Relative Density-Ratio Estimation for Robust Distribution Comparison

Makoto Yamada
Tokyo Institute of Technology
yamada@sg.cs.titech.ac.jp

Taiji Suzuki
The University of Tokyo
s-taiji@stat.t.u-tokyo.ac.jp

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

Hirotaka Hachiya **Masashi Sugiyama**
Tokyo Institute of Technology
{hachiya@sg. sugi@}cs.titech.ac.jp

Abstract

Divergence estimators based on direct approximation of density-ratios without going through separate approximation of numerator and denominator densities have been successfully applied to machine learning tasks that involve distribution comparison such as outlier detection, transfer learning, and two-sample homogeneity test. However, since density-ratio functions often possess high fluctuation, divergence estimation is still a challenging task in practice. In this paper, we propose to use *relative divergences* for distribution comparison, which involves approximation of *relative density-ratios*. Since relative density-ratios are always smoother than corresponding ordinary density-ratios, our proposed method is favorable in terms of the non-parametric convergence speed. Furthermore, we show that the proposed divergence estimator has asymptotic variance *independent* of the model complexity under a parametric setup, implying that the proposed estimator hardly overfits even with complex models. Through experiments, we demonstrate the usefulness of the proposed approach.

1 Introduction

Comparing probability distributions is a fundamental task in statistical data processing. It can be used for, e.g., *outlier detection* [1, 2], *two-sample homogeneity test* [3, 4], and transfer learning [5, 6].

A standard approach to comparing probability densities $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ would be to estimate a divergence from $p(\boldsymbol{x})$ to $p'(\boldsymbol{x})$, such as the *Kullback-Leibler (KL) divergence* [7]:

$$\mathrm{KL}[p(\boldsymbol{x}), p'(\boldsymbol{x})] := \mathbb{E}_{p(\boldsymbol{x})}\left[\log r(\boldsymbol{x})\right], \quad r(\boldsymbol{x}) := p(\boldsymbol{x})/p'(\boldsymbol{x}),$$

where $\mathbb{E}_{p(\boldsymbol{x})}$ denotes the expectation over $p(\boldsymbol{x})$. A naive way to estimate the KL divergence is to separately approximate the densities $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ from data and plug the estimated densities in the above definition. However, since density estimation is known to be a hard task [8], this approach does not work well unless a good parametric model is available. Recently, a divergence estimation approach which directly approximates the *density-ratio* $r(\boldsymbol{x})$ without going through separate approximation of densities $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ has been proposed [9, 10]. Such density-ratio approximation methods were proved to achieve the optimal non-parametric convergence rate in the mini-max sense.

However, the KL divergence estimation via density-ratio approximation is computationally rather expensive due to the non-linearity introduced by the 'log' term. To cope with this problem, another divergence called the *Pearson (PE) divergence* [11] is useful. The PE divergence is defined as

$$\mathrm{PE}[p(\boldsymbol{x}), p'(\boldsymbol{x})] := \tfrac{1}{2}\mathbb{E}_{p'(\boldsymbol{x})}\left[(r(\boldsymbol{x})-1)^2\right].$$

The PE divergence is a squared-loss variant of the KL divergence, and they both belong to the class of the *Ali-Silvey-Csiszár divergences* (which is also known as the *f-divergences*, see [12, 13]). Thus, the PE and KL divergences share similar properties, e.g., they are non-negative and vanish if and only if $p(\boldsymbol{x}) = p'(\boldsymbol{x})$.

Similarly to the KL divergence estimation, the PE divergence can also be accurately estimated based on density-ratio approximation [14]: the density-ratio approximator called *unconstrained least-squares importance fitting* (uLSIF) gives the PE divergence estimator *analytically*, which can be computed just by solving a system of linear equations. The practical usefulness of the uLSIF-based PE divergence estimator was demonstrated in various applications such as outlier detection [2], two-sample homogeneity test [4], and dimensionality reduction [15].

In this paper, we first establish the non-parametric convergence rate of the uLSIF-based PE divergence estimator, which elucidates its superior theoretical properties. However, it also reveals that its convergence rate is actually governed by the 'sup'-norm of the true density-ratio function: $\max_{\boldsymbol{x}} r(\boldsymbol{x})$. This implies that, in the region where the denominator density $p'(\boldsymbol{x})$ takes small values, the density-ratio $r(\boldsymbol{x}) = p(\boldsymbol{x})/p'(\boldsymbol{x})$ tends to take large values and therefore the overall convergence speed becomes slow. More critically, density-ratios can even diverge to infinity under a rather simple setting, e.g., when the ratio of two Gaussian functions is considered [16]. This makes the paradigm of divergence estimation based on density-ratio approximation unreliable.

In order to overcome this fundamental problem, we propose an alternative approach to distribution comparison called *α-relative divergence estimation*. In the proposed approach, we estimate the *α-relative divergence*, which is the divergence from $p(\boldsymbol{x})$ to the *α-mixture density*:

$$q_\alpha(\boldsymbol{x}) = \alpha p(\boldsymbol{x}) + (1-\alpha)p'(\boldsymbol{x}) \quad \text{for } 0 \leq \alpha < 1.$$

For example, the α-relative PE divergence is given by

$$\mathrm{PE}_\alpha[p(\boldsymbol{x}), p'(\boldsymbol{x})] := \mathrm{PE}[p(\boldsymbol{x}), q_\alpha(\boldsymbol{x})] = \tfrac{1}{2}\mathbb{E}_{q_\alpha(\boldsymbol{x})}\left[(r_\alpha(\boldsymbol{x}) - 1)^2\right], \tag{1}$$

where $r_\alpha(\boldsymbol{x})$ is the *α-relative density-ratio* of $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$:

$$r_\alpha(\boldsymbol{x}) := p(\boldsymbol{x})/q_\alpha(\boldsymbol{x}) = p(\boldsymbol{x})/\Big(\alpha p(\boldsymbol{x}) + (1-\alpha)p'(\boldsymbol{x})\Big). \tag{2}$$

We propose to estimate the α-relative divergence by direct approximation of the *α-relative density-ratio*.

