Direct Divergence Approximation between Probability Distributions and Its Applications in Machine Learning

Masashi Sugiyama
Tokyo Institute of Technology, Japan.
sugi@cs.titech.ac.jp http://sugiyama-www.cs.titech.ac.jp/˜sugi
Song Liu
Tokyo Institute of Technology, Japan.
song@sg.cs.titech.ac.jp
Marthinus Christoffel du Plessis
Tokyo Institute of Technology, Japan.
christo@sg.cs.titech.ac.jp
Masao Yamanaka
Tokyo Institute of Technology, Japan.
yamanaka@sp.dis.titech.ac.jp
Makoto Yamada
NTT Corporation, Japan.
yamada.makoto@lab.ntt.co.jp
Taiji Suzuki
The University of Tokyo, Japan.
s-taiji@stat.t.u-tokyo.ac.jp
Takafumi Kanamori
Nagoya University, Japan.
kanamori@is.nagoya-u.ac.jp

Abstract

Approximating a divergence between two probability distributions from their samples is a fundamental challenge in statistics, information theory, and machine learning. A divergence approximator can be used for various purposes such as two-sample homogeneity testing, change-point detection, and class-balance estimation. Furthermore, an approximator of a divergence between the joint distribution and the product of marginals can be used for independence testing, which has a wide range of applications including feature selection and extraction, clustering, object matching, independent component analysis, and causal direction estimation. In this paper, we review recent advances in divergence approximation. Our emphasis is that directly approximating the divergence without estimating probability distributions is more sensible than a naive two-step approach of first estimating probability distributions and then approximating the divergence. Furthermore, despite the overwhelming popularity of the Kullback-Leibler divergence as a divergence measure, we argue that alternatives such as the Pearson divergence, the relative Pearson divergence, and the L^2-distance are more useful in practice because of their computationally efficient approximability, high numerical stability, and superior robustness against outliers.

Keywords

Machine learning, probability distributions, Kullback-Leibler divergence, Pearson divergence, L^2-distance.

1 Introduction

Let us consider the problem of approximating a divergence D between two probability distributions P and P' on $\mathbb{R}^d$ from two sets of independent and identically distributed samples $\mathcal{X} := \{\boldsymbol{x}_i\}_{i=1}^n$ and $\mathcal{X}' := \{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ following P and P'.

A divergence approximator can be used for various purposes such as two-sample testing [1, 2], change detection in time-series [3], class-prior estimation under class-balance change [4], salient object detection in images [5], and event detection from movies [6] and Twitter [7]. Furthermore, an approximator of the divergence between the joint distribution and the product of marginal distributions can be used for solving a wide range of machine learning problems [8], including independence testing [9], feature selection [10, 11], feature extraction [12, 13], canonical dependency analysis [14], object matching [15], independent component analysis [16], clustering [17, 18], and causal direction learning [19]. For this reason, accurately approximating a divergence between two probability distributions from their samples has been one of the challenging research topics in the statistics, information theory, and machine learning communities.

A naive way to approximate the divergence from P to P', denoted by $D(P\|P')$, is to first obtain estimators $\widehat{P}_{\mathcal{X}}$ and $\widehat{P}'_{\mathcal{X}'}$ of the distributions P and P' separately from their samples $\mathcal{X}$ and $\mathcal{X}'$, and then compute a plug-in approximator $D(\widehat{P}_{\mathcal{X}}\|\widehat{P}'_{\mathcal{X}'})$. However, this naive two-step approach violates *Vapnik's principle* [20]:

> *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.*

More specifically, if we know the distributions P and P', we can immediately know their divergence $D(P\|P')$. However, knowing the divergence $D(P\|P')$ does not necessarily imply knowing the distributions P and P', because different pairs of distributions can yield the same divergence value. Thus, estimating the distributions P and P' is more general than estimating the divergence $D(P\|P')$. Following Vapnik's principle, direct divergence approximators $\widehat{D}(\mathcal{X}, \mathcal{X}')$ that do not involve the estimation of distributions P and P' have been developed recently [21, 22, 23, 24, 25].

The purpose of this article is to give an overview of the development of such direct divergence approximators. In Section 2, we review the definitions of the Kullback-Leibler divergence, the Pearson divergence, the relative Pearson divergence, and the L^2-distance, and discuss their pros and cons. Then, in Section 3, we review direct approximators of these divergences that do not involve the estimation of probability distributions. In Section 4, we show practical usage of divergence approximators in unsupervised change-detection in time-series, semi-supervised class-prior estimation under class-balance change, salient object detection in an image, and evaluation of statistical independence between random variables. Finally, we conclude in Section 5.

2 Divergence Measures

A function $d(\cdot, \cdot)$ is called a *distance* if and only if the following four conditions are satisfied:

- Non-negativity: $\forall x, y, \quad d(x, y) \geq 0$
- Non-degeneracy: $d(x, y) = 0 \iff x = y$
- Symmetry: $\forall x, y, \quad d(x, y) = d(y, x)$
- Triangle inequality: $\forall x, y, z \quad d(x, z) \leq d(x, y) + d(y, z)$

A divergence is a pseudo-distance that still acts like a distance, but it may violate some of the above conditions. In this section, we introduce useful divergence and distance measures between probability distributions.

2.1 Kullback-Leibler (KL) Divergence

The most popular divergence measure in statistics and machine learning is the KL divergence [26] defined as

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

where $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ are probability density functions of P and P', respectively.

Advantages of the KL divergence are that it is compatible with maximum likelihood estimation, it is invariant under input metric change, its Riemannian geometric structure is well studied [27], and it can be approximated accurately via *direct density-ratio estimation* [21, 22, 28]. However, it is not symmetric, it does not satisfy the triangle inequality, its approximation is computationally expensive due to the log function, and it is sensitive to outliers and numerically unstable because of the strong non-linearity of the log function and possible unboundedness of the density-ratio function p/p' [29, 24].

2.2 Pearson (PE) Divergence

The PE divergence [30] is a squared-loss variant of the KL divergence defined as

$$\mathrm{PE}(p\|p') := \int p'(\boldsymbol{x}) \left(\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})} - 1 \right)^2 \mathrm{d}\boldsymbol{x}. \tag{1}$$

Because both the PE and KL divergences belong to the class of Ali-Silvey-Csiszár divergences (which is also known as f-divergences) [31, 32], they share similar theoretical properties such as the invariance under input metric change.

The PE divergence can also be accurately approximated via direct density-ratio estimation in the same way as the KL divergence [23, 28]. However, its approximator can be obtained *analytically* in a computationally much more efficient manner than the KL divergence, because the quadratic function the PE divergence adopts is compatible with least-squares estimation. Furthermore, the PE divergence tends to be more robust against outliers than the KL divergence [33]. However, other weaknesses of the KL divergence such as asymmetry, violation of the triangle inequality, and possible unboundedness of the density-ratio function p/p' remain unsolved in the PE divergence.

2.3 Relative Pearson (rPE) Divergence

To overcome the possible unboundedness of the density-ratio function p/p', the rPE divergence was recently introduced [24]. The rPE divergence is defined as

$$\begin{aligned} \mathrm{rPE}(p\|p') &:= \mathrm{PE}(p\|q_\alpha) \\ &= \int q_\alpha(\boldsymbol{x}) \left(\frac{p(\boldsymbol{x})}{q_\alpha(\boldsymbol{x})} - 1 \right)^2 \mathrm{d}\boldsymbol{x}, \end{aligned} \tag{2}$$

where, for $0 \leq \alpha < 1$, q_α is defined as the α-mixture of p and p':

$$q_\alpha = \alpha p + (1-\alpha)p'.$$

When $\alpha = 0$, the rPE divergence is reduced to the plain PE divergence. The quantity p/q_α is called the *relative density ratio*, which is always upper-bounded by $1/\alpha$ for $\alpha > 0$ because

$$\frac{p(\boldsymbol{x})}{q_\alpha(\boldsymbol{x})} = \frac{1}{\alpha + (1-\alpha)\frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})}} < \frac{1}{\alpha}.$$

Thus, it can overcome the unboundedness problem of the PE divergence, while the invariance under input metric change is still maintained.

