Sparse Mixture Conditional Density Estimation by Superficial Regularization

Peter Krauthausen, Patrick Ruoff, and Uwe D. Hanebeck
Intelligent Sensor-Actuator-Systems Laboratory (ISAS),
Institute for Anthropomatics,
Karlsruhe Institute of Technology, Germany.
{Peter.Krauthausen, Patrick Ruoff}@kit.edu,
Uwe.Hanebeck@ieee.org

***Abstract*—In this paper, the estimation of conditional densities between continuous random variables from noisy samples is considered. The conditional densities are modeled as heteroscedastic Gaussian mixture densities allowing for closed-form solution of Bayesian inference with full-densities. The main contributions of this paper are an improved generalization quality of the estimates by the introduction of a superficial regularizer, the consideration of model uncertainty relative to local data densities by means of adaptive covariances, and the proposition of an efficient distance-based estimation algorithm. This algorithm corresponds to an iterative nested optimization scheme, optimizing hyper-parameters, component placement, and mixture weights. The obtained solutions are sparse, smooth, and generalize well as benchmark experiments, e.g., in nonlinear filtering show.**

Keywords: Conditional density estimation, nonlinear filtering, Gaussian mixture density, Regularization.

I. Introduction

Conditional densities lie at the heart of any Bayesian estimation framework. The conditional densities' quality directly impacts the estimation performance of any probabilistic graphical model [14] in terms of achievable accuracy and runtime. If the true conditional density is not accessible, but samples distributed according to the true conditional density are given, the conditional density function may be estimated. In this paper, the problem of estimating conditional densities relating continuous random variables from samples is considered. This problem is challenging, because estimating a continuous function from a finite set of samples allows for an infinite number of solutions rendering the problem ill-posed. Yet, the obtained continuous conditional densities are essential, e.g., for nonlinear filtering.

There are two fundamentally different approaches towards conditional density estimation (CDE): estimation of an assumed generative model and its uncertainty, composing a probabilistic model out of both, or direct estimation of the probabilistic model. Exemplary methods, for estimating the underlying generative and augmenting it *error bars* are, e.g, Gaussian Processes (GP) [21] or probabilistic splines. The advantages of this approach are the regularization of the generative model by means well-known linear/kernel smoothing in the case of GP or the curvature minimization of splines, which yield smooth generative model that generalize well. The major drawback is that the generative model is used as an argument of noise density capturing the model uncertainty. In general, this prohibits exact solution to Bayesian inference. For example, Gaussian Process Regression (GPR) [21] based (nonlinear) filter as GP-EKF/UKF [11]–[13] or the analytic moment based GP (AM-GPF) [1] are based on the density approximations per recursive calculation step. Regarding the direct estimation of probabilistic model, all density estimators [23], e.g., expectation maximization (EM) for GMM [2], [17] or kernel density estimation (KDE) [19], [24], may be employed as a conditional density function may be trivially obtained by $f(\underline{y}|\underline{x}) = \frac{f(\underline{y},\underline{x})}{f(\underline{x})}$. As this operation is based on the estimation of the (joint) densities' parameters, the "wrong" parameters are optimized. Only little research has been performed, for estimating $f(\underline{y}|\underline{x})$ in the form of a Gaussian mixture density (GMM) [5], [6], [17], directly [8]. The most important contributions are an SVM-inspired CDE [26], conditional density approximation for fast evaluations [8], and regularized distance-based CDE [15], [16]. The advantages of estimating a GMM conditional density are, that it allows for exact Bayesian inference [3], [9] and the versatility of the representation. One disadvantage of this approach is that the GMM only approximates a valid conditional density as the condition $\int_\Omega f(\underline{y}|\hat{x})\,\mathrm{d}y = 1$ can not be met by a finite mixture. Furthermore, generalization of the the GMM is challenging. In particular, the above direct CDE approaches are similar to KDE as they optimize the mixture weights for components centered at the data points only and generalization in parts of the state space scarcely populated by data is not guaranteed. Additionally, the GMM's uncertainty is not captured.

The present paper is an extension of the approaches in [15], [16]. The main contributions are:

- The introduction of a superficial **regularization**, penalizing the roughness of the probabilistic model interpreted as a curvature of a surface, leading to an improved generalization.
- Consideration of the **model's uncertainty** by adaptive kernel widths yielding large or small component covariances relative to the amount of data populating the components' surrounding state space.
- An **efficient solution** of the arising optimization problem by a two-step scheme, i.e., iteratively optimizing the

hyper-parameters, the component placement and weights. The CDE is phrased as a minimization of a target function comprised of a term quantifying the distance of the localized cumulative distributions and a regularizer. The superficial regularizer avoids the fallacies of entropy-based regularization and allows for an optimization of the components' positions. The regularizer is shown to be an upper bound on the curvature of the generative model for additive normal noise. Thereby, regularization of the probabilistic model corresponds to a regularization of the generative model. The CDE is formalized as a quadratic program for the GMM's weight optimization embedded in a nonlinear function minimization determining hyper-parameters and positions. The improved performance of the proposed approach is demonstrated using benchmark and synthetic data. The resulting conditional densities are shown to be sparse and of high quality.

The rest of this paper is structured as follows. Sec. II gives the mathematical problem definition. Sec. III describes an overview over the nested optimization scheme and the outer and inner loop of optimization. The employed distance measure and the novel regularization term are explained in Sec. IV and Sec. V. The algorithm is summarized in Sec. VI and validated by experiments in Sec. VIII.

