Supplementary Materials
AdaReg: Data Adaptive Robust Estimation in Linear Regression with Application in GTEx Gene Expressions

by Meng Wang, Lihua Jiang, and Michael P. Snyder

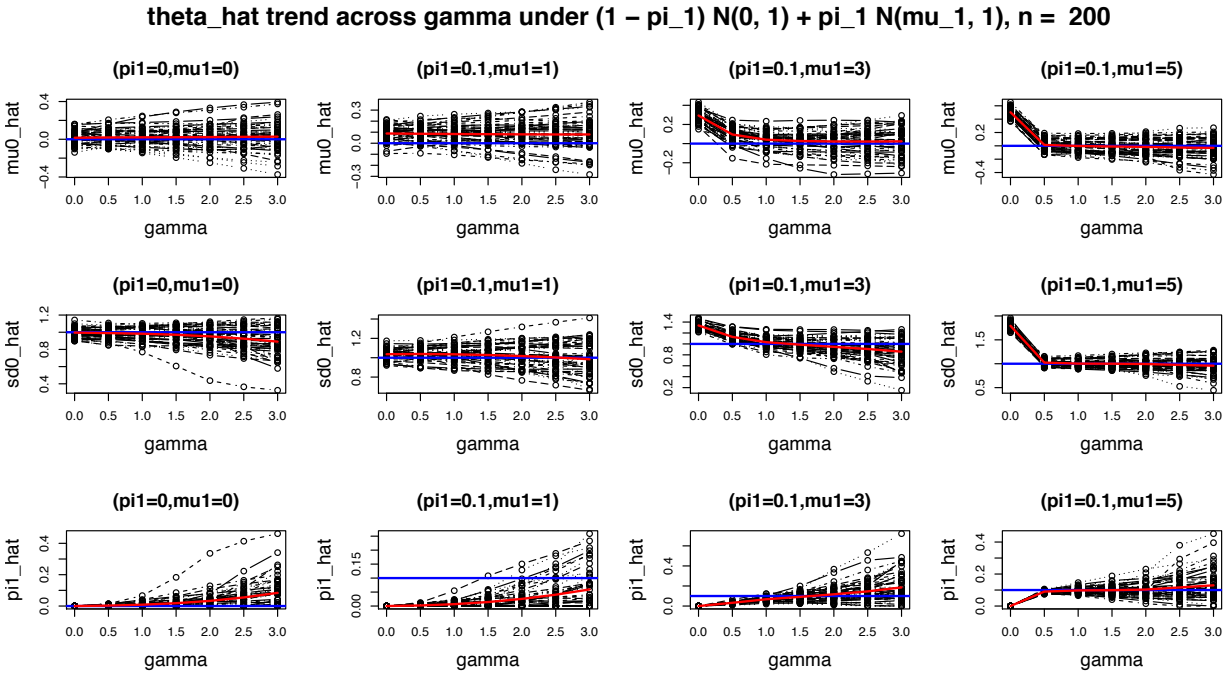

Figure S1: Trends of $\hat{\mu}_0$, $\hat{\sigma}_0$ and $\hat{\pi}_1$ across $\gamma = 0, 0.5, 1, 2, 3$ under Gaussian mixture model $(1 - \pi_1)\mathcal{N}(0, 1) + \pi_1\mathcal{N}(\mu_1, 1)$ where $n = 200$ and $(\mu_1, \pi_1) = (0, 0), (1, 0.1), (3, 0.1), (5, 0.1)$. Each black curve is from one sample realization. The red curve is the average of 50 black curves at each γ and the blue line is the underlying parameter value.

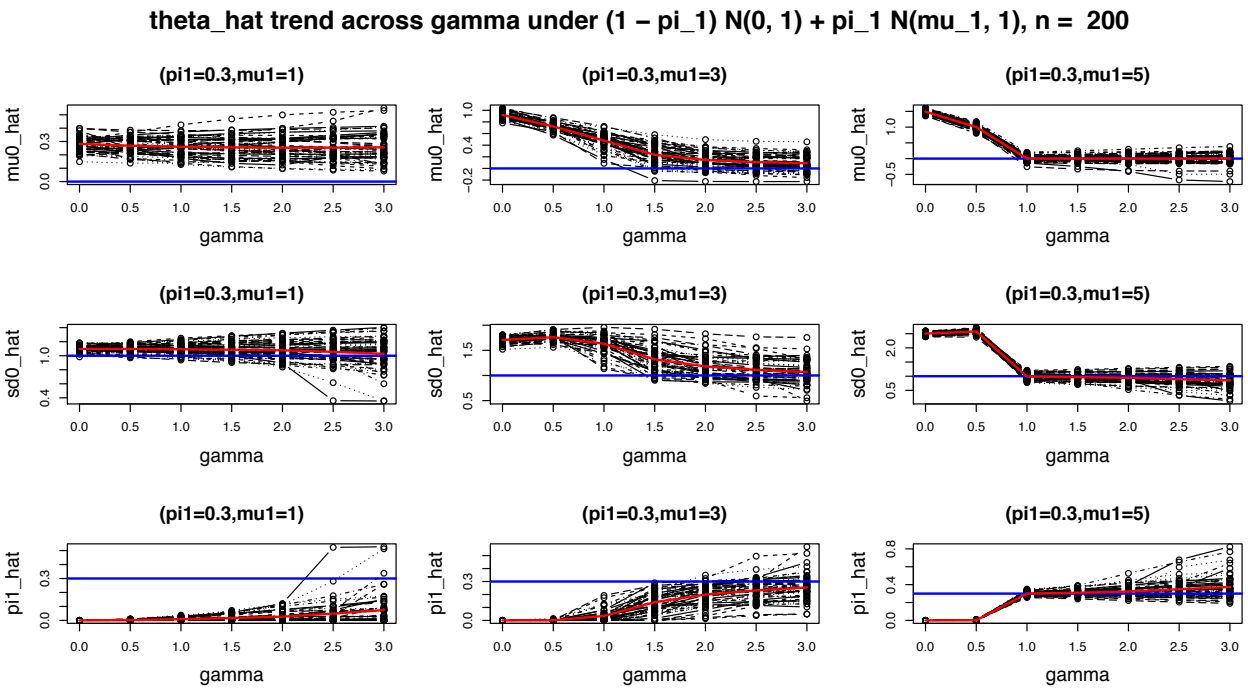

Figure S2: The same caption as Figure S1 under $n = 200$, $(\mu_1, \pi_1) = (1, 0.3), (3, 0.3), (5, 0.3)$.