The rPE divergence is still compatible with least-squares estimation, and it can be approximated in almost the same way as the PE divergence via *direct relative density-ratio estimation* [24]. Indeed, an rPE-divergence approximator can still be obtained analytically in an accurate and computationally efficient manner. However, it still violates symmetry and the triangle inequality in the same way as the KL and PE divergence. Furthermore, the choice of α may not be straightforward in some applications.

2.4 L^2-Distance

The L^2-distance is another standard distance measure between probability distributions defined as

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

The L^2-distance is a proper distance measure, and thus it is symmetric and satisfies the triangle inequality. Furthermore, the density difference $p(\boldsymbol{x}) - p'(\boldsymbol{x})$ is always bounded as long as each density is bounded. Therefore, the L^2-distance is stable, without the need of tuning any control parameter such as α in the rPE divergence.

The L^2-distance is also compatible with least-squares estimation, and it can be accurately and analytically approximated in a computationally efficient and numerically stable manner via *direct density-difference estimation* [25]. However, the L^2-distance is not invariant under input metric change, which is a unique property inherent to ratio-based divergences.

3 Direct Divergence Approximation

In this section, we review recent advances in direct divergence approximation.

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 approximate a divergence between from p to p' from samples $\mathcal{X}$ and $\mathcal{X}'$.

3.1 KL Divergence Approximation

The key idea of direct KL divergence approximation is to estimate the density ratio p/p' without estimating the densities p and p' [21]. More specifically, a density-ratio estimator is obtained by minimizing the KL divergence from p to $r \cdot p'$ with respect to a density-ratio model r, under the constraints that the density-ratio function is non-negative and $r \cdot p'$ is integrated to one:

$$\min_r \mathrm{KL}(p \| r \cdot p')$$
$$\text{subject to } r \geq 0 \text{ and } \int r(\boldsymbol{x}) p'(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1.$$

Its empirical optimization problem, where an irrelevant constant is ignored and the expectations are approximated by the sample averages, is given by

$$\max_r \frac{1}{n} \sum_{i=1}^{n} \log r(\boldsymbol{x}_i)$$
$$\text{subject to } r \geq 0 \text{ and } \frac{1}{n'} \sum_{i'=1}^{n'} r(\boldsymbol{x}'_{i'}) = 1.$$

Let us consider the following Gaussian density-ratio model:

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

where $\|\cdot\|$ denotes the ℓ_2-norm. We define the vector of parameters $\{\theta_\ell\}_{\ell=1}^{n}$ as

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

where $^\top$ denotes the transpose. In this model, the Gaussian kernels are located on numerator samples $\{\boldsymbol{x}_i\}_{i=1}^{n}$ because the density ratio p/p' tends to take large values in the regions where the numerator samples $\{\boldsymbol{x}_i\}_{i=1}^{n}$ exist. Alternatively, Gaussian kernels may be located on both numerator and denominator samples, but this seems not to further improve the accuracy [21]. When n is very large, a (random) subset of numerator samples $\{\boldsymbol{x}_i\}_{i=1}^{n}$ may be chosen as Gaussian centers, which can reduce the computational cost.

For the Gaussian density-ratio model (3), the above optimization problem is expressed as

$$\begin{aligned}&\max_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n}\log\left(\sum_{\ell=1}^{n}\theta_\ell \exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_\ell\|^2}{2\sigma^2}\right)\right)\\ &\text{subject to } \theta_1,\ldots,\theta_n \geq 0\\ &\text{and } \frac{1}{n'}\sum_{i'=1}^{n'}\sum_{\ell=1}^{n}\theta_\ell \exp\left(-\frac{\|\boldsymbol{x}'_{i'}-\boldsymbol{x}_\ell\|^2}{2\sigma^2}\right)=1.\end{aligned}$$

This is a convex optimization problem and thus the global optimal solution can be obtained easily, e.g., by gradient-projection iterations. Furthermore, the global optimal solution tends to be *sparse* (i.e., many parameter values become exactly zero), which can be utilized for reducing the computational cost.

The Gaussian width σ is a tuning parameter in this algorithm, and it can be systematically optimized by *cross-validation* with respect to the objective function. More specifically, the numerator samples $\mathcal{X} := \{\boldsymbol{x}_i\}_{i=1}^n$ are divided into T disjoint subsets $\{\mathcal{X}_t\}_{t=1}^T$ of (approximately) the same size. Then a density-ratio estimator $\widehat{r}_t(\boldsymbol{x})$ is obtained using $\mathcal{X}\backslash\mathcal{X}_t$ and $\mathcal{X}' := \{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ (i.e., all numerator samples without $\mathcal{X}_t$ and all denominator samples), and its objective value for the hold-out numerator samples $\mathcal{X}_t$ is computed:

$$\frac{1}{|\mathcal{X}_t|}\sum_{\boldsymbol{x}\in\mathcal{X}_t}\log\widehat{r}_t(\boldsymbol{x}),$$

where $|\mathcal{X}_t|$ denotes the number of elements in the set $\mathcal{X}_t$. This procedure is repeated for $t=1,\ldots,T$, and the σ value that maximizes the average of the above hold-out objective values is chosen as the best one.

Given a density-ratio estimator $\widehat{r}$, a KL-divergence approximator $\widehat{\mathrm{KL}}(\mathcal{X}\|\mathcal{X}')$ can be constructed as

$$\widehat{\mathrm{KL}}(\mathcal{X}\|\mathcal{X}') := \frac{1}{n}\sum_{i=1}^{n}\log\widehat{r}(\boldsymbol{x}_i).$$

A MATLAB® implementation of the above KL divergence approximator (called the *KL importance estimation procedure*; KLIEP) is available from

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

Variations of this procedure for other density-ratio models have been developed, including the log-linear model [34], the Gaussian mixture model [35], and the mixture of probabilistic principal component analyzers [36]. Also, an unconstrained variant, which corresponds to approximately maximizing the *Legendre-Fenchel lower bound* of the KL divergence [37], was proposed [22]:

$$\widetilde{\mathrm{KL}}(\mathcal{X}\|\mathcal{X}') := \max_{r}\left[\frac{1}{n}\sum_{i=1}^{n}\log r(\boldsymbol{x}_i)-\frac{1}{n'}\sum_{i'=1}^{n'}r(\boldsymbol{x}'_{i'})+1\right].$$

3.2 PE Divergence Approximation

The PE divergence can also be directly approximated without estimating the densities p and p' via direct estimation of the density ratio p/p' [23]. More specifically, a density-ratio estimator is obtained by minimizing the p'-weighted squared difference between a density-ratio model r and the true density-ratio function p/p':

$$\min_r \int p'(\boldsymbol{x}) \left(r(\boldsymbol{x}) - \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})} \right)^2 \mathrm{d}\boldsymbol{x}.$$

Its empirical criterion where an irrelevant constant is ignored and the expectations are approximated by the sample averages is given by

$$\min_r \left[\frac{1}{n'} \sum_{i'=1}^{n'} r^2(\boldsymbol{x}'_{i'}) - \frac{2}{n} \sum_{i=1}^{n} r(\boldsymbol{x}_i) \right].$$

