Mach Learn (2013) 91:189–209
DOI 10.1007/s10994-013-5329-8

Semi-supervised learning with density-ratio estimation

Masanori Kawakita · Takafumi Kanamori

Received: 17 April 2012 / Accepted: 14 January 2013 / Published online: 9 February 2013

Abstract In this paper we study statistical properties of semi-supervised learning, which is considered to be an important problem in the field of machine learning. In standard supervised learning only labeled data is observed, and classification and regression problems are formalized as supervised learning. On the other hand, in semi-supervised learning, unlabeled data is also obtained in addition to labeled data. Hence, the ability to exploit unlabeled data is important to improve prediction accuracy in semi-supervised learning. This problem is regarded as a semiparametric estimation problem with missing data. Under discriminative probabilistic models, it was considered that unlabeled data is useless to improve the estimation accuracy. Recently, the weighted estimator using unlabeled data achieves a better prediction accuracy compared to the learning method using only labeled data, especially when the discriminative probabilistic model is misspecified. That is, improvement under the semiparametric model with missing data is possible when the semiparametric model is misspecified. In this paper, we apply the density-ratio estimator to obtain the weight function in semi-supervised learning. Our approach is advantageous because the proposed estimator does not require well-specified probabilistic models for the probability of the unlabeled data. Based on statistical asymptotic theory, we prove that the estimation accuracy of our method outperforms supervised learning using only labeled data. Some numerical experiments present the usefulness of our methods.

Keywords Semi-supervised learning · Density ratio · Semiparametric model · Missing data

Editor: Xiaojin Zhu.

M. Kawakita
Department of Informatics, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan
e-mail: kawakita@inf.kyushu-u.ac.jp

T. Kanamori (✉)
Department of Computer Science and Mathematical Informatics, Nagoya University, Furocho,
Chikusaku, Nagoya 464-8603, Japan
e-mail: kanamori@is.nagoya-u.ac.jp

1 Introduction

In this paper, we analyze statistical properties of semi-supervised learning. In standard supervised learning, only the labeled data (x, y) is observed, and the goal is to estimate the relation between x and y, i.e., the conditional expectation of y given x, or the conditional probability of y given x. In semi-supervised learning (Chapelle et al. 2006), the unlabeled data x' is also obtained in addition to the labeled data. In real-world data such as text data, we can often obtain both labeled and unlabeled data. A typical example is that x and y denote the text and tag of the article, respectively. Significant effort is required to tag the article; hence, the labeled data is scarce, while the unlabeled data is abundant. In semi-supervised learning, the study of methods that exploit unlabeled data is important.

In standard semi-supervised learning, statistical models of the joint probability $p(x, y)$, i.e., generative models, are often used to incorporate the information present in the unlabeled data into the estimation. For example, under the statistical model $p(x, y; \boldsymbol{\beta})$ having the parameter $\boldsymbol{\beta}$, the information involved in the unlabeled data is used to estimate the parameter $\boldsymbol{\beta}$ via the marginal probability $p(x; \boldsymbol{\beta}) = \int p(x, y; \boldsymbol{\beta})dy$. The amount of information in unlabeled samples was studied by Castelli and Cover (1996), Dillon et al. (2010), Sinha and Belkin (2007). This approach was developed to deal with various data structures. For example, semi-supervised learning with manifold assumption or cluster assumption has been studied along this line (Belkin and Niyogi 2004; Lafferty and Wasserman 2007). By making some assumptions regarding generative models, it was revealed that unlabeled data is useful for improving prediction accuracy.

Statistical models of the conditional probability $p(y|x)$, i.e., discriminative models, are also used in semi-supervised learning. It appears that the unlabeled data is not very useful to estimate the conditional probability, because the marginal probability does not have any information on $p(y|x)$ (Lasserre et al. 2006; Seeger 2001; Zhang and Oles 2000). The maximum likelihood estimator (MLE) obtained using a parametric model of $p(y|x)$ is not affected by the unlabeled data. However, Sokolovska et al. (2008) proved that even for discriminative models, the unlabeled data is still useful for improving the prediction accuracy of the learning method using only labeled data.

Semi-supervised learning methods work well under some assumptions regarding the population distribution and statistical models. However, semi-supervised learning has the possibility of degrading the estimation accuracy, especially when a misspecified model is applied; see Cozman et al. (2003), Grandvalet and Bengio (2005), Nigam et al. (1999). Hence, *safe* semi-supervised learning is desired. The learning algorithms proposed by Sokolovska et al. (2008) and Li and Zhou (2011) have a theoretical guarantee such that the unlabeled data does not degrade the estimation accuracy.

In this paper, we developed a learning method proposed by Sokolovska et al. (2008). To incorporate the information present in the unlabeled data into the estimator, Sokolovska et al. (2008) used the weighted estimator. In the estimation of the weight function, a well-specified model for the marginal probability $p(x)$ was assumed. This is a strong assumption for semi-supervised learning. To overcome the drawback, we applied the density-ratio estimator for the estimation of the weight function (Sugiyama and Kawanabe 2012; Sugiyama et al. 2012). We proved that semi-supervised learning with the density-ratio estimator improves standard supervised learning. Our method is available for not only classification problems but also regression problems, while many semi-supervised learning methods focus on binary classification problems.

This paper is organized as follows. In Sect. 2, we show the problem setup. In Sect. 3, we introduce the weighted estimator investigated by Sokolovska et al. In Sect. 4, we briefly explain the density-ratio estimation. In Sect. 5, we discuss the asymptotic variance of the considered semi-supervised learning methods. In Sect. 6, we prove that the weighted estimator using labeled and unlabeled data outperforms supervised learning using only labeled data. In Sect. 7, we discuss our numerical experiments, and we conclude the paper in Sect. 8.

2 Problem setup

Here, we introduce the problem setup. We assume that the probability distribution of training samples is given as

$$(x_i, y_i) \overset{\text{i.i.d.}}{\sim} p(y|x)p(x), \quad i = 1, \dots, n, \qquad x'_j \overset{\text{i.i.d.}}{\sim} q(x), \quad j = 1, \dots, n', \tag{1}$$

where $p(y|x)$ is the conditional probability of $y \in \mathcal{Y}$ given $x \in \mathcal{X}$, and $p(x)$ and $q(x)$ are marginal probabilities on $\mathcal{X}$. Here, $q(x)$ is regarded as the probability in the testing phase, i.e., the test data (x, y) is distributed from the joint probability $p(y|x)q(x)$, and the estimation accuracy is evaluated under the test probability. The paired sample (x_i, y_i) is called labeled data, and the unpaired sample x'_j is called unlabeled data. Our goal is to estimate the conditional probability $p(y|x)$ or the conditional expectation $E[y|x]$ based on the labeled and unlabeled data in (1). When $\mathcal{Y}$ is a finite set, the problem is called a classification problem. For $\mathcal{Y} = \mathbb{R}$ (the set of real numbers), the estimation of $E[y|x]$ is referred to as the regression problem.

In common learning problems including semi-supervised learning, marginal distributions of training and test samples are the same, i.e., $p(x) = q(x)$ is assumed. Below we introduce a covariate shift adaptation in which $p(x) = q(x)$ is not guaranteed. Learning methods used in the covariate shift adaptation are applied to our problem. Hence, in the problem setup in (1), the marginal distributions $p(x)$ and $q(x)$ are separately represented.

In the context of *covariate shift* adaptation (Shimodaira 2000), the assumption that $p(x) \neq q(x)$ is generally employed. As shown in Sect. 3, the weighted estimator with the weight function $q(x)/p(x)$ is used to correct the estimation bias induced by the covariate shift; see Sugiyama et al. (2007, 2012), Sugiyama and Kawanabe (2012), for details. Hence, the estimation of the weight function $q(x)/p(x)$ is important to achieve good prediction accuracy in the test phase.

On the other hand, in *semi-supervised learning* (Chapelle et al. 2006), the equality $p(x) = q(x)$ is assumed, and often n' is often significantly greater than n. This setup is also quite practical. For example, in text data mining, the labeled data is scarce, while the unlabeled data is abundant. In this paper, we assume that the equality

$$p(x) = q(x) \tag{2}$$

holds. Even under the assumption that $p(x) = q(x)$, it is important to consider problem (1) in which the marginal distributions $p(x)$ and $q(x)$ can be different. The reason will be clarified in Sect. 3.

We define the following semiparametric model,

$$\mathcal{M} = \left\{ p(y|x; \boldsymbol{\alpha}) r(x) : \boldsymbol{\alpha} \in A \subset \mathbb{R}^d, \ r \in \mathcal{P} \right\}, \tag{3}$$

for the estimation of the conditional probability $p(y|x)$, where $\mathcal{P}$ is the set of all probability densities of the covariate x. The parameter of interest is $\boldsymbol{\alpha}$, and $r(x) \in \mathcal{P}$ is regarded as the nuisance parameter. The model $\mathcal{M}$ does not necessarily include the true test probability

$p(y|x)q(x)$, i.e., the parameter $\boldsymbol{\alpha}$ may not exist such that $p(y|x) = p(y|x;\boldsymbol{\alpha})$ holds. This is a significant condition when we consider the improvement of the inference in semi-supervised learning.

Our target is to estimate the parameter $\boldsymbol{\alpha}^*$ that satisfies

$$\max_{\boldsymbol{\alpha}\in A} E\big[\log p(y|x;\boldsymbol{\alpha})\big] = E\big[\log p\big(y|x;\boldsymbol{\alpha}^*\big)\big], \tag{4}$$

where $E[\cdot]$ denotes the expectation with respect to the distribution of test samples, $p(y|x)q(x)$. If the model $\mathcal{M}$ includes the true probability, we have $p(y|x;\boldsymbol{\alpha}^*) = p(y|x)$ due to the non-negativity of the Kullback-Leibler divergence (Cover and Thomas 2006). However, in the misspecified setup, the equality $p(y|x;\boldsymbol{\alpha}^*) = p(y|x)$ is not guaranteed.

3 Weighted estimator in semi-supervised learning

In this section, we introduce weighted estimators. First, we consider the MLE for the estimation of $p(y|x)$ under the model $p(y|x;\boldsymbol{\alpha})$.