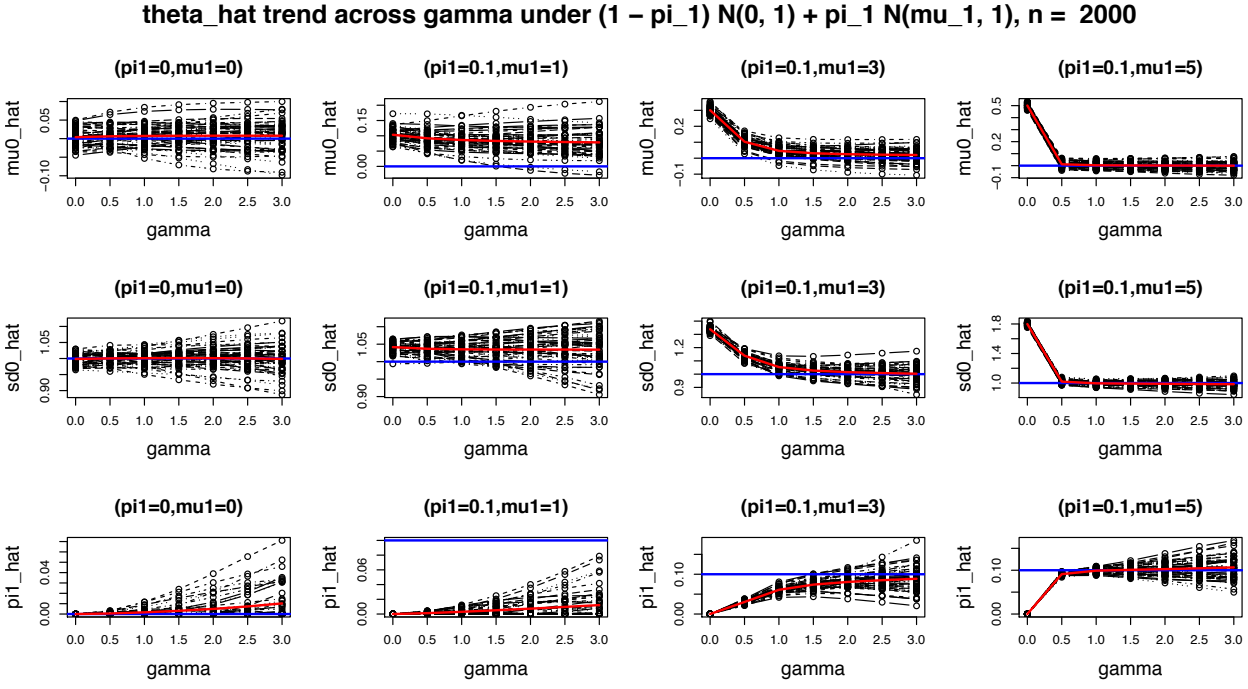

Figure S3: The same caption as Figure S1 under $n = 2000$, $(\mu_1, \pi_1) = (0, 0), (1, 0.1), (3, 0.1), (5, 0.1)$.

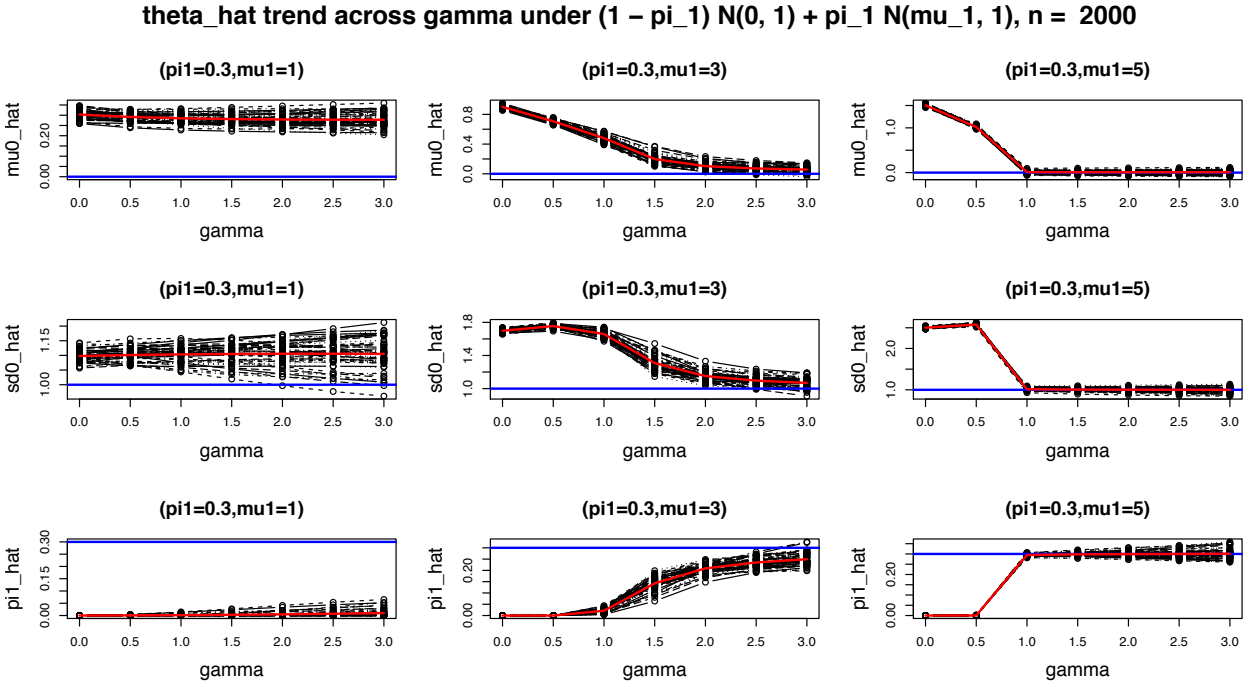

Figure S4: The same caption as Figure S1 under $n = 2000$, $(\mu_1, \pi_1) = (1, 0.3), (3, 0.3), (5, 0.3)$.

