Multivariate Parametric Density Estimation Based On The Modified Cramér-von Mises Distance

Peter Krauthausen, Henning P. Eberhardt, and Uwe D. Hanebeck

***Abstract*— In this paper, a novel distance-based density estimation method is proposed, which considers the overall density function in the goodness-of-fit. In detail, the parameters of Gaussian mixture densities are estimated from samples, based on the distance of the cumulative distributions over the entire state space. Due to the ambiguous definition of the standard multivariate cumulative distribution, the Localized Cumulative Distribution and a modified Cramér-von Mises distance measure are employed. A further contribution is the derivation of a simple-to-implement optimization procedure for the optimization problem. The proposed approach's good performance in estimating arbitrary Gaussian mixture densities is shown in an experimental comparison to the Expectation Maximization algorithm for Gaussian mixture densities.**

I. Introduction

In many technical applications in robotics, computer vision, and machine learning, probabilistic information fusion is fundamental. Typical applications range from vehicle or person localization and tracking [1] to speech recognition and non-verbal communication using Hidden Markov Models [2] or Bayesian Networks [3]. Central to all these applications is the information fusion according to probabilistic models given in the form of conditional densities, i.e., $f(y|x) = \frac{f(x,y)}{f(x)}$. There are essentially three ways how these densities can be obtained: (a) domain knowledge, i.e., an expert quantifies the uncertainty, (b) a functional dependency underlying the density in conjunction with a noise description is given, and (c) samples of a random variable are given and the density is estimated from these samples only. In this paper, the latter problem is addressed for estimating continuous density functions, as commonly used in probabilistic information fusion, from sparse sets of samples.

Density estimation methods can be categorized [4, p. 33] into non-parametric [4], [5], [6] and parametric approaches [7]. The most prominent non-parametric method is the Parzen-Window approach [8], also known as kernel density estimation. There, a density function is obtained from placing a kernel at each sample point and then optimizing the kernels' parameters. This approach to density estimation suffers from the fact that all data points are stored in the density function representation. This renders the use of these densities for many applications in information fusion, but especially for recursive state estimation, impractical without further sparsification.

In contrast, in parametric density estimation the data is assumed to be generated from a specific, typically sparse model and density estimation corresponds to estimating this model's parameters. Typically, finite mixture densities are estimated [7] by *maximizing the likelihood* of the data employing the Expectation Maximization (EM) algorithm [9]. This means that, e.g., a Gaussian mixture density's weights, means, and covariances are estimated. Both frequentistic and Bayesian approaches to parametric density estimation exist [7], [10]. Yet, this advantage comes along with new challenges: overfitting, singularities, and model selection. Overfitting, as the overconfident estimation of a density, may be alleviated by penalizing roughness [7]. Singularities caused by the coincidence of a sample position with a component position may be avoided by a good initialization of the algorithm. Regarding model selection, there exists a wealth of criteria to determine the appropriate number of components, e.g., Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or Minimum Message Length (MML). Typically, parametric models are determined by evaluating the density's fit at the sample points only [9]. The goodness-of-fit is not tested over the entire density, but only at distinct points.

In this paper, an approach is proposed that estimates densities based on the minimization of a distance over the entire state space and not at distinct points only. For this reason, a squared integral distance of distributions is employed. Since for the multivariate case, the definition of the cumulative distribution function is not well defined, the Localized Cumulative Distributions [11], [12] and the corresponding modified Cramér-von Mises distance measure are employed. The results of the proposed method are Gaussian mixture densities with arbitrary weights, means, and covariances. In contrast to non-parametric approaches, the components are not centered about the samples.

The rest of this paper is structured as follows. Initially, the mathematical problem formulation of the density estimation problem is given. In Sec. III, the Localized Cumulative Distributions in accordance with [11], [12] are derived and in Sec. IV, the efficient minimization of the modified Cramér-von Mises distance measure is devised. In Sec. V, the overall algorithm is summarized and the algorithm's properties are discussed. After giving some advice towards an efficient implementation, the approach is validated by comparing it against EM in terms of estimation accuracy.

