International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No. 2 ,Issue No.9, pp : 900-904 1 Sept.2013

Semi-Supervised Least-Squares Conditional Density Estimation

Rubaiya Rahtin Khan[a], Masashi Sugiyama[b]
[a]United International University, [b]Tokyo Institute of Technology
[a]rubaiya@cse.uiu.ac.bd, [b]sugi@cs.titech.ac.jp

Abstract----*Conditional density estimation is an useful alternative to regression to learn an input-output relationship under multi-modality, asymmetry, and heteroscedasticity. The supervised learning method called least-squares conditional density estimation (LSCDE) is the state-of-the-art method that directly estimates the conditional density using a linear model. In this paper, we extend the supervised LSCDE method to a semi-supervised scenario so that unlabelled data can be utilized, and numerically illustrates its usefulness.*

Keywords---- Semi-supervised learning, Conditional density, Least squares, Direct density ratio estimation.

I. Introduction

Machine learning is broadly concerned with designing computer algorithms that learn from experience or automatically discover useful patterns from data. Traditionally, there have been two types of tasks in machine learning, namely unsupervised learning and supervised learning [i]. Semi-supervised learning is a special branch of machine learning that tries to combine these two fundamental tasks. In addition to unlabelled data, the algorithm is provided with some supervision information but not necessarily for all examples. In this case, the data set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{n+n'}$ can be divided into two parts: the points $\mathbf{X}_l = (\mathbf{x}_1,..., \mathbf{x}_n)$, for which labels $\mathbf{Y}_l = (\mathbf{y}_1,..., \mathbf{y}_n)$ are provided, and the points $\mathbf{X}_u = (\mathbf{x}_{n+1},..., \mathbf{x}_{n+n'})$, the labels of which are not known. This is the standard setting for semi-supervised learning.

The goal of regression analysis is to estimate the conditional mean of output $\mathbf{y}$ given input $\mathbf{x}$ [ii]. Regression is suitable for analysing an input-output relation if the conditional density of the relation $p(\mathbf{y}\,|\,\mathbf{x})$ is unimodal, symmetric, and homoscedastic. However, under multimodality, asymmetry, and heteroscedasticity, regression analysis is not informative enough, and it is preferable to estimate the conditional distribution itself.

There are already many existing methods for estimating conditional densities. Conditional density estimation was introduced by Rosenblatt [iii]. A bias correction was proposed by Hyndman, Bashtannyk and Grunwald [iv]. Fan, Yao and Tong [v] proposed a direct estimator based on local polynomial estimation. Bandwidth selection rules have been proposed by Bashtannyk and Hyndman [vi], Fan and Yim [vii], and Hall, Racine and Li [viii]. The conditional distribution estimation problem is examined in Hall, Wolff and Yao [ix]. Other papers have used conditional density estimates as an input to other problems, including Robinson [x], Tjostheim [xi], Polonik and Yao [xii], and Hyndman and Yao [xiii].

Table I: Existing methods for conditional density estimation

Name	Disadvantage
ϵ-neighbour kernel density estimation (ϵ-KDE)	It does not perform well in high dimensional problems.
Mixture density network (MDN)[xiv]	Its training is time-consuming and only a local solution may be obtained due to the non-convexity of neural network learning.
Kernel quantile regression(KQR)[xv][xvi]	The range of applications is limited to only one dimensional output y . Also, additional errors may occur while converting conditional cumulative distributions into conditional densities.
Ratio of kernel density estimators (RKDE)	Taking the ratio of two estimated quantities increases the estimation error.
Least-squares conditional density estimation (LSCDE)[xvii]	Unlabelled samples cannot be utilized.

Representative methods for estimating conditional densities are summarized in Table I, showing that LSCDE is the state-of-the-art method under the supervised learning setup. On the other hand, in many real-world situations, in addition to labelled samples, many unlabelled samples are also available for which only the input values are known without any information about the associated output values. In this work, we extend LSCDE to the semi-supervised learning setting so that such unlabelled samples can be utilized, and show its usefulness experimentally.

II. Proposed Method

In this section, we first review the existing LSCDE method and then explain its extension to semi-supervised settings.

