Open access • Posted Content • DOI:10.1101/869362

AdaReg: Data Adaptive Robust Estimation in Linear Regression with Application in GTEx Gene Expressions — Source link

Meng Wang, Lihua Jiang, Michael Snyder

Institutions: Stanford University

Published on: 10 Dec 2019 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Outlier, Mean squared error, Mixture model and Population

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

Meng Wang, Lihua Jiang, Michael P. Snyder

Department of Genetics
Stanford University

Abstract

With the development of high-throughput RNA sequencing (RNA-seq) technology, the Genotype Tissue-Expression (GTEx) project (Consortium et al., 2015) generated a valuable resource of gene expression data from more than 11,000 samples. The large-scale data set is a powerful resource for understanding the human transcriptome. However, the technical variation, sequencing background noise and unknown factors make the statistical analysis challenging. To eliminate the possibility that outliers might affect the estimation of population distribution, we need a more robust estimation method, a method that will adapt to heterogeneous genes and further optimize the estimate for each gene. We followed the approach of the robust estimation based on γ-density-power-weight (Fujisawa and Eguchi, 2008; Windham, 1995), where γ is the exponent of density weight which controls the balance between bias and variance. As far as we know, our work is the first to propose a procedure to tune the parameter γ to balance the bias-variance trade-off under the mixture distributions. We constructed a robust likelihood criterion based on weighted densities in the mixture model of Gaussian population distribution mixed with unknown outlier distribution, and developed a data-adaptive γ-selection procedure embedded into the robust estimation. We provided a heuristic analysis on the selection criterion and found that our practical selection trend under various γ's in average performance has similar capability to capture minimizer γ as the inestimable Mean Squared Error (MSE) trend from our simulation studies under a series of settings. Our data-adaptive robustifying procedure in the linear regression problem (AdaReg) shows a significant advantage in both simulation studies and real data application of heart samples from the GTEx project compared to the fixed γ procedure and other robust methods. This paper discusses some limitations of this method, and future work.

Keywords: robust estimation, density power weight, data-adaptive, selection criterion, linear regression, GTEx project

1 Introduction

The emergence of high-throughput RNA sequencing (RNA-seq) technology dramatically increases the development of gene expression analysis. The Genotype Tissue-Expression (GTEx) project (Consortium et al., 2015) generated the gene expression data from RNA-seq from more than $11,000$ samples of 53 tissues (up to the version 7 release from GTEx Portal), providing a valuable resource to study tissue variation in the human transcriptome. However, the technical variation, sequencing background noise and unknown factors make statistical analysis challenging. As an example, consider the expression for gene MYH7 (myosin heavy chain 7) in heart atrial appendage and heart left ventricle. Figure 1 (in the left panel) shows sample densities in these two parts of heart. We can see a long tail in the low expression end of the distribution of left ventricle samples. These outliers could be caused by technical noise, or could represent an abnormality in the sample expression. This is only for one gene, while different genes have different types of outliers. The

outlier proportions can be large or small, and their magnitudes vary from gene to gene. Given the presence of various outliers, there is a great need to have a robust estimation method, especially adaptive to different cases.

The literature of robust estimation in linear regression is very rich (Hampel et al., 2011; Huber, 2011; Maronna et al., 2018; Rousseeuw and Leroy, 1987). One approach is based on selecting a subset less influenced by the leverage points such as the least median squares (lms) (Rousseeuw, 1984), finding the narrowest hyperplanes covering half of the observations, and the least trimmed squares (lts) (Rousseeuw, 1984, 1985), removing the topmost most leverage points. The other approach is based on down-weighting the outliers by choosing various weight functions forming the M-estimate. Classical seminal works include Huber's weight (Huber, 1964),, Hampel's three-part weight, Tukey's bisquare weight, S-estimation (Rousseeuw and Yohai, 1984), and others. Windham (1995) proposed a robustifying procedure based on the density power weight f_0^γ ($\gamma \geq 0$) to robustly fit the population model f_0 (extended details in Subsection 2.2), which is related to the density weight divergence in Basu et al. (1998). Jones et al. (2001) gave a comparison of these two methods. Fujisawa and Eguchi (2008) revisited these old works and constructed a γ-cross entropy robust criterion, assuming under a proper $\gamma(\geq 0)$, the outliers go to the tails of density power f_0^γ and thus do not contribute much in the population estimation. Recently, the γ-cross entropy criterion has gained much attention and there are a series of variant works including robust estimation using an unnormalized model (Kanamori and Fujisawa, 2015), robust clustering (Chen et al., 2014), Gaussian graphical modeling (Katayama et al., 2018; Miyamura and Kano, 2006), and others.

One useful property of both the γ-cross entropy criterion and the γ-density-power-weight is that the estimation procedure is controlled by only one system parameter: γ. Many previous papers pointed out that γ controls the trade-off between robustness and efficiency of the estimate. In Subsection 2.4, we investigate the trade-off in terms of bias and variance. However, how to choose a proper γ in practice is still unknown. Previous γ-robustifying procedures are based on a fixed preselected γ. In the example of gene MYH7, Figure 1 (in the right panel) shows the density plot of standardized residuals under various γ's. When $\gamma = 0$, the estimation is the same as from Ordinary Least Squares (OLS). Ideally, if γ is the properly selected, the residual peak should be around zero. Under $\gamma = 0, 0.5, 1$, the residual peaks are obviously greater than zero, which means such γ is not large enough so that there is still some information left in the residuals, while under $\gamma = 2, 3$, the density peaks are close to zero. We can see this gene prefers a large γ. As we have pointed out, the outliers vary from gene to gene and thus a proper γ also varies among different genes. This motivates us to develop a procedure to select a proper γ to adapt to each gene.

Our contribution is that following the approach of using density power weight, we construct a robust weighted likelihood estimation criterion based on weighted densities in the mixture model, and develop a data-adaptive γ-selection procedure in linear regression (AdaReg). We provide a heuristic analysis on the selection criterion, and find that our practical selection trend under various γ's in average performance has similar capability to capture minimizer γ as the inestimable Mean Squared Error (MSE) trend from our simulation studies under a series of settings. In this paper, we mainly discuss the estimation problem in linear regression. We believe our data-adaptive robustifying procedure will broaden the direction of density power weight in more applications.

The outline of the rest of the paper is as follows: In Section 2, we develop and analyze our algorithm, AdaReg. In Subsection 2.1, we first set up the regression problem in a mixture model. In Subsection 2.2, we consider weighting the model to purify the mixture. In Subsection 2.3, using our weighted mixture model, we construct our robust likelihood criterion based on the weighted densities, and provide robust estimation under a fixed γ. In Subsection 2.4, we investigate the bias and variance trade-off in terms of γ. We develop a novel γ selection criterion in Subsection 2.5. In Subsection 2.6, we summarize our data-adaptive algorithm (AdaReg) in the linear regression problem. Section 3 covers simulations and applications. We apply AdaReg in our simulation studies in Subsection 3.1, and in a real dataset of heart samples from RNA-seq in the GTEx project in Subsection 3.2. We compare our method to the fixed γ procedures and other robust regression methods. Finally, in Section 4, we summarize our work, point out a few limitations, and suggest some further work.

2 Method

2.1 Problem setup

Suppose we would like to investigate the relationship between a response variable $\mathbf{y} \in \mathbb{R}^n$ and design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ from n samples. For example, we are interested in investigating gene expression variation under different tissue types. In the similar notation as in (Katayama et al., 2018), consider the response variable $\mathbf{y}$ coming from a mixture model,

$$\mathbf{y} = (\mathbf{I} - \mathbf{B_1})\mathbf{y_0} + \mathbf{B_1}\mathbf{y_1}, \tag{1}$$

where $\mathbf{I} \in \mathbb{R}^{n\times n}$ is the identity matrix and $\mathbf{B_1}$ is a diagonal matrix where the diagonal elements are Bernoulli variables with probability $\pi_1 \in [0, 0.5)$ to be 1, and $\mathbf{y_0}$ is the clean part forming the population, and $\mathbf{y_1}$ is the outlier. The 0-1 elements in the diagonal of $\mathbf{B_1}$ indicate the samples coming from the clean part $\mathbf{y_0}$ or the outlier $\mathbf{y_1}$. We assume independence among all the components of $\mathbf{y}$. We only consider the outliers in the response variable not in the covariates here. Suppose the clean part of $\mathbf{y}$ comes from the ordinary linear model,

$$\mathbf{y}_0 = \mathbf{X}\boldsymbol{\beta_0} + \boldsymbol{\epsilon_0}, \tag{2}$$

where $\boldsymbol{\beta_0} \in \mathbb{R}^p$ is the coefficient vector, and the Gaussian independent noise $\boldsymbol{\epsilon_0} \sim \mathcal{N}(\mathbf{0}, \mathbf{D_0})$, $\mathbf{D_0} = \text{diag}(\sigma_{01}^2\mathbf{1}_{n_1\times 1}, \ldots, \sigma_{0K}^2\mathbf{1}_{n_K\times 1})$, K is the group number, n_k is the sample size in group k, and each group can have its own variance σ_{0k}^2. We further parameterize the population distribution as Gaussian. For the RNA-seq data, as a convention, we take logarithm transformation on the standardized expression of Transcripts Per Kilobase Million (TPM) or Reads Per Kilobase Million (RPKM) to make the density more symmetric, more Gaussian distributed. For the intensity data such as from microarray or mass spectrometry platform, the expression data is conventionally assumed as Gaussian distributed in the log scale. Under unknown distribution of the outliers and non-vanishing outlier proportion, our goal is to robustly estimate the population parameters $\boldsymbol{\theta_0} = (\boldsymbol{\beta_0}, \sigma_{01}^2, \ldots, \sigma_{0K}^2)$ and π_0 for each gene.

In the following analysis, we first consider there is only one group, i.e., the design matrix $\mathbf{X}$ is simply $\mathbf{1_{n\times 1}}$, and develop a criterion to data-adaptively select γ then go back to embed our selection procedure into a general linear regression setting.

2.2 Weighted mixture density

Suppose the response $\mathbf{y}$ is from one group. Based on the setting (1)-(2), the problem is reduced to each component y_i from a mixture density

$$y_i \overset{\text{iid}}{\sim} f(y) := (1-\pi_1)f_{0,\boldsymbol{\theta_0}}(y) + \pi_1 f_1(y), \quad i = 1, \ldots, n, \tag{3}$$

where f is the mixture density function, $f_{0,\boldsymbol{\theta_0}}(\cdot) = \phi(\cdot; \mu_0, \sigma_0^2)$ is for the population modeled as Gaussian density ϕ with mean μ_0 and variance σ_0^2, and f_1 is the unknown outlier density.

In the early work, Windham (1995) considered to attach a weight $w_i = w(y_i) \propto \phi^\gamma(y_i; \mu_0, \sigma_0^2)$ to each y_i, where $\gamma \geq 0$ is the exponent parameter and the weights are self-standardized, that is, $\sum_{i=1}^n w_i = 1$. In this way, weighting the samples by the power of the population density, the points coming from the population gain more weights, while the outliers gain less weights so that the outliers do not contribute much in the population estimation. To estimate the score function $s(y, \boldsymbol{\theta}) := \partial \log f_{0,\boldsymbol{\theta}}(y)/\partial\boldsymbol{\theta}$, matching its empirical average from the weighted samples and the corresponding theoretical expectation gives the estimating equation for the Windham's procedure,

$$\sum_{i=1}^n s(y_i, \boldsymbol{\theta})w_i = \int s(y, \boldsymbol{\theta})\phi\left(y; \mu, \frac{\sigma^2}{1+\gamma}\right)\mathrm{d}y. \tag{4}$$

When $\gamma = 0$, the solution to (4) is the Maximum Likelihood Estimation (MLE).

Now we formulate the process of Windham's procedure in terms of weighting the mixture density. Consider multiplying $f_{0,\boldsymbol{\theta_0}}^{1+\gamma}$ on both sides of (3), then re-standardizing each product to a density function gives our weighted mixture model,

$$f^{(w_0)}(y) = \pi_0^{(w_0)} f_0^{(w_0)}(y) + \pi_1^{(w_0)} f_1^{(w_0)}(y), \tag{5}$$

where

$$\begin{cases} f^{(w_0)}(y) = \frac{f(y)f^{\gamma}_{0,\boldsymbol{\theta_0}}(y)}{\int f(x)f^{\gamma}_{0,\boldsymbol{\theta_0}}(x)\mathrm{d}x}, \\ f_0^{(w_0)}(y) = \frac{f^{1+\gamma}_{0,\boldsymbol{\theta_0}}(y)}{\int f^{1+\gamma}_{0,\boldsymbol{\theta_0}}(x)\mathrm{d}x}, \\ f_1^{(w_0)}(y) = \frac{f_1(y)f^{\gamma}_{0,\boldsymbol{\theta_0}}(y)}{\int f_1(x)f^{\gamma}_{0,\boldsymbol{\theta_0}}(x)\mathrm{d}x}, \\ \pi_0^{(w_0)} = \frac{\pi_0}{\pi_0+\pi_1\delta^{(w_0)}}, \\ \pi_1^{(w_0)} = 1-\pi_0^{(w_0)}, \\ \delta^{(w_0)} = \int \frac{f_1(x)}{f_{0,\boldsymbol{\theta_0}}(x)} f_0^{(w_0)}(x)\mathrm{d}x. \end{cases}$$

In the weighted mixture model (5), the function with superscript (w_0) means the function depends on the weight $f^{\gamma}_{0,\boldsymbol{\theta_0}}$. When the population and outlier distributions, the null and the alternative, are not hard to distinguish, under a proper γ, the integral of the likelihood ratio of f_1 to $f_{0,\boldsymbol{\theta_0}}$ is small in the probability measure of the weighted null density. In such scenario, $(\pi_1/\pi_0)\delta^{(w_0)} \approx 0$ and thus $\pi_0^{(w_0)} \approx 1$. Therefore, our weighted mixture model can be approximately purified to

$$f^{(w_0)}(y) \approx f_0^{(w_0)}(y). \tag{6}$$

In the case that $f_{0,\boldsymbol{\theta_0}}(\cdot) = \phi(\cdot;\mu_0,\sigma_0^2)$, the empirical mass density of $f^{(w_0)}$ at y_i can be estimated by the weight

$$\bar{f}^{(w_0)}(y_i) = w_i = \frac{\phi^{\gamma}(y_i;\mu_0,\sigma_0^2)}{\sum_{j=1}^{n}\phi^{\gamma}(y_j;\mu_0,\sigma_0^2)}, \tag{7}$$

and $f_0^{(w_0)}$ is the density of Gaussian distribution $\mathcal{N}(\mu_0, \frac{\sigma_0^2}{1+\gamma})$ with mean μ_0 and variance $\frac{\sigma_0^2}{1+\gamma}$, shrinking the variance to $1/(1+\gamma)$ of the original variance, and thus the outliers go to the tails of the variance-shrunk density.

Remark: unlike other weighting procedures such as Huber, Hampel, Tukey's methods where the weights are basically loss penalty, the weights in the γ-robustifiying procedure take the role of transforming samples from a mixture distribution to a more purified distribution.

2.3 Robust estimation under a fixed γ

From the estimation equation in (4), the Windham's procedure essentially relies on the weighted density approximation (6). Here we take a direct approach, measuring the ordinary cross entropy of the weighted mixture density to the weighted theoretical null density,

$$\begin{aligned} d_w(f^{(w)}, f_0^{(w)};\boldsymbol{\theta},\gamma) &= -\int f^{(w)}(y)\log f_0^{(w)}(y)\mathrm{d}y \\ &= -\int \frac{f(y)\phi^{\gamma}(y;\mu,\sigma^2)}{\int f(x)\phi^{\gamma}(x;\mu,\sigma^2)\mathrm{d}x}\phi\left(y;\mu,\frac{\sigma^2}{1+\gamma}\right)\mathrm{d}y, \end{aligned} \tag{8}$$

where $f^{(w)}$ and $f_0^{(w)}$ are defined in (5) (replacing $\boldsymbol{\theta_0}$ by $\boldsymbol{\theta}$). The empirical cross entropy on the samples is

$$\begin{aligned} \bar{d}_w(\bar{f}^{(w)}, f_0^{(w)};\boldsymbol{\theta},\gamma) &= -\sum_{i=1}^{n}\bar{f}^{(w)}(y_i)\log f_0^{(w)}(y_i) \\ &= -\sum_{i=1}^{n} w_i \log\phi\left(y_i;\mu,\frac{\sigma^2}{1+\gamma}\right) \\ &= \frac{1}{2}\left(\log\left(\frac{2\pi\sigma^2}{1+\gamma}\right) + \sum_{i=1}^{n} w_i\frac{(x_i-\mu)^2}{\sigma^2/(1+\gamma)}\right), \end{aligned} \tag{9}$$

where w_i is defined in (7). Given a fixed γ, define

$$\tilde{\boldsymbol{\theta}}_{\boldsymbol{0\gamma}} = \arg\min_{\boldsymbol{\theta}} d_w(f^{(w)}, f_0^{(w)};\boldsymbol{\theta},\gamma), \tag{10}$$

and its M-estimate

$$\hat{\boldsymbol{\theta}}_{\mathbf{0}\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\theta}} \bar{d}_w(\bar{f}^{(w)}, f_0^{(w)}; \boldsymbol{\theta}, \gamma). \tag{11}$$

Since the weights reply on the parameter we would like to estimate, we can iteratively update the weights and $\boldsymbol{\theta}$. The minimizer $\hat{\boldsymbol{\theta}}_{\mathbf{0}\boldsymbol{\gamma}}$ is the fixed point. From our practical experience and previous works, the convergence of the algorithm is not a problem. In the step of given the weights w_i's to update $\boldsymbol{\theta}$, setting the derivative of $\bar{d}_w(\bar{f}^{(w)}, f_0^{(w)}; \boldsymbol{\theta}, \gamma)$ with respect to $\boldsymbol{\theta}$ to be zero gives

$$\hat{\mu}_0 = \sum_{i=1}^{n} w_i y_i, \tag{12}$$

$$\hat{\sigma}_0^2 = (1+\gamma)\sum_{i=1}^{n} w_i (y_i - \hat{\mu}_0)^2, \tag{13}$$

which is the exact estimation equation in (4). Rearranging the terms in (4) gives

$$\frac{1}{n}\sum_{i=1}^{n} \psi_\gamma(y_i, \boldsymbol{\theta}) = 0, \tag{14}$$

where $\psi_\gamma(y, \boldsymbol{\theta}) = f_{0,\boldsymbol{\theta}}^{\gamma}(y) s(y, \boldsymbol{\theta}) \int f_{0,\boldsymbol{\theta}}^{1+\gamma}(x)\mathrm{d}x - f_{0,\boldsymbol{\theta}}^{\gamma}(y) \int f_{0,\boldsymbol{\theta}}^{1+\gamma}(x) s(x, \boldsymbol{\theta})\mathrm{d}x$ and $s(y, \boldsymbol{\theta}) = \frac{\partial \ln f_{0,\boldsymbol{\theta}}(y)}{\partial \boldsymbol{\theta}}$, which is the estimating equation for γ-cross entropy (Fujisawa and Eguchi, 2008). Our M-estimate from the cross entropy criterion (9) agrees with the estimator from Windham's procedure and γ-cross entropy criterion, and thus they share the same consistency property and asymptotic normality property stated in Proposition 1. In the case that underlying density is purely Gaussian without outliers ($\pi_1 = 0$), the asymptotic variances for $\hat{\mu}_0$ and $\hat{\sigma}_0^2$ are

$$(1 + \frac{\gamma^2}{1+2\gamma})^{3/2}\sigma_0^2, \quad \text{and} \quad \frac{(1+\gamma)^3(3\gamma^2+4\gamma+2)}{(1+2\gamma)^{5/2}}\sigma_0^4, \quad \text{respectively,}$$

also shown in (Basu et al., 1998; Jones et al., 2001). From Proposition 1, setting $\gamma = 0$ gives that the asymptotic variance for $\hat{\mu}_0$ is σ_0^2 and for $\hat{\sigma}_0^2$ is $2\sigma_0^4$, which are the most efficient asymptotic variances for MLEs. If the predefined $\gamma > 0$, their asymptotic variances increase with γ.

