Communications for Statistical Applications and Methods
2018, Vol. 25, No. 6, 633–645

https://doi.org/10.29220/CSAM.2018.25.6.633
Print ISSN 2287-7843 / Online ISSN 2383-4757

Linear regression under log-concave and Gaussian scale mixture errors: comparative study

Sunyul Kim[a], Byungtae Seo[1,a]

[a]Department of Statistics, Sungkyunkwan University, Korea

Abstract

Gaussian error distributions are a common choice in traditional regression models for the maximum likelihood (ML) method. However, this distributional assumption is often suspicious especially when the error distribution is skewed or has heavy tails. In both cases, the ML method under normality could break down or lose efficiency. In this paper, we consider the log-concave and Gaussian scale mixture distributions for error distributions. For the log-concave errors, we propose to use a smoothed maximum likelihood estimator for stable and faster computation. Based on this, we perform comparative simulation studies to see the performance of coefficient estimates under normal, Gaussian scale mixture, and log-concave errors. In addition, we also consider real data analysis using Stack loss plant data and Korean labor and income panel data.

Keywords: NPMLE, log-concave, Gaussian scale mixture

1. Introduction

In the traditional regression model $y = m(x) + \epsilon$, there have been significant research efforts to find the mean function $m(x)$ to give meaningful statistical insight. This is because $m(x)$ itself is the function of interest while ϵ or the distribution of ϵ is a nuisance factor. When $m \in \mathcal{M}$ is assumed, where $\mathcal{M}$ is a certain family of mean functions, the estimation of $m(x)$ can be achieved by using either the least square estimator (LSE) or the maximum likelihood estimator (MLE). The MLE is identical with the LSE under the normality assumption of ϵ; therefore, the common choice for the distribution of ϵ is the normal distribution. However, there are few research papers that consider non-normal error distributions. In most practical applications, the normality assumption seems suitable unless there exist severe outliers or the error distribution is severely skewed. It is well known that the LSE or MLE under the normality assumption is very sensitive to outliers. For this, Holland and Welsch (1977) proposed the M-estimation and there are many subsequent research papers. These basically modified the object function in order to reduce the effect of outliers; however these are not the MLE.

In the ML framework, if the error distribution is assumed to be a family of heavy tailed distributions, the MLE of the regression parameters is known to be robust (Lange *et al.*, 1989; Lange and Sinsheimer, 1993). In this sense, Seo *et al.*, (2017) proposed a family of Gaussian scale mixture distributions for the error distribution. They showed that the resulting estimator is as robust as many well-known M-estimators without specifying any tuning parameter. In addition, their estimator is automatically tuned from normal to any heavy tailed distribution so that the efficiency loss resulting

[1] Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2 Sungkyunkwan-ro, Jongno-Gu, Seoul 03063, Korea. E-mail: seobt@skku.edu

Published 30 November 2018 / journal homepage: http://csam.or.kr
© 2018 The Korean Statistical Society, and Korean International Statistical Society. All rights reserved.

from a larger class is minimized. A potential drawback of their method is the inefficiency when the true error distribution is skewed.

A skewed error distribution may not be a pure interest because there should be a suitable transformation to make the error distribution symmetric. However, transformation could blur the statistical inference for the mean function and such transformation is not always available. In addition, in some cases, the error distribution itself could be of great interest. For example, the quality of Value-at-Risk estimator in the time series model is severely affected by the quality of the estimated error distribution.

We consider a family of log-concave distributions for the error distribution due to their many attractive features. It contains many well-known parametric distributions such as normal, t, and gamma distributions. It is also known that marginal distributions, convolutions, and product measures of log-concave distributions preserves log-concavity, see Dharmadhikari and Joag-Dev (1988). Dümbgen *et al.* (2007) showed the consistency for the linear and isotonic regression estimator with log-concave error distributions. Although they showed some numerical results, there is no rigorous study for the performance of the method.

In this paper, we investigate the performance of estimators under the normality, log-concavity, and Gaussian scale mixture with some real data examples. In addition, we propose the use of a smoothed log-concave distribution to estimate of the regression parameters that can make the computation more reliable and faster than raw estimators. We also derive a formula for an iterative reweighted least squares based on the smoothed log-concave density. This paper is organized as follows. In Section 2, we review some literature related to the continuous Gaussian scale mixture and log-concave densities with their computational aspects. In Section 3, we propose an estimation method in which the error distribution is assumed to be log-concave. In Section 4, we also adopt the smoothed log-concave MLE for the inference of regression parameters. Some numerical studies including real data analysis are given in Section 5. We then end this paper with some concluding remarks in Section 6.

2. Gaussian scale mixture and log-concave distributions

2.1. Gaussian scale mixture distribution

A family of Gaussian scale mixture densities is given as

$$\mathcal{F}_{SM} = \left\{ f(x;Q) = \int_0^\infty \frac{1}{\sigma}\phi\left(\frac{x}{\sigma}\right) dQ(\sigma) |\ Q \text{ is a probability measure on } \mathbb{R}^+ \right\}, \tag{2.1}$$

where $\phi(\cdot)$ is the density of the standard normal distribution and Q is a probability measure defined on positive real numbers. When Q is discrete with finite support points, $f(x;Q)$ is simply a finite mixture model. But, $\mathcal{F}_{SM}$ is a more flexible family than finite mixture models because Q is not restricted to be finitely discrete. A specific choice of Q results in some well-known distributions. For example, f can be reduced to the t distribution when Q is a Gamma distribution. Hence, this family is a quite large class of densities in the sense that it contains most unimodal symmetric distributions.

The advantage of the use of this family is that a smooth density estimator is obtained without any tuning parameter. If we assume that the error distribution in the regression analysis belongs to this family, it is known that the MLE of regression parameters is robust to traditional outliers (Seo *et al.*, 2017). In addition, when there is no outlier, the parameter estimate is almost the same as the ordinary least square estimator or the MLE based on the normality assumption.