For the Gaussian density-ratio model (3) together with the ℓ_2-regularizer, the above optimization problem is expressed as

$$\min_{\boldsymbol{\theta}} \left[\boldsymbol{\theta}^\top \widehat{\boldsymbol{G}}' \boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \widehat{\boldsymbol{h}} + \lambda \|\boldsymbol{\theta}\|^2 \right], \tag{4}$$

where $\lambda \geq 0$ denotes the regularization parameter, $\widehat{\boldsymbol{G}}'$ is the $n \times n$ matrix with the (ℓ, ℓ')-th element defined by

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

and $\widehat{\boldsymbol{h}}$ is the n-dimensional vector with the ℓ-th element defined by

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

This is a convex optimization problem, and the global optimal solution can be computed *analytically* as

$$(\widehat{\boldsymbol{G}}' + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}},$$

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

The Gaussian width σ and the regularization parameter λ are the tuning parameters in this algorithm, and they can be systematically optimized by cross-validation with respect to the objective function as follows: First, the numerator and denominator samples $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^n$ and $\mathcal{X}' = \{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ are divided into T disjoint subsets $\{\mathcal{X}_t\}_{t=1}^T$ and $\{\mathcal{X}'_t\}_{t=1}^T$, respectively. Then a density-ratio estimator $\widehat{r}_t(\boldsymbol{x})$ is obtained using $\mathcal{X} \backslash \mathcal{X}_t$ and $\mathcal{X}' \backslash \mathcal{X}'_t$ (i.e.,

all samples without $\mathcal{X}_t$ and $\mathcal{X}'_t$), and its objective value for the hold-out samples $\mathcal{X}_t$ and $\mathcal{X}'_t$ is computed:

$$\frac{1}{|\mathcal{X}'_t|}\sum_{\boldsymbol{x}'\in\mathcal{X}'_t}\widehat{r}_t(\boldsymbol{x}')^2-\frac{2}{|\mathcal{X}_t|}\sum_{\boldsymbol{x}\in\mathcal{X}_t}\widehat{r}_t(\boldsymbol{x}). \tag{5}$$

This procedure is repeated for $t=1,\ldots,T$, and the σ and λ values that maximize the average of the above hold-out objective values are chosen as the best ones.

By expanding the squared term $\left(\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}-1\right)^2$ in Eq.(1), the PE divergence can be expressed as

$$\mathrm{PE}=\int p(\boldsymbol{x})\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}-1 \tag{6}$$

$$=-\int p'(\boldsymbol{x})\left(\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}\right)^2\mathrm{d}\boldsymbol{x}+2\int p(\boldsymbol{x})\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}-1. \tag{7}$$

Note that Eq.(7) can also be obtained via *Legendre-Fenchel convex duality* of the divergence functional [38]. Based on these expressions, PE divergence approximators are obtained using a density-ratio estimator $\widehat{r}$ as

$$\widehat{\mathrm{PE}}(\mathcal{X}\|\mathcal{X}'):=\frac{1}{n}\sum_{i=1}^{n}\widehat{r}(\boldsymbol{x}_i)-1, \tag{8}$$

$$\widetilde{\mathrm{PE}}(\mathcal{X}\|\mathcal{X}'):=-\frac{1}{n'}\sum_{i'=1}^{n'}\widehat{r}(\boldsymbol{x}'_{i'})^2+\frac{2}{n}\sum_{i=1}^{n}\widehat{r}(\boldsymbol{x}_i)-1. \tag{9}$$

Eq.(8) is suitable for algorithmic development because this would be the simplest expression, while Eq.(9) is suitable for theoretical analysis because this corresponds to the negative of the objective function in Eq.(4).

A MATLAB® implementation of the above method (called *unconstrained least-squares importance fitting*; uLSIF) is available from

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

If the ℓ_2-regularizer

$$\|\boldsymbol{\theta}\|^2:=\sum_{\ell=1}^{n}\theta_\ell^2$$

in Eq.(4) is replaced with the ℓ_1-regularizer

$$\|\boldsymbol{\theta}\|_1:=\sum_{\ell=1}^{n}|\theta_\ell|,$$

the solution tends to be sparse [39]. Then the solution can be obtained in a computationally more efficient way [40], and furthermore a regularization path tracking algorithm [41] is available for efficiently computing solutions with different regularization parameter values.

3.3 rPE Divergence Approximation

The rPE divergence can be directly estimated in the same way as the PE divergence [24]:

$$\min_r \int q_\alpha(\boldsymbol{x}') \left(r(\boldsymbol{x}) - \frac{p(\boldsymbol{x})}{q_\alpha(\boldsymbol{x})} \right)^2 \mathrm{d}\boldsymbol{x}.$$

Its empirical criterion where an irrelevant constant is ignored and the expectations are approximated by sample averages is given by

$$\min_r \left[\frac{\alpha}{n} \sum_{i=1}^{n} r^2(\boldsymbol{x}_i) + \frac{1-\alpha}{n'} \sum_{i'=1}^{n'} r^2(\boldsymbol{x}'_{i'}) - \frac{2}{n} \sum_{i=1}^{n} r(\boldsymbol{x}_i) \right].$$

For the Gaussian density-ratio model (3) together with the ℓ_2-regularizer, the above optimization problem is expressed as

$$\min_{\boldsymbol{\theta}} \left[\boldsymbol{\theta}^\top (\alpha \widehat{\boldsymbol{G}} + (1-\alpha)\widehat{\boldsymbol{G}}')\boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \widehat{\boldsymbol{h}} + \lambda \|\boldsymbol{\theta}\|^2 \right],$$

where $\widehat{\boldsymbol{G}}$ is the $n \times n$ matrix with the (ℓ, ℓ')-th element defined by

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

This is a convex optimization problem, and the global optimal solution can be computed analytically as

$$(\alpha \widehat{\boldsymbol{G}} + (1-\alpha)\widehat{\boldsymbol{G}}' + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}}.$$

Cross-validation for tuning the Gaussian width σ and the regularization parameter λ can be carried out in the same way as the PE-divergence case, with Eq.(5) replaced by

$$\frac{\alpha}{|\mathcal{X}_t|} \sum_{\boldsymbol{x} \in \mathcal{X}_t} \widehat{r}_t(\boldsymbol{x})^2 + \frac{1-\alpha}{|\mathcal{X}'_t|} \sum_{\boldsymbol{x}' \in \mathcal{X}'_t} \widehat{r}_t(\boldsymbol{x}')^2 - \frac{2}{|\mathcal{X}_t|} \sum_{\boldsymbol{x} \in \mathcal{X}_t} \widehat{r}_t(\boldsymbol{x}).$$

By expanding the squared term $\left(\frac{p(\boldsymbol{x})}{q_\alpha(\boldsymbol{x})} - 1 \right)^2$ in Eq.(2), the rPE divergence can be expressed as

$$\mathrm{rPE} = \int p(\boldsymbol{x}) \frac{p(\boldsymbol{x})}{q_\alpha(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} - 1 \tag{10}$$

$$= -\int q_\alpha(\boldsymbol{x}) \left(\frac{p(\boldsymbol{x})}{q_\alpha(\boldsymbol{x})} \right)^2 \mathrm{d}\boldsymbol{x} + 2 \int p(\boldsymbol{x}) \frac{p(\boldsymbol{x})}{q_\alpha(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} - 1. \tag{11}$$

Based on these expressions, rPE divergence approximators are given using the relative density-ratio estimator $\widehat{r}_\alpha$ as

$$\widehat{\mathrm{rPE}}_\alpha(\mathcal{X}\|\mathcal{X}') := \frac{1}{n}\sum_{i=1}^{n}\widehat{r}_\alpha(\boldsymbol{x}_i) - 1, \tag{12}$$

$$\widetilde{\mathrm{rPE}}_\alpha(\mathcal{X}\|\mathcal{X}') := -\frac{\alpha}{n}\sum_{i=1}^{n}\widehat{r}_\alpha(\boldsymbol{x}_i)^2 - \frac{(1-\alpha)}{n'}\sum_{i'=1}^{n'}\widehat{r}_\alpha(\boldsymbol{x}'_{i'})^2 + \frac{2}{n}\sum_{i=1}^{n}\widehat{r}_\alpha(\boldsymbol{x}_i) - 1. \tag{13}$$

A MATLAB® implementation of this algorithm (called *relative uLSIF*; RuLSIF) is available from

"`http://sugiyama-www.cs.titech.ac.jp/~yamada/RuLSIF.html`".

