2011 American Control Conference
on O'Farrell Street, San Francisco, CA, USA
June 29 - July 01, 2011

Incorporating Prior Knowledge into Nonparametric Conditional Density Estimation

Peter Krauthausen, Masoud Roschani, and Uwe D. Hanebeck

Abstract— **In this paper, the problem of sparse nonparametric conditional density estimation based on samples *and* prior knowledge is addressed. The prior knowledge may be restricted to parts of the state space and given as generative models in form of mean-function constraints or as probabilistic models in the form of Gaussian mixture densities. The key idea is the introduction of additional constraints and a modified kernel function into the conditional density estimation problem. This approach to using prior knowledge is applicable to all nonparametric conditional density estimation approaches phrased as constrained optimization problems. The quality of the estimates, their sparseness, and the achievable improvements by using prior knowledge are shown in experiments for both Support-Vector Machine-based and integral distance-based conditional density estimation algorithms.**

I. Introduction

Many applications of control, such as the control of chemical or mechanical processes, involve state estimation for noise-corrupted nonlinear systems. For all these applications, probabilistic state estimation, e.g., with Dynamic Bayesian Networks [1], is fundamental and high-quality conditional densities representing the systems are necessary in order to obtain accurate state estimates. In this paper, conditional density estimation based on samples *and* prior knowledge is investigated. This is an especially important problem for the practitioner, as in the *real world* hardly any data acquisition is flawless, but typically an iterative approach of mutually alternating data acquisition and evaluation steps is adopted. For example, in Human-Robot Cooperation, it is crucial for the robot's control to have a position and posture estimate of the user in order to assist him and to avoid collisions. Since robots are trained in limited scenarios only, e.g., *cooking*, there exists plenty of data about some typical movements, but hardly any knowledge about movement patterns exceeding this scenario, e.g., *cleaning*. When training a robot for a wide range of scenarios, it would be beneficial to reuse the prior knowledge in form of already compiled models, in order to avoid time-consuming (batch) retraining. This paper aims at alleviating this problem by addressing the conditional density estimation problem based on samples *and* prior knowledge. The prior knowledge may be represented in the form of a generative model and/or an additional probabilistic model. To the best of our knowledge, the present work is the first to address sparse kernel conditional density estimation from samples, when additional prior knowledge is given.

P. Krauthausen and U. D. Hanebeck are with the Intelligent Sensor-Actuator-Systems Laboratory (ISAS), {Peter.Krauthausen@kit.edu, Uwe.Hanebeck@ieee.org}, M. Roschani is with the Vision and Fusion Laboratory (IES), Masoud.Roschani@kit.edu, Institute for Anthropomatics, Karlsruhe Institute of Technology, Germany. 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*".

II. Related Work

The related work may be categorized into approaches to *1) conditional density estimation* and *2) the incorporation of prior knowledge* for each of these approaches:

1) Conditional Densities Estimation (CDE): There are essentially three approaches to CDE for continuous random variables: (a) the conditional density, i.e., the probabilistic model, is directly provided by the user, (b) a generative model and noise specifications are available, or (c) only samples are given and the conditional density needs to be estimated. In contrast to *density estimation*, e.g., with EM or Parzen-Window / kernel density estimation algorithms [2], [3], [4], only little research has been performed on *conditional density estimation*, cf. [5] for an overview on nonparametric CDE. The few approaches may be categorized according to the representation of the conditional density estimate f: f is given as a fraction $f(\underline{y}|\underline{x}) = \frac{f(\underline{y},\underline{x})}{f(\underline{x})}$ [5], i.e., as a straight-forward extension to density estimation, or in the form of a sparse conditional mixture density [6], [7], [8]. The former approach will always yield valid conditional densities, which is not the case for the latter approach. Yet, the latter approach's mixture representation of the conditional densities will allow for efficient closed-form calculations, when using the estimated conditional densities for Bayesian inference, e.g., in Bayesian networks [9]. For this reason, the latter approach is considered only.

2) Incorporating Prior Knowledge: None of the above approaches to CDE considers the inclusion of prior knowledge in the form of a generative or probabilistic model even if restricted to parts of the state space only. For the sparse nonparametric CDE approaches, prior knowledge may be introduced in a similar way to Support-Vector Machines (SVM) [10]. An overview about incorporating prior knowledge into SVMs is provided in [11], which can be summarized by two approaches: incorporating information by changing the kernel function or not, i.e., changing the representation of the solution. The latter may correspond to the addition of *virtual* SVs/data or the use of "knowledge bases" in form of *if-then-else* rules, which add constraints to the optimization problem [12]. A change in the kernel as well as in the regularizer is proposed in [13] to emphasize local features and incorporate invariances. Other approaches include extending the optimization problem with a term penalizing the distance between a prior knowledge function,

978-1-4577-0079-8/11/$26.00

the data, and the estimate [14]. Note that the incorporation of prior knowledge for conditional density estimation is especially hard as additional constraints have to be asserted to obtain valid conditional densities, i.e., the probability mass of the conditional density estimate has to be non-negative and integrate to one for all fixed input values. Typically, these conditions are relaxed and met only approximately.

