IEICE Transactions on Information and Systems,
vol.E93-D, no.3, pp.583–594, 2010.

Least-Squares Conditional Density Estimation

Masashi Sugiyama (sugi@cs.titech.ac.jp)
Tokyo Institute of Technology
and
Japan Science and Technology Agency

Ichiro Takeuchi (takeuchi.ichiro@nitech.ac.jp)
Nagoya Institute of Technology

Taiji Suzuki (s-taiji@stat.t.u-tokyo.ac.jp)
The University of Tokyo

Takafumi Kanamori (kanamori@is.nagoya-u.ac.jp)
Nagoya University

Hirotaka Hachiya (hachiya@sg.cs.titech.ac.jp)
Tokyo Institute of Technology

Daisuke Okanohara (daisuke.okanohara@gmail.com)
The University of Tokyo

Abstract

Estimating the conditional mean of an input-output relation is the goal of regression. However, regression analysis is not sufficiently informative if the conditional distribution has multi-modality, is highly asymmetric, or contains heteroscedastic noise. In such scenarios, estimating the conditional distribution itself would be more useful. In this paper, we propose a novel method of conditional density estimation that is suitable for multi-dimensional continuous variables. The basic idea of the proposed method is to express the conditional density in terms of the density ratio and the ratio is directly estimated without going through density estimation. Experiments using benchmark and robot transition datasets illustrate the usefulness of the proposed approach.

Keywords

Conditional density estimation, multimodality, heteroscedastic noise, direct density ratio estimation, transition estimation

1 Introduction

Regression is aimed at estimating the conditional *mean* of output $\boldsymbol{y}$ given input $\boldsymbol{x}$. When the conditional density $p(\boldsymbol{y}|\boldsymbol{x})$ is unimodal and symmetric, regression would be sufficient for analyzing the input-output dependency. However, estimating the conditional mean may not be sufficiently informative, when the conditional distribution possesses multimodality (e.g., inverse kinematics learning of a robot [4]) or a highly skewed profile with heteroscedastic noise (e.g., biomedical data analysis [13]). In such cases, it would be more informative to estimate the conditional distribution itself. In this paper, we address the problem of estimating conditional densities when $\boldsymbol{x}$ and $\boldsymbol{y}$ are continuous and multi-dimensional.

When the conditioning variable $\boldsymbol{x}$ is discrete, estimating the conditional density $p(\boldsymbol{y}|\boldsymbol{x} = \widetilde{\boldsymbol{x}})$ from samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ is straightforward—by only using samples $\{\boldsymbol{y}_i\}_{i=1}^n$ such that $\boldsymbol{x}_i = \widetilde{\boldsymbol{x}}$, a standard density estimation method gives an estimate of the conditional density. However, when the conditioning variable $\boldsymbol{x}$ is continuous, conditional density estimation is not straightforward since no sample exactly matches the condition $\boldsymbol{x}_i = \widetilde{\boldsymbol{x}}$. A naive idea for coping with this problem is to use samples $\{\boldsymbol{y}_i\}_{i=1}^n$ that *approximately* satisfy the condition: $\boldsymbol{x}_i \approx \widetilde{\boldsymbol{x}}$. However, such a naive method is not reliable in high-dimensional problems. Slightly more sophisticated variants have been proposed based on weighted kernel density estimation [10, 47], but they still share the same weakness.

The mixture density network (MDN) [4] models the conditional density by a mixture of parametric densities, where the parameters are estimated by a neural network. MDN was shown to work well, although its training is time-consuming and only a local optimal solution may be obtained due to the non-convexity of neural network learning. Similarly, a mixture of Gaussian processes was explored for estimating the conditional density [42]. The mixture model is trained in a computationally efficient manner by an expectation-maximization algorithm [8]. However, since the optimization problem is non-convex, one may only access to a local optimal solution in practice.

The kernel quantile regression (KQR) method [40, 25] allows one to predict percentiles of the conditional distribution. This implies that solving KQR for all percentiles gives an estimate of the entire conditional cumulative distribution. KQR is formulated as a convex optimization problem, and therefore a unique global solution can be obtained. Furthermore, the entire solution path with respect to the percentile parameter, which was shown to be piece-wise linear, can be computed efficiently [41]. However, the range of applications of KQR is limited to one-dimensional output and solution path tracking tends to be numerically rather unstable in practice.

In this paper, we propose a new method of conditional density estimation named *least-squares conditional density estimation* (LS-CDE), which can be applied to multi-dimensional inputs and outputs. The proposed method is based on the fact that the conditional density can be expressed in terms of unconditional densities as $p(\boldsymbol{y}|\boldsymbol{x}) = p(\boldsymbol{x}, \boldsymbol{y})/p(\boldsymbol{x})$. Our key idea is that we do not estimate the two densities $p(\boldsymbol{x}, \boldsymbol{y})$ and $p(\boldsymbol{x})$ separately, but we *directly* estimate the density ratio $p(\boldsymbol{x}, \boldsymbol{y})/p(\boldsymbol{x})$ without going through

density estimation. Experiments using benchmark and robot transition datasets show that our method compares favorably with existing methods in terms of the accuracy and computational efficiency.

The rest of this paper is organized as follows. In Section 2, we present our proposed method LS-CDE and investigate its theoretical properties. In Section 3, we discuss the characteristics of existing and proposed approaches. In Section 4, we compare the experimental performance of the proposed and existing methods. Finally, in Section 5, we conclude by summarizing our contributions and outlook.

2 A New Method of Conditional Density Estimation

In this section, we formulate the problem of conditional density estimation and give a new method.

2.1 Conditional Density Estimation via Density Ratio Estimation

Let $\mathcal{D}_{\mathrm{X}}$ ($\subset \mathbb{R}^{d_{\mathrm{X}}}$) and $\mathcal{D}_{\mathrm{Y}}$ ($\subset \mathbb{R}^{d_{\mathrm{Y}}}$) be input and output data domains, where d_{X} and d_{Y} are the dimensionality of the data domains, respectively. Let us consider a joint probability distribution on $\mathcal{D}_{\mathrm{X}} \times \mathcal{D}_{\mathrm{Y}}$ with probability density function $p(\boldsymbol{x}, \boldsymbol{y})$, and suppose that we are given n independent and identically distributed (i.i.d.) paired samples of input $\boldsymbol{x}$ and output $\boldsymbol{y}$:

$$\{\boldsymbol{z}_i \mid \boldsymbol{z}_i = (\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{D}_{\mathrm{X}} \times \mathcal{D}_{\mathrm{Y}}\}_{i=1}^{n}.$$

The goal is to estimate the conditional density $p(\boldsymbol{y}|\boldsymbol{x})$ from the samples $\{\boldsymbol{z}_i\}_{i=1}^{n}$.

Our primal interest is in the case where both variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are continuous. In this case, conditional density estimation is not straightforward since no sample exactly matches the condition.

Our proposed approach is to consider the ratio of two densities:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})} := r(\boldsymbol{x}, \boldsymbol{y}),$$

where we assume $p(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{D}_{\mathrm{X}}$. However, naively estimating two densities and taking their ratio can result in large estimation error. In order to avoid this, we propose to estimate the *density ratio function* $r(\boldsymbol{x}, \boldsymbol{y})$ directly without going through density estimation of $p(\boldsymbol{x}, \boldsymbol{y})$ and $p(\boldsymbol{x})$.

2.2 Linear Density-ratio Model

We model the density ratio function $r(\boldsymbol{x}, \boldsymbol{y})$ by the following linear model:

$$\widehat{r}_{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{\alpha}^{\top} \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}), \tag{1}$$

where $^\top$ denotes the transpose of a matrix or a vector,

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_b)^\top$$

are parameters to be learned from samples, and

$$\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) = (\phi_1(\boldsymbol{x}, \boldsymbol{y}), \phi_2(\boldsymbol{x}, \boldsymbol{y}), \ldots, \phi_b(\boldsymbol{x}, \boldsymbol{y}))^\top$$

are basis functions such that

$$\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) \geq \mathbf{0}_b \quad \text{for all } (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}_{\mathrm{X}} \times \mathcal{D}_{\mathrm{Y}}.$$

$\mathbf{0}_b$ denotes the b-dimensional vector with all zeros. The inequality for vectors is applied in an element-wise manner.