P. Krauthausen, H. P. Eberhardt, and U. D. Hanebeck are with the Intelligent Sensor-Actuator-Systems Laboratory (ISAS), Institute for Anthropomatics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany. peter.krauthausen@kit.edu, henning.eberhardt@kit.edu, uwe.hanebeck@ieee.org. This work was supported partially by theGe rman Research Foundation (DFG) within the Collaborative Research Center SFB 588 on "*Humanoid robots – Learning and Cooperating Multimodal Robots*".

II. Problem Definition

It is assumed that data $\mathcal{D}$ generated from an underlying density $\tilde{f}$ is interpreted as a Dirac mixture density, also known as the *empirical probability density function* [4]

$$f_{\mathcal{D}}(\underline{x}) = \sum_{i=1}^{|\mathcal{D}|} w_i \, \delta(\underline{x} - \underline{x}_i) \,, \tag{1}$$

with $\underline{x}_i := [x_i^{(1)} \; \dots \; x_i^{(N)}]^{\mathrm{T}} \in \mathbb{R}^N$, $\underline{x} \in \mathbb{R}^N$, identical weights for all components $w_i := 1/|\mathcal{D}|$ and $\delta(.)$, the Dirac distribution. The problem addressed in this paper is the determination of a density f_{GM} having minimal distance to the *true* underlying density function $\tilde{f}$ that generated the samples. Since $\tilde{f}$ is not accessible, the distance is approximated by using the distance to $f_{\mathcal{D}}$ instead of $\tilde{f}$. Throughout this paper, a solution in the form of mixtures of multivariate normal distributed densities is sought

$$f_{\mathrm{GM}}(\underline{x}) = \sum_{i=1}^{M} \alpha_i \, \mathcal{N}(\underline{x} - \underline{\mu}_i, \Sigma_i) \,, \tag{2}$$

with

$$\mathcal{N}(\underline{x} - \underline{\mu}, \Sigma) = \tfrac{1}{\sqrt{|2\pi\Sigma|}} \, \exp\left[-\tfrac{1}{2}(\underline{x} - \underline{\mu})^{\mathrm{T}} \Sigma^{-1} (\underline{x} - \underline{\mu})\right]$$

and $\sum_{i=1}^{M} \alpha_i = 1$, $0 \leq \alpha_i \leq 1$, mean vector $\underline{\mu} \in \mathbb{R}^N$, covariance matrix Σ, and $|\,.\,|$ denoting the determinant. Estimating f_{GM} corresponds to determining the weights α_i, means $\underline{\mu}_i$, and covariance matrices Σ_i for all components $0 \leq i \leq M$. In order to avoid the pitfall of comparing probabilities at a set of distinct points, the distance between the cumulative distributions of f_{GM} and $f_{\mathcal{D}}$ is considered.

III. Localized Cumulative Distributions (LCD)

The cumulative distribution is commonly used for comparing univariate densities. The standard cumulative distribution for multivariate density functions is non-unique as the definition allows for several directions of integration. For N dimensions a total of 2^N possible functions exist [11]. Even worse is the fact, that these cumulative distributions are non-symmetric, causing the estimates to be biased w.r.t. the choice of the cumulative distribution function [11]. The application of the LCD [11], as an alternative representation, resolves this issue. For the sake of self-containedness, the following definition is restated from [11].

Definition 1 (Localized Cumulative Distribution, [11]): Given a random vector $\mathbf{x} \in \mathbb{R}^N$ and the corresponding probability density function $f(\underline{x}) : \mathbb{R}^N \to \mathbb{R}_+$. The Localized Cumulative Distribution is defined as

$$F(\underline{m}, \underline{b}) = \int_{\mathbb{R}^N} f(\underline{x}) \cdot \mathcal{K}(\underline{x} - \underline{m}, \underline{b}) \, \mathrm{d}\underline{x} \tag{3}$$

with $\Omega \subset \mathbb{R}^N \times \mathbb{R}_+^N$, $F : \Omega \to [0,1]$, $\underline{b} \in \mathbb{R}_+^N$, $\mathcal{K}(\underline{x} - \underline{m}, \underline{b})$ a suitable kernel [11] centered at $\underline{m} = [\, m^{(1)} \cdots m^{(N)} \,]^T$ with width $\underline{b}$ and $\mathcal{K} : \Omega \to [0,1]$.

