Journal of Machine Learning Research 20 (2019) 1-33 Submitted 7/18; Revised 2/19; Published 4/19

Quantification Under Prior Probability Shift: the Ratio Estimator and its Extensions

Afonso Fernandes Vaz AFONSOFVAZ@GMAIL.COM
Rafael Izbicki RAFAELIZBICKI@GMAIL.COM
Rafael Bassi Stern RBSTERN@GMAIL.COM
Department of Statistics
Federal University of São Carlos
São Carlos, SP 13565-905, Brazil

Editor: Charles Elkan

Abstract

The quantification problem consists of determining the prevalence of a given label in a target population. However, one often has access to the labels in a sample from the training population but not in the target population. A common assumption in this situation is that of prior probability shift, that is, once the labels are known, the distribution of the features is the same in the training and target populations. In this paper, we derive a new lower bound for the risk of the quantification problem under the prior shift assumption. Complementing this lower bound, we present a new approximately minimax class of estimators, ratio estimators, which generalize several previous proposals in the literature. Using a weaker version of the prior shift assumption, which can be tested, we show that ratio estimators can be used to build confidence intervals for the quantification problem. We also extend the ratio estimator so that it can: (i) incorporate labels from the target population, when they are available and (ii) estimate how the prevalence of positive labels varies according to a function of certain covariates.

Keywords: quantification, prior probability shift, data set shift, domain shift, semi-supervised learning

1. Introduction

In several applications of binary classifiers, predicting the labels of individual observations *per se* is less important than evaluating the proportion of each label on an unlabeled target data set. The latter task is called quantification (Forman, 2008). For example, a company may be interested in evaluating the proportion of users who like each of their products, without access to labeled reviews of these products.

A common approach to such a problem is to (i) train a classifier for the user's evaluation based on labeled reviews of other products, and (ii) apply this classifier to the unlabeled target set and use the proportion of users who are classified as liking the product as an estimator. However, it is known that this two-step approach, known as "classify and count", fails because of domain shift (Forman, 2006; Tasche, 2016). In order to deal with this problem, several improvements have been proposed under an assumption named prior shift (Saerens et al., 2002; Forman, 2008; Bella et al., 2010; Barranquero et al., 2015). A particular estimator that successfully performs quantification is the adjusted count (AC) estimator

(Gart and Buck, 1966; Saerens et al., 2002; Forman, 2008). Part of the success of the AC estimator is explained in Tasche (2017) by showing that it is Fisher consistent. However, there are more properties one might desire of an estimator.

In order to investigate these properties, Vaz et al. (2017) introduces the ratio estimator, which is a generalization of the AC estimator. Vaz et al. (2017) derives the asymptotic mean squared error of the ratio estimator. Here, we show that the ratio estimator is approximately minimax and consistent under the prior probability shift assumption. In order to derive this result, we prove a new lower bound for the risk of the quantification problem under the prior shift assumption. This lower bound is general and applies to every method under the prior probability shift assumption. We also derive a central limit theorem for the ratio estimator which helps to explain its good performance and leads to a method for building confidence intervals for the quantification problem. This result also allows us to propose a new type of ratio estimator based on Reproducing Kernel Hilbert spaces. Since the AC estimator and the method in Bella et al. (2010) are special cases of the ratio estimator, they benefit from all of the results above.

It is important to evaluate whether the prior probability shift assumption indeed holds, otherwise the AC method can perform poorly (Tasche, 2017). We show that the ratio estimator works under an assumption that is less stringent than the prior shift assumption. Moreover, we show how this assumption can be tested. We are not aware of other methods to test the prior shift and related assumptions.

We also generalize the ratio estimator to two extensions of the quantification problem. In the first scenario, some labels are available in the target population. The combined estimator extends the ratio estimator in order to incorporate these labels and obtain a larger effective sample size. The second scenario considers that the prevalence of each label varies according to additional covariates. This generalization allows one to use unlabeled data to identify e.g. how the approval of a product varies with age. In this scenario, we introduce the regression ratio estimator, which offers improvements over the standard methods that are used in sentiment analyses (Wang et al., 2012).

Section 2 discusses the standard quantification problem under the prior probability shift assumption. Subsection 2.1 provides new lower bounds for the risk in this scenario. Subsection 2.2 introduces the ratio estimator, uses the result from the previous subsection to show that it is approximately minimax and also derives its convergence rate and a central limit theorem. Subsection 2.3 uses the asymptotic behavior of the ratio estimator to propose a new type of ratio estimator based on Reproducing Kernel Hilbert spaces. Finally, the ratio estimator requires a weaker version of prior probability shift to obtain consistency. Subsection 2.4 discusses a new algorithm for testing this assumption.

Section 3 proposes extensions of the ratio estimator to scenarios which are more general than the standard quantification problem. Subsection 3.1 proposes the combined estimator, for cases in which some labels are available in the population of interest. Subsection 3.2 proposes the ratio regression estimator, for the situation in which the prevalence of a given label varies according to a covariate. All proofs are presented in the appendix; code and data used for the experiments is available at https://github.com/afonsofvaz/ratio_estimator.

2. Quantification under prior probability shift

In order to formally approach the quantification problem, we use the same notation as in Wasserman (2006). If $\mathbf{Z}_1 \in \mathbb{R}^{d_1}$ and $\mathbf{Z}_2 \in \mathbb{R}^{d_2}$ are random vectors and $R \subset \mathbb{R}^{d_1}$, then $\mathbb{P}(\mathbf{Z}_1 \in R|\mathbf{Z}_2)$ is the conditional probability that $\mathbf{Z}_1$ is in R given $\mathbf{Z}_2$. Using $\mathbb{P}$, one can obtain $F_{\mathbf{Z}_1|\mathbf{Z}_2}$, $f_{\mathbf{Z}_1|\mathbf{Z}_2}$, $\mathbb{E}[\mathbf{Z}_1|\mathbf{Z}_2]$, and $\mathbb{V}[\mathbf{Z}_1|\mathbf{Z}_2]$ which are, respectively, the conditional distribution, density, expected value and variance of $\mathbf{Z}_1$ given $\mathbf{Z}_2$. Marginal properties of $\mathbf{Z}_1$ are indicated by omitting the conditioning random variable. Also, if $(\mathbf{Z}_n)_{n\in\mathbb{N}}$ is a sequence of random vectors, then $\mathbf{Z}_n \overset{a.s.}{\to} \mathbf{Z}$, $\mathbf{Z}_n \overset{\mathbb{P}}{\to} \mathbf{Z}$, and $\mathbf{Z}_n \rightsquigarrow \mathbf{Z}$ indicate respectively, that $\mathbf{Z}_n$ converges almost surely, in probability, and in distribution to $\mathbf{Z}$. In order to express the rate at which convergence occurs, it is useful to use $\mathcal{O}$ and Ω notation. If $(a_n)_{n\in\mathbb{N}}$ is a sequence in $\mathbb{R}$, then $a_n = \mathcal{O}(g(n))$ if there exists c such that, for every n, $a_n \leq c \cdot g(n)$ and $a_n = \Omega(g(n))$ if there exists c such that, for every n, $a_n \geq c \cdot g(n)$. Finally, $\mathbb{I}$ is the indicator function. An expression such as $\mathbb{I}(g(\mathbf{X}) \in A)$ is equal to 1 when $g(\mathbf{X}) \in A$ and to 0 when $g(\mathbf{X}) \notin A$.

In the quantification problem, for each sample instance $i \in \{1, \ldots, n\}$, $(\mathbf{X}_i, Y_i, S_i)$ is a vector of random variables such that $\mathbf{X}_i \in \mathbb{R}^d$ are features, $Y_i \in \{0, 1\}$ is a label of interest and $S_i \in \{0, 1\}$ is the indicator that this instance has been labeled. That is, whenever $S_i = 0$, then Y_i is not observed. Note that S_i can be random.

In the above framework, some subsets of the instances are frequently used. The sets $A_k := \{i \in \{1, \ldots, n\} : S_i = k\}$ represent the labeled ($k = 1$) and unlabeled ($k = 0$) instances. Similarly, $A_{k,j} := \{i \in \{1, \ldots, n\} : S_i = k \text{ and } Y_i = j\}$ represent the instances that are labeled ($k = 1$) or unlabeled ($k = 0$) and have a positive ($j = 1$) or a zero ($j = 0$) label. Also the number of instances that are unlabeled, labeled or that have label j are denoted, respectively, by $n_U := |A_0|$, $n_L := |A_1|$ and $n_j := |A_{1,j}|$.

In a quantification problem, one wishes to estimate $\theta := \mathbb{P}(Y = 1|S = 0)$, that is, the prevalence of positive labels among unlabeled samples. This prevalence is not assumed to be the same as the one over labeled sets, $\mathbb{P}(Y = 1|S = 1)$. The estimator for θ can depend only on the available data, that is, the features of all instances and the labels that were obtained. Formally, letting Z_i^j denote $(Z_i, \ldots, Z_j)$, a valid estimator is a function of $\mathbf{X}_1^n$, S_1^n and $(Y_i)_{i\in A_1}$. The set of all such valid estimators is denoted by $\mathcal{S}$.

In the standard formulation of the prior probability shift problem, $\{(\mathbf{X}_i, Y_i)\}_{i\in A_0}$ is called the *target population* (since the labels are unavailable), and $\{(\mathbf{X}_i, Y_i)\}_{i\in A_1}$ is called the training population (Tasche, 2017). It is common for both populations to be i.i.d.,

Assumption 1

- *$(S_1, \mathbf{X}_1, Y_1), \ldots, (S_n, \mathbf{X}_n, Y_n)$ are independent.*
- *For every $s \in \{0, 1\}$, $(\mathbf{X}_1, Y_1)|S_1 = s, \ldots, (\mathbf{X}_n, Y_n)|S_n = s$ are identically distributed.*

Unless additional assumptions are made, it is not possible to learn about θ using solely the observed data. One assumption that allows learning about θ is the prior probability shift, which states that "the class-conditional feature distributions of the training and test sets are the same" (Fawcett and Flach, 2005). Prior shift is formalized in Assumption 2.

Assumption 2 *[**Prior probability shift**] For every $(y_1, \ldots, y_n) \in \{0, 1\}^n$, $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is stochastically independent of $(S_1, \ldots, S_n)$ conditionally on $(Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)$.*

Although Assumption 2 is written in a different way than in papers such as Moreno-Torres et al. (2012), the content is similar. While Moreno-Torres et al. (2012) uses a subscript on the probability function to determine which is the reference population, we perform this task using the random variable, S. For instance, the probability that an instance from the target population has the label "1" is referred in previous notation and in this paper, respectively, as $\mathbb{P}_{tg}(Y_i = 1)$ and $\mathbb{P}(Y_i = 1|S_i = 0)$. Using this translation, Assumption 2 is the same as the prior probability shift in Moreno-Torres et al. (2012). Assumption 2 holds if and only if $f_{\mathbf{X}|Y,S=0} \equiv f_{\mathbf{X}|Y,S=1}$, that is, $\mathbb{P}_{tg}(\mathbf{x}|y) \equiv \mathbb{P}_{tr}(\mathbf{x}|y)$.

2.1. Lower bound on the risk for quantification under prior probability shift

Under Assumptions 1 and 2 it is possible to learn about θ from the features and labels that are available in the quantification problem. For example, one can use the features and labels in the training population to learn about $f_{\mathbf{X}|Y=0}$ and $f_{\mathbf{X}|Y=1}$. Also, if these densities are sufficiently different, then one can combine the information about them to the features in the target population to learn about the unknown labels in this population and, therefore, about θ. Definition 1 formally presents two classes in which the possible values of $f_{\mathbf{X}|Y=0}$ and $f_{\mathbf{X}|Y=1}$ are separable.

Definition 1 *Let $f_i(\mathbf{x}) = f_{\mathbf{X}|Y=i}(\mathbf{x})$, $\epsilon, K > 0$ and $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be a non-constant function.*

$$\begin{cases} \mathcal{F}_{\mathcal{L}1,\epsilon} := \{(f_0, f_1) : \|f_0 - f_1\|_1 \geq \epsilon\} \\ \mathcal{F}_{g,K,\epsilon} := \left\{(f_0, f_1) : \mathbb{E}_{f_i}[g(\mathbf{X})^2|Y = i] \leq K, \ \textit{and} \ |\mathbb{E}_{f_1}[g(\mathbf{X})|Y = 1] - \mathbb{E}_{f_0}[g(\mathbf{X})|Y = 0]| \geq \epsilon\right\} \end{cases}$$

Under the classes in Definition 1 it is possible to learn about θ and the learning rate depends on both the number of labeled and unlabeled instances. A lower bound for how these sample sizes affect the rate at which one learns about θ is presented in Theorem 3.

Definition 2 *Let $\mathcal{F}$ be a collection of (f_0, f_1). The minimax rate, $M(\mathcal{F})$, for estimating θ under the squared loss, $\mathcal{F}$, and Assumptions 1 and 2 is*

$$M(\mathcal{F}) = \inf_{\widehat{\theta} \in \mathcal{S}} \sup_{(f_0,f_1) \in \mathcal{F}; \theta \in [0,1]} \mathbb{E}_{f_0,f_1,\theta}\left[(\widehat{\theta} - \theta)^2 \middle| S_1^n\right]$$

Theorem 3 *$M(\mathcal{F}_{\mathcal{L}1,\epsilon}) \geq \Omega(\max(n_L^{-1}, n_U^{-1}))$ and $M(\mathcal{F}_{g,K,\epsilon}) \geq \Omega(\max(n_L^{-1}, n_U^{-1}))$.*

Theorem 3 shows that it is not possible to obtain an estimator for θ which has convergence rate faster than $\Omega(\max(n_L^{-1}, n_U^{-1}))$. In particular, it is not possible to learn θ by observing solely a limited amount of labels. The following subsection introduces the ratio estimator for θ, which achieves the lower bound in Theorem 3 under $\mathcal{F}_{g,K,\epsilon}$.

2.2. The ratio estimator and its theoretical properties

Definition 4 (Ratio estimator) *Let $g : \mathbb{R}^d \longrightarrow \mathbb{R}$. The untrimmed ratio estimator for θ based on g, $\widehat{\theta}_{UR}$, is*

$$\widehat{\theta}_{UR} := \frac{\frac{\sum_{i \in A_0} g(\mathbf{X}_i)}{n_U} - \frac{\sum_{i \in A_{1,0}} g(\mathbf{X}_i)}{n_0}}{\frac{\sum_{i \in A_{1,1}} g(\mathbf{X}_i)}{n_1} - \frac{\sum_{i \in A_{1,0}} g(\mathbf{X}_i)}{n_0}}$$

Since $\theta \in [0, 1]$, the ratio estimator, $\widehat{\theta}_R$, is

$$\widehat{\theta}_R = \max(0, \min(1, \widehat{\theta}_R))$$

The ratio estimator generalizes estimators which were previously proposed in the literature. This fact follows from observing that the terms in the untrimmed ratio estimator are sample averages of $g(\mathbf{X})$ among three groups of instances: unlabeled instances, instances labeled as 0, and instances labeled as 1. For instance, the adjusted count (AC) estimator (Gart and Buck, 1966; Saerens et al., 2002; Forman, 2008; Tasche, 2017) is the a ratio estimator when $g(\mathbf{x}) \in \{0, 1\}$, that is, $g(\mathbf{x})$ is the output of a classifier for Y. Also, the estimator in Bella et al. (2010) is a ratio estimator when $g(\mathbf{x}) = \widehat{\mathbb{P}}(Y = 1|\mathbf{x})$, that is, $g(\mathbf{x})$ is a soft classifier for Y.