The existence and uniqueness of the MLE of Q is well studied in Lindsay (1983). In addition, the MLE of Q is discrete with a finite number of support points. The computation of the MLE for Q in (2.1) is achieved by utilizing the gradient characterization of the mixture likelihood. The gradient function of Q is defined as a directional derivative of the log-likelihood at Q toward a degenerate distribution δ_s having all mass at s. This gradient function can be represented as

$$D_Q(s) = \sum_{i=1}^{n} \frac{f(x;\delta_s)}{f(x;Q)} - n.$$

If Q is the MLE, $D_Q(s)$ should be nonpositive for all $s \in \mathbb{R}^+$ because positive $D_Q(s^*)$ for a certain $s^* > 0$ implies that we can increase the likelihood by adding s^* to the set of support points of Q.

Based on this observation, the algorithm finds s^* such that $D_Q(s^*) > 0$ for a given Q. If such s^* is found, the support of Q is updated by adding s^* to the current set of support points in Q. The probability mass for this new support points are then updated with the expectation and maximization (EM) algorithm or standard optimization procedures like the Newton-Raphson method. This algorithm is called vertex direction method; in addition, there are also several relatives such as the vertex exchange method (Böhning, 1995), the intra simplex direction method (Lesperance and Kalbfleisch, 1992), and the constrained Newton-method for multiple supports (Wang, 2007).

2.2. Log-concave distribution

A family of log-concave densities can be expressed as

$$\mathcal{F}_{LC} = \left\{ f(x) = c_\varphi \exp\varphi(x) |\ \varphi \text{ is a concave function on } \mathbb{R} \right\},$$

where $c_\varphi = 1/\int \exp\varphi(x)dx$ is the normalizing constant. This family is a subset of the family of unimodal densities, but it contains most of the well-known parametric distributions. Karlin (1968) showed that it has subexponential tails and nondecreasing hazard rates. From this richness of $\mathcal{F}_{LC}$, it can be an important alternative to model potentially skewed unimodal densities.

Walther (2009) showed that the NPMLE of a log-concave density exists and explain how one can develop an algorithm to estimate the NPMLE. Suppose that $X_1 < X_2 < \cdots < X_n$ is ordered random observations obtained from a population having a log-concave density $f(x) \in \mathcal{F}_{LC}$. The log-likelihood function is then

$$\ell_n(\varphi) = \int \log f(x) d\hat{\mathbb{F}}_n(x) = \int \varphi(x) d\hat{\mathbb{F}}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i),$$

where $\hat{\mathbb{F}}_n$ is the empirical cumulative distribution function (CDF) based on the sample. Since f itself is a density, we have to maximize $\ell(\varphi)$ with the constraint $\int \exp\varphi(x)dx = 1$ over all concave functions φ. Under this constraint, maximizing $\ell(\varphi)$ is equivalent to maximizing

$$\ell_n^*(\varphi) = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i) - \int \exp\varphi(x)dx,$$

see Silverman (1982, Theorem 3.1). It was proven that the NPMLE $\hat{\varphi}_n$ uniquely exists and is piecewise linear on $[X_1, X_n]$ with the set of knots contained in $\{X_1, \ldots, X_n\}$, and $\hat{\varphi}_n = -\infty$ on $\mathbb{R}\backslash [X_1, X_n]$, see Walther (2002) or Pal *et al.* (2007).

That is, $\hat{\varphi}_n$ should be a concave piecewise linear function on $[X_1, X_n]$ with the knots $S_n(\hat{\varphi}_n) = \{t \in (X_1, X_n) : \hat{\varphi}_n'(t-) \neq \hat{\varphi}_n'(t+)\} \bigcup \{X_1, X_n\}$, where $\hat{\varphi}_n'(t-)$ and $\hat{\varphi}_n'(t+)$ are the left and right derivatives of $\hat{\varphi}_n$ at t, respectively. Since a piecewise linear function can be drawn by connecting the points at each knot, the maximization of the log-likelihood function ℓ_n^* can be solved by finding the vector $\hat{\boldsymbol{\varphi}} = \{\hat{\varphi}_n(x_1), \ldots, \hat{\varphi}_n(x_n)\}$ instead of finding the function $\hat{\varphi}_n$ directly. Since $\hat{\boldsymbol{\varphi}}$ should be concave and piecewise linear, the estimation problem can be reduced to

$$\hat{\boldsymbol{\varphi}} = \arg\max_{\varphi \in \mathcal{K}} \ell_n^*(\varphi),$$

where

$$\mathcal{K} = \left\{ \boldsymbol{\varphi} \in \mathbb{R}^n : \frac{\varphi_i - \varphi_{i-1}}{x_i - x_{i-1}} \geq \frac{\varphi_{i+1} - \varphi_i}{x_{i+1} - x_i} \text{ for } i = 2, \ldots, n-1 \right\}.$$

That is, we can turn our estimation problem into a linearly constrained optimization problem. There are several algorithms to solve this linearly constrained optimization problem such as the Iterative Convex Minorant Algorithm (ICMA) and the active set algorithm. It was shown that the active set algorithm is generally more efficient than other existing algorithms, see Rufibach (2007).

