Vol.2009-MPS-76 No.4
Vol.2009-BIO-19 No.4
2009/12/17

Conditional Density Estimation Based on Density Ratio Estimation

Masashi Sugiyama ,[†1,†2] Ichiro Takeuchi ,[†3]
Taiji Suzuki ,[†4] Takafumi Kanamori ,[†5]
Hirotaka Hachiya [†1] and Daisuke Okanohara [†4]

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.

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 multi-modality (e.g., inverse kinematics learning of a robot[3]) or a highly skewed profile with heteroscedastic noise (e.g., biomedical data analysis[10]). 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.

†1 Tokyo Institute of Technology
†2 Japan Science and Technology Agency
†3 Nagoya Institute of Technology
†4 The University of Tokyo
†5 Nagoya University

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[8),32)], but they still share the same weakness.

The mixture density network (MDN)[3)] 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[28)]. The mixture model is trained in a computationally efficient manner by an expectation-maximization algorithm[6)]. 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[17),26)] 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[27)]. 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.

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

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

$\boldsymbol{0}_b$ 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

$$\widehat{\boldsymbol{H}} := \frac{1}{n}\sum_{i=1}^{n} \overline{\boldsymbol{\Phi}}(\boldsymbol{x}_i), \qquad (4)$$

$$\widehat{\boldsymbol{h}} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\phi}(\boldsymbol{x}_i, \boldsymbol{y}_i).$$

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

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

where a regularizer $\lambda\boldsymbol{\alpha}^\top\boldsymbol{\alpha}$ ($\lambda > 0$) 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

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

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

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

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

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

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)*[12),13)] 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[⋆2]). 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

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

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,

⋆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 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[14)].

and $\mathcal{H}_{[]}$ is the *bracketing entropy* of $\mathcal{G}_M$ with respect to the $L_2(p_{\rm x} \times \mu_{\rm Y})$-norm[31].

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[31] 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 $(d_{\rm X} + d_{\rm Y})/p$, where p is the order of differentiability of the Sobolev space (see page 105 of the book[7] 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 $\gamma > 0$[34]. On the other hand, in the above setup, the bracketing entropy is lower-bounded by $K(M/\epsilon)^{(d_{\rm X}+d_{\rm Y})/p}$ with a constant K depending only on p, $d_{\rm X}$, and $d_{\rm Y}$[15]. Therefore, if the dimension of the domains $\mathcal{D}_{\rm X}$ and $\mathcal{D}_{\rm Y}$ is so large that $(d_{\rm X} + d_{\rm Y})/p > 2$, γ should be larger than 2. This means that a situation where p is small and $d_{\rm X}$ and $d_{\rm Y}$ 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 $d_{\rm X}$ and $d_{\rm Y}$ 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[19),25]).

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 $L_2(p_{\rm x} \times \mu_{\rm Y})$-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

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

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 $C_{\rm SQ}$ is the constant defined by

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

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

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

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 $C_{\rm KL}$ is the constant defined by

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

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

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

$$\widehat{\mathrm{KL}}_{\mathcal{Z}_k} := -\frac{1}{|\mathcal{Z}_k|}\sum_{(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}})\in\mathcal{Z}_k}\log\widehat{p}_{\mathcal{Z}\setminus\mathcal{Z}_k}(\widetilde{\boldsymbol{y}}|\widetilde{\boldsymbol{x}}),$$

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[18),22)].

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

$$\begin{aligned}\overline{\Phi}_{\ell,\ell'}(\boldsymbol{x}) &= \int\phi_\ell(\boldsymbol{x},\boldsymbol{y})\phi_{\ell'}(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{y}\\ &= (\sqrt{\pi}\sigma)^{d_{\mathcal{Y}}}\exp\left(-\frac{\xi_{\ell,\ell'}(\boldsymbol{x})}{4\sigma^2}\right),\end{aligned}$$

$$\int\widehat{\boldsymbol{\alpha}}^\top\boldsymbol{\phi}(\widetilde{\boldsymbol{x}},\boldsymbol{y})\mathrm{d}\boldsymbol{y} = (\sqrt{2\pi}\sigma)^{d_{\mathcal{Y}}}\sum_{\ell=1}^{b}\widehat{\alpha}_\ell\exp\left(-\frac{||\boldsymbol{x}-\boldsymbol{u}_\ell||^2}{2\sigma^2}\right),$$