Remark 5 *The ratio estimator can be generalized to the case in which $Y_i \in \{0, 1, \ldots, k\}$. In this case, let $g : \mathbb{R}^d \to \mathbb{R}^k$ be a fixed function. By defining G as a $k \times (k+1)$ matrix such that $G_{i,j} = \mathbb{E}[g_i(\mathbf{X})|Y = j - 1, S = 1]$, $p \in \mathbb{R}^{k+1}$ such that $p_i = \mathbb{P}(Y = j - 1|S = 0)$, and $g \in \mathbb{R}^k$ such that $g_i = \mathbb{E}[g_i(\mathbf{X})|S = 0]$, $\widehat{\theta}_{UR}$ is obtained by solving the linear system*

$$\begin{cases} \widehat{g} & = \widehat{G} \cdot \widehat{\theta}_{UR} \\ 1 & = \mathbf{1}^t \cdot \widehat{\theta}_{UR} \end{cases}, \text{ where } \widehat{g}_i = \frac{\sum_{k \in A_0} g_i(\mathbf{X}_k)}{n_U} \text{ and } \widehat{G}_{i,j} = \frac{\sum_{k \in A_{1,j}} g_i(\mathbf{X}_k)}{n_j}$$

Since $\widehat{\theta}_{UR}$ might have negative components, it is generally inadmissible according to the squared error (de Finetti, 2017)[p.90-91] that is, there exist estimators which have a squared error strictly smaller than $\widehat{\theta}_{UR}$. The ratio estimator, $\widehat{\theta}_R$ satisfies this property and is the projection of $\widehat{\theta}_{UR}$ onto the simplex (Michelot, 1986): $\widehat{\theta}_R = \arg\min_{\hat{p}:\hat{p}\geq 0, \sum_i \hat{p}_i = 1} \|\widehat{\theta}_{UR} - \hat{p}\|_2^2$.

Similarly to the AC estimator (Tasche, 2017), the ratio estimator is Fisher consistent under weak assumptions.[1]. They are described in Assumptions 3 and 4.

Assumption 3 (Weak prior shift) *The function, g, is such that $g(\mathbf{X})_1^n$ is stochastically independent of $\mathbf{S}_1^n$ conditionally on $\mathbf{Y}_1^n = y_1^n$.*

Assumption 4 (Separability) *The function, g, is such that*

1. *$\mathbb{E}[g(\mathbf{X}_i)|Y_i = j, S_i = 1]$ are defined, for $j \in \{0, 1\}$.*
2. *$\mathbb{E}[g(\mathbf{X}_i)|Y_i = 1, S_i = 1] - \mathbb{E}[g(\mathbf{X}_i)|Y_i = 0, S_i = 1] \neq 0$*

The condition in Assumption 3 is a relaxed type of prior probability shift that is strictly weaker than Assumption 2. Assumption 4 requires two more conditions of $g(\mathbf{x})$. According to condition 1, the population versions of the expectations in Definition 4 are defined. Condition 2 states that the ratio estimator calculated on these population parameters is defined, that is, there is no division by 0.

Theorem 6 *Under Assumptions 1, 3 and 4, $\widehat{\theta}_{UR}$ and $\widehat{\theta}_R$ are Fisher consistent for θ.*

[1] Although Fisher consistency is typically not equivalent to consistency in probability (Gerow, 1989; Kass and Vos, 2011), in the sequence we show that the ratio estimator is consistent in both senses.

It is also possible to guarantee a finite population bound on the mean squared error of $\widehat{\theta}_R$. This result is obtained in Theorem 7, which substitutes Assumption 4 by the stronger condition that $(f_0, f_1) \in \mathcal{F}_{g,K,\epsilon}$.

Theorem 7 *Under Assumptions 1 and 3,*

$$\sup_{(f_0,f_1)\in\mathcal{F}_{g,K,\epsilon}} \mathbb{E}_{f_0,f_1}\left[\left(\widehat{\theta}_R - \theta\right)^2 \middle| S_1^n\right] \leq \mathcal{O}(\max(n_L^{-1}, n_U^{-1}))$$

Under the assumptions of Theorem 7, if $n_U \gg n_L$, then the convergence of the mean squared error of the ratio estimator is the same as the one that would have been obtained if one observed solely n_L labels from the target population and used the sample's label proportions to estimate θ. The same type of result cannot generally be obtained for the untrimmed ratio estimator, since the trimming is necessary to guarantee that the ratio of random variables does not have infinite variance. While these conclusions are similar to the ones obtained from Theorem 3 in Lipton et al. (2018), there exist two main differences. First, while the former assumes that there are 2 labels only, the latter applies to an arbitrary number of labels. Second, Theorem 7 upper bounds the squared error by $\mathcal{O}(\max(n_L^{-1}, n_U^{-1}))$, which is slightly tighter than the bound of $\mathcal{O}\left(\max\left(\frac{\log n_L}{n_L}, \frac{\log n_U}{n_U}\right)\right)$ in Lipton et al. (2018).

It follows from Theorem 3 and Theorem 7 that the ratio estimator satisfies several desirable properties. These properties are presented in Definition 8 and Corollary 9.

Definition 8 *Let $\mathcal{S}$ and $\mathcal{F}$ be, respectively, the classes of estimators and distributions over the data under consideration. An estimator $\widehat{\theta}^* \in \mathcal{S}$ is approximately minimax for estimating θ under the squared error loss if*

$$\mathcal{O}\left(\sup_{(f_0,f_1)\in\mathcal{F}_{g,K,\epsilon}} \mathbb{E}_{f_0,f_1}\left[\left(\widehat{\theta}^* - \theta\right)^2 \middle| S_1^n\right]\right) = \Omega\left(\inf_{\widehat{\theta}\in\mathcal{S}} \sup_{(f_0,f_1)\in\mathcal{F};\theta\in[0,1]} \mathbb{E}_{f_0,f_1,\theta}\left[(\widehat{\theta} - \theta)^2 \middle| S_1^n\right]\right)$$

That is, the squared error of $\widehat{\theta}^*$ attains the optimal rate of convergence.

Corollary 9 *Under Assumptions 1 and 3, if there exists $\epsilon, K > 0$ such that $(f_0, f_1) \in \mathcal{F}_{g,K,\epsilon}$, then $\widehat{\theta}_R$ is consistent for θ in probability and in $\mathcal{L}_2$ as $n_U \xrightarrow{\mathbb{P}} \infty$ and $n_L \xrightarrow{\mathbb{P}} \infty$. Also, under Assumptions 1, 3, and $\mathcal{F}_{g,K,\epsilon}$, $\widehat{\theta}_R$ is approximately minimax.*

Corollary 9 shows that the ratio estimator converges to θ under a weaker version of the prior probability shift assumption and that the rate of this convergence is minimax (i.e., it is the same rate as that of the minimax estimator). Since the estimators from Gart and Buck (1966); Saerens et al. (2002); Forman (2008); Bella et al. (2010) are particular cases of the untrimmed ratio estimator, their trimmed versions also converge to θ under the weak prior shift.

The ratio estimator also satisfies a central limit theorem. In order to obtain this result, besides requiring Assumptions 1, 3 and 4, it is also necessary to require that conditionally on Y, $g(\mathbf{X})$ has bounded variance and that the number of labeled samples goes to infinity. These conditions are described in Assumption 5. The central limit theorem is presented in Theorem 10.

Assumption 5

1. *$\mathbb{V}[g(\mathbf{X}_i)|Y_i = j] < \infty$, for every $j \in \{0, 1\}$.*

2. *There exists $h(n) \geq 0$ such that $\lim_{n\to\infty} \frac{h(n)}{n} < 1$, $\lim_{n\to\infty} h(n) = \infty$, and $\frac{n_L}{h(n)} \xrightarrow{\mathbb{P}} 1$.*

Theorem 10 *Define $\mu_j := \mathbb{E}[g(\mathbf{X}_1)|Y_1 = j]$, $\sigma_j^2 := \mathbb{V}[g(\mathbf{X}_1)|Y_1 = j]$, $p_L := \lim_{n\to\infty} \frac{h(n)}{n}$, and $p_{j|L} := \mathbb{P}(Y = j|S = 1)$. Under Assumptions 1, 3, 4 and 5,*

1. *If $p_L \neq 0$, then*

$$\sqrt{n}(\widehat{\theta}_R - \theta) \rightsquigarrow N\left(0, \frac{\frac{(1-\theta)\sigma_0^2+\theta\sigma_1^2+(\mu_1-\mu_0)^2\theta(1-\theta)}{1-p_L} + \frac{(1-\theta)^2\sigma_0^2}{p_L p_{0|L}} + \frac{\theta^2\sigma_1^2}{p_L p_{1|L}}}{(\mu_1 - \mu_0)^2}\right)$$

2. *If $p_L = 0$, then*

$$\sqrt{h(n)}(\widehat{\theta}_R - \theta) \rightsquigarrow N\left(0, \frac{\frac{(1-\theta)^2\sigma_0^2}{p_{0|L}} + \frac{\theta^2\sigma_1^2}{p_{1|L}}}{(\mu_1 - \mu_0)^2}\right)$$

It is possible to use Theorem 10 to obtain an approximate confidence interval for θ. This interval is obtained by inverting the convergence results in Theorem 10, and substituting θ for $\widehat{\theta}_R$ and the population parameters, μ_0, μ_1, σ_0^2, σ_1^2, p_L, $p_{0|L}$ and $p_{1|L}$, by their respective empirical averages. This confidence interval may also be used to test hypothesis such as $H_0 : \theta \in \Theta_0$.

Theorem 10 also provides an approximation for the mean squared error of $\widehat{\theta}_R$. This approximation for the common case in which $n_U \gg n_L$ is presented in the following corollary.

Corollary 11 *Under Assumptions 1, 3, 4 and 5, if $p_L = 0$ ($n_U \gg n_L$), then*

$$MSE(\widehat{\theta}_R) \approx \frac{1}{n_L(\mu_1 - \mu_0)^2}\left(\frac{\sigma_0^2(1-\theta)^2}{p_{0|L}} + \frac{\sigma_1^2\theta^2}{p_{1|L}}\right) \tag{1}$$

Corollary 11 brings some insights on how g should be chosen in order for $\widehat{\theta}_R$ to be an accurate estimator of θ. For instance, it shows that one should choose g such that $|\mu_1 - \mu_0|$ is large and both σ_0^2 and σ_1^2 are small. This implies that the distributions of $g(\mathbf{X})|Y = 1$ and $g(\mathbf{X})|Y = 0$ should place most of their masses in regions that are far apart. This conclusion explains the success of the methods in Forman (2008), in which $g(\mathbf{x})$ is a classifier, and Bella et al. (2010), in which $g(\mathbf{x})$ is an estimate of $\mathbb{P}(Y = 1|\mathbf{x})$.

One of the main deficiencies of the standard AC estimator is that its denominator can be very close to zero, which makes it very unstable (due to a large variance). In order to handle this, we can explicitly use the approximation of the MSE (Corollary 11) to choose better functions g. This procedure is discussed in the following subsection.

2.3. Choosing g via approximate MSE minimization

One possible criterion for the choice of g is the minimization of $MSE(\widehat{\theta}_R)$, defined in Corollary 11. However, the latter depends on unobservable quantities. An alternative is to minimize an estimate of $MSE(\widehat{\theta}_R)$. This estimate is presented in Definition 12.

Definition 12 *Let $\hat{\theta}$ be an estimator of θ and, for each $i \in \{0,1\}$, let*

$$\widehat{\mu}_i = n_i^{-1} \sum_{A_{1,i}} g(\mathbf{X}_i) \qquad \widehat{\sigma}_i^2 = n_i^{-1} \sum_{A_{1,i}} (g(\mathbf{X}_i) - \widehat{\mu}_i)^2 \qquad \widehat{p}_{i|L} = \frac{n_i}{n_0 + n_1}$$

The empirical MSE of the ratio estimator induced by g, $\widehat{MSE}(g)$, is

$$\widehat{MSE}(g) \approx \frac{1}{n_L(\widehat{\mu}_1 - \widehat{\mu}_0)^2} \left(\frac{\widehat{\sigma}_0^2 (1-\widehat{\theta})^2}{\widehat{p}_{0|L}} + \frac{\widehat{\sigma}_1^2 \widehat{\theta}^2}{\widehat{p}_{1|L}} \right)$$

In order to avoid overfitting, we perform the minimization of $\widehat{MSE}(g)$ on a Reproducing Kernel Hilbert Space (RKHS; Wahba (1990)). More precisely, if K is a Mercer kernel and $\mathcal{H}_K$ is the RKHS associated to K, then we choose g^* as

$$g^* := \arg\min_{g \in \mathcal{H}_K} \widehat{\text{MSE}}(g) \tag{2}$$

In the following, Theorem 13 presents a characterization of g^* in eq. 2.

Theorem 13 *Let K be a Mercer kernel and $\mathcal{H}_K$ the corresponding RKHS. Also,*

- *$\mathbb{K}$: the Gram matrix defined for $(i,j) \in A_1^2$ and such that $(\mathbb{K})_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$.*
- *m_i: A vector of size $|A_1|$ and such that, for each $k \in A_1$, $m_{i,k} = \frac{\sum_{j \in A_{1,i}} K(\mathbf{x}_j, \mathbf{x}_k)}{n_i}$.*
- *$M = (m_1 - m_0)(m_1 - m_0)^t$.*
- *$\widehat{\Sigma}_i$: a $|A_1| \times |A_1|$ matrix such that $(\widehat{\Sigma}_i)_{k,l}$ is the sample covariance between $(K(\mathbf{x}_j, \mathbf{x}_k))_{j \in A_{1,i}}$ and $(K(\mathbf{x}_j, \mathbf{x}_l))_{j \in A_{1,i}}$.*
- *N: a $|A_1| \times |A_1|$ matrix such that $N = \frac{\widehat{\theta}^2}{\widehat{p}_{1|L}} \widehat{\Sigma}_1 + \frac{(1-\widehat{\theta})^2}{\widehat{p}_{0|L}} \widehat{\Sigma}_0$.*
- *$\mathbf{w}^* = \arg\min_{w \in \mathbb{R}^{n_L}} \frac{\mathbf{w}^t N \mathbf{w}}{\mathbf{w}^t M \mathbf{w}}$*

The function g^* in eq. 2 satisfies $g^*(\mathbf{x}) = \sum_{i \in A_1} w_i^* K(\mathbf{x}, \mathbf{x}_i)$.

The vector, $\mathbf{w}^*$ in Theorem 13 is the eigenvector associated to the largest eigenvalue in absolute value, λ^*, of the generalized eigenvalue problem, $M\mathbf{w}^* = \lambda^* N \mathbf{w}^*$. If N is invertible, $\mathbf{w}^*$ is the eigenvector associated to the largest eigenvalue in absolute value of $N^{-1}M$. Alternatively, if N is not invertible one can substitute N in Theorem 12 by $(N + \gamma\mathbb{1})^{-1}$, where $\mathbb{1}$ is the identity matrix and γ is a small number that makes $N + \gamma\mathbb{1}$ invertible. Adding γ to the diagonal also also adds regularization and can therefore lead to an improved solution. In practice we choose γ via data-splitting.

The results in this and in the previous section rely on the weak prior shift assumption. As shown in the next subsection, one of the advantages of this assumption to the regular prior shift is that it is easier to test.

2.4. Testing the weak prior shift assumption

The following proposition is useful for testing the weak prior shift assumption:

Proposition 14 *Under Assumption 3, there exists* $0 \leq p \leq 1$ *such that*

$$pF_{g(\mathbf{X})|S=1,Y=1} + (1-p)F_{g(\mathbf{X})|S=1,Y=0} = F_{g(\mathbf{X})|S=0}.$$

It follows from Proposition 14 that Assumption 3 entails the hypothesis:

$$H_0 : \exists 0 \leq p \leq 1 \text{ such that } pF_{g(\mathbf{X})|S=1,Y=1} + (1-p)F_{g(\mathbf{X})|S=1,Y=0} = F_{g(\mathbf{X})|S=0}$$

In the following, we show how to construct an hypothesis test for H_0 when $g(\mathbf{x})$ is continuous. Since the weak prior shift entails H_0, if this test is used to test the weak prior shift, it will have the correct type I error. However, H_0 may hold when Assumption 3 is false. A specific example in which H_0 is satisfied but Assumption 3 is not satisfied is given in the following example.