The above definition shows how the LCD is obtained by multiplying $f(\underline{x})$ with the kernel $\mathcal{K}$ and integrating over $\underline{x}$. In the rest of this paper, the following kernel is employed

$$\mathcal{K}(\underline{x} - \underline{m}, \underline{b}) = \sqrt{|2\pi\Sigma_b|} \, \mathcal{N}(\underline{x} - \underline{m}, \Sigma_b) \,,$$

typically with $\Sigma_b = \mathrm{diag}(\underline{1}b)$, i.e., identical width b for all dimensions. This yields the LCDs of (1) and (2) given by

$$F_{\mathcal{D}}(\underline{m}, b) = \sum_{i=1}^{L} w_i \sqrt{|2\pi\Sigma_b|} \, \mathcal{N}(\underline{x}_i - \underline{m}, \Sigma_b) \tag{4}$$

and

$$F_{GM}(\underline{m}, b) = \sum_{i=1}^{M} \alpha_i \, \sqrt{|2\pi\Sigma_b|} \, \mathcal{N}(\underline{\mu}_i - \underline{m}, \Sigma_i + \Sigma_b) \,. \tag{5}$$

IV. Minimizing the Distance Measure

As a measure of fit, the squared integral distance of the cumulative distributions is employed. For this reason, the modified Cramér-von Mises distance measure (mCvMD) [11], between the LCDs of the Dirac mixture (4) and the Gaussian mixture density (5), is calculated. This distance is given by

$$D = \int_{\mathbb{R}^+} w(b) \int_{\mathbb{R}^N} \left(F_{\mathrm{GM}}(\underline{m}, b) - F_{\mathcal{D}}(\underline{m}, b)\right)^2 \, \mathrm{d}\underline{m} \, \mathrm{d}b \,. \tag{6}$$

Similar to [11], the function $w(.)$ is selected as

$$w(b) = \begin{cases} \frac{1}{b^{N-1}} & b \in [0, b_{\max}] \\ 0 & \text{elsewhere} \end{cases} ,$$

as it ensures convergence of the integral.

Evaluation of the distance measure

Minimization of (6) requires the solution of integrals about the kernel position $\underline{m}$ and kernel width b. Closed-form solutions to the $\underline{m}$-integrals in (6) exist

$$D = \int_{\mathbb{R}^+} w(b) \, |2\pi\Sigma_b| \, \left(\mathrm{P}_1 - 2\,\mathrm{P}_2 + \mathrm{P}_3\right) \, \mathrm{d}b \,, \tag{7}$$

where $\mathrm{P}_1 - \mathrm{P}_3$ denote the solutions to the $\underline{m}$-integral over the kernel positions, for each summand of the resolved term $\left(F_{\mathrm{GM}}(\underline{m}, b) - F_{\mathcal{D}}(\underline{m}, b)\right)^2$ in (6). In particular, these are

$$\mathrm{P}_1 = \sum_{i=1}^{M} \sum_{j=1}^{M} \alpha_i \, \alpha_j \, \mathcal{N}(\underline{\mu}_i - \underline{\mu}_j, 2\Sigma_b + \Sigma_i + \Sigma_j) \,,$$

$$\mathrm{P}_2 = \sum_{i=1}^{M} \sum_{j=1}^{L} \alpha_i \, w_j \, \mathcal{N}(\underline{\mu}_i - \underline{x}_j, 2\Sigma_b + \Sigma_i) \,,$$

$$\mathrm{P}_3 = \sum_{i=1}^{L} \sum_{j=1}^{L} w_i \, w_j \, \mathcal{N}(\underline{x}_i - \underline{x}_j, 2\Sigma_b) \,.$$

Note that P_3 is irrelevant for the minimization of D, as it is constant w.r.t. the parameters of the Gaussian mixture density to be determined. Yet, it needs to be calculated to obtain the absolute value of the distance measure, e.g., for experimental comparisons. For the integral in (6), which ranges over the kernel width b, no closed-form solution

is known up to now. Numerical integration is required to calculate D in (7). In summary, the amount of computation necessary, is governed by $(M \cdot L + M \cdot M) \cdot e \cdot s$ evaluations of a multivariate Gaussian density. Here, M and L are defined as in (4) and (5), e are the number of evaluation points used by the numerical integration and s are the number of steps to convergence. In the next section, the overall algorithm is described in detail.

