Appendices

Appendix A. Synthetic data examples to illustrate the methods when individual genotype data is known

Appendix A.1. Synthetic data generation

Here we describe the design and the data generation process for the synthetic data examples used for illustrating the methods in section 4.1 and Appendix A.2 below. SNP data was generated using R package *scrime*[67], which is a package that contains tools for the analysis of high-dimensional data, especially with a focus on SNP data. SNPs generated by this package were independent of each other such that they were unlinked but some minimal correlation existed among SNPs post-generation. In all the synthetic data, we had two GWASs with one phenotype vector and M SNPs in each, and m SNPs of shared effect (SSEs) across both GWASs. Each GWAS had N subjects. In order to get these two simulated synthetic GWASs, we first generated a large SNP matrix with N rows and $(2m+4M)$ columns, and a corresponding binary phenotype vector y, using the *simulateSNPglm*[67,68] function in *scrime*. The minor allele frequencies for the simulated SNPs ranged from 1.01% to 49.9%. The first $2m$ columns in this matrix were set apart to select m SSEs. Out of the remaining $4M$ columns, the first $2M$ columns were set apart to select M SNP columns for the first phenotype and the last $2M$ columns for selecting M SNP columns for the second phenotype. The binary y vector generated by *simulateSNPglm* was used to generate two quantitative phenotypes. For the first quantitative phenotype, $y = 1$ values were replaced by values from a $N(25, 1)$ distribution, and $y = 0$ values were replaced by values from a $N(5, 1)$ distribution. So, if one were to imagine that the first simulated phenotype represents psychotic symptom scores on a scale ranging from 0 to 30, then the 'symptom scores' for healthy controls will have a mean of 5 and that for cases (e.g. patients with schizophrenia) will have a mean of 25. For the second phenotype, $y = 1$ and $y = 0$ values were replaced by values from $N(15, 1)$ and $N(10, 1)$ distributions, respectively. So, the second phenotype could be imagined as scores from a depression scale with mean 15 for cases and mean 10 for controls. The two quantitative phenotypes were generated from the same binary vector in the above fashion in order to ensure that the two sets of polygenic scores obtained at a later step (-one set for each phenotype-) are correlated (genetic correlation). For synthetic data generation in Appendix A.2, we varied the mean-pair for the second phenotype from $(15, 10)$ to $(14, 11)$ or $(13, 12)$ or $(12.6, 12.4)$ to get different values of genetic correlation, r_g, between the two phenotypes.

Let us call the first and second quantitative phenotypes generated above as y_1 and y_2, respectively. In order to select the M SNPs (for the simulated GWAS) for the first quantitative phenotype, we ran $2M$ univariate GLMs with y_1 regressed against each of $2M$ SNPs set apart from the original $N \times (2m + 4M)$ matrix. Typically polygenic models consist of only small to moderate effect sizes. In order to ensure that our simulated GWASs also had only small to moderate effect sizes, M SNPs from $2M$ SNPs were selected based on their t-values in the following way.

SNP-selection-method: M t-values were randomly selected from among the t-values between the 5^{th} percentile and 95^{th} percentile of all the $2M$ t-values obtained, and the corresponding M SNPs were chosen to be included in the first GWAS. The 5^{th} percentile and 95^{th} percentile

were typically around -1.5 and 1.5, respectively, and the 5^{th} percentile and 95^{th} percentile p-values typically around 0.035 and 0.96 respectively. Thus most of the M SNPs generated in this scenario had small effect sizes, and some of them with moderate effect sizes, but none with large effect sizes or p-values meeting the 5×10^{-8} threshold for GWAS significance. Similar procedure was applied to select M small-and-moderate-effect-sized-SNPs for the second phenotype, y_2: the M SNPs with t-values between the 5^{th} and 95^{th} percentiles were randomly selected from among all $2M$ t-values obtained from univariate GLMs of y_2 against each of the last $2M$ columns from the original $N \times (2m + 4M)$ SNP matrix. As in the case of the first phenotype, the t-values for the selected SNPs for the second GWAS also ranged from -1.5 to 1.5 and p-values from 0.035 to 0.96, approximately.