Example 1 *If* $F_{g(\mathbf{X})|S=0,Y=1} = F_{g(\mathbf{X})|S=0,Y=0}$ *and* $F_{g(\mathbf{X})|S=0,Y=1} = F_{g(\mathbf{X})|S=1,Y=1}$, *then there exists p such that* $pF_{g(\mathbf{X})|S=1,Y=1} + (1-p)F_{g(\mathbf{X})|S=1,Y=0} = F_{g(\mathbf{X})|S=0}$ *(namely,* $p = 1$*) even if* $F_{g(\mathbf{X})|S=1,Y=0} \neq F_{g(\mathbf{X})|S=0,Y=0}$. *In this case, Assumption 3 does not hold and* H_0 *is satisfied.*

The following statistic, T, measures disagreement with H_0:

$$T = \inf_{0 \leq p \leq 1} d\left(p\widehat{F}_{g(\mathbf{X})|S=1,Y=1} + (1-p)\widehat{F}_{g(\mathbf{X})|S=1,Y=0}, \widehat{F}_{g(\mathbf{X})|S=0}\right),$$

where d is a distance between cumulative distributions, such as the Kolmogorov distance, and $\widehat{F}$ are the empirical cumulative distributions (Wasserman, 2013):

$$\widehat{F}_{g(\mathbf{X})|S=1,Y=i}(w) = \frac{1}{|A_{1,i}|}\sum_{i \in A_{1,i}} \mathbb{I}(g(\mathbf{X}_i) \leq w), \quad \widehat{F}_{g(\mathbf{X})|S=0}(w) = \frac{1}{|A_0|}\sum_{i \in A_0} \mathbb{I}(g(\mathbf{X}_i) \leq w).$$

Algorithm 1, which is presented below, obtains a p-value for H_0 based on T. The algorithm uses kernel smoothers (Wasserman, 2013) to estimate the conditional densities of $g(\mathbf{X})$ given Y, $f(g(\mathbf{x})|Y=0)$ and $f(g(\mathbf{x})|Y=1)$, by $\widehat{f}(g(\mathbf{x})|Y=0)$ and $\widehat{f}(g(\mathbf{x})|Y=1)$.

Note that our test is different from those proposed by Saerens et al. (2002) and Lipton et al. (2018). While our test evaluates whether the prior shift assumption is reasonable for the observed data, the above tests *assume prior shift* and evaluate whether the prevalence of positive labels in the unlabeled sample is the same as that in the labeled sample, that is, $\mathbb{P}(Y=1|S=0) = \mathbb{P}(Y=1|S=1)$.

The following subsection performs several experiments to test the performance of Algorithm 1 and of the ratio estimators which were discussed in previous subsections.

Algorithm 1 p-value for testing the weak prior shift assumption

```
Input: Labeled and unlabeled sample, number of Monte Carlo simulations, $B$ Output: p-value
Compute $T_{\text{obs}}$, the test statistic for the observed data
Compute $\widehat{\theta}$, an estimate of $\theta$
Compute $\widehat{f}(g(\mathbf{x})|Y=1)$ and $\widehat{f}(g(\mathbf{x})|Y=0)$.
for all $i \in \{1, \ldots, B\}$ do
    Sample $W_1^{(0)}, \ldots, W_{n_0}^{(0)} \sim \widehat{f}(g(\mathbf{x})|Y=0)$ and $W_1^{(1)}, \ldots, W_{n_1}^{(1)} \sim \widehat{f}(g(\mathbf{x})|Y=1)$
    Sample $W_1^{(U)}, \ldots, W_{n_U}^{(U)} \sim \widehat{\theta}\widehat{f}(g(\mathbf{x})|Y=1) + (1-\widehat{\theta})\widehat{f}(g(\mathbf{x})|Y=0)$
    Compute $T_i$, the test statistic based on the labeled and unlabeled samples:
        $\{(W_i^{(0)}, Y_i = 0)\}_{i=1}^{n_0} \cup \{(W_i^{(1)}, Y_i = 1)\}_{i=1}^{n_1}; \quad \{W_1^{(U)}, \ldots, W_{n_u}^{(U)}\}$
end for
return $\frac{1}{B}\sum_{i=1}^{B} \mathbb{I}\left(T_{\text{obs}} \geq T_i\right)$
```

Estimator class	Specific method	Criteria for choosing $g(x)$
Classify and count		Logistic regression (LR), k-NN, random forest (RF).
Ratio	Forman (2006)	Logistic regression (LR), k-NN, random forest (RF).
	Bella et al. (2010)	Logistic regression (LR), k-NN, random forest (RF).
	RKHS	Linear kernel (Linear), Gaussian kernel (Gauss).
	EM (Saerens et al., 2002)	Logistic regression (LR), k-NN, random forest (RF).

Table 1: Methods compared in the experiments.

2.5. Experiments

Next, we compare the errors of the ratio estimator and of the classify and count estimator based on the estimator g when using various methods for obtaining g. We also include comparisons with the EM methods by Saerens et al. (2002). Table 1 summarizes all the variants that were tested.

We compare the above methods in five data sets: Candles (Freeman et al., 2013; Izbicki and Stern, 2013), Bank Marketing (Moro et al., 2011), SPAM e-mail (Blake, 1998), Wisconsin Breast Cancer (Mangasarian, 1990) and Blocks Classification (Malerba et al., 1996). Each database was transformed into a prior shift problem by choosing at random n_1 (n_0) instances among the ones labeled as 1 (0) to be marked as labeled instances and by choosing n_U instances to be marked as unlabeled. Each unlabeled unit is taken with probability θ randomly among the instances labeled as 1 in the original data set and with probability $1-\theta$ among those labeled as 0. The quantification sample sizes used in each of these data sets are described in Table 2. For all data sets we let θ vary in $\{0.1; 0.2; 0.3; 0.4; 0.5\}$ and repeated the generation and testing 100 times for each of the 11 methods in Table 1.

Figure 1 represents the average of the mean squared error (MSE; red point) and a confidence interval for the MSE (vertical blue bar) for each setting and method[2]. Figure 2 shows the number of experiments in which each method had the best average MSE,

[2] For the sake of visualization, we omit all estimators based on K-NN on this plot. Figure 9 in Appendix B contains all methods.

	data set	n_U	n_L	n_1	n_0
1	cancer	100	300	150	150
2	candles	300	300	150	150
3	block	800	300	150	150
4	spam	2000	300	150	150
5	bank	10000	300	150	150

Table 2: Sample sizes for each data set.

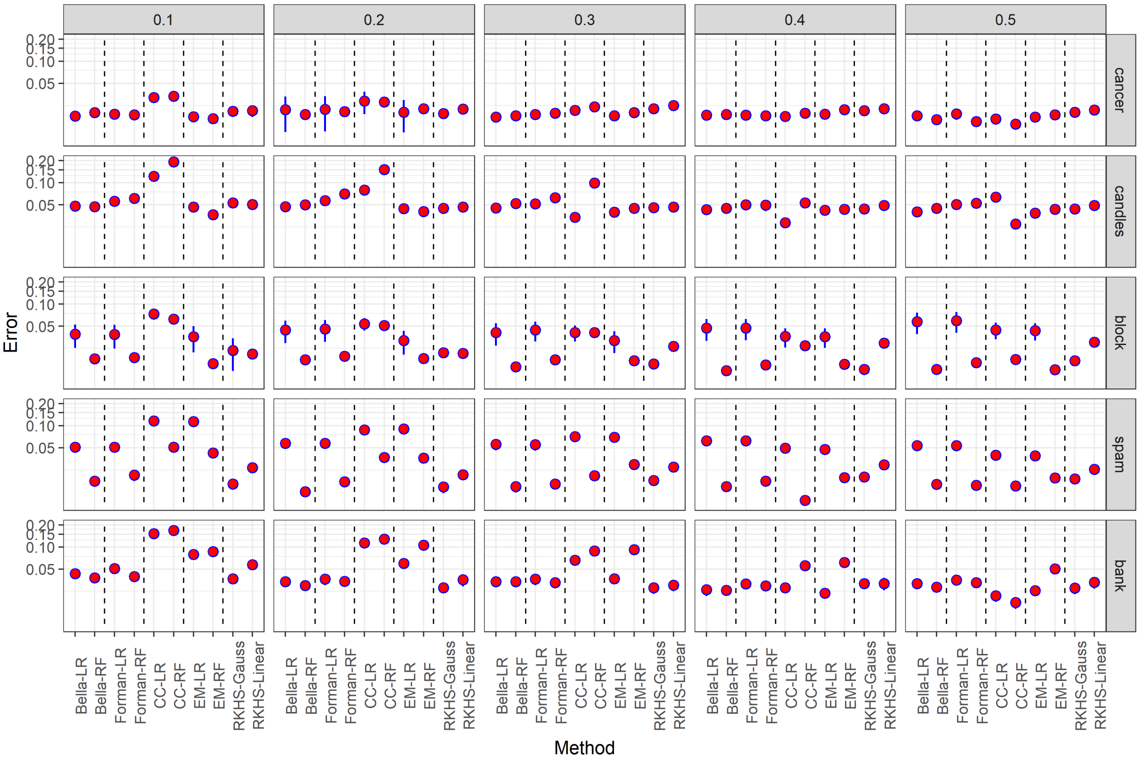

Figure 1: Root mean square deviation of each method by setting in logarithmic scale.

considering all data sets and θ values. These plots indicate that the ratio estimator generally performs better than the classify and count estimator for all choices of g. The main exception to this rule occurs when $\theta \approx 0.5$ and hence there is no prior shift. Also, the method in Bella et al. (2010) performs better than the one in Forman (2006) in essentially all scenarios. This suggests that soft classifiers might lead to better ratio estimators than hard classifiers. Moreover, the best performance is usually achieve when g is based on Random Forest, which corroborates that choosing a good classifier is key to having a good estimate of θ. The EM method was found to be very competitive in general, although the ratio estimators had better performance in some cases (e.g., for the bank data set). Finally, the RKHS approach is a competitive method, especially when using the Gaussian kernel. For additional figures related to this experiment see Appendix B.

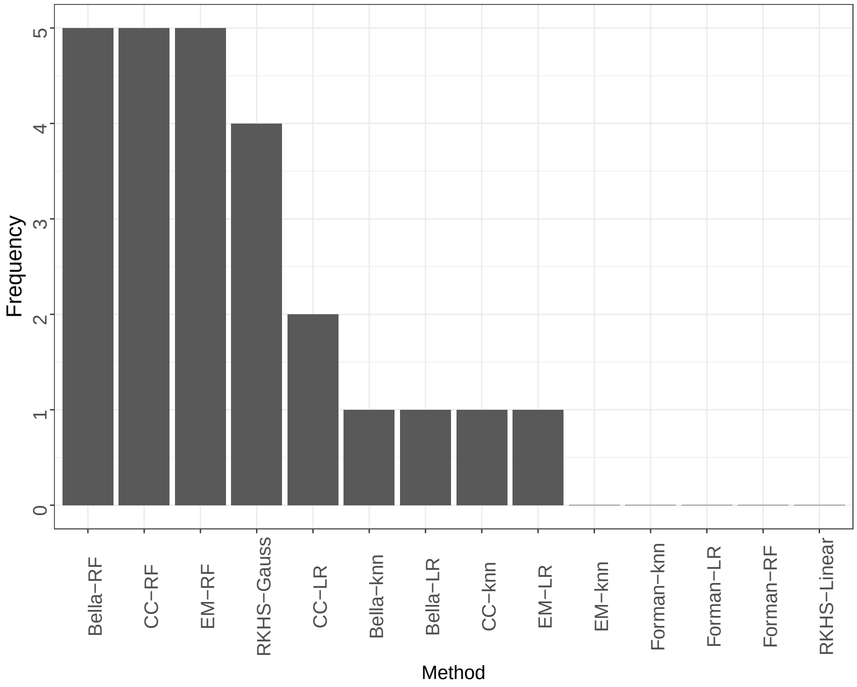

Figure 2: Number of times in which each method presented best MSE.

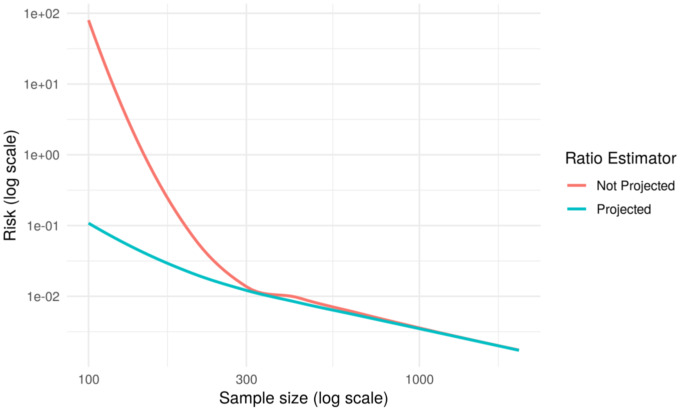

Figure 3: Mean squared error of the ratio estimator for the multiclass problem.

For all of ratio estimators, data sets and values for θ above, we construct confidence intervals for θ based on Theorem 10. We find that in all but one scenario the empirical coverage was at least as high as the specified value of 95%. The empirical coverage in the exception was 94%. The intervals constructed using Theorem 10 seem to be conservative.

Next, we simulate data using the following multiclass setting: $\mathbf{X}|Y = y \sim N(\mu_y, \Sigma)$ with $\Sigma = I_{10}$, $\mu_1 = (0, \ldots, 0)$, $\mu_2 = (0.75, \ldots, 0.75)$, $\mu_3 = (1.25, \ldots, 1.25)$, $\mathbb{P}(Y = 1|S = 1) = 0.2$,

Gaussian	**Exponential**
$g(\mathbf{x})\|S=1,Y=0 \sim \mathrm{N}(0,1)$,	$g(\mathbf{x})\|S=1,Y=0 \sim \mathrm{Exp}(1)$,
$g(\mathbf{x})\|S=1,Y=1 \sim \mathrm{N}(2,1)$,	$g(\mathbf{x})\|S=1,Y=1 \sim \mathrm{Exp}(5)$,
$g(\mathbf{x})\|S=0,Y=0 \sim \mathrm{N}(\gamma,1)$	$g(\mathbf{x})\|S=0,Y=0 \sim \mathrm{Exp}(\gamma)$
$g(\mathbf{x})\|S=0,Y=1 \sim \mathrm{N}(2,1)$,	$g(\mathbf{x})\|S=0,Y=1 \sim \mathrm{Exp}(5)$,
$\mathbb{P}(Y=1\|S=0)=0.6$,	$\mathbb{P}(Y=1\|S=0)=0.6$,
$\mathbb{P}(Y=1\|S=1)=0.2$	$\mathbb{P}(Y=1\|S=1)=0.2$
Gaussian-Exponential	**Beta**
$g(\mathbf{x})\|S=1,Y=0 \sim \mathrm{N}(1,1)$,	$g(\mathbf{x})\|S=1,Y=0 \sim \mathrm{Beta}(1,1)$,
$g(\mathbf{x})\|S=1,Y=1 \sim \mathrm{Exp}(1)$,	$g(\mathbf{x})\|S=1,Y=1 \sim \mathrm{Beta}(1,10)$,
$g(\mathbf{x})\|S=0,Y=0 \sim \mathrm{N}(\gamma,1)$	$g(\mathbf{x})\|S=0,Y=0 \sim \mathrm{Beta}(\gamma,1)$
$g(\mathbf{x})\|S=0,Y=1 \sim \mathrm{Exp}(1)$,	$g(\mathbf{x})\|S=0,Y=1 \sim \mathrm{Beta}(1,10)$,
$\mathbb{P}(Y=1\|S=0)=0.6$,	$\mathbb{P}(Y=1\|S=0)=0.6$,
$\mathbb{P}(Y=1\|S=1)=0.2$	$\mathbb{P}(Y=1\|S=1)=0.2$

Table 3: Scenarios used for testing the weak prior shift assumption.

