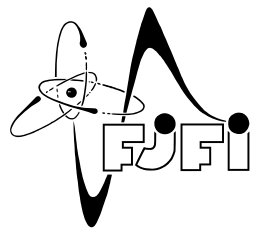
Robust Estimation and Inference in Poisson Regression Models

Jana Novotná

FNSPE, Czech Technical University in Prague

20. 9. 2020



Outline

Goals

- Define a new robust estimator for Poisson regression models – modified median estimator
  - Use of robust estimators for logistic regression models
  - Use of median estimator for Poisson regression models
- Examining the accuracy of the new estimator
  - Presence of outliers
  - Presence of leverage points
- Hypothesis testing

Poisson Regression Model

- Generalized linear model
- It allows us to model a quantity with a Poisson distribution
  - Number of complaints about doctors per year
  - Number of unemployed in an area

Poisson Regression Model

- $Y_1, Y_2, \ldots, Y_n$ independent variables, $Y_i \sim \mathrm{Po}(\mu_i)$

$$\ln \mu_i = \ln s_i + \ln \lambda_i = \ln s_i + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}, \quad i = 1, 2, \ldots, n$$

- $g(\mu_i) = \ln(\mu_i)$ – link function
- $\mu_i = s_i \lambda_i$ – including of a sample size

Median Estimator – Introduction

- When working with discrete random variables, there is a problem with detecting small changes in the product $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$
- We convert discrete random variables to continuous random variables using statistical smoothing

$$Z = Y + U$$

$$Y \sim \mathrm{Po}(\lambda), \quad U \sim \mathcal{U}(0,1)$$

- The median function for the quantity Z is obtained by expressing z from the equation

$$F_Z(z;\lambda) = \frac{1}{2}$$

Median Estimator – Median Function

$$m(\lambda) = \begin{cases} \dfrac{1}{2}e^{\lambda}, & 0 < \lambda < C_0, \\ k + \dfrac{k!}{\lambda^k}\left(\dfrac{1}{2}e^{\lambda} - \displaystyle\sum_{j=0}^{k-1}\dfrac{\lambda^j}{j!}\right), & \lambda \in \langle C_{k-1}, C_k), \ k \in \mathbb{N}, \end{cases}$$

where C_k is the positive solution of the equation

$$e^{-\lambda}\sum_{j=0}^{k}\frac{\lambda^j}{j!} = \frac{1}{2}$$

for unknown λ and parameter $k \in \mathbb{N}_0$.

Median Estimator

- Median estimator is defined by

$$\hat{\boldsymbol{\beta}}_{\mathrm{M}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} |Y_i + U_i - m(\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}))|,$$

where

$$U_i \sim \mathcal{U}(0,1), \quad i = 1, 2, \ldots, n,$$

$$\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) = \exp\{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\}, \quad i = 1, 2, \ldots, n.$$

Modified Median Estimator – Derivation

- We first define the k-enhanced median estimator by

$$\hat{\boldsymbol{\beta}}_{\mathrm{M},k} = \arg\min_{\boldsymbol{\beta}} \frac{1}{k} \sum_{i=1}^{n} \sum_{j=1}^{k} |Y_i + U_{ij} - m(\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}))|.$$

- Next we use the law of large numbers and for $k \to +\infty$ we get

$$\frac{1}{k} \sum_{j=1}^{k} |Y_i + U_{ij} - m(\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}))| \xrightarrow{\mathrm{P}} \mathrm{E}_U |Y_i + U - m(\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}))|$$

$$= \int_0^1 |Y_i + u - m(\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}))| \, \mathrm{d}u.$$

Modified Median Estimator

- Modified median estimator is defined by

$$\hat{\boldsymbol{\beta}}_{\mathrm{MM}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \int_0^1 |Y_i + u - m(\lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}))| \, \mathrm{d}u$$

Asymptotic Properties

Theorem

Under the assumption the modified median estimator $\hat{\boldsymbol{\beta}}_{\mathrm{MM}}$ is consistent estimator of $\boldsymbol{\beta}$, its asymptotic distribution is given by

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{\beta}^0\right) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\left(\mathbf{0}_p, \boldsymbol{V}(\boldsymbol{\beta}^0)\right).$$

