Limitations of the Wasserstein MDE for univariate data

Yannis G. Yatracos

Received: 20 July 2022 / Accepted: 3 September 2022 / Published online: 17 October 2022
© The Author(s) 2022

Abstract

Minimum Kolmogorov and Wasserstein distance estimates, \( \hat{\theta}_{MKD} \) and \( \hat{\theta}_{MW D} \), respectively, of model parameter, \( \theta \), are empirically compared, obtained assuming the model is intractable. For the Cauchy and Lognormal models, simulations indicate both estimates have expected values nearly equal to \( \theta \), but \( \hat{\theta}_{MKD} \) has in all repetitions of the experiments smaller SD than \( \hat{\theta}_{MW D} \), and \( \hat{\theta}_{MKD} \)’s relative efficiency with respect to \( \hat{\theta}_{MW D} \) improves as the sample size, \( n \), increases. The minimum expected Kolmogorov distance estimate, \( \hat{\theta}_{MEKD} \), has eventually bias and SD both smaller than the corresponding Wasserstein estimate, \( \hat{\theta}_{MEWD} \), and \( \hat{\theta}_{MEWD} \)’s relative efficiency improves as \( n \) increases. These results hold also for stable models with stability index \( \alpha < 2 \), which have finite first moment. Similar results are expected to hold for multivariate heavy tail models. Combined with existing results in the literature, the findings do not support the use of Wasserstein distance in statistical inference, especially for intractable and Black Box models with unverifiable heavy tails.

Keywords Bias · Black-Box model · Coarsened posterior · Data science · Intractable model · Kolmogorov distance · Model with heavy tails · Minimum distance estimate (MDE) · Rate of convergence of estimate · Relative efficiency · Robustness · Sampler · Wasserstein distance

1 Introduction

Kantorovich (1942) introduced a measure for the work needed in the translocation of two mass distributions, \( \mu \) and \( \mu^* \), that has led to Wasserstein (1969) distances, \( W_p(\mu, \mu^*) \), \( p \geq 1 \). \( W_p \) has been used recently in several research areas, including Machine Learning and Statistics; see, e.g., Villani (2008) and Kolouri et al. (2017).

Bassetti et al. (2006a) introduced for a tractable statistical model the minimum Wasserstein distance estimate, \( \hat{\theta}_{n,MWD} \), of a parameter, \( \theta \), using \( W_p \) and the empirical distribution, \( \hat{\mu}_\theta \), of the observed data; \( \theta \in \Theta \), \( n \) is the sample size, \( p \geq 1 \). For data in \( \mathbb{R}^d \), \( d > 2 \), the weakness of \( \hat{\mu}_\theta \) in approximating the underlying distribution (or probability measure), \( \mu_\theta \), in \( W_p \)-distance (Talagrand 1994) is a known drawback; see also, e.g., Dudley (1968) and Weed and Bach (2019, p. 2). In addition, \( W_p \) requires the existence of the \( p \)th moment of the model, it is hard to compute when \( d > 1 \), it is not a smooth functional and it is not robust; see, e.g., Wasserman (2019). There are no tools to do inference with \( W_p \), except when the underlying model has bounded support or strong model assumptions hold, which cannot be verified for Black Box or intractable models; see, e.g., Sommerfeld and Munk (2018, p. 220), Bernton et al. (2019b, p. 245) and Wasserman (2019). These drawbacks limit the use of \( \hat{\theta}_{n,MWD} \) and \( W_p \) in multivariate statistical inference; \( p \geq 1 \).

Siaperembudur et al. (2012) studied integral probability metrics and their use in nonparametric two-sample testing, and showed for the empirical estimate of the kernel distance, \( \kappa(P, Q) \), between probabilities \( P \) and \( Q \): (a) that it converges at a faster rate than the empirical estimate of the Kantorovich-Wasserstein distance, \( W_1(P, Q) \), and (b) that its rate of convergence is independent of the dimension, \( d \), of the observations. The authors concluded that the kernel distance, \( \kappa \), “is better suited for use in statistical inference applications” (p. 1569, lines -4 to -1). Despite this conclusion and the draw-
backs in the previous paragraph, the Wasserstein distance has been used recently in statistical inference for intractable univariate models, e.g., in Bernton et al. (2019a,b).

In simulations herein, using \( W_1 \) and univariate heavy tail models\(^1\) treated as intractable, the Minimum Wasserstein Distance Estimate, \( \hat{\theta}_{MWD} \), and the Minimum Expected Wasserstein Distance Estimate, \( \hat{\theta}_{MEWD} \) (Bernton et al. 2019a), are obtained using several samples of synthetic data. These estimates are compared, respectively, with the corresponding versions of Kolmogorov Minimum Distance Estimate (MDE, Wolfowitz 1957), \( \hat{\theta}_{MKD} \) and \( \hat{\theta}_{MEKD} \), introduced in Yatracos (2021). It is observed that the empirical efficiencies of \( \hat{\theta}_{MKD} \) and \( \hat{\theta}_{MEKD} \) improve more and more as \( n \) increases, with regard to \( \hat{\theta}_{MWD} \) and \( \hat{\theta}_{MEWD} \), respectively. Explanations and theoretical confirmation are provided for these findings, which are also expected to hold for multivariate models with heavy tails. These results, along with the previously presented drawbacks and additional limitations of statistical procedures derived with \( W_1 \) and presented in the sequel (in this section), do not support the use of the Wasserstein distances in statistical inference, especially for Black Box or intractable models in \( \mathbb{R}^d \) with unverifiable heavy tails; \( d \geq 1 \).