$\mathbb{P}(Y=2|S=1)=0.3$ $\mathbb{P}(Y=1|S=0)=0.25$, and $\mathbb{P}(Y=2|S=0)=0.10$. We use a multivariate logistic regression to compute $g_1(\mathbf{x})=\widehat{P}(Y=1|\mathbf{x},S=1)$ and $g_2(\mathbf{x})=\widehat{P}(Y=2|\mathbf{x},S=1)$. Figure 3 indicates that the mean squared error of the multiclass ratio estimator goes to zero as the sample size increases. Moreover, it shows that projecting the raw estimator to the simplex improves the convergence, especially for small sample sizes.

We also evaluate the power of the weak prior shift test in Section 2.4. In order to test the weak prior shift, we generate data according to 4 scenarios, which are presented in table 2.5. In all of these scenarios, the weak prior shift assumption holds for a single value of γ. Figure 4 presents the power of the weak prior shift test in each scenario using a level of significance of $\alpha=5\%$. Besides the test achieving the level of 5% when weak prior shift holds, it also has a high power whenever the marginal distribution of $g(\mathbf{X})$ differs over the labeled and over the unlabeled data. The reason why such test presents minima when the weak priori shift assumption does not hold is described in Example 1.

The following section discusses extensions of the ratio estimator to scenarios that are more general than the standard quantification problem.

3. Extensions of the quantification problem

3.1. Combined estimator

Sometimes, a few labels are available in the target population ($S=0$). Let $A_0^* \subset A_0$ denote the indices of these labeled sample instances. In this scenario, it is possible to obtain an estimate of θ that combines the ratio estimator with the additional labels which are available. The labeled estimator of θ is defined as:

$$\widehat{\theta}_L := \frac{1}{|A_0^*|}\sum_{i\in A_0^*} Y_i$$

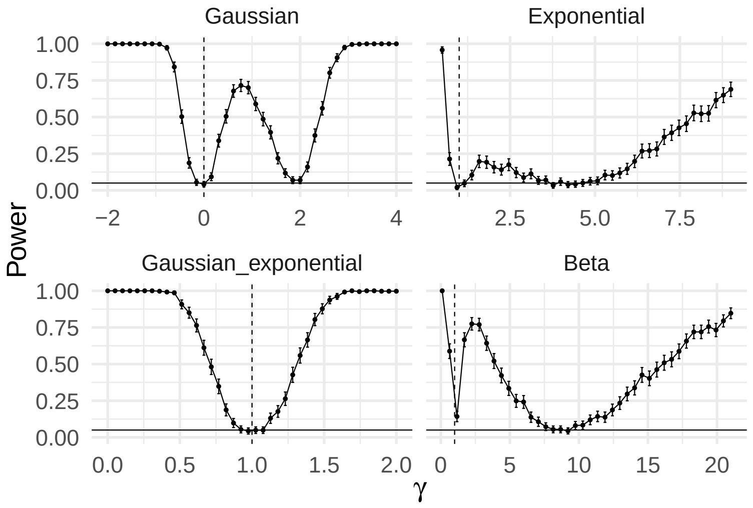

Figure 4: Power of the weak prior shift test at level $\alpha = 5\%$. The dashed vertical lines indicate the value of γ for which the weak prior shift holds.

It is possible to better estimate θ by combining the labeled estimator, $\widehat{\theta}_L$, with the ratio estimator, $\widehat{\theta}_R$. We propose to combine these estimators by means of a convex combination,

$$\widehat{\theta}_C = w\widehat{\theta}_R + (1-w)\widehat{\theta}_L, \tag{3}$$

which we name the *combined* estimator. The following theorem provides an insight on how to choose w.

Theorem 15 *Under Assumptions 1, 3, 4 and 5, the value of w that minimizes the mean squared error of the combined estimator (Equation 3) is*

$$w^* = MSE[\widehat{\theta}_L] \times (MSE[\widehat{\theta}_L] + MSE[\widehat{\theta}_R])^{-1}.$$

In practice, $MSE[\widehat{\theta}_L]$ and $MSE[\widehat{\theta}_R]$ need to be estimated. Note that $MSE[\widehat{\theta}_L] = \theta(1-\theta) \times |A_0^*|^{-1}$ and $MSE[\widehat{\theta}_R]$ is given by Theorem 10. We therefore use $\widehat{w} = \widehat{MSE}[\widehat{\theta}_L] \times (\widehat{MSE}[\widehat{\theta}_L] + \widehat{MSE}[\widehat{\theta}_R])^{-1}$, where $\widehat{MSE}[\widehat{\theta}_L]$ and $\widehat{MSE}[\widehat{\theta}_R]$ are obtained by substituting the parameters in $MSE[\widehat{\theta}_L]$ and $MSE[\widehat{\theta}_R]$ by their corresponding empirical averages.

We evaluate the combined estimator under the same scenarios used for the ratio estimator in Section 2.5 and using $\theta = 0.3$. For each scenario, we consider 10, 20, 30, 40 or 50 available labels from the target population. Figure 5 presents the errors for each setting scenario and number of available labels in the target population.[3] When one of $\widehat{\theta}_L$ and $\widehat{\theta}_R$ has an error which is much lower than the other, than this lowest error is comparable to that of the combined estimator. Also, when $\widehat{\theta}_L$ and $\widehat{\theta}_R$ have similar errors, then the error of the combined estimator is approximately $\sqrt{2}^{-1}$ times this common error. These results indicate that a few labels from the target population can improve the estimation of θ.

[3] Similar plots (with similar conclusions) for $\theta \in \{0.1, 0.2, 0.4, 0.5\}$ can be found in Figures 12—15 in Appendix B.

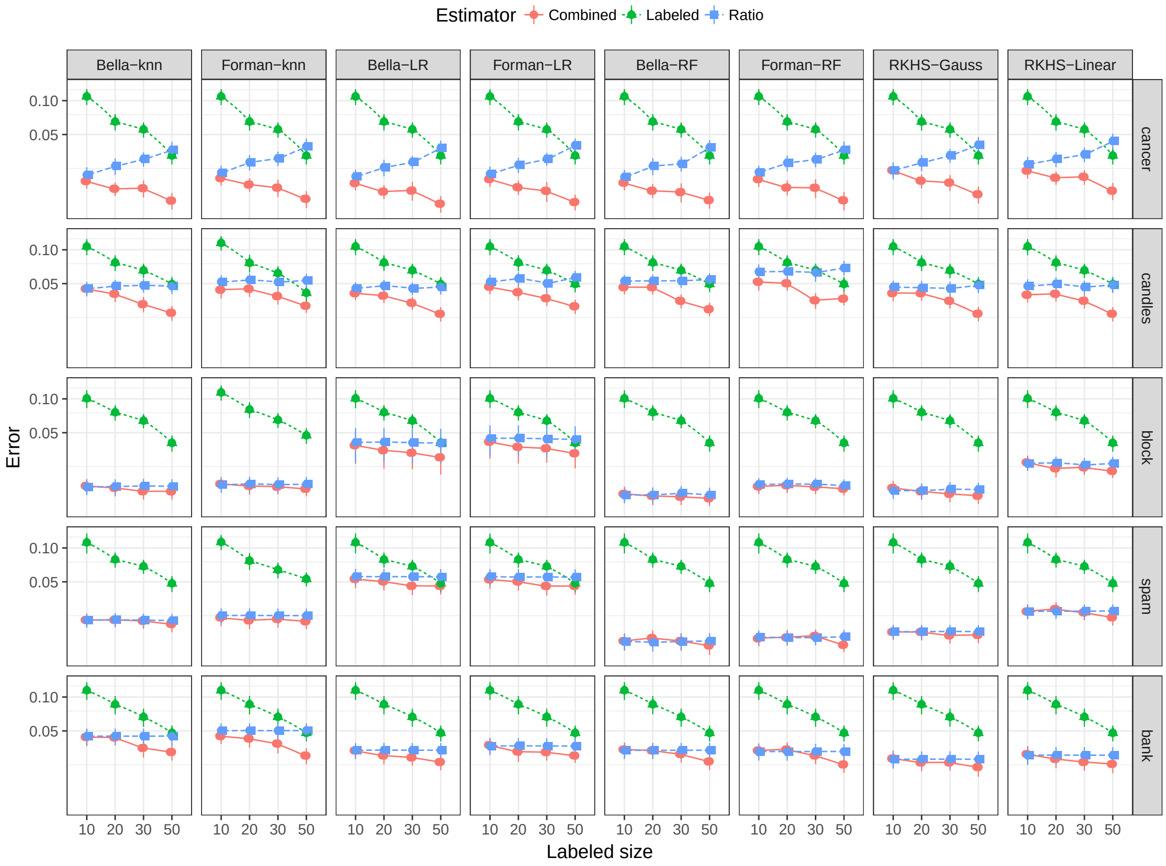

Figure 5: Root mean square deviation in logarithmic scale for each data set, method and estimator by size of the labeled sample and using $\theta = 0.3$

3.2. Regression quantification

As a generalization of the quantification problem, one might be interested on how the prevalence of Y in the target population varies according to a new set of covariates, $\mathbf{Z}$. For example, suppose that a company implements a program of continuous improvement for one of its products. In order to measure the effects of the program, it is necessary to evaluate how the proportion of positive reviews for the product varies over time. This problem fits into the generalization of the quantification problem when taking $\mathbf{Z}$ to be the date at which each review was posted. This problem is often called sentiment analysis (Wang et al., 2012) and is usually solved using a classify and count approach, which suffers from the same downsides as the ones discussed in the standard quantification problem. Other approaches can be found in Hofer and Krempl (2013).

The ratio estimator can be extended to this regression setting. Let the new sample instances be $(\mathbf{X}_1, \mathbf{Z}_1, Y_1, S_1), \ldots, (\mathbf{X}_n, \mathbf{Z}_n, Y_n, S_n)$, where $\mathbf{X}$, Y and S have the same interpretation as in the quantification problem and $\mathbf{Z}$ is the new covariate of interest. The goal in the new setting is to estimate

$$\theta(\mathbf{z}) := \mathbb{P}(Y=1|S=0,\mathbf{z}),$$

the proportion of positive labels in the target population when $\mathbf{Z}=\mathbf{z}$. In order to estimate $\theta(\mathbf{z})$, it is necessary to make additional assumptions on how $\mathbf{Z}$ relates to the other variables in the quantification problem. One such assumption is presented below.

Assumption 6 *$g(\mathbf{X})$ is stochastically independent of $\mathbf{Z}$ conditionally on Y and S.*

In the scenario in which $\mathbf{X}$ are written reviews of products and Z is time, Assumption 6 states that, if the label of a product is known, then the time at which the label was given does not affect the written review. This assumption motivates the definition of the regression ratio estimator.

Definition 16 *The untrimmed regression ratio estimator, $\widehat{\theta}_{URR}(\mathbf{z})$, is*

$$\widehat{\theta}_{URR}(\mathbf{z}) = \frac{\widehat{\mathbb{E}}[g(\mathbf{X})|S=0,\mathbf{z}] - \widehat{\mathbb{E}}[g(\mathbf{X})|Y=0,S=1]}{\widehat{\mathbb{E}}[g(\mathbf{X})|Y=1,S=1] - \widehat{\mathbb{E}}[g(\mathbf{X})|Y=0,S=1]},$$

where $\widehat{\mathbb{E}}[g(\mathbf{X})|Y=0,S=1]$ and $\widehat{\mathbb{E}}[g(\mathbf{X})|Y=1,S=1]$ are the same empirical averages as in Definition 4 and $\widehat{\mathbb{E}}[g(\mathbf{X})|S=0,\mathbf{z}]$ is an estimate of the regression function $\mathbb{E}[g(\mathbf{X})|S=0,\mathbf{z}]$. For instance $\widehat{\mathbb{E}}[g(\mathbf{X})|S=0,\mathbf{z}]$ could be the Nadaraya-Watson regression estimator (Nadaraya, 1964) based on the target population for $g(\mathbf{X})$ given $\mathbf{Z}$. The regression ratio estimator, $\widehat{\theta}_{RR}(\mathbf{z})$, is $\max(0,\min(1,\widehat{\theta}_{URR}(\mathbf{z})))$.

Next, we derive an upper bound on the rate of convergence of the regression ratio estimator.

Theorem 17 *Under Assumptions 1, 3 and 6,*

$$\mathbb{E}\left[\left(\widehat{\theta}_{RR}(\mathbf{Z})-\theta(\mathbf{Z})\right)^2\middle|S_1^n\right] \leq \mathcal{O}\left(\max\left(\mathbb{E}\left[(\widehat{\mathbb{E}}[g(\mathbf{X})|S=0,\mathbf{Z}]-\mathbb{E}[g(\mathbf{X})|S=0,\mathbf{Z}])^2|S_1^n\right], n_L^{-1}\right)\right)$$

Theorem 17 shows that the integrated mean squared error of $\widehat{\theta}_{RR}$ depends both on n_L and on the integrated mean squared error of $\widehat{\mathbb{E}}[g(\mathbf{X})|S=0,\mathbf{Z}]$ with respect to the regression function, $\mathbb{E}[g(\mathbf{X})|S=0,\mathbf{Z}]$. If one uses standard nonparametric methods to estimate $\mathbb{E}[g(\mathbf{X})|S=0,\mathbf{Z}]$, then it is possible to prove the consistency of $\widehat{\theta}_{RR}$ under weak additional assumptions. For example, if in the target population the pairs $(g(\mathbf{X}_1),\mathbf{Z}_1),\ldots,(g(\mathbf{X}_{n_U}),\mathbf{Z}_{n_U})$ are i.i.d., the regression function $\mathbb{E}[g(\mathbf{X})|S=0,\mathbf{z}]$ is sufficiently smooth over $\mathbf{z}$ and $\widehat{\mathbb{E}}[g(\mathbf{X})|S=0,\mathbf{z}]$ is the Nadaraya-Watson kernel estimator with bandwidth $h_{n_U}=O(n_U^{-1/5})$, then it follows (Wasserman, 2006)[p.73] that

$$\mathbb{E}\left[\left(\widehat{\theta}_{RR}(\mathbf{Z})-\theta(\mathbf{Z})\right)^2\middle|S_1^n\right] \leq \mathcal{O}\left(\max(n_U^{-4/5}, n_L^{-1})\right)$$

In the equation above, the rate of convergence over n_U is slower than the one obtained in Theorem 7. This is expected, since the rates of convergence of nonparametric estimators are typically slower than those of sample means.

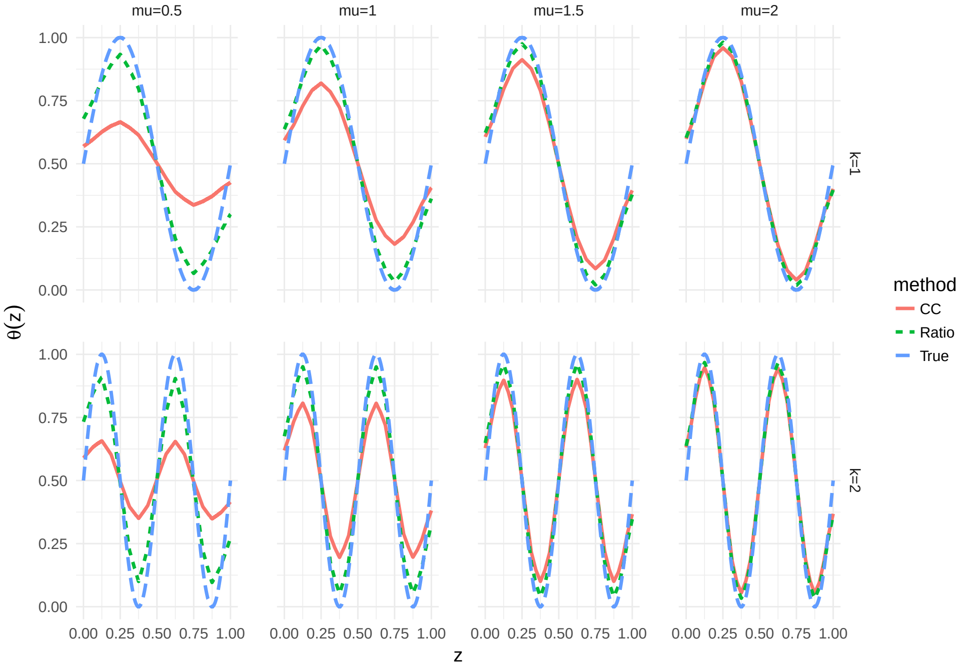

Figure 6: Average of the fitted regression in each setting.