A notable advantage of this approach is that the α-relative density-ratio is always bounded above by $1/\alpha$ when $\alpha > 0$, even when the ordinary density-ratio is unbounded. Based on this feature, we theoretically show that the α-relative PE divergence estimator based on α-relative density-ratio approximation is more favorable than the ordinary density-ratio approach in terms of the non-parametric convergence speed.

We further prove that, under a correctly-specified parametric setup, the asymptotic variance of our α-relative PE divergence estimator does not depend on the model complexity. This means that the proposed α-relative PE divergence estimator hardly overfits even with complex models.

Through experiments on outlier detection, two-sample homogeneity test, and transfer learning, we demonstrate that our proposed α-relative PE divergence estimator compares favorably with alternative approaches.

2 Estimation of Relative Pearson Divergence via Least-Squares Relative Density-Ratio Approximation

Suppose we are given independent and identically distributed (i.i.d.) samples $\{\boldsymbol{x}_i\}_{i=1}^n$ from a d-dimensional distribution P with density $p(\boldsymbol{x})$ and i.i.d. samples $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$ from another d-dimensional distribution P' with density $p'(\boldsymbol{x})$. Our goal is to compare the two underlying distributions P and P' only using the two sets of samples $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$.

In this section, we give a method for estimating the α-relative PE divergence based on direct approximation of the α-relative density-ratio.

Direct Approximation of α-Relative Density-Ratios: Let us model the α-relative density-ratio $r_\alpha(\boldsymbol{x})$ (2) by the following kernel model $g(\boldsymbol{x};\boldsymbol{\theta}) := \sum_{\ell=1}^{n}\theta_\ell K(\boldsymbol{x},\boldsymbol{x}_\ell)$, where $\boldsymbol{\theta} := (\theta_1,\ldots,\theta_n)^\top$ are parameters to be learned from data samples, $^\top$ denotes the transpose of a matrix or a vector, and $K(\boldsymbol{x},\boldsymbol{x}')$ is a kernel basis function. In the experiments, we use the Gaussian kernel.

The parameters $\boldsymbol{\theta}$ in the model $g(\boldsymbol{x};\boldsymbol{\theta})$ are determined so that the following expected squared-error J is minimized:

$$\begin{aligned} J(\boldsymbol{\theta}) &:= \tfrac{1}{2}\mathbb{E}_{q_\alpha(\boldsymbol{x})}\left[\left(g(\boldsymbol{x};\boldsymbol{\theta}) - r_\alpha(\boldsymbol{x})\right)^2\right] \\ &= \tfrac{\alpha}{2}\mathbb{E}_{p(\boldsymbol{x})}\left[g(\boldsymbol{x};\boldsymbol{\theta})^2\right] + \tfrac{(1-\alpha)}{2}\mathbb{E}_{p'(\boldsymbol{x})}\left[g(\boldsymbol{x};\boldsymbol{\theta})^2\right] - \mathbb{E}_{p(\boldsymbol{x})}\left[g(\boldsymbol{x};\boldsymbol{\theta})\right] + \text{Const.}, \end{aligned}$$

where we used $r_\alpha(\boldsymbol{x})q_\alpha(\boldsymbol{x}) = p(\boldsymbol{x})$ in the third term. Approximating the expectations by empirical averages, we obtain the following optimization problem:

$$\widehat{\boldsymbol{\theta}} := \mathop{\mathrm{argmin}}_{\boldsymbol{\theta}\in\mathbb{R}^n}\left[\tfrac{1}{2}\boldsymbol{\theta}^\top\widehat{\boldsymbol{H}}\boldsymbol{\theta} - \widehat{\boldsymbol{h}}^\top\boldsymbol{\theta} + \tfrac{\lambda}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right], \tag{3}$$

where a penalty term $\lambda\boldsymbol{\theta}^\top\boldsymbol{\theta}/2$ is included for regularization purposes, and λ (≥ 0) denotes the regularization parameter. $\widehat{\boldsymbol{H}}$ and $\widehat{\boldsymbol{h}}$ are defined as

$$\widehat{H}_{\ell,\ell'} := \tfrac{\alpha}{n}\sum_{i=1}^{n}K(\boldsymbol{x}_i,\boldsymbol{x}_\ell)K(\boldsymbol{x}_i,\boldsymbol{x}_{\ell'}) + \tfrac{(1-\alpha)}{n'}\sum_{j=1}^{n'}K(\boldsymbol{x}'_j,\boldsymbol{x}_\ell)K(\boldsymbol{x}'_j,\boldsymbol{x}_{\ell'}),\ \widehat{h}_\ell := \tfrac{1}{n}\sum_{i=1}^{n}K(\boldsymbol{x}_i,\boldsymbol{x}_\ell).$$

It is easy to confirm that the solution of Eq.(3) can be *analytically* obtained as $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{H}} + \lambda\boldsymbol{I}_n)^{-1}\widehat{\boldsymbol{h}}$, where $\boldsymbol{I}_n$ denotes the n-dimensional identity matrix. Finally, a density-ratio estimator is given as

$$\widehat{r}_\alpha(\boldsymbol{x}) := g(\boldsymbol{x};\widehat{\boldsymbol{\theta}}) = \sum_{\ell=1}^{n}\widehat{\theta}_\ell K(\boldsymbol{x},\boldsymbol{x}_\ell).$$

When $\alpha = 0$, the above method is reduced to a direct density-ratio estimator called *unconstrained least-squares importance fitting* (uLSIF) [14]. Thus, the above method can be regarded as an extension of uLSIF to the α-relative density-ratio. For this reason, we refer to our method as *relative uLSIF* (RuLSIF).