V. Algorithm

The proposed algorithm is an easy-to-implement optimization scheme, consisting of two steps:

1) Choose starting parameters θ;
2) **while** (gradient(θ) $\neq \underline{0}$ and D too large)
   change θ along search direction;

In step 1), the number of Gaussian mixture density components M needs to be determined. This is a well known problem, when estimating finite mixture models. Typically, criteria like AIC, BIC, or MML are employed to solve this problem [7], [10], [13]. Besides selecting M, an initial parameter set for each component $\{\alpha_i,\, \underline{\mu}_i,\, \Sigma_i\}$ for $1 \leq i \leq M$ has to be selected. Identical weights $\alpha_i = \frac{1}{M}$ are appropriate. The mean μ_i and covariance matrix Σ_i of each component are chosen by the following, more elaborate method.

First, $M \cdot (N+1) \leq |\mathcal{D}|$ points representing the given N-dimensional data $f_{\mathcal{D}}$ are chosen using either the cluster centers of k-means or, as a deterministic approximation of the data, a Dirac mixture reduction, as proposed in [14]. Random selection is an option too, but the proposed methods provide a better coverage of the dataset.

In order to obtain the parameters μ_i and Σ_i the points are randomly assigned to $(N+1)$-tuples $P_i := \{P_{i,1}, ..., P_{i,N+1}\}$ with $1 \leq i \leq M$. Using the generated $(N+1)$-tuples, the values of μ_i and Σ_i are set to the sample mean and covariance of the respective tuple P_i. By construction, all points lie on the the same covariance ellipsoid of the associated Gaussian component of the mixture. This method is intuitive and guarantees valid covariance matrices for different $P_{i,j}$, which will come in very handy in the next section. We refer the interested reader to [7] for a description of alternative initialization methods.

In step 2), a quasi-Newton optimization procedure (limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm [15]) is used for minimizing the distance measure. This algorithm estimates the Hessian matrix and therefore allows for high-dimensional optimization. This optimization method requires a gradient, which is calculated numerically. In order to evaluate (7), a numerical integration needs to be performed too. Here, the use of an adaptive Lobatto Quadrature is proposed as the distance measure is reasonably smooth when integrated over b.

In order to obtain a valid probability density from the optimization process, several constraints have to be fulfilled:

- *All Σ_i need to be valid:*
  Constraint is achieved by optimizing P_i and calculation the sample statistics.
- *All weights must sum to one, $\sum_{i=1}^{M} \alpha_i = 1$, and $f_{GM}(\underline{x}) \geq 0$ for all $x \in \Omega$:*
  Minimizing mCvMD penalizes deviations in probability mass between $f_{\mathcal{D}}$ and f_{GM}. These constraints are automatically enforced due to $\int_{\mathbb{R}^N} f_{\mathcal{D}}(\underline{x})\, \mathrm{d}\underline{x} = 1$.

Finally, the occurrence of singular covariance matrices may be prevented by adding a penalty for very small distances within the $(N+1)$-tuples P_i. This may be understood as lower-bounding the allowed covariance dimensions. Note, for all of the following experiments, no constraint was necessary.

VI. Experiments

In this section, an experimental comparison of the proposed approach to the standard parametric density estimation method EM for Gaussian mixture densities is provided.

Geyser Data Set

As a benchmark data set, a rescaled version of the *Old Faithful Geyser* data obtained from the website of [13] is used. It consists of 272 data points. Each point corresponds to the duration time of the current and waiting time until the next eruption of a Geyser. The density estimate obtained by the proposed approach is depicted in Fig. 1 (a). The red points are the original samples and the contours show the obtained Gaussian mixture density. In Fig. 1 (b), the Gaussian mixture density estimated by EM is given. These results show that EM and the presented approach yield *visually* very similar results.

Two Components - Weakly and Strongly Overlapping

For these experiments, samples were drawn from a Gaussian mixture density (2) with $M = 2$ components,

$$\underline{\alpha} = [0.5\,;\ 0.5]^{\mathrm{T}}, \qquad \underline{\mu}_1 = [\beta \cdot 0.35\,;\ 0.0]^{\mathrm{T}} = (-1) \cdot \underline{\mu}_2\,,$$

and

$$\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \begin{bmatrix} 0.05 & 0.0 \\ 0.0 & 3.0 \end{bmatrix}.$$

The samples were generated for $\beta = \{1, 2\}$. The true density, the respective samples and the estimated densities using the proposed approach and EM for both β are given in Fig. 1 (c-h). As can be seen in Fig. 1 (c-e) and Tab. I, the proposed algorithm can estimate the Gaussian mixture density with their distinct clusters well. In contrast to the well separable case, the visual results in Fig. 1 (f-h) as well as the results in Tab. I show the increasing difficulty for both algorithms.