Besides showing the consistency of the regression ratio estimator, we also show that it outperforms the classify and count method in artificial data sets. Specifically, we run the regression ratio estimator using the Nadaraya-Watson kernel estimator and generate data sets under the following specifications:

- $Z \sim U(0,1)$
- $\mathbb{P}(Y=1|S=1,Z=z)=0.5$
- $\mathbb{P}(Y=1|S=0,Z=z)=0.5(\sin(2zk\pi)+1)$, for $k \in \{1,2\}$
- $X|Y=0 \sim N(\mu,1)$ and $X|Y=1 \sim N(-\mu,1)$, for $\mu \in \{0.5,1,1.5,2\}$
- $n_L = n_U = 10^3$
- $g(X)$ is the Bayes classifier, i.e. $g(X)=\mathbb{I}(X>0)$

For each combination of k and μ, 400 independent data sets were generated. Figure 6 presents the average curve fitted for $\theta(\mathbf{z})$ using the regression ratio and classify and count estimators. One can observe that, while for small values of μ the ratio regression outperforms the classify and count estimator, for large values of μ both estimators are similar. This occurs because when μ is large, the classification problem of determining the value of Y is easier and both methods perform well. One can also observe from Figure 6 that the classify and count method performs worse than the regression ratio estimator

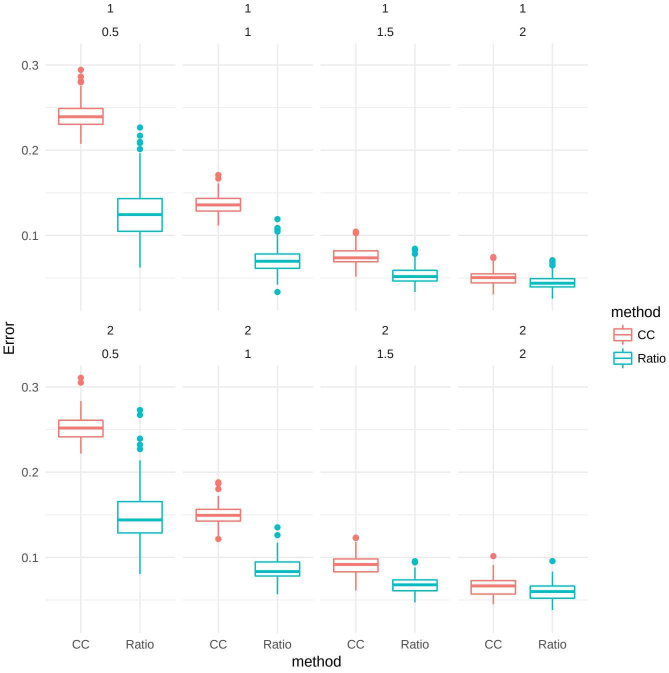

Figure 7: Boxplots of the root mean square deviation in each setting.

because its fit is generally smoother than the true regression curve. Figure 7 summarizes the mean squared error of each method in each setting. The regression ratio estimator leads to substantial improvements over the classify and count method.

4. Final remarks

We present the ratio estimator for the problem of quantification, show that it is approximately minimax under the prior probability shift assumption, and provide an hypothesis test for this assumption. Since the methods in Forman (2006) and Bella et al. (2010) are particular instances of the ratio estimator, it follows that they are also approximately minimax. The lower bound on the risk that we derive is of independent interest and can be used to investigate the optimality of other quantification methods. We also derive the limiting distribution of the ratio estimator, which allows the derivation of a ratio estimator based on Reproducing Kernel Hilbert Spaces and of confidence intervals for the quantification problem. A simulation study shows that the ratio estimator based on Reproducing Kernel Hilbert spaces is a competitive new alternative.

Besides the above results, we also generalize the ratio estimator to two other scenarios. In the first one, we consider the case in which some labels are available in the target population. The combined estimator uses these labels and the ratio estimator to obtain a larger effective sample size than the ratio estimator. In the second scenario, we consider the prevalence of positive labels varies according to a new variable, $\mathbf{Z}$. We show that, under Assumption 6, the regression ratio estimator can be made consistent for $\theta(\mathbf{Z})$. A still unresolved issue is how much it is possible to relax Assumption 6 while still being able to learn $\theta(\mathbf{Z})$.

Acknowledgments

This work was partially supported by FAPESP grant 2017/03363-8 and CNPq grant 306943/2017-4. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Appendix A. Proofs

Lemma 18 *Let $\mathbb{X}_U = (\mathbf{X}_u)_{u \in A_0}$, $\mathbf{Y}_U = (Y_u)_{u \in A_0}$, $\mathbb{X}_L = (\mathbf{X}_l)_{l \in A_1}$, and $\mathbf{Y}_L = (Y_l)_{l \in A_1}$. Under Assumptions 1 and 2, if $\theta \sim$ Uniform$(0,1)$, then $\mathbb{E}[\mathbb{V}[\theta|\mathbb{X}_U, \mathbb{X}_L, \mathbf{Y}_L]] \geq \Omega(n_U^{-1})$. Under the same assumptions, if $\mathbb{P}(\theta = 0.5 - n_L^{-1}) = \mathbb{P}(\theta = 0.5 + n_L^{-1}) = 0.5$, $f_0 =$ Bernoulli(0), $\alpha \sim$ Uniform$(\epsilon^*, 1)$ $f_1|\alpha =$ Bernoulli(α), then $\mathbb{E}[\mathbb{V}[\theta|\mathbb{X}_U, \mathbb{X}_L, \mathbf{Y}_L]] \geq \Omega(n_L^{-1})$.*

Proof It follows from Assumptions 1 and 2 that the dependency relations between data and parameters can be represented by figure 8.

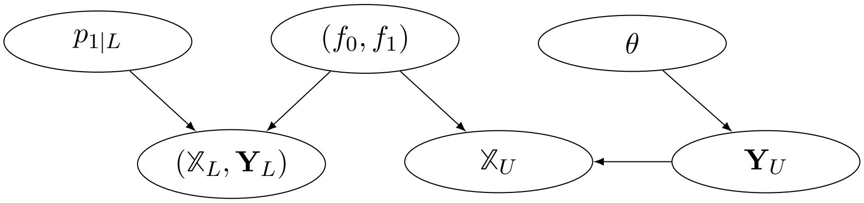

Figure 8: Dependency relations between data and parameters in the prior shift model.

If $\theta \sim U(0,1)$, then

$$\begin{aligned}
\mathbb{E}[\mathbb{V}[\theta|\mathbb{X}_U, \mathbb{X}_L, \mathbf{Y}_L]|S_1^n] &\geq \mathbb{E}[\mathbb{V}[\theta|\mathbb{X}_U, \mathbb{X}_L, \mathbf{Y}_L, \mathbf{Y}_U]|S_1^n] \\
&= \mathbb{E}[\mathbb{V}[\theta|\mathbf{Y}_U]|S_1^n] && \text{fig. 8} \\
&= \Omega(n_U^{-1}) && \mathbf{Y}_U|\theta \text{ i.i.d. Bernoulli}(\theta)
\end{aligned} \tag{4}$$

Next, let $\mathbb{P}(\theta = 0.5 - n_L^{-0.5}) = \mathbb{P}(\theta = 0.5 + n_L^{-0.5}) = 0.5$, $f_0 =$ Bernoulli(0) $\alpha \sim U(\epsilon, 1)$ and $f_1|\alpha =$ Bernoulli(α). Define $\mathbb{X}_{L,1} = (\mathbf{X}_l)_{l \in A_{1,1}}$ and note that

$$f(\alpha|\mathbb{X}_L, \mathbf{Y}_L) \propto \alpha^{n_{L,1}\bar{\mathbb{X}}_{L,1}}(1-\alpha)^{n_{L,1}(1-\bar{\mathbb{X}}_{L,1})}\mathbb{I}(\alpha \in (\epsilon, 1)) \tag{5}$$

Let $\lambda := \theta\alpha$ and observe that, for every $\lambda \in \left(\epsilon(0.5 + n_L^{-0.5}), 0.5 - n_L^{-0.5}\right]$,

$$\begin{aligned}
\mathbb{P}(\theta = 0.5 + n_L^{-0.5}|\lambda, \mathbb{X}_L, \mathbf{Y}_L) &= \frac{f_\alpha(\lambda(0.5 + n_L^{-0.5})^{-1}|\mathbb{X}_L, \mathbf{Y}_L)}{f_\alpha(\lambda(0.5 + n_L^{-0.5})^{-1}|\mathbb{X}_L, \mathbf{Y}_L) + f_\alpha(\lambda(0.5 - n_L^{-0.5})^{-1}|\mathbb{X}_L, \mathbf{Y}_L)} \\
&= \frac{1}{1 + \frac{\left(\frac{\lambda}{0.5 - n_L^{-0.5}}\right)^{n_{L,1}\bar{\mathbb{X}}_{L,1}}\left(1 - \frac{\lambda}{0.5 - n_L^{-0.5}}\right)^{n_{L,1}(1-\bar{\mathbb{X}}_{L,1})}}{\left(\frac{\lambda}{0.5 + n_L^{-0.5}}\right)^{n_{L,1}\bar{\mathbb{X}}_{L,1}}\left(1 - \frac{\lambda}{0.5 + n_L^{-0.5}}\right)^{n_{L,1}(1-\bar{\mathbb{X}}_{L,1})}}} & eq.5 \\
&= \frac{1}{1 + \left(1 + \frac{4}{n_L^{0.5} - 2}\right)^{n_{L,1}}\left(1 - \frac{4}{(1-2\theta\alpha)n_L^{0.5} + 2}\right)^{n_{L,1}(1-\bar{\mathbb{X}}_{L,1})}} & \lambda = \theta\alpha
\end{aligned} \tag{6}$$

Let $\gamma_{n_L} := \mathbb{P}(\theta = 0.5 + n_L^{-1}|\lambda, \mathbb{X}_L, \mathbf{Y}_L)$. Note that $\frac{n_{L,1}}{n_L} \overset{a.s.}{\rightarrow} p_{1|L}$ and $\bar{\mathbb{X}}_{L,1} \overset{a.s.}{\rightarrow} \alpha$. Therefore, since $|\theta - 0.5| \leq n_L^{-0.5}$, it follows from eq. 6 that γ_{n_L} converges a.s. to a quantity between $\frac{1}{1+\exp\left(\frac{-8\alpha p_{1|L}}{1-\alpha}\right)}$ and $\frac{1}{1+\exp\left(\frac{8\alpha p_{1|L}}{1-\alpha}\right)}$. That is,

$$\mathbb{E}[\gamma_{n_L}(1 - \gamma_{n_L})|S_1^n] \geq \Omega(1). \tag{7}$$

Note that $\bar{\mathbb{X}}_U$ is sufficient for (θ, α) and $\bar{\mathbb{X}}_U$ converges a.s. to λ. Therefore,

$$\begin{aligned}
\mathbb{E}[\mathbb{V}[\theta|\mathbb{X}_U, \mathbb{X}_L, \mathbf{Y}_L]|S_1^n] &= \mathbb{E}[\mathbb{V}[\theta|\bar{\mathbb{X}}_U, \mathbb{X}_L, \mathbf{Y}_L]|S_1^n] \\
&\geq \mathbb{E}[\mathbb{V}[\theta|\lambda, \mathbb{X}_L, \mathbf{Y}_L]|S_1^n] \\
&\geq 4n_L^{-1}\mathbb{E}\left[\gamma_{n_L}(1 - \gamma_{n,L})\mathbb{I}\left(\lambda \in \left(\epsilon(0.5 + n_L^{-1}), 0.5 - n_L^{-1}\right]\right)|S_1^n\right]
\end{aligned} \tag{8}$$

Since $\mathbb{P}(\lambda \in \epsilon(0.5 + n_L^{-1}), 0.5 - n_L^{-1}) \geq 0.5$, it follows from eqs. 7 and 8 that

$$\mathbb{E}[\mathbb{V}[\theta|\mathbb{X}_U, \mathbb{X}_L, \mathbf{Y}_L]|S_1^n] \geq \Omega(n_L^{-1}) \tag{9}$$

The proof follows from combining eqs. 4 and 9. ■

Proof [Theorem 3] We wish to find a lower bound for the minimax risk given a constraint, $\mathcal{F}$. In order to do so, we use the result that the minimax risk is lower bounded by the Bayes risk of any Bayes estimator associated to a prior with support in $\mathcal{F}$ (Wasserman, 2006; Esteves et al., 2017). Since we consider the squared error loss, the Bayes risk of the Bayes estimator is $\mathbb{E}[\mathbb{V}[\theta|\mathbb{X}_U, \mathbb{X}_L, \mathbf{Y}_L]]$. Hence, if there exists two priors with support in $\mathcal{F}$ such that the first one satisfies $\mathbb{E}[\mathbb{V}[\theta|\mathbb{X}_U, \mathbb{X}_L, \mathbf{Y}_L]] \geq \Omega(n_L^{-1})$ and the second one satisfies $\mathbb{E}[\mathbb{V}[\theta|\mathbb{X}_U, \mathbb{X}_L, \mathbf{Y}_L]] \geq \Omega(n_U^{-1})$, then we can conclude that the minimax risk is lower bounded by $\Omega(\max(n_L^{-1}, n_U^{-1}))$. Lemma 18 can be used to determine these priors for the classes $\mathcal{F}_{\mathcal{L}_1,\epsilon}$ and $\mathcal{F}_{g,k,\epsilon}$. The proof for $\mathcal{F}_{\mathcal{L}_1,\epsilon}$ follows from taking $\epsilon^* = \epsilon$ in Lemma 18. Next, if $\mathcal{F}_{g,k,\epsilon} \neq \emptyset$, then there exist a and b such that $\frac{\epsilon}{|g(a)-g(b)|} < 1$. Without loss of generality, let $a = 0$ and $b = 1$. The proof follows from taking $\epsilon^* = \frac{\epsilon}{|g(1)-g(0)|}$ in Lemma 18. ■

Lemma 19 *For every function, g, under Assumption 3,*

$$\theta := \mathbb{P}(Y = 1|S = 0) = \frac{\mathbb{E}[g(\mathbf{X})|S = 0] - \mathbb{E}[g(\mathbf{X})|Y = 0, S = 1]}{\mathbb{E}[g(\mathbf{X})|Y = 1, S = 1] - \mathbb{E}[g(\mathbf{X})|Y = 0, S = 1]}$$

Proof Let $f(\mathbf{x})$ denote the density of $\mathbf{X}$. Note that

$$\begin{aligned}
g(\mathbf{x})f(\mathbf{x}|S = 0) &= \sum_{j=0}^{1} g(\mathbf{x})f(\mathbf{x}|Y = j, S = 0)\mathbb{P}(Y = j|S = 0) && \text{Law of total prob.} \\
\mathbb{E}[g(\mathbf{X})|S = 0] &= \sum_{j=0}^{1} \mathbb{E}[g(\mathbf{X})|Y = j, S = 0]\mathbb{P}(Y = j|S = 0) && \text{Integration over } \mathbf{x} \\
&= \sum_{j=0}^{1} \mathbb{E}[g(\mathbf{X})|Y = j, S = 1]\mathbb{P}(Y = j|S = 0) && \text{Assumption 3} \quad (10)
\end{aligned}$$

Isolating $\mathbb{P}(Y = 1|S = 0)$ in equation 10 yields

$$\theta := \mathbb{P}(Y = 1|S = 0) = \frac{\mathbb{E}[g(\mathbf{X})|S = 0] - \mathbb{E}[g(\mathbf{X})|Y = 0, S = 1]}{\mathbb{E}[g(\mathbf{X})|Y = 1, S = 1] - \mathbb{E}[g(\mathbf{X})|Y = 0, S = 1]}$$

■

Proof [Theorem 6] Follows directly from the definition of $\widehat{\theta}_{UR}$ and $\widehat{\theta}_{R}$, and Lemma 19. ■

