NIPS 2008

Efficient Direct Density Ratio Estimation for Non-stationarity Adaptation and Outlier Detection

Takafumi Kanamori (Nagoya University)
Shohei Hido (IBM)
Masashi Sugiyama (Tokyo Institute of Technology)

Density Ratio

$$\left.\begin{array}{r} x_1, \ldots, x_n \sim_{i.i.d.} \; p(x) \\ x'_1, \ldots, x'_m \sim_{i.i.d.} \; q(x') \end{array}\right\} \overset{\text{estimate}}{\Longrightarrow} \quad w(x) = \frac{p(x)}{q(x)},$$

Density Ratio (Importance)

Density ratio can be used for various succeeding tasks:

Feature selection (Suzuki, et al., ECML workshop 2008), Multi-task learning (Bickel, et al., ICML 2008), Domain Adaptation (Storkey and Sugiyama, NIPS 2006, Tsuboi et al., SDM 2008), etc.

Our main applications: Covariate shift adaptation, Outlier detection

Table of Contents

Estimation of Density Ratio

Samples : $$x_1^{(\mathrm{tr})}, \ldots, x_n^{(\mathrm{tr})} \sim p^{(\mathrm{tr})}(x),$$
$$x_1^{(\mathrm{te})}, \ldots, x_m^{(\mathrm{te})} \sim p^{(\mathrm{te})}(x)$$

- $\dim x$ **is small** $\Longrightarrow$ **Naive methods are available, e.g. separately estimate** $\widehat{p}^{(\mathrm{tr})}, \widehat{p}^{(\mathrm{te})}$ **by kernel density estimator, and obtain** $\widehat{w}(x) = \widehat{p}^{(\mathrm{te})}/\widehat{p}^{(\mathrm{tr})}$**.**

- $\dim x$ **is not small** $\Longrightarrow$ **direct density ratio estimation**
  - **Kernel Mean Matching** **(Huang, et al., NIPS2006): mean matching in RKHS.**
  - **Logistic Regression** **(Bickel et al., ICML2008): binary classification approach**
  - **KLEIP** **(Sugiyama, et al., NIPS2007): inference using KL-divergence**
  - **Proposed Method: Least squares approach**
    - **Efficient computation of estimator and leave-one-out cross validation**

Least Squares Approach to Importance Estimation

Square error of density ratio:

$$\frac{1}{2}\int\left(w(x)-\frac{p^{(\mathrm{te})}(x)}{p^{(\mathrm{tr})}(x)}\right)^2 p^{(\mathrm{tr})}(x)dx$$

$$=\underline{\frac{1}{2}\int w(x)^2 p^{(\mathrm{tr})}(x)dx-\int w(x)p^{(\mathrm{te})}(x)dx}+(\textbf{const.})$$

note: In succeeding tasks, $w(x^{(\mathrm{tr})})$ is often used. Thus, the expectation with $p^{(\mathrm{tr})}$ is valid.

Estimator: $w(x)=\sum_{\ell=1}^{b}\alpha_\ell\varphi_\ell(x),\quad \varphi_\ell(x)=e^{-\gamma\|x-c_\ell\|^2}>0,\quad c_\ell$**: kernel center**

empirical loss: $\underline{\frac{1}{2n}\sum_{i=1}^{n}w(x_i^{(\mathrm{tr})})^2-\frac{1}{m}\sum_{j=1}^{m}w(x_j^{(\mathrm{te})})} = \frac{1}{2}\alpha^\top H\alpha-g^\top\alpha \longrightarrow \min_{\alpha}$

$$H_{\ell\ell'}:=\frac{1}{n}\sum_{i=1}^{n}\varphi_\ell(x_i^{(\mathrm{tr})})\varphi_{\ell'}(x_i^{(\mathrm{tr})}),\quad g_\ell:=\frac{1}{m}\sum_{j=1}^{m}\varphi_\ell(x_j^{(\mathrm{te})})$$

