# ROBUST PARAMETER ESTIMATION BASED ON THE K-DIVERGENCE

*Yair Sorek and Koby Todros*

Ben-Gurion University of the Negev

## ABSTRACT

In this paper we present a new divergence, called $\mathcal{K}$-divergence, that involves a weighted version of the hypothesized log-likelihood function. To down-weight low density areas, attributed to outliers, the corresponding weight function is a convolved version of the underlying density with a strictly positive smoothing "$\mathcal{K}$"ernel function parameterized by a bandwidth parameter. The resulting minimum $\mathcal{K}$-divergence estimator (M$\mathcal{K}$DE) operates by minimizing the empirical $\mathcal{K}$-divergence w.r.t. the vector parameter of interest. The M$\mathcal{K}$DE utilizes Parzen's non-parametric kernel density estimator, arising from the nature of the weight function, to suppress outliers. By proper selection of the kernel's bandwidth parameter we show that the M$\mathcal{K}$DE can gain enhanced estimation performance along with implementation simplicity as compared to other robust estimators.

***Index Terms—*** Divergences, estimation theory, robust statistics.

## 1. INTRODUCTION

The maximum likelihood estimator (MLE) [1], [2] is a well-established tool for multivariate estimation that operates by minimizing the empirical Kullback-Leibler divergence (KLD) [3] between the underlying probability distribution of the data and a hypothesized class of parametric distributions (the model) [4]. When the assumed probability model is correctly specified, i.e., it encompasses the true underlying distribution, the MLE is asymptotically efficient [1], [2]. However, when the hypothesized score-function is unbounded over the observation space, the MLE may be highly sensitive to small model misspecification inflicted by outliers [5], [6].

To overcome this limitation, several robust alternatives to the MLE have been proposed. The class of M-estimators [5]-[7] is a popular family of robust estimators that are generally applied to location-scatter models [8]-[12]. Another robust alternative has been recently developed in [13], [14] that operates by fitting a Gaussian model to a transformed version of the underlying probability distribution. This estimator can gain resilience against outliers while maintaining the implementation simplicity of the Gaussian MLE. However, it is mostly designated to location-scatter models.

To handle more flexible distributional models, such as Gaussian mixture model (GMM) [15], several robust MLE alternatives have been developed that utilize divergence measures other than the KLD. In [16]-[19] it was proposed to minimize the Hellinger distance [20] between the assumed parametric density function and Parzen's non-parametric kernel density estimator [21] of the underlying distribution. The rational is that low-density areas, corresponding to distant outliers, barely affect the distance, thus, leading to robust parameter estimation. Although the minimum Hellinger distance estimator (MHDE) may be successful in mitigating the effect of outliers [16], [22], it may have the following drawbacks. First, consistency of the kernel density estimator is required to guarantee a consistent parameter estimation at the true model, i.e., when the hypothesized parametric family is correctly specified. Therefore, the kernel's bandwidth parameter must vary with the sample size at a certain rate. Second, asymptotic efficiency at the true model, holds only under strong regularity conditions on the kernel function and the underlying distribution [20]. Additionally, the MHDE involves numerical integration over the observation space which can be computationally demanding and inaccurate in the multivariate case.

The Hellinger distance is a special case of the $\alpha$-divergence [23] that involves powered versions of the compared densities. Interestingly, in [24], [25] was shown that indirect $\alpha$-divergence, between some power transformed version of the underlying distribution and the hypothesized parametric family, can be successfully applied to robust parameter estimation that does not require integration and density estimation. The resulting minimum $\alpha$-divergence estimator (M$\alpha$DE) incorporates density power weights that suppress outliers. Nevertheless, while under the specific regression models considered in [24], [25], consistency of the parameter estimate is maintained under the power transform, this may not be the case in general.

The $\beta$-divergence [26]-[31] and the $\gamma$-divergence [32]-[34] are other disparity measures that also involve density powers to induce outlier resilince. Unlike the MHDE, the minimum $\beta$-divergence estimator (M$\beta$DE) and the minimum $\gamma$-divergence estimator (M$\gamma$DE) do not require non-parametric density estimation. However, they incorporate integration over powered versions of the assumed density. Under some flexible parametric probability models, such as GMM, the corresponding integral cannot be solved analytically when the power levels are fractional. In these cases, numerical integration is required which may be computationally demanding and inaccurate under a multivariate observation space.

**Main contribution:** In this paper, we propose a new divergence, called $\mathcal{K}$-divergence, that can gain simple implementation in flexible distributional models by avoiding the use of density powers. To induce outlier resilience, the $\mathcal{K}$-divergence involves a weighted version of the hypothesized log-likelihood function. In order to down-weight low density areas, attributed to outlying measurements, the corresponding weight function is a convolved version of underlying probability density function with a strictly positive smoothing "$\mathcal{K}$"ernel function parameterized by a bandwidth parameter. The resulting robust estimator is generated by minimizing an empirical version of the $\mathcal{K}$-divergence w.r.t. the vector parameter of interest. This estimator is referred here to as the minimum $\mathcal{K}$-divergence estimator (M$\mathcal{K}$DE). The M$\mathcal{K}$DE utilizes Parzen's non-parametric kernel density estimator to effectively suppress outliers. Here, Parzen's estimator arises from the empirical estimate of the weight function, as being the expectation of the translated kernel function w.r.t. the underlying distribution of the data.

