RESEARCH ARTICLE

A Hybrid Method for Density Power Divergence Minimization with Application to Robust Univariate Location and Scale Estimation

Andrews T. Anum[a] and Michael Pokojovy[b]

[a]ORCID:0000-0001-7287-280X: Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, Texas 79968, USA; [b]ORCID:0000-0002-2122-2572: Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, Texas 79968, USA

ARTICLE HISTORY
Compiled February 20, 2023

ABSTRACT
We develop a new globally convergent optimization method for solving a constrained minimization problem underlying the minimum density power divergence estimator for univariate Gaussian data in the presence of outliers. Our hybrid procedure combines classical Newton's method with a gradient descent iteration equipped with a step control mechanism based on Armijo's rule to ensure global convergence. Extensive simulations comparing the resulting estimation procedure with the more prominent robust competitor, Minimum Covariance Determinant (MCD) estimator, across a wide range of breakdown point values suggest improved efficiency of our method. Application to estimation and inference for a real-world dataset is also given.

KEYWORDS
Minimum density power divergence estimator; Rousseeuw's Minimum Covariance Determinant estimator (MCD); gradient descent; Armijo rule; Newton's method

1. Introduction

Robustness of an estimator is a multifaceted statistical concept that does not have universal definition. Martin (1979) defined an estimator to be robust if its performance remains reasonably good when the actual distribution of the data deviates from the assumed one. We will refer to an estimator as robust if it is not too sensitive to outliers which means that the estimator draws a valid conclusion about underlying population parameters even in the presence of outliers in the dataset. Due to this attractive property, robust estimators have consequently been used successfully in numerous applications. In fact, it is possible to tell if an estimator is heavily affected by deviations based on the influence function (Hampel, 1974), which basically quantifies the impact of an outlier on the estimator. Another measure developed by Hampel (1971) (cf. also Donoho and Huber (1983)) is the breakdown point. Martin (1979) defined the breakdown point of an estimator as the largest fraction of contaminated data, over all combinations for each fraction, for which the bias remains bounded. See also Fox et al. (2002).

CONTACT Michael Pokojovy. Email: mpokojovy@utep.edu

For non-robust estimators, such as sample mean $\bar{x}$, even a single observation approaching infinity may cause the estimator to break down. A much more robust location estimator is the sample median. An advantage of the sample median over the sample mean is that it is not influenced by outlying observations. Just like the sample mean, the sample standard deviation, $\sqrt{s^2}$, which is supposed to measure how widely spread a dataset is, faces the challenges posed by extreme values in the dataset. The effect of outliers motivated some statisticians to completely eliminate them from the data because of the risk of severely biased conclusions drawn from the analysis of such datasets. This approach became very popular and many practitioners handled outliers as such. It is acceptable to drop an outlier from the data if it is obvious that it was entered wrongly. However, it is oftentimes unclear whether it is a true outlier or a data instance entered incorrectly. Therefore, removing an observation from the dataset typically results in information loss. Other data analysts perform "data curation" in the event that the supposed outlier is suspected to be an incorrect input. Values used in lieu of this outlier are usually imputed, e.g., using the sample mean or more sophisticated imputation approaches (van Buuren, 2012).

A wide variety of robust estimators have been developed over the past decades. Typically, these estimators can be classified as M-, L-, R- or S-estimators, etc. (Huber, 2009). Vast portion of robust statistics literature specifically focuses on location and/or scale estimation. The more prominent methodologies include Stahel-Donoho estimator (Donoho, 1982; Stahel, 1981), Minimum Covariance Determinant (MCD) estimator of Rousseeuw (1984) (see also Rousseeuw and van Driessen (1999); Hubert et al. (2012) for FastMCD and DetMCD computational algorithms), Minimum Volume Ellipsoid (MVE) estimator (Rousseeuw, 1985), Constrained M-estimators (Kent and Tyler, 1996), (generalized) S-estimators Croux and Hössjer (1994), (generalized) τ-estimator Croux and Hössjer (1999), cluster-based estimators (Jobe and Pokojovy, 2015) and many others. See Maronna and Yohai (2006) for a more systematic review. Many (if not most) algorithmic implementations of multivariate robust estimators are based on "elemental concentration," als known as "C-step," which, in the presence of outliers, can lead to inconsistency, even in moderate space dimensions (Hawkins and Olive, 2002). Alternative procedures, which are computationally feasible, yet affine-equivariant, were proposed by Peña and Prieto (2001a); Reyen et al. (2009). In higher dimensions, the robustness is typically only maintained in large samples.

Other authors developed robust estimators for general continuous parametric models by paralleling the well-known maximum likelihood approach through adoption of minimum distance (or, more generally, divergence) estimation. One of such approaches was first proposed by Beran (1977). The author used Hellinger distance as a robust approximation to Kullback-Leibler divergence associated with the ML-approach to put forth his robust estimators. Under appropriate regularity conditions, the proposed estimators were shown to have full asymptotic efficiency at the model. The methodology relies upon bandwidth selection and, thus, may be subject to some adverse effects in continuous models in the sense that the estimators require utilization of nonparametric smoothing procedures, which are known to severely affect kernel-based procedures. As a matter of fact, kernel density estimation is a nonparametric smoothing technique that is relied upon to yield an estimate of the population density. A crucial amendment to the original procedure of Beran (1977) was made by Basu and Lindsay (1994). To alleviate the reliance on bandwidth selection, they smoothed both the model and the data with the same kernel.

A class of minimum density power divergence estimators was introduced by Basu et al. (1998) for robust estimation in general parametric models. They defined the divergence functional $d_\alpha(f, g)$ between two p-variate probability density functions $g(\boldsymbol{x})$ and $f(\boldsymbol{x})$ as

$$d_\alpha(f, g) := \int_{\mathbb{R}^p} \left(f^{1+\alpha}(\boldsymbol{x}) - \left(1 + \frac{1}{\alpha}\right) g(\boldsymbol{x}) f^\alpha(\boldsymbol{x}) + \frac{1}{\alpha} g^{1+\alpha}(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x} \tag{1}$$

for $\alpha > 0$. On the strength of (Basu et al., 1998, Theorem 1), $d_\alpha(\cdot, \cdot)$ is a divergence in the sense that it is a premetric satisfying the identity of indiscernibles. As $\alpha \to 0$, the divergence function reduces to the usual Kullback-Leibler divergence function

$$d_0(f, g) := \lim_{\alpha \to 0} d_\alpha(f, g) = \int_{\mathbb{R}^p} g(\boldsymbol{z}) \left(\log \frac{g(\boldsymbol{z})}{f(\boldsymbol{z})} \right) \mathrm{d}\boldsymbol{z}. \tag{2}$$

Assuming $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{\text{i.i.d.}}{\sim} g(\boldsymbol{x})$ (with $g(\boldsymbol{x})$ being unknown) and invoking the law of large numbers, we get

$$\int_{\mathbb{R}^p} f^\alpha(\boldsymbol{x}) g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \approx \frac{1}{n} \sum_{i=1}^n f^\alpha(\boldsymbol{x}_i).$$

Further, observing that the latter term in Equation (1) does no depend on $f(\boldsymbol{x})$, one can easily conclude that minimizing $d_\alpha\big(f(\cdot|\boldsymbol{\theta}), g(\cdot)\big)$ over $\boldsymbol{\theta} \in \Theta$ is asymptotically equivalent with minimizing the *density power divergence* function

$$H_n(\boldsymbol{\theta}) = \int_{\mathbb{R}^p} f^{1+\alpha}(\boldsymbol{x}|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} - \left(1 + \frac{1}{\alpha}\right) n^{-1} \sum_{i=1}^n f^\alpha(\boldsymbol{x}_i|\boldsymbol{\theta}). \tag{3}$$

Thus, the *minimum density power divergence* (MDPD) estimator is given as

$$\hat{\boldsymbol{\theta}}_n^{\text{DPD}} := \arg\min_{\boldsymbol{\theta} \in \Theta} H_n(\boldsymbol{\theta}). \tag{4}$$

Note that minimizing $H_n(\cdot)$ does not require knowledge about actual distribution of $\boldsymbol{x}_i$'s or involve kernel density estimation and, thus, in contrast to Basu and Lindsay (1994), does not rely on bandwidth selection. Under appropriate conditions, Basu et al. (1998) showed that the MDPD estimator has a number of attractive statistical properties such as asymptotic normality, affine equivariance (in case of location and scatter estimation), high breakdown point, high efficiency and so on (cf. Section 2). Due to these and other attractive properties, the MDPD estimator is very useful for robust inference (Basu et al., 2013, 2017, 2018; Ghosh et al., 2016).

Computing the MDPD estimator $\hat{\boldsymbol{\theta}}_n^{\text{DPD}}$ amounts to solving the optimization problem in Equation (4) or, under very liberal smoothness conditions, solving the associated estimating equations (viz. (Basu and Lindsay, 1994, Equation (2.3))). Either approach requires numerical optimization or nonlinear equation solving. In maximum likelihood estimation, these problems are typically tackled with Newton–Raphson method or Fisher scoring algorithm (Avriel, 2003; Small and Wang, 2003). Further approaches include various quasi-Newton methods, sequential linear-quadratic programming (SLQP), etc. (Hinze et al., 2008; Boyd and Vandenberghe, 2004). All of these methods exhibit local superlinear (typically, quadratic) convergence but heavily rely on good “warmstarts” or initial guesses to converge. In fact, (undamped) (quasi-)Newton-like methods are known to induce chaotic dynamics or may “escape” the parameter set Θ if the initial guess is selected too far from a local minimum (see Supplemental Section S1.4). While a good warmstart may be relatively easy to come up with in non-robust estimation, the problem becomes challenging, if not unfeasible, in the presence of multivariate outliers. Due to the very nature of “masking” through outliers, the chance to randomly come up with a good “warmstart” (especially when simultaneously estimating multivariate scatter parameters) becomes virtually zero. Cf. Hawkins and Olive (2002); Peña and Prieto (2001b). Therefore, it is crucial to adopt optimization methods that never fail to converge to (at least) a local optimum. One of the mechanisms behind the compromised convergence of (quasi-)Newton-type methods is the lack of a step control mechanism. A wide variety of numerical optimization methods with step control are known in the literature. The more prominent ones include multiple variants of gradient descent schemes, accelerated gradient methods, Frank-Wolfe-type iterations, etc. (Boyd and Vandenberghe, 2004; Varadhan and Roland, 2008). In various scenarios, square-root up to quadratic convergence can be achieved. Step control mechanisms for (quasi-)Newton methods have also been investigated (Grippo et al., 1986; Potschka, 2014). These and other locally-adaptive Newton-type methods can fail to converge when the objective becomes locally non-convex as it is typically the case with $H_n(\cdot)$ in Equation (3).

