Regularized Non-Parametric Multivariate Density and Conditional Density Estimation

Peter Krauthausen and Uwe D. Hanebeck

***Abstract*— In this paper, a distance-based method for both multivariate non-parametric density and conditional density estimation is proposed. The contributions are the formulation of both density estimation problems as weight optimization problems for Gaussian mixtures centered about samples with identical parameters. Furthermore, the minimization is based on the modified Cramér-von Mises distance of the Localized Cumulative Distributions, removing the ambiguity of the definition of the multivariate cumulative distribution function. The minimization problem is amended with a regularization term penalizing the densities' roughness to avoid overfitting. The resulting estimation problems for both densities and conditional densities are shown to be phrasable in the form of readily implementable quadratic programs. Experimental comparison against EM, SVR, and GPR based on the log-likelihood and performance in benchmark recursive filtering applications show high quality of the densities and good performance at less computational cost, i.e., the density representations are sparser.**

I. INTRODUCTION

Probabilistic information fusion is based on the availability of efficiently represented high-quality probabilistic models in form of densities or conditional densities. This is especially true for, but not limited to, the most important probabilistic graphical models: Bayesian Networks, Dynamic Bayesian Networks, and Hidden Markov Models [1], [2]. There are essentially two different ways of deriving probabilistic models: a given generative model of the phenomena at hand is compiled into a probabilistic model or, given i.i.d. random samples, the probabilistic models are estimated. In this paper, the latter problem is solved, i.e., given a set of samples, the corresponding density or conditional density is estimated. (Conditional) Density estimation may be categorized into parametric and non-parametric approaches, cf. [3], [4], [5], [6] for a detailed overview. In parametric density estimation, a compressed description of the data in form of a parametric model based on the data [7] is sought. The most prominent approach is the *maximization of the likelihood* of the samples given the parameters by application of the expectation maximization (EM) algorithm [8]. This approach has several drawbacks: the number of components in the mixture has to be chosen, singularities may occur, and the resulting densities are prone to overfit the data. In contrast, in non-parametric density estimation not a parameter estimate shall be determined, but the closest density estimate to the entire true density function [3]. The simplest non-parametric approach is kernel density estimation, i.e., mixture densities with components centered about data points are employed. The two main challenges to this approach are the determination of component parameters, which impact the densities' shape and smoothness, and the complex function representations entailing the entire data set. Especially for nonlinear filters and Bayesian Networks, densities with many components are too expensive to evaluate during online Bayesian inference.

In this paper, both challenges will be addressed for density and conditional density estimation. We propose the use of weighted kernel densities, i.e., Gaussian mixtures (GM) [9] with identical axis-aligned components, but variable weights. For minimizing the squared integral distance between the cumulative distributions of the empirical density function and the GM in conjunction with a penalization of the GM's smoothness, optimization problems in form of a quadratic program are derived. Solving the optimization problem results in sparse densities, which are well applicable for online inference, and smoothness w.r.t. the chosen regularization, and therefore less prone to overfitting. The rest of this paper is structured as follows. Initially, the mathematical problem formulation is given, which is common to both density and conditional density estimation. In Sec. III, density estimation and in Sec. IV, conditional density estimation is considered respectively. The similarities and specific differences between density and conditional density estimation will be considered in the latter section. In Sec. V, the choice of kernel parameters is discussed and in Sec. VI, the approach is tested against EM, Support Vector Regression (SVR) [10], [11], and Gaussian Process Regression (GPR) [12] on density estimation and conditional density.

II. MATHEMATICAL PROBLEM FORMULATION

Given a set of i.i.d. random samples $\underline{x}_i \in \mathcal{D}$, with $\underline{x}_i := [x_i^{(1)} \ \dots \ x_i^{(N)}]^{\mathrm{T}} \in \mathbb{R}^N$, represented in the form of a mixture of Dirac distributions $\delta(.)$, also known as the *empirical probability density function* [3],

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

the density function $\tilde{f}$ underlying the data shall be estimated. Disregarding whether $\tilde{f}$ is a density or conditional density, estimates are sought in form of an axis-aligned GM [9]

$$f_{\mathrm{GM}}(\underline{x}) = \sum_{i=1}^{M} \alpha_i \prod_{k=1}^{N} \mathcal{N}(x^{(k)} - \mu_i^{(k)}, \sigma_i^{(k)}) \,, \tag{2}$$

P. Krauthausen 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, uwe.hanebeck@ieee.org. This work was supported partially by the German Research Foundation (DFG) within the Collaborative Research Center SFB 588 on "*Humanoid robots – Learning and Cooperating Multimodal Robots*".

with components $\mathcal{N}(x-\mu,\sigma) := \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}$. The above mixture contains M components with one Gaussian for each of the N dimensions with mean $\mu_i^{(k)}$ and standard deviation $\sigma_i^{(k)}$. In f_{GM} the parameters are set to $\mu_i^{(k)} = x_i^{(k)}$ and $\sigma_i^{(k)} = \sigma_j^{(k)}$. This means that the mixture components are centered about the samples. Note that the variances for each dimension are fixed and determined *a priori*. Selecting appropriate hyper-parameters is discussed in Sec. V. In summary, the only remaining variables are the weights, i.e., in vector form $\underline{\alpha} = [\alpha_1, \ldots, \alpha_M]^{\mathrm{T}}$. In the rest of this section, the optimization of the weights based on a distance measure and a roughness penalty is described. In the following sections, the general formulations will be instantiated for density and conditional density estimation.

A. Distance Measure

In order to compare $f_{\mathcal{D}}$ with f_{GM}, the distance between their cumulative distributions is used. The cumulative distribution is widely employed for comparing discrete and continuous random variables. Yet, the conventional measure is not well defined for multivariate density functions, as it is not unique and non-symmetric [13], [14]. For this reason, the cumulative distribution function, the Localized Cumulative Distribution (LCD) [13] is used in this paper, as it is unique and symmetric. The LCD is an alternative representation of the cumulative distribution function obtained from integration with symmetric kernels for all positions and widths.