The performance of RuLSIF depends on the choice of the kernel function (the kernel width in the case of the Gaussian kernel) and the regularization parameter λ. Model selection of RuLSIF is possible based on cross-validation (CV) with respect to the squared-error criterion J.

Using an estimator of the α-relative density-ratio $r_\alpha(\boldsymbol{x})$, we can construct estimators of the α-relative PE divergence (1). After a few lines of calculation, we can show that the α-relative PE divergence (1) is equivalently expressed as

$$\mathrm{PE}_\alpha = -\tfrac{\alpha}{2}\mathbb{E}_{p(\boldsymbol{x})}\left[r_\alpha(\boldsymbol{x})^2\right] - \tfrac{(1-\alpha)}{2}\mathbb{E}_{p'(\boldsymbol{x})}\left[r_\alpha(\boldsymbol{x})^2\right] + \mathbb{E}_{p(\boldsymbol{x})}\left[r_\alpha(\boldsymbol{x})\right] - \tfrac{1}{2} = \tfrac{1}{2}\mathbb{E}_{p(\boldsymbol{x})}\left[r_\alpha(\boldsymbol{x})\right] - \tfrac{1}{2}.$$

Note that the middle expression can also be obtained via *Legendre-Fenchel convex duality* of the divergence functional [17].

Based on these expressions, we consider the following two estimators:

$$\widehat{\mathrm{PE}}_\alpha := -\tfrac{\alpha}{2n}\sum_{i=1}^{n}\widehat{r}_\alpha(\boldsymbol{x}_i)^2 - \tfrac{(1-\alpha)}{2n'}\sum_{j=1}^{n'}\widehat{r}_\alpha(\boldsymbol{x}'_j)^2 + \tfrac{1}{n}\sum_{i=1}^{n}\widehat{r}_\alpha(\boldsymbol{x}_i) - \tfrac{1}{2}, \tag{4}$$

$$\widetilde{\mathrm{PE}}_\alpha := \tfrac{1}{2n}\sum_{i=1}^{n}\widehat{r}_\alpha(\boldsymbol{x}_i) - \tfrac{1}{2}. \tag{5}$$

We note that the α-relative PE divergence (1) can have further different expressions than the above ones, and corresponding estimators can also be constructed similarly. However, the above two expressions will be particularly useful: the first estimator $\widehat{\mathrm{PE}}_\alpha$ has superior theoretical properties (see Section 3) and the second one $\widetilde{\mathrm{PE}}_\alpha$ is simple to compute.

3 Theoretical Analysis

In this section, we analyze theoretical properties of the proposed PE divergence estimators. Since our theoretical analysis is highly technical, we focus on explaining practical insights we can gain from the theoretical results here; we describe all the mathematical details in the supplementary material.

For theoretical analysis, let us consider a rather abstract form of our relative density-ratio estimator described as

$$\mathrm{argmin}_{g \in \mathcal{G}} \left[\tfrac{\alpha}{2n} \textstyle\sum_{i=1}^{n} g(\boldsymbol{x}_i)^2 + \tfrac{(1-\alpha)}{2n'} \sum_{j=1}^{n'} g(\boldsymbol{x}'_j)^2 - \tfrac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{x}_i) + \tfrac{\lambda}{2} R(g)^2 \right], \qquad (6)$$

where $\mathcal{G}$ is some function space (i.e., a statistical model) and $R(\cdot)$ is some regularization functional.

Non-Parametric Convergence Analysis: First, we elucidate the non-parametric convergence rate of the proposed PE estimators. Here, we practically regard the function space $\mathcal{G}$ as an infinite-dimensional *reproducing kernel Hilbert space* (RKHS) [18] such as the Gaussian kernel space, and $R(\cdot)$ as the associated RKHS norm.

Let us represent the complexity of the function space $\mathcal{G}$ by γ ($0 < \gamma < 2$); the larger γ is, the more complex the function class $\mathcal{G}$ is (see the supplementary material for its precise definition). We analyze the convergence rate of our PE divergence estimators as $\bar{n} := \min(n, n')$ tends to infinity for $\lambda = \lambda_{\bar{n}}$ under

$$\lambda_{\bar{n}} \to o(1) \ \text{ and } \ \lambda_{\bar{n}}^{-1} = o(\bar{n}^{2/(2+\gamma)}).$$

The first condition means that $\lambda_{\bar{n}}$ tends to zero, but the second condition means that its shrinking speed should not be too fast.

Under several technical assumptions detailed in the supplementary material, we have the following asymptotic convergence results for the two PE divergence estimators $\widehat{\mathrm{PE}}_\alpha$ (4) and $\widetilde{\mathrm{PE}}_\alpha$ (5):

$$\widehat{\mathrm{PE}}_\alpha - \mathrm{PE}_\alpha = \mathcal{O}_p(\bar{n}^{-1/2} c \|r_\alpha\|_\infty + \lambda_{\bar{n}} \max(1, R(r_\alpha)^2)), \qquad (7)$$

$$\begin{aligned} \widetilde{\mathrm{PE}}_\alpha - \mathrm{PE}_\alpha = \mathcal{O}_p\Big(& \lambda_{\bar{n}}^{1/2} \|r_\alpha\|_\infty^{1/2} \max\{1, R(r_\alpha)\} \\ & + \lambda_{\bar{n}} \max\{1, \|r_\alpha\|_\infty^{(1-\gamma/2)/2}, R(r_\alpha)\|r_\alpha\|_\infty^{(1-\gamma/2)/2}, R(r_\alpha)\}\Big), \end{aligned} \qquad (8)$$

where $\mathcal{O}_p$ denotes the asymptotic order in probability,

$$c := (1+\alpha)\sqrt{\mathbb{V}_{p(\boldsymbol{x})}[r_\alpha(\boldsymbol{x})]} + (1-\alpha)\sqrt{\mathbb{V}_{p'(\boldsymbol{x})}[r_\alpha(\boldsymbol{x})]},$$

and $\mathbb{V}_{p(\boldsymbol{x})}$ denotes the variance over $p(\boldsymbol{x})$:

$$\mathbb{V}_{p(\boldsymbol{x})}[f(\boldsymbol{x})] = \textstyle\int \left(f(\boldsymbol{x}) - \int f(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right)^2 p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