Lemma 20 *Let Z_1 and Z_2 be random variables such that $\mathbb{E}[Z_2] \neq 0$ and $\frac{\mathbb{E}[Z_1]}{\mathbb{E}[Z_2]} \in [0, 1]$. Define $T = \max\left(0, \min\left(1, \frac{Z_1}{Z_2}\right)\right)$. For every random variable, S, and $\epsilon_1, \epsilon_2 \in (0, 1)$.*

$$\mathbb{E}\left[\left(T - \frac{\mathbb{E}[Z_1|S]}{\mathbb{E}[Z_2|S]}\right)^2 \middle| S\right] \leq \frac{4(|\mathbb{E}[Z_1|S]| + \epsilon_1)\max\left(\mathbb{V}[Z_1|S], \mathbb{V}[Z_2|S]\right)}{\min(1, (1-\epsilon_2)^4\mathbb{E}[Z_2|S]^4)} + \epsilon_1^{-2}\mathbb{V}[Z_1|S] + (\epsilon_2\mathbb{E}[Z_2|S])^{-2}\mathbb{V}[Z_2|S]$$

Proof It follows from Taylor's expansion of $\frac{Z_1}{Z_2}$ that there exists $Z_{1,*}$ bounded between $\mathbb{E}[Z_1|S]$ and Z_1, and $Z_{2,*}$ between $\mathbb{E}[Z_2|S]$ and Z_2 such that

$$\frac{Z_1}{Z_2} = \frac{\mathbb{E}[Z_1|S]}{\mathbb{E}[Z_2|S]} + \frac{1}{Z_{2,*}}(Z_1 - \mathbb{E}[Z_1|S]) - \frac{Z_{1,*}}{Z_{2,*}^2}(Z_2 - \mathbb{E}[Z_2|S])$$

Therefore, by letting $A = \{|Z_1 - \mathbb{E}[Z_1|S]| \leq \epsilon_1, |Z_2 - \mathbb{E}[Z_2|S]| \leq \epsilon_2 \mathbb{E}[Z_2|S]\}$, obtain

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{Z_1}{Z_2} - \frac{\mathbb{E}[Z_1|S]}{\mathbb{E}[Z_2|S]}\right)^2 \middle| A, S\right]\mathbb{P}(A|S) \\
=&\mathbb{E}\left[\left(\frac{1}{Z_{2,*}}(Z_1 - \mathbb{E}[Z_1|S]) - \frac{Z_{1,*}}{Z_{2,*}^2}(Z_2 - \mathbb{E}[Z_2|S])\right)^2 \middle| A, S\right]\mathbb{P}(A|S) \\
\leq& 4 \max\left(\mathbb{E}\left[\frac{1}{Z_{2,*}^2}(Z_1 - \mathbb{E}[Z_1|S])^2 \middle| A, S\right], \mathbb{E}\left[\frac{Z_{1,*}^2}{Z_{2,*}^4}(Z_2 - \mathbb{E}[Z_2|S])^2 \middle| A, S\right]\right)\mathbb{P}(A|S) \\
\leq& \frac{4(|\mathbb{E}[Z_1|S]| + \epsilon_1)\max\left(E\left[(Z_1 - \mathbb{E}[Z_1|S])^2 | A, S\right], E\left[(Z_2 - \mathbb{E}[Z_2|S])^2 \middle| A, S\right]\right)\mathbb{P}(A|S)}{\min(1, (1-\epsilon_2)^4 \mathbb{E}[Z_2|S]^4)} \\
\leq& \frac{4(|\mathbb{E}[Z_1|S]| + \epsilon_1)\max\left(\mathbb{V}[Z_1|S], \mathbb{V}[Z_2|S]\right)}{\min(1, (1-\epsilon_2)^4 \mathbb{E}[Z_2|S]^4)} \qquad (11)
\end{aligned}$$

Finally, obtain that

$$\begin{aligned}
E\left[\left(T - \frac{\mathbb{E}[Z_1|S]}{\mathbb{E}[Z_2|S]}\right)^2 \middle| S\right] &= \mathbb{E}\left[\mathbb{E}\left[\left(T - \frac{\mathbb{E}[Z_1|S]}{\mathbb{E}[Z_2|S]}\right)^2 \middle| \mathbb{I}_A, S\right] \middle| S\right] \\
&\leq \mathbb{E}\left[\left(\frac{Z_1}{Z_2} - \frac{\mathbb{E}[Z_1|S]}{\mathbb{E}[Z_2|S]}\right)^2 \middle| A, S\right]\mathbb{P}(A|S) + \mathbb{P}(A^c|S) \qquad T, \frac{\mathbb{E}[Z_1]}{\mathbb{E}[Z_2]} \in [0,1] \\
&\leq \frac{4(|\mathbb{E}[Z_1|S]| + \epsilon_1)\max\left(\mathbb{V}[Z_1|S], \mathbb{V}[Z_2|S]\right)}{\min(1, (1-\epsilon_2)^4 \mathbb{E}[Z_2|S]^4)} + \mathbb{P}(A^c|S) \qquad \text{eq. } 11
\end{aligned}$$

The result follows from applying the union bound and Chebyshev's inequality to obtain

$$\begin{aligned}
\mathbb{P}(A^c|S) &\leq \mathbb{P}(|Z_1 - \mathbb{E}[Z_1|S]| > \epsilon_1 | S) + \mathbb{P}(|Z_2 - \mathbb{E}[Z_2|S]| > \epsilon_2 \mathbb{E}[Z_2|S] | S) \\
&\leq \epsilon_1^{-2}\mathbb{V}[Z_1|S] + (\epsilon_2 \mathbb{E}[Z_2|S])^{-2}\mathbb{V}[Z_2|S]
\end{aligned}$$

■

Proof [Theorem 7] Define $Z_1 = \frac{\sum_{i \in A_0} g(\mathbf{X}_i)}{|A_0|} - \frac{\sum_{i \in A_{1,0}} g(\mathbf{X}_i)}{|A_{1,0}|}$ and also $Z_2 = \frac{\sum_{i \in A_{1,1}} g(\mathbf{X}_i)}{|A_{1,1}|} - \frac{\sum_{i \in A_{1,0}} g(\mathbf{X}_i)}{|A_{1,0}|}$. Note that

$$\begin{aligned}
\mathbb{E}\left[\frac{\sum_{i \in A_0} g(\mathbf{X}_i)}{|A_0|} \middle| \mathbf{S}_1^n\right] &= \mathbb{E}[g(\mathbf{X})|S = 0] \\
\mathbb{E}\left[\frac{\sum_{i \in A_{1,j}} g(\mathbf{X}_i)}{|A_{1,j}|} \middle| \mathbf{S}_1^n\right] &= \mathbb{E}[g(\mathbf{X})|Y = j, S = 1]
\end{aligned}$$

It follows from Lemma 19 that $\theta = \frac{\mathbb{E}[Z_1|\mathbf{S}_1^n]}{\mathbb{E}[Z_2|\mathbf{S}_1^n]}$. With $T := \widehat{\theta}_R = \max\left(0, \min\left(1, \frac{Z_1}{Z_2}\right)\right)$, obtain

$$\begin{aligned}
\mathbb{E}\left[\left(\widehat{\theta}_R - \theta\right)^2 \middle| \mathbf{S}_1^n\right] &= \mathbb{E}\left[\left(T - \frac{\mathbb{E}[Z_1|\mathbf{S}_1^n]}{\mathbb{E}[Z_2|\mathbf{S}_1^n]}\right)^2 \middle| \mathbf{S}_1^n\right] \\
&\leq \frac{4(|\mathbb{E}[Z_1|\mathbf{S}_1^n]| + \epsilon_1)\max\left(\mathbb{V}[Z_1|\mathbf{S}_1^n], \mathbb{V}[Z_2|\mathbf{S}_1^n]\right)}{\min(1, (1-\epsilon_2)^4\mathbb{E}[Z_2|\mathbf{S}_1^n]^4)} \\
&\quad + \epsilon_1^{-2}\mathbb{V}[Z_1|\mathbf{S}_1^n] + (\epsilon_2\mathbb{E}[Z_2|\mathbf{S}_1^n])^{-2}\mathbb{V}[Z_2|\mathbf{S}_1^n] \qquad \text{Lemma 20}
\end{aligned}$$

The result follows from observing that, under $\mathcal{F}_{g,K,\epsilon}$, $\mathbb{E}[Z_1|\mathbf{S}_1^n]$ and $\mathbb{E}[Z_2|\mathbf{S}_1^n]$ are bounded by constants, $\mathbb{V}[Z_1|\mathbf{S}_1^n] = \mathcal{O}(\max(n_L^{-1}, n_U^{-1}))$, and $\mathbb{V}[Z_2|\mathbf{S}_1^n] = \mathcal{O}(n_L^{-1})$. ■

Proof [Theorem 10] The proof strategy is divided into two parts. The first part consists of proving a joint central limit theorem for the three sample averages that appear in the untrimmed ratio estimator. The second part uses this central limit theorem and the delta method to complete the proof for each case that is considered in the theorem.

The main challenge appears when proving the central limit theorem for the sample averages that appear in the ratio estimator. This occurs since these averages are not marginally independent. However, they are independent conditional on the values of $\mathbf{Y}$ and $\mathbf{S}$. This conditional independence can be used to calculate the limiting behavior of the characteristic function of the standardized averages, which completes this part of the proof.

We tidy the proof by using the following notation: $\mu_U := \mathbb{E}[g(\mathbf{X}_i)|S_i = 0]$, $\sigma_U^2 := \mathbb{V}[g(\mathbf{X}_i)|S_i = 0]$, $Z_{U,n} := \frac{\sqrt{n_U}}{\sigma_U}\left(\frac{\sum_{i=1}^n g(\mathbf{X}_i)\mathbb{I}(S_i=0)}{n_U} - \mu_U\right)$, $Z_{j,n} := \frac{\sqrt{n_j}}{\sigma_j}\left(\frac{\sum_{i=1}^n g(\mathbf{X}_i)\mathbb{I}(S_i=0,Y_i=j)}{n_j} - \mu_j\right)$, $F_i = \mathbb{I}(S_i = 1)(Y_i + 1)$, $A_U = \{F_1 = 0\}$, $A_0 = \{F_1 = 1\}$, and $A_1 = \{F_1 = 2\}$. Note that

$$\begin{aligned}
\lim_{n\to\infty} \phi_{Z_{U,n},Z_{0,n},Z_{1,n}}(t_U, t_0, t_1) &= \lim_{n\to\infty} \mathbb{E}\left[\mathbb{E}\left[\exp\left(\sum_{j\in\{U,0,1\}} it_j Z_{j,n}\right) \middle| F_1, \ldots, F_n\right]\right] \\
&= \lim_{n\to\infty} \mathbb{E}\left[\prod_{j\in\{U,0,1\}} \mathbb{E}\left[\exp\left(it_j Z_{j,n}\right) \middle| F_1, \ldots, F_n\right]\right] \\
&= \lim_{n\to\infty} \mathbb{E}\left[\prod_{j\in\{U,0,1\}} \left(\phi_{\frac{g(\mathbf{x}_1)-\mu_j}{\sigma_j}\big|A_j}(t_j n_j^{-0.5})\right)^{n_j}\right] \qquad (12)
\end{aligned}$$

It follows from the Central Limit Theorem for i.i.d. random variables that, for every $j \in \{U, 0, 1\}$, $\phi^{n_j}_{\frac{g(\mathbf{x}_1)-\mu_j}{\sigma_j}\big|A_j}(t_j n_j^{-0.5}) \to \exp(-0.5t_j^2)$ as $n_j \to \infty$. Since $n_j \overset{a.s.}{\to} \infty$, conclude from eq. 12 and the dominated convergence theorem that

$$\lim_{n\to\infty} \phi_{Z_{U,n},Z_{0,n},Z_{1,n}}(t_U, t_0, t_1) = \prod_{j\in\{U,0,1\}} \exp(-0.5t_j^2)$$

and, using $\mathbb{1}$ as the identity matrix, obtain

$$(Z_{U,n}, Z_{0,n}, Z_{1,n}) \rightsquigarrow N(0, \mathbb{1}) \qquad (13)$$

Assume that $p_L \neq 0$. In this case, since $\frac{n_L}{n} \xrightarrow{\mathbb{P}} p_L$, it follows from eq. 13 that

$$\sqrt{n}\left(\frac{\sum_{i=1}^n g(\mathbf{X}_i)\mathbb{I}(S_i=0)}{n_U} - \mu_U, \frac{\sum_{i=1}^n g(\mathbf{X}_i)\mathbb{I}(S_i=0,Y_i=0)}{n_0} - \mu_0, \frac{\sum_{i=1}^n g(\mathbf{X}_i)\mathbb{I}(S_i=0,Y_i=1)}{n_1} - \mu_1\right)$$

converges in distribution to $N\left(0, diag\left(\frac{\sigma_U^2}{1-p_L}, \frac{\sigma_0^2}{p_L p_{0|L}}, \frac{\sigma_1^2}{p_L p_{1|L}}\right)\right)$. Since $\theta = \frac{\mu_U - \mu_0}{\mu_1 - \mu_0}$ (Lemma 19) and $\widehat{\theta}_{UR} = \frac{\frac{\sum_{i=1}^n g(\mathbf{X}_i)\mathbb{I}(S_i=0)}{n_U} - \frac{\sum_{i=1}^n g(\mathbf{X}_i)\mathbb{I}(S_i=0,Y_i=0)}{n_0}}{\frac{\sum_{i=1}^n g(\mathbf{X}_i)\mathbb{I}(S_i=0,Y_i=1)}{n_1} - \frac{\sum_{i=1}^n g(\mathbf{X}_i)\mathbb{I}(S_i=0,Y_i=0)}{n_0}}$, it follows from the delta method (Casella and Berger, 2002) that

$$\sqrt{n}(\widehat{\theta}_{UR} - \theta) \rightsquigarrow N\left(0, \frac{\sigma_U^2(1-p_L)^{-1}}{(\mu_1-\mu_0)^2} + \frac{(\mu_U-\mu_1)^2\sigma_0^2(p_L p_{0|L})^{-1}}{(\mu_1-\mu_0)^4} + \frac{(\mu_U-\mu_0)^2\sigma_1^2(p_L p_{1|L})^{-1}}{(\mu_1-\mu_0)^4}\right)$$

Since $\mu_U = (1-\theta)\mu_0 + \theta\mu_1$ and $\sigma_U^2 = (1-\theta)\sigma_0^2 + \theta\sigma_1^2 + (\mu_1-\mu_0)^2\theta(1-\theta)$ obtain that

$$\sqrt{n}(\widehat{\theta}_{UR} - \theta) \rightsquigarrow N\left(0, \frac{\frac{(1-\theta)\sigma_0^2 + \theta\sigma_1^2 + (\mu_1-\mu_0)^2\theta(1-\theta)}{1-p_L} + \frac{(1-\theta)^2\sigma_0^2}{p_L p_{0|L}} + \frac{\theta^2\sigma_1^2}{p_L p_{1|L}}}{(\mu_1-\mu_0)^2}\right)$$

Next, assume that $p_L = 0$. Obtain that $\sqrt{h(n)}(Z_{U,n} - \mu_U) \xrightarrow{\mathbb{P}} 0$ and

$$\sqrt{h(n)}\left(\frac{\sqrt{p_{0|L}}}{\sigma_0}(Z_{0,n} - \mu_0), \frac{\sqrt{p_{1|L}}}{\sigma_1}(Z_{1,n} - \mu_1)\right) \rightsquigarrow N(0, \mathbb{1})$$

It follows from the delta method and Slutsky's theorem that

$$\sqrt{h(n)}(\widehat{\theta}_{UR} - \theta) \rightsquigarrow N\left(0, \frac{\frac{(1-\theta)^2\sigma_0^2}{p_{0|L}} + \frac{\theta^2\sigma_1^2}{p_{1|L}}}{(\mu_1-\mu_0)^2}\right)$$

The same convergence results hold for $\widehat{\theta}_R$ since the derivative of the trimming function is 1 around θ. ■

