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

\[ x_1, \ldots, x_n \sim i.i.d. \quad p(x) \quad \text{estimate} \quad w(x) = \frac{p(x)}{q(x)}, \]

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

1. Estimation of Density Ratio
   - Least Squares Approach
   - Computation of Leave-one-out Cross Validation
   - Numerical Results

2. Applications of Density Ratio
   - Covariate Shift Adaptation
   - Outlier Detection
Estimation of Density Ratio

\[
x_{1}^{(\text{tr})}, \ldots, x_{n}^{(\text{tr})} \sim p^{(\text{tr})}(x),
\]
\[
x_{1}^{(\text{te})}, \ldots, x_{m}^{(\text{te})} \sim p^{(\text{te})}(x)
\]

\textbf{dim} \ x \text{ is small} \implies \text{Naive methods are available, e.g. separately estimate } \hat{p}^{(\text{tr})}, \hat{p}^{(\text{te})} \text{ by kernel density estimator, and obtain } \hat{w}(x) = \hat{p}^{(\text{te})}/\hat{p}^{(\text{tr})}.

\textbf{dim} \ x \text{ is not small} \implies \text{direct density ratio estimation}

- \textbf{Kernel Mean Matching} (Huang, et al., NIPS2006): mean matching in RKHS.
- \textbf{Logistic Regression} (Bickel et al., ICML2008): binary classification approach
- \textbf{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:

\[
\frac{1}{2} \int \left( w(x) - \frac{p^{(te)}(x)}{p^{(tr)}(x)} \right)^2 p^{(tr)}(x) dx = \frac{1}{2} \int w(x)^2 p^{(tr)}(x) dx - \int w(x) p^{(te)}(x) dx + (\text{const.})
\]

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

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

Empirical loss:

\[
\frac{1}{2n} \sum_{i=1}^n w(x_i^{(tr)})^2 - \frac{1}{m} \sum_{j=1}^m w(x_j^{(te)}) = \frac{1}{2} \alpha^\top H \alpha - g^\top \alpha \rightarrow \min_{\alpha}
\]

\[
H_{\ell\ell'} := \frac{1}{n} \sum_{i=1}^n \varphi_\ell(x_i^{(tr)}) \varphi_{\ell'}(x_i^{(tr)}), \quad g_\ell := \frac{1}{m} \sum_{j=1}^m \varphi_\ell(x_j^{(te)})
\]
Non-Negativity Condition and Regularization

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

(I) \[ \min_{\alpha} \frac{1}{2} \alpha^\top H \alpha - g^\top \alpha + \lambda R(\alpha) \rightarrow \tilde{\alpha}, \quad \tilde{\alpha}_\ell = \max\{\tilde{\alpha}_\ell, 0\} \]

(II) \[ \min_{\alpha} \frac{1}{2} \alpha^\top H \alpha - g^\top \alpha + \lambda R(\alpha), \text{ s.t. } \alpha \geq 0 \rightarrow \hat{\alpha} \]

- Proposed estimator: (I) with \( L_2 \)-regularization \( R(\alpha) = \|\alpha\|^2 \) called uLSIF (unconstrained Least-square Importance Inference)
  - Estimator \( \hat{\alpha} \) is analytically computed: \( \hat{\alpha} = \max\{(H + \lambda I)^{-1} g, 0\} \)
  - Leave-one-out cross validation is analytically computed.
    \[ \text{LOOCV} = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{1}{2}(\hat{w}^{(i)}(x^{(tr)}_i))^2 - \hat{w}^{(i)}(x^{(te)}_i) \right] : \text{directly computed via } \hat{\alpha}. \]
    \( \hat{w}^{(i)} \): estimator obtained without the samples \( x^{(tr)}_i, x^{(te)}_i (i = 1, \ldots, n, n \leq m) \).

