CQE Center for Quantitative Economics

Accounting for Asymmetry in M-Estimation

Manuel Stapper[†]

109/2024

[†] Department of Economics, University of Münster, Germany and
Department of Infectious Disease Epidemiology and Dynamics, London School of Hygiene and Tropical Medicine

wissen.leben
Universität Münster

Accounting for Asymmetry in M-Estimation

Manuel Stapper[1,2*]

[1*]Institute of Econometrics and Economic Statistics, University of Münster.

[2]Department of Infectious Disease Epidemiology and Dynamics, London School of Hygiene and Tropical Medicine.

Abstract

Standard M-Estimation techniques are biased if an asymmetric distribution is assumed. This article proposes a novel approach that uses an adaptive asymmetric loss function to tackle the bias. Its consistency and asymptotic normality are proven. The robustness properties are assessed in a simulation study showing similar performance compared to existing approaches. Its versatility is demonstrated in three applications to time series data, an instrumental regression and a classification task.

Keywords: Robust Statistics, M-Estimation, Computational Statistics

1 Introduction

In statistical analysis it is common to make assumptions about an underlying model and use suitable techniques to estimate model parameters. If these assumptions hold true, optimal estimators can often be derived. According to Huber and Ronchetti (2009), the field of robust statistics presents alternative methods for situations where reality and observed data deviate from model assumption. Sometimes, these deviations are minor and do not affect classical estimation much. This article focuses on cases where extreme values are observed, posing a thread to the validity of model assumptions. In this context, an extreme observation is one that is highly unlikely based on an assumed distribution, thus likely to violate underlying assumptions.

Extreme observations are common in practice and can arise from various causes. In some cases, rare events are of particular interest and should be addressed explicitly. For instance, financial data often includes extreme observations such as large stock

returns. In meteorology, extreme events may manifest as hurricanes, floods or earthquakes. In insurance, large claim amounts following natural disasters are common. When modelling the spread of infectious diseases, epidemics or pandemics are examples for unusual events. Various approaches have proven useful to incorporate outlying observations, including heavy-tailed and mixture distributions, or regime-switching models. The model assumptions are formulated to explain the extreme observations we see in reality rather than applying robust techniques.

In other situations, it is sensible to limit the influence of extreme observations. For instance, when studying the spread of contagious diseases, focusing on typical spread patterns may be more relevant than unusual epidemic behaviour. Another common reason for disregarding extreme observations are measurement errors. These errors can arise in physics from faulty sensors or disturbances during an experiment. In survey analysis, measurement errors may occur due to participants misunderstanding a question or due to errors when reading in questionnaires. One obvious approach that we often see in applications is removing observations from a sample before estimation. However, the removal of outliers relies on the practitioner's judgment and the definition of threshold values for considering an observation outlying is arbitrary. Moreover, extreme observations are sometimes concealed. In multivariate applications, an observation may be usual based on the marginal distributions but an outlier when considering the joint distribution. Finally, outlier removal is a discrete decision. Methods should be preferred that allow for a smooth transition between usual and outlying observations. One of the fundamental robust estimation techniques is M-estimation, which allows such a smooth transition.

The concept of M-estimation dates back to the 1960s with fundamental work by Tukey (1960) and Huber (1964). Initially, M-Estimators intended to provide a robust estimate of a location parameter with no assumption of a distribution. Nowadays M-Estimators are applied in both non-parametric and parametric approaches. This article discusses M-Estimation in parametric frameworks, starting from simple random samples and then generalising to more complex applications.

The rest of the article is organised as follows. Section 2 introduces the concept of M-Estimation for simple random samples. It focuses on the assumption of an asymmetric distribution, where standard M-Estimation is biased, presents an established correction approach as well as a novel correction approach. Consistency and asymptotic normality are proven for the new approach under mild assumptions. Section 3 assesses the estimators' robustness properties through simulation studies. Section 4 discusses the generalisation from simple random samples to regression and time series frameworks and multivariate classification in real life data applications. Finally, Section 5 gives an outlook and concludes.

2 Methodology

As an introductory example, let $X_1, ..., X_n$ be a independent and identically distributed (iid) sample following a Poisson distribution with mean λ_0 we want to

estimate robustly. With a loss function $\rho_k(z)$, we estimate λ_0 by minimising

$$\sum_{i=1}^{n} \rho_k \left(\frac{X_i - \lambda}{\sqrt{\lambda}} \right) \tag{1}$$

with respect to λ. The loss function is assumed to be a symmetric loss functions yielding a robust estimate, typically defined in terms of a tuning constant k that controls the trade-off between robustness and efficiency. Analytically equivalent, the estimate can be found by solving

$$\sum_{i=1}^{n} \psi_k \left(\frac{X_i - \lambda}{\sqrt{\lambda}} \right) \stackrel{!}{=} 0 \ , \tag{2}$$

where ψ_k is the simple derivative of ρ_k. Defining the weight function $w_k(z) := \psi_k(z)/z$ and rearranging the estimation function above, we obtain

$$\lambda = \frac{\sum_{i=1}^{n} w_k(Z_i) X_i}{\sum_{i=1}^{n} w_k(Z_i)} \ , \tag{3}$$

where $Z_i = \frac{X_i - \lambda}{\sqrt{\lambda}}$. In above fixed point equation, λ is found by iterating between computing weights and plugging the weights in to update the estimate of λ. We refer to the three approaches as ρ-type, ψ-type and weight-type M-Estimators respectively.

2.1 Consistency Correction

In above example of a Poisson distribution, we see that the ψ-type estimator solves equation (2), however for the true parameter λ_0, we have

$$\mathrm{E}_{\lambda_0} \left[\psi_k \left(\frac{X_i - \lambda_0}{\sqrt{\lambda_0}} \right) \right] \neq 0$$

due to the skewness of the Poisson distribution and the symmetry of ψ_k. That induces a bias in the estimation which does not vanish with increasing sample size. To tackle the bias and make the estimator consistent, a frequently used approach corrects the estimation equation by subtracting above expectation in the estimation equation, as for example discussed by Cantoni and Ronchetti (2001). Instead of solving (2), the estimator is found by finding the root of

$$\sum_{i=1}^{n} \psi_k \left(\frac{X_i - \lambda}{\sqrt{\lambda}} \right) - c(\lambda, k) \tag{4}$$

where $c(\lambda, k) = \mathrm{E}_\lambda \left[\psi_k \left(\frac{X_i - \lambda}{\sqrt{\lambda}} \right) \right]$ is referred to as correction term in the following and depends on the distribution of X_i and the tuning constant k. The estimation equation is a natural generalisation of (2); whenever X_i is symmetrically distributed and ψ_k is

symmetric, the correction term is zero and the estimation equation reduces to equation (2).

