Constrained Least-Squares Density-Difference Estimation

Tuan Duong Nguyen
Tokyo Institute of Technology
nguyen@sg.cs.titech.ac.jp

Marthinus Christoffel du Plessis
Tokyo Institute of Technology
christo@sg.cs.titech.ac.jp

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

Masashi Sugiyama
Tokyo Institute of Technology
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/~sugi

Abstract

We address the problem of estimating the *difference* between two probability densities. A naive approach is a two-step procedure that first estimates two densities separately and then computes their difference. However, such a two-step procedure does not necessarily work well because the first step is performed without regard to the second step and thus a small error in the first stage can cause a big error in the second stage. Recently, a single-shot method called the *least-squares density-difference* (LSDD) estimator has been proposed. LSDD directly estimates the density difference without separately estimating two densities, and it was demonstrated to outperform the two-step approach. In this paper, we propose a variation of LSDD called the *constrained least-squares density-difference* (CLSDD) estimator, and theoretically prove that CLSDD improves the accuracy of density difference estimation for correctly specified parametric models. The usefulness of the proposed method is also demonstrated experimentally.

Keywords

density difference, asymptotic variance, L^2-distance, bias, class-balance change, two-sample homogeneity test.

1 Introduction

We address the problem of estimating the *difference* between two probability densities. A density-difference estimator is useful in solving various machine learning tasks such as class-balance estimation under class-prior change [6], image segmentation and registration [13, 2], target object detection and recognition [9, 22], feature selection and extraction [20, 19], and change-point detection in time series [11, 14, 21].

A naive approach to density-difference estimation is a two-step procedure of first estimating two densities separately and then computing their difference. However, such a two-step procedure does not necessarily work well because the first step is performed without regard to the second step and thus a small error incurred in the first stage can cause a big error in the second stage. Recently, a single-shot method called the *least-squares density-difference* (LSDD) estimator has been proposed [18]. LSDD directly estimates the density difference without separately estimating two densities, and it was demonstrated to outperform the two-step approach.

In this paper, we propose a variation of LSDD called the *constrained least-squares density-difference* (CLSDD) estimator, and theoretically prove that CLSDD improves the accuracy of density difference estimation for correctly specified parametric models. The usefulness of the proposed method is also demonstrated experimentally.

The remainder of this paper is structured as follows. In Section 2, we formulate the problem of density-difference estimation and review the original LSDD method. In Section 3, we describe our proposed CLSDD method, and theoretically prove its superiority for correctly specified parametric models. In Section 4, we apply CLSDD to approximating the L^2-distance between two probability densities and theoretically prove its superiority for correctly specified parametric models. In Section 5, the usefulness of the proposed CLSDD method is demonstrated experimentally. Finally, we conclude in Section 6.

2 Density-Difference Estimation

In this section, we formulate the problem of density-difference estimation and review the original LSDD method [18].

2.1 Problem Formulation

Suppose that we are given two sets of independent and identically distributed samples $\mathcal{X} := \{\boldsymbol{x}_i\}_{i=1}^n$ and $\mathcal{X}' := \{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ drawn from probability distributions on $\mathbb{R}^d$ with densities $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$, respectively:

$$\mathcal{X} := \{\boldsymbol{x}_i\}_{i=1}^n \overset{i.i.d}{\sim} p(\boldsymbol{x}),$$
$$\mathcal{X}' := \{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'} \overset{i.i.d}{\sim} p'(\boldsymbol{x}).$$

Our goal is to estimate the difference $f(\boldsymbol{x})$ between $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ from the samples $\mathcal{X}$ and $\mathcal{X}'$:

$$f(\boldsymbol{x}) := p(\boldsymbol{x}) - p'(\boldsymbol{x}).$$

2.2 Least-Squares Density-Difference Estimation

In LSDD, a density-difference model $g(\boldsymbol{x})$ is fitted to the true density-difference function $f(\boldsymbol{x})$ under the squared loss:

$$\underset{g}{\operatorname{argmin}} \int \Big(g(\boldsymbol{x}) - f(\boldsymbol{x})\Big)^2 \mathrm{d}\boldsymbol{x}.$$

We use the following linear-in-parameter model as $g(\boldsymbol{x})$:

$$g(\boldsymbol{x}) = \sum_{l=1}^{b} \theta_l \psi_l(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{\psi}(\boldsymbol{x}), \tag{1}$$

where b denotes the number of basis functions,

$$\boldsymbol{\psi}(\boldsymbol{x}) = (\psi_1(\boldsymbol{x}), \ldots, \psi_b(\boldsymbol{x}))^\top$$

is a b-dimensional linearly independent basis function vector,

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_b)^\top$$

is a b-dimensional parameter vector, and $^\top$ denotes the transpose. In practice, we use the following Gaussian kernel model as $g(\boldsymbol{x})$:

$$g(\boldsymbol{x}) = \sum_{l=1}^{n+n'} \theta_l \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}_l\|^2}{2\sigma^2}\right), \tag{2}$$

where

$$(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n, \boldsymbol{c}_{n+1}, \ldots, \boldsymbol{c}_{n+n'}) := (\boldsymbol{x}_1, \ldots \boldsymbol{x}_n, \boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{n'})$$

are Gaussian kernel centers. If $n + n'$ is large, we may only use a subset of $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n, \boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{n'}$ as Gaussian kernel centers.