- (II) + \( L_1 \)-regularization \( \Rightarrow \) regularization path in term of \( \lambda \) is obtained. Numerically rather unstable.
<table>
<thead>
<tr>
<th>Methods</th>
<th>Density estimation</th>
<th>Optimization</th>
<th>Out-of-sample prediction (CV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KDE</td>
<td>Necessary</td>
<td>Analytic</td>
<td>Possible</td>
</tr>
<tr>
<td>KMM</td>
<td>Not necessary</td>
<td>Convex quadratic</td>
<td>Not possible</td>
</tr>
<tr>
<td>LogReg</td>
<td>Not necessary</td>
<td>Convex non-linear</td>
<td>Possible</td>
</tr>
<tr>
<td>KLIEP</td>
<td>Not necessary</td>
<td>Convex non-linear</td>
<td>Possible</td>
</tr>
<tr>
<td>uLSIF</td>
<td>Not necessary</td>
<td>Analytic</td>
<td>Possible</td>
</tr>
</tbody>
</table>

- KMM estimates not the function $w(x)$ but the values $w(x^{(tr)}_i), i = 1, \ldots, n$. 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 $w(x) = p^{(te)}(x)/p^{(tr)}(x)$.

$$x^{(tr)}_1, \ldots, x^{(tr)}_n \sim N_d(0, I_d)$$

$$x^{(te)}_1, \ldots, x^{(te)}_{1000} \sim N_d(e_1, I_d), \quad e_1 = (1, 0, \ldots, 0)\top \in \mathbb{R}^d$$

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

■ Evaluation of accuracy: square error on $x^{(tr)}_1, \ldots, x^{(tr)}_n$

normalized-MSE (NMSE) = $\frac{1}{n} \sum_{i=1}^{n} \left( \frac{\hat{w}(x^{(tr)}_i)}{\sum_{j=1}^{n} \hat{w}(x^{(tr)}_j)} - \frac{w(x^{(tr)}_i)}{\sum_{j=1}^{n} w(x^{(tr)}_j)} \right)^2$
Square Errors of Density Ration Estimation

Computation Time for Fixed Model Parameters
- 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:

![Graphs showing computation time over input dimension (dim x) and number of training samples (#samples).]
Application 1: Covariate Shift Adaptation

\[ Y = f^*(X) + \varepsilon, \quad \text{training data} \ (x^{(tr)}, y^{(tr)}) \sim p(y|x)p^{(tr)}(x) \]

\[ \text{test data} \ (x^{(te)}, y^{(te)}) \sim p(y|x)p^{(te)}(x) \]

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^{(tr)}, y_1^{(tr)}), \ldots, (x_n^{(tr)}, y_n^{(tr)}) \}, \{ x_1^{(te)}, \ldots, x_m^{(te)} \}

1. Importance estimation: \( x^{(tr)}, x^{(te)}, \) estimate \( \hat{w}(x) \approx \frac{p^{(te)}(x)}{p^{(tr)}(x)} \)

2. Weighted least-square estimation:

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

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

Note:

\[
\frac{1}{n} \sum_{i=1}^{n} \hat{w}(x_i^{(tr)})(y_i^{(tr)} - f(x_i^{(tr)}))^2 \approx \int \frac{p^{(te)}(x)}{p^{(tr)}(x)} (y - f(x))^2 \cdot p(y|x)p^{(tr)}(x)dx
\]

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

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

Data set \{ (x_i, y_i) \}_{i=1}^{n} \colon x_i = (x_i^{(1)}, \ldots, x_i^{(d)}) \in [0, 1]^d \text{ (normalized)}.

\{ (x_i, y_i) \}_{i=1}^{n} \xrightarrow{\text{random}} (x_k, y_k) \text{; accepted as } (x^{(te)}, y^{(te)}) \text{ with probability } \min\{1, (x_k^{(c)})^2\}.

The coordinate \( c \) is randomly determined and fixed in each trial.

\((x^{(tr)}, y^{(tr)})\): uniformly sampled from the rest.

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

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.

Model dataset (no outliers): \( x_1^{(te)}, \ldots, x_m^{(te)} \sim p^{(te)}(x) \)

Evaluation dataset: \( x_1^{(tr)}, \ldots, x_n^{(tr)} \sim p^{(tr)}(x) \)

\[
\text{estimate} \quad w(x) = \frac{p^{(te)}(x)}{p^{(tr)}(x)}
\]

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

\[
w(x^{(tr)}) = \frac{p^{(te)}(x^{(tr)})}{p^{(tr)}(x^{(tr)})} < 1, \quad (p^{(te)}(x^{(tr)}) < p^{(tr)}(x^{(tr)}))
\]

- \( w(x) \) 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
    \[ + \rho \% \text{ negative labeled test samples} \quad (\rho = 1, 2, 5) \]

- 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.

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