For all analyses using synthetic data, m SSEs were generated in two different ways. In both cases, a quantitative phenotype similar to y_1 was generated: $N(25, 1)$ distribution values replacing $y = 1$ values and $N(5, 1)$ values replacing $y = 0$ values. This simulated phenotype was regressed against $2m$ SNPs set apart from from the original $N \times (2m + 4M)$ matrix.

shared-effect-SNP-selection-method-1: In the first case, m t-values between the 5^{th} percentile and 95^{th} percentile of all t-values from the $2m$ GLMs were randomly selected, and the corresponding m SNPs were selected as the SSEs. t-values of the selected SSEs ranged between -1.5 and 1.5. Thus in this case, the selected SSEs had mostly small effect sizes, but some had moderate effect sizes.

shared-effect-SNP-selection-method-2: For the second case 5 of the m SSEs were replaced by SNPs with moderate effect sizes (p-values between 0.015 and 0.05 approximately, $-\log_{10}$(p-values) between 1.25 and 1.75) selected from the first quartile of t-values from all the t-values obtained from the $2m$ univariate GLMs.

shared-effect-SNP-selection-method-3: In the first case, m t-values between the 30^{th} percentile and 70^{th} percentile of all t-values from the $3m$ GLMs were randomly selected, and the corresponding m SNPs were selected as the SSEs. Thus in this case, the selected SSEs had only small effect sizes.

m indices were randomly selected from indices 1 to M (and fixed for later comparison) and the SNPs at these m locations, for both phenotypes, were replaced by SSEs.

Appendix A.2. Results from larger synthetic data examples

The synthetic data example used in section 4.1 to illustrate our approach, had small sample size ($N = 3000$) and small number of SNPs ($M = 1000$). In order to check whether the method and results were valid for larger N and M, we analyzed synthetic data with $M = 20000$ and $N = 60000$. The results from this analysis are given in the supplementary file available at github-link, https://github.com/mjohn5/sharpen/tree/main/appendix_A2

Appendix B: Further illustration using synthetic data

B.1. Comparison between Lasso and Elastic Nets (with $\alpha \in (0, 1)$)

We illustrate the theoretical concepts mentioned in the supplementary subsection $S.2$ with synthetic data. Two GWASs with 1000 SNPs and 3000 subjects each were generated using the method mentioned in appendix A.1, with mean-pair for the second phenotype selected

as (15,10). 100 SNPs selected using the *shared-effect-SNP-selection-method-1* were imputed at the same locations for both GWASs; these 100 SNPs were the shared SNPs. The locations chosen for this specific synthetic data example were

$$91-100, 191-200, 291-300, 391-400, 491-500, 591-600,$$

$$691-700, 791-800, 891-900 \text{ and } 991-1000.$$

SNPs at first 20 of these locations (i.e. 91-100 and 191-200) were made highly correlated among each other simply by sorting them, independently from each other. This technique was the same as the one used in Waldmann *et. al.*[54] for the simulated examples presented in that paper. In our case, this procedure resulted in an average correlation coefficient of 0.98 between each pair of the first 20 of the shared SNPs. Plots similar to those in Figure 2 were created for the current example (plot available upon request). Lasso based approach with PGS-adapted λ gave the following results for the shared locations:

$$91, 95, 99, 175, 195, 197, 199, 200, 476, 985.$$

Three out of the 10 were incorrectly selected, but more pertinent to our current discussion, among the first 20 locations with highly correlated shared SNPs, only 7 (91, 95, 99, 195, 197, 199, 200) were selected. QRR based approach with the PGS-adapted λ selected the following locations:

$$91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 398, 476, 985.$$

In this case, all of the first 20 locations were selected. The above results are consistent with the fact that QRR is a grouping effect method while Lasso is not. When QRR selects one SNP from a group of highly correlated SNPs, it selects the entire group, while Lasso does not have this property. For an intermediate α, the grouping effect will be in between that of Lasso and QRR. For the record, among the 571 locations selected by the Lasso based approach at a much smaller λ, those among first 20 locations (i.e. the highly correlated ones) were

$$91, 92, 94, 95, 98, 99, 100, 192, 193, 195, 197, 198, 199, 200.$$

Even in this case, we see that Lasso does not select all 20 of the highly correlated shared SNPs, but all of them are included in 686 SNPs selected by QRR. Thus the choice between QRR approach and the Lasso approach is essentially based on whether a grouping effect is desirable or not. SNPs sharing the same biological pathway could be thought of as from the same group. Lasso will select only one representative SNP from such a group while as elastic nets with $\alpha \in (0, 1)$ will select multiple SNPs from the group or may be even select the entire group).

