Citation: O. Kenig, K. Todros and T. Adali, "Robust GMM Parameter Estimation via the K-BM Algorithm," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10094602.

DOI: https://doi.org/10.1109/ICASSP49357.2023.10094602

ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094602

ROBUST GMM PARAMETER ESTIMATION VIA THE K-BM ALGORITHM

Ori Kenig, Koby Todros

Ben-Gurion University of the Negev

Tülay Adali

University of Maryland, Baltimore County

ABSTRACT

In this paper, we develop an expectation-maximization (EM)-like scheme, called $\mathcal{K}$-BM, for iterative numerical computation of the minimum $\mathcal{K}$-divergence estimator (M$\mathcal{K}$DE). This estimator utilizes Parzen's non-parameteric $\mathcal{K}$ernel density estimate to down weight low density areas attributed to outliers. Similarly to the standard EM algorithm, the $\mathcal{K}$-BM involves successive **M**aximizations of lower **B**ounds on the objective function of the M$\mathcal{K}$DE. Differently from EM, these bounds do not rely on conditional expectations only. The proposed $\mathcal{K}$-BM algorithm is applied to robust parameter estimation of a finite-order multivariate Gaussian mixture model (GMM). Simulation studies illustrate the performance advantage of the $\mathcal{K}$-BM as compared to other state-of-the-art robust GMM estimators.

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

1. INTRODUCTION

Parameter estimation of a Gaussian mixture model (GMM) is an important problem, encountered in many applications that involve parametric density estimation and data clustering [1]-[7].

The maximum likelihood estimator (MLE) is a popular tool for this problem that operates by minimizing the empirical Kullback-Leibler divergence (KLD) [8] between the underlying probability distribution of the data and a hypothesized probability model [9]. Direct implementation of the MLE for GMM parameter estimation is highly cumbersome. Therefore, an iterative numerical procedure, called expectation-maximization (EM) [10], is applied instead that successively maximizes tractable lower bounds on the log-likelihood function [11]. These bounds involve conditional expectation of a complete log-likelihood function that, in addition to the observations, incorporates latent data. Proper specification of the latent data results in bounds that are easier to maximize than the (incomplete) log-likelihood function, leading to a simplified implementation of the MLE. Nevertheless, the resulting EM algorithm for GMM parameter estimation [12] is highly sensitive to small model misspecification inflicted by outliers [13].

To overcome this limitation, several robust EM-like alternatives where proposed. In [14]-[17, Sec. 5.2], robust GMM estimators for univariate data were developed. Robust alternatives for multivariate data were proposed in [18] and [19], that account for outliers by replacing the hypothesized GMM with a mixture of heavy-tailed t-distributions. These multivariate estimators may be successful in mitigating the effect of outliers. However, the estimation accuracy may be defected when the outliers are not associated with the tails of the assumed t-distributions, e.g., when they are non symmetrically scattered about the centroids. A different multivariate technique, called robust improper maximum likelihood estimator (RIMLE), was proposed in [20]. To account for outliers, this method utilizes a convex combination between a nominal GMM and an improper uniform contaminating distribution with a certain level. An optimally level-tuned version of this estimator, called OTRIMLE, was developed in [21]. Nevertheless, we note that deviation from the assumption of uniformly distributed outliers may lead to inaccurate estimation. Lastly, a class of robust multivariate methods, called TCLUST, that fit trimmed GMMs with non-overlapping components was developed in [22]-[26]. In this context, it is important to note that improper trimming or the presence of overlapping Gaussian components may degrade the estimation performance.

Main contribution: In this paper, we develop a new EM-like scheme for robust multivariate GMM parameter estimation. Unlike the methods described above, we do not alter the hypothesized GMM nor do we assume a specific contaminating distribution. Furthermore, we do not apply trimming or restrict the Gaussian mixture components to be non-overlapping.

The proposed method provides numerical computation of the minimum $\mathcal{K}$-divergence estimator (M$\mathcal{K}$DE) [27]. This estimator utilizes Parzen's non-parametric kernel density estimate [28] to down-weight low density domains attributed to outliers. In other words, the M$\mathcal{K}$DE applies intrinsic model-free weighting to suppress outliers. In [27], it was shown that when the hypothesized model is correctly specified, the M$\mathcal{K}$DE is consistent for any fixed value of the kernel's bandwidth parameter. Hence, this parameter is a degree-of-freedom that can be tuned to enhance the estimation accuracy of the model parameters. Additional properties of the M$\mathcal{K}$DE appear in [27].

We begin by developing a general EM-like scheme, called $\mathcal{K}$-BM, for iterative numerical computation of the M$\mathcal{K}$DE. Similarly to the standard EM, the $\mathcal{K}$-BM involves successive **M**aximizations of lower **B**ounds on the objective function of the M$\mathcal{K}$DE. However, in difference to the EM, these bounds do not rely on conditional expectations only.

The proposed $\mathcal{K}$-BM is then applied to the GMM parameter estimation problem at hand. We obtain explicit estimation update equations for the GMM parameters. Furthermore, a data-driven procedure for selection of the kernel's bandwidth parameter is developed that minimizes an instantaneous stochastic approximation of the mean-integrated-squared-error (MISE). Lastly, the $\mathcal{K}$-BM is examined in a simulation study that illustrates its advantages over the non-robust EM and other state-of-the-art robust alternatives.

