Density-Difference Estimation*

Masashi Sugiyama
Tokyo Institute of Technology, Japan.
sugi@cs.titech.ac.jp http://sugiyama-www.cs.titech.ac.jp/~sugi
Takafumi Kanamori
Nagoya University, Japan.
kanamori@is.nagoya-u.ac.jp
Taiji Suzuki
The University of Tokyo, Japan.
s-taiji@stat.t.u-tokyo.ac.jp
Marthinus Christoffel du Plessis
Tokyo Institute of Technology, Japan.
christo@sg.cs.titech.ac.jp
Song Liu
Tokyo Institute of Technology, Japan.
song@sg.cs.titech.ac.jp
Ichiro Takeuchi
Nagoya Institute of Technology, Japan.
takeuchi.ichiro@nitech.ac.jp

Abstract

We address the problem of estimating the *difference* between two probability densities. A naive approach 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 estimation error incurred in the first stage can cause a big error in the second stage. In this paper, we propose a single-shot procedure for directly estimating the density difference without separately estimating two densities. We derive a non-parametric finite-sample error bound for the proposed single-shot density-difference estimator and show that it achieves the optimal convergence rate. We then show how the proposed density-difference estimator can be utilized in L^2-distance approximation. Finally, we experimentally demonstrate the usefulness of the proposed method in robust distribution comparison such as class-prior estimation and change-point detection.

Keywords

density difference, L^2-distance, robustness, Kullback-Leibler divergence, kernel density estimation.

*An earlier version of this paper is presented at the Neural Information Processing Systems (NIPS) conference, 2012 (Sugiyama et al., 2012c).

1 Introduction

When estimating a quantity consisting of two elements, a two-stage approach of first estimating the two elements separately and then approximating the target quantity based on the estimates of the two elements often performs poorly, because the first stage is carried out without regard to the second stage and thus a small estimation error incurred in the first stage can cause a big error in the second stage. To cope with this problem, it would be more appropriate to directly estimate the target quantity in a single-shot process without separately estimating the two elements.

A seminal example that follows this general idea is pattern recognition by the *support vector machine*[1] (Boser et al., 1992; Cortes & Vapnik, 1995; Vapnik, 1998): Instead of separately estimating two probability distributions of positive and negative patterns, the support vector machine directly learns the boundary between the positive and negative classes that is sufficient for pattern recognition. More recently, a problem of estimating the ratio of two probability densities was tackled in a similar fashion (Qin, 1998; Sugiyama et al., 2008; Gretton et al., 2009; Kanamori et al., 2009; Nguyen et al., 2010; Kanamori et al., 2012; Sugiyama et al., 2012b; Sugiyama et al., 2012a): The ratio of two probability densities is directly estimated without going through separate estimation of the two probability densities.

In this paper, we further explore this line of research, and propose a method for directly estimating the *difference* between two probability densities in a single-shot process. Density ratios and density differences can both be used for comparing probability densities via approximation of divergences such as the *Kullback-Leibler (KL) divergence* (Kullback & Leibler, 1951) and the L^2-distance. A divergence estimator can be used for solving various machine learning tasks including class-balance estimation under class-prior change (Saerens et al., 2002; du Plessis & Sugiyama, 2012), image segmentation and registration (Liu et al., 2010; Atif et al., 2003), target object detection and recognition (Gray & Principe, 2010; Yamanaka et al., 2013b), feature selection and extraction (Torkkola, 2003; Suzuki & Sugiyama, 2013), and change-point detection in time series (Kawahara & Sugiyama, 2012; Liu et al., 2013; Yamanaka et al., 2013a). In this divergence approximation scenario, density differences are more advantageous than density ratios in several aspects. For example, density ratios can be unbounded even for simple cases (Cortes et al., 2010; Yamada et al., 2013), whereas density differences are always bounded as long as both densities are bounded. Thus, density differences are expected to be learned more easily than density ratios. Also, density ratios are asymmetric and thus the "direction" needs to be determined by a user, whereas density differences are symmetric and thus there is no need to think about the direction. These are our primal motivations to develop a density-difference estimator.

[1] More precisely, Vapnik (1998) said as follows: *If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem.* Two-stage density-difference estimation corresponds to solving a more general problem of separate density estimation in the first stage.

Note that density ratios have their own applications beyond divergence approximation to which density differences may not be applied, such as importance sampling and conditional probability estimation (Sugiyama et al., 2012a). On the other hand, density differences also have their own unique applications to which density ratios may not be applied, such as the estimation of highest density-difference regions in flow cytometric data analysis (Duong et al., 2009) and unsupervised labeling (du Plessis, 2013). This implies that, for density ratios and density differences, neither of them includes the other in terms of ranges of applications. This is our additional motivation to pursue a practical algorithm for density-difference estimation.

For this density-difference estimation problem, we propose a single-shot method, called the *least-squares density-difference* (LSDD) estimator, that directly estimates the density difference without separately estimating two densities. LSDD is derived within the framework of kernel regularized least-squares estimation, and its solution can be computed *analytically* in a computationally efficient and stable manner. Furthermore, LSDD is equipped with cross-validation, and thus all tuning parameters such as the kernel width and the regularization parameter can be systematically and objectively optimized. We derive a finite-sample error bound for the LSDD estimator in a non-parametric setup and show that it achieves the optimal convergence rate.

We also apply LSDD to L^2-distance estimation and show that it is more accurate than the difference of KDEs, which tends to severely underestimate the L^2-distance (Anderson et al., 1994). Compared with the KL divergence, the L^2-distance is more robust against outliers (Basu et al., 1998; Scott, 2001; Besbeas & Morgan, 2004). We experimentally demonstrate the usefulness of LSDD in robust distribution comparison such as semi-supervised class-prior estimation and unsupervised change detection.

The rest of this paper is structured as follows. In Section 2, we derive the LSDD method and investigate its theoretical properties. In Section 3, we show how LSDD can be utilized for L^2-distance approximation. In Section 4, we illustrate the numerical behavior of LSDD. Finally, we conclude in Section 5.

2 Density-Difference Estimation

In this section, we propose a single-shot method for estimating the difference between two probability densities from samples, and analyze its theoretical properties.

2.1 Problem Formulation and Naive Approach

First, we formulate the problem of density-difference estimation.

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

and $p'(\boldsymbol{x})$, respectively:

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

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

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

A naive approach to density-difference estimation is to use *kernel density estimators* (KDEs) (Silverman, 1986). For Gaussian kernels, the KDE-based density-difference estimator is given by

$$\widetilde{f}(\boldsymbol{x}) := \widehat{p}(\boldsymbol{x}) - \widehat{p}'(\boldsymbol{x}),$$

where

$$\widehat{p}(\boldsymbol{x}) := \frac{1}{n(2\pi\sigma^2)^{d/2}} \sum_{i=1}^{n} \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{2\sigma^2}\right),$$
$$\widehat{p}'(\boldsymbol{x}) := \frac{1}{n'(2\pi\sigma'^2)^{d/2}} \sum_{i'=1}^{n'} \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'_{i'}\|^2}{2\sigma'^2}\right).$$

The Gaussian widths σ and σ' may be determined based on cross-validation (Härdle et al., 2004).

However, we argue that the KDE-based density-difference estimator is not the best approach because of its two-step nature: Small estimation error incurred in each density estimate can cause a big error in the final density-difference estimate. More intuitively, good density estimators tend to be smooth and thus a density-difference estimator obtained from such smooth density estimators tends to be over-smoothed (Hall & Wand, 1988; Anderson et al., 1994, see also numerical experiments in Section 4.1.1).

To overcome this weakness, we give a single-shot procedure of directly estimating the density difference $f(\boldsymbol{x})$ without separately estimating the densities $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$.

2.2 Least-Squares Density-Difference Estimation

In our proposed approach, we fit a density-difference model $g(\boldsymbol{x})$ to the true density-difference function $f(\boldsymbol{x})$ under the squared loss[2]:

$$\operatorname*{argmin}_{g} \int \Big(g(\boldsymbol{x}) - f(\boldsymbol{x})\Big)^2 \mathrm{d}\boldsymbol{x}. \tag{1}$$

[2]Hall and Wand (1988) used a leave-one-out variant of this criterion for jointly determining the bandwidths of two KDEs. See Section 4 for its numerical behavior.

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

$$g(\boldsymbol{x}) = \sum_{\ell=1}^{b} \theta_\ell \psi_\ell(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{\psi}(\boldsymbol{x}), \tag{2}$$

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 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 non-parametric Gaussian kernel model as $g(\boldsymbol{x})$:

$$g(\boldsymbol{x}) = \sum_{\ell=1}^{n+n'} \theta_\ell \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}_\ell\|^2}{2\sigma^2}\right), \tag{3}$$

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 use only a subset of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{n'}\}$ as Gaussian kernel centers.

For the model (2), the optimal parameter $\boldsymbol{\theta}^*$ is given by

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

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

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

Note that, for the Gaussian kernel model (3), the integral in $\boldsymbol{H}$ can be computed analytically as

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

where d denotes the dimensionality of $\boldsymbol{x}$. This is a part of the reason why we chose the Gaussian kernel model in practice. Another reason for this choice is its theoretical superiority, as discussed in Section 2.3.2.

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:

$$\widehat{\boldsymbol{\theta}} := \operatorname*{argmin}_{\boldsymbol{\theta}} \left[\boldsymbol{\theta}^\top \boldsymbol{H} \boldsymbol{\theta} - 2\widehat{\boldsymbol{h}}^\top \boldsymbol{\theta} + \lambda \boldsymbol{\theta}^\top \boldsymbol{\theta} \right], \tag{4}$$

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 objective function in Eq.(4) and equating it to zero, we can obtain the solution $\widehat{\boldsymbol{\theta}}$ analytically as

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{H} + \lambda \boldsymbol{I}_b)^{-1} \widehat{\boldsymbol{h}},$$

where $\boldsymbol{I}_b$ denotes the b-dimensional identity matrix.

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

$$\widehat{f}(\boldsymbol{x}) = \widehat{\boldsymbol{\theta}}^\top \boldsymbol{\psi}(\boldsymbol{x}). \tag{5}$$

We call this the *least-squares density-difference* (LSDD) estimator.

2.3 Theoretical Analysis

Here, we theoretically investigate the behavior of the LSDD estimator.

2.3.1 Parametric Convergence

First, we consider a linear parametric setup where basis functions in our density-difference model (2) are fixed.

Suppose that $n/(n+n')$ converges to $\eta \in [0,1]$, and let $\lambda = o(\sqrt{1/n}, \sqrt{1/n'})$. Then the *central limit theorem* (Rao, 1965) asserts that $\sqrt{\frac{nn'}{n+n'}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ converges in law to the normal distribution with mean $\boldsymbol{0}$ and covariance matrix

$$\boldsymbol{H}^{-1}((1-\eta)\boldsymbol{V}_p + \eta \boldsymbol{V}_{p'})\boldsymbol{H}^{-1},$$

where $\boldsymbol{V}_p$ denotes the covariance matrix of $\boldsymbol{\psi}(\boldsymbol{x})$ under the probability density $p(\boldsymbol{x})$:

$$\boldsymbol{V}_p := \int \left(\boldsymbol{\psi}(\boldsymbol{x}) - \boldsymbol{\psi}_p\right) \left(\boldsymbol{\psi}(\boldsymbol{x}) - \boldsymbol{\psi}_p\right)^\top p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{6}$$

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

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

This result implies that the LSDD estimator has asymptotic normality with asymptotic order $\sqrt{1/n + 1/n'}$, which is the optimal convergence rate in the parametric setup.

2.3.2 Non-Parametric Error Bound

Next, we consider a non-parametric setup where a density-difference function is learned in a Gaussian *reproducing kernel Hilbert space* (RKHS) (Aronszajn, 1950).

Let $\mathcal{H}_\gamma$ be the Gaussian RKHS with width γ:

$$k_\gamma(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{\gamma^2}\right).$$

Let us consider a slightly modified LSDD estimator that is more suitable for non-parametric error analysis[3]: For $n' = n$,

$$\widehat{f} := \operatorname*{argmin}_{g \in \mathcal{H}_\gamma} \left[\|g\|_{L^2}^2 - 2\left(\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{x}_i) - \frac{1}{n}\sum_{i'=1}^{n} g(\boldsymbol{x}'_{i'})\right) + \lambda \|g\|_{\mathcal{H}_\gamma}^2 \right],$$

where $\|\cdot\|_{L^2}$ denotes the L^2-norm and $\|\cdot\|_{\mathcal{H}_\gamma}$ denotes the norm in RKHS $\mathcal{H}_\gamma$.

Then we can prove that, for all $\rho, \rho' > 0$, there exists a constant K such that, for all $\tau \geq 1$ and $n \geq 1$, the non-parametric LSDD estimator with appropriate choice of λ and γ satisfies[4]

$$\|\widehat{f} - f\|_{L^2}^2 + \lambda \|\widehat{f}\|_{\mathcal{H}_\gamma}^2 \leq K\left(n^{-\frac{2\alpha}{2\alpha+d}+\rho} + \tau n^{-1+\rho'}\right) \tag{7}$$

with probability not less than $1 - 4e^{-\tau}$. Here, d denotes the dimensionality of input vector $\boldsymbol{x}$, and $\alpha \geq 0$ denotes the regularity of Besov space to which the true density-difference function f belongs (smaller/larger α means f is "less/more complex"; see Appendix A for its precise definition). Because $n^{-\frac{2\alpha}{2\alpha+d}}$ is the optimal learning rate in this setup (Eberts & Steinwart, 2011), the above result shows that the non-parametric LSDD estimator achieves the optimal convergence rate.

It is known that, if the naive KDE with a Gaussian kernel is used for estimating a probability density with regularity $\alpha > 2$, the optimal learning rate cannot be achieved (Farrell, 1972; Silverman, 1986). To achieve the optimal rate by KDE, we should choose a kernel function specifically tailored to each regularity α (Parzen, 1962). However, such a kernel function is not non-negative and it is difficult to implement it in practice. On the other hand, our LSDD estimator can always achieve the optimal learning rate for a Gaussian kernel without regard to regularity α.

2.4 Model Selection by Cross-Validation

The above theoretical analyses showed the superiority of LSDD in terms of the convergence rates. However, the practical performance of LSDD depends on the choice of models (i.e.,

[3]More specifically, the regularizer is replaced from the squared ℓ_2-norm of parameters to the squared RKHS-norm of a learned function, which is necessary to establish consistency. Nevertheless, we use the squared ℓ_2-norm of parameters in experiments because it is simpler and performs well in practice.

[4]Because our theoretical result is highly technical, we only describe a rough idea here. More precise statement of the result and its complete proof are provided in Appendix A, where we utilize the mathematical technique developed in Eberts and Steinwart (2011) for a regression problem.

the kernel width σ and the regularization parameter λ). Here, we show that 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 $\mathcal{X}\backslash\mathcal{X}_t$ and $\mathcal{X}'\backslash\mathcal{X}'_t$ (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.

A MATLAB® implementation of LSDD is available from

'`http://sugiyama-www.cs.titech.ac.jp/~sugi/software/LSDD/`'.

