Supplement for ”Enabling Privacy Preserving GWAS in Heterogenous Human Populations”

Sean Simmons [1,2,3], Cenk Sahinalp [3,4], and
Bonnie Berger [1,2*]

[1]Department of Mathematics, [2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA
[3] School of Computing Science, Simon Fraser University, Burnaby, BC, Canada and
[4] School of Informatics and Computing, Indiana University, Bloomington, IN

[*]to whom correspondence should be addressed: bab@mit.edu

1 Proofs

Theorem 1. *The modified versions of the score and noise based methods for picking high scoring SNPs given in the manuscript are ϵ-differentially private.*

Proof. The proofs are the same as those given in previous works [4], where the score function for returning SNPs $s_1, \cdots, s_{m_{ret}}$ equals $\sum_{i=1}^{m_{ret}} |\mu_{s_i} y|$ □

Theorem 2. *Algorithm 1 returns the correct value of $d_i(c)$.*

Proof. Let U_k, L_k, l_k and u_k be as in Algorithm 1.

Assume that y and y' differ in at most k coordinates, then

$$\mu_i y - \mu_i y' = \sum_{j, y_j \neq y'_j} \mu_{ij}(y_j - y'_j) \leq -(l_1 + \cdots + l_k)$$

so

$$\mu_i y' \geq \mu_i y - \sum_{i=1}^{k} l_k = L_k$$

Similarly

$$\mu_i y' \leq \mu_i y + \sum_{i=1}^{k} u_k = U_k$$

so if $d_i(c) \leq k$ than $L_k \leq c \leq U_k$. It is easy to see, however, that if $L_k \leq c \leq U_k$ than $d_i(c) \leq k$, so $d_i(c) = k$ if and only if $c \in [L_k, L_{k-1}) \cup (U_{k-1}, U_k]$. Therefore Algorithm 1 correctly calculates $d_i(c)$.

□

2 Generation of Simulated GWAS data

In order to produce simulated data, we used PLINK [3]. The code used to generate this data is available on our website.

We generated two populations of individuals. For each set we first used plink to choose the MAF for 10000 SNPs, each uniformally at random from [.05,.5]. 9900 of the SNPs had no effect on phenotype, 100 had an odds ratio of 1.1. We then generated 5000 people from each of the populations, half

Algorithm 1 Calculates the neighbor distance

```
Require: $y, \mu_i, c$
Ensure: The neighbor distance, $d_i$.
  Let $\hat{u}_j = \max(\mu_{ij}(1 - y_j), \mu_{ij}(0 - y_j))$
  Let $\hat{l}_j = \min(\mu_{ij}(1 - y_j), \mu_{ij}(0 - y_j))$
  Let $i_1, \cdots, i_n$ be a permutation on $1, \ldots, n$ such that $\hat{u}_{i_1} \geq \cdots \geq \hat{u}_{i_n}$. Let $u_j = \hat{u}_{i_j}$ for all $j$.
  Let $j_1, \cdots, j_n$ be a permutation on $1, \ldots, n$ such that $\hat{l}_{j_1} \leq \cdots \leq \hat{l}_{j_n}$. Let $l_k = \hat{l}_{j_k}$ for all $k$.
  Let $U_k = \sum_{j=1}^{k} u_j + \mu_i y$ and $L_k = \sum_{j=1}^{k} l_j + \mu_i y$, $k = 1, \cdots, n$.
  Return $k$ such that $c \in [L_{k+1}, L_k) \cup (U_k, U_k + 1]$
```

of whom where cases, the other half controls. We then combined these two populations to produce our simulated dataset.