We propose a new hybrid method that implements a switching mechanism between a gradient descent method with monotone step control based on Armijo’s rule and undamped Newton’s method. Starting in the gradient descent mode, the algorithm switches to full-step Newton’s method in a vicinity of a “convex” minimum. Additionally, a convex projector is implemented to incorporate constraints on $\boldsymbol{\theta}$. To illustrate the advantage of our hybrid method over the plain gradient descent with Armijo’s rule, we apply it to DPD minimization in connection with robust estimation of univariate Gaussian location and scale parameters.

Two major considerations motivated the development of our new hybrid optimization method. Firstly, there is a wide variety of optimization methods in the literature that do not always converge. Unlike our new method, the convergence of these methods depends on how close the initial value is chosen to the true solution. In most practical scenarios though, we do not know the true solution to guide our choice of the initial value. Hence, an optimization procedure, like ours, that always converges at least to a local minimum, regardless of how close the initial value is to the true solution, is necessary. Secondly, practitioners prefer optimization techniques that exhibit a rapid convergence speed. Therefore, we equipped our proposed method with a step control mechanism that takes the largest step size at each iteration, rather than using a constant step size. To further improve the convergence speed and render it locally quadratic, our proposed method transitions from the step-controlled gradient descent to the Newton’s method once the objective becomes locally convex, which typically happens in a vicinity of a local minimum. This switch mechanism is implemented as to adapt to local convexity of the objective in the interior of the feasible set and ensure a generalized form of Armijo’s condition remains true. Intuitively, at every iteration, our optimization method checks if the eigenvalues of the objective’s Hessian matrix are all positive and satisfy appropriate smallness/largeness constraints. In this case, a full Newton step is made provided the iteration does not escape the feasible set. If any of the previous checks fail, the usual projected gradient descent with Armijo’s rule is performed instead. This typically happens either when the objective becomes locally non-convex or if the minimum is attained at the boundary of the feasible set. Being both theoretically justified (viz. Section 4) and extensively tested empirically (viz. Section 5), our new hybrid method (viz. Algorithm 4.1) offers an excellent alternative to conventional optimization techniques used in statistics.

The rest of the paper is structured as follows. Section 2 gives a summary of theoretical properties of the MDPD estimator. Section 3 provides an overview of robust statistical inference about the location and scale parameters via robustified test statistics, including rejection regions, p-values and robust confidence intervals. In Section 4, our proposed optimization method is presented and protocoled along with appropriate theoretical results. An empirical run-time analysis of our proposed method using simulated data, with and without contamination, is provided in Section 5. Also, the empirical convergence rate of our proposed method as well as the empirical breakdown point of the MDPD estimator are analyzed and reported. Additionally, three families of contaminated models are considered to benchmark the performance of our proposed method against several state-of-the-art competitors. Section 6 presents an application example of our proposed otpimization method in the context of analyzing and forecasting monthly chlamydia case numbers in El Paso, Texas, USA.

2. Theoretical Properties of the MDPD Estimator

Suppose the data $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$ are independently sampled from a common population with a cumulative distribution function $G(\boldsymbol{x})$ and a propability density function $g(\boldsymbol{x})$, not necessarily belonging to the family $\{f(\cdot|\tilde{\boldsymbol{\theta}})\}_{\tilde{\boldsymbol{\theta}}\in\Theta}$. Further, suppose

$$\boldsymbol{\theta} = \arg\min_{\tilde{\boldsymbol{\theta}}\in\Theta} d_\alpha\big(g, f(\cdot|\tilde{\boldsymbol{\theta}})\big) \text{ for some } \alpha \geq 0$$

exists uniquely. In case $g(\cdot)$ belongs to the family $\{f(\cdot|\tilde{\boldsymbol{\theta}})\}_{\tilde{\boldsymbol{\theta}}\in\Theta}$, one trivially has $g(\cdot) = f(\cdot|\boldsymbol{\theta})$. Introducing the score function $\boldsymbol{u}(\boldsymbol{x}|\tilde{\boldsymbol{\theta}}) := \partial \log f(\boldsymbol{x}|\tilde{\boldsymbol{\theta}})/\partial\tilde{\boldsymbol{\theta}}$, Basu and Lindsay (1994) derived the estimating equations

$$\boldsymbol{U}_n(\tilde{\boldsymbol{\theta}}) \equiv \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{u}(\boldsymbol{x}_i|\tilde{\boldsymbol{\theta}}) f^\alpha(\boldsymbol{x}|\boldsymbol{\theta}) - \int_{\mathbb{R}^p} \boldsymbol{u}(\boldsymbol{x}|\tilde{\boldsymbol{\theta}}) f^{1+\alpha}(\boldsymbol{x}|\tilde{\boldsymbol{\theta}})\mathrm{d}\boldsymbol{x} = \boldsymbol{0} \tag{5}$$

for the MDPD estimator. Interpreting $\hat{\boldsymbol{\theta}}^n_{\text{MDPD}}$ as an M-estimator, the following asymptotic consistency and normality result was further established.

Theorem 2.1 ((Basu and Lindsay, 1994, Theorem 2))**.** *Under appropriate conditions, the estimator $\hat{\boldsymbol{\theta}}^n_{\text{MDPD}}$ is well-defined. Further, as $n \to \infty$,*

(i) $\hat{\boldsymbol{\theta}}^n_{\text{MDPD}}$ is a consistent estimator of $\boldsymbol{\theta}$,

(ii) $n^{1/2}\big(\hat{\boldsymbol{\theta}}^n_{\text{MDPD}} - \boldsymbol{\theta}\big) \overset{\mathcal{D}}{\to} \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J})$, where

$$\begin{aligned}\boldsymbol{J} &= \int_{\mathbb{R}^p} \big(\boldsymbol{u}(\boldsymbol{x}|\boldsymbol{\theta}) \otimes \boldsymbol{u}(\boldsymbol{x}|\boldsymbol{\theta})\big) f^{1+\alpha}(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{x} \\ &\quad + \int_{\mathbb{R}^p} \big(\boldsymbol{i}(\boldsymbol{x}|\boldsymbol{\theta}) - \alpha\boldsymbol{u}(\boldsymbol{x}|\boldsymbol{\theta}) \otimes \boldsymbol{u}(\boldsymbol{x}|\boldsymbol{\theta})\big)\big(g(\boldsymbol{x}) - f(\boldsymbol{x}|\boldsymbol{\theta})\big) f^\alpha(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{x}, \\ \boldsymbol{K} &= \int_{\mathbb{R}^p} \big(\boldsymbol{u}(\boldsymbol{x}|\boldsymbol{\theta}) \otimes \boldsymbol{u}(\boldsymbol{x}|\boldsymbol{\theta})\big) f^{2\alpha}(\boldsymbol{x}|\boldsymbol{\theta}) g(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \boldsymbol{\xi} \otimes \boldsymbol{\xi}\end{aligned}$$

with $\boldsymbol{\xi} = \int_{\mathbb{R}^p} \boldsymbol{u}(\boldsymbol{x}|\boldsymbol{\theta}) f^{\alpha}(\boldsymbol{x}|\boldsymbol{\theta}) g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ *and the information function*

$$\boldsymbol{i}(\boldsymbol{x}|\tilde{\boldsymbol{\theta}}) := -\partial \boldsymbol{u}(\boldsymbol{x}|\tilde{\boldsymbol{\theta}})/\partial\tilde{\boldsymbol{\theta}}.$$

Basu and Lindsay (1994) also derived the influence function of the MDPD estimator and discussed sufficient conditions for it to be bounded. It was further shown that the estimator is equivariant with respect to reparametrization. For $\alpha > 0$, the estimator is also invariant under non-singular affine-linear transformations of the data.

Since we chose the univariate Gaussian model to illustrate our proposed hybrid optimization method, the rest of this section presents a discussion of MDPD estimation of univariate Gaussian location and scale parameters. Consider the univariate Gaussian location-scale family $\big(f(x|\tilde{\mu},\tilde{\sigma})\big)_{(\tilde{\mu},\tilde{\sigma})\in\Theta}$ consisting of univariate probability densities

$$f(x|\tilde{\mu},\tilde{\sigma}) = \frac{1}{\tilde{\sigma}\sqrt{2\pi}} \exp\Big(-\frac{(x-\tilde{\mu})^2}{2\tilde{\sigma}^2}\Big) \text{ for } x \in \mathbb{R} \text{ and } (\tilde{\mu},\tilde{\sigma}) \in \Theta$$

with $\Theta := \big\{\tilde{\boldsymbol{\theta}} \equiv (\tilde{\mu},\tilde{\sigma}) \in \mathbb{R}^2 \,|\, \tilde{\sigma} > 0\big\}$. In this case, performing an obvious substitution to evaluate the first integral in Equation (3), the density power divergence function can be explicitly expressed as

$$H_n(\tilde{\mu},\tilde{\sigma}) = \psi_\alpha(\tilde{\sigma}) - \left(1+\frac{1}{\alpha}\right) n^{-1} \sum_{i=1}^{n} f^{\alpha}(x_i|\tilde{\mu},\tilde{\sigma}) \tag{6}$$

$$\text{where } \psi_\alpha(\tilde{\sigma}) = c_\alpha \tilde{\sigma}^{-\alpha} \text{ with } c_\alpha := (2\pi)^{-\alpha/2}(1+\alpha)^{-1/2}.$$

See Supplemental Section S1.1 for gradient vector and Hessian matrix of H_n.

Evaluating the integrals in Theorem 2.1, Basu and Lindsay (1994) obtained the following asymptotic normality and efficiency results for the univariate Gaussian location and scale MDPD estimators:

$$n^{1/2}\big(\hat{\mu}_n^{\text{MDPD}} - \mu\big) \overset{\mathcal{D}}{\to} \mathcal{N}\Big(0, \Big(1+\frac{\alpha^2}{1+2\alpha}\Big)^{3/2}\sigma^2\Big), \tag{7}$$

$$n^{1/2}\big(\hat{\sigma}_n^{\text{MDPD}} - \sigma\big) \overset{\mathcal{D}}{\to} \mathcal{N}\Big(0, \frac{(1+\alpha)^2}{(2+\alpha^2)^2}\Big\{\frac{2(1+\alpha)^3(1+2\alpha^2)}{(1+2\alpha)^{5/2}} - \alpha^2\Big\}\sigma^2\Big) \tag{8}$$

as $n \to \infty$. Equations (7)–(8) allow for a Wald-style asymptotic inference theory for $\hat{\mu}_n^{\text{MDPD}}$ and $\hat{\sigma}_n^{\text{MDPD}}$ presented in Section 3 below.

Basu and Lindsay (1994) investigated the gross-error breakdown point of the MPDP estimator for univariate Gaussian location and scale parameters under what is referred to as "point" contamination. For the density $g(x)$ of a $\mathcal{N}(\mu,\sigma^2)$ random variable and some $\varepsilon > 0$, they considered the contaminated model

$$q(x) = (1-\varepsilon)g(x) + \varepsilon\delta_{x_0}(x) \tag{9}$$

with a point contamination supported at some $x_0 \in \mathbb{R}$, where $\delta_{x_0}(x)$ denotes Dirac's delta-"function." Note that $\delta_{x_0}(x)$ is a generalized function (i.e., a continuous functional) and not an actual integrable function. Nonetheless, since positive powers of Gaussian densities are elements of Schwartz space $\mathcal{S}(\mathbb{R})$ of rapidly decaying smooth functions, the integrals in Equation (1) can be continuously extended to accommodate for Dirac's $\delta_{x_0}(x)$.