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 practice, we may 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[4)]—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[9]. 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[8),32)], 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[3]. 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[17),26)]. 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[27]. 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). In our preliminary experiments, we found that 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[11] and the *logistic regression* based method[2),5),21)] 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[19),20),24),25),29),30),33)] employs the *Kullback-Leibler divergence*[16)] 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[12),13)]. Furthermore, in the context of density estimation, the squared-loss is often preferred to the Kullback-Leibler loss[1),23)].

4. Conclusions

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.

Acknowledgments

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

References

1) Basu, A., Harris, I.R., Hjort, N.L. and Jones, M.C.: Robust and Efficient Estimation by Minimising a Density Power Divergence, *Biometrika*, Vol.85, No.3, pp. 540–559 (1998).
2) Bickel, S., Brückner, M. and Scheffer, T.: Discriminative Learning for Differing Training and Test Distributions, *Proceedings of the 24th International Conference on Machine Learning*, pp.81–88 (2007).
3) Bishop, C.M.: *Pattern Recognition and Machine Learning*, Springer, New York, NY, USA (2006).
4) Chapelle, O., Schölkopf, B. and Zien, A.(eds.): *Semi-Supervised Learning*, MIT Press, Cambridge (2006).
5) Cheng, K.F. and Chu, C.K.: Semiparametric Density Estimation under a Two-sample Density Ratio Model, *Bernoulli*, Vol.10, No.4, pp.583–604 (2004).
6) Dempster, A.P., Laird, N.M. and Rubin, D.B.: 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).
7) Edmunds, D. and Triebel, H.(eds.): *Function Spaces, Entropy Numbers, Differential Operators*, Cambridge Univ Press (1996).
8) Fan, J., Yao, Q. and Tong, H.: Estimation of Conditional Densities and Sensitivity Measures in Nonlinear Dynamical Systems, *Biometrika*, Vol.83, No.1, pp.189–206 (1996).
9) Härdle, W., Müller, M., Sperlich, S. and Werwatz, A.: *Nonparametric and Semiparametric Models*, Springer, Berlin (2004).
10) Hastie, T., Tibshirani, R. and Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York (2001).
11) Huang, J., Smola, A., Gretton, A., Borgwardt, K.M. and Schölkopf, B.: Correcting Sample Selection Bias by Unlabeled Data, *Advances in Neural Information Processing Systems 19* (Schölkopf, B., Platt, J. and Hoffman, T., eds.), MIT Press, Cambridge, MA, pp.601–608 (2007).
12) Kanamori, T., Hido, S. and Sugiyama, M.: Efficient Direct Density Ratio Estimation for Non-stationarity Adaptation and Outlier Detection, *Advances in Neural Information Processing Systems 21* (Koller, D., Schuurmans, D., Bengio, Y. and Botton, L., eds.), Cambridge, MA, MIT Press, pp.809–816 (2009).
13) Kanamori, T., Hido, S. and Sugiyama, M.: A Least-squares Approach to Direct Importance Estimation, *Journal of Machine Learning Research*, Vol.10, pp.1391–1445 (2009).
14) Kimeldorf, G.S. and Wahba, G.: Some Results on Tchebycheffian Spline Functions, *Journal of Mathematical Analysis and Applications*, Vol.33, No.1, pp.82–95 (1971).
15) Kolmogorov, A.N. and Tikhomirov, V.M.: ε-entropy and ε-capacity of Sets in Function Spaces, *American Mathematical Society Translations*, Vol.17, No.2, pp. 277–364 (1961).
16) Kullback, S. and Leibler, R.A.: On Information and Sufficiency, *Annals of Mathematical Statistics*, Vol.22, pp.79–86 (1951).
17) Li, Y., Liu, Y. and Zhu, J.: Quantile Regression in Reproducing Kernel Hilbert Spaces, *Journal of the American Statistical Association*, Vol.102, No.477, pp.255–268 (2007).
18) Luntz, A. and Brailovsky, V.: On Estimation of Characters Obtained in Statistical Procedure of Recognition, *Technicheskaya Kibernetica*, Vol.3 (1969). in Russian.
19) Nguyen, X., Wainwright, M. and Jordan, M.: Estimating Divergence Functionals and the Likelihood Ratio by Penalized Convex Risk Minimization, *Advances in Neural Information Processing Systems 20* (Platt, J.C., Koller, D., Singer, Y. and Roweis, S., eds.), MIT Press, Cambridge, MA, pp.1089–1096 (2008).
20) Nguyen, X., Wainwright, M.J. and Jordan, M.I.: Nonparametric Estimation of the Likelihood Ratio and Divergence Functionals, *Proceedings of IEEE International Symposium on Information Theory*, Nice, France, pp.2016–2020 (2007).
21) Qin, J.: Inferences for Case-control and Semiparametric Two-sample Density Ratio Models, *Biometrika*, Vol.85, No.3, pp.619–639 (1998).