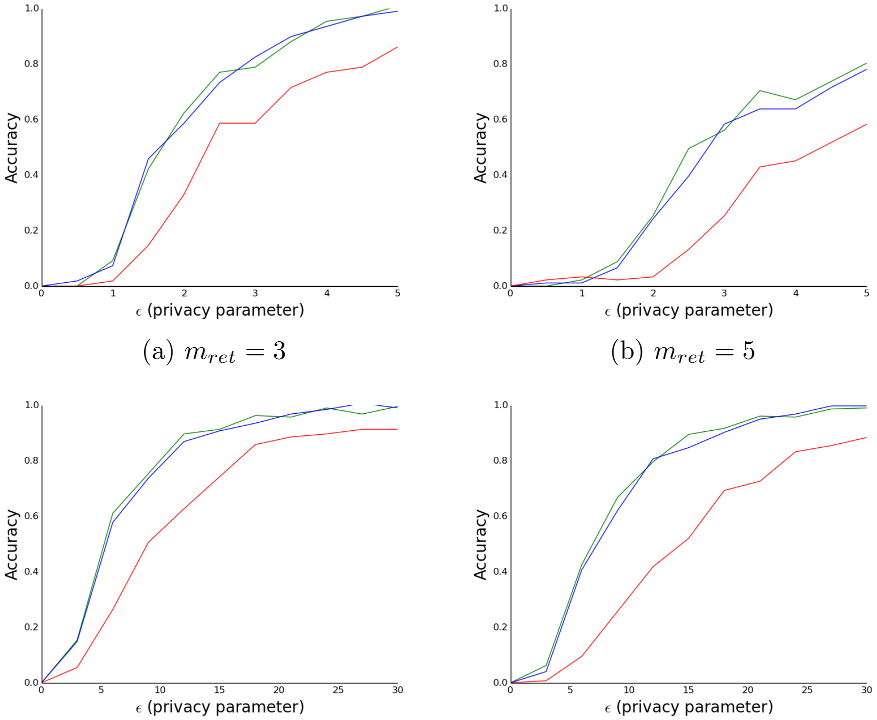
The code to do this is present online, as is the simulated data generated in this way.

3 Simulated Data PrivStrat

In the manuscript we only showed the result of using PrivStrat to pick high scoring SNPs on real GWAS data, not on our simulated data. The results, however, are similar–namely accuracy decreases as m_{ret} increases and and accuracy increases as ϵ increases. More than that, the noise and score based methods outperform the neighbor based method. We picture the results in Figure 1.

4 Testing PrivLMM

Due to space constraints we did not show the results of testing our methods using LMM based statistics. Therefore we show those results here. In particular, we look at how well our differentially private version of the LMM statistic does at picking high scoring SNPs in our real GWAS. We assume that we are given an estimate of σ_e and σ_g ahead of time (details of how to get these estimates are given in the following section)–in particular we applied GCTA to get an estimate of these quantities. The results, pictured in Figure 2, are very similar to those we had for PrivStrat–namely accuracy decreases as m_{ret} increases and and accuracy increases as ϵ increases. More



(a) $m_{ret} = 3$

(b) $m_{ret} = 5$

(c) $m_{ret} = 10$

(d) $m_{ret} = 15$

Figure 1: We measure the accuracy (the percentage of the top SNPs correctly returned) of the three methods for picking top SNPs on simulated GWAS data using score (blue), neighbor (red) and noise (green) based methods with m_{ret} (the number of SNPs being returned) equal to a. 3 b. 5 c. 10 and d. 15 for varying values of the privacy parameter ϵ. We see that in all four graphs that score and noise based methods outperform the neighbor method. These results are averaged over 20 iterations.

than that, the noise and score based methods outperform the neighbor based method, except for when $m_{ret} = 10$ in which case all three are comparable.

4.1 Estimating Heritability

A final problem to consider is the estimation of σ_e and σ_g. This, however, can be done using a sample-and-aggregate based framework [1]. In particular, the works by choosing some integer $K > 1$, and dividing the set of participants into K disjoint sets of equal size. On each of these subsets we can estimate $h^2 = \frac{\sigma_g^2}{\sigma_e^2+\sigma_g^2}$ using GCTA [5] or a similar tool. This gives us K estimates of h^2, namely $h_1^2, \ldots, h_K^2$. Let $\tilde{h}^2$ be the average of these K values. Our ϵ-differentially private estimate of h^2 is then given by $\tilde{h}^2 + Lap(0, \frac{1}{\epsilon})$.