According to Hampel et al. (1986), the breakdown occurs if the location estimate goes to infinity and/or the scale estimate goes to 0 or positive infinity as $|x_0| \to \infty$. Following this definition and assuming the model in Equation (9), the asymptotic breakdown of the MDPD estimator was shown to occur if

$$\varepsilon > \alpha/(1+\alpha)^{3/2} \text{ for any } \alpha > 0 \tag{10}$$

with the maximum of $2/(3\sqrt{3}) \approx 0.385$ attained at $\alpha^* = 2$. No results about other types of contaminations (such as cluster or "radial" contamination) are presently known in the literature and require future investigation.

The maximal breakdown point in Equation (10) (which can be attained at the price of reduced efficiency) exceeds that of classical estimators such as winsorized mean and standard deviation. Other comptetitors such as the univariate version of the MCD estimator (Rousseeuw, 1984) can attain the "optimal" asymptotic breakdown point of 50%. Nonetheless, when adjusting MDPD and MCD for equal breakdown points up to 38.5%, our simulations in Section 5 show that the efficiency of MDPD estimators (of both location and scale) is typically significantly higher than that of "raw" (unreweighted) MCD across all sample sizes which futher legitimizes practical applicability of MDPD.

3. Robust One-Sample Statistical Inference

We present a concise summary of robust inference theory based on the MDPD estimators subsequently used in empirical estimation of breakdown point in Section 5.3 and real-world example in Section 6. Consider a sample $x_1, x_2, \ldots, x_n$ from a univariate normal population with some unknown location parameter μ and scale parameter σ, possibly containing a certain fraction ε of outlers. Let $\hat{\mu}_{\text{MDPD}}$ and $\hat{\sigma}_{\text{MDPD}}$ be the MDPD location and scale estimates, respectively, computed from the data $x_1, x_2, \ldots, x_n$.

3.1. *Robust Inference About the Location Parameter*

The usual two-sided hypothesis testing framework for the location parameter μ is

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_a : \mu \neq \mu_0$$

for some given null-value of the location parameter μ_0. Since the normality assumption can be severely compromised due the presence of outliers, the power of usual z- and t-tests can be rendered virtually zero. Therefore, instead of relying on conventional test statistics, we consider the robust z-statistic

$$z_{\text{rob}} = \frac{\hat{\mu}_{\text{MDPD}} - \mu_0}{\left(1 + \frac{\alpha^2}{1+2\alpha}\right)^{3/2} \frac{\hat{\sigma}_{\text{MDPD}}}{n^{1/2}}},$$

where the constant in the denominator is given by the asymptotic variance in Equation (7). Under the null hypothesis, $z_{\text{rob}} \xrightarrow{d} \mathcal{N}(0,1)$ as $n \to \infty$.

Thus, for a test size $\gamma \in (0,1)$ of interest, the rejection region is given by

$$|z_{\text{rob}}| > z_{1-\gamma/2} \tag{11}$$

where $z_{1-\gamma/2}$ is the $(1-\gamma/2)$-quantile of a standard normal random variable z. (To avoid confusion with the exponent α in the DPD function definition, we employ a less common notation γ to denote the test size/type-I error.) The associated p-value is computed as $p = 2\mathbb{P}\big\{z > |\hat{z}_{\text{rob}}|\big\}$ where $\hat{z}_{\text{rob}}$ is the observed statistic. Similarly, the two-sided asymptotic $(1-\gamma)$-confidence interval for μ is given by

$$\hat{\mu}_{\text{MDPD}} \pm z_{1-\frac{\gamma}{2}} \left(1 + \frac{\alpha^2}{1+2\alpha}\right)^{3/4} \frac{\hat{\sigma}_{\text{MDPD}}}{n^{1/2}}. \tag{12}$$

The presence of outliers typically "shifts" the location estimate while inflating the scale and causing skewness. MDPD estimators do not suffer from these type of assumption violations rendering the interval in Equation (12) significantly tighter for contaminated samples compared to non-robust counterparts.

3.2. *Robust Inference About the Scale Parameter*

The two-sided hypothesis testing framework for the scale parameter σ is given by

$$H_0 : \sigma = \sigma_0 \quad \text{vs.} \quad H_a : \sigma \neq \sigma_0$$

for some given null-value σ_0. The robust χ^2-like test statistic we consider is

$$d^2_{\text{rob}} = \frac{\hat{\sigma}^2_{\text{MDPD}}/\sigma^2_0}{\frac{(1+\alpha)^2}{(2+\alpha^2)^2}\left\{\frac{(1+\alpha)^4(1+4\alpha^2)}{(1+2\alpha)^3} - \alpha^2\right\}\frac{1}{n}}.$$

Under the null hypothesis, $d^2_{\text{rob}} \xrightarrow{d} \chi^2_1$ as $n \to \infty$.

For a prescribed test size $\gamma \in (0,1)$, the rejection region is given by

$$d^2_{\text{rob}} < \chi^2_{1,\gamma/2} \quad \text{or} \quad d^2_{\text{rob}} > \chi^2_{1,1-\gamma/2}$$

where $\chi^2_{\gamma/2}$ and $\chi^2_{1,1-\gamma/2}$ are respective lower/upper quantiles of the χ^2_1-distribution. The respective p-value is

$$p = 2\min\Big\{\mathbb{P}\big\{\chi^2_1 < \hat{d}^2_{\text{rob}}\big\}, \mathbb{P}\big\{\chi^2_1 > \hat{d}^2_{\text{rob}}\big\}\Big\} \tag{13}$$

whith he observed statistic $\hat{d}^2_{\text{rob}}$. The two-sided asymptotic confidence interval for σ is

$$\frac{\hat{\sigma}_{\text{MDPD}}}{n^{1/2}} \sqrt{\frac{(1+\alpha)^2}{(2+\alpha^2)^2}\left\{\frac{(1+\alpha)^4(1+4\alpha^2)}{(1+2\alpha)^3} - \alpha^2\right\}} \times \big(\sqrt{\chi^2_{1,\gamma/2}}, \sqrt{\chi^2_{1,1-\gamma/2}}\big). \tag{14}$$

Similar to Equation (12), the confidence interval in Equation (14) is likely to be tighter for contaminated data than the one obtained with conventional non-robust approach.

4. New Hybrid Optimization Method

"The" gradient descent method (Hinze et al., 2008, Algorithm 2.1, p. 99) is a first-order iterative optimization procedure that is used to compute or locate minima of a scalar objective function. The method takes steps proportional to the negative gradient (anti-gradient) of the function as the (locally) steepest descent direction. To assure convergence, a typical gradient descent implementation without step control takes "baby steps" of size $\frac{k}{k+1}$ in the descent direction at the k-th iteration which may take a considerably long amount of time to achieve desired accuracy. To accelerate convergence and enforce convex constraints, we employ the well-known projected Armijo's rule (Hinze et al., 2008, Algorithm 2.5, p. 107) which provides an effective step selection mechanism. Plainly speaking, Armijo's rule selects the largest step size at each iteration as to guarantee a "uniformly linear" decay of the objective function which speeds up the convergence, while maintaining a monotonic decrease in the objective, and reduces the amount of gradient function evaluations.

Whereas gradient descent with Armijo's rule proves efficacious at initial stages (also due to potentially compromised regularity in a vicinity of the boundary making higher-order methods pointless), the convergence speed can be significantly improved if a second-order method is applied in a vicinity of a (convex) local minimum. To this end, we developed a new hybrid optimization scheme involving both gradient descent with step control and Newton's method. The procedure starts with gradient descent and then switches to Newton's method – provided the Hessian matrix remains uniformly convex and the induced sequence in the feasible set starts exhibiting numerical "Cauchy convergence." Based on simulations reported later in this section, this hybrid procedure takes significantly less time to run compared to the gradient descent with Armijo's rule as baseline.

We first formulate a general version of the proposed method as Algorithm 4.1 and prove a global convergence Theorem 4.2. Adopting the usual terminology in numerical optimization, when referring to a globally convergent scheme, we mean the ability to converge to a stationary point from any warmstart, but not necessarily to a global minimum. We later apply this general result to the univariate Gaussian DPD function $H_n(\cdot)$ by verifying the conditions of Theorem 4.2. Based on simulation results reported in Section 5, the local minimum our algorithm converges to is either a global minimum or is located remarkably close to the former, as measured by the empirical breakdown point and mean square error. In contrast to the convential approach used in robust statistics that involves sampling of multiple warmstarms, a single warmstart originating from nonrobust estimation (sample mean and standard deviation in this case) was employed to illustrate the power of the proposed approach. If more robust "pivot" estimators (such as sample median and $\hat{\sigma}_{Q_n}$-estimator of Rousseeuw and Croux (1993)) are employed, the runtime can further be reduced.

Being local optimization methods, the projected gradient descent algorithm, Newton's method and our proposed hybrid optimization scheme do not necessarily converge to the global minimum. This naturally holds for other state-of-the-art "gradient-based" competitors such as interior point, sequential quadratic programming (SQP) and active set methods implemented in the `fmincon` function of `Matlab`®. Despite this "shortcoming," local optimization methods, most notably expectation maximization (EM)-type techniques, are still among the most commonly used ones in computational statistics for a variety of reasons. On the one hand, in many situations, global optimizations algorithms are known to be at least NP-hard as the number of variables increases making them computationally prohibitive. See, e.g., Bernholt and Fischer (2004) for multivariate MCD estimator of Rousseeuw (1984). On the other hand, rigorous large-sample theories can be established for estimators computed as local minimizers of non-convex objectives (see, e.g., (McLachlan and Peel, 2000, Chapter 2) for Gaussian mixtures) so that "good" local minimizers are sufficient for both theoretical and practical purposes.