3.4 L^2-Distance Approximation

The key idea is to directly estimate the density difference $p - p'$ without estimating each density [25]. More specifically, a density-difference estimator is obtained by minimizing the squared difference between a density-difference model f and the true density-difference function $p - p'$:

$$\min_f \int \Big(f(\boldsymbol{x}) - \big(p(\boldsymbol{x}) - p'(\boldsymbol{x})\big)\Big)^2 \mathrm{d}\boldsymbol{x}.$$

Its empirical criterion where an irrelevant constant is ignored and the expectation is approximated by the sample average is given by

$$\min_f \left[\int f(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x} - \left(\frac{2}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i) - \frac{2}{n'}\sum_{i'=1}^{n'} f(\boldsymbol{x}'_{i'})\right)\right].$$

Let us consider the following Gaussian density-difference model:

$$f(\boldsymbol{x}) = \sum_{\ell=1}^{n+n'} \xi_\ell \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{c}_\ell\|^2}{2\sigma^2}\right), \tag{14}$$

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 centers. Then the above optimization problem is expressed as

$$\min_{\boldsymbol{\xi}=(\xi_1,\ldots,\xi_{n+n'})^\top} \Big[\boldsymbol{\xi}^\top \boldsymbol{U}\boldsymbol{\xi} - 2\boldsymbol{\xi}^\top \widehat{\boldsymbol{v}} + \lambda\|\boldsymbol{\xi}\|^2\Big],$$

where the ℓ_2-regularizer $\lambda\|\boldsymbol{\xi}\|^2$ is included, $\boldsymbol{U}$ is the $(n+n')\times(n+n')$ matrix with the (ℓ,ℓ')-th element defined by

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

d denotes the dimensionality of $\boldsymbol{x}$, and $\widehat{\boldsymbol{v}}$ is the $(n+n')$-dimensional vector with the ℓ-th element defined by

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

This is a convex optimization problem, and the global optimal solution can be computed *analytically* as

$$(\boldsymbol{U}+\lambda\boldsymbol{I})^{-1}\widehat{\boldsymbol{v}}.$$

The above optimization problem is essentially the same form as least-squares density-ratio approximation for the PE divergence, and therefore least-squares density-difference approximation can enjoy all the computational properties of least-squares density-ratio approximation.

Cross-validation for tuning the Gaussian width σ and the regularization parameter λ can be carried as follows: First, the numerator and denominator samples $\mathcal{X}=\{\boldsymbol{x}_i\}_{i=1}^{n}$ and $\mathcal{X}'=\{\boldsymbol{x}'_{i'}\}_{i'=1}^{n'}$ are divided into T disjoint subsets $\{\mathcal{X}_t\}_{t=1}^{T}$ and $\{\mathcal{X}'_t\}_{t=1}^{T}$, respectively. Then a density-difference estimator $\widehat{f}_t(\boldsymbol{x})$ is obtained using $\mathcal{X}\backslash\mathcal{X}_t$ and $\mathcal{X}'\backslash\mathcal{X}'_t$ (i.e., all samples without $\mathcal{X}_t$ and $\mathcal{X}'_t$), and its objective value for the hold-out samples $\mathcal{X}_t$ and $\mathcal{X}'_t$ is computed:

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

Note that the first term can be computed analytically for the Gaussian density-difference model (14):

$$\int \widehat{f}_t(\boldsymbol{x})^2\mathrm{d}\boldsymbol{x} = \widehat{\boldsymbol{\xi}}_t^\top\boldsymbol{U}\widehat{\boldsymbol{\xi}}_t,$$

where $\widehat{\boldsymbol{\xi}}_t$ is the parameter vector learned from $\mathcal{X}\backslash\mathcal{X}_t$ and $\mathcal{X}'\backslash\mathcal{X}'_t$.

For an equivalent expression of the L^2-distance,

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

if f is replaced with a density-difference estimator $\widehat{f}$ and approximate the expectations by empirical averages, the following L^2-distance approximator can be obtained:

$$\widehat{\boldsymbol{v}}^\top \widehat{\boldsymbol{\xi}}. \tag{15}$$

Similarly, for another expression

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

replacing f with a density-difference estimator $\widehat{f}$ gives another L^2-distance approximator:

$$\widehat{\boldsymbol{\xi}}^\top \boldsymbol{U} \widehat{\boldsymbol{\xi}}. \tag{16}$$

Eq.(15) and Eq.(16) themselves give valid approximations to $L^2(p, p')$, but their linear combination

$$\widehat{L}^2(\mathcal{X}, \mathcal{X}') := 2\widehat{\boldsymbol{v}}^\top \widehat{\boldsymbol{\xi}} - \widehat{\boldsymbol{\xi}}^\top \boldsymbol{U} \widehat{\boldsymbol{\xi}},$$

was shown to have a smaller bias than than Eq.(15) and Eq.(16).

A MATLAB® implementation of the above algorithm (called *least-squares density difference*; LSDD) is available from

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

4 Usage of Divergence Approximators in Machine Learning

In this section, we show applications of divergence approximators in machine learning.

4.1 Class-Prior Estimation under Class-Balance Change

In real-world pattern recognition tasks, changes in class balance are often observed between the training and test phases. In such cases, naive classifier training produces significant estimation bias because the class balance in the training dataset does not properly reflect that in the test dataset. Here, let us consider a binary pattern recognition task of classifying pattern $\boldsymbol{x}$ to class $y \in \{+1, -1\}$. The goal is to learn the class balance in a test dataset under a semi-supervised learning setup where unlabeled test samples are provided in addition to labeled training samples [42].

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

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

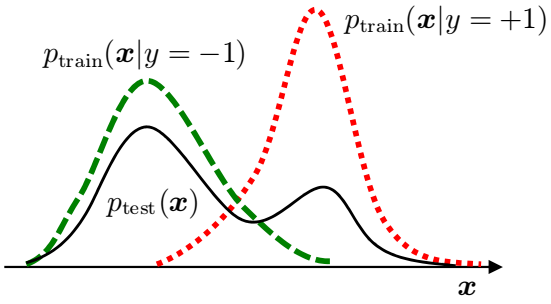


Figure 1: Schematic of class-prior estimation under class-balance change.

to the test input density $p_{\mathrm{test}}(\boldsymbol{x})$ under some divergence measure [4]. Here, $\pi \in [0,1]$ is a mixing coefficient to be learned to minimize the divergence (see Figure 1).

We use four *UCI benchmark datasets*[1] for numerical 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 graphs on the left-hand side of Figure 2 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 the two-step procedure of first estimating probability densities by kernel density estimation (KDE) and then learning π.