A. Least-Squares Conditional Density Estimation

1) Direct Density Ratio Estimation: Let, $D_X (\subset R^{d_X})$ and $D_Y (\subset R^{d_Y})$ be input and output data domains, where d_X and d_Y are dimensionality of the input and output data domains. Now, considering a joint probability distribution on $D_x \times D_Y$ with probability density function $p(\mathbf{x},\mathbf{y})$, $p(\mathbf{x},\mathbf{y})$ independent and identically distributed

(i.i.d.) paired samples of input $\mathbf{x}$ and output $\mathbf{y}$ are assumed to be given as follows:
$\{\mathbf{z}_i | \mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i) \in D_X \times D_Y\}_{i=1}^{n}$.

The goal is to estimate the conditional density $p(\mathbf{y}|\mathbf{x})$ from the given samples $\{\mathbf{z}_i\}_{i=1}^{n}$. $p(\mathbf{y}|\mathbf{x})$ can be expressed as the ratio of two densities:

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

where it is assumed that the marginal density $p(\mathbf{x}) > 0$ for all $\mathbf{x} \in D_X$.

Generally, estimating the two densities separately and taking the ratio can result in large estimation error. To avoid this problem, the density ratio function $r(\mathbf{x},\mathbf{y})$ is estimated directly without going through estimating the two densities $p(\mathbf{x},\mathbf{y})$ and $p(\mathbf{x})$ separately.

2) Linear Density-Ratio Model: The density ratio function $r(\mathbf{x},\mathbf{y})$ is modelled by the following linear model:

$$\hat{r}_\alpha(\mathbf{x},\mathbf{y}) := \boldsymbol{\alpha}^T \phi(\mathbf{x},\mathbf{y}).$$

Here, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_b)^T$ are parameters to be learnt from samples and $\phi(\mathbf{x},\mathbf{y}) = (\phi_1(\mathbf{x},\mathbf{y}), \phi_2(\mathbf{x},\mathbf{y}), ..., \phi_b(\mathbf{x},\mathbf{y}))^T$ are basis functions such that $\phi(\mathbf{x},\mathbf{y}) \geq \mathbf{0}_b$ for all $(\mathbf{x},\mathbf{y}) \in D_X \times D_Y$ where b is the number of basis functions and $\mathbf{0}_b$ denotes the b-dimensional vector with all zeros.

3) Least-Squares Approach: In LSCDE, the parameter $\boldsymbol{\alpha}$ is determined so that the following squared error J_0 is minimized:

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

This can be expressed as

$$\begin{aligned} J_0(\boldsymbol{\alpha}) &:= \frac{1}{2}\iint \hat{r}_\alpha(\mathbf{x},\mathbf{y})^2 p(\mathbf{x}) d\mathbf{x} d\mathbf{y} \\ &\quad - \iint \hat{r}_\alpha(\mathbf{x},\mathbf{y}) r(\mathbf{x},\mathbf{y}) p(\mathbf{x}) d\mathbf{x} d\mathbf{y} + C \\ &= \frac{1}{2}\iint (\boldsymbol{\alpha}^T \phi(\mathbf{x},\mathbf{y}))^2 p(\mathbf{x}) d\mathbf{x} d\mathbf{y} \\ &\quad - \iint \boldsymbol{\alpha}^T \phi(\mathbf{x},\mathbf{y}) p(\mathbf{x},\mathbf{y}) d\mathbf{x} d\mathbf{y} + C, \end{aligned}$$

where

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

is a constant and therefore can be safely ignored. We denote the first two terms by J:

$$J(\boldsymbol{\alpha}) := J_0(\boldsymbol{\alpha}) - C = \frac{1}{2}\boldsymbol{\alpha}^T \mathbf{H} \boldsymbol{\alpha} - \mathbf{h}^T \boldsymbol{\alpha}, \qquad (1)$$

where

$$\mathbf{H} := \int \overline{\boldsymbol{\Phi}}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x},$$

$$\mathbf{h} := \iint \phi(\mathbf{x},\mathbf{y}) p(\mathbf{x},\mathbf{y}) d\mathbf{x} d\mathbf{y},$$

$$\overline{\boldsymbol{\Phi}}(\mathbf{x}) := \int \phi(\mathbf{x},\mathbf{y}) \phi(\mathbf{x},\mathbf{y})^T d\mathbf{y}. \qquad (2)$$