3 L^2-Distance Estimation by LSDD

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

$$L^2(p, p') := \int \Big(p(\boldsymbol{x}) - p'(\boldsymbol{x})\Big)^2 \mathrm{d}\boldsymbol{x}, \tag{8}$$

from samples $\mathcal{X} := \{\boldsymbol{x}_i\}_{i=1}^n$ and $\mathcal{X}' := \{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ drawn independently from the probability distributions with densities $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$, respectively.

3.1 Basic Forms

For an equivalent expression

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

if we replace $f(\boldsymbol{x})$ with an LSDD estimator $\widehat{f}(\boldsymbol{x})$ and approximate the expectations by empirical averages, the following L^2-distance estimator can be obtained:

$$L^2(p, p') \approx \widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\theta}}. \tag{9}$$

Similarly, for another expression

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

replacing $f(\boldsymbol{x})$ with an LSDD estimator $\widehat{f}(\boldsymbol{x})$ gives another L^2-distance estimator:

$$L^2(p, p') \approx \widehat{\boldsymbol{\theta}}^\top \boldsymbol{H} \widehat{\boldsymbol{\theta}}. \tag{10}$$

3.2 Reduction of Bias Caused by Regularization

Eq.(9) and Eq.(10) themselves give approximations to $L^2(p, p')$. Nevertheless, we argue that the use of their combination, defined by

$$\widehat{L}^2(\mathcal{X}, \mathcal{X}') := 2\widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}^\top \boldsymbol{H} \widehat{\boldsymbol{\theta}}, \tag{11}$$

is more sensible.

To explain the reason, let us consider a generalized L^2-distance estimator of the following form:

$$\beta \widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\theta}} + (1-\beta) \widehat{\boldsymbol{\theta}}^\top \boldsymbol{H} \widehat{\boldsymbol{\theta}}, \tag{12}$$

where β is a real scalar. If the regularization parameter λ (≥ 0) is small, then Eq.(12) can be expressed as

$$\beta \widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\theta}} + (1-\beta) \widehat{\boldsymbol{\theta}}^\top \boldsymbol{H} \widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{h}}^\top \boldsymbol{H}^{-1} \widehat{\boldsymbol{h}} - \lambda(2-\beta) \widehat{\boldsymbol{h}}^\top \boldsymbol{H}^{-2} \widehat{\boldsymbol{h}} + o_p(\lambda), \tag{13}$$

where o_p denotes the probabilistic order (the derivation of Eq.(13) is given in Appendix B). Thus, up to $O_p(\lambda)$, the bias introduced by regularization (i.e., the second term in the right-hand side of Eq.(13) that depends on λ) can be eliminated if $\beta = 2$, which yields Eq.(11). Note that, if no regularization is imposed (i.e., $\lambda = 0$), both Eq.(9) and Eq.(10) yield $\widehat{\boldsymbol{h}}^\top \boldsymbol{H}^{-1} \widehat{\boldsymbol{h}}$, the first term in the right-hand side of Eq.(13).

Eq.(11) is actually equivalent to the negative of the optimal objective value of the LSDD optimization problem without regularization (i.e., Eq.(4) with $\lambda = 0$). This can be naturally interpreted through a lower bound of $L^2(p, p')$ obtained by *Legendre-Fenchel convex duality* (Rockafellar, 1970):

$$L^2(p, p') = \sup_g \left[2 \left(\int g(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int g(\boldsymbol{x}) p'(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right) - \int g(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} \right],$$

where the supremum is attained at $g = f$. If the expectations are replaced by empirical estimators and the linear-in-parameter model (2) is used as g, the above optimization problem is reduced to the LSDD objective function without regularization (see Eq.(4)). Thus, LSDD corresponds to approximately maximizing the above lower bound and Eq.(11) is its maximum value.

Through eigenvalue decomposition of $\boldsymbol{H}$, we can show that

$$2\widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}^\top \boldsymbol{H} \widehat{\boldsymbol{\theta}} \geq \widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\theta}} \geq \widehat{\boldsymbol{\theta}}^\top \boldsymbol{H} \widehat{\boldsymbol{\theta}}.$$

Thus, our approximator (11) is not less than the plain approximators (9) and (10).

3.3 Further Bias Correction

$\widehat{\boldsymbol{h}}^\top \boldsymbol{H}^{-1}\widehat{\boldsymbol{h}}$, the first term in Eq.(13), is an essential part of the L^2-distance estimator (11). However, it is actually a slightly biased estimator of the target quantity $\boldsymbol{h}^\top \boldsymbol{H}^{-1}\boldsymbol{h}$ $(= \boldsymbol{\theta}^{*\top}\boldsymbol{H}\boldsymbol{\theta}^* = \boldsymbol{h}^\top\boldsymbol{\theta}^*)$:

$$\mathbb{E}[\widehat{\boldsymbol{h}}^\top \boldsymbol{H}^{-1}\widehat{\boldsymbol{h}}] = \boldsymbol{h}^\top \boldsymbol{H}^{-1}\boldsymbol{h} + \mathrm{tr}\left(\boldsymbol{H}^{-1}\left(\frac{1}{n}\boldsymbol{V}_p + \frac{1}{n'}\boldsymbol{V}_{p'}\right)\right), \tag{14}$$

where $\mathbb{E}$ denotes the expectation over all samples $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^n$ and $\mathcal{X}' = \{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$, and $\boldsymbol{V}_p$ and $\boldsymbol{V}_{p'}$ are defined by Eq.(6) (its derivation is given in Appendix C).

The second term in the right-hand side of Eq.(14) is an estimation bias that is generally non-zero. Thus, based on Eq.(14), we can construct a bias-corrected L^2-distance estimator as

$$\widetilde{L}^2(\mathcal{X},\mathcal{X}') := 2\widehat{\boldsymbol{h}}^\top\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}^\top\boldsymbol{H}\widehat{\boldsymbol{\theta}} - \mathrm{tr}\left(\boldsymbol{H}^{-1}\left(\frac{1}{n}\widehat{\boldsymbol{V}}_p + \frac{1}{n'}\widehat{\boldsymbol{V}}_{p'}\right)\right), \tag{15}$$

where $\widehat{\boldsymbol{V}}_p$ is an empirical estimator of covariance matrix $\boldsymbol{V}_p$,

$$\widehat{\boldsymbol{V}}_p := \frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{\psi}(\boldsymbol{x}_i) - \widehat{\boldsymbol{\psi}}_p\right)\left(\boldsymbol{\psi}(\boldsymbol{x}_i) - \widehat{\boldsymbol{\psi}}_p\right)^\top,$$

and $\widehat{\boldsymbol{\psi}}_p$ is an empirical estimator of the expectation $\boldsymbol{\psi}_p$:

$$\widehat{\boldsymbol{\psi}}_p := \frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}(\boldsymbol{x}_i).$$

The true L^2-distance is non-negative by definition (see Eq.(8)), but the above bias-corrected estimate can take a negative value. Following the same line as Baranchik (1964), the *positive-part* estimator may be more accurate:

$$\overline{L}^2(\mathcal{X},\mathcal{X}') := \max\left\{0, \widetilde{L}^2(\mathcal{X},\mathcal{X}')\right\}.$$

However, in our preliminary experiments, $\overline{L}^2(\mathcal{X},\mathcal{X}')$ does not always perform well particularly when $\boldsymbol{H}$ is ill-conditioned. For this reason, we practically propose to use $\widehat{L}^2(\mathcal{X},\mathcal{X}')$ defined by Eq.(11).

4 Experiments

In this section, we experimentally evaluate the performance of LSDD.

4.1 Numerical Examples

First, we show numerical examples using artificial datasets.

4.1.1 LSDD vs. KDE

Let

$$p(\boldsymbol{x}) = N(\boldsymbol{x}; (\mu, 0, \ldots, 0)^\top, (4\pi)^{-1}\boldsymbol{I}_d),$$
$$p'(\boldsymbol{x}) = N(\boldsymbol{x}; (0, 0, \ldots, 0)^\top, (4\pi)^{-1}\boldsymbol{I}_d),$$

where $N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multi-dimensional normal density with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ with respect to $\boldsymbol{x}$, and $\boldsymbol{I}_d$ denotes the d-dimensional identity matrix. Before feeding data samples to algorithms, we pre-normalize them to have unit variance in the element-wise manner.

In LSDD, the Gaussian width σ and the regularization parameter λ are chosen by 5-fold cross-validation in terms of the LSDD criterion (see Section 2.4) from the following grid values:

$$\sigma \in \{10^{-2}, 10^{-1.5}, 10^{-1}, 10^{-0.5}, 10^{0}\},$$
$$\lambda \in \{10^{-1}, 10^{-0.5}, 10^{0}, 10^{0.5}, 10^{1}\}.$$

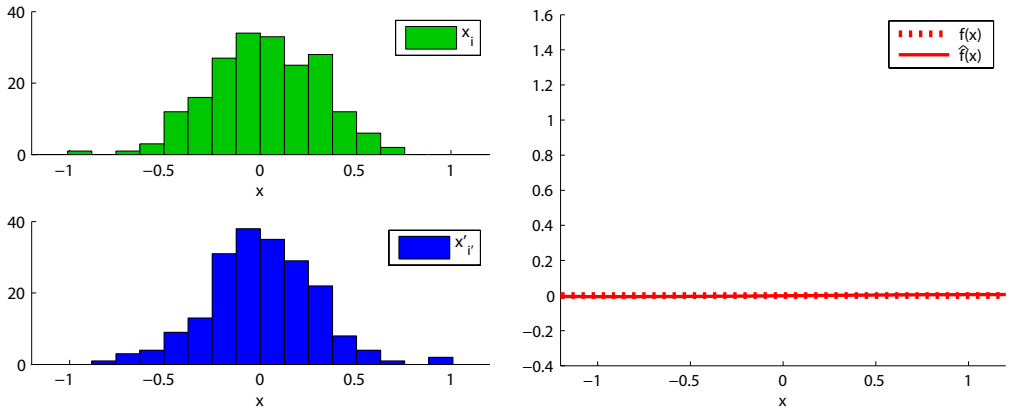
We experimentally compare the behavior of LSDD with two methods based on the difference of kernel density estimators (KDEs):

KDEi: Two Gaussian widths are *independently* chosen from the above candidate values based on 5-fold least-squares cross-validation. That is, for each density, we perform cross-validation in terms of the L^2-distance between estimated and true densities so that the density is optimally approximated (Härdle et al., 2004).

KDEj Two Gaussian widths are *jointly* chosen from the above candidate values based on 5-fold cross-validation in terms of the LSDD criterion (Hall & Wand, 1988). That is, we compute the cross-validated LSDD criterion as a function of two Gaussian widths and find the best pair that minimizes the criterion.

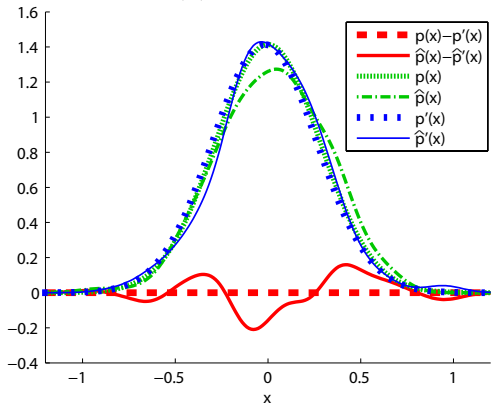
We first illustrate the behavior of the LSDD and KDE-based methods under $d = 1$ and $n = n' = 200$. Figure 1 depicts the data samples and density-difference estimation results obtained by LSDD, KDEi, and KDEj for $\mu = 0$ (i.e., $f(x) = p(x) - p'(x) = 0$). Cross-validation scores of LSDD are also included at the bottom of the figure. The figure shows that LSDD and KDEj give accurate estimates of the true density difference $f(x) = 0$. On the other hand, the density-difference estimate obtained by KDEi is rather fluctuated, although both densities are reasonably well approximated by KDEs. This illustrates an advantage of directly estimating the density difference without going through separate estimation of each density. Figure 2 depicts the results for $\mu = 0.5$ (i.e., $f(x) \neq 0$), showing again that LSDD performs well. KDEi and KDEj give the same estimation result for this dataset, which slightly underestimates the peaks.

Next, we compare the performance of L^2-distance approximation based on LSDD, KDEi, and KDEj. For $\mu = 0, 0.2, 0.4, 0.6, 0.8$ and $d = 1, 5$, we draw $n = n' = 200$ samples from the above $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$. Figure 3 depicts the mean and standard error of

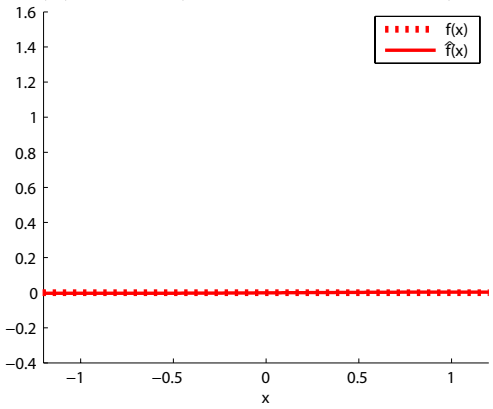


(a) Samples

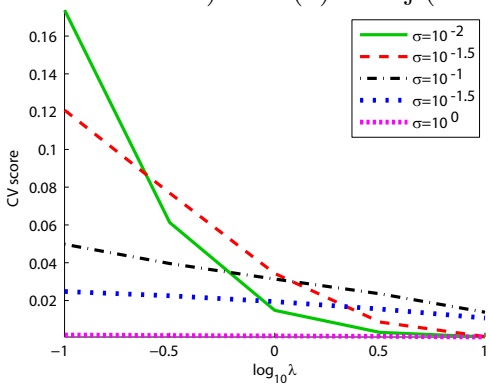
(b) LSDD ($\sigma = 10^0$ and $\lambda = 10^1$)



(c) KDEi ($\sigma = 10^{-1}$ and $\sigma' = 10^{-1}$)

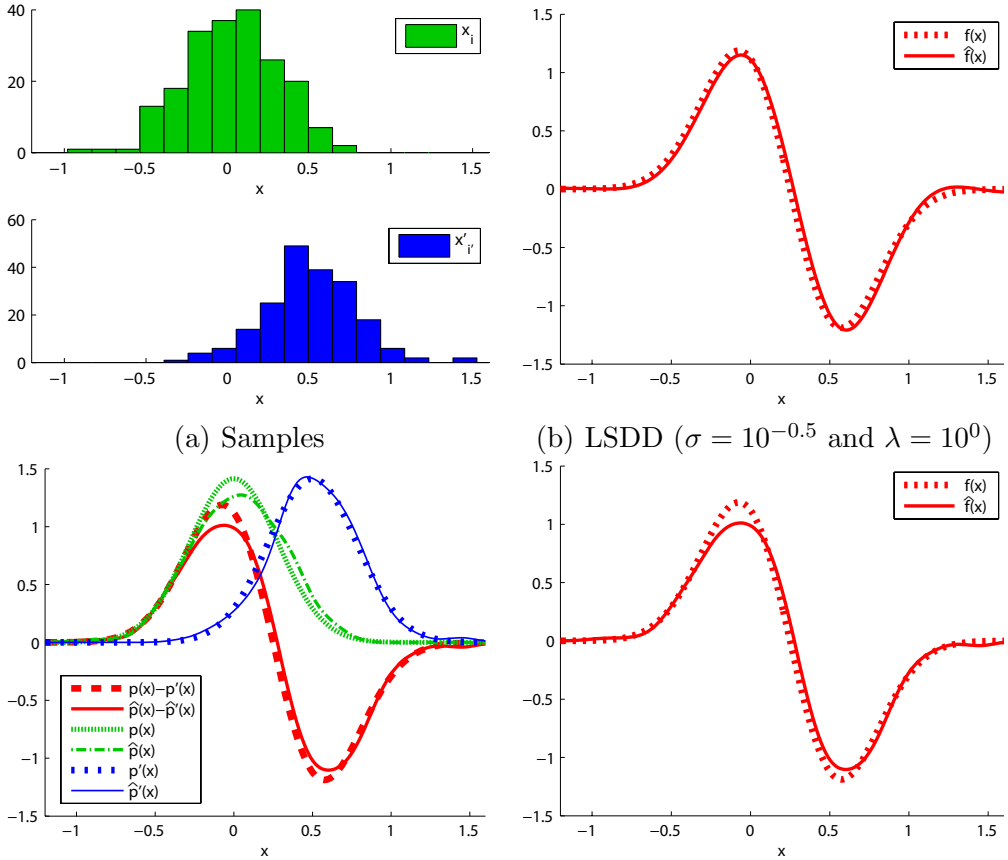


(d) KDEj ($\sigma = 10^0$ and $\sigma' = 10^0$)



(e) CV scores for LSDD

Figure 1: Estimation of density difference when $\mu = 0$ (i.e., $f(x) = p(x) - p'(x) = 0$). In the captions of (b), (c), and (d), model parameters chosen by cross-validation are also described. σ and σ' in (c) and (d) denote Gaussian widths for $p(x)$ and $p'(x)$, respectively.