In the paper we show that, under some regularity conditions, the M$\mathcal{K}$DE is consistent, asymptotically normal and unbiased w.r.t. the maximizer of the $\mathcal{K}$-divergence over the hypothesized parametric class. We obtain a closed form expression for the asymptotic mean-squared-error (MSE) matrix. Unlike the MHDE, consistency

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

of the non-parametric density estimator is not required to guarantee consistency of the parameter estimate at the true model. This is due to the fact that, unlike the MHDE, the M$\mathcal{K}$DE does not operate by fitting the hypothesized parametric distribution to the kernel density estimator, which only serves here as a data-weighting function. In the paper we show that asymptotic efficiency at the true model is attained when the bandwidth parameter approaches infinity. Additionally, we derive a condition on the kernel function that guarantees outlier resilience. In the paper we develop a data-driven procedure for selection of the kernel's bandwidth parameter. This procedure stochastically approximates the minimizer of a lower bound on the asymptotic weighted MSE (WMSE) at a nominal vector parameter.

The proposed M$\mathcal{K}$DE is illustrated for robust estimation of a two-component GMM. We show that the M$\mathcal{K}$DE outperforms the MLE and the minimum divergence estimators discussed above.

Lastly, we note that proofs for the theorems stated throughout the paper will be provided in the full length journal version.

## 2. THE $\mathcal{K}$-DIVERGENCE

In this section, we present the $\mathcal{K}$-divergence. Let $G$ and $F$ be probability distributions of a continuous random vector $\mathbf{x} \in \mathbb{R}^p$. Assume that $G$ and $F$ posses density functions, w.r.t. Lebesgue's measure $\lambda$, that are denoted by $g(\cdot)$ and $f(\cdot)$, respectively. The $\mathcal{K}$-divergence between $G$ and $F$ is defined as:

$$\mathcal{K}_h[G||F] \triangleq \mathbb{E}\left[\psi_G(\mathbf{x},h)\log\frac{g(\mathbf{x})}{f(\mathbf{x})};G\right] + \log\mathbb{E}\left[\psi_G(\mathbf{x},h);F\right], \tag{1}$$

where $\mathbb{E}[\cdot;P]$ denotes the statistical expectation w.r.t. some probability distribution $P$. The weight function $\psi_G(\cdot;\cdot)$ that appears in both summands is defined as:

$$\psi_G(\mathbf{r},h) \triangleq \frac{(K_h * g)(\mathbf{r})}{\mathbb{E}[(K_h * g)(\mathbf{x});G]}, \tag{2}$$

where $K_h(\mathbf{r}) \triangleq h^{-p}K(h^{-1}\mathbf{r})$, $K(\mathbf{r})$ is a strictly positive, bounded, continuous and integrable kernel function and $h \in \mathbb{R}_{++}$ is a bandwidth parameter. Without loss of generality, we shall assume that $\int_{\mathbb{R}^p} K(\mathbf{r})d\lambda(\mathbf{r}) = 1$. Here, $(K_h{*}g)(\mathbf{r}) \triangleq \int_{\mathbb{R}^p} K_h(\mathbf{r}-\boldsymbol{s})g(\boldsymbol{s})d\lambda(\boldsymbol{s})$ denotes the convolution operator between $K_h(\cdot)$ and $g(\cdot)$ evaluated at $\mathbf{r} \in \mathbb{R}^p$. The following theorem states a basic non-negativity property of the $\mathcal{K}$-divergence.

**Theorem 1.** *$\mathcal{K}_h[G||F] \geq 0$, where equality holds if and only if $G = F$.*

## 3. THE MINIMUM $\mathcal{K}$-DIVERGENCE ESTIMATOR

In this section, we derive the M$\mathcal{K}$DE, analyze its asymptotic performance and study its resilience against outliers.

### 3.1. Derivation of the M$\mathcal{K}$DE

Consider a parametric family of probability distributions $\{F_{\boldsymbol{\theta}}\}$, indexed by a vector parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^m$. Given a sequence of mutually independent samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ from a probability distribution $G$, which does not necessarily belong to $\{F_{\boldsymbol{\theta}}\}$, the minimum $\mathcal{K}$-divergence estimator is generated by minimizing the empirical version of the $\mathcal{K}$-divergence between $G$ and $F_{\boldsymbol{\theta}}$. Using (1) and (2), it can be shown that minimization of the empirical $\mathcal{K}$-divergence w.r.t. $\boldsymbol{\theta}$ amounts to maximization of the following objective:

$$\mathcal{J}_h(\boldsymbol{\theta}) \triangleq \sum\nolimits_{n=1}^{N} w(\mathbf{x}_n;h)\log f(\mathbf{x}_n;\boldsymbol{\theta}) - \log\hat{u}(\boldsymbol{\theta},h), \tag{3}$$

where the weight function $w(\mathbf{r};h) \triangleq \tilde{g}(\mathbf{r};h)/\sum_{m=1}^{N}\tilde{g}(\mathbf{x}_m;h)$, the function $\tilde{g}(\mathbf{r};h) \triangleq \hat{g}(\mathbf{r};h) - N^{-1}K_h(0)$ and

$$\hat{g}(\mathbf{r};h) \triangleq \frac{1}{N}\sum\nolimits_{m=1}^{N} K_h(\mathbf{r}-\mathbf{x}_m) \tag{4}$$

is Parzen's kernel density estimator [21] of $g(\mathbf{r})$. The function

$$\hat{u}(\boldsymbol{\theta},h) \triangleq \int_{\mathbb{R}^p} \hat{g}(\mathbf{r};h)f(\mathbf{r};\boldsymbol{\theta})d\lambda(\mathbf{r}). \tag{5}$$