As $\mathbf{H}$ and $\mathbf{h}$ included in $J(\boldsymbol{\alpha})$ contain the expectations over unknown densities $p(\mathbf{x})$ and $p(\mathbf{x},\mathbf{y})$, they are approximated by sample averages as follows:

$$\hat{J}(\boldsymbol{\alpha}) := \frac{1}{2}\boldsymbol{\alpha}^T \hat{\mathbf{H}} \boldsymbol{\alpha} - \hat{\mathbf{h}}^T \boldsymbol{\alpha},$$

where

$$\hat{\mathbf{H}} := \frac{1}{n}\sum_{i=1}^{n} \overline{\boldsymbol{\Phi}}(\mathbf{x}_i), \qquad (3)$$

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

Now, the optimization criterion is summarized as

$$\tilde{\boldsymbol{\alpha}} := \underset{\alpha \in R_b}{\arg\min} \left[\hat{J}(\boldsymbol{\alpha}) + \frac{\lambda}{2} \boldsymbol{\alpha}^T \boldsymbol{\alpha} \right].$$

Here, the parameter $\lambda > 0$ is included as a regularization parameter for stabilization purposes. Taking the derivative of the objective function and equating it to zero, the solution $\tilde{\boldsymbol{\alpha}}$ can be obtained by solving the following system of linear equations:

$$(\hat{\mathbf{H}} + \lambda \mathbf{I}_b)\boldsymbol{\alpha} = \hat{\mathbf{h}}.$$

Here, $\mathbf{I}_b$ denotes the b-dimensional identity matrix. The solution $\tilde{\boldsymbol{\alpha}}$ is given analytically as

$$\tilde{\boldsymbol{\alpha}} = (\hat{\mathbf{H}} + \lambda \mathbf{I}_b)^{-1} \hat{\mathbf{h}}.$$

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

$$\hat{\boldsymbol{\alpha}} = \max(\mathbf{0}_b, \tilde{\boldsymbol{\alpha}}).$$

We renormalize the solution in the test phase to assure that the obtained density-ratio function is a conditional density. More specifically, for test input point $\tilde{\mathbf{x}}$, the final solution is given as

$$\hat{p}(\mathbf{y}|\mathbf{x} = \tilde{\mathbf{x}}) = \frac{\hat{\boldsymbol{\alpha}}^T \phi(\tilde{\mathbf{x}},\mathbf{y})}{\int \hat{\boldsymbol{\alpha}}^T \phi(\tilde{\mathbf{x}},\mathbf{y}') d\mathbf{y}'}. \qquad (4)$$

B. Semi-Supervised LSCDE

In many real-world problems, in addition to labelled samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$, unlabelled samples $\{\mathbf{x}_i\}_{i=n+1}^{n+n'}$ which are drawn independently from the marginal density $p(\mathbf{x})$ are also available. In this section, we extend LSCDE to utilize such unlabelled samples to further improve the estimation accuracy.

From Equation (3), the matrix $\mathbf{H}$ can be approximated in the semi-supervised setup as follows:

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

Inclusion of unlabelled samples may decrease the estimation variance, while increasing the estimation bias. Due to this bias increase, there is a possibility that the use of unlabelled samples degrades the overall performance. To avoid this problem, we introduce a parameter $0 \le \gamma \le 1$, which controls the weight of the unlabelled samples to be used in the estimation of $\mathbf{H}$. The modified equation is given as follows:

$$\hat{\mathbf{H}}' = (1-\gamma)\hat{\mathbf{H}}'_l + \gamma\hat{\mathbf{H}}'_u,$$

where

$$\hat{\mathbf{H}}'_l := \frac{1}{n}\sum_{i=1}^{n}\overline{\mathbf{\Phi}}(\mathbf{x}_i),$$

$$\hat{\mathbf{H}}'_u := \frac{1}{n'}\sum_{i=n+1}^{n+n'}\overline{\mathbf{\Phi}}(\mathbf{x}_i).$$