Definition 1 (Localized Cumulative Distribution, [13]): Given a multivariate random vector $\underline{\boldsymbol{x}} \in \mathbb{R}^N$ and the corresponding probability density function $f(\underline{x}) : \mathbb{R}^N \to \mathbb{R}_+$. The Localized Cumulative Distribution is defined as

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

with $\Omega \subset \mathbb{R}^N \times \mathbb{R}_+^N$, $F : \Omega \to [0,1]$, $\underline{b} \in \mathbb{R}_+^N$, $\mathcal{K}_{\underline{b}}(\underline{x},\underline{m})$ a suitable kernel [13] centered at $\underline{m} = [\, m^{(1)} \cdots m^{(N)}\,]^T$ with width $\underline{b}$ and $\mathcal{K} : \Omega \to [0,1]$.

For the rest of this paper, only separable Gaussian kernels [13] with mean $\underline{m}$ and identical width b for all dimensions are considered

$$\mathcal{K}_{\underline{b}}(\underline{x},\underline{m}) = \prod_{k=1}^{N} \exp\left(-\frac{1}{2}\frac{(x^{(k)}-m^{(k)})^2}{b^2}\right).$$

The LCD of $f_{\mathcal{D}}$ is given by

$$F_{\mathcal{D}}(\underline{m},b) = \sum_{i=1}^{|\mathcal{D}|} w_i \prod_{k=1}^{N} \exp\left(-\frac{1}{2}\frac{\left(x_i^{(k)}-m^{(k)}\right)^2}{b^2}\right). \tag{4}$$

Similar to $F_{\mathcal{D}}$, the LCD of f_{GM}, corresponding to the density or conditional density to be estimated, will be derived in the respective sections. Given the LCDs of the $f_{\mathcal{D}}$ and the estimation function, the distance between the *Localized Cumulative Distributions* (LCD) [13] shall be minimized. For this purpose, the modified Cramér-von Mises distance measure (MCvMD) [13] is employed

$$\mathrm{D} = \int_{\mathbb{R}^+} w(b) \int_{\mathbb{R}^N} \left(F_{\text{GM}}(\underline{m},b) - F_{\mathcal{D}}(\underline{m},b)\right)^2 \mathrm{d}\underline{m}\,\mathrm{d}b\,. \tag{5}$$

B. Regularization

Minimizing the distance between the empirical and the estimate's distributions is performed by minimizing D w.r.t. $\underline{\alpha}$. In order to penalize the roughness of the estimate f_{GM} a penalty term is added to the optimization problem

$$\begin{aligned}\mathrm{R} &:= \int_{\mathbb{R}^N} f(\underline{x})^2\,\mathrm{d}\underline{x} \\ &= \sum_{i=1}^{L}\sum_{j=1}^{L} \alpha_i\alpha_j \prod_{k=1}^{N} \mathcal{N}(\mu_i^{(k)} - \mu_j^{(k)}, \sqrt{2}\sigma^{(k)})\,. \end{aligned} \tag{6}$$

The term R may be calculated in closed-form and conveniently in vector/matrix formulation $\mathrm{R} = \underline{\alpha}^{\mathrm{T}}\,\mathbf{E}\,\underline{\alpha}$ for the specific type of density (2). Note that for f_{GM} to be a valid density, the constraint $\sum_{i=1}^{M}\alpha_i = \mathrm{Z}$, with constant Z and $0 \le \alpha_i \le 1$ needs to be asserted. The value of Z depends on whether a density or a conditional density is estimated and will be discussed in the respective sections. Minimizing R can be understood as penalizing the distance to a *flat* function, as the term (6) is related to an inner product in function space. It is noteworthy, that the above term is also related to the (negative) Renyi entropy of order $r = 2$ [15], i.e., an entropy measure for continuous random variables. Besides penalizing the overall density, an additional constraint $\alpha_i \le \nu$ is introduced to avoid large weights. In summary, the overall optimization problem comprises a term corresponding to the data fit and a term penalizing non-smooth densities over the state space.

In the following two sections, the instantiations of the optimization problem D + R for density and conditional density estimation are presented.

III. DENSITY ESTIMATION

Given samples $\mathcal{D}$ in form of the empirical probability density function $f_{\mathcal{D}}$ (1), the density estimation problem corresponds to determining the weights $\underline{\alpha}$ of the Gaussian mixture f_{GM} (2), by minimizing D + R, as given in (5) and (6), w.r.t. fixed sets of $\mu_i^{(k)}$ and $\sigma_i^{(k)}$ in the sense of Sec. II. In the rest of this section, the formulation of the MCvMD (5) based on the LCD of (2) will be simplified as only the weights are variable. In conjunction with the penalty, a readily solvable quadratic program will be derived.

A. Simplifying the Distance Measure

In order to estimate f_{GM}, the distance between the LCD of $f_{\mathcal{D}}$ and f_{GM} shall be minimized. The LCD $F_{\mathcal{D}}$ of $f_{\mathcal{D}}$ is

(4) and the LCD of (2) is given by [13]

$$F_{GM}(\underline{m}, b) = \sum_{i=1}^{M} \alpha_i \prod_{k=1}^{N} \frac{b}{\sqrt{(\sigma_i^{(k)})^2 + b^2}} \exp\left(-\frac{1}{2} \frac{\left(\mu_i^{(k)} - m^k\right)^2}{(\sigma_i^{(k)})^2 + b^2} \right) . \tag{7}$$

