Density-Difference Estimation

Masashi Sugiyama[1] **Takafumi Kanamori**[2] **Taiji Suzuki**[3]
Marthinus Christoffel du Plessis[1] **Song Liu**[1] **Ichiro Takeuchi**[4]
[1]Tokyo Institute of Technology, Japan [2]Nagoya University, Japan
[3]University of Tokyo, Japan [4]Nagoya Institute of Technology, Japan

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.

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]: Instead of separately estimating two probability distributions of patterns for positive and negative classes, the support vector machine directly learns the boundary between the two 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 [2, 3]: 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 differences would be more desirable than density ratios because density ratios can diverge to infinity even under a mild condition (e.g., two Gaussians [4]), whereas density differences are always finite as long as each density is bounded. Density differences can be used for solving various machine learning tasks such as class-balance estimation under class-prior change [5] and change-point detection in time series [6].

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 with in the framework of kernel regularized least-squares estimation, and thus it inherits various useful properties: For example, the LSDD

solution can be computed *analytically* in a computationally efficient and stable manner, and all tuning parameters such as the kernel width and the regularization parameter can be systematically and objectively optimized via cross-validation. We derive a finite-sample error bound for the LSDD estimator and show that it achieves the optimal convergence rate in a non-parametric setup.

We then apply LSDD to L^2-distance estimation and show that it is more accurate than the difference of KDEs, which tends to severely under-estimate the L^2-distance [7]. Because the L^2-distance is more robust against outliers than the *Kullback-Leibler divergence* [8], the proposed L^2-distance estimator can lead to the paradigm of robust distribution comparison. We experimentally demonstrate the usefulness of LSDD in semi-supervised class-prior estimation and unsupervised change detection.

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.

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. Our goal is to estimate the density difference,

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

from the samples $\mathcal{X}$ and $\mathcal{X}'$.

A naive approach to density-difference estimation is to use *kernel density estimators* (KDEs). However, we argue that the KDE-based density-difference estimator is not the best approach because of its two-step nature. Intuitively, good density estimators tend to be smooth and thus the difference between such smooth density estimators tends to be over-smoothed as a density-difference estimator [9]. 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 $p'(\boldsymbol{x})$.

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:

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

We use the following Gaussian kernel model as $g(\boldsymbol{x})$:

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

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

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

where $\boldsymbol{H}$ is the $(n + n') \times (n + n')$ matrix and $\boldsymbol{h}$ is the $(n + n')$-dimensional vector defined as

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

$$h_\ell := \int \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}_\ell\|^2}{2\sigma^2}\right) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int \exp\left(-\frac{\|\boldsymbol{x}' - \boldsymbol{c}_\ell\|^2}{2\sigma^2}\right) p'(\boldsymbol{x}') \mathrm{d}\boldsymbol{x}'.$$

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

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

where λ (≥ 0) is the regularization parameter and $\widehat{\boldsymbol{h}}$ is the $(n+n')$-dimensional vector defined as

$$\widehat{h}_\ell := \frac{1}{n}\sum_{i=1}^{n}\exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{c}_\ell\|^2}{2\sigma^2}\right) - \frac{1}{n'}\sum_{i'=1}^{n'}\exp\left(-\frac{\|\boldsymbol{x}'_{i'} - \boldsymbol{c}_\ell\|^2}{2\sigma^2}\right).$$

Taking the derivative of the objective function in Eq.(2) and equating it to zero, we can obtain the solution analytically as

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

where $\boldsymbol{I}$ denotes the identity matrix.

Finally, a density-difference estimator $\widehat{f}(\boldsymbol{x})$, which we call the *least-squares density-difference* (LSDD) estimator, is given as

$$\widehat{f}(\boldsymbol{x}) = \sum_{\ell=1}^{n+n'}\widehat{\theta}_\ell \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}_\ell\|^2}{2\sigma^2}\right).$$

Non-Parametric Error Bound: Here, we theoretically analyze an estimation error of LSDD.