Hypothesis Testing

- We would like to test

$$H_0 : \boldsymbol{K}^{\mathrm{T}}\boldsymbol{\beta}^0 = \boldsymbol{m} \quad \text{vs.} \quad H_1 : \boldsymbol{K}^{\mathrm{T}}\boldsymbol{\beta}^0 \neq \boldsymbol{m}$$

- $r \leq d$ is number of independent rows in $\boldsymbol{K}^{\mathrm{T}}$
- We define a Wald-type test statistic

$$W_n(\hat{\boldsymbol{\beta}}_{\mathrm{MM}}) = n\left(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{m}\right)^{\mathrm{T}}\left(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{V}}_n(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})\boldsymbol{K}\right)^{-1}\left(\boldsymbol{K}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{MM}} - \boldsymbol{m}\right)$$

- The asymptotic distribution of $W_n(\hat{\boldsymbol{\beta}}_{\mathrm{MM}})$ is $\chi^2(r)$

Simulation Experiments

- We observed in simulations:
  - Maximum likelihood estimator
  - Median estimator
  - Modified median estimator
  - Mallows estimator
- We considered two models and two types of contamination

Basic Model

- The vector of regression coefficients is equal to

$$\boldsymbol{\beta} = (3;\, 1)^{\mathrm{T}}$$

- It holds for vectors of explanatory variables

$$x_{i1} = 1, \quad x_{i2} \sim \mathcal{N}(1;\, 0{,}6), \quad x_{i2} \text{ independent}, \quad i = 1, 2, \ldots, n$$

- It holds for independent response variables

$$Y_i \sim \mathrm{Po}(\lambda_i), \quad \lambda_i = \exp\{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}\}, \quad i = 1, 2, \ldots, n$$

Observed Quantities

- We get $N = 1000$ estimates for each choice of the number of observations n and the contamination level ε.
- We observe quantities

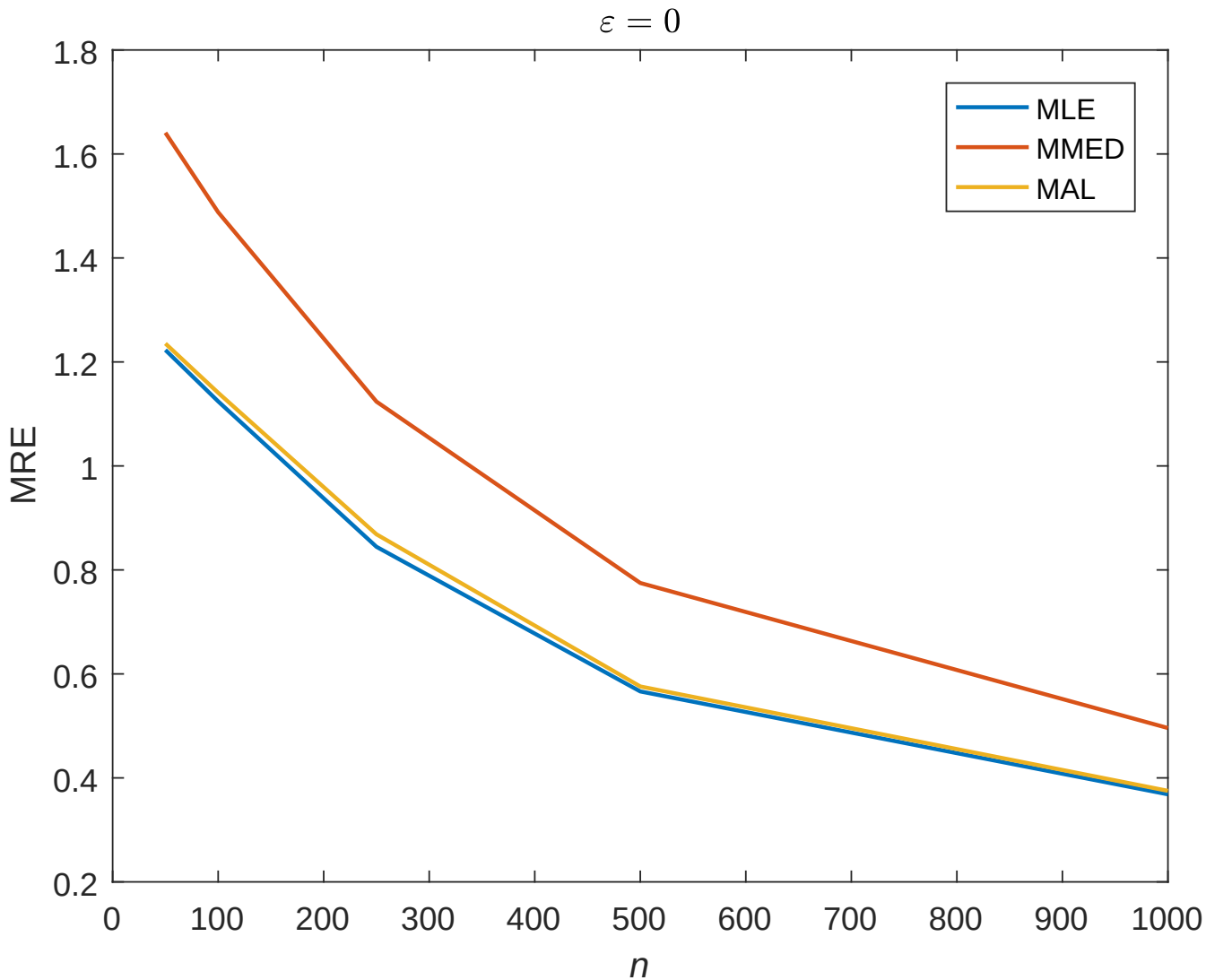
$$Z_j = \frac{1}{2}\left(\frac{|\hat{\beta}_1^{(j)} - \beta_1|}{|\beta_1|} + \frac{|\hat{\beta}_2^{(j)} - \beta_2|}{|\beta_2|}\right) \cdot 100, \quad j = 1, 2, \ldots, N$$