For the statistical model of the conditional probability $p(y|x;\boldsymbol{\alpha})$, let $\boldsymbol{u}(x,y;\boldsymbol{\alpha}) \in \mathbb{R}^d$ be the score function

$$\boldsymbol{u}(x,y;\boldsymbol{\alpha}) = \nabla \log p(y|x;\boldsymbol{\alpha}),$$

where ∇ denotes the gradient with respect to the model parameter $\boldsymbol{\alpha}$. It is well known that the score function satisfies the equality

$$\int \boldsymbol{u}(x,y;\boldsymbol{\alpha})p(y|x;\boldsymbol{\alpha})q(x)dxdy = \mathbf{0}$$

for any $\boldsymbol{\alpha} \in A$. Considering the extremal condition of (4), we see that the parameter $\boldsymbol{\alpha}^*$ in (4) is a solution of the following equation:

$$\int \boldsymbol{u}(x,y;\boldsymbol{\alpha})p(y|x)q(x)dxdy = \mathbf{0}, \quad \boldsymbol{\alpha} \in A. \tag{5}$$

When the target probability $p(y|x)$ is realized by the model, i.e., $p(y|x) = p(y|x;\boldsymbol{\alpha}^*)$, the solution of (5) is given as $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$. Otherwise, $p(y|x;\boldsymbol{\alpha}^*)$ is regarded as an approximation of $p(y|x)$ in the sense of the Kullback-Leibler divergence.

Our purpose is to estimate the parameter $\boldsymbol{\alpha}^*$ that satisfies Eq. (5) from training samples. By replacing the expectation with respect to $p(y|x)q(x)$ with the empirical distribution of labeled training samples, we obtain the following equation with respect to the parameter $\boldsymbol{\alpha}$,

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{u}(x_i,y_i;\boldsymbol{\alpha}) = \mathbf{0}, \quad \boldsymbol{\alpha} \in A. \tag{6}$$

Let $\widehat{\boldsymbol{\alpha}}$ be a solution of Eq. (6). Because $p(x) = q(x)$ is assumed, $p(y|x;\widehat{\boldsymbol{\alpha}})$ is expected to approximate $p(y|x,\boldsymbol{\alpha}^*)$. This is because the law of large numbers yields that for each $\boldsymbol{\alpha}$, the empirical mean (6) converges in probability to (5). The estimator $\widehat{\boldsymbol{\alpha}}$ is the MLE with the statistical model $p(y|x;\boldsymbol{\alpha})$. Under the regularity condition, the MLE has statistical consistency, i.e., the estimated parameter $\widehat{\boldsymbol{\alpha}}$ converges in probability to $\boldsymbol{\alpha}^*$ when the sample size n tends to infinity. See van der Vaart (1998) for details of statistical consistency.

In addition, the score function $\boldsymbol{u}$ is the optimal choice among Z-estimators (van der Vaart 1998, Chap. 5), when the true conditional probability density $p(y|x)$ is realized by the model $p(y|x;\boldsymbol{\alpha})$. Here the asymptotic variance-covariance matrix of the estimated parameter is employed to compare the estimation accuracy. Suppose that only labeled samples are available, then the optimality of the score function $\boldsymbol{u}$ implies that information regarding the marginal distribution $q(x)$ is not useful to improve the estimation accuracy. Intuitively, this is because the score function is orthogonal to the tangent space spanned by infinitesimal shifts of the marginal probability $q(x)$. See Amari and Kawanabe (1997) for details pertaining to the geometrical interpretation of the semiparametric inference.

We consider the setup of semi-supervised learning in which unlabeled data is available. When the model $\mathcal{M}$ is specified, i.e., $p(y|x) = p(y|x;\boldsymbol{\alpha}^*)$, the estimator (6) obtained using only the labeled data is efficient. This is obtained from numerous studies about the semiparametric inference with missing data; see Nan et al. (2009), Robins et al. (1994) and references therein.

In contrast, suppose that in semi-supervised learning, the model $\mathcal{M}$ is misspecified, i.e., $p(y|x)$ is not realized by the model $p(y|x;\boldsymbol{\alpha})$. In this case, Sokolovska et al. proved that it is possible to improve the MLE in (6) by using the so-called weighted MLE. This result implies that unlabeled data is useful in semi-supervised learning, when the model is misspecified. Before we give an explanation of the study by Sokolovska et al., we briefly introduce the weighted MLE.

The weighted MLE is defined as a solution of the following equation:

$$\frac{1}{n}\sum_{i=1}^{n} w(x_i)\boldsymbol{u}(x_i, y_i;\boldsymbol{\alpha}) = \boldsymbol{0}, \quad \boldsymbol{\alpha} \in A, \tag{7}$$

where $w(x)$ is a weight function. Suppose that $w(x) = q(x)/p(x)$ under the problem setup (1). Then the law of large numbers leads to probabilistic convergence,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} w(x_i)\boldsymbol{u}(x_i, y_i;\boldsymbol{\alpha}) &\xrightarrow{p} \int \frac{q(x)}{p(x)}\boldsymbol{u}(x, y;\boldsymbol{\alpha})p(y|x)p(x)dx \\ &= \int \boldsymbol{u}(x, y;\boldsymbol{\alpha})p(y|x)q(x)dx.\end{aligned}$$

Hence the estimator $p(y|x;\widehat{\boldsymbol{\alpha}})$ based on (7) will provide a good estimation of $p(y|x)$ under the marginal probability $q(x)$. This indicates that $p(y|x;\widehat{\boldsymbol{\alpha}})$ is expected to approximate $p(y|x)$ over the region on which $q(x)$ takes a large value. The weight function $w(x)$ serves to adjust the bias of the estimator. Hence, the weighted MLE is useful especially when $p(x) \neq q(x)$ holds. If we cannot directly access probability densities $p(x)$ and $q(x)$, we need to estimate the weight function $w(x) = q(x)/p(x)$. However, in semi-supervised learning, the weight function is given as $w(x) = q(x)/p(x) = 1$, and it is known beforehand. While it may be thought that there is no need to estimate the weight function, Sokolovska et al. showed that the estimation of the weight function is useful, even though it is already known in semi-supervised learning.

Here, we briefly introduce the result obtained by Sokolovska et al. (2008). Let the set $\mathcal{X}$ be finite. Then, $\mathcal{P}$ is a finite-dimensional parametric model. Suppose that the sample size of the unlabeled data is enormous and that the probability function $q(x)$ on $\mathcal{X}$ is known with a high degree of accuracy. The probability $p(x)$ is estimated by the MLE $\widehat{p}(x)$ from training samples $\{x_i\}_{i=1}^{n}$ in the labeled data. Then, Sokolovska et al. showed that the weighted MLE

(7) obtained with the estimated weight function $w(x) = q(x)/\widehat{p}(x)$ improves the naive MLE when the true conditional probability $p(y|x)$ is not realized by the model $p(y|x; \boldsymbol{\alpha})$.

The phenomenon above is similar to the statistical paradox analyzed by Henmi and Eguchi (2004), Henmi et al. (2007). In a semiparametric estimation, Henmi and Eguchi (2004) pointed out that the estimation accuracy of the parameter of interest can be improved by the estimation of the nuisance parameter, even when the nuisance parameter is known beforehand. Hirano et al. (2003) also pointed out that the estimator with the estimated propensity score is more efficient than that using the true propensity score for estimation of average treatment effects. Here the propensity score corresponds to the weight function $w(x)$ in our context.

We illustrate the statistical paradox according to Henmi et al. (2007). Let $f(x)$ be a non-negative function on $\mathbb{R}^p$, and θ be the integral of $f(x)$, i.e., $\theta = \int f(x)dx$. Our goal is to compute θ based on the Monte Carlo method. First, generate $x_1, \ldots, x_n$ from a probability $p(x; \boldsymbol{\alpha}_0)$ with a preliminarily fixed $\boldsymbol{\alpha}_0$. Then, let $\widetilde{\theta}$ be an importance sampling estimator $\widetilde{\theta} = \frac{1}{n}\sum_{i=1}^n f(x_i)/p(x_i; \boldsymbol{\alpha}_0)$. The law of large numbers ensures the statistical consistency of $\widetilde{\theta}$. As an alternative estimate of θ, we define $\widehat{\theta}$ as $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n f(x_i)/p(x_i; \widehat{\boldsymbol{\alpha}})$, where $\widehat{\boldsymbol{\alpha}}$ is the MLE of $\boldsymbol{\alpha}_0$ under a specified model $p(x; \boldsymbol{\alpha})$. Henmi et al. (2007) proved that the asymptotic variance of $\widehat{\theta}$ is smaller than or equal to that of $\widetilde{\theta}$. This result appears to be paradoxical, since the estimation of the known parameter $\boldsymbol{\alpha}_0$ improves the ordinary importance sampling $\widetilde{\theta}$.

For the estimation of the weight function $w(x)$ in (7), we apply the density-ratio estimator instead of estimating the probability densities separately. Recently, estimation methods of the weight function obtained from training samples have been intensively studied under the name of *density-ratio estimation*; see Sugiyama et al. (2012) and Sugiyama and Kawanabe (2012) for details. We show that the density-ratio estimator provides a practical method for semi-supervised learning. In the next section, we introduce density-ratio estimation.

4 Density-ratio estimation

Density-ratio estimators are available to estimate the weight function $w(x) = q(x)/p(x)$. Recently, methods to directly estimate density-ratios have been developed in the machine learning community (Sugiyama and Kawanabe 2012; Sugiyama et al. 2012). We apply the density-ratio estimator to estimate the weight function $w(x)$ instead of using the estimator of each probability density.

We briefly introduce the density-ratio estimator according to Qin (1998). Suppose that the following training samples are observed:

$$x_i \overset{\text{i.i.d.}}{\sim} p(x), \quad i = 1, \ldots, n, \qquad x'_j \overset{\text{i.i.d.}}{\sim} q(x), \quad j = 1, \ldots, n'. \tag{8}$$

Our goal is to estimate the density-ratio $w(x) = q(x)/p(x)$. Let $w(x; \boldsymbol{\theta})$ be an r-dimensional parametric model for the density-ratio defined by

$$w(x; \boldsymbol{\theta}) = \exp\{\theta_1\phi_1(x) + \cdots + \theta_r\phi_r(x)\}, \tag{9}$$

where $\phi_1(x)$ is given as $\phi_1(x) = 1$. Though a more general parametric model, say $\bar{w}(x; \boldsymbol{\theta}) = \exp\{\bar{\phi}(x; \boldsymbol{\theta})\}$ with a nonlinear function $\bar{\phi}(x; \boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$, is available for the density-ratio estimation, we use (9) for simplicity. In the first-order asymptotic theory, we only need the quantity in the first-order Taylor expansion of $\bar{\phi}(x; \boldsymbol{\theta})$ in the model $\bar{w}(x; \boldsymbol{\theta})$ if the model $\bar{w}(x; \boldsymbol{\theta})$ is closed under multiplication of positive constants. Thus, assuming the model (9) does not lose generality. See the paper by Qin (1998) for a more general description.