Two probability models with heavy tails, the Cauchy (\( \theta, \theta \)) and the Lognormal (\( \theta, \sigma \)), with unknown parameter, \( \theta \), are used initially in the simulations for various sample sizes; \( \sigma \) is assumed known. The results for both models indicate negligible empirical bias for \( \hat{\theta}_{MKD}, \hat{\theta}_{MEKD} \) and \( \hat{\theta}_{MWD} \). However, large bias is introduced in all repetitions of the simulations using the empirical average, \( \hat{\theta}_{MEWD} \), due to the right skewed distribution of \( \hat{\theta}_{MEWD} \), that is more concentrated in the smaller values of the parameter space \( \Theta \), unlike the distribution of \( \hat{\theta}_{MEKD} \). The empirical SDs of \( \hat{\theta}_{MKD} \) and \( \hat{\theta}_{MEKD} \), for sample size \( n \geq 300 \), always smaller, respectively, than those of \( \hat{\theta}_{MWD} \) and \( \hat{\theta}_{MEWD} \), and the corresponding SDs ratios decrease to zero as \( n \) increases. Thus, the empirical relative efficiencies of \( \theta_{MKD} \) with respect to \( \theta_{MWD} \), and of \( \theta_{MEKD} \) with respect to \( \theta_{MEWD} \), improve more and more as \( n \) increases. Similar simulations results for \( \theta_{MKD}, \theta_{MWD}, \theta_{MEKD} \) and \( \theta_{MEWD} \) are observed for the heavy tail, univariate stable models with index of stability \( \alpha \in (1, 2) \); for these models, only the first moment exists and \( W_1 \) is well defined. Under mild assumptions, the asymptotic distributions of \( n^{1/2} (\hat{\theta}_{MKD} - \theta) \) and \( n^{1/2} (\hat{\theta}_{MEKD} - \theta) \) are derived, and the ratio of the mean squared errors, \( n^{1/2} (\hat{\theta}_{MKD} - \theta)^2 / n^{1/2} (\hat{\theta}_{MEKD} - \theta)^2 \), is for large \( n \) proportional to \( [n^{1/2} \theta / \sqrt{n}]^2 \), which converges to zero when \( n \) increases to infinity. Similar results are expected to hold for \( \hat{\theta}_{MKD} \) and \( \hat{\theta}_{MWD} \), but also for \( \hat{\theta}_{MEKD} \) and \( \hat{\theta}_{MEWD} \).

\( \hat{\theta}_{n,MWD} \) is usually studied under restrictive assumptions for which the asymptotic distribution of \( \sqrt{n}(\hat{\theta}_{n,MWD} - \theta) \) is obtained; see, e.g., Bassetti and Regazzini (2006b, Proposition 4.1) and Bernton et al. (2019a, Theorem 2.3). One assumption used is \( \int_0^\infty \sqrt{\Phi(|X| - r)} dr < \infty \), which is equivalent to \( \int_{-\infty}^{\infty} \sqrt{g(x)} (1 - F_X(x)) dx < \infty \), and both imply \( E(X^2) < \infty; \) \( X \) is random variable with the underlying model. Del Barrio et al. (1999, p. 1012, lines 9–14) observe however that there are several models with finite 2nd moments for which this assumption does not hold: “…the previous theorem is far from covering the basic case \( E(X^2) < \infty \).” This indicates also the limitations on the applicability of the \( \sqrt{n} \)-rate of convergence for \( \hat{\theta}_{n,MWD} \) in models with finite second moment, unlike \( \hat{\theta}_{n,MEKD} \).

The findings indicate also that a coarsened \( W_1 \)-approximate posterior (Bernton et al. 2019b) will be less concentrated at \( \theta \) for heavy tail models. The reason is that when one of \( \hat{\mu}_n \)‘s summands with the observed data depends on an outlier, synthetic data, \( \hat{\mu}_n^* \), obtained from a model with parameter \( \theta^* \) far from \( \theta \) is included in the \( \epsilon \)-Wasserstein neighborhood, \( N_\epsilon(\hat{\mu}_n) \), with center, \( \hat{\mu}_n \), as explained in Sect. 4 for \( \theta_{MWD} ; \epsilon > 0 \). This makes the approximate posterior less concentrated at \( \theta \). When \( \hat{\mu}_n \)’s summands do not depend on an outlier, the number of \( \theta^* \)-values in the \( W_1 \)-approximate posterior is larger than those obtained via other distances, e.g., Total Variation, Hellinger, Kolmogorov and \( L_2 \)-distances, since the Wasserstein distance takes into consideration the underlying geometry of the space and brings probability models closer (Wasserman 2019, p. 1).

The distances and the definitions of stable random variables and models are in Sect. 2. The estimates, the simulation method and the numerical results appear in Sect. 3. Justifications of the findings are in Sect. 4, followed by a conclusion. The reader may proceed directly to Figs. 1, 2, 3, 4, 5 and

\(^{1}\) Have tails heavier than those of the exponential model.
with the histograms and compare their supports obtained with the Kolmogorov and the Wasserstein distances.

2 Distances-Stable random variables and models

Definition 2.1 For cumulative distribution functions $F$, $G$ in $\mathbb{R}^d$, the Kolmogorov distance

$$d_K(F, G) = \sup_{y \in \mathbb{R}^d} |F(y) - G(y)|.$$  \hspace{1cm} (1)

Definition 2.2 For any sample $U = (U_1, \ldots, U_n)$ of random vectors in $\mathbb{R}^d$, $n\hat{F}_n(u)$ denotes the number of $U_i$’s with all their components smaller or equal to the corresponding components of $u(\in \mathbb{R}^d)$. $\hat{F}_n$ is denoted by $\hat{F}_n$ and is the empirical cumulative distribution function (c.d.f.) of $U$; $\hat{F}_n$ denotes the empirical c.d.f. of $U^*$.

The empirical distribution,

$$\hat{\mu}_n = \hat{\mu}_{n,U} = \frac{1}{n} \sum_{i=1}^{n} \delta_{U_i},$$  \hspace{1cm} (2)

with $\delta_{U_i}$ Dirac distribution, $i = 1, \ldots, n$; $\hat{\mu}_n^*$ denotes the empirical distribution of $U^*$.