We assume $n' = n$, and let $\mathcal{H}_\gamma$ be the reproducing kernel Hilbert space (RKHS) corresponding to the Gaussian kernel with width γ: $k_\gamma(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/\gamma^2\right)$. Let us consider a slightly modified LSDD estimator that is more suitable for non-parametric error analysis[1]:

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

Then we have the following theorem:

Theorem 1. *Suppose that there exists a constant M such that $\|p\|_\infty \leq M$ and $\|p'\|_\infty \leq M$. Suppose also that 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*

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

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 [10] for the definitions). Then, for all $\epsilon > 0$ and $p \in (0,1)$, there exists a constant $K > 0$ depending on M, c, ϵ, and 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 $1 - 4e^{-\tau}$.

If we set $\lambda = n^{-\frac{2\alpha+d}{(2\alpha+d)(1+p)+(\epsilon-p+\epsilon p)}}$ and $\gamma = n^{-\frac{1}{(2\alpha+d)(1+p)+(\epsilon-p+\epsilon p)}}$, and take ϵ and p sufficiently small, then we immediately have the following corollary.

Corollary 1. *Suppose that the same assumptions as Theorem 1 hold. 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*

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

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

[1]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 seems to perform well in practice.

Note that $n^{-\frac{2\alpha}{2\alpha+d}}$ is the optimal learning rate to estimate a function in $B^{\alpha}_{2,\infty}$. 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.

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 [11, 12]. To achieve the optimal rate by KDE, we should choose a kernel function specifically tailored to each regularity α [13]. 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.

Model Selection by Cross-Validation: The above theoretical analysis showed the superiority of LSDD. However, in practice, the performance of LSDD depends on the choice of models (i.e., 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. Finally, we choose the model that minimizes the average hold-out error.

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

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

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

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, we obtain $L^2(p, p') \approx \widehat{\boldsymbol{h}}^\top\widehat{\boldsymbol{\theta}}$. 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 $L^2(p, p') \approx \widehat{\boldsymbol{\theta}}^\top\boldsymbol{H}\widehat{\boldsymbol{\theta}}$.

Although $\widehat{\boldsymbol{h}}^\top\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}^\top\boldsymbol{H}\widehat{\boldsymbol{\theta}}$ themselves give approximations to $L^2(p, p')$, we argue that the use of their combination, defined by

$$\widehat{L}^2(\mathcal{X}, \mathcal{X}') := 2\widehat{\boldsymbol{h}}^\top\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}^\top\boldsymbol{H}\widehat{\boldsymbol{\theta}}, \tag{3}$$

is more sensible. To explain the reason, let us consider a generalized L^2-distance estimator of the form $\beta\widehat{\boldsymbol{h}}^\top\widehat{\boldsymbol{\theta}} + (1-\beta)\widehat{\boldsymbol{\theta}}^\top\boldsymbol{H}\widehat{\boldsymbol{\theta}}$, where β is a real scalar. If the regularization parameter λ (≥ 0) is small, this can be expressed as

$$\beta\widehat{\boldsymbol{h}}^\top\widehat{\boldsymbol{\theta}} + (1-\beta)\widehat{\boldsymbol{\theta}}^\top\boldsymbol{H}\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{h}}^\top\boldsymbol{H}^{-1}\widehat{\boldsymbol{h}} - \lambda(2-\beta)\widehat{\boldsymbol{h}}^\top\boldsymbol{H}^{-2}\widehat{\boldsymbol{h}} + o_p(\lambda), \tag{4}$$

where o_p denotes the probabilistic order. Thus, up to $O_p(\lambda)$, the bias introduced by regularization (i.e., the second term in the right-hand side of Eq.(4) that depends on λ) can be eliminated if $\beta = 2$, which yields Eq.(3). Note that, if no regularization is imposed (i.e., $\lambda = 0$), both $\widehat{\boldsymbol{h}}^\top\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}^\top\boldsymbol{H}\widehat{\boldsymbol{\theta}}$ yield $\widehat{\boldsymbol{h}}^\top\boldsymbol{H}^{-1}\widehat{\boldsymbol{h}}$, the first term in the right-hand side of Eq.(4).