Non-Negativity Condition and Regularization

Impose the constraint $w(x) \geq 0$ to the estimator:

$$\text{(I)} \quad \min_{\alpha} \frac{1}{2}\alpha^\top H\alpha - g^\top\alpha + \lambda R(\alpha) \;\rightarrow\; \widetilde{\alpha}, \qquad \widehat{\alpha}_\ell = \max\{\widetilde{\alpha}_\ell,\ 0\}$$

$$\text{(II)} \quad \min_{\alpha} \frac{1}{2}\alpha^\top H\alpha - g^\top\alpha + \lambda R(\alpha),\ \text{s.t.}\ \alpha \geq 0 \;\rightarrow\; \widehat{\alpha}$$

- **Proposed estimator: (I) with L_2-regularization $R(\alpha) = \|\alpha\|_2^2$**
  called uLSIF (unconstrained Least-square Importance Inference)
  - **Estimator $\widehat{\alpha}$ is analytically computed : $\widehat{\alpha} = \max\{(H+\lambda I)^{-1}g, 0\}$**
  - **Leave-one-out cross validation is analytically computed.**

    $$\text{LOOCV} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2}(\widehat{w}^{(i)}(x_i^{(\text{tr})}))^2 - \widehat{w}^{(i)}(x_i^{(\text{te})})\right]$$ **: directly computed via $\widehat{\alpha}$.**

    $\widehat{w}^{(i)}$: estimator obtained without the samples $x_i^{(\text{tr})}, x_i^{(\text{te})}$ $(i = 1, \dots, n,\ n \leq m)$.

- **(II) + L_1-regularization $\Longrightarrow$ regularization path in term of λ is obtained.**
  Numerically rather unstable.

Relation between Proposed and Existing Methods

Methods	Density estimation	Optimization	Out-of-sample prediction (CV)
KDE	Necessary	Analytic	Possible
KMM	Not necessary	Convex quadratic	Not possible
LogReg	Not necessary	Convex non-linear	Possible
KLIEP	Not necessary	Convex non-linear	Possible
uLSIF	Not necessary	Analytic	Possible

- **KMM estimates not the function $w(x)$ but the values $w(x_i^{(\mathrm{tr})}), i = 1, \ldots, n$. Thus, cross validation will not be possible.**
- **Model selection of uLSIF is computationally much more efficient than LogReg and KLIEP, because analytic formula of leave-one-out cv is available.**

Numerical Examples: Estimation of Density Ratio

- **Estimate** $w(x) = p^{(\mathrm{te})}(x)/p^{(\mathrm{tr})}(x)$.

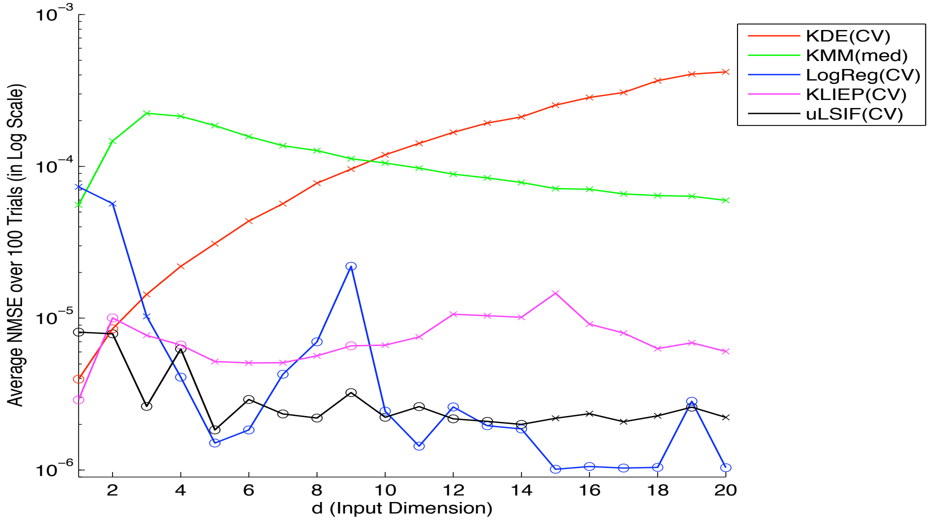
$$x_1^{(\mathrm{tr})}, \ldots, x_n^{(\mathrm{tr})} \sim N_d(\mathbf{0}, \ \boldsymbol{I}_d)$$
$$x_1^{(\mathrm{te})}, \ldots, x_{1000}^{(\mathrm{te})} \sim N_d(\boldsymbol{e}_1, \ \boldsymbol{I}_d), \quad \boldsymbol{e}_1 = (1, 0, \ldots, 0)^\top \in \mathbb{R}^d$$