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

In this section, we briefly review the M$\mathcal{K}$DE [27]. We begin with some definitions and assumptions. Let $G_{\mathbf{x}}$ denote an unknown probability distribution of a random vector $\mathbf{x} \in \mathbb{R}^p$. Furthermore, consider a parametric class $\{F_{\mathbf{x};\boldsymbol{\theta}}\}$ comprising hypothesized probability distributions of $\mathbf{x}$, which does not necessarily contain the true distribution $G_{\mathbf{x}}$. This class is indexed by a vector parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta} \subseteq \mathbb{R}^m$ is a parameter space. We assume that $G_{\mathbf{x}}$ and $F_{\mathbf{x};\boldsymbol{\theta}}$ posses density functions, w.r.t. Lebesgue's measure λ, denoted by $g_{\mathbf{x}}(\cdot)$ and $f_{\mathbf{x}}(\cdot;\boldsymbol{\theta})$, respectively. The latter density is called the hypothesized likelihood function.

Given a sequence of mutually independent samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ from $G_{\mathbf{x}}$, the M$\mathcal{K}$DE is generated by minimizing the empirical $\mathcal{K}$-divergence [27, Eq. (3)] between $G_{\mathbf{x}}$ and $\{F_{\mathbf{x};\boldsymbol{\theta}}\}$. In [27, Sec. III-A], it was shown that this minimization amounts to maximization of the following objective w.r.t. $\boldsymbol{\theta}$:

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

where the weight function

$$w(\mathbf{r}; h) \triangleq \frac{\tilde{g}_{\mathbf{x}}(\mathbf{r}; h)}{\sum_{m=1}^{N} \tilde{g}_{\mathbf{x}}(\mathbf{x}_m; h)}, \quad (2)$$

the function $\tilde{g}_{\mathbf{x}}(\mathbf{r}; h) \triangleq \hat{g}_{\mathbf{x}}(\mathbf{r}; h) - N^{-1} K_h(0)$ and

$$\hat{g}_{\mathbf{x}}(\mathbf{r}; h) \triangleq \frac{1}{N} \sum\nolimits_{m=1}^{N} K_h(\mathbf{r} - \mathbf{x}_m) \quad (3)$$

is Parzen's kernel density estimator [28] of the true underlying density $g_{\mathbf{x}}(\mathbf{r})$. Here, $K_h(\mathbf{r}) \triangleq h^{-p} K(h^{-1}\mathbf{r})$, where $K(\mathbf{r})$ is a strictly positive, bounded, continuous and integrable kernel function and $h \in \mathbb{R}_{>0}$ is a bandwidth parameter. One can verify that under this formulation the weights $w(\mathbf{x}_n; h)$, $n = 1, \ldots, N$ are non-negative. Lastly, the function

$$\hat{u}(\boldsymbol{\theta}, h) \triangleq \int_{\mathbb{R}^p} \hat{g}_{\mathbf{x}}(\mathbf{r}; h) f_{\mathbf{x}}(\mathbf{r}; \boldsymbol{\theta}) d\lambda(\mathbf{r}). \quad (4)$$

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

$$\hat{\boldsymbol{\theta}}_h \triangleq \arg\max\nolimits_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{J}_h(\boldsymbol{\theta}). \quad (5)$$

Notice that the hypothesized log-likelihood realizations in (1) are weighted in accordance to the approximate underlying density of the data. This results in intrinsic model-free suppression of outlying observations, corresponding to low density areas. A rigorous outlier robustness analysis of the M$\mathcal{K}$DE (5) appears in [27, Sec. III-C].

3. THE $\mathcal{K}$-BM ALGORITHM

In this section, we develop a general EM-like scheme, called $\mathcal{K}$-BM for iterative numerical computation of the M$\mathcal{K}$DE (5). Similarly to the standard EM algorithm [10], [11], this method successively maximizes tractable lower bounds on the objective $\mathcal{J}_h(\boldsymbol{\theta})$ (1).

3.1. A lower bound on $\mathcal{J}_h(\boldsymbol{\theta})$

To develop a lower bound on $\mathcal{J}_h(\boldsymbol{\theta})$, we begin by bounding the hypothesized log-likelihood function $\log f_{\mathbf{x}}(\mathbf{r}; \boldsymbol{\theta})$. To that sake, we introduce a latent random vector $\mathbf{y} \in \mathbb{R}^q$, which is assumed to be related to $\mathbf{x}$ through some non-invertible mapping. The hypothesized distribution $F_{\mathbf{x};\boldsymbol{\theta}}$ is then a marginalized version of the joint distribution $F_{\mathbf{x},\mathbf{y};\boldsymbol{\theta}}$. Hence, one can verify that for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$

$$\begin{aligned}\log f_{\mathbf{x}}(\mathbf{r}; \boldsymbol{\theta}) = \log f_{\mathbf{x}}(\mathbf{r}; \boldsymbol{\theta}') + v(\mathbf{r}; \boldsymbol{\theta}, \boldsymbol{\theta}') \\ + \mathcal{D}_{\mathrm{KL}}[F_{\mathbf{y}|\mathbf{x}=\mathbf{r};\boldsymbol{\theta}'} || F_{\mathbf{y}|\mathbf{x}=\mathbf{r};\boldsymbol{\theta}}],\end{aligned}$$

