Appendix A-2 continued

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 below and plotted in the figures A.1 to A.6.

First, we generated a synthetic data set with everything kept the same as in our illustrative example in section 4.1, except that the M, N and m were larger - 20K, 60K and 2K, respectively. Figure A.1 gives the results for this example. The correlation between the polygenic scores for the two simulated GWASs based on the entire SNP matrix (i.e. our estimate of r_g plotted as the green horizontal line in the third panel) was 0.219, comparable to the one obtained for the smaller example presented in section 4; the results seen in Figure A.1 are very similar to that seen for that example. Next we considered examples with smaller r_g values (0.119 in Figure A.2 and 0.0247 in Figure A.3). These smaller r_g values were obtained by varying the mean-pairs for the simulated second phenotype. The results seen in Figures A.2 and A.3 are very similar to that seen in A1.

Next we considered a scenario where r_g was large (close to 0.6, for example). With $M = 20K$ and $N = 60K$, it was impossible to generate such a large r_g with $m = 2K$, by varying the mean-pair for the second phenotype alone. Thus, in order to get such a larger r_g we increased m from $2K$ to $10K$. r_g for the simulation scenario for figure A.4 was 0.58. The results in this case also are very similar to those seen in the previous figures, except in the third panel from left. In this panel, r_g estimated on the SNP-subsets gets close to the original r_g only towards the right end of the λ grid. So, for this case, it would be preferable to pick a λ towards the right end of the grid.

In order to assess the performance of the elastic nets, when the number of SSEs was much lower, we considered the case with $m = 400$. The results for this case are plotted in Figure A.5, and are very similar to the results from previous scenarios. In the five simulation scenarios that we considered so far, the SSEs were selected using *shared-effect-SNP-selection-method-1* mentioned above. That is, all the common SNPs had very small effects. In the last scenario, we considered the case where a few of the common SNPs (- 5 to be exact -) had moderate effects. For this scenario, *shared-effect-SNP-selection-method-2* mentioned above, was used. The results for this final scenario, plotted in Figure A.6, are essentially the same as for the previous scenarios. Thus, based on all the synthetic data examples presented in this section, we may conclude that the strategies presented in section 4 hold for data sets with larger sample sizes as well.

The above M and N (20K and 60K, respectively) were the largest that we could run on a cluster machine with 252GB of memory. We were constrained by several factors specific to the simulation procedure: 1) the fact that we needed N to be 3 times M, 2) for simulating and $N \times M$ SNP matrix, the design of our simulations required simulating a $N \times (2m + 4M)$ matrix first. Note that this is not an issue for real data. With the same amount of available memory and other computational hardware capacity, we could easily consider GWASs with $M = 40K$ and $N = 120K$ or larger.