Proposition 1. *If there is no outlier that $f = f_{0,\boldsymbol{\theta}_0}$, then $\tilde{\boldsymbol{\theta}}_{\mathbf{0}\boldsymbol{\gamma}} = \boldsymbol{\theta}_{\mathbf{0}}$ and thus the estimate is Fisher consistent. Under the mild conditions, applying the theorems for M-estimate in Van der Vaart (2000), shown in Fujisawa (2013); Fujisawa and Eguchi (2008),*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathbf{0}\boldsymbol{\gamma}} - \tilde{\boldsymbol{\theta}}_{\mathbf{0}\boldsymbol{\gamma}}) \to^d \mathcal{N}(\mathbf{0}, \Sigma_f(\tilde{\boldsymbol{\theta}}_{\mathbf{0}\boldsymbol{\gamma}})), \quad \text{as } n \to \infty,$$

where $\Sigma_f(\boldsymbol{\theta}) = J_f(\boldsymbol{\theta})^{-1} I_f(\boldsymbol{\theta}) J_f^{\top}(\boldsymbol{\theta})^{-1}$, $J_f(\boldsymbol{\theta}) = \mathbb{E}_f(\frac{\partial \psi_\gamma(Y,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\top}})$ and $I_f(\boldsymbol{\theta}) = \mathbb{E}_f\, \psi_\gamma(Y, \boldsymbol{\theta})\psi_\gamma^{\top}(Y, \boldsymbol{\theta})$.

The population proportion can be simply estimated from the view of weighting the mixture model. Consider taking the integral of unstandardized density-power-weighted mixture model,

$$\int f(y) f_{0,\boldsymbol{\theta}_0}^{\gamma}(y)\mathrm{d}y = \pi_0 \int f_{0,\boldsymbol{\theta}_0}^{1+\gamma}(y)\mathrm{d}y + \pi_1 \int f_1(y) f_{0,\boldsymbol{\theta}_0}^{\gamma}(y)\mathrm{d}y. \tag{15}$$

Rearranging (15) gives

$$\pi_0 \quad = \quad \pi_0^{\ddagger} - (1-\pi_0)\delta^{(w_0)}, \quad \text{where } \pi_0^{\ddagger} = \frac{\int f(y) f_{0,\boldsymbol{\theta}_0}^{\gamma}(y)\mathrm{d}y}{\int f_{0,\boldsymbol{\theta}_0}^{1+\gamma}(y)\mathrm{d}y} \text{ and } \delta^{(w_0)} \text{ is defined in (5).}$$

Hence, an upward bias estimator for π_0 is

$$\hat{\pi}_0 = \frac{\frac{1}{n}\sum_{i=1}^{n} f_{0,\hat{\boldsymbol{\theta}}_0}^{\gamma}(y_i)}{\int f_{0,\hat{\boldsymbol{\theta}}_0}^{1+\gamma}(y)\mathrm{d}y}, \tag{16}$$

which agrees with the result in Kanamori and Fujisawa (2015) by minimizing their unnormalized density power score.

2.4 Bias and variance trade-off in γ

So far we have seen that the M-estimate $\hat{\boldsymbol{\theta}}_{\mathbf{0\gamma}}$ and its asymptotic limit $\tilde{\boldsymbol{\theta}}_{\mathbf{0\gamma}}$ depend on a predefined γ. Proposition 1 shows that $\hat{\boldsymbol{\theta}}_{\mathbf{0\gamma}}$ is $\sqrt{n}$-consistent to $\tilde{\boldsymbol{\theta}}_{\mathbf{0\gamma}}$. However, there is still a latent bias from $\tilde{\boldsymbol{\theta}}_{\mathbf{0\gamma}}$ to the target parameter $\boldsymbol{\theta}_{\mathbf{0}}$, also pointed out in Fujisawa (2013). The weighted mixture model (5) is the ideal model where the $\boldsymbol{\theta}$ in the weighting density $f^{\gamma}_{0,\boldsymbol{\theta}}$ is the underlying $\boldsymbol{\theta}_{\mathbf{0}}$. However, actually the weighting density we fit is under $\tilde{\boldsymbol{\theta}}_{\mathbf{0\gamma}}$ (supposed we have an infinite amount of samples), and the fitted weighted mixture density is

$$\begin{aligned} f^{(\tilde{w})}(y) &:= \frac{f(y)f^{\gamma}_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0\gamma}}}(y)}{\int f(x)f^{\gamma}_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0\gamma}}}(x)\mathrm{d}x} = \tilde{\pi}_0 \frac{f_{0,\boldsymbol{\theta}_{\mathbf{0}}}(y)f^{\gamma}_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0\gamma}}}(y)}{\int f_{0,\boldsymbol{\theta}_{\mathbf{0}}}(x)f^{\gamma}_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0\gamma}}}(x)\mathrm{d}x} + \tilde{\pi}_1 \frac{f_1(y)f^{\gamma}_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0\gamma}}}(y)}{\int f_1(x)f^{\gamma}_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0\gamma}}}(x)\mathrm{d}x} \\ &=: \tilde{\pi}_0 f_0^{(\tilde{w})}(y) + \tilde{\pi}_1 f_1^{(\tilde{w})}(y), \end{aligned}$$

where $\tilde{\pi}_0 = \frac{\pi_0}{\pi_0+\pi_1\delta_\gamma^{(\tilde{w})}}$, $\delta_\gamma^{(\tilde{w})} = \int \frac{f_1(x)}{f_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0\gamma}}}(x)} f_0^{(\tilde{w})}(x)\mathrm{d}x$ and $\tilde{\pi}_1 = 1-\tilde{\pi}_0$, which are in the similar notations as in (5) except the superscript is changed from (w_0) to $(\tilde{w})$. Consider the Gaussian mixture model, $f(y) = \pi_0\phi(y;\mu_0,\sigma_0^2) + \pi_1\phi(y;\mu_1,\sigma_1^2)$. The asymptotic estimates for (μ_0,σ_0^2) from minimizing (8) are

$$\begin{aligned} \tilde{\mu} &= \int x f^{(\tilde{w})}(y)\mathrm{d}y = \tilde{\pi}_0\tilde{\mu}_0 + \tilde{\pi}_1\tilde{\mu}_1, \\ \tilde{\sigma}^2 &= (1+\gamma)\int (y-\tilde{\mu})^2 f^{(\tilde{w})}(y)\mathrm{d}y = (1+\gamma)(\tilde{\pi}_0\tilde{\sigma}_0^2 + \tilde{\pi}_1\tilde{\sigma}_1^2 + \tilde{\pi}_0\tilde{\pi}_1(\tilde{\mu}_1-\tilde{\mu}_0)^2), \end{aligned}$$

where $(\tilde{\mu}_0,\tilde{\sigma}_0^2)$ are the parameters in $f_0^{(\tilde{w})}$ (density for $\mathcal{N}(\mu_0,\sigma_0^2)$) and $(\tilde{\mu}_1,\tilde{\sigma}_1^2)$ in $f_1^{(\tilde{w})}$ (density for $\mathcal{N}(\mu_1,\sigma_1^2)$). Their M-estimates are in (12) and (13). Based on some basic calculations,

$$\frac{\phi(x;\mu,\sigma^2)\phi^{\gamma}(x;\nu,\tau^2)}{\int \phi(y;\mu,\sigma^2)\phi^{\gamma}(y;\nu,\tau^2)\mathrm{d}y} = \phi\Big(x;\frac{\mu+\nu\frac{\gamma\sigma^2}{\tau^2}}{1+\frac{\gamma\sigma^2}{\tau^2}},\frac{\sigma^2}{1+\frac{\gamma\sigma^2}{\tau^2}}\Big), \tag{17}$$

and

$$\delta_\gamma^{(\tilde{w})} = \Big(\sqrt{\frac{\sigma_0^2}{\sigma_1^2}}\Big)^{\gamma}\sqrt{\frac{1+\gamma\frac{\sigma_0^2}{\tilde{\sigma}^2}}{1+\gamma\frac{\sigma_1^2}{\tilde{\sigma}^2}}}\exp\Big(-\frac{\gamma}{2}\Big(\frac{(\mu_1-\tilde{\mu})^2}{\tilde{\sigma}^2+\gamma\sigma_1^2} - \frac{(\mu_0-\tilde{\mu})^2}{\tilde{\sigma}^2+\gamma\sigma_0^2}\Big)\Big),$$

we get $(\tilde{\mu},\tilde{\sigma}^2)$ satisfying

$$\begin{aligned} \tilde{\mu} &= \frac{\mu_0 + r\mu_1}{1+r}, \\ \frac{\frac{\tilde{\sigma}^2}{\sigma_0^2}+\gamma}{1+\gamma} &= \tilde{\pi}_0\Big(1 + r\frac{\sigma_1^2}{\sigma_0^2} + \tilde{\pi}_1 a\frac{(\frac{b}{a}(\mu_1-\tilde{\mu}) - (\mu_0-\tilde{\mu}))^2}{\sigma_0^2}\Big), \end{aligned}$$

where $a = 1/(1+\gamma\frac{\sigma_0^2}{\tilde{\sigma}^2})$, $b = 1/(1+\gamma\frac{\sigma_1^2}{\tilde{\sigma}^2})$, and

$$r = \frac{\tilde{\pi}_1}{\tilde{\pi}_0}\frac{b}{a} = \frac{\pi_1}{\pi_0}\Big(\sqrt{\frac{\sigma_0^2}{\sigma_1^2}}\Big)^{\gamma}\Big(\frac{1+\gamma\frac{\sigma_0^2}{\tilde{\sigma}^2}}{1+\gamma\frac{\sigma_1^2}{\tilde{\sigma}^2}}\Big)^{3/2}\exp\Big(-\frac{\gamma}{2}\Big(\frac{(\mu_1-\tilde{\mu})^2}{\tilde{\sigma}^2+\gamma\sigma_1^2} - \frac{(\mu_0-\tilde{\mu})^2}{\tilde{\sigma}^2+\gamma\sigma_0^2}\Big)\Big).$$

Suppose $\tilde{\sigma}^2 = \sigma_0^2 = \sigma_1^2$ known. Increasing γ accelerates $\tilde{\mu}$ approaching to μ_0. The form of $\tilde{\sigma}^2$ is much complicated, which depends not only on γ but also how close of $\tilde{\mu}$ to μ_0 and the underlying parameters $(\mu_0,\mu_1,\sigma_0^2,\sigma_1^2,\pi_0)$. Hence, the latent bias is model dependent, relying on the choice of γ and usually hard to quantify.

From a classical point of view in robust estimation, we investigate the Influence Function (IF) to see how it varies under different γ's. Previous works also studied the IF of density weight divergence in (Basu et al., 1998; Jones et al., 2001), but here we more focus on how it is affected by γ. The IF for normal

mean estimate $\tilde{\mu}_{0\gamma}$ and the IF for normal variance estimate $\tilde{\sigma}^2_{0\gamma}$ from γ-robustifying procedure at normal distribution $\mathcal{N}(\mu_0, \sigma_0^2)$ are

$$IF_\gamma(y_1; \tilde{\mu}_{0\gamma}(\mathbf{y}), \phi(\cdot; \mu_0, \sigma_0^2)) = (y_1 - \mu_0)(1+\gamma)^{3/2} e^{-\frac{\gamma(y_1-\mu_0)^2}{2\sigma_0^2}},$$
$$IF_\gamma(y_1; \tilde{\sigma}^2_{0\gamma}(\mathbf{y}), \phi(\cdot; \mu_0, \sigma_0^2)) = \left(\frac{(y_1-\mu_0)^2}{\sigma_0^2} - \frac{1}{1+\gamma}\right)\sigma_0^2(1+\gamma)^{5/2} e^{-\frac{\gamma(y_1-\mu_0)^2}{2\sigma_0^2}},$$

where $\tilde{\mu}_{0\gamma}(\mathbf{y})$ and $\tilde{\sigma}^2_{0,\gamma}(\mathbf{y})$ are the minimizers defined in (10). The IF provides us a heuristic tool considering if there are an infinitesimal amount of outliers lying at point y_1, how the outliers affect the asymptotic bias of the estimate (Hampel et al., 2011). We demonstrate the IF curves at $\mathcal{N}(0,1)$ under various γ's in Figure 2. When $\gamma = 0$, both IFs of the mean estimate and the variance estimate are unbounded, while when $\gamma > 0$, the IFs are redescending and thus the estimates are robust. In the case that the outlier y_1 is more than 2 standard deviation away from the underlying mean zero, a larger γ has a better performance to down weight the outliers, whereas in the case that the outlier is near zero, a large γ can cause even bigger bias than that of MLE (corresponding to $\gamma = 0$). When $\gamma \geq 0.5$, the maximums of both the gross error sensitivities of the IF of mean estimate and the IF of variance estimate under various γ's in the comparison are at $\gamma = 3$. The IF approach only considers an infinitesimal amount of outliers, while in reality, there could be a heavy proportion of outliers in various magnitudes. Besides, there needs a balance to choose a γ to minimize the estimation error for the mean and also the error for the variance, which makes the problem of γ selection more difficult.

We demonstrate how γ balances bias and variance of the estimation in simulation studies. Consider $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} (1-\pi_1)\mathcal{N}(0,1) + \pi_1\mathcal{N}(\mu_1, 1)$, where $n = 200, 2000$ and under each n, consider $(\mu_1, \pi_1) = (0,0), (1, 0.1), (3, 0.1), (5, 0.1), (1, 0.3), (3, 0.3), (5, 0.3)$. The population parameters are set as $\mu_0 = 0, \sigma_0^2 = 1$. We evaluate the MSE for estimating μ_0 and σ_0 under $\gamma = 0, 0.5, 1, 2, 3$ from $B = 50$ realizations, i.e., $MSE = \frac{1}{B}\sum_{b=1}^{B} \|\hat{\boldsymbol{\theta}}^{(b)}_{\boldsymbol{0\gamma}} - \boldsymbol{\theta_0}\|_2^2 = \frac{1}{B}\sum_{b=1}^{B}(\hat{\mu}^{(b)}_{0\gamma} - \mu_0)^2 + (\hat{\sigma}^{(b)}_{0\gamma} - \sigma_0)^2$, where the superscript (b) represents the b^{th} realization. From Figure 3 — Figure 6, the MSE trends in the top row in each figure show that a too small γ leads a large bias and a too big γ leads large variance. Such trade-off phenomenon is very clear under $\mu_1 = 3, 5$ and $\pi_1 = 0.1, 0.3$ especially under large sample size, while under cases such as $\mu = 1$ and $\pi_1 = 0.1, 0.3$, the MSE trends are almost flat. In the simulations, each realization can vary a lot and also is affected by sample size. The estimation trends of $\hat{\mu}_0$, $\hat{\sigma}_0^2$ and $\hat{\pi}_0$ varying with γ are summarized in Figure S1 — Figure S4. From these simulation studies, we can see the performance of the estimation depends on γ, and the optimal γ, the turning point of the MSE curve, varies from case to case relying on the proportion and the magnitude of outliers and very data-dependent.

2.5 Selection criterion on γ

Windham (1995) provided a selection method for γ from a view of asymptotic efficiency, relating the convergence rate of the iterative algorithm to the asymptotic variance. However, it did not consider the latent bias from the true population parameter. The analysis in Fujisawa and Eguchi (2008) is based on the assumption that under some γ, the outliers lie in the tails of the population density raised to power γ. Under this assumption, they claimed the latent bias is small. However, this assumption is not guaranteed for each γ. As far as we know, how to select a "proper" γ is still unknown for the mixture model.