In both Eq.(7) and Eq.(8), the coefficients of the leading terms (i.e., the first terms) of the asymptotic convergence rates become smaller as $\|r_\alpha\|_\infty$ gets smaller. Since

$$\|r_\alpha\|_\infty = \left\| \left(\alpha + (1-\alpha)/r(\boldsymbol{x}) \right)^{-1} \right\|_\infty < \tfrac{1}{\alpha} \quad \text{for } \alpha > 0,$$

larger α would be more preferable in terms of the asymptotic approximation error. Note that when $\alpha = 0$, $\|r_\alpha\|_\infty$ can tend to infinity even under a simple setting that the ratio of two Gaussian functions is considered [16]. Thus, our proposed approach of estimating the α-relative PE divergence (with $\alpha > 0$) would be more advantageous than the naive approach of estimating the plain PE divergence (which corresponds to $\alpha = 0$) in terms of the non-parametric convergence rate.

The above results also show that $\widehat{\mathrm{PE}}_\alpha$ and $\widetilde{\mathrm{PE}}_\alpha$ have different asymptotic convergence rates. The leading term in Eq.(7) is of order $\bar{n}^{-1/2}$, while the leading term in Eq.(8) is of order $\lambda_{\bar{n}}^{1/2}$, which is slightly slower (depending on the complexity γ) than $\bar{n}^{-1/2}$. Thus, $\widehat{\mathrm{PE}}_\alpha$ would be more accurate than $\widetilde{\mathrm{PE}}_\alpha$ in large sample cases. Furthermore, when $p(\boldsymbol{x}) = p'(\boldsymbol{x})$, $\mathbb{V}_{p(\boldsymbol{x})}[r_\alpha(\boldsymbol{x})] = 0$ holds and thus $c = 0$ holds. Then the leading term in Eq.(7) vanishes and therefore $\widehat{\mathrm{PE}}_\alpha$ has the even faster convergence rate of order $\lambda_{\bar{n}}$, which is slightly slower (depending on the complexity γ) than $\bar{n}^{-1}$. Similarly, if α is close to 1, $r_\alpha(\boldsymbol{x}) \approx 1$ and thus $c \approx 0$ holds.

When $\bar{n}$ is not large enough to be able to neglect the terms of $o(\bar{n}^{-1/2})$, the terms of $O(\lambda_{\bar{n}})$ matter. If $\|r_\alpha\|_\infty$ and $R(r_\alpha)$ are large (this can happen, e.g., when α is close to 0), the coefficient of the $O(\lambda_{\bar{n}})$-term in Eq.(7) can be larger than that in Eq.(8). Then $\widetilde{\mathrm{PE}}_\alpha$ would be more favorable than $\widehat{\mathrm{PE}}_\alpha$ in terms of the approximation accuracy.

See the supplementary material for numerical examples illustrating the above theoretical results.

Parametric Variance Analysis: Next, we analyze the asymptotic variance of the PE divergence estimator $\widehat{\mathrm{PE}}_\alpha$ (4) under a parametric setup.

As the function space $\mathcal{G}$ in Eq.(6), we consider the following parametric model: $\mathcal{G} = \{g(\boldsymbol{x};\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^b\}$ for a finite b. Here we assume that this parametric model is *correctly specified*, i.e., it includes the true relative density-ratio function $r_\alpha(\boldsymbol{x})$: there exists $\boldsymbol{\theta}^*$ such that $g(\boldsymbol{x};\boldsymbol{\theta}^*) = r_\alpha(\boldsymbol{x})$. Here, we use RuLSIF without regularization, i.e., $\lambda = 0$ in Eq.(6).

Let us denote the variance of $\widehat{\mathrm{PE}}_\alpha$ (4) by $\mathbb{V}[\widehat{\mathrm{PE}}_\alpha]$, where randomness comes from the draw of samples $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{x}'_j\}_{j=1}^{n'}$. Then, under a standard regularity condition for the asymptotic normality [19], $\mathbb{V}[\widehat{\mathrm{PE}}_\alpha]$ can be expressed and upper-bounded as

$$\begin{aligned}\mathbb{V}[\widehat{\mathrm{PE}}_\alpha] &= \mathbb{V}_{p(\boldsymbol{x})}\left[r_\alpha - \alpha r_\alpha(\boldsymbol{x})^2/2\right]/n + \mathbb{V}_{p'(\boldsymbol{x})}\left[(1-\alpha) r_\alpha(\boldsymbol{x})^2/2\right]/n' + o(n^{-1}, n'^{-1}) && (9)\\ &\leq \|r_\alpha\|_\infty^2/n + \alpha^2\|r_\alpha\|_\infty^4/(4n) + (1-\alpha)^2\|r_\alpha\|_\infty^4/(4n') + o(n^{-1}, n'^{-1}). && (10)\end{aligned}$$