III. Experiments

In this section, we experimentally compare the performance of the supervised and semi-supervised LSCDE methods. The accuracy of $\hat{p}(\mathbf{y}\,|\,\mathbf{x})$ is measured by the negative log-likelihood for $\tilde{n}$ test samples $\{\tilde{\mathbf{z}}_i \,|\, \tilde{\mathbf{z}}_i = (\tilde{\mathbf{x}}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^{\tilde{n}}$:

$$NLL := -\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\log \hat{p}(\tilde{\mathbf{y}}_i \,|\, \tilde{\mathbf{x}}_i).$$

A. Illustration

Experiments were performed on one-dimensional toy data sets, that is, $d_X = d_Y = 1$. Inputs $\{x_i\}_{i=1}^{n}$ were independently drawn from the uniform distribution on $(-1,1)$. Outputs $\{y_i\}_{i=1}^{n}$ were generated by a heteroscedastic noise model as follows:

$$y_i = \sin c(2\pi x_i) + \frac{1}{8}\exp(1-x_i)\varepsilon_i,$$

where ε_i is the noise for the i th sample. 10000 test samples were used to evaluate the test performance.

Figure 1 shows data samples with Gaussian noise:

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

where $N(\mu,\sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2. Figure 2 shows NLL of semi-supervised LSCDE in comparison with supervised LSCDE. It is observed that introducing unlabelled samples tends to decrease NLL when the number of labelled samples is small.

Figure 3 shows data samples with the following skewed noise:

$$\varepsilon_i \overset{i.i.d}{\sim} \frac{3}{4}N(0,1) + \frac{1}{4}N\left(\frac{2}{3},\frac{1}{9}\right),$$

Figure 4 shows NLL of semi-supervised LSCDE in comparison with supervised LSCDE. It is observed again that the use of unlabelled samples tends to decrease NLL.

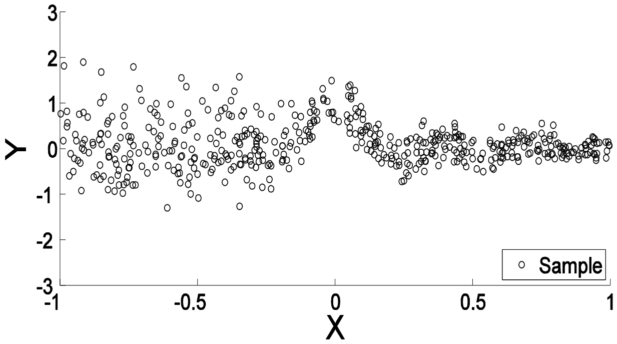

Figure 1: Heteroscedastic Gaussian noise data

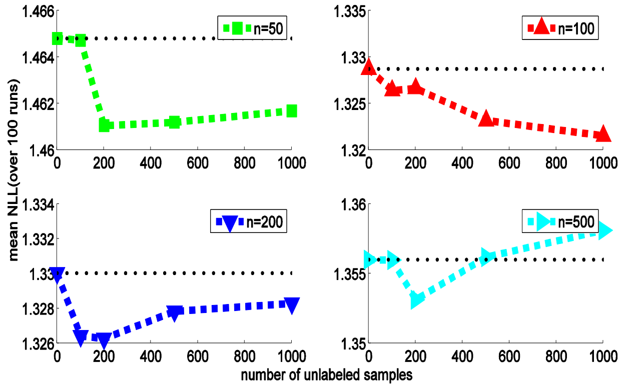

Figure 2: NLL of semi-supervised LSCDE for heteroscedastic Gaussian noise data as functions of the number of unlabelled samples. Dotted lines denote NLL of supervised LSCDE (i.e., semi-supervised LSCDE with no unlabelled samples).

