Truncated Regression for Heavy-Tailed Data

Reza Alizadeh Noughabi and Adel Mohammadpour

Department of Statistics
Amirkabir University of Technology (Tehran polytechnic)

16 February 2022
23rd workshop on Applied Stochastic Processes

به نام او

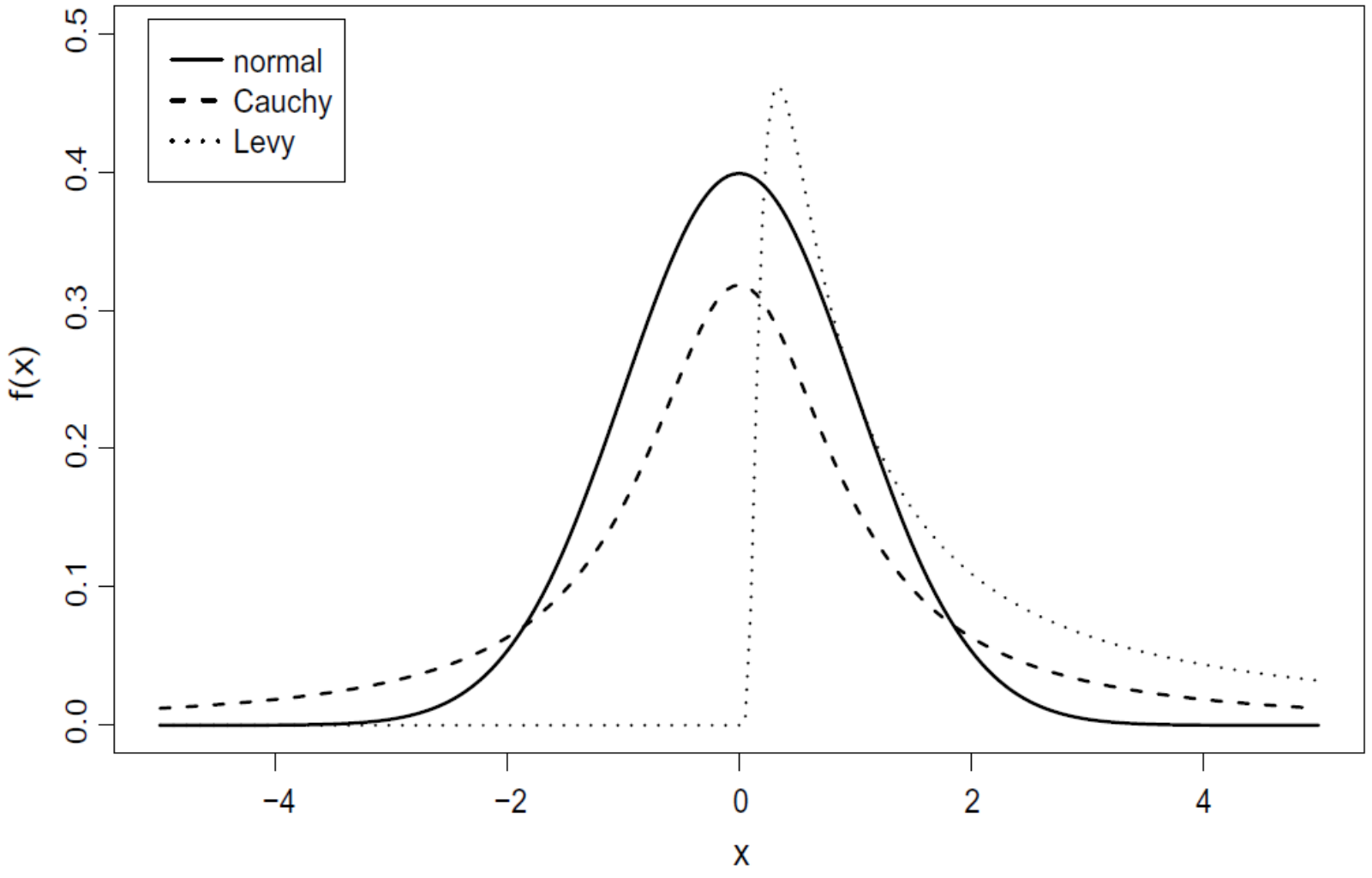
رگرسیون دم‌بریده برای داده‌های دم‌کلفت

رضا علیزاده نوقابی و عادل محمدپور

دانشگاه صنعتی امیرکبیر

روش‌های سنتی رگرسیون نسبت به داده‌های دورافتاده یا اصطلاحاً دم‌کلفت استوار نیستند. وجود داده‌های دورافتاده ناشی از دم‌کلفت بودن توزیع خطای مدل رگرسیون است. تاکنون روش‌های متنوعی با عنوان رگرسیون استوار ارائه شده است. برخی از آن‌ها با فرض غیرگاوسی بودن خطای مدل سعی به حل پارامتری مسئله کرده‌اند. ولی هیچ‌یک از آن‌ها به‌صورت تحلیلی نتوانسته‌اند مسئله را در حالت کلی حل کنند یا آنکه برای همه‌ی بخش‌های الگوریتم ارائه شده دلایل ریاضی ارائه کنند. هدف ما در این سخنرانی ارائه راه‌حلی تحلیلی برای مسئله رگرسیون پایدار است. به‌نحوی که همه اجزاء الگوریتم ارائه شده را به‌صورت ریاضی اثبات کنیم. برای این منظور ما از خواص جالب آمارهای ترکیبی توزیع‌های پایدار بهره گرفته‌ایم، که به ما اجازه می‌دهد امید ریاضی یا واریانس بسیاری از آن‌ها را محاسبه کنیم. در صورتی که واریانس توزیع پایدار غیر گاوسی اصلاً وجود ندارد. برای این کار ابتدا یک رگرسیون معمولی با استفاده از داده‌ها محاسبه و داده‌ها را بر اساس مانده‌ها مرتب می‌کنیم. سپس داده‌هایی را در نظر می‌گیریم که واریانس آمارهای ترتیبی آن‌ها متناهی است. حال با استفاده از تابع چگالی آمارهای ترتیبی، به برآورد پارامترها می‌پردازیم و بهترین برآوردگر خطی نااریب را محاسبه می‌کنیم. این راه حل را می‌توانیم به حالت چند متغیر نیز توسعه دهیم.