where $v(\mathbf{r}; \boldsymbol{\theta}, \boldsymbol{\theta}') \triangleq \mathbb{E}\left[\log \frac{f_{\mathbf{x},\mathbf{y}}(\mathbf{r},\mathbf{y};\boldsymbol{\theta})}{f_{\mathbf{x},\mathbf{y}}(\mathbf{r},\mathbf{y};\boldsymbol{\theta}')}; F_{\mathbf{y}|\mathbf{x}=\mathbf{r};\boldsymbol{\theta}'}\right]$, $\mathbb{E}[\cdot; P]$ denotes the statistical expectation w.r.t. a probability distribution P, $F_{\mathbf{y}|\mathbf{x};\boldsymbol{\theta}}$ is the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ parameterized by $\boldsymbol{\theta}$ and $\mathcal{D}_{\mathrm{KL}}[\cdot||\cdot]$ denotes the KLD [8]. Therefore, the non-negativity of the KLD implies that

$$\log f_{\mathbf{x}}(\mathbf{r}; \boldsymbol{\theta}) \geq \log f_{\mathbf{x}}(\mathbf{r}; \boldsymbol{\theta}') + v(\mathbf{r}; \boldsymbol{\theta}, \boldsymbol{\theta}') \;\; \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}. \quad (6)$$

Next, we derive an upper bound on the term $\log \hat{u}(\boldsymbol{\theta}, h)$ in (1). Notice that $\log \hat{u}(\boldsymbol{\theta}, h) = \log \hat{u}(\boldsymbol{\theta}', h) + \log \frac{\hat{u}(\boldsymbol{\theta}, h)}{\hat{u}(\boldsymbol{\theta}', h)}$ $\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$. Therefore applying the inequality, $\log a \leq a - 1$, that holds for any $a > 0$, yields

$$\log \hat{u}(\boldsymbol{\theta}, h) \leq \log \hat{u}(\boldsymbol{\theta}', h) + \frac{\hat{u}(\boldsymbol{\theta}, h)}{\hat{u}(\boldsymbol{\theta}', h)} - 1 \;\; \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}. \quad (7)$$

Hence, by (1), (6) and (7) it follows that a lower bound on $\mathcal{J}_h(\boldsymbol{\theta})$ takes the form:

$$\mathcal{J}_h(\boldsymbol{\theta}) \geq \mathcal{B}_h(\boldsymbol{\theta}, \boldsymbol{\theta}') \triangleq \mathcal{J}_h(\boldsymbol{\theta}') + \mathcal{Q}_h(\boldsymbol{\theta}, \boldsymbol{\theta}') \;\; \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}, \quad (8)$$

where $\mathcal{Q}_h(\boldsymbol{\theta}, \boldsymbol{\theta}') \triangleq \sum_{n=1}^{N} w(\mathbf{x}_n; h) v(\mathbf{x}_n; \boldsymbol{\theta}, \boldsymbol{\theta}') - \frac{\hat{u}(\boldsymbol{\theta}, h)}{\hat{u}(\boldsymbol{\theta}', h)} + 1$. Note that equality holds when $\boldsymbol{\theta} = \boldsymbol{\theta}'$.

3.2. The $\mathcal{K}$-BM procedure

The $\mathcal{K}$-BM algorithm successively alternates between maximization of the bound in (8) (M-step) w.r.t. $\boldsymbol{\theta}$ and its reevaluation (B-step) carried out by replacing the argument $\boldsymbol{\theta}'$ with the attained maximum point. Discarding the $\boldsymbol{\theta}$-independent terms comprised in $\mathcal{B}_h(\boldsymbol{\theta}, \boldsymbol{\theta}')$ results in the following equivalent objective:

$$\mathcal{Q}_h^*(\boldsymbol{\theta}, \boldsymbol{\theta}') \triangleq \sum\nolimits_{n=1}^{N} w(\mathbf{x}_n; h) v^*(\mathbf{x}_n; \boldsymbol{\theta}, \boldsymbol{\theta}') - \frac{\hat{u}(\boldsymbol{\theta}, h)}{\hat{u}(\boldsymbol{\theta}', h)}, \quad (9)$$

where $v^*(\mathbf{r}; \boldsymbol{\theta}, \boldsymbol{\theta}') \triangleq \mathbb{E}\left[\log f_{\mathbf{x},\mathbf{y}}(\mathbf{r}, \mathbf{y}; \boldsymbol{\theta}); F_{\mathbf{y}|\mathbf{x}=\mathbf{r};\boldsymbol{\theta}'}\right]$. Note that unlike the objective in (1), the considered objective (9) involves a conditional expectation of the joint (complete) log-likelihood $\log f_{\mathbf{x},\mathbf{y}}(\cdot, \cdot)$, which upon proper choice of the latent data, should be easier to handle as compared to the marginal (incomplete) log-likelihood $\log f_{\mathbf{x}}(\cdot)$. Furthermore, the objective (9) does not involve a non-linear logarithm function applied to the integral term $\hat{u}(\boldsymbol{\theta}, h)$ (4), which may further simplify the maximization.

To conclude, the $\mathcal{K}$-BM update $\hat{\boldsymbol{\theta}}_h^{(l)} \rightarrow \hat{\boldsymbol{\theta}}_h^{(l+1)}$, where $l \in \mathbb{N}$ denotes an iteration index, consists of the following two steps:

B-step: Compute the lower bound equivalent $\mathcal{Q}_h^*(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_h^{(l)})$ in (9).

M-step: Obtain $\hat{\boldsymbol{\theta}}_h^{(l+1)} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{Q}_h^*(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_h^{(l)})$.

