Statistics and Its Interface ()

Variable Selection in ROC Curve Analysis with Focused Information Criteria(Supplementary material: Some simulation results)

Baoying Yang[‡] , Xin Huang , and Gengsheng Qin

1. SIMULATION STUDIES

In this section, based on the placement value model, we conduct simulation studies to evaluate the finite sample performances of the AIC, BIC and FIC in terms of the Mean Square Error (MSE) and the Mean Absolute Deviation (MAD) of the estimators for AUC index.

For the diseased sample, the AIC and BIC under a submodel S can be expressed as (See Hjort and Claeskens, 2003):

$$\text{(1.1)}\quad \text{AIC}_S^D = -\widehat{\boldsymbol{\delta}}_{\text{full}}(K^D)^{-\frac{1}{2}}H_S^D(K^D)^{-\frac{1}{2}}\widehat{\boldsymbol{\delta}}_{\text{full}} + 2|S|,$$

$$\text{(1.2)}\quad \text{BIC}_S^D = -\widehat{\boldsymbol{\delta}}_{\text{full}}(K^D)^{-\frac{1}{2}}H_S^D(K^D)^{-\frac{1}{2}}\widehat{\boldsymbol{\delta}}_{\text{full}} + log(n)|S|,$$

respectively, where $|S|$ is the number of elements in S.

Based on expressions (1.1) and (1.2), we choose the models with the smallest AIC and BIC value as the best one. Using the FIC criteria, we choose the model with the smallest FIC value focused on AUC as the best one. In simulation studies, we compare performances of the AIC, BIC, and FIC criteria through comparing the estimates $\widehat{AUC}(\mathbf{Z}_0)$ of AUC at the given covariates $\mathbf{Z}_0$, the MSE and MAD of $\widehat{AUC}(\mathbf{Z}_0)$ over M=1000 simulation runs under each simulation setting, where $\text{MSE}(\mathbf{Z}_0) = \frac{1}{M}\sum_{m=1}^{M}(\widehat{AUC}_m(\mathbf{Z}_0) - AUC(\mathbf{Z}_0))^2$, $\text{MAD}(\mathbf{Z}_0) = \frac{1}{M}\sum_{m=1}^{M}|\widehat{AUC}(\mathbf{Z}_0)_m - AUC(\mathbf{Z}_0)|$, and $\widehat{AUC}(\mathbf{Z}_0)_m$ is the estimate for $AUC(\mathbf{Z}_0)$ based on the m-th simulated sample.

We use the following placement value model in examples 1-5:

$$\Phi^{-1}(U^D|\mathbf{Z}^D) = -\eta_0 X^D - \boldsymbol{\eta}^t\mathbf{Z}^D + \varepsilon,$$

where $\varepsilon \sim N(0,\sigma^2)$. The simulated data are generated from the models with different simulation settings.

Example 1: We set $X^D = 1$. The q dimension covariates $\mathbf{Z}^{\bar{D}}$ are generate from $\mathbf{Z}^D \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (1,\cdots,1)$, and the covariance matrix $\boldsymbol{\Sigma} = (\Sigma_{ij})$ with $\Sigma_{ii} = \rho^{|i-j|}$, $1 \le i \ne j \le q$. The correlation coefficient ρ is chosen to be 0.5, and 0.8, respectively. We choose $\boldsymbol{\theta} = (\eta_0,\sigma) = (0.8, 0.1)$, and $\boldsymbol{\eta} = (0.5, 0.3, 0.2, 0)$ with $q = 4$. The diseased sample size is $n_1 = 300$;

Example 2: The model is the same as that in example 1 except that the sample size is $n_1 = 500$;

Example 3: The model is the same as that in example 1 except that the sample size is $n_1 = 1000$;

Example 4: The model is the same as that in example 1 except that $\boldsymbol{\eta} = (0.5, 0.3, 0, 0, 0, 0, 0.4, 0)$ with $q = 8$;

Example 5: To consider the robustness of the proposed method, we consider a case in which the error term doesn't follow the normal distribution, but the simulation is still conducted under the assumption that the error follows the normal distribution. The placement values are generated from the following model:

$$\Phi^{-1}(U^D|\mathbf{Z}^D) = -\eta_{\mathbf{0}}\mathbf{X^D} - \boldsymbol{\eta}^{\mathbf{t}}\mathbf{Z^D} + \varepsilon,$$