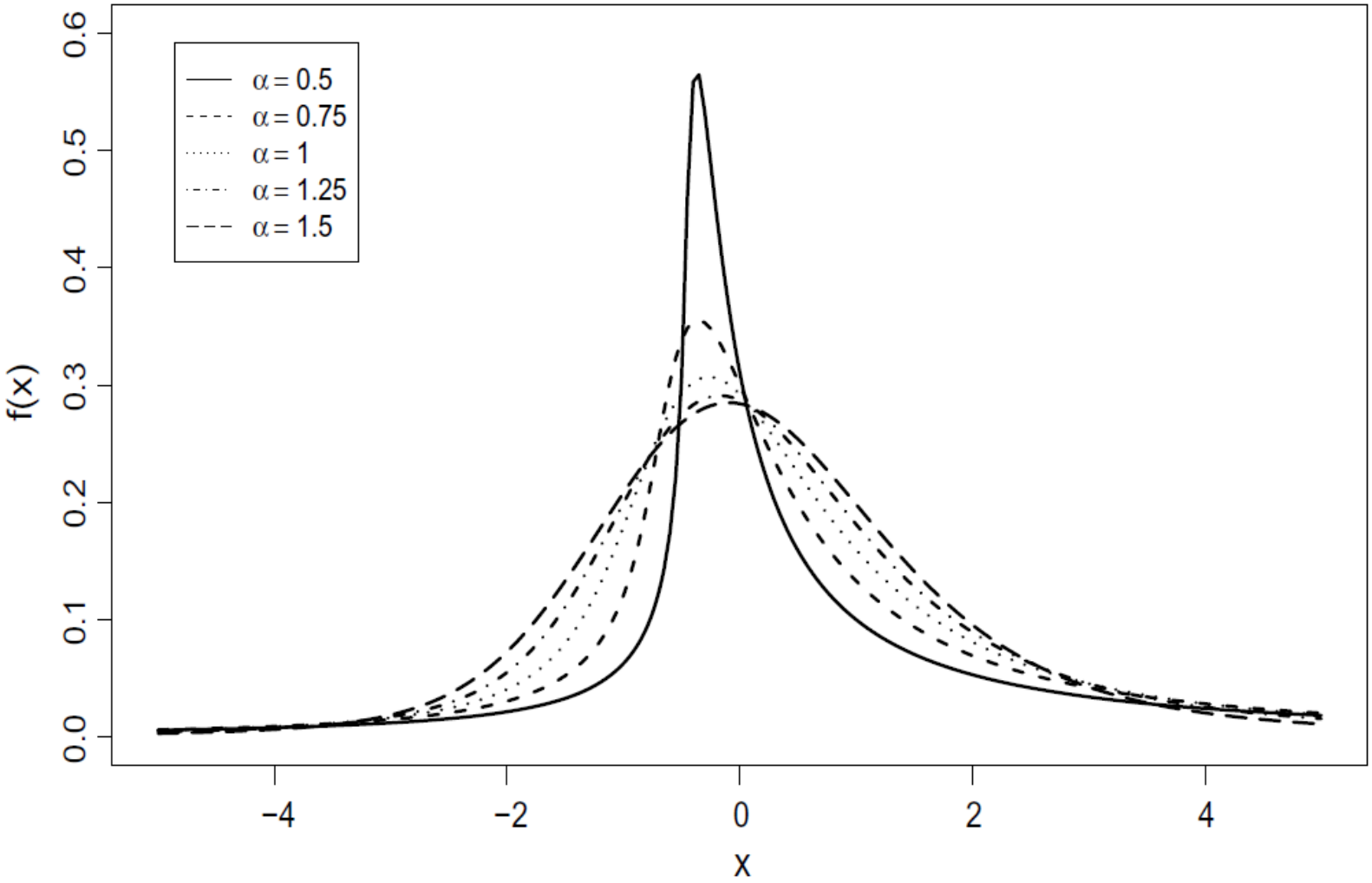
Tail



Tail index, Nolan (2020)

		symmetric	non-symmetric
All moments exist	no tails	Uniform(−1,1)	Uniform(0,1)
	light tails	normal finite mixtures of normals symmetric Laplace	exponential, skewed normal skewed Laplace
Limited moments exist	finite variance	t (d.f. > 2)	Pareto, $\alpha > 2$
	infinite variance	sym. stable ($\alpha < 2$) t (d.f. < 2)	non-sym. stable ($\alpha < 2$) Pareto, $\alpha < 2$