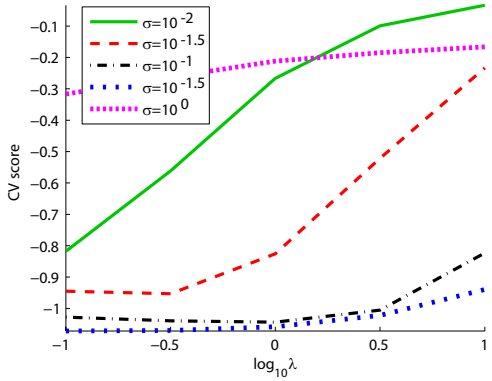


(a) Samples

(b) LSDD ($\sigma = 10^{-0.5}$ and $\lambda = 10^{0}$)

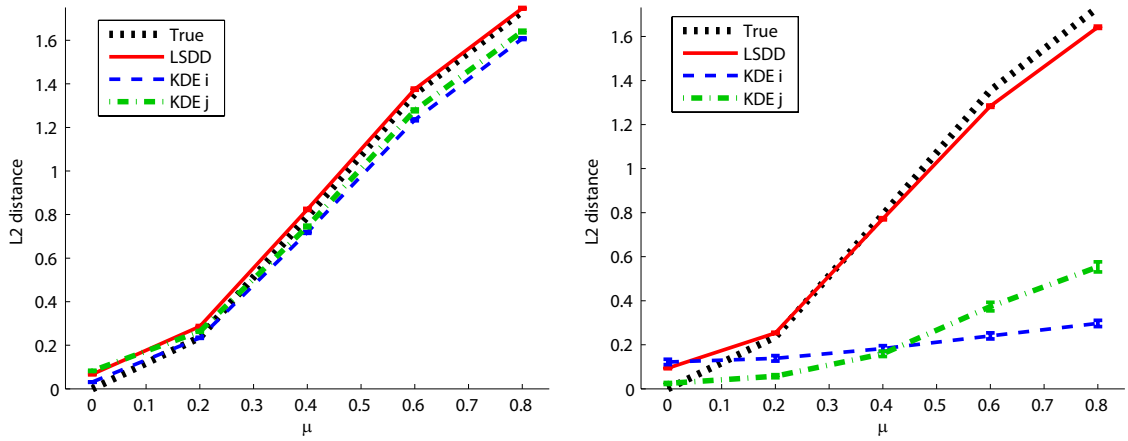
(c) KDEi ($\sigma = 10^{-1}$ and $\sigma' = 10^{-1}$)

(d) KDEj ($\sigma = 10^{-1}$ and $\sigma' = 10^{-1}$)



(e) CV scores for LSDD

Figure 2: Estimation of density difference when $\mu = 0.5$ (i.e., $f(x) = p(x) - p'(x) \neq 0$). In the captions of (b), (c), and (d), model parameters chosen by cross-validation are also described. σ and σ' in (c) and (d) denote Gaussian widths for $p(x)$ and $p'(x)$, respectively.



(a) $d = 1$ (b) $d = 5$

Figure 3: L^2-distance estimation by LSDD, KDEi, and KDEj for $n = n' = 200$ as functions of the Gaussian mean μ. Means and standard errors over 1000 runs are plotted.

estimated L^2-distances over 1000 runs as functions of mean μ. When $d = 1$ (Figure 3(a)), the LSDD-based L^2-distance estimator gives the most accurate estimates of the true L^2-distance, whereas the KDEi-based L^2-distance estimator slightly underestimates the true L^2-distance for large μ. This is caused by the fact that KDE tends to provide smooth density estimates (see Figure 2(c) again): Such smooth density estimates are accurate as density estimates, but the difference of smooth density estimates yields a small L^2-distance estimate (Anderson et al., 1994). More specifically, the density $p'(x)$ is estimated accurately at around $x = 0.5$, but negative values of the density difference $f(x)$ are underestimated there because $\widehat{p}(x)$ is smoother than the true density $p(x)$ and thus its tail values at around $x = 0.5$ are larger. The KDEj-based L^2-distance estimator tends to improve this drawback of KDEi to some extent, but it still slightly underestimates the true L^2-distance when μ is large.

When $d = 5$ (Figure 3(b)), the KDE-based L^2-distance estimators even severely underestimate the true L^2-distance for large μ. On the other hand, the LSDD-based L^2-distance estimator still gives reasonably accurate estimates of the true L^2-distance even when $d = 5$. However, we note that LSDD also slightly underestimates the true L^2-distance when μ is large, because slight underestimation tends to yield smaller variance and thus such stabilized solutions are more accurate in terms of the bias-variance trade-off.

In Figure 1, we have illustrated that LSDD and KDEj work better than KDEi when the Gaussian mean is $\mu = 0$. However, in Figure 3(a), KDEi is actually shown to be the best performing method for $\mu = 0$ in terms of the average over 1000 runs. To fill this gap, let us depict in Figure 4 histograms of L^2-distance estimates obtained by LSDD, KDEi, and KDEj over 1000 runs for the Gaussian mean $\mu = 0$. This graph shows that LSDD and KDEj give exactly correct solutions (i.e., zero) about 300 times, whereas KDEi gives estimates about 0.01 more than 300 times. The graphs plotted in Figure 1 correspond to such typical results where LSDD and KDEj outperform KDEi. On the other hand, KDEi

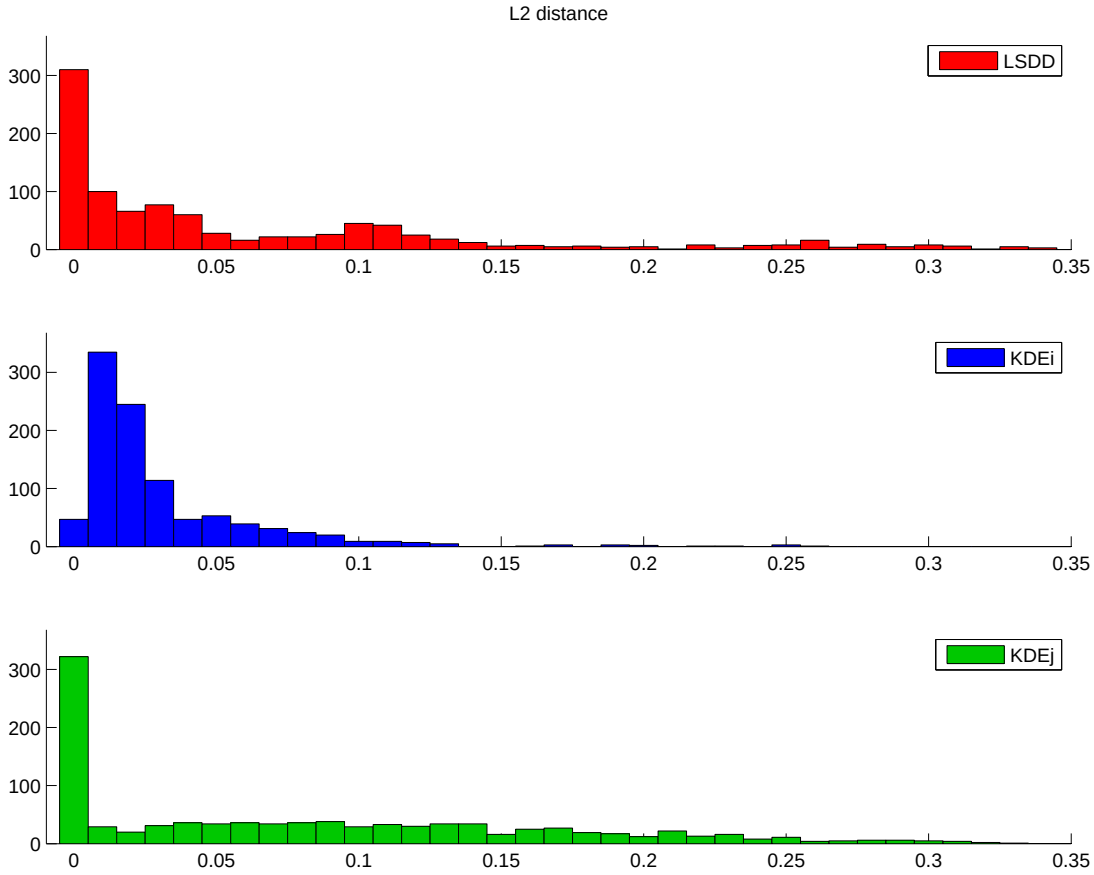
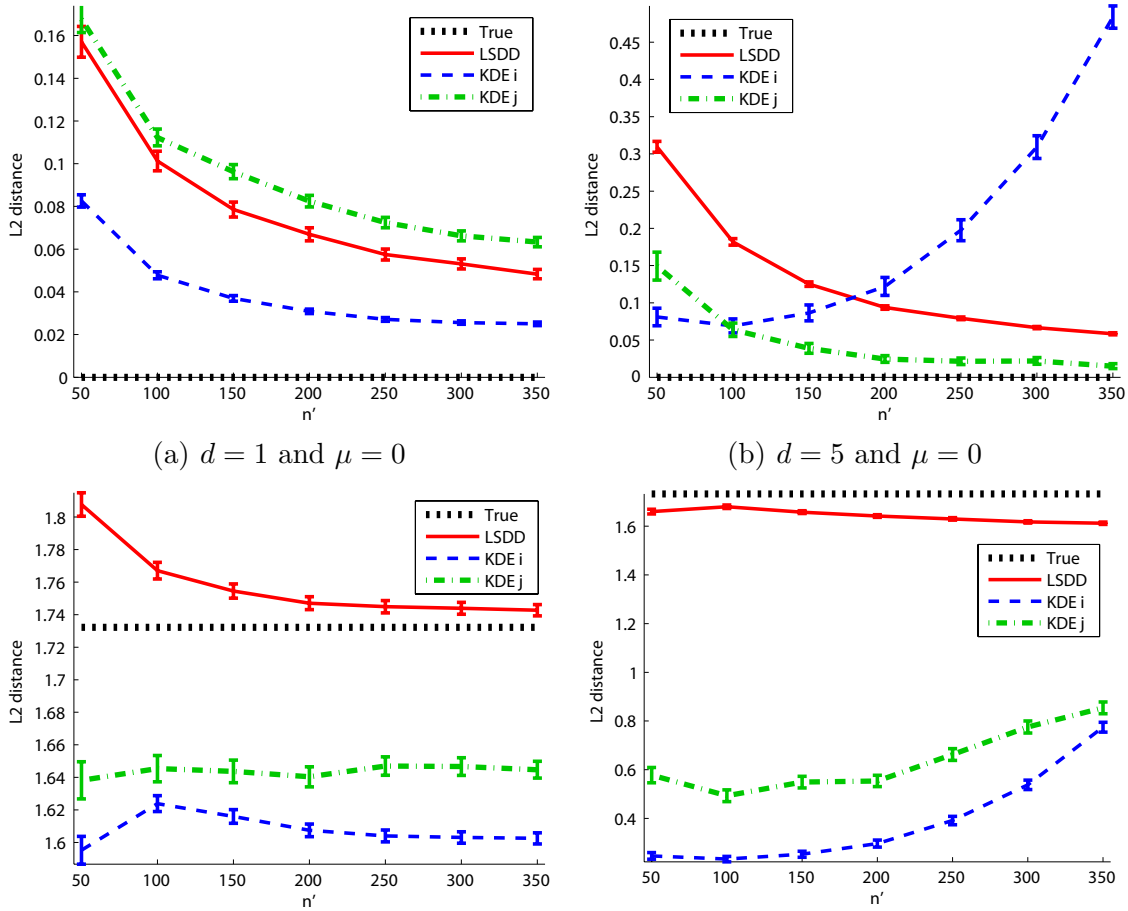


Figure 4: Histograms of L^2-distance estimation by LSDD, KDEi, and KDEj over 1000 runs for $n = n' = 200$ and the Gaussian mean $\mu = 0$.

stably gives estimates less than 0.1 almost always, whereas LSDD and KDEj occasionally give large estimates. This rather unstable behavior of LSDD and KDEj, which was caused by inappropriate choice of the Gaussian width (and the regularization parameter for LSDD) by cross-validation, led to larger mean values of LSDD and KDEj than KDEi in Figure 3(a).

Finally, we investigate the behavior of LSDD, KDEi, and KDEj in L^2-distance estimation when the numbers of samples from two distributions are imbalanced. Figure 5 plots the means and standard errors of the L^2-distance estimated by LSDD, KDEi, and KDEj for $d = 1, 5$ and $\mu = 0, 0.8$ over 1000 runs as functions of n' when n is fixed to 200.

When $d = 1$ and $\mu = 0$ (Figure 5(a)), all three methods behave similarly and the accuracy tends to be improved as n' increases. However, improvement when $n' > n = 200$ is moderate. When the input dimensionality is increased to $d = 5$ (Figure 5(b)), LSDD and KDEj still have the same tendency. However, KDEi behaves differently and the approximation error tends to grow as n' increases. This implies that improving the accuracy of one of the density estimates does not necessarily improve the overall estimation accuracy of the density difference.