$$\text{MRE} = \bar{Z}_N = \frac{1}{N}\sum_{j=1}^{N} Z_j$$

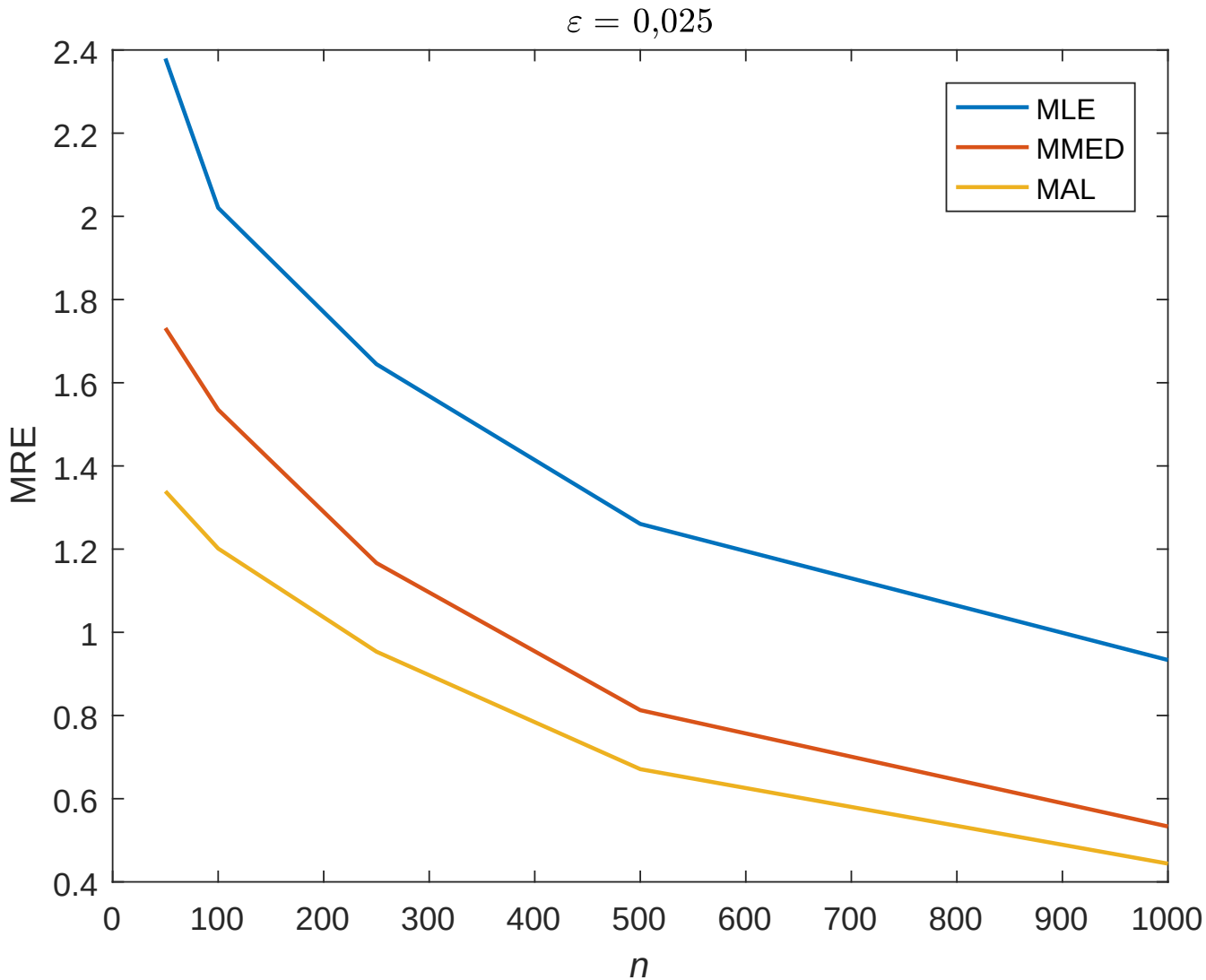
$$\text{s.e.} = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(Z_j - \bar{Z}_N)^2}$$