For any function $\boldsymbol{\eta}(x;\boldsymbol{\theta}) \in \mathbb{R}^r$ that may depend on the parameter $\boldsymbol{\theta}$, one has the equality

$$\int \boldsymbol{\eta}(x;\boldsymbol{\theta}) w(x) p(x) dx - \int \boldsymbol{\eta}(x;\boldsymbol{\theta}) q(x) dx = \mathbf{0}, \tag{10}$$

since $w(x)p(x) = q(x)$ holds. Hence, the empirical approximation of the above equation is expected to provide an estimation equation of the density-ratio. The empirical approximation of the above equality under the parametric model of $w(x;\boldsymbol{\theta})$ is given as

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\eta}(x_i;\boldsymbol{\theta}) w(x_i;\boldsymbol{\theta}) - \frac{1}{n'}\sum_{j=1}^{n'} \boldsymbol{\eta}\big(x'_j;\boldsymbol{\theta}\big) = \mathbf{0}. \tag{11}$$

Let $\widehat{\boldsymbol{\theta}}$ be a solution of (11), then $w(x;\widehat{\boldsymbol{\theta}})$ is an estimator of $w(x)$. Since the law of large numbers yields the probabilistic convergence of (11) to (10), the existence of a solution of (11) is guaranteed for the large sample limit under a specified density-ratio model. Note that we do not need to separately estimate probability densities $p(x)$ and $q(x)$. The estimation Eq. (11) provides a direct estimator of the density-ratio based on moment matching with the function $\boldsymbol{\eta}(x;\boldsymbol{\theta})$.

Qin (1998) proved that the optimal choice of $\boldsymbol{\eta}(x;\boldsymbol{\theta})$ is given as

$$\boldsymbol{\eta}(x;\boldsymbol{\theta}) = \frac{1}{1 + w(x;\boldsymbol{\theta}) \cdot n'/n} \nabla \log w(x;\boldsymbol{\theta}) = \frac{1}{1 + w(x;\boldsymbol{\theta}) \cdot n'/n} \boldsymbol{\phi}(x), \tag{12}$$

where $\boldsymbol{\phi}(x) = (\phi_1(x), \dots, \phi_r(x))^T$. The optimal function $\boldsymbol{\eta}(x;\boldsymbol{\theta})$ above is exactly the same as the score function of the logistic regression model

$$\Pr\big(y = \text{“}p\text{”} \,|\, x;\boldsymbol{\theta}\big) = \frac{w(x;\boldsymbol{\theta}) \cdot n'/n}{1 + w(x;\boldsymbol{\theta}) \cdot n'/n},$$

implying that the observations (8) are regarded as labeled training samples with label “p” or “q”, although the assignment of the label is not random in the density-ratio estimation. Since the estimation equation with the optimal function (12) is represented as a minimization problem, the existence of a solution is guaranteed under a mild assumption. By using $\boldsymbol{\eta}(x;\boldsymbol{\theta})$ above, the asymptotic variance matrix of $\widehat{\boldsymbol{\theta}}$ is minimized from among the set of moment matching estimators, when $w(x)$ is realized by the model $w(x;\boldsymbol{\theta})$. Hence, (12) is regarded as the counterpart of the score function for parametric probability models.

5 Semi-supervised learning with density-ratio estimation

We study the asymptotics of the weighted MLE (7) using the estimated density-ratio. The estimation equation is given as

$$\begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} w(x_i;\boldsymbol{\theta}) \boldsymbol{u}(x_i, y_i;\boldsymbol{\alpha}) = \mathbf{0}, \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \boldsymbol{\eta}(x_i;\boldsymbol{\theta}) w(x_i;\boldsymbol{\theta}) - \dfrac{1}{n'}\displaystyle\sum_{j=1}^{n'} \boldsymbol{\eta}\big(x'_j;\boldsymbol{\theta}\big) = \mathbf{0}. \end{cases} \tag{13}$$

Here the statistical models (3) and (9) are employed. The first equation is used for the estimation of the parameter $\boldsymbol{\alpha}$ of the model $p(y|x;\boldsymbol{\alpha})$, and the second equation is used for the estimation of the density-ratio $w(x;\boldsymbol{\theta})$. The estimator defined by (13) is referred to as the density-ratio-estimation-based semi-supervised learning, or *DRESS* for short.

In Sokolovska et al. (2008), the marginal probability density $p(x)$ is estimated by using a well-specified parametric model. Clearly, preparing the well-specified parametric model is not practical when $\mathcal{X}$ is not a finite set. On the other hand, it is easy to prepare a specified model of the density-ratio $w(x)$, whenever $p(x)=q(x)$ holds in (1). The model (9) is an example. Indeed, $w(x;\mathbf{0})=1$ holds. Hence, the assumption that the true weight function is realized by the model $w(x;\boldsymbol{\theta})$ is not an obstacle in semi-supervised learning.

We show the asymptotic expansion of the estimation equation (13). Recall that $p(x)=q(x)$ is assumed. Let $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\theta}}$ be solutions of (13). In addition, define $\boldsymbol{\alpha}^*$ to be a solution of

$$\int \boldsymbol{u}(x,y;\boldsymbol{\alpha})p(y|x)q(x)dxdy=\mathbf{0}$$

and $\boldsymbol{\theta}^*$ be the parameter such that $w(x;\boldsymbol{\theta}^*)=1$, i.e., $\boldsymbol{\theta}^*=\mathbf{0}$. We prepare some notations: $\boldsymbol{u}=\boldsymbol{u}(x,y;\boldsymbol{\alpha}^*)$, $\boldsymbol{\eta}=\boldsymbol{\eta}(x;\boldsymbol{\theta}^*)$, $\boldsymbol{u}_i=\boldsymbol{u}(x_i,y_i;\boldsymbol{\alpha}^*)$, $\boldsymbol{\eta}_i=\boldsymbol{\eta}(x_i;\boldsymbol{\theta}^*)$, $\boldsymbol{\eta}'_j=\boldsymbol{\eta}(x'_j;\boldsymbol{\theta}^*)$, $\delta\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*$ and $\delta\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*$. The Jacobian matrix of the score function $\boldsymbol{u}$ with respect to the parameter $\boldsymbol{\alpha}$ is denoted as $\nabla\boldsymbol{u}$, i.e., the d by d matrix whose element is given as $(\nabla\boldsymbol{u}(x,y;\boldsymbol{\alpha}))_{ik}=\frac{\partial^2}{\partial\alpha_i\partial\alpha_k}\log p(y|x;\boldsymbol{\alpha})$. The variance matrix and the covariance matrix under the probability $p(y|x)p(x)$ are denoted as $V[\cdot]$ and $\mathrm{Cov}[\cdot,\cdot]$, respectively.

In the estimation of density ratios, functions $\boldsymbol{\eta}(x;\boldsymbol{\theta})$ and $A\boldsymbol{\eta}(x;\boldsymbol{\theta})$ with any invertible matrix A produce the same estimator. This is clear from the second expression of (13). Thus, without loss of generality we assume that $\boldsymbol{\eta}$ at $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ is represented as

$$\boldsymbol{\eta}\big(x;\boldsymbol{\theta}^*\big)=\boldsymbol{\phi}(x)+\widetilde{\boldsymbol{\phi}}(x),$$

where $\widetilde{\boldsymbol{\phi}}(x)$ is an arbitrary function orthogonal to $\boldsymbol{\phi}(x)$, i.e., $E[\boldsymbol{\phi}\widetilde{\boldsymbol{\phi}}^T]$ is equal to the zero matrix. If $\boldsymbol{\eta}(x;\boldsymbol{\theta}^*)$ does not have any component that is represented as a linear transformation of $\boldsymbol{\phi}(x)$, the estimator would be degenerated. Under the regularity condition, the estimated parameters $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\theta}}$ converge to $\boldsymbol{\alpha}^*$ and $\boldsymbol{\theta}^*$, respectively. The asymptotic expansion of (13) around $(\boldsymbol{\alpha},\boldsymbol{\theta})=(\boldsymbol{\alpha}^*,\boldsymbol{\theta}^*)$ leads to

$$E[\nabla\boldsymbol{u}]\delta\boldsymbol{\alpha}+E\big[\boldsymbol{u}\boldsymbol{\phi}^T\big]\delta\boldsymbol{\theta}=-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i+o_p\big(n^{-1/2}\big),$$

$$E\big[\boldsymbol{\phi}\boldsymbol{\phi}^T\big]\delta\boldsymbol{\theta}=\frac{1}{n'}\sum_{j=1}^{n'}\boldsymbol{\eta}'_j-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\eta}_i+o_p\big(n^{-1/2}\big).$$

Hence, we have

$$\begin{aligned}E[\nabla\boldsymbol{u}]\delta\boldsymbol{\alpha}={}&\frac{1}{n}\sum_{i=1}^{n}\big\{E\big[\boldsymbol{u}\boldsymbol{\phi}^T\big]E\big[\boldsymbol{\phi}\boldsymbol{\phi}^T\big]^{-1}\boldsymbol{\eta}_i-\boldsymbol{u}_i\big\}\\&-\frac{1}{n'}\sum_{j=1}^{n'}E\big[\boldsymbol{u}\boldsymbol{\phi}^T\big]E\big[\boldsymbol{\phi}\boldsymbol{\phi}^T\big]^{-1}\boldsymbol{\eta}'_j+o_p\big(n^{-1/2}\big).\end{aligned}$$

Therefore, we obtain the asymptotic variance of the estimator $\widehat{\boldsymbol{\alpha}}$ defined from (13) as follows:

$$
\begin{aligned}
&n \cdot E[\nabla \boldsymbol{u}]V[\delta\boldsymbol{\alpha}]E[\nabla \boldsymbol{u}]^T \\
&\quad = V[\boldsymbol{u}] + \left(1+\frac{n}{n'}\right)E\bigl[\boldsymbol{u}\boldsymbol{\phi}^T\bigr]E\bigl[\boldsymbol{\phi}\boldsymbol{\phi}^T\bigr]^{-1}V[\boldsymbol{\eta}]E\bigl[\boldsymbol{\phi}\boldsymbol{\phi}^T\bigr]^{-1}E\bigl[\boldsymbol{\phi}\boldsymbol{u}^T\bigr] \\
&\qquad - E\bigl[\boldsymbol{u}\boldsymbol{\phi}^T\bigr]E\bigl[\boldsymbol{\phi}\boldsymbol{\phi}^T\bigr]^{-1}\mathrm{Cov}[\boldsymbol{\eta},\boldsymbol{u}] - \mathrm{Cov}[\boldsymbol{u},\boldsymbol{\eta}]E\bigl[\boldsymbol{\phi}\boldsymbol{\phi}^T\bigr]^{-1}E\bigl[\boldsymbol{\phi}\boldsymbol{u}^T\bigr] + o(1).
\end{aligned}
$$

On the other hand, the variance of the naive MLE, $\widetilde{\boldsymbol{\alpha}}$, defined as a solution of (6) is given as

$$
n \cdot E[\nabla \boldsymbol{u}]V[\delta\widetilde{\boldsymbol{\alpha}}]E[\nabla \boldsymbol{u}]^T = V[\boldsymbol{u}] + o(1),
$$

where $\delta\widetilde{\boldsymbol{\alpha}} = \widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*$.