For the model (1), the optimal parameter $\boldsymbol{\theta}^*$ is given by

$$\begin{aligned}
\boldsymbol{\theta}^* &:= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \int \Big(g(\boldsymbol{x}) - f(\boldsymbol{x})\Big)^2 \mathrm{d}\boldsymbol{x} \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[\int g(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} - 2\int g(\boldsymbol{x}) f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[\boldsymbol{\theta}^\top \boldsymbol{H} \boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \boldsymbol{h}\right] \\
&= \boldsymbol{H}^{-1} \boldsymbol{h},
\end{aligned}$$

where $\boldsymbol{H}$ is the $b \times b$ matrix and $\boldsymbol{h}$ is the b-dimensional vector defined as

$$\boldsymbol{H} := \int \boldsymbol{\psi}(\boldsymbol{x})\boldsymbol{\psi}(\boldsymbol{x})^\top \mathrm{d}\boldsymbol{x},$$
$$\boldsymbol{h} := \int \boldsymbol{\psi}(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int \boldsymbol{\psi}(\boldsymbol{x}')p'(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}'.$$

Note that, for the Gaussian kernel model (2), the integral in $\boldsymbol{H}$ can be computed analytically as

$$\begin{aligned} H_{l,l'} &= \int \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{c}_l\|^2}{2\sigma^2}\right)\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{c}_{l'}\|^2}{2\sigma^2}\right)\mathrm{d}\boldsymbol{x} \\ &= (\pi\sigma^2)^{d/2}\exp\left(-\frac{\|\boldsymbol{c}_l-\boldsymbol{c}_{l'}\|^2}{4\sigma^2}\right), \end{aligned}$$

where d denotes the dimensionality of $\boldsymbol{x}$. Replacing the expectations in $\boldsymbol{h}$ by empirical estimators and adding an ℓ_2-regularizer to the objective function, we arrive at the following optimization problem:

$$\widehat{\boldsymbol{\theta}} := \operatorname*{argmin}_{\boldsymbol{\theta}} \left[\boldsymbol{\theta}^\top \boldsymbol{H}\boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \widehat{\boldsymbol{h}} + \lambda\boldsymbol{\theta}^\top\boldsymbol{\theta}\right],$$

where λ (> 0) is the regularization parameter and $\widehat{\boldsymbol{h}}$ is the b-dimensional vector defined as

$$\widehat{\boldsymbol{h}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(\boldsymbol{x}_i) - \frac{1}{n'}\sum_{i'=1}^{n'}\boldsymbol{\psi}(\boldsymbol{x}'_{i'}).$$

Taking the derivative of the above objective function and set it to zero, we can obtain the solution $\widehat{\boldsymbol{\theta}}$ analytically as

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{H}_\lambda^{-1}\widehat{\boldsymbol{h}},$$

where

$$\boldsymbol{H}_\lambda = \boldsymbol{H} + \lambda\boldsymbol{I}_b.$$

$\boldsymbol{I}_b$ denotes the b-dimensional identity matrix. Finally, a density-difference estimator $\widehat{f}(\boldsymbol{x})$ is given as

$$\widehat{f}(\boldsymbol{x}) = \widehat{\boldsymbol{\theta}}^\top\boldsymbol{\psi}(\boldsymbol{x}).$$

This is called the *least-squares density-difference* (LSDD) estimator.

The LSDD estimator possesses superior theoretical properties: It has asymptotic normality with asymptotic order $\sqrt{1/n + 1/n'}$, which is known to be the optimal convergence

rate in the parametric setup. Also, a non-parametric LSDD estimator achieves the optimal convergence rate [18].

The practical performance of LSDD depends on the choice of models (i.e., the kernel width σ and the regularization parameter λ). The model can be optimized by *cross-validation* (CV). More specifically, we first divide the samples $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$ and $\mathcal{X}' = \{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ into T disjoint subsets $\{\mathcal{X}_t\}_{t=1}^{T}$ and $\{\mathcal{X}'_t\}_{t=1}^{T}$, respectively. Then we obtain a density-difference estimate $\widehat{f}_t(\boldsymbol{x})$ from $\mathcal{X} \setminus \mathcal{X}_t$ and $\mathcal{X}' \setminus \mathcal{X}'_t$ (i.e., all samples without $\mathcal{X}_t$ and $\mathcal{X}'_t$), and compute its hold-out error for $\mathcal{X}_t$ and $\mathcal{X}'_t$ as

$$\mathrm{CV}_t := \int \widehat{f}_t(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} - \frac{2}{|\mathcal{X}_t|} \sum_{\boldsymbol{x} \in \mathcal{X}_t} \widehat{f}_t(\boldsymbol{x}) + \frac{2}{|\mathcal{X}'_t|} \sum_{\boldsymbol{x}' \in \mathcal{X}'_t} \widehat{f}_t(\boldsymbol{x}'),$$

where $|\mathcal{X}|$ denotes the number of elements in the set $\mathcal{X}$. We repeat this hold-out validation procedure for $t = 1, \ldots, T$, and compute the average hold-out error as

$$\mathrm{CV} := \frac{1}{T} \sum_{t=1}^{T} \mathrm{CV}_t.$$

Finally, we choose the model that minimizes CV.

3 Constrained Least-Squares Density-Difference Estimation

In this section, we propose a variation of LSDD called the *constrained least-squares density-difference* (CLSDD) estimator, and theoretically prove that CLSDD improves the accuracy of density difference estimation for correctly specified parametric models.