3. Application to linear regression

In the linear regression model,

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad \text{for } i = 1, \ldots, n,$$

where $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$, if we assume that ϵ_i's are independent normal random variables with mean zero and variance σ^2, it is known that the MLE of $\boldsymbol{\beta}$ is the same as the ordinary least square estimator. Instead of the normality assumption, we suppose that ϵ_i's are a random sample from a population having a density contained in $\mathcal{F}_{LC}$. That is, we adopt the methodology to estimate $\boldsymbol{\beta}$ under the assumption that the distribution of ϵ_i is a log-concave distribution. The estimator of $(\varphi, \boldsymbol{\beta})$ is then the maximizer of the log-likelihood function

$$\ell_n^*(\varphi, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \varphi\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right) - \int \exp \varphi(x) dx.$$

There is no direct way to find the maximizers $\hat{\varphi}$ and $\hat{\boldsymbol{\beta}}$ simultaneously. For this, we reform the idea of an alternating algorithm which was introduced in Dümbgen *et al.* (2011, Section 3.3). An alternating algorithm is as follows. First, it computes the log-concave MLE $\varphi^{(0)}$ for the residuals $y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^{(0)}$, $i = 1, \ldots, n$ with any initial parameter $\boldsymbol{\beta}^{(0)}$ that satisfies $\sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^{(0)}) = 0$ such as the ordinary least square estimator. And it proceeds the following steps for $k \geq 1$ iteratively until it converges:

(a) Determine

$$\tilde{\boldsymbol{\beta}}^{(k)} \in \arg\max_{\beta} \sum_{i=1}^n \varphi^{(k-1)} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^{(k-1)}\right).$$

(b) Adjust the parameter $\tilde{\boldsymbol{\beta}}^{(k)}$ to make the sum of the residuals zero. That is, replace $\tilde{\beta}_0^{(k)}$ in $\tilde{\boldsymbol{\beta}}^{(k)} = (\tilde{\beta}_0^{(k)}, \tilde{\beta}_1^{(k)}, \ldots, \tilde{\beta}_p^{(k)})$ with $\tilde{\beta}_0^{(k)} - c^{(k)}$ where $c^{(k)} = n^{-1} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \tilde{\boldsymbol{\beta}}^{(k)})$, and set this adjusted $\tilde{\boldsymbol{\beta}}^{(k)}$ to $\boldsymbol{\beta}^{(k)}$.

(c) Determine

$$\varphi^{(k)} = \arg\max_{\varphi} \left\{ \frac{1}{n} \sum_{i=1}^n \varphi\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^{(k)}\right) - \int \exp \varphi(x) dx \right\}.$$

In (a), if $\varphi^{(k-1)}$ is known, $\boldsymbol{\beta}^{(k)}$ can be obtained using the general-purpose optimizer function `optim` in `R` (Dümbgen *et al.*, 2011). This function was found to be fast and reliable enough for their first research. However, it is not stable for estimating regression coefficients in multiple linear regression, and the efficiency of the algorithm becomes poor when the error distribution is severely heavy-tailed or highly skewed. In addition, when the number of covariates are very large, `optim()` function often returns an error due to a nonsmooth objective function. For this reason, we suggest a more reliable algorithm that produces an iterative reweighted least square expression in Section 4. Once $\boldsymbol{\beta}^{(k)}$ is obtained, the maximization problem in step (c) can be solved efficiently by an active set algorithm. We use the implementation in the contributed package `logcondens` by Rufibach and Dümbgen (2010) in `R`.

4. Estimation of regression coefficients with smoothed log-concave density

As mentioned in Section 3, a characteristic feature of the MLEs of log-concave densities is that they are not smooth and not differentiable at each knot. Dümbgen *et al.* (2011) provided an idea how to apply log-concave density estimation to regression analysis; however, they did not provide a specific way to maximize the likelihood function in a given situation of an error distribution. We here propose a method to find the MLE of regression parameters using a smoothed version of the log-concave density estimator which is smooth and differentiable anywhere. Chen and Samworth (2013) investigated a smoothed version of the log-concave MLE given by

$$\tilde{f}_n(x) = \hat{f}_n(x) * \phi_h(x),$$

where $\hat{f}_n(\cdot)$ is the obtained log-concave MLE and $\phi_h(\cdot)$ is the Gaussian kernel with tuning parameter h. Note that $\tilde{f}_n$ is a log-concave density as long as $\hat{f}_n$ is a log-concave density, see Dharmadhikari and Joag-Dev (1988, Theorem 2.8). According to Dümbgen *et al.* (2011, Remark 2.3), $\hat{f}_n(\cdot)$ underestimates the variance of the true density. Based on this observation, Chen and Samworth (2013) proposed $h = \hat{\sigma} - \tilde{\sigma}$, where $\hat{\sigma}^2 = \int (x-\bar{x})^2 \hat{f}_n(x)dx$ and $\tilde{\sigma}^2 = 1/(n-1)\sum_{i=1}^n (x_i - \bar{x})^2$ so that the estimator has an exact size of the variance. From the piecewise linear property of $\hat{\varphi}_n$, we can write $\hat{\varphi}_n$ as

$$\hat{\varphi}_n(x) = \sum_{i=1}^{c-1} (\alpha_{0i} + \alpha_{1i}x)\, I\left[\ell_i \le x \le u_i\right],$$

for some real numbers α_{01}, α_{1i}, and $l_i < u_i$, $i = 1, \ldots, c-1$. This $\hat{\varphi}_n(x)$ can be obtained from the ICMA or the active set algorithm as described in Section 2.2. Then, a smoothed version of the log-concave MLE can be explicitly computed as

$$\begin{aligned}
\tilde{f}_n(x) &= \int \exp\hat{\varphi}_n(t)\frac{1}{\sqrt{2\pi h}}e^{-\frac{1}{2h}(x-t)^2}dt \\
&= \sum_{i=1}^{c-1}\int_{\ell_i}^{u_i} \exp(\alpha_{0i} + \alpha_{1i}t)\frac{1}{\sqrt{2\pi h}}e^{-\frac{1}{2h}(x-t)^2}dt \\
&= \sum_{i=1}^{c-1}\exp\left(\alpha_{1i}x + \frac{h}{2}\alpha_{1i}^2 + \alpha_{0i}\right)\left(\Phi\left(\frac{u_i - x - h\alpha_{1i}}{\sqrt{h}}\right) - \Phi\left(\frac{\ell_i - x - h\alpha_{1i}}{\sqrt{h}}\right)\right), \qquad (4.1)
\end{aligned}$$

where $\Phi(\cdot)$ is the CDF of the standard normal density.