Note that the number b of basis functions is not necessarily a constant; it can depend on the number n of samples. Similarly, the basis functions $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$ could be dependent on the samples $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$. This means that *kernel* models (i.e., $b = n$ and $\phi_i(\boldsymbol{x}, \boldsymbol{y})$ is a kernel function 'centered' at $(\boldsymbol{x}_i, \boldsymbol{y}_i)$) are also included in the above formulation. We explain how the basis functions $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$ are practically chosen in Section 2.6.

2.3 A Least-squares Approach to Conditional Density Estimation

We determine the parameter $\boldsymbol{\alpha}$ in the model $\widehat{r}_{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{y})$ so that the following squared error J_0 is minimized:

$$J_0(\boldsymbol{\alpha}) := \frac{1}{2} \iint \left(\widehat{r}_{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{y}) - r(\boldsymbol{x}, \boldsymbol{y})\right)^2 p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y}.$$

This can be expressed as

$$\begin{aligned} J_0(\boldsymbol{\alpha}) &= \frac{1}{2} \iint \widehat{r}_{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{y})^2 p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} - \iint \widehat{r}_{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{y}) r(\boldsymbol{x}, \boldsymbol{y}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} + C \\ &= \frac{1}{2} \iint \left(\boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})\right)^2 p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} - \iint \boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) p(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} + C, \end{aligned} \tag{2}$$

where

$$C := \frac{1}{2} \iint r(\boldsymbol{x}, \boldsymbol{y}) p(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y}$$

is a constant and therefore can be safely ignored. Let us denote the first two terms of Eq.(2) by J:

$$\begin{aligned} J(\boldsymbol{\alpha}) &:= J_0(\boldsymbol{\alpha}) - C \\ &= \frac{1}{2} \boldsymbol{\alpha}^\top \boldsymbol{H} \boldsymbol{\alpha} - \boldsymbol{h}^\top \boldsymbol{\alpha}, \end{aligned}$$

where

$$\begin{aligned}\boldsymbol{H} &:= \int \overline{\boldsymbol{\Phi}}(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \\ \boldsymbol{h} &:= \iint \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) p(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y}, \\ \overline{\boldsymbol{\Phi}}(\boldsymbol{x}) &:= \int \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})^\top \mathrm{d}\boldsymbol{y}. \end{aligned} \tag{3}$$

$\boldsymbol{H}$ and $\boldsymbol{h}$ included in $J(\boldsymbol{\alpha})$ contain the expectations over unknown densities $p(\boldsymbol{x})$ and $p(\boldsymbol{x}, \boldsymbol{y})$, so we approximate the expectations by sample averages. Then we have

$$\widehat{J}(\boldsymbol{\alpha}) := \frac{1}{2} \boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}} \boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha},$$

where

$$\begin{aligned}\widehat{\boldsymbol{H}} &:= \frac{1}{n} \sum_{i=1}^{n} \overline{\boldsymbol{\Phi}}(\boldsymbol{x}_i), \\ \widehat{\boldsymbol{h}} &:= \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\phi}(\boldsymbol{x}_i, \boldsymbol{y}_i). \end{aligned} \tag{4}$$

Note that the integral over $\boldsymbol{y}$ included in $\overline{\boldsymbol{\Phi}}(\boldsymbol{x})$ (see Eq.(3)) can be computed in principle since it does not contain any unknown quantity. As shown in Section 2.6, this integration can be computed analytically in our basis function choice.

Now our optimization criterion is summarized as

$$\widetilde{\boldsymbol{\alpha}} := \operatorname*{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^b} \left[\widehat{J}(\boldsymbol{\alpha}) + \frac{\lambda}{2} \boldsymbol{\alpha}^\top \boldsymbol{\alpha} \right], \tag{5}$$

where a regularizer $\lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha} / 2$ ($\lambda > 0$) is included for stabilization purposes[1]. Taking the derivative of the above objective function and equating it to zero, we can see that the solution $\widetilde{\boldsymbol{\alpha}}$ can be obtained just by solving the following system of linear equations.

$$(\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_b) \boldsymbol{\alpha} = \widehat{\boldsymbol{h}},$$

where $\boldsymbol{I}_b$ denotes the b-dimensional identity matrix. Thus, the solution $\widetilde{\boldsymbol{\alpha}}$ is given analytically as

$$\widetilde{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_b)^{-1} \widehat{\boldsymbol{h}}. \tag{6}$$

Since the density ratio function is non-negative by definition, we modify the solution $\widetilde{\boldsymbol{\alpha}}$ as[2]

$$\widehat{\boldsymbol{\alpha}} := \max(\boldsymbol{0}_b, \widetilde{\boldsymbol{\alpha}}), \tag{7}$$

[1] We may also use $\lambda \boldsymbol{\alpha}^\top \boldsymbol{R} \boldsymbol{\alpha}$ as a regularizer for an arbitrary positive symmetric matrix $\boldsymbol{R}$ without sacrificing the computational advantage.

[2] A variant of the proposed method would be to include the positivity constraint $\boldsymbol{\alpha} \geq \boldsymbol{0}_n$ directly in Eq.(6). Our preliminary experiments showed that the estimation accuracy of this modified algorithm turned out to be comparable to Eq.(7), while the constrained version was computationally less efficient than Eq.(7) since we need to use a numerical quadratic program solver for computing the solution. For this reason, we only consider Eq.(7) in the rest of this paper.

where the 'max' operation for vectors is applied in an element-wise manner. Thanks to this rounding-up processing, the solution $\widehat{\boldsymbol{\alpha}}$ tends to be sparse, which contributes to reducing the computation time in the test phase.

In order to assure that the obtained density-ratio function is a conditional density, we renormalize the solution in the test phase—given a test input point $\widetilde{\boldsymbol{x}}$, our final solution is given as

$$\widehat{p}(\boldsymbol{y}|\boldsymbol{x}=\widetilde{\boldsymbol{x}}) = \frac{\widehat{\boldsymbol{\alpha}}^\top \boldsymbol{\phi}(\widetilde{\boldsymbol{x}}, \boldsymbol{y})}{\int \widehat{\boldsymbol{\alpha}}^\top \boldsymbol{\phi}(\widetilde{\boldsymbol{x}}, \boldsymbol{y}')\mathrm{d}\boldsymbol{y}'}. \tag{8}$$

We call the above method *Least-Squares Conditional Density Estimation (LS-CDE)*. LS-CDE can be regarded as an application of the direct density ratio estimation method called the *unconstrained Least-Squares Importance Fitting (uLSIF)* [17, 18] to the problem of density ratio estimation.

A MATLAB® implementation of the LS-CDE algorithm is available from

`http://sugiyama-www.cs.titech.ac.jp/~sugi/software/LSCDE/`

2.4 Convergence Analysis

Here, we show a non-parametric convergence rate of the LS-CDE solution. Those who are interested in practical issues of the proposed method may skip this subsection.

Let $\mathcal{G}$ be a general set of functions on $\mathcal{D}_{\mathrm{X}} \times \mathcal{D}_{\mathrm{Y}}$. Note that $\mathcal{G}$ corresponds to the span of our model, which could be non-parametric (i.e., an infinite dimensional linear space[3]). For a function g ($\in \mathcal{G}$), let us consider a non-negative function $R(g)$ such that

$$\max\left\{\sup_{\boldsymbol{x}}\left[\int g(\boldsymbol{x}, \boldsymbol{y})\mathrm{d}\boldsymbol{y}\right], \sup_{\boldsymbol{x},\boldsymbol{y}}[g(\boldsymbol{x}, \boldsymbol{y})]\right\} \leq R(g).$$

Then the problem (5) can be generalized as

$$\widehat{r} := \operatorname*{argmin}_{g\in\mathcal{G}}\left[\frac{1}{2n}\sum_{i=1}^{n}\int g(\boldsymbol{x}_i, \boldsymbol{y})^2\mathrm{d}\boldsymbol{y} - \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{x}_i, \boldsymbol{y}_i) + \lambda_n R(g)^2\right],$$

where λ_n is the regularization parameter depending on n. We assume that the true density ratio function $r(\boldsymbol{x}, \boldsymbol{y})$ is contained in $\mathcal{G}$ and there exists M (> 0) such that $R(r) < M$. We also assume that there exists γ ($0 < \gamma < 2$) such that

$$\mathcal{H}_{[]}(\mathcal{G}_M, \epsilon, L_2(p_{\mathrm{x}} \times \mu_{\mathrm{Y}})) = \mathcal{O}\left(\left(\frac{M}{\epsilon}\right)^{\gamma}\right),$$

where

$$\mathcal{G}_M := \{g \in \mathcal{G} \mid R(g) \leq M\}.$$

[3]If a reproducing kernel Hilbert space is chosen as $\mathcal{G}$ and the regularization term $R(g)$ is chosen appropriately, the optimization problem in the infinite dimensional space is reduced to a finite dimensional one. Then the optimal approximation can be found in the form of $\widehat{r}_{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{y})$ when kernel functions centered at the training samples are used as the basis functions [20].