3.1 Derivation of CLSDD

By definition, the true density-difference function $f(\boldsymbol{x})$ satisfies

$$\int f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int p'(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1 - 1 = 0.$$

Thus, it would be preferable that our model $g(\boldsymbol{x})$ satisfies the following constraint:

$$\int g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0.$$

For the linear-in-parameter model (1), this constraint is expressed as

$$\boldsymbol{\theta}^\top \overline{\boldsymbol{\psi}} = 0,$$

where $\overline{\boldsymbol{\psi}}$ is the b-dimensional vector defined by

$$\overline{\boldsymbol{\psi}} := \int \boldsymbol{\psi}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

For the Gaussian kernel model (2), this constraint can be further simplified as

$$\boldsymbol{\theta}^\top \mathbf{1}_b = 0,$$

where $\mathbf{1}_b$ denotes the b-dimensional vector with all ones, because

$$\int \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{c}_l\|^2}{2\sigma^2}\right)\mathrm{d}\boldsymbol{x} = (2\pi\sigma^2)^{d/2}.$$

Based on the above constraint, the CLSDD optimization problem is defined as

$$\begin{aligned}\widetilde{\boldsymbol{\theta}} := \ &\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \ \left[\boldsymbol{\theta}^\top \boldsymbol{H}\boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \widehat{\boldsymbol{h}} + \lambda\boldsymbol{\theta}^\top\boldsymbol{\theta}\right] \\ &\text{subject to } \boldsymbol{\theta}^\top\overline{\boldsymbol{\psi}} = 0.\end{aligned}$$

The *Karush-Kuhn-Tucker* (KKT) optimality conditions [4] for the CLSDD optimization problem are given as

$$\begin{bmatrix} 2\boldsymbol{H}_\lambda & \overline{\boldsymbol{\psi}} \\ \overline{\boldsymbol{\psi}}^\top & 0 \end{bmatrix}\begin{bmatrix} \widetilde{\boldsymbol{\theta}} \\ \nu \end{bmatrix} = \begin{bmatrix} 2\widehat{\boldsymbol{h}} \\ 0 \end{bmatrix},$$

where $\nu \in \mathbb{R}$ is a Lagrangian multiplier. Using a *block-matrix inversion formula* [1], we have

$$\begin{bmatrix} 2\boldsymbol{H}_\lambda & \overline{\boldsymbol{\psi}} \\ \overline{\boldsymbol{\psi}}^\top & 0 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{2}\boldsymbol{H}_\lambda^{-1} - \frac{1}{2}\boldsymbol{D}_\lambda & \frac{\boldsymbol{H}_\lambda^{-1}\overline{\boldsymbol{\psi}}}{\overline{\boldsymbol{\psi}}^\top\boldsymbol{H}_\lambda^{-1}\overline{\boldsymbol{\psi}}} \\ \frac{\overline{\boldsymbol{\psi}}^\top\boldsymbol{H}_\lambda^{-1}}{\overline{\boldsymbol{\psi}}^\top\boldsymbol{H}_\lambda^{-1}\overline{\boldsymbol{\psi}}} & -\frac{1}{2}\overline{\boldsymbol{\psi}}^\top\boldsymbol{H}_\lambda^{-1}\overline{\boldsymbol{\psi}} \end{bmatrix},$$

where

$$\boldsymbol{D}_\lambda := \frac{\boldsymbol{H}_\lambda^{-1}\overline{\boldsymbol{\psi}\boldsymbol{\psi}}^\top\boldsymbol{H}_\lambda^{-1}}{\overline{\boldsymbol{\psi}}^\top\boldsymbol{H}_\lambda^{-1}\overline{\boldsymbol{\psi}}}.$$

Then, the CLSDD solution $\widetilde{\boldsymbol{\theta}}$ is given as

$$\widetilde{\boldsymbol{\theta}} = (\boldsymbol{H}_\lambda^{-1} - \boldsymbol{D}_\lambda)\widehat{\boldsymbol{h}} = \widehat{\boldsymbol{\theta}} - \boldsymbol{D}_\lambda\widehat{\boldsymbol{h}}.$$

Finally, a density-difference estimator $\widetilde{f}(\boldsymbol{x})$ is given as

$$\widetilde{f}(\boldsymbol{x}) = \widetilde{\boldsymbol{\theta}}^\top\boldsymbol{\psi}(\boldsymbol{x}).$$

3.2 Parametric Convergence Analysis of CLSDD

We consider a linear parametric setup where basis functions in our density-difference model (1) are fixed and correctly specified, i.e., there exists $\boldsymbol{\theta}^* \in \mathbb{R}^b$ such that

$$f(\boldsymbol{x}) = \boldsymbol{\theta}^{*\top} \boldsymbol{\psi}(\boldsymbol{x}).$$

Let $\boldsymbol{V}_p$ be the variance-covariance matrix of $\boldsymbol{\psi}(\boldsymbol{x})$ under the probability density $p(\boldsymbol{x})$:

$$\boldsymbol{V}_p := \int \Big(\boldsymbol{\psi}(\boldsymbol{x}) - \overline{\boldsymbol{\psi}}_p\Big)\Big(\boldsymbol{\psi}(\boldsymbol{x}) - \overline{\boldsymbol{\psi}}_p\Big)^\top p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$

where $\overline{\boldsymbol{\psi}}_p$ denotes the expectation of $\boldsymbol{\psi}(\boldsymbol{x})$ under the probability density $p(\boldsymbol{x})$:

$$\overline{\boldsymbol{\psi}}_p := \int \boldsymbol{\psi}(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Suppose that $\frac{n}{n+n'}$ converges to $\eta \in [0,1]$, and let $\lambda = o(\sqrt{1/n + 1/n'})$. Then the *central limit theorem* [15] asserts that $\sqrt{\frac{nn'}{n+n'}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ converges in law to the normal distribution with mean $\boldsymbol{0}_b$ and variance-covariance matrix