Eq.(3) is actually equivalent to the negative of the optimal objective value of the LSDD optimization problem without regularization (i.e., Eq.(2) with $\lambda = 0$). This can be naturally interpreted through a lower bound of $L^2(p, p')$ obtained by *Legendre-Fenchel convex duality* [14]:

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

where the supremum is attained at $g = f$. If the expectations are replaced by empirical estimators and the Gaussian kernel model (1) is used as g, the above optimization problem is reduced to the LSDD objective function without regularization (see Eq.(2)). Thus, LSDD corresponds to approximately maximizing the above lower bound and Eq.(3) 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 (3) is not less than the plain approximators $\widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}^\top \boldsymbol{H} \widehat{\boldsymbol{\theta}}$.

4 Experiments

In this section, we experimentally demonstrate the usefulness of LSDD. A MATLAB® implementation of LSDD used for experiments is available from

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

Illustration: Let $N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the multi-dimensional normal density with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ with respect to $\boldsymbol{x}$, and let

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

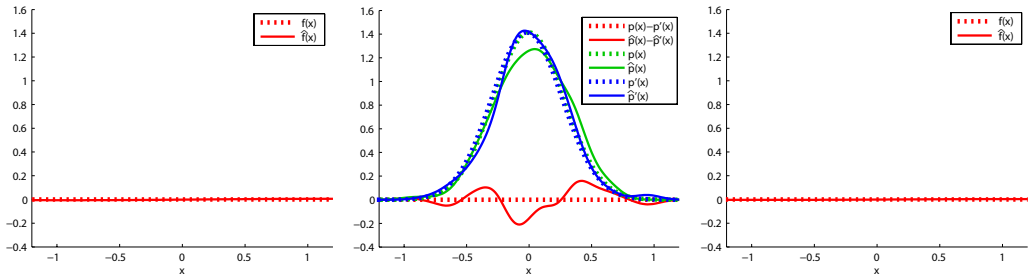
We first illustrate how LSDD behaves under $d = 1$ and $n = n' = 200$. We compare LSDD with KDEi (KDE with two Gaussian widths chosen *independently* by least-squares cross-validation [15]) and KDEj (KDE with two Gaussian widths chosen *jointly* to minimize the LSDD criterion [9]). The number of folds in cross-validation is set to 5 for all methods.

Figure 1 depicts density-difference estimation results obtained by LSDD, KDEi, and KDEj for $\mu = 0$ (i.e., $f(x) = p(x) - p'(x) = 0$). The figure shows that LSDD and KDEj give accurate estimates of the density difference $f(x) = 0$. On the other hand, the 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 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 when μ is large. This is caused by the fact that KDE tends to provide smooth density estimates (see Figure 2(b) again): Such smooth density estimates are accurate as density estimates, but the difference of smooth density estimates yields a small L^2-distance estimate [7]. The KDEj-based L^2-distance estimator tends to improve this drawback of KDEi, but it still slightly underestimates the true L^2-distance when μ is large.

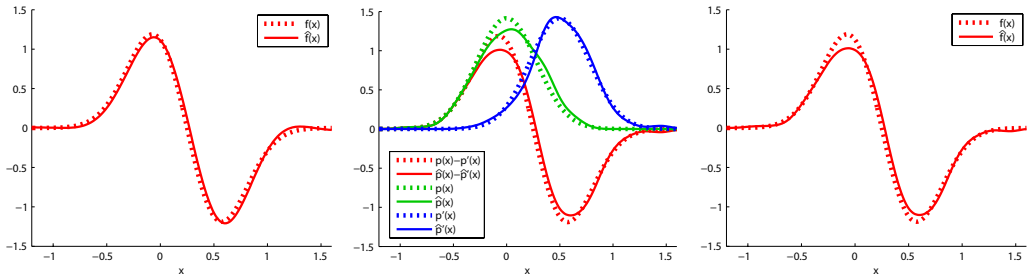
When $d = 5$ (Figure 3(b)), the KDE-based L^2-distance estimators even severely underestimate the true L^2-distance when μ is large. On the other hand, the LSDD-based L^2-distance estimator still gives reasonably accurate estimates of the true L^2-distance even when $d = 5$. However, we note that LSDD 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.