In this paper, an approach to including prior knowledge in the form of mean constraints of a generative model and/or a probabilistic model given as a Gaussian mixture density (GM) is proposed—abstracting from the specific formulation of the estimation problem, e.g., error calculation and roughness penalties [6], [7], [8]. The key idea is to perform CDE using samples and prior knowledge simultaneously by introducing additional constraints to the optimization problem and by modifying the kernel function. The proposed approaches are shown to improve the quality of the conditional density estimates, which are given in the form of axis-aligned GMs. Depending on the specific formulation of the optimization problem, f may be very sparse.

The rest of this paper is organized as follows. In Sec. III, a generic formulation of the conditional density estimation problem as an abstract optimization problem will be given. In Sec. IV, the problem is formulated and in Sec. V, the incorporation is derived for the two considered forms of prior knowledge. The improved quality of the conditional density estimates will be shown in Sec. VI.

III. Conditional Density Estimation

This section revises and generalizes recent advances towards sparse kernel conditional density estimation presented in [6], [7], [8]. In general, a set $\mathcal{D}$ of i.i.d. random samples $(\underline{x}_i, \underline{y}_i)$ is given. The corresponding empirical probability density function [3] is given in the form of

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

i.e., a mixture of Dirac distributions $\delta(.)$, with

$$\underline{x}_i := [x_i^{(1)}\ \dots\ x_i^{(M)}]^{\mathrm{T}} \in \mathbb{R}^{\mathrm{M}},\ \ \underline{y}_i \in \mathbb{R}^{\mathrm{N}}\,,$$

and the conditional density function $\tilde{f}$ underlying the data shall be estimated. An estimate f is determined in the form of an axis-aligned GM, i.e., the components' covariances only have non-zero entries on the main-diagonal

$$f(\underline{y}|\underline{x}) = \sum_{i=1}^{|\mathcal{D}|} \alpha_i \underbrace{\prod_{j=1}^{M} k\left(x^{(j)}; \mu_{xi}^{(j)}, \sigma_{xi}^{(j)}\right) \prod_{l=1}^{N} k\left(y^{(l)}; \mu_{yi}^{(l)}, \sigma_{yi}^{(l)}\right)}_{= \mathcal{K}_i\left([\underline{x};\underline{y}]^{\mathrm{T}}, [\underline{\mu}_{xi};\underline{\mu}_{yi}]^{\mathrm{T}}\right)}\,. \tag{2}$$

This mixture comprises $|\mathcal{D}|$ components, each with one Gaussian kernel $k(x; \mu, \sigma) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}$ per in- and output dimension. In (2), the parameters are set to $\mu_{xi}^{(k)} = x_i^{(k)}$, $\mu_{yi}^{(k)} = y_i^{(k)}$, $\sigma_{xi}^{(k)} = \sigma_{xj}^{(k)}$ and $\sigma_{yi}^{(k)} = \sigma_{yj}^{(k)}$, i.e., the mixture components are located at the sample positions. The variances for each dimension are fixed and determined *a priori*. The only remaining free variables are the weights, i.e., $\underline{\alpha} = [\alpha_1\ \dots\ \alpha_{|\mathcal{D}|}]^{\mathrm{T}}$. In order to determine sparse conditional densities, the generic optimization problem [8]

$$\begin{aligned} \underline{\alpha}^* = \arg\min_{\underline{\alpha}} \quad & D(\,\tilde{f},\, f\,) + \lambda\, R(\,f\,) \\ \text{s.t.} \quad & \underline{\alpha}^{\mathrm{T}}\, \underline{1} = \mathrm{c},\ 0 \le \alpha_i \le \nu'\,, \end{aligned} \tag{3}$$

has to be solved. The target function in (3) consists of two parts: D and R. D measures the deviation between the true conditional density $\tilde{f}$, approximated by the data $f_{\mathcal{D}}$, and the estimate f. Typically, the error between the respective cumulative distributions is calculated [6]. Many choices for D are possible, for example the (ε-insensitive) l_1-error at the sample points as in Support-Vector Regression (SVR) [10] or the integral squared error over the entire state space [8].