Let us denote the variance of $\widetilde{\mathrm{PE}}_\alpha$ by $\mathbb{V}[\widetilde{\mathrm{PE}}_\alpha]$. Then, under a standard regularity condition for the asymptotic normality [19], the variance of $\widetilde{\mathrm{PE}}_\alpha$ is asymptotically expressed as

$$\begin{aligned}\mathbb{V}[\widetilde{\mathrm{PE}}_\alpha] = \mathbb{V}_{p(\boldsymbol{x})}&\left[\left(r_\alpha + (1-\alpha r_\alpha)\mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \boldsymbol{H}_\alpha^{-1}\nabla g\right)/2\right]/n\\ &+ \mathbb{V}_{p'(\boldsymbol{x})}\left[\left((1-\alpha) r_\alpha \mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \boldsymbol{H}_\alpha^{-1}\nabla g\right)/2\right]/n' + o(n^{-1}, n'^{-1}), \qquad (11)\end{aligned}$$

where ∇g is the gradient vector of g with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ and

$$\boldsymbol{H}_\alpha = \alpha\mathbb{E}_{p(\boldsymbol{x})}[\nabla g \nabla g^\top] + (1-\alpha)\mathbb{E}_{p'(\boldsymbol{x})}[\nabla g \nabla g^\top].$$

Eq.(9) shows that, up to $O(n^{-1}, n'^{-1})$, the variance of $\widehat{\mathrm{PE}}_\alpha$ depends only on the true relative density-ratio $r_\alpha(\boldsymbol{x})$, not on the estimator of $r_\alpha(\boldsymbol{x})$. This means that the model complexity does not affect the asymptotic variance. Therefore, *overfitting* would hardly occur in the estimation of the relative PE divergence even when complex models are used. We note that the above superior property is applicable only to relative PE divergence estimation, not to relative density-ratio estimation. This implies that overfitting occurs in relative density-ratio estimation, but the approximation error cancels out in relative PE divergence estimation.

On the other hand, Eq.(11) shows that the variance of $\widetilde{\mathrm{PE}}_\alpha$ is affected by the model $\mathcal{G}$, since the factor $\mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \boldsymbol{H}_\alpha^{-1}\nabla g$ depends on the model in general. When the equality $\mathbb{E}_{p(\boldsymbol{x})}[\nabla g]^\top \boldsymbol{H}_\alpha^{-1}\nabla g(\boldsymbol{x};\boldsymbol{\theta}^*) = r_\alpha(\boldsymbol{x})$ holds, the variances of $\widetilde{\mathrm{PE}}_\alpha$ and $\widehat{\mathrm{PE}}_\alpha$ are asymptotically the same. However, in general, the use of $\widehat{\mathrm{PE}}_\alpha$ would be more recommended.

Eq.(10) shows that the variance $\mathbb{V}[\widehat{\mathrm{PE}}_\alpha]$ can be upper-bounded by the quantity depending on $\|r_\alpha\|_\infty$, which is monotonically lowered if $\|r_\alpha\|_\infty$ is reduced. Since $\|r_\alpha\|_\infty$ monotonically decreases as α increases, our proposed approach of estimating the α-relative PE divergence (with $\alpha > 0$) would be more advantageous than the naive approach of estimating the plain PE divergence (which corresponds to $\alpha = 0$) in terms of the parametric asymptotic variance.

See the supplementary material for numerical examples illustrating the above theoretical results.

4 Experiments

In this section, we experimentally evaluate the performance of the proposed method in two-sample homogeneity test, outlier detection, and transfer learning tasks.

Two-Sample Homogeneity Test: First, we apply the proposed divergence estimator to two-sample homogeneity test.

Given two sets of samples $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P$ and $\mathcal{X}' = \{\boldsymbol{x}'_j\}_{j=1}^{n'} \overset{\text{i.i.d.}}{\sim} P'$, the goal of the two-sample homogeneity test is to test the *null hypothesis* that the probability distributions P and P' are the same against its complementary alternative (i.e., the distributions are different). By using an estimator $\widehat{\mathrm{Div}}$ of some divergence between the two distributions P and P', homogeneity of two distributions can be tested based on the *permutation test* procedure [20].

Table 1: Experimental results of two-sample test. The mean (and standard deviation in the bracket) rate of accepting the null hypothesis (i.e., $P = P'$) for IDA benchmark repository under the significance level 5% is reported. Left: when the two sets of samples are both taken from the positive training set (i.e., the null hypothesis is correct). Methods having the mean acceptance rate 0.95 according to the *one-sample t-test* at the significance level 5% are specified by bold face. Right: when the set of samples corresponding to the numerator of the density-ratio are taken from the positive training set and the set of samples corresponding to the denominator of the density-ratio are taken from the positive training set and the negative training set (i.e., the null hypothesis is not correct). The best method having the lowest mean accepting rate and comparable methods according to the *two-sample t-test* at the significance level 5% are specified by bold face.

			$P = P'$				$P \neq P'$			
Datasets	d	$n = n'$	MMD	LSTT (α = 0.0)	LSTT (α = 0.5)	LSTT (α = 0.95)	MMD	LSTT (α = 0.0)	LSTT (α = 0.5)	LSTT (α = 0.95)
banana	2	100	**.96 (.20)**	**.93 (.26)**	**.92 (.27)**	**.92 (.27)**	.52 (.50)	**.10 (.30)**	**.02 (.14)**	**.17 (.38)**
thyroid	5	19	**.96 (.20)**	**.95 (.22)**	**.95 (.22)**	.88 (.33)	**.52 (.50)**	.81 (.39)	**.65 (.48)**	.80 (.40)
titanic	5	21	**.94 (.24)**	**.86 (.35)**	**.92 (.27)**	**.89 (.31)**	**.87 (.34)**	**.86 (.35)**	**.87 (.34)**	**.88 (.33)**
diabetes	8	85	**.96 (.20)**	.87 (.34)	**.91 (.29)**	.82 (.39)	**.31 (.46)**	**.42 (.50)**	.47 (.50)	.57 (.50)
b-cancer	9	29	.98 (.14)	**.91 (.29)**	**.94 (.24)**	**.92 (.27)**	.87 (.34)	**.75 (.44)**	**.80 (.40)**	**.79 (.41)**
f-solar	9	100	**.93 (.26)**	**.91 (.29)**	**.95 (.22)**	**.93 (.26)**	**.51 (.50)**	.81 (.39)	**.55 (.50)**	**.66 (.48)**
heart	13	38	1.00 (.00)	.85 (.36)	**.91 (.29)**	**.93 (.26)**	.53 (.50)	**.28 (.45)**	**.40 (.49)**	.62 (.49)
german	20	100	.99 (.10)	**.91 (.29)**	**.92 (.27)**	**.89 (.31)**	.56 (.50)	.55 (.50)	**.44 (.50)**	.68 (.47)
ringnorm	20	100	**.97 (.17)**	**.93 (.26)**	**.91 (.29)**	.85 (.36)	**.00 (.00)**	**.00 (.00)**	**.00 (.00)**	**.02 (.14)**
waveform	21	66	.98 (.14)	**.92 (.27)**	**.93 (.26)**	**.88 (.33)**	**.00 (.00)**	**.00 (.00)**	**.02 (.14)**	**.00 (.00)**