Instead of using a correction term, an alternative approach is now introduced which makes the estimation equation have zero expectation by stepping back from the restriction of a symmetric ψ function. For that purpose, one can simply introduce a second tuning constant, such that we apply different tuning constants depending on the sign of the input, i.e.

$$\psi_{k_L,k_U}(z) = \begin{cases} \psi_{k_L}(z) & \text{if } z \leq 0 \\ \psi_{k_U}(z) & \text{if } z > 0 \end{cases}$$

where k_L is referred to as lower tuning constant and likewise k_U the upper tuning constant. Then, one of the two tuning constants is held fixed and the respective other is chosen in such a way that

$$\mathrm{E}\left[\psi_{k_L,k_U}\left(\frac{X_i - \lambda}{\sqrt{\lambda}}\right)\right] = 0 \tag{5}$$

in the example of a Poisson distribution. Finding the upper or lower tuning constant that solves above equation (5) is time intensive due to repeated and usually numerical computation of the expectation. Hence, the estimation should be carried out iteratively by first selecting either the lower or upper tuning constant and then iterating between (i) finding the tuning constant that solves (5) given a current estimate of λ and (ii) re-estimating λ given the two tuning constants. In the estimation step, we can select to use the ρ-type, ψ-type or weight-type estimator. Both, the ρ-function and the weight-function are simply defined asymmetrically in the same way as the ψ-function. In the update step (i), it is noted that the solution of (5) is not necessarily existent. To tackle it, it is therefore recommended to fix the upper tuning constant if the distribution is assumed to be right-skewed and the lower tuning constant if it is assumed left-skewed.

2.2 Estimation Procedure

Let us now consider the general case of an independent and identically distributed (iid) random sample $X_1, ..., X_n$ following a distribution parameterised by $\theta \in \Theta \subseteq \mathbb{R}^p$. The estimation approach introduced above is applicable if the parameter space Θ fulfils the following two properties: Θ is a compact subset of $\mathbb{R}^p$, for example there is no integer constraint, and no parameter specifies the support of X_i.

Let for example X_i follow a distribution with a single parameter, $p = 1$, and let $\mu(\theta)$ and $\sigma(\theta)$ be the mean and standard deviation of X_i given parameter θ. Using a loss function $\rho_{k_L,k_U}(z)$, we estimate θ by iterating between estimation and update step. Without loss of generality, assume we specify the upper tuning constant k_U and find k_L in the update step. Denoting the current iteration's estimate as $\hat{\theta}^{(s)}$, we update k_L by solving

$$\mathrm{E}_{\hat{\theta}^{(s)}}\left[\psi_{k_L,k_U}\left(\frac{X_i - \mu\left(\hat{\theta}^{(s)}\right)}{\sigma\left(\hat{\theta}^{(s)}\right)}\right)\right] = 0$$

for k_L. In the estimation step we find $\hat{\theta}^{(s+1)}$ by minimising

$$\sum_{i=1}^{n} \rho_{k_L,k_U} \left(\frac{X_i - \mu(\theta)}{\sigma\left(\hat{\theta}^{(s)}\right)} \right)$$

for θ given the current tuning constants from the update step. Note that the standard deviation is held fixed from the previous iteration for numerical stability in the estimation.

The new parameter estimate can be found in two ways: Either, we can minimise with respect to θ, giving what is referred to as the direct estimator, or we can minimise with respect to $\mu(\theta)$ and then translate the estimated mean to a parameter estimate, which will be referred to as moment based estimator in the following. The minimisation problem can be solved by direct, numerical minimisation, by using the ψ-type estimator or by using the weight-type estimator. The often times faster to compute weight-type estimator is only applicable if we select the moment based approach. However, to be applicable, we need to be able to translate a mean estimate to a parameter estimate and further, we need to ensure that the mean estimate yields a valid parameter estimate. In practice, the problem is found to be negligible.

Considering now the general case of p parameters, the estimation approach is simply extended to higher powers of the sample X_i. Let $\mu_j(\theta)$ and $\sigma_j(\theta)$ be the mean and standard deviation of X_i^j respectively. We can then select potentially different loss functions and tuning constants $\rho^{(j)}_{k_{Lj},k_{Uj}}(z)$. The update step thus consists of solving p equations

$$\mathrm{E}_{\hat{\theta}^{(s)}} \left[\psi^{(j)}_{k_{Lj},k_{Uj}} \left(\frac{X_i^j - \mu_j\left(\hat{\theta}^{(s)}\right)}{\sigma_j\left(\hat{\theta}^{(s)}\right)} \right) \right] = 0$$

for the tuning constants. In the estimation step, we minimise

$$\sum_{j=1}^{p} \sum_{i=1}^{n} \rho^{(j)}_{k_{Lj},k_{Uj}} \left(\frac{X_i - \mu_j\left(\theta\right)}{\sigma_j\left(\hat{\theta}^{(s)}\right)} \right) .$$

Similar as in the single parameter case, we can minimise the loss directly with respect to either the parameter vector θ or with respect to the means $\mu_1, ..., \mu_p$. If we select the moment based approach, we can solve the minimisation problem separately for the powers of X_i, allowing for the application of the ψ-type or weight-type estimator which also facilitates a potentially faster parallel computation.

2.3 Inference

With the estimation procedure established, its asymptotic properties are now discussed. In the following, it is focused on the estimator updating lower tuning constants, derivations for the estimator updating the upper tuning constants are analogous. Let $X_1, ..., X_n$ be an iid sample from a distribution parameterised by $\theta \in \mathbb{R}^p$ and for

notational brevity let

$$\psi\left(X^{(n)},\theta\right)=\left(\psi_1\left(X^{(n)},\theta\right),...,\psi_p\left(X^{(n)},\theta\right)\right)'=0 \tag{6}$$

with

$$\psi_j\left(X^{(n)},\theta\right)=\frac{1}{n}\sum_{i=1}^{n}\psi^{(j)}_{k_{Lj}(\theta),k_{Uj}}\left(\frac{X_i^j-\mu_j(\theta)}{\sigma_j(\theta)}\right) \tag{7}$$

be the estimation function that is solved for θ by the estimator.

Note that above estimation function differs from the one used in each iteration step, because the lower tuning constants k_{Lj} and the standard deviations σ_j both depend on θ. In essence, the iterative estimation solves above equation, keeping terms constants in a step-wise procedure only helps to gain numerical stability. The j-th lower tuning constant in (7), $k_{Lj}(\theta)$, solves

$$\mathrm{E}_\theta\left[\psi^{(j)}_{k_{Lj},k_{Uj}}\left(\frac{X_i^j-\mu_j(\theta)}{\sigma_j(\theta)}\right)\right]=0$$

where the expectation is taken with respect to the distribution of X_i parameterised by θ.

2.3.1 Consistency

The consistency of the estimator can be formulated in the following theorem.

Theorem 4.1:
Let $\hat{\theta}_n$ be the sequence of estimators solving equation (6) for a sample $X_1,...,X_n$, then $\hat{\theta}_n$ is a consistent root of the estimation equation if

(C1) The parameter space Θ is a compact subset of the real numbers and the true parameter vector $\theta_0\in\Theta$ is an interior point.

(C2) $\mu_j(\theta)<\infty$ for $j=1,...,2p$ and all $\theta\in\mathcal{B}(\theta_0,\epsilon)$ and an $\epsilon>0$, where $\mathcal{B}(\theta_0,\epsilon)=\{\theta\in\mathbb{R}^p:\|\theta_0-\theta\|<\epsilon\}$.