The graphs on the right-hand side of Figure 2 plot the test misclassification error obtained by a weighted ℓ_2-regularized kernel least-squares classifier [43] with weighted cross-validation [44]. The results show the LSDD-based method provides lower classification errors, which would be brought by good estimates of test class-balances.

4.2 Change-Detection in Time-Series

The goal 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

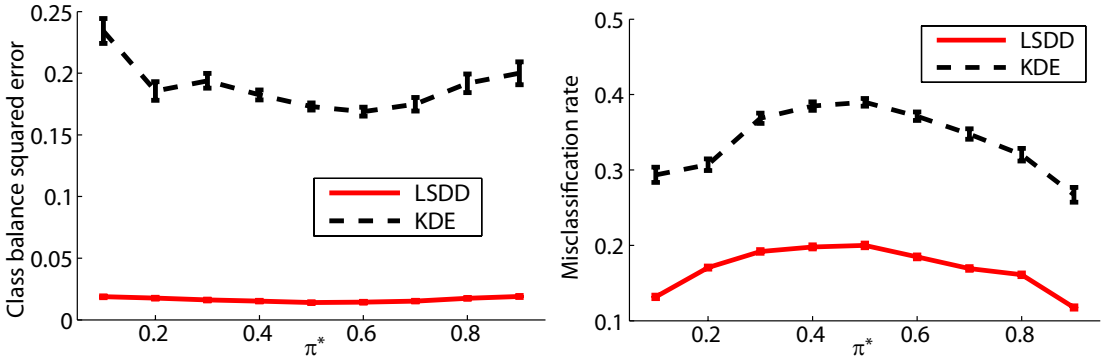
$$\boldsymbol{Y}(t) := [\boldsymbol{y}(t)^\top, \boldsymbol{y}(t+1)^\top, \ldots, \boldsymbol{y}(t+k-1)^\top]^\top \in \mathbb{R}^{km}$$

be a subsequence of time series at time t with length k. Instead of a single point $\boldsymbol{y}(t)$, the subsequence $\boldsymbol{Y}(t)$ is treated as a sample here, because time-dependent information can be naturally incorporated by this trick [3]. Let

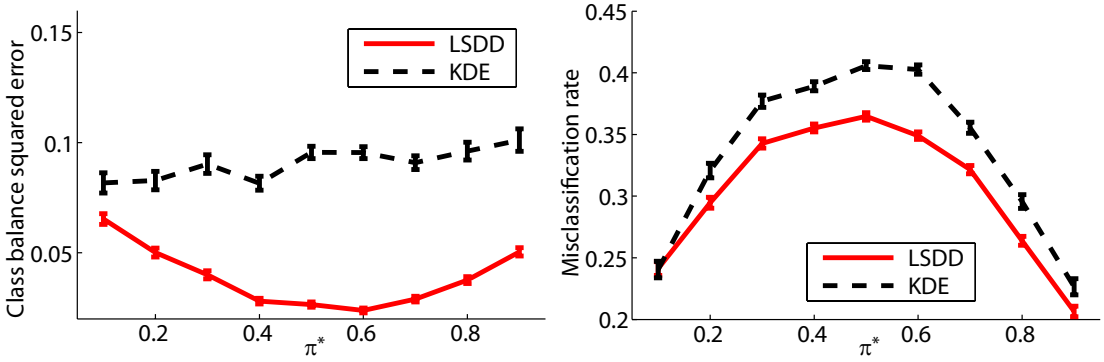
$$\mathcal{Y}(t) := \{\boldsymbol{Y}(t), \boldsymbol{Y}(t+1), \ldots, \boldsymbol{Y}(t+r-1)\}$$

be a set of r retrospective subsequence samples starting at time t. Then a divergence between the probability distributions of $\mathcal{Y}(t)$ and $\mathcal{Y}(t+r)$ may be used as the plausibility of change points (see Figure 3).

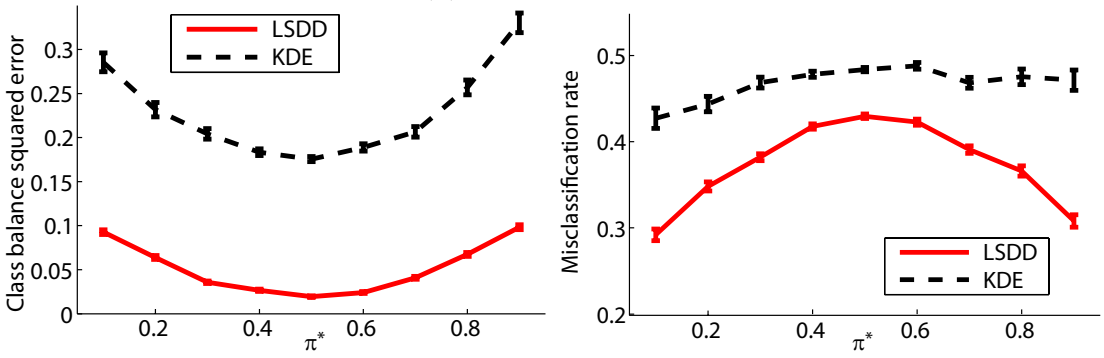
[1] `http://archive.ics.uci.edu/ml/`



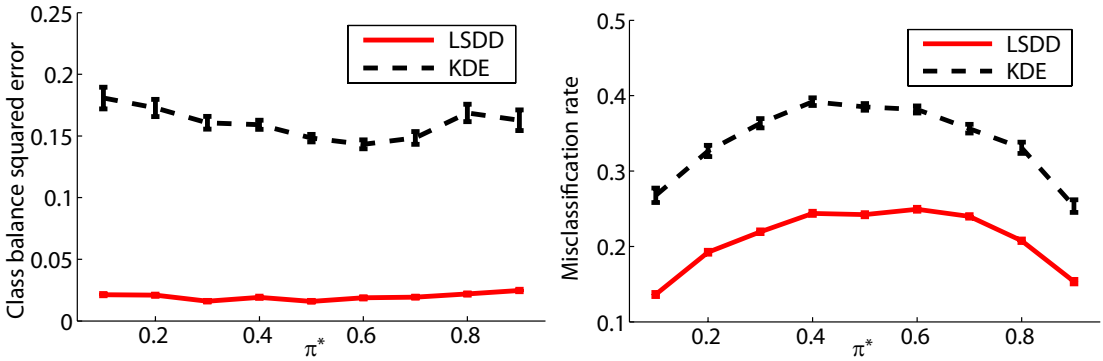
(a) Australian dataset



(b) Diabetes dataset



(c) German dataset



(d) Statlogheart dataset

Figure 2: Left: Squared error of class-prior estimation. Right: Misclassification error by a weighted ℓ_2-regularized kernel least-squares classifier with weighted cross-validation.

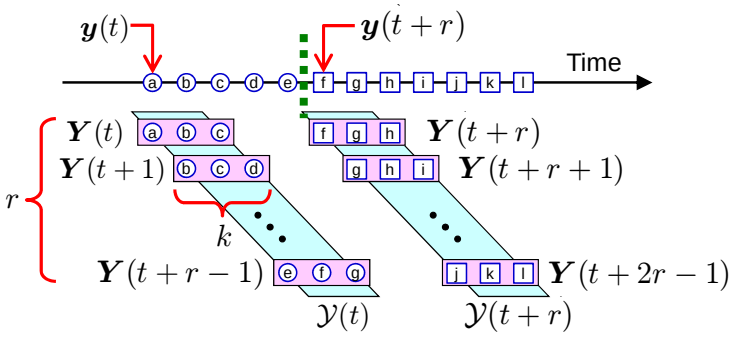


Figure 3: Schematic of change-point detection in time-series.