When an asymmetric divergence such as the KL divergence [7] or the PE divergence [11] is adopted for two-sample test, the test results depend on the choice of *directions*: a divergence from P to P' or from P' to P. [4] proposed to choose the direction that gives a smaller p-value—it was experimentally shown that, when the uLSIF-based PE divergence estimator is used for the two-sample test (which is called the *least-squares two-sample test*; LSTT), the heuristic of choosing the direction with a smaller p-value contributes to reducing the *type-II error* (the probability of accepting incorrect null-hypotheses, i.e., two distributions are judged to be the same when they are actually different), while the increase of the *type-I error* (the probability of rejecting correct null-hypotheses, i.e., two distributions are judged to be different when they are actually the same) is kept moderate.

We apply the proposed method to the binary classification datasets taken from the *IDA benchmark repository* [21]. We test LSTT with the RuLSIF-based PE divergence estimator for $\alpha = 0$, 0.5, and 0.95; we also test the *maximum mean discrepancy* (MMD) [22], which is a kernel-based two-sample test method. The performance of MMD depends on the choice of the Gaussian kernel width. Here, we adopt a version proposed by [23], which automatically optimizes the Gaussian kernel width. The p-values of MMD are computed in the same way as LSTT based on the permutation test procedure.

First, we investigate the rate of accepting the null hypothesis when the null hypothesis is correct (i.e., the two distributions are the same). We split all the positive training samples into two sets and perform two-sample test for the two sets of samples. The experimental results are summarized in the left half of Table 1, showing that LSTT with $\alpha = 0.5$ compares favorably with those with $\alpha = 0$ and 0.95 and MMD in terms of the type-I error.

Next, we consider the situation where the null hypothesis is not correct (i.e., the two distributions are different). The numerator samples are generated in the same way as above, but a half of denominator samples are replaced with negative training samples. Thus, while the numerator sample set contains only positive training samples, the denominator sample set includes both positive and negative training samples. The experimental results are summarized in the right half of Table 1, showing that LSTT with $\alpha = 0.5$ again compares favorably with those with $\alpha = 0$ and 0.95. Furthermore, LSTT with $\alpha = 0.5$ tends to outperform MMD in terms of the type-II error.

Overall, LSTT with $\alpha = 0.5$ is shown to be a useful method for two-sample homogeneity test. See the supplementary material for more experimental evaluation.

Inlier-Based Outlier Detection: Next, we apply the proposed method to outlier detection.

Let us consider an outlier detection problem of finding irregular samples in a dataset (called an "evaluation dataset") based on another dataset (called a "model dataset") that only contains regular samples. Defining the density-ratio over the two sets of samples, we can see that the density-ratio

Table 2: Experimental results of outlier detection. Mean AUC score (and standard deviation in the bracket) over 100 trials is reported. The best method having the highest mean AUC score and comparable methods according to the *two-sample t-test* at the significance level 5% are specified by bold face. The datasets are sorted in the ascending order of the input dimensionality d.