(C3) $\psi^{(j)}_{k_{Lj}(\theta),k_{Uj}}(z)$ is bounded in absolute value and continuous in θ and z.

Proof:
To show consistency, the conditions in Theorem 5.9 of van der Vaart (1998) are checked:

(A) $\sup_{\theta\in\Theta}\left\|\psi\left(X^{(n)},\theta\right)-\psi(\theta)\right\|\xrightarrow{p}0$

(B) $\|\psi(\theta_0)\|=0$

(C) $\inf_{\theta \in \Theta \setminus \mathcal{B}(\theta_0, \epsilon)} \|\psi(\theta)\| > 0$ for any $\epsilon > 0$.

Conditions (A) and (B) are sufficient to show that $\hat{\theta}_n$ is a consistent root of the estimation equation. If additionally (C) is met, then $\hat{\theta}_n \xrightarrow{\text{p}} \theta_0$. If (C) is not met, multiple roots of the estimation equation may exist and the estimator minimises the loss function locally and not necessarily globally.

To show (A), the parameter space Θ is first defined as the intersection of valid parameters and those with finite first $2p$ moments. Condition (C2) ensures that θ_0 is still an interior point of Θ. Given a parameter vector θ, we see that

$$\psi^{(j)}_{k_{Lj}(\theta), k_{Uj}} \left(\frac{X_i^j - \mu_j(\theta)}{\sigma_j(\theta)} \right)$$

is an i.i.d. sample for each j. Its expectation and standard deviation are finite due to the ψ-functions being bounded, see assumption (C3). The Law of Large Numbers thus ensures that

$$\begin{aligned} \psi_j \left(X^{(n)}, \theta \right) =& \frac{1}{n} \sum_{i=1}^{n} \psi^{(j)}_{k_{Lj}(\theta), k_{Uj}} \left(\frac{X_i^j - \mu_j(\theta)}{\sigma_j(\theta)} \right) \\ & \xrightarrow{\text{p}} \text{E}_{\theta_0} \left[\psi^{(j)}_{k_{Lj}(\theta), k_{Uj}} \left(\frac{X_i^j - \mu_j(\theta)}{\sigma_j(\theta)} \right) \right] \end{aligned}$$

for every $j = 1, ..., p$. Let $\psi_j(\theta)$ denote the limit and let $\psi(\theta) = (\psi_1(\theta), ..., \psi_p(\theta))'$ be the vector of those limits. Next, we see that all ψ functions are continuous in θ by assumption (C3), such that we can argue along the lines of the comment to Theorem 5.9 of van der Vaart (1998). To show the uniform convergence of condition (A), it suffices to show that the functions $\psi^{(j)}_{k_{Lj}(\theta), k_{Uj}} \left(\frac{X_i^j - \mu_j(\theta)}{\sigma_j(\theta)} \right)$ are dominated by an integrable function, which is guaranteed by the boundedness of ψ functions. The dominated convergence theorem is applicable and condition (A) is met. Note that the boundedness of ψ functions in condition (C3) simplifies the proof of condition (A). Although we focus on bounded ψ functions and thus robust estimators, condition (C3) can be relaxed for non-robust estimators. Instead of boundedness, we can assume that ψ functions are dominated by integrable functions.

Condition (B) can be verified by looking at the elements of the vector $\psi(\theta_0)$

$$\psi_j(\theta_0) = \text{E}_{\theta_0} \left[\psi^{(j)}_{k_{Lj}(\theta_0), k_{Uj}} \left(\frac{X_i^j - \mu_j(\theta_0)}{\sigma_j(\theta_0)} \right) \right]$$

which are zero by the definition of the lower tuning constant and $\hat{\theta}_n$ is therefore a consistent root.

For some cases, it can be easy to verify condition (C) as well, for example if loss functions are convex. For redecscending M-functions, where the ψ function goes to zero for extreme values, other approaches can be used to gain confidence that the estimator minimises the loss function globally. One can check an estimate by computing the weights of observations. If less than 50% of the weights are zero, it indicates that the estimate minimises the loss function globally. Further, one can precede a consistent and robust estimation step and use the estimate as starting values for a second estimation, for example by using Huber's functions ijn the first step.

2.3.2 Asymptotic Normality

With the consistency being discussed, the asymptotic normality is established next.

Theorem 4.2:
If assumptions (C1) - (C3) hold and additionally

(N1) The density $f_X(x)$ is continuously differentiable in θ for every $x \in \mathbb{R}$ and every $\theta \in \mathcal{B}(\theta_0, \epsilon)$ for an $\epsilon > 0$.

(N2) If X_i is discrete, then there exists an $\epsilon > 0$ such that $P_{\theta_0}\left(X_i = d\sigma_j(\theta) + \mu_j(\theta)\right) = 0$ for all $j = 1, ..., p$, all $\theta \in \mathcal{B}(\theta_0, \epsilon)$ and $d \in D_j(\theta, k_{U,j})$ where $D_j(\theta, k_{U,j})$ are points of non-differentiability of $\psi_{k_{Lj}(\theta), k_{Uj}}(z)$ in z.

(N3) $\lim_{n\to\infty} \mathrm{E}_{\theta_0}\left(\frac{\partial}{\partial\theta'}\psi\left(X^{(n)}, \theta\right)\big|_{\theta_0}\right)$ is not singular.

then the sequence of estimators $\hat{\theta}_n$ is asymptotically normal and

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, A(\theta_0)^{-1}B(\theta_0)\left(A(\theta_0)^{-1}\right)'\right)$$

where

$$A(\theta) = \mathrm{E}\left(\frac{\partial}{\partial\theta'}\psi\left(X^{(n)}, \theta\right)\right)$$
$$B(\theta) = n \cdot \mathrm{E}\left(\psi\left(X^{(n)}, \theta\right)\psi\left(X^{(n)}, \theta\right)'\right)$$

Proof of Theorem (4.2):
The theorem is proven by showing that all three conditions of Theorem 4.1.3 of Amemiya (1985) are met given the assumptions above. The conditions are

(A) $\frac{\partial}{\partial \theta'}\psi\left(X^{(n)},\theta\right)$ exists and is continuous in θ in an open, convex neighbourhood of θ_0.

(B) $\frac{\partial}{\partial \theta'}\psi\left(X^{(n)},\theta\right)\big|_{\theta_n^*}$ converges to a finite non-singular matrix $A(\theta_0) = \lim_{n\to\infty} \mathrm{E}_{\theta_0}\left(\frac{\partial}{\partial \theta'}\psi(X^{(n)},\theta)\big|_{\theta_0}\right)$ in probability for any consistent θ_n^*.

(C) $\sqrt{n}\psi\left(X^{(n)},\theta_0\right)$ converges in distribution to a mean zero Normal distribution with covariance matrix

$$B(\theta_0) = \lim_{n\to\infty} n \cdot \mathrm{E}_{\theta_0}\left(\psi\left(X^{(n)},\theta_0\right)\psi\left(X^{(n)},\theta_0\right)'\right)$$

.

