UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA
UFSCar-USP

Improved quantification under dataset shift

Afonso Fernandes Vaz

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA
UFSCar-USP

Afonso Fernandes Vaz

Quantificação em problemas com mudança de domínio

Dissertação apresentada ao Departamento de Estatística – Des/UFSCar e ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Estatística - Programa Interinstitucional de Pós-Graduação em Estatística UFSCar-USP.

Orientador: Prof. Dr. Rafael Izbicki

São Carlos
Julho de 2018

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA
UFSCar-USP

Afonso Fernandes Vaz

Improved quantification under dataset shift

Master dissertation submitted to the Departamento de Estatística – Des/UFSCar and to Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Joint Graduate Program in Statistics - UFSCar-USP.

Advisor: Prof. Dr. Rafael Izbicki

São Carlos
July 2018

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Afonso Fernandes Vaz, realizada em 17/05/2018:

Prof. Dr. Rafael Izbicki
UFSCar

Prof. Dr. Caio Lucidius Naberezny Azevedo
UNICAMP

Prof. Dr. Francys Andrews de Souza
UNICAMP

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Caio Lucidius Naberezny Azevedo, Francys Andrews de Souza e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Prof. Dr. Rafael Izbicki

This work is dedicated to my family and friends who have unconditionally supported me in its development.

ACKNOWLEDGEMENTS

I would first like to thank my advisor and friend Dr. Rafael Izbicki. You supported me greatly and were always willing to help me.

I would also like to thank my parents and girlfriend for their wise counsel and affection. You are always there for me.

Finally, I would like to thank the CAPES - Coordination for the Improvement of Higher Education Personnel - for their financial support and CEMEAI - Center for Mathematical Sciences Applied to Industry - for their computational support.

“Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.”

(George Box)

ABSTRACT

VAZ, A. F. **Improved quantification under dataset shift**. 2018. 49 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Several machine learning applications use classifiers as a way of quantifying the prevalence of positive class labels in a target dataset, a task named quantification. For instance, a naive way of determining what proportion of positive reviews about given product in the Facebook with no labeled reviews is to (i) train a classifier based on Google Shopping reviews to predict whether a user likes a product given its review, and then (ii) apply this classifier to Facebook posts about that product. Unfortunately, it is well known that such a two-step approach, named Classify and Count, fails because of data set shift, and thus several improvements have been recently proposed under an assumption named prior shift. However, these methods only explore the relationship between the covariates and the response via classifiers and none of them take advantage of the fact that one often has access to a few labeled samples in the target set. Moreover, the literature lacks in approaches that can handle a target population that varies with another covariate; for instance: How to accurately estimate how the proportion of new posts or new webpages in favor of a political candidate varies in time? We propose novel methods that fill these important gaps and compare them using both real and artificial datasets. Finally, we provide a theoretical analysis of the methods.

Keywords: Quantification, Data set shift, Prior shift, Machine Learning.

RESUMO

VAZ, A. F. **Quantificação em problemas com mudança de domínio**. 2018. 49 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Muitas aplicações de aprendizado de máquina usam classificadores para determinar a prevalência da classe positiva em um conjunto de dados de interesse, uma tarefa denominada quantificação. Por exemplo, uma maneira ingênua de determinar qual a proporção de postagens positivas sobre um determinado protuto no Facebook sem ter resenhas rotuladas é (i) treinar um classificador baseado em resenhas do Google Shopping para prever se um usuário gosta de um produto qualquer, e então (ii) aplicar esse classificador às postagens do Facebook relacionados ao produtos de interesse. Infelizmente, é sabido que essa técnica de dois passos, denominada classificar e contar, falha por não levar em conta a mudança de domínio. Assim, várias melhorias vêm sendo feitas recentemente sob uma suposição denominada prior shift. Entretanto, estes métodos exploram a relação entre as covariáveis apenas via classificadores e nenhum deles aproveitam o fato de que, em algumas situações, podemos rotular algumas amostras do conjunto de dados de interesse. Além disso, a literatura carece de abordagens que possam lidar com uma população-alvo que varia com outra covariável; por exemplo: Como estimar precisamente como a proporção de novas postagens ou páginas web a favor de um candidato político varia com o tempo? Nós propomos novos métodos que preenchem essas lacunas importantes e os comparamos utilizando conjuntos de dados reais e similados. Finalmente, nós fornecemos uma análise teórica dos métodos propostos.