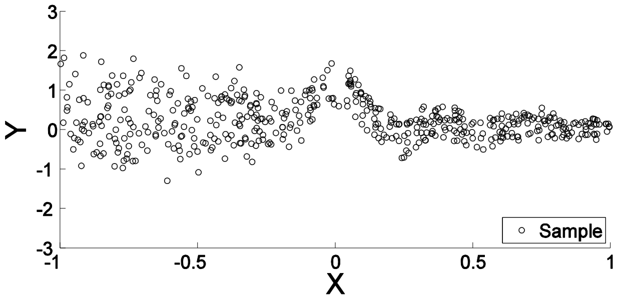

Figure 3: Heteroscedastic skewed noise data.

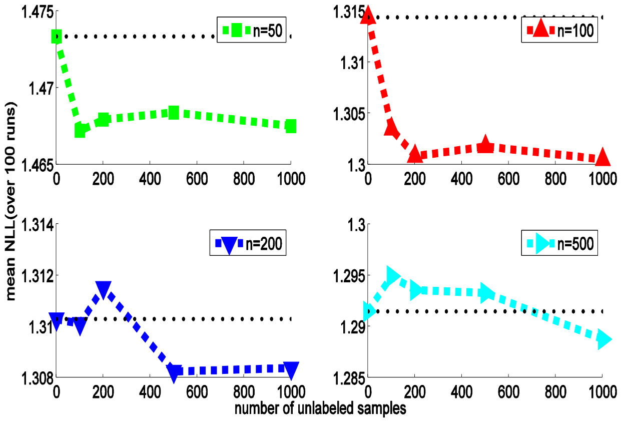

Figure 4: NLL of semi-supervised LSCDE for heteroscedastic skewed noise data as functions of the number of unlabelled samples. Dotted lines denote NLL of supervised LSCDE (i.e., semi-supervised LSCDE with no unlabelled samples).

B. Robot Data

The ball-batting robot data [xviii] was used for experiments. Figure 5 illustrates a ball-batting robot which consists of two links and two joints. Input variables are angles of Joint 1 and 2, angular Velocities of Joint 1 and 2, and torques applied to Joint 1 and 2, and the output variable is the carry of the ball. 1000 test samples were used to evaluate the test performance.

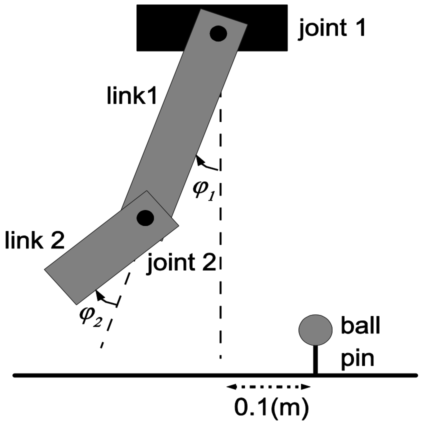

Figure 5: A ball-batting robot.

The mean NLL over 100 runs for the supervised and semi-supervised LSCDE methods are shown in Figure 6. Table II presents the mean and standard deviation of NLL over 100 runs. The boldfaced results were shown to be significantly better based on the Wilcoxon signed rank test [xix] at the 5% significance level. This shows that the proposed semi-supervised LSCDE method tends to outperform the existing supervised LSCDE method.

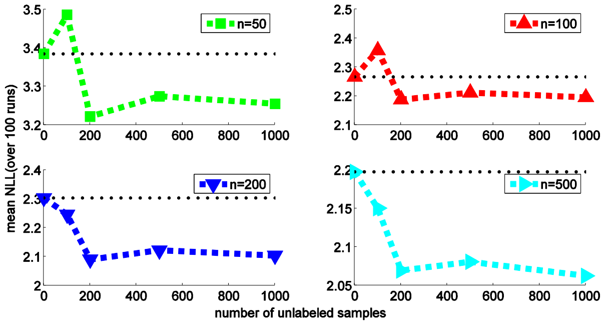

Figure 6: NLL of semi-supervised LSCDE for ball-batting robot data as functions of the number of unlabelled samples. Dotted lines denote NLL of supervised LSCDE (i.e., semi-supervised LSCDE with no unlabelled samples).

Table II: Mean and standard deviation of NLL for ball-batting robot data over 100 runs.