Recall that the γ-robustifying procedure actually approximates $f^{(w)}$ to be $f_0^{(w)}$. What we expect is to weight more on the population points while less on the outliers. However, only letting $f^{(w)}$ close to $f_0^{(w)}$ does not guarantee that this approximation is overall good for the population points. As we have seen in Subsection 2.4, there is a bias-variance trade-off controlled by γ. Since we are mainly interested in estimating the distribution of the population points, we consider a goodness-of-fit (GOF) measurement on how close of $f_0^{(w)}$ to $f^{(w)}$ in the probability measure of the population distribution and thus give an overall GOF on the

population points,

$$\int \frac{f_0^{(w)}(y)}{f^{(w)}(y)} \cdot f_{0,\boldsymbol{\theta}}(y)\mathrm{d}y = \int \left(\frac{\int f(x) f_{0,\boldsymbol{\theta}}^{\gamma}(x)\mathrm{d}x}{\int f_{0,\boldsymbol{\theta}}^{1+\gamma}(x)\mathrm{d}x} \right) \frac{f_{0,\boldsymbol{\theta}}(y)}{f(y)} f_{0,\boldsymbol{\theta}}(y)\mathrm{d}y \tag{18}$$

$$= \int \frac{\pi_0^{\ddagger}(\boldsymbol{\theta},\gamma) f_{0,\boldsymbol{\theta}}(y)}{f(y)} f_{0,\boldsymbol{\theta}}(y)\mathrm{d}y =: \mathbb{E}_{f_{0,\boldsymbol{\theta}}}\, fdr^{\ddagger}(Y;\gamma), \tag{19}$$

where $\pi_0^{\ddagger}(\boldsymbol{\theta},\gamma) := \pi_0^{\ddagger}$ defined in (16) and $fdr(y) := \mathbb{P}(y \text{ is from the null}|y) = \pi_0 f_0(y)/f(y)$ is the local false discovery rate introduced in Efron (2005). Here we replace the true π_0 by $\pi_0^{\ddagger}$ and add the same superscript $\ddagger$ on fdr. The oracle value for $\mathbb{E}_{f_{0,\boldsymbol{\theta}}}\, fdr^{\ddagger}(Y;\gamma)$ is the ordinary expected fdr under the null, i.e., $\mathbb{E}_{f_{0,\boldsymbol{\theta_0}}}\, fdr(Y)$ when $\pi_0^{\ddagger}(\boldsymbol{\theta_\gamma},\gamma) = \pi_0$ and $\boldsymbol{\theta_\gamma} = \boldsymbol{\theta_0}$.

Consider under some γ, $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}_{\boldsymbol{0\gamma}}$ which is the asymptotic limit of the M-estimate. We have $\mathbb{E}_{f_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0}}}}\, fdr^{\ddagger}(Y;\gamma) = \frac{\pi_0^{\ddagger}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}},\gamma)}{\pi_0} \cdot c(\tilde{\boldsymbol{\theta}}_{\mathbf{0}}, f)$, where

$$\frac{\pi_0^{\ddagger}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}},\gamma)}{\pi_0} = \int \frac{f_{0,\boldsymbol{\theta_0}}(y)}{f_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0}}}(y)} \left(\frac{f_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0}}}^{1+\gamma}(y)}{\int f_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0}}}^{1+\gamma}(x)\mathrm{d}x} \right) \mathrm{d}y + \frac{\pi_1}{\pi_0} \int \frac{f_1(y)}{f_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0}}}(y)} \left(\frac{f_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0}}}^{1+\gamma}(y)}{\int f_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0}}}^{1+\gamma}(x)\mathrm{d}x} \right) \mathrm{d}y,$$

$$c(\tilde{\boldsymbol{\theta}}_{\mathbf{0}}, f) = \int \frac{f_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0}}}(y)}{\frac{f_{0,\boldsymbol{\theta_0}}(y)}{f_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0}}}(y)} + \frac{\pi_1}{\pi_0} \frac{f_1(y)}{f_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0}}}(y)}} \mathrm{d}y.$$

If the γ is well selected such that $\tilde{\boldsymbol{\theta}}_{\boldsymbol{0\gamma}}$ close to $\boldsymbol{\theta_0}$ and $\pi_0^{\ddagger}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}},\gamma)$ close to π_0, then $\mathbb{E}_{f_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0}}}}\, fdr^{\ddagger}(Y;\gamma)$ will be close to the ordinary $\mathbb{E}_{f_{0,\boldsymbol{\theta_0}}}\, fdr(Y)$. For example, under the Gaussian mixture model $f(y) = \pi_0\phi(y;\mu_0,\sigma_0^2) + \pi_1\phi(y;\mu_1,\sigma_1^2)$, we have

$$\begin{aligned} \frac{\pi_0^{\ddagger}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}},\gamma)}{\pi_0} &= \int \frac{f_{0,\boldsymbol{\theta_0}}(y)}{f_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0}}}(y)} \phi(y;\tilde{\mu}_0, \frac{\tilde{\sigma}_0^2}{1+\gamma})\mathrm{d}y + \frac{\pi_1}{\pi_0} \int \frac{f_1(y)}{f_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0}}}(y)} \phi(y;\tilde{\mu}_0, \frac{\tilde{\sigma}_0^2}{1+\gamma})\mathrm{d}y \\ &= \sqrt{\frac{1+\gamma}{1+a_0\gamma}} \exp\left(-\frac{b_0\gamma}{2(1+a_0\gamma)} \right) + \frac{\pi_1}{\pi_0} \sqrt{\frac{1+\gamma}{1+a_1\gamma}} \exp\left(-\frac{b_1\gamma}{2(1+a_1\gamma)} \right), \end{aligned} \tag{20}$$

where $a_0 = \frac{\sigma_0^2}{\tilde{\sigma}_0^2}$, $b_0 = (\frac{\mu_0-\tilde{\mu}_0}{\tilde{\sigma}_0})^2$, $a_1 = \frac{\sigma_1^2}{\tilde{\sigma}_0^2}$, and $b_1 = (\frac{\mu_1-\tilde{\mu}_0}{\tilde{\sigma}_0})^2$. Under a proper γ, $\tilde{\mu}_0 \approx \mu_0$ and $\tilde{\sigma}_0^2 \approx \sigma_0^2$ so that $a_0 \approx 1$ and $b_0 \approx 0$, thus the first term in the Right Hand Side (RHS) of (20) is close to 1. Further if π_1 is much smaller than π_0 or $\tilde{\mu}_0$ is far away from μ_1 relative to the scale $\tilde{\sigma}_0$, i.e., b_1 is large, then the second term of the RHS of (20) is close to 0. Hence, under such γ, $\pi_0^{\ddagger}(\tilde{\boldsymbol{\theta}}_{\mathbf{0}},\gamma)/\pi_0 \approx 1$ and $c(\tilde{\boldsymbol{\theta}}_{\mathbf{0}}, f) \approx \mathbb{E}_{f_{0,\boldsymbol{\theta_0}}}\, fdr(Y;\gamma)$. In a special case that $\pi_1 = 0$, by Fisher consistency of $\tilde{\boldsymbol{\theta}}_{\mathbf{0}}$ from Proposition 1, for all $\gamma \geq 0$, $\tilde{\boldsymbol{\theta}}_{\boldsymbol{0\gamma}} = \boldsymbol{\theta_0}$ and thus $\mathbb{E}_{f_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0}}}}\, fdr^{\ddagger}(Y;\gamma)$ reaches to its oracle $\mathbb{E}_{f_{0,\boldsymbol{\theta_0}}}\, fdr(Y)$, which is one in this case. In reality, since we do not know the underlying mixture model, we would like to approximate the oracle $\mathbb{E}_{f_{0,\boldsymbol{\theta_0}}}\, fdr(Y)$ to be one, although this approximation may not be accurate in some cases where f_0 and f_1 are close to each other and the outlier proportion is not small. By comparing $\mathbb{E}_{f_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0}}}}\, fdr^{\ddagger}(Y;\gamma)$ to one tries to capture the latent bias hidden in $\tilde{\boldsymbol{\theta}}_{\boldsymbol{0\gamma}}$.

To estimate $\mathbb{E}_{f_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0}}}}\, fdr^{\ddagger}(Y;\gamma)$ in practice, since the expectation of fdr is with respect to the population, here we do not care too much of the tail effect when estimating the density functions. We discretize the integral into bins and estimate the density by its histogram in each bin. In details, we first partition the range of the observed y's to S non-overlap intervals, i.e., $[y_{(1)}, y_{(n)}] = \cup_{s=1}^{S} I_s$ where $y_{(i)}$ is the i^{th} smallest value, and $I_s \cap I_t = \emptyset$ for $s \neq t$. Define $|I_s| = \#\{y : y \in I_s\}$. We then merge the adjacent bins such that for all $|I_k|$ strictly positive. In one bin I_k, we approximate the integral of mixture density f over I_s by sample proportion $|I_s|/n =: \mathbb{P}_{\bar{f}}(I_s)$ and the integral of f_0 over I_s is $\int_{I_s} f_{0,\boldsymbol{\theta}}(y)\mathrm{d}y =: \mathbb{P}_{f_{0,\boldsymbol{\theta}}}(I_s)$. For the most left and right intervals, we extend them to $-\infty$ and $+\infty$, respectively, when calculating $\mathbb{P}_{f_{0,\theta}}$. Finally, we get an empirical estimate for $\mathbb{E}_{f_{0,\tilde{\boldsymbol{\theta}}_{\mathbf{0}}}}\, fdr^{\ddagger}(Y;\gamma)$ from

$$\hat{\mathbb{E}}_{f_{0,\hat{\boldsymbol{\theta}}_{\mathbf{0\gamma}}}} fdr^{\ddagger}(Y;\gamma) = \sum_{s=1}^{S} \frac{\hat{\pi}_0\, \mathbb{P}^2_{f_{0,\hat{\boldsymbol{\theta}}_{\mathbf{0\gamma}}}}(I_s)}{\mathbb{P}_{\bar{f}}(I_s)}. \tag{21}$$

Since this empirical estimate (21) depends on $\hat{\boldsymbol{\theta}}_{\mathbf{0}\boldsymbol{\gamma}}$, this term tries to capture the variation of $\hat{\boldsymbol{\theta}}_{\mathbf{0}\boldsymbol{\gamma}}$ under different γ's.

Our selection criterion is to select a γ minimizing the distance between the empirical estimate for $\mathbb{E}_{f_{0,\tilde{\boldsymbol{\theta}}_\mathbf{0}}} fdr^\ddagger$ and the approximation one for the oracle expectation, i.e.,

$$\hat{\gamma}^* = \arg\min_{\gamma} |\hat{\mathbb{E}}_{f_{0,\hat{\boldsymbol{\theta}}_{\mathbf{0}\boldsymbol{\gamma}}}} fdr^\ddagger(Y;\gamma) - 1|, \tag{22}$$

where $\hat{\boldsymbol{\theta}}_\mathbf{0}$ is our robust estimate from (11). In the ideal case, the optimal γ^* is the minimizer of $\mathbb{E}_f$ MSE$(\hat{\boldsymbol{\theta}}_{\mathbf{0}\boldsymbol{\gamma}})(= \mathbb{E}_f \|\hat{\boldsymbol{\theta}}_{\mathbf{0}\boldsymbol{\gamma}} - \boldsymbol{\theta}_\mathbf{0}\|_2^2)$. But it is impractical since MSE depends on unknown $\boldsymbol{\theta}_\mathbf{0}$. Our contribution is to transfer the inestimable criterion to an estimable criterion in order to select a proper γ in practice. Observe that

$$\begin{aligned}
&|\hat{\mathbb{E}}_{f_{0,\hat{\boldsymbol{\theta}}_\mathbf{0}}} fdr^\ddagger(Y;\gamma) - 1| \\
&\le |\hat{\mathbb{E}}_{f_{0,\hat{\boldsymbol{\theta}}_\mathbf{0}}} fdr^\ddagger(Y;\gamma) - \mathbb{E}_{f_{0,\tilde{\boldsymbol{\theta}}_\mathbf{0}}} fdr^\ddagger(Y;\gamma)| + |\mathbb{E}_{f_{0,\tilde{\boldsymbol{\theta}}_\mathbf{0}}} fdr^\ddagger(Y;\gamma) - \mathbb{E}_{f_{0,\boldsymbol{\theta}_\mathbf{0}}} fdr(Y)| + |\mathbb{E}_{f_{0,\boldsymbol{\theta}_\mathbf{0}}} fdr(Y) - 1|.
\end{aligned} \tag{23}$$

We expect around the optimal γ^*, the first term in the RHS of (23), which controls the variance, is increasing with γ, and the second term in the RHS, which controls the bias, is decreasing with γ. Hence, there would be a turning point, which indicates the occurring of a proper γ. The third term in the RHS is the unavoidable bias from the selection procedure. Since the estimate in (21) not only depends on the estimate for $\boldsymbol{\theta}_\mathbf{0}$ and π_0, but also on the estimate for the density ratio, due to its complicated form, we study the behavior of our selection trend via simulations. In the same setting as in the simulations in Subsection 2.4, we compare the inestimable Squared Error (SE) trend SE$(\hat{\boldsymbol{\theta}}_{\mathbf{0}\boldsymbol{\gamma}})$ versus our practical selection trend $|\hat{\mathbb{E}}_{f_{0,\hat{\boldsymbol{\theta}}_{\mathbf{0}\boldsymbol{\gamma}}}} fdr^\ddagger(Y;\gamma) - 1|$ across various γ's in Figure 3 – Figure 6. Overall, comparing the average performances, our selection trends have similar capability in average performance as the MSE trends to capture the minimizer γ under various settings and under both small and large sample sizes. In the case that there is no outlier ($\mu_1 = 0, \pi_1 = 0$), both MSE trend and the average selection trend increase in γ. In the case of well-separated mixture densities under ($\mu_1 = 5, \pi_1 = 0.1$), there is a clear turning point at $\gamma = 1$. In the case of hardly separated mixture densities under ($\mu_1 = 1, \pi_1 = 0.3$), where the oracle $\mathbb{E}_0\, fdr$ is around 0.76, although approximating $\mathbb{E}_0\, fdr$ to be 1 is not quite accurate, the SE trends and our selection trends are almost flat in average. In the cases of moderately hard separated mixture densities under ($\mu_1 = 3, 5, \pi_1 = 0.3$), our average selection trend can have two local minimums, one in a large γ and another one near $\gamma = 0$. When we look at the $\hat{\sigma}_0$ under different γ's in Figure S2 and Figure S4, the trends first increase from $\gamma = 0$ then decrease, indicating our selection trends are more likely affected by $\hat{\sigma}_0$ in these cases. And under ($\mu_1 = 3, \pi_1 = 0.3$) especially in small sample size, the realization curves have large variations.

(Huber, 2011) pointed that the redescending M-estimate could have multiple minima and can be locally trapped. When there are multiple comparable and well-separated density bumps, our selection may pick up the wrong bump as the population especially in the small sample size, but such case is easy to detect. In the real application, we add one diagnosis step on the fitted residuals to flag the possible problematic fitting. We detail it in Subsection 2.6.