Let $\mathcal{H}$ be a Hilbert space endowed with some inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Typically, $\mathcal{H}$ is chosen isomorphically isometric (or identical) to $\mathbb{R}^m$ endowed with the usual Euclidean norm (Johnson and Wichern, 2007, Section 5.3). Let $\Theta \subset \mathcal{H}$ denote the parameter set. Usually, Θ is an open, convex set, e.g., $\Theta := \big\{\tilde{\boldsymbol{\theta}} \equiv (\tilde{\mu}, \tilde{\sigma}) \in \mathbb{R}^2 \,|\, \tilde{\sigma} > 0\big\}$ for robust univariate Gaussian estimation. Given a sample $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, let $H_n(\boldsymbol{\theta}) \equiv H_n(\boldsymbol{\theta}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ denote a smooth objective function (viz. DPD function in our paper). Select a closed, convex set $\Theta_n \subset \Theta$ (which may vary from sample to sample) such that all local minima are (likely) contained in Θ_n. Such choice is possible for robust univariate Gaussian parameter estimation on the strength of Theorem 4.3. Following (Hinze et al., 2008, p. 67), we further define the orthogonal projector

$$\Pi_{\Theta_n} \colon \mathcal{H} \to \Theta_n, \quad \tilde{\boldsymbol{\theta}} \mapsto \underset{\tilde{\tilde{\boldsymbol{\theta}}} \in \Theta_n}{\operatorname{argmin}} \|\tilde{\boldsymbol{\theta}} - \tilde{\tilde{\boldsymbol{\theta}}}\|_{\mathcal{H}}. \tag{15}$$

According to Hinze et al. (2008), $\boldsymbol{\theta}^* \in \Theta_n$ is a local minimizer of $H_n \colon \Theta_n \subset \mathcal{H} \to \mathbb{R}$ if and only if

$$\epsilon(\boldsymbol{\theta}^*) = 0 \quad \text{with} \quad \epsilon(\tilde{\boldsymbol{\theta}}) := \tilde{\boldsymbol{\theta}} - \Pi_{\Theta_n}\big(\tilde{\boldsymbol{\theta}} - \nabla H_n(\tilde{\boldsymbol{\theta}})\big). \tag{16}$$

We are now in position to state our proposed hybrid algorithm.

Algorithm 4.1. *Let $\varepsilon > 0$, $\gamma > 0$ and $\vartheta > 0$ be chosen small. e.g., $\varepsilon := 10^{-8}$, $\gamma := 10^{-4}$, and $\vartheta := 10^{-6}$.*

0. Choose $\boldsymbol{\theta}^{(0)} \in \Theta_n$.

For $k = 0, 1, 2, 3, \dots$, iterate the steps:

(1) Compute the Hessian matrix $\boldsymbol{H}^{(k)} := \nabla^2 H_n(\boldsymbol{\theta}^{(k)})$. If

$$\lambda_{\min}\big(\boldsymbol{H}^{(k)}\big) \geq \vartheta \quad \textit{and} \quad \lambda_{\max}\big(\boldsymbol{H}^{(k)}\big) \leq 1/\vartheta,$$

compute the (undamped) Newton's descent direction

$$\boldsymbol{d}_k^N = -\big(\boldsymbol{H}^{(k)}\big)^{-1} \nabla H_n(\boldsymbol{\theta}^k).$$

(2) If $\Pi_{\Theta_n}(\boldsymbol{\theta}^{(k)} + \boldsymbol{d}_k^N) = \boldsymbol{\theta}^{(k)} + \boldsymbol{d}_k^N$ (or, equivalently, $\boldsymbol{\theta}^{(k)} + \boldsymbol{d}_k^N \in \Theta_n$)

$$\lambda_{\min}\big(\boldsymbol{H}^{(k)}\big) \geq \vartheta, \quad \lambda_{\max}\big(\boldsymbol{H}^{(k)}\big) \leq 1/\vartheta \quad \textit{and}$$
$$H_n\big(\boldsymbol{\theta}^{(k)} + \boldsymbol{d}_k^N\big) - H_n\big(\boldsymbol{\theta}^{(k)}\big) \leq -\vartheta \big\|\nabla H_n(\boldsymbol{\theta}^{(k)})\big\|_{\mathcal{H}}^2,$$

update $\boldsymbol{\theta}^{(k+1)} := \boldsymbol{\theta}^{(k)} + \boldsymbol{d}_k^N$. Otherwise,

(a) Compute the alternative descent direction $\boldsymbol{d}_k^G = -\nabla H_n(\boldsymbol{\theta}^{(k)})$.

(b) *Choose the largest step size* $s_k \in \{1, \frac{1}{2}, \frac{1}{4}, \dots\}$, *for which*

$$H_n\big(\boldsymbol{\theta}^{(k)} + s_k \boldsymbol{d}_G^{(k)}\big) - H_n\big(\boldsymbol{\theta}^{(k)}\big) \le -\frac{\gamma}{s_k} \big\| \Pi_{\Theta_n}\big(\boldsymbol{\theta}^{(k)} + s_k \boldsymbol{d}_G^{(k)}\big) - \boldsymbol{\theta}^{(k)} \big\|_{\mathcal{H}}^2.$$

(c) *Update* $\boldsymbol{\theta}^{(k+1)} := \Pi_{\Theta_n}\big(\boldsymbol{\theta}^{(k)} + s_k \boldsymbol{d}_k^G\big)$.

(3) *Practically, the algorithm is terminated as soon as*

$$H_n\big(\boldsymbol{\theta}^{(k+1)}\big) - H_n\big(\boldsymbol{\theta}^{(k)}\big) < \varepsilon \quad \text{or} \quad \|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}\|_{\mathcal{H}} < \varepsilon.$$

The warmstart $\boldsymbol{\theta}^{(0)}$ can be obtained with another robust or non-robust pivot estimator of choice. We explored both options. See Section 5 for details. The convergence of Algorithm 4.1 in given in Theorem 4.2 below. See proof in Supplemental Section S1.2.

Theorem 4.2. *Let* $\Theta_n \subset \Theta \neq \emptyset$ *be an open set and let* $H_n \colon \Theta \subset \mathcal{H} \to \mathbb{R}$ *be*

a) twice continuously (Fréchet-)differentiable in Θ,
b) bounded from below over Θ_n *and*
c) possess a Lipschitz continuous gradient ∇H_n *over* Θ_n *i.e., for some* $L > 0$

$$\big\|\nabla H_n(\boldsymbol{\theta}_1) - \nabla H_n(\boldsymbol{\theta}_2)\big\|_{\mathcal{H}} \le L \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_{\mathcal{H}} \quad \text{for any} \quad \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta.$$

For any sequence $(\boldsymbol{\theta}^{(k)})_{k\in\mathbb{N}} \subset \Theta_n$ *generated by Algorithm 4.1, the following holds true:*

(1) *The method is globally convergent for any choice of the "warmstart"* $\boldsymbol{\theta}^{(0)}$, *i.e.,* $\varepsilon(\boldsymbol{\theta}^{(k)}) \to 0$ *as* $k \to \infty$.

(2) *If the sequence* $(\boldsymbol{\theta}^{(k)})_{k\in\mathbb{N}}$ *converges to some internal point* $\boldsymbol{\theta}^* \in \Theta_n$ *such that*

$$\lambda_{\min}\big(\nabla^2 H_n(\boldsymbol{\theta}^*)\big) \ge 2\vartheta \quad \text{and} \quad \lambda_{\max}\big(\nabla^2 H_n(\boldsymbol{\theta}^*)\big) \le 1/(2\vartheta),$$

then there exists $k_0 \in \mathbb{N}$ *so that Algorithm 4.1 switches to the usual Newton's method for* $k \ge k_0$ *and the superlinear convergence holds*

$$\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^*\|_{\mathcal{H}} = o\big(\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^*\|_{\mathcal{H}}\big) \quad \text{as } k_0 \le k \to \infty.$$

If additionally $\nabla^2 H_n(\cdot)$ *is Lipschitzian, the convergence becomes quadratic*

$$\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^*\|_{\mathcal{H}} = O\big(\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^*\|_{\mathcal{H}}^2\big) \quad \text{as } k_0 \le k \to \infty.$$

We turn now to the DPD function (viz. Equation (6)) in the univariate Gaussian case. In this situation, $\mathcal{H} := \mathbb{R}^2$ and $\Theta := \mathbb{R} \times (0, \infty)$. We select

$$\Theta_n := \big\{(\tilde{\mu}, \tilde{\sigma}) \in \mathbb{R}^2 \,|\, \tilde{\sigma} \ge \underline{\tilde{\sigma}}_n\big\} \text{ for some small } \underline{\tilde{\sigma}}_n = \underline{\tilde{\sigma}}_n(x_1, \dots, x_n) > 0. \tag{17}$$

The orthogonal projector $\Pi_{\Theta_n} \colon \mathcal{H} \to \Theta_n$ (cf. Equation (15)) reads then as

$$\Pi_{\Theta_n} \colon (\tilde{\mu}, \tilde{\sigma}) \mapsto \big(\tilde{\mu}_n, \max\{\tilde{\sigma}_n, \underline{\tilde{\sigma}}_n\}\big).$$

The threshold $\underline{\tilde{\sigma}}_n$ should be selected sufficiently small. We use $\underline{\tilde{\sigma}}_n = 10^{-20}$ in our simulations reported in Section 5. Practically, this choice can be assessed (and if necessary adjusted *a posteriori*): unless the iteration in Algorithm 4.1 converges at some

$\boldsymbol{\theta}^* = (\mu^*, \sigma^*)$ with $\sigma^* = \underline{\tilde{\sigma}}_n$, the threshold $\underline{\tilde{\sigma}}_n$ was likely chosen sufficiently small. On the strength of Theorem 4.3 below conditions a)–c) of Theorem 4.2 are satisfied implying our hybrid Algorithm 4.1 is applicable. See proof in Supplemental Section S1.3.

Theorem 4.3. *The univariate Gaussian DPD function $H_n(\cdot)$ in Equation* (6) *satisfies the conditions of Theorem 4.2.*

5. Simulation and Comparisons

The goals of this section are two-fold. First, we compare the runtime of our new hybrid Algorithm 4.1 with that of the "usual" gradient descent with Armijo's rule to show a speed-up of up to 10 times (cf. Section 5.1). Next, we empirically verify the $n^{-1/2}$-convergence rate and empirical breakdown point of the resulting algorithmic implementation of the MDPD estimator. As pointed out by some authors (see, e.g., Hawkins and Olive (2002)), algorithmic implementations of "brand-name" robust estimators can exhibit severe departures from their theoretical counterparts so that various desirable statistical properties theoretically established for the latter may fail to be true for the numerical algorithm implementing respective estimators. In this sense, the algorithmic implementation and not its theoretical prototype is *the* estimator. Sections 5.2 and 5.3 demonstrate that the MDPD implementation based on Algorithm 4.1 "inherits" the aforementioned theoretical properties. Finally, in Section 5.2 we compare the empirical MSE of the algorithmic MDPD with that of the univariate "raw" (i.e., unreweighted) MCD estimator ruling in favor of the former.

We implemented Algorithm 4.1 in `Matlab`®. The set of codes are provided in Supplement. For the MCD estimator of Rousseeuw (1984), we used the algorithmic implementation provided in the `FSDA` toolbox (Riani et al., 2012) in `Matlab`®. For fairness reasons, only the raw (unreweighted) MCD was used for comparisons reported throughout this section. (Iterative) reweighting is generally known to increase the efficiency of robust estimators. It was successfully applied to boost the performance of the MCD estimator and can also be adopted for the MDPD estimator. This is beyond the scope of the present work and will be part of future investigations.