Now, for our regression problem, we first compute the residuals $\hat{\epsilon}_i = y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}$ for given $\hat{\boldsymbol{\beta}}$. Based on $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$, we can compute the smoothed MLE of the error distribution $\tilde{f}_n(\epsilon)$ as in (4.1). With this estimated $\tilde{f}_n(\epsilon)$, the log-likelihood function $\ell_n(\boldsymbol{\beta})$ can be defined as $\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^n \log \tilde{f}_n(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})$. The maximizer of $\ell_n(\boldsymbol{\beta})$ can be then obtained by the Newton-Raphson algorithm. The one-step Newton-Raphson update can be written as

$$\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)} - \frac{\ell_n'\left(\boldsymbol{\beta}^{(t-1)}\right)}{\ell_n''\left(\boldsymbol{\beta}^{(t-1)}\right)},$$

where

$$\frac{\partial \ell_n(\boldsymbol{\beta}|\mathbf{x},\mathbf{y})}{\partial \boldsymbol{\beta}} = -\sum_{i=1}^n \mathbf{x}_i \frac{\tilde{f}_n'\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)}{\tilde{f}_n\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)},$$

$$\frac{\partial^2 \ell_n(\boldsymbol{\beta}|\mathbf{x},\mathbf{y})}{\partial \boldsymbol{\beta}^2} = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \left(\frac{\tilde{f}_n\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right) \tilde{f}_n''\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right) - \tilde{f}_n'\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2}{\tilde{f}_n\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2} \right).$$

Further, if we denote

$$W_1 = \mathrm{diag}(a_1, \ldots, a_n), \quad \text{where } a_i = \frac{\tilde{f}_n'\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)}{\tilde{f}_n\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)}$$

and

$$W_2 = \mathrm{diag}(b_1, \ldots, b_n), \quad \text{where } b_i = \frac{\tilde{f}_n\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right) \tilde{f}_n''\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right) - \tilde{f}_n'\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2}{\tilde{f}_n\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2},$$

$\boldsymbol{\beta}^{(t)}$ can be rephrased as

$$\boldsymbol{\beta}^{(t)} = \left(X^\top W_2 X\right)^{-1} X^\top \left(W_2 X \boldsymbol{\beta}^{(t-1)} - W_1\right), \tag{4.2}$$

which is the form of iterative reweighted least squares.

We summarize the above as:

1. For given $\hat{\boldsymbol{\beta}}$, compute $\hat{\epsilon}_i = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$.
2. Compute $\tilde{f}_n(\epsilon)$ based on $\hat{\epsilon}_i, i, \ldots, n$ by using (4.1).
3. For given $\tilde{f}_n$, compute $\hat{\boldsymbol{\beta}}$ by using (4.2).
4. Repeat (1)–(3) until a stopping rule is satisfied.

5. Numerical examples

5.1. Simulations

In this section, we conduct some Monte Carlo simulation studies to see the performance of two different methods. We generate 200 simulated samples from the model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$. In here, x_{1i} and x_{2i} are independently generated from $B(1, 0.5)$ and $N(0, 1)$ respectively for $i = 1, \ldots, n$.

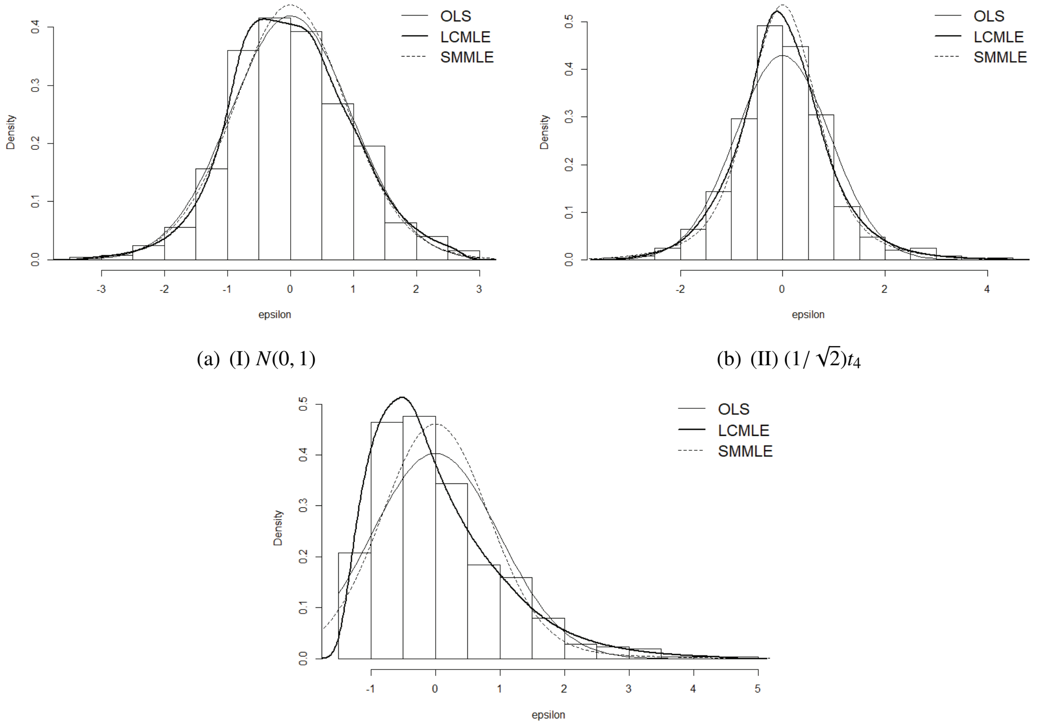
Table 1: Empirical MSE × 100 (empirical bias × 100) for $n = 250$ and 500