(a) $d = 1$ and $\mu = 0$

(b) $d = 5$ and $\mu = 0$

(c) $d = 1$ and $\mu = 0.8$

(d) $d = 5$ and $\mu = 0.8$

Figure 5: L^2-distance estimation by LSDD, KDEi, and KDEj for $n = 200$ as functions of n'. Means and standard errors over 1000 runs are plotted.

When $d = 1$ and $\mu = 0.8$ (Figure 5(c)), LSDD tends to provide better estimates as n' increases, whereas KDEi and KDEj keep underestimating the true L^2-distance even when n' is increased. Finally, when $d = 5$ and $\mu = 0.8$ (Figure 5(d)), LSDD stably provides reasonably good results and its performance does not change significantly when n' is increased. On the other hand, KDEi and KDEj tend to give better results as n' increases.

Overall, LSDD sometimes gives slightly better results for $n' > n$, but its performance is not significantly different from those for $n' = n$. On the other hand, the accuracy of KDEi and KDEj when n' is increased gets better or worse depending on the situation. Thus, having more data samples from one of the distributions does not seem to always improve the estimation accuracy in density-difference estimation.

4.1.2 L^2-Distance vs. KL-Divergence

The *Kullback-Leibler* (KL) divergence (Kullback & Leibler, 1951) is a popular divergence measure for comparing probability distributions. The KL-divergence from $p(\boldsymbol{x})$ to $p'(\boldsymbol{x})$ is defined as

$$\mathrm{KL}(p\|p') := \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}.$$

First, we illustrate the difference between the L^2-distance and the KL-divergence. For $d = 1$, let

$$p(x) = (1-\eta)N(x; 0, 1^2) + \eta N(x; \mu, 1/4^2),$$
$$p'(x) = N(x; 0, 1^2).$$

Implications of the above densities are that samples drawn from $N(x; 0, 1^2)$ are inliers, whereas samples drawn from $N(x; \mu, 1/4^2)$ are outliers. We set the outlier rate at $\eta = 0.1$ and the outlier mean at $\mu = 0, 2, 4, \ldots, 10$ (see Figure 6).

Figure 7(a) depicts the L^2-distance and the KL-divergence for outlier mean $\mu = 0, 2, 4, \ldots, 10$. This shows that both the L^2-distance and the KL-divergence increase as μ increases. However, the L^2-distance is bounded from above, whereas the KL-divergence diverges to infinity as μ tends to infinity. This result implies that the L^2-distance is less sensitive to outliers than the KL-divergence, which well agrees with the observation given in Basu et al. (1998).

Next, we draw $n = n' = 100$ samples from $p(x)$ and $p'(x)$, and estimate the L^2-distance by LSDD and the KL-divergence by the *Kullback-Leibler importance estimation procedure*[5] (KLIEP) (Sugiyama et al., 2008). Figure 7(b) depicts estimated L^2-distance and KL-divergence for outlier mean $\mu = 0, 2, 4, \ldots, 10$ over 100 runs. This shows that both LSDD and KLIEP reasonably capture the profiles of the true L^2-distance and the KL-divergence, although the scale of KLIEP values is much different from the true values (see Figure 7(a)) because the estimated normalization factor was unreliable.

Finally, based on the *permutation test* procedure (Efron & Tibshirani, 1993), we conduct hypothesis testing of the null hypothesis whether densities p and p' are the same. More specifically, we first compute a distance estimate for the original datasets $\mathcal{X}$ and $\mathcal{X}'$ and obtain a distance/divergence estimate $\widehat{D}(\mathcal{X}, \mathcal{X}')$. Next, we randomly permute the $|\mathcal{X} \cup \mathcal{X}'|$ samples, and assign the first $|\mathcal{X}|$ samples to a set $\widetilde{\mathcal{X}}$ and the remaining $|\mathcal{X}'|$ samples to another set $\widetilde{\mathcal{X}}'$. Then we compute a distance/divergence estimate again using the randomly permuted datasets $\widetilde{\mathcal{X}}$ and $\widetilde{\mathcal{X}}'$ and obtain $\widetilde{D}(\widetilde{\mathcal{X}}, \widetilde{\mathcal{X}}')$. Because $\widetilde{\mathcal{X}}$ and $\widetilde{\mathcal{X}}'$ can be regarded as being drawn from the same distribution, $\widetilde{D}(\widetilde{\mathcal{X}}, \widetilde{\mathcal{X}}')$ would take a value close

[5]Estimation of the KL-divergence from data has been extensively studied recently (Wang et al., 2005; Sugiyama et al., 2008; Pérez-Cruz, 2008; Silva & Narayanan, 2010; Nguyen et al., 2010). Among them, KLIEP was shown to possess a superior convergence property and demonstrated to work well in practice (Sugiyama et al., 2008). KLIEP is based on direct estimation of density ratio $p(\boldsymbol{x})/p'(\boldsymbol{x})$ without density estimation of $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$. See also Nguyen et al. (2010), which proposes essentially the same procedure.

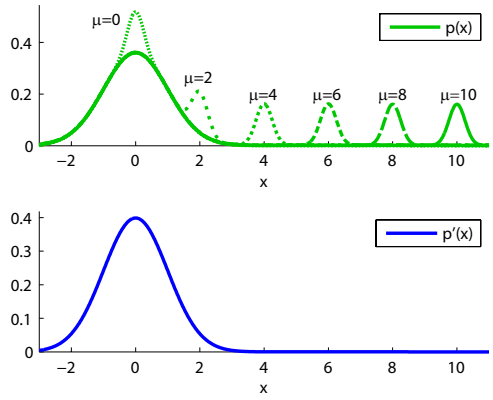
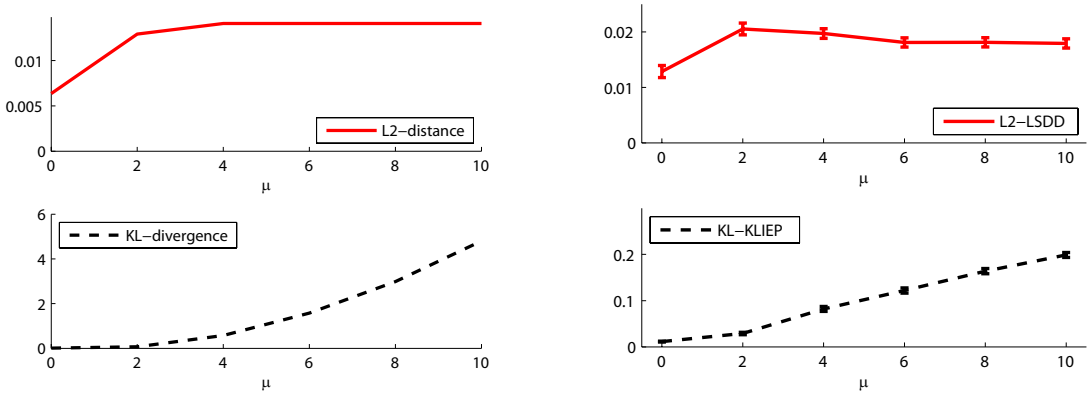


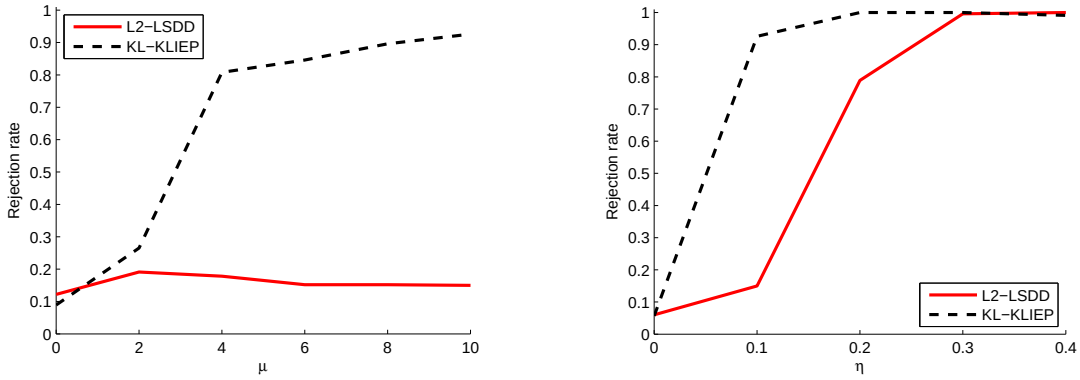
Figure 6: Comparing two densities in the presence of outliers. $p(x)$ includes outliers at $\mu = 0, 2, 4, \ldots, 10$.



(a) True L^2-distance and true KL-divergence

(b) Means and standard errors of L^2-distance estimation by LSDD and KL-divergence estimation by KLIEP over 100 runs.

Figure 7: Comparison of L^2-distance and KL-divergence for outlier rate $\eta = 0.1$ as functions of outlier mean μ.



(a) $\eta = 0.1$

(b) $\mu = 10$

Figure 8: Results of two-sample test over 1000 runs for (a) outlier rate $\eta = 0.1$ as functions of outlier mean μ and for (b) outlier mean $\mu = 10$ as functions of outlier rate η.

to zero. This random permutation procedure is repeated many times (100 times in the following experiments), and the distribution of $\widetilde{D}(\widetilde{\mathcal{X}}, \widetilde{\mathcal{X}}')$ under the null hypothesis (i.e., the two distributions are the same) is constructed. Finally, the p-value is approximated by evaluating the relative ranking of $\widehat{D}(\mathcal{X}, \mathcal{X}')$ in the histogram of $\widetilde{D}(\widetilde{\mathcal{X}}, \widetilde{\mathcal{X}}')$. We set the significance level at 5%.

Figure 8(a) depicts the rejection rate of the null hypothesis (i.e., $p = p'$) over 1000 runs for outlier rate $\eta = 0.1$ and outlier mean $\mu = 0, 2, 4, \ldots, 10$, based on the L^2-distance estimated by LSDD and the KL-divergence estimated by KLIEP. This shows that the KLIEP-based test rejects the null hypothesis more frequently for large μ, whereas the rejection rate of the LSDD-based test is almost unchanged with respect to μ.

Figure 8(b) depicts the rejection rate of the null hypothesis for outlier mean $\mu = 10$ and outlier rate $\eta = 0, 0.1, 0.2, \ldots, 0.4$. When $\eta = 0$ (i.e., no outliers), both the LSDD-based test and the KLIEP-based test accept the null hypothesis with the designated significance level approximately. When $\eta = 0.1$, the LSDD-based test still keeps a low rejection rate, whereas the KLIEP-based test tends to reject the null hypothesis more frequently. When $\eta \geq 0.3$, both the LSDD-based test and the KLIEP-based test always reject the null hypothesis.

Overall, the above results imply that the two-sample homogeneity test by LSDD is more robust against outliers (i.e., two distributions tend to be regarded as the same even in the presence of outliers) than the KLIEP-based test.

4.2 Applications

Here, we apply LSDD to semi-supervised class-balance estimation under class-prior change and change-point detection in time series.

4.2.1 Semi-Supervised Class-Balance Estimation

In real-world pattern recognition tasks, changes in class balance between the training and test phases are often observed. In such cases, naive classifier training produces significant estimation bias because the class balance in the training dataset does not properly reflect that of the test dataset. Here, we consider the problem of learning the class balance of a test dataset in a semi-supervised learning setup where unlabeled test samples are provided in addition to labeled training samples (Chapelle et al., 2006).

More formally, we consider the binary classification problem of classifying pattern $\boldsymbol{x} \in \mathbb{R}^d$ to class $y \in \{+1, -1\}$ under class-prior change, where the class-prior probability for training data $p_{\mathrm{train}}(y)$ and that for test data $p_{\mathrm{test}}(y)$ are different:

$$p_{\mathrm{train}}(y) \neq p_{\mathrm{test}}(y).$$

However, we assume that the class-conditional density for training data $p_{\mathrm{train}}(\boldsymbol{x}|y)$ and that for test data $p_{\mathrm{test}}(\boldsymbol{x}|y)$ are unchanged:

$$p_{\mathrm{train}}(\boldsymbol{x}|y) = p_{\mathrm{test}}(\boldsymbol{x}|y).$$

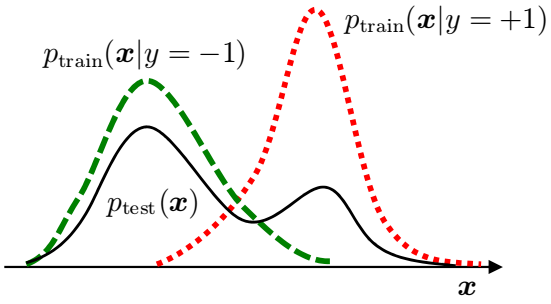


Figure 9: Schematic illustration of semi-supervised class-balance estimation.

Note that training and test joint densities $p_{\mathrm{train}}(\boldsymbol{x}, y)$ and $p_{\mathrm{test}}(\boldsymbol{x}, y)$ as well as training and test input densities $p_{\mathrm{train}}(\boldsymbol{x})$ and $p_{\mathrm{test}}(\boldsymbol{x})$ are generally different under this setup.

Here, our objective is to estimate $p_{\mathrm{test}}(y)$ from labeled training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ drawn independently from $p_{\mathrm{train}}(\boldsymbol{x}, y)$ and unlabeled test samples $\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ drawn independently from $p_{\mathrm{test}}(\boldsymbol{x})$. Given test labels $\{y'_{i'}\}_{i'=1}^{n'}$, $p_{\mathrm{test}}(y)$ can be naively estimated by n'_y/n', where n'_y is the number of test samples in class y. Here, however, we want to estimate $p_{\mathrm{test}}(y)$ *without* $\{y'_{i'}\}_{i'=1}^{n'}$.

The class balance in the test set can be estimated by matching a mixture of class-wise training input densities,

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

to the test input density $p_{\mathrm{test}}(\boldsymbol{x})$ (Saerens et al., 2002), where $\pi \in [0, 1]$ is a mixing coefficient to learn. See Figure 9 for schematic illustration. Here, we use the L^2-distance estimated by LSDD, KDEi, and KDEj (see Section 4.1.1) for this distribution matching. Note that, when LSDD is used to estimate the L^2-distance, separate estimation of $p_{\mathrm{train}}(\boldsymbol{x}|y=\pm 1)$ is not involved, but the difference between $p_{\mathrm{test}}(\boldsymbol{x})$ and $q_{\mathrm{test}}(\boldsymbol{x}; \pi)$ is directly estimated.

As an additional baseline, we include the EM-based class-prior estimation method (Saerens et al., 2002), which actually corresponds to distribution matching under the KL divergence. More specifically, in the EM-based algorithm, test class-prior estimate $\widehat{p}_{\mathrm{test}}(y)$ and test class-posterior estimate $\widehat{p}_{\mathrm{test}}(y|\boldsymbol{x})$ are iteratively estimated as follows:

1. Obtain an estimate of the training class-posterior probability, $\widehat{p}_{\mathrm{train}}(y|\boldsymbol{x})$, from labeled training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, for example, by kernel logistic regression (Hastie et al., 2001) or its squared-loss variant (Sugiyama, 2010).