R is a term penalizing the density's roughness, e.g., the norm in the RKHS induced by the kernel [6], [7], the squared l_2-norm of $\underline{\alpha}$, or a term related to the Renyi entropy of f [8]. The trade-off between these two factors is found by λ. Additionally, constraints have to be asserted for f to be a valid conditional density [8]. As Gaussian kernels are used, non-negativity of (2) is achieved by $\alpha_i \ge 0$. Asserting the condition $\int_{\mathbb{R}^{\mathrm{N}}} f(\underline{y}|\hat{\underline{x}})\ \mathrm{d}y = 1$ for all $\hat{x} \in \mathbb{R}$ may be relaxed to hold for all sample points only [6]. A constraint based on the volume c of the state space spanned by the data is proposed in [7], [8]. The optimization problem requires setting all parameters, except for $\underline{\alpha}$, in advance. For this purpose, an algorithm has been devised in [7]. In summary, the problem is generic and allows for several choices with regard to the error measure, the roughness penalty, and the constraints. In the rest of this paper, a generic approach to introducing prior knowledge into this optimization problem is devised, abstracting from the specific optimization problem.

IV. Problem Formulation

The problem considered in this paper is conditional density estimation based on i.i.d. random samples $(\underline{x}_i, \underline{y}_i) \in \mathcal{D}$ and given prior knowledge. The prior knowledge for some input range may be given as a generative model

$$\underline{\boldsymbol{y}} = \underline{a}(\underline{\boldsymbol{x}}) + \underline{\boldsymbol{v}} \ \text{ with } \underline{\boldsymbol{v}} \sim \mathcal{N}(0, \boldsymbol{\Sigma})\,,$$

or as a probabilistic model

$$f(\underline{y}|\underline{x})\,,\ \text{e.g.,}\ f(\underline{y}|\underline{x}) = \mathcal{N}(\underline{y}; \underline{a}(\underline{x}), \boldsymbol{\Sigma}(\underline{x}))\,,$$

as shown in Fig. 1 (a). These two forms of prior knowledge resemble the design steps described in the Sec. II and arise from prior analysis, i.e., regression of the mean function or estimation of a probabilistic model for a specific purpose, e.g., *cooking*. This paper is restricted to specific approximations of these forms. The proposed incorporation is extendible to other, more general representations, e.g., a mean function with a desired σ-bound. For simplicity, the case that prior knowledge is restricted to parts of the state space is considered from now on only.

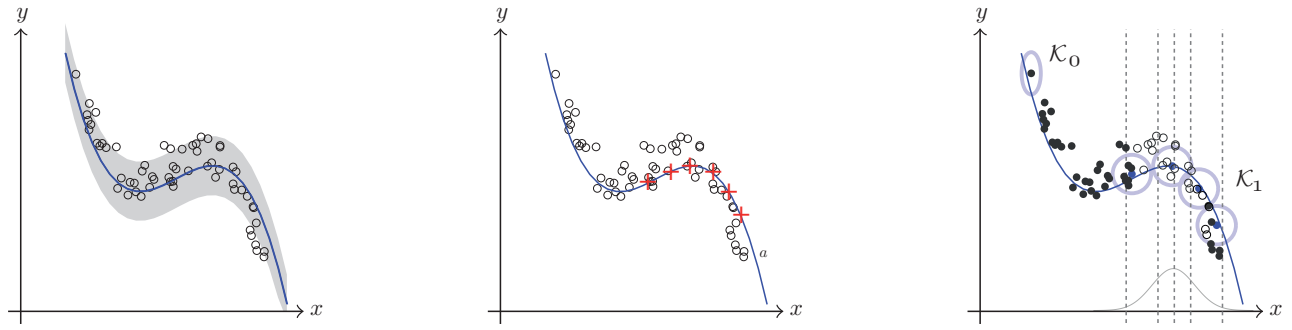


Fig. 1. (a) True probabilistic model ($\mu \pm \sigma$) and samples, (b) prior knowledge in form of mean constraints obtained from the generative model (red), and (c) prior knowledge in form of an additional GM $\mathcal{K}_1$, with components (light blue) centered at fixed points (dark blue). The mixture of kernels $\mathcal{K}_0$ (depicted top left) and $\mathcal{K}_1$ is governed by the x-location weighted by the overlaid Gaussian weighting function over $\boldsymbol{x}$ and its σ-bounds (dashed lines).

V. Incorporation of Prior Knowledge

Depending on the type of prior knowledge, the introduction of the prior knowledge into (3) differs. In the next sections, prior knowledge in the form of a generative model, Fig. 1 (b), and a probabilistic model, Fig. 1 (c), will be considered.

