Robust Algorithms for Change-Point Regressions Using the $t$-Distribution

Kang-Ping Lu and Shao-Tung Chang

Abstract: Regression models with change-points have been widely applied in various fields. Most methodologies for change-point regressions assume Gaussian errors. For many real data having longer-than-normal tails or atypical observations, the use of normal errors may unduly affect the fit of change-point regression models. This paper proposes two robust algorithms called EMT and FCT for change-point regressions by incorporating the $t$-distribution with the expectation and maximization algorithm and the fuzzy classification procedure, respectively. For better resistance to high leverage outliers, we introduce a modified version of the proposed method, which fits the $t$ change-point regression model to the data after moderately pruning high leverage points. The selection of the degrees of freedom is discussed. The robustness properties of the proposed methods are also analyzed and validated. Simulation studies show the effectiveness and resistance of the proposed methods against outliers and heavy-tailed distributions. Extensive experiments demonstrate the preference of the $t$-based approach over normal-based methods for better robustness and computational efficiency. EMT and FCT generally work well, and FCT always performs better for less biased estimates, especially in cases of data contamination. Real examples show the need and the practicability of the proposed method.

Keywords: change-point regression; robust; $t$-distribution; outlier; EMT algorithm; FCT algorithm

1. Introduction

In regression analyses, the relationship between the response and explanatory variables often changes abruptly at some positions, termed change-points (CPs) or thresholds. Change-point (CP) locations signify the place of changes in the data structure, which is usually pivotal for decision making or for realizing some scientific issues. Hence, CP regression models have been widely used for uncovering the meaning of CPs and regression coefficients in many fields, such as engineering, economics, medicine, climatology, genomics and linguistics. There are some motivating examples of CP detection in reality. In medicine, CPs often denote threshold values that are altered by the effect of some risk factor on the response. For example, the risk of a baby having Down Syndrome (DS) at birth is low and unrelated to the mother’s age until an age threshold, but then afterward, the DS risk grows steeply with the mother’s age [1]. Another use of CP models in medical studies is for monitoring patients’ health, such as detecting heart rate trend [2] or inspecting the activity of epileptics [3]. Besides, in econometrics, CPs often signal the time of market structure changes as a result of government interference or policy changes [4]. More applications include discovering the time of changes in the temperature trend for studying the global warming effect in climatology [5] and finding changes in the DNA copy number in genomics [6].
Outliers frequently occur in regression data due to measurement errors or improperly sampling a fraction of data from a different population. Outliers can lead to distortions in the model fitting. Outliers even have greater impacts on CP regressions since they may not only distort the estimations of model parameters but also twist the segmented structure of data. Fearnhead and Rigaill [6] provided a real example to illustrate the CP detection errors caused by outliers. They remarked that many CP methods are sensitive to outliers because of the modeling assumption of Gaussian noises, with much work on CP regressions being carried out for Gaussian models, such as in [7,10,11]. Despite that the distortion very likely resulted from outliers and the considerable concern for robust procedures in the literature of CP detection, there are few studies on robust estimations for CP regressions.

In regression analysis, the two common methods for solving outlier problems are outlier diagnostic and robust procedures. The robust approach is more recommended [12,13] since the diagnostic approach has issues such as it may fail due to the masking effect, the probability distribution of the resulting statistics is unknown and the variations of estimators may be underestimated. Hence, in this study, we aim at finding a new robust procedure for CP regressions. Here, we regard an estimation method to be robust in the sense of insensitivity to outliers and heavy-tailed distributions.

In this paper, we propose a robust CP regression model using $t$-distributions. The longer tail of a $t$-distribution makes it a more robust approach to CP regression data fitting by giving less weight to observations that are atypical in the computation of CPs and regression parameters. With the $t$ CP regression approach, the normal distribution in each regime is embedded in a broader class of symmetric distributions, with an additional parameter called the degrees of freedom, $r$. As commented by Lange et al. [14] and Peel and McLachlan [15], the parameter $r$ is helpful for tuning the robustness of the $t$-based method so that it is crucial for achieving resistant estimations. The proposed method treats CPs as latent class variables representing data points lying in respective regimes so that the new method is free of the difficulty of non-differentiability of the log-likelihood function with respect to CPs regarded as parameters. The computations of $t$ log-likelihood are simplified by writing the $t$-distribution as a scale mixture of normal distributions. Through two latent variables introduced into the observed data, the expectation and maximization (EM) algorithm is employed to estimate the CP model.

Considering the indefiniteness of the demarcation between two adjacent pieces, we extend the crisp CP model to a fuzzy class model. We create a fuzzy classification $t$ (FCT) method which is more robust than the $t$-based EM-type algorithm. To make the $t$ CP model more tolerant of high leverage outliers, we further propose a modified version to first moderately prune high leverage points and then employ the new method to the pruned dataset. The rules for deciding the proportion of pruning and the degrees of freedom of $t$-distributions are also discussed. Through numerous simulations and real applications, we demonstrate the effectiveness and superiority of the proposed methods by comparing them with existing methods.

The rest of this paper is organized as follows. In Section 2, we present some related work in the literature. In Section 3, we develop the new methods and illustrate the determination of the degrees of freedom of the $t$-distribution, the resistance to high leverage outliers and the robust properties of the proposed methods. In Section 4, experiments with numeric and real examples are provided to show the effectiveness and superiority of the proposed methods. Finally, the discussion and conclusions are stated in Section 5.

2. Related Work

Since change-point (CP) problems are often encountered in practice, numerous methodological approaches have been implemented for estimating CP models, such as likelihood-based estimation, Bayesian analyses, grid-searching approaches and clustering methods. A main difficulty with the likelihood estimations for CP regressions is the non-differentiability of the likelihood function with respect to the CPs viewed as parameters.
Researchers have tried to overcome this problem through various approximations or smooth transitions between two adjoining regimes. For example, Muggeo [16] created a segmented (SEG) regression method by means of a simple linearization technique. This method allows simultaneous inference on CPs and regression coefficients, but it is not resistant to outliers and only workable for continuous models. Recently, Lu and Chang [10] conquered the non-differentiable problem by transforming the CP regression model into a mixture regression model and introduced an EM-BP algorithm to derive estimates. There have been more recent methods for solving CP problems. Chakar et al. [17] consider the infeasibility of the dynamic programming algorithm for dependent series. They proposed a robust estimator of the autocorrelation parameter to de-correlate the autoregressive time series. Then, the classical inference for independent series can be utilized to estimate the CPs in the autoregressive series.

There are some recent Bayesian CP analyses. For instance, Ko et al. [18] proposed the Dirichlet process hidden Markov model (DPHMM) to detect CPs without requiring the specification of the number of CPs a priori. Bardwell and Fearnhead [19] presented a Bayesian approach to detect abnormal regions in multiple time series, allowing the presence of changes in a small subset of the time series.

Nonparametric CP methods were also suggested for the data, which cannot be well-approximated by any parametric distribution, such as in the works of Zou et al. [20] and Haynes et al. [21]. Moreover, Frick et al. [7] considered the question of search accuracy in CP analysis. They introduced a simultaneous multiscale CP estimator (SMUCE) to provide the estimates of an unknown step function with the probability of over- and under-estimating the number of CPs being controlled. Pein et al. [8] further proposed a heterogeneous simultaneous multiscale change-point estimator (H-SMUCE) for a heterogeneous piecewise constant model. More recently, various researchers have focused on developing computationally efficient methods [22,23]. For more reviews on CP work, we refer to articles [24,25] and references therein.

In spite of recent rapid developments on various aspects of CP detection, there is less work on the robustness to the presence of outliers [6,24]. To alleviate the disturbance of outliers, data may be pre-processed to remove outliers. After editing suspected outliers, the subsequent analysis is still based on the normality assumption. A serious problem with this approach is that the resulted influences may fail to reflect uncertainty in the “cleaned” data; particularly, standard errors may be too small.

The least absolute deviations (LAD) estimation has been used to estimate CP regression models robustly by a few researchers [26,27]. A robust approach connected with the LAD procedure is assuming that regression errors have a Laplace distribution. Yang [28] and Jafari et al. [29] employ Laplace distributions to estimate CP regression models robustly. The restriction of a single CP detection limits their applicability because multiple CPs often exist in reality. Although the LAD method is more robust than the least squares (LS) estimation, it is computationally intensive to optimize the objective function [25].

M-estimation is another popular robust approach in linear regressions, which replaces the least squares criterion with a resistant criterion. Considering the robustness of M-estimation, Fearnhead and Rigaill [6] proposed to associate the penalized cost approaches with a less sensitive loss function for estimating CPs robustly. The authors claimed that a bounded loss, such as the bi-weight loss, is favorable for robust estimations. However, their method relies on a constant variance assumption which is unsuitable in many applications [8]. Recently, a robust estimator based on the Wilcoxon statistic was proposed for estimating the CP of the mean [30], but this approach is restricted to a single CP detection.

The t-distribution has also been employed for robust fitting of a statistical model to the data structure. For example, Lange et al. [14] proposed to use t-distributions for the robust fit of a traditional linear regression model to data. Peel and McLachlan [15] proposed a mixture model of t-distributions for robust clustering. Yao et al. [31] presented a t mixture regression model for estimating model parameters robustly. Lin et al. [32,33]
presented the tests on heteroscedasticity in Student \( t \) and skew-\( t \)-normal regression models, respectively. Although \( t \)-distributions have been used in mixture regressions for making resistant estimations, they were barely considered for CP regression problems. Though locating CPs in a piecewise regression model is similar to classifying data into classes of similar objects, CP regression problems are different from mixture regressions in some aspects. For example, data points in CP regressions are ordered according to the order of some partitioning variable, and hence there are restrictions on the memberships of data points belonging to each individual segment. Subsequently, the robust mixture regressions of \( t \)-distributions cannot be applied to CP regressions directly. The rare implementation of \( t \)-distributions in CP regressions may be because the \( t \) likelihood function is non-differentiable with respect to CPs and there is no explicit form for the maximizer of the \( t \)-log-likelihood.

A few studies have utilized \( t \)-distributions to solve CP regression problems. Osorio and Galea [34] located a single CP for \( t \) regression models based on Schwarz Information Criterion (SIC). Their method is restricted to a single CP detection, and the degrees of freedom of \( t \)-distributions are not well-estimated. Lin et al. [35] adopted a Bayesian approach to identify the position of changes in variance for a \( t \) regression model with the regression coefficients assumed to be constant. The assumption of constant regression coefficients is restrictive and unrealistic. Furthermore, the burden of their computations is heavy because six steps are required in one pass of the Gibbs sampler for locating a single CP only, not mentioned for detecting multiple CPs. That is why they suggested detecting multiple CPs one by one in a hierarchical binary way instead of simultaneous detection. However, the hierarchical binary method for locating multiple CPs may not result in a correct optimum [11]. Moreover, these researchers did not illustrate the resistance of their method against high leverage outliers. It is known that \( t \)-distributions are sensitive to high leverage outliers, and these kinds of outliers usually have severe effects on the fit of a regression model to the data.