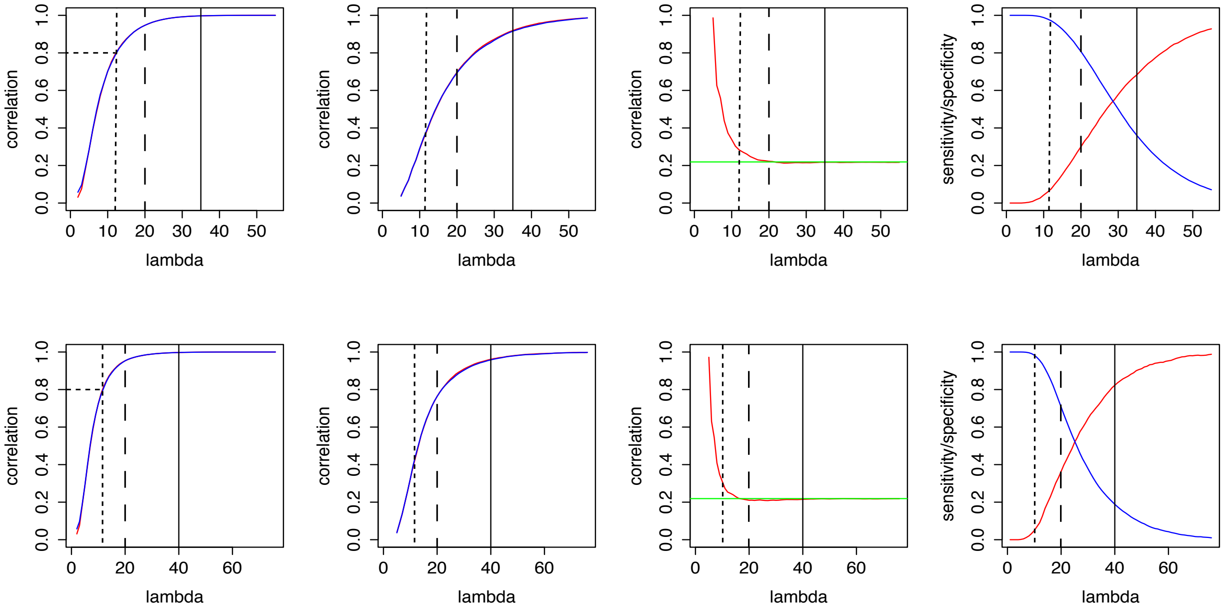

Figure A.1: M = 20000, N = 60000, m = 2000 (all common SNPs had small-effect sizes), r_g = 0.219

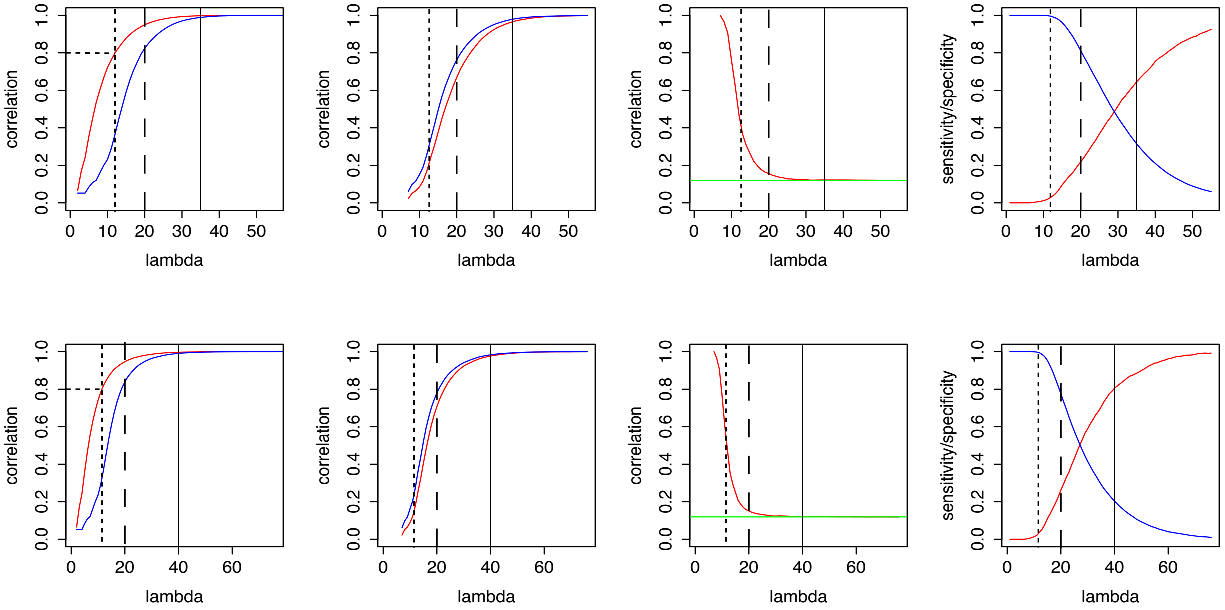

Figure A.2: M = 20000, N = 60000, m = 2000 (all common SNPs had small-effect sizes), r_g = 0.119