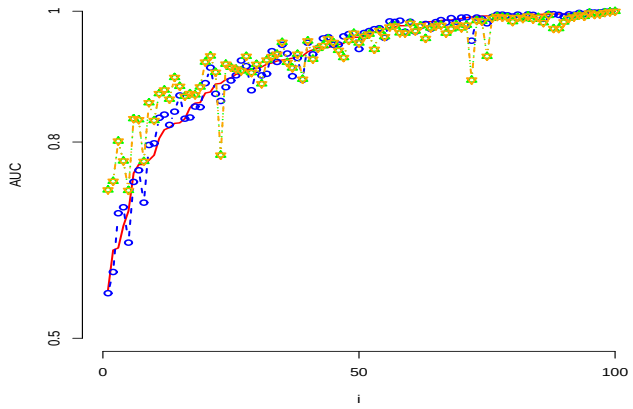
where $\mathbf{X}^D = (1,\xi)$ with $\xi \sim N(0,1)$, $\mathbf{Z}^D$ are generated from the same distribution as that in example 1. The coefficients $\eta_{\mathbf{0}} = (\mathbf{0.2}, \mathbf{0.1})$ and $\eta = (\mathbf{0.5}, \mathbf{0}, \mathbf{0.3}, \mathbf{0.2}, \mathbf{0}, \mathbf{0})$. The true distribution of error term is $\varepsilon \sim 0.1t(3)$, where $t(3)$ is a t-distribution with 3 degree of freedom.

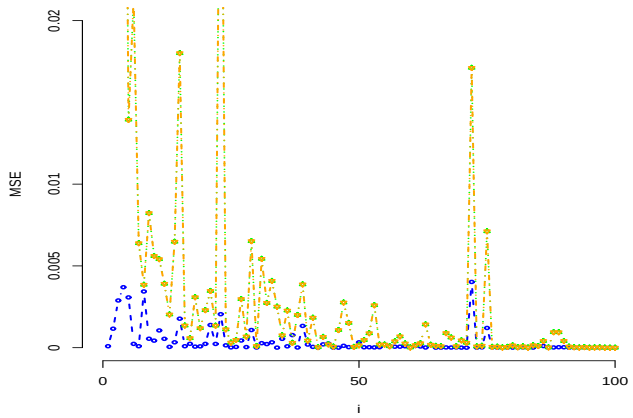
For given $\mathbf{Z}_0$, AUC can be expressed as $AUC(\mathbf{Z}_0) = g(\boldsymbol{\theta}^0, \boldsymbol{\eta}^0 + \boldsymbol{\delta}/\sqrt{n_1}|\mathbf{Z}_0) = E\left(1 - U^D|\mathbf{Z}_0\right)$. Using the simulated data from the true placement value models described in examples 1-5, we estimate AUCs at 100 different covariates $\mathbf{Z}_0$ and the corresponding MSE($\mathbf{Z}_0$)'s and MAD($\mathbf{Z}_0$)'s under the selected models by using AIC, BIC and FIC over M=1000 simulation runs, respectively.

Figures $1-10$ display the results for AUC, MSE amd MAD comparisons by using the AIC, BIC and FIC. From these figures, we can see that the true AUC is varying with $\mathbf{Z}_0$, and the estimates of AUC based on FIC are much closer to the true AUC than the AIC and BIC based estimates. Figures $1-12$ show that the MSE($\mathbf{Z}_0$) and the MAD($\mathbf{Z}_0$) based on the FIC selected models are smaller than those based on the AIC and BIC selected models, which indicates that the FIC has better finite sample performances than the AIC and BIC in variable selection of placement value model.

[‡]Corresponding author

[‡]The research of Baoying Yang was supported by National Natural Science Foundation of China (NNSFC, No. 11501472) and the Soft Science Research Program in Sichuan Province of China (No. 2015ZR0211).
[‡]We would like to thank the editors and referees for their comments which led to substantial improvements in this article.