II. Problem Statement

The true conditional density $\tilde{f}$ shall be estimated from a set $\mathcal{D}$ of i.i.d. random samples $(\underline{x}_i, \underline{y}_i)$. The empirical probability density function [23] is a mixture of Dirac distributions $\delta(.)$

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

with

$$\underline{x}_i := [x_i^{(1)} \; \dots \; x_i^{(M)}]^{\mathrm{T}} \in \mathbb{R}^{\mathrm{M}}, \; \underline{y}_i \in \mathbb{R}^{\mathrm{N}}, w_i = \tfrac{1}{|\mathcal{D}|} \,.$$

The obtained estimate f shall have the form of an axis-aligned GMM, i.e., the components' covariances are

$$\mathbf{\Sigma}_i = \operatorname{diag}\left(\left[\left(\sigma_i^{(1)}\right)^2 \dots \left(\sigma_i^{(M+N)}\right)^2 \right]^{\mathrm{T}} \right) \tag{2}$$

The target function may be simplified as follows

$$\begin{aligned} &f(\underline{y}|\underline{x}) = \\ &\sum_{i=1}^{L} \alpha_i \, \mathcal{N}\left(\begin{bmatrix} \underline{y}_i \\ \underline{x}_i \end{bmatrix}; \begin{bmatrix} \underline{\mu}_{y,i} \\ \underline{\mu}_{x,i} \end{bmatrix}, \mathbf{\Sigma}_i \right) = \\ &\sum_{i=1}^{L} \alpha_i \prod_{k=1}^{M} \mathcal{N}\left(x^{(k)}; \mu_{x,i}^{(k)}, \sigma_{x,i}^{(k)} \right) \prod_{l=1}^{N} \mathcal{N}\left(y^{(l)}; \mu_{y,i}^{(l)}, \sigma_{y,i}^{(l)} \right) . \end{aligned} \tag{3}$$

This mixture consists of L components with a distinct normal density for each in- and output dimension. Restricting the estimate type to this specific GMM not only facilitates the latter calculations, but also allows for the implementation of a constant time nonlinear filter.

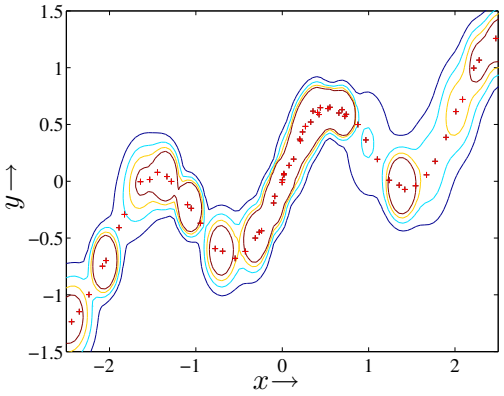

Figure 1. Conditional density function $f(y|x)$ (contour) estimated from $\mathcal{D}$ (red crosses) with twice the number of observations in the input interval $[0, 1]$.

III. Conditional Density Estimation

Determining the conditional density estimate f from $\mathcal{D}$ is an ill-posed problem. The estimation problem is therefore phrased as an optimization problem involving a data fit term D and a regularization term R reflecting the user chosen function preference for *smooth* densities. The trade-off between D and R is the target function used for determining all parameters of (3) optimized:

$$\theta = (\underbrace{\{\alpha_i\}_{1\le i\le L}, \left\{\underline{\mu}_i, \mathbf{\Sigma}_i(\phi)\right\}_{1\le i\le L}}_{\text{Components' parameters}}, \underbrace{\phi, \lambda}_{\text{Hyper-parameters}}) \tag{4}$$

with weights $\alpha_i \in [0, c]$, means $\underline{\mu}_i \in \mathbb{R}^N$, covariance matrices $\mathbf{\Sigma}_i(\phi) \in \mathbb{R}^{N\times N}$, and parameters ϕ thereof. The generic optimization problem is simply

$$\theta^* = \arg\min_{\theta} \; D + \lambda R \,. \tag{5}$$

For the efficient solution of (5) a nested optimization scheme is adapted:

1) **Outer loop:** Determine $\eta := \left(\{\underline{\mu}\}, \phi, \lambda\right)$,
2) **Inner loop:** Determine $\underline{\alpha}$, given fixed η.

Combined into one optimization problem one arrives at

$$\theta^* = \arg \min_{(\lambda,\phi,\{\underline{\mu}\})} \left(\min_{\underline{\alpha}_\eta} \; D(\mathcal{D}, \alpha_\eta, \eta) + \lambda|_\eta \, R(\alpha_\eta, \eta) \right) \tag{6}$$

$$\text{s.t.} \quad \mathbf{A}\,\underline{\alpha}_\eta \le \underline{b} \tag{7}$$

$$\underline{1}\,\underline{\alpha}_\eta = d \tag{8}$$

$$0 \le \underline{\alpha}_{\eta,i} \le c_\eta \tag{9}$$

The target function (6) corresponds to the trade-off between data fit term D and regularization term R. The constraint (7) enforces an approximately uniform mass distribution w.r.t. the input dimension. The term (8) asserts that approximately $\underline{1}^{\mathrm{T}}\,\underline{\alpha}_\eta = \mathrm{I}$, i.e., the probability mass is approximately normalized. Positivity of the conditional density is assured by (9).

In the rest of this section the inner and outer loop of the optimization scheme will be described in more detailed.

A. Outer Loop

In the outer loop the hyper-parameters and components means are optimized.

a) Determining $\{\underline{\mu}\}$: - In general all components' positions may be optimized in one large optimization problem. Since determining all positions corresponds to solving an $|\{\underline{\mu}\}| \cdot (M+N)$-dimensional optimization problem this becomes intractable already for small problems [25]. In order to obtain a tractable solution, it is proposed, to reduce number of components by setting $\{\underline{\mu}\} = \mathcal{D} \cup \{\underline{\mu}_v\}$ and only iteratively optimizing each $\{\underline{\mu}_v\}$ in turn. The model selection problem, i.e, the number of components and their starting values, should be solved w.r.t. the model's uncertainty. Roughly speaking, place $\{\underline{\mu}_v\}$ where model uncertainty is low, e.g., where the distance of the marginal density $f(\underline{x})$ to a uniform density is high. Because an exact optimization would be too costly for just determining starting values, a greedy iterative approach is adapted. In each step, an additional component is sampled from the "largest" components' covariance until a user defined upper bound on the number of components or the covariance size falls below a given threshold. The $\underline{y}$-locations may then be obtained by standard regression procedures. Note, that $|\{\underline{\mu}\}|$ is only an upper bound on the number of components in the mixture, as obsolete components are removed by the inner loop QP solution. This is a great simplification of the inherent model selection problem. Given the components $\{\underline{\mu}\}$, the $\{\underline{\mu}_v\}$ are optimized by iteratively fixing all but one component, then solving the inner loop for this position and repeating this process until convergence or given tolerance is achieved.