Datasets	d	OSVM ($\nu = 0.05$)	OSVM ($\nu = 0.1$)	RuLSIF ($\alpha = 0$)	RuLSIF ($\alpha = 0.5$)	RuLSIF ($\alpha = 0.95$)
IDA:banana	2	**.668 (.105)**	**.676 (.120)**	.597 (.097)	.619 (.101)	.623 (.115)
IDA:thyroid	5	.760 (.148)	**.782 (.165)**	**.804 (.148)**	**.796 (.178)**	.722 (.153)
IDA:titanic	5	**.757 (.205)**	**.752 (.191)**	**.750 (.182)**	.701 (.184)	.712 (.185)
IDA:diabetes	8	**.636 (.099)**	.610 (.090)	.594 (.105)	.575 (.105)	**.663 (.112)**
IDA:breast-cancer	9	**.741 (.160)**	.691 (.147)	**.707 (.148)**	**.737 (.159)**	**.733 (.160)**
IDA:flare-solar	9	.594 (.087)	.590 (.083)	**.626 (.102)**	**.612 (.100)**	.584 (.114)
IDA:heart	13	.714 (.140)	.694 (.148)	**.748 (.149)**	**.769 (.134)**	.726 (.127)
IDA:german	20	**.612 (.069)**	**.604 (.084)**	**.605 (.092)**	**.597 (.101)**	**.605 (.095)**
IDA:ringnorm	20	**.991 (.012)**	**.993 (.007)**	.944 (.091)	.971 (.062)	**.992 (.010)**
IDA:waveform	21	.812 (.107)	.843 (.123)	**.879 (.122)**	**.875 (.117)**	**.885 (.102)**
Speech	50	.788 (.068)	**.830 (.060)**	.804 (.101)	**.821 (.076)**	**.836 (.083)**
20News (‘rec’)	100	.598 (.063)	.593 (.061)	.628 (.105)	.614 (.093)	**.767 (.100)**
20News (‘sci’)	100	.592 (.069)	.589 (.071)	.620 (.094)	.609 (.087)	**.704 (.093)**
20News (‘talk’)	100	.661 (.084)	.658 (.084)	.672 (.117)	.670 (.102)	**.823 (.078)**
USPS (1 vs. 2)	256	.889 (.052)	**.926 (.037)**	.848 (.081)	.878 (.088)	.898 (.051)
USPS (2 vs. 3)	256	.823 (.053)	.835 (.050)	.803 (.093)	.818 (.085)	**.879 (.074)**
USPS (3 vs. 4)	256	.901 (.044)	.939 (.031)	.950 (.056)	.961 (.041)	**.984 (.016)**
USPS (4 vs. 5)	256	.871 (.041)	.890 (.036)	.857 (.099)	.874 (.082)	**.941 (.031)**
USPS (5 vs. 6)	256	.825 (.058)	.859 (.052)	.863 (.078)	.867 (.068)	**.901 (.049)**
USPS (6 vs. 7)	256	.910 (.034)	.950 (.025)	.972 (.038)	.984 (.018)	**.994 (.010)**
USPS (7 vs. 8)	256	.938 (.030)	.967 (.021)	.941 (.053)	.951 (.039)	**.980 (.015)**
USPS (8 vs. 9)	256	.721 (.072)	.728 (.073)	.721 (.084)	.728 (.083)	**.761 (.096)**
USPS (9 vs. 0)	256	.920 (.037)	.966 (.023)	.982 (.048)	.989 (.022)	**.994 (.011)**

values for regular samples are close to one, while those for outliers tend to be significantly deviated from one. Thus, density-ratio values could be used as an index of the degree of outlyingness [1, 2].

Since the evaluation dataset usually has a wider support than the model dataset, we regard the evaluation dataset as samples corresponding to the denominator density $p'(\boldsymbol{x})$, and the model dataset as samples corresponding to the numerator density $p(\boldsymbol{x})$. Then, outliers tend to have smaller density-ratio values (i.e., close to zero). Thus, density-ratio approximators can be used for outlier detection.

We evaluate the proposed method using various datasets: IDA benchmark repository [21], an in-house French speech dataset, the 20 Newsgroup dataset, and the USPS hand-written digit dataset (the detailed specification of the datasets is explained in the supplementary material).

We compare the *area under the ROC curve* (AUC) [24] of RuLSIF with $\alpha = 0$, 0.5, and 0.95, and *one-class support vector machine (OSVM)* with the Gaussian kernel [25]. We used the *LIBSVM* implementation of OSVM [26]. The Gaussian width is set to the median distance between samples, which has been shown to be a useful heuristic [25]. Since there is no systematic method to determine the tuning parameter ν in OSVM, we report the results for $\nu = 0.05$ and 0.1.

The mean and standard deviation of the AUC scores over 100 runs with random sample choice are summarized in Table 2, showing that RuLSIF overall compares favorably with OSVM. Among the RuLSIF methods, small α tends to perform well for low-dimensional datasets, and large α tends to work well for high-dimensional datasets.

Transfer Learning: Finally, we apply the proposed method to transfer learning.

Let us consider a transductive transfer learning setup where labeled training samples $\{(\boldsymbol{x}_j^{\mathrm{tr}}, y_j^{\mathrm{tr}})\}_{j=1}^{n_{\mathrm{tr}}}$ drawn i.i.d. from $p(y|\boldsymbol{x})p_{\mathrm{tr}}(\boldsymbol{x})$ and unlabeled test samples $\{\boldsymbol{x}_i^{\mathrm{te}}\}_{i=1}^{n_{\mathrm{te}}}$ drawn i.i.d. from $p_{\mathrm{te}}(\boldsymbol{x})$ (which is generally different from $p_{\mathrm{tr}}(\boldsymbol{x})$) are available. The use of *exponentially-weighted importance weighting* was shown to be useful for adaptation from $p_{\mathrm{tr}}(\boldsymbol{x})$ to $p_{\mathrm{te}}(\boldsymbol{x})$ [5]:

$$\min_{f \in \mathcal{F}} \left[\frac{1}{n_{\mathrm{tr}}} \sum_{j=1}^{n_{\mathrm{tr}}} \left(\frac{p_{\mathrm{te}}(\boldsymbol{x}_j^{\mathrm{tr}})}{p_{\mathrm{tr}}(\boldsymbol{x}_j^{\mathrm{tr}})} \right)^{\tau} \mathrm{loss}(y_j^{\mathrm{tr}}, f(\boldsymbol{x}_j^{\mathrm{tr}})) \right],$$

where $f(\boldsymbol{x})$ is a learned function and $0 \leq \tau \leq 1$ is the exponential flattening parameter. $\tau = 0$ corresponds to plain empirical-error minimization which is statistically efficient, while $\tau = 1$ corresponds to importance-weighted empirical-error minimization which is statistically consistent; $0 < \tau < 1$ will give an intermediate estimator that balances the trade-off between statistical efficiency and consistency. τ can be determined by *importance-weighted cross-validation* [6] in a data dependent fashion.

Table 3: Experimental results of transfer learning in human activity recognition. Mean classification accuracy (and the standard deviation in the bracket) over 100 runs for human activity recognition of a new user is reported. We compare the plain *kernel logistic regression* (KLR) without importance weights, KLR with relative importance weights (RIW-KLR), KLR with exponentially-weighted importance weights (EIW-KLR), and KLR with plain importance weights (IW-KLR). The method having the highest mean classification accuracy and comparable methods according to the *two-sample t-test* at the significance level 5% are specified by bold face.