Hence, the proposed M$\mathcal{K}$DE is given by:

$$\hat{\boldsymbol{\theta}}_h \triangleq \arg\max\nolimits_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{J}_h(\boldsymbol{\theta}). \tag{6}$$

**Remark 1.** *Note that $\hat{\boldsymbol{\theta}}_h$ approaches the MLE as $h \to \infty$. Also note that the log-likelihood realizations in (3) are weighted in accordance to the approximate underlying density of the data. This results in intrinsic suppression of outlying measurements, corresponding to low density areas. Additionally, one sees that unlike the $\beta$-divergence [26], [27] and the $\gamma$-divergence [32] based objectives, the integral term (5), in the considered objective (3), does not involve powers of the hypothesized parametric density $f(\cdot;\boldsymbol{\theta})$. In fact, it follows from (4) that the integral (5) can be computed when $\int_{\mathbb{R}^p} K_h(\mathbf{r}-\boldsymbol{s})f(\mathbf{r};\boldsymbol{\theta})d\lambda(\mathbf{r})$ has analytical solution. One can verify that in the GMM case, this integral is easily calculated when the kernel function $K(\cdot)$ belongs to the class of Gaussian shaped functions.*

### 3.2. Asymptotic performance analysis

Here, we utilize the theory of $U$-statistics [7] to study the asymptotic performance of the proposed M$\mathcal{K}$DE. Throughout the analysis we assume a symmetric kernel function in the density estimator (4).

In the following theorem, we show that $\hat{\boldsymbol{\theta}}_h$ (6) is a consistent estimator of $\boldsymbol{\theta}_h^*$ that represents the best fitting parameter, in the sense of minimum $\mathcal{K}$-divergence, i.e.,

$$\boldsymbol{\theta}_h^* \triangleq \arg\min\nolimits_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{K}_h[G||F_{\boldsymbol{\theta}}]. \tag{7}$$

**Theorem 2** (Consistency)**.** *Assume that the regularity conditions stated in [35, Sec. II] are satisfied. Then,*

$$\hat{\boldsymbol{\theta}}_h \xrightarrow[N\to\infty]{p} \boldsymbol{\theta}_h^*, \tag{8}$$

*where "$\xrightarrow{p}$" denotes convergence in probability [36].*

Next, we show that the M$\mathcal{K}$DE is asymptotically normal and unbiased. Furthermore, we derive a closed form expression of the asymptotic MSE matrix.

**Theorem 3** (Asymptotic normality and unbiasedness)**.** *Assume that the regularity conditions stated in [35, Sec. III] hold. Then,*

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h^*) \xrightarrow[N\to\infty]{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}_h^*,h)\right), \tag{9}$$

*where "$\xrightarrow{d}$" denotes convergence in distribution [36]. The covariance matrix in (9) takes the form:*

$$\boldsymbol{\Sigma}(\boldsymbol{\theta},h) = \mathbf{C}^{-1}(\boldsymbol{\theta},h)\mathbf{D}(\boldsymbol{\theta},h)\mathbf{C}^{-1}(\boldsymbol{\theta},h), \tag{10}$$

*where* $\mathbf{C}(\boldsymbol{\theta},h) \triangleq \mathbb{E}[\psi_G(\mathbf{x},h)\nabla^2_{\boldsymbol{\theta}}\log f(\mathbf{x};\boldsymbol{\theta});G] - \nabla^2_{\boldsymbol{\theta}}\log u(\boldsymbol{\theta},h)$,

$$\mathbf{D}(\boldsymbol{\theta},h) \triangleq \mathbb{E}[\mathbf{v}(\mathbf{x},\boldsymbol{\theta},h)\mathbf{v}^T(\mathbf{x},\boldsymbol{\theta},h);G],$$
$$\mathbf{v}(\mathbf{r},\boldsymbol{\theta},h) \triangleq \psi_G(\mathbf{r},h)\mathbf{c}(\mathbf{r},\boldsymbol{\theta},h) + \mathbf{d}(\mathbf{r},\boldsymbol{\theta},h) - \mathbf{z}(\mathbf{r},\boldsymbol{\theta},h),$$
$$\mathbf{c}(\mathbf{r},\boldsymbol{\theta},h) \triangleq \nabla_{\boldsymbol{\theta}}\log f(\mathbf{r};\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}}\log u(\boldsymbol{\theta},h),$$
$$\mathbf{d}(\mathbf{r},\boldsymbol{\theta},h) \triangleq \mathbb{E}[\mathbf{c}(\mathbf{x},\boldsymbol{\theta},h)\varphi_h(\mathbf{r}-\mathbf{x});G],$$
$$\mathbf{z}(\mathbf{r},\boldsymbol{\theta},h) \triangleq \frac{v(\mathbf{r},\boldsymbol{\theta},h)}{u(\boldsymbol{\theta},h)}\nabla_{\boldsymbol{\theta}}\log\frac{v(\mathbf{r},\boldsymbol{\theta},h)}{u(\boldsymbol{\theta},h)},$$
$$v(\mathbf{r},\boldsymbol{\theta},h) \triangleq (K_h * f)(\mathbf{r};\boldsymbol{\theta}), \quad u(\boldsymbol{\theta},h) \triangleq \mathbb{E}[(K_h * g)(\mathbf{x});F_{\boldsymbol{\theta}}]$$