3. Robust Change-Point Regression Using \( t \)-Distributions

3.1. An EMT Algorithm for Change-Point Regression Models

Assume a dataset, \( \{(x_{i1}, y_i, \ldots, x_{ip}, y_i)\} \), is randomly drawn from a population of CP regressions, with \( x_j = (x_{j1}, \ldots, x_{jp})' \in \mathbb{R}^p \) (\( \cdot \) represents the transpose of a vector or a matrix) and \( y_j \in \mathbb{R} \), \( j = 1, \ldots, n \) expressing the \( j \)th independent observation and response, respectively. Suppose that \( X_d^* = (x_{d1}, x_{d2}, \ldots, x_{dp}) \), for some \( d \in \{1, \ldots, p\} \), is an independent variable that splits the domain of independent variables into sub-domains, such that regression models vary from sub-domain to sub-domain. Thus, we may assume that the dataset \( \{(x'_i, y_i), \ldots, (x'_i, y_i)\} \) has been positioned in the increasing order of \( X_d^* \). Accordingly, CPs can be stated by the rank of the ordered data. Suppose a CP regression model contains \( c - 1 \) CPs at \( \tau_1, \tau_2, \ldots, \tau_{c-1} \), with \( 1 \leq \tau_1 < \tau_2 < \cdots < \tau_{c-1} \leq n-1 \) (\( \tau_{c-1} < n \) since the last observation cannot be a CP). Set \( \tau_0 = 0 \) and \( \tau_c = n \). Then, the CP regression model can be represented as follows. Given \( (x'_i, y_i) \) lying in the \( i \)th segment, i.e., \( \tau_{i-1} < j \leq \tau_i \),

\[
y_{ji} = \beta_0 + \beta_{j1} x_{j1} + \cdots + \beta_{jp} x_{jp} + e_{ji}, \quad i = 1, \ldots, c,
\]
where $e_{ij}$ are independent with mean 0 and variance $\sigma_i^2, i=1,\ldots,c$. Since CPs can appear randomly among $\{1,\ldots,n-1\}$, the $c-1$ CPs can be viewed as any possible combination of $c-1$ numbers drawn from $\{1,\ldots,n-1\}$. Let $\tau=(\tau_1,\tau_2,\ldots,\tau_{c-1})'$ with $1 \leq \tau_1 < \tau_2 < \cdots < \tau_{c-1} \leq n-1$ be a collection of $c-1$ CPs. Define a random variable $w_\tau$ as:

$$w_\tau = \begin{cases} 
1, & \text{if } 1 \leq \tau_1 < \tau_2 < \cdots < \tau_{c-1} \leq n-1 \text{ are CPs,} \\
0, & \text{otherwise},
\end{cases}$$

with:

$$P(w_\tau = 1) = \epsilon_\tau$$

and

$$\sum_{1 \leq \tau_1 < \tau_2 < \cdots < \tau_{c-1} \leq n-1} \epsilon_\tau = 1.$$  

Then, define $z_{ij}$ as:

$$z_{ij} = \begin{cases} 
\sum_{1 \leq \tau_1 < \tau_2 < \cdots < \tau_{c-1} \leq n-1} w_\tau, & 1 \leq i \leq c, 1 \leq j \leq n, \\
0, & \text{otherwise.}
\end{cases}$$

Note that $z_{ij}, i=1,\ldots,c, j=1,\ldots,n$ are actually the class variables, where $z_{ij} = \begin{cases} 
1, & \text{if } (x_j,y_j) \text{ is from segment } i, \\
0, & \text{otherwise.}
\end{cases}$

Thus, the sample $(x,y) = ((x'_1, y'_1),\ldots,(x'_n, y'_n))'$ is incomplete with missing data $z = (z_{11}, z_{12}, \ldots, z_{cn})'$. The EM algorithm is used for estimations. The complete likelihood function is:

$$L' (\theta | (x,y), z) = \prod_{i=1}^{c} \prod_{j=1}^{n} \left[ \pi_i \pi_j (y_j | x_j) \right]^{z_{ij}},$$

where $\pi_i$ are mixing proportions and $\theta$ is the vector of parameters included in regressions, and $f_j(y_j | x_j)$ is the probability density function of response $Y_j$ given by regressors $X_j$ in the $i$th segment. The complete log-likelihood function is given by:

$$l' (\theta | (x,y), z) = \sum_{i=1}^{c} \sum_{j=1}^{n} z_{ij} \log \left( \pi_i \pi_j (y_j | x_j) \right)$$

For the E step of EM, we need the conditional expectation of $z_j$ given by $(x,y)$ under $\hat{\theta}$.

$$E(z_{ij} | (x,y), \theta^0) = \sum_{1 \leq \tau_1 < \tau_2 < \cdots < \tau_{c-1} \leq n-1} E(w_\tau | (x,y), \theta^0)$$
Based on Bayes’ theorem, 

\[ P\left(w_t = 1 | (x, y), \theta^0 \right) \]

is estimated by:

\[ \mathcal{E}_t^* = \frac{\prod_{i=1}^{r_i} f_i^0 \left( y_j | x \right) \prod_{j=1}^{r_j} f_j^0 \left( y_j | x \right) \mathcal{E}_t^*}{\sum_{i=1}^{c} \sum_{j=1}^{c} \prod_{i=1}^{r_i} f_i^0 \left( y_j | x \right) \prod_{j=1}^{r_j} f_j^0 \left( y_j | x \right) \mathcal{E}_t^*}, \quad (5) \]

where the superscript 0 indicates that parameters are given as \( \theta^0 \). Then, \( E\left(z_j | (x, y), \theta^0 \right) \) is estimated by:

\[ p^*_{ij} = \begin{cases} \sum_{i=1}^{c} \sum_{j=1}^{c} \mathcal{E}_t^*, & i = 1, \ldots, c, j = 1, \ldots, n, \\ 0, & \text{otherwise}. \end{cases} \quad (6) \]

Replace \( z_j \) in (4) with \( p^*_{ij} \) in Equation (6) and denote the resulted expression as \( Q\left(\theta | (x, y), \theta^0 \right) \). The M step is to maximize \( Q\left(\theta | (x, y), \theta^0 \right) \):

\[ Q\left(\theta | (x, y), \theta^0 \right) = \sum_{i=1}^{c} \sum_{j=1}^{n} p^*_{ij} \ln \left( \pi_i f_i \left( y_j | x \right) \right). \quad (7) \]

Repeating the steps E and M until the stop criterion is met, we then obtain the estimator of CPs as:

\[ \hat{\tau} = \arg \max_{\tau^*} \left( \mathcal{E}_t^* \right), \quad \tau^* = \left( \tau_1, \tau_2, \ldots, \tau_{c-1} \right)'. \]

To robustly estimate the CP regression model given by Equation (1), assume that the regression errors \( e_{ji} \), \( j = 1, \ldots, n \), follow a \( t \)-distribution, with \( r_i \) degrees of freedom (DF), scale parameter \( \sigma_i \), and the density:

\[ f(e | \sigma, r) = \frac{1}{\sigma h\left( \frac{e}{\sigma} | r \right), \quad h(t | r) = \frac{\Gamma\left( \frac{r+1}{2} \right)}{\left( \pi r \right)^{1/2} \Gamma\left( \frac{r}{2} \right) \left[ 1 + \left( \frac{e}{\sigma} \right)^2 \right]^{(r+1)/2}}, \quad (8) \]

where \( h(t | r) \) is the density of a standard \( t \)-distribution with \( r \) DF, \( t(r) \). We first suppose \( r_i \)s are known, and we discuss the decision of \( r_i \)s later. The complete log-likelihood is:

\[ l^*(\theta | (x, y), z) = \sum_{i=1}^{c} \sum_{j=1}^{n} z_{ij} \ln \left( \pi_i f_i \left( y_j - X' \beta_i | x_j, r_i, \sigma_i \right) \right), \]

where \( f_i(\cdot) \) is the density of Equation (8), \( X'_j = (1, x'_j) = (1, x_{j1}, \ldots, x_{jp}) \), \( \beta_i = (\beta_{i1}, \ldots, \beta_{ip})' \), \( \theta = (\theta'_1, \ldots, \theta'_c)' \). To simplify the computations, we express the \( t \)-distribution as a scale mixture of normal distributions (c.f. [12,28]). We introduce into the data vector additional missing data \( v = (v_1, \ldots, v_p)' \), satisfying that given \( z_{ij} = 1 \),
\[ y_j - X_j \beta | x_j, v_j, z_{ij} = 1 \sim N(0, \sigma_i^2 / v_j), \quad v_j | z_{ij} = 1 \sim \text{gamma} \left( \frac{r_j}{2}, \frac{r_j}{2} \right), \quad (9) \]

where \( \text{gamma}(\alpha, \lambda) \) has density
\[
g(v; \alpha, \lambda) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha v^{\alpha-1} e^{-\lambda v}, \quad \alpha, \lambda > 0.
\]

Then, \( y_j - X_j \beta | x_j, z_{ij} = 1 \) has a \( t(\sigma_i) \) distribution with a scale parameter \( \sigma_i \). The complete log likelihood function becomes:
\[
\ell^c(\theta | (x, y), z, v)
\]
\[
= \sum_{j=1}^{n} \sum_{i=1}^{c} z_{ij} \ln \left( \pi_i \sqrt{v_j} \phi \left( \frac{\sqrt{v_j} (y_j - X_j \beta_i)}{\sigma_i} \right) g(v_j | \sigma_i) \right)
\]
\[
= \sum_{j=1}^{n} \sum_{i=1}^{c} z_{ij} \ln \left( \pi_i \sqrt{v_j} \exp \left( -\frac{v_j (y_j - X_j \beta_i)^2}{2\sigma_i^2} \right) \frac{r_j}{2} v_j^{r_j/2-1} \exp(-v_j/2) \right)
\]
\[
= \sum_{j=1}^{n} \sum_{i=1}^{c} z_{ij} \ln \left( \pi_i \sqrt{v_j} \sqrt{2\pi\sigma_i} \right) + \sum_{j=1}^{n} \sum_{i=1}^{c} z_{ij} \ln(\pi_i) - \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{c} z_{ij} \ln(\sigma_i^2) - \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{c} z_{ij} \frac{v_j (y_j - X_j \beta_i)^2}{\sigma_i^2}
\]
\[
+ \sum_{j=1}^{n} \sum_{i=1}^{c} z_{ij} \ln \left( \frac{\sqrt{v_j}}{\sqrt{2\pi}} \frac{r_j}{2} \Gamma \left( \frac{r_j}{2} \right) \right) - \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{c} z_{ij} v_j r_j,
\]
where \( \phi(\cdot) \) is the density of \( N(0,1) \). To use EM, given the current estimate \( \theta^{(k)} \), since the last two terms do not contain unknown parameters \( \theta \), the computation of \( E \left( \ell^c(\theta | (x, y), z, v) | (x, y), \theta^{(k)} \right) \) in step E simplifies to computing \( E \left( z_{ij} | (x, y), \theta^{(k)} \right) \) and \( E \left( v_j | (x, y), \theta^{(k)} , z_{ij} = 1 \right) \). Based on Equation (6):
\[
E \left( z_{ij} | (x, y), \theta^{(k)} \right) = \rho_{ij}^{(k+1)} = \left\{ \begin{array}{ll}
\sum_{I \in \mathcal{I}} \sum_{C_{ij} < C_{ij}^{*} < C_{ij}^{*} \cdots} \varepsilon^{(k)}_{ij}, & i = 1, \cdots, c, \ j = 1, \cdots, n,
0, & \text{otherwise},
\end{array} \right. \quad (10)
\]
where \( \varepsilon^{(k)}_{ij} \) is computed by Equation (5) with the density \( f_j(y_j - X_j \beta_i^{(k)} | \sigma_i^{(k)}, r_j) \) given by Equation (8).