Definition 2.3 (e.g., see Villani 2008, Chapter 6. p. 105) Let $(\mathcal{X}, \rho)$ be a Polish metric space and let $p \in [1, \infty)$. For any two probability measures, $\mu$ and $\mu^*$ on $\mathcal{X}$, let $\Pi(\mu, \mu^*)$ denote all joint probabilities $\pi$ in $\mathcal{X} \times \mathcal{X}$ that have marginals $\mu$ and $\mu^*$. The Wasserstein distance of order $p$ between $\mu$ and $\mu^*$ is

$$W_p(\mu, \mu^*) = \inf_{\pi \in \Pi(\mu, \mu^*)} \left[\int_{\mathcal{X} \times \mathcal{X}} \rho(x, y)^p d\pi(x, y)\right]^{1/p}$$

$$= \inf_{\pi \in \Pi(\mu, \mu^*)} \left\{\left[\mathbb{E}_\pi \rho(X, Y)^p\right]^{1/p}\right\},$$  \hspace{1cm} (3)

$E_\pi$ denotes expected value with respect to $\pi$.

According to Villani (2008, p. 106): “…, $W_p$ is still not a distance in the strict sense, because it might take the value $+\infty$, but otherwise it does satisfy the axioms of a distance, …”. This justifies the assumption for $\mathcal{X}$ to be compact, which often accompanies the use of $W_p$, or that when $\mathcal{X} = R$, $\hat{\rho}(x, y) = |x - y|$, the random variables $X, Y$ in Definition 2.3 have finite moments of order $p \geq 1$.

For $p = 1$ and $\mathcal{X} = R$, (3) is also called Kantorovich distance; see, e.g., Villani (2008, p. 120). If $\hat{\rho}(x, y) = |x - y|$, $\hat{\mu}_n$ is the empirical distribution of i.i.d. r.v.s. $X_1, \ldots, X_n$ from $\mu_0$ and $\hat{\mu}_n^*$ is the empirical distribution of i.i.d. r.v.s. $X_1^*, \ldots, X_n^*$ from $\mu_*$, then

$$W_1(\hat{\mu}_n, \hat{\mu}_n^*) = n^{-1} \sum_{i=1}^{n} |X_{(i)} - X_{(i)}^*|.$$  \hspace{1cm} (4)

with $X_{(i)}, X_{(i)}^*, i = 1, \ldots, n$, denote order statistics.

Remark 2.1 If $\mu_\theta$ has unbounded support, then $W_1(\hat{\mu}_n, \hat{\mu}_n^*)$ is unbounded.

The distribution of stable random variable and stable distribution (or model) from Nolan (2020, Definition 1.3) is used.

Definition 2.4 A random variable $X$ is stable if and only if $X$ has the same distribution as the random variable $aV + b$, with $a \neq 0$ and $b \in \mathbb{R}$, and the random variable $V$ has characteristic function

$$Ee^{itV} = \begin{cases} \exp(-|t|^\alpha \{1 - \beta \tan \frac{\pi \alpha}{2} (\text{sign}(t))\}), & \text{if } \alpha \neq 1; \\
\exp(-|t|[1 + i \beta \frac{2}{\pi} (\text{sign}(t)) \log |t|]), & \text{if } \alpha = 1,
\end{cases}$$

where $0 < \alpha \leq 2$, $-1 \leq \beta \leq 1$, $t \in \mathbb{R}, i = \sqrt{-1}$ and

$$\text{sign}(t) = \begin{cases} 1, & \text{if } t > 0; \\
0, & \text{if } t = 0; \\
-1, & \text{if } t < 0.
\end{cases}$$

The distribution of $X$ is symmetric around 0 when $\beta = b = 0$.

Definition 2.4 indicates that four parameters are needed to define a general univariate stable distribution: an index of stability or characteristic exponent, $\alpha \in (0, 2)$, skewness parameter, $\beta \in [-1, 1]$, a scale parameter $\gamma(\geq 0)$ and a location parameter, $\delta \in \mathbb{R}$. We use herein this parametrisation that is suggested for estimation problems and the distribution is denoted by $S(\alpha, \beta, \gamma, \delta; pm = 0)$; “$pm$” determines the parametrisation number. There are several such parametrisations of stable distributions (Nolan 2020, p. 5) to tackle different problems. The characteristic function for $S(\alpha, \beta, \gamma, \delta; pm = 0)$ has the simplest form and is continuous in all parameters. The interested reader may consult Nolan (2020, Sect. 1.3) for more details on the parametrisations and the properties of stable random variables and models.

Among the models used in simulations herein, the stable with $\alpha = .5$ and $\beta = 0$, and the Cauchy have no first moment, the stable with $\alpha = 1.1$ and $\beta = 0$ has first moment and the Lognormal has all moments. These are all heavy tails models. The Normal and Uniform have all moments and are not heavy tail models.
3 Simulation results

3.1 \( \hat{\theta}_{MKD}, \hat{\theta}_{MWD} \) and their empirical relative efficiency

We start with the description of the classical Wolfowitz (1957) MDE, \( \hat{\theta}_{n,MKD} \), using the Kolmogorov distance, \( d_K \), as for known and tractable models.

For the MDEs herein exist, see, e.g., Yatracos (1985, p. 769). There may be several parameter values where the minimum in (5) is achieved and then their average is reported as \( \hat{\theta}_{n,MKD} \). Replacing in (5), \( d_K \) by \( W_p, \hat{F}_n \) by \( \hat{\mu}_n \) and \( F_0 \) by \( \mu_\theta \), estimate \( \tilde{\theta}_{n,MWD} \) is obtained.

With intractable or unknown models we cannot use \( F_0 \) in (5), so we use instead the empirical c.d.f., \( \hat{F}_{n,s} \), obtained with a sample drawn from the sampler and input parameter \( s(\in \Theta) \); see, e.g., Yatracos (2021). The luxury of having a sampler allows to repeat the process by drawing \( N_{rep} \) samples with input \( s \), and then calculate the

\[
\hat{\theta}_{n,MKD} = \arg \min_{s \in \Theta} \left\{ d_K (\hat{F}_n, F_s) \right\}; \tag{5}
\]

\[
\hat{\theta}_{n,MWD} = \arg \min_{s \in \Theta} \left\{ d_K (\hat{F}_n, \hat{F}_{n,s,i}) \right\}; \tag{7}
\]

For sample size \( n \), the metric space \( \Theta \) in the simulations is covered by \( d_\theta \)-balls of radius \( \epsilon_\theta \) with centers \( \Theta_n \); \( \epsilon_\theta_\theta > 0 \). We repeat the process leading to (6) for all \( s \) from \( \Theta_n \). There are several minima obtained via (6) for \( s \in \Theta_n \) to compare, and the Minimum Distance Estimate (MDE) is the s-value achieving the global minimum, \( \hat{\theta}_{MKD} \) and \( \hat{\theta}_{MWD} \) obtained for the Normal and the Uniform models.