In Figure 4, we illustrate results of unsupervised change detection for the *IPSJ SIG-SLP Corpora and Environments for Noisy Speech Recognition* (CENSREC) dataset[2] that records human voice in noisy environments such as a restaurant, and the *Human Activity Sensing Consortium (HASC) challenge 2011* dataset[3] that provides human activity information collected by portable three-axis accelerometers. These graphs show that the KL-based method is excessively sensitive to noise and thus change points are not clearly detected. On the other hand, the L^2-based method more clearly indicates the existence of change points.

It was also demonstrated that divergence-based change-detection methods are useful in event detection from movies [6] and Twitter [7].

4.3 Salient Object Detection in an Image

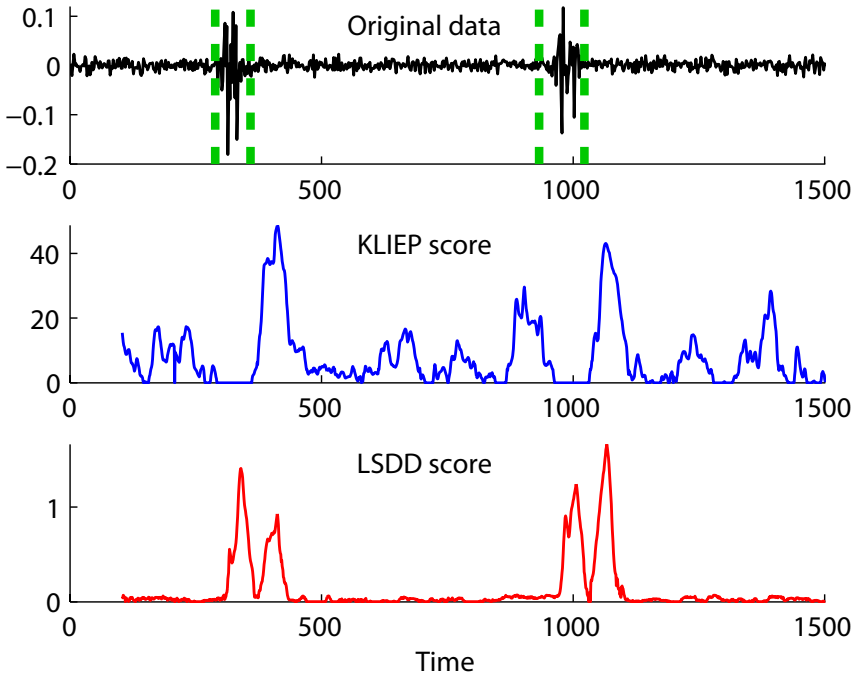
The goal is to find salient objects in an image. This can be achieved by computing a divergence between the probability distributions of image features (such as brightness, edges, and colors) in the center window and its surroundings [5]. This divergence computation is swept over the entire image with changing scales (Figure 5).

The object detection results on the *MSRA salient object database* [45] by the rPE divergence with $\alpha = 0.1$ are described in Figure 6, where pixels in gray-scale saliency maps take brighter color if the estimated divergence value is large. The results show that visually salient objects can be successfully detected by the divergence-based approach.

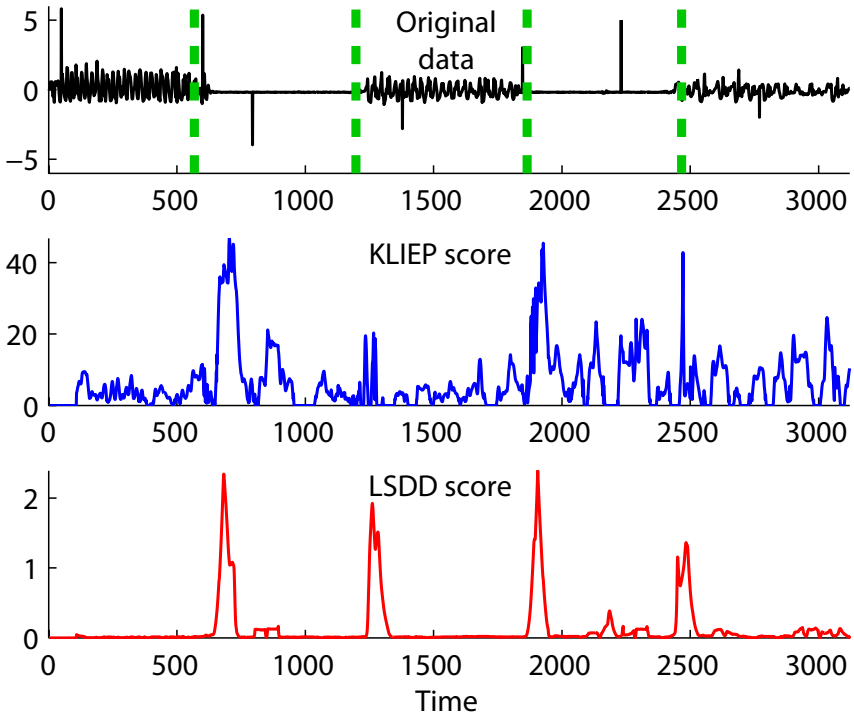
4.4 Measuring Statistical Independence

The goal is to measure how strongly two random variables $\boldsymbol{U}$ and $\boldsymbol{V}$ are statistically dependent, from paired samples $\{(\boldsymbol{u}_i, \boldsymbol{v}_i)\}_{i=1}^n$ drawn independently from the joint distribution with density $p_{\mathbf{U},\mathbf{V}}(\boldsymbol{u}, \boldsymbol{v})$. Let us consider a divergence between the joint density $p_{\mathbf{U},\mathbf{V}}$ and the product of marginal densities $p_{\mathbf{U}} \cdot p_{\mathbf{V}}$. This actually serves as a measure of statistical independence, because $\boldsymbol{U}$ and $\boldsymbol{V}$ are independent if and only if the divergence

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



(a) CENSREC dataset



(b) HASC dataset

Figure 4: Results of change-point detection. Original time-series data is plotted in the top graphs, and change scores obtained by KLIEP (Section 3.1) and LSDD (Section 3.4) are plotted in the bottom graphs.

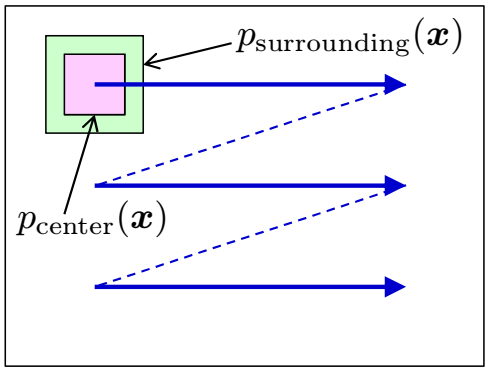


Figure 5: Schematic of salient object detection in an image.

is zero (i.e., $p_{\mathbf{U},\mathbf{V}} = p_{\mathbf{U}} \cdot p_{\mathbf{V}}$), and the dependence between $\boldsymbol{U}$ and $\boldsymbol{V}$ is stronger if the divergence is larger.