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



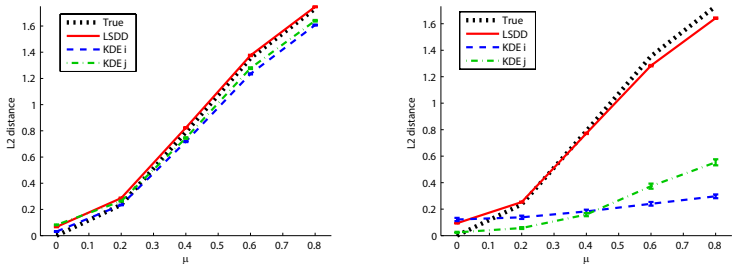
(a) LSDD (b) KDEi (c) KDEj

Figure 1: Estimation of density difference when $\mu = 0$ (i.e., $f(x) = p(x) - p'(x) = 0$).



(a) LSDD (b) KDEi (c) KDEj

Figure 2: Estimation of density difference when $\mu = 0.5$ (i.e., $f(x) = p(x) - p'(x) \neq 0$).



(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.

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 a binary pattern recognition task of classifying pattern $\boldsymbol{x} \in \mathbb{R}^d$ to class $y \in \{+1, -1\}$. Our goal is to learn 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 [16]. 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 $p_{\mathrm{test}}(\boldsymbol{x})$ [5], where $\pi \in [0, 1]$ is a mixing coefficient to learn. See Figure 4 for schematic illustration. Here, we use the L^2-distance estimated by LSDD and the difference of KDEs 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.

We use four UCI benchmark datasets (`http://archive.ics.uci.edu/ml/`), where we randomly choose 10 labeled training samples from each class and 50 unlabeled test samples following true class-prior $\pi^* = 0.1, 0.2, \ldots, 0.9$. Figure 6 plots the mean and standard error of the squared difference between true and estimated class-balances π and the misclassification error by a weighted ℓ_2-regularized least-squares classifier [17] with weighted cross-validation [18] over 1000 runs. The results show that LSDD tends to provide better class-balance estimates than the KDEi-based, the KDEj-based, and the EM-based methods [5], which are translated into lower classification errors.

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 [6]. 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 5). As a dissimilarity measure, we use the L^2-distance estimated by LSDD and the Kullback-Leibler (KL) divergence estimated by the *KL importance estimation procedure* (KLIEP) [2, 3]. We set $k = 10$ and $r = 50$.

First, we use the *IPSJ SIG-SLP Corpora and Environments for Noisy Speech Recognition* (CENSREC) dataset (`http://research.nii.ac.jp/src/en/CENSREC-1-C.html`). This dataset is provided by the *National Institute of Informatics, Japan* that records human voice in a noisy environment such as a restaurant. The top graphs in Figure 7(a) display the original time-series (true change points were manually annotated) and change scores obtained by KLIEP and LSDD. The graphs show that the LSDD-based change score indicates the existence of change points more clearly than the KLIEP-based change score.

Next, we use a dataset taken from the *Human Activity Sensing Consortium (HASC) challenge 2011* (`http://hasc.jp/hc2011/`), 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 added zero-mean Gaussian noise with standard deviation 5 at each time point with probability 0.005. The top graphs in Figure 7(b) display the original time-series for a sequence of actions "jog", "stay", "stair down", "stay", and "stair up" (there exists 4 change points at time 540, 1110, 1728, and 2286) and the change scores obtained by KLIEP and LSDD. The graphs show that the LSDD score is much more stable and interpretable than the KLIEP score.

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 test the L^2-distance estimated by KDEi and KDEj and native change detection methods based on autoregressive models (AR) [19], subspace identification (SI) [20], singular spectrum transformation (SST) [21], one-class support vector machine (SVM) [22], kernel Fisher discriminant analysis (KFD) [23], and kernel change-point detection (KCP) [24]. Tuning parameters included in these methods were manually optimized. For 10 datasets taken from each of the CENSREC and HASC data collections, mean ROC curves and AUC values are displayed at the bottom of Figure 7(b). The 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 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 derived a finite-sample error bound for LSDD in a non-parametric setup, and showed that it achieves the optimal convergence rate. We also proposed an L^2-distance estimator based on LSDD, which nicely cancels a bias caused by regularization. Through experiments on class-prior estimation and change-point detection, the usefulness of the proposed LSDD was demonstrated.