A. Mean Function of the Generative Model

Given the mean function $\underline{a}(\underline{x})$ of a generative model, the conditional density estimate obtained from solving (3) shall have conditional expectations identical to the mean function for all fixed inputs $\hat{\underline{x}}$, i.e.,

$$\int_{\mathbb{R}^N} \underline{y}\, f(\underline{y}|\, \underline{x} = \hat{\underline{x}})\ \mathrm{d}\underline{y} \overset{!}{=} \hat{\underline{y}} = \underline{a}(\hat{\underline{x}})\,. \tag{4}$$

Incorporating (4) into the optimization problem corresponds to penalizing deviations of the conditional expectations to the mean function, i.e., solutions with a lower deviation are preferred. This approach for incorporating knowledge about the generative model's mean function may be understood as a "knowledge base" rule [12], but with real-valued (in)equality consequence. Even if only prior knowledge for a restricted interval, e.g., $[\underline{x}_{\min}, \underline{x}_{\max}]$, is considered, this corresponds to an infinite number of constraints. Therefore, an approximation in the form of

$$l_1^\varepsilon \left(\int_{\mathbb{R}^N} \underline{y}\, f(\underline{y}|\, \underline{x} = \hat{\underline{x}}_i)\ \mathrm{d}\underline{y}\,,\ \hat{\underline{y}}_i \right) \leq \xi_i\,, \quad \forall \hat{\underline{x}}_i \in \mathcal{I}\,, \tag{5}$$

is proposed, where $\hat{\underline{y}}_i = \underline{a}(\hat{\underline{x}}_i)$ are the mean function's values at a set of distinct points $\mathcal{I} \subset [\underline{x}_{\min}, \underline{x}_{\max}]$, cf. Fig. 1 (b), and l_1^ε is the ε-insensitive loss-function allowing for small violations of the constraints. Many choices for the $\hat{\underline{x}}_i$ are possible, e.g., the sample locations or equidistant $\underline{x}$-positions in $[\underline{x}_{\min}, \underline{x}_{\max}]$, as used for the experiments in Sec. VI. The constraints restrict the optimization variables $\underline{\alpha}$, i.e., one obtains for each point $\hat{\underline{x}}_i \in \mathcal{I}$

$$\int_{\mathbb{R}^N} \underline{y}\, f(\underline{y}|\, \underline{x} = \hat{\underline{x}}_i)\ \mathrm{d}\underline{y} = \underline{\alpha}^{\mathrm{T}} \underline{k}_y(\underline{x} = \hat{\underline{x}}_i)\,, \quad \forall \hat{\underline{x}}_i \in \mathcal{I}\,, \tag{6}$$

where $\underline{k}_y(\underline{x})$ is the vector of the components of (2)

$$k_y^{(i)}(\underline{x}) = \prod_{j=1}^{M} k\left(x^{(j)}; \mu_{xi}^{(j)}, \sigma_{xi}^{(j)}\right)\,, \tag{7}$$

after integrating $\boldsymbol{y}$. With (6) and the following reformulations of (5), one arrives at the constraints for the weights for each point $\hat{\underline{x}}_i \in \mathcal{I}$

$$\left| \int_{\mathbb{R}^N} \underline{y}\, f(\underline{y}|\, \underline{x} = \hat{\underline{x}}_i)\ \mathrm{d}\underline{y} \ - \ \hat{\underline{y}}_i \right| = \left| \underline{\alpha}^{\mathrm{T}} \underline{k}_y(\hat{\underline{x}}_i) - \hat{\underline{y}}_i \right| < \varepsilon + \xi_i\,, \quad \forall \hat{\underline{x}}_i \in \mathcal{I}\,. \tag{8}$$

Insertion of (8) into (3) is then performed by replacing the absolute value in the loss-function by a positive and a negative constraint.

Restrictions & Properties: The proposed approach requires only minor changes to (3) and yields conditional density estimates in the form of (2) with at most $|\mathcal{D}|$ components, but depending on the algorithm used, significantly less. The approach suffers from the discretization of $\underline{a}(\underline{x})$, which may be crude. The results in Sec. VI show the approximation's insignificance under reasonable conditions. By construction, the accuracy with which the constraints may be met, depends on the distribution and the number of data points. If only a few data points are distributed over the state space, the mean function will be met only vaguely—irrelevant of the number of constraints. As will be shown in Sec. VI, this is typically not a problem and the approach yields favorable results. The parameter ε may be determined as an additional hyperparameter, cf. [7], [8].

B. Location-Based Mixture Kernel

When a probabilistic model is given, it is more challenging to incorporate this prior knowledge into the CDE. In the following, the key idea, its implementation, and the properties of this approach are explained.

Key Idea: The key idea is to perform CDE not with one fixed default kernel for all samples, but to combine the default kernel with a modified kernel encoding the prior knowledge per sample. The combination is weighted by the confidence in the prior model, which is modeled as a *weighting function* w.r.t. the location in the state space. The resulting combined kernel replaces the default kernel in the calculation of D, R, and the constraints for (3).