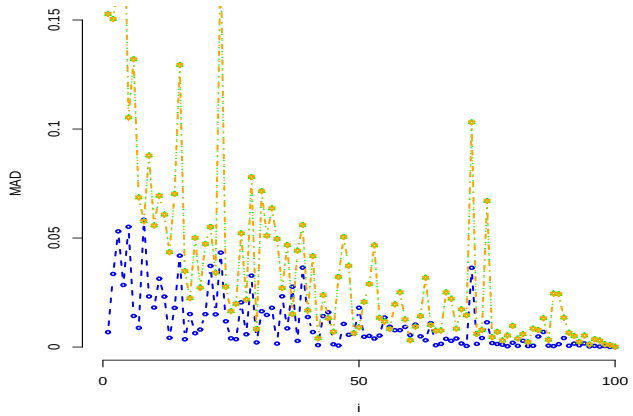
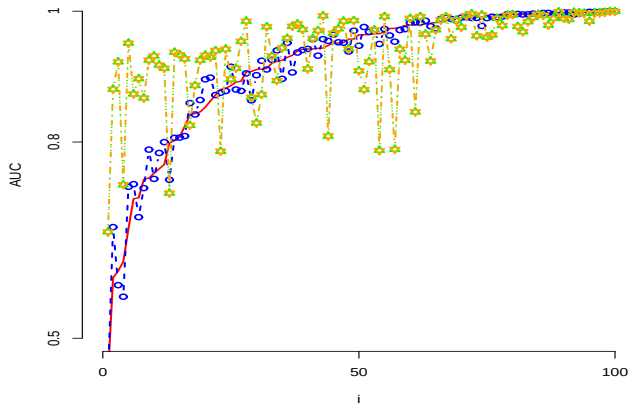
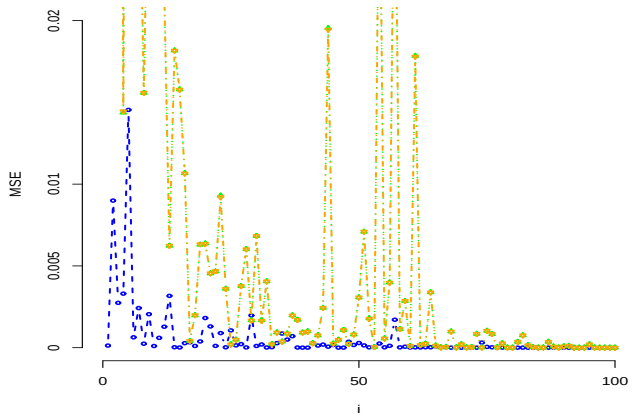