The described experiment is repeated \( N \) times in order to report for sample size \( n \) the proportion of \( \hat{\theta}_{MKD} \) closer to \( \theta \) than \( \hat{\theta}_{MWD} \), as well as the minimum and the maximum of the \( N \) ratios, SDK/SDW, in the interval [\( \min(\text{SDK/SDW}), \max(\text{SDK/SDW}) \)].

R-functions ks.test and wasserstein1d are used in all the simulations. Results follow in Table 1 for the Cauchy, Uniform \( (\theta = 5, \sigma = \theta = 5) \) and Lognormal \( (\theta = 5, \sigma = 5) \) models, with \( N = 100 \) and sample sizes \( n = 100, 200, 300, 600, 1000, 2000, 5000, 10000 \). For every \( n \), we use NSIEVE=M=50 and \( N_{rep} = 100 \). The results indicate that, as \( n \) increases, the \( \hat{\theta}_{MKD} \) distribution concentrates faster around its mean than the \( \hat{\theta}_{MWD} \) distribution. Simply observe the evolution of the ends of intervals, [\( \min(\text{SDK/SDW}), \max(\text{SDK/SDW}) \)], with both upper and lower bound decreasing to zero as \( n \) increases. In addition, the percentage of \( \hat{\theta}_{MKD} \) closer to \( \theta \) becomes larger than that of \( \hat{\theta}_{MWD} \) and increases with \( n \). In practice, one of the MDEs is used and the larger SDW may lead to a less accurate \( \hat{\theta}_{MWD} \).

Simulation results for the Normal \( (\theta = 5, \sigma = \theta = 5) \) and Uniform \( (3, \theta = 5) \) follow in Table 2 for \( n = 1000 \) only and all the other specifications of the experiment unchanged. For the Normal, \( \hat{\theta}_{MKD} \) is closer to \( \theta \) than \( \hat{\theta}_{MWD} \) in 58% of the repetitions and SDK is smaller than SDW in 36% of the repetitions. For the Uniform, \( \hat{\theta}_{MWD} \) is always closer to \( \theta \) than \( \hat{\theta}_{MKD} \), but SDK is always smaller than SDW. For the Normal model the lower and upper bounds on the ratio SDK/SDW are, respectively, smaller and larger than 1, and one is nearly inverse of the other. For both models the differences between \( \hat{\theta}_{MKD} \) and \( \hat{\theta}_{MWD} \) and also SDK and SDW are at the second decimal. Thus, in practice, there is no difference between \( \hat{\theta}_{MKD} \) and \( \hat{\theta}_{MWD} \) obtained for the Normal and the Uniform models.

3.2 \( \hat{\theta}_{MEKD}, \hat{\theta}_{MEWD} \) and their empirical relative efficiency

The Minimum Expected Distance estimate was introduced for intractable models; see, e.g., Bernton et al. (2019a). For the Kolmogorov distance, the average of \( d_K (\hat{F}_n, \hat{F}_{n,s,i}, i = 1, \ldots, N_{rep}) \), in (6) is obtained for each \( s \in \Theta_n \), and

\[
\hat{\theta}_{MEKD} = \arg \min_{s \in \Theta_n} \left\{ N_{rep}^{-1} \sum_{i=1}^{N_{rep}} d_K (\hat{F}_n, \hat{F}_{n,s,i}) \right\}; \tag{8}
\]

The Min MEKD is similarly obtained. The data used for \( \hat{\theta}_{MEKD} \) and \( \hat{\theta}_{MEWD} \) is used again. Results for the Cauchy and the Lognormal models are in Table 3 and are similar to the results in Table 1 for the SDs but, in
\[ \gamma = \alpha \] for the Normal, for which \( \beta \) is unknown in parameter space \( \Theta = [3, 8] \).

The results in Table 1 for \( n = 1000 \) only. For the Normal, \( \hat{\theta}_{MKD} \) is closer to \( \theta \) than \( \hat{\theta}_{MEWD} \) in 43% of the repetitions and SDEK is always larger than SDEW. For the Uniform, \( \hat{\theta}_{MKD} \) is closer to \( \theta \) than \( \hat{\theta}_{MEWD} \) in 32% of the repetitions and SDEK is smaller than SDEW in 89% of the repetitions. For both models the difference between \( \hat{\theta}_{MEWD} \) and \( \hat{\theta}_{MEKD} \) is at the second decimal and the difference in the SDs is at the second decimal for the Normal and the third decimal for the Uniform. Thus, in practice, there is no difference between \( \hat{\theta}_{MKD} \) and \( \hat{\theta}_{MEWD} \) for the Normal and the Uniform models.

| \( n \) | \( \hat{\theta}_{MKD} > \hat{\theta}_{MEWD} \) | \( \hat{\theta}_{MKD} > \hat{\theta}_{MEWD} \) | \( \hat{\theta}_{MKD} > \hat{\theta}_{MEWD} \) |
|-------|--------------------------------|--------------------------------|--------------------------------|
| 100   | 57 [.47, .89]                  | 75 [.37, .75]                 | \_                             |
| 200   | 73 [.38, .81]                  | 73 [.3, .59]                  | \_                             |
| 300   | 78 [.31, .78]                  | 72 [.28, .54]                 | \_                             |
| 600   | 85 [.27, .65]                  | 76 [.23, .47]                 | \_                             |
| 1000  | 79 [.21, .55]                  | 85 [.19, .38]                 | \_                             |
| 2000  | 82 [.18, .43]                  | 84 [.16, .34]                 | \_                             |
| 5000  | 88 [.13, .37]                  | 90 [.12, .28]                 | \_                             |
| 10,000| 92 [.09, .3]                   | 93 [.09, .22]                 | \_                             |