Moreover, the gamma distribution is the conjugate prior distribution for \( v_j \). It is easy to show that:
\[
v_j | (x, y), z_{ij} = 1, \theta^{(k)} \sim \text{gamma} \left( \frac{r_j + 1}{2}, \frac{1}{2} \left( r_j + \left( \frac{y_j - X_j \beta_i^{(k)}}{\sigma_i^{(k)}} \right)^2 \right) \right), \quad (11)
\]
\[
E\left(v_j \mid (x, y), \theta^{(k)}\right), z_j = 1 = q^{(k+1)}_j = \frac{\tilde{r}_j + 1}{\tilde{r}_j + \left[\frac{(y_j - X_j^i \beta^{(k)})}{\sigma^{(k)}_j}\right]^2}.
\]  
(12)

Then, step M is to maximize:

\[
E\left(f\left(0 \mid (x, y), z, v\right) \mid (x, y), \theta^{(k)}\right)
\]

\[
\propto Q(\theta \mid (x, y), \theta^{(k)}) = \sum_{i=1}^{n} \sum_{j=1}^{c} p_{j}^{(k+1)} \ln \pi_{j} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{c} p_{j}^{(k+1)} \ln \left(\sigma^{(k)}_j\right)^2 - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{c} p_{j}^{(k+1)} \left(\frac{y_j - X_j^i \beta^{(k)}_j}{\sigma^{(k)}_j}\right)^2
\]

Taking partial derivatives of \(Q(\theta \mid (x, y), \theta^{(k)})\) with respect to all parameters with the Lagrange multiplier under constraint \(\sum_{i=1}^{n} \pi_{j} = 1\), we obtain the updated estimates:

\[
\beta^{(k+1)}_j = \left(X_j^U \beta^{(k+1)}_j \right)^{-1} X_j^U Y_j,
\]

(13)

\[
\left(\sigma^{(k+1)}_j\right)^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{c} p_{j}^{(k+1)} q_{j}^{(k+1)} \left(y_j - X_j^i \beta^{(k+1)}_j\right)^2}{\sum_{j=1}^{n} \sum_{j=1}^{c} p_{j}^{(k+1)} q_{j}^{(k+1)}},
\]

(14)

where \(X \in \mathbb{R}^{n \times (p+1)}\) is a matrix with the \(j\)th row \(X_j = (1, x_j', \ldots, x_{j,p})', Y \in \mathbb{R}^n\) is a vector with the \(j\)th component \(y_j\) and \(U_i^{(k+1)}\) is a diagonal matrix with the \(j\)th diagonal element equal to \(p_{j}^{(k+1)} q_{j}^{(k+1)}\). Based on the above, we propose an EMT Algorithm 1 for estimating CP regression models, as follows.

**Algorithm 1. EMT**

Set \(s = \mathbb{C}_{c-1}^{0}, \quad \epsilon^{(k)} = \left[\epsilon^{(k)}_1: \tau = (\tau_1, \tau_2, \ldots, \tau_{c-1})', 1 \leq \tau_1 < \tau_2 < \cdots < \tau_{c-1} \leq n-1\right]_{\tau_{\text{rand}}} , \)

\[
p^{(k+1)} = \left[p_{j}^{(k+1)}: 1 \leq i \leq c, 1 \leq j \leq n\right]_{\text{rand}} \in \mathbb{R}^{c \times n},
\]

\[
q^{(k+1)} = \left[q_{j}^{(k+1)}: 1 \leq i \leq c, 1 \leq j \leq n\right]_{\text{rand}} \in \mathbb{R}^{c \times n}.
\]

Let \(\eta > 0, k = 0, \epsilon^{(0)} = 1/\mathbb{C}_{c-1}^{0}, \) for each possible \(\tau, \) and \(q^{(0)}_j = 1, i = 1, \cdots, c, j = 1, \cdots, n. \)

Step 1. Compute \(\mathbf{p}^{(0)}\) with Equation (6) using \(\epsilon^{(0)}\). Then, compute \(\mathbf{p}^{(0)}_i, \)

\[
\left(\sigma^{(0)}_j\right)^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{c} p_{j}^{(0)} q_{j}^{(0)} \left(y_j - X_j^i \beta^{(0)}_j\right)^2}{\sum_{j=1}^{n} \sum_{j=1}^{c} p_{j}^{(0)} q_{j}^{(0)}},
\]

(15)

where \(\eta > 0, \) for each possible \(\tau.\)

Step 2. Compute \(\mathbf{p}^{(k+1)}\) and \(\mathbf{q}^{(k+1)}\) with Equations (10) and (12), respectively.

Step 3. Compute \(\mathbf{p}^{(k+1)} = \left(\beta^{(k+1)}_1, \cdots, \beta^{(k+1)}_c\right)' \in \mathbb{R}^{c \times p+1} \) with \(\mathbf{p}^{(k+1)}\) and \(\mathbf{q}^{(k+1)}\) by Equation (13).

Step 4. Compute \(\sigma^{(k+1)} = \left(\sigma^{(k+1)}_1, \cdots, \sigma^{(k+1)}_c\right)' \) with \(\mathbf{p}^{(k+1)}, \mathbf{p}^{(k+1)}\) and \(\mathbf{q}^{(k+1)}\) by Equation (14).
Step 5. Update \( \epsilon^{(k+1)} \) with \( \beta^{(k+1)} \), \( \sigma^{(k+1)} \) and \( \epsilon^{(k)} \) by Equation (5) with the density given by Equation (8).

Step 6. If \( \| \mathbf{e}^{(k)} - \mathbf{e}^{(k)} \| < \eta \), then stop; else, \( k = k + 1 \) and go to Step 2.

After the EMT algorithm converges, the estimates of \( c - 1 \) CPs are given by

\[
\hat{\tau} = \arg \max_{i \in I \cup T, j \in \mathbb{T}, \sigma \in \mathbb{G}, \pi \in \Pi} \left( \epsilon_i \right), \quad \tau = (\tau_1, \tau_2, \ldots, \tau_{c-1})'.
\]

3.2. A FCT Algorithm for \( t \) CP Regression Models

Since the EM algorithm converges slowly when data involve noises (c.f. [36]), we extend the crisp model to a fuzzy class model for improving the convergence rate and the precision of estimations. Similar to Lu and Chang [10], using the fuzzy partitioning concept, we regard each probable CP combination as a partition of data with a fuzzy membership and then transform them into the pseudo memberships of data belonging to each individual segment. As in previous descriptions, consider a regression model with \( c - 1 \) CPs, as given in Equation (1). Suppose \( \tau = (\tau_1, \tau_2, \tau_{c-1})' \), \( 1 \leq \tau_1 < \tau_2 < \cdots < \tau_{c-1} \leq n - 1 \), is a collection of CPs with a fuzzy membership \( \alpha_i \) under the conditions \( 0 \leq \alpha_i \leq 1 \) and \( \sum_{i=1}^{c-1} \sum_{j=1}^{c} \alpha_i = 1 \). Now, let \( z_j \) be the fuzzy membership of observation \((x_j, y_j)\) belonging to the \( i \)th piece. Then, \( z_j \) can be expressed by \( \alpha_i \) as follows (c.f. [10]):

\[
z_j = \begin{cases} 
\sum_{i=1}^{c-1} \sum_{j=1}^{c} \alpha_i, & i = 1, \ldots, c - 1, \quad 1 \leq j \leq n - c + i, \\
0, & i = 1, \ldots, c - 1, \quad 1 \leq j < i \text{ or } n - c + i < j \leq n, \\
1 - \sum_{i=1}^{c-1} \mu_j, & i = c, \quad 1 \leq j \leq n.
\end{cases}
\]  

(15)

Now, the CP model is a fuzzy class model with a fuzzy \( c \)-partition, \( \{z_1, \ldots, z_n\} \), \( z_i = (z_{i1}, \ldots, z_{in})' \), \( i = 1, \ldots, c \), and \( \sum_{j=1}^{n} z_j = 1 \), \( j = 1, \ldots, n \). The regression parameters can be estimated by maximizing the fuzzy classification maximum likelihood (FCML) function (c.f. [10,37]):

\[
J(\theta) = \sum_{i=1}^{c} \sum_{j=1}^{n} z_{ij} \ln f_i((x_j, y_j) | \theta) + w \sum_{i=1}^{c} \sum_{j=1}^{n} z_{ij} \ln \pi_i
\]  

(16)

subject to \( \sum_{i=1}^{c} \pi_i = 1 \), \( \pi_i > 0 \), where \( m, c > 1 \), \( r \geq 0 \), \( z_j \in [0,1] \), \( \sum_{i=1}^{c} z_j = 1 \), \( i = 1, \ldots, c \), \( j = 1, \ldots, n \) and \( \theta = (\beta_1', \ldots, \beta_r', \sigma_{\epsilon}, \sigma_{\beta}, \sigma_{\pi}, \ldots) \)' .

As stated, assume that the CP regression errors, \( e_j = y_j - X j \beta_j \), \( j = 1, \ldots, n \), have a \( t(r) \)-distribution. With the scale mixture expression of \( t(r) \) in Equation (9), the FCML objective function in Equation (16) is:

\[
T_{m,w}(z, \pi, \beta, \sigma) = \sum_{i=1}^{c} \sum_{j=1}^{n} z_{ij} \ln f_j(y_j | \beta, \sigma) + w \sum_{i=1}^{c} \sum_{j=1}^{n} z_{ij} \ln \pi_i
\]
Based on Equation (12), replace $v_j$ with $q_j$ in $T_{m,w}(z, \pi, \beta, \sigma)$. Take the partial derivatives of the Lagrange expression with $T_{m,w}(z, \pi, \beta, \sigma)$ under constraint $\sum_{i=1}^{n} \pi_i = 1$, then, we obtain the updated estimates:

$$
\beta^{(k+1)} = (X U^{(k)} X')^{-1} X U^{(k)} Y,
$$

$$
\sigma^{(k+1)} = \frac{\sum_{j=1}^{n} z_{ij}^{(k)} q_{ij} (y_j - X' \beta^{(k)})^2}{\sum_{j=1}^{n} q_{ij}^{(k)}},
$$

$$
\pi^{(k+1)} = \frac{\sum_{j=1}^{n} z_{ij}^{(k)}}{n},
$$

where $U^{(k)}_j$ is a diagonal matrix with the jth diagonal element equal to $z_{ij}^{(k)} q_{ij}$. Having the regression estimates $\hat{\theta}$, we update the CP memberships $\alpha, \tau$ as follows. Given a CP collection $\tau$, define:

$$
d^{(k)}_i = -\sum_{j=1}^{c} \sum_{\tau_j \in \tau} \ln f_i(y_j | x_j, \hat{\theta})
$$

We then optimize the objective function:

$$
J(\omega) = \sum_{i} \cdots \sum_{\tau \in \tau} \omega^{\tau} \exp(d^{(k)}_i)
$$

subject to $\sum_{\tau \in \tau} \omega^{\tau} = 1, 0 \leq \omega^{\tau} \leq 1$. The updated $\omega^{\tau}$ is:

$$
\omega^{\tau} = \frac{\exp(-d^{(k)}_i)^{-1/(m-1)}}{\sum_{i} \cdots \sum_{\tau \in \tau} \exp(-d^{(k)}_i / (m-1))}
$$

We propose an FCT Algorithm 2 for robustly fitting CP regression models, as follows.
Algorithm 2. FCT

Set $s = C^{-1}_{c-1}$, $\alpha^{(k)} = \left[ \alpha_{i}^{(k)} : \tau = (\tau_{1}, \tau_{2}, \ldots, \tau_{c-1})', \tau_{1} \leq \tau_{2} < \cdots < \tau_{c-1} \leq n-1 \right]_{\text{mod}}$

Let $\eta > 0$, $k = 0$, $\alpha_{c}^{(0)} = 1/C_{c-1}$, for each possible $\tau$, and $q_{y}^{(0)} = 1$, $i = 1, \ldots, c$, $j = 1, \ldots, n$.