*and* $\varphi_h(\mathbf{r}) \triangleq K_h(\mathbf{r})/\mathbb{E}[(K_h * g)(\mathbf{x});G]$

Theorem 3 implies that similarly to the standard MLE [1], [2], the proposed M$\mathcal{K}$DE converges at a rate of $1/\sqrt{N}$. By (9), the asymptotic MSE matrix of $\hat{\boldsymbol{\theta}}_h$ at $\boldsymbol{\theta}^*_h$ is given by:

$$\mathbf{R}(\boldsymbol{\theta}^*_h,h) \triangleq N^{-1}\boldsymbol{\Sigma}(\boldsymbol{\theta}^*_h,h). \tag{11}$$

**Remark 2** (Consistency and asymptotic efficiency)**.** *Note that when the hypothesized parametric family is correctly specified, we have that* $G = F_{\boldsymbol{\theta}_0}$, *where* $\boldsymbol{\theta}_0$ *is the true underlying vector parameter. In this case, when* $\{F_{\boldsymbol{\theta}}\}$ *is identifiable, it follows from Theorem 1, that* $\boldsymbol{\theta}^*_h = \boldsymbol{\theta}_0$ *is the unique mimimizer of the* $\mathcal{K}$*-divergence. Hence, when the regularity conditions stated in [35, Sec. II] are satisfied, it follows from Theorem 2 that the M*$\mathcal{K}$*DE is consistent at the true model for any fixed value of the bandwidth parameter* $h$*. Additionally, one can verify that in this case, it follows from Theorem 3 that the asymptotic MSE* $\mathbf{R}(\boldsymbol{\theta}^*_h,h)$ *approaches the Cramér-Rao lower bound [37], [38] as* $h \to \infty$*. In other words, at the true model, the M*$\mathcal{K}$*DE approaches asymptotic efficiency as* $h$ *becomes larger.*

### 3.3. Robustness analysis

In robust parameter estimation, we consider the case where the underlying distribution $G$ is a contaminated version of a nominal distribution $F_{\boldsymbol{\theta}_0}$, i.e.,

$$G = (1-\epsilon)F_{\boldsymbol{\theta}_0} + \epsilon Q, \tag{12}$$

where $Q$ is a contaminating probability distribution and $0 \leq \epsilon \leq 1$ denotes a small contamination ratio. A common tool to study the robustness of Fisher-consistent estimators [39] is the influence function [6]. This function, quantifies the bias effect caused by an infinitesimal amount of contamination at a single point $\mathbf{r}$ in the observation space that represents an outlier. Note that in this case, of a single point contamination, the contaminating distribution $Q$ in (12) is considered to be a Dirac measure [40] concentrated at $\mathbf{r}$. An estimator is said to be B-robust, if it's influence function is bounded [6]. Under some mild regularity assumptions, it can be shown that the M$\mathcal{K}$DE is Fisher-consistent with influence function taking the form:

$$\mathbf{IF}(\mathbf{r};\boldsymbol{\theta}_0,h) = \bar{\mathbf{D}}^{-1}(\boldsymbol{\theta}_0,h)\bar{\mathbf{c}}(\mathbf{r},\boldsymbol{\theta}_0,h)\bar{\psi}_F(\mathbf{r},\boldsymbol{\theta}_0,h), \tag{13}$$

where $\bar{\mathbf{D}}(\boldsymbol{\theta},h) \triangleq \mathbb{E}[\bar{\psi}_F(\mathbf{x},\boldsymbol{\theta},h)\bar{\mathbf{c}}(\mathbf{x},\boldsymbol{\theta},h)\bar{\mathbf{c}}^T(\mathbf{x},\boldsymbol{\theta},h);F_{\boldsymbol{\theta}}]$ is assumed to be invertible, the function $\bar{\mathbf{c}}(\mathbf{r},\boldsymbol{\theta},h) \triangleq \nabla_{\boldsymbol{\theta}}\log f(\mathbf{r};\boldsymbol{\theta}) - \mathbb{E}[\bar{\psi}_F(\mathbf{x},\boldsymbol{\theta},h)\nabla_{\boldsymbol{\theta}}\log f(\mathbf{x};\boldsymbol{\theta});F_{\boldsymbol{\theta}}]$, and the function $\bar{\psi}_F(\mathbf{r},\boldsymbol{\theta},h) \triangleq (K_h * f)(\mathbf{r};\boldsymbol{\theta})/\mathbb{E}[(K_h * f)(\mathbf{x};\boldsymbol{\theta});F_{\boldsymbol{\theta}}]$. Hence, one can verify that the influence function (13) is bounded whenever there exists a positive constant $M$, such that:

$$\|\nabla_{\boldsymbol{\theta}}\log f(\mathbf{r};\boldsymbol{\theta})\|(K_h * f)(\mathbf{r};\boldsymbol{\theta}) \leq M \quad \forall \mathbf{r} \in \mathbb{R}^p. \tag{14}$$

Note that the condition (14) can be used for selection of a kernel function to guarantee B-robustness.