Assumption (C3) guarantees that all ψ functions are bounded and continuous. The derivative $\frac{\partial}{\partial \theta'}\psi\left(X^{(n)},\theta\right)$, however, can still have points of discontinuity. There are three possible sources of discontinuity in the derivative. Firstly, the ψ function itself can have points of non-differentiability. That would lead to a violation of condition (A) if we observed an X_i such that $\left(X_i^j - \mu_j(\theta_0)\right)/\sigma_j(\theta_0)$ is one of the discontinuity points. If X_i is continuous, the probability is zero. If X_i is discrete, then assumption (N2) guarantees that the derivative $\frac{\partial}{\partial \theta'}\psi\left(X^{(n)},\theta\right)$ is continuous around θ_0. Secondly, the moment function $\mu_j(\theta)$ can have jumps in θ. With assumption (N1), we are restricting to distributions without discontinuities in the derivative of $\mu_j(\theta)$. Finally, the function $k_{Lj}(\theta)$ is not necessarily continuously differentiable, but both of the above arguments again guarantee continuity in the derivative and thus condition (A) is met.

Condition (B) can be replaced by the following alternative and sufficient condition

$$\sup_{\theta\in\mathcal{B}(\theta_0,\epsilon)}\left\|\frac{\partial}{\partial\theta}\psi\left(X^{(n)},\theta\right) - A(\theta)\right\| \xrightarrow{\mathrm{p}} 0$$

which is common practice when proving normality of extremum Estimators. For a detailed discussion and a proof of sufficiency, see Shi (2010).

All elements of $\psi\left(X^{(n)},\theta\right)$ are i.i.d., continuously differentiable in the neighbourhood of θ_0 and bounded in absolute value. Hence, each derivative is finite on $\mathcal{B}(\theta_0,\epsilon)$ and converges to its expectation, which coincides with the elements of $A(\theta_0)$. Thus, condition (B) is met.

Finally, for condition (C) we see that $\mathrm{E}_{\theta_0}\left(\psi_j\left(X^{(n)},\theta_0\right)\right) = 0$. For every $j = 1, ..., p$, we have i.i.d. samples

$$\psi^{(j)}_{k_{Lj}(\theta_0),k_{Uj}}\left(\frac{X_i^j - \mu_j(\theta_0)}{\sigma_j(\theta_0)}\right)$$

with finite mean due to the boundedness of ψ functions. Application of the multivariate Central Limit Theorem gives

$$\sqrt{n}\psi\left(X^{(n)},\theta_0\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, n\,\mathrm{Cov}_{\theta_0}\left(\psi\left(X^{(n)},\theta_0\right)\right)\right)$$

where

$$\mathrm{Cov}_{\theta_0}\left(\psi\left(X^{(n)},\theta_0\right)\right) = \lim_{n\to\infty} \mathrm{E}\left(\psi\left(X^{(n)},\theta_0\right)\psi\left(X^{(n)},\theta_0\right)'\right)$$

which completes the proof.

After estimating the parameters of a distribution, the asymptotic variance in Theorem 4.2 can be estimated in two ways. Either by deriving the expectations $A(\theta)$ and $B(\theta)$ and plugging in the estimate $\hat{\theta}$, or by estimating the expectations based on observations $x_1, ..., x_n$. Both are common practice when using a sandwich estimator for covariance matrix estimation. For a detailed discussion, see for example Stefanski and Boos (2002).

Above conditions for consistency and asymptotic normality are often met. However, assumption (N2) can be problematic. If a ψ-function is selected which is not continuously differentiable everywhere, a potential violation of (N2) can be circumvented by replacing the ψ-function with a continuously differentiable approximation.

3 Simulation Study

The estimation approaches discussed above are implemented in the Julia Programming Language, see Bezanson et al (2012), and made available in the RobustMFit.jl package, see Stapper (2023). It is now used to assess the estimator's properties through simulation studies.

First, we quantify the bias that occurs when not accounting for the asymmetry of the underlying distribution. For that purpose, simple random samples of length $n = 100$ are generated following a Poisson(λ) distribution with λ ranging from 1 to 50. For each λ, a total of 1000 samples is generated and then fitted using standard M-estimation techniques without bias correction and with the novel approach that updates the lower tuning constant. The loss function is specified as Huber's ρ-function with tuning constant $k = 1.5$ for the former and with $k_U = 1.5$ for the latter. Figure 1 displays the relative bias of the two estimators, i.e. the absolute bias scaled by the true parameter λ. It shows that the uncorrected approach exhibits a bias that is especially pronounced for small λ with larger skewness. The absolute bias of the corrected approach is small compared to the uncorrected estimation as desired.

In addition to the bias comparison for a Poisson distribution, let X_i now follow a standard Normal distribution. Again, we generate 1000 samples of length $n = 100$ and estimate with an uncorrected and a corrected estimator. The M-functions are again selected to be Huber's functions with $k = 1.5$ and $k_U = 1.5$. We use the same loss function for both powers of X_i. Figure 2 shows kernel density estimates of the parameter estimates. We see that the mean parameter is estimated well by both approaches, since the distribution of X_i is symmetric and there is no need for correcting the standard M-estimation approach. The estimation of σ involves the squared sample, which follows a skewed distribution. Thus, the uncorrected estimation exhibits a bias that is not present in the corrected approach.

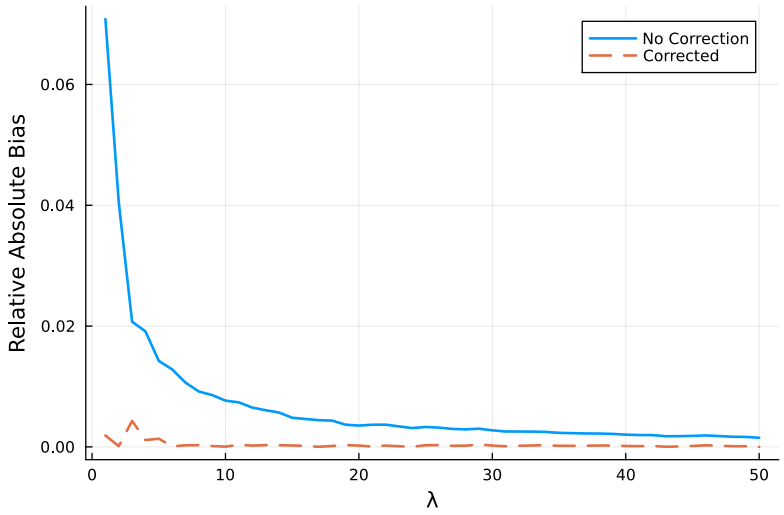

Fig. 1 Relative Absolute Bias of the Uncorrected M-Estimator and the Corrected Estimator for a Poisson Distribution.