  - **setup 1:** $d = 1, 2, \ldots, 20, \quad n = 100$
  - **setup 2:** $n = 50, 60, \ldots, 150, \quad d = 10$

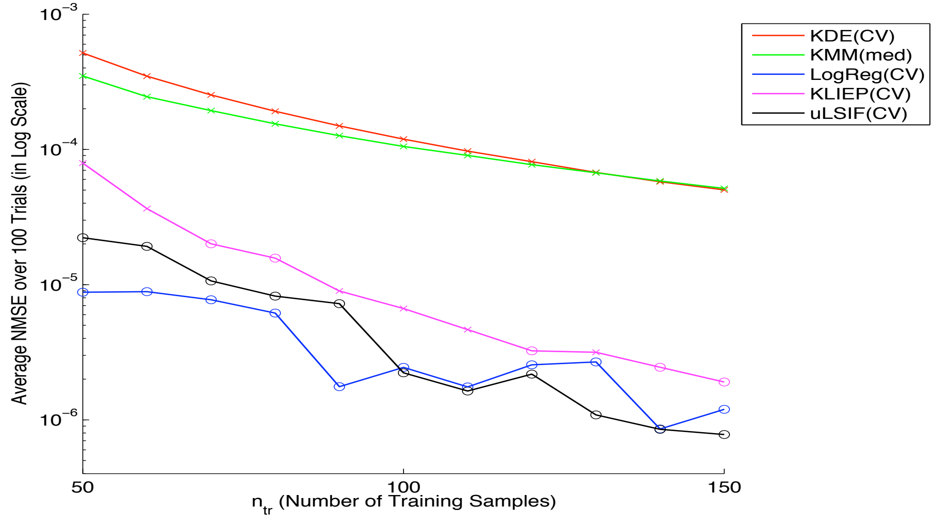
- **Evaluation of accuracy: square error on** $x_1^{(\mathrm{tr})}, \ldots, x_n^{(\mathrm{tr})}$

normalized-MSE (NMSE) $$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\widehat{w}(x_i^{(\mathrm{tr})})}{\sum_{j=1}^{n}\widehat{w}(x_j^{(\mathrm{tr})})} - \frac{w(x_i^{(\mathrm{tr})})}{\sum_{j=1}^{n} w(x_j^{(\mathrm{tr})})}\right)^2$$