The distributions $F_{\mathcal{D}}$ and F_{GM} are compared using the MCvMD (5). Solving (5) equals solving the integrals over the kernel position $\underline{m}$ and the kernel width b. Solving the inner integral over $\underline{m}$ may be performed analytically. Since only $\underline{\alpha}$ is variable, the result of solving (5) can be written compactly in vector/matrix form as

$$\mathrm{D} = \int_{\mathbb{R}^+} w(b) \left(\underline{\alpha}^{\mathrm{T}} \mathbf{A}_1 \, \underline{\alpha} - 2 \, \underline{\alpha}^{\mathrm{T}} \, \underline{\mathrm{A}}_2 + \mathrm{A}_3 \right) \mathrm{d}b \,, \tag{8}$$

Here, $\mathbf{A}_1$-A_3 denote the solutions for the terms of the binomial w.r.t. $\underline{\alpha}$. Note that the integrals over b in (8) can be solved numerically. The result for A_3 is omitted, as it is constant w.r.t. $\underline{\alpha}$. This omission and using compact vector/matrix notation allows the calculation of D as

$$\mathrm{D} := \underline{\alpha}^{\mathrm{T}} \, \mathbf{P}_1 \, \underline{\alpha} - 2 \, \underline{\alpha}^{\mathrm{T}} \, \underline{\mathrm{P}}_2 \,, \tag{9}$$

with

$$\mathbf{P}_1 := ((\, p_1^{ij} \,)) = \int_{\mathbb{R}^+} w(b) \, \mathrm{a}_1^{(i,j)}(b) \, \mathrm{d}b \,,$$
$$\underline{\mathrm{P}}_2 := ((\, p_2^{i} \,)) = \int_{\mathbb{R}^+} w(b) \, \mathrm{a}_2^{(i)}(b) \, \mathrm{d}b \,,$$

The matrix P_1 and the vector $\underline{\mathrm{P}}_2$ consist of elements $\mathrm{a}_1^{(i,j)}(b)$ and $\mathrm{a}_2^{(i)}(b)$ given in Sec. VIII.

B. Quadratic Program

Considering the reformulation of D in (9) and the regularization term R in (6), one notes that the expressions are only quadratic in $\underline{\alpha}$, since the $\sigma_i^{(k)}$ for each dimension are determined *a priori*, which is discussed in Sec. V. By combining the quadratic terms in D (9) and R (6) one obtains

$$\underline{\alpha}^{\mathrm{T}} \, \mathbf{P}_1 \, \underline{\alpha} + c \, \underline{\alpha}^{\mathrm{T}} \; E \, \underline{\alpha} = \underline{\alpha}^{\mathrm{T}} \, \mathbf{Q} \, \underline{\alpha} \,, \tag{10}$$

with $\mathbf{Q} = \mathbf{P}_1 + c\mathbf{E}$ and $c \in \mathbb{R}$, a trade-off parameter reflecting one's belief in the need for regularization. Using (10) with the necessary constraints yields a readily implementable quadratic program

$$\begin{aligned} \underline{\alpha}^* = \arg\min_{\underline{\alpha}} \quad & \underline{\alpha}^{\mathrm{T}} \mathbf{Q} \; \underline{\alpha} - 2 \cdot \underline{\alpha}^{\mathrm{T}} \, \underline{\mathrm{P}}_2 \\ \text{s.t.} \quad & \underline{\alpha}^{\mathrm{T}} \, \underline{1} = 1 \,, \\ & 0 \leq \alpha_i \leq \max(\, \min(\, \nu, 1 \,), 1/|\mathcal{D}| \,) \,, \end{aligned} \tag{11}$$

with $i = 1, \ldots, |\mathcal{D}|$, the weights $\underline{\alpha}$, the trade-off parameter c, and a maximum allowed weight ν. The additional constraint $\sum_{i=1}^{M} \alpha_i = 1$ enforces that a convex combination of densities is obtained, which asserts that the integral of the density is 1. Additionally, positive weights are required so that the resulting density cannot be negative. The maximum weight ν was introduced to avoid overly large weights. Note that the maximum weight ν is in itself constrained to be larger than $1/|\mathcal{D}|$ to allow for valid parameter assignments. The first constraint could not be met otherwise. The quadratic program (11) can be solved with any standard solver. In this section, the density estimation problem in form of a quadratic program, based on the minimization of the MCvMD and the localized cumulative distributions, was derived. In the next section, the estimation of conditional densities will be considered.

IV. CONDITIONAL DENSITY ESTIMATION

In this section, conditional densities shall be estimated. The problem setting differs from Sec. II, in that the data is represented in the form of tuples

$$f_{\mathcal{D}}(\underline{x}, \underline{y}) = \textstyle\sum_{i=1}^{|\mathcal{D}|} w_i \, \delta(\underline{x} - \underline{x}_i) \; \delta(\underline{y} - \underline{y}_i) \,, \tag{12}$$

which were generated from an underlying conditional density $\tilde{f}(y|x)$. An estimate of $\tilde{f}(y|x)$ is sought in the form of

$$\begin{aligned} & f_{\text{CGM}}(\underline{y}|\underline{x}) \\ & \quad = \textstyle\sum_{i=1}^{M} \alpha_i \, \mathcal{N}(\underline{y} - \underline{\mu}_{yi}, \Sigma_{yi}) \, \mathcal{N}(\underline{x} - \underline{\mu}_{xi}, \Sigma_{xi}) \,. \end{aligned} \tag{13}$$

By assuming axis-aligned Gaussian mixtures for $\underline{\boldsymbol{x}}$ and $\underline{\boldsymbol{y}}$, the covariance matrices are given by $\Sigma = \mathrm{diag}(\, [(\sigma^{(1)})^2 \cdots (\sigma^{(N)})^2]^{\mathrm{T}} \,)$ and the multivariate Gaussians in (13) may be expressed as

$$\begin{aligned} & \mathcal{N}(\underline{y} - \underline{\mu}_{yi}, \Sigma_{yi}) \, \mathcal{N}(\underline{x} - \underline{\mu}_{xi}, \Sigma_{xi}) = \\ & \quad \prod_{k_y=1}^{N_y} \mathcal{N}(x - \mu_{yi}^{(k_y)}, \sigma_{yi}^{(k_y)}) \prod_{k_x=1}^{N_x} \mathcal{N}(x - \mu_{xi}^{(k_x)}, \sigma_{xi}^{(k_x)}) \,. \end{aligned}$$