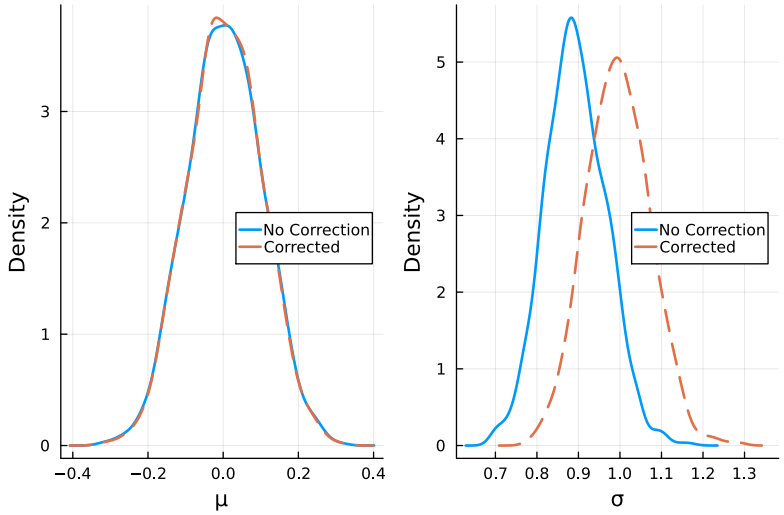

Fig. 2 Kernel Density Estimate of the Uncorrected M-Estimator and the Corrected Estimator for a Normal Distribution.

Let us now focus in the robustness properties of two estimators: the estimator that updates the lower tuning constant and the established estimator that involves a correction term. Following the definition of Huber and Ronchetti (2009), a robust estimator has limited consequences if the model assumptions deviate from reality. Hence, to quantify robustness, we make assumptions regarding the data generating process and a model, then measure the consequences. As model we assume that X_i

follows an i.i.d. Poisson sample. As data generating process we assume that X_i follows a Poisson distribution with mean $\lambda = 5$ but with a probability of 1%, we observe $X_i + \delta$, where δ is an integer. We assess the robustness for different corruption values of δ between 0 and 25. For each δ, a total of 1000 samples is generated, each of length $n = 100$. We select Huber's functions for the estimation and choose the tuning constants such that both approaches yield a 95% relative efficiency compared to Maximum Likelihood estimation, allowing for a fair comparison.

Figure 3 shows the average estimate against the corruption size δ. As intended, both approaches yield limited consequences. The correction term approach exhibits a slightly smaller bias.

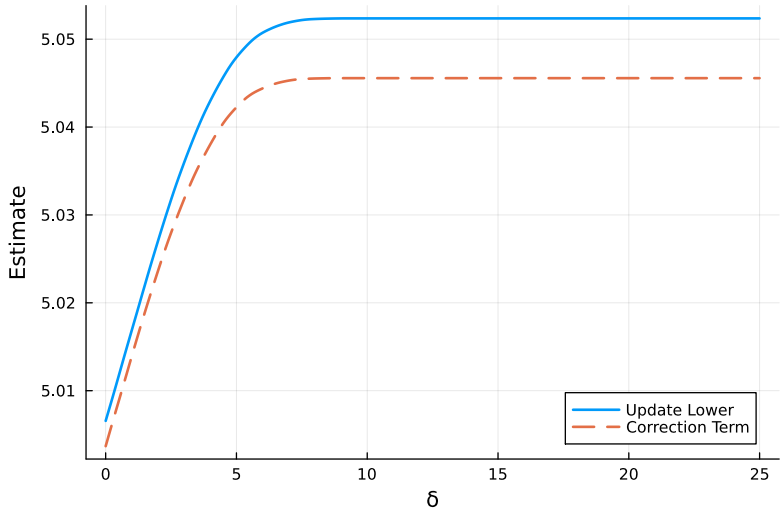

Fig. 3 Estimates of Robust Approaches in the Presence of a Single Additive Outliers.

Next, we assess the breakdown point, which is loosely defined as the maximum percentage of observations that can be corrupted while still having limited consequences on the estimate. We stick to the Poisson distribution and Huber's functions with 95% efficiency. Instead of letting δ vary, we set $\delta = 100$ and corrupt between 0 and 50 of the $n = 100$ observations. Figure 4 shows that the correction term approach exhibits a breakdown point just shy of 50%, white the approach with updated lower tuning constants appears to break down at around 45% corrupted observations.

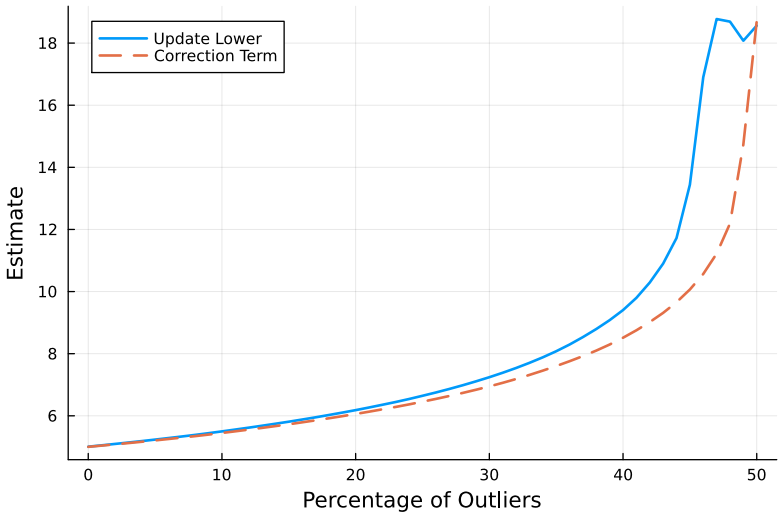

Fig. 4 Estimates of Robust Approaches in the Presence of Multiple Additive Outliers.

Above results suggest that the novel approach has slightly poorer robustness properties while matching efficiency. However, it shall be noted that only the Poisson distribution has been assumed. Considering that the differences between the two approaches are rather small, additional simulation studies would be necessary to inspect whether the differences are systematic.

4 Application

In real life applications, an i.i.d. assumptions if often not justified. The novel approach has been discussed for simple random samples so far, we now adapt it to more complex model frameworks.

4.1 Corporate Insolvencies

Consider a data set first analysed by Weiß and Feld (2019) containing the number of monthly corporate insolvencies in the districts of Rhineland-Palatinate between 2008 and 2016. We focus on one of the 36 areas, the district of Kaiserslautern. Figure 5 displays the number of corporate insolvencies and we see that there is one month in 2014 with an exceptionally large number of insolvencies.

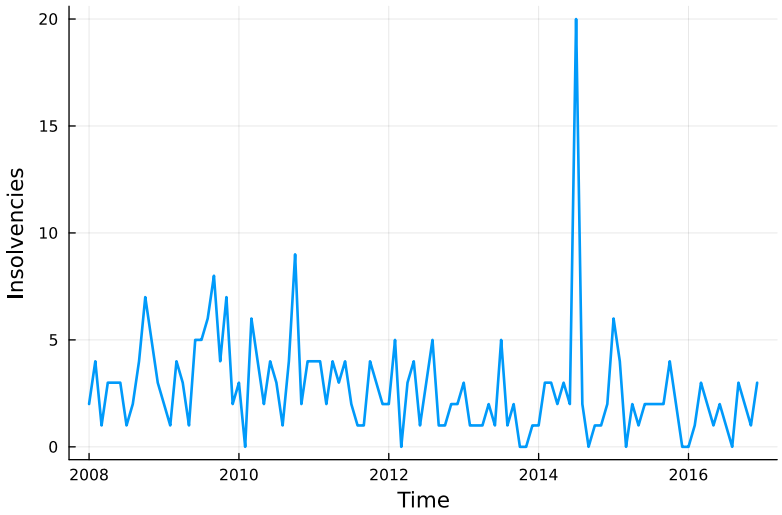

Fig. 5 Number of monthly corporate insolvencies in Kaiserslautern

Instead of incorporating the possibility of extreme observations in the model, we fit a simple log-linear INARCH(1) model