2. Obtain an estimate of the training class-prior probability from training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ as $\widehat{p}_{\mathrm{train}}(y) = n_y/n$, where n_y is the number of training samples in class y.

3. Set the initial estimate of the test class-posterior probability as

$$\widehat{p}_{\mathrm{test}}(y) \longleftarrow \widehat{p}_{\mathrm{train}}(y).$$

4. Compute a new test class-posterior estimate $\widehat{p}_{\mathrm{test}}(y|\boldsymbol{x})$ based on the current test class-prior estimate $\widehat{p}_{\mathrm{test}}(y)$ as

$$\widehat{p}_{\mathrm{test}}(y|\boldsymbol{x}) \longleftarrow \frac{\widehat{p}_{\mathrm{test}}(y)\widehat{p}_{\mathrm{train}}(y|\boldsymbol{x})/\widehat{p}_{\mathrm{train}}(y)}{\sum_{y'=1}^{c}\widehat{p}_{\mathrm{test}}(y')\widehat{p}_{\mathrm{train}}(y'|\boldsymbol{x})/\widehat{p}_{\mathrm{train}}(y')}.$$

5. Compute a new test class-prior estimate $\widehat{p}_{\mathrm{test}}(y)$ based on the current test class-prior estimate $\widehat{p}_{\mathrm{test}}(y|\boldsymbol{x})$ as

$$\widehat{p}_{\mathrm{test}}(y) \longleftarrow \frac{1}{n'}\sum_{i'=1}^{n'}\widehat{p}_{\mathrm{test}}(y|\boldsymbol{x}'_{i'}).$$

6. Iterate 4. and 5. until convergence.

We use four UCI benchmark datasets[6] for experiments, where we randomly choose 10 labeled training samples from each class and 50 unlabeled test samples following true class-prior

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

The left graphs in Figure 10 plot the mean and standard error of the squared difference between true and estimated class-balances π. These graphs show that LSDD tends to provide better class-balance estimates than alternative approaches.

Next, we use the estimated class balance to train a classifier. We use a weighted ℓ_2-regularized least-squares classifier (Rifkin et al., 2003). That is, a class label $\widehat{y}$ for a test input $\boldsymbol{x}$ is estimated by

$$\widehat{y} = \mathrm{sign}\left(\sum_{\ell=1}^{n}\widehat{\alpha}_\ell K(\boldsymbol{x}, \boldsymbol{x}_\ell)\right),$$

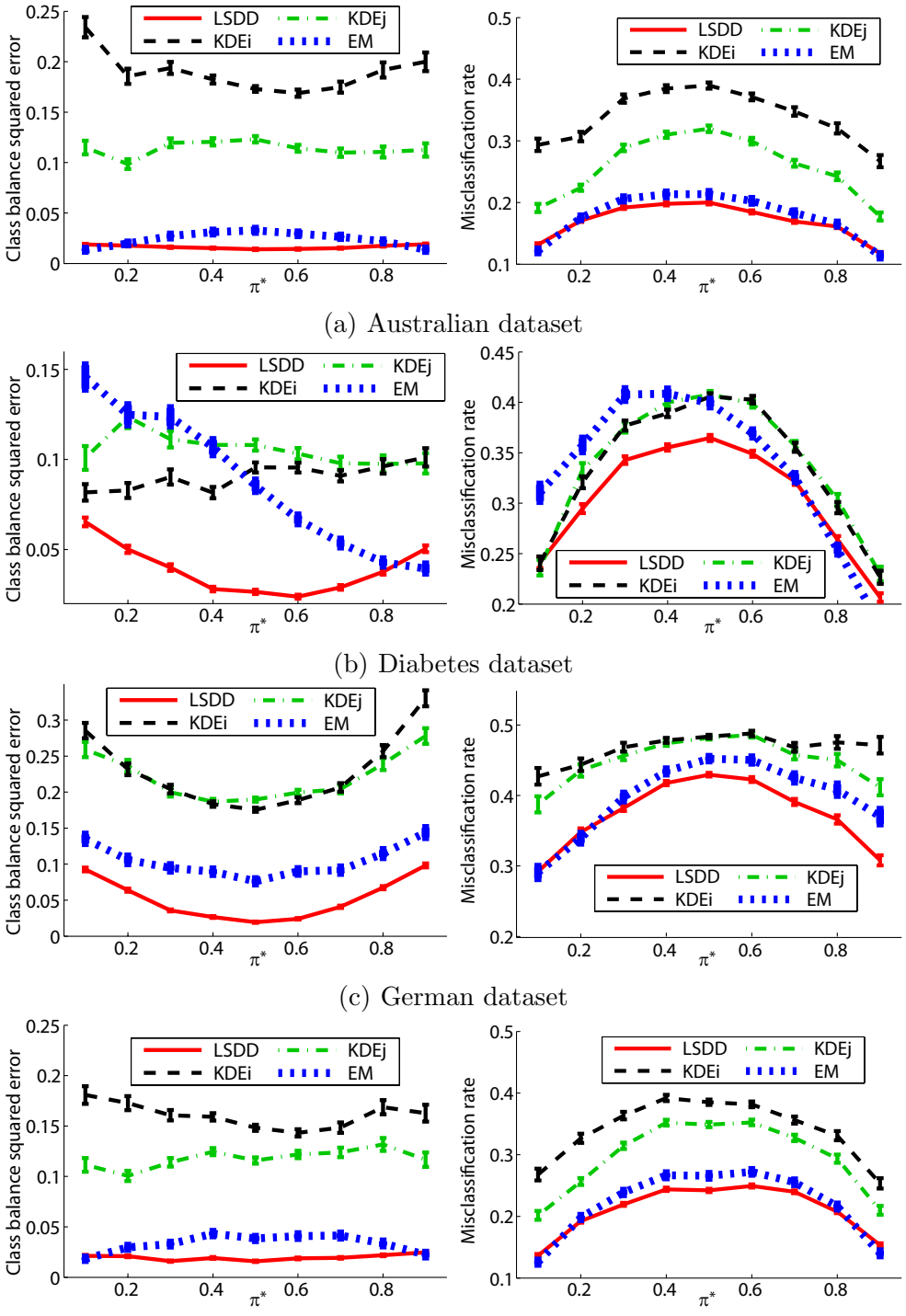
where $K(\boldsymbol{x}, \boldsymbol{x}')$ is the Gaussian kernel function with kernel width κ. $\{\widehat{\alpha}_\ell\}_{\ell=1}^{n}$ are learned parameters given by

$$(\widehat{\alpha}_1, \ldots, \widehat{\alpha}_n) := \underset{\alpha_1,\ldots,\alpha_n}{\mathrm{argmin}}\left[\sum_{i=1}^{n}\frac{\pi_{y_i}}{n_{y_i}/n}\left(\sum_{\ell=1}^{n}\alpha_\ell K(\boldsymbol{x}_i, \boldsymbol{x}_\ell) - y_i\right)^2 + \delta\sum_{\ell=1}^{n}\alpha_\ell^2\right],$$

where $\pi_{+1} = \pi$, $\pi_{-1} = 1 - \pi$, and δ (≥ 0) is the regularization parameter. The Gaussian width κ and the regularization parameter δ are chosen by 5-fold weighted cross-validation (Sugiyama et al., 2007) in terms of the misclassification error.

The right graphs in Figure 10 plot the test misclassification error over 1000 runs. The results show the LSDD-based method provides lower classification errors, which would be brought by good estimates of test class-balances.

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



(a) Australian dataset

(b) Diabetes dataset

(c) German dataset

(d) Statlogheart dataset

Figure 10: Results of semi-supervised class-balance estimation. Left: Squared error of class balance estimation. Right: Misclassification error by a weighted ℓ_2-regularized least-squares classifier with weighted cross-validation.

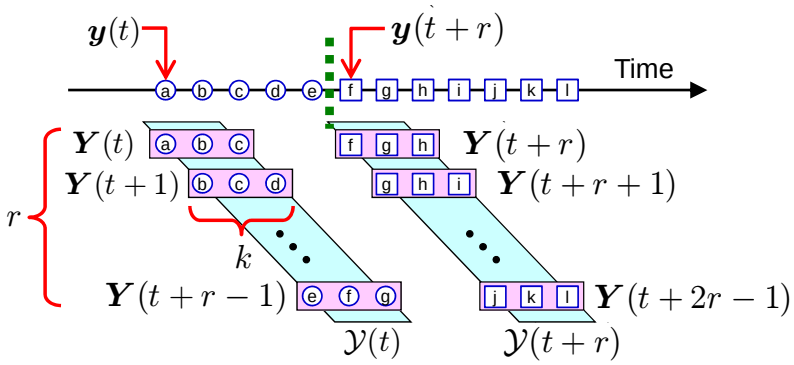


Figure 11: Schematic illustration of unsupervised change detection.

4.2.2 Unsupervised Change Detection

The objective of change detection is to discover abrupt property changes behind time-series data.

Let $\boldsymbol{y}(t) \in \mathbb{R}^m$ be an m-dimensional time-series sample at time t, and let

$$\boldsymbol{Y}(t) := [\boldsymbol{y}(t)^\top, \boldsymbol{y}(t+1)^\top, \ldots, \boldsymbol{y}(t+k-1)^\top]^\top \in \mathbb{R}^{km}$$

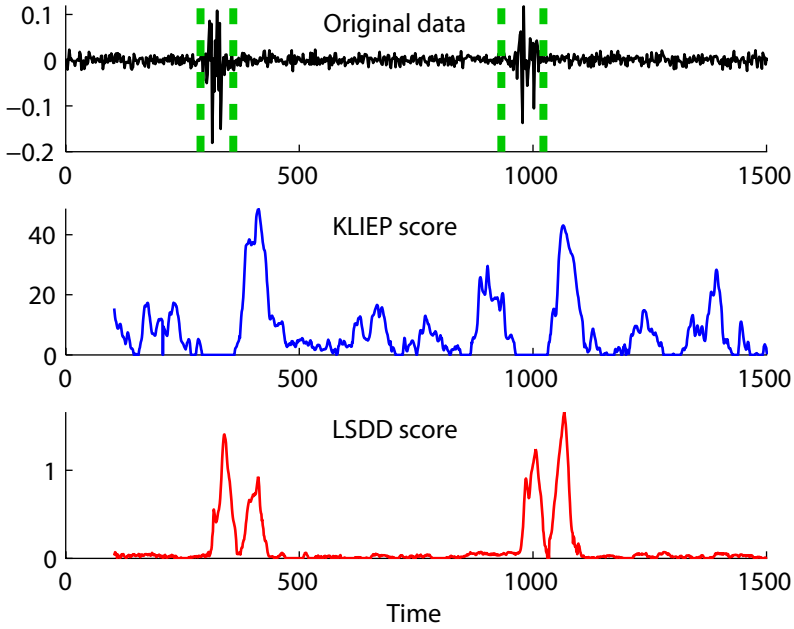
be a subsequence of time series at time t with length k. We treat the subsequence $\boldsymbol{Y}(t)$ as a sample, instead of a single point $\boldsymbol{y}(t)$, by which time-dependent information can be incorporated naturally (Kawahara & Sugiyama, 2012). Let $\mathcal{Y}(t)$ be a set of r retrospective subsequence samples starting at time t:

$$\mathcal{Y}(t) := \{\boldsymbol{Y}(t), \boldsymbol{Y}(t+1), \ldots, \boldsymbol{Y}(t+r-1)\}.$$

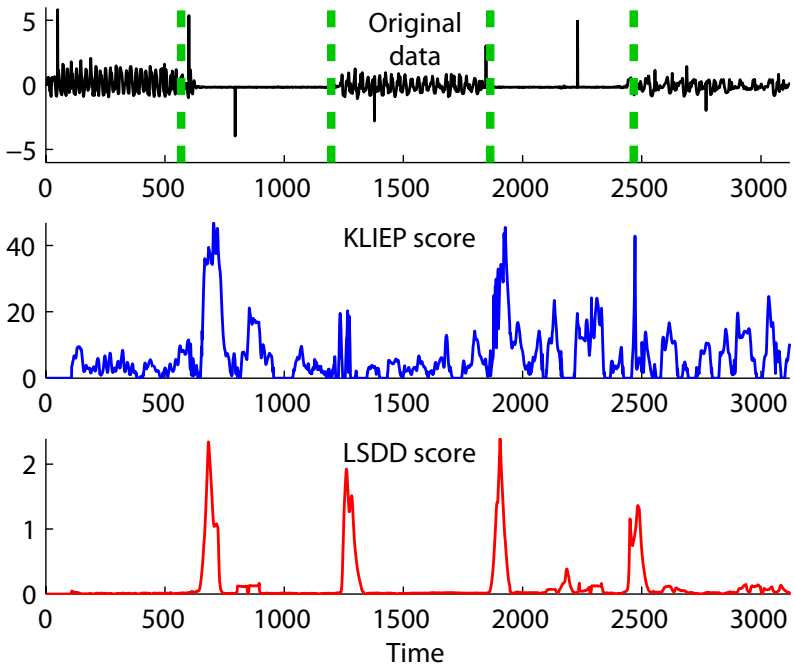
Our strategy is to compute a certain dissimilarity measure between two consecutive segments $\mathcal{Y}(t)$ and $\mathcal{Y}(t+r)$, and use it as the plausibility of change points (see Figure 11). As a dissimilarity measure, we use the L^2-distance estimated by LSDD and the KL-divergence estimated by the *KL importance estimation procedure* (KLIEP) (Sugiyama et al., 2008). We set $k = 10$ and $r = 50$.

We use two datasets. One is the *IPSJ SIG-SLP Corpora and Environments for Noisy Speech Recognition* (CENSREC) dataset[7]. This dataset was provided by the *National Institute of Informatics, Japan* that records human voice in a noisy environment such as a restaurant. Another dataset is the *Human Activity Sensing Consortium (HASC) challenge 2011*[8], which provides human activity information collected by portable three-axis accelerometers. Because the orientation of the accelerometers is not necessarily fixed, we take the ℓ_2-norm of the 3-dimensional data. The HASC dataset is relatively simple, so we artificially add zero-mean Gaussian noise with standard deviation 5 at each time point with probability 0.005.

[7] `http://research.nii.ac.jp/src/en/CENSREC-1-C.html`
[8] urlhttp://hasc.jp/hc2011/



(a) CENCREC dataset



(b) HASC dataset

Figure 12: Illustrative results of unsupervised change detection for (a) CENSREC speech data and (b) HASC dataset. Original time-series data is plotted in the top graphs, and change scores obtained by KLIEP and LSDD are plotted in the bottom graphs.

The top graphs in Figure 12 display the original time-series, where true change points were manually annotated. The time-series data in Figure 12(a) corresponds to a sequence of "noise", "speech", "noise", "speech", and "noise", whereas that in Figure 12(b) corresponds to a sequence of actions "jog", "stay", "stair down", "stay", and "stair up".

The bottom graphs in Figure 12 plot change scores obtained by each method. The results show that the LSDD-based change score indicates the existence of change points more clearly than the KLIEP-based approach. The superior performance of LSDD over the KLIEP-based change score would be brought by its robustness against outliers (see Section 4.1.2).