Tail index



Heavy-Tailed Stable Distributions

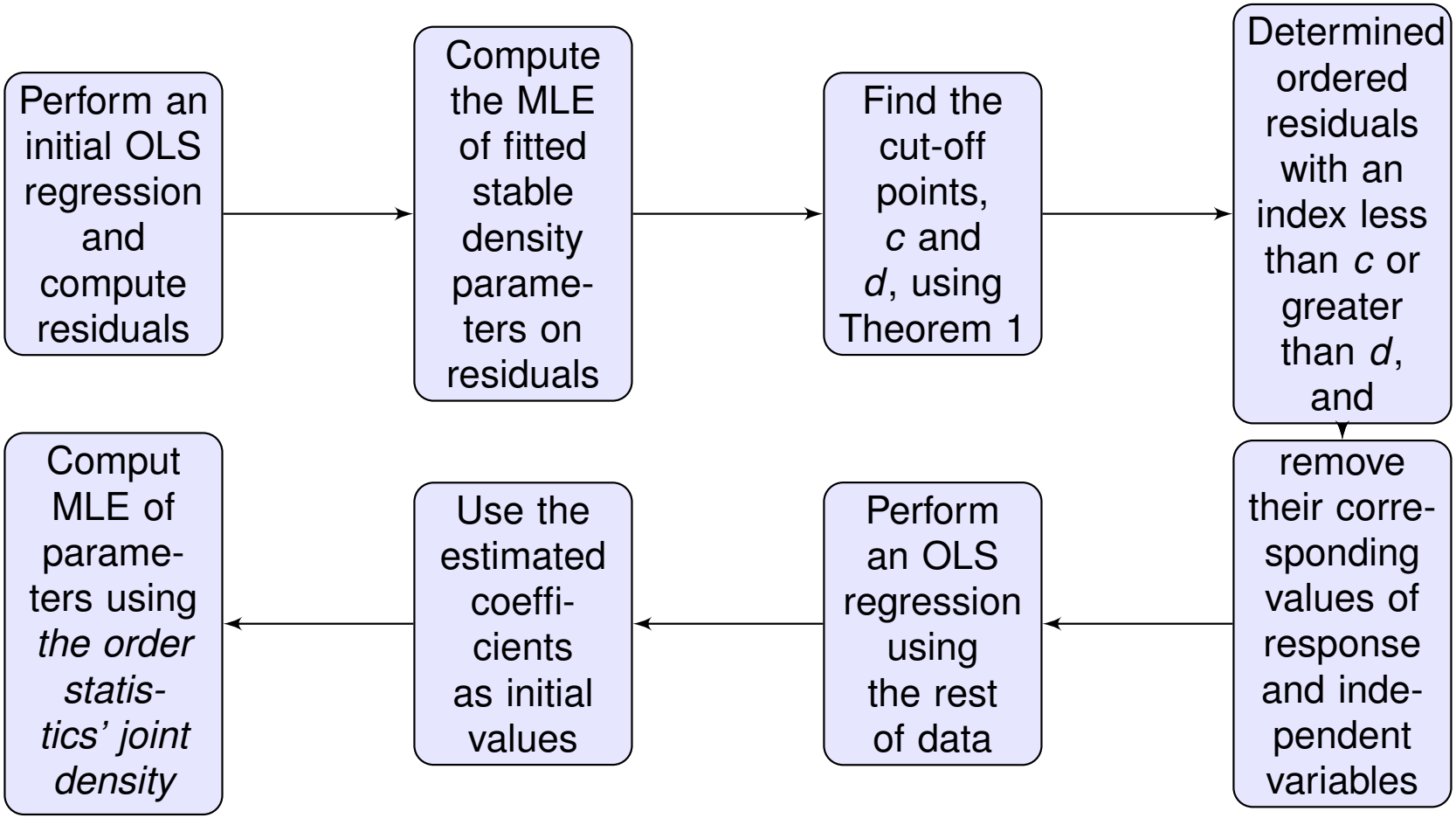
Definition

A random variable X is stable if and only if $\varphi_U(t)$ as follows

$$\varphi_U(t) = \begin{cases} \exp\left\{-\gamma^\alpha |t|^\alpha \left[1 - i\beta\left(\tan\frac{\pi\alpha}{2}\right)(\mathit{sign}\, t)\right] + i\delta t\right\}; & \alpha \neq 1, \\ \exp\left\{-\gamma\,|t|\left[1 + i\beta\frac{2}{\pi}(\mathit{sign}\, t)\log|t|\right] + i\delta t\right\}; & \alpha = 1, \end{cases}$$

where $0 < \alpha \leq 2$, $-1 \leq \beta \leq 1$, $\gamma > 0$, $\delta \in \mathbb{R}$.