Step 1. Compute $z_{ijy}^{(0)}$ $1 \leq i \leq c$, $1 \leq j \leq n$ by Equation (15).

Step 2. Compute $p_{i}^{(k+1)}$, $\sigma_{i}^{(k+1)}$, $\theta_{i}^{(k+1)}$, $i = 1, \ldots, c$, by Equations (17)–(19), respectively.

Step 3. Update $d_{ijy}^{(k+1)}$ by (20) using $p_{i}^{(k+1)}$, $\sigma_{i}^{(k+1)}$, $i = 1, \ldots, c$, and Equation (8).

Step 4. Compute $z_{ijy}^{(k+1)}$ and $d_{ijy}^{(k+1)}$ by Equations (12) and (15), respectively.

Step 5. If $\left| d_{ijy}^{(k+1)} - d_{ijy}^{(k)} \right| < \eta$, then stop; else, $k = k + 1$ and go to step 2.

After FCT converges, the estimates of $c - 1$ CPs are given by $\hat{\tau} = \arg \max_{1 \leq \tau_{1} < \tau_{2} < \cdots < \tau_{c-1} \leq n} \left( e_{i} \right)$, $\tau = (\tau_{1}, \tau_{2}, \ldots, \tau_{c-1})'$.

3.3. Decide the Degrees of Freedom of t-Distributions

In this section, we discuss the decision of the degrees of freedom (DF), $r = (r_{1}, \ldots, r_{c})'$. Typically, one can derive the maximum likelihood estimator (MLE) of $r$ by maximizing the log-likelihood over $r$. Another solution is to compute the likelihood over a grid set, $G = \{1, 2, \ldots, r_{\text{max}}\}$, and choose DF from $G$ which has the largest likelihood (c.f. [28]). As is known, $t(r)$ approaches normal distributions for a large enough $r$. In practice, $r$ is not necessary to be very large for achieving normality. Similar to [31], we set $r_{\text{max}} = 15$. After performing lots of simulations, we found that choosing $r$ based on a grid set often fails to provide CP estimates that are precise enough. However, numerous simulation results show that both EMT and FCT with $r = 1$ (EMT1 and FCT1) always produce rather precise CP estimates, no matter whether data are contaminated or not. Accurate CP estimates are critical to solving CP regression problems. To demonstrate that EMT1 and FCT1 are generally suitable to all kinds of data, we employ EMT and FCT with $r = 1, 5, 10$ and 15 to diversified datasets randomly generated from the models of 1 and 2 CPs, with data including 0, 5, 10 and 15 noises, respectively. The models considered for generating data are shown in Figure 1 with regression errors following a normal distribution of $N(0, 0.5^2)$ (except for the cases in Figure 1, more situations and different kinds of models were also examined, but results are similar, and hence, we only report four cases here). We evaluate the performance based on $R$ replications in each individual circumstance with $R$ equal to 1000 and 200 for models of 1 and 2 CPs, respectively. The evaluation criteria includes (1) the mean of CP estimates ($\overline{CP}$) along with the standard errors (SECP), and (2) the sum of the mean squared errors (MSES) of model coefficient estimates computed as follows:

$$\text{MSES} = \sum_{i,j} \text{MSE}(\beta_{ij}) = \frac{1}{R} \sum_{r=1}^{R} \left( \overline{\beta}_{ij} - \beta_{ij} \right)^2$$
where \( \hat{\beta}_{ij(r)} \) is the estimate of \( \beta_{ij} \) in the rth replication. The normal-based EM-BP algorithm for CP regressions presented in [10] is also employed and renamed EMN for comparisons. For illustrative purposes, the true regression coefficients and one of the simulated series are shown in Figure 1. Table 1 summarizes the simulation results.

Figure 1. Models used to show EMT1 performs better, \( \beta \): *(1, 0.5), +: (6, -0.5) show the coefficients of the model from which data were generated; here, \( \beta \): *: (1, 0.5), +: (6, -0.5) indicate that the data marked * and + in the figure were generated from models \( E(y) = 1 + 0.5 x \) and \( E(y) = 6 - 0.5 x \), respectively. These notations are also used for the figures below.

Table 1. The performance of EMT and FCT with different degrees of freedom.

| (I) 1 CP Cases | CP = 25 | CP = 33 | CP = 26 | CP = 56 |
|----------------|---------|---------|---------|---------|
| **(a) 0 outliers** | Method | EMN | EMT1 | EMT5 | EMT10 | EMT15 | FCT1 | FCT5 | FCT10 | FCT15 |
| CP1hat | 25.18 | 25.15 | 25.26 | 25.20 | 25.19 | 25.21 | 25.10 | 25.11 | 25.13 | 25.13 |
| SECP  | 0.168 | 0.186 | 0.169 | 0.166 | 0.166 | 0.120 | 0.109 | 0.109 | 0.109 | 0.109 |
| MSES  | 0.76  | 1.21  | 0.80  | 0.78  | 0.78  | 0.98  | 0.58  | 0.56  | 0.55  |
| **(b) 15 noises** | Method | EMN | EMT1 | EMT5 | EMT10 | EMT15 | FCT1 | FCT5 | FCT10 | FCT15 |
| CP1hat | 18.80 | 33.74 | 21.75 | 24.41 | 21.75 | 33.51 | 29.86 | 29.70 | 28.70 |
| SECP  | 0.137 | 0.154 | 0.237 | 0.272 | 0.237 | 0.111 | 0.130 | 0.123 | 0.109 |
| MSES  | 14.17 | 0.98  | 10.68 | 8.24  | 10.68 | 0.77  | 0.82  | 1.65  | 2.38  |
| **(c) 1 extreme outlier (2, 8)** | Method | EMN | EMT1 | EMT5 | EMT10 | EMT15 | FCT1 | FCT5 | FCT10 | FCT15 |
| CP1hat | 18.55 | 25.79 | 25.56 | 24.70 | 23.34 | 26.06 | 26.06 | 25.94 | 25.85 |
| SECP  | 0.112 | 0.167 | 0.147 | 0.149 | 0.152 | 0.064 | 0.063 | 0.063 | 0.063 |
| MSES  | 1.73  | 1.02  | 0.70  | 0.69  | 0.82  | 0.75  | 0.49  | 0.47  | 0.47  |

Table 1 demonstrates that EMT1 and FCT1 generally perform better than those with different degrees of freedom (DF) by virtue of more precise CP estimates and less MSE in the estimations of regression coefficients in all cases, especially the cases of data with contaminations. For clean data such as in Figure 1a, all EMTs and FCTs with different DF provide sufficient and comparable estimates of CPs and regression coefficients. However, for data containing noisy points, such as the cases of Figure 1b,d, only EMT1 and FCT1
work well in all, but the other EMTs and FCTs mostly fail to provide sufficient CP estimates. As illustrated by Peel and McLachlan [15], the use of $t$ components provides less extreme estimates of the posterior probabilities of segment membership of the CP model. The longer tail of $t$-distributions makes the fitted CP models more robust by giving reduced weight to atypical observations in the calculation of CP estimates. Furthermore, the $t$ model-based approach embeds the normal distribution for each segment into a wider class of elliptically symmetric distributions, with an additional parameter called DF which is key for tuning robustness. This illustrates that the long tail of $t(1)$-distributions makes EMT1 and FCT1 more robust than those with higher DF. Even with normal data involving no contaminations, EMT1 and FCT1 still perform as well as EMN because of the similarity of symmetry between normal and $t$-distributions. Moreover, the best robustness of EMT1 and FCT1 among all EMTs and FCTs can be seen from the robustness properties of the proposed methods illustrated in Section 3.5. Thus, one may first fit data using EMN, EMT and FCT with several different DF, such as $r = 1, 5, 10$ and 15, respectively. If the estimates of EMN, EMTs and FCTs are close, one may use EMN, considering that the data are clean and normally distributed. Otherwise, EMT1 and FCT1 should be adopted.

### 3.4. Resistance against High Leverage Outliers

Similar to the M-estimates for linear regressions, the proposed EMT and FCT are still susceptible to high leverage outliers. Here, high leverage outliers refer to outliers in the space of independent variables, since points far away from the main body of points in the $X$ space have high leverage. Modifications for solving this problem are needed. In regression analyses, the leverage score of an observation $(x_j, y_j)$ is often defined as:

$$h_j = n^{-1} + (n-1)^{-1} \text{MD}_j,$$

where

$$\text{MD}_j = (x_j - \bar{x})'(S^{-1}(x_j - \bar{x})), \quad \bar{x} = (x_1, \ldots, x_p)^', \quad x_j = (x_{j1}, \ldots, x_{jp})'.$$

$\bar{x}$ and $S$ are the sample mean and the sample covariance of $x_i$s, respectively. However, $\bar{x}$ and $S$ are nonresistant to outliers, and leverage points may not be recognized due to the effect of other high leverage points [34]. To lessen the mask effect, we consider replacing $\bar{x}$ and $S$ with the minimum variance determinant (MCD) estimators [38] among all the subsets of $\{x_1, x_2, \ldots, x_n\}$ having $h$ elements, $1 \leq h \leq n$. In choosing $h$, we adopt the suggestion proposed by Rousseeuw and Leroy [39] and Lopuhaa and Rousseeuw [40], that $h = (n + p + 1)/2$ should be utilized to have the largest breakdown value for the MCD estimators. Through the fast MCD (FastMCD) algorithm of Rousseeuw and Van Driessen [41] with the small sample corrections proposed by Pison [42], the dataset is trimmed by removing the top $\alpha \times 100\%$ of the data sorted in the descending order of the leverages of data. Then, EMT/FCT is applied to the trimmed data. Since the proposed EMT/FCT is only sensitive to outliers with rather high leverages, a small value of $\alpha$ such as 0.05 is usually enough in practice. To advance the ability of the proposed algorithms in withstanding outliers, for data involving doubtful high leverage outliers, we suggest employing EMT/FCT after trimming those observations with a modified MD ranking in the top $\alpha \times 100\%$ of MDs arranged in declining order using a small $\alpha$.

### 3.5. The Robust Properties of EMT and FCT

In this section, we employ the influence function and gross error sensitivity to illustrate the ability of the proposed algorithms in withstanding outliers. These two criteria have been frequently used to measure robustness in statistics [13]. Suppose $y_1, \ldots, y_n$ are
observed data for estimating the parameter $\theta$. An M-estimator is created by minimizing the form:

$$
\sum_{j=1}^{n} \rho(y_j; \theta),
$$

(22)

where $\rho$ is a function that measures the difference between the assumed distribution and the true one. An M-estimator is derived by solving the equation:

$$
\sum_{j=1}^{n} \varphi(y_j; \theta) = 0,
$$

(23)

where $$
\varphi(y_j; \theta) = \left( \frac{\partial}{\partial \theta} \right) \rho(y_j; \theta).
$$

For example, the sample mean and median are the M-estimators for the loss function $\rho(y_j; \theta) = (y_j - \theta)^2$ and $\rho(y_j; \theta) = |y_j - \theta|$, respectively. Consider a random sample, $Y_1, \ldots, Y_n$, from a population of density function $f_j(y; \theta)$ and let $\rho(y_j; \theta) = -\log f_j(y_j; \theta)$. Then, the MLE is also an M-estimator.