Similarly to [29], it can be shown that the sequence $\{\mathcal{J}_h(\hat{\boldsymbol{\theta}}_h^{(l)})\}$ converges to a local maximum or to a saddle point as $l \rightarrow \infty$.

4. THE $\mathcal{K}$-BM FOR GMM PARAMETER ESTIMATION

In this section, the $\mathcal{K}$-BM algorithm is applied to the problem of GMM parameter estimation.

4.1. Problem formulation

Consider the case where the underlying probability distribution $G_{\mathbf{x}}$ is a contaminated version of a nominal distribution $F_{\mathbf{x};\boldsymbol{\theta}_0}$, i.e.,

$$G_{\mathbf{x}} = (1 - \epsilon) F_{\mathbf{x};\boldsymbol{\theta}_0} + \epsilon Q_{\mathbf{x}}, \quad (10)$$

where $0 \leq \epsilon \leq 1$ is a small unknown contamination ratio and $Q_{\mathbf{x}}$ is an unknown contaminating probability distribution. Here, $F_{\mathbf{x};\boldsymbol{\theta}}$ is a finite-order Gaussian mixture distribution with density function

$$f_{\mathbf{x}}(\mathbf{r}; \boldsymbol{\theta}) = \sum\nolimits_{m=1}^{M} \alpha_m \phi(\mathbf{r}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \quad (11)$$

where $\{\alpha_m\}_{m=1}^M$ are non-negative mixing proportions that sum up to unity and $\phi(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes a multivariate normal density function with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Similarly to [21], we shall assume here that the model order M is known. Hence, given a sequence of N mutually independent samples from $G_{\mathbf{x}}$, the problem at hand is to estimate $\boldsymbol{\theta} \triangleq [\boldsymbol{\alpha}^T, \mathbf{b}^T, \mathbf{c}^T] \in \mathbb{R}^{\frac{1}{2}M(p+1)(p+2)}$, where $\boldsymbol{\alpha} \triangleq [\alpha_1, \ldots, \alpha_M]^T$, $\mathbf{b} \triangleq [\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_M^T]^T$, $\mathbf{c} \triangleq [\mathbf{c}_1^T, \ldots, \mathbf{c}_M^T]^T$, $\mathbf{c}_m \triangleq \mathrm{vech}[\boldsymbol{\Sigma}_m]$ and $\mathrm{vech}[\cdot]$ denotes the half-vectorization operator of a symmetric matrix [30].

4.2. Derivation of the $\mathcal{K}$-BM for GMM parameter estimation

We begin with specification of the latent random vector $\mathbf{y}$, required for obtaining the objective (9). Similarly to the standard EM algorithm for GMM parameter estimation [12], we set $\mathbf{y}$ as an indicator to the nominal Gaussian component generating $\mathbf{x}$, i.e., $\mathbf{x}$ and $\mathbf{y}$ are assumed to be related through the non-invertible mapping:

$$\mathbf{x} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_M]\mathbf{y}, \tag{12}$$

where $\boldsymbol{s}_m$ is a Gaussian random vector with mean $\boldsymbol{\mu}_m$ and covariance $\boldsymbol{\Sigma}_m$, $\mathbf{y} = \mathbf{e}_m$ w.p. α_m end $\mathbf{e}_m \in \mathbb{R}^M$ is the m-th unit vector.

Next, we choose the following Gaussian kernel function:

$$K_h(\mathbf{r}) \triangleq \phi(\mathbf{r}; \mathbf{0}, h^2\mathbf{I}). \tag{13}$$

Under this kernel and the nominal GMM density (11), the integral term (4) has analytical solution and the outlier robustness condition stated in [27, Eq. (24)] is satisfied.

Hence, by (12) and (13) it follows that the functions $v^*(\cdot;\cdot,\cdot)$ and $\hat{u}(\cdot,\cdot)$ in the objective $\mathcal{Q}_h^*(\boldsymbol{\theta},\boldsymbol{\theta}')$ (9), that is calculated in the B-step, take the forms:

$$v^*(\mathbf{r};\boldsymbol{\theta},\boldsymbol{\theta}') = \sum_{m=1}^M \gamma_m(\mathbf{r};\boldsymbol{\theta}') \log(\alpha_m \phi(\mathbf{r};\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m)), \tag{14}$$

$$\hat{u}(\boldsymbol{\theta},h) = \frac{1}{N}\sum_{n=1}^N \sum_{m=1}^M \alpha_m \phi(\mathbf{x}_n;\boldsymbol{\mu}_m,\bar{\boldsymbol{\Sigma}}_m), \tag{15}$$

where $\gamma_m(\mathbf{r};\boldsymbol{\theta}) \triangleq \frac{\alpha_m \phi(\mathbf{r},\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m)}{\sum_{k=1}^M \alpha_k \phi(\mathbf{r},\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}$, $\bar{\boldsymbol{\Sigma}}_m \triangleq \boldsymbol{\Sigma}_m + h^2\mathbf{I}_p$ and $\mathbf{I}_p$ denotes a $p \times p$ identity matrix.