μ_{Y} is the Lebesgue measure on $\mathcal{D}_{\mathrm{Y}}$, $p_{\mathrm{x}} \times \mu_{\mathrm{Y}}$ is a product measure of p_{x} and μ_{Y}, and $\mathcal{H}_{[]}$ is the *bracketing entropy* of $\mathcal{G}_M$ with respect to the $L_2(p_{\mathrm{x}} \times \mu_{\mathrm{Y}})$-norm [45].

Intuitively, the bracketing entropy $\mathcal{H}_{[]}(\mathcal{G}_M, \epsilon, L_2)$ expresses the complexity of the model $\mathcal{G}_M$, and ϵ is a precision measure of the model complexity. The larger the bracketing entropy $\mathcal{H}_{[]}(\mathcal{G}_M, \epsilon, L_2)$ is for a certain precision ϵ, the more complex the model is for that precision level. As the precision is increased (i.e., $\epsilon \to 0$), the bracketing entropy measured with precision ϵ typically diverges to infinity. The "dimension" of the model is reflected in the divergence rate of the bracketing entropy when $\epsilon \to 0$. See the book [45] for details.

When the set $\mathcal{G}_M$ is the closed ball of radius M centered at the origin of a Sobolev space, γ is given by $(d_{\mathrm{X}} + d_{\mathrm{Y}})/p$, where p is the order of differentiability of the Sobolev space (see page 105 of the book [9] for details). Hence, γ is small for a set of smooth functions with few variables. The reproducing kernel Hilbert spaces with Gaussian kernel

$$\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\sigma^2}\right)\exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{y}'\|^2}{2\sigma^2}\right),$$

which we will use in our practical implementation (see Section 2.6) satisfy the above entropy condition for any small $\gamma > 0$ [49]. On the other hand, in the above setup, the bracketing entropy is lower-bounded by $K(M/\epsilon)^{(d_{\mathrm{X}}+d_{\mathrm{Y}})/p}$ with a constant K depending only on p, d_{X}, and d_{Y} [21]. Therefore, if the dimension of the domains $\mathcal{D}_{\mathrm{X}}$ and $\mathcal{D}_{\mathrm{Y}}$ is so large that $(d_{\mathrm{X}} + d_{\mathrm{Y}})/p > 2$, γ should be larger than 2. This means that a situation where p is small and d_{X} and d_{Y} are large is not covered in our analysis; such a model is too complex to deal with in our framework. Fortunately, it is known that the Gaussian kernel satisfies $\gamma \in (0, 2)$. Hence, the Gaussian kernel as well as Sobolev spaces with large p and small d_{X} and d_{Y} is included in our analysis.

Under the above assumptions, we have the following theorem (its proof is omitted since it follows essentially the same line as the references [28, 38]).

Theorem 1 *Under the above setting, if $\lambda_n \to 0$ and $\lambda_n^{-1} = o(n^{2/(2+\gamma)})$, then*

$$\|\widehat{r} - r\|_2 = \mathcal{O}_p(\lambda_n^{1/2}),$$

where $\|\cdot\|_2$ denotes the $L_2(p_{\mathrm{x}} \times \mu_{\mathrm{Y}})$-norm and $\mathcal{O}_p$ denotes the asymptotic order in probability.

Note that the conditions $\lambda_n \to 0$ and $\lambda_n^{-1} = o(n^{2/(2+\gamma)})$ intuitively means that λ_n should converge to zero as n tends to infinity but the speed of convergence should not be too fast.

2.5 Cross-validation for Model Selection

We elucidated the convergence rate of the LS-CDE solution. However, its practical performance still depends on the choice of model parameters such as the basis functions $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$ and the regularization parameter λ.

Here we show that cross-validation (CV) is available for model selection. CV should be carried out in terms of the error metric used for evaluating the test performance. Below, we investigate two cases: the *squared (SQ) error* and the *Kullback-Leibler (KL) error*. The SQ error for a conditional density estimator $\widehat{p}(\boldsymbol{y}|\boldsymbol{x})$ is defined as

$$\begin{aligned}\mathrm{SQ}_0 &:= \frac{1}{2}\iint \left(\widehat{p}(\boldsymbol{y}|\boldsymbol{x}) - p(\boldsymbol{y}|\boldsymbol{x})\right)^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \\ &= \mathrm{SQ} + C_{\mathrm{SQ}},\end{aligned}$$

where

$$\mathrm{SQ} := \frac{1}{2}\iint \left(\widehat{p}(\boldsymbol{y}|\boldsymbol{x})\right)^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} - \iint \widehat{p}(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y},$$

and C_{SQ} is the constant defined by

$$C_{\mathrm{SQ}} := \frac{1}{2}\iint p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}.$$

The KL error for a conditional density estimator $\widehat{p}(\boldsymbol{y}|\boldsymbol{x})$ is defined as

$$\begin{aligned}\mathrm{KL}_0 &:= \iint p(\boldsymbol{x},\boldsymbol{y}) \log \frac{p(\boldsymbol{x},\boldsymbol{y})}{\widehat{p}(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \\ &= \mathrm{KL} + C_{\mathrm{KL}},\end{aligned}$$

where

$$\mathrm{KL} := -\iint p(\boldsymbol{x},\boldsymbol{y}) \log \widehat{p}(\boldsymbol{y}|\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y},$$

and C_{KL} is the constant defined by

$$C_{\mathrm{KL}} := \iint p(\boldsymbol{x},\boldsymbol{y}) \log p(\boldsymbol{y}|\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}.$$

The smaller the value of SQ or KL is, the better the performance of the conditional density estimator $\widehat{p}(\boldsymbol{y}|\boldsymbol{x})$ is.

For the above performance measures, CV is carried out as follows. First, the samples

$$\mathcal{Z} := \{\boldsymbol{z}_i \mid \boldsymbol{z}_i = (\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$$

are divided into K disjoint subsets $\{\mathcal{Z}_k\}_{k=1}^K$ of approximately the same size. Let $\widehat{p}_{\mathcal{Z}\backslash\mathcal{Z}_k}$ be the conditional density estimator obtained using $\mathcal{Z}\backslash\mathcal{Z}_k$ (i.e., the estimator obtained without $\mathcal{Z}_k$). Then the target error values are approximated using the hold-out samples $\mathcal{Z}_k$ as

$$\begin{aligned}\widehat{\mathrm{SQ}}_{\mathcal{Z}_k} &:= \frac{1}{2|\mathcal{Z}_k|}\sum_{\widetilde{\boldsymbol{x}}\in\mathcal{Z}_k}\int \left(\widehat{p}_{\mathcal{Z}\backslash\mathcal{Z}_k}(\boldsymbol{y}|\widetilde{\boldsymbol{x}})\right)^2 \mathrm{d}\boldsymbol{y} - \frac{1}{|\mathcal{Z}_k|}\sum_{(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}})\in\mathcal{Z}_k}\widehat{p}_{\mathcal{Z}\backslash\mathcal{Z}_k}(\widetilde{\boldsymbol{y}}|\widetilde{\boldsymbol{x}}), \\ \widehat{\mathrm{KL}}_{\mathcal{Z}_k} &:= -\frac{1}{|\mathcal{Z}_k|}\sum_{(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}})\in\mathcal{Z}_k}\log \widehat{p}_{\mathcal{Z}\backslash\mathcal{Z}_k}(\widetilde{\boldsymbol{y}}|\widetilde{\boldsymbol{x}}),\end{aligned}$$

where $|\mathcal{Z}_k|$ denotes the number of elements in the set $\mathcal{Z}_k$. This procedure is repeated for $k = 1, 2, \ldots, K$ and its average is computed:

$$\widehat{\mathrm{SQ}} := \frac{1}{K}\sum_{k=1}^{K} \widehat{\mathrm{SQ}}_{\mathcal{Z}_k},$$

$$\widehat{\mathrm{KL}} := \frac{1}{K}\sum_{k=1}^{K} \widehat{\mathrm{KL}}_{\mathcal{Z}_k}.$$

We can show that $\widehat{\mathrm{SQ}}$ and $\widehat{\mathrm{KL}}$ are almost unbiased estimators of the true costs SQ and KL, respectively; the 'almost'-ness comes from the fact that the number of samples is reduced in the CV procedure due to data splitting [26, 33].