For the sake of clarity, the rest of this section is concerned with the case of $N_y = 1$ and $N_x = 1$ only, i.e., the input and output dimensions are scalar. Note, that this does not solve the ambiguity in the definition of the cumulative distribution as (12) still requires a multivariate cumulative distribution. Given the definition of the empirical probability density function (12) and the form of the target density function (13), solving the conditional density estimation problem appears to be very similar to the (unconditional) density estimation problem. Yet, there are distinct differences:

- The *constraints differ*, as it needs to be asserted that
$$\textstyle\int_{-\infty}^{\infty} f(y|\hat{x}) \, \mathrm{d}y = 1 \,, \qquad f(y|\hat{x}) \geq 0 \,. \tag{14}$$
- $\mathcal{D}$ is represented as a joint density. A conditional LCD is not defined, i.e., in contrast to classical density estimation *no direct comparison* is possible.

The key idea is to reduce this problem to the comparison of the LCDs of joint densities [10]. For (13), one calculates

$$f_{\overline{\text{GM}}}(y, x) = f_{\text{CGM}}(y|x) \, f_{\mathcal{D}}(x) \,. \tag{15}$$

Yet, this problem transformation comes along with a new challenge: In (15), f_{CGM} and $f_{\mathcal{D}}$ have $|\mathcal{D}|$ components each. Using the LCD of (15) with the squared distance measure, $F^2_{\overline{\text{GM}}}(.)$ will have $|\mathcal{D}|^4$ components. This is too expensive to calculate for many problems and can be avoided by

approximating $f_{\mathcal{D}}(x)$. In order to extend the conditional density (13) to a joint density according to (15), a density $\bar{f}$ substituting $f_{\mathcal{D}}$ with $L \ll |\mathcal{D}|$ components is determined. Using the approximative $\bar{f}$

$$f_{\mathcal{D}}(\underline{x}) \approx \sum_{l=1}^{L} \alpha_l \prod_{k=1}^{N} \mathcal{N}(x^{(k)} - \mu_l^{(k)}, \sigma_l^{(k)}) =: \bar{f}(\underline{x}) \,. \quad (16)$$

Here, it is assumed that $f_{\mathcal{D}}(x)$ is well posed in the sense, that the data is almost evenly distributed over a fixed interval $x \in [x_{min}, x_{max}]$ to allow for a robust estimation of $\bar{f}(x)$.

For the experiments in Sec. VI, $\bar{f}(\underline{x}) = \mathcal{N}(\underline{x} - \hat{\underline{x}}, a\,\hat{\mathbf{C}})$, with sample mean $\hat{\underline{x}}$, sample covariance $\hat{\mathbf{C}}$, and a large constant factor a was chosen to model our concentration of $\mathcal{D}$ over $\boldsymbol{x}$. Regarding computational complexity, the use of $\bar{f}(x)$ will cost an amount of computation in the order of the conditional density estimation and therefore is negligible in the overall asymptotic complexity consideration.

A. Simplifying the Distance Measure

In order to compare the distributions for $f_{\overline{\mathrm{GM}}}(x, y)$ and $f_{\mathcal{D}}(x, y)$, their respective LCDs need to be determined. For $\bar{f}(\underline{x})$ in form of a GM, the LCD [13] of $f_{\overline{\mathrm{GM}}}(x, y)$ using (16) with $\underline{m} = [m_x \; m_y]^{\mathrm{T}}$ is given by

$$F_{\overline{\mathrm{GM}}}(\underline{m}, b) = \sum_{i=1}^{M} \alpha_i \sum_{l=1}^{L} \frac{\alpha_l \, c_x \, b^2}{\sqrt{(\sigma_{il}^x)^2 + b^2}\sqrt{(\sigma_i^y)^2 + b^2}} \cdot \exp\left(-\frac{1}{2}\frac{(\mu_i^y - m_y)^2}{(\sigma_i^y)^2 + b^2} \; - \frac{1}{2}\frac{(\mu_{il}^x - m_x)^2}{(\sigma_{il}^x)^2 + b^2} \right) . \quad (17)$$

The LCD in (17) resembles the LCD in (7), but has L times more components and the components about x are weighted according to their distance to $\bar{f}$. The modified Cramér-von Mises distance (5) may be employed as in the unconditional case. In analogy, one obtains

$$\mathrm{D} = \underline{\alpha}^{\mathrm{T}} \mathbf{P}_1 \, \underline{\alpha} - 2 \, \underline{\alpha}^{\mathrm{T}} \underline{\mathrm{P}}_2 \,, \quad (18)$$

but with differing $\mathbf{P}_1$ and P_2 w.r.t. (12) and (17),

$$\mathbf{P}_1 := ((\, p_1^{ij} \,)) = \int_{\mathbb{R}^+} w(b) \, \mathrm{d}_1^{(i,j)}(b) \; \mathrm{d}b \,,$$
$$\underline{\mathrm{P}}_2 := ((\, p_2^{i} \,)) = \int_{\mathbb{R}^+} w(b) \, \mathrm{d}_2^{(i)}(b) \; \mathrm{d}b \,.$$

The terms $\mathrm{d}_1^{(i,j)}(b)$ and $\mathrm{d}_2^{(i)}(b)$ are given in the Appendix.