Such a dependence measure can be approximated in the same way as ordinary divergences by using the two datasets formed as $\mathcal{X} = \{(\boldsymbol{u}_i, \boldsymbol{v}_i)\}_{i=1}^n$ and $\mathcal{X}' = \{(\boldsymbol{u}_i, \boldsymbol{v}_j)\}_{i,j=1}^n$. The dependence measure based on the KL divergence is called *mutual information* (MI) [46]:

$$\mathrm{MI} := \iint p_{\mathbf{U},\mathbf{V}}(\boldsymbol{u}, \boldsymbol{v}) \log \frac{p_{\mathbf{U},\mathbf{V}}(\boldsymbol{u}, \boldsymbol{v})}{p_{\mathbf{U}}(\boldsymbol{u}) p_{\mathbf{V}}(\boldsymbol{v})} \mathrm{d}\boldsymbol{u}\mathrm{d}\boldsymbol{v}.$$

MI plays a central role in information theory [47].

On the other hand, its PE-divergence variant is called the *squared-loss mutual information* (SMI):

$$\mathrm{SMI} := \iint p_{\mathbf{U}}(\boldsymbol{u}) p_{\mathbf{V}}(\boldsymbol{v}) \left(\frac{p_{\mathbf{U},\mathbf{V}}(\boldsymbol{u}, \boldsymbol{v})}{p_{\mathbf{U}}(\boldsymbol{u}) p_{\mathbf{V}}(\boldsymbol{v})} - 1 \right)^2 \mathrm{d}\boldsymbol{u}\mathrm{d}\boldsymbol{v}.$$

SMI is useful for solving various machine learning tasks [8], including independence testing [9], feature selection [10, 11], feature extraction [12, 13], canonical dependency analysis [14], object matching [15], independent component analysis [16], clustering [17, 18], and causal direction estimation [19].

An L^2-distance variant of the dependence measure is called *quadratic mutual information* (QMI) [48]:

$$\mathrm{QMI} := \iint \Big(p_{\mathbf{U},\mathbf{V}}(\boldsymbol{u}, \boldsymbol{v}) - p_{\mathbf{U}}(\boldsymbol{u}) p_{\mathbf{V}}(\boldsymbol{v}) \Big)^2 \mathrm{d}\boldsymbol{u}\mathrm{d}\boldsymbol{v}.$$

QMI is also a useful dependence measure in practice [49].

5 Conclusions

In this article, we reviewed recent advances in direct divergence approximation. Direct divergence approximators theoretically achieve optimal convergence rates both in para-

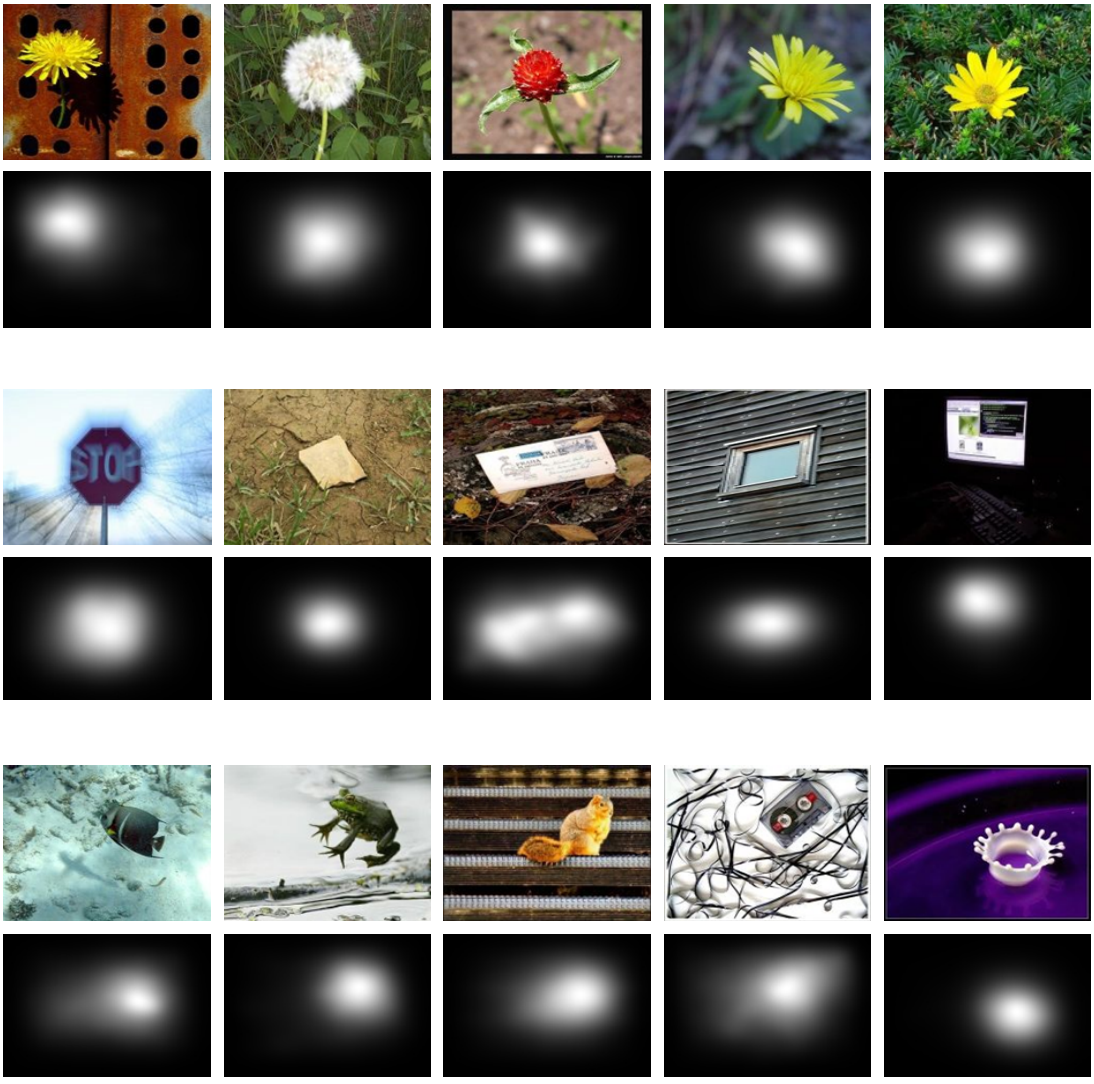


Figure 6: Results of salient object detection in an image. Upper: Original images. Lower: Obtained saliency maps (brighter color means more salient).

metric and non-parametric cases and experimentally compare favorably with the naive density-estimation counterparts [22, 21, 23, 24, 25].

However, direct divergence approximators still suffer from the *curse of dimensionality*. A possible cure for this problem is to combine them with dimensionality reduction, based on the hope that two probability distributions share some commonality [50, 51, 52]. Further investigating this line would be a promising future direction.

Acknowledgments

MS was supported by the JST PRESTO program, the FIRST program, and AOARD, SL was supported by the JST PRESTO program, MCdP was supported by the MEXT scholarship and the JST PRESTO program, TS was supported by MEXT KAKENHI 22700289. and TK was supported by JSPS KAKENHI 24500340.

References

[1] M. Sugiyama, T. Suzuki, Y. Itoh, T. Kanamori, and M. Kimura, "Least-squares two-sample test," *Neural Networks*, vol. 24, no. 7, pp. 735–751, 2011.

[2] T. Kanamori, T. Suzuki, and M. Sugiyama, "f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 708–720, 2012.

[3] Y. Kawahara and M. Sugiyama, "Sequential change-point detection based on direct density-ratio estimation," *Statistical Analysis and Data Mining*, vol. 5, no. 2, pp. 114–127, 2012.