In the Lognormal, \( \sigma \) is assumed known. \( \hat{\theta}_{MKD} \) and \( \hat{\theta}_{MEWD} \) are averages of \( \hat{\theta}_{MKD} \) and \( \hat{\theta}_{MEWD} \). “\( \hat{\theta}_{MKD} > \hat{\theta}_{MEWD} \)” denotes the % of \( \hat{\theta}_{MKD} \) closer to \( \theta \) than \( \hat{\theta}_{MEWD} \). SDK = SD of \( \hat{\theta}_{MKD} \) in \( \hat{\theta}_{MKD} \). SDW = SD of \( \hat{\theta}_{MEWD} \) in \( \hat{\theta}_{MEWD} \). Results based on \( N = 100 \) repetitions.

\[ \hat{\theta}_{MKD} > \hat{\theta}_{MEWD} \] indicates.

\[ \hat{\theta}_{MKD} > \hat{\theta}_{MEWD} \] denote the % of \( \hat{\theta}_{MKD} \) closer to \( \theta \) than \( \hat{\theta}_{MEWD} \). SDK = SD of \( \hat{\theta}_{MKD} \) in \( \hat{\theta}_{MKD} \). SDW = SD of \( \hat{\theta}_{MEWD} \) in \( \hat{\theta}_{MEWD} \). Results based on \( N = 100 \) repetitions.

The results for the Normal and Uniform follow in Table 4 for \( n = 1000 \) only. For the Normal, \( \hat{\theta}_{MEKD} \) is closer to \( \theta \) than \( \hat{\theta}_{MEWD} \) in 43% of the repetitions and SDEK is always larger than SDEW. For the Uniform, \( \hat{\theta}_{MEKD} \) is closer to \( \theta \) than \( \hat{\theta}_{MEWD} \) in 32% of the repetitions and SDEK is smaller than SDEW in 89% of the repetitions. For both models the difference between \( \hat{\theta}_{MEKD} \) and \( \hat{\theta}_{MEWD} \) is at the second decimal and the difference in the SDs is at the second decimal for the Normal and the third decimal for the Uniform. Thus, in practice, there is no difference between \( \hat{\theta}_{MEKD} \) and \( \hat{\theta}_{MEWD} \) for the Normal and the Uniform models.

### 3.3 \( \hat{\theta}_{MKD}, \hat{\theta}_{MEKD}, \hat{\theta}_{MKD} \) and \( \hat{\theta}_{MEWD} \) for Stable models

The estimates introduced in the previous sub-sections are now compared in simulations for the stable model, \( S(\alpha, \beta = 0, \gamma, \delta; \text{pm} = 0) \). The R-function \( rstable \) is used from the library \( \text{stabledist} \) to generate data with parameter values \( \gamma = \delta = \theta = 5; \alpha = .5 \) and \( \alpha = 1.1 \) with tails heavier, respectively, than those of the Cauchy, for which \( \alpha = 1 \), and the Normal, for which \( \alpha = 2 \). \( \theta \) is the unknown parameter to be estimated and \( \Theta = [3, 8] \). The results appear in Tables 5 and 6, with the notation described in Tables 1 and 3. \( \hat{\theta}_{MKD} \) and \( \hat{\theta}_{MKD} \) improve more \( \hat{\theta}_{MKD} \) and \( \hat{\theta}_{MEWD} \), respectively, for the heavier tails model with \( \alpha = .5 \), as the % of smaller bias for \( \hat{\theta}_{MKD} \) and the intervals of \( \text{SDK/SDW} \) indicate. Both models have no closed form expressions for their densities and it is not assumed known whether the first moment of each model exists.

### 3.4 Histograms

The histograms of \( \hat{\theta}_{MKD}, \hat{\theta}_{MEKD}, \hat{\theta}_{MEWD} \) and \( \hat{\theta}_{MEWD} \) for \( n = 1000 \) are presented in Figs. 1, 2, 3, 4, 5 and 6 for the Cauchy, the Lognormal, the Normal, the Uniform and Stable models with \( \alpha = .05, 1.1 \). The estimates in these Figures are denoted by their indices, MKD, MWD, MEKD and MEWD. For the models with heavy tails, the distributions of \( \hat{\theta}_{MEWD} \) and \( \hat{\theta}_{MEKD} \) are, respectively, more spread than those of \( \hat{\theta}_{MKD} \) and \( \hat{\theta}_{MEKD} \), and those for \( \hat{\theta}_{MEWD} \) are skewed. Most disturbing is the finding for the Lognormal model for which all moments exist but do not determine the model. Similar are the findings in histograms for the Cauchy and Lognormal models and \( n = 10000, 20000, 50000 \). \( \hat{\theta}_{MKD} \) and \( \hat{\theta}_{MEWD} \) perform best for the Normal model, but will deteriorate in high dimension due to the curse of dimensionality affecting the Wasserstein distance.

From these histograms it is observed, for models with heavy tails, that the Wasserstein distance allows realizations of \( \hat{\theta}_{MKD} \) to be more distant from \( \theta = 5 \) than realizations of \( \hat{\theta}_{MKD} \) and \( \hat{\theta}_{MEKD} \). This phenomenon is explained in the next section.
Table 3 Comparisons of $\hat{\theta}_{MEKD}$ and $\hat{\theta}_{MEWD}$. $\theta$ is unknown in parameter space $\Theta = [3,8]$

| $n$ | $\hat{\theta}_{MEKD} > \hat{\theta}_{MEWD}$ (%) | SDEK/SDEW | $\hat{\theta}_{MEKD} > \hat{\theta}_{MEWD}$ (%) | SDEK/SDEW |
|-----|---------------------------------|-----------|---------------------------------|-----------|
| 100 | 100                             | [.45,.98] | 100                             | [.44,1.7] |
| 200 | 100                             | [.35,.81] | 100                             | [.31,1.05] |
| 300 | 100                             | [.25,.58] | 100                             | [.26,.95] |
| 600 | 100                             | [.21,.41] | 100                             | [.16,.59] |
| 1000| 100                             | [.17,.33] | 100                             | [.13,.43] |
| 2000| 100                             | [.13,.30] | 100                             | [.1,.34] |
| 5000| 100                             | [.08,.19] | 100                             | [.06,.16] |
| 10,000| 100                         | [.07,.15] | 100                             | [.05,.12] |