Truncated Regression



Tail index and Mean

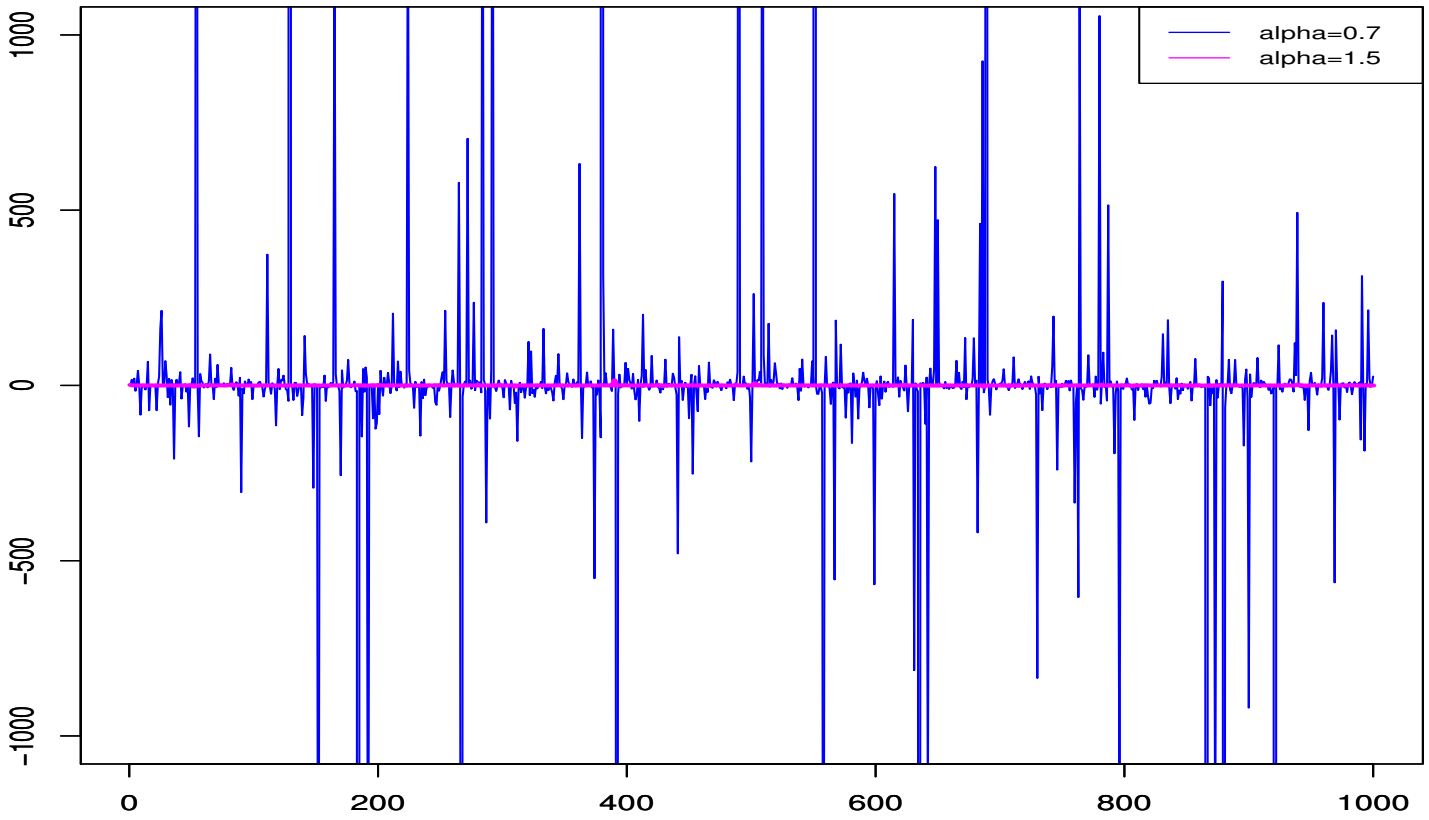


Figure: Mean of random samples from stable distribution in 1000 iterations with sample size $n = 100$.

Tail index and Standard Deviation

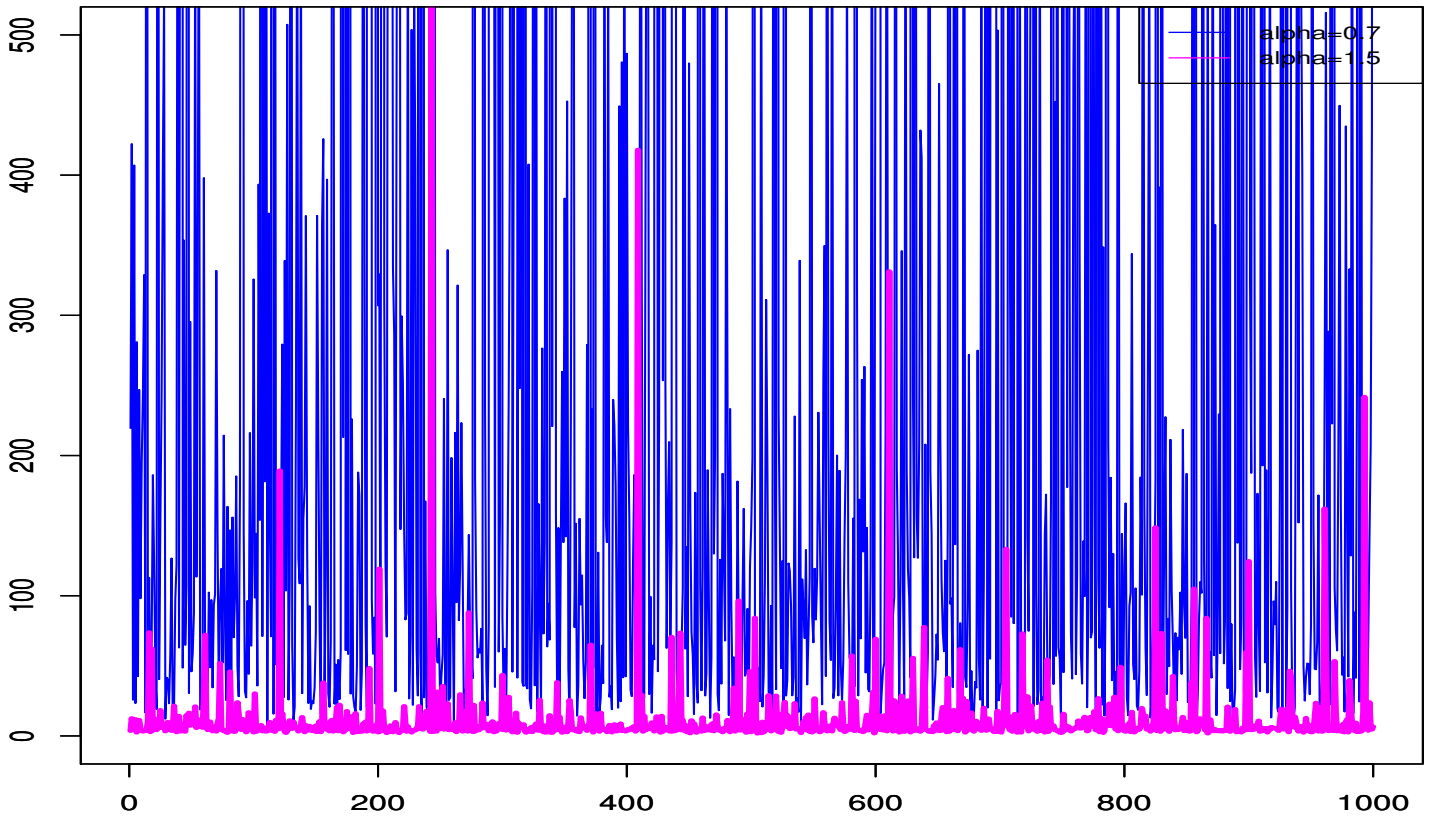


Figure: Standard deviation of random samples from stable distribution in 1000 iterations with sample size $n = 100$.