b) Determining $\{\phi\}$: In order to capture the **model's uncertainty**, it is proposed to parametrize each components' covariance $\mathbf{\Sigma}_i$ by a value $\phi_i \in \mathbb{R}$ relating the covariance size to the local data density around each components' mean, $(\underline{x}, \underline{y}) \in \mathbb{R}^{N \times M}$ by the distance to $\underline{d}_N = (\underline{x}_N, \underline{y}_N) \in \mathcal{D} \subset \mathbb{R}^N \times \mathbb{R}^M$, the nearest neighboring data point, according to

$$\mathbf{\Sigma}_i = \phi_i \, \mathbf{K} + \tau \, \mathbf{I} \,, \tag{10}$$

with the average sample variance of all k-nearest samples $\left(\hat{\sigma}_i^{(j)}\right)^2$ in each dimension, one defines for all components

$$\mathbf{K} = \operatorname{diag}\left(\left[\left(\hat{\sigma}_i^{(1)}\right)^2 \dots \left(\hat{\sigma}_i^{(M+N)}\right)^2 \right]^{\mathrm{T}} \right) .$$

The second summand in (10) adds a floor value τ to avoid singular $\mathbf{\Sigma}_i$, as ϕ_i is calculated by

$$\phi_i = \left(\begin{bmatrix} \underline{y}_i \\ \underline{x}_i \end{bmatrix} - \underline{d}_N \right)^{\mathrm{T}} \mathbf{K} \left(\begin{bmatrix} \underline{y}_i \\ \underline{x}_i \end{bmatrix} - \underline{d}_N \right) , \tag{11}$$

may be zero if two points coincide. The ϕ_i corresponds to the Mahalanobis distance to the nearest neighbor w.r.t. to the sample variance $\hat{\sigma}_i^{(j)}$ in each dimension for the i-the component. The effect of this adaptive kernel width is depicted in Fig. 1, where samples for the system $\boldsymbol{y} = \boldsymbol{x} + \sin(\boldsymbol{x}) + \varepsilon$ are given. In Fig. 1 the number of samples was doubled in the interval $[0, 1]$, reducing the component's covariances. Note, that in the case of an infinite amount of samples neglecting the floor value in (10) will degrade the covariances to Dirac distributions. In the case of no noise, the exact underlying functional dependency would be recovered.

c) Determining λ: This parameter governs the trade-off between the data fit term D and the regularization term R, i.e., for small values the data fit is emphasized and for higher values smoother conditional densities are preferered. Estimating λ from the samples may be done by a grid search and selecting the value yielding the lowest function value for the inner loop's optimization or direct minimization of the inner loop's value by standard function minimization methods. We refer the interested reader to the literature on calculating λ for *support vector machines* [22] for further reading.

Summarizing, the outer loop of the estimation method corresponds to a constrained nonlinear optimization problem, which might be solved with a standard solver, and uses the inner loop as a subfunction in its calculations. The joint algorithm of the inner and outer loop is presented in Sec. VI.

B. Inner Loop

After determining $\eta := \left(\{\underline{\mu}\}, \phi, \lambda\right)$ in the outer loop, it remains to determine the GMM weights $\underline{\alpha}$ w.r.t. to the given fixed η. The target function consists of the data fit term D and the regularization term R.

- $\mathbf{D}(\mathcal{D}, \alpha_\eta, \eta)$ - measures the similarity between the smooth estimate f and the unsmooth EPDF $f_\mathcal{D}$. Typical choices include the data likelihood or a distance measure, e.g., the squared integral distance between the cumulative distributions of f and $f_\mathcal{D}$. In this paper, the modified Cramér-von Mises distance measure is employed, as it compares localized cumulative distributions [7], which avoid the non-uniqueness and non-symmetry of the default definition of the multivariate cumulative distribution functions, cf. Sec. IV for more details.
- $\mathbf{R}(\alpha_\eta, \eta)$ - the regularization term encodes our preference for smooth density surfaces, as these generalize better in scarcly populated parts of the state space. Since component positions shall be optimized the default regularization by an entropy-related penalty function [15], [16], [26] is not applicable. In order to avoid trivial minimization the pdf's roughness is interpreted as the integral squared curvature of the surface. A derivation of the **superficial regularization** is given in Sec. V.

Note, that both components of the target function may be written as a quadratic function of $\underline{\alpha}$, i.e.,

$$D(\mathcal{D}, \alpha_\eta, \eta) = \underline{\alpha}^{\mathrm{T}} \, \mathbf{D} \, \underline{\alpha} + \underline{d}^{\mathrm{T}} \underline{\alpha} \,, \tag{12}$$

$$R(\alpha_\eta, \eta) = \underline{\alpha}^{\mathrm{T}} \, \mathbf{R} \, \underline{\alpha} \,. \tag{13}$$

The target function assesses the quality of the estimate f. Additionally, the following conditions have to met, in order for f to be a valid conditional density for any fixed input value $\hat{\underline{x}}$

$$f(\underline{y}|\hat{\underline{x}}) \geq 0 \,, \qquad \int_{-\infty}^{\infty} f(\underline{y}|\hat{\underline{x}}) \, \mathrm{d}y = 1 \,. \tag{14}$$