Palavras-chave: Quantificação, Mudança de domínio, Aprendizado de máquina.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

CHAPTER

1

INTRODUCTION

In several statistical learning applications, predicting the labels[1] of individual observations *per se*, based on its features[2], is less important than evaluating the proportion of each class labels on an unlabeled dataset[3]. The latter task called is called quantification (FORMAN, 2008). In the following we present an example which will be used throughout this work. Consider that Company A is interested in evaluating the proportion of positive reviews in a social network like Facebook about each of their products in order to improve its control system of customer satisfaction. In order to achieve this, the company invests in a technology to automatically collect Facebook reviews about its products. A possible strategy to analyze these data is read each review and manually label it. Unfortunately, this process is very expansive and time consuming, and hence unpractical. Therefore, this company needs to deploy a tool which yields an accurate estimation of the proportion of positive reviews using the unlabeled data, i.e., without manually labeling the dataset. That is, it needs a good quantification tool.

In this setting, is impossible use any algorithm of supervised leaning, once we do not have labels. Because of this, a common strategy in quantification problems is to take a labeled dataset which has similar properties to the dataset of interest, and then use it to learn something about the relationship between labels and features. For example, consider again Company A's problem. Suppose that it is in a Benchmark[4] program with the Company B which develop the same type of product than A. After some meetings, Company A discovers that B has a large amount of labeled reviews about its own products. Although they are different products, Company A decides to use these labeled data to learn something about the reviews it is interested on.

In this setting, a common and intuitive approach to solve the quantification problem is

[1] Essentially, a label is the value of a categorical variable. For example, if an e-mail is or is not a spam.
[2] A feature is a vector of covariates related to the label. For example, the text of an e-mail.
[3] An unlabeled dataset is a dataset in which we do not have access to labels, but only to features. For example, a dataset in which we have the e-mail texts but we do not know which ones are spam or not.
[4] Benchmark is a quality tool where two or more companies shares informations in order to improve its products and process.

to (i) adjust a classifier for the reviews evaluation based on labeled reviews from Company B, and (ii) apply this classifier to the unlabeled dataset from Company A and use the proportion of reviews which are classified as positive as an estimate of the quantity of interest. However, it is known that this two-step approach, known as “classify and count" (FORMAN, 2008) may lead to inconsistent results, especially when the proportion of positive reviews changes substantially from Company A to B. This occurs because the basic assumption to apply standard classification techniques is that the distribution of the labeled dataset is the same as the distribution of the dataset in which the classifier will be applied to (i.e., the i.i.d. assumption) (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). In particular, in the quantification setting, this assumption is almost never realistic. Consider our example: even if companies A and B have similar products, these products are developed by different process. Therefore, it is natural the customer satisfaction to be different and consequently the proportion of positive reviews will be different.

This difference between the probability distribution of labeled/unlabeled data is known as *dataset shift* (or *dataset drift*, or *domain shift*) (FORMAN, 2006; TASCHE, 2016). In order to be able to learn something about Company A reviews, some assumption about the relationship between datasets A and B must be made. A common assumption used in quantification problems which relaxes the i.i.d. assumption is the prior probability shift assumption. Several quantification methods have been developed under such assumption (SAERENS; LATINNE; DECAESTECKER, 2002; FORMAN, 2008; BELLA *et al.*, 2010; BARRANQUERO; DÍEZ; COZ, 2015), which states that although the proportion of positive labels can change over the datasets, the label conditional distribution of the features are the same.

A particular estimator that successfully performs quantification under prior shift assumption is the adjusted count (AC) estimator (GART; BUCK, 1966; SAERENS; LATINNE; DECAESTECKER, 2002; FORMAN, 2008). Part of the success of the AC estimator is explained in Tasche (2017), namely that it is Fisher consistent.

This work introduces and explores the properties of a generalization of the AC estimator, which we call the *ratio estimator*. We derive convergence rates for the ratio estimator that lead to further insights on why the AC estimator and also the method from Bella *et al.* (2010) yield good results. These rates also suggest alternative procedures that have good performance, and show how to build confidence intervals for the proportion of interest. To the best of our knowledge, the latter task was still unsolved.

We also generalize our formulation of the prior probability shift problem to a more general setting of the quantification problem which allows for additional covariates. This generalization allows one to answer question such as “how does the proportion of people that like a given product vary with age?” and “how does the proportion of positive tweets about Donald Trump vary with time?” using unlabeled data. We show how a generalization of the ratio estimator is able to answer these questions and can offer improved solutions than standard sentiment analyses (WANG *et al.*, 2012).