$$\begin{aligned} Y_t|\mathcal{F}_{t-1} &\sim \text{Poisson}(\lambda_t) \\ \log(\lambda_t) &= \beta_0 + \alpha_t \log(Y_{t-1}+1) \end{aligned}$$

where $\mathcal{F}_t$ is the information set available at time t. Three estimators are applied and compared: a Maximum Likelihood (ML) estimator, an M-Estimator without bias correction and the novel M-Estimator based on an asymmetric loss function. For the latter two, we choose Tukey's functions with tuning constants $k = 3$ and $k_U = 3$ respectively. The bias corrected M-Estimator is found by iterating between estimation and update step as in the i.i.d. case. Given current iterations parameter estimates $\hat{\theta}^{(s)}$, a lower tuning constant needs to be found for each of the observations, since the conditional distributions differ. We solve

$$\text{E}\left[\psi_{k_L,k_U}\left(\frac{Y_t-\lambda_t\left(\hat{\theta}^{(s)}\right)}{\sqrt{\lambda_t\left(\hat{\theta}^{(s)}\right)}}\right)\middle|\,\mathcal{F}_{t-1}\right]=0$$

for each $t = 2, ..., T$ where $T = 108$ and $\lambda_t(\theta)$ are conditional means based on parameters θ. That yields lower tuning constants k_{Lt}. New estimates in iteration step $s+1$

Table 1 Estimates and 95% confidence intervals for insolvency counts

	MLE	M-Estimation (corr.)	M-Estimation (uncorr.)
β_0	0.7328	0.6566	0.5310
	[0.4560, 1.0096]	[0.2795, 1.0304]	[0.1003, 0.9506]
α_1	0.2300	0.1916	0.2152
	[0.0251, 0.4349]	[-0.1067, 0.4856]	[-0.1099, 0.5464]

are then found by minimising

$$\sum_{t=2}^{T} \rho_{k_{Lt},k_U} \left(\frac{Y_t - \lambda_t(\theta)}{\sqrt{\lambda_t\left(\hat{\theta}^{(s)}\right)}} \right) .$$

In contrast to the corrected M-Estimator, the uncorrected estimator carries out the estimation step repeatedly using tuning constant $k = 3$. Standard errors for the two robust estimators are found in a bootstrap approach by repeatedly generating data from the fitted model and re-estimating parameters.

Table 1 summarises the estimation results. The parameter α_1 is especially interesting, since it described the time dependence in insolvency numbers. We see that the three estimates differ, however the magnitude is not large compared to the uncertainty in the estimation. Further, the parameter is significant at the 5% level only for ML estimation. This can either be caused by extreme observations in the ML estimation giving a false impression of time dependence or by the larger uncertainty in M-estimation unable to uncover time dependence.

4.2 Doctor Visits - Instrumental Regression

Next, we apply the novel M-Estimation technique in a regression context. Cameron and Trivedi (2013) analysed the number of doctor visits of elderly patients in the US answering whether an additional insurance increases the number of visits. The cross-sectional data stems from the 2003 US Medical Expenditure Panel Survey and contains the following variables: *docvis* gives the number of doctor visits, *private* and *medicaid* are binary variables indicating whether a patient has a private health insurance or is eligible for low-income medicaid coverage respectively. *age* and *age2* contain the (squared) age, *educyr* gives the years of education, *actlim* is binary and has value one if the patient has an activity limitation, *totchr* gives the number of chronic diseases. The number of doctor visits is regressed on seven regressors in a Poisson and Negative Binomial regression. Following Cameron and Trivedi (2013), we take potential endogeneity into account by considering three cases: (i) assuming that all regressors are exogenous (ii) assuming that *private* is endogenous and (iii) *private* and *medicaid* are endogenous. If we assume endogeneity, we use the two instruments *income* and *ssiratio*. The former gives the patient's household income and the latter describes the ratio of social security income to total income.

In line with Cameron and Trivedi (2013), we remove all patients from the data with private and medicaid insurance. Further, they exclude three outlying observations where the number of visits exceeds 70. Here, we consider two additional cases: Not excluding extreme observations at all and applying the novel M-Estimator. When accounting for endogeneity, we follow the control function approach described in section 10.4.2 of Cameron and Trivedi (2013). In a two step procedure, the endogenous regressors are first regressed on all exogenous regressors and the two instruments by OLS and the residuals of the first stage regression are then included as regressor in the second stage count regression. For the second stage we apply both, a Poisson and a Negative Binomial regression.

The M-estimator of the second stage in case of a Poisson distribution carries out the two steps iteratively. Let Y_i be the number of doctor visits and x_i all regressors considered in the second stage. Starting with initial values for parameters we then update the lower tuning constants by solving

$$\mathrm{E}\left(\psi_{k_{Li},k_U}\left(\frac{Y_i-\mu^{(i)}\left(\hat{\theta}^{(s)}\right)}{\sigma^{(i)}\left(\hat{\theta}^{(s)}\right)}\right)\right)=0$$

for k_{Li} for each observation Y_i where $\mu^{(i)}\left(\hat{\theta}^{(s)}\right)$ is the expectation of Y_i assuming the current parameters $\hat{\theta}^{(s)}$. Likewise, $\sigma^{(i)}\left(\hat{\theta}^{(s)}\right)$ denotes its standard deviation. Then, new parameter estimates are found by minimising

$$\sum_{i=1}^{n}\rho_{k_{Li},k_U}\left(\frac{Y_i-\mu^{(i)}\left(\theta\right)}{\sigma^{(i)}\left(\hat{\theta}^{(s)}\right)}\right)\ .$$

Above steps only hold true for the Poisson case. If we want to fit a Negative Binomial distribution in the second stage, we introduce tuning constants for the squared observations and find the lower tuning constants

$$\mathrm{E}\left(\psi_{k_{L2i},k_{U2}}\left(\frac{Y_i^2-\mu_2^{(i)}\left(\hat{\theta}^{(s)}\right)}{\sigma_2^{(i)}\left(\hat{\theta}^{(s)}\right)}\right)\right)=0$$

where μ_2 and σ_2 are mean and standard deviation of Y_i^2 respectively. The estimation step then also includes squared observations

$$\sum_{i=1}^{n}\rho_{k_{Li},k_U}\left(\frac{Y_i-\mu^{(i)}\left(\theta\right)}{\sigma^{(i)}\left(\hat{\theta}^{(s)}\right)}\right)+\rho_{k_{L2i},k_{U2}}\left(\frac{Y_i^2-\mu_2^{(i)}\left(\theta\right)}{\sigma_2^{(i)}\left(\hat{\theta}^{(s)}\right)}\right)\ .$$

Table 2 summarises the estimation results. Estimates for all remaining regressors are left out for clarity. Standard errors are computed for all approaches by bootstrap

to take the first stage uncertainty into account. For a fair comparison, the bootstrap samples are drawn from the original sample including patients with more than 70 doctor visits and each drawn observation with more than 70 visits is then removed from the bootstrap sample. The M-Estimator is computed using Huber's functions with fixed upper tuning constants $k_U = k_{U2} = 1.5$.