2.6 Basis Function Design

A good model may be chosen by CV, given that a family of promising model candidates is prepared. As model candidates, we propose to use a Gaussian kernel model: for $\boldsymbol{z} = (\boldsymbol{x}^\top, \boldsymbol{y}^\top)^\top$,

$$\begin{aligned}\phi_\ell(\boldsymbol{x}, \boldsymbol{y}) &= \exp\left(-\frac{\|\boldsymbol{z} - \boldsymbol{w}_\ell\|^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{u}_\ell\|^2}{2\sigma^2}\right)\exp\left(-\frac{\|\boldsymbol{y} - \boldsymbol{v}_\ell\|^2}{2\sigma^2}\right), \end{aligned} \tag{9}$$

where

$$\{\boldsymbol{w}_\ell \mid \boldsymbol{w}_\ell = (\boldsymbol{u}_\ell^\top, \boldsymbol{v}_\ell^\top)^\top\}_{\ell=1}^{b}$$

are center points randomly chosen from

$$\{\boldsymbol{z}_i \mid \boldsymbol{z}_i = (\boldsymbol{x}_i^\top, \boldsymbol{y}_i^\top)^\top\}_{i=1}^{n}.$$

We may use different Gaussian widths for $\boldsymbol{x}$ and $\boldsymbol{y}$. However, for simplicity, we decided to use the common Gaussian width σ for both $\boldsymbol{x}$ and $\boldsymbol{y}$ under the setting where the variance of each element of $\boldsymbol{x}$ and $\boldsymbol{y}$ is normalized to one.

An advantage of the above Gaussian kernel model is that the integrals over $\boldsymbol{y}$ in matrix $\overline{\boldsymbol{\Phi}}$ (see Eq.(3)) and in the normalization factor (see Eq.(8)) can be computed analytically; indeed, a simple calculation yields

$$\begin{aligned}\overline{\Phi}_{\ell,\ell'}(\boldsymbol{x}) &= \int \phi_\ell(\boldsymbol{x}, \boldsymbol{y})\phi_{\ell'}(\boldsymbol{x}, \boldsymbol{y})\mathrm{d}\boldsymbol{y} \\ &= (\sqrt{\pi}\sigma)^{d_{\mathrm{Y}}} \exp\left(-\frac{\xi_{\ell,\ell'}(\boldsymbol{x})}{4\sigma^2}\right), \\ \int \widehat{\boldsymbol{\alpha}}^\top \boldsymbol{\phi}(\widetilde{\boldsymbol{x}}, \boldsymbol{y})\mathrm{d}\boldsymbol{y} &= (\sqrt{2\pi}\sigma)^{d_{\mathrm{Y}}} \sum_{\ell=1}^{b} \widehat{\alpha}_\ell \exp\left(-\frac{\|\widetilde{\boldsymbol{x}} - \boldsymbol{u}_\ell\|^2}{2\sigma^2}\right), \end{aligned}$$

where

$$\xi_{\ell,\ell'}(\boldsymbol{x}) := 2\|\boldsymbol{x}-\boldsymbol{u}_\ell\|^2 + 2\|\boldsymbol{x}-\boldsymbol{u}_{\ell'}\|^2 + \|\boldsymbol{v}_\ell - \boldsymbol{v}_{\ell'}\|^2.$$

In the experiments, we fix the number of basis functions to

$$b = \min(100, n),$$

and choose the Gaussian width σ and the regularization parameter λ by CV from

$$\sigma, \lambda \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10\}.$$

2.7 Extention to Semi-supervised Scenarios

Another potential advantage of LS-CDE lies in the semi-supervised learning setting [5]—in addition to the labeled samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$, unlabeled samples $\{\boldsymbol{x}'_i\}_{i=n+1}^{n+n'}$ which are drawn independently from the marginal density $p(\boldsymbol{x})$ are available.

In conditional density estimation, unlabeled samples $\{\boldsymbol{x}'_i\}_{i=n+1}^{n+n'}$ are not generally useful since they are irrelevant to the conditional density $p(\boldsymbol{y}|\boldsymbol{x})$. However, in LS-CDE, unlabeled samples could be used for improving the estimation accuracy of the matrix $\boldsymbol{H}$. More specifically, instead of Eq.(4), the following estimator may be used:

$$\widehat{\boldsymbol{H}} = \frac{1}{n+n'} \sum_{i=1}^{n+n'} \overline{\boldsymbol{\Phi}}(\boldsymbol{x}_i).$$

3 Discussions

In this section, we discuss the characteristics of existing and proposed methods of conditional density estimation.

3.1 ϵ-neighbor Kernel Density Estimation (ϵ-KDE)

For estimating the conditional density $p(\boldsymbol{y}|\boldsymbol{x})$, ϵ-neighbor kernel density estimation (ϵ-KDE) employs the standard kernel density estimator using a subset of samples, $\{\boldsymbol{y}_i\}_{i\in\mathcal{I}_{\boldsymbol{x},\epsilon}}$ for some threshold ϵ (≥ 0), where $\mathcal{I}_{\boldsymbol{x},\epsilon}$ is the set of sample indices such that

$$\|\boldsymbol{x}_i - \boldsymbol{x}\| \leq \epsilon.$$

In the case of Gaussian kernels, ϵ-KDE is expressed as

$$\widehat{p}(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{|\mathcal{I}_{\boldsymbol{x},\epsilon}|} \sum_{i\in\mathcal{I}_{\boldsymbol{x},\epsilon}} N(\boldsymbol{y}; \boldsymbol{y}_i, \sigma^2 \boldsymbol{I}_{d_{\mathrm{Y}}}),$$

where $N(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The threshold ϵ and the bandwidth σ may be chosen based on CV [12]. ϵ-KDE is simple and easy to use, but it may not be reliable in high-dimensional problems. Slightly more sophisticated variants have been proposed based on weighted kernel density estimation [10, 47], but they may still share the same weakness.

3.2 Mixture Density Network (MDN)

The mixture density network (MDN) models the conditional density by a mixture of parametric densities [4]. In the case of Gaussian densities, MDN is expressed as

$$\widehat{p}(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\ell=1}^{t} \pi_\ell(\boldsymbol{x}) N(\boldsymbol{y}; \boldsymbol{\mu}_\ell(\boldsymbol{x}), \sigma_\ell^2(\boldsymbol{x}) \boldsymbol{I}_{d_{\mathrm{Y}}}),$$

where $\pi_\ell(\boldsymbol{x})$ denotes the mixing coefficient such that

$$\sum_{\ell=1}^{t} \pi_\ell(\boldsymbol{x}) = 1 \quad \text{and} \quad 0 \leq \pi_\ell(\boldsymbol{x}) \leq 1 \quad \text{for all } \boldsymbol{x} \in \mathcal{D}_{\mathrm{X}}.$$

All the parameters $\{\pi_\ell(\boldsymbol{x}), \boldsymbol{\mu}_\ell(\boldsymbol{x}), \sigma_\ell^2(\boldsymbol{x})\}_{\ell=1}^t$ are learned as a function of $\boldsymbol{x}$ by a neural network with regularized maximum likelihood estimation. The number t of Gaussian components, the number of hidden units in the neural network, and the regularization parameter may be chosen based on CV. MDN has been shown to work well, although its training is time-consuming and only a local solution may be obtained due to the non-convexity of neural network learning.