5.1. *Run-Time Analysis*

We considered the following uncontaminated and contaminated scenarios:

1) clean (uncontaminated) standard Gaussian data: $x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$,
2) standard Gaussian data subject to a point contamination of size $\varepsilon > 0$ concentrated at $x_0 = 300$ (cf. Equation (9)):

$$x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \text{ for } i = 1, \ldots, \lceil (1-\varepsilon)n \rceil, \quad x_i \overset{\text{i.i.d.}}{\sim} \delta_{x_0} \text{ for } i = \lfloor \varepsilon n \rfloor + 1, \ldots, n. \tag{18}$$

Throughout this subsection, we used $\alpha = 0.5$ in the definition of the MDPD estimator. We also considered two sets of warmstarts for both algorithms: a non-robust one given by sample mean and sample variance as well as a robust warmstart given by the sample median and $\widehat{\mathrm{IQR}}/1.349$. To compare the runtime of Algorithm 4.1 (denoted "GD/NM") to that of the gradient descent method with Amijo's step control (abbreviated as "GD"), we performed extensive simulations with 50,000 replications for the first scenario (clean data) across various samples size (n = 50, 100, 500, 1,000, 5,000,

	Robust warmstart		Non-robust warmstart	
n	MDPD (GD)	MDPD (GD/NM)	MDPD (GD)	MDPD (GD/NM)
50	0.00074 (<1E-5)	0.00025 (<1E-5)	0.00070 (<1E-5)	0.00020 (<1E-5)
100	0.00080 (<1E-5)	0.00028 (<1E-5)	0.00076 (<1E-5)	0.00024 (<1E-5)
500	0.00383 (<1E-5)	0.00067 (<1E-5)	0.00364 (<1E-5)	0.00056 (<1E-5)
1,000	0.00470 (<1E-5)	0.00075 (<1E-5)	0.00454 (<1E-5)	0.00062 (<1E-5)
5,000	0.00711 (<1E-5)	0.00121 (<1E-5)	0.00678 (<1E-5)	0.00117 (<1E-5)
10,000	0.01564 (<1E-5)	0.00255 (<1E-5)	0.01512 (<1E-5)	0.00239 (<1E-5)
20,000	0.02059 (<1E-5)	0.00358 (<1E-5)	0.02000 (<1E-5)	0.00374 (<1E-5)
50,000	0.03559 (<1E-5)	0.00605 (<1E-5)	0.03496 (<1E-5)	0.00603 (<1E-5)

Table 1. Runtime (in seconds) for MDPD (GD) and hybrid MDPD (GD/NM) on uncontaminated data (with $N = 50{,}000$ replications).

	Robust warmstart		Non robust warmstart	
n	MDPD(GD)	MDPD(GD/NM)	MDPD(GD)	MDPD(GD/NM)
50	0.00122 (1E-5)	0.00041 (<1E-5)	14.83865 (0.25340)	12.81805 (0.05992)
100	0.00142 (1E-5)	0.00051 (<1E-5)	16.32262 (0.31518)	14.94859 (0.20328)
500	0.00791 (1E-5)	0.00141 (<1E-5)	80.82424 (0.27504)	37.97573 (0.29705)
1,000	0.01000 (1E-5)	0.00168 (<1E-5)	101.50674 (0.21162)	46.86788 (0.30575)
5,000	0.01658 (1E-5)	0.00308 (<1E-5)	169.16763 (0.43474)	92.81577 (0.38869)
10,000	0.03768 (1E-5)	0.00651 (<1E-5)	392.46923 (0.69275)	204.26582 (0.75015)
20,000	0.05216 (1E-5)	0.00980 (<1E-5)	538.58643 (0.59402)	307.16256 (0.89321)
50,000	0.09782 (3E-5)	0.01775 (0.00025)	989.95375 (2.68287)	555.75601 (0.99069)

Table 2. Average runtime and runtime standard deviation (in parentheses), both in seconds, for MDPD (GD) and hybrid MDPD (GD/NM) applied to contaminated data (with $N = 50{,}000$ and $N = 50$ replications for robust and non-robust warmstarts, respectively).

10,000, 20,000 and 50,000) for both non-robust and robust warmstarts (cf. Table 1). For the second scenario (point contamination), we used $N = 50{,}000$ replications for robust warmstarts and $N = 50$ replications for non-robust warmstarts due to substantially longer runtimes of GD (cf. Table 2). In view of this empirical evidence, it can be seen from Tables 1 and 2 that GD/NM runs up to 10 times faster than GD demonstrating the advantage of our proposed algorithmic MDPD implementation. Thoughout the rest of this paper, the latter will be used in lieu of the usual gradient descent (GD) method.

5.2. *Empirical Convergence Rate*

We analyzed the empirical convergence rate of the algorithmic MDPD estimator in terms of empirical MSE decay rate, which is theoretically expected to be $O(n^{-1})$ (cf. Section 3). The MCD estimator, being probably one of the toughest robust competitors, was used as reference in all simulations. To ensure adequate comparison, we matched nominal breakdown points (abbreviated as "bdp" in tables below) for both estimators using the asymptotic Equation (10) to express the breakdown point of the MDPD estimator for given α.

We performed a Monte Carlo simulation by drawing $N = 1{,}000{,}000$ univariate standard Gaussian samples ("clean" data) of size $n = 30, 50, 100, 200, 300, 400$. We introduced outlying values into randomly generated datasets to evaluate the performance of the estimators. For each sample, we computed the MDPD and MCD estimates $\hat{\mu}_{\text{MDPD}}$, $\hat{\sigma}_{\text{MDPD}}$ (using robust warmstarts) with $\alpha = 0.1$, 0.25, 0.5, 0.75, 1.0 and $\hat{\mu}_{\text{MCD}}$, $\hat{\sigma}_{\text{MCD}}$ adjusted to have identical breakdown points across all α's considered. Averaging over

N replications, we obtained the empirical MSE values

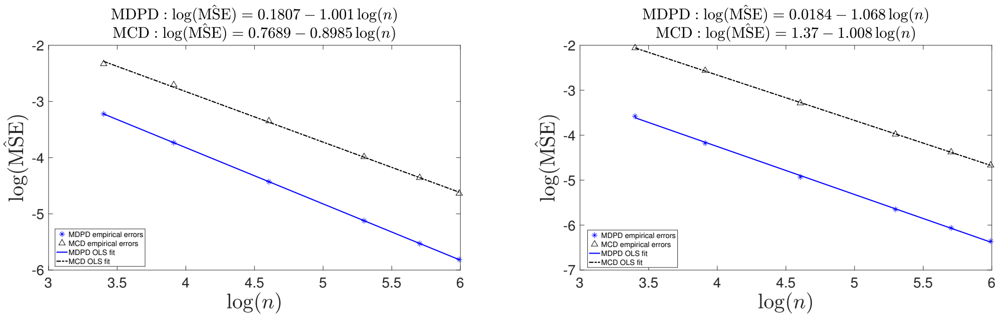
$$\hat{\mathrm{MSE}}(\hat\mu_{\mathrm{MDPD}}) = \frac{1}{N}\sum_{i=1}^{N}\hat\mu^2_{\mathrm{MDPD},i}, \qquad \hat{\mathrm{MSE}}(\hat\mu_{\mathrm{MCD}}) = \frac{1}{N}\sum_{i=1}^{N}\hat\mu^2_{\mathrm{MCD},i},$$

$$\hat{\mathrm{MSE}}(\hat\sigma_{\mathrm{MDPD}}) = \frac{1}{N}\sum_{i=1}^{N}\big(\log(\hat\sigma_{\mathrm{MDPD},i})\big)^2, \quad \hat{\mathrm{MSE}}(\hat\sigma_{\mathrm{MCD}}) = \frac{1}{N}\sum_{i=1}^{N}\big(\log(\hat\sigma_{\mathrm{MCD},i})\big)^2,$$

where $\hat\mu_i$ and $\hat\sigma_i$ denote respective estimates obtained from the i-th sample. The natural logarithm was applied to tranform scale estimators in order to account for the Riemannian nature of the variance space (Pourahmadi, 2013). Regression lines were fitted using ordinary least squares for each of the models

$$\log\big(\hat{\mathrm{MSE}}(\hat\theta)\big) = \beta_{0,\hat\theta} + \beta_{1,\hat\theta}\log(n) + \varepsilon,$$

with θ being $\hat\mu_{\mathrm{MDPD}}$, $\hat\sigma_{\mathrm{MDPD}}$, $\hat\mu_{\mathrm{MCD}}$ or $\hat\sigma_{\mathrm{MCD}}$. The slope $\beta_{1,\theta}$ corresponds to estimated convergence rate of respective estimator $\hat\theta$.



(a) $\hat\mu_{\mathrm{MDPD}}$ vs. $\hat\mu_{\mathrm{MCD}}$

(b) $\hat\sigma_{\mathrm{MDPD}}$ vs. $\hat\sigma_{\mathrm{MCD}}$

Figure 1. Empirical MSE vs. sample size n in log-log-coordinates.

Figure 1 displays the empirical MSE vs sample size n for respective location (panel (a)) and scale estimators (panel (b)) in the log-log-coordinates for $\alpha = 0.5$ (nominal breakdown point of 0.272). Table 1 documents estimated convergence rates for selected values of α.

		MDPD				MCD			
		Robust warmstart		Non-robust warmstart		Robust warmstart		Non-robust warmstart	
α	Nominal bdp	$\hat\beta_{0\hat\mu}$	$\hat\beta_{1\hat\mu}$	$\hat\beta_{0\hat\mu}$	$\hat\beta_{1\hat\mu}$	$\hat\beta_{0\hat\mu}$	$\hat\beta_{1\hat\mu}$	$\hat\beta_{0\hat\mu}$	$\hat\beta_{1\hat\mu}$
0.10	0.087	0.0051	-0.9985	0.0051	-0.9985	0.4294	-0.9863	0.4294	-0.9863
0.25	0.179	0.0528	-0.9987	0.0528	-0.9987	0.6785	-0.9566	0.6785	-0.9566
0.50	0.272	0.1834	-1.0012	0.1834	-1.0012	0.7709	-0.8986	0.7709	-0.8986
0.75	0.324	0.3293	-1.0043	0.3293	-1.0043	0.8842	-0.8786	0.8842	-0.8786
1.00	0.354	0.4660	-1.0058	0.4660	-1.0058	0.7473	-0.8292	0.7473	-0.8292

Table 3. Estimated convergence rates for both methods applied to location estimation with matched nominal breakdown point values for various α's.