Table 4 Comparisons of $\hat{\theta}_{MEKD}$ and $\hat{\theta}_{MEWD}$

| $n$ | $\hat{\theta}_{MEKD} > \hat{\theta}_{MEWD}$ (%) | SDEK/SDEW | $\hat{\theta}_{MEKD} > \hat{\theta}_{MEWD}$ (%) | SDEK/SDEW |
|-----|---------------------------------|-----------|---------------------------------|-----------|
| 1000| 43%                             | [1.03,1.42] | 32%                             | [.85,1.07] |

Table 5 Comparisons of $\hat{\theta}_{MKD}$ and $\hat{\theta}_{MW D}$ for Stable models

| $n$ | $\hat{\theta}_{MKD} > \hat{\theta}_{MW D}$ (%) | SDK/SDW | $\hat{\theta}_{MKD} > \hat{\theta}_{MW D}$ (%) | SDK/SDW |
|-----|---------------------------------|---------|---------------------------------|---------|
| 300 | 99                              | [.12,.25] | 78                             | [.34,.80] |
| 1000| 100                             | [.07,.13] | 76                             | [.23,.64] |
| 5000| 100                             | [.03,.06] | 86                             | [.16,.49] |

Table 6 Comparisons of $\hat{\theta}_{MEKD}$ and $\hat{\theta}_{MEWD}$ for Stable models

| $n$ | $\hat{\theta}_{MEKD} > \hat{\theta}_{MEWD}$ (%) | SDEK/SDEW | $\hat{\theta}_{MEKD} > \hat{\theta}_{MEWD}$ (%) | SDEK/SDEW |
|-----|---------------------------------|-----------|---------------------------------|-----------|
| 300 | 100                             | [.14,.30] | 100                             | [.33,.62] |
| 1000| 100                             | [.07,.17] | 100                             | [.23,.39] |
| 5000| 100                             | [.04,.07] | 100                             | [.15,.29] |

4 Justification of the empirical findings

The disturbing findings for $\hat{\theta}_{MW D}$ in univariate data simulations are due to non-robustness of the Wasserstein distance, i.e., $W_1(\hat{\mu}_n, \hat{\mu}^*_n) = n^{-1} \sum_{i=1}^n |X_i - X^*_i|$, increases to infinity for unbounded observations since, as seen in (4).

$$W_1(\hat{\mu}_n, \hat{\mu}^*_n) = n^{-1} \sum_{i=1}^n |X_i - X^*_i|,$$  \hspace{1cm} (9)

is unbounded for fixed $n$; $X_1, X^*_1, \ldots, X_n, X^*_n, i = 1, \ldots, n$, denote order statistics. If $\hat{\mu}_n$ and $\hat{\mu}^*_n$ are obtained, respectively, from $F_\theta$ and $F_n$, with $s$ element of the sieve $\Theta_n$ used to obtain $\hat{\theta}_{MW D}$, a very extreme observation from $F_\theta$ in $\hat{\mu}_n$ will cause a large increase in the value of the $W_1$-distance (9) for $s$ near $\theta$ and, at least for location models, an element $s^*$ from $\Theta_n$ far from $\theta$ will be $\hat{\theta}_{MW D}$. Elements like $s^*$ are observed for models with heavy tails in the extremes of the supports of the $\hat{\theta}_{MW D}$-histograms; compare with the extremes of the supports of the $\hat{\theta}_{MKD}$-histograms in Figs. 1, 2, 5 and 6.
Fig. 1  Histograms of $\hat{\theta}_{MKD}, \hat{\theta}_{MWD}, \hat{\theta}_{MEKD}, \hat{\theta}_{MEWD}$ denoted by their indices for the Cauchy model and sample size $n = 1000$, with $\theta = 5$ and parameter space $\Theta = [3, 8]$. Compare the supports of the histograms.

This phenomenon influences also the $\hat{\theta}_{MEWD}$-histograms. The corresponding Kolmogorov distance, $d_K(\hat{F}_n, \hat{F}^*_n)$, is bounded, an extreme observation from $F_\theta$ may alter its value for $s$ in $\Theta_n$ near $\theta$ by no more than $1/n$ and, e.g., for location models, $\hat{\theta}_{MKD}$ is not going to be very distant from $\theta$.

To realize that extreme observations occur with relatively high probability for heavy tail models, note that in samples from the Cauchy and the Normal models with location and scale parameters $(\theta, \sigma) = (0, 1)$, there will be (on average) approximately 100 times more values above 3 in the Cauchy model than in the Normal model (Nolan 2020, p. 3).

The convergence of the ratio $SDK/SDW$ to 0 as $n$ increases in Table 5 for $\alpha = 1.1$ is confirmed theoretically for any stable model $F_\theta$ with index $1 < \alpha < 2$, using the ratio of the mean squared errors since the bias is negligible. Results in Del Barrio et al. (1999, Theorems 1.1 (b) and 2.3) are used in the derivations; $\theta \in \Theta (\subset \mathbb{R})$. For $\hat{\theta}_{n, MKD}, \hat{\theta}_{n, MWD}$ obtained as in (5), the asymptotic distributions of $k_n(\hat{\theta}_{n, MKD} - \theta)$ and $k_n(\hat{\theta}_{n, MWD} - \theta)$ are determined using the approach in Pollard (1980). Under moment conditions, the mean squared errors $E(\hat{\theta}_{n, MKD} - \theta)^2$ and $E(\hat{\theta}_{n, MWD} - \theta)^2$ are approximated using the corresponding limit distributions,
Fig. 3 Histograms of \( \hat{\theta}_{MKD}, \hat{\theta}_{MWD}, \hat{\theta}_{MEKD}, \hat{\theta}_{MEWD} \) denoted by their indices for the Normal model and sample size \( n = 1000 \), with \( \theta = 5 \) and parameter space \( \Theta = [3,8] \). Compare the supports of the histograms.