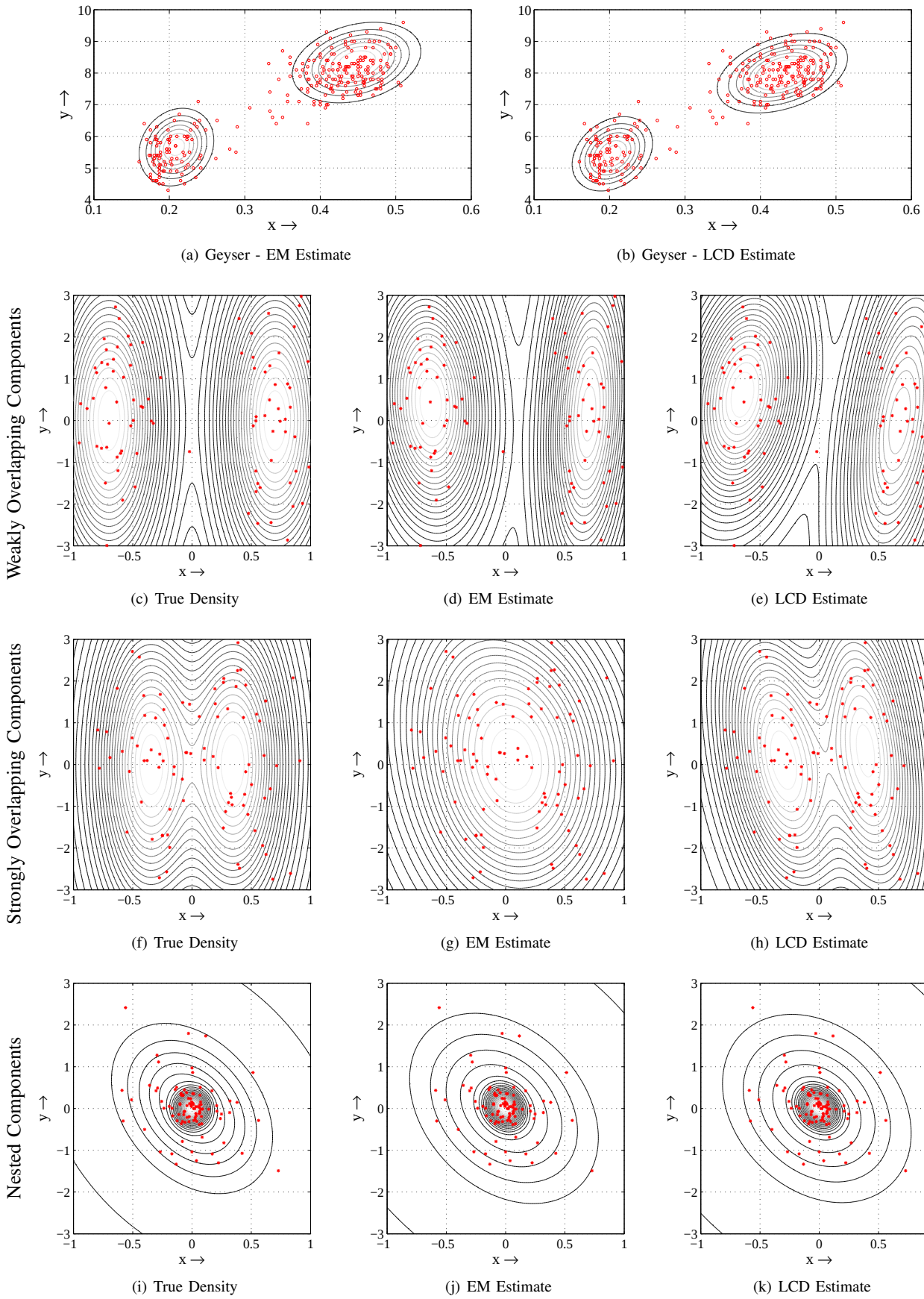
Coinciding means

To test the performance of the algorithm in estimating Gaussian mixture densities with identical means, samples were generated using the following density with $M = 2$