Proof [Theorem 13] It follows from the Representer Theorem (Wahba, 1990) that, for every $g \in \mathcal{H}_K$, $g(\mathbf{x}) = \sum_{k \in A_1} w_k K(\mathbf{x}, \mathbf{x}_k)$. Using this fact, for every $i \in \{0, 1\}$,

$$\widehat{\mu}_i = \frac{\sum_{j \in A_{1,i}} g(\mathbf{x}_j)}{n_i} = \frac{\sum_{k \in A_1} w_k \sum_{j \in A_{1,i}} K(\mathbf{x}_j, \mathbf{x}_k)}{n_i} = \mathbf{w}^t m_i$$

$$\widehat{\sigma}_i^2 = \frac{\sum_{j \in A_{1,i}} (g(\mathbf{x}_j) - \widehat{\mu}_i)^2}{n_i} = \mathbf{w}^t \widehat{\Sigma}_i \mathbf{w}$$

Therefore, for every $g \in \mathcal{H}_K$,

$$\widehat{\text{MSE}}(g) = \frac{\mathbf{w}^t N \mathbf{w}}{\mathbf{w}^t M \mathbf{w}}.$$

■

Proof [Proposition 14]

$$\begin{aligned} F_{g(\mathbf{X})|S=0} &= \theta F_{g(\mathbf{X})|S=0,Y=1} + (1-\theta)F_{g(\mathbf{X})|S=0,Y=0} \\ &= \theta F_{g(\mathbf{X})|S=1,Y=1} + (1-\theta)F_{g(\mathbf{X})|S=1,Y=0} && 3 \end{aligned}$$

Thus, there exists $0 \leq p \leq 1$ such that $F_{g(\mathbf{X})|S=0} = pF_{g(\mathbf{X})|S=1,Y=1} + (1-p)F_{g(\mathbf{X})|S=1,Y=0}$. ■

Proof [Theorem 15]

$$\begin{aligned} MSE[\widehat{\theta}_C] = \mathbb{E}\left[\left((w\widehat{\theta}_R + (1-w)\widehat{\theta}_L) - \theta\right)^2\right] &= \mathbb{E}\left[\left((w(\widehat{\theta}_R - \theta) + (1-w)(\widehat{\theta}_L - \theta)\right)^2\right] \\ &= w^2 MSE[\widehat{\theta}_R] + (1-w)^2 MSE[\widehat{\theta}_L] + 2w(1-w)\mathbb{E}[(\widehat{\theta}_R - \theta)(\widehat{\theta}_L - \theta)] && \widehat{\theta}_R \text{ indep. } \widehat{\theta}_L, \mathbb{E}[\widehat{\theta}_L] = \theta \\ &= w^2 MSE[\widehat{\theta}_R] + (1-w)^2 MSE[\widehat{\theta}_L] \end{aligned}$$

It follows that $MSE[\widehat{\theta}_C]$ is minimized by taking $w = MSE[\widehat{\theta}_L] \times (MSE[\widehat{\theta}_L] + MSE[\widehat{\theta}_R])^{-1}$.
■

Lemma 21 *For every function, g, under Assumptions 3 and 6*

$$\theta(\mathbf{z}) = \frac{\mathbb{E}[g(\mathbf{X})|S=0,\mathbf{z}] - \mathbb{E}[g(\mathbf{X})|Y=0,S=1]}{\mathbb{E}[g(\mathbf{X})|Y=1,S=1] - \mathbb{E}[g(\mathbf{X})|Y=0,S=1]}$$

Proof For every $\mathbf{z} \in \mathbb{R}^{d_z}$, Let $f(\mathbf{x})$ denote the density of $\mathbf{X}$. Note that

$$\begin{aligned} g(\mathbf{x})f(\mathbf{x}|S=0,\mathbf{z}) &= \sum_{j=0}^{1} g(\mathbf{x})f(\mathbf{x}|Y=j,S=0,\mathbf{z})\mathbb{P}(Y=j|S=0,\mathbf{z}) && \text{Law of total prob.} \\ \mathbb{E}[g(\mathbf{X})|S=0,\mathbf{z}] &= \sum_{j=0}^{1} \mathbb{E}[g(\mathbf{X})|Y=j,S=0,\mathbf{z}]\mathbb{P}(Y=j|S=0,\mathbf{z}) && \text{Integration over } \mathbf{x} \\ &= \sum_{j=0}^{1} \mathbb{E}[g(\mathbf{X})|Y=j,S=0]\mathbb{P}(Y=j|S=0,\mathbf{z}) && \text{Assumption 6} \\ &= \sum_{j=0}^{1} \mathbb{E}[g(\mathbf{X})|Y=j,S=1]\mathbb{P}(Y=j|S=0,\mathbf{z}) && \text{Assumption 3} \end{aligned} \tag{14}$$

Isolating $\mathbb{P}(Y=1|S=0,\mathbf{z})$ in equation 14 yields

$$\theta(\mathbf{z}) := \mathbb{P}(Y=1|S=0,\mathbf{z}) = \frac{\mathbb{E}[g(\mathbf{X})|S=0,\mathbf{z}] - \mathbb{E}[g(\mathbf{X})|Y=0,S=1]}{\mathbb{E}[g(\mathbf{X})|Y=1,S=1] - \mathbb{E}[g(\mathbf{X})|Y=0,S=1]}$$

■

Proof [Theorem 17] The main difference between this proof and the one of Theorem 7 is that $\widehat{\mathbb{E}}[g(\mathbf{X})|S=0,\mathbf{z}]$ is usually biased for $\mathbb{E}[g(\mathbf{X})|S=0,\mathbf{z}]$. The proof strategy consists of isolating this bias term from the squared error and then replicating steps which are similar to the ones in Theorem 7.

In order to present the proof in a compact form, some special notation is used. Specifically, $h_0(\mathbf{z}) = \mathbb{E}[g(\mathbf{X})|S=0,\mathbf{z}]$, $\hat{h}_0(\mathbf{z}) = \widehat{\mathbb{E}}[g(\mathbf{X})|S=0,\mathbf{z}]$, $h_{1,i} = \mathbb{E}[g(\mathbf{X})|S=1,Y=i]$, and $\hat{h}_{1,i} = \widehat{\mathbb{E}}[g(\mathbf{X})|S=1,Y=i]$. Using this notation and Definition 16, note that $\hat{\theta}_{RR}(\mathbf{Z}) = \max\left(0, \min\left(1, \frac{\hat{h}_0(\mathbf{z})-\hat{h}_{1,0}}{\hat{h}_{1,1}-\hat{h}_{1,0}}\right)\right)$. Therefore,

$$
\begin{aligned}
&\mathbb{E}\left[\left(\widehat{\theta}_{RR}(\mathbf{Z}) - \theta(\mathbf{Z})\right)^2 \middle| S_1^n\right] \\
=&\mathbb{E}\left[\left(\widehat{\theta}_{RR}(\mathbf{Z}) - \frac{h_0(\mathbf{Z}) - h_{1,0}}{h_{1,1} - h_{1,0}}\right)^2 \middle| S_1^n\right] && \text{Lemma 21} \\
\leq&\mathcal{O}\left(\mathbb{E}\left[\left(\widehat{\theta}_{RR}(\mathbf{Z}) - \frac{\mathbb{E}[\hat{h}_0(\mathbf{Z})|S_1^n] - h_{1,0}}{h_{1,1} - h_{1,0}}\right)^2 + \left(\frac{\mathbb{E}[\hat{h}_0(\mathbf{Z})|S_1^n] - h_{1,0}}{h_{1,1} - h_{1,0}} - \frac{h_0(\mathbf{Z}) - h_{1,0}}{h_{1,1} - h_{1,0}}\right)^2 \middle| S_1^n\right]\right) \\
\leq&\mathcal{O}\left(\mathbb{V}[\hat{h}_{1,0}|S_1^n] + \mathbb{V}[\hat{h}_{1,1}|S_1^n] + \mathbb{V}[\hat{h}_0(\mathbf{Z})|S_1^n] + \left(\mathbb{E}[\hat{h}_0(\mathbf{Z})|S_1^n] - h_0(\mathbf{Z})\right)^2\right) && \text{Lemma 20} \\
=&\mathcal{O}\left(\max\left(\mathbb{V}[\hat{h}_{1,0}|S_1^n], \mathbb{V}[\hat{h}_{1,1}|S_1^n], \mathbb{E}\left[(\hat{h}_0(\mathbf{Z}) - h_0(\mathbf{z}))^2|S_1^n\right]\right)\right) \\
=&\mathcal{O}\left(\max\left(n_L^{-1}, \mathbb{E}\left[(\hat{h}_0(\mathbf{Z}) - h_0(\mathbf{z}))^2|S_1^n\right]\right)\right)
\end{aligned}
$$

The last equality follows from observing that $\hat{h}_{1,0}$ and $\hat{h}_{1,1}$ are sample averages. ■

Appendix B. Additional Figures

These are additional figures to the experiments of Sections 2.5 and 3.1.

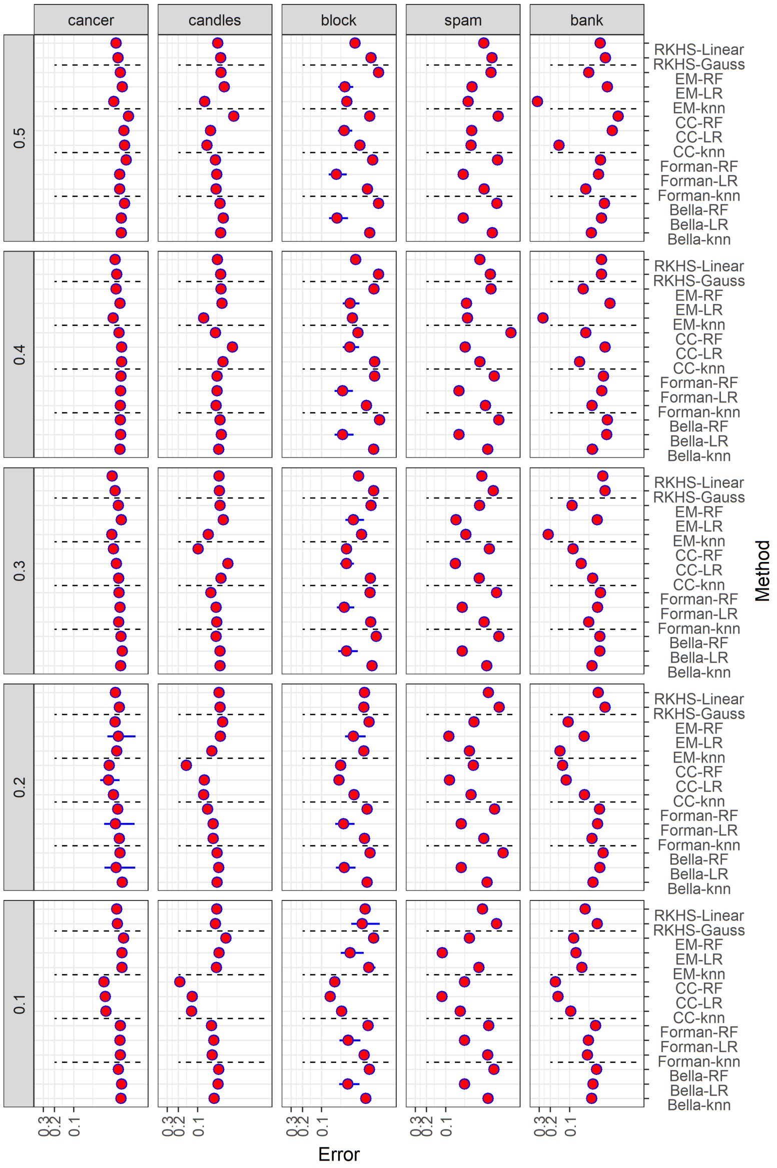

Figure 9: Root mean square deviation of each method by setting in logarithmic scale (including K-NN estimator).

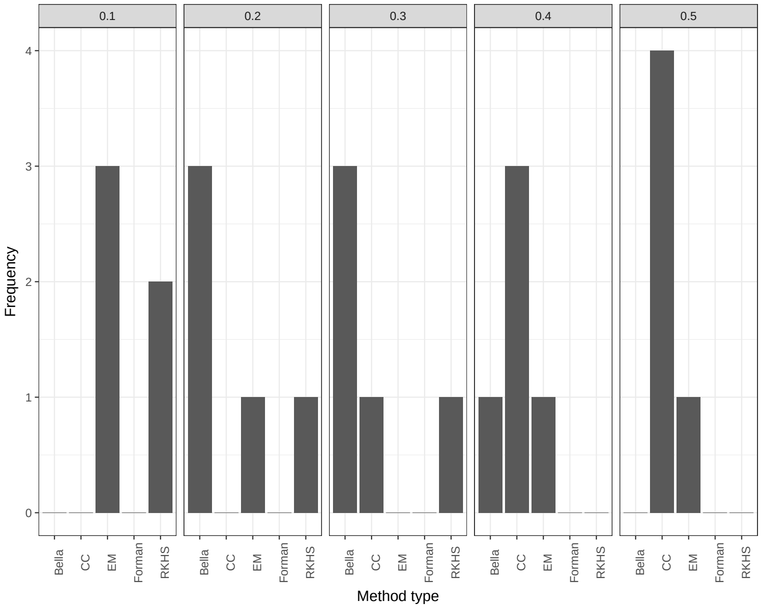

Figure 10: Number of times in which each specific method presents smaller MSE by θ values.

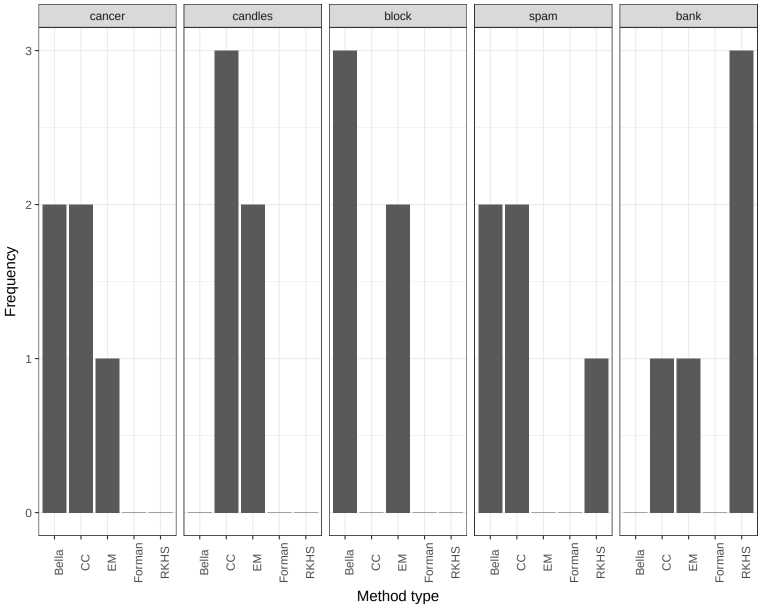

Figure 11: Number of times in which each specific method presents smaller MSE by data set.