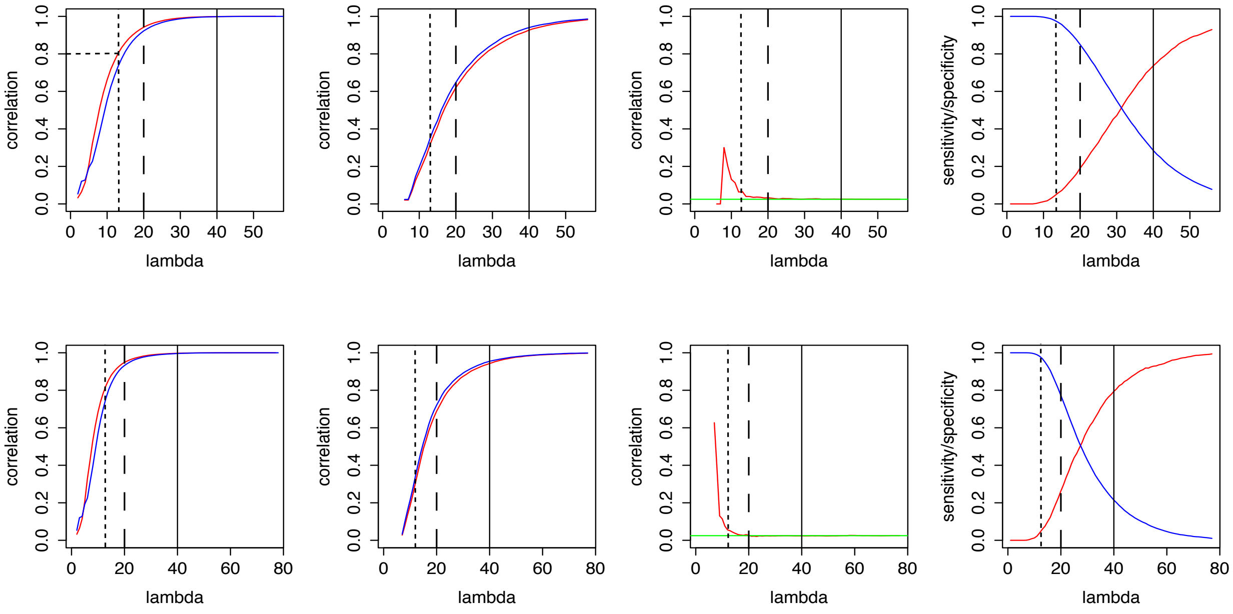

Figure A.3: M = 20000, N = 60000, m = 2000 (all common SNPs had small-effect sizes), r_g = 0.0247

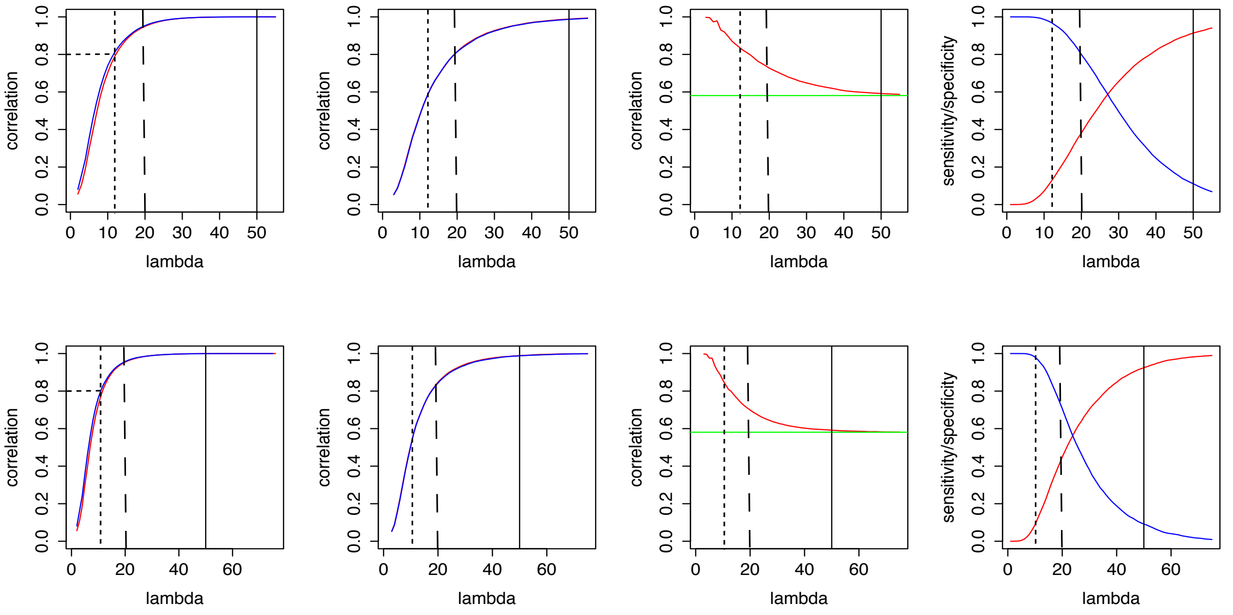

Figure A.4: M = 20000, N = 60000, m = 10000 (all common SNPs had small-effect sizes), r_g = 0.58