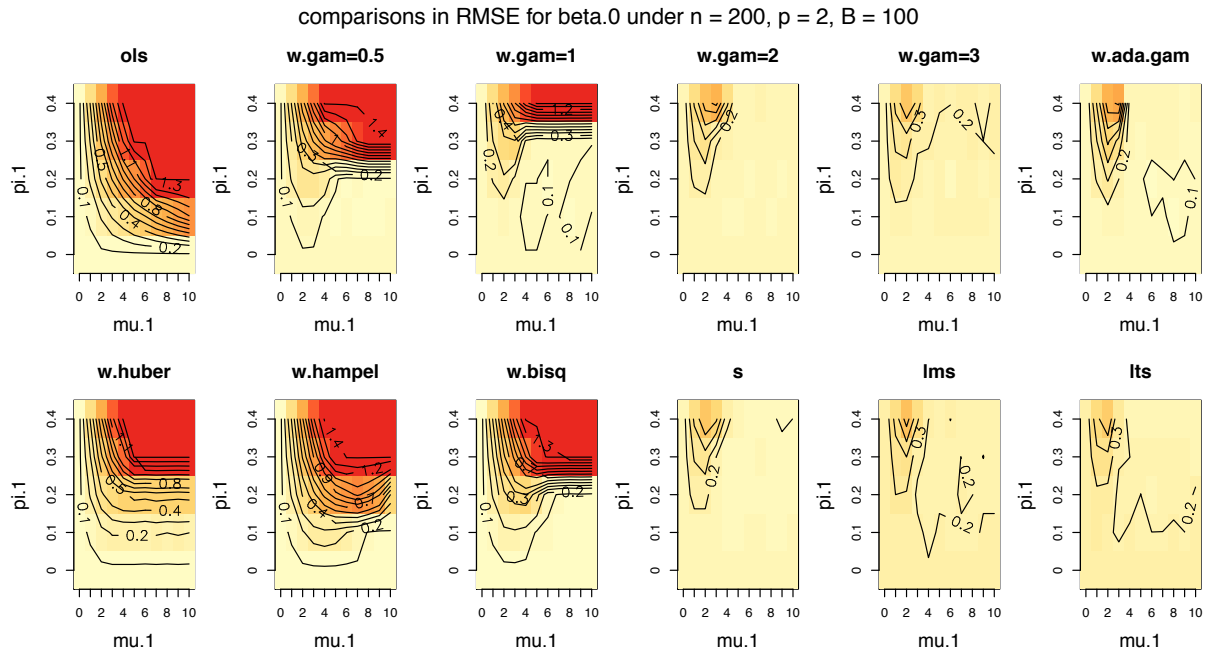

Figure S5: Heat map of RMSE comparisons for estimating $\boldsymbol{\beta_0}$ under the regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta_0} + \boldsymbol{\epsilon}$ where sample size $n = 200$ and $p = 2$, $\boldsymbol{\beta_0} = \mathbf{1_{p\times 1}}$ and $X_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 10^2)$ and the noise is i.i.d. from $(1 - \pi_1)\mathcal{N}(0, 1) + \pi_1\mathcal{N}(\mu_1, 1)$, and $(\pi_1, \mu_1) = (0, 0) \cup (0.1, 0.2, \ldots, 0.4) \times (1, 2, \ldots, 10)$. The RMSEs are truncated by 2.

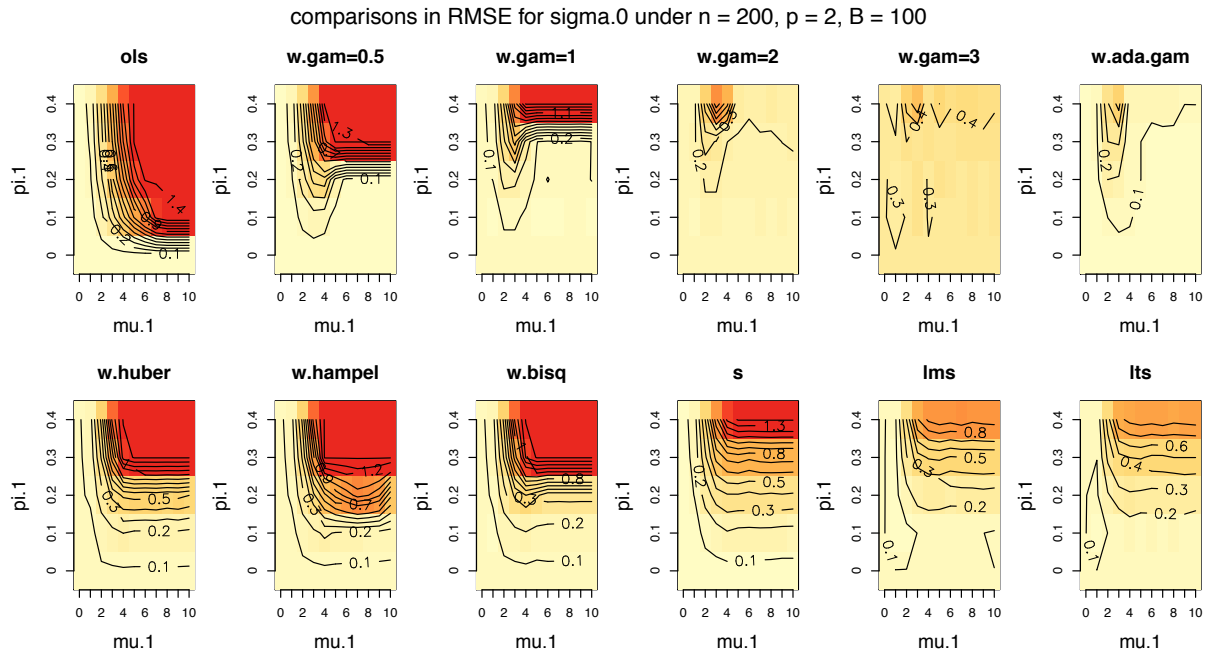

Figure S6: Heat map of RMSE comparisons for estimating σ_0 under the same setting as in Figure S5.