Implementation: The given weighting function, denoted by $\mathrm{P} : (\{0,1\}, \mathbb{R}^{\mathrm{M}}, \mathbb{R}^{\mathrm{N}}) \mapsto [0,1]$, determines the convex combination of the default kernel $\mathcal{K}_0$ and the *a priori* obtained probabilistic model (PM), i.e., kernel $\mathcal{K}_1$. The sum of the values of P over the discrete events is required to be one. This convex combination of kernels is again a valid kernel [15]. In general, any function P producing valid convex combinations over the kernels is allowed, e.g., hybrid conditional densities. The default kernel $\mathcal{K}_0$ corresponds to one component of (2) and the kernel $\mathcal{K}_1$, encoding the prior knowledge, is given in the form of (2). For each sample, the weighting function gives the mixture proportions $\underline{p}$ of the default kernel $\mathcal{K}_0$ and the PM kernel $\mathcal{K}_1$ to obtain the new mixture component $\mathcal{K}_i$ of (2) as

$$\mathcal{K}_i([\underline{x};\underline{y}]^{\mathrm{T}}, [\underline{u};\underline{v}]^{\mathrm{T}}) \;=\; \begin{bmatrix} \mathcal{K}_0([\underline{x};\underline{y}]^{\mathrm{T}}, [\underline{u};\underline{v}]^{\mathrm{T}}) \\ \mathcal{K}_1([\underline{x};\underline{y}]^{\mathrm{T}}, [\underline{u};\underline{v}]^{\mathrm{T}}) \end{bmatrix}^{\mathrm{T}} \cdot \underline{p} \;. \quad (9)$$

Based on the location in the state space $[\underline{u};\underline{v}]^{\mathrm{T}}$, P assigns a value to the event $k = i$, that kernel $\mathcal{K}_i$ produced this sample

$$\underline{p} := [\mathrm{P}(k=0|\underline{u},\underline{v}) \;\; \mathrm{P}(k=1|\underline{u},\underline{v})]^{\mathrm{T}} \;.$$

The variation in kernel mixture proportions depending on the location in state space can be seen in Fig. 1 (c), where the dashed lines indicate the σ-bounds of P in the form of a Gaussian function over x only. Following the idea that prior knowledge is restricted to an interval of the state space, e.g., a soft weighting function

$$\mathrm{P}(k=1|\underline{u},\underline{v}) = \sqrt{\det(2\pi\,\boldsymbol{\Sigma}_{uv})}\,\mathcal{N}([\underline{u};\underline{v}]^{\mathrm{T}}; \underline{m}, \boldsymbol{\Sigma}_{uv}) \,, \quad (10)$$

with $\mathrm{P}(k=0|\underline{u},\underline{v}) = 1 - \mathrm{P}(k=1|\underline{u},\underline{v})$ or a hard weighting function, e.g., rectangle functions, may be considered. Note that arbitrary weighting functions representing the confidence in the prior knowledge are permissible, e.g., arbitrary-shaped GM, as long as they adhere to the aforementioned conditions. The resulting model may be understood as a blend of a product probability kernel [16], as the location-based kernel combination corresponds to a causal dependency, and a multiple kernel approach [17]. In contrast to [17], the mixture is determined according to a hard-/soft-mapping function *a priori* for each data point and only component weights are determined in the optimization problem (3).

Restrictions & Properties: Some of the restrictions in the above description may be relaxed. The prior knowledge need not be restricted to parts of the state space. Even though only one PM encoding the prior knowledge was assumed above, more PMs are permissible, given the weighting function gives rise to a valid kernel. The quality of the prior knowledge depends on the number of components in the PM, impacting training time and the size of the conditional density estimate. In order to use the mixture kernel (9) with (3), the error and the roughness penalty as well as the constraints have to be recalculated. The respective roughness penalty has to be computed only once. Much of the calculation can be saved by rearranging and combining identical terms. More general P, e.g., GM, may be used. Many calculations may be saved for the soft transition model by neglecting the mixture kernel for small $\mathrm{P}(k=i|\underline{u},\underline{v})$. It is an important property that the conditional density estimates, obtained from solving the changed (3), will in the worst case contain no more than $|\mathcal{D}|+L$ components. This follows from

$$\begin{aligned} f(\underline{y}|\underline{x}) &= \sum_{i=1}^{|\mathcal{D}|} \alpha_i [\underbrace{\mathrm{P}(k=0|\underline{u}_i,\underline{v}_i)}_{p_{0,i}} \underbrace{\mathcal{K}_0([\underline{x};\underline{y}]^{\mathrm{T}}, [\underline{u}_i;\underline{v}_i]^{\mathrm{T}})}_{\mathcal{K}_0^i} \\ &\qquad + \underbrace{\mathrm{P}(k=1|\underline{u}_i,\underline{v}_i)}_{p_{1,i}} \underbrace{\mathcal{K}_1([\underline{x};\underline{y}]^{\mathrm{T}}, [\underline{u}_i';\underline{v}_i']^{\mathrm{T}})}_{\sum_{j=1}^{L} \beta_j \mathcal{K}_1^j}] \\ &= \sum_{i=1}^{|\mathcal{D}|} \alpha_i\, p_{0,i}\, \mathcal{K}_0^i + \sum_{j=1}^{L} \underbrace{\sum_{i=1}^{|\mathcal{D}|} \alpha_i\, p_{1,i}\, \beta_j}_{\beta_j'} \mathcal{K}_1^j \\ &= \sum_{i=1}^{|\mathcal{D}|} \alpha_i\, p_{0,i}\, \mathcal{K}_0^i + \sum_{j=1}^{L} \beta_j' \mathcal{K}_1^j \,, \end{aligned}$$