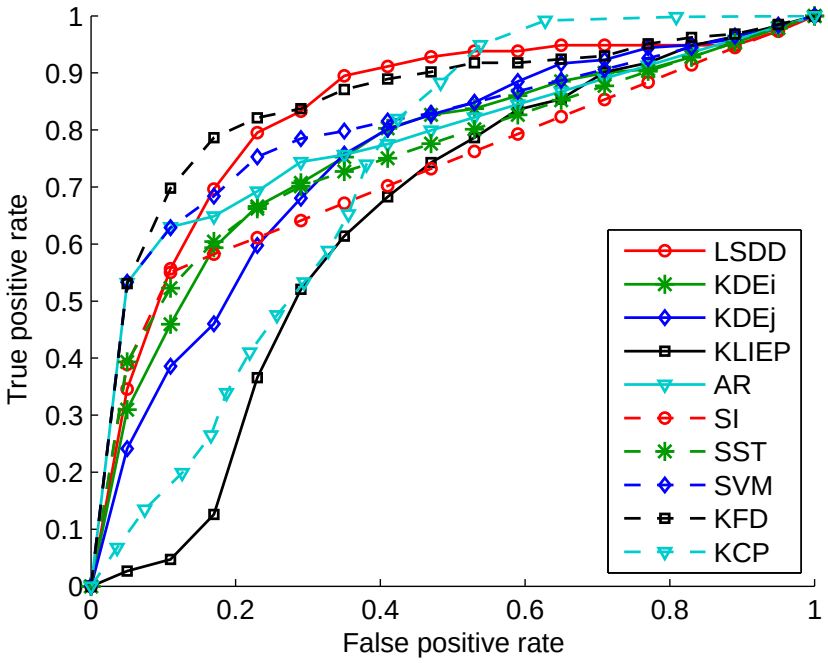
Finally, we compare the change-detection performance more systematically using the *receiver operating characteristic (ROC) curves* (i.e., the false positive rate vs. the true positive rate) and the *area under the ROC curve (AUC) values*. In addition to LSDD and KLIEP, we also test the L^2-distance estimated by KDEi and KDEj (see Section 4.1.1). Moreover, in our comparison, we also include native change detection methods based on autoregressive models (AR) (Takeuchi & Yamanishi, 2006), subspace identification (SI) (Kawahara et al., 2007), singular spectrum transformation (SST) (Moskvina & Zhigljavsky, 2003), one-class support vector machine (SVM) (Desobry et al., 2005), kernel Fisher discriminant analysis (KFD) (Harchaoui et al., 2009), and kernel change-point detection (KCP) (Arlot et al., 2012). Tuning parameters included in these methods were manually optimized.

We use 10 datasets taken from each of the CENSREC and HASC data collections. Mean ROC curves are plotted in Figure 13 and AUC values are described in Table 1. The experimental results show that LSDD tends to outperform other methods and is comparable to state-of-the-art native change-detection methods.

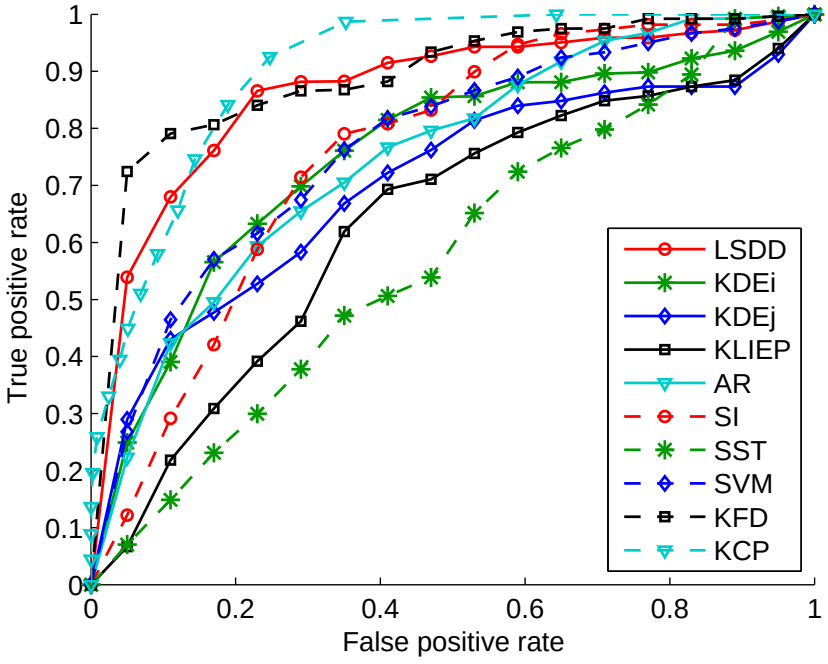
5 Conclusions

In this paper, we proposed a method for directly estimating the difference between two probability density functions without density estimation. The proposed method, called the *least-squares density-difference* (LSDD), was derived within the framework of kernel regularized least-squares estimation, and its solution can be computed analytically in a computationally efficient and stable manner. Furthermore, LSDD is equipped with cross-validation, and thus all tuning parameters such as the kernel width and the regularization parameter can be systematically and objectively optimized. We showed the asymptotic normality of LSDD in a parametric setup and derived a finite-sample error bound for LSDD in a non-parametric setup. In both cases, LSDD was shown to achieve the optimal convergence rates.

We also proposed an L^2-distance estimator based on LSDD, which nicely cancels the bias caused by regularization. The LSDD-based L^2-distance estimator was experimentally shown to be more accurate than differences of kernel density estimators and more robust against outliers than a Kullback-Leibler divergence estimator. However, we also experimentally observed that cross-validation of LSDD is sometimes rather unstable when



(a) CENCREC dataset



(b) HASC dataset

Figure 13: Mean ROC curves of unsupervised change detection.

Table 1: AUC values of unsupervised change detection. The best method and comparable ones in terms of mean AUC values by the *t-test* (Henkel, 1976) at the significance level 5% are indicated with boldface. "SE" stands for "Standard error".

(a) CENSREC dataset

Data ID	LSDD	KDEi	KDEj	KLIEP	AR	SI	SST	SVM	KFD	KCP
1	.888	.737	.731	.437	.769	.739	.507	.604	.881	.917
2	.871	.803	.706	.618	.777	.736	.541	.612	.912	.879
3	.910	.753	.690	.744	.762	.821	.616	.886	.876	.743
4	.936	.823	.578	.683	.776	.816	.723	.871	.981	.826
5	.878	.712	.799	.667	.768	.701	.625	.843	.880	.945
6	.830	.732	.711	.696	.679	.727	.484	.781	.841	.947
7	.813	.727	.737	.513	.727	.733	.612	.779	.938	.968
8	.889	.841	.734	.691	.783	.775	.526	.698	.934	.935
9	.828	.739	.586	.609	.776	.770	.609	.819	.922	.980
10	.943	.687	.773	.692	.670	.747	.551	.835	.889	.984
Mean	**.879**	.755	.705	.635	.749	.756	.580	.773	**.905**	**.913**
SE	.014	.016	.023	.030	.013	.012	.023	.032	.013	.024

(b) HASC dataset

Data ID	LSDD	KDEi	KDEj	KLIEP	AR	SI	SST	SVM	KFD	KCP
1	.792	.823	.753	.650	.860	.690	.806	.800	.885	.874
2	.842	.665	.741	.712	.733	.800	.745	.725	.904	.826
3	.773	.605	.536	.708	.910	.899	.807	.932	.707	.641
4	.921	.839	.837	.587	.816	.735	.685	.751	.903	.759
5	.838	.849	.859	.565	.831	.823	.809	.840	.961	.725
6	.834	.755	.781	.676	.868	.740	.736	.838	.871	.800
7	.841	.763	.598	.657	.807	.759	.797	.829	.770	.532
8	.878	.833	.857	.581	.629	.704	.682	.800	.852	.661
9	.864	.850	.866	.693	.738	.744	.781	.790	.842	.697
10	.847	.663	.680	.554	.796	.725	.790	.850	.866	.787
Mean	**.843**	.764	.751	.638	**.799**	.762	.764	**.815**	**.856**	.730
SE	.013	.029	.036	.020	.026	.020	.016	.018	.023	.032

the target density difference is zero (i.e., two distributions are equivalent). This can potentially cause performance degradation in two-sample homogeneity testing because estimation of zero density-difference is repeatedly executed when approximating the null distribution in the permutation-test framework. Stabilizing cross-validation and improving the accuracy of density-difference estimation when the target density-difference is zero is a remaining future work.

It is straightforward to extend the proposed LSDD method to the difference of weighted densities

$$\nu p(\boldsymbol{x}) - \nu' p'(\boldsymbol{x}),$$

where ν and ν' are scalars. Also, LSDD can be easily extended to estimate the weighted L^2-distance:

$$\int \Big(p(\boldsymbol{x}) - p'(\boldsymbol{x})\Big)^2 w(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$

where $w(\boldsymbol{x}) > 0$ is a weight function.

A related line of research to density-difference estimation is density-ratio estimation (Sugiyama et al., 2012a), which directly estimates the ratio of probability densities without separate density estimation (Qin, 1998; Huang et al., 2007; Bickel et al., 2007; Sugiyama et al., 2008; Kanamori et al., 2009; Sugiyama et al., 2012b). Potential weaknesses of density-ratio estimation are that density ratios can be unbounded even for simple cases (Cortes et al., 2010) and their estimation may suffer from outliers (Basu et al., 1998; Scott, 2001; Besbeas & Morgan, 2004).

To mitigate these weaknesses, the concept of *relative* density ratios was introduced recently, which "flatten" the density ratio $\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}$ as $\frac{p(\boldsymbol{x})}{\beta p(\boldsymbol{x}) + (1-\beta) p'(\boldsymbol{x})}$ for $0 \leq \beta < 1$ (Yamada et al., 2013). Even when the plain density ratio is unbounded, the relative density ratio is always bounded by $\frac{1}{\beta}$ for $\beta > 0$. Although estimation of relative density ratios as well as approximation of relative divergences was demonstrated to be more reliable (Yamada et al., 2013), there is no systematic method to choose the relativity parameter β, which is a critical limitation in practice.

On the other hand, density-difference estimation is more advantageous than density-ratio estimation in the senses that density differences are always bounded as long as each density is bounded, their estimation is robust against outliers (Basu et al., 1998; Scott, 2001; Besbeas & Morgan, 2004), and there exist no tuning parameters such as the relativity parameter β. However, a potential weakness of density differences is that they cannot be used for importance sampling (Sugiyama & Kawanabe, 2012) and conditional probability estimation (Sugiyama et al., 2010b; Sugiyama, 2010), which are promising usages of density-ratio estimation. Thus, further exploring usages of density-difference estimation, particularly in the tasks that density-ratio estimation cannot be used for, is a promising future work.

A simple application of density-difference estimation would be probabilistic pattern recognition, because the sign of the density difference gives the Bayes-optimal decision (Duda et al., 2001). Furthermore, in the context of pattern recognition with a reject option, the density difference can be used for finding the optimal rejection threshold

(Chow, 1970). In the future work, we will investigate the behavior of LSDD in probabilistic pattern recognition theoretically and experimentally.

Density-difference estimation is a novel research paradigm in machine learning, and we have proposed a simple but useful method for this emerging topic. Our future work will develop more powerful algorithms for density-difference estimation. For example, considering more general loss functions than the squared loss (Sugiyama et al., 2012b) and incorporating dimension reduction (von Bünau et al., 2009; Sugiyama et al., 2010a; Sugiyama et al., 2011; Yamada & Sugiyama, 2011) would be interesting directions to pursue. Exploring a wide variety of real-world applications is also an important future work.

Acknowledgments

The authors would like to thank Wittawat Jitkrittum and John Quinn for their comments and Zaïd Harchaoui for providing us a program code of kernel change-point detection. Masashi Sugiyama was supported by MEXT KAKENHI 23300069 and AOARD, Takafumi Kanamori was supported by MEXT KAKENHI 24500340, Taiji Suzuki was supported by MEXT KAKENHI 22700289, the Aihara Project, the FIRST program from JSPS initiated by CSTP, and the Global COE Program "The research and training center for new development in mathematics", MEXT, Japan, Marthinus Christoffel du Plessis was supported by MEXT Scholarship, Song Liu was supported by the JST PRESTO program, and Ichiro Takeuchi was supported by MEXT KAKENHI 23700165.

A Technical Details of Non-Parametric Convergence Analysis in Section 2.3.2

First, we define the linear operators P_n, P, P'_n, P', Q_n, Q as

$$P_n f := \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i), \quad Pf := \int_{\mathbb{R}^d} f(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$
$$P'_n f := \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}'_i), \quad P'f := \int_{\mathbb{R}^d} f(\boldsymbol{x})p'(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$
$$Q_n f := P_n f - P'_n f, \quad Qf := Pf - P'f.$$

Let $\mathcal{H}_\gamma$ be an RKHS endowed with the Gaussian kernel with width γ:

$$k_\gamma(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{\gamma^2}\right).$$

A density-difference estimator $\widehat{f}$ is obtained as

$$\widehat{f} := \operatorname*{argmin}_{f \in \mathcal{H}_\gamma} \left[\|f\|^2_{L^2(\mathbb{R}^d)} - 2Q_n f + \lambda \|f\|^2_{\mathcal{H}_\gamma}\right].$$

We assume the following conditions:

Assumption 1. *The densities are bounded: There exists a constant M such that*

$$\|p\|_\infty \leq M \quad \textit{and} \quad \|p'\|_\infty \leq M.$$

The density difference $f = p - p'$ is a member of Besov space with regularity α. That is, $f \in B^\alpha_{2,\infty}$ where $B^\alpha_{2,\infty}$ is the Besov space with regularity α, and

$$\|f\|_{B^\alpha_{2,\infty}} := \|f\|_{L_2(\mathbb{R}^d)} + \sup_{t>0}(t^{-\alpha}\omega_{r,L_2(\mathbb{R}^d)}(f,t)) < c \quad \textit{for } r = \lfloor\alpha\rfloor + 1,$$

where $\lfloor\alpha\rfloor$ denotes the largest integer less than or equal to α and $\omega_{r,L_2(\mathbb{R}^d)}$ is the r-th modulus of smoothness (see Eberts and Steinwart (2011) for the definitions).

Then we have the following theorem.

Theorem 2. *Suppose Assumption 1 is satisfied. Then, for all $\epsilon > 0$ and $p \in (0,1)$, there exists a constant $K > 0$ depending on M, c, ϵ, p such that for all $n \geq 1$, $\tau \geq 1$, and $\lambda > 0$, the LSDD estimator $\widehat{f}$ in $\mathcal{H}_\gamma$ satisfies*

$$\|\widehat{f} - f\|^2_{L^2(\mathbb{R}^d)} + \lambda\|\widehat{f}\|^2_{\mathcal{H}_\gamma} \leq K\left(\lambda\gamma^{-d} + \gamma^{2\alpha} + \frac{\gamma^{-(1-p)(1+\epsilon)d}}{\lambda^p n} + \frac{\gamma^{-\frac{2(1-p)d}{1+p}(1+\epsilon+\frac{1-p}{4})}}{\lambda^{\frac{3p-p^2}{1+p}} n^{\frac{2}{1+p}}} + \frac{\tau}{n^2\lambda} + \frac{\tau}{n}\right),$$

with probability not less than $1 - 4e^{-\tau}$.