For two estimators with identical convergence rates β_1, the one with a smaller β_0 value has higher relative efficiency. Refer to Table 3 and Table 4. As expected, all estimated slopes are close to -1 confirming theoretical $n^{-1/2}$-convergence (in root-MSE) of respective estimators. At the same time, the estimated intercepts for the MDPD

		MDPD				MCD			
		Robust warmstart		Non-robust warmstart		Robust warmstart		Non-robust warmstart	
α	Nominal bdp	$\hat{\beta}_{0\hat{\sigma}}$	$\hat{\beta}_{1\hat{\sigma}}$	$\hat{\beta}_{0\hat{\sigma}}$	$\hat{\beta}_{1\hat{\sigma}}$	$\hat{\beta}_{0\hat{\sigma}}$	$\hat{\beta}_{1\hat{\sigma}}$	$\hat{\beta}_{0\hat{\sigma}}$	$\hat{\beta}_{1\hat{\sigma}}$
0.10	0.087	-0.43640	-1.03912	-0.43640	-1.039120	1.07914	-1.0288	1.07914	-1.02880
0.25	0.179	-0.32901	-1.04148	-0.32901	-1.04148	1.26577	-1.02650	1.26577	-1.02650
0.50	0.272	0.01571	-1.06732	0.01533	-1.06725	1.36584	-1.00766	1.36584	-1.00766
0.75	0.324	0.35483	-1.09579	0.35456	-1.09574	1.47279	-1.00689	1.47279	-1.00689
1.00	0.354	0.60271	-1.11609	0.60279	-1.11611	1.41822	-0.98520	1.41822	-0.98520

Table 4. Estimated convergence rates for both methods applied to scale estimation with matched nominal breakdown point values for various α's.

estimator are typically smaller than their counterparts from the MCD estimator implying higher relative efficiency of the former. The efficiency of both estimator pairs can be graphically compared using Figure S(3) and Figure S(4) in the Supplement. The latter display boxplots for empirical MSE values for both location estimators (Supplemental Figure S(3)) and scale estimators (Supplemental Figure S(4)). In most cases, it can be observed the boxplots for the MDPD estimator are contained in those for the MCD estimator suggesting advantage of MPDP over MCD.

5.3. *Empirical Breakdown Point*

Plainly speaking, the breakdown point of an estimator is defined as the maximal proportion of "bad" points an estimator can handle before it becomes uninformative due to bias "explosion." Since algorithmic implementations of robust estimators can oftentimes exhibit suboptimal breakdown points, our goal is to empirically verify the nominal breakdown point. Again, the MDPD estimator will be compared side-by-side with the MCD estimator with the same nominal breakdown point. To alleviate undue advantages of the MDPD estimator, (non-robust) sample mean and sample standard deviation were used as a warmstart for Algorithm 4.1. We arbitrarily chose $\alpha = 0.5$ and used Equation (10) to compute the asymptotic breakdown point of DPD as $0.5/((1+0.5)^{\frac{3}{2}}) = 0.272$.

Motivated by the theoretical definition of breakdown, we declare a location estimator $\hat{\mu}$ or a scale estimator $\hat{\sigma}$ of population parameters μ and σ to breakdown if

$$|\hat{\mu} - \mu| \leq k\sigma \quad \text{or} \quad (1/k) \leq |\hat{\sigma}/\sigma| \leq k \tag{19}$$

fails for some large k, respectively. We chose $k = 10$ in our simulations. Due to affine equivariance, we let $\mu = 0$, $\sigma = 1$.

		$\hat{\mu}_n$		$\hat{\sigma}_n$	
n	Nominal bdp	MCD	MDPD	MCD	MDPD
30	8	9	10	9	10
50	13	15	16	15	16
100	27	28	31	28	31
200	54	56	61	55	61
300	81	85	91	83	91
400	108	112	121	110	121

Table 5. Empirical breakdown point comparisons.

For this simulation, a total of $N = 5{,}000$ standard normal samples of different sizes $n = 30, 50, 100, 200, 300, 400$ were drawn. "Bad" points were then introduced to each sample under the point contamination model (viz. Equation (18)). Both DPD (with $\alpha = 0.5$) and MCD (with a nominal breakdown point of 0.272) estimators to estimate the mean and the standard deviation for each sample. If for a given number of outliers 0.1% or more estimates failed to satisfy the empirical no-breakdown condition in Equation (19), the estimator was declared to breakdown in respective situation. The results are displayed in Table 5. See also Supplemental Figure S(5). In sum, both estimators pairs empirically performed close to their theoretically predicted behavior with MDPD allowing for more outliers before breakdown occurs as compared to MCD.

5.4. *Additional Comparisons*

We present additional simulation studies to benchmark the performance of our hybrid algorithm against that of projected gradient descent (GD), Newton's method and the three state-of-the-art local optimizers (i.e., interior point, SQP and active set methods) from the `fmincon` routine of `Matlab`®. In Section 5.4.1 below, we report extensive simulations for three contaminated statistical models where the global minimizer is unknown so that the average objective value is the sole metric available. In contrast, in Sections 5.4.2 and 5.4.3, Rosenbrock's banana and Mishra's bird functions with known global minima are considered allowing for evaluating global convergence properties. As expected, unless the problem is convex, no convergence to global minimum can be guaranteed in general.

5.4.1. Gaussian, Laplace and Exponential Contaminated Models

Consider three families of contaminated models:

$$\begin{aligned}
&\text{(I):}\ \ (1-\varepsilon)\mathcal{N}(0,1)+\varepsilon_1\mathcal{N}(-\sqrt{50},\psi^2)+\varepsilon_2\mathcal{N}(\sqrt{50},\psi^2)+\varepsilon_3\mathcal{N}(\sqrt{100},\psi^2),\\
&\text{(II):}\ \ (1-\varepsilon)\mathrm{Laplace}(0,1)+\varepsilon_1\mathrm{Laplace}(-\sqrt{50},\psi)\\
&\qquad\quad +\varepsilon_2\mathrm{Laplace}(\sqrt{50},\psi)+\varepsilon_3\mathrm{Laplace}(\sqrt{100},\psi),\\
&\text{(III):}\ \ (1-\varepsilon)\mathrm{Exp}(1)+\big(50+\varepsilon_1\mathrm{Exp}(1/0.02)+\varepsilon_2\mathrm{Exp}(1/0.05)+\varepsilon_3\mathrm{Exp}(1/0.1)\big),
\end{aligned}$$

where $\varepsilon = 0.2, 0.3$, $\psi = 0.1, 3.0$ and $\varepsilon_1, \varepsilon_2, \varepsilon_3$ are randomly selected non-negative numbers summing up to ε. The number 50 in (III) shifts the mean of the three latter exponential distributions by 50 to the right. Models (I), (II) and (III) correspond to standard Gaussian, Laplace and exponential "bulks" contaminated by three clusters of distant outliers. While Gaussian and exponential distributions have smooth density functions and, thus, are allowed by our convergence theory, the Laplace density is not differentiable at the origin, yet semismooth. Therefore, the generalized gradient was considered instead. Our hybrid optimization method worked properly despite this assumption violation.

Selecting $\alpha = 0.25, 0.5, 0.75$, in each of these scenarios, we generated 50,000 independent samples of size $n = 50, 100, 200, 300, 500$ from respective model and compared the performance of our hybrid optimization method to the aforementioned competitors in terms of the H_n objective value at respective numerical (local) minimum using usual MLE estimators as warmstarts. The active set method was excluded for the contaminated Laplace model since the routine produced sporadic exceptions and failed to converge (probably, due to semismoothness of Laplace density). Sample means and standard deviations were recorded. A total of 12 + 12 + 6 = 30 tables were obtained. See Supplemental Sections S3.1, S3.2 and S3.3. One of these supplemental tables, namely Supplemental Table S3.1.3, is reproduced below as Table 6. Analyzing estimated average objective function values, our hybrid optimization algorithm performed head-to-head with projected GD, interior point, SQP and active set methods across all scenarios considered, while Newton's method performed worst.

Method	$n = 50$	$n = 100$	$n = 200$	$n = 300$	$n = 500$
Hybrid	-0.3405 (0.0314)	-0.3362 (0.0216)	-0.3342 (0.0149)	-0.3336 (0.0121)	-0.3332 (0.0094)
NM	-0.2102 (0.0026)	-0.2127 (0.0020)	-0.2147 (0.0014)	-0.2137 (0.0011)	-0.2142 (0.0009)
GM	-0.3405 (0.0314)	-0.3362 (0.0216)	-0.3342 (0.0149)	-0.3336 (0.0121)	-0.3332 (0.0094)
fmcIP	-0.3405 (0.0314)	-0.3362 (0.0216)	-0.3342 (0.0149)	-0.3336 (0.0121)	-0.3332 (0.0094)
fmcSQP	-0.3405 (0.0314)	-0.3362 (0.0216)	-0.3342 (0.0149)	-0.3336 (0.0121)	-0.3332 (0.0094)
fmcAS	-0.3405 (0.0314)	-0.3362 (0.0216)	-0.3342 (0.0149)	-0.3336 (0.0121)	-0.3332 (0.0094)

Table 6. H_n objective estimated averages and standard deviations (in parentheses) with $\alpha = 0.75$, $\varepsilon = 0.20$ and $\psi = 0.10$

5.4.2. Rosenbrock's Banana Function

As a test example, we considered the "banana function" of Rosenbrock (1960)

$$f(x, y) = (a - x)^2 + b(y - x^2)^2 \text{ with } a = 1, b = 100.$$

The function attains a unique global minimum of 0 over $(x, y) \in \mathbb{R}^2$ at $(1, 1)$, but has a parabolic flat "valley" making the minimization problem challenging for non-adaptive optimization techniques. Based on 50,000 replications with random warmstarts (uniform over $[-50, 50]^2$), Table 7 reports estimated average and standard deviation values along with percentages of cases for respective estimators to converge to the (global) minimum. The first four methods performed very well with respect to all metrics with the hybrid method exhibiting perfect performance.

Method	Avg H_n	Std H_n	Convergence %
Hybrid	0	0	100
NM	0.0074	0.58	99.98
GD	0.16	8.99	99.92
fmcIP	1.81	49.70	99.85
fmcSQP	1271.23	9818.66	7.60
fmcAS	360.59	5787.10	25.78

Table 7. Estimated mean and standard deviations of H_n along with convergence percentages

5.4.3. Mishra's Bird Problem

The constrained optimization problem of Mishra (2006) reads as

$$f(x, y) \to \min \text{ over } (x, y) \text{ with } (x + 5)^2 + (y + 5)^2 \leq 25, -10 \leq x \leq 0, -6.5 \leq y \leq 0$$

where $f(x, y) = \sin(y) \exp(1 - \cos(x))^2 + \cos(x) \exp(1 - \sin(y))^2 + (x - y)^2$. Pos-

Method	Avg H_n	Std H_n	Convergence %
Hybrid	-41.48	47.55	34.63
NM	1.81	24.96	1.96
GD	-43.85	47.57	36.12
fmcIP	-46.54	51.04	41.44
fmcSQP	-41.88	49.47	36.52
fmcAS	-40.91	49.64	36.01

Table 8. Estimated mean and standard deviations of H_n along with percentages of convergence to global mininum

sessing multiple local minima, the function attains a global minimum at $(x^*, y^*) \approx (-3.1302468, -1.5821422)$ with $f(x^*, y^*) \approx -106.7645367$. Based on 50,000 replications with random warmstarts (uniform over $[-10, 0]^2$), Table 8 reports estimated average and standard deviation values along with percentages of cases for respective estimators to converge to the (global) minimum. All methods perform fairly poorly with Newton's method exhibiting the worst performance. In sum, as expected, none of the local optimizers considered, including our hybrid technique, is consistently able to converge at global minimum. Nonetheless, in situations where the objective function is nearly convex, as it is the case in Supplemental Figures S(1) and S(2) in the context of DPD minimization, our hybrid scheme as well as other state-of-the-art local optimization techniques are expected to converge to the global minimum.