obtained by simple applications of the distributive law. This result is important for all applications sensitive to the size of the representation, e.g., when modeling dynamic systems. Because more components due to the insertion of the mixture kernel will yield lower error in the respective part of the state space, it needs to be asserted that the probability mass is not concentrated on the prior knowledge. This problem may be avoided by the introduction of additional normalization constraints, which force the probability mass to spread more evenly over the state space.

VI. Experiments

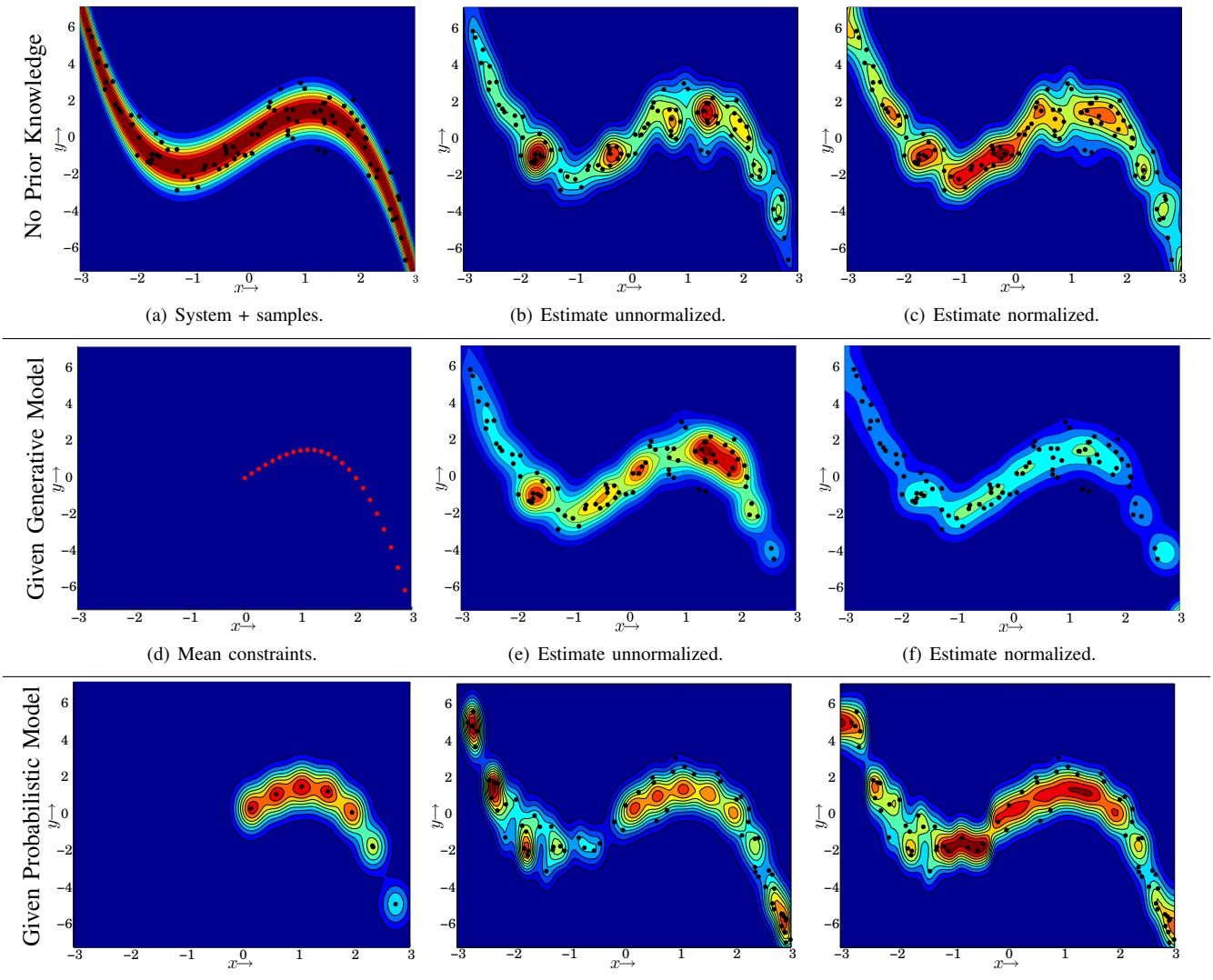
In this section, the achievable improvements of the conditional density estimates are evaluated. For the comparison, the SVM-based conditional density estimation approach [7] as well as the LCD-based approach [8] are employed with and without prior knowledge. Regarding the regularization term, all LCD-based approaches and all location-based kernels were used with the squared l_2-norm of $\underline{\alpha}$. For the other SVM experiments, the norm in the RKHS was used. In order to assess the quality of the approaches, the error in estimating the probabilistic model of the following cubic system, perturbed by additive white Gaussian noise is employed

$$\boldsymbol{y} = 2\,\boldsymbol{x} - 0.5\,\boldsymbol{x}^3 + \boldsymbol{w} \,, \;\; \boldsymbol{w} \sim \mathcal{N}(0, 0.9) \,. \quad (11)$$

