A New Semiparametric Approach to Finite Mixture of Regressions using Penalized Regression via Fusion

Erin Austin[1], Wei Pan[2], Xiaotong Shen[3]

[1] *Department of Mathematical and Statistical Sciences, University of Colorado Denver, 80204*

[2] *Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455*

[3] *School of Statistics, University of Minnesota, Minneapolis, MN 55455*

Supplementary Material

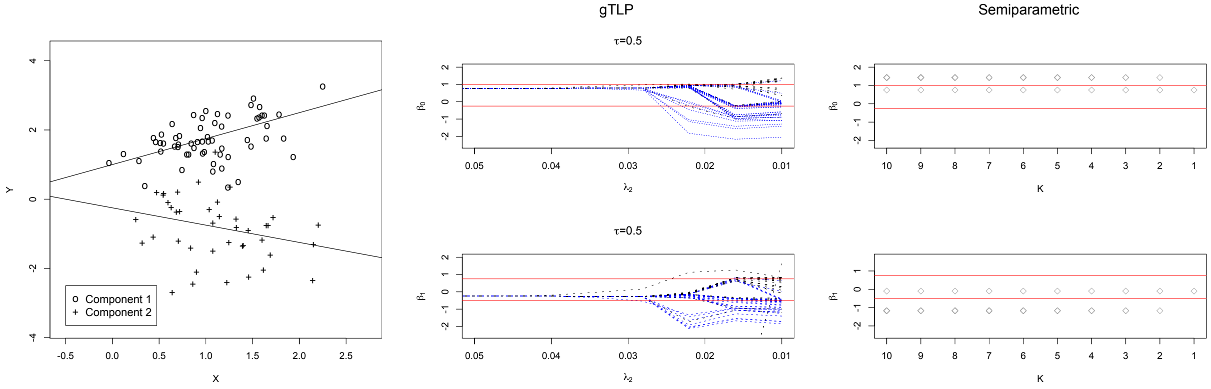
The following is the full detail for the second simulation example referenced in Section 3.2.

S1 Additional Simulation Results

Our second simulation considers the case of partially overlapping responses from both components. Here, we again let X_i represent a continuous covariate, generating it from a normal distribution with mean 1 and standard deviation 0.5. Set $\beta_{01} = 1$ and $\beta_{11} = \frac{3}{4}$ for component one, and $\beta_{02} = \frac{-1}{4}$ and $\beta_{12} = \frac{-1}{2}$ for component two. Next, we let $\epsilon_{i1} \sim \ln \mathcal{N}(0, 0.3)$ and $\epsilon_{i2} \sim \mathrm{t}_{10}$. Lastly, we reduce our sample size in half for explorative purposes

($n = 100$). Figure 1(a) shows the resulting scatterplot.

Figure1(a) reveals how our second simulation extends the first scenario in three important ways: (1) we removed the clear distinction in responses between components by overlapping the left tail of component one with the right tail of component two, (2) we added weight to the tails of our previous symmetric normal distribution, and (3) we added a skewed distribution for one of the component's errors. Figures 1(b) and (c) present the comparison of the gTLP to the semiparametric approach. Note that results from the other penalized methods are not presented because the gTLP performed the best among those methods (and is the focus of this article).



(a) (b) (c)

Figure 1: (a) Y_i and X_i scatterplot with true regression lines, and β_0 (row 1) and β_1 (row 2) estimates using (b) gTLP and (c) SP

The results of the semiparametric approach in Figure 1(c) show it performed similarly to simulation one. Like before, it appears to reveal two groups, but it is not able to find the correct parameter values. Moreover, the evidence of two groups comes largely from looking at the results of multiple prespecified component numbers, an approach counterintuitive to application. Figure 1(b) provides the performance of the gTLP. There is a tradeoff of sightly less confidence in component number, but with significant gains in estimating the model parameters. In particular, in examining the β_1 paths, there are possibly three groups as λ nears 0.02, but then they trend towards two groups: the darker lines reflecting a left-skewed distribution with a majority of values close to the true parameter as λ nears 0.01, and the lighter lines reflecting a more symmetric coefficient distribution for those in component two, but one with noticeably more spread. In both groups, the gTLP are, at worst, strongly trending towards the true model coefficients. To be specific, if you assign samples for $\lambda = 0.01$ by absolute distance to the closest true component slope, approximately 85% of the lognormal and 98% of the t samples are correct. Overall, the gTLP was able to contribute both some sense of component count and estimates of their respective FMR coefficients in this setting, where responses were not distinguishable by component (Figure1(a)). The SP method was not able to provide this same information.