n	Error	Method	β_0	β_1	β_2
250	(I)	OLS	0.81 (1.74)	1.74 (−1.41)	0.39 (−0.24)
		SMMLE	0.82 (1.66)	1.84 (−1.26)	0.40 (−0.27)
		LCMLE	0.85 (1.55)	2.00 (−1.03)	0.42 (−0.27)
	(II)	OLS	0.77 (0.28)	1.72 (−0.94)	0.40 (−0.16)
		SMMLE	0.71 (0.21)	1.30 (−0.80)	0.28 (−0.02)
		LCMLE	0.71 (0.16)	1.40 (−0.72)	0.30 (−0.11)
	(III)	OLS	0.76 (−0.01)	1.49 (−0.12)	0.35 (0.44)
		SMMLE	0.67 (0.07)	1.21 (−0.23)	0.26 (0.39)
		LCMLE	0.51 (0.30)	0.50 (−0.74)	0.11 (0.22)
500	(I)	OLS	0.34 (−0.19)	0.81 (0.40)	0.19 (−0.83)
		SMMLE	0.35 (−0.18)	0.84 (0.38)	0.19 (−0.72)
		LCMLE	0.35 (−0.11)	0.88 (0.25)	0.21 (−0.87)
	(II)	OLS	0.40 (0.20)	0.69 (0.13)	0.18 (−0.02)
		SMMLE	0.33 (0.31)	0.54 (−0.11)	0.13 (0.10)
		LCMLE	0.34 (0.27)	0.55 (−0.02)	0.14 (0.07)
	(III)	OLS	0.40 (0.49)	0.83 (−0.54)	0.21 (0.46)
		SMMLE	0.34 (0.62)	0.67 (−0.81)	0.14 (0.49)
		LCMLE	0.27 (0.51)	0.28 (−0.56)	0.06 (0.22)

MSE = mean squared error; OLS = Ordinary least square; SMMLE = smoothed MLE; LCMLE = log-concave MLE; MLE = maximum likelihood estimator.

The true $(\beta_0, \beta_1, \beta_2)$ is set to be $(2, 1, 1)$. To investigate how the estimates are changed over different error distributions, we consider (I) $N(0, 1)$, (II) $(1/\sqrt{2})t_4$, and (III) $(1/\sqrt{8})(\chi^2_4 - 4)$. For each error distribution, Table 1 shows empirical mean squared error (MSE) and empirical bias for different true error distributions and estimation methods for sample sizes $n = 250$ and $n = 500$ based on 200 replications. Ordinary least square (OLS), smoothed MLE (SMMLE), and log-concave MLE (LCMLE) represent the ordinary least square estimator, the MLE under the Gaussian scale mixture error, and the MLE under the log-concave error, respectively. All estimators are obtained using R program. For SMMLE, both the initial value of $(\beta_0, \beta_1, \beta_2)$ and the initial distribution of ϵ_i are obtained from the OLS estimation. For LCMLE, the OLS estimator is used for the initial value of $(\beta_0, \beta_1, \beta_2)$, but the initial distribution of ϵ_i is re-estimated from its corresponding residuals.

When the true error distribution is $N(0, 1)$, although there is no significant difference among the three estimators, OLS has the best performance followed by SMMLE. This is natural because the assumed classes of error distributions contain the normal distributions and the size of the class of the normal distributions, which produces OLS, is the smallest. For t case, OLS breaks down while other methods still give reasonable estimators, and SMMLE shows the best performance. For χ^2 case, i.e., skewed distribution case, both OLS and SMMLE break down, but LCMLE gives reasonable estimators. In addition to parameter estimates, we also provide estimated error distributions. For this, we generate one simulated sample with size $n = 500$ for each error distributions (I)–(III) and apply both SMMLE and LCMLE. Figure 1 shows estimated error distributions along with the probability histogram of simulated errors for normal, t, and χ^2 errors. In this figure, the light solid line represents the true error probability density function and the dashed line shows the estimated error density based on SMMLE. The thick solid line represents the estimated error density based on LCMLE. Estimated distributions from both SMMLE and LCMLE are close to the true distribution when the true error distribution is the normal or student's t. When the true error distribution is χ^2, LCMLE produces a reasonable error distribution while the estimated error distribution based on SMMLE shows a severe bias.



(a) (I) $N(0, 1)$

(b) (II) $(1/\sqrt{2})t_4$

(c) (III) $(1/\sqrt{8})(\chi^2_4 - 4)$

Figure 1: *Estimated error densities based on one simulated sample of size* $n = 500$ *for (I)–(III).*

5.2. Real data example

For the first real data example, we consider Stack Loss Plant data (Brownlee, 1960). This dataset has been used in many robust regression literature because it has some severe outliers. Bellio and Ventura (2005) showed that observations 1, 3, 4, and 21 are those outliers. The dataset contains four variables (Air Flow, Water Temp, Acid Conc, and Stack Loss). For this dataset, we consider a linear regression model using Stack Loss variable as a response and other variables as covariates.