In the sequel, we study the difference between the variance of $\widehat{\boldsymbol{\alpha}}$ and that of $\widetilde{\boldsymbol{\alpha}}$.

6 Maximum improvement by semi-supervised learning

Given the model for the density-ratio $w(x;\boldsymbol{\theta})$, we compare the asymptotic variance-covariance matrices of the estimators $\widetilde{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\alpha}}$. First, let us define

$$
\bar{\boldsymbol{u}}(x) = \int \boldsymbol{u}\bigl(x,y;\boldsymbol{\alpha}^*\bigr)p(y|x)dy,
$$

i.e., $\bar{\boldsymbol{u}}(x)$ is the projection of the score function $\boldsymbol{u}(x,y;\boldsymbol{\alpha}^*)$ onto the subspace consisting of all functions depending only on x, where the inner product is defined by the expectation under the joint probability $p(y|x)p(x)$. Note that the equality $E[\bar{\boldsymbol{u}}] = \boldsymbol{0}$ holds, since $p(x) = q(x)$ holds. Let the matrix B be

$$
B = E\bigl[\bar{\boldsymbol{u}}\boldsymbol{\phi}^T\bigr]E\bigl[\boldsymbol{\phi}\boldsymbol{\phi}^T\bigr]^{-1}.
$$

Then, a simple calculation yields that the difference of the variance matrix between $\widetilde{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\alpha}}$ is equal to

$$
\begin{aligned}
\mathrm{Diff}[\boldsymbol{u}] &:= n \cdot E[\nabla \boldsymbol{u}]V[\delta\widetilde{\boldsymbol{\alpha}}]E[\nabla \boldsymbol{u}]^T - n \cdot E[\nabla \boldsymbol{u}]V[\delta\boldsymbol{\alpha}]E[\nabla \boldsymbol{u}]^T \\
&= \frac{n'}{n+n'}E\bigl[\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}^T\bigr] - \left(1+\frac{n}{n'}\right)V\left[B\boldsymbol{\eta} - \frac{n'}{n+n'}\bar{\boldsymbol{u}}\right] + o(1). \qquad (14)
\end{aligned}
$$

In the second equality, we suppose that n'/n converges to a positive constant. When $\mathrm{Diff}[\boldsymbol{u}]$ is positive definite, the estimator $\widehat{\boldsymbol{\alpha}}$ using the labeled and unlabeled data improves the estimator $\widetilde{\boldsymbol{\alpha}}$ using only the labeled data. It is straightforward to see that the improvement is not attained if $\bar{\boldsymbol{u}} = \boldsymbol{0}$ holds. In general, the score function $\boldsymbol{u}(x,y;\boldsymbol{\alpha}) = \nabla \log p(y|x;\boldsymbol{\alpha})$ satisfies $\bar{\boldsymbol{u}} = \boldsymbol{0}$ if the model is specified. However, when the model of the conditional probability $p(y|x)$ is misspecified, there is a possibility that the proposed estimator (13) outperforms the MLE $\widetilde{\boldsymbol{\alpha}}$.

We derive the optimal moment function $\boldsymbol{\eta}$ for the estimation of the parameter $\boldsymbol{\alpha}^*$. The optimal $\boldsymbol{\eta}$ can be different from (12). We prepare some notations. Let $\Pi_\phi\bar{\boldsymbol{u}}$ be the $\mathbb{R}^d$-valued function on $\mathcal{X}$, each element of which is the projection of each element of $\bar{\boldsymbol{u}}$ onto the subspace spanned by $\{\phi_1(x),\dots,\phi_r(x)\}$. Here, the inner product is defined by the expectation under the marginal probability $p(x)$. In addition, let $\Pi_\phi^\perp\bar{\boldsymbol{u}}$ be the projection of $\bar{\boldsymbol{u}}$ onto the orthogonal complement of the subspace, i.e., $\Pi_\phi^\perp\bar{\boldsymbol{u}} = \bar{\boldsymbol{u}} - \Pi_\phi\bar{\boldsymbol{u}}$.

Theorem 1 *We assume that the model of the density-ratio is defined as*

$$w(x;\boldsymbol{\theta}) = \exp\{\boldsymbol{\phi}(x)^T\boldsymbol{\theta}\}$$

with the basis functions $\boldsymbol{\phi}(x) = (\phi_1(x), \dots, \phi_r(x))$ *satisfying* $\phi_1(x) = 1$. *Suppose that* $E[\boldsymbol{\phi}\boldsymbol{\phi}^T] \in \mathbb{R}^{r\times r}$ *is invertible and that the rank of* $E[\bar{\boldsymbol{u}}\boldsymbol{\phi}^T]E[\boldsymbol{\phi}\boldsymbol{\phi}^T]^{-1}$ *is equal to the dimension of the parameter* $\boldsymbol{\alpha}$, *i.e., row full rank. We assume that the moment function* $\boldsymbol{\eta}(x;\boldsymbol{\theta})$ *at* $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ *is represented as*

$$\boldsymbol{\eta}(x;\boldsymbol{\theta}^*) = \boldsymbol{\phi}(x) + \widetilde{\boldsymbol{\phi}}(x) \tag{15}$$

where $\widetilde{\boldsymbol{\phi}}(x)$ *is a function orthogonal to* $\boldsymbol{\phi}(x)$, *i.e.,* $E[\boldsymbol{\phi}(x)\widetilde{\boldsymbol{\phi}}(x)^T] = O$ *is equal to the zero matrix. Then, an optimal* $\widetilde{\boldsymbol{\phi}}$ *is given as*

$$\widetilde{\boldsymbol{\phi}} = \frac{n'}{n+n'}B^T(BB^T)^{-1}\varPi_{\boldsymbol{\phi}}^{\perp}\bar{\boldsymbol{u}}. \tag{16}$$

For the optimal choice of $\boldsymbol{\eta}$, *the maximum improvement is given as*

$$\begin{aligned}\mathrm{Diff}[\boldsymbol{u}] &= \frac{n'}{n+n'}E[\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}^T] - \frac{n^2}{n'(n+n')}E[\varPi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}}(\varPi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}})^T] + o(1)\\ &= \frac{n'}{n+n'}E[\varPi_{\boldsymbol{\phi}}^{\perp}\bar{\boldsymbol{u}}(\varPi_{\boldsymbol{\phi}}^{\perp}\bar{\boldsymbol{u}})^T] + \frac{n'-n}{n'}E[\varPi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}}(\varPi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}})^T] + o(1).\end{aligned} \tag{17}$$

Proof Because $\phi_1(x) = 1$, one has $E[\widetilde{\boldsymbol{\phi}}] = \mathbf{0}$ and $E[\varPi_{\boldsymbol{\phi}}^{\perp}\bar{\boldsymbol{u}}] = E[1\cdot\varPi_{\boldsymbol{\phi}}^{\perp}\bar{\boldsymbol{u}}] = \mathbf{0}$. Hence, one has $E[\varPi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}}] = E[\bar{\boldsymbol{u}}] - E[\varPi_{\boldsymbol{\phi}}^{\perp}\bar{\boldsymbol{u}}] = \mathbf{0}$. Our goad is to find $\widetilde{\boldsymbol{\phi}}$ that minimizes $V[B\boldsymbol{\eta} - \frac{n'}{n+n'}\bar{\boldsymbol{u}}]$ in (14) in the sense of positive definiteness. The orthogonal decomposition leads to

$$V\left[B\boldsymbol{\eta} - \frac{n'}{n+n'}\bar{\boldsymbol{u}}\right] = V\left[B\boldsymbol{\phi} - \frac{n'}{n+n'}\varPi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}}\right] + V\left[B\widetilde{\boldsymbol{\phi}} - \frac{n'}{n+n'}\varPi_{\boldsymbol{\phi}}^{\perp}\bar{\boldsymbol{u}}\right],$$

because of the orthogonality between $B\boldsymbol{\phi} - \frac{n'}{n+n'}\varPi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}}$ and $B\widetilde{\boldsymbol{\phi}} - \frac{n'}{n+n'}\varPi_{\boldsymbol{\phi}}^{\perp}\bar{\boldsymbol{u}}$, and the equality $E[B\widetilde{\boldsymbol{\phi}} - \frac{n'}{n+n'}\varPi_{\boldsymbol{\phi}}^{\perp}\bar{\boldsymbol{u}}] = \mathbf{0}$. Hence, $\widetilde{\boldsymbol{\phi}}$ that satisfies

$$B\widetilde{\boldsymbol{\phi}} = \frac{n'}{n+n'}\varPi_{\boldsymbol{\phi}}^{\perp}\bar{\boldsymbol{u}}$$

is an optimal choice. Because the matrix B is row full rank, a solution of the above equation is given by

$$\widetilde{\boldsymbol{\phi}} = \frac{n'}{n+n'}B^T(BB^T)^{-1}\varPi_{\boldsymbol{\phi}}^{\perp}\bar{\boldsymbol{u}}.$$

We obtain the maximum improvement of $\mathrm{Diff}[\boldsymbol{u}]$ by using the equalities $V[\varPi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}}] = E[\varPi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}}(\varPi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}})^T]$ and $B\boldsymbol{\phi} = E[\bar{\boldsymbol{u}}\boldsymbol{\phi}^T]E[\boldsymbol{\phi}\boldsymbol{\phi}^T]^{-1}\boldsymbol{\phi} = \varPi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}}$. □

Suppose that the optimal moment function $\boldsymbol{\eta} = \boldsymbol{\phi} + \widetilde{\boldsymbol{\phi}}$ presented in Theorem 1 is used with the score function $\boldsymbol{u}(x,y;\boldsymbol{\alpha})$. Then, the improvement (17) is maximized when $E[\varPi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}}(\varPi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}})^T]$ is minimized. Hence, the model $w(x;\boldsymbol{\theta})$ with the lower dimensional parameter $\boldsymbol{\theta}$ is preferred as long as the assumption in Theorem 1 is satisfied. This is intuitively understandable. Indeed, the statistical perturbation of the density-ratio estimator is minimized when the smallest model is employed.