2.6 Data-adaptive algorithm in linear regression

So far, we consider the response variable $\mathbf{y}$ only from one group, mainly affected by one covariate variable as in (3), i.e.,

$$y_i \overset{\text{iid}}{\sim} \pi_0 \mathcal{N}(\mu_0, \sigma_0^2) + \pi_1 F_1.$$

Now we go back to the general setting in linear regression (1)-(2), i.e.,

$$y_i \sim \pi_0 \mathcal{N}(\mathbf{x_i}^\top \boldsymbol{\beta}_\mathbf{0}, \sigma_{0k(i)}^2) + \pi_1 F_{1i}(y_i), \quad i = 1, \ldots, n, \tag{24}$$

We still assume independence among y_i's but now y_i comes from different populations since $(\mathbf{x_i}^\top \boldsymbol{\beta}_\mathbf{0}, \sigma_{0k(i)}^2)$ varies with i. Hence, we need first to construct the null samples from the same distribution by transforming y_i to a standardized expression, Consider the standardized residual for the i^{th} sample,

$$r_i = \frac{y_i - \mathbf{x_i}^\top \boldsymbol{\beta}_\mathbf{0}}{\sigma_{0k(i)}}, \tag{25}$$

where $\mathbf{x_i^\top}$ is the i^{th} row of matrix $\mathbf{X}$ and the index $k(i)$ indicates the group of the i^{th} sample. Under the mixture model of $\mathbf{y}$, the standardized residuals are i.i.d. from the mixture model

$$r_i \overset{\text{iid}}{\sim} (1-\pi_1)\mathcal{N}(0,1) + \pi_1 F_1, \quad i = 1, \ldots, n.$$

To get empirical cross entropy between $\bar{f}^{(w)}$ and $f^{(w)}_{0,\boldsymbol{\theta}}$ defined in (9) on sample y_i's, by changing variables from the residuals to $\mathbf{y}$, we have

$$\bar{f}^{(w)}_r(r_i)\mathrm{d}r_i = w_i = w_i(\mathbf{r}) = \frac{\phi^\gamma(y_i; \mathbf{x_i^\top}\boldsymbol{\beta_0}, \sigma^2_{0k(i)})}{\sum_{j=1}^n \phi^\gamma(y_j; \mathbf{x_j^\top}\boldsymbol{\beta_0}, \sigma^2_{0k(j)})} = w_i(\mathbf{y}) = \bar{f}^{(w)}(y_i)\mathrm{d}y_i,$$

and

$$f^{(w)}_{r0}(r_i)\mathrm{d}r_i = \frac{1}{\sigma_{0k(i)}} f^{(w)}_{r0}\left(\frac{y_i - \mathbf{x_i^\top}\boldsymbol{\beta_0}}{\sigma_{0k(i)}}\right)\mathrm{d}y_i = \phi(y_i; \mathbf{x_i^\top}\boldsymbol{\beta_0}, \sigma^2_{0k(i)}) = f^{(w)}_0(y_i)\mathrm{d}y_i.$$

Hence, in linear regression problem, our robust likelihood criterion in terms of cross entropy on the weighted samples is

$$\begin{aligned}
& d_w(\bar{f}^{(w)}(\mathbf{y}), f^{(w)}_0(\mathbf{y}); \boldsymbol{\theta}, \gamma) \\
= & -\sum_{i=1}^n w_i \log\phi\left(y_i; \mathbf{x_i^\top}\boldsymbol{\beta_0}, \frac{\sigma^2_{0k(i)}}{1+\gamma}\right) \\
= & -\sum_{k=1}^K \sum_{j=1}^{n_k} w_{(k,j)} \log\phi\left(y_{(k,j)}; \mathbf{x^\top_{(k,j)}}\boldsymbol{\beta_0}, \frac{\sigma^2_{0k}}{1+\gamma}\right) \\
= & \frac{1}{2}\left(\sum_{k=1}^K (\sum_{j=1}^{n_k} w_{(k,j)}) \log(\sigma^2_{0k}) + \sum_{k=1}^K \frac{\sum_{j=1}^{n_k} w_{(k,j)}(y_{(k,j)} - \mathbf{x^\top_{(k,j)}}\boldsymbol{\beta})^2}{\sigma^2_{0k}/(1+\gamma)}\right) + \text{constant},
\end{aligned}$$

where the pair index (k, j) indicates the j^{th} sample in group k, $k = 1, \ldots, K$ and $j = 1, \ldots, n_K$. Therefore, given γ, the M-estimate for $\theta_\mathbf{0}$ is

$$\hat{\boldsymbol{\theta}}_\mathbf{0} = \arg\min_{\boldsymbol{\theta}} d_w(\bar{f}^{(w)}(\mathbf{y}), f^{(w)}_0(\mathbf{y}); \boldsymbol{\theta}, \gamma).$$

The minimizer can be found by iterating the weights and the estimate for $\boldsymbol{\theta_0}$. After updating the standardized residuals based on the estimate for $\boldsymbol{\theta_0}$ in each iteration, we apply our data-adaptive γ selection procedure of **Algorithm 1** to select a proper γ to further optimize the weights. The whole procedure is summarized in **Algorithm 2**.

3 Simulation Studies and real data application

In this section, we apply our algorithm AdaReg in the simulation studies to compare with other robust procedures then apply it in the real dataset of heart samples from RNA-seq in GTEx project.

3.1 Comparisons of robust procedures in simulation studies

In the simulation studies, we compare the performances of several robust estimation methods in a series of settings to evaluate their estimation accuracy. The methods under comparisons include the OLS ($\gamma = 0$), γ-robustifying methods under $\gamma = 0.5, 1, 2, 3$ and our data-adaptive robustifying procedure (AdaReg), other M-estimation methods based on Huber's weight (Huber, 2011), Hampel's weight (Hampel et al., 2011), Tukey's bisquare weight, and S-estimation (Rousseeuw and Yohai, 1984), and resistant robust methods including the least median squares (lms) (Rousseeuw, 1984), and the least trimmed squares (lts) (Rousseeuw, 1984, 1985). The OLS is implemented by function $lm(\cdot)$ in R (Team et al., 2013). The fixed γ and data-adaptive robustifying procedures are from **Algorithm 1-2**. The other M-estimations are from function $rlm(\cdot)$ in

Algorithm 1: Data-adaptive γ selection procedure

Data: Sample residuals $r_1, r_2, \ldots, r_n$, error tolerance ϵ (default 10^{-4}), step index $t = 1$.
Result: Population parameters $(\hat{\pi}_0(\mathbf{r}), \hat{\boldsymbol{\theta}}_{\mathbf{0}}(\mathbf{r}))$ where $\hat{\boldsymbol{\theta}}_{\mathbf{0}}(\mathbf{r}) = (\hat{\mu}_0, \hat{\sigma}_0^2)$.

Initialization: take sample mean and sample variance of r_i's for $\theta_0^{(0)}$.
forall γ *in the sequence* Γ *form* 0 *to* 3 *with increment* 0.1 **do**
  while $\|\boldsymbol{\theta}_{\mathbf{0}}^{(\mathbf{t})} - \boldsymbol{\theta}_{\mathbf{0}}^{(\mathbf{t-1})}\|_1 \geq \epsilon$ *or* $t < 50$ **do**
    Update $\mathbf{w}^{(t)}$ from ${\boldsymbol{\theta}_{\mathbf{0}}}^{(t-1)}$ based on (7) ;
    Update ${\boldsymbol{\theta}_{\mathbf{0}}}^{(t+1)}$ from $\mathbf{w}^{(t)}$ based on (12)-(13) ;
    $t = t + 1$;
  end
  Get fixed point $\hat{\boldsymbol{\theta}}_{\mathbf{0}}(\gamma)$ under γ ;
  Calculate $\hat{\pi}_0(\gamma)$ from $\hat{\boldsymbol{\theta}}_{\mathbf{0}}(\gamma)$ based on (16) ;
  Evaluate $\hat{\mathbb{E}}_{f_{0,\hat{\boldsymbol{\theta}}_{\mathbf{0}}\boldsymbol{\gamma}}} f dr^{\ddagger}(r;\gamma)$ from $(\hat{\pi}_0(\gamma), \hat{\boldsymbol{\theta}}_{\mathbf{0}}(\gamma))$ based on (21);
end
Report: the selected γ from $\hat{\gamma}^* = \arg\min_{\gamma\in\Gamma} |\hat{\mathbb{E}}_{f_{0,\hat{\boldsymbol{\theta}}_{\mathbf{0}}\boldsymbol{\gamma}}} f dr^{\ddagger}(r;\gamma) - 1|$ and report $(\hat{\pi}_0(\hat{\gamma}^*), \hat{\boldsymbol{\theta}}_{\mathbf{0}}(\hat{\gamma}^*))$.

Algorithm 2: Data-adaptive robust estimation in linear regression

Data: Response variable $\mathbf{y}$, design matrix $\mathbf{X}$, group indexes, and error tolerance ϵ (default 10^{-4}), step index $t = 1$.
Result: Robust estimate for $\boldsymbol{\theta}_{\mathbf{0}} = (\boldsymbol{\beta}_{\mathbf{0}}, \sigma_{01}^2, \ldots, \sigma_{0K}^2)$ and residual fitting information.

Initialization: Get ${\boldsymbol{\beta}_{\mathbf{0}}}^{(0)} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ from OLS, obtain standardized residuals $\mathbf{r}^{(0)}$ from (25), and initialize $(\sigma_{0k}^2)^{(0)} = \mathrm{Var}({\mathbf{r}_{(\mathbf{k}.)}}^{(0)})$ from the sample residuals in group k.
while $\|{\theta_{\mathbf{0}}}^{(t)} - {\theta_{\mathbf{0}}}^{(t-1)}\|_1 \geq \epsilon$ *or* $t < 50$ **do**
  Apply data-adaptive γ-selection procedure on $\mathbf{r}^{(t-1)}$ to obtain the selected $\gamma^{*(t-1)}$ from **Algorithm 1**;
  Obtain the diagonal weight matrix $\mathbf{W}^{(t-1)}$ with diagonal element $w_i(\mathbf{r}^{(t-1)}; \gamma^{*(t-1)})/(\sigma_{0k(i)}^2)^{(t-1)}$;
  Update ${\boldsymbol{\beta}_{\mathbf{0}}}^{(t)} = (\mathbf{X}^\top\mathbf{W}^{(t-1)}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}^{(t-1)}\mathbf{y}$;
  Update $(\sigma_{0k}^2)^{(t)} = (1+\gamma)(\mathbf{y}_{(\mathbf{k}.)} - \mathbf{X}{\boldsymbol{\beta}_{\mathbf{0}}}^{(t)})^\top\mathbf{W}_{(\mathbf{k},.)}(\mathbf{y}_{(\mathbf{k}.)} - \mathbf{X}{\boldsymbol{\beta}_{\mathbf{0}}}^{(t)})$, where $\mathbf{W}_{(\mathbf{k},.)} = \mathrm{diag}(\frac{w_{(k,1)}}{\sum_{j=1}^{n_k} w_{(k,j)}}, \ldots, \frac{w_{(k,n_k)}}{\sum_{j=1}^{n_k} w_{(k,j)}})$;
  Update the residuals $\mathbf{r}^{(t)}$ based on ${\boldsymbol{\theta}_{\mathbf{0}}}^{(t)}$;
  $t = t + 1$;
end
Diagnosis: After iterations, check whether the absolute value of the robustly fitted mean or the estimated mode from kernel density fitting on the standardized residuals is ≥ 1. If it is true, we flag it as "need further verify".
Report: $\hat{\boldsymbol{\theta}}_{\mathbf{0}}$ and robust mean, variance and outlier proportion of the residuals, and diagnosis result.

package "MASS" (Venables and Ripley, 2013). The lms, lts and S estimation are from function $lqs(\cdot)$ in package "WRS2" (Mair and Wilcox, 2016).

We consider three cases: (i) a small sample size and small dimension case of $n = 200$ and $p = 2$ (including the intercept), (ii) a moderately large sample size and small dimension case of $n = 2000$ and $p = 2$, and (iii) a moderately large sample size and moderately large dimension case of $n = 2000$ and $p = 20$. In each case, we set $\boldsymbol{\beta_0} = \mathbf{1_{p\times 1}}$. For the design matrix $\mathbf{X}$, its first column is 1's and we generate its other elements i.i.d. from $\mathcal{N}(0, 10^2)$. We consider the elements of the mixture noise $\boldsymbol{\epsilon}$ i.i.d. from $(1-\pi_1)\mathcal{N}(0,1) + \pi_1\mathcal{N}(\mu_1, 1)$ and set (π_1, μ_1) from the grid $(0.1, 0.2, \ldots, 0.4) \times (1, 2, \ldots, 10)$ and $(0, 0)$. In each case, we evaluate the square root of the mean square error (RMSE) for $\boldsymbol{\beta_0}$ and σ_0^2 for each method from $B = 100$ independent repeated procedures, where RMSE $= \sqrt{\text{MSE}} = \sqrt{\frac{1}{B}\sum_{b=1}^{B} \|\hat{\boldsymbol{\theta}}^{(\boldsymbol{b})} - \boldsymbol{\theta_0}\|_2^2}$. The results are summarized in Figure S5-Figure S6 for case (i) $(n = 200, p = 2)$, Figure S8-Figure S9 for case (ii) $(n = 2000, p = 2)$, Figure S11-Figure S12 for case (iii) $(n = 2000, p = 20)$. We further investigate method detection ability when $\pi_1 > 0$ by evaluating the True Positive Rate (TPR, the proportion of the rejections over the true positives) and the False Positive Rate (FPR, the proportion of the non-rejections over the true negatives). We reject a sample if its standardized residual score $\frac{|y_i - \mathbf{x_i}^\top \hat{\boldsymbol{\beta}}|}{\hat{\sigma}} \geq 2.5$. We report the average TPR and average FPR for each method under each case from 100 independent repeated procedures summarized in Figure S7 for case (i) $(n = 200, p = 2)$, Figure S10 for case (ii) $(n = 2000, p = 2)$, and Figure S13 for case (iii) $(n = 2000, p = 20)$. For the cases of large proportion of outliers under $\pi_1 = 0.4$ and large $\mu_1 \geq 3$, it could happen that our algorithm picks the wrong bump as the population distribution, especially in small sample size as mentioned in Subsection 2.5. Hence, we filter the cases that the absolute value of the fitted mean for the standardized residuals ≥ 1 for AdaReg in the comparisons under the setting of $n = 200$, in total less than 20 realizations in the simulations.

To compare the estimations for $\boldsymbol{\beta_0}$ from Figure S5 ($n = 200$, $p = 2$), Figure S8 ($n = 2000$, $p = 2$) and Figure S11 ($n = 2000$, $p = 20$), in each case we can see the fixed γ-robustifying procedure from $\gamma = 0$ (OLS) to $\gamma = 3$ gradually perform better and better, that is, the high RMSE region is smaller and smaller. The Huber and Hampel weighted robust methods perform worse than the γ-robustifying procedures. The Tukey's bisquare estimation has similar performance as the $\gamma = 0.5$ robustifying procedure. The Rousseeuw's methods (lms, lts, S estimation) perform similar to the large $\gamma \geq 2$ robustifying procedure in the low dimension case $(p = 2)$ but become worse in the case of moderately high dimension case $(p = 20)$. Our AdaReg performs better than the fixed γ-procedure overall, slightly worse in the situations of the null and the alternative hardly separable such as in $(\pi_1, \mu_1) = (0.3, 2), (0.4, 3)$. To compare the convergence rate, the small $\gamma(\leq 1)$ procedures and Huber, Tukey and Hampel's estimations have similar performances from case $(n = 200, p = 2)$ to $(n = 2000, p = 2)$, respectively, while for the large $\gamma(> 1)$ and data-adaptive robustifying procedures have better performances when sample size increases, which indicates that their estimation convergence rates are relatively slow. To compare the dimension effect of the explanatory variables from case $(n = 2000, p = 2)$ to $(n = 2000, p = 20)$, the number of explanatory variables does not affect much for γ-robustifying procedures, Huber, Tukey and Hampel methods, since they down weight the outliers in the residuals which is essentially in one dimension. Although the S-estimator performs well in low dimension, it loses its advantage in high dimension. However, for Rousseeuw's methods (lms, lts), they search the population points in the $(1 + p)$-dimensional space, and thus high dimension dramatically diminishes their performances.

We also compare the estimations for population variance σ_0^2 of the noise and outlier detection abilities, summarized in Figure S6-Figure S7 (for $n = 200$, $p = 2$), Figure S9-Figure S10 (for $n = 2000$, $p = 2$) and Figure S12-Figure S13 (for $n = 2000$, $p = 20$). To robustly estimate the population variance fo the noise is important for the outlier detection. An upwards biased estimate can lead high false negative rate and a downwards biased estimate can have high false positive rate. From the simulations, the $\gamma(> 0)$ robustitfying procedures perform better than the traditional robust estimation methods to estimate σ_0. In the case of $(n = 2000, p = 20)$, when $\gamma = 3$, the variance estimator is more likely to be affected by the robustness of $\hat{\boldsymbol{\beta}}_{\boldsymbol{0}}$ in high dimensions. For a fixed γ-robustifying procedure, it can be easily locally trapped. Our AdaReg shows its advantage in all cases to estimate the residual variance. From the plots of FPR versus TPR, our AdaReg preserves low FPR well in the hard detectable cases when $\mu_1 \leq 3$ and achieves high TPR in the relatively easily detectable cases when $\mu_1 > 3$ over the other methods, no matter the sample size is large or small, the dimension is high or low, in our simulation studies.

3.2 Real Data Application in GTEx

We apply our AdaReg on the human gene expression data from RNA-seq in the GTEx project (Consortium et al., 2015) in version 7 (data downloaded from GTEx Portal). There are in total 600 heart samples from atrial appendage (297 samples) and left ventricle (303 samples). We consider the protein-coding genes whose either atrial sample median or left ventricle sample median is > 1 inTPM, in total $12,422$ genes. Here, we take logarithm transformation on $(\text{TPM} + 1)$ making the data more symmetric.

We assume the expressions for each gene are independently from a mixture model $Y_{ij} \sim \pi_{01}\mathcal{N}(\mu_{01}, \sigma_{01}^2) + \pi_{02}\mathcal{N}(\mu_{02}, \sigma_{02}^2) + \pi_1 F_1$, $j = 1, \ldots, 600$, where $\pi_1 = 1 - \pi_{01} - \pi_{02}$. The population distribution for heart atrial appendage is $\mathcal{N}(\mu_{01}, \sigma_{01}^2)$ and for heart left ventricle is $\mathcal{N}(\mu_{02}, \sigma_{02}^2)$. The outlier proportion is π_1 and the outliers come from unknown distribution F_1. We are interested in estimating $(\mu_{01}, \sigma_{01}^2, \mu_{02}, \sigma_{02}^2, \pi_1)$ under the presence of outliers in the gene expressions. We apply a series γ-robustifying procedures under $\gamma = 0, 0.5, 1, 2, 3$ and our data-adaptive procedure. Since from the simulation studies, the other methods cannot adapt to various settings, we do not compare them in this subsection.

Figure 7 shows the densities of $\hat{\pi}_1$ from the series γ-robustifying procedures under $\gamma = 0.5, 1, 2, 3$ and data-adaptively selected γ. Under $\gamma = 0$, which is OLS, since in this method $\hat{\pi}_1 = 0$, we do not show it in the density of plot of $\hat{\pi}_1$. We can see under $\gamma = 0.5$, the outlier proportions are more concentrated near 0, while under $\gamma = 2, 3$, the procedures claim more outliers. The density of $\hat{\pi}_1$ from AdaReg is comparable to the part of the density under $\gamma = 1$ in the estimated small outlier proportion region then approaches to the tail of the density under $\gamma = 2$ in the estimated high outlier proportion region, which indicates our data-adaptive procedure really tries to adapt to different outlier scenarios.