The positivity constraint is trivial to assure for a mixture density by restricting each mixture weight $\alpha_i \geq 0$. The normalization constraint can only be met approximately for a finite mixture. In order to safeguard this property, the probability mass contained in the relevant ROI of the input dimension is required to calculate to the ROI size, i.e.,

$$\underline{1}^{\mathrm{T}} \underline{\alpha} = d\,, \tag{15}$$

with $d = (x_{\max} - x_{\min})$ as proposed in [15], [16]. Because this constraints only asserts the normalization over the (potentially large) interval spanned by data, additional constraints limiting the probability mass over smaller partitions are introduced,

$$\mathbf{A}\,\underline{\alpha}_\eta \leq \underline{b}\,. \tag{16}$$

In (16), $\mathbf{A}$ assigns the samples to a partition according to their location and $\underline{b}$ contains the probability mass assigned to the partitions. These constraints are less strict as (15), because they do not enforce the exact mass constraint on the partitions, in order to allow for some variation. The optimization problem consisting of the target function and the constraints, may be summarized as the following quadratic program (QP),

$$\begin{aligned} \alpha^* = \min_{\underline{\alpha}} \;\; & \underline{\alpha}^{\mathrm{T}}\,\mathbf{Q}\,\underline{\alpha} + \underline{q}^{\mathrm{T}}\underline{\alpha} \\ \text{s.t.} \;\; & \mathbf{A}\,\underline{\alpha} \leq \underline{b}\,, \\ & \underline{1}^{\mathrm{T}}\,\underline{\alpha} = \underline{d}\,, \\ & 0 \leq \underline{\alpha}_i \leq c\,. \end{aligned} \tag{17}$$

Efficient solutions to above QP may be obtained by the application of standard solvers. In the next sections, the components of the target function and their formulation as quadratic functions are presented.

IV. Distance Measure

The first term of the target function (17) is the distance between the conditional density estimate f and the epdf $f_{\mathcal{D}}$. Because f is a conditional density, the marginal density $f_x(\underline{x})$ is estimated or the empirical density $f_{\mathcal{D}}(\underline{x})$ used to calculate

$$f'(\underline{x}, \underline{y}) = f(\underline{y}|\underline{x}) \cdot f_x(\underline{x})\,. \tag{18}$$

The resulting joint density f' is compared to the epdf, which by definition is a joint density [15], [16], [26]. As a distance measure the squared integral distance of the cumulative distributions over the entire state space shall be employed. Due to the asymmetry and non-uniqueness of the standard cumulative distributions, the *localized cumulative distributions* (LCD) are compared using the modified Cramér-von Mises distance measure (mCvMD) [7]. The LCD computes a cumulative distribution based on local probability masses computed as a result the multiplication of a density with all a given kernel function at all state space positions and for all kernel positions. The mCvMD compares two LCD on the basis of their corresponding local probability masses. Similar to [15] only axis-aligned kernels are used and the efficient calculation of the distance in form of a quadratic function of the weights was adopted.

V. Superficial Regularization

The second component of the target function (17) is the regularization term. The key idea of this regularizer is that the roughness of density's surface corresponds to the generalization of the estimate. Roughly speaking, a highly oscillating surface is less likely to correctly generalize well in parts of the state with little data. Mathematically, the roughness interpreted as the curvature of the surface, i.e., integral of the curvature over entire surface. It is noteworthy, that this form of regularization does not assume or require any underlying functional dependency. As this curvature calculation is in general not solvable in closed form an approximation of the curvature by an upper bound is proposed. In the following the regularization terms for the scalar case, i.e., the case of scalar in- and output dimensions, and for the multivariate will be stated and their properties discussed.

A. Scalar Case

In the scalar case, the curvature of a 2D surface embedded in a 3D space is considered. The proposed regularizer is therefore an approximation of the standard curvature definition from the differential geometry. For the sake of brevity, the shorthand $D^{(m)} f(\underline{p}) = f_m(\underline{p})$ for $f(y|x) \equiv f(\underline{p})$ for point $\underline{p} = (y, x)$, thus $D^{(xy)} f(\underline{p}) = f_{xy}(\underline{p})$ is used. The squared Gaussian curvature for the conditional density f at point $\underline{p}$ is defined as

$$\kappa(\underline{p})^2 = \left(\frac{f_{xx}(\underline{p}) f_{yy}(\underline{p}) - f_{xy}^2(\underline{p})}{\left(1 + f_x^2(\underline{p}) + f_y^2(\underline{p})\right)^2} \right)^2 \tag{19}$$

$$\leq \left(f_{xx}(\underline{p})\, f_{yy}(\underline{p}) - f_{xy}^2(\underline{p}) \right)^2 . \tag{20}$$

As the denominator in (19) is always positive, (20) is an upper bound to the point-wise squared Gaussian curvature. The curvature over the entire state space $\mathcal{A}$ containing all input and output values is obtained by calculating

$$K = \int_{\mathcal{P}} \kappa(\underline{p})^2 \, \mathrm{d}\underline{p}\,. \tag{21}$$

Since K is used in the inner loop, we assume the component's means and covariances fixed and optimize over the weights only. Therefore, the following regularizer is proposed.

Definition 1 (Scalar Superficial Regularizer)**.** *For a conditional density $f(y|x)$ with $x, y \in \mathbb{R}$, given in the form of*

$$f(y|x) = \sum_{i=1}^{L} \alpha_i\, k^{(i)}(x, y) = \underline{\alpha}^T \underline{k}\,. \tag{22}$$

the superficial regularizer R is defined w.r.t. $\underline{\alpha}$ as

$$R := c\, \underline{\alpha}^T \mathbf{K}\, \underline{\alpha}\,, \tag{23}$$

with constant c and

$$\mathbf{K}_{ij} = \sum_{k=1}^{L} \int_{\mathbb{R}^2} k_{xx}^{(i,k)}(\underline{p})\; k_{yy}^{(i,k)}(\underline{p})\; k_{xx}^{(k,j)}(\underline{p})\; k_{yy}^{(k,j)}(\underline{p})\; d\underline{p}\,. \tag{24}$$

For the purposes of this paper, the constant c in the quadratic form (23) is typically neglected. The most import properties of the (23) are given in the following lemma.