Square Errors of Density Ration Estimation

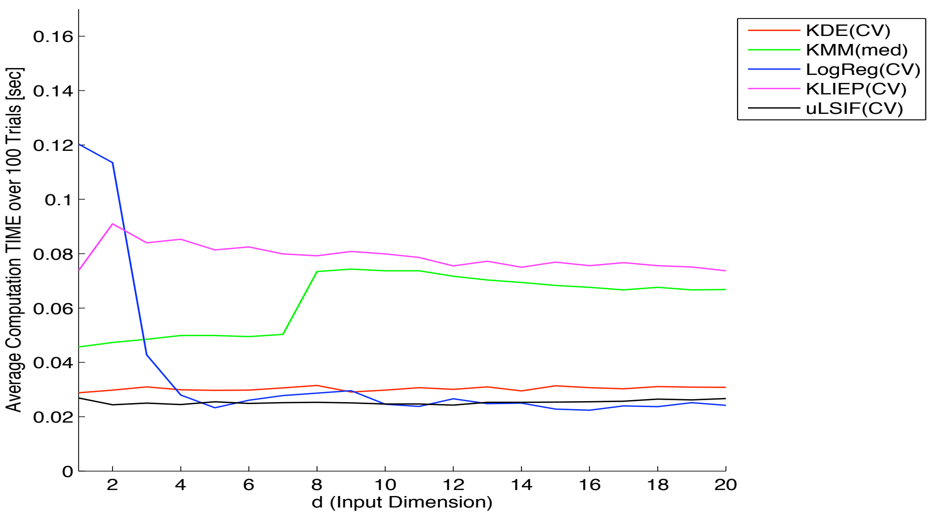


$\dim x$

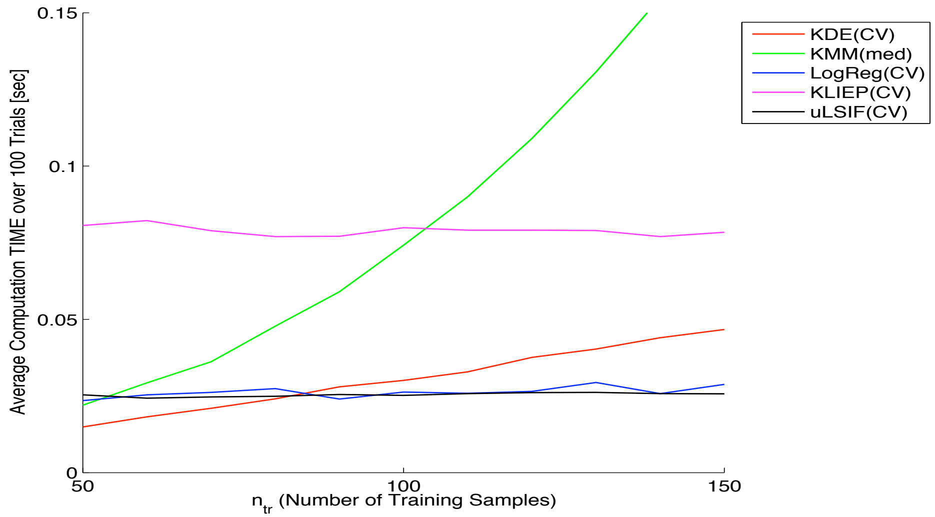


$\#samples$

Computation Time for Fixed Model Parameters



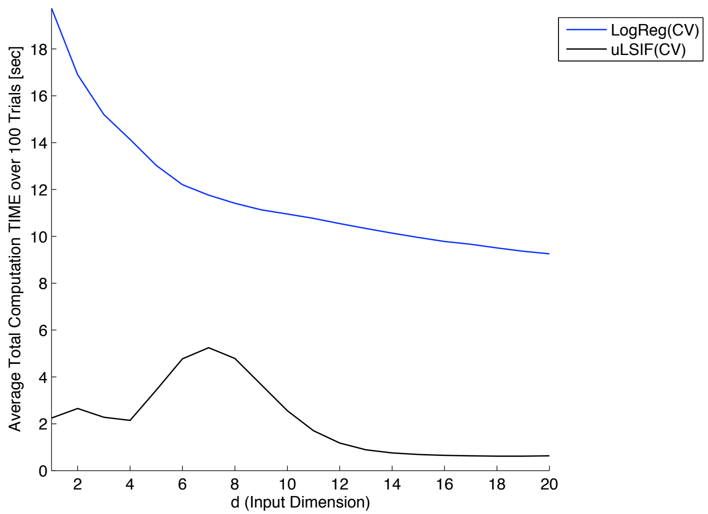
$\dim x$



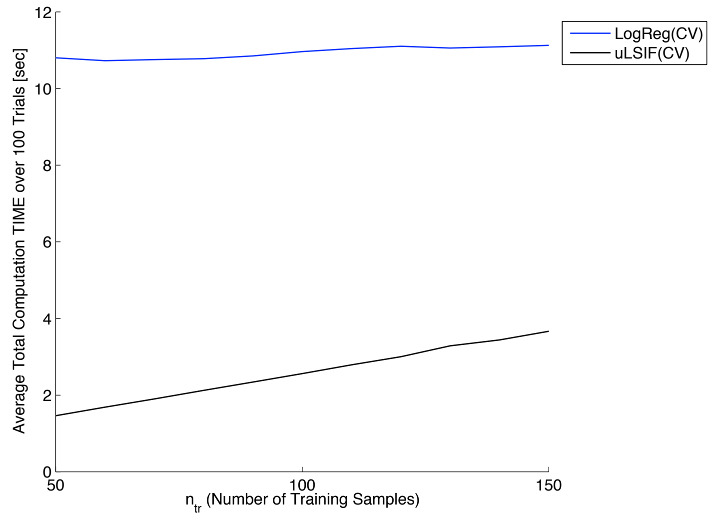
$\#samples$

- **Least-square error: uLSIF, Logistic reg., KLIEP $<$ KDE, KMM**
- **Computation time: uLSIF, Logistic reg. $<$ KDE $<$ KMM, KLIEP**