## 4. SELECTION OF THE BANDWIDTH PARAMETER

In this section, we develop a data-driven procedure for selecting the bandwidth parameter $h$ to control the estimation accuracy of the nominal vector parameter $\boldsymbol{\theta}_0$ in (12), or a subset of $\boldsymbol{\theta}_0$ containing specific parameters of interest. The proposed method is based on a lower bound on an asymptotic approximation to the weighted MSE (WMSE) $\mathbb{E}[\|\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_0\|^2_{\mathbf{W}};G]$, where $\|\mathbf{a}\|_{\mathbf{C}} \triangleq \sqrt{\mathbf{a}^T\mathbf{C}\mathbf{a}}$ denotes the weighted Euclidean semi-norm of a vector $\mathbf{a}$ with positive-semidefinite weighting matrix $\mathbf{C}$. Note that the weighting matrix $\mathbf{W}$ can be chosen to compensate for possibly different units of the coordinates of $\boldsymbol{\theta}_0$. Also note that in the presence of nuisance parameters, when one is interested in controlling the estimation accuracy of only a subvector of $\boldsymbol{\theta}_0$, $\mathbf{W}$ is chosen to include zeros in the entries corresponding to the nuisance parameters. The lower bound on the WMSE is derived in the following manner. First, observe that

$$\begin{aligned}\mathbb{E}[\|\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_0\|^2_{\mathbf{W}};G] &= \mathbb{E}[\|\hat{\boldsymbol{\theta}}_h - \mathbb{E}[\hat{\boldsymbol{\theta}}_h;G]\|^2_{\mathbf{W}};G] \\ &+ \|\mathbb{E}[\hat{\boldsymbol{\theta}}_h;G] - \boldsymbol{\theta}_0\|^2_{\mathbf{W}}.\end{aligned} \tag{15}$$

Therefore, by (15) and Theorem 3, it follows that

$$\begin{aligned}\mathbb{E}[\|\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_0\|^2_{\mathbf{W}};G] &\overset{a}{\approx} \text{tr}[\mathbf{R}(\boldsymbol{\theta}^*_h,h)\mathbf{W}] + \|\boldsymbol{\theta}^*_h - \boldsymbol{\theta}_0\|^2_{\mathbf{W}} \\ &\geq \text{tr}[\mathbf{R}(\boldsymbol{\theta}^*_h,h)\mathbf{W}],\end{aligned} \tag{16}$$

where tr$[\cdot]$ denotes the trace operator, $\boldsymbol{\theta}^*_h$ (7) is the minimizer of the $\mathcal{K}$-divergence and $\mathbf{R}(\boldsymbol{\theta}^*_h,h)$, defined in (11), is the asymptotic MSE matrix at $\boldsymbol{\theta}^*_h$. Hence, minimization of the objective tr$[\mathbf{R}(\boldsymbol{\theta}^*_h,h)\mathbf{W}]$ w.r.t. $h$ amounts to minimization of a lower bound on the asymptotic WMSE at $\boldsymbol{\theta}_0$. Note that the bound in (16) becomes tighter as the bias term $\|\boldsymbol{\theta}^*_h - \boldsymbol{\theta}_0\|_{\mathbf{W}}$ becomes smaller, which is the case when the contamination ratio $\epsilon$ in (12) is getting smaller.

In practice, the asymptotic MSE $\mathbf{R}(\boldsymbol{\theta}^*_h,h)$ is unknown. Therefore, instead, we use its empirical estimate, obtained by replacing $\boldsymbol{\theta}^*_h$ with $\hat{\boldsymbol{\theta}}_h$ and the unknown expectations in (10) with their empirical estimates. The resulting empirical asymptotic MSE takes the form:

$$\hat{\mathbf{R}}(\hat{\boldsymbol{\theta}}_h,h) \triangleq N^{-1}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}}_h,h), \tag{17}$$

where

$$\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta},h) \triangleq \hat{\mathbf{C}}^{-1}(\boldsymbol{\theta},h)\hat{\mathbf{D}}(\boldsymbol{\theta},h)\hat{\mathbf{C}}^{-1}(\boldsymbol{\theta},h), \tag{18}$$
$$\hat{\mathbf{C}}(\boldsymbol{\theta},h) \triangleq \sum\nolimits_{n=1}^{N} w(\mathbf{x}_n;h)\nabla^2_{\boldsymbol{\theta}}\log f(\mathbf{x}_n;\boldsymbol{\theta}) - \nabla^2_{\boldsymbol{\theta}}\log\hat{u}(\boldsymbol{\theta},h),$$

$w(\mathbf{r};h)$ and $\hat{u}(\boldsymbol{\theta},h)$ are defined below (3) and in (5), respectively,

$$\hat{\mathbf{D}}(\boldsymbol{\theta},h) \triangleq \frac{1}{N}\sum\nolimits_{n=1}^{N}\hat{\mathbf{v}}(\mathbf{x}_n,\boldsymbol{\theta},h)\hat{\mathbf{v}}^T(\mathbf{x}_n,\boldsymbol{\theta},h),$$
$$\hat{\mathbf{v}}(\mathbf{r},\boldsymbol{\theta},h) \triangleq \hat{\psi}_G(\mathbf{r},h)\hat{\mathbf{c}}(\mathbf{r},\boldsymbol{\theta},h) + \hat{\mathbf{d}}(\mathbf{r},\boldsymbol{\theta},h) - \hat{\mathbf{z}}(\mathbf{r},\boldsymbol{\theta},h),$$
$$\hat{\psi}_G(\mathbf{r},h) \triangleq (N-1)\sum\nolimits_{n=1}^{N}(\hat{\varphi}_h(\mathbf{x}_n-\mathbf{r}) - N^{-1}\hat{\varphi}_h(0)),$$
$$\hat{\varphi}_h(\mathbf{r}) \triangleq N^{-1}K_h(\mathbf{r})/\sum\nolimits_{n=1}^{N}\tilde{g}(\mathbf{x}_n;h),$$
$$\hat{\mathbf{d}}(\mathbf{r},\boldsymbol{\theta},h) \triangleq (N-1)\sum_{n=1}^{N}\left(\hat{\varphi}_h(\mathbf{x}_n-\mathbf{r})\hat{\mathbf{c}}(\mathbf{x}_n,\boldsymbol{\theta},h) - \tfrac{\hat{\varphi}_h(0)\hat{\mathbf{c}}(\mathbf{r},\boldsymbol{\theta},h)}{N}\right)$$