In the M-step, maximization of $\mathcal{Q}_h^*(\boldsymbol{\theta},\boldsymbol{\theta}')$ w.r.t. the mixing proportions, subject to the unit sum constraint, is performed via Lagrange multipliers method [31]. Maximization of $\mathcal{Q}_h^*(\boldsymbol{\theta},\boldsymbol{\theta}')$ w.r.t. the mean vectors and the covariance matrices is carried out by equating the corresponding partial gradients to zero. The maximum points satisfy the following fixed-point equations for $m = 1, \ldots, M$:

$$\begin{aligned}
\alpha_m &= (1+\xi_m)^{-1}(\lambda_m + \alpha_m \eta) \\
\boldsymbol{\mu}_m &= \boldsymbol{\Gamma}_m^{-1}(\boldsymbol{\Sigma}_m^{-1}\hat{\boldsymbol{\mu}}_{w,m} - \rho_m \lambda_m^{-1}\bar{\boldsymbol{\Sigma}}_m^{-1}\hat{\boldsymbol{\mu}}_{\phi,m}) \\
\boldsymbol{\Sigma}_m &= \hat{\boldsymbol{\Sigma}}_{w,m} + \rho_m \lambda_m^{-1}\boldsymbol{\Sigma}_m\bar{\boldsymbol{\Sigma}}_m^{-1}(\bar{\boldsymbol{\Sigma}}_m - \hat{\boldsymbol{\Sigma}}_{\phi,m})\bar{\boldsymbol{\Sigma}}_m^{-1}\boldsymbol{\Sigma}_m,
\end{aligned} \tag{16}$$

where $\eta \triangleq \frac{\hat{u}(\boldsymbol{\theta},h)}{\hat{u}(\boldsymbol{\theta}',h)}$, $\lambda_m \triangleq \sum_{n=1}^N w(\mathbf{x}_n;h)\gamma_m(\mathbf{x}_n;\boldsymbol{\theta}')$ and the term $\xi_m \triangleq \frac{1}{\hat{u}(\boldsymbol{\theta}',h)}\sum_{n=1}^N \phi(\mathbf{x}_n,\boldsymbol{\mu}_m,\bar{\boldsymbol{\Sigma}}_m)$. Here, $\rho_m \triangleq \alpha_m\xi_m$ and the matrix term $\boldsymbol{\Gamma}_m \triangleq \boldsymbol{\Sigma}_m^{-1} - \rho_m\lambda_m^{-1}\bar{\boldsymbol{\Sigma}}_m^{-1}$. The empirical mean vectors $\hat{\boldsymbol{\mu}}_{w,m}$ and $\hat{\boldsymbol{\mu}}_{\phi,m}$ take the forms:

$$\hat{\boldsymbol{\mu}}_{w,m} \triangleq \frac{\sum_{n=1}^N w(\mathbf{x}_n;h)\gamma_m(\mathbf{x}_n;\boldsymbol{\theta}')\mathbf{x}_n}{\sum_{n=1}^N w(\mathbf{x}_n;h)\gamma_m(\mathbf{x}_n;\boldsymbol{\theta}')}$$

$$\hat{\boldsymbol{\mu}}_{\phi,m} \triangleq \frac{\sum_{n=1}^N \phi(\mathbf{x}_n;\boldsymbol{\mu}_m,\bar{\boldsymbol{\Sigma}}_m)\mathbf{x}_n}{\sum_{n=1}^N \phi(\mathbf{x}_n;\boldsymbol{\mu}_m,\bar{\boldsymbol{\Sigma}}_m)}.$$

Lastly, the empirical covariances $\hat{\boldsymbol{\Sigma}}_{w,m}$ and $\hat{\boldsymbol{\Sigma}}_{\phi,m}$ are given by:

$$\hat{\boldsymbol{\Sigma}}_w \triangleq \frac{\sum_{n=1}^N w(\mathbf{x}_n;h)\gamma_m(\mathbf{x}_n;\boldsymbol{\theta}')(\mathbf{x}_n-\boldsymbol{\mu}_m)(\mathbf{x}_n-\boldsymbol{\mu}_m)^T}{\sum_{n=1}^N w(\mathbf{x}_n;h)\gamma_m(\mathbf{x}_n;\boldsymbol{\theta}')}$$

$$\hat{\boldsymbol{\Sigma}}_{\phi,m} \triangleq \frac{\sum_{n=1}^N \phi(\mathbf{x}_n;\boldsymbol{\mu}_m,\bar{\boldsymbol{\Sigma}}_m)(\mathbf{x}_n-\boldsymbol{\mu}_m)(\mathbf{x}_n-\boldsymbol{\mu}_m)^T}{\sum_{n=1}^N \phi(\mathbf{x}_n;\boldsymbol{\mu}_m,\bar{\boldsymbol{\Sigma}}_m)}.$$

The solutions of (16) are obtained via fixed-point iterations.

4.3. Choice of the kernel's bandwidth parameter

The bandwidth parameter h of the kernel function (13) plays an important role in the estimation accuracy of the nominal GMM. Here, we develop a data-driven procedure for selection of this parameter. We begin by introducing the MISE between the underlying distribution density $g_{\mathbf{x}}(\cdot)$ and the estimated one $f_{\mathbf{x}}(\cdot;\hat{\boldsymbol{\theta}}_h)$. This quantity takes the form:

$$\mathcal{I}[g_{\mathbf{x}}(\cdot), f_{\mathbf{x}}(\cdot;\hat{\boldsymbol{\theta}}_h)] \triangleq \mathbb{E}\left[\int_{\mathbb{R}^p}\left(g_{\mathbf{x}}(\mathbf{r}) - f_{\mathbf{x}}(\mathbf{r};\hat{\boldsymbol{\theta}}_h)\right)^2 d\lambda(\mathbf{r}); P_{\hat{\boldsymbol{\theta}}_h}\right]. \tag{17}$$