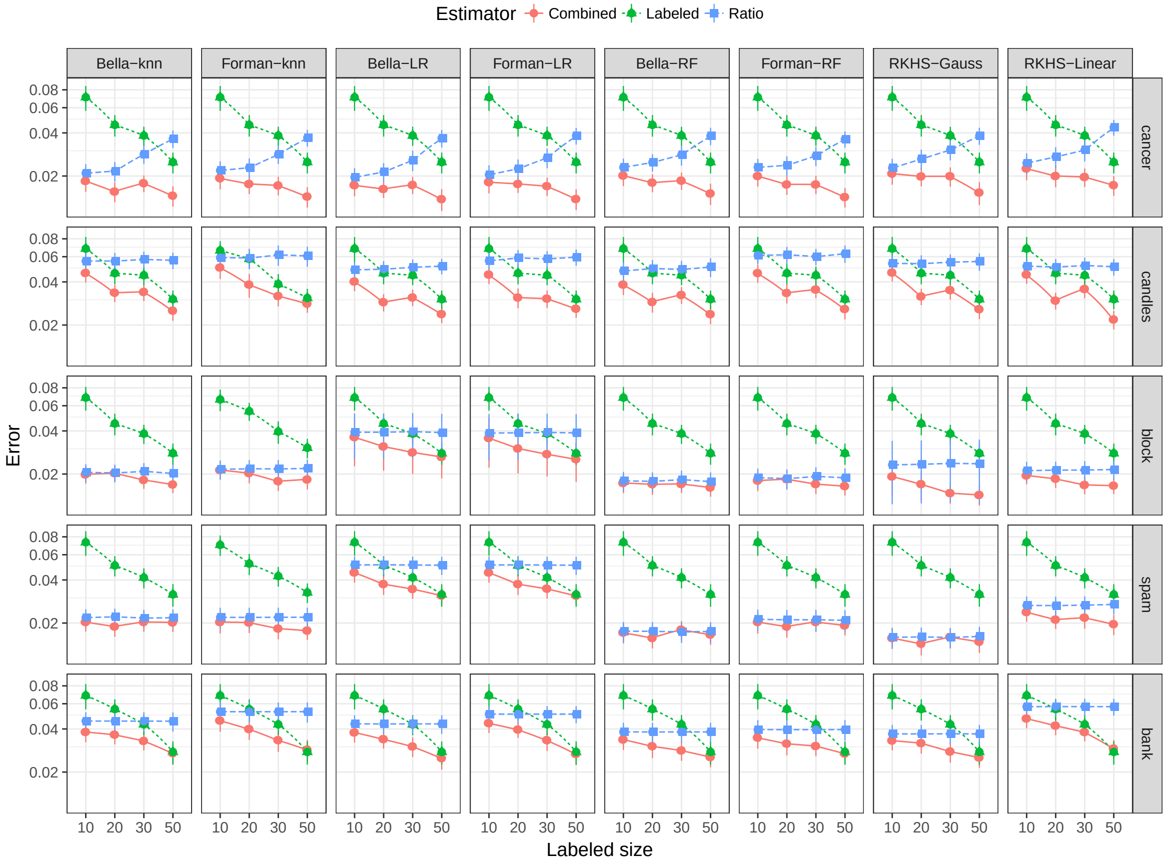

Figure 12: Root mean square deviation in logarithmic scale for each data set, method and estimator by size of the labeled sample and using $\theta = 0.1$.

References

J. Barranquero, J. Díez, and J. J. del Coz. Quantification-oriented learning based on reliable classifiers. *Pattern Recognition*, 48(2):591–604, 2015.

A. Bella, C. Ferri, J. Hernández-Orallo, and M. J. Ramirez-Quintana. Quantification via probability estimators. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 737–742. IEEE, 2010.

C. Blake. Uci repository of machine learning databases. *http://www. ics. uci. edu/˜ mlearn/MLRepository. html*, 1998.

G. Casella and R. L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

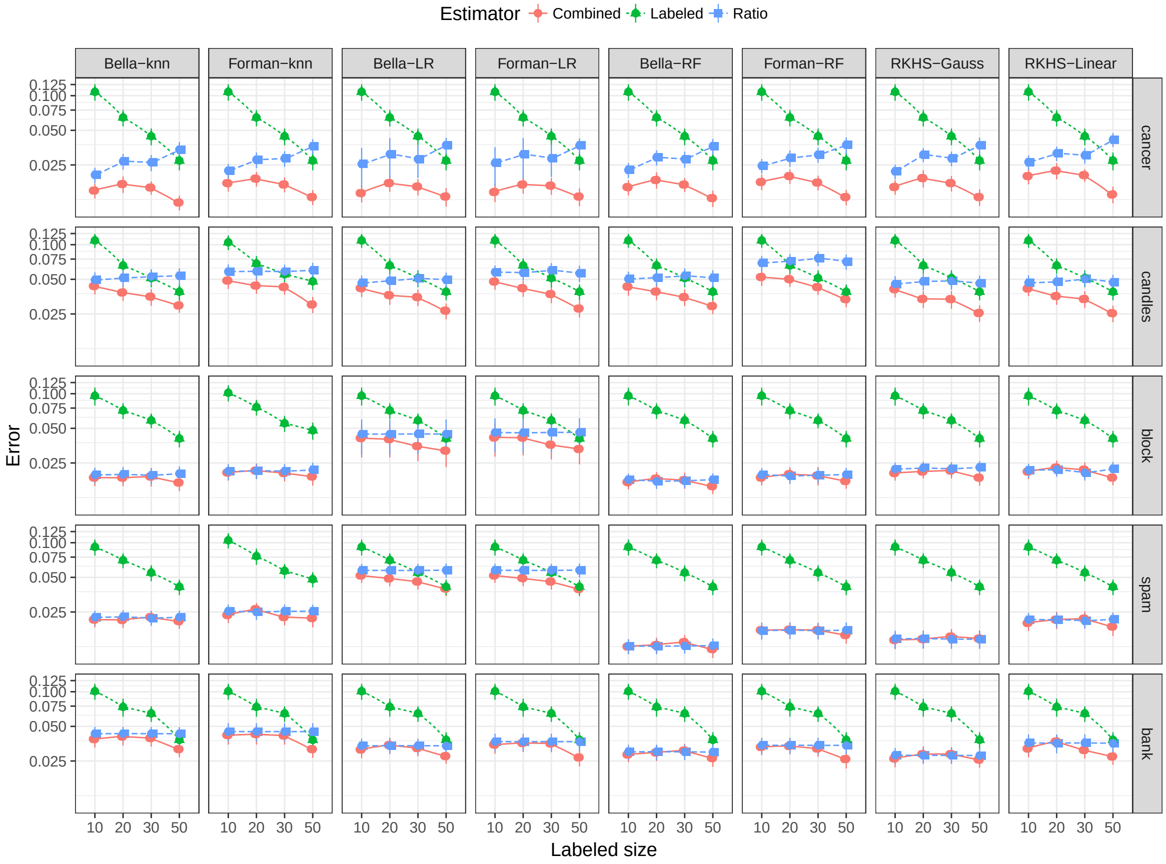

Figure 13: Root mean square deviation in logarithmic scale for each data set, method and estimator by size of the labeled sample and using $\theta = 0.2$.

B. de Finetti. *Theory of probability: a critical introductory treatment*, volume 6. John Wiley & Sons, 2017.

L. G. Esteves, R. Izbicki, and R. B. Stern. Teaching decision theory proof strategies using a crowdsourcing problem. *The American Statistician*, 71(4):336–343, 2017.

T. Fawcett and P. A. Flach. A response to Webb and Ting's *On the application of ROC analysis to predict classification performance under varying class distributions*. *Machine Learning*, 58(1):33–38, 2005.

G. Forman. Quantifying trends accurately despite classifier error and class imbalance. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 157–166. ACM, 2006.

G. Forman. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206, 2008.

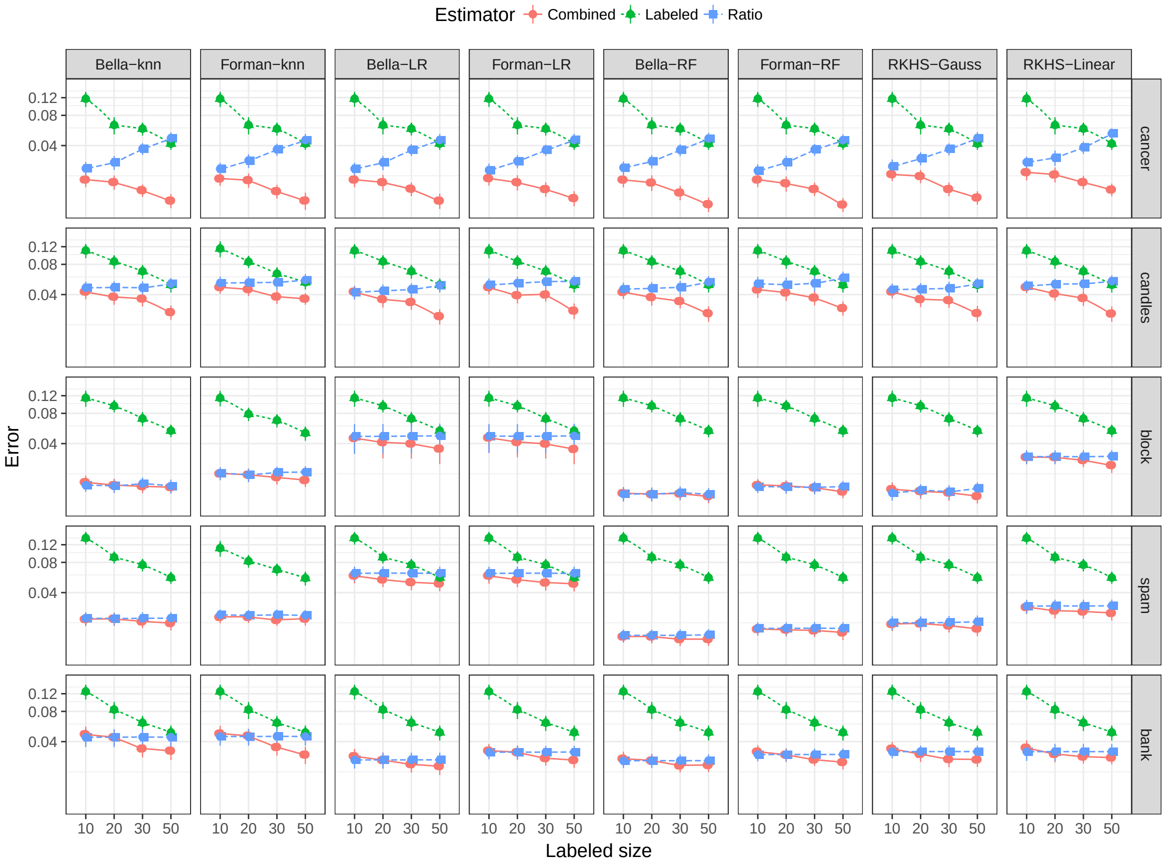

Figure 14: Root mean square deviation in logarithmic scale for each data set, method and estimator by size of the labeled sample and using $\theta = 0.4$.

PE Freeman, R Izbicki, AB Lee, JA Newman, CJ Conselice, AM Koekemoer, JM Lotz, and M Mozena. New image statistics for detecting disturbed galaxy morphologies at high redshift. *Monthly Notices of the Royal Astronomical Society*, 434(1):282–295, 2013.

J. J. Gart and A. A. Buck. Comparison of a screening test and a reference test in epidemiologic studies. ii. a probabilistic model for the comparison of diagnostic tests. *American Journal of Epidemiology*, 83(3):593–602, 1966.

K. Gerow. Fisher consistency-the evolution of a concept: It's hard to get it right the first time, 1989.

V. Hofer and G. Krempl. Drift mining in data: A framework for addressing drift in classification. *Computational Statistics & Data Analysis*, 57(1):377–391, 2013.

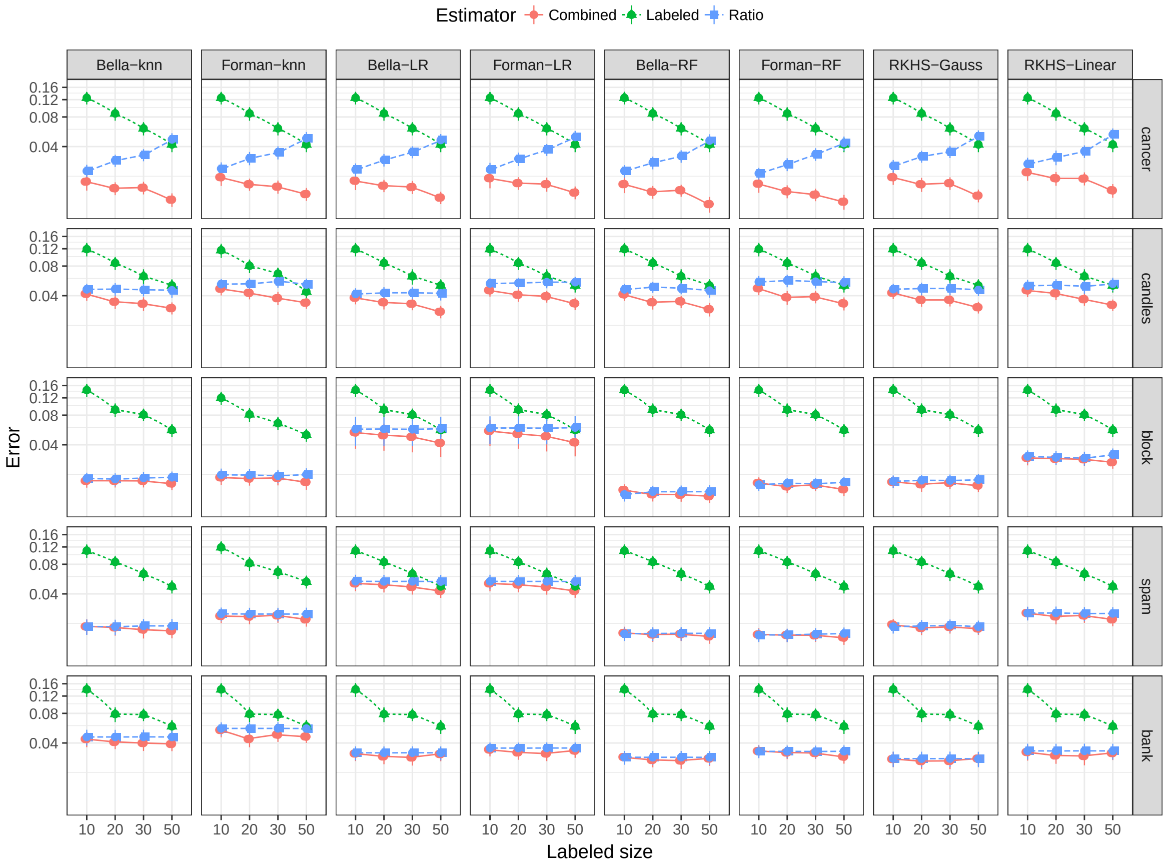

Figure 15: Root mean square deviation in logarithmic scale for each data set, method and estimator by size of the labeled sample and using $\theta = 0.5$.

R. Izbicki and R. B. Stern. Learning with many experts: model selection and sparsity. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(6):565–577, 2013.

R. E. Kass and P. W. Vos. *Geometrical foundations of asymptotic inference*, volume 908. John Wiley & Sons, 2011.

Z. C. Lipton, Y. X. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. *International Conference on Machine Learning (ICML)*, 2018.

D. Malerba, F. Esposito, and G. Semeraro. A further comparison of simplification methods for decision-tree induction. In *Learning from data*, pages 365–374. Springer, 1996.

O. L. Mangasarian. Cancer diagnosis via linear programming. *SIAM news*, 23(5):18, 1990.

C. Michelot. A finite algorithm for finding the projection of a point onto the canonical simplex of n. *Journal of Optimization Theory and Applications*, 50(1):195–200, 1986.

J. G. Moreno-Torres, T. Raeder, R. Alaiz-RodríGuez, N. V. Chawla, and F. Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.

S. Moro, R. Laureano, and P. Cortez. Using data mining for bank direct marketing: An application of the crisp-dm methodology. In *Proceedings of European Simulation and Modelling Conference-ESM'2011*, pages 117–121. Eurosis, 2011.

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964.

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

D. Tasche. Does quantification without adjustments work? *arXiv preprint arXiv:1602.08780*, 2016.

D. Tasche. Fisher consistency for prior probability shift. *Journal of Machine Learning Research*, 18(95):1–32, 2017. URL `http://jmlr.org/papers/v18/17-048.html`.

A. Vaz, R. Izbicki, and R. B. Stern. Prior shift using the ratio estimator. In *International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, pages 25–35. Springer, 2017.

G. Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan. A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, pages 115–120, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

L. Wasserman. *All of nonparametric statistics.* Springer, 2006.

L. Wasserman. *All of statistics: a concise course in statistical inference.* Springer Science & Business Media, 2013.