and $\hat{\mathbf{c}}(\mathbf{r},\boldsymbol{\theta},h)$ and $\hat{\mathbf{z}}(\mathbf{r},\boldsymbol{\theta},h)$ are obtained from $\mathbf{c}(\mathbf{r},\boldsymbol{\theta},h)$ and $\mathbf{z}(\mathbf{r},\boldsymbol{\theta},h)$, that are defined below (10), by replacing $u(\boldsymbol{\theta},h)$ with its empirical estimate $\hat{u}(\boldsymbol{\theta},h)$ in (5).

We emphasize that the estimator (17) is constructed by the same sequence of samples used for obtaining the M$\mathcal{K}$DE (6). To conclude, following the discussion above, the proposed selection rule for the bandwidth parameter $h$ is given by:

$$h_{\text{opt}} \triangleq \arg\min\nolimits_{h\in I}\text{tr}[\hat{\mathbf{R}}(\hat{\boldsymbol{\theta}}^*_h,h)\mathbf{W}], \tag{19}$$

where $I$ is some predefined search interval.

## 5. NUMERICAL EXAMPLES

In this section, the proposed M$\mathcal{K}$DE (6) is applied for estimating the parameters of a two-component GMM. The performance of the M$\mathcal{K}$DE are compared to those of the non-robust (mismatched) MLE, and the robust M$\alpha$DE, M$\beta$DE and M$\gamma$DE. Implementation of the MHDE is highly cumbersome under the considered scenario due to multivariate numerical integration. Therefore, comparison to the MHDE is omitted.

**Simulation settings:** In this study, the density function of the nominal probability distribution $F_{\boldsymbol{\theta}}$ in (12) is given by:

$$f(\mathbf{r};\boldsymbol{\theta}) = 0.5\phi(\mathbf{r};\mathbf{0},\sigma^2\mathbf{I}) + 0.5\phi(\mathbf{r};\boldsymbol{\eta},\sigma^2\mathbf{I}), \tag{20}$$

where $\phi(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes a normal density function, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\mathbf{I}$ is the identity matrix. Here, the vector parameter $\boldsymbol{\theta} \triangleq [\boldsymbol{\eta}^T, \sigma^2]^T$, where $\boldsymbol{\eta}$ is the vector parameter of interest and $\sigma^2$ is a nuisance parameter. The components of the true vector parameter $\boldsymbol{\theta}_0 \triangleq [\boldsymbol{\eta}_0^T, \sigma_0^2]^T$ were set to $\boldsymbol{\eta}_0 = 6 \times \mathbf{1}$, where $\mathbf{1}$ denotes a vector with unit entries and $\sigma_0^2 = 2$. The contaminating distribution $Q$ in (12) was considered to be normal with density:

$$q(\mathbf{r}) = \phi(\mathbf{r};\boldsymbol{\mu}_c,\sigma_c^2\mathbf{I}), \tag{21}$$

where $\boldsymbol{\mu}_c$ and $\sigma_c^2$ were set to $10 \times \mathbf{1}$ and 5, respectively. In all simulation examples that follow, we assume that a sequence of $N = 300$ mutually independent samples from the contaminated distribution $G$ in (12) is available. The dimensionality was set to $p = 10$.

**Implementation details:** The proposed M$\mathcal{K}$DE (6) was implemented with a Gaussian kernel function

$$K_h(\mathbf{r}) \triangleq \phi(\mathbf{r};\mathbf{0},h^2\mathbf{I}), \tag{22}$$

under which the robustness condition (14) holds for the nominal density function (20). The M$\alpha$DE was implemented similarly to [24], by minimizing the empirical $\alpha$-divergence between the power-transformed density $g_\alpha(\mathbf{x}) \triangleq g^{\frac{1}{\alpha}}(\mathbf{x})/\int_{\mathbb{R}^p} g^{\frac{1}{\alpha}}(\mathbf{x})d\lambda(\mathbf{x})$ and $f(\mathbf{x};\boldsymbol{\theta})$. The tuning parameter $\alpha$ was set to 0.5. One can verify that, under the nominal density function (20), the M$\beta$DE and the M$\gamma$DE have extremely difficult implementations under fractional values of the tuning parameters $\beta$ and $\gamma$. This is due to the fact that under these values the involved multivariate integrals cannot be solved analytically. Therefore, these estimators were implemented here with $\beta = \gamma = 1$. We note that for $\beta = 1$, the M$\beta$DE reduces to the minimum $L_2$-distance estimator, which is B-robust, but may be inefficient at the true model. All compared estimators were implemented via iterative fixed-point algorithms. These are presented in [35, Sec. VI] along with exact details regarding the initialization, maximum number of iterations and stopping rule.

**Results:** Throughout the examples that follow we examine the estimation performance of the parameter of interest $\boldsymbol{\eta}_0$. Therefore, the weight matrix $\mathbf{W}$ in (15) was set to be diagonal, where the first $p$ diagonal entires were set to one while the last one was set to zero.