6. Example

Our real-world data illustration is based on an excerpt from the notifiable conditions report provided by the City of El Paso Department of Public Health (2021) containing monthly records of the total number of new cases of various notifiable conditions in El Paso, Texas (USA) starting from January 2004. All 2004–2017 records were manually tabulated as part of a student research project supervised by the second author in 2018. Focusing on Sexually Transmitted Diseases (STDs), we chose to analyze the recorded chlamydia cases because the associated time series appeared to exhibit anomalous peaks and the counts were sufficiently large to be treated as a continuous variable. Figure 2 is a plot of monthly new chlamydia cases recorded during this time period where the red circles show some "suspicious" peaks.

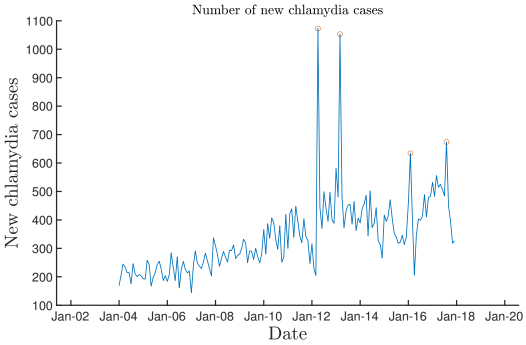


Figure 2. Reported monthly new chlamydia cases from 2004 to 2017.

Figure 2 suggests the variability in the dataset changes across the recorded time frame. As the number of chlamydia cases increase, volatility increases and vice versa. To account for this (potential) heteroscedasticity, we chose to adopt the well-known Geometric Brownian Motion (GBM) model given by stochastic differential equation

$$\mathrm{d}S(t) = rS(t)\mathrm{d}t + \sigma S(t)\mathrm{d}W(t), \quad S(0) = S_0, \tag{20}$$

where S_0 is typically constant and $\big(W(t)\big)_{t\geq 0}$ is a standard Wiener process. The scalar parameters, $r \in \mathbb{R}$ and $\sigma > 0$, are referred to as drift and volatility, respectively. Being probably the simplest heteroscedastic continuous-time model widely applied in quantitative finance (Brigo et al., 2009), Equation (20) is well-suited for modeling continuous non-negative quantities as the solution process (verified using Itô's calculus)

$$S(t) = S_0 \exp\left\{\Big(r - \frac{1}{2}\sigma^2\Big)t + \sigma W(t)\right\} \tag{21}$$

(almost surely) remains positive as long as the initial value S_0 is selected positive.

Our process dataset contains 14 years' worth of montly observations adding up to 168 ordered data points. We used the first 12 years (144 observations) to calibrate the model using three different sets of location/scale estimators (viz. MDPD, MCD and usual) and performed a *post hoc* model diagnostic (Section 6.1), while the latter 2 years (24 observations) served for "backtesting" purposes to assess the performance for each of the three methods in terms of empirical MSE (Section 6.2). The usual risk neutrality requirements typical for financial forecasting are not relevant in our context. Therefore, both estimation and prediction can be performed under the physical measure associated with the Wiener process $\big(W(t)\big)_{t\geq 0}$ instead of the martingale measure widely employed in quantitative finance (Matlsev and Pokojovy, 2021).

6.1. *Model Calibration and Diagnostic Plots*

Assume the process $\big(S(t)\big)_{t\geq 0}$ following Equation (20) is observed over an equispaced discrete time grid $\{t_0, t_1, \ldots, t_n\}$ with $t_k - t_{k-1} \equiv \Delta t > 0$. Introducing the log-differences

$$x(t_k) := \ln(S(t_k)) - \ln(S(t_{k-1})) = \Big(r - \frac{1}{2}\sigma^2\Big)\Delta t + \sigma\big(W(t_k) - W(t_{k-1})\big),$$

we conclude

$$x(t_k) \overset{\text{i.i.d.}}{\sim} \mathcal{N}\Big((\Delta t)\Big(r - \frac{1}{2}\sigma^2\Big), (\Delta t)\sigma^2\Big). \tag{22}$$

Letting $\hat{\mu}_x$ and $\hat{\sigma}_x$ denote location and scale estimates obtained from $x(t_k)$'s, e.g., $\bar{x}$ and $\sqrt{s^2}$ or their robust MDPD or MCD counterparts, Equation (22) implies

$$\hat{\mu}_x = (\Delta t)\Big(\hat{r} - \frac{1}{2}\hat{\sigma}^2\Big) \quad \text{and} \quad \hat{\sigma}_x^2 = (\Delta t)\hat{\sigma}^2$$

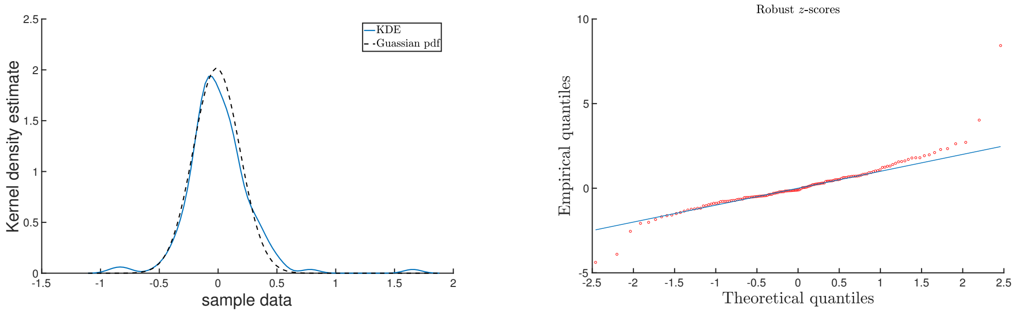
or, solving for $\hat{r}$ and $\hat{\sigma}$, the drift and volatility can be estimated via

$$\hat{r} = \frac{\hat{\mu}_x}{\Delta t} + \frac{1}{2}\frac{\hat{\sigma}_x^2}{\Delta t} \quad \text{and} \quad \hat{\sigma} = \frac{\hat{\sigma}_x}{\sqrt{\Delta t}}.$$

Estimates	$\hat{\mu}_x$	$\hat{\sigma}_x$	drift $\hat{r}$	volatility $\hat{\sigma}$
MDPD($\alpha = 0.5$)	-0.0142	0.1986	0.0055	0.1986
MCD	-0.0245	0.0391	-0.0237	0.0391
Usual	0.0043	0.2697	0.0407	0.2697

Table 9. GBM parameter estimation based on 10 years of monthly chlamydia data ($\Delta t = 1$ month). MCD and MDPD are matched to both have the breakdown point of 0.272.

Applied to 2004–2015 chlamydia time series data (with $\Delta t = 1$ month), Table 9 lists respective estimates obtained with the usual sample mean $\bar{x}$/standard deviation $\sqrt{s^2}$, MDPD($\alpha = 0.5$) and MCD(bdp $= 0.272$) estimators. To assess the quality of fit, we prepared robust diagnostic plots using the estimates obtained with MDPD($\alpha = 0.5$). As can be seen from Figure 3(a), the distribution of the log-differences is nearly normal with some outliers on both tails. Likewise, the robust normal Q-Q plot in Figure 3(b) (based on $\hat{\mu}_{x,\mathrm{MDPD}}$ and $\hat{\sigma}_{x,\mathrm{MDPD}}$) exhibits a good quality of fit over the "bulk region."



(a) Kernel density estimate for $x(t_k)$'s

(b) Robust Gaussian Q-Q-plot for $x(t_k)$'s

Figure 3. Diagnostic plots based on log-differences $x(t_k) = \log\big(S(t_k)\big) - \log\big(S(t_{k-1})\big)$.

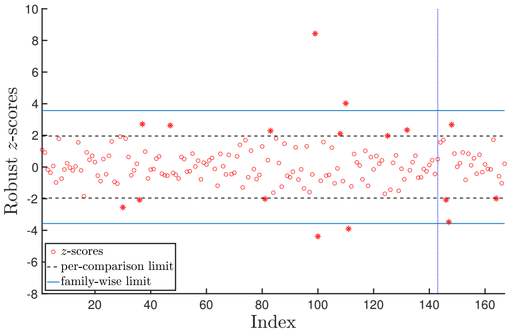


Figure 4. Outlier detection plot.

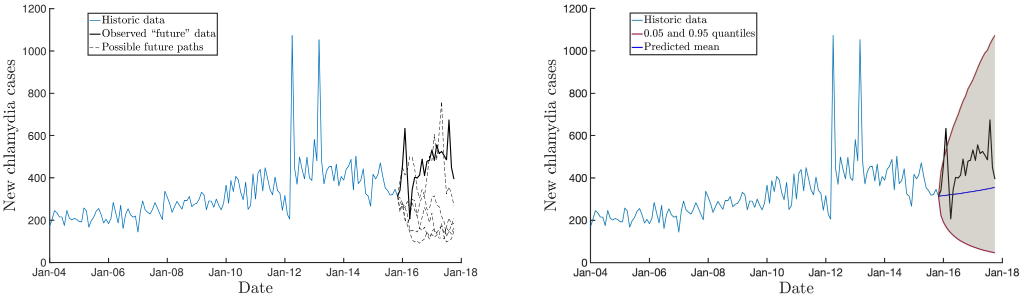
We adopted a formal two-stage testing procedure akin to Fisher's protected test to detect outlying $x(t_k)$'s. Applying MDPD($\alpha = 0.5$) estimation (to 2004–2015 data), robust z-scores $z_k = \frac{x(t_k) - \hat{\mu}_{x,\mathrm{MDPD}}}{\hat{\sigma}_{x,\mathrm{MDPD}}}$ were computed and plotted in Figure 4. Assuming indepedence amongst $z(t_k)$'s, a family-wise test of size 0.05 for the presence of outlier was performed using the cut-off values $\pm z_{0.0975^{1/n}} = \pm 3.57$ with $n = 144 - 1 = 143$ (cf. (Jobe and Pokojovy, 2015, p. 1543)). Since the family-wise test was statistically significant, the sample was declared to contain outliers, which were then detected on a per-comparison basis using a cut-off values $\pm z_{0.975} = \pm 1.96$. This resulted in 13 outliers detected in historic data and another four outliers in backtesting data suggesting the importance of robust estimation in our scenario. In sum, the GBM model with robustly estimated parameters appears to be an appropriate choice for analyzing our dataset.