Figure 1. Example 1: Comparison with $n_1 = 300$ and $\rho = 0.5$





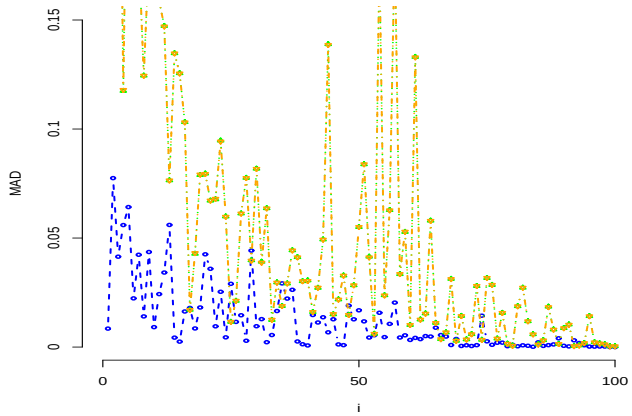


Figure 2. Example 1: Comparison with $n_1 = 300$ and $\rho = 0.8$

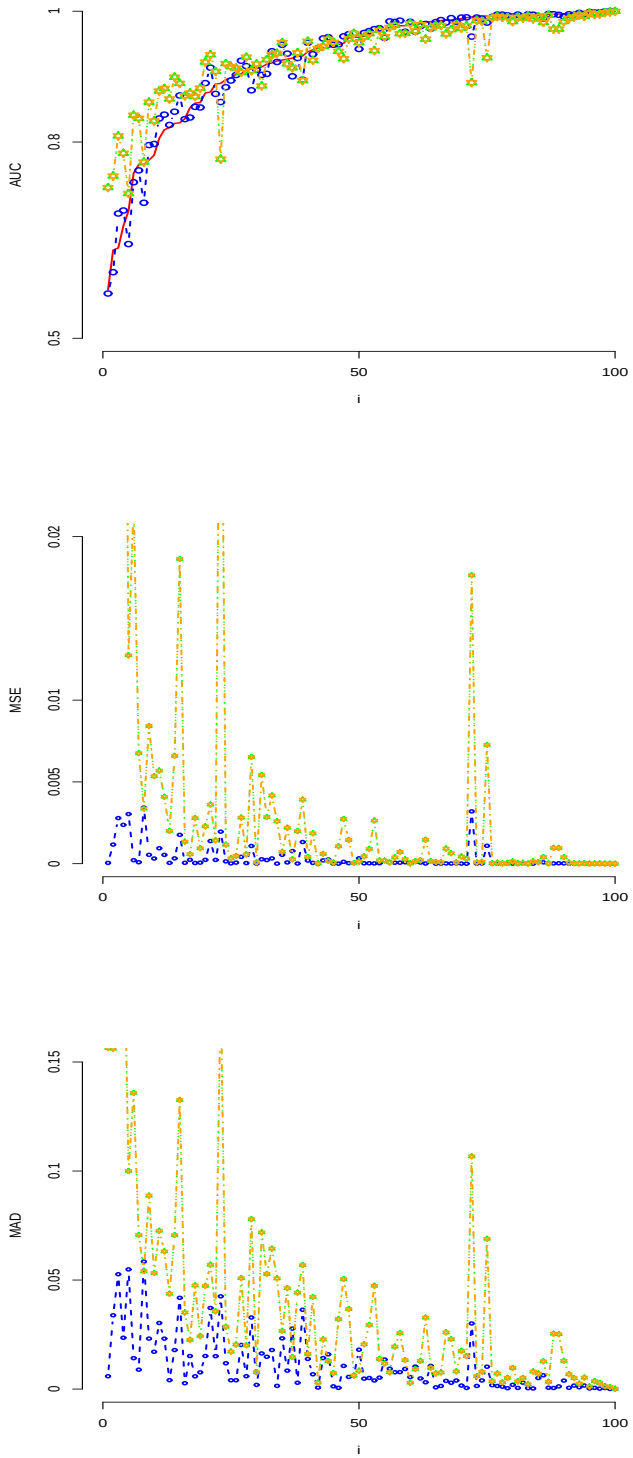