$$(\boldsymbol{H}^{-1} - \boldsymbol{D})\boldsymbol{B}_\eta(\boldsymbol{H}^{-1} - \boldsymbol{D}),$$

where $\boldsymbol{0}_b$ denotes the b-dimensional vector with all zeros and

$$\boldsymbol{D} := \frac{\boldsymbol{H}^{-1}\overline{\boldsymbol{\psi}\boldsymbol{\psi}}^\top \boldsymbol{H}^{-1}}{\overline{\boldsymbol{\psi}}^\top \boldsymbol{H}^{-1}\overline{\boldsymbol{\psi}}},$$
$$\boldsymbol{B}_\eta := (1-\eta)\boldsymbol{V}_p + \eta \boldsymbol{V}_{p'}.$$

Note that $\boldsymbol{H}^{-1} - \boldsymbol{D}$ is positive semi-definite because

$$\boldsymbol{H}^{-1} - \boldsymbol{D} = \boldsymbol{H}^{-1/2}\boldsymbol{P}\boldsymbol{H}^{-1/2},$$

where $\boldsymbol{P}$ denotes the orthogonal projection matrix onto the orthogonal complement of $\boldsymbol{H}^{-1/2}\overline{\boldsymbol{\psi}}$. This result implies that the CLSDD estimator has asymptotic normality with asymptotic order $\sqrt{1/n + 1/n'}$, which is the optimal convergence rate in the parametric setup.

In the same setup, the original LSDD estimator $\widehat{\boldsymbol{\theta}}$ possesses asymptotic normality with variance-covariance matrix $\boldsymbol{H}^{-1}\boldsymbol{B}_\eta\boldsymbol{H}^{-1}$ [18]. Because of the positive semi-definiteness of $\boldsymbol{H}^{-1} - \boldsymbol{D}$, we can confirm that CLSDD generally possesses smaller asymptotic variance than LSDD. This immediately implies

$$\mathbb{E}\left\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|^2 \leq \mathbb{E}\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|^2,$$

if terms with $o_p(\sqrt{1/n + 1/n'})$ coming from the central limit theorem are ignored, where $\mathbb{E}$ denotes the expectation over all samples and o_p denotes the asymptotic order in probability, Thus, CLSDD is provably more accurate than LSDD for correctly specified models.

4 L^2-Distance Estimation by CLSDD

In this section, we consider the problem of approximating the L^2-distance between $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$,

$$L^2(p, p') := \int \Big(p(\boldsymbol{x}) - p'(\boldsymbol{x})\Big)^2 \mathrm{d}\boldsymbol{x}, \tag{3}$$

from samples $\mathcal{X} := \{\boldsymbol{x}_i\}_{i=1}^n$ and $\mathcal{X}' := \{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$.

Let us consider the following equivalent expression of Eq.(3):

$$L^2(p, p') = \int f(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int f(\boldsymbol{x}') p'(\boldsymbol{x}') \mathrm{d}\boldsymbol{x}'.$$

If we replace $f(\boldsymbol{x})$ with the CLSDD estimator $\widetilde{f}(\boldsymbol{x})$ and approximate the expectations by empirical averages, we obtain the following L^2-distance estimator:

$$\widetilde{L}^2(\mathcal{X}, \mathcal{X}') := \widetilde{\boldsymbol{\theta}}^\top \widehat{\boldsymbol{h}} = \widehat{\boldsymbol{h}}^\top (\boldsymbol{H}_\lambda^{-1} - \boldsymbol{D}_\lambda) \widehat{\boldsymbol{h}}. \tag{4}$$

On the other hand, its LSDD counterpart is given by

$$\widehat{L}^2(\mathcal{X}, \mathcal{X}') := \widehat{\boldsymbol{\theta}}^\top \widehat{\boldsymbol{h}} = \widehat{\boldsymbol{h}}^\top \boldsymbol{H}_\lambda^{-1} \widehat{\boldsymbol{h}}.$$

Suppose that our linear-in-parameter model (1) is correctly specified, i.e., there exist $\boldsymbol{\theta}^* \in \mathbb{R}^b$ such that

$$f(\boldsymbol{x}) = \boldsymbol{\theta}^{*\top} \boldsymbol{\psi}(\boldsymbol{x}).$$

Then the true L^2-distance is expressed as

$$L^2(p, p') = \boldsymbol{h}^\top (\boldsymbol{H}^{-1} - \boldsymbol{D}) \boldsymbol{h} = \boldsymbol{h}^\top \boldsymbol{H}^{-1} \boldsymbol{h},$$

where the second equality follows from

$$\int f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \boldsymbol{\theta}^{*\top} \overline{\boldsymbol{\psi}} = 0.$$

If the regularization parameter is set at $\lambda = o(1/n + 1/n')$ and terms with $o_p(1/n + 1/n')$ are ignored, we have

$$\begin{aligned}
\mathbb{E}[\widetilde{L}^2(\mathcal{X}, \mathcal{X}')] &= L^2(p, p') + \mathrm{tr}\left((\boldsymbol{H}^{-1} - \boldsymbol{D})\boldsymbol{C}\right), \\
\mathbb{E}[\widehat{L}^2(\mathcal{X}, \mathcal{X}')] &= L^2(p, p') + \mathrm{tr}\left(\boldsymbol{H}^{-1}\boldsymbol{C}\right),
\end{aligned}$$

where

$$\boldsymbol{C} := \frac{1}{n}\boldsymbol{V}_p + \frac{1}{n'}\boldsymbol{V}_{p'}.$$

From the positive semi-definiteness of $\boldsymbol{H}^{-1} - \boldsymbol{D}$, we have

$$\mathrm{tr}\left((\boldsymbol{H}^{-1} - \boldsymbol{D})\boldsymbol{C}\right) \leq \mathrm{tr}\left(\boldsymbol{H}^{-1}\boldsymbol{C}\right).$$

This means that, for correctly specified models, the L^2-distance estimator $\widetilde{L}^2(\mathcal{X}, \mathcal{X}')$ based on CLSDD possesses smaller bias than its LSDD counterpart $\widehat{L}^2(\mathcal{X}, \mathcal{X}')$.