Acknowledgments: We would like to thank Wittawat Jitkrittum for his comments and Zaïd Harchaoui for providing us a program code of kernel change-point detection. MS was supported by MEXT KAKENHI 23300069 and AOARD, TK was supported by MEXT KAKENHI 24500340, TS 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, MCdP was supported by MEXT Scholarship, SL was supported by the JST PRESTO program, and IT was supported by MEXT KAKENHI 23700165.

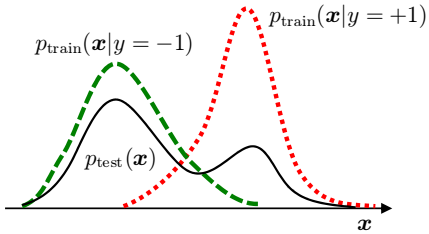


Figure 4: Class-balance estimation.

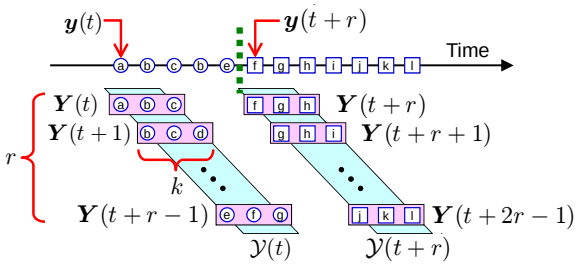
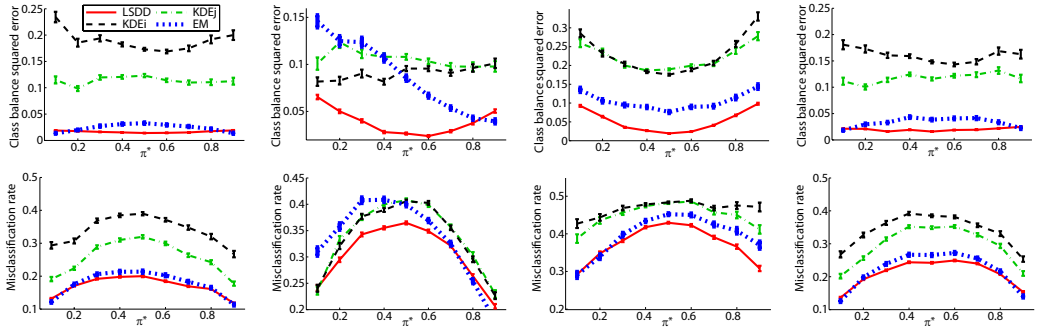
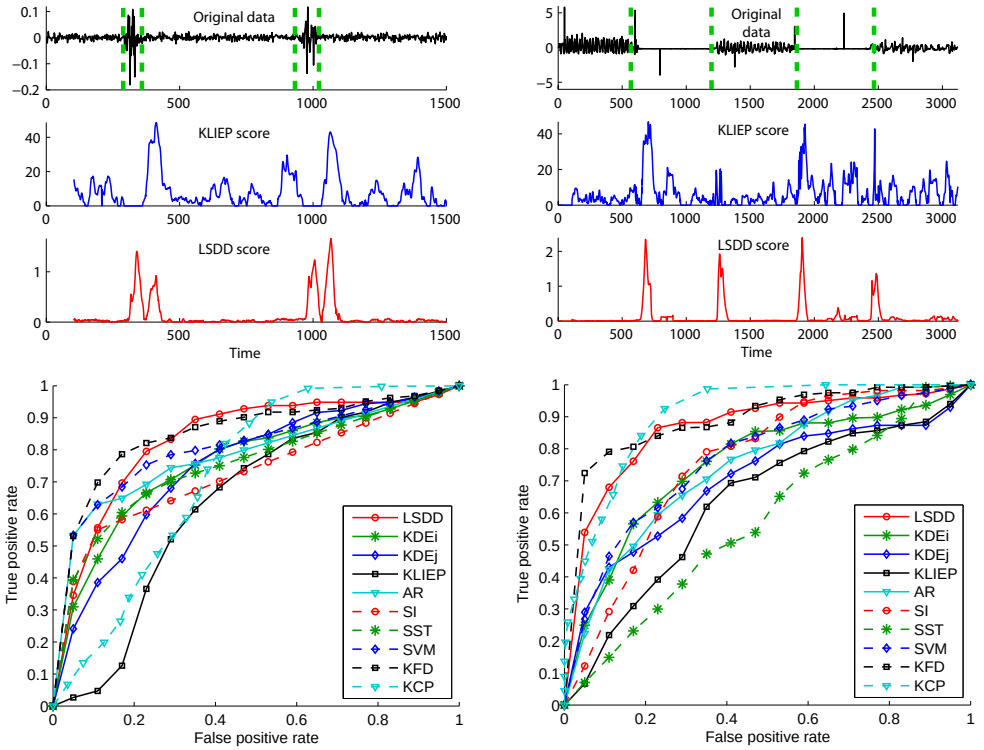