[4] M. C. du Plessis and M. Sugiyama, "Semi-supervised learning of class balance under class-prior change by distribution matching," in *Proceedings of 29th International Conference on Machine Learning (ICML2012)*, J. Langford and J. Pineau, Eds., Edinburgh, Scotland, Jun. 26–Jul. 1 2012, pp. 823–830.

[5] M. Yamanaka, M. Matsugu, and M. Sugiyama, "Salient object detection based on direct density-ratio estimation," *IPSJ Transactions on Mathematical Modeling and Its Applications*, 2013, to appear.

[6] ——, "Detection of activities and events without explicit categorization," *IPSJ Transactions on Mathematical Modeling and Its Applications*, 2013, to appear.

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

[8] M. Sugiyama, "Machine learning with squared-loss mutual information," *Entropy*, vol. 15, no. 1, pp. 80–112, 2013.

[9] M. Sugiyama and T. Suzuki, "Least-squares independence test," *IEICE Transactions on Information and Systems*, vol. E94-D, no. 6, pp. 1333–1336, 2011.

[10] T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese, "Mutual information estimation reveals global associations between stimuli and biological processes," *BMC Bioinformatics*, vol. 10, no. 1, p. S52 (12 pages), 2009.

[11] W. Jitkrittum, H. Hachiya, and M. Sugiyama, "Feature selection via ℓ_1-penalized squared-loss mutual information," *IEICE Transactions on Information and Systems*, vol. E95-D, no. 7, 2013, to appear.

[12] T. Suzuki and M. Sugiyama, "Sufficient dimension reduction via squared-loss mutual information estimation," *Neural Computation*, vol. 3, no. 25, pp. 725–758, 2013.

[13] M. Yamada, G. Niu, J. Takagi, and M. Sugiyama, "Computationally efficient sufficient dimension reduction via squared-loss mutual information," in *Proceedings of the Third Asian Conference on Machine Learning (ACML2011)*, ser. JMLR Workshop and Conference Proceedings, C.-N. Hsu and W. S. Lee, Eds., vol. 20, Taoyuan, Taiwan, Nov. 13-15 2011, pp. 247–262.

[14] M. Karasuyama and Sugiyama, "Canonical dependency analysis based on squared-loss mutual information," *Neural Networks*, vol. 34, pp. 46–55, 2012.

[15] M. Yamada and M. Sugiyama, "Cross-domain object matching with model selection," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS2011)*, ser. JMLR Workshop and Conference Proceedings, G. Gordon, D. Dunson, and M. Dudík, Eds., vol. 15, Fort Lauderdale, Florida, USA, Apr. 11-13 2011, pp. 807–815.

[16] T. Suzuki and M. Sugiyama, "Least-squares independent component analysis," *Neural Computation*, vol. 23, no. 1, pp. 284–301, 2011.

[17] M. Sugiyama, M. Yamada, M. Kimura, and H. Hachiya, "On information-maximization clustering: Tuning parameter selection and analytic solution," in *Proceedings of 28th International Conference on Machine Learning (ICML2011)*, L. Getoor and T. Scheffer, Eds., Bellevue, Washington, USA, Jun. 28–Jul. 2 2011, pp. 65–72.

[18] M. Kimura and M. Sugiyama, "Dependence-maximization clustering with least-squares mutual information," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 15, no. 7, pp. 800–805, 2011.

[19] M. Yamada and M. Sugiyama, "Dependence minimizing regression with model selection for non-linear causal inference under non-Gaussian noise," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI2010)*. Atlanta, Georgia, USA: The AAAI Press, Jul. 11–15 2010, pp. 643–648.

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

[21] 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*, vol. 60, no. 4, pp. 699–746, 2008.

[22] 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*, vol. 56, no. 11, pp. 5847–5861, 2010.

[23] T. Kanamori, S. Hido, and M. Sugiyama, "A least-squares approach to direct importance estimation," *Journal of Machine Learning Research*, vol. 10, pp. 1391–1445, Jul. 2009.

[24] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama, "Relative density-ratio estimation for robust distribution comparison," *Neural Computation*, vol. 25, no. 5, pp. 1324–1370, 2013.

[25] M. Sugiyama, T. Suzuki, T. Kanamori, M. C. du Plessis, S. Liu, and I. Takeuchi, "Density-difference estimation," *Neural Computation*, 2013, to appear.

[26] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.

[27] S. Amari and H. Nagaoka, *Methods of Information Geometry*. Providence, RI, USA: Oxford University Press, 2000.

[28] M. Sugiyama, T. Suzuki, and T. Kanamori, *Density Ratio Estimation in Machine Learning*. Cambridge, UK: Cambridge University Press, 2012.

[29] C. Cortes, Y. Mansour, and M. Mohri, "Learning bounds for importance weighting," in *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. K. I. Williams, R. Zemel, J. Shawe-Taylor, and A. Culotta, Eds., 2010, pp. 442–450.

[30] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philosophical Magazine Series 5*, vol. 50, no. 302, pp. 157–175, 1900.

[31] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society, Series B*, vol. 28, no. 1, pp. 131–142, 1966.

[32] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 229–318, 1967.

[33] M. Sugiyama, T. Suzuki, and T. Kanamori, "Density ratio matching under the Bregman divergence: A unified framework of density ratio estimation," *Annals of the Institute of Statistical Mathematics*, vol. 64, no. 5, pp. 1009–1044, 2012.

[34] Y. Tsuboi, H. Kashima, S. Hido, S. Bickel, and M. Sugiyama, "Direct density ratio estimation for large-scale covariate shift adaptation," *Journal of Information Processing*, vol. 17, pp. 138–155, 2009.

[35] M. Yamada and M. Sugiyama, "Direct importance estimation with Gaussian mixture models," *IEICE Transactions on Information and Systems*, vol. E92-D, no. 10, pp. 2159–2162, 2009.

[36] M. Yamada, M. Sugiyama, G. Wichern, and J. Simm, "Direct importance estimation with a mixture of probabilistic principal component analyzers," *IEICE Transactions on Information and Systems*, vol. E93-D, no. 10, pp. 2846–2849, 2010.

[37] A. Keziou, "Dual representation of ϕ-divergences and applications," *Comptes Rendus Mathématique*, vol. 336, no. 10, pp. 857–862, 2003.

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

[39] R. Tibshirani, "Regression shrinkage and subset selection with the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[40] R. Tomioka, T. Suzuki, and M. Sugiyama, "Super-linear convergence of dual augmented Lagrangian algorithm for sparsity regularized estimation," *Journal of Machine Learning Research*, vol. 12, pp. 1537–1586, May 2011.

[41] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

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

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

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

[45] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2011.

[46] C. Shannon, "A mathematical theory of communication," *Bell Systems Technical Journal*, vol. 27, pp. 379–423, 1948.

[47] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2006.

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

[49] J. Sainui and M. Sugiyama, "Direct approximation of quadratic mutual information and its application to dependence-maximization clustering," *IEICE Transactions on Information and Systems*, 2013, submitted.

[50] M. Sugiyama, M. Kawanabe, and P. L. Chui, "Dimensionality reduction for density ratio estimation in high-dimensional spaces," *Neural Networks*, vol. 23, no. 1, pp. 44–59, 2010.

[51] M. Sugiyama, M. Yamada, P. von Bünau, T. Suzuki, T. Kanamori, and M. Kawanabe, "Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search," *Neural Networks*, vol. 24, no. 2, pp. 183–198, 2011.

[52] M. Yamada and M. Sugiyama, "Direct density-ratio estimation with dimensionality reduction via hetero-distributional subspace analysis," in *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI2011)*. San Francisco, California, USA: The AAAI Press, Aug. 7–11 2011, pp. 549–554.