Remark 1 Suppose that the basis functions, $\phi_1(x), \dots, \phi_r(x)$, are closely orthogonal to $\bar{\boldsymbol{u}}$, i.e., $E[\bar{\boldsymbol{u}}\boldsymbol{\phi}^T]$ is close to the null matrix. Then, the improvement $\mathrm{Diff}[\boldsymbol{u}]$ is close to $\frac{n'}{n+n'}E[\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}^T]$. As a result, we have $\sup_{\boldsymbol{\phi}} \mathrm{Diff}[\boldsymbol{u}] = \frac{n'}{n+n'}E[\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}^T]$ in which the supremum is taken over the basis of the density-ratio model satisfying the assumption in Theorem 1. However, the basis functions satisfying the exact equality $E[\bar{\boldsymbol{u}}\boldsymbol{\phi}^T] = O$ are useless, where O is the zero matrix. Because the equality $E[\bar{\boldsymbol{u}}\boldsymbol{\phi}^T] = O$ leads to $B = O$, the equality (14) is thus reduced to

$$\mathrm{Diff}[\boldsymbol{u}] = \frac{n'}{n+n'}E\big[\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}^T\big] - \frac{n+n'}{n'}V\left[\frac{n'}{n+n'}\bar{\boldsymbol{u}}\right] + o(1) = o(1).$$

This result implies that there is a singularity at the basis function $\boldsymbol{\phi}$ such that $E[\bar{\boldsymbol{u}}\boldsymbol{\phi}^T] = O$.

Example 1 Let $\boldsymbol{u}(x, y; \boldsymbol{\alpha})$ be the score function of the model $y = \boldsymbol{\alpha}^T\boldsymbol{b}(x) + Z$, $Z \sim N(0, \sigma^2)$, where $\boldsymbol{b}(x) = (b_1(x), \dots, b_d(x))$ is the vector consisting of basis functions and σ^2 is a known parameter. Then, one has $\boldsymbol{u}(x, y; \boldsymbol{\alpha}) = (y - \boldsymbol{\alpha}^T\boldsymbol{b}(x))\boldsymbol{b}(x)$. Suppose that the true conditional probability leads to the regression function $y = f(x) + Z$, where $E[Z|x] = 0$ for all x. Then, one has $\bar{\boldsymbol{u}}(x; \boldsymbol{\alpha}) = (f(x) - \boldsymbol{\alpha}^T\boldsymbol{b}(x))\boldsymbol{b}(x)$ and $E[\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}^T] = E[(f(x) - \boldsymbol{\alpha}^T\boldsymbol{b}(x))^2\boldsymbol{b}(x)\boldsymbol{b}(x)^T]$. Hence, the upper bound of the improvement is governed by the degree of the model misspecification $(f(x) - \boldsymbol{\alpha}^T\boldsymbol{b}(x))^2$. According to Theorem 1, an optimal moment function $\boldsymbol{\eta}(x; \boldsymbol{\theta})$ is given as

$$\boldsymbol{\eta}(x; \boldsymbol{\theta}^*) = \boldsymbol{\phi}(x) + \frac{n'}{n+n'}B^T\big(BB^T\big)^{-1}\big(\big(f(x) - \boldsymbol{\alpha}^{*T}\boldsymbol{b}(x)\big)\boldsymbol{b}(x) - B\boldsymbol{\phi}(x)\big)$$

at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, where $B = E[(f - \boldsymbol{\alpha}^{*T}\boldsymbol{b})\boldsymbol{b}\boldsymbol{\phi}^T]E[\boldsymbol{\phi}\boldsymbol{\phi}^T]^{-1}$.

It is not practical to apply the optimal function $\boldsymbol{\eta}(x; \boldsymbol{\theta})$ defined by (16). The optimal moment function depends on $\bar{\boldsymbol{u}}$, and one needs information on the probability $p(y|x)$ to obtain the explicit form of $\bar{\boldsymbol{u}}$. The estimation of $\bar{\boldsymbol{u}}$ needs a non-parametric estimation since the model misspecification of $\mathcal{M}$ is significant in our setup. Thus, we consider a more practical estimator for the density ratio. Suppose that $\widetilde{\boldsymbol{\phi}} = \boldsymbol{0}$ holds for the moment function $\boldsymbol{\eta}(x; \boldsymbol{\theta}^*)$. For example, the optimal moment function (12) satisfies $\boldsymbol{\eta}(x; \boldsymbol{\theta}^*) = \frac{n}{n+n'}\boldsymbol{\phi}(x)$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, i.e., $\widetilde{\boldsymbol{\phi}} = \boldsymbol{0}$. For the density-ratio model $w(x; \boldsymbol{\theta}) = \exp\{\boldsymbol{\phi}(x)^T\boldsymbol{\theta}\}$ with $\phi_1(x) = 1$ and the moment function satisfying $\boldsymbol{\eta}(x; \boldsymbol{\theta}^*) = \boldsymbol{\phi}(x)$, a brief calculation yields that

$$\mathrm{Diff}[\boldsymbol{u}] = \frac{n' - n}{n'}E\big[\Pi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}}(\Pi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}})^T\big] + o(1). \tag{18}$$

Hence, an improvement is realized when $n < n'$ holds. We note that the larger model $w(x; \boldsymbol{\theta})$ attains a better improvement in (18). In fact, $\Pi_{\phi}\bar{\boldsymbol{u}}$ approaches $\bar{\boldsymbol{u}}$ when the density-ratio model $w(x; \boldsymbol{\theta}) = \exp\{\boldsymbol{\theta}^T\boldsymbol{\phi}(x)\}$ becomes large. Hence, the non-parametric estimation of the density-ratio may be a good choice for realizing a large improvement in the estimation of the parameter $\boldsymbol{\alpha}^*$. This is totally different from the case in which the optimal $\widetilde{\boldsymbol{\phi}}$ presented in Theorem 1 is used in the density-ratio estimation. The relation between $\mathrm{Diff}[\boldsymbol{u}]$ using the optimal $\widetilde{\boldsymbol{\phi}}$ and $\mathrm{Diff}[\boldsymbol{u}]$ with $\widetilde{\boldsymbol{\phi}} = \boldsymbol{0}$ is illustrated in Fig. 1. With the limit of the dimension of $\boldsymbol{\theta}$, both variance matrices monotonically converge to $\frac{n'-n}{n'}E[\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}^T]$.

In semi-supervised learning, the size of unlabeled data is often large. When the size of unlabeled samples tends to infinity, the asymptotic variance of each estimator is simplified

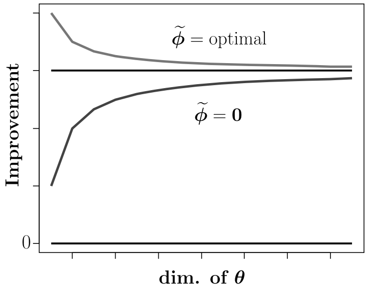

Fig. 1 The improvement Diff[$\boldsymbol{u}$] is depicted as the function of the dimension of the density-ratio model. Because the improvement is represented by a matrix, the *vertical axis* of the figure shows the partial ordering defined from positive semi-definiteness. When the dimension of $\boldsymbol{\theta}$ tends to infinity and $n' > n$ holds, the two curves converge to the common positive definite matrix $\frac{n'-n}{n'} E[\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}^T]$

as follows:

$$\begin{aligned}
\text{MLE:}\quad & \frac{1}{n} J^{-1} E\big[\boldsymbol{u}\boldsymbol{u}^T\big]\big(J^{-1}\big)^T + o(1/n),\\
\text{DRESS with optimal } \widetilde{\boldsymbol{\phi}}\text{:}\quad & \frac{1}{n} J^{-1}\big(E\big[\boldsymbol{u}\boldsymbol{u}^T\big] - E\big[\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}^T\big]\big)\big(J^{-1}\big)^T + o(1/n),\\
\text{DRESS with } \widetilde{\boldsymbol{\phi}} = \mathbf{0}\text{:}\quad & \frac{1}{n} J^{-1}\big(E\big[\boldsymbol{u}\boldsymbol{u}^T\big] - E\big[\Pi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}}(\Pi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}})^T\big]\big)\big(J^{-1}\big)^T + o(1/n),
\end{aligned}$$

where $J = E[\nabla \boldsymbol{u}]$. In the semi-supervised algorithm proposed by Sokolovska et al. (2008), the size of unlabeled data is assumed to be infinite. A simple calculation yields that its estimation accuracy is asymptotically the same as DRESS with optimal $\widetilde{\boldsymbol{\phi}}$ above. Hence, when a density-ratio model is applied instead of a specified model of the marginal distribution, the information loss of DRESS with $\widetilde{\boldsymbol{\phi}} = \mathbf{0}$ is quantified by $E[\Pi_{\boldsymbol{\phi}}^{\perp}\bar{\boldsymbol{u}}(\Pi_{\boldsymbol{\phi}}^{\perp}\bar{\boldsymbol{u}})^T]$, which is obtained as the difference between $E[\boldsymbol{u}\boldsymbol{u}^T] - E[\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}^T]$ and $E[\boldsymbol{u}\boldsymbol{u}^T] - E[\Pi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}}(\Pi_{\boldsymbol{\phi}}\bar{\boldsymbol{u}})^T]$. The information loss becomes smaller when higher dimensional density-ratio models are used. This tendency is the same as the estimation of the integration by the Monte Carlo method studied by Henmi et al. (2007).

7 Numerical experiments

Here we show numerical experiments in which supervised learning and semi-supervised learning algorithms are compared. Both regression and classification problems are presented.

7.1 Regression problems

We consider the following regression problem with d-dimensional covariate variables.

Labeled data:

$$\begin{aligned}
& y_i = \mathbf{1}^T \boldsymbol{x}_i + \varepsilon \frac{\|\boldsymbol{x}_i\|^2}{d} + z_i, \qquad z_i \sim N\big(0, \sigma^2\big), \quad i = 1, \ldots, n,\\
& \boldsymbol{x}_i \sim N_d(\mathbf{0}, I_d), \qquad \mathbf{1}^T = (1, \ldots, 1) \in \mathbb{R}^d.
\end{aligned} \tag{19}$$

Unlabeled data: $\boldsymbol{x}'_j \sim N_d(\mathbf{0}, I_d),\ j = 1, \dots, n'$.
Regression model: $y = \boldsymbol{\alpha}^T \boldsymbol{x} + z, \quad \boldsymbol{\alpha} \in \mathbb{R}^d, \quad z \sim N(0, s^2)$.
Score function for $\boldsymbol{\alpha}$ parameter: $\boldsymbol{u}(\boldsymbol{x}, y; \boldsymbol{\alpha}) = (y - \boldsymbol{\alpha}^T \boldsymbol{x})\boldsymbol{x}$.