3.3 Kernel Quantile Regression (KQR)

Kernel quantile regression (KQR) allows one to predict the 100τ-percentile of conditional distributions for a given τ ($\in (0,1)$) when y is one-dimensional [40, 25]. For the Gaussian kernel model

$$\widehat{f}_\tau(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_{i,\tau} \phi_i(\boldsymbol{x}) + b_\tau,$$

where

$$\phi_i(\boldsymbol{x}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{2\sigma^2}\right),$$

the parameters $\{\alpha_{i,\tau}\}_{i=1}^n$ and b_τ are learned by

$$\min_{\{\alpha_{i,\tau}\}_{i=1}^n, b_\tau} \left[\sum_{i=1}^{n} \psi_\tau(y_i - \widehat{f}_\tau(\boldsymbol{x}_i)) + \lambda \sum_{i,j=1}^{n} \phi_i(\boldsymbol{x}_j) \alpha_{i,\tau} \alpha_{j,\tau} \right],$$

where $\psi_\tau(r)$ denotes the pin-ball loss function defined by

$$\psi_\tau(r) = \begin{cases} (1-\tau)|r| & (r \leq 0), \\ \tau |r| & (r > 0). \end{cases}$$

Thus, solving KQR for all $\tau \in (0,1)$ gives an estimate of the entire conditional distribution. The bandwidth σ and the regularization parameter λ may be chosen based on CV.

A notable advantage of KQR is that the solution of KQR is piece-wise linear with respect to τ, so the entire solution path can be computed efficiently [41]. This implies that the conditional cumulative distribution can be computed efficiently. However, solution path tracking tends to be numerically rather unstable and the range of applications of KQR is limited to one-dimensional output y. Furthermore, some heuristic procedure is needed to convert conditional cumulative distributions into conditional densities, which can cause additional estimation errors.

3.4 Other Methods of Density Ratio Estimation

A naive method for estimating the density ratio $p(\boldsymbol{x}, \boldsymbol{y})/p(\boldsymbol{x})$ is to first approximate the two densities $p(\boldsymbol{x}, \boldsymbol{y})$ and $p(\boldsymbol{x})$ by standard kernel density estimation and then taking the ratio of the estimated densities. We refer to this method as the ratio of kernel density estimators (RKDE). As we will show through experiments in the next section, RKDE does not work well since taking the ratio of estimated quantities significantly magnifies the estimation error.

To overcome the above weakness, we decided to directly estimate the density ratio without going through density estimation under the squared-loss (see Section 2.3). The *kernel mean matching* method [14] and the *logistic regression* based method [30, 6, 3] also allow one to directly estimate a density ratio $q(\boldsymbol{x})/q'(\boldsymbol{x})$. However, the derivation of these methods heavily relies on the fact that the two density functions $q(\boldsymbol{x})$ and $q'(\boldsymbol{x})$ share the same domain, which is not fulfilled in the current setting. For this reason, these methods may not be employed for conditional density estimation.

Other methods of direct density ratio estimation [37, 38, 28, 29, 43, 44, 48] employs the *Kullback-Leibler divergence* [22] as the loss function, instead of the squared-loss. It is possible to use these methods for conditional density estimation in the same way as the proposed method, but it is computationally rather inefficient [17, 18]. Furthermore, in the context of density estimation, the squared-loss is often preferred to the Kullback-Leibler loss [2, 34].

4 Numerical Experiments

In this section, we investigate the experimental performance of the proposed and existing methods.

4.1 Illustrative Examples

Here we illustrate how the proposed LS-CDE method behaves using toy datasets.

Let $d_{\mathrm{X}} = d_{\mathrm{Y}} = 1$. Inputs $\{x_i\}_{i=1}^n$ were independently drawn from $U(-1, 1)$, where $U(a, b)$ denotes the uniform distribution on (a, b). Outputs $\{y_i\}_{i=1}^n$ were generated by the following heteroscedastic noise model:

$$y_i = \mathrm{sinc}(2\pi x_i) + \frac{1}{8}\exp(1 - x_i) \cdot \varepsilon_i.$$

We tested the following three different distributions for $\{\varepsilon_i\}_{i=1}^n$:

(a) Gaussian: $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1)$,

(b) Bimodal: $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \frac{1}{2}N(-1,\frac{4}{9}) + \frac{1}{2}N(1,\frac{4}{9})$,

(c) Skewed: $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \frac{3}{4}N(0,1) + \frac{1}{4}N(\frac{3}{2},\frac{1}{9})$,

where '$\overset{\text{i.i.d.}}{\sim}$' denotes 'independent and identically distributed' and $N(\mu,\sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2. See Figure 1(a)–Figure 1(c) for their profiles. The number of training samples was set to $n = 200$. The numerical results were depicted in Figure 1(a)–Figure 1(c), illustrating that LS-CDE well captures heteroscedasticity, bimodality, and asymmetricity.

We have also investigated the experimental performance of LS-CDE using the following real datasets:

(d) Bone Mineral Density dataset: Relative spinal bone mineral density measurements on 485 North American adolescents [13], having a heteroscedastic asymmetric conditional distribution.

(e) Old Faithful Geyser dataset: The durations of 299 eruptions of the Old Faithful Geyser [46], having a bimodal conditional distribution.

Figure 1(d) and Figure 1(e) depict the results, showing that heteroscedastic and multimodal structures were nicely revealed by LS-CDE.

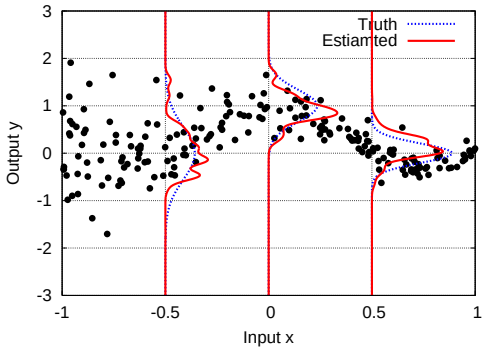
4.2 Benchmark Datasets

We applied the proposed and existing methods to the benchmark datasets accompanied with the *R* package [31] (see Table 1) and evaluate their experimental performance.

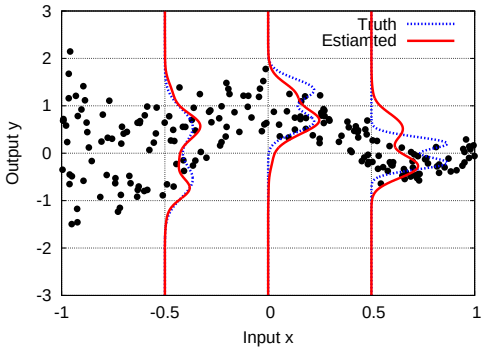
In each dataset, 50% of samples were randomly chosen for conditional density estimation and the rest was used for computing the estimation accuracy. The accuracy of a conditional density estimator $\widehat{p}(\boldsymbol{y}|\boldsymbol{x})$ was measured by the negative log-likelihood for test samples $\{\widetilde{\boldsymbol{z}}_i \mid \widetilde{\boldsymbol{z}}_i = (\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{y}}_i)\}_{i=1}^{\widetilde{n}}$:

$$\text{NLL} := -\frac{1}{\widetilde{n}} \sum_{i=1}^{\widetilde{n}} \log \widehat{p}(\widetilde{\boldsymbol{y}}_i|\widetilde{\boldsymbol{x}}_i). \tag{10}$$

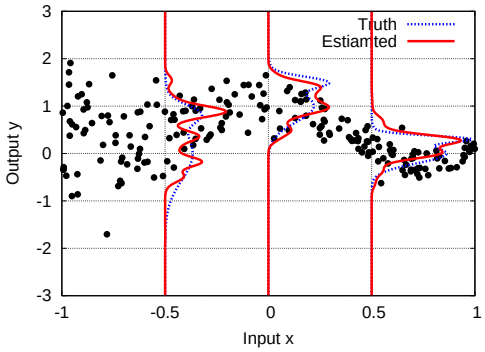
Thus, the smaller the value of NLL is, the better the performance of the conditional density estimator $\widehat{p}(\boldsymbol{y}|\boldsymbol{x})$ is.