Note that, in the original LSDD paper [18], a slightly more sophisticated L^2-distance estimator was proposed:

$$2\widehat{\boldsymbol{\theta}}^\top \widehat{\boldsymbol{h}} - \widehat{\boldsymbol{\theta}}^\top \boldsymbol{H} \widehat{\boldsymbol{\theta}}, \tag{5}$$

which was shown to possess smaller bias than the naive approximator $\widehat{\boldsymbol{\theta}}^\top \widehat{\boldsymbol{h}}$. However, its improvement is of order $\mathcal{O}(\lambda)$, which is ignorable in the current context.

5 Experiments

In this section, we experimentally evaluate the performance of CLSDD and LSDD. We focus on using the Gaussian kernel model (2) here.

5.1 Illustration

First, we numerically illustrate the behavior of CLSDD and LSDD using one-dimensional artificial data. Let

$$p(x) = \mathcal{N}(x; \mu_1, \sigma_1^2),$$
$$p'(x) = \mathcal{N}(x; \mu_2, \sigma_2^2),$$

where $\mathcal{N}(x; \mu, \sigma^2)$ denotes the normal density with mean μ and variance σ^2 with respect to x. We draw $n = n' = 200$ samples from $p(x)$ and $p'(x)$.

Figure 1 depicts the density-difference estimation results for $\mu_1 = 0$, $\mu_2 = 1.4$, and $\sigma_1 = \sigma_2 = 1$. This shows that CLSDD gives a slightly better estimate of the density difference $f(x)$ than LSDD.

Next, we investigate the squared difference between true and estimated density differences. Figure 2 depicts means and standard errors of squared differences between the true density difference and LSDD/CLSDD estimators as functions of

$$\mu_2 = -1, -0.9, -0.8, \ldots, 1$$

over 1000 runs. The results show that CLSDD tends to be more accurate than LSDD.

Now, we compare the CLSDD-based L^2-distance estimator given by Eq.(4) with the LSDD-based estimator given by Eq.(5). Figure 3 depicts means and standard errors of estimated L^2-distances over 1000 runs for $\mu_1 = 0$, $\sigma_1 = 3$, and $\sigma_2 = 5$ as functions of mean

$$\mu_2 = -1, -0.9, -0.8, \ldots, 1.$$

This shows that CLSDD tends to outperform LSDD.

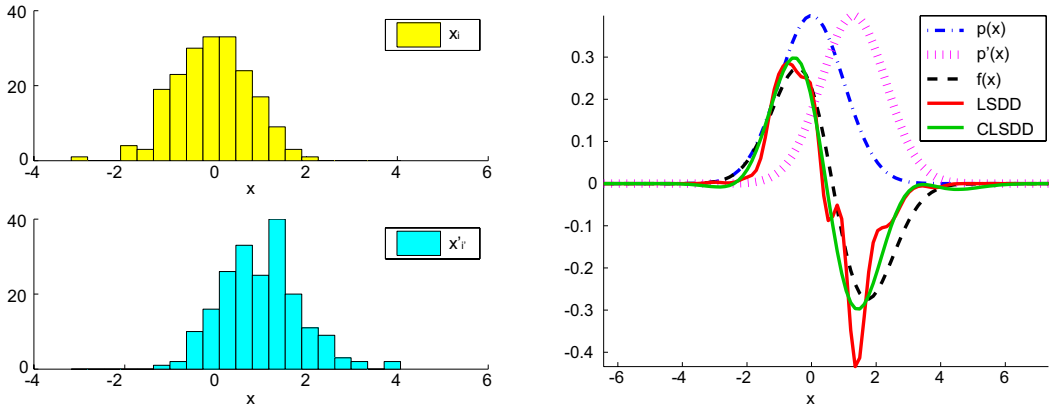
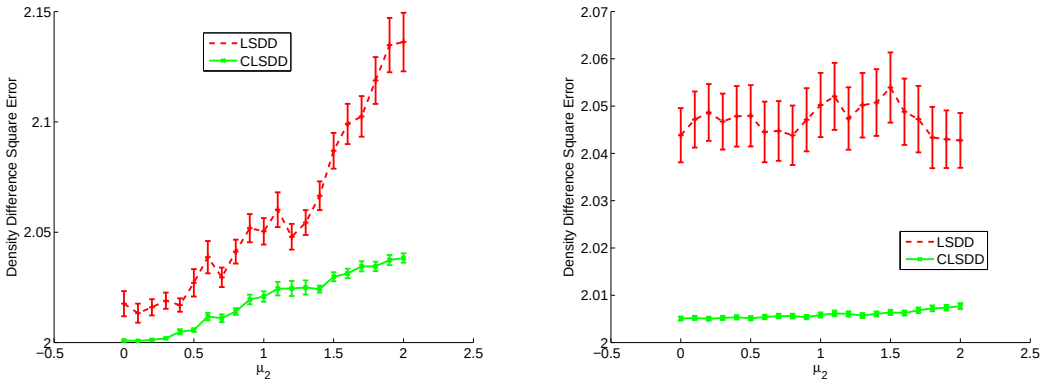


Figure 1: Estimation of density difference for $\mu_1 = 0$, $\mu_2 = 1.4$, and $\sigma_1 = \sigma_2 = 1$. Left: Histograms of samples drawn from $p(x)$ (top) and $p'(x)$ (bottom). Right: True densities $p(x)$ and $p'(x)$, true density difference $f(x) = p(x) - p'(x)$, and its estimates by LSDD and CLSDD.