Figure 3. Example 2: Comparison with $n_1 = 500$ and $\rho = 0.5$

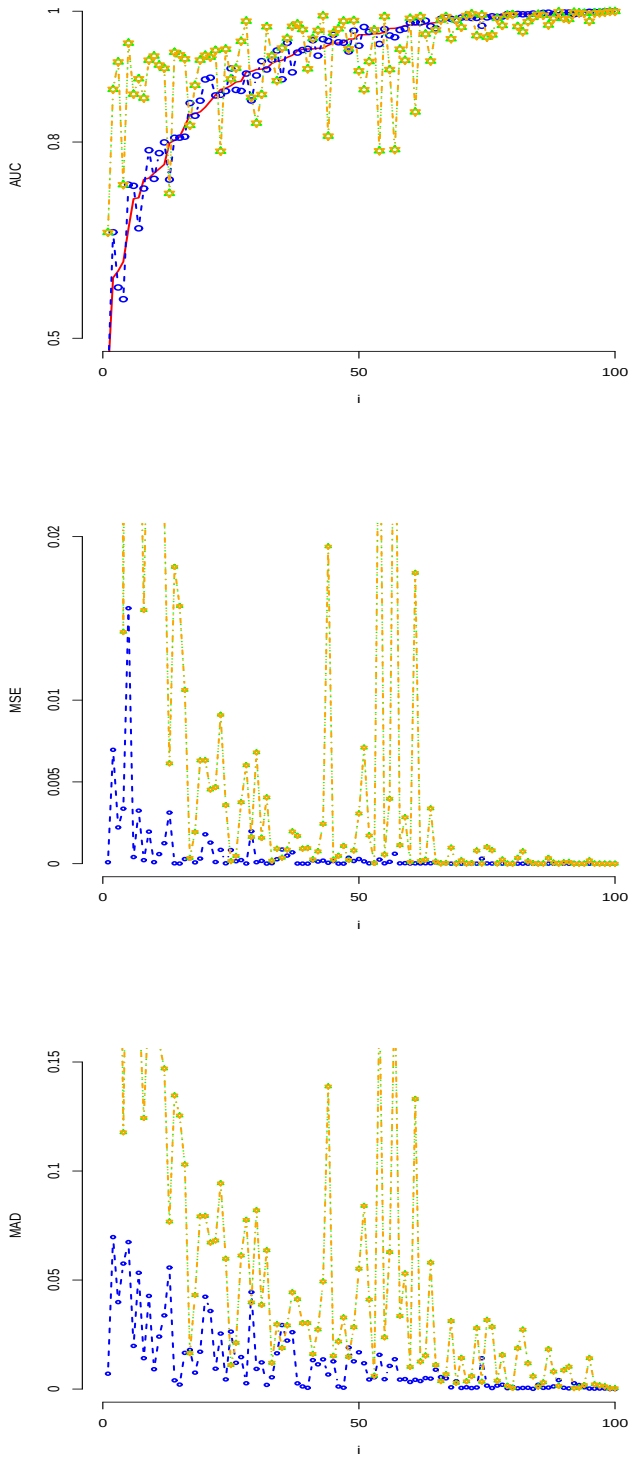
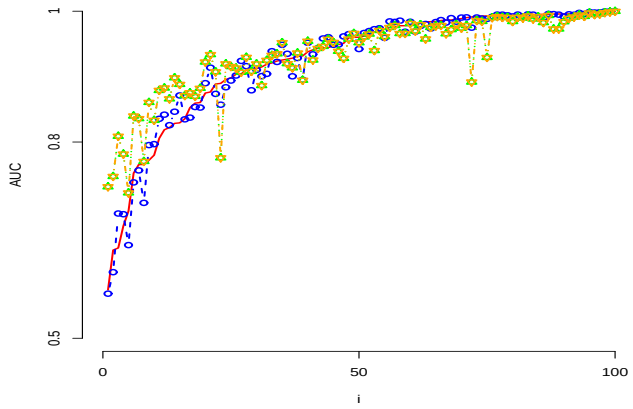
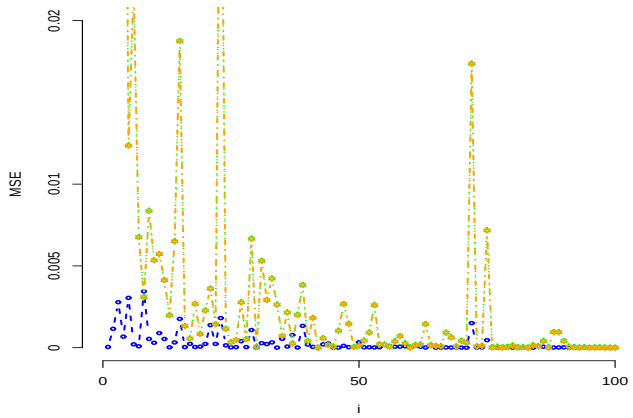