Theorem 1. *The superficial regularizer R, as introduced in Def. 1, has the following properties:*

1) *R is an upper bound to K, as defined in (20).*
2) *For a generative model perturbed by zero-mean Gaussian additive noise, the superficial regularizer of the probabilistic model, is an upper bound on a linear transform of the squared curvature of the generative model.*

The proofs are given in the appendix. In Fig. 2 an example demonstrating the effects of Theorem 1 is given. In this example a progression from a sinusoidal to a constant generative model is given, cf. Fig. 2 (a), is used to demonstrate the change in integral curvature of the generative function, i.e., a curve in the plane, and the surface curvature. Fig. 2 (b) demonstrates that both, the generative model's *function* curvature and the probabilistic model's *surface* curvature, descrease when progressing towards the constant function, showing the impact of the second property of Theorem 1. Therefore, minimizing this regularization of the surface curvature implicitly regularizes the generative model too. Note, that this property is of course bound to the fact that the model is capable of modeling the generative model well, i.e., has an appropriate number of components in the GMM.

Other Properties:

- In contrast to the entropy-based regularization terms, e.g., the norm in the reproducing kernel hilbert space induced by the components' kernel [16], [26] or a Renyi entropy-based regularizer for arbitrary GMM [15], doesn't prefer spread in the components' mean distributions. For the superficial regularizer **no trivial minimization** of the regularization term by spreading means is possible.
- **Non-bijective functions** may be straightforwardly learned, e.g., a motion model based on two overlapping tracks. GPR would require a meta structure and the solution of data association problem or an additional estimation of bimodal noise. In contrast, Fig. 3 shows the result of estimating such a mixture of tracks with the proposed algorithm. Note, the increase in uncertainty where locally only one track is present $x \in [-1, 1]$ compared to the rest of the state space.

B. Multivariate Case

For the multivariate case with multivariate input and output dimensions, the following superficial regularizer is proposed.

Definition 2 (Multivariate Superficial Regularizer)**.** *For a conditional density $f(\underline{y}|\underline{x})$ with $\underline{x} \in \mathbb{R}^N$, $\underline{y} \in \mathbb{R}^M$, given in the form of*

$$f(\underline{y}|\underline{x}) = \underline{\alpha}^T \left(\underline{k}_1 \circ \ldots \circ \underline{k}_N \circ \underline{k}_{N+1} \circ \ldots \circ \underline{k}_{N+M} \right) , \quad (25)$$

the superficial regularizer R is defined w.r.t. $\underline{\alpha}$ as

$$R := c\, \underline{\alpha}^T \mathbf{K}\, \underline{\alpha} , \quad (26)$$

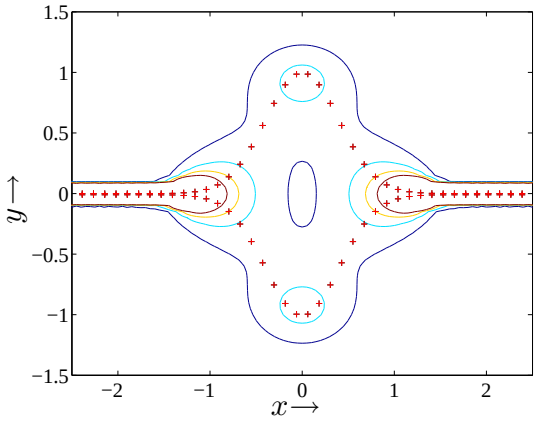

Figure 3. Conditional density function $f(y|x)$ representing a non-bijective function, e.g., as arising from data collected for two overlapping tracks.

with constant c and $\mathbf{K}_{ij} = \int_{\mathbb{R}^2} \prod_{d=1}^{N+M} k_{dd}^{(i,j)}(\underline{p})\ d\underline{p}$.

In Def. 2, the symbol $\circ$ denotes the Hadamard product. The definition makes the strong assumption of separability along the dimensions. The multivariate superficial regularizer subsumes the scalar superfical regularizer as a special case, but is not proven to be an upper bound on the surface curvature for higher dimensions. Yet, the estimation of non-bijective functions is trivially possible for the multivariate case.

VI. Algorithm

In this section, an overview over the entire optimization algorithm including both loops is given. The pseudo-code of the algorithm is given in Alg. 1.

Algorithm 1 Conditional Density Estimation

Input: $\mathcal{D}$
Calculate $\theta_{k=0}$ as in Sec. III-A ▷ Initial Values
repeat ▷ Outer Loop
  Calculate $\{\phi\}_k$ from $\mathcal{D}, \{\underline{\mu}\}_k$
  $(\alpha_k, \nu_k) \leftarrow$ OptimizeWeights$(\mathcal{D}, \eta_k)$ ▷ Inner Loop
  $\eta_{k+1} \leftarrow$ Update(η_k) ▷ Update $\{\underline{\mu}\}$, $\mathbf{K}$, λ
until $\Delta(\nu_{k-1}, \nu_k) > \varepsilon$
function OptimizeWeights$(\mathcal{D}, \eta)$ ▷ Inner Loop
  Calculate $D(\mathcal{D}, \alpha_\eta, \eta)$ and $R(\alpha_\eta, \eta)$
  Calculate constraints from (14)
  Compose and solve QP (17)
  return Weights α_k, Value of (17) ν_k
end function
Output: $\theta^* \leftarrow \theta_k$

VII. Assumptions and Computational Complexity

In summary, the limitations of the presented algorithm are:

- The number of components of the GMM needs to be high enough to approximate a conditional density function sufficiently well.