CHAPTER

2

QUANTIFICATION UNDER PRIOR PROBABILITY SHIFT

Let $(\mathbf{X}_1, Y_1, S_1), \ldots, (\mathbf{X}_n, Y_n, S_n)$ be a sample such that $\mathbf{X}_i \in \mathbb{R}^d$ are features, $Y_i \in \{0,1\}$ is the label of interest and $S_i \in \{0,1\}$ is the indicator that the i-th sample unit is labeled. 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, $\mathbb{P}(Y = 1|S = 1)$. In the standard formulation of the prior probability shift problem, $\{(\mathbf{X}_i, Y_i)\}_{S_i=0}$ is called the *target population* (since the labels are unavailable), and $\{(\mathbf{X}_i, Y_i)\}_{S_i=1}$ is called the *training population* (TASCHE, 2017). For instance, in the example introduced in Chapter 1 $S_i = 0$ if the the i-th review comes from Company A and $S_i = 1$ if the the i-th review comes from Company B; $Y_i = 0$ if the i-th review is negative and $Y_i = 1$ if the i-th review is positive; $\mathbf{X}_i$ is some vectorial representation of i-th text review; finally, θ is the proportion of positive reviews about the product of Company A.

It is common for both populations to be i.i.d., that is,

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; 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 independent of $(S_1, \ldots, S_n)$ conditionally on $(Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)$.

In our example, Assumption 1 means that the $\mathbf{X}|Y=i$ $(i \in \{0,1\})$ distribution is the same on both $S=0$ and $S=1$. That is, given that a review is positive, the distribution of words used on it is the same for Company A and B.

In the following analysis, some subsets of the data and their sizes are used several times. $A_k := \{i \in \{1,\ldots,n\} : S_i = k\}$ is the set of indexes of the labeled $(k=1)$ or of the unlabeled $(k=0)$ samples, $n_U := |A_0|$, and $n_L := |A_1|$. Also, $A_{k,j} := \{i \in \{1,\ldots,n\} : S_i = k \text{ and } Y_i = j\}$ is the set of indexes of the labeled $(k=1)$ or unlabeled $(k=0)$ samples, and positive label $(j=1)$ or zero label $(j=0)$, and $n_j := |A_{1,j}|$. Also, for every vector, $\mathbf{Z}$, $\mathbf{Z}_i^j := (Z_i,\ldots,Z_j)$.

2.1 The ratio estimator and its theoretical properties

In this section, we introduce our estimator which we call by Ratio estimator and is motived by Lemma 1.

Lemma 1. 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

$$\begin{aligned}
g(\mathbf{x})f(\mathbf{x}|S=0) &= \sum_{j=0}^{1} g(\mathbf{x})f(\mathbf{x}|Y=j,S=0)\mathbb{P}(Y=j|S=0) && \text{Law of total prob.} \\
\mathbb{E}[g(\mathbf{X})|S=0] &= \sum_{j=0}^{1} \mathbb{E}[g(\mathbf{X})|Y=j,S=0]\mathbb{P}(Y=j|S=0) && \text{Integration over } \mathbf{x} \\
&= \sum_{j=0}^{1} \mathbb{E}[g(\mathbf{X})|Y=j,S=1]\mathbb{P}(Y=j|S=0) && \text{Assumption 3} \qquad (2.1)
\end{aligned}$$

Isolating $\mathbb{P}(Y=1|S=0)$ in Equation 2.1 yields

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

□

In words Lemma 1 means that, given a real function of the covariates, the probability of interested θ may be written as the ratio of expected values. Motived by this result, we propose the Ratio estimator which is presented in Definition 1.

Definition 1 (Ratio estimator)**.** Let $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ be an arbitrary function. The ratio estimator for θ based on g, $\widehat{\theta}_R$, is such that

$$\widehat{\theta}_R := \frac{\widehat{\mathbb{E}}[g(\mathbf{X})|S=0] - \widehat{\mathbb{E}}[g(\mathbf{X})|Y=0,S=1]}{\widehat{\mathbb{E}}[g(\mathbf{X})|Y=1,S=1] - \widehat{\mathbb{E}}[g(\mathbf{X})|Y=0,S=1]}, \text{ where}$$
$$\widehat{\mathbb{E}}[g(\mathbf{X})|S=0] = \frac{\sum_{i\in A_0} g(\mathbf{X}_i)}{n_U} \text{ and } \widehat{\mathbb{E}}[g(\mathbf{X})|Y=j,S=1] = \frac{\sum_{i\in A_{1,j}} g(\mathbf{X}_i)}{n_j}$$

Since $\theta \in [0,1]$, the trimmed ratio estimator, $\widehat{\theta}_{TR}$, is such that

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

The ratio estimator generalizes the adjusted count (AC) estimator (GART; BUCK, 1966; SAERENS; LATINNE; DECAESTECKER, 2002; FORMAN, 2008), which consists of a particular case in which $g(\mathbf{x}) \in \{0,1\}$, that is, $g(\mathbf{x})$ is the output of a classifier for Y. Another particular case of the ratio estimator is the estimator introduced by Bella *et al.* (2010), which consists of taking $g(\mathbf{x}) = \widehat{\mathbb{P}}(Y=1|\mathbf{x})$. In this case, $g(\mathbf{x})$ is a soft classifier for Y.