B.2. Comparison between PGS-adapted λ and cross-validation based λ

In this subsection we illustrate the difference in the results obtained using λ-thresholds based on the traditional cross-validation approach and our PGS-adapted approach. Synthetic data for two GWASs with 1000 SNPs and 300 subjects in each, and 100 shared SNPs were generated using the method mentioned in appendix A.1, with mean-pair for the second

phenotype selected as (15,10). The 100 shared SNPs were selected using the *shared-effect-SNP-selection-method-1* as in the previous subsections.

cv.glmnet function in the *glmnet* R package was used for selecting a λ-threshold, λ^*_{cv}, within each GWAS based on 10-fold cross validation (CV) method. Specifically, the λ-value corresponding to minimum value of the mean cross-validated error was selected. Intersecting the two subsets of SNPs obtained (one from each from each synthetic GWAS at the corresponding λ^*_{cv}), yielded 6 SNPs for Lasso and 933 SNPs for QRR. Out of the 6 'shared' SNPs obtained with Lasso based on λ^*_{cv}, none belonged to the true underlying 100 shared SNPs. Thus, the sensitivity and true positive rate using CV-based λ-threshold for Lasso are both zero. The sensitivity in the CV-based QRR case is 95.5% and the true positive rate 10.2%. However, these values (obtained in the QRR case) are quite meaningless because the shared subset size of 933 is almost as large as the total number of SNPs, 1000. Hence, such a large sensitivity value is not surprising. We may as well have used all the 1000 SNPs which would correspond to 100% sensitivity. The true positive rate of 10.2% obtained by CV-based QRR is also very close to the original true positive rate of 10% based on all 1000 SNPs.

Based on correlation plots (see Figure B.1 below) similar to the right-panel in Figure 1 and left-most panel in Figure 2, a λ-threshold corresponding to 60 shared SNPs was selected for Lasso and a λ-threshold corresponding to 64 shared SNPs was selected for QRR applying the PGS-adapted criterion proposed in this paper. Note that shared subset sizes of 60 and 64 are reasonable based on feasibility considerations as well. Sensitivity in the Lasso case based on PGS-adapted λ is 13%, implying that 13 out of 60 SNPs selected belonged to the 100 true shared SNPs. Thus the true positive rate with the 60 SNPs selected in this case is 21.7%, more than double the true positive rate of 10% that could be obtained with all 1000 SNPs. The sensitivity in the QRR case is 11% and true positive rate equals to 17.2%. In this particular synthetic example, although the true positive rate for PGS-adapted λ based QRR is not as high as that of the corresponding Lasso approach, it is still substantially higher compared to 10% true positive rate that could be accomplished using all 1000 SNPs. This substantial increase in true positive rate based on a small and feasible subset of selected SNPs is the main justification for our approach. Further validation of the improvement over the CV-based approach is done using more extensive simulations in section 5.5.