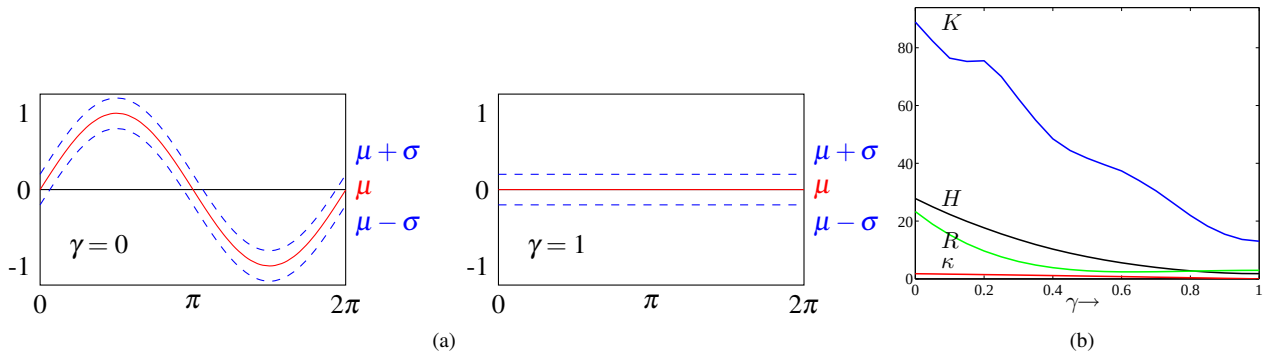

Figure 2. Correlation of the curvature of the surface and the generative model. The progression from a constant to a sine function depending on the value of γ (a) and numerically calculated curvature measures for varying progression parameters γ: Gaussian curvature K (blue), mean curvature H (black), the superficial regularizer's value R (multiplied by 10^{-15}, green), and curvature of the generative model κ (red) are shown in (b).

- Similar to GPRs [25], the bottleneck of the constrained nonlinear optimization is $\{\underline{\mu}\}$, because the optimization problem scales with $(N + M) \cdot |\{\underline{\mu}\}|$, e.g., for 20 scalar components a 20-dimensional optimization approach has to be solved.
- For training time, only vague statements about the complexity of the algorithm are possible as these are implementation dependent. Solving a generic QP involves is $\mathcal{O}(n^3)$, with n the number of variables. Yet, the given problem lends itself to local decomposition, e.g., by chunking. During run-time only the cost for evaluation or multiplication of a GMM is necessary.

VIII. Experiments

In this section the proposed algorithm is compared with indirect and direct methods for obtaining conditional densities based on the estimate's quality for synthetic noisy functional dependency and the performance in nonlinear filtering, i.e., the potential field of use for the proposed method.

A. Comparison of Conditional Density Function Estimates

For comparing the estimation quality of the proposed approach, a sinusoidal system with additive noise was used

$$\boldsymbol{y} = \tfrac{2}{3}\, \boldsymbol{x}^2\, \sin(\boldsymbol{x}) + \boldsymbol{w}\,, \qquad \boldsymbol{w} \sim \mathcal{N}(0, 0.2)\,. \tag{27}$$

For training, 100 random samples were drawn according to (27) in the x-interval of $[-\pi, \pi]$ and additional 100 random samples were generated for validation. The quality of conditional density function estimates is assessed by comparing the negative log-likelihood (NL) scores for the validation given the estimates. Lower NL values indicate better performance. The proposed approach is compared with GPR, SVR, several variants of EM, as well as the entropy-regularized approach [15] (CDE-ENT), resembling the proposed approach most. The standard GPR implementation from [21] was used, the SVR code is identical to [16] as is the CDE-ENT code to [15]. The EM implementation of Matlab was used and restricted to estimate only covariances with zero off-diagonal entries. EM1/2 estimate homoscedastic/heteroscedastic GMM with the same number of components as CDE-ENT, where EM3/4 estimate homoscedastic/heteroscedastic GMM with AIC chosen number of components. The proposed approach (CDE-SF) according to Alg. 1 was used with no variable means and $\lambda = 0.5$ fixed for D and R normalized to by their respective values for uniform weights. The results of the comparison are given in Tab. VIII-A. The results show the **sparse, high quality of the conditional densities** produced by the proposed approach.

B. Comparison for Nonlinear Filtering Problems

For comparing the CDE as substrate to nonlinear filtering, the Kitagawa growth process [10], as presented in [1], comprised of the nonlinear system and measurement equations

$$\begin{aligned} \boldsymbol{x}_{k+1} &= 0.5\, \boldsymbol{x}_k + \tfrac{25\, \boldsymbol{x}_k}{1+\boldsymbol{x}_k^2} + \boldsymbol{w}_k\,, \\ \boldsymbol{y}_{k+1} &= 5\, \sin(2\, \boldsymbol{x}_{k+1}) + \boldsymbol{v}_{k+1}\,, \end{aligned} \tag{28}$$

was used. Identical to [1], $\boldsymbol{w}_k \sim \mathcal{N}(w, 0.2)$ and $\boldsymbol{v}_{k+1} \sim \mathcal{N}(v_{k+1}, 0.01)$ were used and randomly distributed 100 points in $[-10, 10]$ were generated for training. The estimation quality is compared using a prior normal density with $\mu_0 \in [-10, 10]$ and $\sigma_0 = 0.5$. The successive states were estimated for 200 independent $x_0^{(i)}$ and $y_1^{(i)}$. The estimation quality is given for three quartiles of the NL for the true state given the respective models and the Mahalanobis distance $\mathcal{M}(x)$ between the true and the estimated state. The NL distribution over the quantiles give an simple insight into the distribution of the NL and $\mathcal{M}(x)$ shows the error relative to the state estimate's uncertainty. For both scores lower values indicate better performance. The averaged results over ten runs trials are given in Tab. II. The proposed CDE approach was used to estimate the probabilistic models of a stamdard Gaussian mixture filter (GMF+SF). For GMF+SF the identical setup was used as in the prior experiment, whereas (GMF+SF+XV) was trained with an additional five meta-optimized components. Both GMF yield good negative log-likelihood results for the growth process in comparison to the EKF, UKF, GP-UKF, and GP-ADF. Additionally, a clear **improvement in the performance can be observed for GMF+SF+XV compared to GMF+SF** can be observed. The reason for this improvement is that the system model of the growth process induces a strong nonlinearity around the $(0, 0)$. Due to random sampling only

Table I
NEG. LOG-LIKELIHOOD SCORES AND COMPONENTS NUMBERS FOR THE RESULTS OF EM1-4, GPR, SVR, CDE-ENT AND CDE-SF.