First, we examine the relation between the empirical estimate of the WMSE at the true vector parameter, i.e., $\mathbb{E}[\|\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_0\|^2_{\mathbf{W}}; G]$, to the empirical asymptotic version of the WMSE at $\boldsymbol{\theta}_h^*$, i.e., $\mathbb{E}[\|\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h^*\|^2_{\mathbf{W}}; G]$, as a function of the kernel's bandwidth parameter $h$. The first, was obtained via $10^4$ independent Monte-Carlo trials. The latter, is given by $\mathrm{tr}[\hat{\mathbf{R}}(\hat{\boldsymbol{\theta}}_h, h)\mathbf{W}]$, where $\hat{\mathbf{R}}(\hat{\boldsymbol{\theta}}_h, h)$ is the empirical asymptotic MSE matrix defined in (17). We note that this quantity was obtained from a single realization of $N = 300$ i.i.d. snapshots generated from the contaminated distribution $G$. Here, the contamination ratio parameter $\epsilon$ in (12) was set to 0.01. Observing Fig. 1(a), one can notice that the empirical estimate of the asymptotic WMSE at $\boldsymbol{\theta}_h^*$ accurately predicts the minimum of the empirical WMSE at $\boldsymbol{\theta}_0$. This illustrates the reliability of the criterion (19) for selection of the Kernel's bandwidth parameter $h$.

Second, we compared the empirical version of the WMSE $\mathbb{E}[\|\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_0\|^2_{\mathbf{W}}]$ attained by the proposed M$\mathcal{K}$DE to those of the other compared algorithms versus the contamination ratio parameter $\epsilon$ in (12). All empirical WMSE curves were obtained via $10^4$ independent Monte-Carlo trials. Here, the optimal bandwidth parameter $h_{\text{opt}}$ was selected according to (19), where the minimization was carried out over 30 equally spaced grid points of the interval $[1, 30]$. Observing Fig. 1(b), one sees that the proposed M$\mathcal{K}$DE, which involves optimization of a performance related objective, outperforms the all compared algorithms with considerable performance gap even for relatively large values of the contamination ratio $\epsilon$.

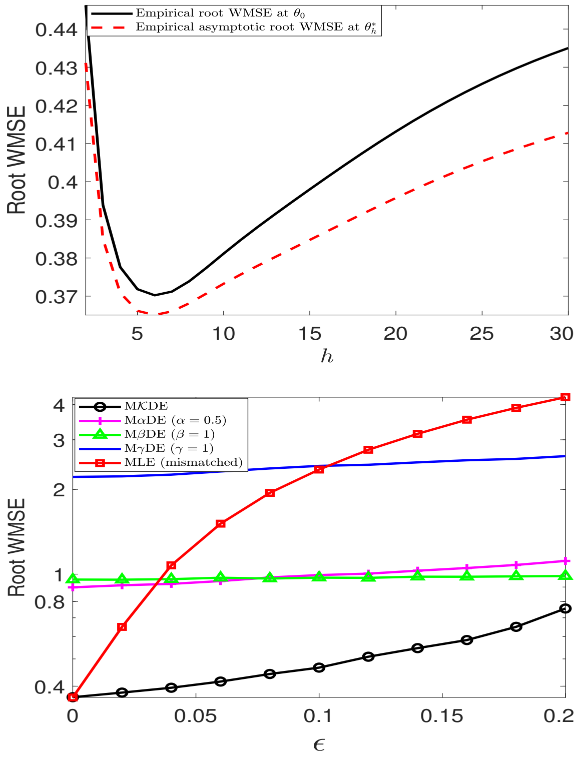

**Fig. 1**. (**a**) The empirical root WMSE (RWMSE) at $\boldsymbol{\theta}_0$ and the empirical asymptotic RMSE at $\boldsymbol{\theta}_h^*$ versus the bandwidth parameter $h$ of the kernel function (22). (**b**) The empirical RWMSE's of the compared estimators versus the contamination ratio parameter $\epsilon$.

## 6. CONCLUSION

In this paper, a new divergence, called $\mathcal{K}$ divergence, was presented. Unlike other state-of-the-art divergences it avoids the use of density powers that may consequence cumbersome implementation in flexible distributional models. Based on the $\mathcal{K}$ divergence, a new robust estimator was developed that utilizes Parzen's non-parametric kernel density estimator to suppress outlying measurements. We have shown that proper selection of the kernel's bandwidth parameter, to optimize a performance related objective, can lead to enhanced estimation performance along with implementation simplicity.

## 7. REFERENCES

[1] E. L. Lehmann, and G. Casella, *Theory of Point Estimation,* Springer, 1998.

[2] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory,* Prentice-Hall, 1998.

[3] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics,* vol. 22, pp. 79-86, 1951.

[4] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica: Journal of the Econometric Society,* pp. 1-25, 1982.

[5] P. J. Huber, *Robust statistics,* Springer, 2011.

[6] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel, *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2011.

[7] R. J. Serfling, *Approximation theorems of mathematical statistics.* John Wiley & Sons, 1980.

[8] R. A. Maronna, "Robust $M$-estimators of multivariate location and scatter," *The Annals of Statistics,* vol. 4, no. 1, pp. 51-67, 1976.

[9] D. E. Tyler, "A distribution-free M-estimator of multivariate scatter," *Annals of statistics,* vol. 15, no. 1, pp. 234-251, 1987.