For model fitting, we use OLS, SMMLE, and LCMLE. Tables 2 and 3 show the estimated regression coefficients and the corresponding standard errors in the parentheses with the data excluding outliers and the original data, respectively. The standard errors of SMMLE and LCMLE were obtained by the bootstrap method. As it is known that OLS is very sensitive to outliers, OLS shows large differences between estimated parameters for each case. LCMLE also shows large differences in the regression parameters. Unlike SMMLE, LCMLE is not so robust to severe outliers with a small sample. Especially when the outliers are highly skewed, LCMLE also becomes skewed and this results in non-robust regression parameter estimates. Figure 2 shows the histograms of residuals and the corresponding estimated error distributions for each method. In this figure, the solid line represents the estimated error distribution from the original data and the dashed line shows the estimated error distribution from Stack Loss Plant data without outliers. Both OLS and LCMLE show large differences between estimated distributions from each data. However, SMMLE shows small difference. That is, outliers have a great influence on the estimation of the parameters and distribution in OLS

Table 2: Estimated parameters and standard errors for the Stack Loss Plant data excluding outliers

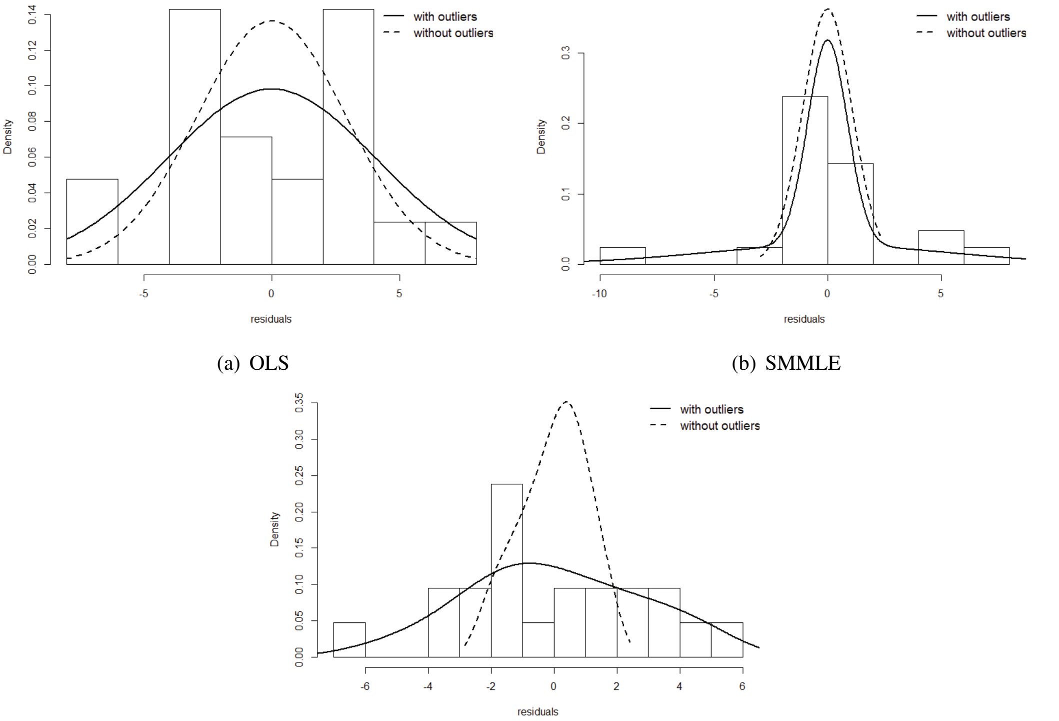
Method	β_0	β_1	β_2	β_3
OLS	−37.6525 (4.7321)	0.7977 (0.0674)	0.5773 (0.1660)	−0.0671 (0.0616)
SMMLE	−37.5401 (5.2553)	0.8055 (0.1006)	0.5558 (0.1970)	−0.0682 (0.0807)
LCMLE	−39.0334 (5.1879)	0.7958 (0.0993)	0.6017 (0.1819)	−0.0555 (0.0734)

OLS = Ordinary least square; SMMLE = smoothed MLE; LCMLE = log-concave MLE; MLE = maximum likelihood estimator.

Table 3: Estimated parameters and standard errors for the Stack Loss Plant data including outliers

Method	β_0	β_1	β_2	β_3
OLS	−39.9197 (11.8960)	0.7156 (0.1349)	1.2953 (0.3680)	−0.1521 (0.1563)
SMMLE	−36.0085 (9.7895)	0.8416 (0.2043)	0.4838 (0.6130)	−0.0872 (0.1433)
LCMLE	−37.5101 (10.7084)	0.6759 (0.1879)	1.3638 (0.5470)	−0.1690 (0.1404)

OLS = Ordinary least square; SMMLE = smoothed MLE; LCMLE = log-concave MLE; MLE = maximum likelihood estimator.



(a) OLS

(b) SMMLE

(c) LCMLE

Figure 2: *The histograms of residuals and the corresponding estimated error densities. OLS = Ordinary least square; SMMLE = smoothed MLE; LCMLE = log-concave MLE; MLE = maximum likelihood estimator.*

and LCMLE.

Second, we consider 19th KLIPS (Korean Labor & Income Panel Study) Data. KLIPS is a longitudinal survey of the labor market/income activities of households and individuals residing in urban areas. With KLIPS data, we try to analyze the effect of personal properties on income. For this, we

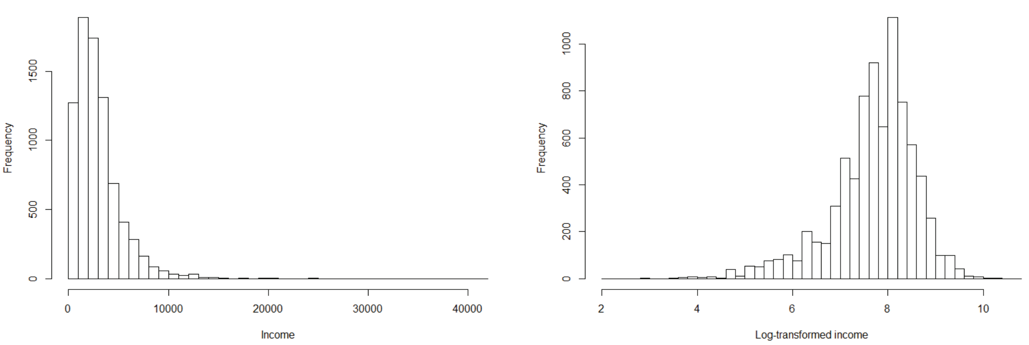


Figure 3: *The histograms of income and log-transformed income variable in the 19th KLIPS data.*