Vol.2009-MPS-76 No.4
Vol.2009-BIO-19 No.4
2009/12/17

22) Schölkopf, B. and Smola, A.J.: *Learning with Kernels*, MIT Press, Cambridge, MA (2002).
23) Scott, D.W.: Remarks on Fitting and Interpreting Mixture Models, *Computing Science and Statistics*, Vol.31, pp.104–109 (1999).
24) Sugiyama, M., Nakajima, S., Kashima, H., von Bünau, P. and Kawanabe, M.: Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation, *Advances in Neural Information Processing Systems 20* (Platt, J.C., Koller, D., Singer, Y. and Roweis, S., eds.), Cambridge, MA, MIT Press, pp. 1433–1440 (2008).
25) Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P. and Kawanabe, M.: Direct Importance Estimation for Covariate Shift Adaptation, *Annals of the Institute of Statistical Mathematics*, Vol.60, No.4, pp.699–746 (2008).
26) Takeuchi, I., Le, Q.V., Sears, T.D. and Smola, A.J.: Nonparametric Quantile Estimation, *Journal of Machine Learning Research*, Vol.7, pp.1231–1264 (2006).
27) Takeuchi, I., Nomura, K. and Kanamori, T.: Nonparametric Conditional Density Estimation Using Piecewise-linear Solution Path of Kernel Quantile Regression, *Neural Computation*, Vol.21, No.2, pp.533–559 (2009).
28) Tresp, V.: Mixtures of Gaussian Processes, *Advances in Neural Information Processing Systems 13* (Leen, T.K., Dietterich, T.G. and Tresp, V., eds.), MIT Press, pp.654–660 (2001).
29) Tsuboi, Y., Kashima, H., Hido, S., Bickel, S. and Sugiyama, M.: Direct Density Ratio Estimation for Large-scale Covariate Shift Adaptation, *Proceedings of the Eighth SIAM International Conference on Data Mining (SDM2008)* (Zaki, M.J., Wang, K., Apte, C. and Park, H., eds.), Atlanta, Georgia, USA, pp.443–454 (2008).
30) Tsuboi, Y., Kashima, H., Hido, S., Bickel, S. and Sugiyama, M.: Direct Density Ratio Estimation for Large-scale Covariate Shift Adaptation, *Journal of Information Processing*, Vol.17, pp.138–155 (2009).
31) vander Vaart, A.W. and Wellner, J.A.: *Weak Convergence and Empirical Processes with Applications to Statistics*, Springer, New York, NY, USA (1996).
32) Wolff, R. C.L., Yao, Q. and Hall, P.: Methods for Estimating a Conditional Distribution Function, *Journal of the American Statistical Association*, Vol.94, No.445, pp.154–163 (1999).
33) Yamada, M. and Sugiyama, M.: Direct Importance Estimation with Gaussian Mixture Models, *IEICE Transactions on Information and Systems*, Vol.E92-D, No. 10, pp.2159–2162 (2009).
34) Zhou, D.-X.: The Covering Number in Learning Theory, *Journal of Complexity archive*, Vol.18, No.3, pp.739–767 (2002).