Order Statistics Moments

Theorem

Let $X_1, \ldots, X_n$ be a random sample from non-Gaussian standard stable distribution and $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ be its corresponding order statistics. Take $c = \frac{2}{\alpha}$ and $d = n + 1 - \frac{2}{\alpha}$.

(I) Suppose $-1 < \beta < 1$. In order that $E\left(X_{k:n}^2\right)$ exists, it is necessary and sufficient that $c < k < d$.

(II) Suppose $\alpha \geq 1$ and $\beta = 1$ or $\beta = -1$ so that $E\left(X_{k:n}^2\right)$ exists, then it is sufficient that $c < k < d$.

(III) Suppose $\alpha < 1$ and $\beta = 1$ or $\beta = -1$. In order that $E\left(X_{k:n}^2\right)$ exists, it is necessary and sufficient that $k < d$ or $c < k$, respectively.

Regression

Consider the standard regression model

$$Y_{ij} = \theta x_i + U_{ij} \; ; i = 1, \ldots, n \, , \, j = 1, \ldots m,$$

where Y_{ij} is dependent variable (or response variable) with m replication, x_i is independent variable (or predictor variable), θ is an unknown parameter and U_{ij}'s are independent identically distributed random variables.

Regression with Stable Errors, Blattberg and Sargent Method

Blattberg and Sargent (1971) use another method (except OLS) for estimating θ with stable's errors when $\beta = 0$. They show that the Best Linear Unbiased Estimator (BLUE) of θ has the following form:

$$\hat{\theta}(\alpha) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} |X_i|^{\frac{1}{\alpha-1}} sign(x_i)\, Y_{ij}}{\sum_{i=1}^{n} |X_i|^{\frac{\alpha}{\alpha-1}}}, \qquad 1 < \alpha < 2.$$

Ranked Set Sampling

Table: Display of m random sample for constructing RSS for each x_i

V_1:	$\boldsymbol{Y_{i,1(1)}}$	$Y_{i,1(2)}$	...	$Y_{i,1(m)}$
V_2:	$Y_{i,2(1)}$	$\boldsymbol{Y_{i,2(2)}}$	...	$Y_{i,2(m)}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
V_m:	$Y_{i,m(1)}$	$Y_{i,m(2)}$	...	$\boldsymbol{Y_{i,m(m)}}$

For simplicity, we denoted $Y_{i,j(j)}$ by $Y_{i(j)}$.

BLUE Using Ranked Set Sampling, Dorniani et al.

Theorem

Consider model (1) and $Y_{i(j)}$ be a RSS from $S(\alpha, 0, 1, 0)$ distribution, Then the BLUE of the θ is given by

$$\tilde{\theta}(\alpha) = \sum_{i=1}^{n} \frac{c^* + e^* x_i}{\Delta^*} \sum_{j \in J} \frac{Y_{ij}}{\tau_j},$$

$$c^* = -n \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{j \in J} \frac{1}{\tau_j} \right) \left(\sum_{j \in J} \frac{\eta_j^2}{\tau_j} \right),$$

$$e^* = n^2 \left(\sum_{j \in J} \frac{1}{\tau_j} \right) \left(\sum_{j \in J} \frac{\eta_j^2}{\tau_j} \right),$$

$$\Delta^* = n \left(\sum_{j \in J} \frac{1}{\tau_j} \right)^2 \left(\sum_{j \in J} \frac{\eta_j^2}{\tau_j} \right) \left[n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right]$$

Regression

The linear regression model can be written as

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

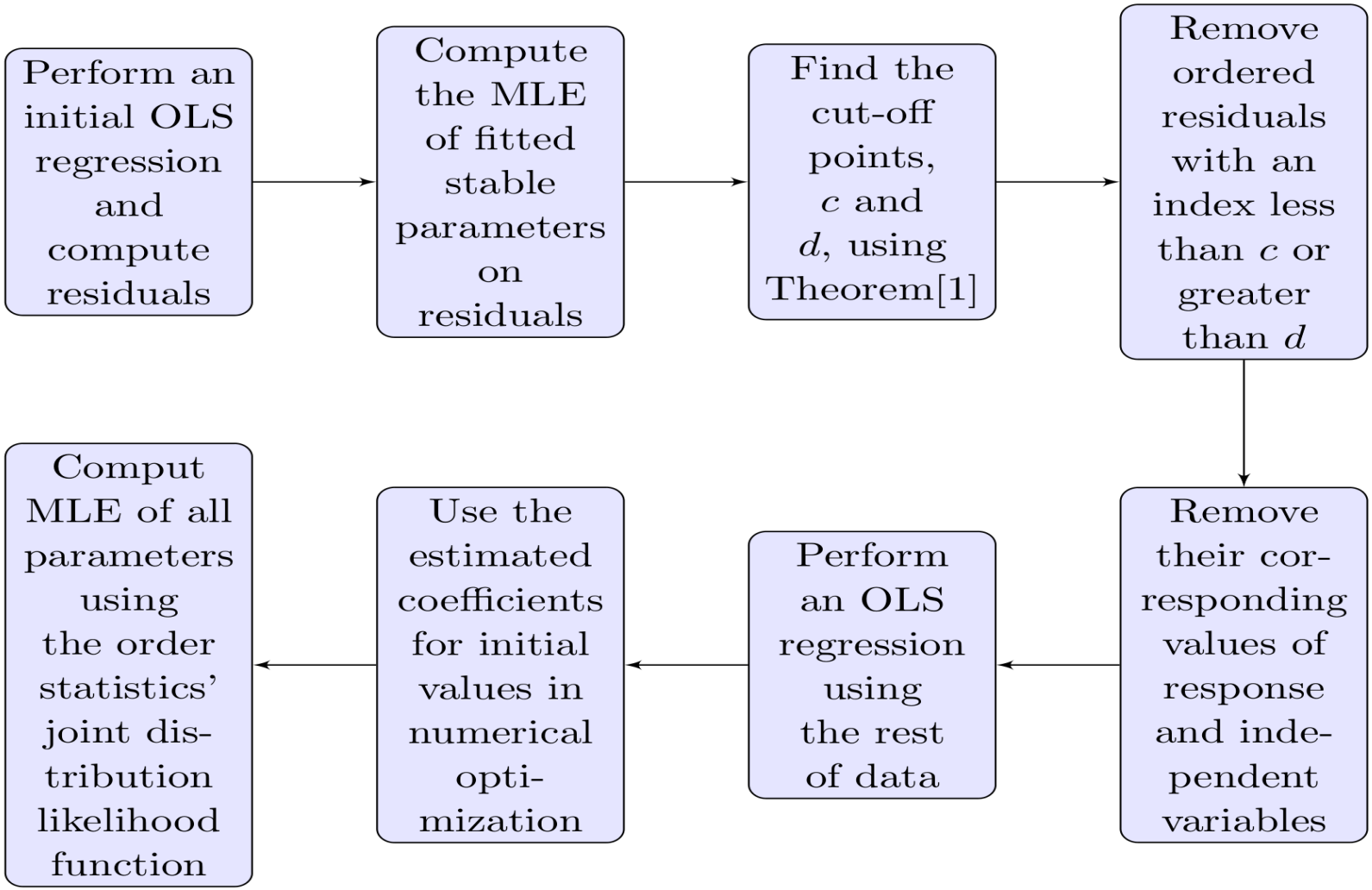
where $\boldsymbol{x} = (x_{i,j})_{n\times k}$ is a design matrix, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ are the regression coefficients and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ are independent and identically distributed random variables. In OLS, ϵ_i must be normally distributed, but we assume that ϵ_i follows a non-Gaussian stable distribution as U.