The influence function (IF) is useful for evaluating the relative influence of individual observations toward the value of an estimate. The IF of an M-estimator is proportional to its $\varphi$ function with the form:

$$
\text{IF}(y; \theta) = -\frac{\varphi(y; \theta)}{\int \frac{\partial \varphi(y; \theta)}{\partial \theta} dF_j(y)},
$$

(24)

where $F_j(y)$ denotes the distribution function of $Y$. An outlier might bring about issues when the influence function of an estimator is unbounded. Many useful robustness measures can be investigated through IF. For example, the gross error sensitivity ($g^*$) denoted by

$$
g^* = \sup_y |\text{IC}(y; F, \theta)|
$$

(25)

is one important robustness measure. $g^*$ can be thought of as the approximate magnitude of the worst influence that the addition of an infinitesimal point mass may have on the value of the associated estimator.

For the proposed EMT, the considered estimator is the MLE associated with the $t$ density having $r$ degrees of freedom:

$$
f_j(y; \theta) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{(\pi r)^{\frac{r}{2}} \Gamma\left(\frac{r}{2}\right)} \left(1 + \frac{(y - \theta)^2}{r}\right)^{-\frac{r+1}{2}}.
$$

Maximizing $\log f_j(y; \theta)$ is equivalent to minimize

$$
\ln \left(1 + \frac{(y - \theta)^2}{r}\right).
$$

The loss function associated with the MLE obtained by EMT$r$ ($r$ denotes DF) is equivalent to:
\[
\rho(y; \theta) = \ln \left(1 + \frac{(y - \theta)^2}{r} \right)
\]

and hence the \( \varphi(y; \theta) \) is:

\[
\varphi(y; \theta) = \frac{-2(y - \theta)}{r + (y - \theta)^2}.
\]

We only need to analyze \( \varphi(y; \theta) \) because IF is proportional to \( \varphi(y; \theta) \) according to Equation (24). Note that:

\[
\lim_{y \to \infty} \varphi(y; \theta) = 0.
\]

Equation (28) indicates that IF approaches to 0 as \( y \) approaches to positive or negative infinity. The maximum and minimum values of \( \varphi(y; \theta) \) can also be derived by solving

\[
(\partial / \partial y) \varphi(y; \theta) = 0.
\]

Therefore, the \( \varphi(y; \theta) \) in Equation (27) is bounded and continuous, as exhibited in Figure 2a. Furthermore, Figure 2a shows that for all \( y \) values, the IF of EMT1 is closest to 0 among all EMTr, which means that EMT1 is the most robust among all EMTr. Besides, Figure 2a demonstrates that for any given observation \( y \), the IF of EMTr increases with \( r \), and tends to be unbounded as \( r \to \infty \) and \( |y| \to \infty \). This investigation reflects that the IF associated with EMN is unbounded because \( t(r) \) approaches to a normal distribution, as \( r \to \infty \). In fact, the loss function corresponding to normal regressions is equivalent to \( \rho(y; \theta) = (y - \theta)^2 \) with \( \varphi(y; \theta) = -2(y - \theta) \). Thus, the IF with EMN is unbounded and the gross error sensitivity \( g^* = \infty \). This is the reason why the normal-based MLE is not robust to outliers.

![Figure 2](image_url)

**Figure 2.** (a) \( \varphi \) function of EMTr and (b) \( \varphi \) function of the median.
The plot of IF for the median is also displayed in Figure 2b. The influence of an excessively large or small \( y \) on the median is constant. However, the influence of an extremely large or small \( y \) on the estimator derived by EMT1 is tiny according to Equation (28) and Figure 2a. In fact, Equation (28) implies that an extremely large or small \( y \) can be viewed as an observation which may come from a different but unknown distribution and have no influence (i.e., \( IC(y; F, \theta) = 0 \)) on our estimator. The maximum influence of observations, \( y \), on our estimators can be derived by solving Equation (29). Hence, the proposed estimator has a bounded and continuous IF, and also a finite gross error sensitivity. Thus, the \( t \) CP model-based approach is robust from the point of view of robust statistics.

EMT1 and FCT1 are more robust than those with DF greater than 1.

4. Experiments and Examples

4.1. Simulation Studies

We used numerical examples to demonstrate the effectiveness and the superiority of the proposed method. In all simulations, the stop criterion and initial values are set as \( \eta = 0.000005 \) and \( \epsilon^{(0)}_r = 1/C_\tau^{\alpha-1} \), for each possible combination of \( \tau - 1 \) CPs, \( \tau = (\tau_1, \ldots, \tau_{\tau-1})' \), \( 1 \leq \tau_1 < \tau_2 < \cdots < \tau_{\tau-1} \leq n-1 \). A sample of size \( n \) is randomly generated by selecting \( n \) values of the partitioning variable uniformly distributed over the interval \([0, 10]\). Then, divide the interval \([0, 10]\) evenly into \( c \) nonoverlapping subintervals, \((d_{i-1}, d_i) \), \( i = 1, \ldots, c \), with \( d_0 = 0 \) and \( d_c = 10 \). The values of the response variable \( y \) are computed according to respective regression settings. Seven methods used for comparisons include:

1. Using EM with errors having normal distributions [10] (EMN)
2. Rank-based segmented regression method [43] (RSE)
3. EMro-t in [44], re-denoted as EM-Tk
4. The proposed EMT with DF equal to 1 (EMT1)
5. The trimmed version of EMT1 with trimmed proportion, \( \alpha = 5\% \) (EMT1-tm)
6. The proposed FCT with DF equal to 1 (FCT1)
7. The trimmed version of FCT1 with trimmed proportion, \( \alpha = 5\% \) (FCT1-tm)

The comparisons are based on three models under four respective scenarios using sample sizes of 50 and 75 with 1000 and 200 replicates for 1 and 3 CPs models in each simulation run, respectively. For illustrative purposes, Figure 3 displays the true coefficients of the models used, and one of the simulated series including generated data, fitted lines and noises. Four respective scenarios considered for three models are similar and include the cases of data having (1) 0 atypical observations, (2) several background noises, (3) high leverage outliers and (4) heavy-tailed distributions. The evaluation measures for comparisons are (1) the MSE of CP estimates (MSE-CP) described as

\[
\text{MSE}(\hat{\tau}) = \frac{1}{R} \sum_{k=1}^{R} (\hat{\tau}_k - \tau)^2
\]

where \( \hat{\tau}_k \) is the CP estimate at \( k \)th replication, and (2) the sum of MSE of the regression coefficient estimates (MSES-reg) as given by Equation (21). Table 2 reports the results.
Figure 3. Plots of models used for comparing the proposed approaches with existing methods.

Table 2. Comparisons of 7 methods with models a–c under 4 different situations, respectively.

| Model a: 1 CP Discontinuous Model | (a.1) 0 outliers | EMN | RSE | EM-Tk | EMT1 | EMT1-tm | FCT1 | FCT1-tm |
|-----------------------------------|-----------------|-----|-----|-------|------|---------|------|---------|
| MSE-CP                            | 0.91            | 37.75 | 0.50 | 2.84  | 2.68 | 0.43    | 0.40 |         |
| MSES-reg                          | 0.57            | 0.50  | 0.52 | 0.67  | 0.71 | 0.76    | 0.79 |         |
| (a.2) 8 (14%) noises              | EMN             | RSE  | EM-Tk | EMT1  | EMT1-tm | FCT1 | FCT1-tm |
| MSE-CP                            | 66.98           | 111.16| 135.25| 0.87  | 1.19 | 0.98    | 0.92 |         |
| MSES-reg                          | 2.26            | x    | 0.82  | 0.65  | 0.82 | 0.74    | 0.98 |         |
| (a.3) leverage outliers (10, 10)  | EMN             | RSE  | EM-Tk | EMT1  | EMT1-tm | FCT1 | FCT1-tm |
| MSE-CP                            | 569.16          | x    | 54.68 | 573.70| 2.80 | 574.47  | 0.39 |         |
| MSES-reg                          | 11556.91        | x    | x    | 10,900.66| 0.71 | x      | 0.75 |         |
| (a.4) errors~t (2)               | EMN             | RSE  | EM-Tk | EMT1  | EMT1-tm | FCT1 | FCT1-tm |
| MSE-CP                            | 6.79            | 44.96| 1.17  | 2.69  | 2.95 | 0.41    | 0.41 |         |
| MSES-reg                          | 12603.61        | x    | 2.98  | 5066.66| 3395.77| 1.54 | 0.75 |         |

Model b: 1 CP Continuous Model

| (b.1) 0 outliers | EMN | RSE | EM-Tk | EMT1 | EMT1-tm | FCT1 | FCT1-tm |
|------------------|-----|-----|-------|------|---------|------|---------|
| MSE-CP           | 0.024| 0.004| 0.014 | 0.009| 0.022   | 0.004| 0.008   |
| MSES-reg         | 0.600| 0.620| 0.574 | 0.831| 1.020   | 0.784| 0.875   |
| (b.2) 10 (17%) noises | EMN | RSE | EM-Tk | EMT1 | EMT1-tm | FCT1 | FCT1-tm |
| MSE-CP           | 201.574| 0.535| 208.182| 0.573| 1.245   | 0.277| 0.143   |
| MSES-reg         | 14.174| 2.939| 9.714 | 0.978| 1.745   | 0.771| 1.215   |
| (b.3) leverage outliers (10, 15) | EMN | RSE | EM-Tk | EMT1 | EMT1-tm | FCT1 | FCT1-tm |
| MSE-CP           | 335.479| x    | 2.428 | 477.948| 0.024 | 347.620| 0.004 |
| MSES-reg         | 7451.722| x    | 1.518 | 9694.875| 0.986 | 7884.283| 0.853 |
| (b.4) errors~t (2) | EMN | RSE | EM-Tk | EMT1 | EMT1-tm | FCT1 | FCT1-tm |
| MSE-CP           | 0.038| 0.002| 0.041 | 0.017| 0.019   | 0.013| 0.011   |
| MSES-reg         | 1346.945| 1.742| 2.834 | 1.640| 1.825   | 1.072| 1.218   |

Model c: 2 CPs Model

| (c.1) 0 outliers | EMN | RSE | EM-Tk | EMT1 | EMT1-tm | FCT1 | FCT1-tm |
|------------------|-----|-----|-------|------|---------|------|---------|
| MSE-CP1          | 0.223| 0.170| 0.490 | 0.490| 0.228   | 0.271| 0.457   |
| MSE-CP2          | 5.745| 0.041| 6.142 | 6.142| 5.361   | 3.218| 5.194   |
| MSES-reg         | 4.798| x    | 6.019 | 6.019| 8.255   | 7.580| 8.587   |

Model: $\hat{y} = \beta_0 + \beta_1 x_i$, $\beta_i = (\beta_{i1}, \beta_{i2})$, $i = 1, \ldots, c$

(a.1) 1CP, discontinuous model
$\beta_i: (1, 0.5), *: (4, -0.5)$

(b.1) 1CP, continuous model
$\beta_i: (1, 1.5), *: (11, -0.5)$

(c.1) 2CPs, continuous model
$\beta_i: (0, 0.5), *: (3, -0.5), \cdot (-3, 0.5)$
Table 2 shows that EMT1 and FCT1 perform better or at least comparable to the other methods in all cases of all three models, except the case of data contaminated by leverage outliers. However, for the cases with high leverage outliers, the trimmed versions, EMT1-tm and FCT1-tm, work well. Particularly, FCT1 usually performs better than EMT1 because the influence of outliers can be greatly reduced by moderately increasing the power of the fuzzy index $m$ used in FCT1. More findings are stated as follows:

1. For datasets including no atypical observations, all EM-type and FCT methods (including EMN, EM-Tk, EM-T, EMT1, EMT1-tm, FCT1 and FCT1-tm) work well for all three models, with FCT1 performing better than the others for less biased CP estimates and less MSES-reg with regression coefficient estimates.