Figure 4. Example 2: Comparison with $n_1 = 500$ and $\rho = 0.8$





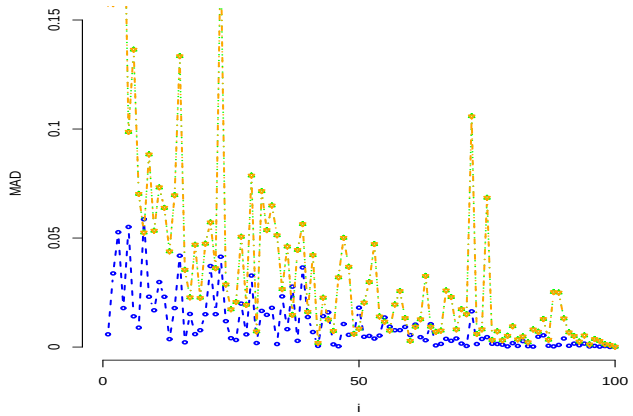
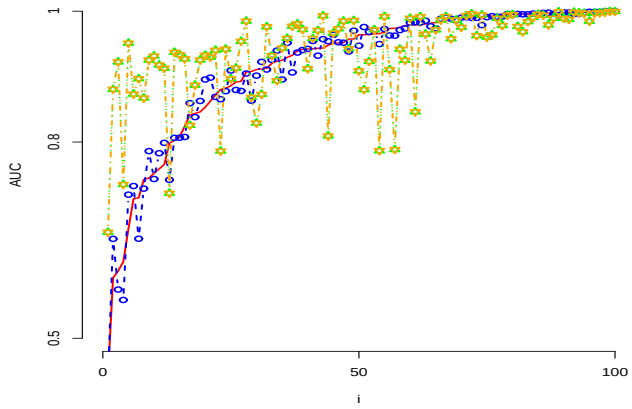
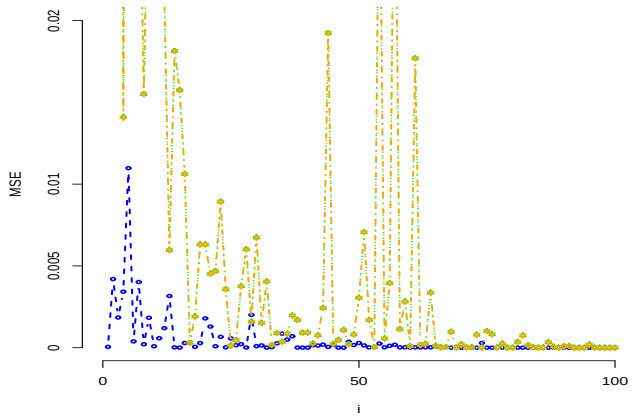