$$\underline{\alpha} = [0.5\,;\ 0.5]^{\mathrm{T}}, \qquad \underline{\mu}_1 = \underline{\mu}_2 = [0.0\,;\ 0.0]^{\mathrm{T}},$$

and

$$\mathbf{\Sigma}_1 = \begin{bmatrix} 6 & -2 \\ -2 & 6 \end{bmatrix}, \qquad \mathbf{\Sigma}_1 = \begin{bmatrix} 0.5 & 0.0 \\ 0.0 & 0.5 \end{bmatrix}.$$



(a) Geyser - EM Estimate (b) Geyser - LCD Estimate

(c) True Density (d) EM Estimate (e) LCD Estimate

(f) True Density (g) EM Estimate (h) LCD Estimate

(i) True Density (j) EM Estimate (k) LCD Estimate

Fig. 1. Samples and obtained density estimation results for the Geyser Benchmark data set, weakly and strongly overlapping two components and two components with identical means. More explanation can be found in the respective parts of Sec. VI.

The mixture contains two components with identical means, but the first component's covariance is more conservative than the second component's in all dimensions, i.e., the variances are larger in all directions. The drawn samples and the estimated densities using the proposed approach are given in Fig. 1 (i-k). This experiment is hard due to overlapping components, as the densities have identical means. The proposed approach still produces good results.

Statistics

In this section, the statistics for the experiments described above and depicted in Fig. 1 are reported. The estimation performance of the proposed LCD method shall be compared to EM method. For the comparison, the mCvMD described in Sec. IV and the *Kolmogorov-Smirnov* distance (KSD) are employed. As was shown in [12], the mCvMD is well suited for comparing sample densities to continuous densities. The KSD is employed to show the performance using a well known benchmark distance.

Experiment 1 - MCvMD: The mCvMD results for the artificial data sets described above are given in Tab. I. The results are averages over 330 experiments with varying samples size. The sample sizes ranged from 10 to 40 per experiment. The results in Tab. I are given in the form of mean distances μ and standard deviations σ averaged over all respective experiments. Tab. I (d) gives the results averaged over all experiments. In each table, the distances in the mCvMD sense between the given samples, the Gaussian mixture density, obtained using the LCD and the EM method, is given w.r.t. the given samples and the true generating mixture density.

Regarding the distance between the estimates and the *true underlying density*, a strong correlation of the distances of both mixture estimates and the distance between the training samples and the true density can be observed. Yet, no approach is consistently better than the other. The proposed approach yields comparable distances to the true density on average. Regarding the distance of the LCD method's estimates to the *training samples*, in all experiments, the proposed approach achieves lower distances than EM. Note, that the lower average distances have lower standard-deviations than EM, too. This holds for all data sets.

Experiment 2 - KSD: In the previous experiment, the mCvMD was used for the comparison. This result might have been expected, due to the fact that the minimized distance measure was used for the comparison. In order to perform a *neutral* comparison, the Kolmogorov-Smirnov-Test and the Kolmogorov-Smirnov-Distance (KSD) are employed, which is a commonly used distance given by

$$D_{\mathrm{KS}} = \sup_{x \in \mathcal{D}} \left| F_{\mathcal{D}}(x) - F(x) \right| .$$

Note, the KSD is only defined for scalar random variables. For the presented two-dimensional examples, the cumulative distribution of the scalar marginal densities $f(x_1)$ and $f(x_2)$ are compared to the empirical distribution function.

In Fig. 2 (a-b), the average KSD of the experiments and in Fig. 2 (c-d), the standard deviation over the respective experiments for both marginal densities are given. In each subfigure, the distances between Gaussian mixture densities estimated using the LCD-based method and EM are given w.r.t. the given samples and the true generating mixture density. The results are averages over 240 experiments for each sample size, which are equally split among the three test cases. The sample sizes ranged from 10 to 100 per experiment. Fig. 2 (a-b) show that the average KSD to the samples is lower for the LCD method than for the EM method. This holds for all marginal densities for all experiments. As $f_{\mathcal{D}}$ is an estimate of the true density, the KSD between the data and the LCD method's result is almost always lower or at least comparable to EM, too. There is one exception, shown in Fig. 2 (b). In this case, the distance to the true density is slightly smaller for the EM estimate than for the LCD estimate. Yet, the standard deviation for this case is very high compared to the difference in distance.