In practice, it may be hard to check normality of the residuals without knowing which expression is an outlier. We here consider local property of the residuals. We estimate the density mode from kernel density fitting on the residuals for each estimation procedure. Suppose a procedure can estimate the population distribution well in each heart group then the standardized residuals should have a density peak around zero, as in the illustration example for gene MYH7 in Section 1. Figure 8 shows the scatter plot of $\hat{\pi}_1$ versus estimated standardized residual mode under each γ-robustifying procedure. Under $\gamma = 0$, $\hat{\pi}_1 = 0$, where we extend the values of the modes to a band in the figure. We can see under small $\gamma = 0, 0.5, 1$, there are a bunch of genes having positive residual modes away from zero, which indicates there is still some information left in the residuals, while under $\gamma = 2, 3$ and data-adaptively selected γ, the residual modes are more symmetric around zero. The residual modes from AdaReg are more concentrated at zero in the estimated small outlier proportion region (the red region), the same as under $\gamma = 0.5, 1$, and it still gathers majority of the residual modes being around zero in the estimated large outlier proportion region, the same as under $\gamma = 2, 3$. We pick the genes in the sparse point region (the blue region in Figure 8) as unfitted points. They may come from the result of being locally trapped. We put them aside for further analysis. In the following differential expression analysis, we do not include those unfitted genes. Note that checking from the residual modes does not give information on the normality region, which is harder to evaluate without prior knowledge. Here we only give some criteria for the not-well-fitted point.

We further investigate differentially expressed genes and enrichment analysis of GO terms (Consortium, 2014). From the γ robustifying procedures, we define the samples whose absolute values of the standardized residuals ≥ 2.5 as outliers then apply the two-sample t-test on the filtered samples after removing the outliers for each gene. Figure S14 shows the volcano plot of fold changes in log scale versus p-values from the t-test. We take the hyperbolic curve with curvature parameter 100 and minimum fold change parameter 1 as significance thresholds (Singh et al., 2016). The numbers of positive(negative) significant genes are comparable among all the procedures. In the GO term analysis from "clusterprofiler" R package (Yu et al., 2012), we combine the significant gene lists under $\gamma = 0, 0.5, 1$ as the results from small γ and combine the lists under $\gamma = 2, 3$ as the results from large γ. Figure 9 and Figure 10 show the significant enriched GO terms under threshold 10^{-4} on the adjusted p-values from the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). For the genes highly expressed in left ventricle than in atrial appendage, they are highly enriched in ventricular cardiac muscle tissue functions, as expected. There the p-values from our AdaReg are comparable to the procedures under large γ and they are more significant than the results under small γ procedures. For the genes highly expressed in atrial appendage than in left ventricle, they are highly enriched in extracellular related functions. Again, the results from our AdaReg show the most significance.

4 Discussion and conclusion

In the large-scale analysis of gene expressions, outliers vary from gene to gene. Given the presence of heterogenous outliers, it is important to have a robust estimation method to adapt to various genes. We followed the approach of the robust estimation based on γ-density-power-weight. Taking advantage that its estimation is only controlled by one tuning parameter γ, we developed a novel γ selection procedure to achieve the goal of data-adaptive estimation. We provided a heuristic analysis on the selection criterion, and found that our selection trends under various γ's have similar capability to capture minimizer γ in average performance as the MSE trends from our simulations under a series of settings. Our data-adaptive robustifying procedure shows its advantage in both simulation studies and real data application compared to the fixed γ procedure and other robust methods in the setting of linear regression. However, there are still some limitations and further work to do.

Assumption of the population distribution. In our setting, we mainly focused on the Gaussian population distribution. It is easy to generalize to other distributions like Poisson distribution, or Gamma distribution for different applications. Our robust criterion and selection procedure can be adapted easily to other settings. However, the approach of density-power-weight robust estimation requires the parametric assumption on the population.

Outliers in the design matrix. In our setting, we only considered the outliers in the response variable. If there are outliers in the design matrix, the density-power-weight-based methods will lose their power. However, the resistant methods like the least median squares are still robust.

Theoretical analysis of the selection criterion. To analyze the performance of our selection criterion, we provided a heuristic analysis of the selection criterion and compared our selection trend to the trend of MSE under various of γ's in the simulation studies. We do not give a theoretical error bound for now because of the complicated form of the criterion, which depends not only on the parameter estimates but also the estimate for the density ratio. We think the error bound, either determinant or probabilistic, still depends on the underlying distribution and needs more theoretical work in the future.

Other Selection criteria. The classical model selection criteria like Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) require that their sample distribution is known and thus do not work in our mixture setting with unknown outlier distribution. One may think about selecting γ from the diagnosis on the residuals. However, if the population is fitted wrong, the residuals could mask the true signals and thus could not give a fair selection on γ. This paper provides one way to select γ, but we look forward to other researchers developing other procedures or criteria to select γ.

In conclusion, robust estimation based on γ-density-power-weight is an interesting and important approach. We proposed one data-adaptive robustifying procedure: automatically selecting a proper γ. We believe that combining density-power-weight robust estimation with a data-adaptive γ selection procedure will be applicable to more situations involving the detection of differential gene expression, and the identification of tissue-specific genes.

Acknowledgments

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The gene expression TPM data used for the analyses described in this manuscript were obtained from GTEx Portal in version 7. We acknowledge the discussions with Dr. Hua Tang at Stanford. We would like to thank the funding supports by GTEx grant (5U01HL13104203) and CEGS grant (2RM1HG00773506).

References

Basu, A., I. R. Harris, N. L. Hjort, and M. Jones (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika 85*(3), 549–559.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological) 57*(1), 289–300.

Chen, T.-L., D.-N. Hsieh, H. Hung, I.-P. Tu, P.-S. Wu, Y.-M. Wu, W.-H. Chang, S.-Y. Huang, et al. (2014). gamma-sup: A clustering algorithm for cryo-electron microscopy images of asymmetric particles. *The Annals of Applied Statistics 8*(1), 259–285.

Consortium, G. et al. (2015). The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science 348*(6235), 648–660.

Consortium, G. O. (2014). Gene ontology consortium: going forward. *Nucleic acids research 43*(D1), D1049–D1056.

Efron, B. (2005). *Local false discovery rates*. Division of Biostatistics, Stanford University.

Fujisawa, H. (2013). Normalized estimating equation for robust parameter estimation. *Electronic Journal of Statistics 7*, 1587–1606.

Fujisawa, H. and S. Eguchi (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis 99*(9), 2053–2081.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (2011). *Robust statistics: the approach based on influence functions*, Volume 114. John Wiley & Sons.

Huber, P. J. (1964, 03). Robust estimation of a location parameter. *Ann. Math. Statist. 35*(1), 73–101.

Huber, P. J. (2011). *Robust statistics*. Springer.

Jones, M., N. L. Hjort, I. R. Harris, and A. Basu (2001). A comparison of related density-based minimum divergence estimators. *Biometrika 88*(3), 865–873.

Kanamori, T. and H. Fujisawa (2015). Robust estimation under heavy contamination using unnormalized models. *Biometrika*, asv014.

Katayama, S., H. Fujisawa, and M. Drton (2018). Robust and sparse gaussian graphical modelling under cell-wise contamination. *Stat 7*(1), e181.

Mair, P. and R. Wilcox (2016). Robust statistical methods in r using the wrs2 package. *Harvard Univ.*

Maronna, R. A., R. D. Martin, V. J. Yohai, and M. Salibián-Barrera (2018). *Robust statistics: theory and methods (with R)*. Wiley.

Miyamura, M. and Y. Kano (2006). Robust gaussian graphical modeling. *Journal of Multivariate Analysis 97*(7), 1525–1550.

Rousseeuw, P. and V. Yohai (1984). Robust regression by means of s-estimators. In *Robust and nonlinear time series analysis*, pp. 256–272. Springer.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American statistical association 79*(388), 871–880.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical statistics and applications 8*(283-297), 37.

Rousseeuw, P. J. and A. M. Leroy (1987). *Robust regression and outlier detection*, Volume 1. Wiley Online Library.

Singh, S., M. Y. Hein, and A. F. Stewart (2016). msvolcano: A flexible web application for visualizing quantitative proteomics data. *Proteomics 16*(18), 2491–2494.

Team, R. C. et al. (2013). R: A language and environment for statistical computing.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

Venables, W. N. and B. D. Ripley (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

Windham, M. P. (1995). Robustifying model fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 599–609.

Yu, G., L.-G. Wang, Y. Han, and Q.-Y. He (2012). clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology 16*(5), 284–287.

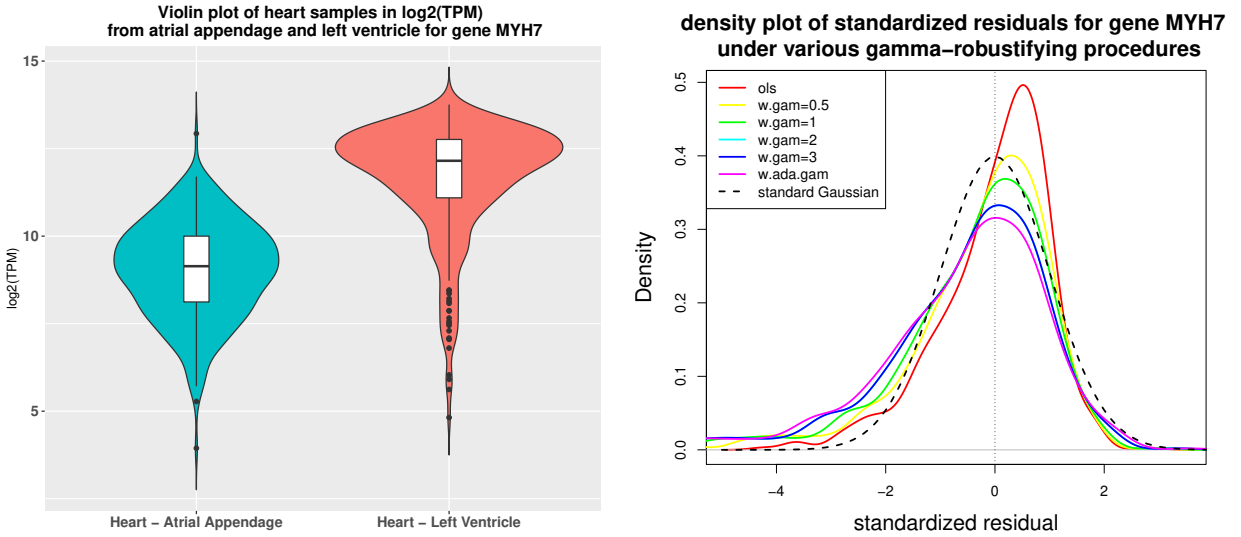

Figure 1: Motivation example of heart samples from atrial appendage and left vertical for gene MYH7 in GTEx RNA-seq data. The left panel is the violin plot of the heart expression in log TPM in two heart groups. The right panel is the density plot of standardized residuals (defined in (25)) from various γ-robustifying procedures and the dashed black curve is the standard Gaussian density without containing any outliers.

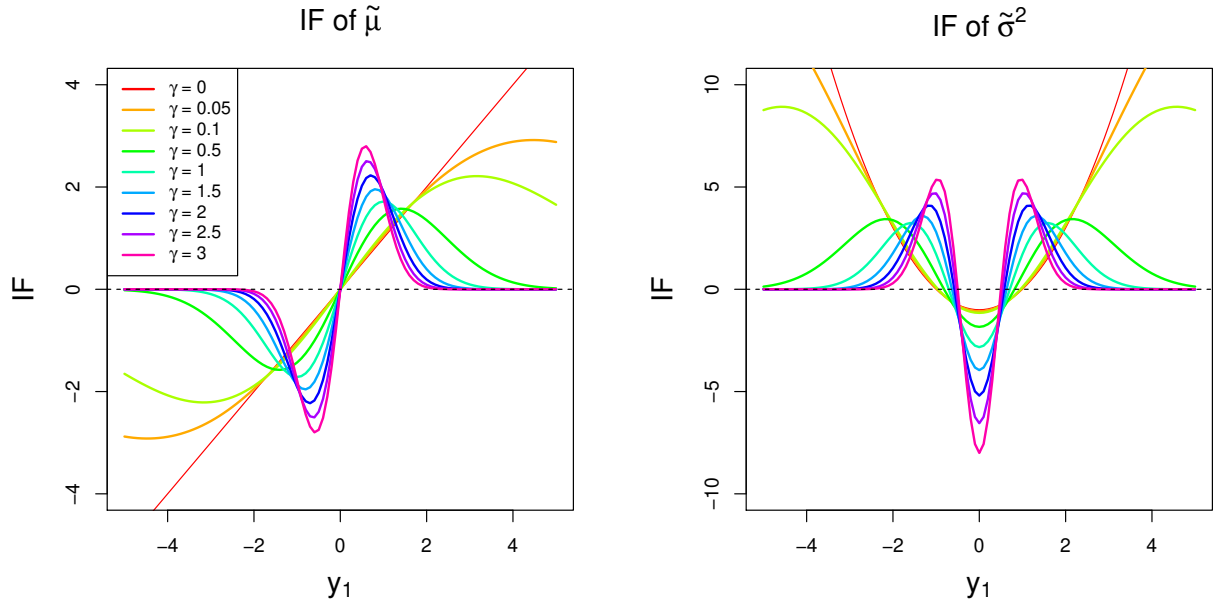

Figure 2: Influence functions under various γ's of $\tilde{\mu}_{0\gamma}$ (in the left panel) and $\tilde{\sigma}^2_{\gamma}$ (in the right panel) at standard normal distribution $\mathcal{N}(0,1)$.

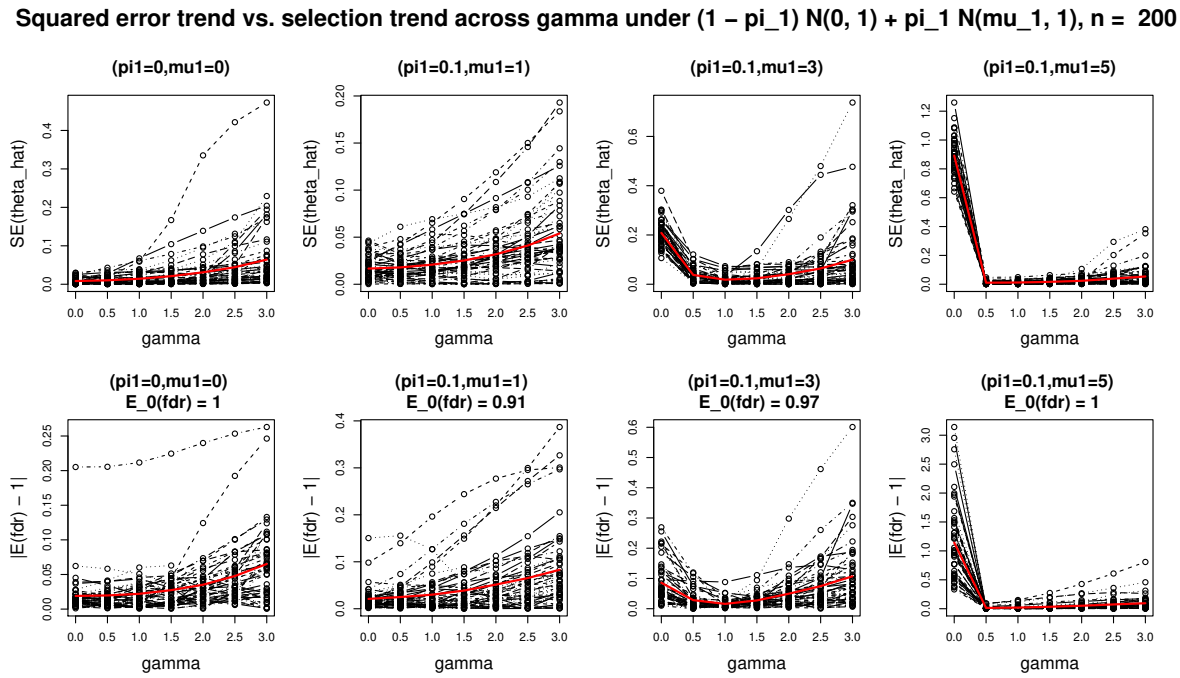

Figure 3: Comparison of squared error trend $\mathrm{SE}(\hat{\boldsymbol{\theta}}_{\mathbf{0}\boldsymbol{\gamma}})$ (in the top row) and selection trend $|\hat{\mathbb{E}}_{f_{0,\hat{\boldsymbol{\theta}}_{\mathbf{0}\boldsymbol{\gamma}}}} fdr^{\ddagger}(Y;\gamma) - 1|$ (in the bottom row) across $\gamma = 0, 0.5, 1, 2, 3$ 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 curves in the top row are the MSE trends and in the bottom row are the average of the selection trends from 50 black curves at each γ. The oracle $\mathbb{E}_0(fdr)$ is indicated in the subtitle for each case.