Hence, under the contaminated model in (10), since the contamination ratio parameter ϵ is assumed to be small, we conclude that

$$\mathcal{I}[g_{\mathbf{x}}(\cdot), f_{\mathbf{x}}(\cdot;\hat{\boldsymbol{\theta}}_h)] \approx \mathcal{I}[f_{\mathbf{x}}(\cdot;\boldsymbol{\theta}_0), f_{\mathbf{x}}(\cdot;\hat{\boldsymbol{\theta}}_h)].$$

This property justifies the use of (17) as a valid criterion for selection of h, which is strongly related to the estimation accuracy of the nominal density $f_{\mathbf{x}}(\cdot;\boldsymbol{\theta}_0)$. Minimization of the MISE (17) w.r.t. h amounts to minimization of the following objective function:

$$\begin{aligned}
\mathcal{V}(h) \triangleq -2\mathbb{E}&\left[\int_{\mathbb{R}^p} f_{\mathbf{x}}(\mathbf{r},\hat{\boldsymbol{\theta}}_h) dG_{\mathbf{x}}(\mathbf{r}); P_{\hat{\boldsymbol{\theta}}_h}\right] \\
&+\mathbb{E}\left[\int_{\mathbb{R}^p} f_{\mathbf{x}}^2(\mathbf{r};\hat{\boldsymbol{\theta}}_h) d\lambda(\mathbf{r}); P_{\hat{\boldsymbol{\theta}}_h}\right].
\end{aligned} \tag{18}$$

Clearly, the objective function (18) is not available, and therefore, an empirical estimate should be used instead. To obtain this estimate, we replace the integral term in the first summand in (18) by an empirical average $\frac{1}{N}\sum_{n=1}^N f_{\mathbf{x}}(\mathbf{x}_n;\hat{\boldsymbol{\theta}}_h)$. The expectations w.r.t. the probability distribution of $\hat{\boldsymbol{\theta}}_h$ can then be estimated via leave-one-out cross-validation [32], i.e., by averaging over a sequence of estimators $\hat{\boldsymbol{\theta}}_{h,-i}$, $i = 1, \ldots, N$, where $\hat{\boldsymbol{\theta}}_{h,-i}$ is the M$\mathcal{K}$DE obtained after excluding the i-th data sample $\mathbf{x}_i$. Nevertheless, this approach may be troublesome since it requires N applications of the $\mathcal{K}$-BM. Therefore, we use instantaneous approximation instead, i.e., the expectations are replaced by their input variables. The resulting instantaneous stochastic approximation of (18) takes the form:

$$\hat{\mathcal{V}}(h) \triangleq -\frac{2}{N}\sum_{n=1}^N f_{\mathbf{x}}(\mathbf{x}_n;\hat{\boldsymbol{\theta}}_h) + \int_{\mathbb{R}^p} f_{\mathbf{x}}^2(\mathbf{r};\hat{\boldsymbol{\theta}}_h) d\lambda(\mathbf{r}). \tag{19}$$

To conclude, the proposed selection rule for the bandwidth parameter h is given by:

$$h_{\text{opt}} \triangleq \arg\min_{h\in I} \hat{\mathcal{V}}(h), \tag{20}$$

where I is some predefined search interval.

5. NUMERICAL EXAMPLES

In this section, the estimation performance of the proposed $\mathcal{K}$-BM are compared to those of the non-robust EM [12] and to the following robust alternatives: t-EM [19], TCLUST [26] and OTRIMLE [21].

Simulation settings: We considered synthetic data generated from the contaminated GMM (10). The sample size, dimensionality and model order were set to $N = 300$, $p = 10$ and $M = 3$, respectively. The nominal mixing proportions and mean vectors were set to $\alpha_{0,1} = \alpha_{0,2} = 0.3$, $\alpha_{0,3} = 0.4$, $\boldsymbol{\mu}_{0,1} = 2 \cdot \mathbf{1}$, $\boldsymbol{\mu}_{0,2} = 10 \cdot \mathbf{1}$ and $\boldsymbol{\mu}_{0,3} = \boldsymbol{\mu}_{0,1} + 6 \cdot \mathbf{v}$, where $\mathbf{1}$ denotes a vector with unit values and $[\mathbf{v}]_i = (-1)^{i-1}$. The nominal covariance matrices $\boldsymbol{\Sigma}_{0,1}$ and $\boldsymbol{\Sigma}_{0,3}$ were considered to have Toeplitz structures, i.e., $[\boldsymbol{\Sigma}_{0,m}]_{i,j} = \sigma_m^2 b_m^{|i-j|}$, $m = 1, 3$, where $b_1 = 0.4$, $b_3 = 0.6$, $\sigma_1^2 = 5$ and $\sigma_3^2 = 3$. The matrix $\boldsymbol{\Sigma}_{0,2}$ was set to $3 \cdot \mathbf{I}_p$. Here, the contaminating distribution $Q_{\mathbf{x}}$ in (10) was considered to be a 2-order GMM with mixing proportions $\alpha_{c,1} = \alpha_{c,2} = 0.5$, mean vectors $\boldsymbol{\mu}_{c,1} = 10 \cdot (\mathbf{1} - \mathbf{v})$, $\boldsymbol{\mu}_{c,2} = -20 \cdot (\mathbf{1} - 0.25\mathbf{v})$ and Toeplitz covariances $[\boldsymbol{\Sigma}_{c,m}]_{i,j} = \sigma_c^2 r^{|i-j|}$, $m = 1, 2$, with $\sigma_c^2 = 10$ and $r = 0.9$.