TABLE I

MCvMD - STATISTICS TO THE EXPERIMENTS DESCRIBED IN SEC. VI.

(a) **MCvMD for Small Overlap**

	Samples		LCD		EM	
	μ	σ	μ	σ	μ	σ
Samples	-	-	0.0066	0.0028	0.013	0.0095
True	0.0930	0.0836	0.0878	0.0826	0.0821	0.0775

(b) **MCvMD for More Overlap**

	Samples		LCD		EM	
	μ	σ	μ	σ	μ	σ
Samples	-	-	0.0042	0.0028	0.0101	0.0077
True	0.0860	0.0821	0.0815	0.0818	0.0780	0.0808

(c) **MCvMD for Nested Components**

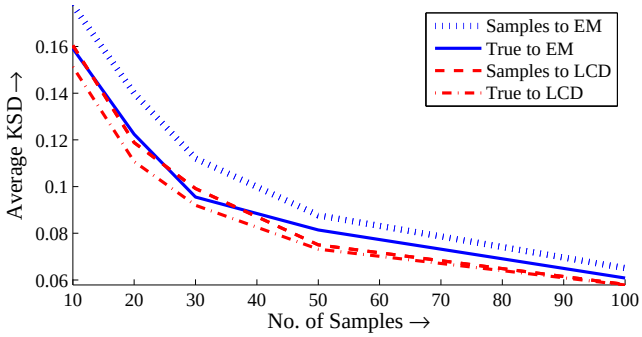
	Samples		LCD		EM	
	μ	σ	μ	σ	μ	σ
Samples	-	-	0.0159	0.0099	0.0291	0.0209
True	0.1817	0.1407	0.1670	0.1383	0.1807	0.1422

(d) **MCvMD for all Experiments**

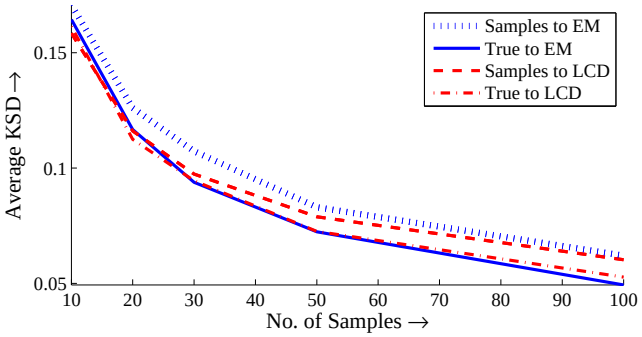
	Samples		LCD		EM	
	μ	σ	μ	σ	μ	σ
Samples	-	-	0.0089	0.0080	0.0174	0.0163
True	0.1204	0.1143	0.1122	0.1114	0.1137	0.1148

VII. CONCLUSIONS

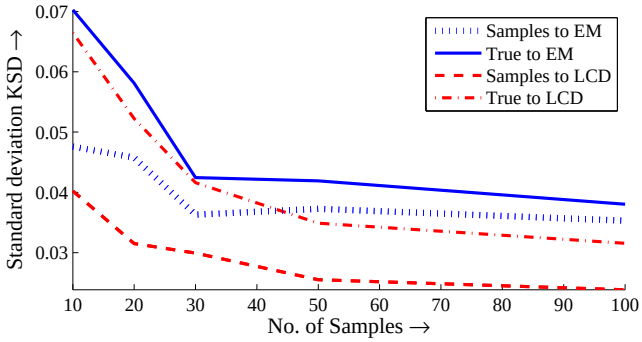
In this paper, a distance-based density estimation algorithm was presented, which considers the overall density function. The parameters of Gaussian mixture densities were determined by minimizing the distances of the Localized Cumulative Distribution of the data and the Gaussian mixture density. The ambiguity of the definition of the standard multivariate cumulative density function was removed by using this distribution and the modified Cramér-von Mises distance. The proposed optimization is easy to implement and experimental comparison to EM shows the excellent performance of the proposed approaches. Regarding the modified Cramér-von Mises distance and the Kolmogorov-Smirnov-Test, the proposed approach yields on average better Gaussian mixture densities with respect to the training samples than EM. Not only is the mean distance closer, but the standard deviations are drastically smaller, i.e., the proposed