- We also monitor the convergence of estimates

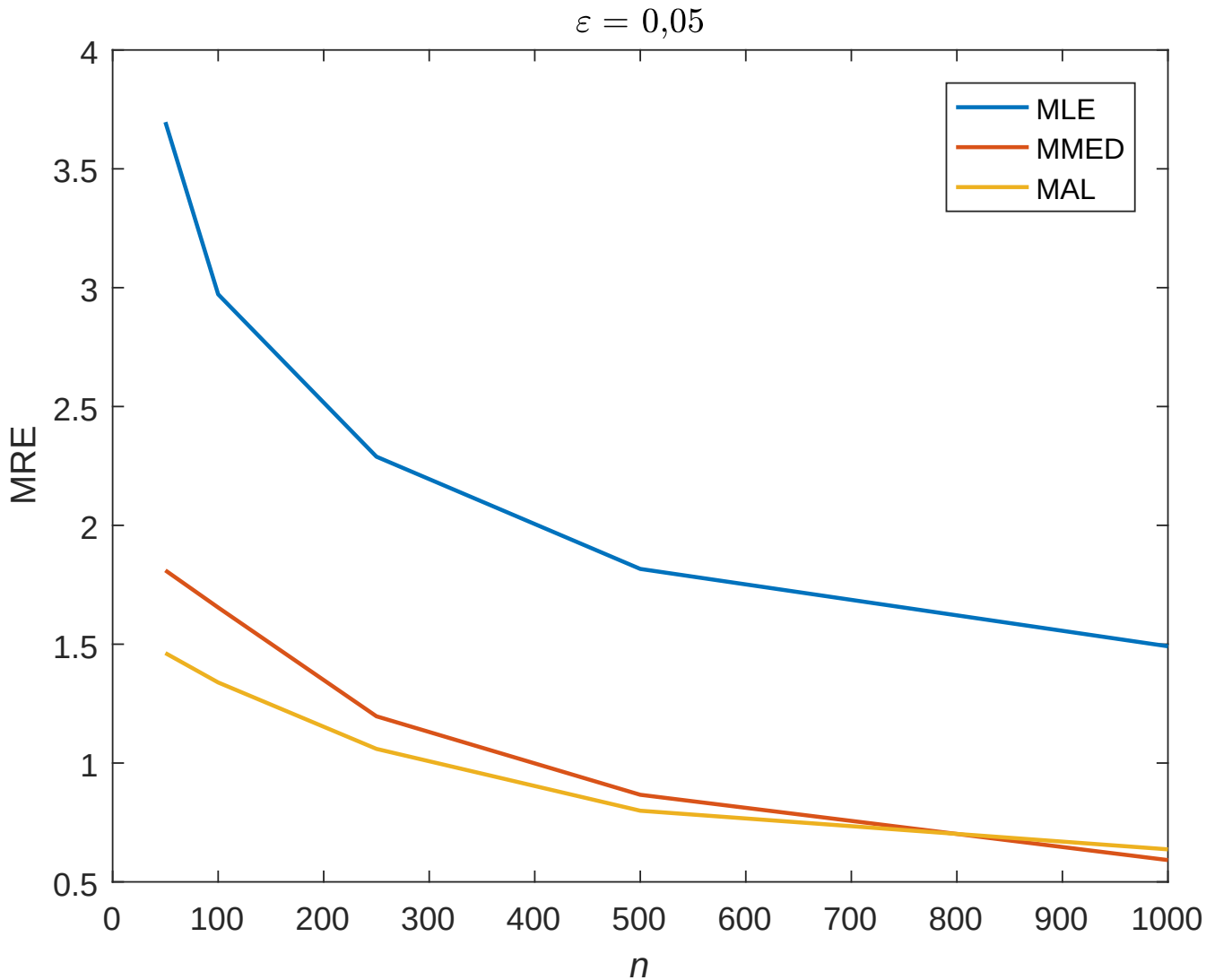
Results – Noncontaminated Data



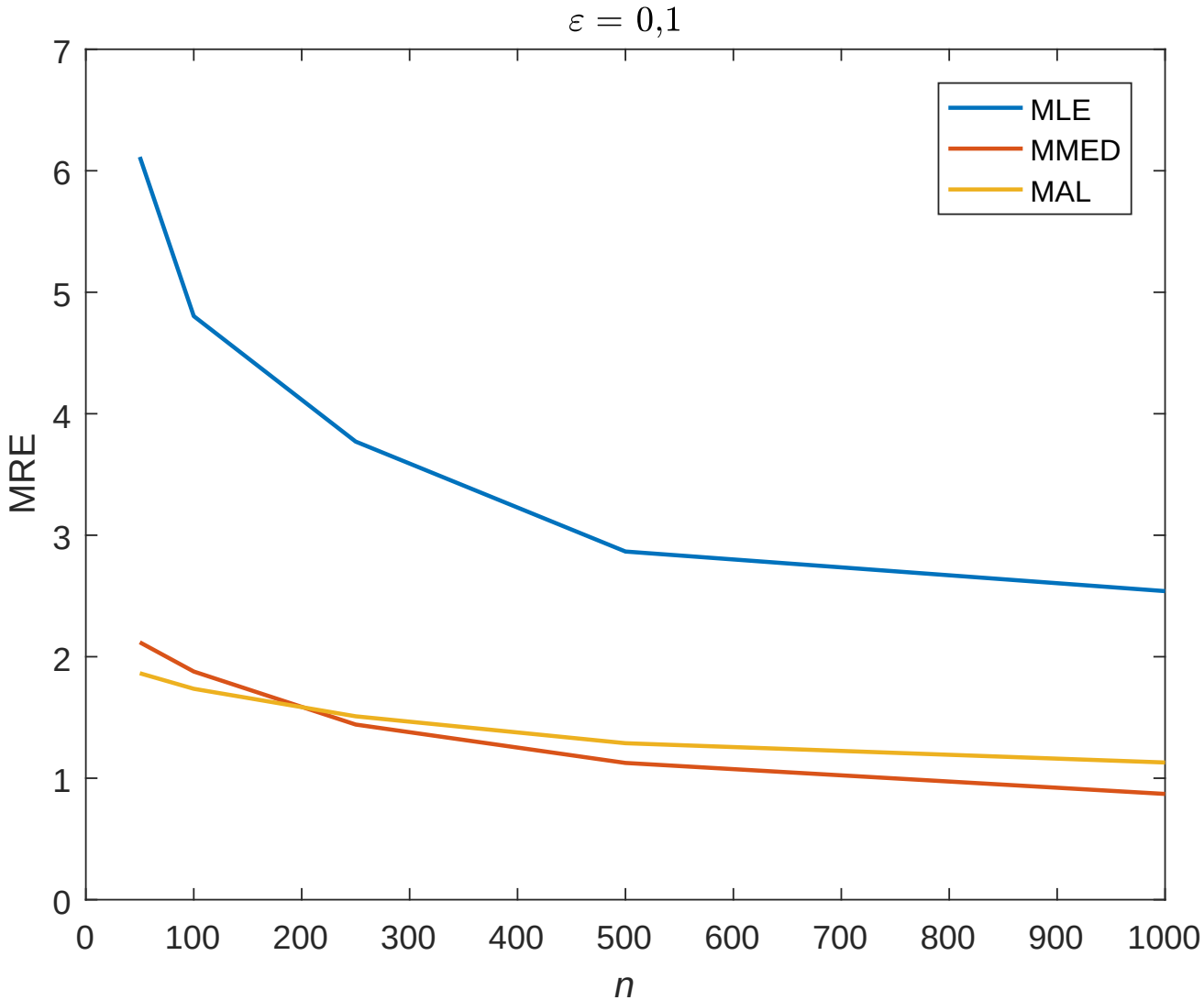
Results – Outliers



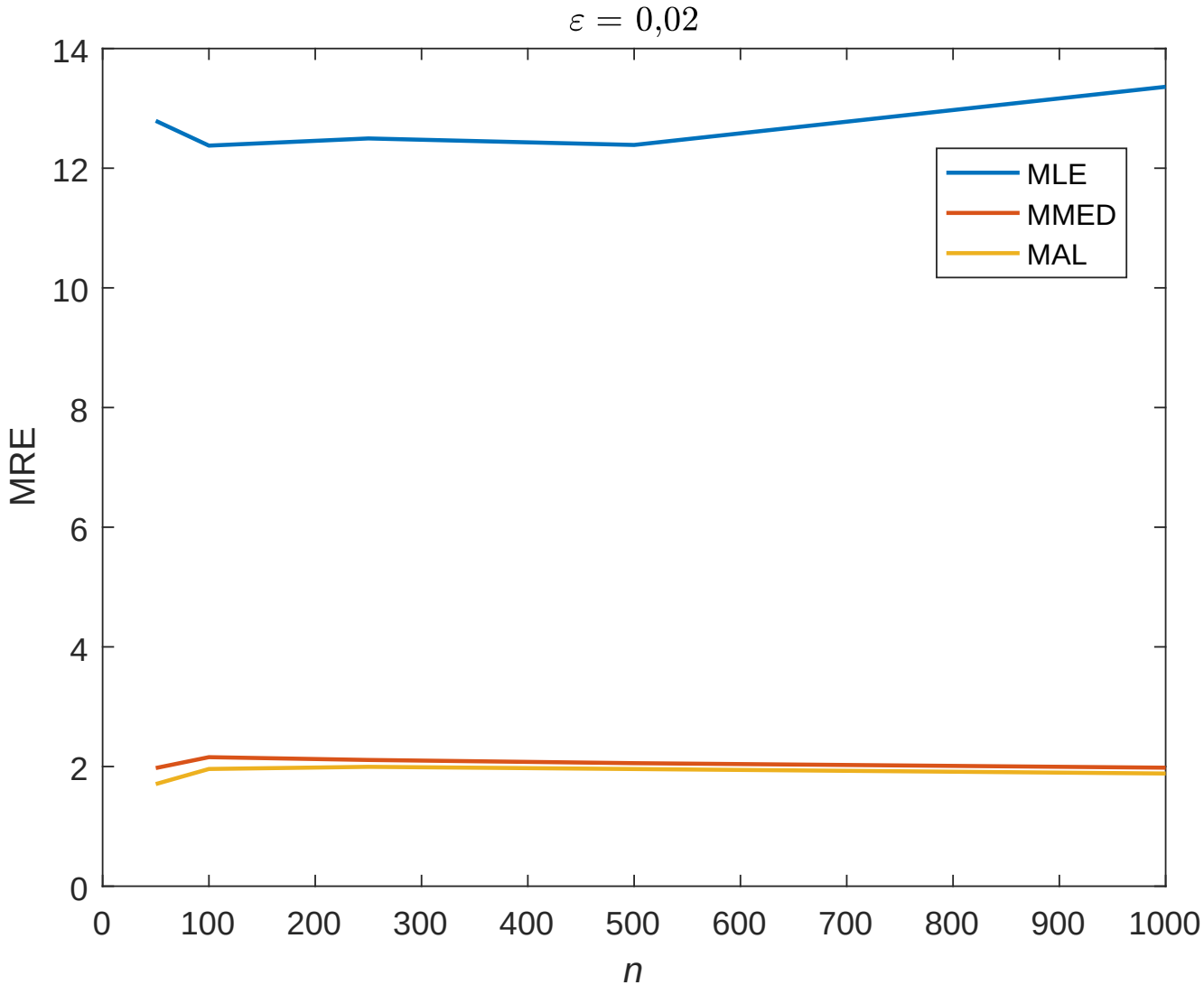