NOR Method

- NOR's method performs an OLS.
- Trims 10% of low and high residuals leverages, and retrieves 80% of observations.
- Based on this new data set, they perform a trimmed OLS fit, and initial values for all parameters are obtained; i.e., initial values for regression coefficients and the initial estimate of the stable parameters, which is considered as an error term.
- Using a numerical optimization, they find the maximum likelihood estimator of parameters.

NOR Method

Their method chooses fixed cut-off points so that at least the lowest and highest 10% of outliers and leverage points are trimmed without considering the values of the tail index and skewness of data.
We propose an effective procedure to calculate cut-off points based on the tail index and skewness parameters.


Perform an initial OLS regression and compute residuals
Compute the MLE of fitted stable parameters on residuals
Find the cut-off points, c and d, using Theorem[1]
Remove ordered residuals with an index less than c or greater than d
Remove their corresponding values of response and independent variables
Perform an OLS regression using the rest of data
Use the estimated coefficients for initial values in numerical optimization
Comput MLE of all parameters using the order statistics' joint distribution likelihood function

Maximum likelihood based on order statistics

According to the above-mentioned Theorem, we propose the following algorithm

- Perform an initial OLS regression and compute residuals.
- Compute the MLE of parameters based on residuals.
- Use above Theorem and estimated parameters to find the cut-off points c and d, i.e., the index of order statistics with finite variance.
- Remove ordered residuals with an index less than c or greater than d and their corresponding values in $\boldsymbol{y}$ and $\boldsymbol{x}$. Perform an OLS regression using this new data.

Maximum likelihood based on order statistics

- Compute MLE of parameters using the following likelihood function based on order statistics. Use the estimated coefficients of the previous step for initial values for numerical optimization:

$$
\begin{aligned}
&\log\left(f_{c,\ldots,d}\left(u_c,\ldots,u_d\right)\right) \\
&\quad = \log\left(\frac{n!}{(c-1)!(n-d)!}\right) + (i-1)\log\left(F_U\left(y_{[c]}-\theta_1-\theta_2 x_{[c]}\right)\right) \\
&\qquad + (n-d)\log\left(1-F_U\left(y_{[d]}-\theta_1-\theta_2 x_{[d]}\right)\right) \\
&\qquad + \sum_{i=c}^{d}\log\left(f_U\left(y_{[i]}-\theta_1-\theta_2 x_{[i]}\right)\right).
\end{aligned}
$$

Trimmed least squares estimator

Using the ordered residuals, we can obtain the trimmed least squares estimators.

$$\begin{bmatrix} Y_{[1]} \\ \vdots \\ Y_{[n]} \end{bmatrix} = \begin{bmatrix} 1 & X_{[1]1} & X_{[1]2} & \cdots & X_{[1]k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{[n]1} & X_{[n]2} & \cdots & X_{[n]k} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_k \end{bmatrix} + \begin{bmatrix} U_{1:n} \\ \vdots \\ U_{n:n} \end{bmatrix},$$

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{V},$$

where $\boldsymbol{Z} = [\mathbf{1}, Z_1, \ldots, Z_k]$, $Z_i = \left(X_{[c]i}, \ldots, X_{[d]i}\right)'$, and $\boldsymbol{V} = \left(Y_{[c]}, \ldots, Y_{[d]}\right)'$.

Trimmed least squares estimator

Since, the moment of order statistics are exist, bias and variance of $\hat{\boldsymbol{\theta}}$ are calculated as follows:

$$\begin{aligned} E\left(\hat{\boldsymbol{\theta}}\right) &= E\left((\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{V}\right) \\ &= (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'E\left(\boldsymbol{Z}\boldsymbol{\theta}+\boldsymbol{U}_t\right) \\ &= \boldsymbol{\theta}+(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'E\left(\boldsymbol{U}_t\right), \end{aligned}$$

$$Var\left(\hat{\boldsymbol{\theta}}\right) = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\Sigma_t\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1},$$

where $\boldsymbol{U}_t = (U_{c:n}, \ldots, U_{d:n})$ and Σ_t are trimmed residual vector and its covariance matrix, respectively.