Computation time of uLSIF and Logistic reg. including cross-validation:



$\dim x$



$\#samples$

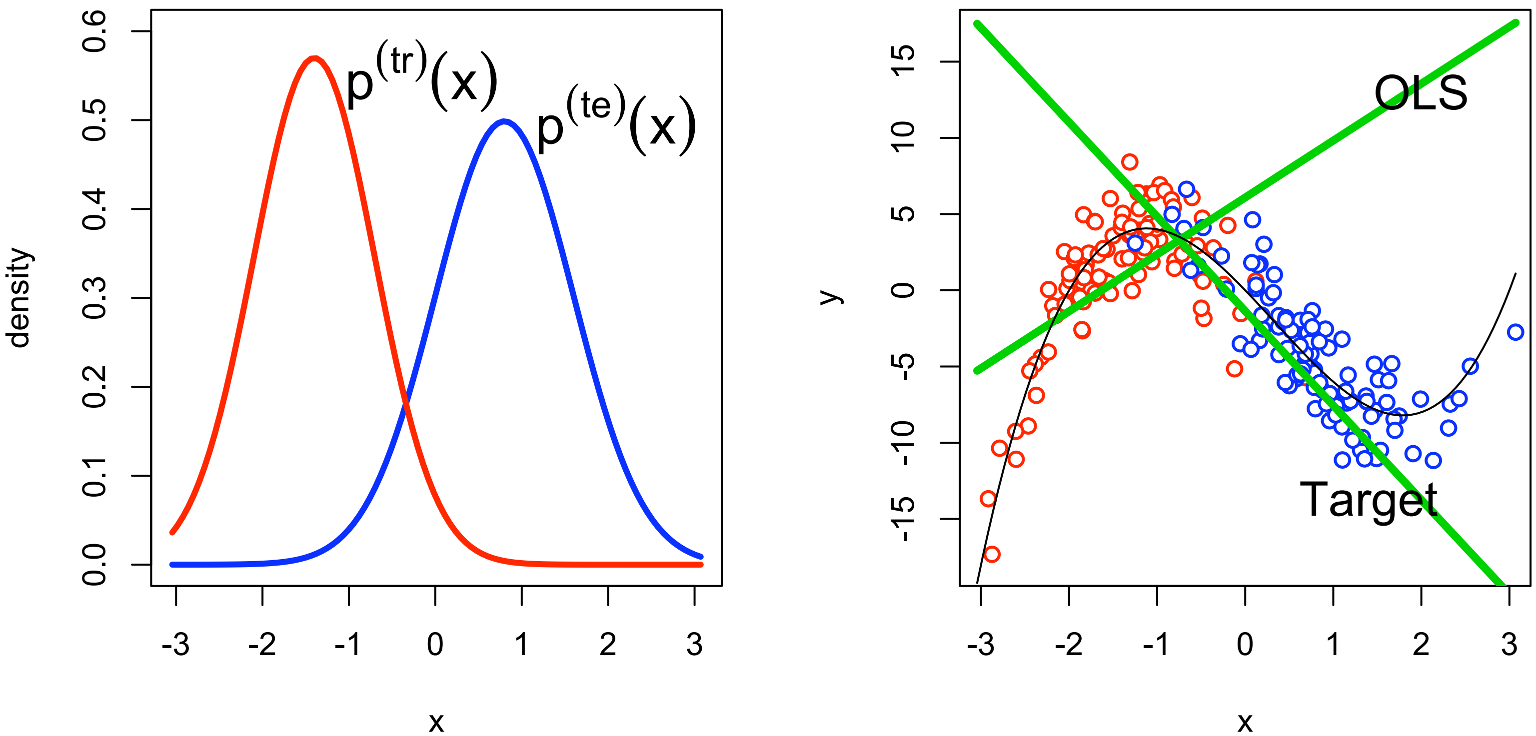
Application 1: Covariate Shift Adaptation