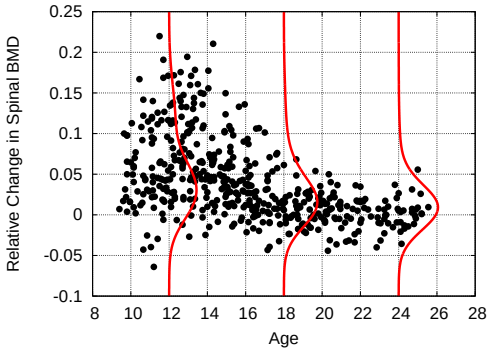
(a) Heteroscedastic Gaussian: $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1)$



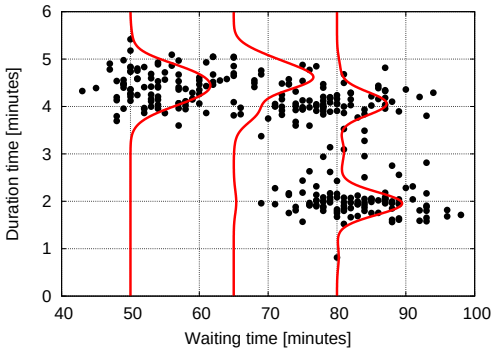
(b) Bimodal: $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \frac{1}{2}N(-1,\frac{4}{9}) + \frac{1}{2}N(1,\frac{4}{9})$



(c) Skewed: $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \frac{3}{4}N(0,1) + \frac{1}{4}N(\frac{3}{2},\frac{1}{9})$



(d) Bone Mineral Density



(e) Old Faithful Geyser

Figure 1: Illustrative examples of LS-CDE. (a) Artificial dataset containing heteroscedastic Gaussian noise, (b) Artificial dataset containing heteroscedastic bimodal Gaussian noise, (c) Artificial dataset containing heteroscedastic asymmetric bimodal Gaussian noise, (d) Relative spinal bone mineral density measurements on North American adolescents [13] having a heteroscedastic asymmetric conditional distribution, and (e) The durations of eruptions of the Old Faithful Geyser [46] having a bimodal conditional distribution.

Table 1: Experimental results on benchmark datasets ($d_{\mathrm{Y}} = 1$). The average and the standard deviation of NLL (see Eq.(10)) over 10 runs are described (smaller is better). The best method in terms of the mean error and comparable methods according to the two-sided paired *t-test* at the significance level 5% are specified by bold face. Mean computation time is normalized so that LS-CDE is one.

Dataset	(n, d_X)	LS-CDE	ϵ-KDE	MDN	KQR	RKDE
caution	(50,2)	**1.24 ± 0.29**	**1.25 ± 0.19**	**1.39 ± 0.18**	**1.73 ± 0.86**	17.11 ± 0.25
ftcollinssnow	(46,1)	**1.48 ± 0.01**	**1.53 ± 0.05**	**1.48 ± 0.03**	**2.11 ± 0.44**	46.06 ± 0.78
highway	(19,11)	**1.71 ± 0.41**	**2.24 ± 0.64**	7.41 ± 1.22	5.69 ± 1.69	15.30 ± 0.76
heights	(687,1)	**1.29 ± 0.00**	1.33 ± 0.01	**1.30 ± 0.01**	**1.29 ± 0.00**	54.79 ± 0.10
sniffer	(62,4)	**0.69 ± 0.16**	**0.96 ± 0.15**	**0.72 ± 0.09**	**0.68 ± 0.21**	26.80 ± 0.58
snowgeese	(22,2)	**0.95 ± 0.10**	1.35 ± 0.17	**2.49 ± 1.02**	**2.96 ± 1.13**	28.43 ± 1.02
ufc	(117,4)	**1.03 ± 0.01**	1.40 ± 0.02	**1.02 ± 0.06**	**1.02 ± 0.06**	11.10 ± 0.49
birthwt	(94,7)	**1.43 ± 0.01**	1.48 ± 0.01	**1.46 ± 0.01**	1.58 ± 0.05	15.95 ± 0.53
crabs	(100,6)	-0.07 ± 0.11	0.99 ± 0.09	**-0.70 ± 0.35**	**-1.03 ± 0.16**	12.60 ± 0.45
GAGurine	(157,1)	**0.45 ± 0.04**	0.92 ± 0.05	**0.57 ± 0.15**	**0.40 ± 0.08**	53.43 ± 0.27
geyser	(149,1)	**1.03 ± 0.00**	1.11 ± 0.02	1.23 ± 0.05	1.10 ± 0.02	53.49 ± 0.38
gilgais	(182,8)	0.73 ± 0.05	1.35 ± 0.03	**0.10 ± 0.04**	0.45 ± 0.15	10.44 ± 0.50
topo	(26,2)	**0.93 ± 0.02**	1.18 ± 0.09	2.11 ± 0.46	2.88 ± 0.85	10.80 ± 0.35
BostonHousing	(253,13)	0.82 ± 0.05	1.03 ± 0.05	**0.68 ± 0.06**	**0.48 ± 0.10**	17.81 ± 0.25
CobarOre	(19,2)	**1.58 ± 0.06**	**1.65 ± 0.09**	**1.63 ± 0.08**	6.33 ± 1.77	11.42 ± 0.51
engel	(117,1)	**0.69 ± 0.04**	1.27 ± 0.05	**0.71 ± 0.16**	N.A.	52.83 ± 0.16
mcycle	(66,1)	**0.83 ± 0.03**	1.25 ± 0.23	1.12 ± 0.10	**0.72 ± 0.06**	48.35 ± 0.79
BigMac2003	(34,9)	**1.32 ± 0.11**	**1.29 ± 0.14**	**2.64 ± 0.84**	**1.35 ± 0.26**	13.34 ± 0.52
UN3	(62,6)	**1.42 ± 0.12**	1.78 ± 0.14	**1.32 ± 0.08**	**1.22 ± 0.13**	11.43 ± 0.58
cpus	(104,7)	1.04 ± 0.07	1.01 ± 0.10	**-2.14 ± 0.13**	N.A.	15.16 ± 0.72
Time		1	0.004	267	0.755	0.089

We compared LS-CDE, ϵ-KDE, MDN, KQR, and RKDE. For model selection, we used CV based on the Kullback-Leibler (KL) error (see Section 2.5), which is consistent with the above NLL. In MDN, CV over three tuning parameters (the number of Gaussian components, the number of hidden units in the neural network, and the regularization parameter; see Section 3.2) was unbearably slow, so the number of Gaussian components was fixed to $t = 3$ and the other two tuning parameters were chosen by CV.

The experimental results are summarized in Table 1. ϵ-KDE was computationally very efficient, but it tended to perform rather poorly. MDN worked well, but it is computationally highly demanding. KQR overall performed well and it was computationally slightly more efficient than LS-CDE. However, its solution path tracking algorithm was numerically rather unstable and we could not obtain solutions for the 'engel' and 'cpus' datasets. RKDE did not perform well for all cases, implying that density ratio estimation via density estimation is not reliable in practice. Overall, the proposed LS-CDE was shown to be a promising method for conditional density estimation in terms of the accuracy and computational efficiency.

4.3 Robot Transition Estimation

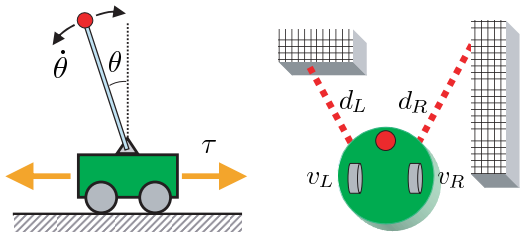
We further applied the proposed and existing methods to the problem of robot transition estimation. We used the pendulum robot and the Khepera robot simulators illustrated in Figure 2.

The pendulum robot consists of wheels and a pendulum hinged to the body. The state of the pendulum robot consists of angle θ and angular velocity $\dot{\theta}$ of the pendulum. The amount of torque τ applied to the wheels can be controlled, by which the robot can move left or right and the state of the pendulum is changed to θ' and $\dot{\theta}'$. The task is to estimate $p(\theta', \dot{\theta}'|\theta, \dot{\theta}, \tau)$, the transition probability density from state $(\theta, \dot{\theta})$ to state $(\theta', \dot{\theta}')$ by action τ.

The Khepera robot is equipped with two infra-red sensors and two wheels. The infra-red sensors d_L and d_R measure the distance to the left-front and right-front walls. The speed of left and right wheels v_L and v_R can be controlled separately, by which the robot can move forward/backward and rotate left/right. The task is to estimate $p(d'_L, d'_R|d_L, d_R, v_L, v_R)$, where d'_L and d'_R are the next state.

The state transition of the pendulum robot is highly stochastic due to slip, friction, or measurement errors with strong heteroscedasticity. Sensory inputs of the Khepera robot suffer from occlusions and contain highly heteroscedastic noise, so the transition probability density may possess multi-modality and heteroscedasticity. Thus transition estimation of dynamic robots is a challenging task. Note that transition estimation is highly useful in model-based reinforcement learning [39].