2. RSE works well for the 1 CP continuous model (model b). However, RSE is not workable in cases of models with discontinuity (model a) or multiple CPs (model c). However, these situations frequently occur in practice.

3. For data with background noises such as cases a.2, b.2 and c.2, only the proposed $t$-based approach works generally well. Particularly, FCT1 provides more precise estimates.

4. For datasets including extremely high leverage outliers such as cases a.3, b.3 and c.3, all methods cannot work well, except EMT1-tm and FCT1-tm. In such situations, both trimmed versions provide effective estimates of CPs and regression coefficients for each model, and FCT1-tm performs better for producing less biased estimates on the whole.

5. From extensive simulations, the iteration time respectively taken by EMT1 and FCT1 in all cases is always less than other EM-based methods. This evidence indicates that the proposed method is most computationally economic. Especially, it is even more timesaving than the normality-based EMN.

Next, we show that the proposed algorithms are more tolerant of extreme outliers in comparison with the existing methods based on model b in Figure 3. Four cases of extreme contaminations occurring at different locations are considered, with a simulated sample shown in Figure 4.

Cases 1 and 4: 5 (9.1%) identical outliers occur at (0, 7) and (10, 12), respectively. The leverage of outliers in these two cases is high.

Cases 2 and 3: 8 (13.8%) identical outliers occur at (3, 0) and (8, 4), respectively. The leverage of outliers in these two cases is moderate.

Note that in all four cases, the distance of the observed $y$-value from the predicted $y$-value is larger than 6 deviations, and the percentage of extreme outliers is more than 9.1% in each case. The estimation results are shown in Table 3.

Table 3 shows that EMT1 and FCT1 perform better or at least comparable to the other methods in all cases of all three models, except the case of data contaminated by leverage outliers. However, for the cases with high leverage outliers, the trimmed versions, EMT1-tm and FCT1-tm, work well. Particularly, FCT1 usually performs better than EMT1 because the influence of outliers can be greatly reduced by moderately increasing the power of the fuzzy index $m$ used in FCT1. More findings are stated as follows:

1. For datasets including no atypical observations, all EM-type and FCT methods (including EMN, EM-Tk, EM-T, EMT1, EMT1-tm, FCT1 and FCT1-tm) work well for all three models, with FCT1 performing better than the others for less biased CP estimates and less MSES-reg with regression coefficient estimates.

2. RSE works well for the 1 CP continuous model (model b). However, RSE is not workable in cases of models with discontinuity (model a) or multiple CPs (model c). However, these situations frequently occur in practice.

3. For data with background noises such as cases a.2, b.2 and c.2, only the proposed $t$-based approach works generally well. Particularly, FCT1 provides more precise estimates.

4. For datasets including extremely high leverage outliers such as cases a.3, b.3 and c.3, all methods cannot work well, except EMT1-tm and FCT1-tm. In such situations, both trimmed versions provide effective estimates of CPs and regression coefficients for each model, and FCT1-tm performs better for producing less biased estimates on the whole.

5. From extensive simulations, the iteration time respectively taken by EMT1 and FCT1 in all cases is always less than other EM-based methods. This evidence indicates that the proposed method is most computationally economic. Especially, it is even more timesaving than the normality-based EMN.

Next, we show that the proposed algorithms are more tolerant of extreme outliers in comparison with the existing methods based on model b in Figure 3. Four cases of extreme contaminations occurring at different locations are considered, with a simulated sample shown in Figure 4.

Cases 1 and 4: 5 (9.1%) identical outliers occur at (0, 7) and (10, 12), respectively. The leverage of outliers in these two cases is high.

Cases 2 and 3: 8 (13.8%) identical outliers occur at (3, 0) and (8, 4), respectively. The leverage of outliers in these two cases is moderate.

Note that in all four cases, the distance of the observed $y$-value from the predicted $y$-value is larger than 6 deviations, and the percentage of extreme outliers is more than 9.1% in each case. The estimation results are shown in Table 3.
Figure 4. Four cases of extreme outliers considered to show the robustness of the proposed methods.

Table 3. Comparing the robustness of 7 methods to extreme outliers using model b under four situations.

| Case | 5 (9.1%) Outliers, CP = 30 | 6 (10.7%) Outliers, CP = 30 | 8 (13.8%) Outliers, CP = 30 | 8 (13.8%) Outliers, CP = 25 | 5 (9.1%) Outliers, CP = 25 | 6 (10.7%) Outliers, CP = 25 |
|------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| Model: | $E(y)=1+1.5x$; $E(y)=1-0.5x$, regression errors, $e \sim N(0,0.5^2)$. |
| Case 1 | 5 outliers at (0,7), CP = 30 |
| MSECP | 25.19 | 13.16 | 32.16 | 36.84 | 5.74 | 14.47 | 8.42 | 25.9 | 85.9 | 32.0 | 113 |
| MSEreg | 1.35 | x | 2.30 | x | 1.42 | 0.46 | 0.07 | 0.41 | 1.27 | x | 2.19 | x |
| EMN | 33.44 | 13.53 | 46.31 | 44.85 | 7.34 | 27.11 | 22.77 | 34.6 | 96.0 | 46.1 | 146 |
| RSE | 0.01 | 1.02 | 3.32 | 4.40 | 4.14 | 1.74 | 4.10 | 34.40 | 0.06 | 0.96 | 3.78 | 30.1 |
| EMT1 | 0.04 | 0.78 | 1.93 | 8.64 | 0.03 | 1.07 | 0.00 | 0.59 | 0.02 | 0.72 | 2.40 | 96.5 |
| FCT1 | 0.11 | 1.59 | 16.38 | 13.88 | 1.79 | 1.78 | 1.83 | 22.05 | 0.16 | 2.37 | 17.3 | 140 |
| FCT1-tm | 0.04 | 1.02 | 5.22 | 13.32 | 0.05 | 1.43 | 0.01 | 3.25 | 0.01 | 2.02 | 5.59 | 143 |

$^1$: MSECP and MSEreg are the MSE of the estimates of CP and regression parameters given by Equations (21) and (30), respectively.

Similar to previous comparisons, Table 3 shows that FCT1 performs best in all four cases of extreme contaminations. Particularly, FCT1 and FCT1-tm are much more resistant against radical outliers than other methods for providing estimates as precise as those obtained from clean data. Specifically, in cases 1 and 4, where the drastic outliers (0, 7) and (10, 12) have rather high leverages and their responses are distant from the predicted values for 12 standard deviations, FCT1 methods are still hardly affected by outliers and produce adequate estimates even if extreme outliers account for almost 10% of the whole dataset. Moreover, as explained earlier, FCT1 is more robust than EMT1 in that the fuzzy partitioning and the fuzziness index, $m$, are helpful for reducing the influence of outliers by giving tiny weights with a large $m$ to the exceptional outliers. Table 3 demonstrates again that the normal-based method, EMN, cannot work well in the presence of extreme outliers. Additionally, while RSE is quite robust to outliers with moderate leverages, it cannot provide satisfied regression estimates for data involving drastic outliers with fairly high leverages.

In summary, the proposed FCT1 method is very robust to outliers, even though the dataset includes extreme outliers with the observed responses far away from the predicted values, by more than 6 standard deviations. Furthermore, FCT1 can still produce sufficient estimates for datasets in the presence of radical outliers with quite high leverages. More importantly, FCT1 is tolerant of the extreme outliers, which account for about 10% of the total data in all four cases of the experiment. In practice, outliers far away from the predicted regression line, by more than 6 standard deviations, are unusual and very likely less than 10% of the whole dataset. Thus, the proposed FCT1 is attainable in real applications, generally.

4.2. Real Data Applications

Example 1. We applied the proposed methods to a dataset involving 107 terrestrial animals gathered by Garland [45] for analyzing the relationship between the maximal running speed (MARS) with the body size. It is understandable that animals of different
sizes can reach different maximal running speeds (MARS). McMahon [46] commented that animals are built for elastic similarity and the MARS of animals is proportional to the fourth root of body mass, \((\text{mass})^{1/4}\), among elastically similar animals. The plot in Figure 5 shows that a slope change occurs in the relationship between \((\text{MARS})^{1/4}\) and \((\text{mass})^{1/4}\). Additionally, several exceptionally low points appear near the place where the regression slope converts. It is appreciable that these suspicious points very likely pull down the regression line and hence have a substantial influence on the regression slope.

We respectively apply EMN, EMT and FCT to the MARS data by setting \(DF = 1\). The fitting results in Figure 5 show that EMT produces a much shorter CP at 25 and the line on the right hand side becomes flatter due to the down pull exerted by those extremely low points (see Figure 5, the blue fitted line by EMT and \(\hat{\beta}_{21}^{EMT} = -0.0456\), flatter than those by EMN and FCT, \(\hat{\beta}_{21}^{EMN} = -0.0892\) and \(\hat{\beta}_{21}^{FCT} = -0.0820\)). The CPs produced by EMR (\(i^{EMR} = 46\)) and FCT (\(i^{FCT} = 41\)) are very close, but the slope of the left line (\(\hat{\beta}_{41}^{EMN} = 0.4973\)) obtained by EMN is much lower than that by FCT1 (\(\hat{\beta}_{41}^{FCT} = 0.9118\)). This evidence indicates that FCT1 is much more resistant against those extraordinarily low points than EMN. Based on previous simulation results showing that FCT1 usually provides less biased estimates than other models, we believe that the estimated model provided by FCT1 is reasonable and best suited to MARS data. The MARS data can be obtained from [45].

\[ E(Y) = \beta_{0} + \beta_{1}x, \hat{\beta}_{i} = (\hat{\beta}_{0i}, \hat{\beta}_{1i}), i = 1, 2 \]

(a) EMN: \(CP = 46\), \((1.8988, 2.4663)\)

\( \hat{\beta}_{1} = (1.6401, 0.4973), \hat{\beta}_{2} = (3.0937, -0.0892) \)

(b) EMT: \(CP = 25\), \((0.8801, 2.1147)\)

\( \hat{\beta}_{1} = (1.4990, 0.7233), \hat{\beta}_{2} = (2.9025, -0.0456) \)

(c) FCT: \(CP = 41\), \((1.6266, 2.7832)\)

\( \hat{\beta}_{1} = (1.4058, 0.9118), \hat{\beta}_{2} = (3.0678, -0.0820) \)

\( a: (1.8988, 2.4663) = (\sqrt{\text{mass}}, \sqrt{\text{MARS}}) \)

Figure 5. Fitting results of EMN, EMT and FCT to MARS data.

Example 2. We analyzed a bedload transport dataset (downloaded from https://doi.org/10.2737/RDS-2007-0004 (accessed on 26 May 2017)) to study the relationship between the discharge rate \((\text{m}^3/\text{s})\) and the rate of bedload transport \((\text{kg/s})\). Bedload transport measures the transportation of particles in flowing fluids along a bed. The bedload transportation is often described as occurring in two phases, starting from a relatively slow stage (Phase I) to a faster stage (Phase II). The transition point is the main interest of researchers in studying bedload transportation. Ryan and Porth [47] had fitted the dataset with a piecewise regression model, and they found it difficult because few data were collected at higher flows. The scatter plots of data in Figure 6 show that the transportation rate tends to increase with the water discharge faster after the discharge rate goes beyond 1.5 \(\text{m}^3/\text{s}\). Furthermore, the two points on the upper right having extraordinarily high transportation rates are leverage points.
Figure 6. Fitting results of EMN, EMT1, EMT1-tm, FCT1 and FCT1-tm to bedload transport data.