$$Y = f^*(X) + \varepsilon, \quad \textbf{training data } (x^{(\mathrm{tr})}, y^{(\mathrm{tr})}) \sim p(y|x)p^{(\mathrm{tr})}(x)$$

$$\textbf{test data } (x^{(\mathrm{te})}, y^{(\mathrm{te})}) \sim p(y|x)p^{(\mathrm{te})}(x)$$

Purpose: estimate $f^*(x) = E[Y|x]$ **based on training data**

Ordinary Least Squares (OLS) has bias if the model is misspecified.



Covariate Shift: Bias Correction using Density Ratio

training data: $\{(x_1^{(\mathrm{tr})}, y_1^{(\mathrm{tr})}), \ldots, (x_n^{(\mathrm{tr})}, y_n^{(\mathrm{tr})})\}, \ \{x_1^{(\mathrm{te})}, \ldots, x_m^{(\mathrm{te})}\}$

1. **Importance estimation:** $x^{(\mathrm{tr})}, x^{(\mathrm{te})} \overset{\text{estimate}}{\Longrightarrow} \widehat{w}(x) \cong \dfrac{p^{(\mathrm{te})}(x)}{p^{(\mathrm{tr})}(x)}$

2. **Weighted least-square estimation:**

$$\min_\theta \frac{1}{n} \sum_{i=1}^n \widehat{w}(x_i^{(\mathrm{tr})})(y_i^{(\mathrm{tr})} - f(x_i^{(\mathrm{tr})}; \theta))^2 + \gamma \|\theta\|_2^2, \quad f(x;\theta) = \sum_{\ell=1}^t \theta_\ell K_h(x, m_\ell).$$

(Hyper-parameters h, γ **are chosen by** *importance weighted CV.* **cf. Sugiyama et al., nips'07)**

Note:

$$\frac{1}{n} \sum_{i=1}^n \widehat{w}(x_i^{(\mathrm{tr})})(y_i^{(\mathrm{tr})} - f(x_i^{(\mathrm{tr})}))^2 \cong \int \frac{p^{(\mathrm{te})}(x)}{p^{(\mathrm{tr})}(x)} (y - f(x))^2 \cdot p(y|x) p^{(\mathrm{tr})}(x) dx$$

$$= \int (y - f(x))^2 \cdot p(y|x) p^{(\mathrm{te})}(x) dx$$

Minimization of weighted square error will provide an asymptotically unbiased estimator of $f^*(x) = E[Y|X]$ **under** $p^{(\mathrm{te})}(x)$**.**

Covariate Shift: Numerical Results

Data set $\{(x_i, y_i)\}_{i=1}^n$**:** $x_i = (x_i^{(1)}, \ldots, x_i^{(d)}) \in [0, 1]^d$ **(normalized).**
$\{(x_i, y_i)\}_{i=1}^n \overset{\text{random}}{\longrightarrow} (x_k, y_k)$**;accepted as** $(x^{(\text{te})}, y^{(\text{te})})$ **with probability** $\min\{1, (x_k^{(c)})^2\}$**.**
The coordinate c **is randomly determined and fixed in each trial.**
$(x^{(\text{tr})}, y^{(\text{tr})})$**: uniformly sampled from the rest.**