Next we want to use the same framework to estimate σ_e^2. Note, however, that this would require a bound on σ_e^2. Note that $\sigma_e^2 \leq Var(y)$, and that we can get a ϵ-differentially private estimate v_{dp} of $Var(y)$ easily using the laplacian mechanism. Then we can easily apply the sample-and-aggregate methodology to $\max\{v_{dp}, \sigma_e^2\}$ to get an ϵ-differentially private estimate. Since $\sigma_g^2 = \sigma_e^2(\frac{1}{1-h^2} - 1)$ this allows us to get a 3ϵ-differentially private estimate of (σ_e^2, σ_g^2). Note that this method relies on a very general methodology, and so it seems likely much more accurate results can be obtained with a little work.

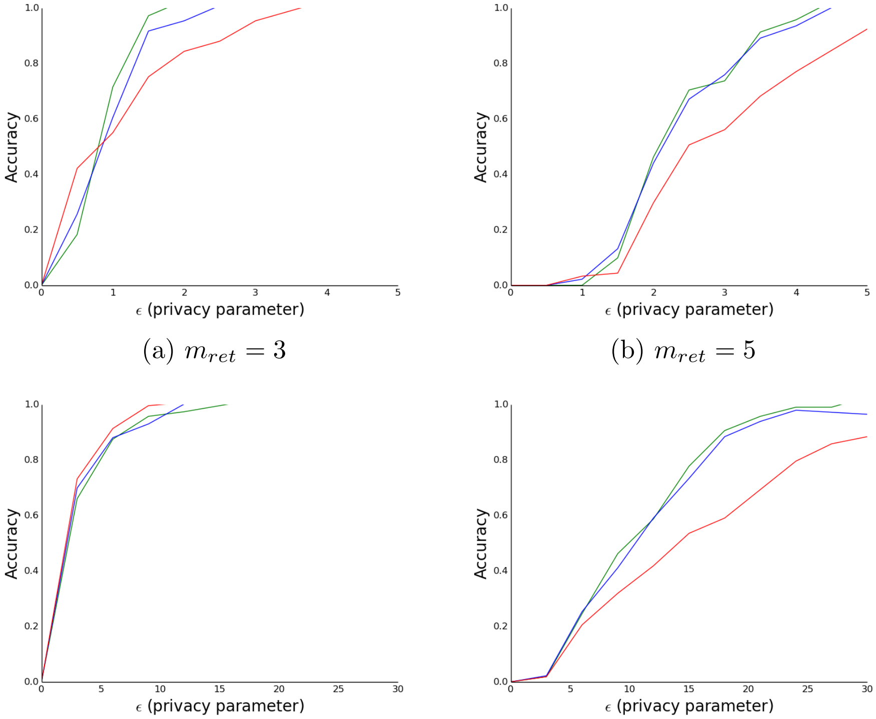
References

[1] J Abowd, M Schneider, and L Vilhuber. Differential privacy applications to bayesian and linear mixed model estimation. *Journal of Privacy and Con*

dentiality, 5(1):73–105, 2013.

[2] F McSherry and K Talwar. Mechanism design via differential privacy. Proceedings of the 48th Annual Symposium of Foundations of Computer Science, 2007.

[3] S Purcell, B Neale, K Todd-Brown, L Thomas, M Ferreira, D Bender, J Maller, P Sklar, P de Bakker, M Daly, and P Sham. Plink: a toolset for whole-genome association and population-based linkage analysis. *AJHG*, 81:559–575, 2007.



(a) $m_{ret} = 3$

(b) $m_{ret} = 5$

(c) $m_{ret} = 10$

(d) $m_{ret} = 15$

Figure 2: We measure the accuracy (the percentage of the top SNPs correctly returned) of the three methods for picking top SNPs using PrivLMM based score (blue), neighbor (red) and noise (green) based methods with m_{ret} (the number of SNPs being returned) equal to a. 3 b. 5 c. 10 and d. 15 for varying values of the privacy parameter ϵ. We see that in three of four graphs that the score and noise based method outperform the neighbor method, while in the final case all perform similarly. These results are averaged over 20 iterations.

[4] C Uhler, S Fienberg, and A Slavkovic. Privacy-preserving data sharing for genome-wide association studies. *Journal of Privacy and Confidentiality*, 5(1):137–166, 2013.

[5] J Yang, S Lee, and M Goddard a P Visscher. Gcta: a tool for genome-wide complex trait analysis. *AJHG*, 88:76–82, 2011.