For the experiments, 100 random samples are generated according to (11) in the range of $[-3,3]$. These are used for training and combined with the respective form of prior knowledge, i.e., knowledge about the generative or probabilistic model. The results are given in Fig. 2, Fig. 3, and Tab. I. For the experiments, the total variation

$$\nu = \tfrac{1}{2(x_{\max}-x_{\min})} \int_{\mathcal{X}} \int_{\mathcal{Y}} |\tilde{f}(y'|x') - f(y'|x')| \; \mathrm{d}y' \; \mathrm{d}x' \quad (12)$$

of the difference between the true conditional density function $\tilde{f}$ and the estimate f is calculated for $\mathcal{X} := [0,\, x_{\max}]$ and $\mathcal{Y} := [y_{\min},\, y_{\max}]$ numerically. The results in Tab. I are averages of ten experiments.

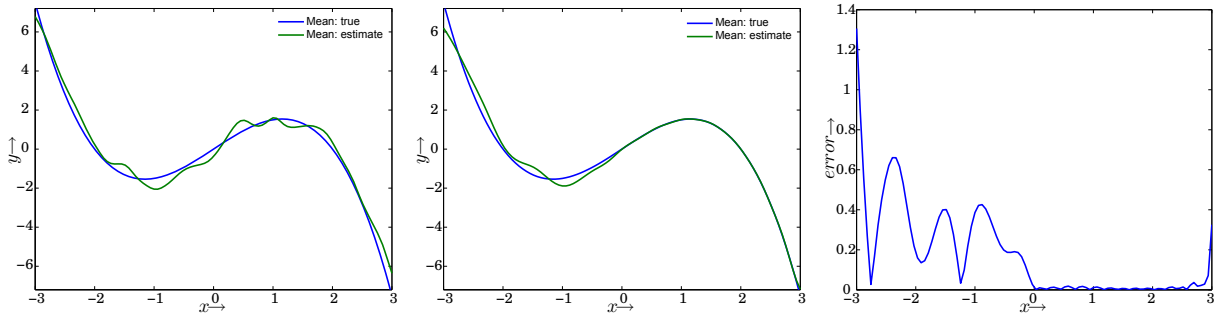


(a) System + samples. (b) Estimate unnormalized. (c) Estimate normalized.

(d) Mean constraints. (e) Estimate unnormalized. (f) Estimate normalized.

(g) $\mathcal{K}_1$ unnormalized. (h) Estimate unnormalized. (i) Estimate normalized.

Fig. 2. (a) True system with samples generated accordingly, (b-c) conditional density estimate (un)normalized without prior knowledge, (d) prior knowledge in the form of mean function values, (e-f) conditional density estimate (un)normalized, (g) prior knowledge in the form of a PM, and (h-i) conditional density estimate (un)normalized. These results were obtained by modification of the LCD-based optimization problem [8].



(a) Samples. (b) Samples + mean constraints. (c) l_1-error.

Fig. 3. Means of the true system (green) and expectations of the normalized conditional density estimate (blue) conditioned on fixed $\boldsymbol{x}$ (a) in case only samples are given, (b) if samples and the means constraints over $[0, 3]$ are given, and (c) the error in terms of the l_1-distance of the conditional expectations to the true mean function. The depicted results were obtained by modification of the LCD-based approach [8].

A. *Mean Function of the Generative Model*

The mean function of (11) is given and 25 constraints are obtained by equidistant sampling of $\underline{a}(\underline{x})$ in $\boldsymbol{x}$-direction over $[0, 3]$, cf. Fig. 2 (d). In Tab. I, the resulting total variation between the true conditional density function $\tilde{f}$ and the estimate f as calculated numerically is given. Additionally, the average l_1-error between the mean function and the conditional expectations as well as the number of components are reported. The difference between the true model and the

TABLE I

TOTAL VARIATION $\nu \pm \sigma$, l_1-ERROR $\pm\sigma$ OF THE MEAN, AND THE NUMBER OF COMPONENTS FOR THE NORMALIZED AND UNNORMALIZED CONDITIONAL DENSITIES OBTAINED BY MODIFICATION OF [8] USED WITHOUT AND WITH PRIOR KNOWLEDGE. THE ERRORS ARE CALCULATED FOR $x \in [0, 3]$, I.E., THE PART OF THE STATE SPACE WITH THE PRIOR KNOWLEDGE.