Implementation details: Full MATLAB implementation of the $\mathcal{K}$-BM is available in [33]. We used R-code implementations of the t-EM, TCLUST and OTRIMLE that were provided by the authors. For the standard EM, a publicly available MATLAB implementation was applied. In all compared algorithms, the model order was set to its true value. All compared algorithms, excluding TCLUST, were initialized by the K-medoids algorithm [34]. The TCLUST, which does not enable user initialization, applies N random initializations. The maximum number of iterations was set to 50. Furthermore, the maximum number of inner fixed-point iterations conducted in each M-step of the $\mathcal{K}$-BM (to obtain the solution of (16)) was also fixed to 50. The bandwidth parameter h of the Gaussian kernel function (13) was selected according to (20). The minimization therein was carried out over 40 equally spaced grid points of the interval $[1, 20]$.

Results: First, we visually inspect the output of the proposed $\mathcal{K}$-BM under a fixed contamination ratio $\epsilon = 0.1$. To that sake, Fig. 1(a) provides a scatter plot of the first two coordinates x_1 and x_2 of the generated snapshots. On top of this plot, we mark the corresponding first two coordinates of the estimated mean vectors. Bisections of the concentration hyper-ellipses, associated with the estimated covariances are also drawn. One sees that the proposed $\mathcal{K}$-BM provides accurate estimates that are merely affected by the contamination. Scatter plots of other coordinates are available through [33]. In Fig. 1(b), we examine the relation between the bandwidth selection objective $\hat{\mathcal{V}}(h)$ (19) and the squared root of the empirical version of the MISE $\mathcal{I}[f_{\mathbf{x}}(\cdot; \boldsymbol{\theta}_0), f_{\mathbf{x}}(\cdot; \hat{\boldsymbol{\theta}}_h)]$ for several values of the bandwidth parameter h. The first quantity was computed based on a single realization of $N = 300$ snapshots, while the latter was obtained by averaging over 10^4 independent Monte-Carlo trials. Also here, the contamination ratio parameter $\epsilon = 0.1$. One sees that $\hat{\mathcal{V}}(h)$ accurately predicts the minimum point of the empirical MISE.

Second, we compared the empirical MISEs (w.r.t. the nominal density $f_{\mathbf{x}}(\cdot; \boldsymbol{\theta}_0)$) of all examined methods versus the contamination ratio parameter ϵ. All empirical MISE curves were obtained by averaging over 10^4 independent Monte-Carlo trials. In Fig. 2, one sees that the proposed $\mathcal{K}$-BM outperforms all compared methods. This performance gap is a consequence of the property that the $\mathcal{K}$-BM involves intrinsic model-free data weighting, which is automatically tuned by optimizing a performance related objective.

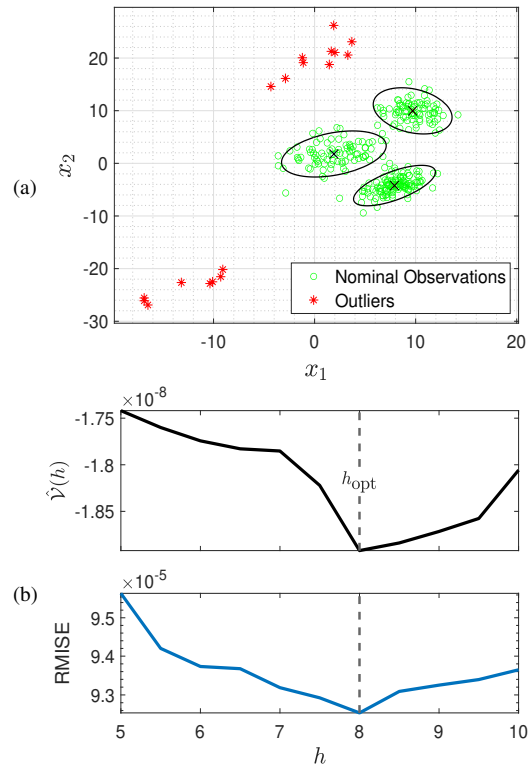

Fig. 1: (**a**) Scatter plot of the first two coordinates x_1 and x_2. The "X" marks and the ellipses represent the mean vectors and the covariance matrices estimated by the $\mathcal{K}$-BM, respectively. (**b**) The bandwidth selection objective $\hat{\mathcal{V}}(h)$ (19) (**Top**) and the root empirical MISE (**Bottom**), versus the bandwidth parameter h. The dashed vertical line represents h_{opt} (20).

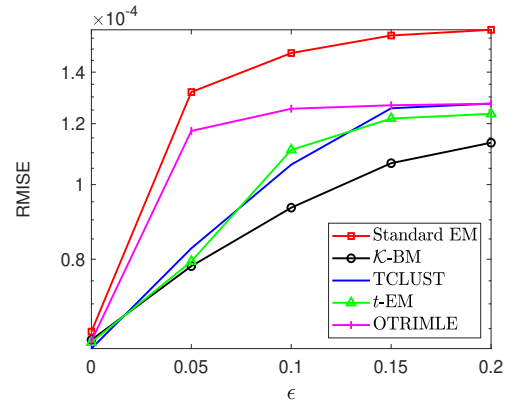

Fig. 2: The root empirical MISEs of the examined estimators versus the contamination ratio parameter ϵ.