B. Quadratic Program

Similar to the unconditional case, combining the quadratic terms in D + R, (18) and (6), simplifies as in (10), yielding the readily implementable quadratic program

$$\begin{aligned} \underline{\alpha}^* = \arg\min_{\underline{\alpha}} \quad & \underline{\alpha}^{\mathrm{T}} \mathbf{Q} \, \underline{\alpha} - 2 \cdot \underline{\alpha}^{\mathrm{T}} \underline{\mathrm{P}}_2 \\ \text{s.t.} \quad & \underline{\alpha}^{\mathrm{T}} \underline{1} = \mathrm{Z} \,, \\ & 0 \le \alpha_i \le \max(\, \min(\, \nu, \mathrm{Z} \,), \mathrm{Z}/|\mathcal{D}| \,) \,, \end{aligned} \quad (19)$$

with $i = 1, \ldots, |\mathcal{D}|$. Here, the weight constraint assumes $x \in [x_{min}, x_{max}]$ with $\mathrm{Z} := x_{max} - x_{min}$. This constraint is approximate and states that the conditional density integrated over x and y - restricted to the above interval - meets the constraint of integrating to one. Due to the different normalization, the constraint on the maximum weight ν changes to being no larger than $\mathrm{Z}/|\mathcal{D}|$ to allow for valid parameter assignments.

The quadratic program (19) can be solved with any standard solver. The derivation of (19) was restricted to the case of scalar input and output dimensions x and y. The formulations hold for multivariate $\underline{x}$ and $\underline{y}$, too. This will be shown in the multivariate helix example in Sec. VI. The only difference is in the term (17). There, a product about the dimensions N_x and N_y would need to be inserted.

V. HYPER-PARAMETER DETERMINATION

Both density and conditional density estimation in Sec. III and in Sec. IV require an *a priori* choice of the hyper-parameters, i.e., the components' variances and the parameters of the optimization problem c and ν. Many approaches to estimating $\sigma^{(i)}$ exist in the kernel density estimation literature [3], [4], [6]. There, the minimization of the least-squares error between the density function depending on $\sigma^{(i)}$ and an unbiased estimate is proposed. Moreover, cross-validation is employed [4], [6]. For optimizing all parameters at once, the literature on GPR [12] and SVR [11] proposes the use of cross-validated maximization of a likelihood score, similar to the least-squares minimization [6]. In our experience, the first approach works best. For this reason, a gradient ascent on the average log-likelihood for a k-fold cross-validation was used to find the best hyper-parameters using randomly partitioned data.

VI. EXPERIMENTS

In order to validate the density estimation (Sec. VI-A) and conditional density estimation (Sec. VI-B) approaches, experimental comparisons against the benchmark parametric algorithm EM, the non-parametric algorithms GPR, and SVR are performed in this section. In order to present robust statistics, all results presented in this section are averages over 10 experiments each. The experimental setup resembles [16], [17] and allows a comparison with these approaches. From an implementation side, EM for GM by Matlab™, the SVR and SVDF implementation from [17], the EKF, UKF, GP-UKF, as well as GP-ADF implementations from [16] were used.

A. Density Estimation

In this section, the proposed approach is compared against EM, the benchmark parametric density estimator for finite mixture models. In this experiment, samples were generated from a mixture density of four Gaussians, depicted in Fig.1 (top). The densities were estimated using 100 training samples generated from this mixture and tested with additional 100 test samples. As a measure of fit, the negative log-likelihood NL_x score of the test samples is employed. In Tab. I, the NL_x results for the LCD approach and four

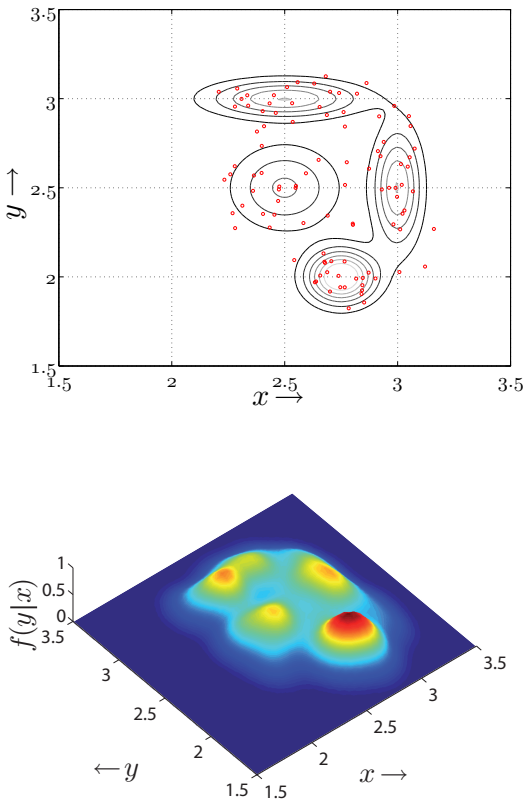


Fig. 1. True density composed of four normal densities with random samples (top) and LCD-based density estimate (bottom).

configurations of EM are given. In all configurations, axis-aligned GMMs were estimated. EM1 denotes the estimation of a mixture with identical variances σ_x and σ_y, only. In EM2, this restriction was lifted. For EM1 and EM2 the number of Gaussians was set to the number of Gaussians obtained by (11). In EM3 and EM4, the same restrictions on the variances as in EM1 and EM2 were used, but the number of components was chosen by optimizing the Akaike information criterion. The results in Tab. I show that all EM approaches yield approximately identical likelihood scores. The results for the LCD-based density estimates, depicted in Fig.1 (bottom), show that the approach in Sec. III may achieve significantly better NL_x results than EM. The number of components given in Tab. I shows that the proposed approach produces densities with more components than the EM, but much less than the maximum number of 100 components. In contrast to classical non-parametric approaches which would use all 100 components - sparser representations are achieved. The results given are averages over ten independent runs of the experiment.