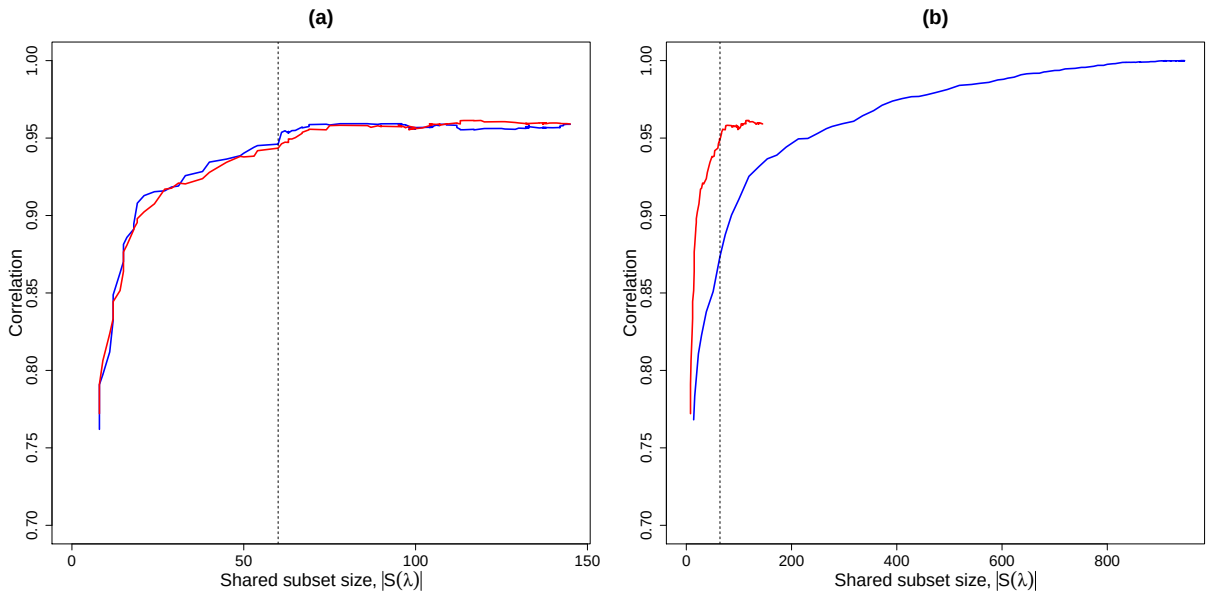

Figure B.1: Correlation between PGS_{full} and $\text{PGS}_{\text{subset}}$ scores plotted against shared subset size. The left panel (a) corresponds to the Lasso case and the right panel (b) corresponds to QRR. A vertical line is plotted at 60 in the left panel and at 64 in the right panel.

Appendix C. Correlated data generation for simulations in section 5

Synthetic data examples used for illustration in section 4 and also in appendix A.2 were generated using R package *scrime*. One issue with data generated via *scrime* is that there is no effective way to incorporate between-SNP correlations. On the average the between-SNP correlation for data generated via *scrime* ranged between 0.10 and 0.15, which would be similar to residual correlations after LD pruning. In order to get a better handle on incorporating between-SNP correlations, we used the R package *SNPSetSimulations* downloaded from https://github.com/fhebert/SNPSetSimulations. *SNPSetSimulations* provides an effective way of generating a matrix of genotypes (i.e. the SNP matrix $\mathbf{X}$) according to given marginal distributions and within-block (e.g. within-gene) dependence structure, mainly by using functions from R package *GenOrd*[69]. Once the SNP matrix was generated, a 'disease status' binary corresponding to these genotypes can be generated using a specified model. The 'disease status' binary variable can then be converted to a continuous variable using the strategies outlined in the first paragraph of appendix A.1.

For most of the simulation scenarios in section 5, SNP matrices at all iterations had 1000 columns (i.e. SNPs). The 1000×1000 correlation matrix used in the simulations had a block-diagonal structure, where block-sizes of 10 and 25 were considered for the simulations in subsection 5.1. (For example, block-sizes of 10 was used for simulations presented in all top panels in Figure 3, and block-sizes of 25 were used for simulations done for the bottom panels.) In all the remaining subsections in section 5, only block-sizes of 10 were utilized.

Within each block, correlations were generated from a $Uniform(r_L, r_U)$ distribution where different values of the range (r_L, r_U) considered were $(0.01, 0.04)$, $(0.05, 0.15)$, $(0.20, 0.40)$, and $(0.40, 0.60)$ (e.g. corresponding to left to right panels within each row in Figure 3 - thus,

in Figure 3, overall correlations on the average increases as we move down and to the right). All the four ranges mentioned above were used only in subsection 5.1. In all other subsections in section 5, only the following two ranges for correlation values were considered: $(0.05, 0.15)$ and $(0.40, 0.60)$. Minor allele frequencies which were used for the marginal distribution of each SNP were generated from a $Uniform(0.2, 0.4)$ distribution. For generating the binary phenotype within each GWAS, 25 SNPs were selected randomly to have non-zero β's. The non-zero β's for phenotype generation were obtained from a $Uniform(0, 0.25)$ distribution (i.e. modest effect sizes) and intercept term was set to -5.25. The mean pairs used for converting the binary phenotype to a continuous phenotype were (25, 5) for the first GWAS and (15, 10) for the second GWAS, similar to the approach mentioned in Appendix A.1 for synthetic datasets. Once correlated SNP data were generated for two GWASs, M_c number of true underlying shared SNPs were inserted using the *shared-effect-SNP-selection-method-1* mentioned in Appendix A.1, at random but fixed locations in both GWASs. $M_c = 100$ was used for simulations with 1000 SNPs and $M_c = 500$ was used for simulations with 10,000 SNPs. All codes used for simulations and synthetic data generation are available upon request.