To prove this, we utilize the technique developed in Eberts and Steinwart (2011) for a regression problem.

Proof. First, note that

$$\|\widehat{f}\|^2_{L^2(\mathbb{R}^d)} - 2Q_n\widehat{f} + \|f\|^2_{L^2(\mathbb{R}^d)} + \lambda\|\widehat{f}\|^2_{\mathcal{H}_\gamma} \leq \|f_0\|^2_{L^2(\mathbb{R}^d)} - 2Q_n f_0 + \|f\|^2_{L^2(\mathbb{R}^d)} + \lambda\|f_0\|^2_{\mathcal{H}_\gamma}.$$

Therefore, we have

$$\begin{aligned}
&\|\widehat{f} - f\|^2_{L^2(\mathbb{R}^d)} + \lambda\|\widehat{f}\|^2_{\mathcal{H}_\gamma} \\
&\quad = \|\widehat{f}\|^2_{L^2(\mathbb{R}^d)} - 2Q_n\widehat{f} + \|f\|^2_{L^2(\mathbb{R}^d)} + 2(Q_n - Q)\widehat{f} + \lambda\|\widehat{f}\|^2_{\mathcal{H}_\gamma} \\
&\quad \leq \|f_0\|^2_{L^2(\mathbb{R}^d)} - 2Q_n f_0 + \|f\|^2_{L^2(\mathbb{R}^d)} + 2(Q_n - Q)\widehat{f} + \lambda\|\widehat{f}\|^2_{\mathcal{H}_\gamma} \\
&\quad = \|f_0\|^2_{L^2(\mathbb{R}^d)} - 2Qf_0 + \|f\|^2_{L^2(\mathbb{R}^d)} + 2(Q_n - Q)(\widehat{f} - f_0) + \lambda\|\widehat{f}\|^2_{\mathcal{H}_\gamma} \\
&\quad = \|f_0 - f\|^2_{L^2(\mathbb{R}^d)} + 2(Q_n - Q)(\widehat{f} - f) + 2(Q_n - Q)(f - f_0) + \lambda\|\widehat{f}\|^2_{\mathcal{H}_\gamma}. \qquad (16)
\end{aligned}$$

Let

$$K(\boldsymbol{x}) := \sum_{j=1}^r \binom{r}{j}(-1)^{1-j}\frac{1}{j^d}\left(\frac{2}{\gamma\sqrt{\pi}}\right)^{\frac{d}{2}} \exp\left(-\frac{2\|\boldsymbol{x}\|^2}{j^2\gamma^2}\right),$$

and $\widetilde{f}(\boldsymbol{x}) := (\gamma\sqrt{\pi})^{-\frac{d}{2}} f$. Using K and $\widetilde{f}$, we define

$$f_0 := K * \widetilde{f} := \int_{\mathbb{R}^d} \widetilde{f}(y) K(x-y) \mathrm{d}y,$$

i.e., f_0 is the convolution of K and $\widetilde{f}$. Because of Lemma 2 in Eberts and Steinwart (2011), we have $f_0 \in \mathcal{H}_\gamma$ and

$$\begin{aligned}
\|f_0\|_{\mathcal{H}_\gamma} &\leq (2^r - 1)\|\widetilde{f}\|_{L^2(\mathbb{R}^d)} \quad (\because \text{Lemma 2 of Eberts and Steinwart (2011)}) \\
&\leq (2^r - 1)(\gamma\sqrt{\pi})^{-\frac{d}{2}} \|f\|_{L^2(\mathbb{R}^d)} \\
&\leq (2^r - 1)(\gamma\sqrt{\pi})^{-\frac{d}{2}} (\|p\|_{L^2(\mathbb{R}^d)} + \|p'\|_{L^2(\mathbb{R}^d)}) \\
&\leq (2^r - 1)(\gamma\sqrt{\pi})^{-\frac{d}{2}} 2\sqrt{M}. \qquad (17)
\end{aligned}$$

Moreover, Lemma 3 in Eberts and Steinwart (2011) gives

$$\|f_0\|_\infty \leq (2^r - 1)\|f\|_\infty \leq (2^r - 1)M, \qquad (18)$$

and Lemma 1 in Eberts and Steinwart (2011) yields that there exists a constant $C_{r,2}$ such that

$$\|f_0 - f\|^2_{L^2(\mathbb{R}^d)} \leq C_{r,2}\omega^2_{r,L^2(\mathbb{R}^d)}(f, \frac{\gamma}{2}) \leq C_{r,2}c^2\gamma^{2\alpha}. \qquad (19)$$

Now, following a similar line to Theorem 3 in Eberts and Steinwart (2011), we can show that, for all $\epsilon > 0$ and $p \in (0, 1)$, there exists a constant $C_{\epsilon,p}$ such that

$$|(P_n - P)(\widehat{f} - f)| \leq \widehat{f} - f.$$

To bound this, we derive the tail probability of

$$(P_n - P)\left(\frac{\widehat{f} - f}{\|\widehat{f} - f\|^2_{L^2(\mathbb{R}^d)} + \lambda\|\widehat{f}\|^2_{\mathcal{H}_\gamma} + r}\right),$$

where $r > 0$ is a positive real number such that $r > r^*$ for

$$r^* = \min_{f \in \mathcal{H}_\gamma} \|f - f\|^2_{L^2(\mathbb{R}^d)} + \lambda\|f\|^2_{\mathcal{H}_\gamma}.$$

Let

$$g_{f,r} = \frac{f - f}{\|f - f\|^2_{L^2(\mathbb{R}^d)} + \lambda\|f\|^2_{\mathcal{H}_\gamma} + r}$$

for $f \in \mathcal{H}_\gamma$ and $r > r^*$. Then we have

$$\begin{aligned}
\|g_{f,r}\|_\infty &\leq \frac{\|f\|_\infty + \|f\|_\infty}{\|f - f\|^2_{L^2(\mathbb{R}^d)} + \lambda\|f\|^2_{\mathcal{H}_\gamma} + r} \leq \frac{\|f\|_{\mathcal{H}_\gamma} + \|f\|_\infty}{\|f - f\|^2_{L^2(\mathbb{R}^d)} + \lambda\|f\|^2_{\mathcal{H}_\gamma} + r} \\
&\leq \frac{1}{\lambda\|f\|_{\mathcal{H}_\gamma} + r/\|f\|_{\mathcal{H}_\gamma}} + \frac{M}{r} \leq \frac{1}{2\sqrt{r\lambda}} + \frac{M}{r},
\end{aligned}$$

and

$$Pg_{f,r}^2 = \frac{P(f-f)^2}{(\|f-f\|_{L^2(\mathbb{R}^d)}^2 + \lambda\|f\|_{\mathcal{H}_\gamma}^2 + r)^2} \leq \frac{M\|f-f\|_{L^2(\mathbb{R}^d)}^2}{(\|f-f\|_{L^2(\mathbb{R}^d)}^2 + \lambda\|f\|_{\mathcal{H}_\gamma}^2 + r)^2} \leq \frac{M}{r}.$$

Here, let

$$\mathcal{F}_r := \{f \in \mathcal{H}_\gamma \mid \|f-f\|_{L^2(\mathbb{R}^d)}^2 + \lambda\|f\|_{\mathcal{H}_\gamma}^2 \leq r\},$$

and we assume that there exists a function such that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_r} |(P_n - P)(f-f)|\right] \leq \varphi_n(r),$$

where $\mathbb{E}$ denotes the expectation over all samples. Then, by the peeling device (see Theorem 7.7 in Steinwart & Christmann, 2008), we have

$$\mathbb{E}\sup_{f\in\mathcal{H}_\gamma} |(P_n - P)g_{f,r}| \leq \frac{8\varphi(r)}{r}.$$

Therefore, by Talagrand's concentration inequality, we have

$$\Pr\left[\sup_{f\in\mathcal{H}_\gamma} |(P_n - P)g_{f,r}| < \frac{10\varphi_n(r)}{r} + \sqrt{\frac{2M\tau}{nr}} + \frac{14\tau}{3n}\left(\frac{1}{2\sqrt{r\lambda}} + \frac{M}{r}\right)\right] \geq 1 - e^{-\tau}, \quad (20)$$

where $\Pr[\cdot]$ denotes the probability of an event.

From here on, we give an upper bound of φ_n. The RKHS $\mathcal{H}_\gamma$ can be embedded in arbitrary Sobolev space $W^m(\mathbb{R}^d)$. Indeed, by the proof of Theorem 3.1 in Steinwart and Scovel (2007), we have

$$\|f\|_{W^m(\mathbb{R}^d)} \leq C_m \gamma^{-\frac{m}{2}+\frac{d}{4}} \|f\|_{\mathcal{H}_\gamma}$$

for all $f \in \mathcal{H}_\gamma$. Moreover, the theories of interpolation spaces give that, for all $f \in W^m(\mathbb{R}^d)$, the supremum norm of f can be bounded as

$$\|f\|_\infty \leq C'_m \|f\|_{L^2(\mathbb{R}^d)}^{1-\frac{d}{2m}} \|f\|_{W^m(\mathbb{R}^d)}^{\frac{d}{2m}},$$

if $d < 2m$. Here we set $m = \frac{d}{2p}$. Then we have

$$\|f\|_\infty \leq C''_p \|f\|_{L^2(\mathbb{R}^d)}^{1-p} \|f\|_{\mathcal{H}_\gamma}^p \gamma^{-\frac{d(1-p)}{4}}.$$

Now, because $\mathcal{F}_r \subset (r/\lambda)^{1/2}\mathcal{B}_{\mathcal{H}_\gamma}$ and

$$P(f-f)^2 \leq M\|f-f\|_{L^2(\mathbb{R}^d)}^2 \leq Mr \quad \text{for } f \in \mathcal{F}_r$$

hold from Theorem 7.16 and Theorem 7.34 in Steinwart and Christmann (2008), we can take

$$\varphi_n(r) = \max\left\{ C_{1,p,\epsilon}\gamma^{-\frac{(1-p)(1+\epsilon)d}{2}} \left(\frac{r}{\lambda}\right)^{\frac{p}{2}} (Mr)^{\frac{1-p}{2}} n^{-1/2}, \right.$$
$$\left. C_{2,p,\epsilon}\gamma^{-\frac{(1-p)(1+\epsilon)d}{1+p}} \left(\frac{r}{\lambda}\right)^{\frac{p}{1+p}} \left[\left(\frac{r}{\lambda}\right)^{\frac{p}{2}} \gamma^{-\frac{d(1-p)}{4}} r^{\frac{1-p}{2}}\right]^{\frac{1-p}{1+p}} n^{-1/(1+p)} \right\},$$

where $\epsilon > 0$ and $p \in (0,1)$ are arbitrary and $C_{1,p,\epsilon}, C_{2,p,\epsilon}$ are constants depending on p, ϵ. In the same way, we can also obtain a bound of $\sup_{f \in \mathcal{H}_\gamma} |(P'_n - P')g_{f,r}|$.

If we set r to satisfy

$$\frac{1}{8} \geq \frac{10\varphi_n(r)}{r} + \sqrt{\frac{2M\tau}{nr}} + \frac{14\tau}{3n}\left(\frac{1}{2\sqrt{r\lambda}} + \frac{M}{r}\right), \tag{21}$$

then we have

$$|(Q_n - Q)(\widehat{f} - f)| \leq \frac{1}{4}\left(r + \|\widehat{f} - f\|^2_{L^2(\mathbb{R}^d)} + \lambda\|\widehat{f}\|_{\mathcal{H}_\gamma}\right) \tag{22}$$

with probability $1 - 2e^{-\tau}$. To satisfy Eq.(21), it suffices to set

$$r = C\left(\frac{\gamma^{-(1-p)(1+\epsilon)d}}{\lambda^p n} + \frac{\gamma^{-\frac{2(1-p)d}{1+p}(1+\epsilon+\frac{1-p}{4})}}{\lambda^{\frac{3p-p^2}{1+p}} n^{\frac{2}{1+p}}} + \frac{\tau}{n^2\lambda} + \frac{\tau}{n}\right), \tag{23}$$

where C is a sufficiently large constant depending on M, ϵ, p.

Finally, we bound the term $(Q_n - Q)(f_0 - f)$. By Bernstein's inequality, we have

$$\begin{aligned}
|(P_n - P)(f_0 - f)| &\leq C\left(\|f - f_0\|_{L_2(P)}\sqrt{\frac{\tau}{n}} + \frac{2^r M\tau}{n}\right) \\
&\leq C\left(\sqrt{2M}\|f - f_0\|_{L^2(\mathbb{R}^d)}\sqrt{\frac{\tau}{n}} + \frac{2^r M\tau}{n}\right) \\
&\leq C\left(\|f - f_0\|^2_{L^2(\mathbb{R}^d)} + \frac{2M\tau}{n} + \frac{2^r M\tau}{n}\right),
\end{aligned} \tag{24}$$

with probability $1 - e^{-\tau}$, where C is a universal constant. In a similar way, we can also obtain

$$|(P'_n - P')(f_0 - f)| \leq C\left(\|f - f_0\|^2_{L^2(\mathbb{R}^d)} + \frac{2M\tau}{n} + \frac{2^r M\tau}{n}\right).$$

Combining these inequalities, we have

$$|(Q_n - Q)(f_0 - f)| \leq C\left(\|f - f_0\|^2_{L^2(\mathbb{R}^d)} + \frac{2^r M\tau}{n}\right), \tag{25}$$

with probability $1 - 2e^{-\tau}$, where C is a universal constant.

Substituting Eqs.(22) and (25) into Eq.(16), we have

$$\begin{aligned}
&\|\widehat{f} - f\|^2_{L^2(\mathbb{R}^d)} + \lambda\|\widehat{f}\|^2_{\mathcal{H}_\gamma} \\
&\quad\leq 2\left\{\|f_0 - f\|^2_{L^2(\mathbb{R}^d)} + C\left(\|f - f_0\|^2_{L^2(\mathbb{R}^d)} + \frac{2^r M\tau}{n}\right) + r + \lambda\|f_0\|_{\mathcal{H}_\gamma}\right\},
\end{aligned}$$

with probability $1 - 4e^{-\tau}$. Moreover, by Eqs.(19) and (17), the right-hand side is further bounded by

$$\|\widehat{f} - f\|^2_{L^2(\mathbb{R}^d)} + \lambda\|\widehat{f}\|^2_{\mathcal{H}_\gamma} \leq C\left\{\gamma^{2\alpha} + r + \lambda\gamma^{-d} + \frac{1+\tau}{n}\right\},$$

Finally, substituting Eq.(23) into the right-hand side, we have

$$\|\widehat{f} - f\|^2_{L^2(\mathbb{R}^d)} + \lambda\|\widehat{f}\|^2_{\mathcal{H}_\gamma} \leq C\left\{\gamma^{2\alpha} + \frac{\gamma^{-(1-p)(1+\epsilon)d}}{\lambda^p n} + \frac{\gamma^{-\frac{2(1-p)d}{1+p}(1+\epsilon+\frac{1-p}{4})}}{\lambda^{\frac{3p-p^2}{1+p}} n^{\frac{2}{1+p}}} + \lambda\gamma^{-d} + \frac{\tau}{\lambda n^2} + \frac{\tau}{n}\right\}$$

with probability $1 - 4e^{-\tau}$ for $\tau \geq 1$. This gives the assertion. □

If we set

$$\lambda = n^{-\frac{2\alpha+d}{(2\alpha+d)(1+p)+(\epsilon-p+\epsilon p)}}, \quad \gamma = n^{-\frac{1}{(2\alpha+d)(1+p)+(\epsilon-p+\epsilon p)}},$$

and take ϵ and p sufficiently small, then we immediately have the following corollary.