The parameter ε in (19) controls the degree of model misspecification. Let f_ε be the target function, i.e., $f_\varepsilon(x) = \mathbf{1}^T \boldsymbol{x} + \varepsilon \|\boldsymbol{x}\|^2/d$, and define

$$e(\varepsilon) = \min_{\boldsymbol{\alpha}} E_{\boldsymbol{x}}\bigl[\bigl|f_\varepsilon(\boldsymbol{x}) - \boldsymbol{\alpha}^T \boldsymbol{x}\bigr|^2\bigr],$$

which implies the squared distance from the true function f_ε to the linear regression model. When the model is specified, the mean square error (MSE) of the MLE $\widetilde{\boldsymbol{\alpha}}$, i.e., $E_{\mathrm{Data}}[E_{\boldsymbol{x}}[|f_0(\boldsymbol{x}) - \widetilde{\boldsymbol{\alpha}}^T \boldsymbol{x}|^2]]$, is asymptotically equal to $\sigma^2 d/n$. Then, as a normalized measure of model misspecification, we use the ratio

$$\delta = \sqrt{e(\varepsilon)} \Big/ \sqrt{\frac{\sigma^2 d}{n}} = \sqrt{\frac{e(\varepsilon) n}{\sigma^2 d}}.$$

When $\delta \gg 1$ holds, misspecification of the model can be statistically detected.

In DRESS, we use a parametric model for density-ratio estimation. For any positive integer k, let $\boldsymbol{x}^{(k)}$ be the d-dimensional vector $(x_1^k, \dots, x_d^k)^T$. The density-ratio model is defined as

$$w(\boldsymbol{x}; \boldsymbol{\theta}) = \exp\bigl\{\theta_0 + \boldsymbol{\theta}_1^T \boldsymbol{x} + \boldsymbol{\theta}_2^T \boldsymbol{x}^{(2)} + \cdots + \boldsymbol{\theta}_L^T \boldsymbol{x}^{(L)}\bigr\}$$

having $Ld + 1$ dimensional parameter $(\theta_0, \boldsymbol{\theta}_1^T, \dots, \boldsymbol{\theta}_L^T)$. We apply the estimator (12) presented by Qin (1998). Because the estimator (12) satisfies $\widetilde{\boldsymbol{\phi}} = \mathbf{0}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, the improvement is asymptotically given by (18). Under the setup of $d = 2, n = 500, n' = 5000$ and $\sigma = 0.2$, we compute the MSEs for the MLE $\widetilde{\boldsymbol{\alpha}}$ and DRESS $\widehat{\boldsymbol{\alpha}}$. The difference in the test error

$$n \cdot \bigl(E\bigl[\bigl(\widetilde{\boldsymbol{\alpha}}^T \boldsymbol{x} - f_\varepsilon(\boldsymbol{x})\bigr)^2\bigr] - E\bigl[\bigl(\widehat{\boldsymbol{\alpha}}^T \boldsymbol{x} - f_\varepsilon(\boldsymbol{x})\bigr)^2\bigr]\bigr),$$

is evaluated for each ε and each dimension of the density ratio, $Ld + 1$, where the expectation is evaluated for the test samples. The MSE is calculated by taking the average over 500 iterations.

Results are shown in Fig. 2. When the model is specified, i.e., $\delta = 0\,(\varepsilon = 0)$, MLE presents better performance than DRESS. However, for a practical setup such as $\delta > 1$, we see that DRESS outperforms MLE. The dependency on the dimension of the density-ratio model is not clearly detected in this experiment. Overall, a larger density-ratio model presents a somewhat unstable result. In fact, in DRESS with large density-ratio model (the right bottom panel in Fig. 2), the MSE can be large, i.e., the improvement is negative, even when the model misspecification δ is greater than 2.

Next, we compare MLE and DRESS with a nonparametric density-ratio estimator. Here we use KuLSIF (Kanamori et al. 2012) as the density-ratio estimator. KuLSIF is a nonparametric estimator of the density-ratio based on the kernel method. The regularization is efficiently conducted to suppress the degree of freedom of the nonparametric model. In KuLSIF, the kernel function of the reproducing kernel Hilbert space corresponds to the basis function $\boldsymbol{\phi}(x)$.

In addition, estimators proposed by Sokolovska et al. (2008) were also examined. Here the learning method is referred to as marginal-probability-based semi-supervised learning (MSSL). In MSSL, the weight function $w(x) = q(x)/p(x)$ is estimated by $q(x)/\widehat{p}(x)$,

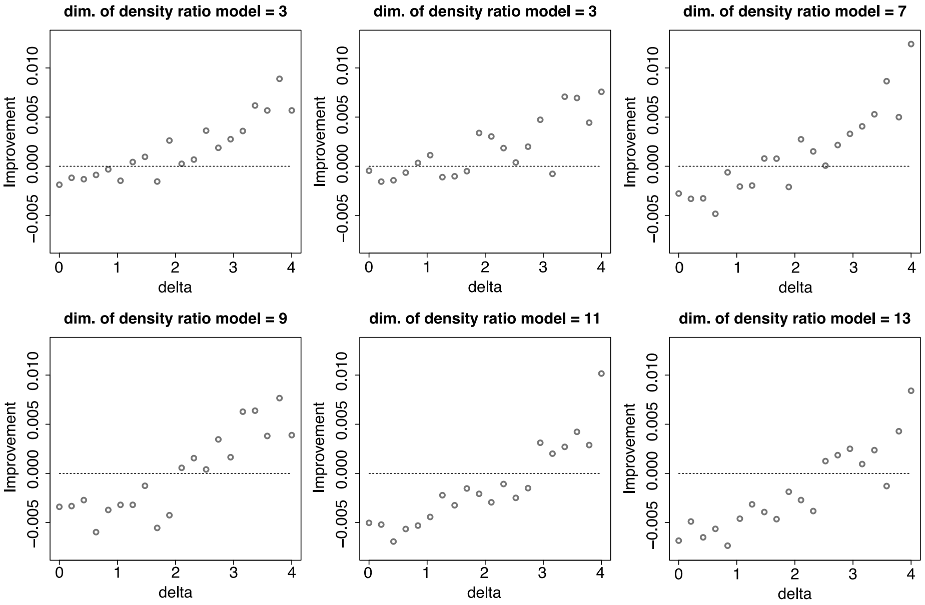

Fig. 2 Differences of MSE are plotted as a function of δ, where δ is the normalized measure of model misspecification. The *vertical axis*, “Improvement”, denotes the difference of the MSEs between MLE and DRESS. A positive improvement denotes that DRESS outperforms MLE

where $q(x)$ is the true probability density of unlabeled data, i.e., $N_d(\mathbf{0}, I_d)$, and $\widehat{p}(x)$ is the MLE of the marginal distribution $p(x)$ of labeled data. We apply two statistical models to estimate $p(x)$: one is $N_d(\boldsymbol{\mu}, I_d)$ with the parameter $\boldsymbol{\mu}$, and the other is $N_d(\boldsymbol{\mu}, \Sigma)$ with parameters $\boldsymbol{\mu}$ and Σ.

Under the setup of $d = 10, n = 50, n' = 20, 1000$ and $\sigma = 0.1, 0.5$, we compute the square root of the MSEs by using the average over 100 iterations. In Fig. 3, results for MLE, DRESS, and MSSL are plotted as a function of $\delta =$ (model error)/(statistical error). In the figures, MSSL-1 (resp. MSSL-2) denotes MSSL with the model $N_d(\boldsymbol{\mu}, I_d)$ (resp. $N_d(\boldsymbol{\mu}, \Sigma)$).

When δ is around 1, it is statistically difficult to detect the model misspecification from the training data of size $n = 50$. For the specified model, i.e., $\varepsilon = 0$, the MLE exhibits a better performance than DRESS and MSSL. However, under a practical setup such as when $\delta > 1$, we see that DRESS with KuLSIF and MSSL-1 outperform MLE. MSSL-2 is observed to be always worse than the others. As shown in the asymptotic analysis, the sample size of the unlabeled data affects the estimation accuracy of DRESS. The numerical results show that DRESS with a large n' attains a smaller error compared with DRESS with small n', especially when $\delta > 1$. The same tendency is observed in MSSL-1. As shown in the last paragraph in Sect. 6, the estimation accuracies of DRESS and MSSL are asymptotically the same when the size of the unlabeled data n' is sufficiently large and when the dimension of the density-ratio model is high. In experiments with $n' = 1000$, the MSEs of DRESS and MSSL-1 are almost the same. This result is consistent with the theoretical analysis. Note that MSSL-1 uses information about the true marginal distribution, although DRESS does not. Hence, DRESS with a nonparametric density-ratio estimator is much more practical than MSSL-1. We also apply MSSL with the weight estimator $\widehat{q}(x)/\widehat{p}(x)$ in which the denom-

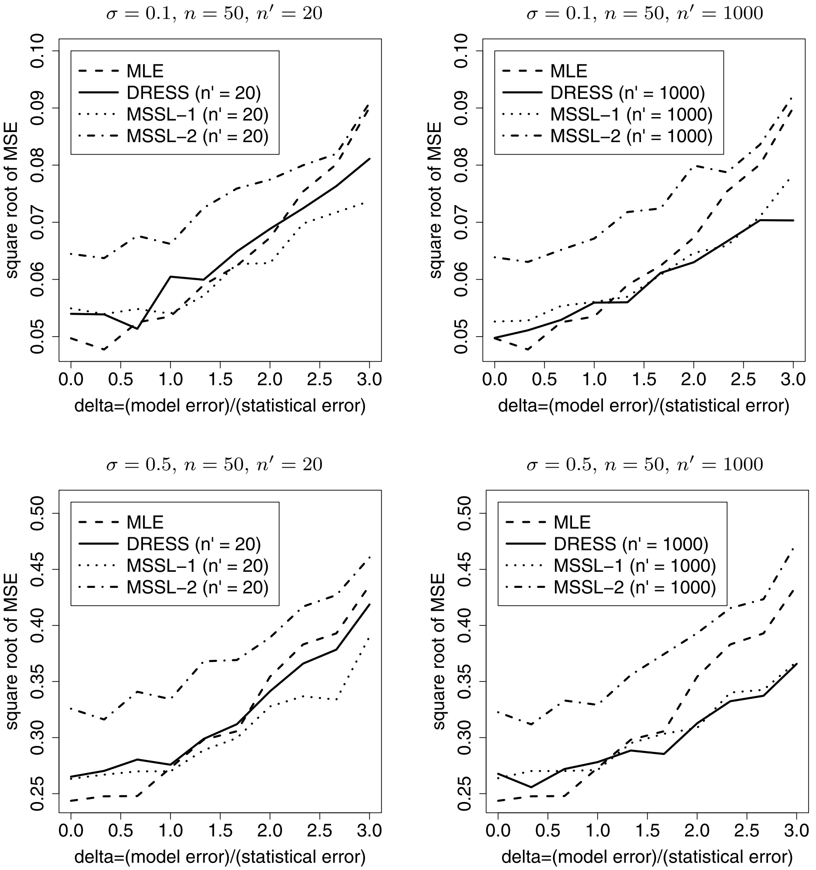

Fig. 3 Square roots of MSEs of MLE, DRESS, MSSL with the model $N_d(\boldsymbol{\mu}, I_d)$ (MSSL-1) and MSSL with the model $N_d(\boldsymbol{\mu}, \Sigma)$ (MSSL-2) are depicted as a function of δ, where δ is a normalized measure of the model misspecification. Dimension of covariates is set to $d = 10$. Size of labeled data is $n = 50$, and size of unlabeled data is $n' = 20$ or $n' = 1000$. The parameter σ is standard deviation of noise involved in dependent variable y, and *upper* (resp. *lower*) *panels* show results for $\sigma = 0.1$ (resp. $\sigma = 0.5$)

inator and the numerator are both estimated from training data using the model $N_d(\boldsymbol{\mu}, \Sigma)$. Its performance was almost always worse than that of MSSL-2.