Table 4: Estimated parameters and standard errors for 19th KLIPS data

Method	OLS	SMMLE	LCMLE
β_0	6.9362 (0.0629)	7.0528 (0.0879)	7.0013 (0.0551)
β_1	−0.5377 (0.0171)	−0.5083 (0.0192)	−0.4827 (0.0141)
α_1	0.5938 (0.0331)	0.4451 (0.0504)	0.3417 (0.0262)
α_2	0.7220 (0.0324)	0.5695 (0.0516)	0.4903 (0.0263)
α_3	0.7051 (0.0324)	0.5388 (0.0527)	0.5049 (0.0308)
α_4	0.1368 (0.0371)	0.0358 (0.0556)	0.0605 (0.0333)
β_2	0.1750 (0.0073)	0.1716 (0.0091)	0.1840 (0.0068)
β_3	0.0590 (0.0166)	0.0418 (0.0155)	0.0279 (0.0128)

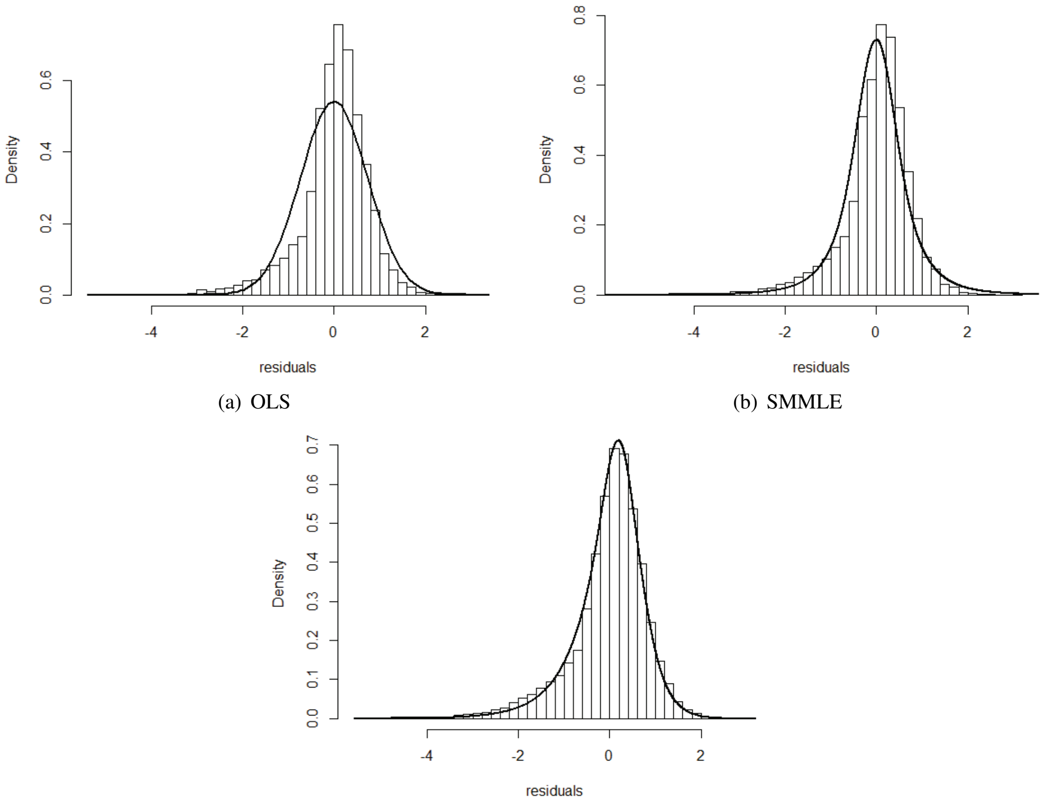
OLS = Ordinary least square; SMMLE = smoothed MLE; LCMLE = log-concave MLE; MLE = maximum likelihood estimator.

consider a linear regression model using pre-tax annual income as a response and gender, age, educational background, and residence area as covariates. Covariates are selected by referring to various research papers on determinants of income. The age variable was divided into five categories: under 30, between 30 and 40, between 40 and 50, between 50 and 60, and over 60. That is, there are 7 independent variables including 4 dummy variables.

Figure 3 shows that the distribution of income in the KLIPS data is severely skewed as it is known that the income distribution tends to be skewed. The left is the histogram of the original income and the right is the histogram of log-transformed income. We use the log-transformed income for our model as we do generally when the data is skewed. Table 4 shows the estimated regression coefficients and the corresponding standard errors in the parentheses. We can check that the standard errors of LCMLE are smaller than those of any other method. Figure 4 shows the histograms of residuals and the corresponding estimated error distributions for each method. In OLS, there is a large gab between the estimated error distribution and the histogram of residuals. In SMMLE, the gap is reduced, but it still shows some difference. However, in LCMLE, they match very well. Table 5 shows 95% bootstrap confidence intervals for the parameters which are calculated based on each method. We can also see that the confidence intervals obtained from LCMLE tend to be shorter than the others.