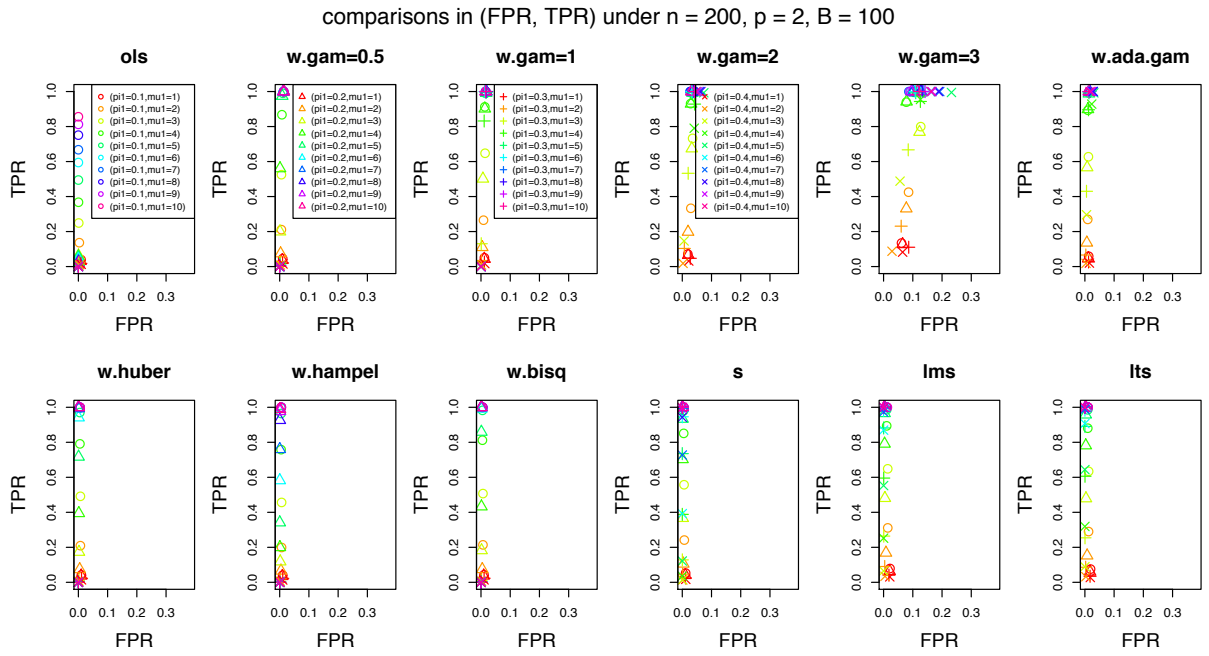

Figure S7: Results of (FPR, TPR) from AdaReg under various residual mixture models labeled in different colors and different shapes. The data were generated from the same model as in Figure S5.

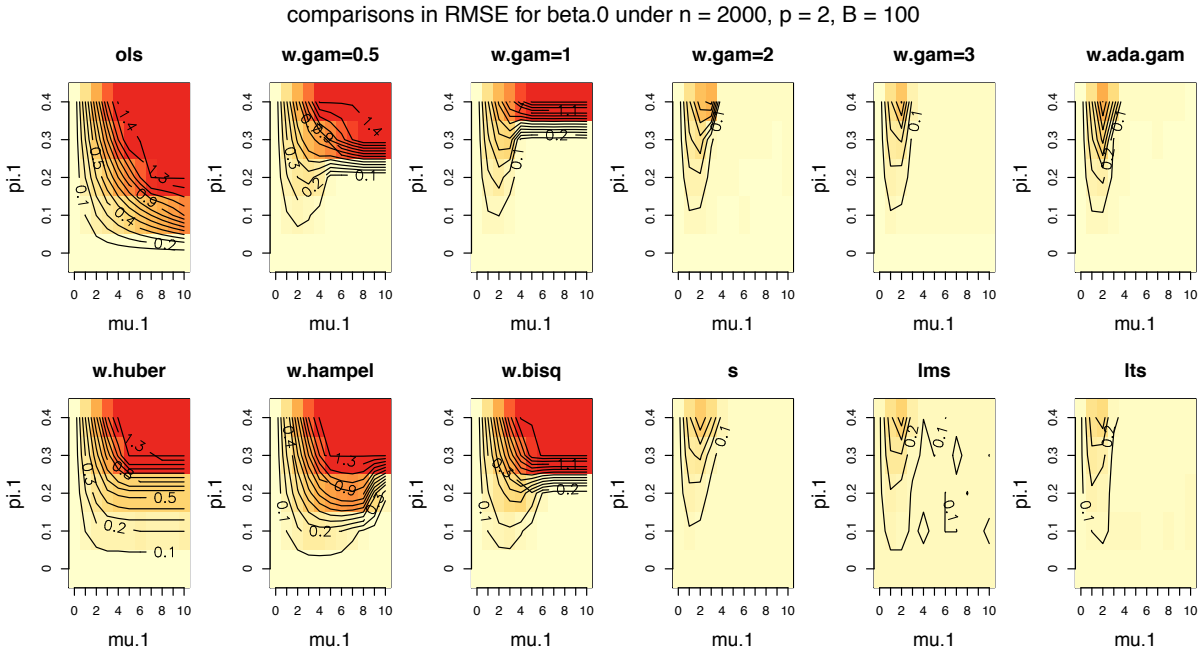

Figure S8: Heat map of RMSE comparisons for estimating $\boldsymbol{\beta_0}$ in the same labels as in Figure S5 except $n = 2000, p = 2$.