In the numerical experiment, even DRESS with $n = 50$ and $n' = 20$ slightly outperforms MLE. This is not supported by the asymptotic analysis, and we need an in-depth theoretical study to understand the statistical features of semi-supervised learning.

7.2 Classification problems

As the first classification task, we use the `spam` data set in "kernlab" of R package (Karatzoglou et al. 2004). The data set includes 4601 samples. The dimension of the covariate is 57, i.e., $\boldsymbol{x} = (x_1, \dots, x_{57})^T$, and the elements represent statistical features of each document. The output y is assigned to "spam" or "nonspam".

For the binary classification problem, we use the logistic model,

$$P\big(y = \text{‘spam’} \mid \boldsymbol{x}; \boldsymbol{\alpha}\big) = \frac{1}{1 + \exp\{-\alpha_0 - \sum_{d=1}^{D} \alpha_d x_d\}},$$

where D is the dimension of the covariate used in the logistic model. In numerical experiments, D varies from 10 to 57, and the first D variables $(x_1, \dots, x_D)$ are incorporated into the logistic model. Hence, the dimension of the model parameter $\boldsymbol{\alpha}$ varies from 11 to 58. We test MSSL, DRESS with KuLSIF (Kanamori et al. 2012), and MLE. In MSSL, the weight function is estimated using Gaussian mixture models for both the denominator and numerator. The size of the labeled training samples is set to $n = 200, 500, 800$, and the size of the unlabeled training samples is set to $n' = 200, 2000$. The remaining samples are served as the test data.

Table 1 shows the prediction errors (%) with the standard deviation. In all cases, the performance of MSSL is worse than that of the others. We also show the p-value of the one-tailed paired t-test for prediction errors of DRESS and MLE. Small p-values denote the superiority of DRESS. We note that the p-value is small when the dimension D is not high. In other words, the numerical results agree with the asymptotic theory in Sect. 6. For relatively high-dimensional models, the prediction error of MLE is smaller than that of DRESS; see the row for $D = 57$ in Table 1. The size of unlabeled data, n', also affects the results. In fact, the p-values become small for large n'. This result is supported by the asymptotic analysis presented in Sect. 6.

In DRESS, we apply the density-ratio estimator to estimate the parameters of logistic models. For several setups, histograms of estimated weights on labeled samples, $w(x_i; \widehat{\boldsymbol{\theta}})$, $i = 1, \dots, n$, are depicted in Fig. 4. Because the true density-ratio is equal to the constant ratio, estimated weights are expected to concentrate around one. When the dimension D of the covariate is low (e.g., $D = 10$), the estimation accuracy of the density-ratio is high. However, for high dimensional data, values of $w(x_i; \widehat{\boldsymbol{\theta}})$ are widely spread. This is a reasonable result because in high-dimensional space, the data structure becomes sparse, and the density-ratio estimation becomes difficult. Because KuLSIF is a kernel-based nonparametric estimator, we need to elaborate the choice of the regularization parameter and the kernel width to obtain stable density-ratio estimates, especially for high dimensional data.

As the second classification problem, we apply semi-supervised learning algorithms to UCI data sets. See Rätsch et al. (2001) and Rätsch et al. (2000) for details of data sets. The original UCI data sets consist of training samples and test samples. In numerical experiments, training and test samples are all merged, and labeled and unlabeled samples for training are randomly chosen from merged samples. Labeled test samples are also randomly chosen from the rest. Let n be the size of labeled training samples, and the size of unlabeled training samples is set to $n' = \lfloor 0.5n \rfloor$ or $n' = 4n$. The size of the labeled test samples is set to the same size of the labeled training samples. Table 2 shows properties of each data set, where "rep." in the table denotes the number of replications for learning to evaluate the average performance.

Table 3 shows test errors of each learning algorithm. Generally, the estimation accuracy of MSSL is much lesser than that of the others. This is because the estimation of the weight function based on marginal probabilities is unstable. To compare DRESS and MLE, we show p-values of the one-tailed paired t-test for prediction errors of DRESS and MLE. Small p-values denote the superiority of DRESS. Figure 5 presents the plot of the ratio $(\dim \boldsymbol{x} + 1)/n$ versus the p-values for each dataset. The quantity $(\dim \boldsymbol{x} + 1)/n$ is the ratio of the number of parameters in logistic models and the sample size. When $(\dim \boldsymbol{x} + 1)/n$

Table 1 Prediction errors (%) of MSSL, DRESS with KuLSIF and MLE are shown. The p-values of the one-tailed paired t-test for DRESS and MLE are also presented. For smaller p-values, the prediction accuracy of DRESS is higher

D	$n = 200,\ n' = 200$				$n = 200,\ n' = 2000$			
	MSSL	DRESS	MLE	p-value	MSSL	DRESS	MLE	p-value
10	34.4 ± 13.9	21.2 ± 1.0	21.6 ± 1.5	0.000	33.9 ± 14.7	21.4 ± 0.9	21.9 ± 1.2	0.000
30	25.5 ± 8.3	14.9 ± 1.8	14.8 ± 1.7	0.976	26.3 ± 8.1	14.7 ± 1.8	14.7 ± 1.7	0.853
57	24.7 ± 9.7	15.0 ± 2.8	15.0 ± 2.8	0.590	22.7 ± 7.9	15.2 ± 3.0	15.0 ± 3.1	0.980
D	$n = 500,\ n' = 200$				$n = 500,\ n' = 2000$			
	MSSL	DRESS	MLE	p-value	MSSL	DRESS	MLE	p-value
10	25.94 ± 7.68	20.74 ± 0.71	20.97 ± 0.92	0.013	29.50 ± 12.34	20.52 ± 0.85	20.93 ± 0.91	0.000
30	26.43 ± 8.25	11.90 ± 0.65	12.02 ± 0.75	0.011	21.27 ± 6.31	11.65 ± 0.87	11.81 ± 0.84	0.001
57	22.18 ± 6.52	10.75 ± 0.99	10.61 ± 0.95	0.996	21.25 ± 6.77	10.53 ± 1.11	10.45 ± 1.01	0.921
D	$n = 800,\ n' = 200$				$n = 800,\ n' = 2000$			
	MSSL	DRESS	MLE	p-value	MSSL	DRESS	MLE	p-value
10	38.6 ± 15.3	20.5 ± 0.7	20.7 ± 0.7	0.009	42.0 ± 16.5	20.5 ± 1.0	20.9 ± 1.1	0.000
30	28.4 ± 8.7	11.3 ± 0.6	11.5 ± 0.6	0.000	28.2 ± 10.7	11.2 ± 0.7	11.5 ± 0.9	0.000
57	26.9 ± 10.7	8.9 ± 0.6	8.9 ± 0.6	0.489	27.3 ± 9.9	8.8 ± 0.8	8.9 ± 0.8	0.000

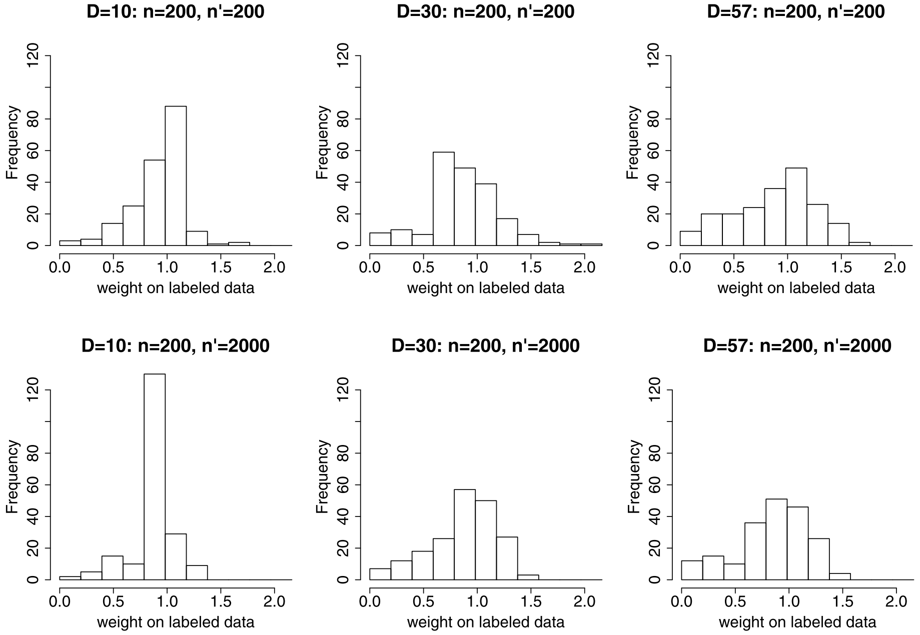

Fig. 4 Estimated weights on labeled data, $w(x_i;\widehat{\boldsymbol{\theta}}), i = 1, \ldots, n$ are plotted as histogram. The size of labeled data is fixed to $n = 500$, and the size of the unlabeled data is set to $n' = 100$ (*top panels*) and $n' = 1000$ (*bottom panels*). The dimension D of the covariate varies from 10 (*left column*) to 57 (*right column*)