6. CONCLUSION

In this paper, a new EM-like scheme, called $\mathcal{K}$-BM, for iterative numerical computation of the M$\mathcal{K}$DE was developed. The $\mathcal{K}$-BM was successfully applied to robust GMM parameter estimation. Unlike other robust estimators, the $\mathcal{K}$-BM does not alter the nominal GMM assumption, nor does it assume a specific contamination model. Furthermore, in the $\mathcal{K}$-BM, the bandwidth parameter of the involved kernel function is automatically tuned to optimize a performance related objective. Simulation studies confirmed that these properties lead to significant performance advantage.

7. REFERENCES

[1] P. D. McNicholas, *Mixture model-based classification,* Chapman and Hall/CRC, 2016.

[2] G. J. McLachlan, S. X. Lee and S. I. Rathnayake, "Finite mixture models," *Annual Review of Statistics and its Application,* vol. 6, pp. 355-378, 2019.

[3] K. Todros and J. Tabrikian, "Blind separation of independent sources using Gaussian mixture model," *IEEE Transactions on Signal Processing,* vol. 55, pp. 3645-3658, 2007.

[4] F. Wang and F. Liao, and Y. Li and H. Wang, "A new prediction strategy for dynamic multi-objective optimization using Gaussian Mixture Model," *Information Sciences,* vol. 580, pp. 331-351, 2021.

[5] N. Bougulia and F. Wentao, *Mixture models and applications,* Springer, 2020.

[6] Y. Wang, T. Adali, S.Y. Kung and Z. Szabo, "Quantification and segmentation of brain tissues from MR images: a probabilistic neural network approach," *IEEE Transactions on Image Processing,* vol. 7, pp. 1165-1181, 1998.

[7] M. Novey and T. Adali, "Complex fixed-point ICA algorithm for separation of QAM sources using Gaussian mixture model," *IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP),* 2007.

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

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

[10] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B,* vol. 39, no. 1, pp. 1-22, 1977.

[11] R. Harpaz and R. Haralick, "The EM algorithm as a lower bound optimization technique," *CUNY Ph. D. Program in Computer Science Technical Reports,*, pp. 1-14, 2006.

[12] Fraley, C. and Raftery A. E, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association,* vol. 97, no. 458, pp. 611-631, 2002.

[13] Christian Hennig, "Breakdown points for maximum likelihood estimators of location-scale mixtures," *The Annals of Statistics,* vol. 32, no. 4, pp. 1313-1340, 2004.

[14] B. R. Clarke and C. R. Heathcote, "Robust estimation of k-component univariate normal mixtures," *Annals of the Institute of Statistical Mathematics,* vol. 46, no. 11, pp. 83-93, 1994.

[15] 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.

[16] 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.

[17] Y. Qin and C. E. Priebe, "Maximum Lq-likelihood estimation via the expectation-maximization algorithm: a robust estimation of mixture models," *Journal of the American Statistical Association,* vol. 108, no. 503, pp. 914-928, 2013.

[18] D. Peel and G. J. McLachlan, "Robust mixture modelling using the t-distribution," *Statistics and Computing,* vol. 10, no. 4, pp. 339-348, 2000.

[19] J. L. Andrews and P. D. McNicholas, "Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions," *Statistics and Computing,* vol. 22, no. 5, pp. 1021-1029, 2012.

[20] P. Coretto and C. Hennig, "Robust improper maximum likelihood: tuning, computation, and a comparison with other methods for robust Gaussian clustering," *Journal of the American Statistical Association,* vol. 111, no. 516, pp. 1648-1659, 2016.

[21] P. Coretto and C. Hennig, "Consistency, breakdown robustness, and algorithms for robust improper maximum likelihood clustering," *Journal of Machine Learning Research,* vol. 18, no. 142, pp. 1-39, 2017.

[22] M. T. Gallegos, "Maximum likelihood clustering with outliers," *Classification, Clustering, and Data Analysis,* pp. 247-255, 2002.

[23] M. T. Gallegos and G. Ritter, "Robust method for cluster analysis," *The Annals of Statistics,* vol. 33, no. 1, pp. 347-380, 2005.

[24] M. T. Gallegos and G. Ritter, " Trimmed ML estimation of contaminated mixtures," *Sankhyā: The Indian Journal of Statistics,* pp. 164-220, 2009.

[25] L. A. García-Escudero *et. al.*, "A general trimming approach to robust cluster analysis," *The Annals of Statistics,* vol. 36, no. 3, pp. 1324-1345, 2008.

[26] L. A. García-Escudero *et. al.*, "An R package for a trimming approach to cluster analysis," *Journal of Statistical Software,* vol. 47, pp. 1-16, 2012.

[27] Y. Sorek and K. Todros, "On the minimum $\mathcal{K}$-divergence estimator," *IEEE Transactions on Signal Processing*, vol. 70, pp. 4337-4352, 2022.

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

[29] C. F. Jeff Wu, "On the convergence properties of the EM algorithm," *The Annals of Statistics,* 22, pp. 95-103, 1983.

[30] J. R. Magnus and H. M. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics (3rd edition)*, Wiley and Sons, 2007.

[31] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods,* New York: Academic Press, 1982.

[32] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning,* pp. 600-601, Springer, 2011.

[33] O. Kenig and K. Todros, "Robust GMM parameter estimation via the $\mathcal{K}$-BM algorithm: MATLAB implementation," *Available online at:* https://github.com/OriKenig1/K_BM

[34] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis,* John Wiley and Sons, 2009.