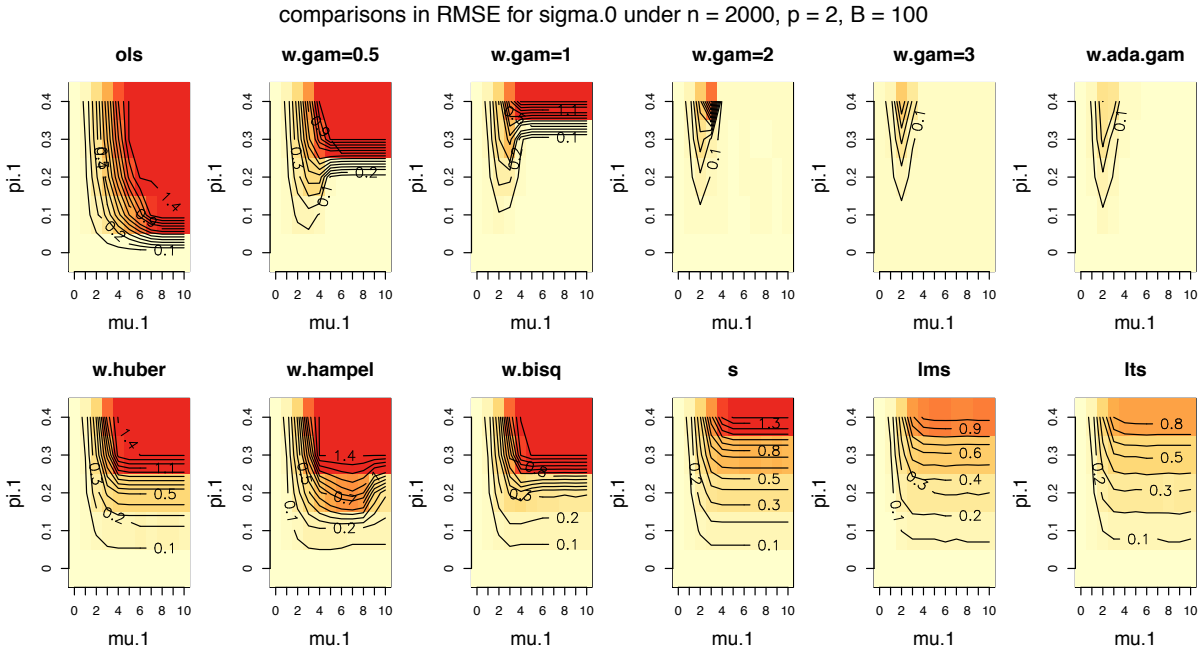

Figure S9: Heat map of RMSE comparisons for estimating σ_0 in the same labels as in Figure S6 except $n = 2000, p = 2$.

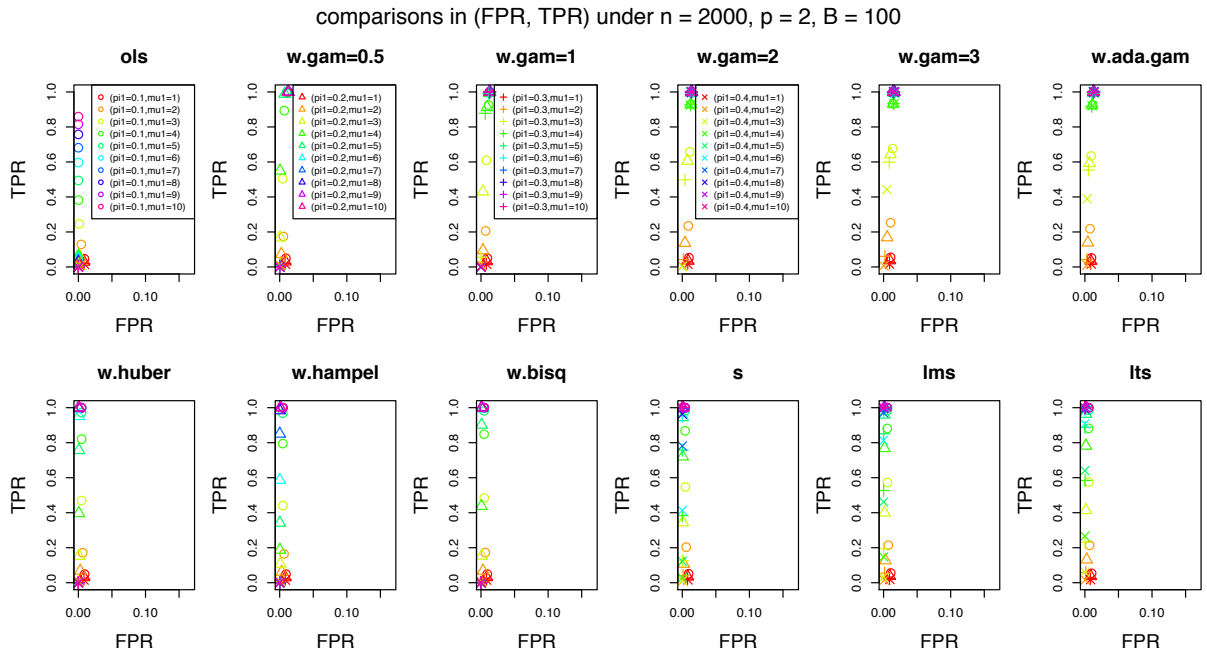

Figure S10: Results of (FPR, TPR) from AdaReg under various residual mixture models labeled in different colors and different shapes. The data were generated from the same model as in Figure S8.