Corollary 1. *Suppose Assumption 1 is satisfied. Then, for all $\rho, \rho' > 0$, there exists a constant $K > 0$ depending on M, c, ρ, and ρ' such that, for all $n \geq 1$ and $\tau \geq 1$, the density-difference estimator $\widehat{f}$ with appropriate choice of γ and λ satisfies*

$$\|\widehat{f} - f\|^2_{L^2(\mathbb{R}^d)} + \lambda\|\widehat{f}\|^2_{\mathcal{H}_\gamma} \leq K\left(n^{-\frac{2\alpha}{2\alpha+d}+\rho} + \frac{\tau}{n^{1-\rho'}}\right)$$

with probability not less than $1 - 4e^{-\tau}$.

Note that $n^{-\frac{2\alpha}{2\alpha+d}}$ is the optimal learning rate to estimate a function in $B^\alpha_{2,\infty}$ (Eberts & Steinwart, 2011). Therefore, the density-difference estimator with a Gaussian kernel achieves the optimal learning rate by appropriately choosing the regularization parameter and the Gaussian width. Because the learning rate depends on α, the LSDD estimator has adaptivity to the smoothness of the true function.

Our analysis heavily relies on the techniques developed in Eberts and Steinwart (2011) for a regression problem. The main difference is that the analysis in their paper involves a clipping procedure, which stems from the fact that the analyzed estimator requires an empirical approximation of the expectation of the square term. The Lipschitz continuity of the square function $f \mapsto f^2$ is utilized to investigate this term, and the clipping procedure is used to ensure the Lipschitz continuity. On the other hand, in the current paper, we can exactly compute $\|f\|^2_{L^2(\mathbb{R}^d)}$ so that we do not need the Lipschitz continuity.

B Derivation of Eq.(13)

When λ (≥ 0) is small, $(\boldsymbol{H} + \lambda\boldsymbol{I}_b)^{-1}$ can be expanded as

$$(\boldsymbol{H} + \lambda\boldsymbol{I}_b)^{-1} = \boldsymbol{H}^{-1} - \lambda\boldsymbol{H}^{-2} + o_p(\lambda),$$

where o_p denotes the probabilistic order. Then Eq.(12) can be expressed as

$$\begin{aligned}
&\beta\widehat{\boldsymbol{h}}^\top\widehat{\boldsymbol{\theta}} + (1-\beta)\widehat{\boldsymbol{\theta}}^\top\boldsymbol{H}\widehat{\boldsymbol{\theta}} \\
&\quad= \beta\widehat{\boldsymbol{h}}^\top(\boldsymbol{H}+\lambda\boldsymbol{I}_b)^{-1}\widehat{\boldsymbol{h}} + (1-\beta)\widehat{\boldsymbol{h}}^\top(\boldsymbol{H}+\lambda\boldsymbol{I}_b)^{-1}\boldsymbol{H}(\boldsymbol{H}+\lambda\boldsymbol{I}_b)^{-1}\widehat{\boldsymbol{h}} \\
&\quad= \beta\widehat{\boldsymbol{h}}^\top\boldsymbol{H}^{-1}\widehat{\boldsymbol{h}} - \lambda\beta\widehat{\boldsymbol{h}}^\top\boldsymbol{H}^{-2}\widehat{\boldsymbol{h}} \\
&\qquad+ (1-\beta)\widehat{\boldsymbol{h}}^\top\boldsymbol{H}^{-1}\widehat{\boldsymbol{h}} - 2\lambda(1-\beta)\widehat{\boldsymbol{h}}^\top\boldsymbol{H}^{-2}\widehat{\boldsymbol{h}} + o_p(\lambda) \\
&\quad= \widehat{\boldsymbol{h}}^\top\boldsymbol{H}^{-1}\widehat{\boldsymbol{h}} - \lambda(2-\beta)\widehat{\boldsymbol{h}}^\top\boldsymbol{H}^{-2}\widehat{\boldsymbol{h}} + o_p(\lambda),
\end{aligned}$$

which concludes the proof.

C Derivation of Eq.(14)

Because $\mathbb{E}[\widehat{\boldsymbol{h}}] = \boldsymbol{h}$, we have

$$\begin{aligned}
\mathbb{E}[\widehat{\boldsymbol{h}}^\top\boldsymbol{H}^{-1}\widehat{\boldsymbol{h}} - \boldsymbol{h}^\top\boldsymbol{H}^{-1}\boldsymbol{h}] &= \mathbb{E}[(\widehat{\boldsymbol{h}}-\boldsymbol{h})^\top\boldsymbol{H}^{-1}(\widehat{\boldsymbol{h}}-\boldsymbol{h})] \\
&= \mathrm{tr}\left(\boldsymbol{H}^{-1}\mathbb{E}[(\widehat{\boldsymbol{h}}-\boldsymbol{h})(\widehat{\boldsymbol{h}}-\boldsymbol{h})^\top]\right) \\
&= \mathrm{tr}\left(\boldsymbol{H}^{-1}\left(\frac{1}{n}\boldsymbol{V}_p[\boldsymbol{\psi}] + \frac{1}{n'}\boldsymbol{V}_{p'}[\boldsymbol{\psi}]\right)\right),
\end{aligned}$$

which concludes the proof.

References

Anderson, N., Hall, P., & Titterington, D. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, *50*, 41–54.

Arlot, S., Celisse, A., & Harchaoui, Z. (2012). *Kernel change-point detection* (Technical Report 1202.3878). arXiv.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, *68*, 337–404.

Atif, J., Ripoche, X., & Osorio, A. (2003). Non-rigid medical image registration by maximisation of quadratic mutual information. *IEEE 29th Annual Northeast Bioengineering Conference* (pp. 32–40).

Baranchik, A. J. (1964). *Multiple regression and estimation of the mean of a multivariate normal distribution* (Technical Report 51). Department of Statistics, Stanford University, Stanford, CA, USA.

Basu, A., Harris, I. R., Hjort, N. L., & Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, *85*, 549–559.

Besbeas, P., & Morgan, B. J. T. (2004). Integrated squared error estimation of normal mixtures. *Computational Statistics & Data Analysis*, *44*, 517–526.

Bickel, S., Brückner, M., & Scheffer, T. (2007). Discriminative learning for differing training and test distributions. *Proceedings of the 24th International Conference on Machine Learning (ICML2007)* (pp. 81–88).

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory* (pp. 144–152). ACM Press.

Chapelle, O., Schölkopf, B., & Zien, A. (Eds.). (2006). *Semi-supervised learning*. Cambridge, MA, USA: MIT Press.

Chow, C. K. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, *16*, 41–46.

Cortes, C., Mansour, Y., & Mohri, M. (2010). Learning bounds for importance weighting. *Advances in Neural Information Processing Systems 23* (pp. 442–450).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Desobry, F., Davy, M., & Doncarli, C. (2005). An online kernel change detection algorithm. *IEEE Transactions on Signal Processing*, *53*, 2961–2974.

du Plessis, M. C. (2013). Class-prior change labeling. Unpublished research memo.

du Plessis, M. C., & Sugiyama, M. (2012). Semi-supervised learning of class balance under class-prior change by distribution matching. *Proceedings of 29th International Conference on Machine Learning (ICML2012)* (pp. 823–830). Edinburgh, Scotland.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. New York, NY, USA: Wiley. Second edition.

Duong, T., Koch, I., & Wand, M. P. (2009). Highest density difference region estimation with application to flow cytometric data. *Biometrical Journal*, *51*, 504–521.

Eberts, M., & Steinwart, I. (2011). Optimal learning rates for least squares SVMs using Gaussian kernels. *Advances in Neural Information Processing Systems 24* (pp. 1539–1547).

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY, USA: Chapman & Hall/CRC.

Farrell, R. H. (1972). On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *The Annals of Mathematical Statistics*, *43*, 170–180.

Gray, D. M., & Principe, J. C. (2010). Quadratic mutual information for dimensionality reduction and classification. *Proceedings of SPIE* (p. 76960D).

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., & Schölkopf, B. (2009). Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning* (pp. 131–160). Cambridge, MA, USA: MIT Press.

Hall, P., & Wand, M. P. (1988). On nonparametric discrimination using density differences. *Biometrika*, *75*, 541–547.

Harchaoui, Z., Bach, F., & Moulines, E. (2009). Kernel change-point analysis. *Advances in Neural Information Processing Systems 21* (pp. 609–616).

Härdle, W., Müller, M., Sperlich, S., & Werwatz, A. (2004). *Nonparametric and semiparametric models*. Berlin, Germany: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY, USA: Springer.

Henkel, R. E. (1976). *Tests of significance*. Beverly Hills, CA, USA.: SAGE Publication.

Huang, J., Smola, A., Gretton, A., Borgwardt, K. M., & Schölkopf, B. (2007). Correcting sample selection bias by unlabeled data. *Advances in Neural Information Processing Systems 19* (pp. 601–608). Cambridge, MA, USA: MIT Press.

Kanamori, T., Hido, S., & Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, *10*, 1391–1445.

Kanamori, T., Suzuki, T., & Sugiyama, M. (2012). Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, *86*, 335–367.

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

Kawahara, Y., Yairi, T., & Machida, K. (2007). Change-point detection in time-series data based on subspace identification. *Proceedings of the 7th IEEE International Conference on Data Mining* (pp. 559–564).

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79–86.

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

Liu, S., Yamada, M., Collier, N., & Sugiyama, M. (2013). Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, *43*, 72–83.

Moskvina, V., & Zhigljavsky, A. A. (2003). An algorithm based on singular spectrum analysis for change-point detection. *Communication in Statistics: Simulation & Computation*, *32*, 319–352.

Nguyen, X., Wainwright, M. J., & Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, *56*, 5847–5861.

Parzen, E. (1962). On the estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, *33*, 1065–1076.

Pérez-Cruz, F. (2008). Kullback-Leibler divergence estimation of continuous distributions. *Proceedings of IEEE International Symposium on Information Theory* (pp. 1666–1670). Toronto, Canada.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, *85*, 619–630.

Rao, C. R. (1965). *Linear statistical inference and its applications*. New York, NY, USA: Wiley.

Rifkin, R., Yeo, G., & Poggio, T. (2003). Regularized least-squares classification. In J. A. K. Suykens, G. Horvath, S. Basu, C. Micchelli and J. Vandewalle (Eds.), *Advances in learning theory: Methods, models and applications*, vol. 190 of *NATO Science Series III: Computer & Systems Sciences*, 131–154. Amsterdam, the Netherlands: IOS Press.

Rockafellar, R. T. (1970). *Convex analysis*. Princeton, NJ, USA: Princeton University Press.

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

Scott, D. W. (2001). Parametric statistical modeling by minimum integrated square error. *Technometrics*, *43*, 274–285.

Silva, J., & Narayanan, S. S. (2010). Information divergence estimation based on data-dependent partitions. *Journal of Statistical Planning and Inference*, *140*, 3180–3198.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London, UK: Chapman and Hall.

Steinwart, I., & Christmann, A. (2008). *Support vector machines*. New York, NY, USA: Springer.

Steinwart, I., & Scovel, C. (2007). Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, *35*, 575–607.

Sugiyama, M. (2010). Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting. *IEICE Transactions on Information and Systems*, *E93-D*, 2690–2701.

Sugiyama, M., & Kawanabe, M. (2012). *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. Cambridge, Massachusetts, USA: MIT Press.

Sugiyama, M., Kawanabe, M., & Chui, P. L. (2010a). Dimensionality reduction for density ratio estimation in high-dimensional spaces. *Neural Networks*, *23*, 44–59.

Sugiyama, M., Krauledat, M., & Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, *8*, 985–1005.

Sugiyama, M., Suzuki, T., & Kanamori, T. (2012a). *Density ratio estimation in machine learning*. Cambridge, UK: Cambridge University Press.

Sugiyama, M., Suzuki, T., & Kanamori, T. (2012b). Density ratio matching under the Bregman divergence: A unified framework of density ratio estimation. *Annals of the Institute of Statistical Mathematics*, *64*, 1009–1044.

Sugiyama, M., Suzuki, T., Kanamori, T., du Plessis, M. C., Liu, S., & Takeuchi, I. (2012c). Density-difference estimation. *Advances in Neural Information Processing Systems 25*.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., & Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, *60*, 699–746.

Sugiyama, M., Takeuchi, I., Suzuki, T., Kanamori, T., Hachiya, H., & Okanohara, D. (2010b). Least-squares conditional density estimation. *IEICE Transactions on Information and Systems*, *E93-D*, 583–594.

Sugiyama, M., Yamada, M., von Bünau, P., Suzuki, T., Kanamori, T., & Kawanabe, M. (2011). Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. *Neural Networks*, *24*, 183–198.

Suzuki, T., & Sugiyama, M. (2013). Sufficient dimension reduction via squared-loss mutual information estimation. *Neural Computation*, *25*, 725–758.

Takeuchi, Y., & Yamanishi, K. (2006). A unifying framework for detecting outliers and change points from non-stationary time series data. *IEEE Transactions on Knowledge and Data Engineering*, *18*, 482–489.

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

Vapnik, V. N. (1998). *Statistical learning theory*. New York, NY, USA: Wiley.

von Bünau, P., Meinecke, F. C., Király, F. J., & Müller, K.-R. (2009). Finding stationary subspaces in multivariate time series. *Physical Review Letters*, *103*, 214101.

Wang, Q., Kulkarmi, S. R., & Verdú, S. (2005). Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, *51*, 3064–3074.

Yamada, M., & Sugiyama, M. (2011). Direct density-ratio estimation with dimensionality reduction via hetero-distributional subspace analysis. *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI2011)* (pp. 549–554). San Francisco, California, USA: The AAAI Press.

Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H., & Sugiyama, M. (2013). Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, *25*, 1324–1370.

Yamanaka, M., Matsugu, M., & Sugiyama, M. (2013a). Detection of activities and events without explicit categorization. *IPSJ Transactions on Mathematical Modeling and Its Applications*. to appear.

Yamanaka, M., Matsugu, M., & Sugiyama, M. (2013b). Salient object detection based on direct density-ratio estimation. *IPSJ Transactions on Mathematical Modeling and Its Applications*. to appear.