Task	KLR $(\alpha = 0, \tau = 0)$	RIW-KLR $(\alpha = 0.5)$	EIW-KLR $(\tau = 0.5)$	IW-KLR $(\alpha = 1, \tau = 1)$
Walks vs. run	0.803 (0.082)	**0.889 (0.035)**	**0.882 (0.039)**	**0.882 (0.035)**
Walks vs. bicycle	0.880 (0.025)	**0.892 (0.035)**	0.867 (0.054)	0.854 (0.070)
Walks vs. train	0.985 (0.017)	**0.992 (0.008)**	0.989 (0.011)	0.983 (0.021)

However, a potential drawback is that estimation of $r(\boldsymbol{x})$ (i.e., $\tau = 1$) is rather hard, as shown in this paper. Here we propose to use *relative importance weights* instead:

$$\min_{f \in \mathcal{F}} \left[\frac{1}{n_{\mathrm{tr}}} \sum_{j=1}^{n_{\mathrm{tr}}} \frac{p_{\mathrm{te}}(\boldsymbol{x}_j^{\mathrm{tr}})}{(1-\alpha)p_{\mathrm{te}}(\boldsymbol{x}_j^{\mathrm{tr}}) + \alpha p_{\mathrm{tr}}(\boldsymbol{x}_j^{\mathrm{tr}})} \mathrm{loss}(y_j^{\mathrm{tr}}, f(\boldsymbol{x}_j^{\mathrm{tr}})) \right].$$

We apply the above transfer learning technique to *human activity recognition* using accelerometer data. Subjects were asked to perform a specific task such as walking, running, and bicycle riding, which was collected by *iPodTouch*. The duration of each task was arbitrary and the sampling rate was 20Hz with small variations (the detailed experimental setup is explained in the supplementary material). Let us consider a situation where a new user wants to use the activity recognition system. However, since the new user is not willing to label his/her accelerometer data due to troublesomeness, no labeled sample is available for the new user. On the other hand, unlabeled samples for the new user and labeled data obtained from existing users are available. Let labeled training data $\{(\boldsymbol{x}_j^{\mathrm{tr}}, y_j^{\mathrm{tr}})\}_{j=1}^{n_{\mathrm{tr}}}$ be the set of labeled accelerometer data for 20 existing users. Each user has at most 100 labeled samples for each action. Let unlabeled test data $\{\boldsymbol{x}_i^{\mathrm{te}}\}_{i=1}^{n_{\mathrm{te}}}$ be unlabeled accelerometer data obtained from the new user.

The experiments are repeated 100 times with different sample choice for $n_{\mathrm{tr}} = 500$ and $n_{\mathrm{te}} = 200$. The classification accuracy for 800 test samples from the new user (which are different from the 200 unlabeled samples) are summarized in Table 3, showing that the proposed method using relative importance weights for $\alpha = 0.5$ works better than other methods.

5 Conclusion

In this paper, we proposed to use a relative divergence for robust distribution comparison. We gave a computationally efficient method for estimating the relative Pearson divergence based on direct relative density-ratio approximation. We theoretically elucidated the convergence rate of the proposed divergence estimator under non-parametric setup, which showed that the proposed approach of estimating the relative Pearson divergence is more preferable than the existing approach of estimating the plain Pearson divergence. Furthermore, we proved that the asymptotic variance of the proposed divergence estimator is independent of the model complexity under a correctly-specified parametric setup. Thus, the proposed divergence estimator hardly overfits even with complex models. Experimentally, we demonstrated the practical usefulness of the proposed divergence estimator in two-sample homogeneity test, inlier-based outlier detection, and transfer learning tasks.

In addition to two-sample homogeneity test, inlier-based outlier detection, and transfer learning, density-ratios can be useful for tackling various machine learning problems, for example, multi-task learning, independence test, feature selection, causal inference, independent component analysis, dimensionality reduction, unpaired data matching, clustering, conditional density estimation, and probabilistic classification. Thus, it would be promising to explore more applications of the proposed relative density-ratio approximator beyond two-sample homogeneity test, inlier-based outlier detection, and transfer learning.

Acknowledgments
MY was supported by the JST PRESTO program, TS was partially supported by MEXT KAKENHI 22700289 and Aihara Project, the FIRST program from JSPS, initiated by CSTP, TK was partially supported by Grant-in-Aid for Young Scientists (20700251), HH was supported by the FIRST program, and MS was partially supported by SCAT, AOARD, and the FIRST program.

References

[1] A. J. Smola, L. Song, and C. H. Teo. Relative novelty detection. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS2009)*, pages 536–543, 2009.

[2] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori. Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems*, 26(2):309–336, 2011.

[3] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, Cambridge, MA, 2007.

[4] M. Sugiyama, T. Suzuki, Y. Itoh, T. Kanamori, and M. Kimura. Least-squares two-sample test. *Neural Networks*, 24(7):735–751, 2011.

[5] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[6] 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.

[7] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

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

[9] 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*, 60:699–746, 2008.

[10] 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.

[11] 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*, 50:157–175, 1900.

[12] 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*, 28:131–142, 1966.

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

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

[15] T. Suzuki and M. Sugiyama. Sufficient dimension reduction via squared-loss mutual information estimation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS2010)*, pages 804–811, 2010.

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

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

[18] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[19] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

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

[21] G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for adaboost. *Machine Learning*, 42(3):287–320, 2001.

[22] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[23] B. Sriperumbudur, K. Fukumizu, A. Gretton, G. Lanckriet, and B. Schölkopf. Kernel choice and classifiability for RKHS embeddings of probability distributions. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1750–1758. 2009.

[24] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159, 1997.

[25] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[26] C.-C. Chang and C.h-J. Lin. *LIBSVM: A Library for Support Vector Machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.