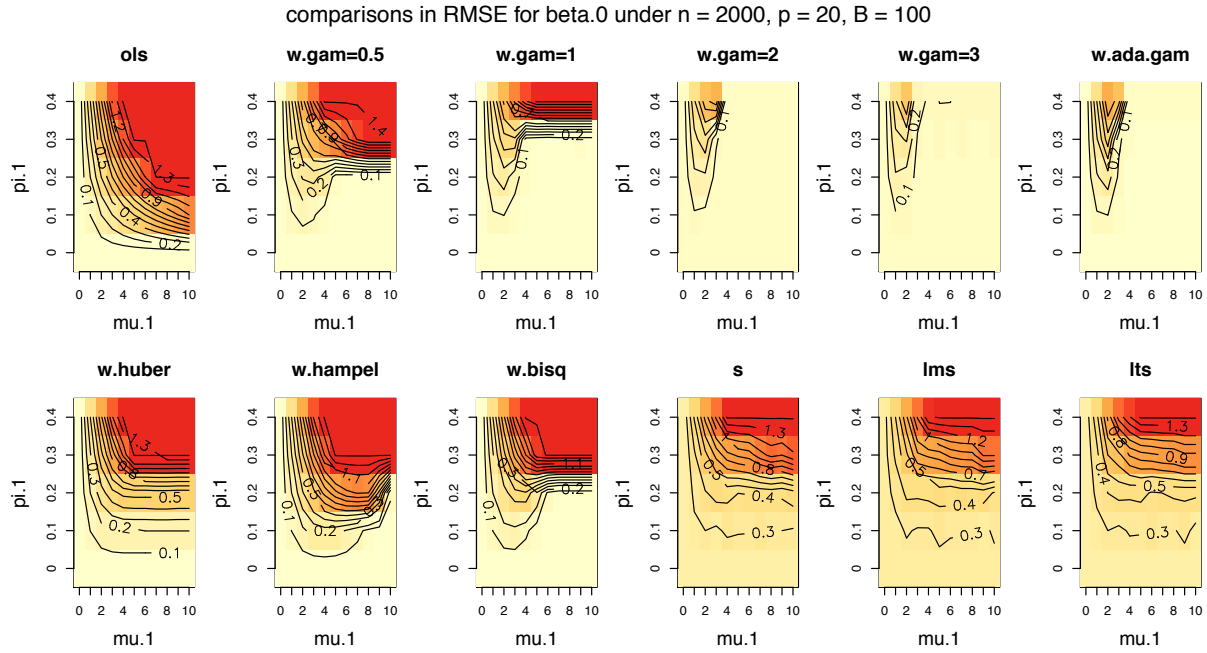

Figure S11: Heat map of RMSE comparisons for estimating $\boldsymbol{\beta_0}$ in the same labels as in Figure S5 except $n = 2000, p = 20$.

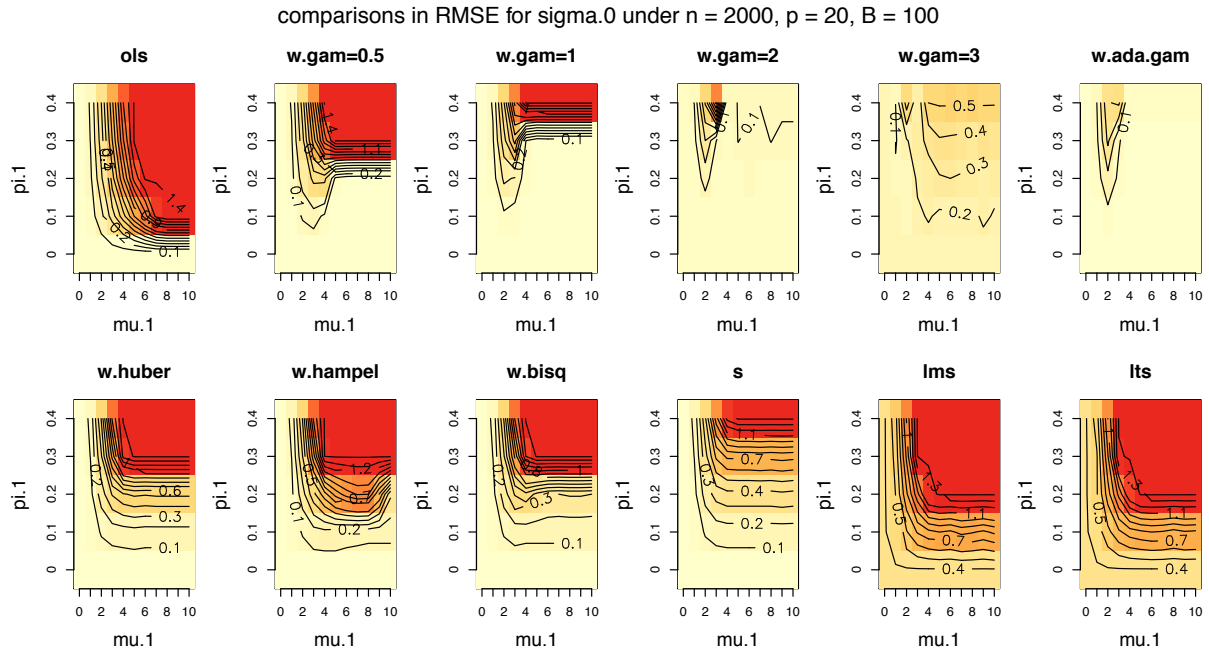

Figure S12: Heat map of RMSE comparisons for estimating σ_0 in the same labels as in Figure S6 except $n = 2000, p = 20$.