Figure 5. Example 3: Comparison with $n_1 = 1000$ *and* $\rho = 0.5$





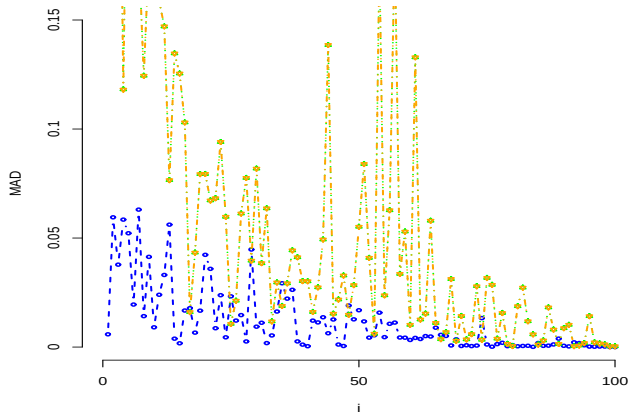
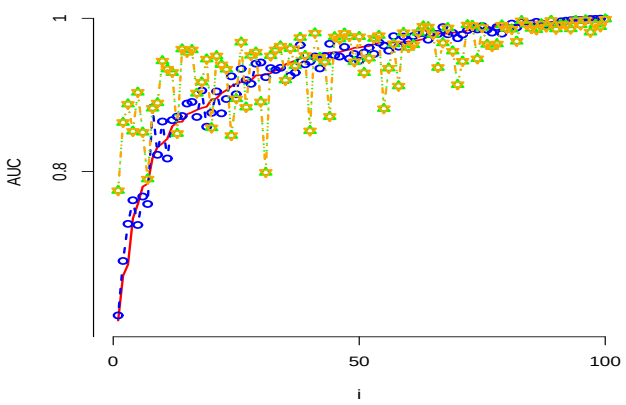
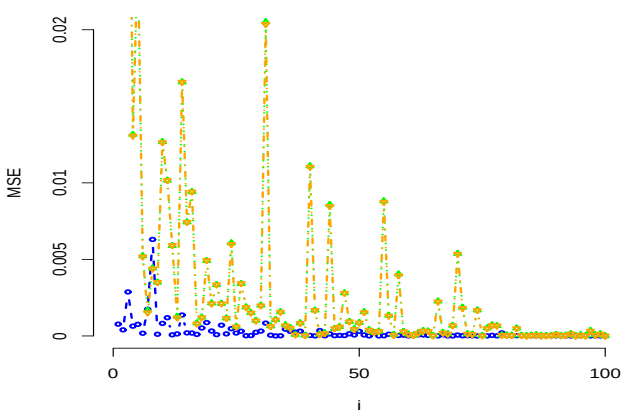


Figure 6. Example 3: Comparison with $n_1 = 1000$ *and* $\rho = 0.8$





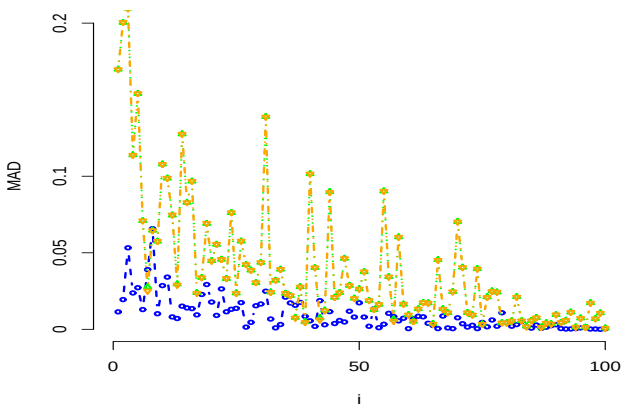
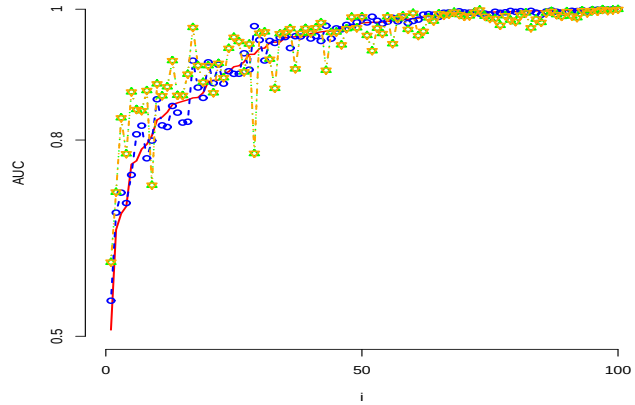
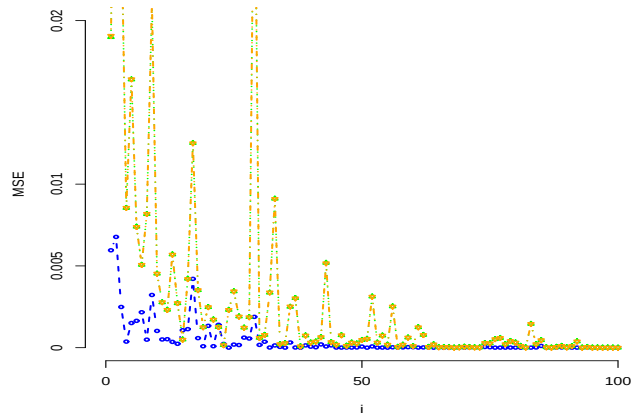