6. Concluding remarks

A regression model based on the Gaussian scale mixture error has a comparable or superior performance to other robust regression estimators. One potential limitation of this model is that the estimation may be unreliable when the true error distribution is not symmetric. The family of log-concave



(a) OLS

(b) SMMLE

(c) LCMLE

Figure 4: *The histograms of residuals and the corresponding estimated error densities.*

Table 5: 95% confidence intervals for estimated parameters for 19th KLIPS data

	OLS		SMMLE		LCMLE	
	2.5 %	97.5 %	2.5 %	97.5 %	2.5 %	97.5 %
β_0	6.8130	7.0595	6.8241	7.1792	6.8942	7.1073
β_1	−0.5713	−0.5041	−0.5370	−0.4592	−0.5095	−0.4570
α_1	0.5290	0.6586	0.3224	0.5354	0.2918	0.3918
α_2	0.6585	0.7855	0.4522	0.6619	0.4386	0.5404
α_3	0.6383	0.7719	0.4217	0.6372	0.4415	0.5619
α_4	0.0640	0.2097	−0.0799	0.1373	−0.0031	0.1256
β_2	0.1606	0.1893	0.1629	0.1979	0.1712	0.1971
β_3	0.0265	0.0916	0.0079	0.0684	0.0043	0.0552

OLS = Ordinary least square; SMMLE = smoothed MLE; LCMLE = log-concave MLE; MLE = maximum likelihood estimator.

densities contains many skewed distributions so that the regression model based on log-concave errors is quite flexible to estimate regression parameters even though the true error distribution is skewed. In this paper, we study the estimation of regression parameters and error distributions with Gaussian scale mixture densities and log-concave densities, as well as compare them by using some numerical examples.

The estimation with log-concave densities can be conducted by a three-step alternating algorithm. In the first step, we proposed the methodology for finding the MLEs for regression parameters with a smoothed version of the log-concave MLE, which produces an iterative reweighted least square expression. We find that the proposed method is stable and efficient to estimate regression coefficients in multiple linear regression even with a large sample size.

Simulation results show that the estimation with log-concave densities is as good as other methods in normal and heavy-tailed cases, and it has a remarkable performance in a skewed case. However, the estimator under log-concave errors is sensitive to outliers when there are severe outliers in a small sample. It seems that our proposed method could still be robust when the outliers occur in a symmetric fashion. However, when there are only extremely large (or small) outliers, the proposed method produces highly skewed log-concave density estimators. This would be the reason why the proposed estimator is not robust in general. On the other hand, since the SMMLE assumes that the error distribution is symmetric, the density estimate is not so heavily affected by skewed outliers regardless of the existence of large (or small) outliers. In this case, the SMMLE produces a symmetric but heavy tailed density estimator. This also explains why the proposed method works well in simulation studies even though it is not robust for the the real data analysis with a small sample size in which some large outliers exist.

Acknowledgements

This research was supported by Basic Science Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (NRF-2016R1A2B4007 373).

References

Bellio R and Ventura L (2005). An introduction to robust estimation with R functions. In *Proceedings of 1st International Work*, 1–57.

Böhning D (1995). A review of reliable maximum likelihood algorithms for semiparametric mixture models, *Journal of Statistical Planning and Inference*, **47**, 5–28.

Brownlee KA (1960). *Statistical Theory and Methodology in Science and Engineering*, John Wiley & Sons, New York.

Chen Y and Samworth RJ (2013). Smoothed log-concave maximum likelihood estimation with applications, *Statistica Sinica*, **23**, 1373–1398.

Dharmadhikari SW and Joag-Dev K (1988). *Unimodality, Convexity, and Applications*, Academic Press, Boston.

Dümbgen L, Hüsler A, and Rufibach K (2007). Active set and EM algorithms for log-concave densities based on complete and censored data, arXiv:0707.4643.

Dümbgen L, Samworth R, and Schuhmacher D (2011). Approximation by log-concave distributions, with applications to regression, *The Annals of Statistics*, **39**, 702–730.

Holland PW and Welsch RE (1977). Robust regression using iteratively reweighted least-squares, *Communications in Statistics - Theory and Methods*, **6**, 813–827.

Karlin S (1968). *Total Positivity*, Stanford University Press, Stanford.

Lange KL, Little RJA, and Taylor JMG (1989). Robust statistical modeling using the *t* distribution, *Journal of the American Statistical Association*, **84**, 881–896.

Lange K and Sinsheimer JS (1993). Normal/independent distributions and their applications in robust regression, *Journal of Computational and Graphical Statistics*, **2**, 175–198.

Lesperance ML and Kalbfleisch JD (1992). An algorithm for computing the nonparametric MLE of a mixing distribution, *Journal of the American Statistical Association*, **87**, 120–126.

Lindsay BG (1983). The geometry of mixture likelihoods: a general theory, *The Annals of Statistics*, **11**, 86–94.

Pal JK, Woodroofe M, and Meyer M (2007). Estimating a Polya frequency function$_2$, **54**, *Lecture Notes-Monograph Series*, 239–249.

Rufibach K (2007). Computing maximum likelihood estimators of a log-concave density function, *Journal of Statistical Computation and Simulation*, **77**, 561–574.

Rufibach K and Dümbgen L (2010). Logcondens: estimate a log-concave probability density from iid observations, *R package version*, **2**.

Seo B, Noh J, Lee T, and Yoon YJ (2017). Adaptive robust regression with continuous Gaussian scale mixture errors, *Journal of the Korean Statistical Society*, **46**, 113–125.

Silverman BW (1982). On the estimation of a probability density function by the maximum penalized likelihood method, *The Annals of Statistics*, **10**, 795–810.

Walther G (2002). Detecting the presence of mixing with multiscale maximum likelihood, *Journal of the American Statistical Association*, **97**, 508–513.

Walther G (2009). Inference and modeling with log-concave distributions, *Statistical Science*, **24**, 319–327.

Wang Y (2007). On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**, 185–198.

Received June 8, 2018; Revised August 6, 2018; Accepted September 4, 2018