We implemented EMN, EMT1, EMT1-tm, FCT1 and FCT1-tm with DF = 1 and trimming percentage $\alpha = 0.05$ to the data. Figure 6 shows that EMN and EMT1 are heavily affected by two high leverage points and both give an identical CP at 74 (discharge rate, DR = 1.81354). This is irrational for only two points included in Phase II. By contrast, the other three methods are much more robust, resulting in the same CP at 62 (DR = 1.5786) and much lower slope estimates for the second line, especially FCT1 and FCT1-tm. Particularly, the slope of the second line, $\hat{\beta}_{21} = 0.0282$, obtained by FCT1-tm is much smaller than those obtained by other methods. These facts illustrate that FCT1 and FCT1-tm are more resistant against unusual observations; especially, FCT1-tm is capable of withstanding high leverage outliers. Our resulting CP estimate, 62, looks reasonable and it is close to the finding of Zhang and Li [48]. As for deriving more accurate estimates for Phase II, just as suggested by Ryan and Porth [47], more data at faster flows are required.

Example 3. We applied the proposed algorithms to the Dow Jones Index closing price ranged from 22 October 2008 to 22 October 2009 (downloaded from https://finance.yahoo.com/quote/%5EDJI/history?ltr=1 (accessed on 20 July 2021)). Due to the financial crisis of 2008 with a bankruptcy of Lehman Brothers, U.S. stock market price series suffered a huge fluctuation from late 2008 to 2009. Thus, the U.S. stock market price series during this period is worthy of investigations to illustrate variance changes in real situations. Adopting an approach common in finance, we consider the stock returns, $R_t = (P_t - P_{t-1})/P_t$, where $P_t$ is the closing price on day $t$. Since the values of $\{R_t\}$ are too small, we applied our methods to the series $\{R'_t\}$, $R'_t = 100\exp(R_t)$. The descriptive statistics of $\{R'_t\}$ in Table 4 and the box plots of the whole data and the 4 sub-segments derived by EMN in Figure 7 reveal that the series $\{R'_t\}$ follows a fat-tailed distribution and includes several extreme outliers. Therefore, $t$-regression models are better-suited to the series $\{R'_t\}$.
Figure 7. (a) Box plots of \( R^c \) with the whole data, and (b) the four sub-segments derived by EMN.

Table 4. Descriptive statistics of \( \{ R^c \} \).

| Sample Size | Mean   | SD    | Minimum | Median  | Maximum | Skewness | Kurtosis |
|-------------|--------|-------|---------|---------|---------|----------|----------|
| 252         | 101.11 | 2.1451| 92.59   | 100.06  | 111.49  | 0.573    | 7.074    |

Lin et al. [35] studied the same data and concluded three CPs at 35, 133 and 182. We set \( c = 4 \) and applied EMN, EMT1 and FCT1 to \( \{ R^c \} \), separately. EMN produces 3 CPs at 32, 122 and 182, which are close to those of Lin et al. [35]. However, the box plots in Figure 7b show that the last three segments obtained by EMN contain extreme outliers. Hence, the CPs estimated by EMN are suspicious due to its sensitivity to outliers and heavy-tails. Moreover, the estimated CPs of Lin et al. are obtained by the binary hierarchic segmentation, which does not generally provide the true optimum (cf. [11,49]). On the other hand, EMT1 and FCT1 produced 3 CPs at 21, 32 and 108 and 21, 32 and 151, in which the second segment contains only 11 observations and the difference of variations between the first two segmentations is not statistically significant. Furthermore, the scree test for choosing the number of CPs (see the discussion in Section 5) suggests that the optimal number of segments is 3. The box plots in Figure 8 show that several outliers appear in the last two segments. According to previous experiments, FCT1 is powerful in resisting atypical observations and it is more robust than EMT1. Thus, the CP estimates, 33 and 151, provided by FCT1 should be the optimal times of variance changes in the return series.
5. Conclusions and Discussion

5.1. Discussion

Computation cost is usually a concern in CP detection methods. The capacity of a robust method in resisting outliers for different sizes of datasets is important. For most statistical methods, it is harder to achieve robustness with a small dataset because estimations are easily affected by anomalous observations for small amounts of data. In order to show the high robustness of the proposed method even for small samples, we used datasets consisting of tens or hundreds of data points in the experiments. When fitting a regression model with \( c - 1 \) CPs to a dataset of size \( n \) by EMT or FCT, we consider \( C_{c-1} \) potential collections of CPs, simultaneously. Therefore, the computational efficiency can be impeded as the size of the dataset or the number of CPs increases. In fact, except for the sample size and the number of CPs, other factors may influence the computational cost, such as the model variations, the model structures, such as continuity or discontinuity at CPs, and the discrepancies of models between two contiguous segments. According to our simulations executed on a computer of an Intel Core i7-7700HQ CPU with 2.80 GHz and 8 Gb of Ram, the proposed methods are suitable for datasets of sizes no more than 2000, 1000 and 500 for locating one, two and three CPs, respectively. For large datasets, a possible approach is to cut the sample size to the favorable size and then derive preliminary estimates by using the reduced data. One possible way for reducing the data size is to remove the data of even rank in the ordered dataset arranged in the increasing order of the partitioning variable, \( \bar{x}_{d} \), then repeatedly reduce the data until the favorable size is reached. After deriving the rudimentary estimates, we can further estimate CPs and regression parameters by employing EMT to the nearby data of the preceding CPs again.

The determination of the number of CPs is important in CP detection. However, it is difficult when lacking likelihood-based tests for the number of CPs because the regularity conditions fail and there is little knowledge about the asymptotic distribution of the test statistics even for Gaussian cases. Most CP studies usually adopted some information criteria or informal tests for deciding the number of segments. Here, we employed the...
idea of Hawkins’s scree test [49] to investigate the progressive changes in the residual sum of squared errors \( RSS(c-1) - RSS(c) \), \( c \) is the number of segments fitted to the data), with the changes in the MSE described as:

\[
MSE(c) = \frac{RSS(c)}{[n-c(p+1)]} = \frac{\sum (y_i - \hat{y}_i)^2}{[n-c(p+1)]},
\]

and the changes in the maximized log-likelihood multiplied by \(-2 (-2\text{lnL})\) (i.e., \( F(c-1) - F(c) \), \( F(c) \) is \(-2\) multiplying the optimized log-likelihood obtained from a model of \( c \) pieces fitted to the data). Additionally, Ciuperca’s [50] generalized Bayesian Information Criterion (BIC) was also utilized.

The suggested indexes were employed to the four models, shown in Figure 9, with the data contaminated by background noises, which includes outliers with a quite high leverage. For data contaminated by high leverage outliers, the trimmed version with \( \alpha = 0.05 \) was used and the results were similar to those of cases including no outliers. Hence, we do not present the results of cases with high leverage outliers here. We report the results of both EMT1 and FCT in Table 5. For models including zero or one single CP, all four suggested indexes (RSS, MSE, \(-2\text{lnL}\) and BIC) provide a correct \( c \) for each respective case. For the models of multiple CPs, the BIC and \(-2\text{lnL}\) index fails to obtain the correct \( c \). The RSS and MSE indexes work well in all circumstances. It can be easily seen from Table 5 that both the indexes RSS and MSE decreased progressively for \( c \) less than the true value, and then the decrease becomes less again once \( c \) goes beyond the true value. These two indexes provide a clear indication of the optimal \( c \). Therefore, the two suggested indexes, RSS and MSE, are useful for determining the number of segments. However, in some applications, the cut-point may not be so clearly indicated by these indices. In such circumstances, one may fit the data with several possible \( c \)’s based on RSS and MSE and then choose the best fitted model on the basis of some criteria in regression analysis. Although there is no solid inferential base for the suggested method, it may suffice as a rough and practical tool.

1. 0 CP, \( c = 1 \), 15 noises
2. 1 CP, \( c = 2 \), 15 noises
3. 2 CPs, \( c = 3 \), 10 noises
4. 3 CPs, \( c = 4 \), noises

![Figure 9. Models used for deciding the number of segments based on the suggested indices.](image)

| try | 0 CP \( c = 1 \), 15 noises | 1 CP \( c = 2 \), 15 noises | 2 CPs \( c = 3 \), 10 noises | 3 CPs \( c = 4 \), 10 noises |
|-----|-----------------|-----------------|-----------------|-----------------|
| \( c \) | RSS | \(-2\text{lnL}\) | MSE | BIC | RSS | \(-2\text{lnL}\) | MSE | BIC | RSS | \(-2\text{lnL}\) | MSE | BIC | RSS | \(-2\text{lnL}\) | MSE | BIC |
| 1 | x | x | x | 115 | x | x | x | 92 | x | x | x | -5 | x | x | 90 | x |
| 2 | -9 | -59 | -1 | 133 | 98 | 997 | 2 | 79 | 18 | 37 | 0.13 | 3 | 190 | 430 | 33 | 3.88 |
| 3 | * | * | -891 | 467 | 11 | -172 | 0 | 108 | 8 | -220 | 0.37 | 11 | 35 | -50 | 30 | 0.72 |
| 4 | 0 | 1 | -43 | 517 | -3 | -217 | 0 | 159 | 3 | 17 | -0.02 | 62 | 4 | -39 | 74 | 0.04 |

(i) Using EMT1

| try | 0 CP \( c = 1 \), 15 noises | 1 CP \( c = 2 \), 15 noises | 2 CPs \( c = 3 \), 10 noises | 3 CPs \( c = 4 \), 10 noises |
|-----|-----------------|-----------------|-----------------|-----------------|
| \( c \) | RSS | \(-2\text{lnL}\) | MSE | BIC | RSS | \(-2\text{lnL}\) | MSE | BIC | RSS | \(-2\text{lnL}\) | MSE | BIC |
| 1 | x | x | x | 115 | x | x | x | 88 | x | x | x | -10 | x | x | 90 | x |
On the other hand, the penalized likelihood approach has been utilized in recent studies, such as [7,8], to infer both the number of CPs and their locations. For large datasets, one may consider the penalized approaches with a penalty added into the objective likelihood to estimate the number and the locations of CPs simultaneously. However, the selection of a proper penalty is not easy because the penalty can have a substantial impact on the estimations for the number of CPs [28]. One may refer to Haynes et al. [51] for information on workable penalties.

In using the trimmed EMT, a small value of $\alpha$ was selected to remove leverage points where the leverage scores rank in the top $\alpha \times 100\%$. Apparently, the choice of $\alpha$ is connected with the results of EMT1-tm and FCT1-tm. A formal statistical test is required for objectively judging whether a datum is a high leverage point based on a cut-point, which is the critical point resulted from a statistical test under a given significance level. Accordingly, high leverage points can be deleted adaptively but not fixed at a subjective value. Moreover, high leverage points may present small residuals and hence provide important information for model fitting. Therefore, more research on incorporating the information of high leverage points with small residuals is also needed to judge high leverage points better. One possible method is to borrow the idea of monitoring introduced by Cerioli et al. [52]. First, the trimmed version with a small $\alpha$ is employed to derive robust estimates of CPs and regression parameters. Then, the technique of monitoring in [52] may be considered for judging the trimmed points to make the most of the dataset. The test statistic for outliers in CP regressions would be better based on the difference in the estimates of CPs and model parameters between different samples. Further research is needed, and it will be covered in our future work.