6.2. *Forecasting Future Cases*

Having estimated r and σ over a time horizon $[t_0, T]$, Equation (21) can be used to simulate future paths of $\big(S(t)\big)_{t\geq 0}$ setting $S_0 = S(T)$. For computational convenience, the Wiener process $\big(W(t)\big)_{t\geq 0}$ is typically replaced with a random walk. Thus, the discrete (recurrent) form of Equation (21) reads as

$$S(t_k) = S(t_{k-1}) \exp\left\{ \Big(r - \frac{1}{2}\sigma^2\Big)(\Delta t) + \sigma(\Delta t)^{1/2}\varepsilon_k \right\}, \quad S(t_0) = S_0 \tag{23}$$

where $\varepsilon_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Note that $\big(\log S(t_k)\big)_{k\geq 0}$ is an AR(1) process. Using Equation (23) with μ and σ replaced by respective empirical estimates $\hat{\mu}$ and $\hat{\sigma}$, both point estimates (via mean, median or other quantiles) and prediction regions for future $S(t_k)$ values can be obtained based on a Monte-Carlo simulation.



(a) Five possible future paths

(b) Pointwise 90% prediction region

Figure 5. Predicting new chlamydia cases over an 24-month time horizon based on MDPD($\alpha = 0.5$) estimators computed from 144 historical observations

In addition to historic observation over the first 150 months, Figure 5(a) displays five possible future realizations of $\big(S(t_k)\big)_{k\geq 0}$, while Figure 5(b) shows the pointwise (i.e., non-simultaneous) 90% prediction region (shaded) and the expected number of future new chlamydia cases $\mathbb{E}\big[S(t_k)\big]$ vs. time t_k estimated based on a Monte-Carlo simulation of size $N = 50{,}000$ with the parameters r and σ robustly estimated using MDPD($\alpha = 0.5$). The "innovations" ε_k's were generated as i.i.d. standard Gaussian, i.e., under the physical measure. It can be seen that the actually observed numbers of new chlamydia cases were mostly contained in the prediction region for the whole duration of the 24-month backtesting period suggesting reliability of the forecast. Additionally, the upper limit of the prediction region provides information about the severity of the "worst-case" scenario under the GBM model employed.

Another important aspect is a comparison between the prediction quality measured in terms of MSE amongst the forecasts obtained with MDPD($\alpha = 0.5$), MCD(bdp $= 0.272$) and usual sample mean/standard deviation. With $S_{\text{obs}}(t_k)$ denoting the observed number of new chlamydia cases over the backtesting period comprised by the latter 24 months, the empirical MSE at time t_k is computed as

$$\widehat{\text{MSE}}\big[S(t_k)\big] = \frac{1}{N}\sum_{i=1}^{N}\big(S^{(i)}(t_k) - S_{\text{obs}}(t_k)\big)^2$$

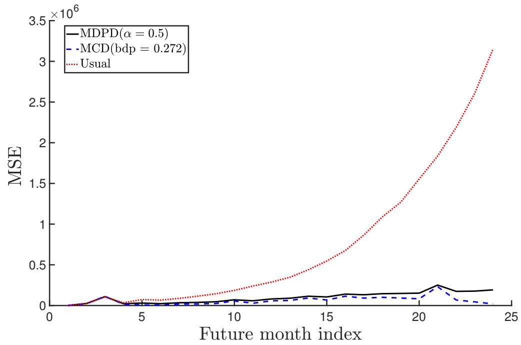


Figure 6. Empirical MSE comparison amongst predictions obtained based MDPD, MCD and usual location/scale estimators.

based on $N = 50{,}000$ independent Monte-Carlo replications $S^{(i)}(t_k)$, $i = 1, 2, \ldots, N$. The empirical MSE curves associated with MDPD and MCD estimators displayed in Figure 6 are head-to-head (with a slight advantage for the MCD curve) but lie way below the MSE curve for the usual estimator indicating the importance of robust estimation under possible violations of model assumptions in the context of forecasting new chlamydia cases.

7. Conclusions

Location and scale estimators are an indispensable instrument in any statistical toolbox. A variety of nonrobust and robust estimators are available for estimating parameters of univariate Gaussian data. Focusing on MDPD estimation, we developed a new hybrid optimization algorithm for proved a global convergence property and demonstrated how our algorithm can be adopted to put forth an empirically robust implementation of the MDPD estimator. We compared it to the more prominent MCD estimator and showed how the MDPD estimator has higher efficiency when matched for the breakdown point. Using a real-world biomedical dataset, we illustrated how MDPD estimator can be applied to calibration, diagnostic and forecasting of a GBM model. Our future work will include extending our methodology to robust estimation of multivariate Gaussian parameters without relying on robust warmstarts.

SUPPLEMENTARY MATERIALS

Supplementary figures, tables and proofs: A pdf document with additional figures, tables and proofs. (Adobe `pdf` file)

Supplementary `Matlab`® codes: A set of programs implementing Algorithm 4.1 as well as computational results reported in this paper. (GNU zipped tar file)

Acknowledgements

The authors thank Mr. Guillermo Lopez-Ramirez for tabulating 2004–2017 City of El Paso (TX) Department of Public Health notifiable conditions used in Section 6. The authors also acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources extensively used for simulations reported in this manuscript. URL: `http://www.tacc.utexas.edu`. Comments and improvement suggestions from the Editor-in-Chief Professor Narayanaswamy Balakrishnan, Associate Editor and an anonymous referee are greatly appreciated.

Disclosure Statement

The authors have no conflict of interests to declare.

Data Availability Statement

All data presented in the article are available in the Supplement.

Funding

The study was funded partially by the National Institute on Minority Health and Health Disparities (NIMHD) grant (U54MD007592) and the National Science Foundation under Grant No 2210929.

References

Avriel, M. (2003). *Nonlinear Programming: Analysis and Methods.* Dover Publishing, Mineola, New York.

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559.

Basu, A. and Lindsay, B. G. (1994). Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, 46(4):683–705.

Basu, A., Mandal, A., Martin, N., and Pardo, L. (2013). Testing statistical hypotheses based on the density power divergence. *Annals of the Institute of Statistical Mathematics*, 65:319–348.

Basu, A., Mandal, A., Martin, N., and Pardo, L. (2017). A Wald-type test statistic for testing linear hypothesis in logistic regression models based on minimum density power divergence estimator. *Electronic Journal of Statistics*, 11(2):2741–2772.

Basu, A., Mandal, A., Martin, N., and Pardo, L. (2018). Density power divergence tests for composite null hypotheses. *Sankhya, Series B*, 80(2):222–262.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, 5(3):445–463.

Bernholt, T. and Fischer, P. (2004). The complexity of computing the MCD-estimator. *Theoretical Computer Science*, 326:383–398.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization.* Cambridge University Press, New York.

Brigo, D., Dalessandro, A., Neugebauer, M., and Triki, F. (2009). A stochastic processes toolkit for risk management: Mean reverting processes and jumps. *Journal of Risk Management in Financial Institutions*, 3(1).

City of El Paso Department of Public Health (2021). Epidemiology: Notifiable conditions by year and disease fact sheets. URL: `https://www.elpasotexas.gov/public-health/services/epidemiology/`.

Croux, C., R. P. and Hössjer, O. (1994). Generalized S-estimators. *Journal of the American Statistical Association*, 89:1271–1281.

Croux, C., R. P. and Hössjer, O. (1999). A class of locally and globally robust regression estimates. *Journal of the American Statistical Association*, 94:174–188.

Donoho, D. (1982). Breakdown properties of multivariate location estimators. Qualifying paper, Harvard University.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. *A Festschrift for Erich L. Lehmann*, 157184.

Fox, J. et al. (2002). Robust Regression. *An R and S-Plus Companion to Applied Regression*, page 91.

Ghosh, A., Mandal, A., Martin, N., and Pardo, L. (2016). Influence analysis of robust Wald-type tests. *Journal of Multivariate Analysis*, 147:102–126.

Grippo, L., Lampariello, F., and Lucidi, S. (1986). A nonmonotone line search technique for Newton's method. *SIAM Journal on Numerical Analysis*, 23:707–716.

Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, pages 1887–1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics*. Wiley Online Library.

Hawkins, D. and Olive, D. (2002). Inconsistency of resampling algorithms for high breakdown regression estimators and a new algorithm. *Journal of the American Statistical Association*, 97:136–148.

Hinze, M., Pinnau, R., Ulbrich, M., and Ulbrich, S. (2008). *Optimization with PDE Constraints*, volume 23. Springer Science & Business Media.

Huber, P. J. (2009). *Robust Statistics*. John Wiley & Sons Inc., Hoboken, NJ, 2nd edition.

Hubert, M., Rousseeuw, P., and Verdonck, T. (2012). A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, 21(3):618–637.

Jobe, J. M. and Pokojovy, M. (2015). A cluster-based outlier detection scheme for multivariate data. *Journal of the American Statistical Association*, 110(512):1543–1551.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, Upper Saddle River, NJ, 6th edition.

Kent, J. and Tyler, D. (1996). Constrained M-estimation for multivariate location and scatter. *The Annals of Statistics*, 24:1346–1370.

Maronna, R.A., M. R. and Yohai, V. (2006). *Robust Statistics: Theory and Methods*. John Wiley & Sons Inc., Hoboken, NJ.

Martin, R. D. (1979). Robust estimation for time series autoregressions. In *Robustness in Statistics*, pages 147–176. Elsevier.

Matlsev, V. and Pokojovy, M. (2021). Applying Heath-Jarrow-Morton model to forecasting the US Treasury daily yield curve rates. *Mathematics*, 9(2):114.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, INC.

Mishra, S. (2006). Some new test functions for global optimization and performance of repulsive particle swarm method. *SSRN Electronic Journal*, pages 1–12.

Peña, D. and Prieto, F. J. (2001a). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, 43(3):286–303.

Peña, D. and Prieto, F. J. (2001b). Multivariate outlier detection and robust covariance matrix estimation: Response. *Technometrics*, 43(3):306–310.

Potschka, A. (2014). Backward step control for global Newton-type methods. *SIAM Journal on Numerical Analysis*, 54(1):361–387.

Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation: With High-Dimensional Data*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ.

Reyen, S. S., Miller, J. J., and Wegman, E. J. (2009). Separating a mixture of two normals with proportional covariances. *Metrika*, 70:297–314.

Riani, M., Perrotta, D., and Torti, F. (2012). FSDA: A MATLAB toolbox for robust analysis and interactive data exploration. *Chemometrics and Intelligent Laboratory Systems*, 116:17–32.

Rosenbrock, H. H. (1960). An automatic method for finding the greatest or least value of a function. *The Computer Journal*, 3(3):175–184.

Rousseeuw, P. (1984). Least median of squares regression. *Journal of American Statistical Association*, 79:310–329.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. V. and Wertz, W., editors, *Mathematical Statistics and Applications*, volume B, pages 283–297. Reidel Publishing, Dordrecht.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283.

Rousseeuw, P. J. and van Driessen, K. (1999). A fast algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41:212–223.

Small, C. and Wang, J. (2003). *Numerical Methods for Nonlinear Estimating Equations*, volume 29 of *Oxford Statistical Science Series*. Oxford University Press, New York.

Stahel, W. (1981). *Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen*. PhD thesis, ETH Zurich.

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC, Boca Raton, FL.

Varadhan, R. and Roland, C. (2008). Simple and globally-convergent methods for accelerating the convergence of any EM algorithm. *Scandinavian Journal of Statistics*, 35(2):335–353.