For both robots, 100 samples were used for conditional density estimation and additional 900 samples were used for computing NLL (see Eq.(10)). The number of Gaussian components was fixed to $t = 3$ in MDN, and all other tuning parameters were chosen by CV based on the Kullback-Leibler (KL) error (see Section 2.5). Experimental results are summarized in Table 2, showing that LS-CDE is still useful in this challenging task of



(a) Pendulum robot. (b) Khepera robot.

Figure 2: Illustration of robots used for experiments.

Table 2: Experimental results on robot transition estimation. The average and the standard deviation of NLL (see Eq.(10)) over 10 runs are described (smaller is better). The best method in terms of the mean error and comparable methods according to the two-sided paired *t-test* at the significance level 5% are specified by bold face. Mean computation time is normalized so that LS-CDE is one.

Dataset	LS-CDE	ϵ-KDE	MDN	RKDE
Pendulum1	**1.27 ± 0.05**	2.04 ± 0.10	**1.44 ± 0.67**	11.24 ± 0.32
Pendulum2	**1.38 ± 0.05**	2.07 ± 0.10	**1.43 ± 0.58**	11.24 ± 0.32
Khepera1	**1.69 ± 0.01**	2.07 ± 0.02	**1.90 ± 0.36**	11.03 ± 0.03
Khepera2	**1.86 ± 0.01**	2.10 ± 0.01	**1.92 ± 0.26**	11.09 ± 0.02
Time	1	0.164	1134	0.431

robot transition estimation.

5 Conclusions and Outlook

We proposed a novel approach to conditional density estimation called LS-CDE. Our basic idea was to directly estimate the ratio of density functions without going through density estimation. LS-CDE was shown to offer a sparse solution in an analytic form and therefore is computationally efficient. A non-parametric convergence rate of the LS-CDE algorithm was also provided. Experiments on benchmark and robot-transition datasets demonstrated the usefulness of LS-CDE.

The validity of the proposed LS-CDE method may be intuitively explained by the fact that it is a "one-shot" procedure (directly estimating the density ratio)—while a naive approach is two-fold (two densities are first estimated and then the ratio is estimated by plugging in the estimated densities). The plug-in approach may not be preferable since the first step is carried out without taking into account how the estimation error produced in the first step influences the second step; indeed, our experiments showed that the two-step procedure did not work properly in all cases. On the other hand, beyond these experimental results, it is important to theoretically investigate how and why the one-shot approach is more suitable than the plug-in approach in conditional density estimation

or in more general context of density ratio estimation. Furthermore, it is important to elucidate the asymptotic distribution of the proposed estimator, e.g., following the line of the papers [11, 27].

We used a regularizer expressed in terms of the density-ratio function in the theoretical analysis in Section 2.4, while the ℓ_2-norm of the parameter vector was adopted in the practical procedure described in Section 2.3. In order to fill the gap between theory and the practical procedure, we may use, for example, the squared norm in the function space as a regularizer. However, in our preliminary experiments, we found that the use of the function space norm as the regularizer was numerically unstable. Thus, an important future topic is to further investigate the role of regularizers in terms of consistency and also numerical stability. *Smoothed analysis* [35, 32, 19] would be a promising approach to addressing this issue.

In the proposed LS-CDE method, a direct density-ration estimation method based on the squared-loss called *unconstrained Least-Squares Importance Fitting (uLSIF)* [17, 18] was applied to conditional density estimation. Similarly, applying the direct density-ration estimation method based on the *log-loss* called the *Kullback-Leibler Importance Estimation Procedure (KLIEP)* [37, 38, 28, 29, 43, 44, 48], we can obtain a *log-loss* variant of the proposed method. A valiant of the KLIEP method explored in the papers [43, 44] uses a *log-linear* model (a.k.a. a *maximum entropy* model [16]) for density ratio estimation:

$$\widehat{r}_{\boldsymbol{\alpha}}(\boldsymbol{x},\boldsymbol{y}) := \frac{\exp(\boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y}))}{\int \exp(\boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y}'))\mathrm{d}\boldsymbol{y}'}.$$

Applying this log-linear KLIEP method to conditional density estimation is actually equivalent to maximum likelihood estimation of conditional densities for log-linear models (for structured output, it is particularly called a *conditional random field* [23]):

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^b} \left[\sum_{i=1}^{n} \log \widehat{r}_{\boldsymbol{\alpha}}(\boldsymbol{x}_i,\boldsymbol{y}_i)\right].$$

A crucial fact regarding maximum-likelihood conditional density estimation is that the normalization factor $\int \exp(\boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y}'))\mathrm{d}\boldsymbol{y}'$ needs to be included in the model; otherwise the likelihood tends to infinity. On the other hand, the proposed method (based on the squared-loss) does not require the normalization factor to be included in the optimization problem. This is evidenced by the fact that, without the normalization factor, the proposed LS-CDE estimator is still consistent (see Section 2.4). This highly contributes to simplifying the optimization problem (see Eq.(5)); indeed, by choosing the linear density ratio model (1), the solution can be obtained analytically, as shown in Eq.(6). This is a significant advantage of the proposed method over standard maximum-likelihood conditional density estimation. An interesting theoretical research direction along this line would be to generalize the loss function to a broader class, for example, the f-divergences [1, 7]. An approach based on the paper [28] would be a promising direction to pursue.

When $d_{\mathrm{X}} = 1$, $d_{\mathrm{Y}} = 1$, and the true ratio r is twice differentiable, the convergence rate of ϵ-KDE is

$$\|\widehat{r} - r\|_2 = \mathcal{O}_p(n^{-1/3}),$$

given that the bandwidth ϵ is optimally chosen [15]. On the other hand, Theorem 1 and the discussions before the theorem in the current paper imply that our method can achieve

$$\mathcal{O}_p(n^{-\frac{1}{(d_{\mathrm{X}}+d_{\mathrm{Y}})/p+2}}\log n) = \mathcal{O}_p(n^{-\frac{1}{2/p+2}}\log n),$$

where $p \geq 2$. Therefore, if we choose a model such that $p > 2$ and the true ratio r is contained, the convergence rate of our proposed method dominates that of ϵ-KDE. This is because ϵ-KDE just takes the average around the target point x and hence it does not capture higher order smoothness of the target conditional density. On the other hand, our method can utilize the information of higher order smoothness by properly choosing the degree of smoothness with cross-validation.

In the experiments, we used the common width for all Gaussian basis functions (9) in the proposed LS-CDE procedure. As shown in Section 2.4, the proposed method is consistent even with the common Gaussian width. However, in practice, it may be more flexible to use different Gaussian widths for different basis functions. Although our method in principle allows the use of different Gaussian widths, this in turn makes cross-validation computationally harder. A possible measure for this issue would be to also learn the Gaussian widths from the data together with coefficients. An expectation-maximization approach to density ratio estimation based on the log-loss for Gaussian mixture models has been explored [48], which allows us to learn the Gaussian covariances in an efficient manner. Developing a squared-loss variant of this method could produce a useful variant of the proposed LS-CDE method.

As explained in Section 2.7, LS-CDE can take advantage of the semi-supervised learning setup [5], although this was not explored in the current paper. Thus it is important to investigate how the performance of LS-CDE is improved under semi-supervised learning scenarios both in theoretically and experimentally.

Although we focused on conditional density estimation in this article, one may have interests in substantially simpler tasks such as error bar estimation and confidential interval estimation. Investigating whether the current line of research can be adapted to solving such simpler problems with higher accuracy is an important future issue.

Even though the proposed approach was shown to work well in experiments, its performance (and also the performance of any other non-parametric approaches) is still poor in high-dimensional problems. Thus, a further challenge is to improve the accuracy in high-dimensional cases. Application of a dimensionality reduction idea in density ratio estimation [36] would be a promising direction to address this issue.

Another possible future work from the application point of view would be the use of the proposed method in reinforcement learning scenarios, since a good transition model can be directly used for solving Markov decision problems in continuous and multi-dimensional domains [24, 39]. Our future work will explore this issue in more detail.

Acknowledgments

We thank fruitful comments from anonymous reviewers. MS was supported by AOARD, SCAT, and the JST PRESTO program.