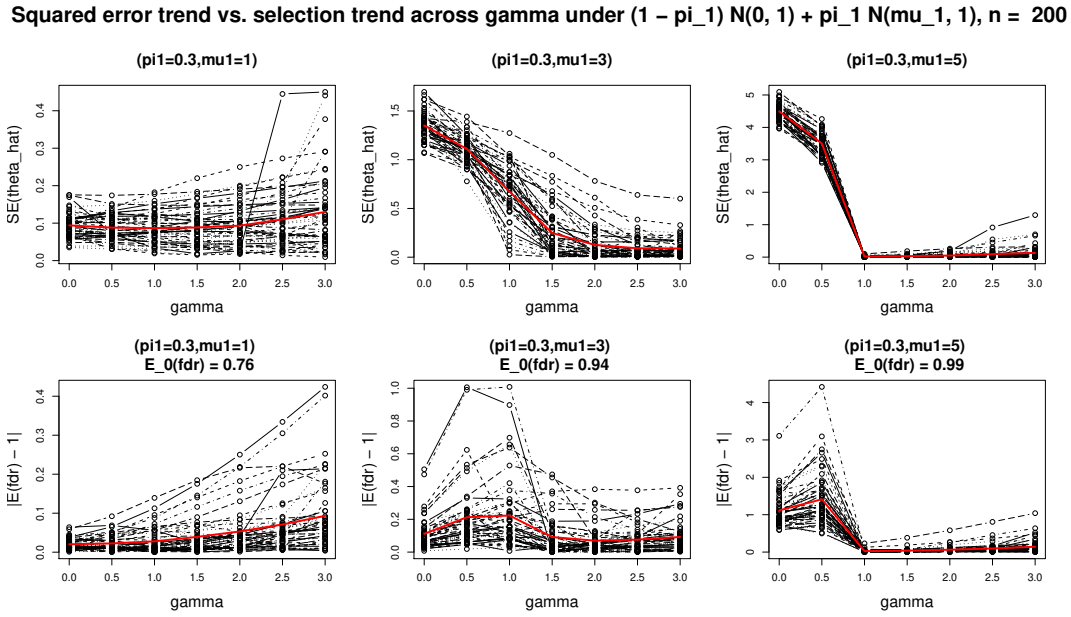

Figure 4: The same caption as Figure 3 under $n = 200$, $(\mu_1, \pi_1) = (1, 0.3), (3, 0.3), (5, 0.3)$.

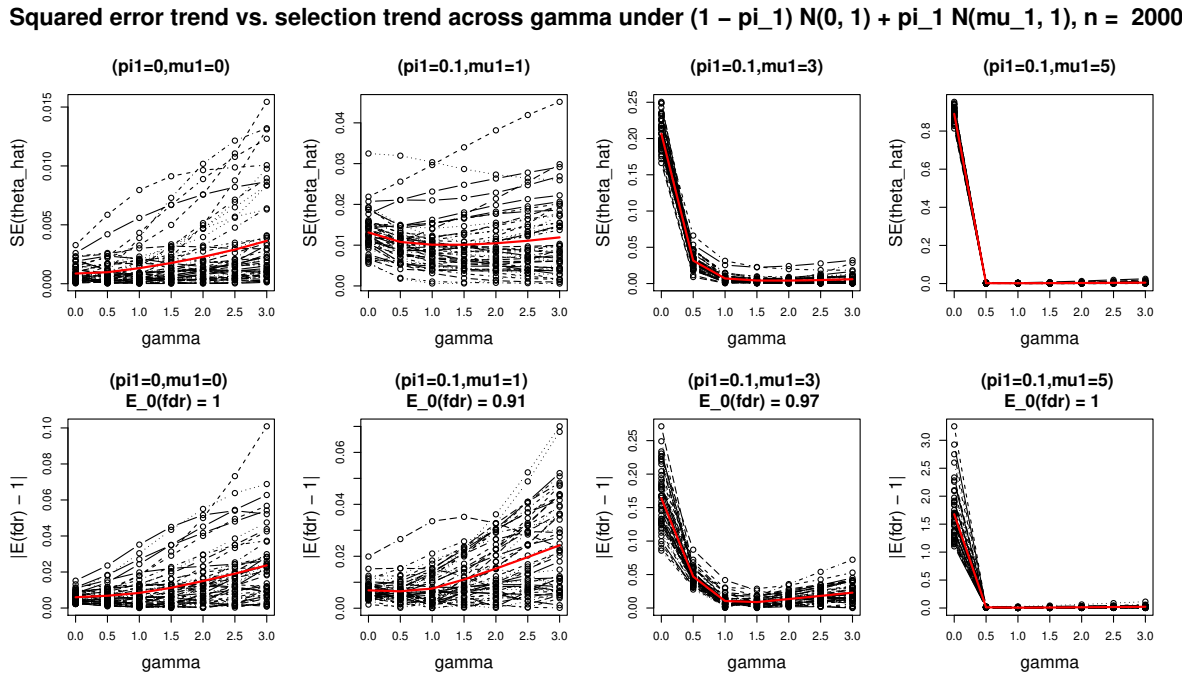

Figure 5: The same caption as Figure 3 under $n = 2000$, $(\mu_1, \pi_1) = (0, 0), (1, 0.1), (3, 0.1), (5, 0.1)$.

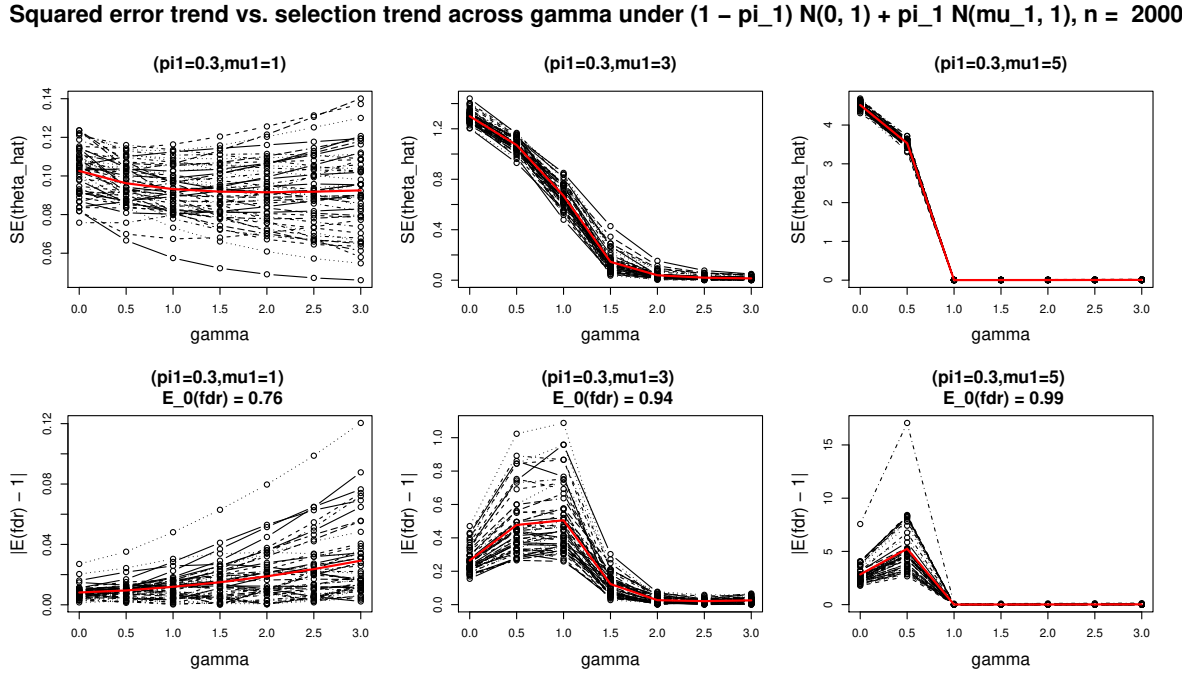

Figure 6: The same caption as Figure 3 under $n = 2000$, $(\mu_1, \pi_1) = (1, 0.3), (3, 0.3), (5, 0.3)$.

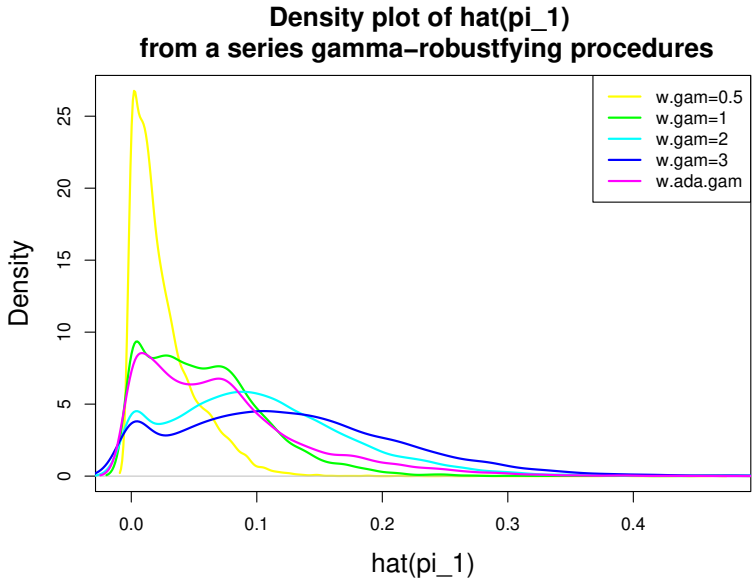

Figure 7: Density plot of $\hat{\pi}_1$ from a series γ-robustifying procedures under $\gamma = 0.5, 1, 2, 3$ and data-adaptive γ applied in heart expression data from GTEx.

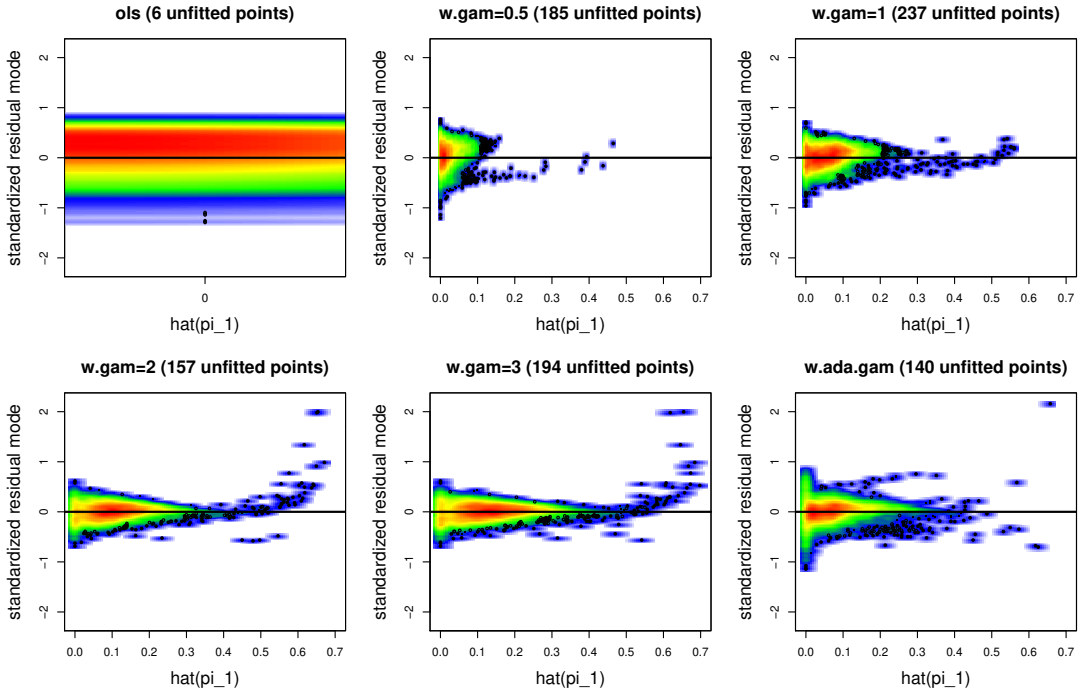

Figure 8: Scatter plot of $\hat{\pi}_1$ versus estimated standardized residual model from kernel density smoothing fitting under γ-robustifying procedures ($\gamma = 0, 0.5, 1, 2, 3$ and data-adaptive γ). The color from red to blue indicates the point density from dense to sparse. The black circled points in the blue region are defined as unfitted points. The horizontal black line is at zero.

Figure 9: Bubble chart of minus of $\log_{10}(p\text{-value})$ for the significantly enriched GO terms (threshold: adjust p-values $\leq 10^{-4}$) under various γ-robustifying procedures including small γ ($\gamma = 0, 0.5, 1$), large $\gamma = 2, 3$ and our data-adaptively selected γ, based on the genes whose heart left ventricle expression significantly greater than heart atrial appendage expression.

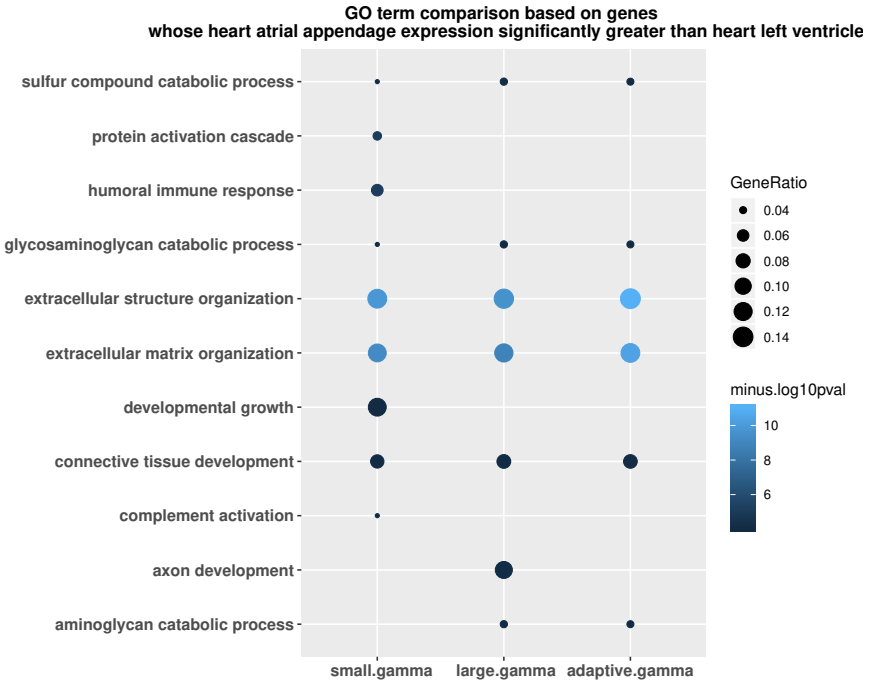

Figure 10: Bubble chart based on the genes whose heart atrial appendage expression significantly greater than heart left ventricle expression.

Supplementary Materials

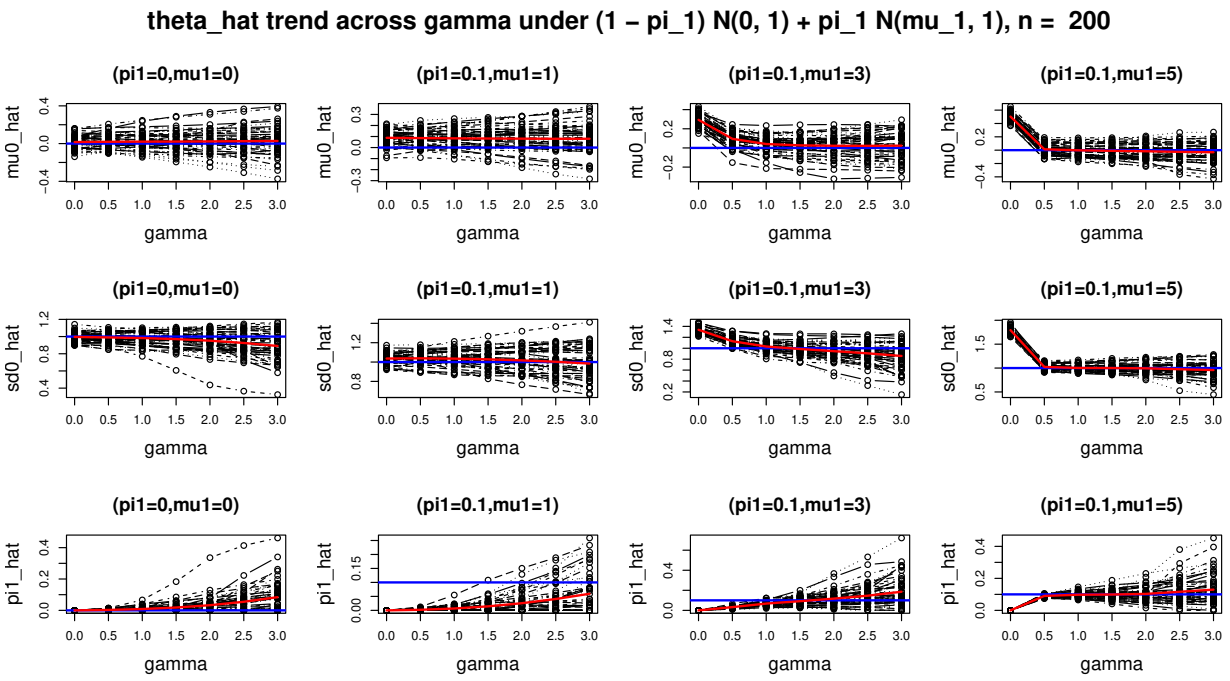

Figure S1: Trends of $\hat{\mu}_0$, $\hat{\sigma}_0$ and $\hat{\pi}_1$ across $\gamma = 0, 0.5, 1, 2, 3$ under Gaussian mixture model $(1 - \pi_1)\mathcal{N}(0, 1) + \pi_1\mathcal{N}(\mu_1, 1)$ where $n = 200$ and $(\mu_1, \pi_1) = (0, 0), (1, 0.1), (3, 0.1), (5, 0.1)$. Each black curve is from one sample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

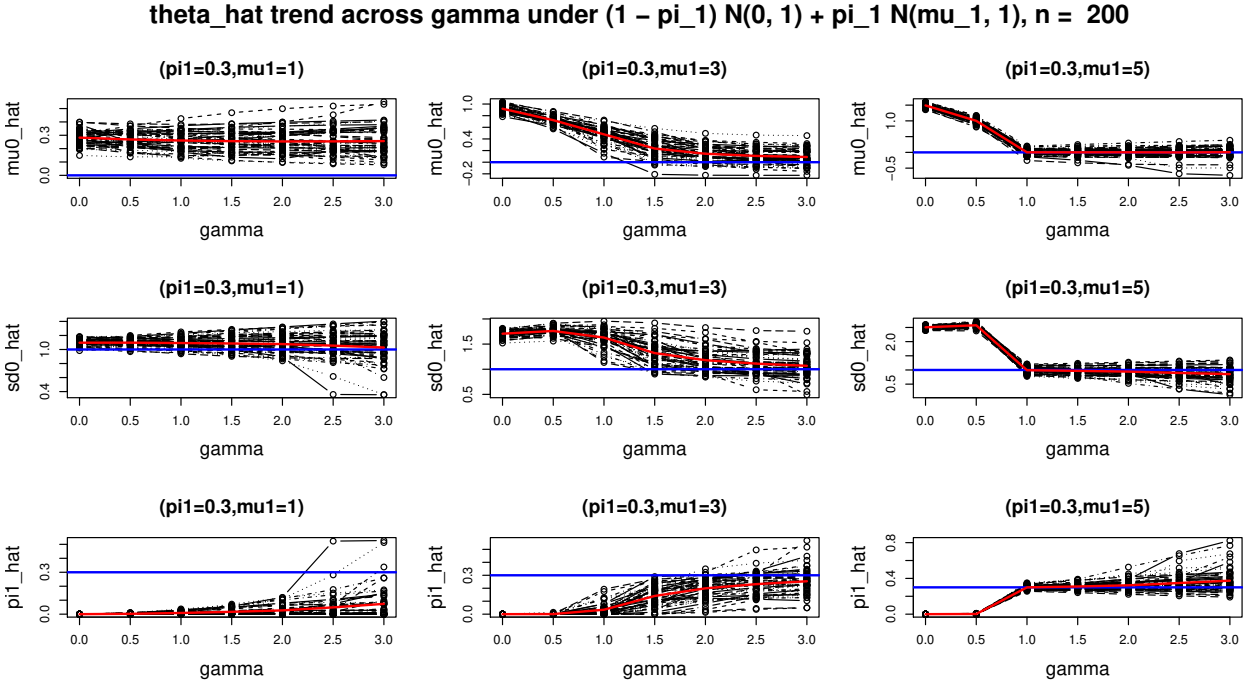

Figure S2: The same caption as Figure S1 under $n = 200$, $(\mu_1, \pi_1) = (1, 0.3), (3, 0.3), (5, 0.3)$.