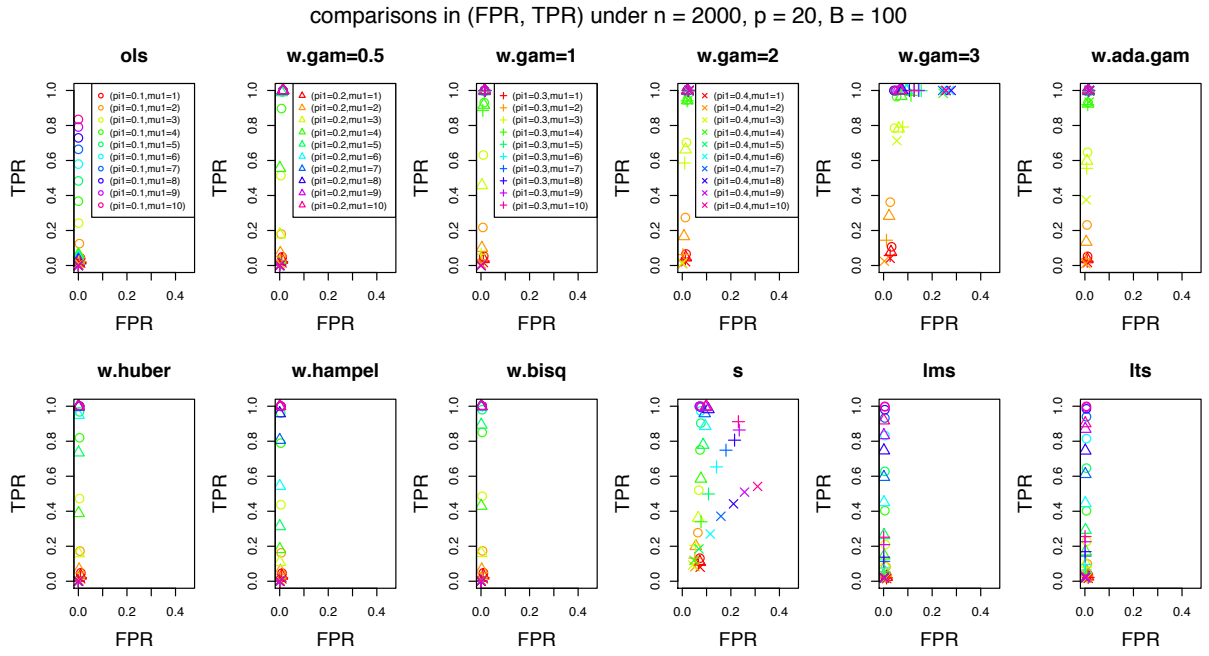

Figure S13: Results of (FPR, TPR) from AdaReg under various residual mixture models labeled in different colors and different shapes. The data were generated from the same model as in Figure S11.

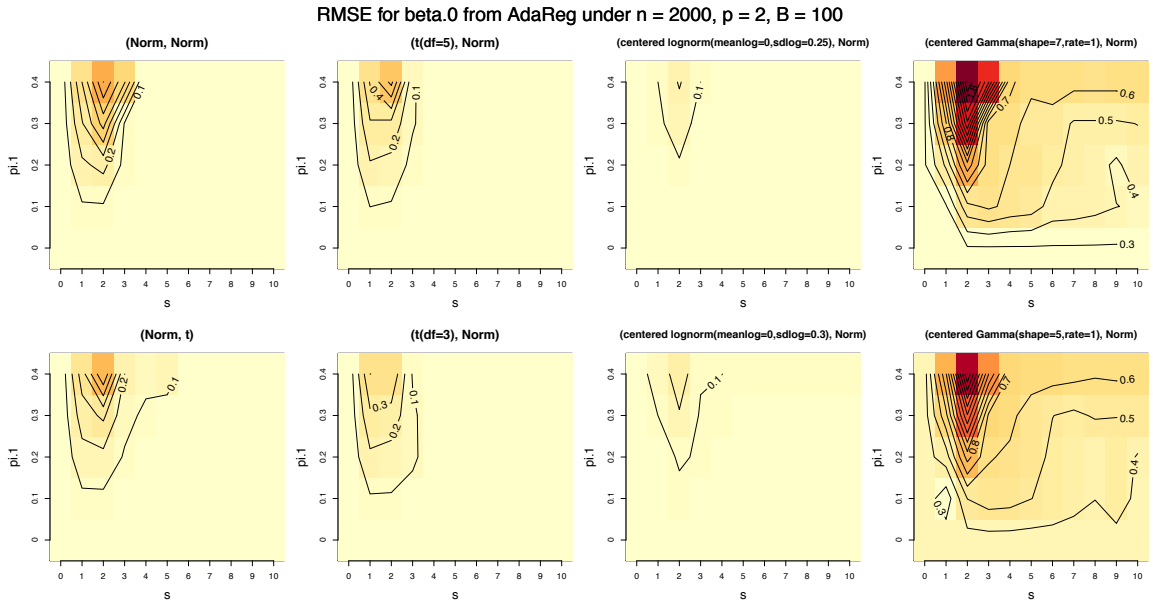

Figure S14: Heatmap of RMSE from AdaReg for estimating $\boldsymbol{\beta_0}$ under various residual mixture models in the regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta_0} + \boldsymbol{\epsilon}$, where the sample size $n = 2000$, the number of coefficients $p = 2$, the repeat time $B = 100$, $\boldsymbol{\beta_0} = \mathbf{1_{p\times 1}}$, and $X_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 10^2)$. The noise ϵ is i.i.d. from $(1-\pi_1)F_0 + \pi_1 F_1$, where F_0 is the population distribution and F_1 is the outlier distribution. The subtitle of each panel represents (F_0, F_1). The x-axis of each panel indicates the re-parameterization factor s, which ranges from 0 to 10 with the increment 1. The y-axis of each panel indicates the outlier proportion π_i ranging from 0 to 0.4 with the increment 0.1. The numbers along the contours indicate the values of RMSE. The RMSE is truncated by 2 for the plot.

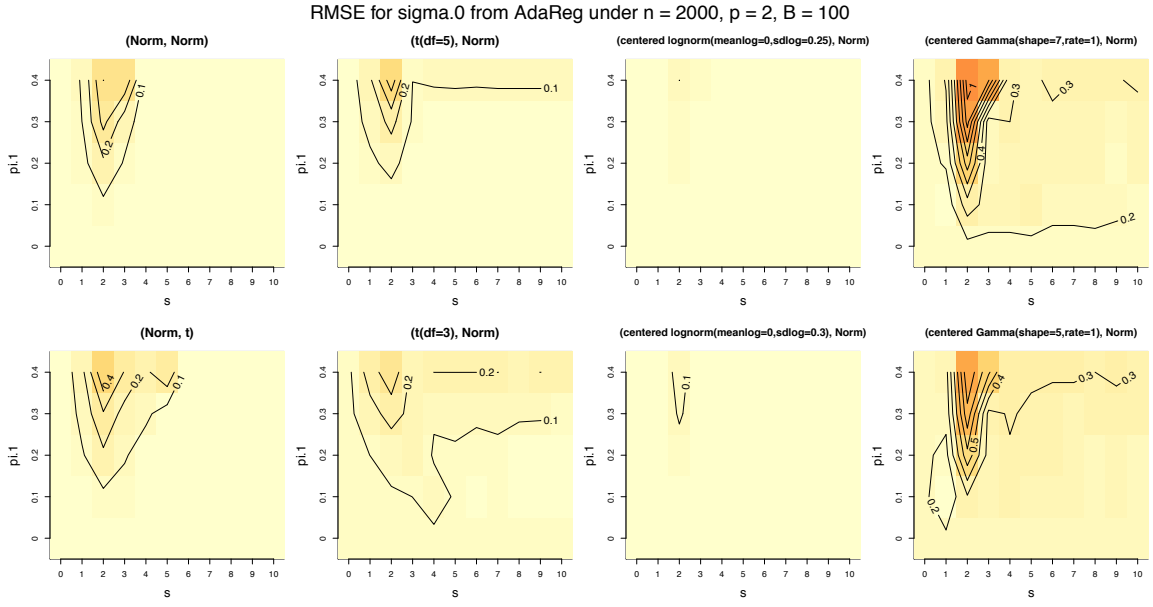

Figure S15: Heatmap of RMSE from AdaReg for estimating σ_0 under various residual mixture models in the same labels as in Figure S14.