Results – Outliers



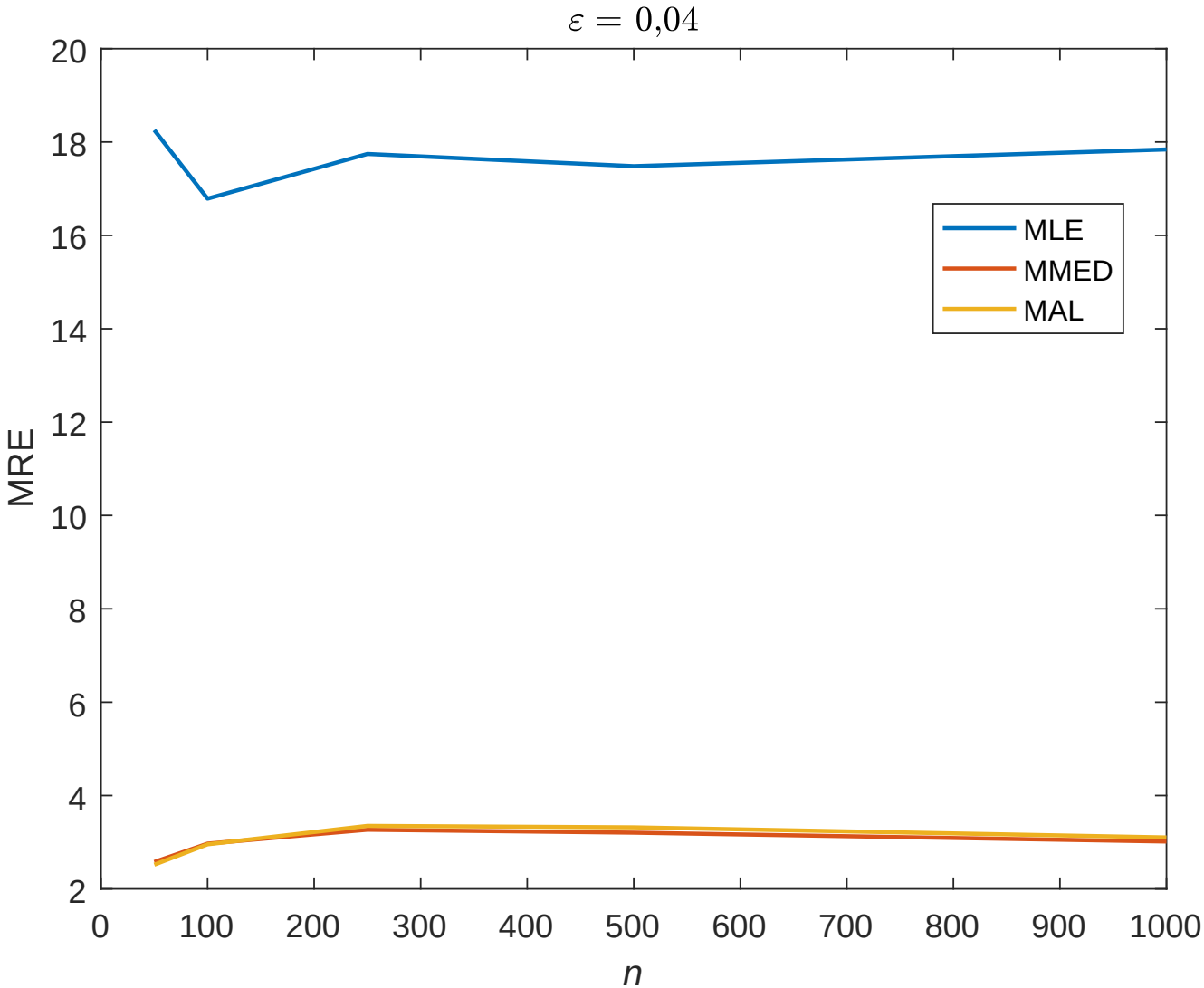
Results – Outliers



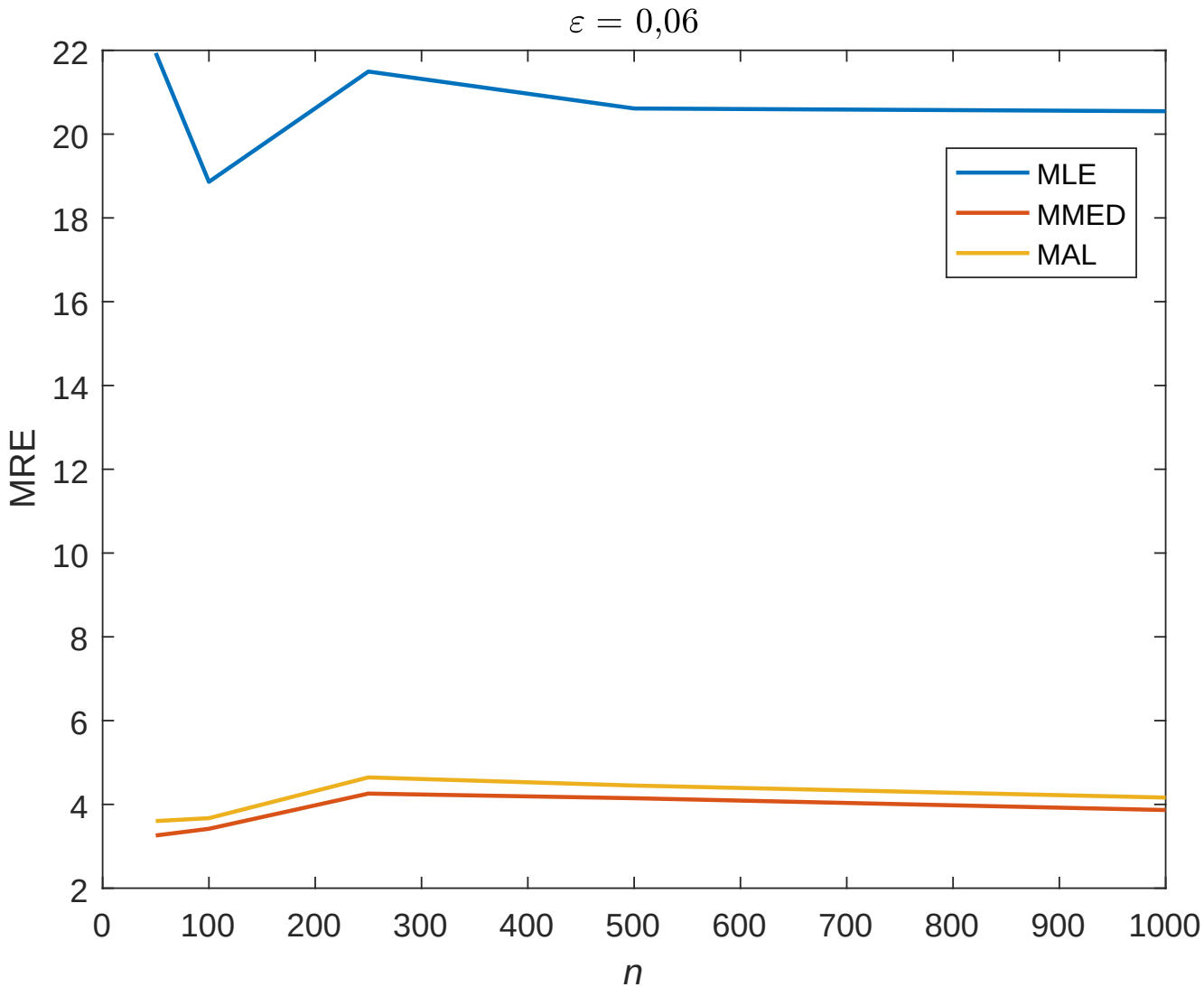
Results – Leverage Points



Results – Leverage Points



Results – Leverage Points



Median Estimator vs. Modified Median Estimator

- For the median estimator, there was a convergence problem in all simulations
- Modified median estimator converged always
- In cases where the median estimator converged, both estimators are similarly accurate
- When using a modified median estimator, we do not have to generate additional random variables

Conclusion

- The newly defined modified median estimator is well applicable in practice
- We improved median estimator in two areas
- The modified median estimator can be more accurate than the Mallows estimator for higher levels of contamination and more observations