	$n=50$	$n=100$	$n=200$	$n=500$
$n'=0$	3.385±3.097	2.267±1.664	2.303±1.212	2.198±0.967
$n'=100$	3.486±3.033	2.358±2.038	**2.246±1.144**	**2.150±0.882**
$n'=200$	3.222±3.033	**2.188±1.917**	**2.090±0.905**	**2.070±0.784**
$n'=500$	3.275±2.760	2.212±1.710	**2.122±0.892**	**2.081±0.770**
$n'=1000$	3.255±2.729	2.196±1.694	**2.104±0.858**	**2.063±0.750**

IV. Conclusion

In this paper, the supervised least-squares conditional density estimation method was extended to a semi-supervised setting. Experiments showed that utilization of unlabelled samples tends to improve the accuracy of LSCDE when the number of labelled samples is small.

Acknowledgement

This work was supported by MEXT KAKENHI 23120004.

References

i. *O. Chapelle, B. Schlkopf, and A. Zien, Semi-supervised Learning, The MIT Press, 2006.*

ii. *F.A. Graybill and H.K. Iyer, Regression Analysis: Concepts and Applications, third ed., National Council on Measurement in Education, April 1995.*

iii. *M. Rosenblatt, "Conditional probability density and regression estimates," Multivariate Analysis II, vol.Ed. P.R. Krishnaiah, pp.25–31, 1969.*

iv. *R.J. Hyndman, D.M. Bashtannyk, and G.K. Grunwald, "Estimating and visualizing conditional densities," Journal of Computational and Graphical Statistics, vol.5, pp.315–336, 1996.*

v. *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.*

vi. *D. Bashtannyk and R. Hyndman, "Bandwidth selection for kernel conditional density estimation," Computational Statistics and Data Analysis, vol.36, pp.279–298, 2001.*

vii. *J. Fan and T. Yim, "A cross-validation method for estimating conditional densities," Biometrika, vol.91(4), pp.819–834, 2004.*

viii. *P. Hall, J. Racine, and Q.Li, "Cross-validation and the estimation of conditional probability densities," Journal of the American Statistical Association, vol.99, No. 468, pp.1015–1026, 2004.*

ix. *P. Hall, R.C.L. Wolff, and Q. Yao, "Methods for Estimating a Conditional Distribution Function," Journal of the American Statistical Association, vol.94, No. 445, pp.154–163, 1999.*

x. *P. Robinson, "Consistent nonparametric entropy-based testing," Review of Economic Studies, vol.58, pp.437–453, 1991.*

xi. *D. Tjostheim, "Non-linear time series: A selective review," Scandinavian Journal of Statistics, vol.21, pp.97–130, 1994.*

xii. *W. Polonik and Q. Yao, "Conditional minimum volume preditive regions for stochastic processes," Journal of the American Statistical Association, vol.95, pp.509–519, 2000.*

xiii. *R. Hyndman and Q. Yao, "Nonparametric estimation and symmetry tests for conditional density functions," Nonparametric Statistics, vol.14, pp.259–278, 2002.*

xiv. *C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.*

xv. *I. Takeuchi, Q.V. Le, T.D. Sears, and A.J. Smola, "Nonparametric quantile estimation," The Journal of Machine Learning Research, vol.7, pp.1231–1264, 2006. [xvi] Y. Li, Y. Liu, and J. Zhu, "Quantile regression in reproducing kernel Hilbert spaces," Journal of the American Statistical Association,vol.102, pp.255–268, 2007.*

xvi. *M. Sugiyama, I. Takeuchi, T. Suzuki, T. Kanamori, H. Hachiya, and D. Okanohara, "Least-Squares Conditional Density Estimation," IEICE Transactions on Information and Systems, vol.E93-D, pp.583–594, 2010.*

xvii. *T. Akiyama, H.Hachiya, and M. Sugiyama, "Efficient exploration through active learning for value function approximation in reinforcement learning," Neural Networks, vol.23, No. 5, pp.639–648, 2010.*

xviii. *F. Wilcoxon, "Individual Comparisons by Ranking Methods," Biometrics Bulletin, vol.Vol. 1, No. 6, pp.80–83, 1945.*