Fig. 4 Histograms of \( \hat{\theta}_{MKD}, \hat{\theta}_{MWD}, \hat{\theta}_{MEKD}, \hat{\theta}_{MEWD} \) denoted by their indices for the Uniform model and sample size \( n = 1000 \), with \( \theta = 5 \) and parameter space \( \Theta = [3,8] \). Compare the supports of the histograms.

and the ratio \( E(\hat{\theta}_{n,MKD} - \theta)^2 / E(\hat{\theta}_{n,MWD} - \theta)^2 \) is proportional to \( (\frac{k_n}{n})^2 \) that converges to 0 as \( n \) increases since \( k_n = n^{1 - \frac{1}{2}} \) and \( k_n^* = \sqrt{n} \). Under mild assumptions, \( E(\hat{\theta}_{MKD} - \theta)^2 / E(\hat{\theta}_{MWD} - \theta)^2 \) will also converge to 0 as \( n \) increases, as observed in simulations for heavy tail models. Similar results are expected to hold for \( \hat{\theta}_{MEKD} \) and \( \hat{\theta}_{MEWD} \) with faster convergence of \( E(\hat{\theta}_{MEKD} - \theta)^2 / E(\hat{\theta}_{MEWD} - \theta)^2 \) to zero as \( n \) increases, because of the additional bias introduced by \( \hat{\theta}_{MEWD} \).

A brief sketch for the derivation of \( k_n \) and \( k_n^* \) follows, assuming parameter identifiability and that \( \hat{\theta}_{n,MWD} \) and \( \hat{\theta}_{n,MKD} \) are unique, consistent and measurable. Assumptions for these conditions to hold appear in Pollard (1980), where \( \{F_s; s \in \Theta\} \) is subset of a normed space, \( \mathcal{Y} \), that includes the sequence of empirical c.d.f.s \( \{\hat{F}_n\} \). Bernton et al. (2019a, Sects. 2.1.1 and 2.1.2) followed the approach under assumptions for which the asymptotic distribution of \( \hat{\theta}_{n,MWD} \) is obtained with scale factor \( k_n = \sqrt{n} \). For the stable model...
Fig. 5 Histograms of $\hat{\theta}_{MKD}, \hat{\theta}_{MWD}, \hat{\theta}_{MEKD}, \hat{\theta}_{MEWD}$ denoted by their indices for the Stable model, $S(5, 0, 5; 0)$ and sample size $n = 1000$, with $\theta = 5$ and parameter space $\Theta = [3, 8]$. Compare the supports of the histograms.

Fig. 6 Histograms of $\hat{\theta}_{MKD}, \hat{\theta}_{MWD}, \hat{\theta}_{MEKD}, \hat{\theta}_{MEWD}$ denoted by their indices for the Stable model, $S(1.1, 0, 5; 0)$ and sample size $n = 1000$, with $\theta = 5$ and parameter space $\Theta = [3, 8]$. Compare the supports of the histograms.

with $1 < \alpha < 2$, $n^{1/\alpha}$ replaces $\sqrt{n}$ in the derivations in these papers. More precisely, since for $\alpha > 1$ the first moment of the underlying model exists,

$$W_1(\mu_{\theta}, \mu_s) = \int_{-\infty}^{\infty} |F_s(t) - F_\theta(t)| dt = ||F_s - F_\theta||,$$  \hspace{1cm} (10)

with $F_s$ and $F_\theta$ c.d.f.s of $\mu_s$ and $\mu_\theta$, respectively. From (10), for the derivations related to $\hat{\theta}_{n,MWD}$, the normed space $Y$ in the context of Pollard (1980) is $L_1(R)$. From Del Barrio et al. (1999), Theorem 1.1, it holds

$$\frac{n}{n^{1/\alpha}} \int_{-\infty}^{\infty} |\hat{F}_n(t) - F_\theta(t)| dt = \frac{n}{n^{1/\alpha}} ||\hat{F}_n - F_\theta|| \xrightarrow{d} G_1,$$ \hspace{1cm} (11)

and from Eqs. (2.21), (2.23) and Theorem 2.3, Eq. (2.7), for the process $\{\hat{F}_n(t)\}$ it holds,

$$\frac{n}{n^{1/\alpha}} [\hat{F}_n - F_\theta] \xrightarrow{w} G_2 \text{ in } L_1(R);$$ \hspace{1cm} (12)

$G_1$ and $G_2$ are determined. The notation for convergence in distribution is “$\xrightarrow{d}$” for random variables and “$\xrightarrow{w}$” for probability measures of random processes, as in Del Barrio.
et al. (1999). (11) and (12) have been obtained using important results in Lawniczak (1983) and the scaling factor in (11) appears clearly in Omelchenko (2012, Theorem 14).

Fundamental in the Pollard (1980, 4.2 Theorem) approach we follow are (11), (12) and the differentiability assumption of the function \( \theta \rightarrow F_\theta \) for the norm in (10), i.e.,

\[
|| F_\theta - F_0 || = o(|| s - \theta ||) \tag{13}
\]

near \( \theta \); \( D \in L_1(R) \). This leads, for \( s \) near \( \theta \), to the differential approximation of \( n^{1-\frac{1}{2}} || \hat{F}_n - F_\theta || \) by \( n^{1-\frac{1}{2}} (\hat{F}_n - F_\theta) - n^{1-\frac{1}{2}} (s-\theta) D \) with error of small order. Then, from (12) and for large \( n \), the distribution of \( n^{1-\frac{1}{2}} \inf_{s \in \Theta} || \hat{F}_n - F_\theta || \) is close to that of \( \inf_{t \in R} || G_2 - tD || \) when the minimum occurs at distance of order \( O_P(n^{-1-\frac{1}{2}}) \) from \( \theta \); it is assumed for simplicity there is a unique \( t \) achieving almost surely the \( \inf_{t \in R} \).