Trimmed least squares estimator

an unbiased estimator (ULS) for regression coefficients is proposed as follows

$$\begin{aligned}\hat{\boldsymbol{\theta}}_u &= (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{V} - (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'E(\boldsymbol{U}_t) \\ &= (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'(\boldsymbol{V} - E(\boldsymbol{U}_t)).\end{aligned}$$

Best linear unbiased estimator

Let $\boldsymbol{\theta}$ be a vector of regression coefficients and γ be a linear combination of $\boldsymbol{\theta}$ components as follows:

$$\gamma = l_0\theta_0 + l_1\theta_1 + \cdots + l_k\theta_k = \boldsymbol{l}'\boldsymbol{\theta},$$

where $\boldsymbol{l}$ is a known vector of constant real values. Then $\hat{\gamma} = \boldsymbol{a}'\boldsymbol{V}$ is the best linear unbiased estimator for γ, where

$$\boldsymbol{a} = \Sigma_t^{-1}\boldsymbol{Z}\boldsymbol{\lambda} + \Sigma_t^{-1}E(\boldsymbol{U}_t)\boldsymbol{\eta},$$

$$\boldsymbol{\eta} = \boldsymbol{A}^{-1}(E(\boldsymbol{U}_t))'\Sigma_t^{-1}\boldsymbol{Z}\left(\boldsymbol{Z}'\Sigma_t^{-1}\boldsymbol{Z}\right)^{-1}\boldsymbol{l},$$

$$\boldsymbol{\lambda} = \left(\boldsymbol{Z}'\Sigma_t^{-1}\boldsymbol{Z}\right)^{-1}\boldsymbol{l} - \left(\boldsymbol{Z}'\Sigma_t^{-1}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\Sigma_t^{-1}E(\boldsymbol{U}_t)\boldsymbol{\eta},$$

$$\boldsymbol{A} = (E(\boldsymbol{U}_t))'\Sigma_t^{-1}\boldsymbol{Z}\left(\boldsymbol{Z}'\Sigma_t^{-1}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\Sigma_t^{-1}E(\boldsymbol{U}_t) - (E(\boldsymbol{U}_t))'\Sigma_t^{-1}E(\boldsymbol{U}_t).$$

Simulation study

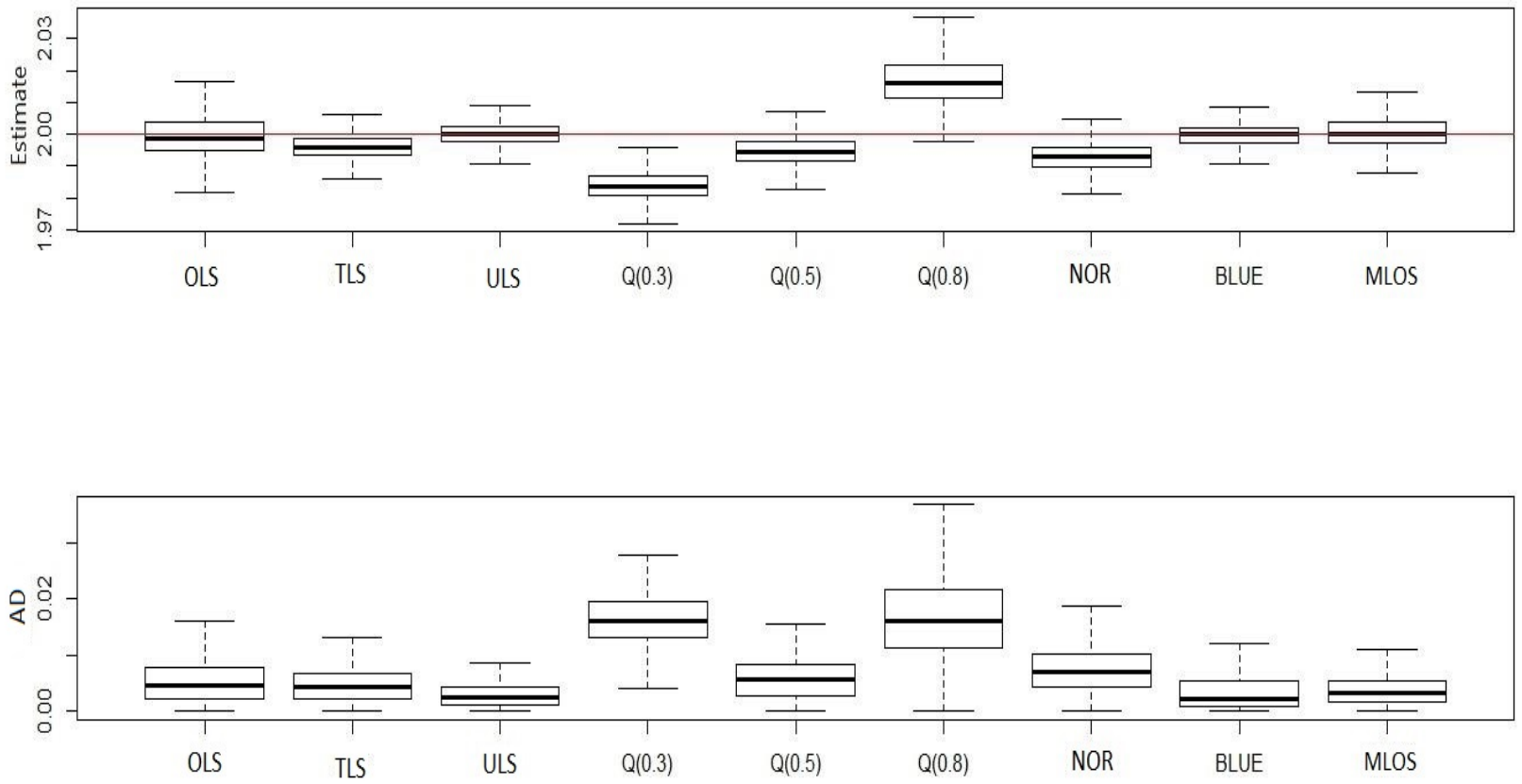


Figure: Box plot of estimation (top) and their AD (bottom) of regression coefficients when errors are simulated from $S(1.5, 0.5)$.