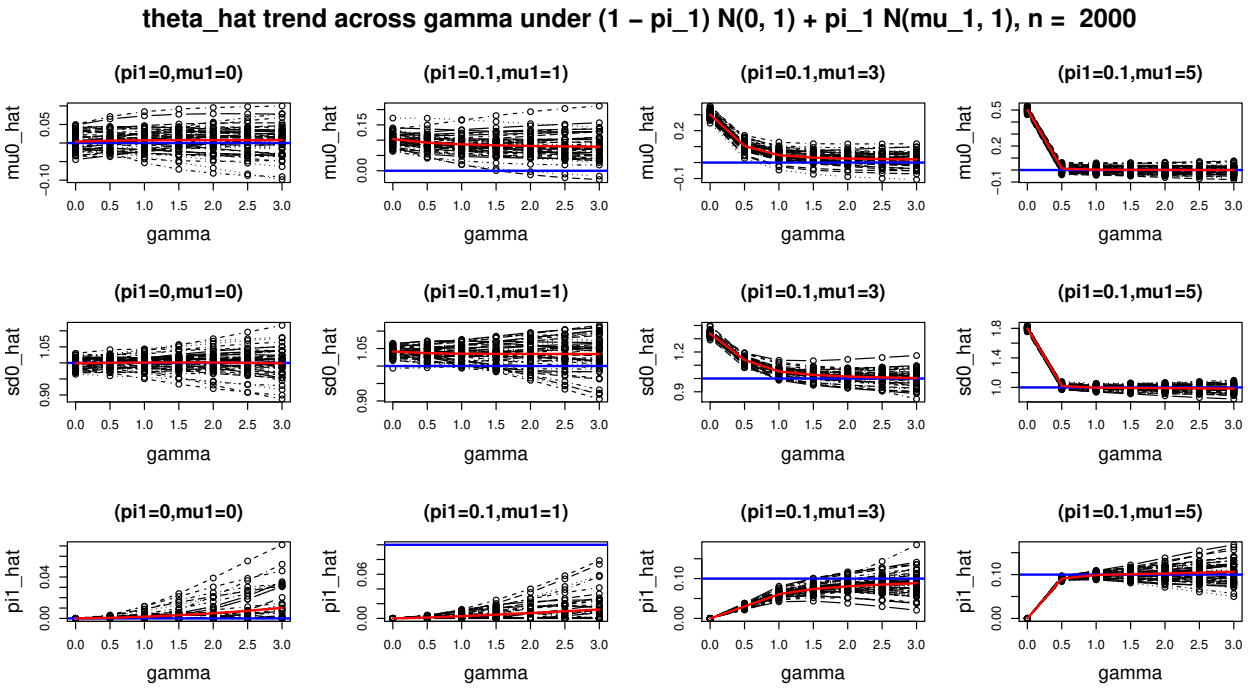

Figure S3: The same caption as Figure S1 under $n = 2000$, $(\mu_1, \pi_1) = (0, 0), (1, 0.1), (3, 0.1), (5, 0.1)$.

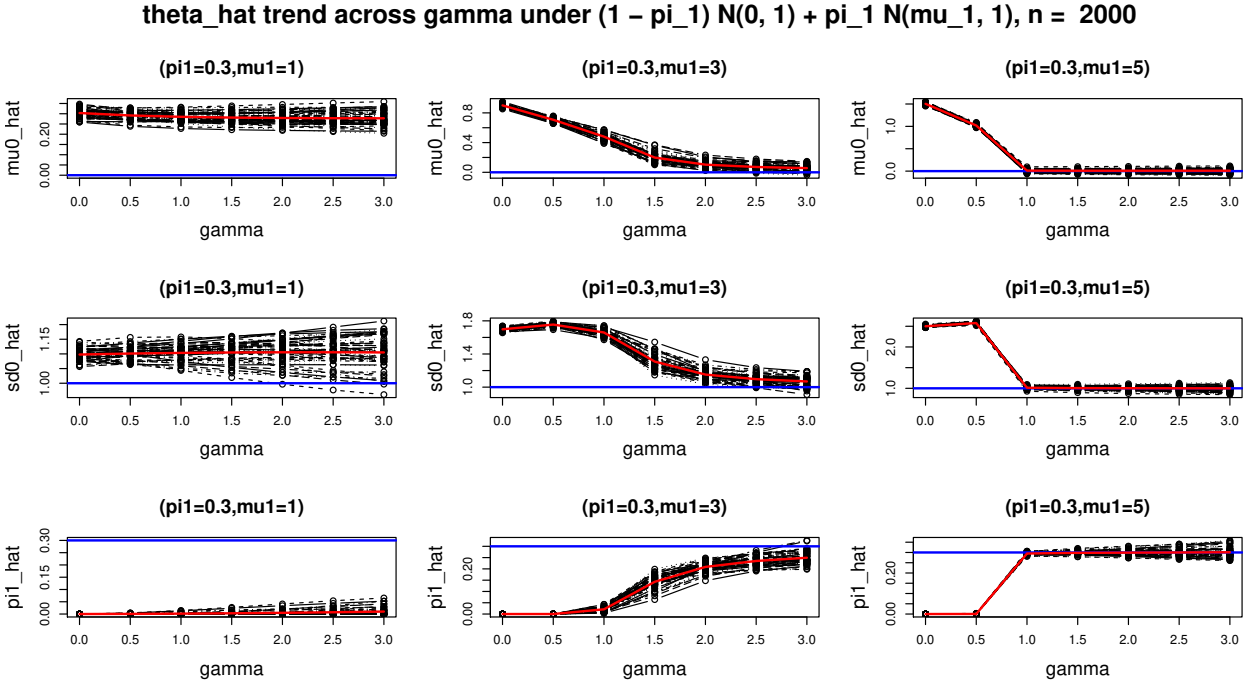

Figure S4: The same caption as Figure S1 under $n = 2000$, $(\mu_1, \pi_1) = (1, 0.3), (3, 0.3), (5, 0.3)$.

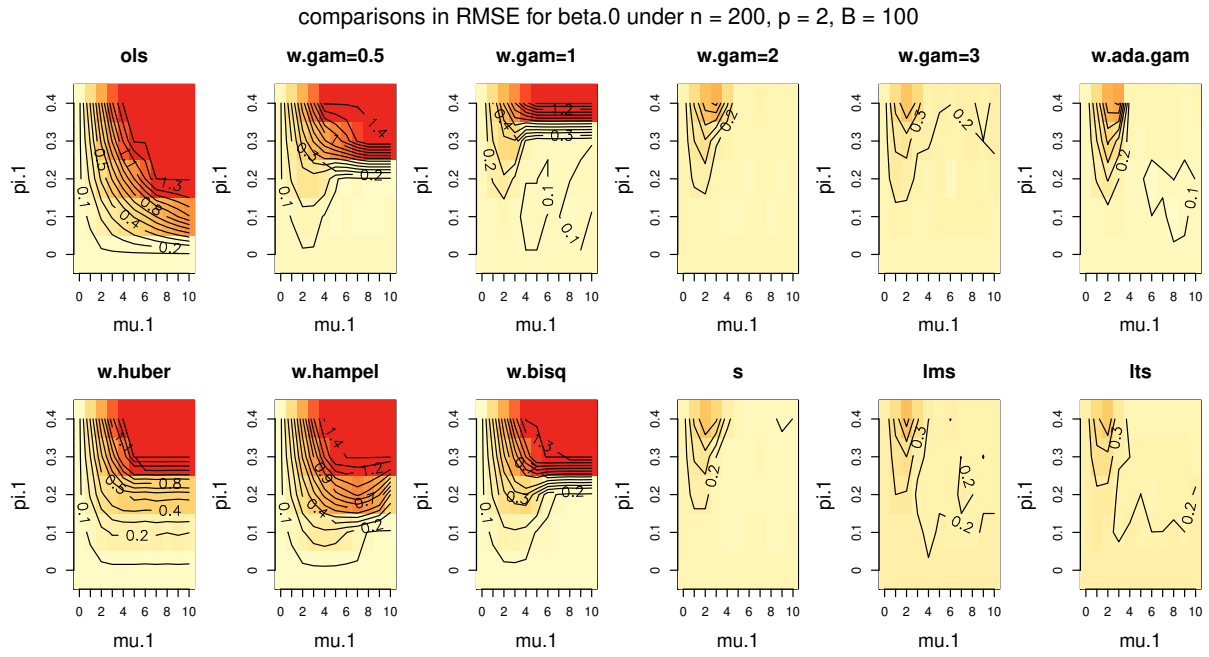

Figure S5: Heat map of RMSE comparisons for estimating $\boldsymbol{\beta_0}$ under the regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta_0} + \boldsymbol{\epsilon}$ where sample size $n = 200$ and $p = 2$, $\boldsymbol{\beta_0} = \mathbf{1_{p\times 1}}$ and $X_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 10^2)$ 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.

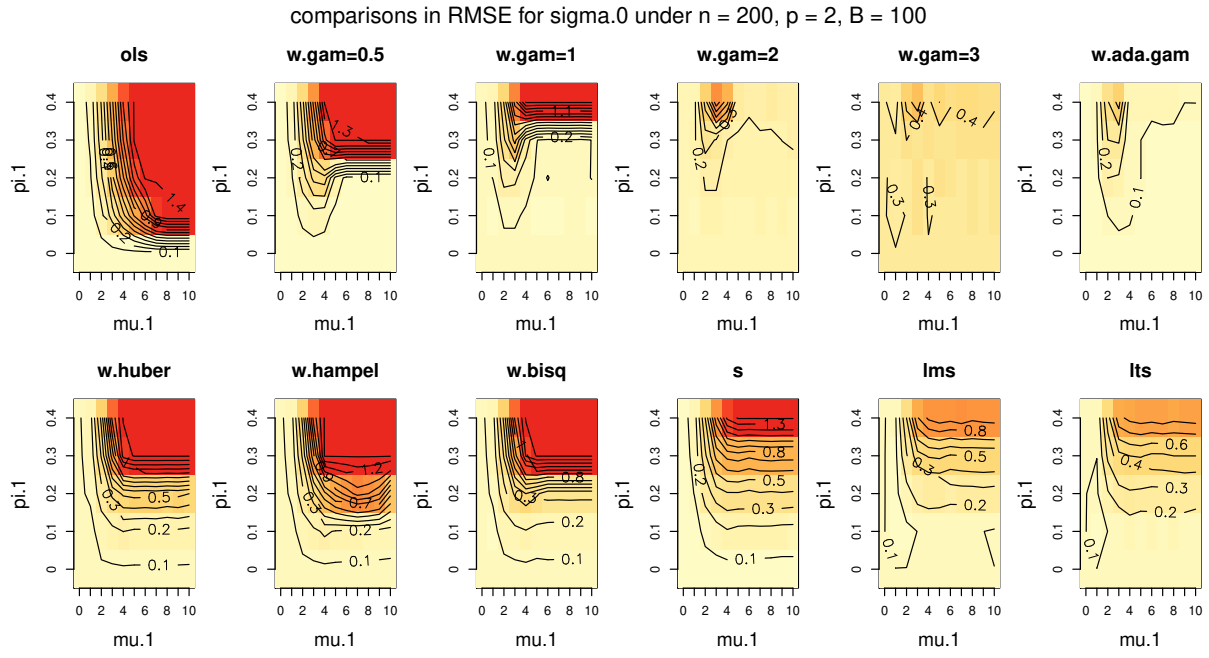

Figure S6: Heat map of RMSE comparisons for estimating σ_0 under the same setting as in Figure S5.

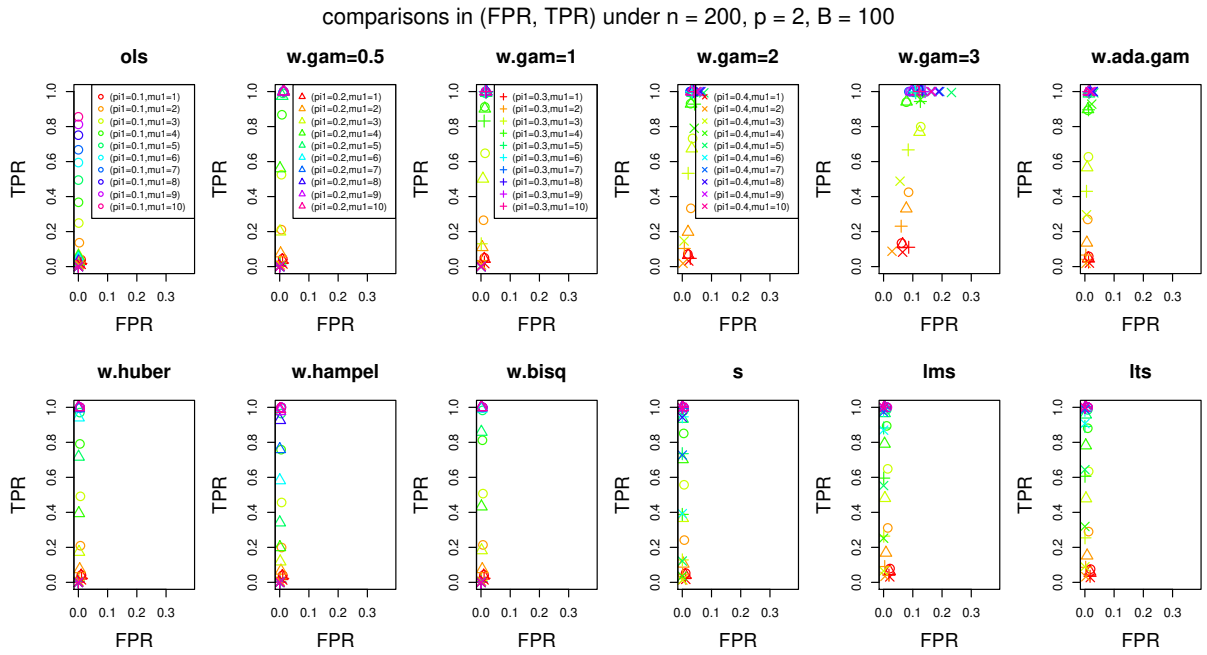

Figure S7: Scatter plot of (FPR, TPR) under the same setting as in Figure S5.

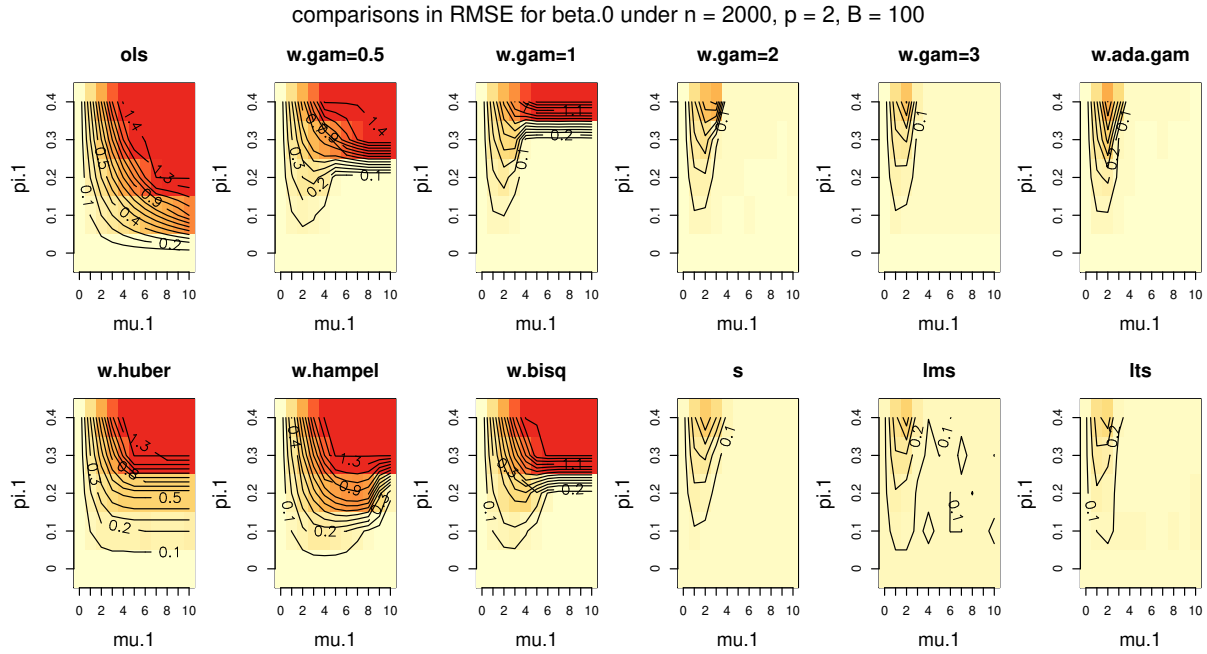

Figure S8: Heat map of RMSE comparisons for estimating $\boldsymbol{\beta_0}$ in the same labels as in Figure S5 except $n = 2000, p = 2$.