References

[1] S.M. Ali and S.D. Silvey, "A general class of coefficients of divergence of one distribution from another," Journal of the Royal Statistical Society, Series B, vol.28, no.1, pp.131–142, 1966.

[2] A. Basu, I.R. Harris, N.L. Hjort, and M.C. Jones, "Robust and efficient estimation by minimising a density power divergence," Biometrika, vol.85, no.3, pp.540–559, 1998.

[3] S. Bickel, M. Brückner, and T. Scheffer, "Discriminative learning for differing training and test distributions," Proceedings of the 24th International Conference on Machine Learning, pp.81–88, 2007.

[4] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, NY, USA, 2006.

[5] O. Chapelle, B. Schölkopf, and A. Zien, eds., Semi-Supervised Learning, MIT Press, Cambridge, 2006.

[6] K.F. Cheng and C.K. Chu, "Semiparametric density estimation under a two-sample density ratio model," Bernoulli, vol.10, no.4, pp.583–604, 2004.

[7] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," Studia Scientiarum Mathematicarum Hungarica, vol.2, pp.229–318, 1967.

[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, vol.39, no.1, pp.1–38, 1977.

[9] D. Edmunds and H. Triebel, eds., Function Spaces, Entropy Numbers, Differential Operators, Cambridge Univ Press, 1996.

[10] J. Fan, Q. Yao, and H. Tong, "Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems," Biometrika, vol.83, no.1, pp.189–206, 1996.

[11] S. Geman and C.R. Hwang, "Nonparametric maximum likelihood estimation by the method of sieves," The Annals of Statistics, vol.10, no.2, pp.401–414, 1982.

[12] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz, Nonparametric and Semiparametric Models, Springer, Berlin, 2004.

[13] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, 2001.

[14] J. Huang, A. Smola, A. Gretton, K.M. Borgwardt, and B. Schölkopf, "Correcting sample selection bias by unlabeled data," in Advances in Neural Information Processing Systems 19, ed. B. Schölkopf, J. Platt, and T. Hoffman, pp.601–608, MIT Press, Cambridge, MA, 2007.

[15] R.J. Hyndman, D.M. Bashtannyk, and G.K. Grunwald, "Estimating and visualizing conditional densities," Journal of Computational and Graphical Statistics, vol.5, no.4, pp.315–336, 1996.

[16] E.T. Jaynes, "Information theory and statistical mechanics," Physical Review, vol.106, no.4, pp.620–630, 1957.

[17] T. Kanamori, S. Hido, and M. Sugiyama, "Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection," Advances in Neural Information Processing Systems 21, ed. D. Koller, D. Schuurmans, Y. Bengio, and L. Botton, Cambridge, MA, pp.809–816, MIT Press, 2009.

[18] T. Kanamori, S. Hido, and M. Sugiyama, "A least-squares approach to direct importance estimation," Journal of Machine Learning Research, vol.10, pp.1391–1445, Jul. 2009.

[19] T. Kanamori, T. Suzuki, and M. Sugiyama, "Condition number analysis of kernel-based density ratio estimation," Tech. Rep. TR09-0006, Department of Computer Science, Tokyo Institute of Technology, Feb. 2009.

[20] G.S. Kimeldorf and G. Wahba, "Some results on Tchebycheffian spline functions," Journal of Mathematical Analysis and Applications, vol.33, no.1, pp.82–95, 1971.

[21] A.N. Kolmogorov and V.M. Tikhomirov, "ε-entropy and ε-capacity of sets in function spaces," American Mathematical Society Translations, vol.17, no.2, pp.277–364, 1961.

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

[23] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," Proceedings of the 18th International Conference on Machine Learning, pp.282–289, 2001.

[24] M.G. Lagoudakis and R. Parr, "Least-squares policy iteration," Journal of Machine Learning Research, vol.4, pp.1107–1149, 2003.

[25] Y. Li, Y. Liu, and J. Zhu, "Quantile regression in reproducing kernel Hilbert spaces," Journal of the American Statistical Association, vol.102, no.477, pp.255–268, 2007.

[26] A. Luntz and V. Brailovsky, "On estimation of characters obtained in statistical procedure of recognition," Technicheskaya Kibernetica, vol.3, 1969. in Russian.

[27] W.K. Newey, "Convergence rates and asymptotic normality for series estimators," Journal of Econometrics, vol.70, no.1, pp.147–168, 1997.

[28] X. Nguyen, M. Wainwright, and M. Jordan, "Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization," in Advances in Neural Information Processing Systems 20, ed. J.C. Platt, D. Koller, Y. Singer, and S. Roweis, pp.1089–1096, MIT Press, Cambridge, MA, 2008.

[29] X. Nguyen, M.J. Wainwright, and M.I. Jordan, "Nonparametric estimation of the likelihood ratio and divergence functionals," Proceedings of IEEE International Symposium on Information Theory, Nice, France, pp.2016–2020, 2007.

[30] J. Qin, "Inferences for case-control and semiparametric two-sample density ratio models," Biometrika, vol.85, no.3, pp.619–639, 1998.

[31] R Development Core Team, The R Manuals, 2008. http://www.r-project.org.

[32] A. Sankar, D.A. Spielman, and S.H. Teng, "Smoothed analysis of the condition numbers and growth factors of matrices," SIAM Journal on Matrix Analysis and Applications, vol.28, no.2, pp.446–476, 2006.

[33] B. Schölkopf and A.J. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002.

[34] D.W. Scott, "Remarks on fitting and interpreting mixture models," Computing Science and Statistics, vol.31, pp.104–109, 1999.

[35] D.A. Spielman and S.H. Teng, "Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time," Journal of the ACM, vol.51, no.3, pp.385–463, 2004.

[36] M. Sugiyama, M. Kawanabe, and P.L. Chui, "Dimensionality reduction for density ratio estimation in high-dimensional spaces," Neural Networks, vol.23, no.1, pp.44–59, 2010.

[37] M. Sugiyama, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," Advances in Neural Information Processing Systems 20, ed. J.C. Platt, D. Koller, Y. Singer, and S. Roweis, Cambridge, MA, pp.1433–1440, MIT Press, 2008.

[38] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe, "Direct importance estimation for covariate shift adaptation," Annals of the Institute of Statistical Mathematics, vol.60, no.4, pp.699–746, 2008.

[39] R.S. Sutton and G.A. Barto, Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA, 1998.

[40] I. Takeuchi, Q.V. Le, T.D. Sears, and A.J. Smola, "Nonparametric quantile estimation," Journal of Machine Learning Research, vol.7, pp.1231–1264, 2006.

[41] I. Takeuchi, K. Nomura, and T. Kanamori, "Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression," Neural Computation, vol.21, no.2, pp.533–559, 2009.

[42] V. Tresp, "Mixtures of gaussian processes," Advances in Neural Information Processing Systems 13, ed. T.K. Leen, T.G. Dietterich, and V. Tresp, pp.654–660, MIT Press, 2001.

[43] Y. Tsuboi, H. Kashima, S. Hido, S. Bickel, and M. Sugiyama, "Direct density ratio estimation for large-scale covariate shift adaptation," Proceedings of the Eighth SIAM International Conference on Data Mining (SDM2008), ed. M.J. Zaki, K. Wang, C. Apte, and H. Park, Atlanta, Georgia, USA, pp.443–454, Apr. 24–26 2008.

[44] Y. Tsuboi, H. Kashima, S. Hido, S. Bickel, and M. Sugiyama, "Direct density ratio estimation for large-scale covariate shift adaptation," Journal of Information Processing, vol.17, pp.138–155, 2009.

[45] A.W. van der Vaart and J.A. Wellner, Weak Convergence and Empirical Processes with Applications to Statistics, Springer, New York, NY, USA, 1996.

[46] S. Weisberg, Applied Linear Regression, John Wiley, New York, NY, USA, 1985.

[47] R.C.L. Wolff, Q. Yao, and P. Hall, "Methods for estimating a conditional distribution function," Journal of the American Statistical Association, vol.94, no.445, pp.154–163, 1999.

[48] M. Yamada and M. Sugiyama, "Direct importance estimation with Gaussian mixture models," IEICE Transactions on Information and Systems, vol.E92-D, no.10, pp.2159–2162, 2009.

[49] D.X. Zhou, "The covering number in learning theory," Journal of Complexity archive, vol.18, no.3, pp.739–767, 2002.