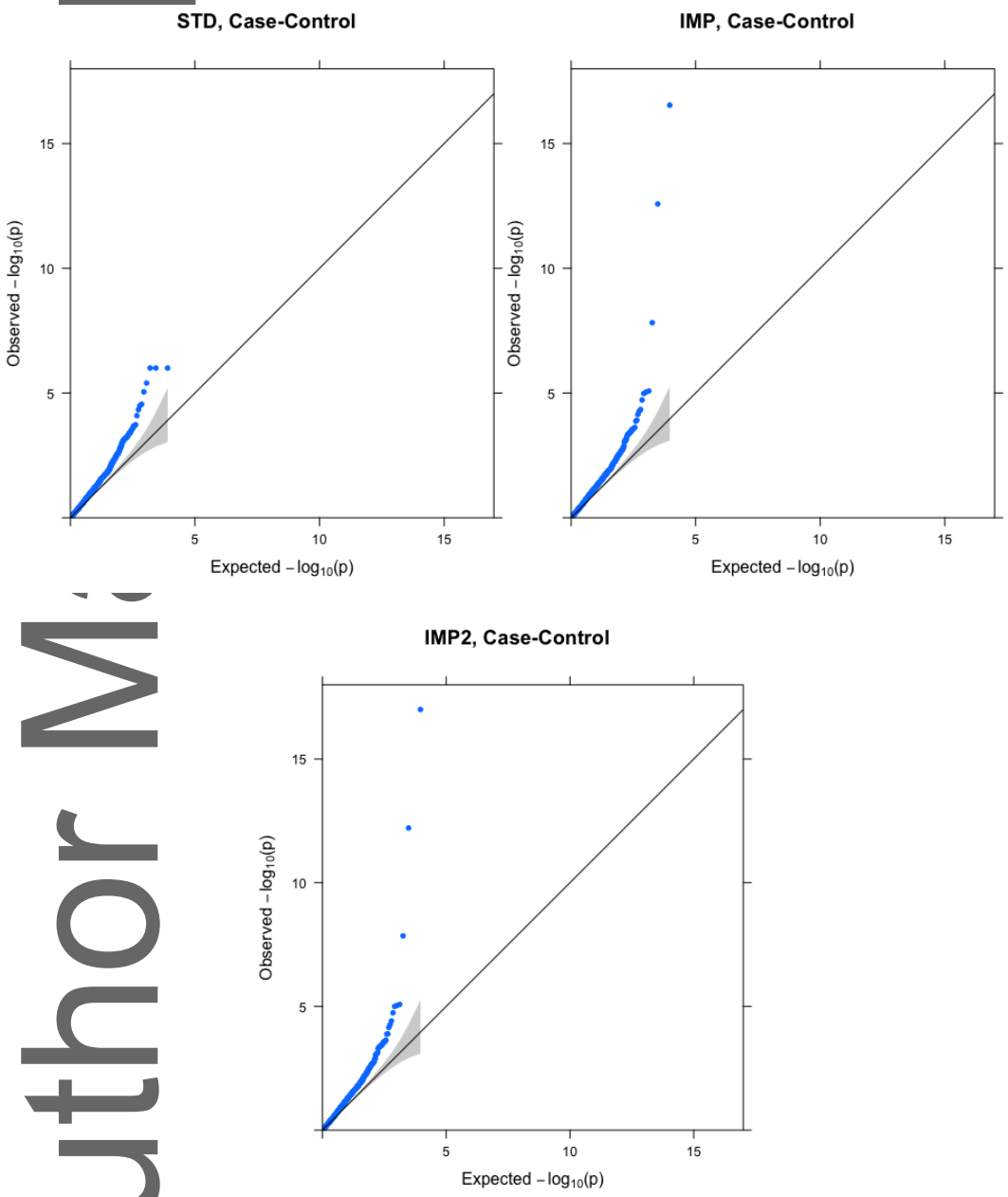


Figure 3 QQ plots of -$\log_{10}$(p-values). Case-control. Top left: STD observed vs. expected. Top right: IMP observed vs. expected. Middle: IMP2 observed vs. expected. Bottom left: STD vs. IMP. Bottom right: STD vs. IMP2.



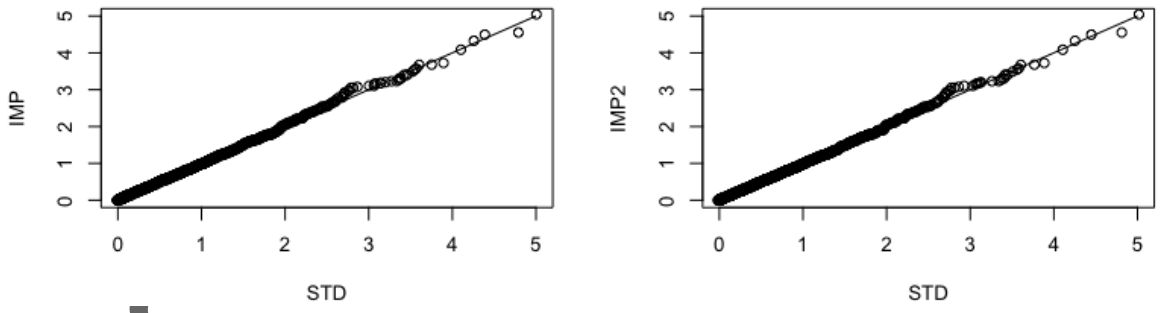


Figure 4 QQ plots of -$\log_{10}$(p-values). Control-control.