The multivariate linear regression model has the form

$$\boldsymbol{Y} = \boldsymbol{Y}\boldsymbol{\Theta} + \boldsymbol{E}$$

where $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m)_{n\times m}$ is the response matrix and $\boldsymbol{Y}_k = (Y_{1k}, \ldots, Y_{nk})'$ is the k-th response vector, $\boldsymbol{X}_{n\times(p+1)}$ is a design matrix, $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ is the matrix of regression coefficients, i.e. $\boldsymbol{\theta}_k = \left(\theta_{0k}, \theta_{1k}, \ldots, \theta_{pk}\right)'$ is the k-th coefficient vector, and $\boldsymbol{E} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_m)$ is the error terms matrix and $\boldsymbol{\epsilon}_i = (\epsilon_{1i}, \ldots, \epsilon_{ni})'$ that usually normally distributed, i.e. E has multivariate normal distribution. In this paper, we assume that the E has a multivariate stable distribution.

Algorithm 1.
(1) Perform an initial multivariate OLS regression and compute residuals.
(2) Sum each row of residual matrix and compute the maximum likelihood estimator (MLE) of parameters based on it.
(3) Ordered summation of each row, and by using Theorem 1, find the cut-off points c and d, i.e., the indices of order statistics with the finite variance.
(4) Remove the ordered residuals with an index less than c or greater than d and their corresponding values in $\boldsymbol{Y}$ and $\boldsymbol{X}$. Trimmed Least Square (TLS), Unbiased Least Square (ULS) and the Best Linear Unbiased Estimator (BLUE).
(5) Compute the biased and the variance of parameters using the density function (4).

Conclusion

Result 1

For estimation of regression coefficients, if the distribution of error term is symmetric, we prefer MLOS, BLUE, $Q(0.5)$, TLS, and then NOR method.

Conclusion

Result 2

- In right skewness distributed error, MLOS, TLS, $Q(0.5)$, and NOR are better than the others, respectively.
- In the cases of left skewness distributed errors, we recommend practitioners to use MLOS, BLUE, NOR, TLS, and $Q(0.5)$, respectively.

Conclusion

Result 3

In general, we can say that simulation results show that in most cases, MLOS and BLUE are better than the other methods. After this methods, $Q(0.5)$, TLS and then NOR method for estimation of regression coefficients.

Conclusion

AUT Journal of Mathematics and Computing

AUT J. Math. Com., 3(1) (2022) 77-91
DOI: 10.22060/AJMC.2021.20246.1062

Comparing regression methods with non-Gaussian stable errors

Reza Alizadeh Noughabi[a], Adel Mohammadpour*[a]

[a]*Department of Statistics, Faculty of Mathematics and Computer Science, Amirkabir University of Technology (Tehran Tehran, Iran*

ABSTRACT: [16] proposed a regression model with heavy-tailed stable errors. In this paper we extend this method for multivariate heavy-tailed errors. Furthermore, A likelihood ratio test (LRT) for tesing significant of regression coefficients is proposed. Also, confidence intervals based on fisher information for [16] method, called NOR, and LRT are computed and compared with well-known methods. At the end we provide some guidance for various error distributions in heavy-tailed cases.

Review Histo
Received:08 July
Accepted:27 Octo
Available Online:

Keywords:

Fluctuation and Noise Letters
Vol. 21, No. 3 (2022) 2250029 (12 pages)
© World Scientific Publishing Company
DOI: 10.1142/S0219477522500298

World Scientific
www.worldscientific.com

Multivariate Regression with Stable Errors Using Order Statistics

Reza Alizadeh Noughabi and Adel Mohammadpour*
Department of Statistics, Faculty of Mathematics and Computer Science Amirkabir University of Technology (Tehran Polytechnic), Iran
**adel@aut.ac.ir*

Received 3 September 2021
Revised 4 January 2022
Accepted 4 January 2022
Published

Communicated by Igor Goychuk

Regression with Stable Errors Based on Order Statistics

Reza Alizadeh Noughabi and Adel Mohammadpour

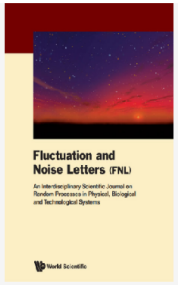
https://doi.org/10.1142/S0219477522500146 | **Cited by:** 0

< Previous Next >

PDF/EPUB Tools Share

Abstract

Classical regression approaches are not robust when errors are heavy-tailed or asymmetric. That may be due to the non-existence of the mean or variance of the error distribution. Estimation based on trimmed data, which ignored outlier or leverage points, has an old history and frequently used. This procedure chooses fixed cut-off points. In this work, we use this idea recently applied for initial estimates of regression coefficients with heavy-tailed stable errors. We propose an effective procedure to calculate the cut-off points based on the tail index and skewness parameters of errors. We use the property of the existence of some moments of stable distribution order statistics. Data are trimmed based on ordered residuals of a least square regression. However, the trimmed data's optimal number is determined based on the number of error order statistics whose variance exists. Then, we use the rest of the ordered data to estimate the regression coefficients. Based on these order statistics' joint



Online Ready

Metrics

Downloaded 1 times

History

bibliography

- Nolan, J. 2020. *Stable distributions: models for heavy-tailed data,* Boston: Birkhauser.
- Nolan, J. P., and Ojeda-Revah, D. 2013. Linear and nonlinear regression with stable errors. *Journal of Econometrics,* 172(2): 186-194.
- Koenker, R., Chernozhukov, V., He, X., Peng, L. 2018. *Handbook of Quantile Regression*, New York: Chapman and Hall/CRC.
- Mohammadi, M., and Mohammadpour, A. 2014. Estimating the parameters of an α-stable distribution using the existence of moments of order statistics. *Statistics and Probability Letters,* 90: 78-84.

Thanks for your attention