	EM1	**EM2**	**EM3**	**EM4**	**GPR**	**SVR**	**CDE-ENT**	**CDE-SF**
NL	3.87	3.97	3.58	3.62	0.23	0.45	0.15	± 0.33
$\pm\,\sigma$	± 1.44	± 1.60	± 1.31	± 1.89	± 0.03	± 0.27	± 0.11	± 0.22
# Comp.	66.7	66.7	11.3	9	N/A	96.2	66.7	67.6

Table II
NEGATIVE LOG-LIKELIHOOD AND MAHALANOBIS DISTANCE RESULTS FOR THE GROWTH PROCESS [10].

	$NL^{0.25}$		$NL^{0.5}$		$NL^{0.75}$		$\mathcal{M}(x)$	
EKF	1063.41	± 378.67	29314.00	± 1103.40	274910.52	± 2068.77	2071897.02	± 2962352.80
UKF	60.50	± 4.12	628.52	± 31.74	2407.02	± 53.32	1030.01	± 4529.04
GP-UKF	65.68	± 4.99	420.06	± 36.26	1769.34	± 177.24	3918.10	± 44902.12
GP-ADF	59.08	± 3.03	261.02	± 17.44	1083.44	± 82.25	26.16	± 48.55
GMF-SF	44.22	± 15.79	254.73	± 78.96	928.08	± 315.01	149.17	± 287.32
GMF-SF+*XV*	41.82	± 14.70	224.07	± 58.28	806.93	± 240.7	98.37	± 181.72

a few data points cover this nonlinearity. Yet, the additional optimized means help fill this "gap" and thereby improve the performance of the GMF+SF. This can be seen in Fig. 4.

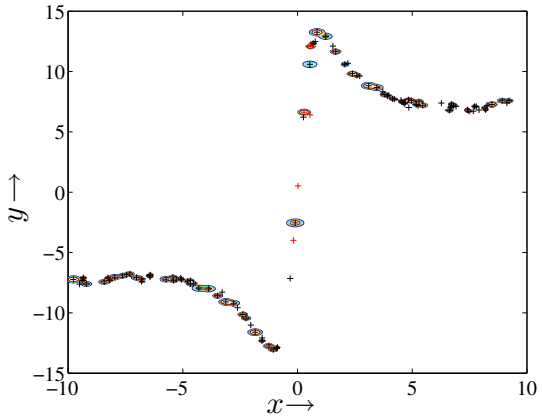

Figure 4. Conditional density function estimated based on an exemplary sample set (black crosses) of (28) and additional optimized components (red crosses).

IX. CONCLUSION

In the present paper an estimation algorithm for conditional densities represented as heteroscedastic Gaussian mixture densities between continuous random variables based on noisy samples is proposed. The algorithm is an iterative nested optimization scheme, optimizing hyper-parameters and all mixture components' parameters. The main contributions are an improved generalization quality of the estimates due to the a novel superficial regularizer, the consideration of model uncertainty by means of adaptive covariances, and an efficient distance-based estimation algorithm. The new superficial regularization term is an affine transformation of the integral squared surface curvature of the conditional density and for additive, zero-mean normal noise is shown to regularize of generative model implicitly. The experiments shows solutions are sparse, smooth, and generalize well especially w.r.t. nonlinear filtering applications.

As future work learning from large data sets possibly by means of distributable/parallelizable computations [18], [20] or approximations [4], [8] should be investigated. Additionally, efficient incremental updating of conditional densities would be interesting.

APPENDIX

Proof of Theorem 1: The proof is given per property: **Property 1:** The square of the surface's curvature of f is simplified by exploiting (22), the linearity of the integral, and the commutativity of the inner product, giving rise to

$$\begin{aligned} K &= \int_{\mathbb{R}^2} \left[f_{xx}(\underline{p})\, f_{yy}(\underline{p}) - f_{xy}^2(\underline{p}) \right]^2 \,\mathrm{d}\underline{p} \\ &= \int_{\mathbb{R}^2} \left[\underline{\alpha}^{\mathrm{T}} \underline{k}_{xx}\, \underline{\alpha}^{\mathrm{T}} \underline{k}_{yy} - (\underline{\alpha}^{\mathrm{T}} \underline{k}_{xy})^2 \right]^2 \,\mathrm{d}\underline{p} \\ &= \int_{\mathbb{R}^2} \left[\underline{\alpha}^{\mathrm{T}} \underline{k}_{xx}\, \underline{k}_{yy} \underline{\alpha}^{\mathrm{T}} - (\underline{\alpha}^{\mathrm{T}} \underline{k}_{xy} k_{xy}^{\mathrm{T}} \underline{\alpha}) \right]^2 \,\mathrm{d}\underline{p} \\ &\leq \int_{\mathbb{R}^2} [\, \underline{\alpha}^{\mathrm{T}} \underbrace{(\underline{k}_{xx}\, \underline{k}_{yy}^{\mathrm{T}})}_{\mathbf{M}}\, \underline{\alpha}]^2 \,\mathrm{d}\underline{p}\,. \end{aligned}$$