Table 2 Properties of each data set are shown, where "rep." denotes the number of replication for learning

data set	dim $\boldsymbol{x}$	n	rep.
banana	2	883	100
breast-cancer	9	46	100
diabetis	8	128	100
flare-solar	9	177	100
german	20	166	100
heart	13	45	100
image	18	385	20
ringnorm	20	1233	100
splice	60	529	20
thyroid	5	35	80
titanic	3	366	100
twonorm	20	1233	100
waveform	21	833	100

is small, the asymptotic analysis in Sect. 6 is considered to be valid, i.e., the sample size is sufficient for the estimation of $\dim \boldsymbol{x} + 1$ parameters in logistic models. On the other hand, large $(\dim \boldsymbol{x} + 1)/n$ implies that higher-order terms in an asymptotic expansion may become significant. In both panels in Fig. 5, DRESS can be significantly better than MLE only when the ratio of the horizontal axis is small. For the case of $n' = 4n$, there is no dataset for which MLE is significantly better than DRESS, though MLE outperforms DRESS for 8 out of 13

Table 3 The p-values of the one-tailed paired t-test for prediction errors of DRESS and MLE. Small p-values denote the superiority of DRESS

data	$n' = 0.5n$		MLE	p-value
	MSSL	DRESS		
banana	47.57 ± 4.90	44.81 ± 4.60	45.95 ± 4.71	0.017
breast-cancer	38.17 ± 10.37	31.09 ± 7.21	30.61 ± 6.86	0.909
diabetis	29.53 ± 5.98	24.50 ± 3.66	24.24 ± 3.57	0.915
flare-solar	38.90 ± 5.65	35.07 ± 4.02	34.54 ± 3.86	0.999
german	37.05 ± 8.16	27.51 ± 3.24	27.08 ± 3.24	0.991
heart	26.31 ± 7.88	24.24 ± 6.80	24.33 ± 6.64	0.391
image	31.95 ± 8.28	17.16 ± 2.25	17.26 ± 2.42	0.397
ringnorm	24.21 ± 1.35	23.03 ± 1.16	24.04 ± 1.21	0.000
splice	19.89 ± 3.56	18.11 ± 1.77	18.16 ± 1.67	0.410
thyroid	22.96 ± 12.01	16.18 ± 7.96	14.68 ± 6.29	0.996
titanic	26.49 ± 5.38	22.29 ± 2.15	22.34 ± 2.15	0.170
twonorm	2.59 ± 0.43	2.57 ± 0.44	2.55 ± 0.44	0.840
waveform	13.69 ± 1.46	12.65 ± 1.02	12.70 ± 1.03	0.094
data	$n' = 4n$		MLE	p-value
	MSSL	DRESS		
banana	47.46 ± 4.35	44.30 ± 3.86	45.95 ± 4.71	0.000
breast-cancer	36.80 ± 7.76	30.98 ± 7.31	30.61 ± 6.86	0.856
diabetis	29.00 ± 5.93	24.48 ± 3.61	24.24 ± 3.57	0.909
flare-solar	40.39 ± 6.40	34.69 ± 3.86	34.54 ± 3.86	0.913
german	33.75 ± 5.92	27.23 ± 3.18	27.08 ± 3.24	0.860
heart	26.09 ± 7.83	24.07 ± 6.73	24.33 ± 6.64	0.127
image	34.87 ± 6.98	17.04 ± 2.24	17.26 ± 2.42	0.199
ringnorm	23.77 ± 1.19	22.98 ± 1.14	24.04 ± 1.21	0.000
splice	32.06 ± 14.05	18.34 ± 1.82	18.16 ± 1.67	0.810
thyroid	15.64 ± 6.95	14.79 ± 6.36	14.68 ± 6.29	0.619
titanic	31.06 ± 13.95	22.36 ± 2.17	22.34 ± 2.15	0.708
twonorm	2.57 ± 0.44	2.56 ± 0.44	2.55 ± 0.44	0.643
waveform	16.00 ± 2.48	12.62 ± 1.00	12.70 ± 1.03	0.010

datasets under the average performance. For some datasets with large $(\dim \boldsymbol{x} + 1)/n$, the discussion based on an asymptotic expansion may not be adequate.

8 Conclusion

In this paper, we investigated semi-supervised learning algorithms using density-ratio estimators. We proved that unlabeled data is useful when the statistical model of the conditional probability $p(y|x)$ is misspecified. This result agrees with the result given by Sokolovska et al. (2008), in which the weight function is estimated using the estimator of the marginal probability under a specified model. The estimator proposed in this paper is useful in prac-

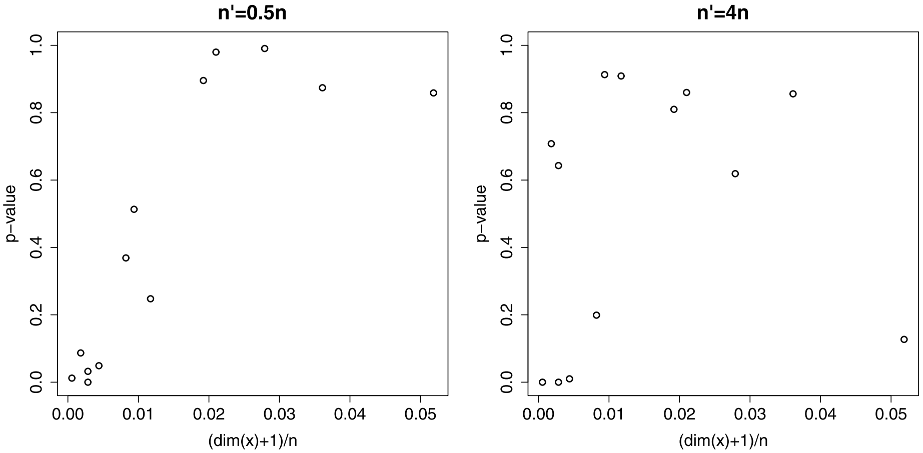

Fig. 5 Plots of p-values versus ratios $(\dim \boldsymbol{x} + 1)/n$ are depicted. *Left* (resp. *Right*) *panel* shows the results for $n' = 0.5n$ (resp. $n' = 4n$)

tice, since our method does not require any well-specified model for the marginal probability. Numerical experiments present the effectiveness of our method.

The theory and numerical experiments both show that when the size of unlabeled samples is too small, the performance of the proposed method can be worse than that of supervised learning with the MLE. In the asymptotic theory, semi-supervised learning outperforms supervised learning when the size of unlabeled data is greater than that of labeled data. However, numerical results in Sect. 7.2 do not necessarily meet the theoretical analysis, though the larger size of unlabeled data improves the prediction performance. From a practical viewpoint, it would be useful to have a "data-driven" way of deciding whether semi-supervised learning should be used or not. A possible method is to develop a resampling method such as cross validation or bootstrap to compare the prediction performance of supervised and semi-supervised learning. As a theoretical approach, a higher-order asymptotic analysis may provide useful insights.

We are also currently investigating semi-supervised learning from the perspective of the semiparametric inference with missing data. The development of a positive application of the statistical paradox in the semiparametric inference is an interesting area for future study in the area of semi-supervised learning.

Acknowledgements The authors are grateful to Dr. Masayuki Henmi, Dr. Hironori Fujisawa and Prof. Shinto Eguchi of the Institute of Statistical Mathematics, as well as the anonymous reviewers for their helpful comments. M.K. was partially supported by JSPS KAKENHI Grant Number 21700308. T.K. was partially supported by JSPS KAKENHI Grant Number 24500340.

References

Amari, S., & Kawanabe, M. (1997). Information geometry of estimating functions in semi-parametric statistical models. *Bernoulli*, *3*, 29–54.

Belkin, M., & Niyogi, P. (2004). Semi-supervised learning on Riemannian manifolds. *Machine Learning*, *56*, 209–239.

Castelli, V., & Cover, T. M. (1996). The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory*, *42*, 2102–2117.

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

Cover, T. M., & Thomas, J. A. (2006). *Wiley series in telecommunications and signal processing*. *Elements of information theory*. New York: Wiley-Interscience.

Cozman, F., Cohen, I., & Cirelo, M. (2003). Semi-supervised learning of mixture models. In *Proceedings of the international conference on machine learning*.

Dillon, J. V., Balasubramanian, K., & Lebanon, G. (2010). Asymptotic analysis of generative semi-supervised learning. In *27th international conference on machine learning* (pp. 295–302).

Grandvalet, Y., & Bengio, Y. (2005). Semi-supervised learning by entropy minimization. In *Neural information processing systems (NIPS 2004)* (Vol. 17, pp. 529–536). Cambridge: MIT Press.

Henmi, M., & Eguchi, S. (2004). A paradox concerning nuisance parameters and projected estimating functions. *Biometrika*, *91*, 929–941.

Henmi, M., Yoshida, R., & Eguchi, S. (2007). Importance sampling via the estimated sampler. *Biometrika*, *94*, 985–991.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*, 1161–1189.

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

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). Kernlab—an S4 package for kernel methods in R. *Journal of Statistical Software*, *11*, 1–20.

Lafferty, J. D., & Wasserman, L. A. (2007). Statistical analysis of semi-supervised regression. In *NIPS*. Rostrevar: Curran Associates, Inc.

Lasserre, J. A., Bishop, C. M., & Minka, T. P. (2006). Principled hybrids of generative and discriminative models. In *CVPR (1)* (pp. 87–94).

Li, Y.-F., & Zhou, Z.-H. (2011). Towards making unlabeled data never hurt. In *ICML* (pp. 1081–1088).

Nan, B., Kalbfleisch, J. D., & Yu, M. (2009). Asymptotic theory for the semiparametric accelerated failure time model with missing data. *The Annals of Statistics*, *37*, 2351–2376.

Nigam, K., Mccallum, A. K., Thrun, S., & Mitchell, T. (1999). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 103–134.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, *85*, 619–639.

Rätsch, G., Schölkopf, B., Smola, A., Mika, S., Onoda, T., & Müller, K.-R. (2000). *Robust ensemble learning* (pp. 207–220). Cambridge: MIT Press.

Rätsch, G., Onoda, T., & Müller, K.-R. (2001). Soft margins for adaboost. *Machine Learning*, *42*, 287–320.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*, 846–866.

Seeger, M. (2001). *Learning with labeled and unlabeled data* (Technical Report). Institute for Adaptive and Neural Computation, University of Edinburgh.

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

Sinha, K., & Belkin, M. (2007). The value of labeled and unlabeled examples when the model is imperfect. In *NIPS*.

Sokolovska, N., Cappé, O., & Yvon, F. (2008). The asymptotics of semi-supervised learning in discriminative probabilistic models. In *Proceedings of the twenty-fifth international conference on machine learning* (pp. 984–991).

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

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

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

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge: Cambridge University Press.

Zhang, T., & Oles, F. J. (2000). A probability analysis on the value of unlabeled data for classification problems. In *17th international conference on machine learning*.