Appendix D. Univariate t-values plot for the'Cam13W' site in the analysis of the first dataset

Note that the p-value threshold for the RO method which yielded a subset of size 153 was 0.0095 and the corresponding t-value at 245 degrees of freedom is approximately ± 2.615 (marked by vertical and horizontal black lines in Figure D.1)

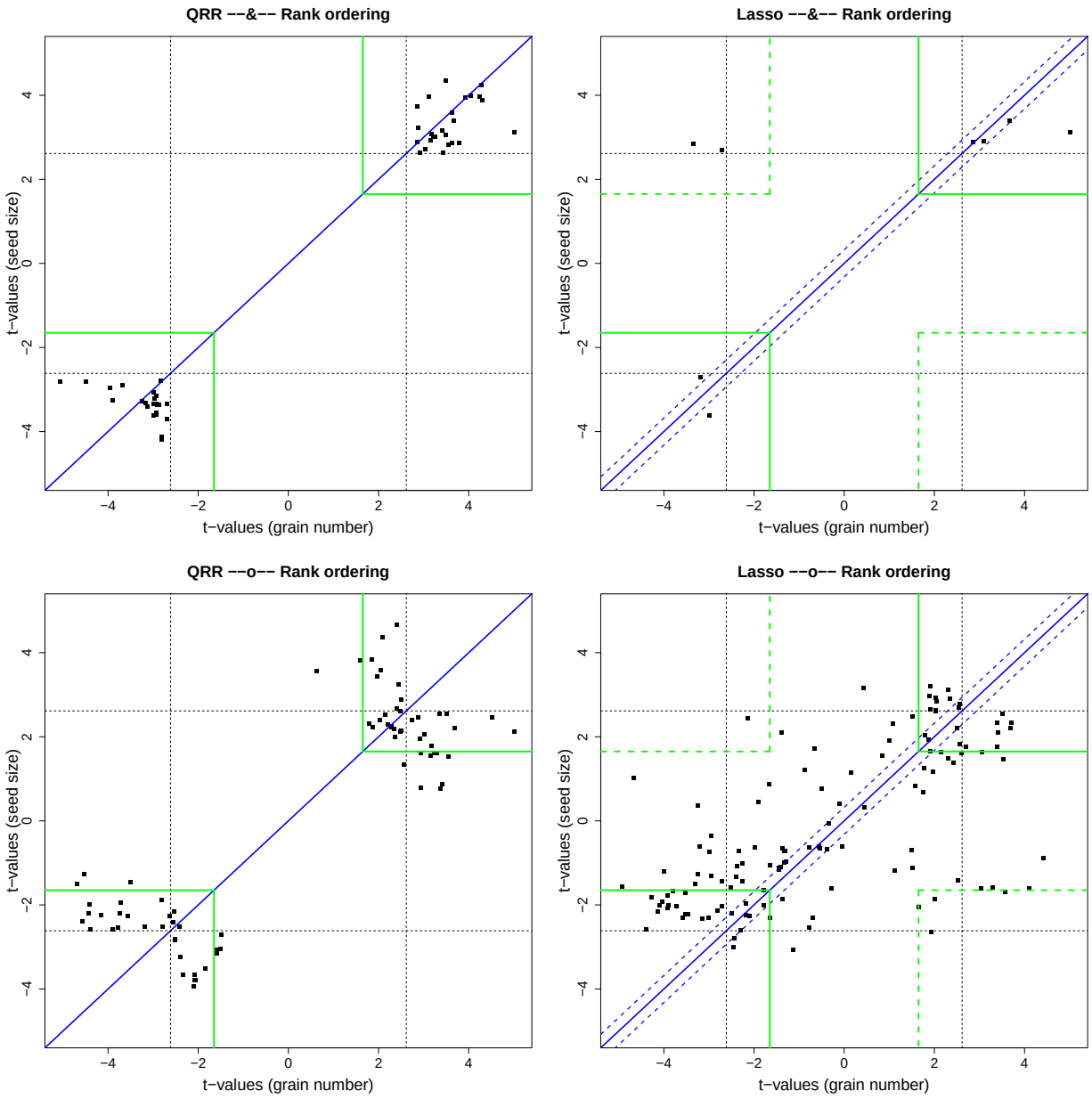

Figure D.1: Univariate t-values of seed size plotted against those for grain number, in all panels. Black and green lines correspond to t-values of ± 2.615 and ± 1.65, respectively. Top left panel: SNPs detected by QRR and RO; top right panel: SNPs detected by Lasso and RO; bottom left panel: SNPs detected by QRR but not by RO; bottom right panel: SNPs detected by Lasso but not by RO. The blue dashed band marks an approximate 95% CI for the solid blue line calculated based on the standard deviation of the differences of the t-values. This figure is based on data from the site 'Cam13W'