	Estimator	Normalized Results		Unnormalized Results		Components
		ν	$l_1(\mu)$	ν	$l_1(\mu)$	
No Prior	SVM	0.24 ± 0.03	0.30 ± 0.08	0.26 ± 0.03	0.50 ± 0.21	99.9
Knowledge	LCD	0.24 ± 0.03	0.37 ± 0.04	0.27 ± 0.04	0.63 ± 0.14	95.3
Mean	SVM-μ	0.21 ± 0.03	0.23 ± 0.08	0.23 ± 0.03	0.27 ± 0.19	71.2
Constraints	LCD-μ	0.22 ± 0.02	0.30 ± 0.09	0.25 ± 0.02	0.59 ± 0.16	100
Prob.	SVM-GM	0.11 ± 0.01	0.12 ± 0.02	0.13 ± 0.02	0.25 ± 0.08	88.5
Model	LCD-GM	0.12 ± 0.01	0.14 ± 0.03	0.13 ± 0.01	0.36 ± 0.13	57.9

estimate is decreased. The error regarding the mean values is shown in Tab. I and depicted for the normalized results in Fig. 3. In Fig. 2 (b-c) f is much smoother between $[0, 3]$ given the prior knowledge. Tab. I shows that the error is reduced for both approaches. For the SVM, this corresponds to a ca. 50% reduction of the error w.r.t. to the considered part of the state space in $\boldsymbol{y}$-direction.

B. Location-Based Mixture Kernel

For the location-based kernel, (11) was approximated in $[0, 3]$ by an axis-aligned GM, cf. Fig. 2 (g). For the approximation, the total variation of the difference between $\tilde{f}$ and the PM was minimized w.r.t. the free variances and mixture weights for user-defined means. In Tab. I, the difference between the true conditional density (11) and the estimate f is calculated numerically according to (12). The statistics show that the deviation of the densities is decreased for normalized and unnormalized f as is the error in the mean values. Fig. 2 (h-i) show the smoothness of f where one is confident of the prior knowledge. Special attention should be paid to the number of components: already the introduction of mean constraints allowed a reduction of components in the density. The introduction of a location-based mixture kernel allows a reduction of components by up to 40%.

VII. CONCLUSION

In this paper, sparse nonparametric conditional density estimation based on given samples *and* prior knowledge is addressed. The key idea is the incorporation of prior knowledge into conditional density estimation problems phrased as constrained optimization problems. For prior knowledge in the form of generative and probabilistic models, the incorporation by mean-function constraints and Gaussian mixture kernels was presented. This approach allows for an efficient incorporation of prior knowledge and is applicable to all algorithms that can be formulated as optimization problems. The estimates using prior knowledge are sparse, of high quality, and the achievable improvements were demonstrated for an SVM-based and an LCD-based conditional density estimation method. The type of the resulting conditional density functions is especially favorable in applications such as, e.g., Bayesian networks, as they allow for closed-form Bayesian inference. It remains future work to compensate for almost uniform conditional densities due to absence of data and to obtain even sparser representations for dynamic systems and recursive inference.

REFERENCES

[1] K. Murphy, "Dynamic Bayesian Network : Representation, Inference and Learning," Ph.D. dissertation, UC Berkeley, 2002.

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

[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] 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*, vol. 54, no. 7, pp. 1707 – 1718, 2010.

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

[7] 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, pp. 1–8.

[8] P. Krauthausen and U. 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, Sept. 2010, pp. 180–186.

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

[11] F. Lauer and G. Bloch, "Incorporating Prior Knowledge in Support Vector Machines for Classification: A Review," *Neurocomputing*, vol. 71, no. 7-9, pp. 1578–1594, 2008.

[12] R. Maclin, J. Shavlik, T. Walker, and L. Torrey, "A Simple and Effective Method for Incorporating Advice into Kernel Methods," in *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006)*, Boston, MA., July 2006.

[13] B. Schölkopf, P. Simard, A. Smola, and V. Vapnik, "Prior Knowledge in Support Vector Kernels," in *Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems 10*. MIT Press, 1998, pp. 640–646.

[14] Z. Sun, Z. Zhang, and H. Wang, "Incorporating Prior Knowledge into Kernel Based Regression," *Acta Automatica Sinica*, vol. 34, no. 12, pp. 1515 – 1521, 2008.

[15] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[16] T. Jebara, R. Kondor, and A. Howard, "Probability Product Kernels," *Journal of Machine Learning Research*, vol. 5, pp. 819–844, 2004.

[17] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple Kernel Learning, Conic Duality, and the SMO Algorithm," in *Proceedings of the Twenty-first International Machine Learning Conference (ICML 2004)*, Banff, Canada, July 2004.