[10] E. Ollila, D. E. Tyler, V. Koivunen and H. V. Poor, "Complex elliptically symmetric distributions: survey, new results and applications," *IEEE Transactions on Signal Processing,* vol. 60, no. 1, pp. 5597-5625, 2012.

[11] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh and M. Muma, "Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts," *IEEE Signal Processing Magazine,* vol. 29, no. 4, pp. 61-80, 2012.

[12] A. M. Zoubir, V. Koivunen, E. Ollila and M. Muma, *Robust statistics for signal processing*, Cambridge University Press, 2018.

[13] K. Todros and A. O. Hero, "Measure-transformed quasi maximum likelihood estimation with application to source localization," *Proceedings of ICASSP 2015,* pp. 3462 - 3466.

[14] K. Todros and A. O. Hero, "Measure-transformed quasi maximum likelihood estimation," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 748-763, 2017.

[15] R. O. Duda and P. E. Hart, *Pattern classification and scene analysis,* Wiley, 1973.

[16] R. Beran, "Minimum Hellinger distance estimates for parametric models," *The Annals of Statistics,* vol. 5, no. 3, pp. 445-463, 1977.

[17] A. Cutler and O. I. Cordero-Braña, "Minimum Hellinger distance estimation for finite mixture models," *Journal of the American Statistical Association,* vol. 91, no. 436, pp. 1716-1723, 1996.

[18] Toma, "Minimum Hellinger distance estimators for some multivariate models: influence functions and breakdown point results," *Comptes Rendus Mathematique,* vol. 345, no. 6, pp. 353-358, 2007.

[19] A. Toma, "Minimum Hellinger distance estimators for multivariate distributions from the Johnson system," *Journal of statistical planning and inference,* vol. 138, no. 3, pp. 803-816, 2008.

[20] A. Basu, H. Shioya and C. Park, *Statistical inference: the minimum distance approach,* CRC press, 2011.

[21] B. W. Silverman, *Density estimation for statistics and data analysis,* CRC press, 1986.

[22] D. L. Donoho and R. C. Liu, "The "automatic" robustness of minimum distance functionals," *The Annals of Statistics,* vol. 16, pp. 552-586, 1988.

[23] A. Cichocki and S. I. Amari, "Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities," *Entropy,* vol. 12, no. 6, pp. 1532-1568, 2010.

[24] A. Iqbal and A-K. Seghouane, "An $\alpha$-divergence-based approach for robust dictionary learning," *IEEE Transactions on Image Processing,* vol. 28, no. 11, pp. 5729-5739, 2019.

[25] A-K. Seghouane and N. Shokouhi, "Adaptive learning for robust radial basis function networks," *IEEE Transactions on Cybernetics,* In Press, 2019.

[26] A. Basu, I. R. Harris, N. L. Hjort and M. C. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika,* vol. 85, no. 3, pp. 549-559, 1998.

[27] M. Mihoko and S. Eguchi, "Robust blind source separation by beta divergence," *Neural computation,* vol. 14, no. 8, pp. 1859-1886, 2002.

[28] M. Mollah, M. Minami and S. Eguchi, "Exploring latent structure of mixture ICA models by the minimum $\beta$-divergence method," *Neural Computation,* vol. 18, no. 1, pp. 166-190, 2006.

[29] H. Fujisawa and S. Eguchi, "Robust estimation in the normal mixture model," *Journal of Statistical Planning and Inference,* vol. 136, no. 11, pp. 3989-4011, 2006.

[30] M. Mollah, S. Eguchi and M. Minami, "Robust prewhitening for ICA by minimizing $\beta$-divergence and its application to FastICA," *Neural Processing Letters,* vol. 25, no. 2, pp. 91-110, 2007.

[31] M. Mollah, N. Sultana, M. Minami and S. Eguchi, "Robust extraction of local structures by the minimum $\beta$-divergence method," *Neural Networks,* vol. 23, no. 2, pp. 226-238, 2010.

[32] H. Fujisawa and S. Eguchi, "Robust parameter estimation with a small bias against heavy contamination," *Journal of Multivariate Analysis,* vol. 99, no. 9, pp. 2053-2081, 2008.

[33] H. Fujisawa and S. Eguchi, "Entropy and divergence associated with power function and the statistical application," *Entropy,* vol. 12, no. 2, pp. 262-274, 2010.

[34] S. Eguchi and S. Kato, "Robust independent component analysis via minimum $\gamma$-divergence estimation," *IEEE Journal of Selected Topics in Signal Processing,* vol. 7, no. 4, pp. 614-624, 2013.

[35] Y. Sorek and K. Todros, "Robust parameter estimation based on the $\mathcal{K}$-divergence: Supplementary Material," *Available online at:* http://www.ee.bgu.ac.il/~todros/ICASSP_SUPP_MATERIAL.pdf, Sep. 2021.

[36] K. B. Athreya and S. N. Lahiri, *Measure theory and probability theory,* Springer-Verlag, 2006.

[37] H. Cramér, "A contribution to the theory of statistical estimation," *Skand. Akt. Tidskr.,* vol. 29, pp. 85-94, 1946.

[38] C. R. Rao, "Information and accuracy attainable in the estimation of statistical parameters," *Bull. Calcutta Math. Soc.,* vol. 37, pp. 81-91, 1945.

[39] D. R. Cox and D. V. Hinkley, *Theoretical Statistics,* Chapman & Hall, 1974.

[40] G. B. Folland, *Real Analysis,* John Wiley and Sons, 1984.