Figure 5: Change-point detection.



(a) Australian dataset (b) Diabetes dataset (c) German dataset (d) Statlogheart dataset

Figure 6: Results of semi-supervised class-balance estimation. Top: Squared error of class balance estimation. Bottom: Misclassification error by a weighted ℓ_2-regularized least-squares classifier.



AUC	LSDD	KDEi	KDEj	KLIEP	AR	SI	SST	SVM	KFD	KCP
Mean	**.879**	.755	.705	.635	.749	.756	.580	.773	**.905**	**.913**
SE	.014	.016	.023	.030	.013	.012	.023	.032	.013	.024

(a) Speech data

AUC	LSDD	KDEi	KDEj	KLIEP	AR	SI	SST	SVM	KFD	KCP
Mean	**.843**	.764	.751	.638	**.799**	.762	.764	**.815**	**.856**	.730
SE	.013	.029	.036	.020	.026	.020	.016	.018	.023	.032

(b) Accelerometer data

Figure 7: Results of unsupervised change detection. From top to bottom: Original time-series, change scores obtained by KLIEP and LSDD, mean ROC curves over 10 datasets, and AUC values for 10 datasets. The best method and comparable ones in terms of mean AUC values by the *t-test* at the significance level 5% are indicated with boldface. "SE" stands for "Standard error".

References

[1] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, USA, 1998.

[2] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

[3] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[4] C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems 23*, pages 442–450, 2010.

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

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

[7] N. Anderson, P. Hall, and D. Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1):41–54, 1994.

[8] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

[9] P. Hall and M. P. Wand. On nonparametric discrimination using density differences. *Biometrika*, 75(3):541–547, 1988.

[10] M. Eberts and I. Steinwart. Optimal learning rates for least squares SVMs using Gaussian kernels. In *Advances in Neural Information Processing Systems 24*, pages 1539–1547, 2011.

[11] R. H. Farrell. On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *The Annals of Mathematical Statistics*, 43(1):170–180, 1972.

[12] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, UK, 1986.

[13] E. Parzen. On the estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[14] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, USA, 1970.

[15] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer, Berlin, Germany, 2004.

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

[17] R. Rifkin, G. Yeo, and T. Poggio. Regularized least-squares classification. In *Advances in Learning Theory: Methods, Models and Applications*, pages 131–154. IOS Press, Amsterdam, the Netherlands, 2003.

[18] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, May 2007.

[19] Y. Takeuchi and K. Yamanishi. A unifying framework for detecting outliers and change points from non-stationary time series data. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):482–489, 2006.

[20] Y. Kawahara, T. Yairi, and K. Machida. Change-point detection in time-series data based on subspace identification. In *Proceedings of the 7th IEEE International Conference on Data Mining*, pages 559–564, 2007.

[21] V. Moskvina and A. A. Zhigljavsky. An algorithm based on singular spectrum analysis for change-point detection. *Communication in Statistics: Simulation & Computation*, 32(2):319–352, 2003.

[22] F. Desobry, M. Davy, and C. Doncarli. An online kernel change detection algorithm. *IEEE Transactions on Signal Processing*, 53(8):2961–2974, 2005.

[23] Z. Harchaoui, F. Bach, and E. Moulines. Kernel change-point analysis. In *Advances in Neural Information Processing Systems 21*, pages 609–616, 2009.

[24] S. Arlot, A. Celisse, and Z. Harchaoui. Kernel change-point detection. Technical Report 1202.3878, arXiv, 2012.