(a) $\mu_1 = 0$ and $\sigma_1 = \sigma_2 = 1$

(b) $\mu_1 = 0$, $\sigma_1 = 3$, and $\sigma_2 = 5$

Figure 2: Means and standard errors of density-difference squared errors for LSDD and CLSDD over 1000 runs as function of mean μ_2.

5.2 Semi-Supervised Class-Balance Estimation

Next, we apply the CLSDD-based L^2-distance estimator to semi-supervised class-balance estimation under class-prior change.

In real-world classification problems, the class balance in the training dataset is often different from that of the test dataset. Such a situation is called *class-prior change* [6]. Because most supervised learning algorithms assume that training data and test data follow the same probability distribution [10, 3], class-prior change can cause significant estimation bias. If the test class balance is known, the estimation bias caused by class-

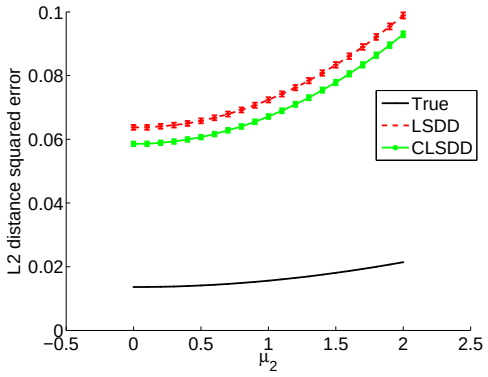


Figure 3: Means and standard errors of true and estimated L^2-distances by LSDD and CLSDD over 1000 runs for $\mu_1 = 0$, $\sigma_1 = 3$, and $\sigma_2 = 5$ as function of mean μ_2.

prior change can be corrected by instance re-weighting or resampling [8, 12]. However, the test class balance is often unknown in practice.

Here, we consider a binary classification problem under a semi-supervised learning setup, where unlabeled test samples are given in addition to labeled training samples [5]. Let $\boldsymbol{x}$ be a pattern to be classified, and let $y \in \{+1, -1\}$ be its class label. We then learn the test class balance by matching a mixture of class-wise training input densities,

$$\pi p_{\mathrm{train}}(\boldsymbol{x}|y = +1) + (1 - \pi) p_{\mathrm{train}}(\boldsymbol{x}|y = -1),$$

with the test input density $p_{\mathrm{test}}(\boldsymbol{x})$ [17, 6]. Here, $\pi \in [0, 1]$ is a class-mixing coefficient that is learned from data. For this distribution matching, we use the L^2-distances estimated by CLSDD and LSDD.

We use UCI binary-classification benchmark datasets[1], where we randomly select 10 labeled training samples from each of the two classes and 50 unlabeled test samples following true class-prior

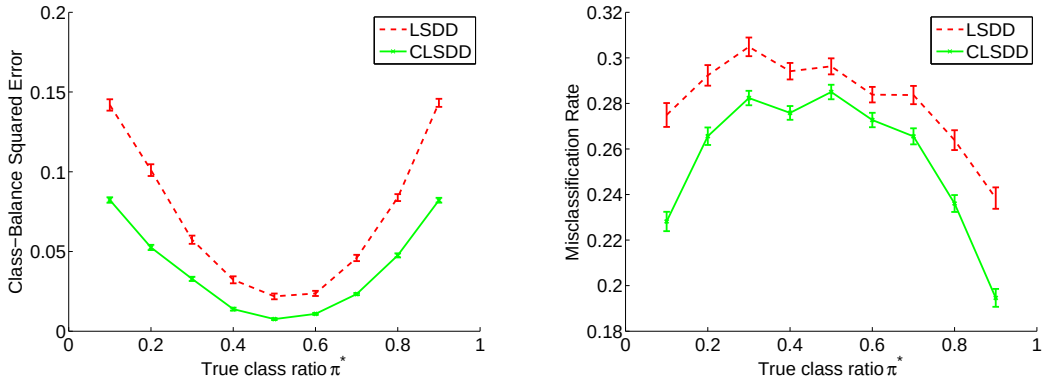
$$\pi^* = 0.1, 0.2, \ldots, 0.9.$$

The left graphs in Figure 4 plot means and standard errors of squared differences between true and estimated class balances over 1000 runs. These graphs show that CLSDD tends to produce better class-balance estimates. The right graphs in Figure 4 plot means and standard errors of misclassification rates by *regularized least-squares classifiers* [16] with class-balance weighting over 1000 runs. The graphs show that better class-balance estimates obtained by CLSDD are translated into lower classification errors.