For the model that does not account for heterogeneity, we see that the standard errors for all parameters except for the overdispersion parameter are much lower compared to models with endogeneity accounted for. Further, we see that estimates from M-Estimation are closer to the estimates based on all observations than they are to the estimates of the restricted sample. When accounting for endogeneity, we see that the relevant parameter estimates change substantially. Estimates of the overdispersion parameter suggest that it is reasonable to assume a Negative Binomial distribution rather than a Poisson. The change in parameter estimates going from one endogenous variable to also treating *medicaid* as endogenous suggest that should assume *medicaid* as endogneous. However, the corresponding residual term is not significant at the 5% level for all three estimation methods. Focusing on the model with two endogenous regressors, a Negative Binomial distribution and looking at the two robust methods, we observe that private insurance increases the number of doctor visits by approximately 108% on average. The expected increase is approximately 119% and 144% for the outlier removal approach and the M-Estimator respectively. Due to the large standard errors in the model and the restrictive selection of socio-economic regressors in the model, the results should be interpreted with care.

4.3 Multivariate Normal - Breast Cancer

In a third and final application, we extend the univariate estimation to a multivariate setting by analysing data on breast cancer, see Wolberg and Street (1995). The data contains 569 observations and 30 standardised features derived from image data of cell samples. Further, it contains a binary variable specifying if an observation stems from a malignant or a benign sample. We aim to classify a new observation into one of the two categories. Since this application serves as an illustrative example, we focus on two features describing the perimeter of cell nuclei shown in Figure 6. The features are assumed to follow a bivariate Normal distribution in both groups, i.e. the random vector $(X_i, Y_i)'$ of perimeter measurements is assumed to follow a $\mathcal{N}(\mu_1, \Sigma_1)$ if the sample is from a benign cell and a $\mathcal{N}(\mu_2, \Sigma_2)$ distribution if it is from a malignant cell. In both dimensions we observe outlying observations, that jeopardise the applicability of non-robust methods.

A new observation is classified by comparing weighted densities of the two fitted distributions where the weights are proportional to the number of observations in the corresponding group. We compare the performance of a non-robust estimator, where the Normal distributions are fitted by Maximum Likelihood, with the novel robust approach in a leave-one-out cross-validation.

Table 2 Instrument Regression Results

Variable		No Endog. Full	No Endog. < 70	No Endog. Mfit	One Endog. Full	One Endog. < 70	One Endog. Mfit	Two Endog. Full	Two Endog. < 70	Two Endog. Mfit
					Poisson					
Private	Coef.	0.157	0.170	0.166	0.591	0.410	0.330	0.751	0.530	0.437
		(0.037)	(0.036)	(0.037)	(0.260)	(0.233)	(0.268)	(0.305)	(0.263)	(0.293)
	Res.				-0.443	-0.246	-0.168	-0.604	-0.366	-0.274
					(0.265)	(0.234)	(0.271)	(0.309)	(0.263)	(0.296)
Medicaid	Coef.	0.128	0.082	0.137	0.339	0.199	0.217	0.961	0.673	0.613
		(0.062)	(0.048)	(0.051)	(0.150)	(0.123)	(0.140)	(0.513)	(0.426)	(0.481)
	Res.							-0.842	-0.598	-0.482
								(0.499)	(0.430)	(0.480)
					Negative Binomial					
Private	Coef.	0.176	0.185	0.164	0.712	0.598	0.564	0.893	0.732	0.733
		(0.038)	(0.037)	(0.038)	(0.265)	(0.250)	(0.238)	(0.310)	(0.280)	(0.298)
	Res.				-0.547	-0.422	-0.407	-0.729	-0.556	-0.577
					(0.268)	(0.252)	(0.237)	(0.312)	(0.280)	(0.296)
Medicaid	Coef.	0.127	0.085	0.104	0.386	0.285	0.298	1.074	0.786	0.890
		(0.061)	(0.048)	(0.052)	(0.149)	(0.131)	(0.125)	(0.538)	(0.459)	(0.530)
	Res.							-0.960	-0.709	-0.797
								(0.525)	(0.462)	(0.538)
Overdisp.		1.575	1.622	1.554	1.578	1.623	1.558	1.580	1.624	1.562
		(0.061)	(0.056)	(0.079)	(0.061)	(0.056)	(0.079)	(0.061)	(0.056)	(0.078)

Standard errors for M-Estimation of Neg.Bin. model are computed using 200 bootstrap samples

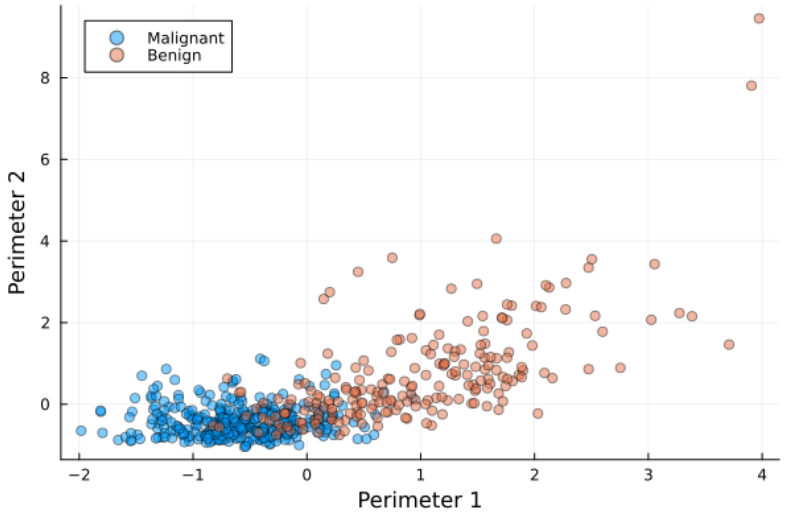

Fig. 6 Perimeter Measurements

To fit the multivariate Normal distribution robustly, we estimate the parameters of the marginals X_i and Y_i separately before estimating the covariance. For that purpose we fit the distribution of $X_i + Y_i$ and identify the covariance by solving $\widehat{\text{Var}}(X_i + Y_i) = \widehat{\text{Var}}(X_i) + \widehat{\text{Var}}(Y_i) + 2\widehat{\text{Cov}}(X_i, Y_i)$.

First we fit the two Normal distributions using the full sample by Maximum Likelihood and by the Robust approach. In the robust case, we select Huber's functions and tuning constants 1.3 for the first moments and 2.5 for second moments of the corresponding distribution. Those tuning constants yield approximately 95% relative efficiency. Figure 7 shows observations together with 95% confidence ellipses split by group and estimation method. We see that the estimates for benign cells are barely affected by outlying observations. For malignant cells, the robust estimation approach results in an ellipse that is more narrow compared to Maximum Likelihood estimation, which highlights that the robust approach downweights extreme observations. In addition to estimates, we compute standard errors in a bootstrap approach by sampling observations in both groups separately and re-estimating parameters. Table 3 summarises estimates and standard errors. We see that standard errors of the robust method in the group of benign cells tend to be larger than standard errors of maximum likelihood estimation. However, for the malignant cells group, we see that the Maximum Likelihood estimator not only estimates the variances to be larger but also the ML estimates exhibit larger standard errors.