Figure 7. Example 4: Comparison with $n_1 = 300$, $\rho = 0.5$, *and* $q = 8$





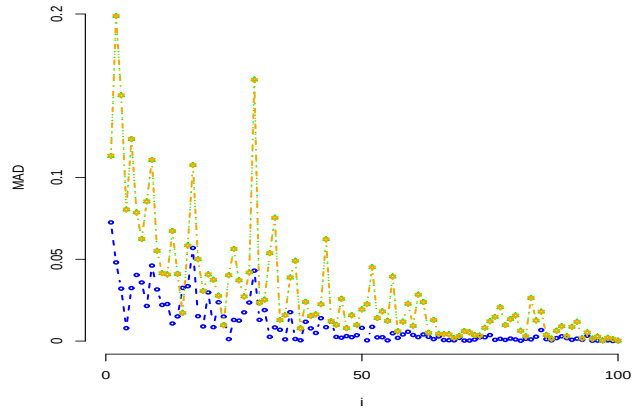


Figure 8. Example 4: Comparison with $n_1 = 300$, $\rho = 0.8$, *and* $q = 8$

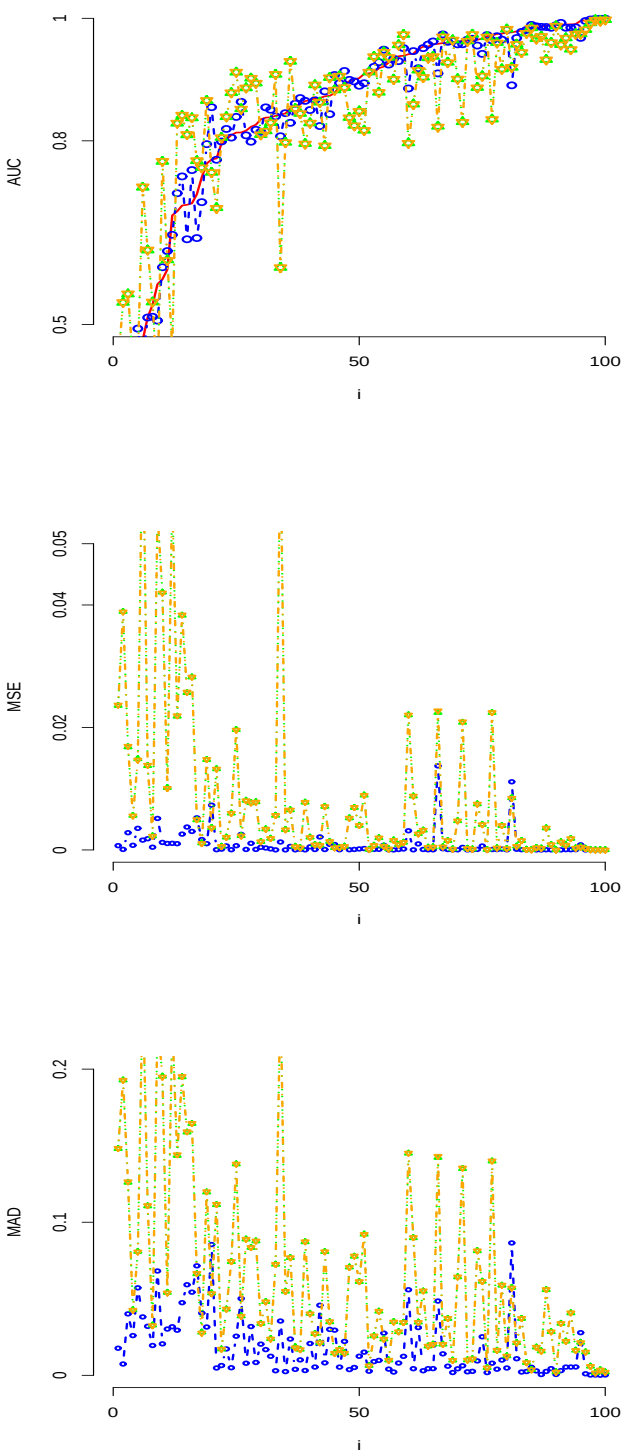


Figure 9. Example 5: Comparison with $n_1 = 300$, $\rho = 0.5$, $q = 6$ *and* $\varepsilon \sim 0.1t(3)$.

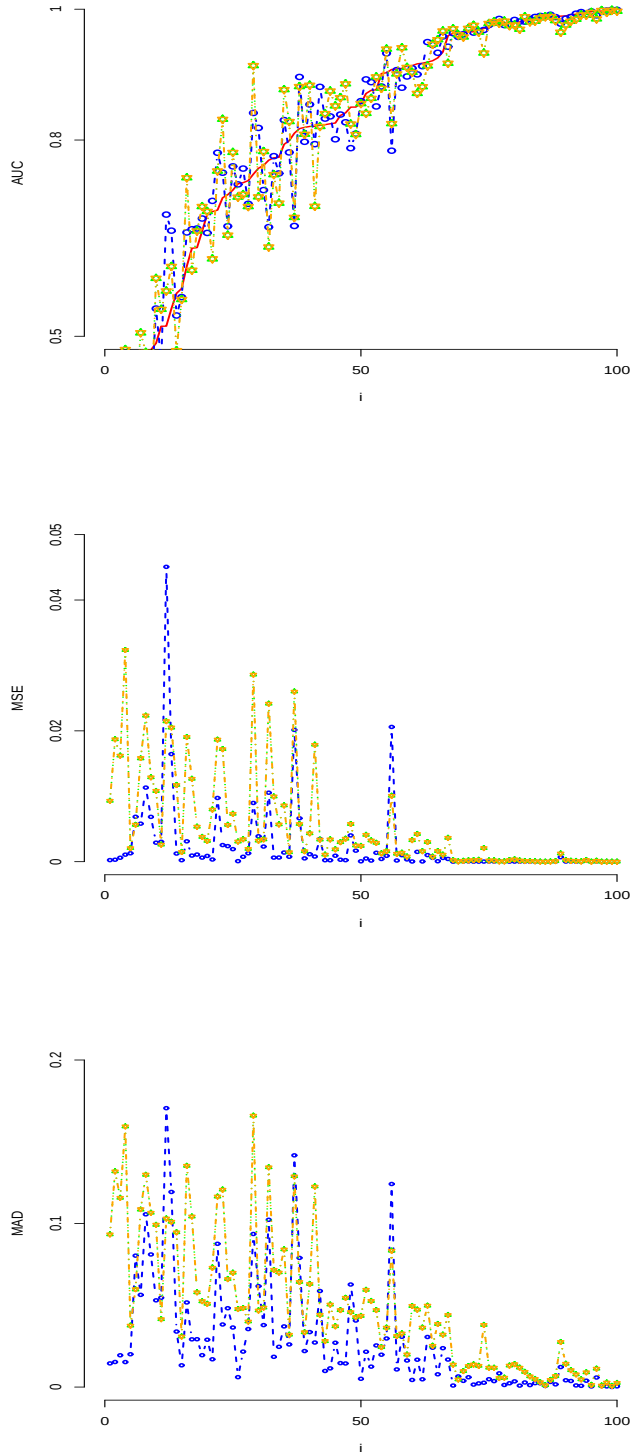


Figure 10. Example 5: Comparison with $n_1 = 300$, $\rho = 0.8$, $q = 6$ *and* $\varepsilon \sim 0.1t(3)$.

Baoying Yang
Department of Statistics, College of Mathematics,
Southwest Jiaotong University, China.
E-mail address: yangbaoying@home.swjtu.edu.cn

Xin Huang
Division of Public Health Sciences,
Fred Hutchinson Cancer Research Center, Seattle, WA, USA
E-mail address: watsonxhuang@gmail.com

Gengsheng Qin
Department of Mathematics and Statistics,
Georgia State University, USA
E-mail address: gqin@gsu.edu