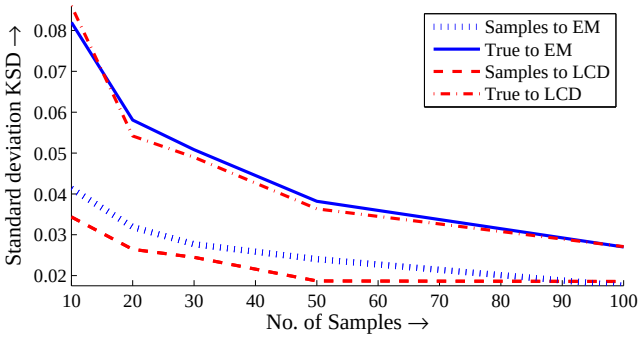
(a) Marginal density of $\mathbf{x}$.



(b) Marginal density of $\mathbf{y}$.



(c) Marginal density of $\mathbf{x}$.



(d) Marginal density of $\mathbf{y}$.

Fig. 2. Average KSD (a-b) of the experiments and standard deviations (c-d) averaged over the respective experiments for both marginal densities are given. In each plot, the distances between the LCD-estimated and the EM-estimated Gaussian mixture density are given with respect to the given samples and the true generating mixture density.

method should be preferred over EM when reliability is important.

It remains future work to investigate consistency and (probabilistic) error bounds. With regard to small sample sizes, the introduction of prior knowledge into the density estimation process as well as the introduction of constraints on the state space appear promising. The former might be introduced by adding soft/hard constraints penalizing the deviation from a given function, e.g., a sine function. The latter would allow for the exclusion of particular areas and direct approximation of density slices, which are especially useful for filtering and prediction purposes. In addition, the execution time could be improved further by implementing an analytic gradient. The proposed procedure's capability to solve high-dimensional problems needs to be tested.

References

[1] P. Rößler, F. Beutler, U. D. Hanebeck, and N. Nitzsche, "Motion Compression Applied to Guidance of a Mobile Teleoperator," in *Proceedings of the 2005 IEEE International Conference on Intelligent Robots and Systems (IROS 2005)*, 2005, pp. 2495–2500.

[2] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, 1989.

[3] D. Koller, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, Cambridge, Massachusetts, 2009.

[4] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley series in probability and mathematical statistics - A Wiley Interscience publication. Wiley, New York, 1992.

[5] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Monographs on Statistics and Applied Probability ; 26. CRC Press, Boca Raton, 1998.

[6] P. B. Eggermont and V. N. LaRiccia, *Maximum Penalized Likelihood Estimation*, vol. 1: Density Estimation, Springer, New York, 2001.

[7] G. McLachlan and D. Peel, *Finite Mixture Models*, Wiley-Interscience, 2000.

[8] E. Parzen, "On Estimation of a Probability Density Function and Mode," *Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[9] 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 (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley & Sons, 2nd edition, 2000.

[11] U. D. Hanebeck and V. Klumpp, "Localized Cumulative Distributions and a Multivariate Generalization of the Cramér-von Mises Distance," in *Proceedings of the 2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2008)*, Seoul, Republic of Korea, Aug. 2008, pp. 33–39.

[12] U. D. Hanebeck, M. F. Huber, and V. Klumpp, "Dirac Mixture Approximation of Multivariate Gaussian Densities," in *Proceedings of the 2009 IEEE Conference on Decision and Control (CDC 2009)*, Shanghai, China, Dec. 2009, pp. 3851–3858.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Science+Business Media, LLC, 2006.

[14] H. Eberhardt, V. Klumpp, and U. D. Hanebeck, "Optimal Dirac Approximation by Exploiting Independencies," in *Proceedings of the 2010 American Control Conference (ACC 2010)*, Baltimore, Maryland, July 2010.

[15] Peihuang Lu, Jorge Nocedal, Ciyou Zhu, and Richard H. Byrd, "A Limited-Memory Algorithm for Bound Constrained Optimization," *SIAM Journal on Scientific Computing*, vol. 16, pp. 1190–1208, 1994.