Further simplification of (29) allows the upper bound by

$$\begin{aligned} \int_{\mathbb{R}^2} [\, \underline{\alpha}^{\mathrm{T}} \mathbf{M}\, \underline{\alpha}\,]^2 \,\mathrm{d}\underline{p} &= \int_{\mathbb{R}^2} \underline{\alpha}^{\mathrm{T}} \mathbf{M}\, \underline{\alpha}\, \underline{\alpha}^{\mathrm{T}} \mathbf{M}\, \underline{\alpha} \,\mathrm{d}\underline{p} \\ &\leq c_{\mathbf{M}} \int_{\mathbb{R}^2} \underline{\alpha}^{\mathrm{T}} \mathbf{M}^2\, \underline{\alpha} \,\mathrm{d}\underline{p} = c_{\mathbf{M}}\, \underline{\alpha}^{\mathrm{T}} \mathbf{K}\, \underline{\alpha}\,, \end{aligned} \quad (29)$$

giving the desired result with an appropriate constant $c_{\mathbf{M}}$ and

$$\begin{aligned} \mathbf{K}_{ij} = \sum_{k=1}^{L} \int_{\mathbb{R}^2} & k_{xx}^{(i,k)}(\underline{p})\, k_{yy}^{(i,k)}(\underline{p}) \\ & \cdot k_{xx}^{(k,j)}(\underline{p})\, k_{yy}^{(k,j)}(\underline{p}) \,\mathrm{d}\underline{p}\,. \end{aligned} \quad (30)$$

Property 2: Considering $x \in \mathbb{R}$, the squared curvature of $y = g(x)$, i.e., a curve in the $xy-$*plane*, is

$$\kappa_g^2(x) = \frac{\left(D^{xx} g(x)\right)^2}{\left[1 + \left(D^x g(x)\right)^2\right]^3} \leq \left(D^{xx} g(x)\right)^2\,, \quad (31)$$

and an upperbound on the integrated squared curvature of $y = g(x)$ is

$$\int_{\mathcal{X}} \kappa_g^2(x)\,\mathrm{d}x \leq \int_{\mathcal{X}} \left(D^{xx} g(x)\right)^2\,\mathrm{d}x\,, \tag{32}$$

The curvature of g is related to the curvature of the surface f

$$\begin{aligned}
\kappa(\underline{p})^2 &\leq \left(f_{xx}(\underline{p})\, f_{yy}(\underline{p}) - f_{xy}^2(\underline{p})\right)^2 \\
&= (\,[D^{xx} f(y|x)][D^x g(x)]^2 - [D^x f(y|x)][D^{xx} g(x)] \\
&\qquad - [D^{xy} f(y|x)]^2\,)^2 \\
&= (\,-[D^x f(y|x)][D^{yy} f(y|x)]\,)^2\,(D^{xx} g(x))^2\,. \qquad (33)
\end{aligned}$$

For Gaussian additive noise, integrating (33) over $\boldsymbol{x}$ yields

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} \kappa(\underline{p})^2\,\mathrm{d}y\,\mathrm{d}x = \int_{\mathcal{X}} c\cdot(D^{xx} g(x))^2\,\mathrm{d}x\,, \tag{34}$$

with $c \in \mathbb{R}^+$ and c independant of g. The result then follows from $K \leq R$. ∎

References

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

[2] 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)*, 39(1):1–38, 1977.

[3] 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)*, pages 134–140, Montreal, Canada, August 1995.

[4] Henning Eberhardt, Vesa Klumpp, and Uwe D. Hanebeck. Density Trees for Efficient Nonlinear State Estimation. In *Proceedings of the 13th International Conference on Information Fusion (Fusion 2010)*, Edinburgh, United Kingdom, July 2010.

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

[6] M.A.T. Figueiredo and A.K. Jain. Unsupervised Learning of Finite Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 381–396, 2002.

[7] 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)*, pages 33–39, Seoul, Republic of Korea, August 2008.

[8] M. P. Holmes, A. G. Gray, and C. L. Isbell. Fast Kernel Conditional Density Estimation: A Dual-Tree Monte Carlo Approach . *Computational Statistics and Data Analysis*, 54(7):1707 –1718, 2010.

[9] M. Huber, D. Brunn, and U. D. Hanebeck. Closed-Form Prediction of Nonlinear Dynamic Systems by Means of Gaussian Mixture Approximation of the Transition Density. In *Proceedings of the 2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2006)*, pages 98–103, Heidelberg, Germany, September 2006.

[10] G. Kitagawa. Monte Carlo Filter and Smoother for non-Gaussian Nonlinear State Space Models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

[11] J. Ko and D. Fox. GP-BayesFilters: Bayesian Filtering Using Gaussian Process Prediction and Observation Models. In *Proceedings of the 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3471–3476, Nice, France, September 2008.

[12] J. Ko, D. Klein, D. Fox, and D. Haehnel. Gaussian Processes and Reinforcement Learning for Identification and Control of an Autonomous Blimp. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 742–747, Rome, Italy, April 2007.

[13] 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)*, pages 1901–1907, San Diego, California, October 2007.

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

[15] Peter Krauthausen and Uwe D. Hanebeck. Regularized Non-Parametric Multivariate Density and Conditional Density Estimation. In *Proceedings of the 2010 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2010)*, Salt Lake City, Utah, September 2010.

[16] Peter Krauthausen, Marco F. Huber, and Uwe 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.

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

[18] E. Osuna, R. Freund, and F. Girosi. An Improved Training Algorithm for Support Vector Machines. In *Proceedings of the 1997 IEEE Workshop on Neural Networks for Signal Processing (IEEE-NNSP)*, Amelia Island, Florida, sep 1997.

[19] E. Parzen. On Estimation of a Probability Density Function and Mode. *Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[20] John C. Platt. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 12, pages 41–65. MIT Press, 1998.

[21] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, Massachusetts, 2006.

[22] B. Schölkopf and A. Smola. *Learning with Kernels: support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning series. MIT Press, Cambridge, Massachusetts, 2002.

[23] 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.

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

[25] E. Snelson and Z. Ghahramani. Sparse Gaussian Processes using Pseudo-inputs. In *Advances in Neural Information Processing Systems 18*, Vancouver, Canada, 2005.

[26] V. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer, New York, 2. edition, 2000.