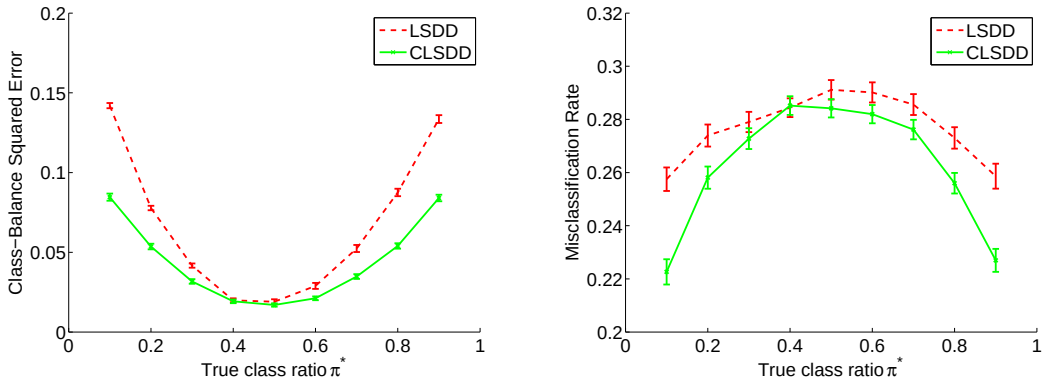
5.3 Two-Sample Test

Finally, we apply the CLSDD-based L^2-distance estimator to distribution comparison by two-sample homogeneity testing.

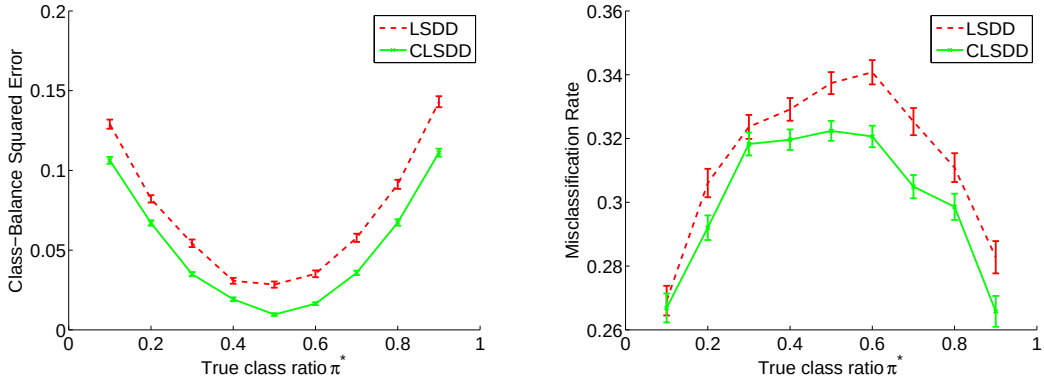
[1] '`http://archive.ics.uci.edu/ml`'.



(a) Adult



(b) Image



(c) Sonar

Figure 4: Class-balance estimation errors (left) and misclassification rates (right).

The goal of two-sample homogeneity test is to determine whether two sets of samples $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$ and $\mathcal{X}' = \{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$, are drawn from the same distribution. In other words, we want to test the null hypothesis $H_0 : P = P'$ against the alternative hypothesis $H_1 : P \neq P'$. Here we use the CLSDD-based and LSDD-based L^2-distance estimators to test the homogeneity of distributions in the framework of *permutation testing* [7]. We again use UCI binary-classification benchmark datasets for experiments.

We first investigate whether the tests can correctly accept the null hypothesis (i.e., $\mathcal{X}$ and $\mathcal{X}'$ follow the same distribution). For each dataset, we randomly split all the positive training samples into two disjoint sets, $\mathcal{X}$ and $\mathcal{X}'$ with $|\mathcal{X}| = |\mathcal{X}'|$. Figure 5(a) shows the rate of accepting the correct null hypothesis under the significance level 0.05, as functions of the relative sample size

$$\eta = 0.2, 0.4, \ldots, 1,$$

i.e., we only use $\eta|\mathcal{X}|$ and $\eta|\mathcal{X}'|$ samples for hypothesis testing. From the results, we can confirm that both the LSDD-based and CLSDD-based methods accept the correct null hypothesis with the pre-specified significance level approximately.

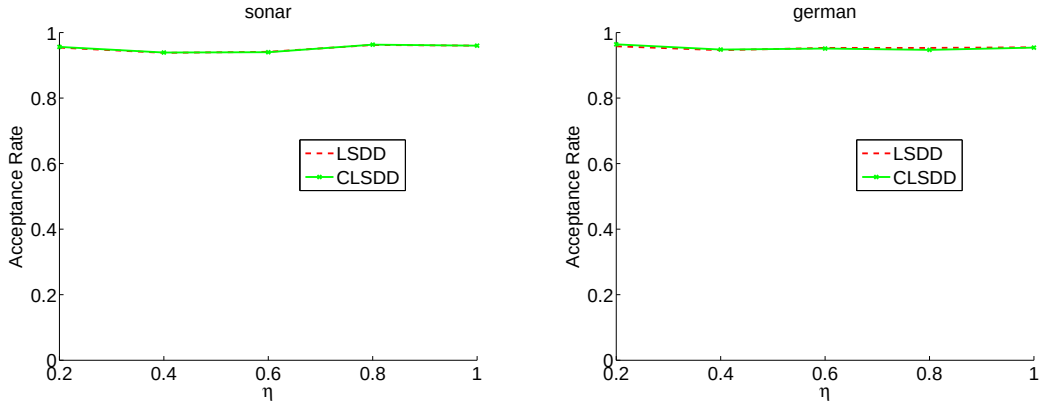
Next, we replace positive samples in the set $\mathcal{X}'$ by randomly chosen negative training samples, yielding $P \neq P'$. Figure 5(b) shows the rate of accepting the incorrect null hypothesis under the significance level 0.05, as functions of the relative sample size η. The results indicate that CLSDD-based method tends to have slightly lower acceptance rates than the LSDD-based method, meaning that the CLSDD-based method tends to have a slightly higher testing power the LSDD-based method.