Data	Uniform	KDE (CV)	KMM (med)	LogReg (CV)	KLIEP (CV)	uLSIF (CV)
kin-8fh	1.00(0.34)	1.22(0.52)	1.55(0.39)	1.31(0.39)	**0.95(0.31)**	**1.02(0.33)**
kin-8fm	1.00(0.39)	1.12(0.57)	1.84(0.58)	1.38(0.57)	**0.86(0.35)**	**0.88(0.39)**
kin-8nh	**1.00(0.26)**	1.09(0.20)	1.19(0.29)	1.09(0.19)	**0.99(0.22)**	**1.02(0.18)**
kin-8nm	**1.00(0.30)**	1.14(0.26)	1.20(0.20)	1.12(0.21)	**0.97(0.25)**	1.04(0.25)
abalone	**1.00(0.50)**	1.02(0.41)	**0.91(0.38)**	**0.97(0.49)**	**0.94(0.67)**	**0.96(0.61)**
image	**1.00(0.51)**	0.98(0.45)	1.08(0.54)	**0.98(0.46)**	**0.94(0.44)**	**0.98(0.47)**
ringnorm	1.00(0.04)	0.87(0.04)	**0.87(0.04)**	0.95(0.08)	0.99(0.06)	0.91(0.08)
twonorm	1.00(0.58)	1.16(0.71)	**0.94(0.57)**	**0.91(0.61)**	**0.91(0.52)**	**0.88(0.57)**
waveform	1.00(0.45)	1.05(0.47)	0.98(0.31)	**0.93(0.32)**	**0.93(0.34)**	**0.92(0.32)**
Average	**1.00**	**1.07**	**1.17**	**1.07**	**0.94**	**0.96**
Comp. time	—	**0.82**	**3.50**	**3.27**	**2.23**	**1.00**

(average on 100 trials. Wilcoxon signed rank test at the significance level 1%**)**

Application 2: Outlier Detection

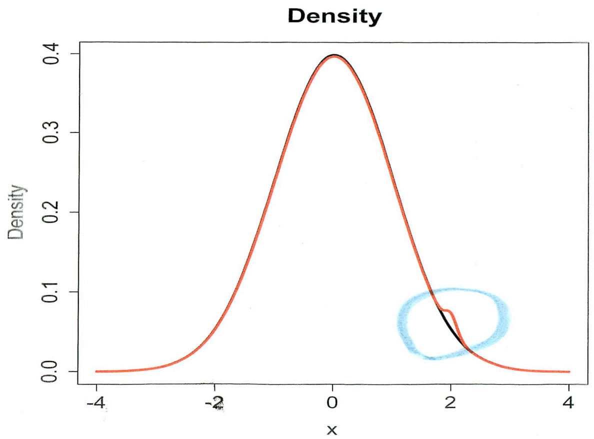
Identify irregular samples in an evaluation dataset.

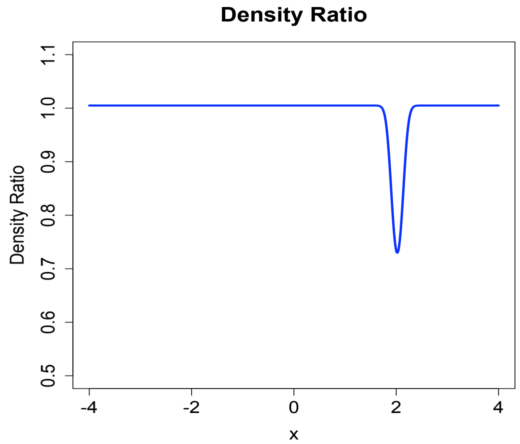
$$\begin{array}{rl} \textbf{Model dataset (no outliers)}: & x_1^{(\mathrm{te})}, \ldots, x_m^{(\mathrm{te})} \sim p^{(\mathrm{te})}(x) \\ \textbf{Evaluation dataset}: & x_1^{(\mathrm{tr})}, \ldots, x_n^{(\mathrm{tr})} \sim p^{(\mathrm{tr})}(x) \end{array} \quad \overset{\textbf{estimate}}{\Longrightarrow} \quad w(x) = \frac{p^{(\mathrm{te})}(x)}{p^{(\mathrm{tr})}(x)}$$

- **For almost all samples in evaluation data:** $w(x^{(\mathrm{tr})}) = \frac{p^{(\mathrm{te})}(x^{(\mathrm{tr})})}{p^{(\mathrm{tr})}(x^{(\mathrm{tr})})} \cong 1.$
- **On outlying samples in evaluation data:**

$$w(x^{(\mathrm{tr})}) = \frac{p^{(\mathrm{te})}(x^{(\mathrm{tr})})}{p^{(\mathrm{tr})}(x^{(\mathrm{tr})})} < 1, \qquad \left(p^{(\mathrm{te})}(x^{(\mathrm{tr})}) < p^{(\mathrm{tr})}(x^{(\mathrm{tr})})\right)$$

