Web-based Supplementary Materials for:

Identifying disease-associated copy number variations by a doubly penalized regression model

Yichen Cheng, James Y. Dai, Xiaoyu Wang, and Charles Kooperberg

Summary: This appendix contains additional details on the simulation study and additional R codes to repeat the results of the simulation studies. For updated versions of the R code or any questions, please email the first author at ycheng11@gsu.edu.

Web Appendix A: Empirical evidence on the choice of α

To study the effect of different α values on the results, we choose α to take a sequence of equally spaced value between 0 and 1 with increment set to 0.05. We simulate data sets with n=500 individuals, of whom 250 are cases ($y = 1$) and 250 are controls ($y = 0$). We generate the LRs for $m = 5000$ markers and randomly select 5 regions with lengths (S) set to 20. The LRs are generated as in Section Results to mimic the correlation between nearby locations.

For each of the 5 regions for each individual, we randomly decide whether to add a CNV to that region. The probability of adding a CNV to the preselected region is set to be $p_0 = 0.2$ for the controls and p_1 for the cases. The average LR for the CNV is μ. So for an individual without a disease, the expected number of CNVs is $5 \times 0.2 = 1$. We consider three scenarios using different combinations of p_1 and μ. Case 1: the individual signal is weak but the proportion of signals is high ($\mu = 0.5, p_1 = 0.5$). Case 2: the individual signal is strong but the proportion of signals is low ($\mu = 1.25, p_1 = 0.3$). Case 3: in between cases 1 and 2 ($\mu = 0.75, p_1 = 0.4$).

We simulate 10 data sets and apply a series of α values for each data set. We compare the MSE, the number of correctly identified CNVs, the number of segments falsely identified as CNVs as well as the recovery rate. For any CNV identified, if more than 50% of the region overlaps with a known CNV, then we say it is a correct identification. If it does not overlap with a known CNV or if the overlap is less than 50% of the length of the region, then we say it is a false discovery. The recovery rate is calculated as the percentage of the true CNV regions identified, with the percentage calculated by the length.

The simulation results are given in Figure 1 with black, red, and blue lines corresponding to Case 1 to Case 3, respectively. The upper left graph shows the percentage of variance explained for different α's. We see an overall decreasing trend for the variance explained, which suggest that the model captures variability in the model better when we have a smaller α value. Recall that α is the weight in front of the penalty term on $\boldsymbol{b}$ and $1 - \alpha$ is the weight in front of the penalty term on the association between the CNV and the phenotype. So giving a higher weight for the association will provide better results in terms of variance explained. Similarly, we observe that when α is between 0.2 and 0.5, the algorithm provides a higher true discovery rate

(number of true CNVs identified) and a lower false discovery rate (number of falsely identified segments). Since the output of our method is regions, those regions might not overlap exactly with the region where true CNVs happen. Instead, we calculate the percentage of true CNVs regions identified; the results are plotted in the lower right graph. We again note that when α is between 0.2 and 0.5, the performance is better. It is worth mentioning that although we see a better performance for $\alpha \in [0.2, 0.5]$, the difference in terms of performance is not large, which indicates that the algorithm is fairly robust to the choice of α. Because of the observations made here, we fix α to be 0.4 for all other analyses in this paper.

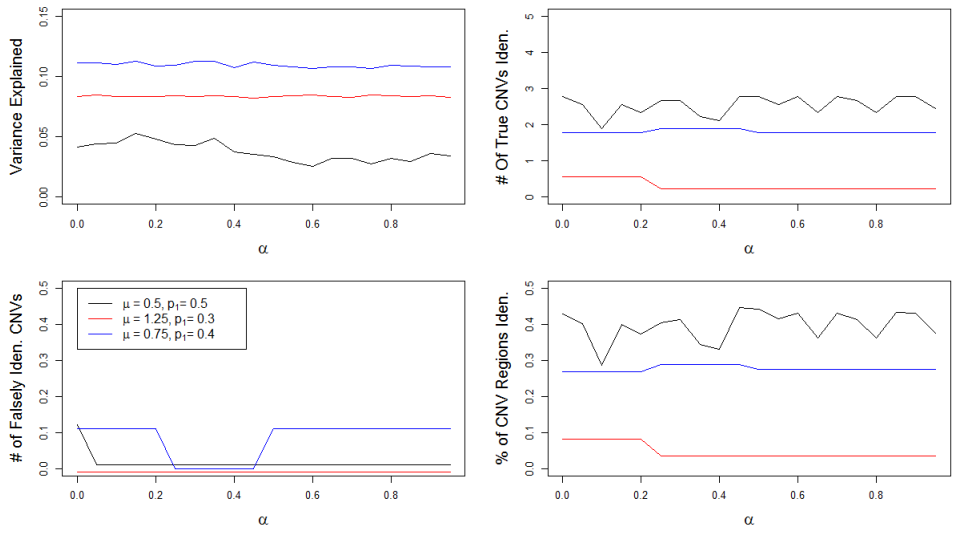

Figure 1: The averaged values for variance explained, number of falsely identified segments, number of the true CNVs identified and the percentage of CNV regressions identified. The black lines are for $\mu = 0.5, p_1 = 0.5$. The red lines are for $\mu = 1.25, p_1 = 0.3$. The blue lines are for $\mu = 0.75, p_1 = 0.4$. The results show that our method is quite robust to different choices of α. And the overall performance is the best when α is within the range of 0.2 to 0.5.

Web Appendix B:

The doubly penalized regression method described in the main text is implemented in R. The R codes needed to reproduce the simulation results are included in the Web-based Supplementary Material. In the zip file that contain the R codes, the main R code is named "calc_eta_mpi.R". This R file mainly consists of two parts. The first part uses the cross-validation to identify the penalty parameter. The second part uses the penalty parameter obtained from the first part and calculate the segmentation results. A read me file is also available for descriptions of the R files and furthur instructions. For updated versions of the R code or any other questions, please email the first author at ycheng11@gsu.edu.