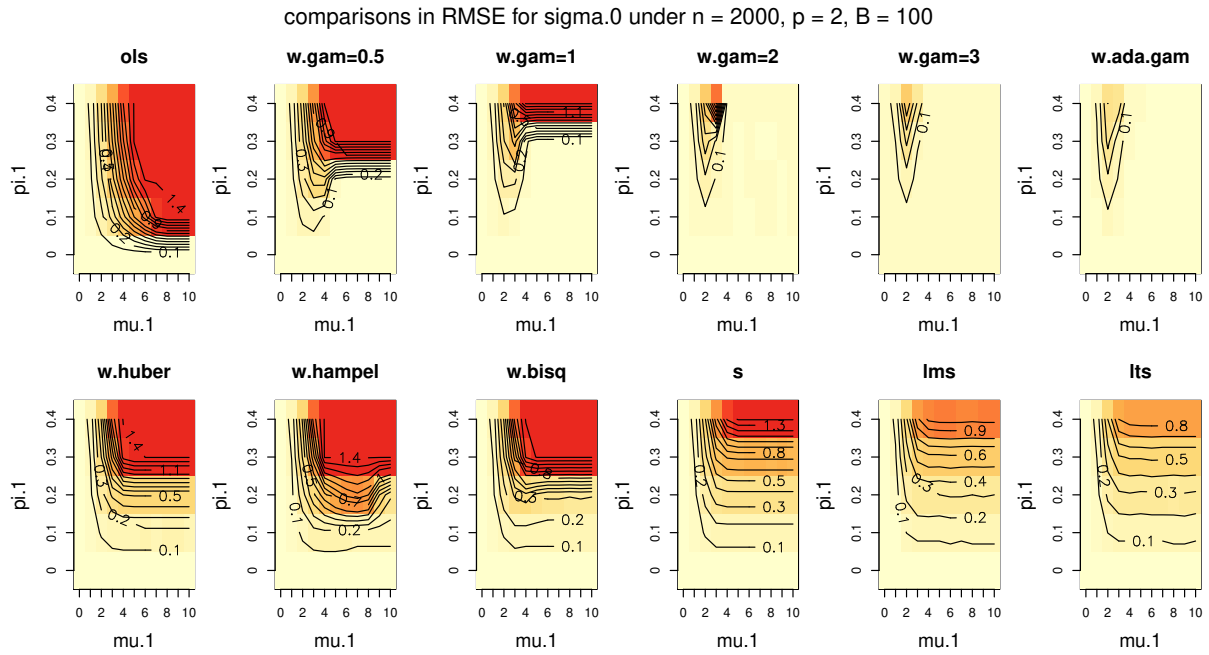

Figure S9: Heat map of RMSE comparisons for estimating σ_0 in the same labels as in Figure S6 except $n = 2000, p = 2$.

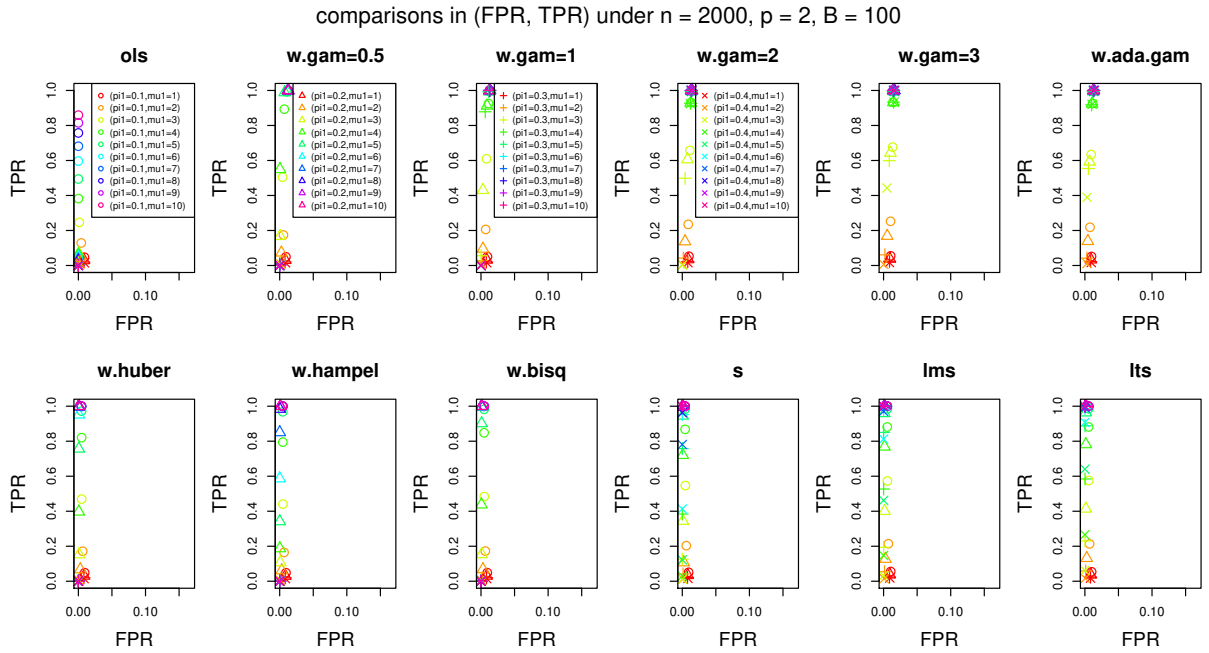

Figure S10: Scatter plot of (FPR, TPR) in the same labels as in Figure S7 except $n = 2000, p = 2$.

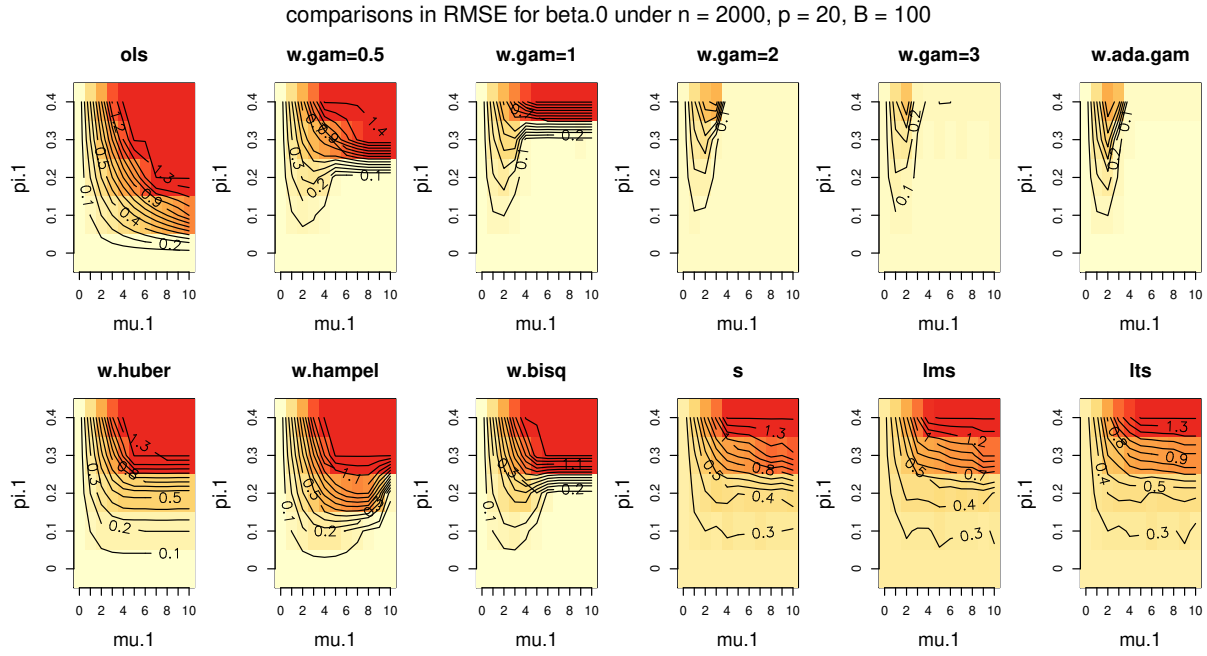

Figure S11: Heat map of RMSE comparisons for estimating $\boldsymbol{\beta_0}$ in the same labels as in Figure S5 except $n = 2000, p = 20$.

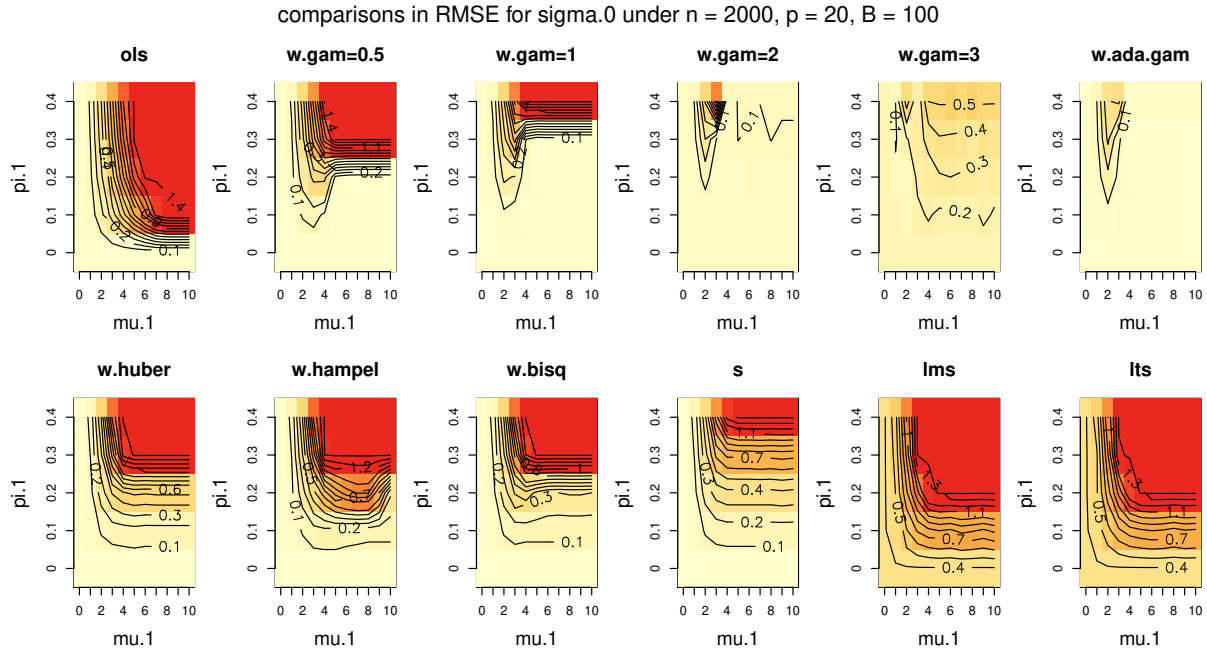

Figure S12: Heat map of RMSE comparisons for estimating σ_0 in the same labels as in Figure S6 except $n = 2000, p = 20$.

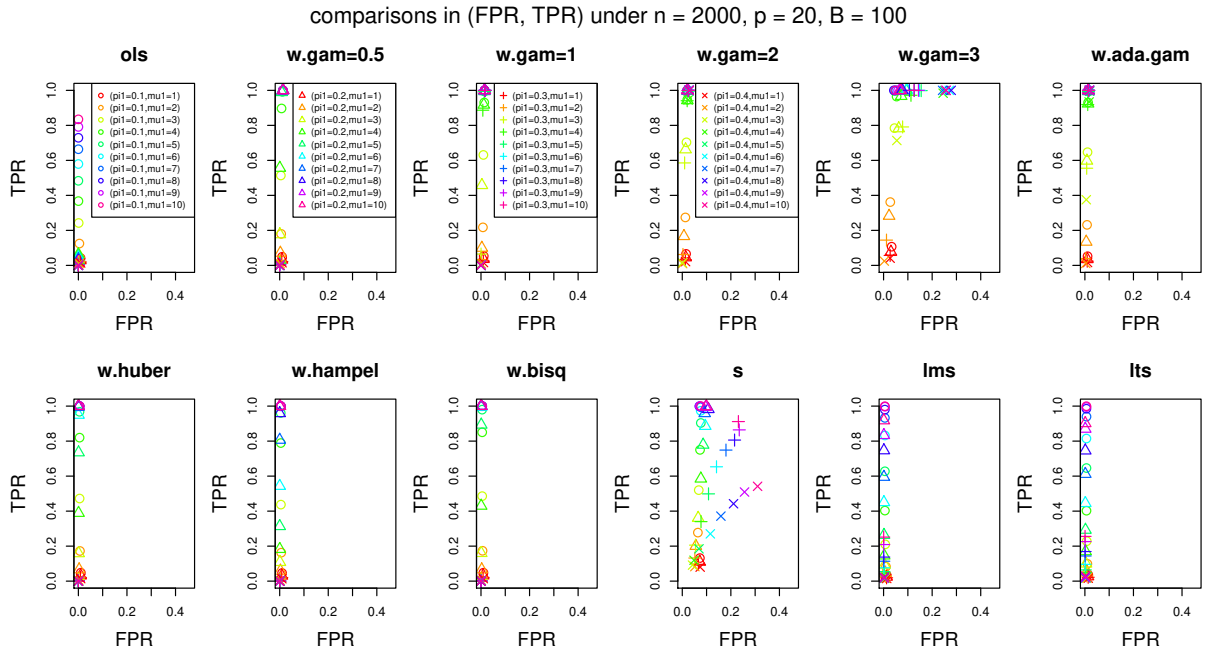

Figure S13: Scatter plot of (FPR, TPR) in the same labels as in Figure S7 except $n = 2000, p = 20$.

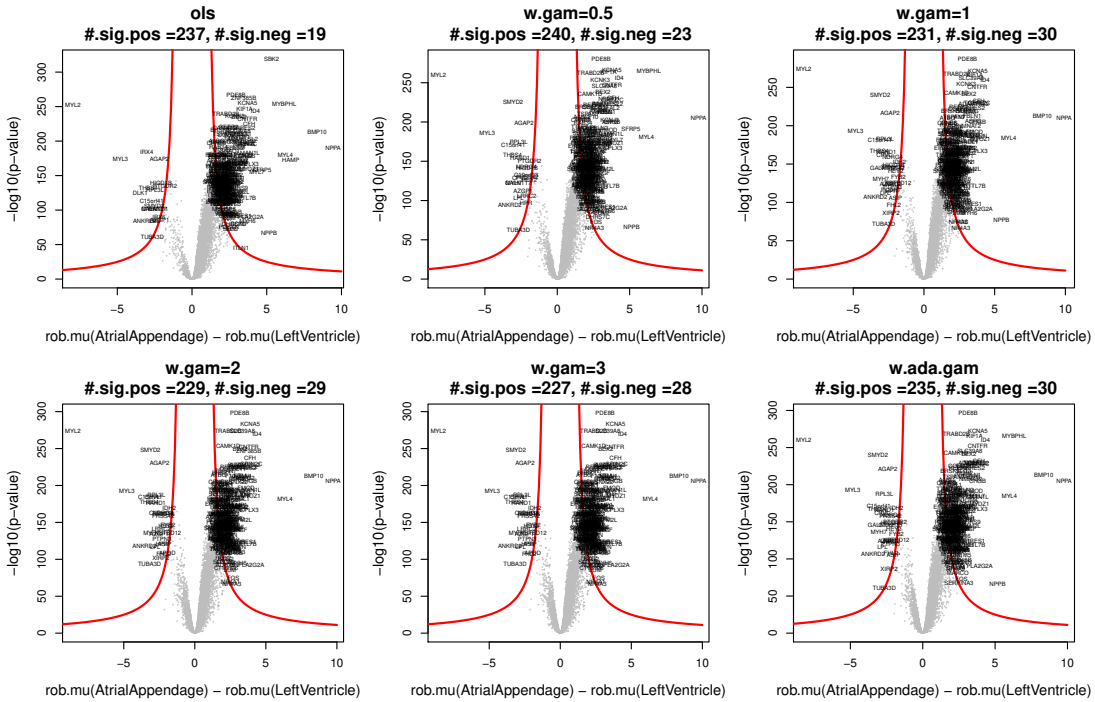

Figure S14: Volcano plot of minus of $\log_{10}(p\text{-value})$ versus log(fold change) ($\hat{\mu}$(atrial appendage) − $\hat{\mu}$(left ventricle)) from two-sample t-test with unequal variances after removing the outliers from various γ-robustifying procedures. The red curve is hyperbolic cut with 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.