- $w(x)$ **can be used as a score of** *outlyingness* **comparing to model dataset.**





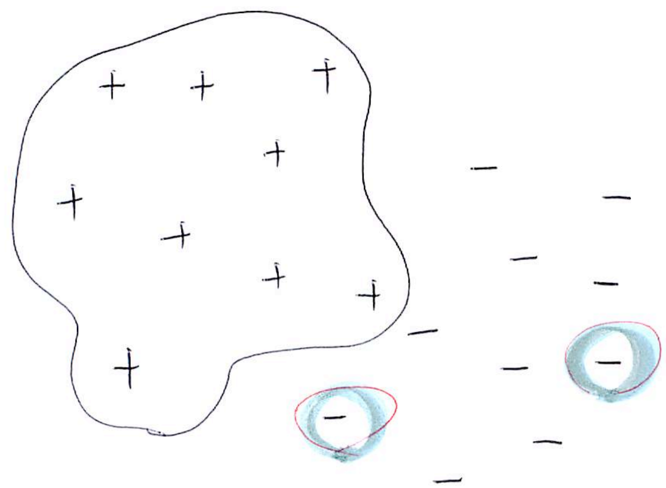
Applications: Intrusion detection in network systems, Topic detection in news documents.

Outlier Detection: Numerical Experiments

- **Benchmark datasets for binary classification problems**
  - **Model data: training samples with positive label.**
  - **Evaluation data: test samples with positive label**

    $+$ $\rho\,\%$ **negative labeled test samples** $(\rho = 1, 2, 5)$

- **Negative labeled samples are randomly chosen from test data set.**
- **On each dataset, results are averaged on 20 trials.**



Outlier Detection: Results on Benchmark Data

Evaluation: Area under the curve (AUC) of ROC curve: larger is better.

Data Name	ρ	uLSIF (CV)	KLIEP (CV)	LogReg (CV)	KMM (med)	OSVM (med)	LOF $k=5$	LOF $k=30$	LOF $k=50$	KDE' (CV)
banana	1	0.851	0.815	0.447	0.578	0.360	0.838	0.915	0.919	0.934
	2	0.858	0.824	0.428	0.644	0.412	0.813	0.918	0.920	0.927
	5	0.869	0.851	0.435	0.761	0.467	0.786	0.907	0.909	0.923
b-cancer	1	0.463	0.480	0.627	0.576	0.508	0.546	0.488	0.463	0.400
	2	0.463	0.480	0.627	0.576	0.506	0.521	0.445	0.428	0.400
	5	0.463	0.480	0.627	0.576	0.498	0.549	0.480	0.452	0.400
diabetes	1	0.558	0.615	0.599	0.574	0.563	0.513	0.403	0.390	0.425
	2	0.558	0.615	0.599	0.574	0.563	0.526	0.453	0.434	0.425
	5	0.532	0.590	0.636	0.547	0.545	0.536	0.461	0.447	0.435
f-solar	1	0.416	0.485	0.438	0.494	0.522	0.480	0.441	0.385	0.378
	2	0.426	0.456	0.432	0.480	0.550	0.442	0.406	0.343	0.374
	5	0.442	0.479	0.432	0.532	0.576	0.455	0.417	0.370	0.346
⋮				⋮				⋮		
other 8 datasets										
⋮				⋮				⋮		
Average		0.661	0.685	0.530	0.608	0.596	0.594	0.629	0.622	0.623
Comp. time		1.00	11.7	5.35	751	12.4	85.5			8.70

Conclusion

- **A new estimator for density ratio has been proposed**
  - **Analytic computation of estimator and LOOCV**
- **Applications: covariate-shift adaptation and outlier detection**

Future works

- **Explore various possible applications of density ratio:**
  - **feature selection, independent component analysis, dimensionality reduction, $\cdots\cdots$.**