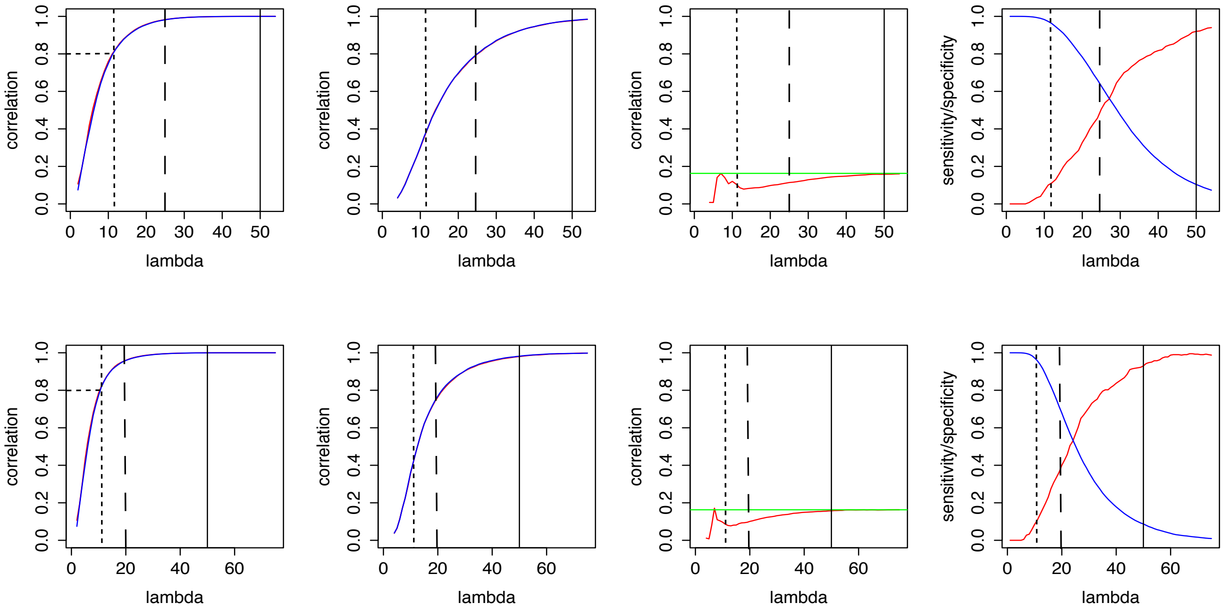

Figure A.5: M = 20000, N = 60000, m = 400 (all common SNPs had small-effect sizes), $r_g = 0.163$

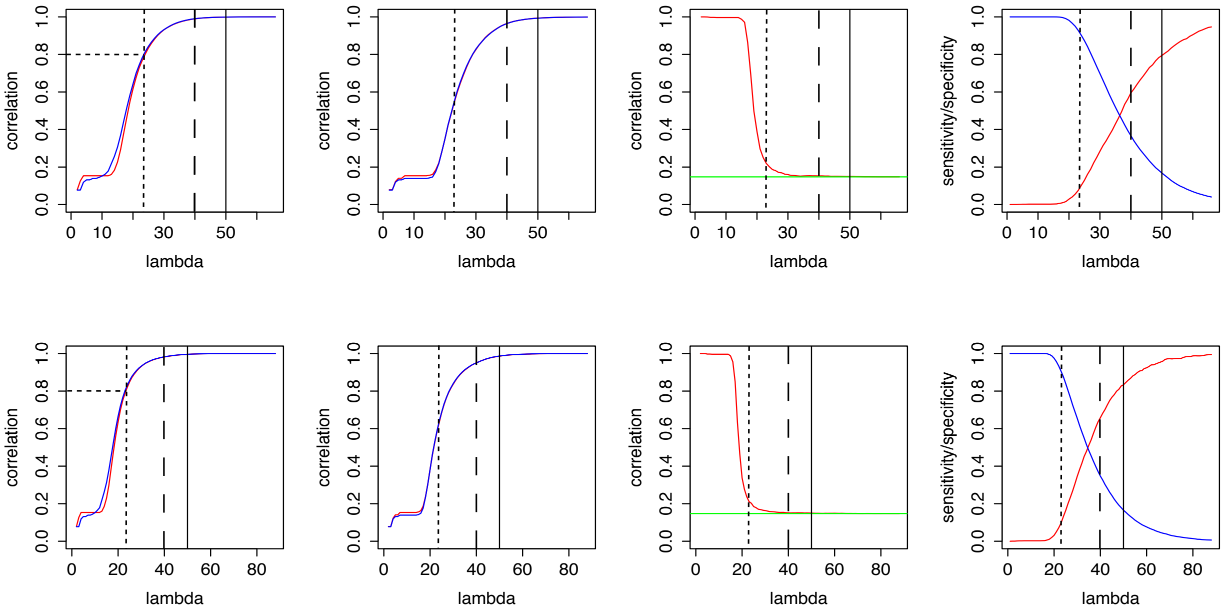

Figure A.6: M = 20000, N = 60000, m = 2000 (1995 SNPs had small-effect sizes, 5 SNPs had large effects), $r_g = 0.148$