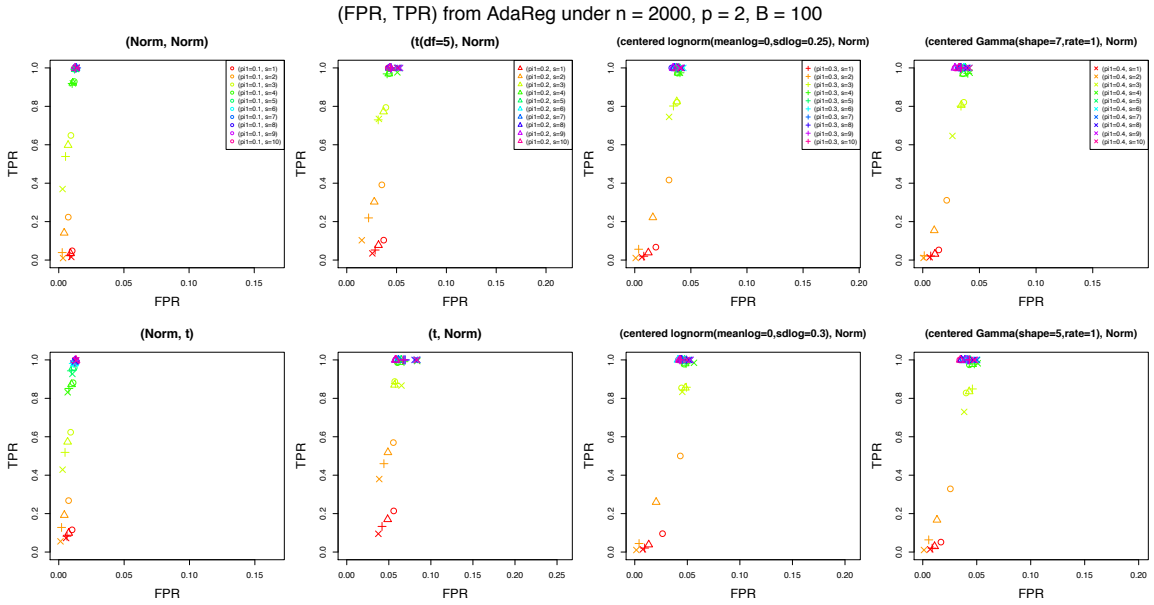

Figure S16: Results of (FPR, TPR) from AdaReg under various residual mixture models labeled in different colors and different shapes. The data were generated from the same model as in Figure S14.

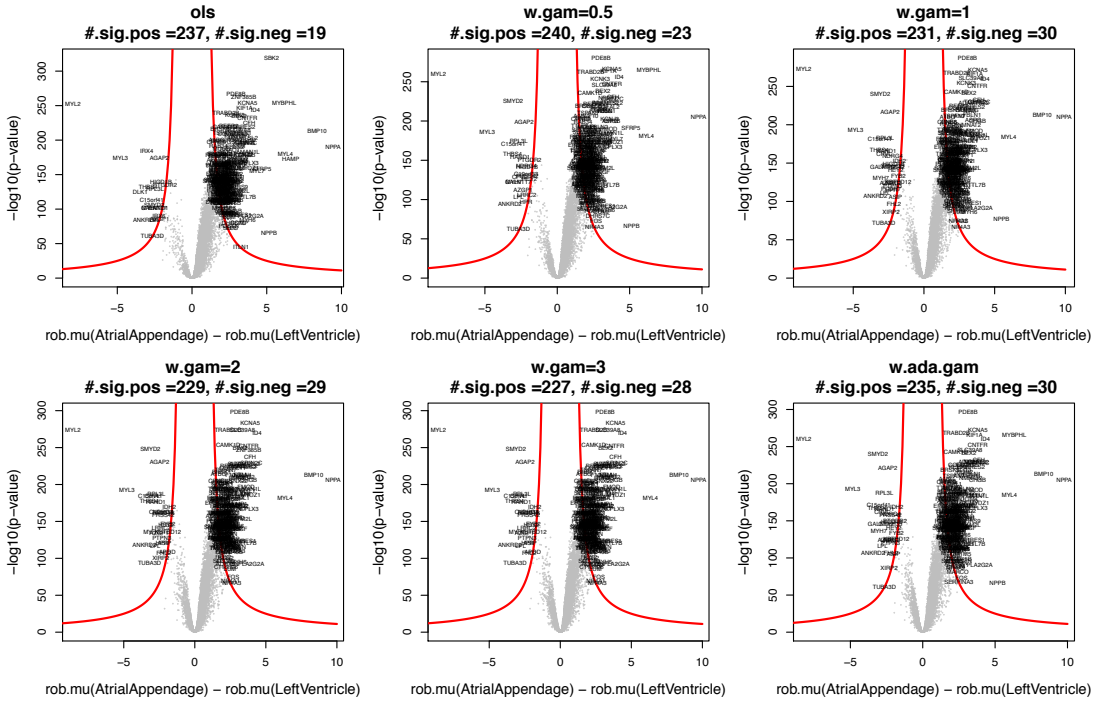

Figure S17: Volcano plot of minus of $\log_{10}(p\text{-value})$ versus log(fold change) ($\hat{\mu}$(atrial appendage) − $\hat{\mu}$(left ventricle)) from two-sample t-test with unequal variances after removing the outliers from various γ-robustifying procedures. The red curve is hyperbolic cut with curvature parameter 100 and minimum fold change parameter 1 (Singh et al., 2016). The gene names are labeled if they are above the red curve.