6 Conclusion

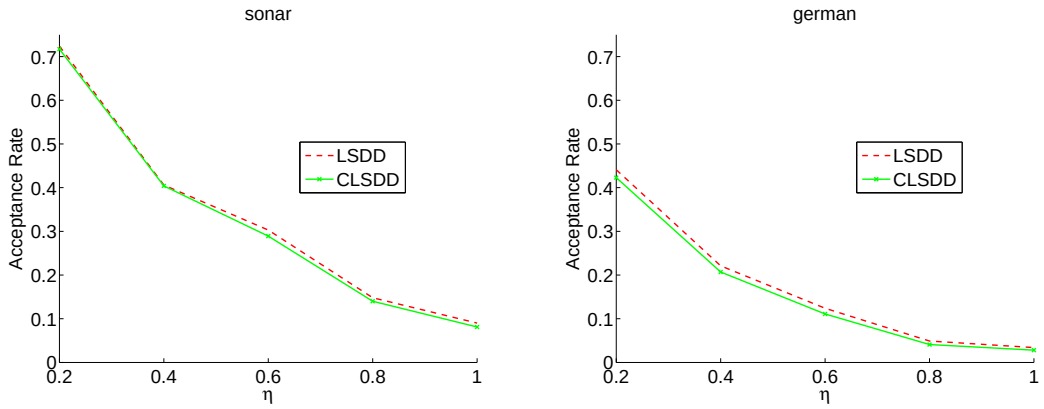
In this paper, we proposed a variation of the *least-squares density-difference* (LSDD) estimator for directly estimating the difference between two probability density functions without density estimation. The proposed method, called the *constrained least-squares density-difference* (CLSDD) estimator, provably improves the estimation accuracy of the density difference and L^2-distance for correctly specified parametric models. We have also experimentally illustrated the usefulness of CLSDD in semi-supervised class-balance estimation and two-sample homogeneity testing.

Acknowledgements

TDN was supported by the MEXT scholarship, MCdP was supported by the MEXT scholarship and MEXT KAKENHI 23120004, TK was supported by MEXT KAKENHI 24500340, and MS was supported by MEXT KAKENHI 25700022 and AOARD.



(a) $P = P'$



(b) $P \neq P'$

Figure 5: The rate of accepting the null hypothesis (i.e., $P = P'$) under a significance level of 0.05. η indicates the relative sample size.

References

[1] A. Albert. *Regression and the Moore-Penrose Pseudoinverse*. Academic Press, New York, NY, USA, 1972.

[2] J. Atif, X. Ripoche, and A. Osorio. Non-rigid medical image registration by maximisation of quadratic mutual information. In *IEEE 29th Annual Northeast Bioengineering Conference*, pages 32–40, 2003.

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

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

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

[6] M. C. du Plessis and M. Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. *Neural Networks*, 50:110–119, 2014.

[7] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York, NY, USA, 1993.

[8] C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI2001)*, pages 973–978, 2001.

[9] D. M. Gray and J. C. Principe. Quadratic mutual information for dimensionality reduction and classification. In *Proceedings of SPIE*, volume 7696, page 76960D, 2010.

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

[11] Y. Kawahara and M. Sugiyama. Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining*, 5(2):114–127, 2012.

[12] Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in non-standard situations. *Machine Learning*, 46(1/3):191–202, 2002.

[13] B. Liu, H. D. Cheng, J. Huang, J. Tian, X. Tang, and J. Liu. Probability density difference-based active contour for ultrasound image segmentation. *Pattern Recognition*, 43(6):2028–2042, 2010.

[14] S. Liu, M. Yamada, N. Collier, and M. Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83, 2013.

[15] C. R. Rao. *Linear Statistical Inference and Its Applications*. Wiley, New York, NY, USA, 1965.

[16] R. Rifkin, G. Yeo, and T. Poggio. Regularized least-squares classification. In J. A. K. Suykens, G. Horvath, S. Basu, C. Micchelli, and J. Vandewalle, editors, *Advances in Learning Theory: Methods, Models and Applications*, volume 190 of *NATO Science Series III: Computer & Systems Sciences*, pages 131–154. IOS Press, Amsterdam, the Netherlands, 2003.

[17] M. Saerens, P. Latinne, and C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002.

[18] M. Sugiyama, T. Suzuki, T. Kanamori, M. C. du Plessis, S. Liu, and I. Takeuchi. Density-difference estimation. *Neural Computation*, 25(10):2734–2775, 2013.

[19] T. Suzuki and M. Sugiyama. Sufficient dimension reduction via squared-loss mutual information estimation. *Neural Computation*, 3(25):725–758, 2013.

[20] K. Torkkola. Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research*, 3:1415–1438, 2003.

[21] M. Yamanaka, M. Matsugu, and M. Sugiyama. Detection of activities and events without explicit categorization. *IPSJ Transactions on Mathematical Modeling and Its Applications*, 6(2):86–92, 2013.

[22] M. Yamanaka, M. Matsugu, and M. Sugiyama. Salient object detection based on direct density-ratio estimation. *IPSJ Transactions on Mathematical Modeling and Its Applications*, 6(2):78–85, 2013.