### 5.2. Conclusions

This paper proposed a new robust method for CP regressions using $t$-distributions. The $t$-distribution provides a longer tailed substitute to the normal distribution and has an additional parameter called DF, which is critical for tuning robustness. These favorable properties make the $t$ CP regression model more tolerant of atypical observations. We considered CPs as random variables and transferred them into latent class variables so that we could implement the EM algorithm to estimate the CPs and regression parameters simultaneously. Moreover, we rewrote the $t$-distribution as a scale mixture of normal distributions to simplify the computation of $t$-log-likelihood. We further extended the latent class model to a fuzzy class model. Incorporating fuzzy clustering methods, we created the FCT algorithm to make more robust estimations for CP regressions by FCT than by EMT. The proposed EMT and FCT are computationally easy and efficient compared to the other existing EM-type methods. The proposed method has no problem with the non-differentiability and requires no restrictions on the models, such as the continuity between two adjoining segments needed, only a single CP allowed, or an unchanged regression variance demanded. Some of these restraints are often required in existing methods. To advance the ability of the proposed methods in resisting high leverage outliers, we proposed the modified versions, called EMT-tm and FCT-tm, by first trimming high leverage points moderately and then employing the proposed method to the reduced data. As for the decision of the degrees of freedom, $r$, the EMT1 and FCT1 (DF $r = 1$) worked well in all experiments, no matter whether the data were contaminated or not. The analysis of robust properties with the proposed $t$-based methods also showed
that EMT1 and FCT1 were most tolerant of atypical observations. In managing CP regression problems, we suggest respectively fitting data with EMN, EMTs and FCTs using values such as \( r = 1, 5, 10 \) and 15. If the estimations of the three methods are close, one may consider that the data are normally distributed, having no abnormal observations, so that EMN can be used for obtaining less varied estimators. Otherwise, EMT1 or FCT1 should be employed. In practice, FCT1 can be given priority for its better ability of withstanding atypical observations. For the concern about the trimmed proportions of data, \( \alpha \), normally, outliers with extremely high leverage are seldom more than 5\%, thus \( \alpha \) can be set as 0.05 at the start and increased in case a higher value is needed (cf. \([28,47]\)). Extensive experiments with numerical and real examples showed the effectiveness and the practicability of EMT1 and FCT1. In particular, only the trimmed versions were workable for data containing extremely high leverage outliers, such as the real data of bedload transportations. This evidence shows the effectiveness and the need of the trimmed versions for real applications. Overall, the proposed EMT and FCT approaches are robust to atypical observations and heavy-tailed distributions. Especially, FCT1 generally works well and performs better than EMT1 for providing less biased estimates of CPs and model parameters.

**Author Contributions:** Conceptualization and methodology, K.-P.L. and S.-T.C.; software, S.-T.C.; validation, K.-P.L.; formal analysis, S.-T.C.; investigation, K.-P.L.; resources, K.-P.L.; data curation, K.-P.L.; writing—original draft preparation, K.-P.L.; writing—review and editing, K.-P.L. and S.-T.C.; visualization, K.-P.L.; supervision, K.-P.L.; project administration, K.-P.L.; funding acquisition, K.-P.L. and S.-T.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was funded by the Ministry of Science and Technology, Taiwan (Grant numbers MOST 108-2118-M-025-001 and MOST 108-2118-M-003-003).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The sources of datasets are respectively stated in each example they are considered (pages 21–23).

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. Muggeo, V.M.R. Segmented: An R package to fit regression models with broken-line relationships. *News R Proj.* **2008**, *8*, 20–25.
2. Yang, P.; Dumont, G.; Ansermino, J.M. Adaptive change detection in heart rate trend monitoring in anesthetized children. *IEEE Trans. Biomed. Eng.* **2006**, *53*, 2211–2219.
3. Schröder, A.L.; Ombao, H. FreSpeD: Frequency-specific change-point detection in Epileptic seizure multi-channel EEG data. *J. Am. Stat. Assoc.* **2019**, *114*, 115–128.
4. Loschi, R.H.; Pontel, J.G.; Cruz, F.R.B. Multiple change-point analysis for linear regression models. *Chil. J. Stat.* **2010**, *1*, 93–112.
5. Werner, R.; Valev, D.; Danov, D.; Guineva, V. Study of structural break points in global and hemispheric temperature series by piecewise regression. *Adv. Space Res.* **2015**, *56*, 2323–2334.
6. Fearnhead, P.; Rigail, G. ChangePoint Detection in the Presence of Outliers. *J. Am. Stat. Assoc.* **2019**, *114*, 169–183.
7. Frick, K.; Munk, A.; Sieling, H. Multiscale change point inference. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2014**, *76*, 495–580.
8. Pein, F.; Sieling, H.; Munk, A. Heterogeneous change point inference. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2017**, *79*, 1207–1227.
9. Zarandi, M.H.F.; Alaeddini, A. A general fuzzy-statistical clustering approach for estimating the time of change in variable sampling control charts. *Inf. Sci.* **2010**, *180*, 3033–3044.
10. Lu, K.P.; Chang, S.T. A fuzzy classification approach to piecewise regression models. *Appl. Soft Comput.* **2018**, *69*, 671–688.
11. Fryzlewicz, P. Wild binary segmentation for multiple change-point detection. *Ann. Stat.* **2014**, *42*, 2243–2281.
12. Maronna, R.A.; Martin, R.D.; Yohai, V.J. *Robust Statistics: Theory and Methods*; Wiley: New York, NY, USA, 2006.
13. Huber, P.J. *Robust Statistics*; Wiley: New York, NY, USA, 1981.
14. Lange, K.L.; Little, R.J.A.; Taylor, J.M.G. Robust statistical modelling using the t distribution. *J. Am. Stat. Assoc.* **1989**, *84*, 881–889.
15. Peel, D.; McLachlan, G.J. Robust mixture modelling using the t distribution. *Stat. Comput.* **2000**, *10*, 339–348.
16. Muggeo, V.M.R. Estimating regression models with unknown breakpoints. *Stat. Med.* **2003**, *22*, 3055–3071.
17. Chakar, S.; Lebarbier, E.; Lévy-Leduc, C.; Robin, S. A robust approach for estimating change-points in the mean of an AR(1) process. *Bernoulli* **2017**, *23*, 1408–1447.
18. Ko, S.I.M.; Chong, T.T.L.; Ghosh, P. Dirichlet process hidden Markov multiple change-point model. *Bayesian Anal.* 2015, 10, 275–296.
19. Bardwell, L.; Fearnhead, P. Bayesian detection of abnormal segments in multiple time series. *Bayesian Anal.* 2015, 12, 193–218.
20. Zou, C.; Yin, G.; Feng, L.; Wang, Z. Nonparametric maximum likelihood approach to multiple change-point problems. *Ann. Stat.* 2014, 42, 970–1002.
21. Haynes, K.; Fearnhead, P.; Eckley, I.A. A computationally efficient nonparametric approach for changepoint detection. *Stat. Comput.* 2017, 27, 1293–1305.
22. Rigail, G. A pruned dynamic programming algorithm to recover the best segmentations with 1 to K_max change-points. *J. Soc. Fr. Stat.* 2015, 156, 180–205.
23. Maidstone, R.; Hocking, T.; Rigail, G.; Fearnhead, P. On optimal multiple changepoint algorithms for large data. *Stat. Comput.* 2017, 27, 519–533.
24. Aminikhanghahi, S.; Cook, D.J. A survey of methods for time series change point detection. *Knowl. Inf. Syst.* 2017, 51, 339–367.
25. Truong, C.; Oudre, L.; Vayatis, N. A review of change point detection methods. *arXiv* 2018, arXiv:1801.00718v1.
26. Ciuperca, G. Estimating nonlinear regression with and without change-points by the LAD method. *Ann. Inst. Stat. Math.* 2011, 63, 717–743.
27. Ciuperca, G. Penalized least absolute deviations estimation for nonlinear model with change-points. *Stat. Pap.* 2011, 52, 371–390.
28. Yang, F. Robust Mean Change-Point Detecting through Laplace Linear Regression Using EM Algorithm. *J. Appl. Math.* 2014, 2014, 856350.
29. Jafari1, A.; Yarmohammadil, M.; Rasekhi, A. A Bayesian analysis to detect change-point in two-phase Laplace model. *Sci. Res. Essay* 2016, 11, 187–193.
30. Gerstenberger, C. Robust Wilcoxon-type estimation of change-point location under short range dependence. *J. Time Ser. Anal.* 2018, 39, 90–104.
31. Yao, W.; Wei., Y.; Yu, C. Robust mixture regression using the t-distribution. *Comput. Stat. Data Anal.* 2014, 71, 116–127.
32. Lin, J.; Zhu, L.X.; Xie, F.C. Heteroscedasticity diagnostics for t linear regression models. *Metrika* 2009, 70, 59–77.
33. Lin, J.G.; Xie, F.; Wei, B.C. Statistical Diagnostics for Skew-t-normal Nonlinear Models. *Commun. Stat. Simul. Comput.* 2009, 38, 2096–2110.
34. Osorio, F.; Galea, M. Detection of a change-point in student-t linear regression models. *Stat. Pap.* 2005, 45, 31–48.
35. Lin, J.G.; Chen, J.; Li, Y. Bayesian Analysis of Student t Linear Regression with Unknown Change-Point and Application to Stock Data Analysis. *Comput. Econ.* 2012, 40, 203–217.
36. Petersen, K.B.; Winther, O.; Hansen, L.K. On the slow Convergence of EM and VBEM in low-noise linear Models. *Neural Comput.* 2005, 17, 1921–1926.
37. Yang, M.S. A survey of fuzzy clustering. *Math. Comput. Model.* 1993, 18, 1–16.
38. Rousseauw, P.J. Least median of squares regression. *J. Am. Stat. Assoc.* 1984, 79, 871–880.
39. Rousseauw, P.J.; Leroy, A.M. *Robust Regression and Outlier Detection*; Wiley-Interscience: New York, NY, USA, 1987.
40. Lopuhaa, H.P.; Rousseuw, P.J. Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices. *Ann. Stat.* 1991, 19, 229–248.
41. Rousseauw, P.J.; Van Driessen, K. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 1999, 41, 212–223.
42. Pison, G.; Van Aelst, S.; Willems, G. Small sample corrections for LTS and MCD. *Metrika* 2002, 55, 111–123.
43. Shi, S.; Li, Y.; Wan, C. Robust continuous piecewise linear regression model with multiple change points. *J. Supercomput.* 2020, 76, 3623–3645.
44. Lu, K.P.; Chang, S.T. Robust algorithms for multiphase regression models. *Appl. Math. Model.* 2020, 77, 1643–1661.
45. Garland, T. The relation between maximal running speed and body mass in terrestrial mammals. *J. Zool.* 1983, 199, 157–170.
46. McMahon, T.A. Using body size to understand the structural design of animals: Quadrupedal locomotion. *J. Appl. Physiol.* 1975, 39, 619–627.
47. Ryan, S.; Porth, L. *A Tutorial on the Piecewise Regression Approach Applied to Bedload Transport Data*; General Technic Report RMRS-GTR-189; U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station: Fort Collins, CO, USA, 2007; p. 41.
48. Zhang, F.; Li, Q. Robust bent line regression. *J. Stat. Plan. Inference* 2017, 183, 41–55.
49. Hawkins, D.M. Fitting multiple change-point models to data. *Comput. Stat. Data Anal.* 2001, 37, 323–341.
50. Ciuperca, G. A general criterion to determine the number of change-points. *Stat. Probab. Lett.* 2011, 81, 1267–1275.
51. Haynes, K.; Eckley, I.A.; Fearnhead, P. Computationally Efficient Changepoint Detection for a Range of Penalties. *J. Comput. Graph. Stat.* 2017, 26, 134–143.
52. Cerioli, A.; Riani, M.; Atkinson, A.C.; Corbellini, A. The power of monitoring: How to make the most of a contaminated multivariate sample. *Stat. Methods Appl.* 2018, 27, 641–649.