From Pollard (1980, Sect. 7, p. 64, l. 2), \( n^{1-\frac{1}{2}} (\hat{\theta}_{n,MDE} - \theta) \) converges weakly to the distribution of a functional \( T(G_2) \) and \( k_n = n^{1-\frac{1}{2}} \). When the infimizer \( t \) is not unique the result still holds, as described in Pollard (1980, Sect. 7).

For Wolfowitz (1957) MDE, \( \hat{\theta}_{n,MDE} \), the rate of convergence, \( k_n^* = \sqrt{n} \), can be obtained as described in Pollard (1980, Sect. 7, p. 63, 64), under assumptions similar to those used to derive \( k_n \). The Dudley (1966, 1967) notion of weak convergence is used, which makes empirical distribution functions measurable in the space \( D[0,1] \), of right continuous functions on \( [0,1] \) with left-hand limits, equipped with sup-norm (i.e. \( d_K \)).

Then, Eqs. (11) and (12) hold with \( d_K \) instead of \( W_1 = || \cdot || \), with \( \sqrt{n} \) instead of \( n^{1-\frac{1}{2}} \) and with the corresponding known limit distributions (Pollard 1980, Sect. 3). If, in addition, the function \( \theta \rightarrow F_\theta \) is \( d_K \)-differentiable, satisfying (13) with \( d_K \) replacing \( || \cdot || \), and assumptions similar to those in the derivation of the scale factor, \( k_n \), for \( \hat{\theta}_{n,MDE} \) hold, then \( \sqrt{n}(\hat{\theta}_{n,MDE} - \theta) \) has an asymptotic distribution and \( k_n^* = \sqrt{n} \) (Pollard 1980, p. 64, l. 2).

5 Conclusion

The results herein for univariate models with heavy tails indicate that \( W_1 \)-MDE is improved by the Kolmogorov-MDE, and this is expected to hold for multivariate heavy tail models. Supplementary findings are that \( W_1 \)-MDE converges to \( \sqrt{n} \)-rate only for some of the models that have second moment, unlike the Kolmogorov-MDE, and that \( W_1 \)-approximate posteriors are less concentrated at \( \theta \). Combining these results with \( W_p \)-drawbacks presented from the literature such as the superiority of the kernel distance over \( W_1 \) in two-sample testing, \( W_p \)'s computational difficulty, the non-existence of tools to do inference except for special cases and the slower \( W_p \)-concentration of \( \hat{\mu}_n \) with high dimensional unbounded data, one concludes that there is no support for the use of \( W_p \) in statistical inference for Black Box and intractable models with unverifiable tail heaviness. These models do not leave space for model specific, alternative versions of \( W_p \) to correct its drawbacks.

Acknowledgements

Many thanks are due to Professor Ajay Jasra, Editor, the AE and the referees, for their encouraging comments that led to an improvement in the last revision of the paper.

Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Bassetti, F., Regazzini, E.: Asymptotic properties and robustness of minimum dissimilarity estimators of location-scale parameters. Theory Probab. Appl. 50, 171–186 (2006)
Bassetti, F., Bodini, A., Regazzini, E.: On minimum Kantorovich distance estimators. Stat. Probab. Lett. 76, 1298–1302 (2006)
Bernton, E., Jacob, P.E., Gerbery, M., Robert, C.P.: On parameter estimation with the Wasserstein distance. Inf. Inference J. IMA 8, 657–676 (2019a)
Bernton, E., Jacob, P.E., Gerbery, M., Robert, C.P.: Approximate Bayesian computation with the Wasserstein distance. JRSS B 81, 235–269 (2019b)
Del Barrio, E., Giné, E., Matrán, C.: Central limit theorems for the Wasserstein distance between the empirical and true distributions. Ann. Prob. 27, 1009–1071 (1999)
Dudley, R.M.: Weak convergence of probabilities on non-separable metric spaces and empirical measures on Euclidean spaces. Ill. J. Math. 10, 109–126 (1966)
Dudley, R.M.: Measures on non-separable metric spaces. Ill. J. Math. 11, 449–453 (1967)
Dudley, R.M.: The speed of mean Glivenko–Cantelli convergence. Ann. Math. Stat. 40, 40–50 (1968)
Kantorovich, L.V.: On the Translocation of Masses. Dokl. Akad. Nauk SSSR 37(7–8), 227–229 (1942). (Translated in Journal of Mathematical Sciences, 133, 4 (2006))
Kiefer, J.: On large deviations of the empiric D. F. of vector chance variables and a law of the iterated logarithm. Pac. J. Math. 11, 649–660 (1961)
Kolouri, S., Park, S., Thorpe, M., Slepčev, D., Rohde, G.K.: Optimal mass transport: signal processing and machine-learning applications. IEEE Signal Process. Mag. 34(4), 43–59 (2017)
Lawniczak, A.: The Lévy–Lindeberg central limit theorem in Orlicz spaces. Proc. Am. Math. Soc. 89, 673–679 (1983)
Nolan, J.: Univariate Stable Distributions: Models for Heavy Tailed Data. Springer Series in Operations Research and Financial Engineering, Springer, Berlin (2020)
Omelchenko, V.: Behavior and convergence of Wasserstein metric in the framework of stable distributions. Bull. Czech Econom. Soc. 30, 124–138 (2012)

Pollard, D.: The minimum distance method of testing. Metrika 27, 43–70 (1980)

Sommerfeld, M., Munk, A.: Inference for empirical Wasserstein distances on finite spaces. JRSS B 80, 219–238 (2018)

Sriperumbudur, B.K., Fukumizu, K., Gretton, A., Scholkopf, B., Lanckriet, G.R.C.: On the empirical estimation of integral probability metrics. Electr. J. Stat. 6, 1550–1599 (2012)

Talagrand, M.: The transportation cost from the uniform measure to the empirical measure in dimension $\geq 3$. Ann. Prob. 22, 919–959 (1994)

Villani, C.: Optimal Transport, Old and New. https://cedricvillani.org/sites/dev/files/old_images/2012/08/preprint-1.pdf (2008)

Wasserman, L.: Optimal Transport and Wasserstein Distance. Lecture Notes on “Statistical Methods for Machine Learning”. https://www.stat.cmu.edu/~larry/=sml/Opt.pdf (2019)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.