B. Conditional Density Estimation

The quality of the conditional densities is validated by comparing log-likelihood scores and by comparing the estimation quality, when using the densities in a Bayesian filter. Furthermore, to demonstrate multivariate estimation, a conditional density corresponding to a helix in 3D is estimated.

TABLE I

AVERAGE NEGATIVE LOG-LIKELIHOOD OF THE TEST DATA WITH σ AND NUMBER OF COMPONENTS FOR THE DENSITY ESTIMATES RETURNED BY EM AND THE PROPOSED DISTANCE-BASED APPROACH.

	EM1	EM2	EM3	EM4	LCD
$NL_x - \mu$	1.54	1.85	1.54	1.90	0.16
$\pm\sigma$	± 0.24	± 0.41	± 0.24	± 0.40	± 0.04
Comp.	48.8	48.8	5.1	5	48.8

TABLE II

AVERAGE NEGATIVE LOG-LIKELIHOOD OF THE TEST DATA WITH σ FOR THE CONDITIONAL DENSITY ESTIMATES RETURNED BY EM, GPR, SVR AND THE PROPOSED DISTANCE-BASED APPROACH.

EM1	EM2	EM3	EM4	GPR	SVR	LCD
3.87	3.97	3.58	3.62	0.23	0.59	0.15
± 1.44	± 1.60	± 1.31	± 1.89	± 0.03	± 0.35	± 0.11

1) Log-Likelihood Score: In order to assess the quality of the conditional density estimation approach, results for estimating the probabilistic model of the cubic system

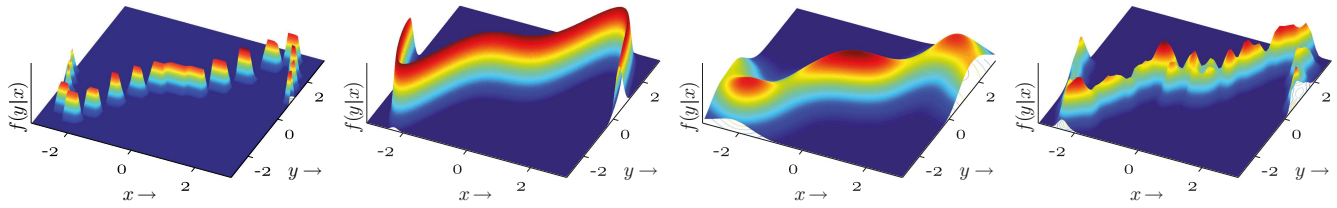
$$\boldsymbol{x}_{k+1} = \tfrac{2}{3}\,\boldsymbol{x}_k^2\,\sin(\boldsymbol{x}_k)\;,\quad \boldsymbol{w}_k \sim \mathcal{N}(0, 0.2)\,, \qquad (20)$$

are reported. The NL_x as calculated in the unconditional case are given for the conditional densities estimated by EM1-4, GPR, SVR, and the LCD-based approach are given in Tab. II. Obviously, all EM results are significantly worse than the results for GPR, SVR, and the LCD-based results. The results given are average values of ten experiments. The corresponding conditional densities of one example run are depicted in Fig. 2 for one EM configuration, GPR, and SVR as well as for LCD.

2) Nonlinear Filtering: To test the conditional densities for state estimation, the derived conditional densities are used with a Gaussian mixture filter, as derived in [9], [19] and compared to the EKF [20], the UKF [21], the GP-UKF [22], the GP-ADF [16], and the mixture filter based on the SVR-derived densities (SVDF). As the state space is almost evenly sampled in the following examples, results of the LCD-based filter (LCDF) neglecting the marginal distribution about x are reported only. The filters are compared by the state estimation quality on the growth model [18] as presented in [16]. The corresponding nonlinear system and measurement equations are

$$\boldsymbol{x}_{k+1} = 0.5\,\boldsymbol{x}_k + \tfrac{25\,\boldsymbol{x}_k}{1+\boldsymbol{x}_k^2} + \boldsymbol{w}_k\,, \qquad \boldsymbol{w}_k \sim \mathcal{N}(w, 0.2)\,,$$
$$\boldsymbol{y}_{k+1} = 5\,\sin(2\,\boldsymbol{x}_{k+1}) + \boldsymbol{v}_{k+1}\,, \quad \boldsymbol{v}_{k+1} \sim \mathcal{N}(v_{k+1}, 0.01)\,.$$

For training, randomly distributed 100 points in $[-10, 10]$ were generated. For the estimation, the prior normal density has $\mu_0 \in [-10, 10]$ and $\sigma_0 = 0.5$. For 200 independent $x_0^{(i)}$, the successive states and $y_1^{(i)}$ were estimated. Statistics are given in Tab. 3 for the Mahalanobis distance $\mathcal{M}(x)$ between the mean of the state estimate and the true value. Additionally, the upper and lower quantiles of NL_x of the hidden state are reported. Smaller values of $\mathcal{M}(x)$ and NL_x show better performance. The results given are average values of ten experiments. Tab. 3 shows that GP-ADF, SVDF,



(a) EM (b) GPR (c) SVR (d) LCD

Fig. 2. Conditional density estimates - from left to right: EM with the same number of Gaussian like the LCD but with all identical and variable variances. The next two results are obtained from GPR and SVR. The last two plots show the LCD-based conditional density estimates for (20) with Gaussian mixture approximation of $f_{\mathcal{D}}$ and neglected prior knowledge. The depicted results are either automatically normalized (GPR) have been numerically normalized.