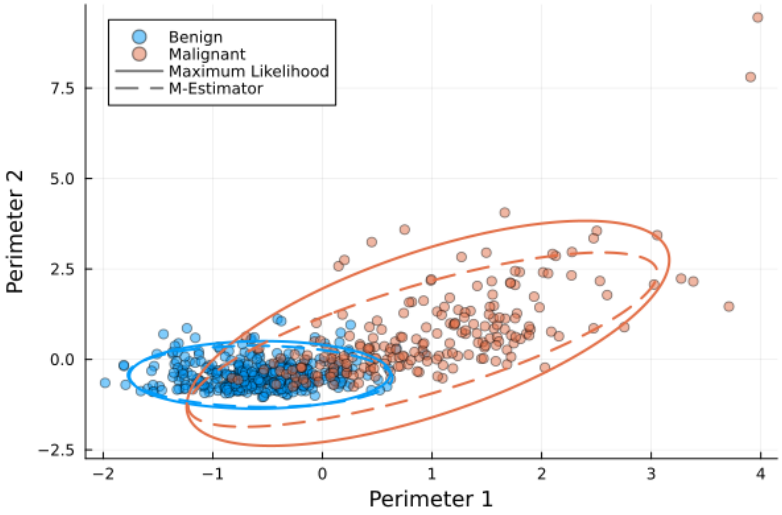

Fig. 7 Perimeter Measurements and Fitted Distributions

Let us now assess whether the differences in estimates between the two methods affect the classification properties. We successively remove a single observation from the data, fit two bivariate Normal distribution and then compare weighted densities of both fitted distributions at the removed observation. Then, the observation is classified as the group with the larger density. Results of the classification are summarised

	Maximum Likelihood			M-Estimation		
	μ_1	Σ_1		μ_2	Σ_2	
Benign	-0.572	0.236	0.006	-0.568	0.243	-0.008
	(0.027)	(0.018)	(0.011)	(0.028)	(0.019)	(0.011)
	-0.428		0.145	-0.470		0.120
	(0.020)		(0.015)	(0.019)		(0.013)
Malignant	0.963	0.809	0.744	0.916	0.765	0.656
	(0.063)	(0.090)	(0.173)	(0.061)	(0.084)	(0.100)
	0.721		1.614	0.546		0.969
	(0.090)		(0.448)	(0.077)		(0.156)

Table 3 Estimated parameters for cell perimeters and standard errors based on 1000 bootstrap repetitions

in Table 4. Using Maximum Likelihood estimation we achieve 499 correctly classified observations whereas the robust estimation yields 511 correct classifications. This improvement in classification accuracy might seem small at first and one can easily assess the outlyingness of an observation. In practice, however, an outlying observation is not easily detected when more than two features are considered. Beyond the scope of this article, it would be interesting to explore whether the robust fitting of multivariate Normal distribution can be generalised to higher dimensions.

Maximum Likelihood					M-Estimation				
		Classification					Classification		
		B	M	Σ			B	M	Σ
True	B	341	16	357	True	B	344	13	357
	M	54	158	212		M	45	167	212
	Σ	395	174	569		Σ	389	180	569

Table 4 Results of leave one out classification

5 Conclusion

This article introduces a new approach to apply M-estimation for asymmetric distributions. Its consistency and asymptotic normality are proven and its robustness properties assessed in a simulation study. The new approach serves as an alternative to an existing bias correction approach. The robustness of both corrected estimators are comparable in the study, with the novel approach showing slightly poorer robustness performance at equal relative efficiencies. In contrast to the correction term approach, the new approach facilitates estimation with ρ-, ψ- and weight-functions for a broad class of distributions. Further, results can be assessed more easily using the asymmetric weight function. Three applications underline the versatility of the new approach for more complex models.

There are limitations for the new approach and its asymptotic properties. Finite moments up to order $2p$ need to be guaranteed, which is also the case for the correction term approach. Using redescending ψ-functions further poses a thread to consistency, which can be addressed with common techniques in M-estimation. Asymptotic normality for discrete distributions is not necessarily given if the ψ-function is not continuously differentiable. Replacing the ψ-function with a differentiable approximation is thus recommended. To ensure that a lower (upper) tuning constant is found while fixing the respective other tuning constant, it is advised that the upper tuning constant is fixed for right-skewed distributions whereas the lower tuning constant should be fixed for left-skewed distributions.

Future research could build on the findings of this article and explore moment estimation without distributional assumptions as well as multivariate distributions. Another interesting research direction is relaxing the need to select one of the tuning constants but instead select a desired relative efficiency. Then, in the updating step, both tuning constants would be updated such that the estimator is unbiased and yields the targeted efficiency. This approach would raise interesting theoretical and computational challenges.

To conclude, the discussed new approach demonstrates its practical use in various application while achieving robustness properties comparable to existing approaches.

References

Amemiya T (1985) Advanced Econometrics. Harvard University Press

Bezanson J, S. K, Shah V, et al (2012) Julia: A fast dynamic language for technical computing. https://doi.org/10.48550/arXiv.1209.5145

Cameron AC, Trivedi PK (2013) Regression Analysis of Count Data, 2nd edn. Econometric Society Monographs, Cambridge University Press

Cantoni E, Ronchetti E (2001) Robust inference for generalized linear models. Journal of the American Statistical Association 96(455):1022–1030. https://doi.org/10.1198/016214501753209004

Huber P (1964) Robust estimation of a location parameter. The Annals of Methematical Statistics 35(1):73–101. https://doi.org/10.1214/aoms/1177703732

Huber PJ, Ronchetti EM (2009) Robust Statistics, 2nd edn. Wiley

Shi X (2010) Asymptotic normality of extremum estimators. Online Ressource, URL https://users.ssc.wisc.edu/~xshi/econ715/Lecture_4_normality.pdf, access: August, 21. 2023

Stapper M (2023) Robustmfit.jl - methods to estimate parameters of a distribution robustly. https://doi.org/10.5281/zenodo.7779288, URL https://github.com/ManuelStapper/RobustMFit.jl, version 0.1.2

Stefanski L, Boos D (2002) The calculus of m-estimation. The American Statistician 56(1):29–38. https://doi.org/10.1198/000313002753631330

Tukey J (1960) A Survey of Sampling from Contaminated Distributions. Stanford University Press, URL https://catalog.princeton.edu/catalog/9944622593506421

van der Vaart A (1998) Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, https://doi.org/10.1017/CBO9780511802256

Weiß CH, Feld M (2019) On the performance of information criteria for model identification of count time series. Studies in Nonlinear Dynamics & Econometrics 24(1). https://doi.org/10.1515/snde-2018-0012, URL https://doi.org/10.1515/snde-2018-0012

Wolberg MOSNWilliam, Street W (1995) Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C5DW2B