	$NL_x^{0.25}$		$NL_x^{0.5}$		$NL_x^{0.75}$		$\mathcal{M}(x)$	
EKF	888.53	$\pm$ 68.30	$2.9e+04$	$\pm$ 772.31	$2.7e+05$	$\pm$ 971.18	$2.0e+06$	$\pm$ 1072.24
UKF	61.35	$\pm$ 0.59	605.76	$\pm$ 18.05	2383.86	$\pm$ 36.25	1042.39	$\pm$ 4.88
GP-UKF	65.10	$\pm$ 4.88	420.78	$\pm$ 34.22	1796.14	$\pm$ 182.14	3523.81	$\pm$ 2973.71
GP-ADF	59.37	$\pm$ 2.20	260.44	$\pm$ 18.93	1093.01	$\pm$ 82.10	27.67	$\pm$ 9.16
SVDF	59.67	$\pm$ 1.69	178.75	$\pm$ 3.98	396.62	$\pm$ 17.10	1.44	$\pm$ 0.17
LCDF	71.45	$\pm$ 9.63	184.42	$\pm$ 10.03	371.63	$\pm$ 61.25	0.78	$\pm$ 0.80

Fig. 3. Average negative Log-Likelihood and Mahalanobis distance results for the growth process [18]. The results are averages over ten experiments.

and the LCDF yield the best performance and compare well to each other. Yet, the uncertainty distribution indicated by the upper quantile of NL_x as well as the $\mathcal{M}(x)$ results are more favorable for the LCD-based filter than for all other approaches. Note that EKF, UKF, and GP-UKF are all drastically outperformed either in $\mathcal{M}(x)$ or NL_x, or both.

3) Multivariate Helix: The above results for conditional density estimation were restricted to scalar in- and output dimensions. To demonstrate the estimation of multivariate conditional densities, a system, mapping a scalar input to bivariate output dimension, is used

$$\begin{bmatrix} z \\ y \end{bmatrix} = \begin{bmatrix} 2\,x^2\,\cos\left(1.5\,x\right) \\ 2\,x^2\,\sin\left(1.5\,x\right) \end{bmatrix} + w\,,$$

with w distributed according to $\sim \mathcal{N}\left(0, \mathrm{diag}\left(\left[0.2^2\, 0.2^2\right]\right)\right)$. The training data consists of uniformly sampled $x \in [0, 3\pi]$ and additively perturbed function values thereof, describing a 3D helix. The data and the obtained estimate are depicted in Fig. 4. In order to capture the multidimensionality, plots of the $x-z$ and $y-z$ planes as well as a rotated 3D plot are given. The obtained estimate is sparse, containing only 66 of the 100 samples as components. Most of the omitted samples are located around the origin, as the data is relatively dense here. The equations presented in Sec. IV were extended to the multivariate case by exploiting the axis-aligned decomposition of (13). The experiment shows how easy the approach generalizes to the multivariate case.

VII. CONCLUSIONS

In this paper, a method for non-parametric density and conditional density estimation was presented based on minimizing the modified Cramér-von Mises distance of the Localized Cumulative Distributions. By minimizing this distance, the proposed approach avoids the problem of the ambiguous definition of the cumulative distribution function. Additionally, overfitting was avoided by adding a regularization term penalizing the densities' roughness. The resulting optimization problems can be phrased in form of readily implementable quadratic programs. Experimental comparison of the arising axis-aligned Gaussian mixture densities and conditional densities against EM, SVR, and GPR show the approaches' good performance w.r.t. the sample log-likelihood scores and the performance in benchmark recursive filtering applications. Yet, the derived densities contain less components (e.g. $< 50\%$ of components) than the other non-parametric GPR and SVR densities.

VIII. APPENDIX

- The detailed results for (5) are

$$\mathrm{a}_1^{(i,j)}(b) = (\sqrt{2\,\pi})^N b^{2\,N} \prod_{k=1}^{N} \frac{1}{\sqrt{{\sigma_{k,j}}^2 + 2\,b^2 + {\sigma_{k,i}}^2}} \cdot \exp\left(-\frac{1}{2}\,\frac{\left(\mu_{k,i} - \mu_{k,j}\right)^2}{{\sigma_{k,j}}^2 + 2\,b^2 + {\sigma_{k,i}}^2}\right),$$

$$\mathrm{a}_2^{(i)}(b) = \sum_{j=1}^{|\mathcal{D}|} w_j\,(\sqrt{2\,\pi})^N b^{2\,N} \cdot \prod_{k=1}^{N} \frac{1}{\sqrt{2\,b^2 + {\sigma_{k,i}}^2}} \exp\left(-\frac{1}{2}\,\frac{\left(x_{k,j} - \mu_{k,i}\right)^2}{2\,b^2 + {\sigma_{k,i}}^2}\right).$$

- The LCD for (15) with (13) and $\underline{m} = [m_x\ m_y]^{\mathrm{T}}$,

$$F_{\overline{\mathrm{GM}}}(\underline{m}, b) = \int_{\mathbb{R}^+}\int_{\mathbb{R}^+} \sum_{i=1}^{M}\sum_{l=1}^{|\mathcal{D}|} \alpha_i\,\alpha_l\,\mathcal{K}_b(y, m_y)\,\mathcal{K}_b(x, m_x) \cdot \mathcal{N}(y - \mu_i^y, \sigma_i^y)\,\mathcal{N}(x - \mu_i^x, \sigma_i^x)\,\mathcal{N}(x - \mu_l^x, \sigma_l^x)\ \mathrm{d}x\ \mathrm{d}y$$

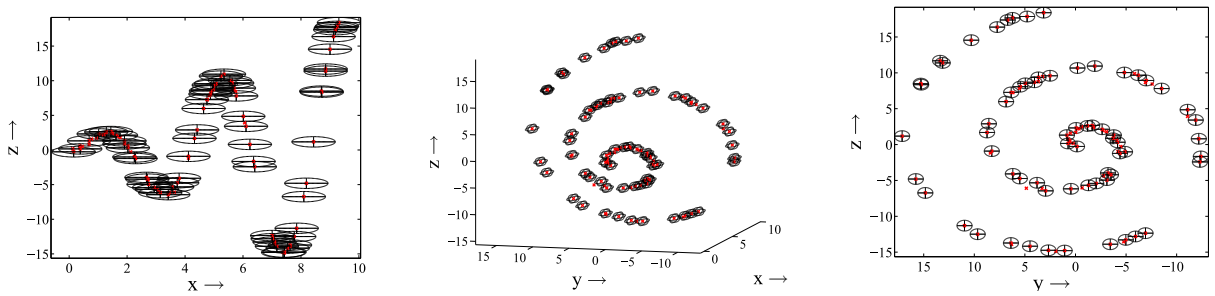


Fig. 4. Three perspectives of $f(z, y|x)$, i.e., a multivariate conditional density: red crosses mark the samples' positions and black axis-aligned ellipsoids correspond to mixture components in $f(z, y|x)$ with $\alpha_i > 0$. The ellipsoids are centered at $\underline{\mu}_i$ and the axial length is σ in the respective directions.

is simplified by employing

$$\int_{\mathbb{R}^+} \mathcal{K}_b(x, m_x)\, \mathcal{N}(x - \mu_i^x, \sigma_i^x)\, \mathcal{N}(x - \mu_l^x, \sigma_l^x)\, \mathrm{d}x = \sqrt{2\,\pi}\, c_x\, b \mathcal{N}(m_x - \mu_{il}^x, \sigma_{il}^x) \quad (21)$$

and setting $c^x := \mathcal{N}(\mu_i^x - \mu_l^x, \sqrt{(\sigma_i^x)^2 + (\sigma_l^x)^2})$ with

$$\mu_{il}^x := \tfrac{\mu_i^x\,(\sigma_l^x)^2 + \mu_l^x\,(\sigma_i^x)^2}{(\sigma_l^x)^2 + (\sigma_i^x)^2}\,, \quad \sigma_{il}^x := \tfrac{(\sigma_l^x)^2\,(\sigma_i^x)^2}{(\sigma_l^x)^2 + (\sigma_i^x)^2}\,.$$

Analog simplification for $\boldsymbol{y}$ and insertion into (21) yields (17)

$$\begin{aligned}
\mathrm{d}_1^{(i,j)}(b) &= \textstyle\sum_{l=1}^{L} \sum_{k=1}^{L} (2\,\pi)^2 b^4\, c_l^x\, c_k^x \\
&\quad \cdot \mathcal{N}(\mu_{il}^x - \mu_{jk}^x, \sqrt{(\sigma_{il}^x)^2 + (\sigma_{jk}^x)^2}) \\
&\quad \cdot \mathcal{N}(\mu_i^y - \mu_j^y, \sqrt{(\sigma_i^y)^2 + 2b^2 + (\sigma_j^y)^2})\,, \\
\mathrm{d}_2^{(i)}(b) &= \textstyle\sum_{l=1}^{L} \sum_{k=1}^{|\mathcal{D}|} w_j\, (2\,\pi)^2 b^4\, c_k^x \\
&\quad \cdot \mathcal{N}(\mu_{il}^x - x_k, \sqrt{(\sigma_{il}^x)^2 + 2b^2}) \\
&\quad \cdot \mathcal{N}(\mu_i^y - y_l, \sqrt{(\sigma_i^y)^2 + 2b^2})\,.
\end{aligned}$$

REFERENCES

[1] D. Koller, *Probabilistic Graphical Models : Principles and Techniques*, N. Friedman, Ed. Cambridge, Massachusetts: MIT Press, 2009.

[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. W. Scott, *Multivariate Density Estimation : Theory, Practice, and Visualization*, ser. Wiley Series in Probability and Mathematical Statistics - A Wiley Interscience publication. New York: Wiley, 1992.

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

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

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

[7] G. McLachlan and D. Peel, *Finite mixture models*. Wiley-Inter., 2004.

[8] 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.

[9] E. Driver and D. Morrell, "Implementation of Continuous Bayesian Networks Using Sums of Weighted Gaussians," in *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, Montreal, Canada, August 1995, pp. 134–140.

[10] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed., ser. Statistics for Engineering and Information Science. New York: Springer, 2000.

[11] B. Schölkopf, A. Smola, R. Williamson, and P. Bartlett, "New Support Vector Algorithms," *Neural Computation*, vol. 12, no. 5, pp. 1207–1245, 2000.

[12] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, T. Dietterich, Ed. Cambridge, Massachusetts: The MIT Press, 2006.

[13] 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.

[14] 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.

[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

[16] M. Deisenroth, M. Huber, and U. Hanebeck, "Analytic Moment-based Gaussian Process Filtering," in *26th International Conference on Machine Learning (ICML)*, Montreal, Canada, June 2009.

[17] P. Krauthausen, M. F. Huber, and U. D. Hanebeck, "Support-Vector Conditional Density Estimation for Nonlinear Filtering," in *Proceedings of the 13th International Conference on Information Fusion (Fusion 2010)*, Edinburgh, United Kingdom, July 2010.

[18] G. Kitagawa, "Monte Carlo Filter and Smoother for non-Gaussian Nonlinear State Space Models," *Journal of Computational and Graphical Statistics*, vol. 5, no. 1, pp. 1–25, 1996.

[19] O. Schrempf and U. D. Hanebeck, "Evaluation of Hybrid Bayesian Networks using Analytical Density Representations," in *Proc. of the 16th IFAC World Congress (IFAC 2005)*, Czech Republic, 2005.

[20] D. Simon, *Optimal State Estimation: Kalman, H-Infinity, and Nonlinear Approaches*, 1st ed. Wiley & Sons, 2006.

[21] S. Julier and J. Uhlmann, "Unscented Filtering and Nonlinear Estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, March 2004.

[22] J. Ko, D. Klein, D. Fox, and D. Haehnel, "GP-UKF: Unscented Kalman Filters with Gaussian Process Prediction and Observation Models," in *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, San Diego, California, October 2007, pp. 1901–1907.