Remark 1. 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 $g_i = \mathbb{E}[g_i(\mathbf{X})|S=0]$, $\widehat{\theta}_R$, is obtained solving the linear system

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

Similarly to the AC estimator (TASCHE, 2017), the ratio estimator is Fisher consistent under weak assumptions. They are described in Assumption 3.

Assumption 3. The function, g, is such that

1. $\mathbb{E}[g(\mathbf{X}_i)|S=0]$ and $\mathbb{E}[g(\mathbf{X}_i)|Y_i=j,S_i=1]$ are defined, for $j \in \{0,1\}$.
2. $\mathbb{E}[g(\mathbf{X}_i)|Y_i=1,S_i=1] - \mathbb{E}[g(\mathbf{X}_i)|Y_i=0,S_i=1] \neq 0$
3. $g(\mathbf{X})_1^n$ is independent of $\mathbf{S}_1^n$ conditionally on $\mathbf{Y}_1^n = y_1^n$.

Assumption 3 requires 3 conditions of $g(\mathbf{x})$. According to condition 1, the populational versions of the expectations in Definition 1 are defined. Condition 2 states that the ratio estimator calculated on these populational quantities is defined, that is, there is no division by 0. Condition 3 is a relaxed type of prior probability shift that is strictly weaker than Assumption 2. In the following, Theorem 1 shows that these conditions guarantee the Fisher consistency of the ratio and the trimmed ratio estimators.

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

Assumption 1 also guarantees a finite population bound on the mean squared error of $\widehat{\theta}_{TR}$. This result is described in Theorem 2.

Theorem 2. Under Assumptions 1 and 3,

$$\mathbb{E}\left[\left(\widehat{\theta}_{TR}-\theta\right)^2 \middle| S_1^n\right] = O(\max(n_L^{-1}, n_U^{-1}))$$

Corollary 1. Under Assumptions 1 and 3, if $n_U \xrightarrow{\mathbb{P}} \infty$ and $n_L \xrightarrow{\mathbb{P}} \infty$, then $\widehat{\theta}_{TR}$ is consistent for θ in probability and in L_2 norm.

It follows from Theorem 2 that, under Assumptions 1 and 3, if $n_U \gg n_L$, then the convergence of the mean squared error of the trimmed 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 standard ratio estimator, since the trimming is necessary to guarantee that the ratio of random variables does not have infinite variance.

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

Assumption 4.

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 3. 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 and 4,

1. If $p_L \neq 0$, then

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

2. If $p_L = 0$, then

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

It is possible to use Theorem 3 in order to obtain an approximate confidence interval for θ. This interval is obtained by inverting the convergence results in Theorem 3, and substituting θ for $\widehat{\theta}_{TR}$ and the populational quantities, μ_0, μ_1, σ_0^2, σ_1^2, p_L, $p_{0|L}$ and $p_{1|L}$, by their respective empirical averages.

Remark 2. This confidence interval may also be used to test hypothesis such as $H_0 : \theta = \theta_0$.

Theorem 3 also provides an approximation for the mean squared error of $\widehat{\theta}_{TR}$. This approximation for the common case in which $n_U \gg n_L$ is presented in the following corollary.

Corollary 2. Under Assumptions 1, 3 and 4, if $p_L = 0$ ($n_U \gg n_L$), then

$$\text{MSE}(\widehat{\theta}_{TR}) \approx \frac{1}{n_L(\mu_1-\mu_0)^2}\left(\frac{\sigma_0^2(1-\theta)^2}{p_{0|L}} + \frac{\sigma_1^2\theta^2}{p_{1|L}}\right) \tag{2.2}$$

Corollary 2 brings some insights on how g should be chosen in order for $\widehat{\theta}_{TR}$ 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 $\mathbb{P}(Y = 1|\mathbf{x})$.

The approximation of the MSE in Corollary 2 can also be used explicitly in the choice of the function, g, in $\widehat{\theta}_{TR}$. Such a procedure is discussed in the following subsection.

2.2 Choosing g via approximate MSE minimization

One possible criterion for the choice of g is the minimization of $MSE(\widehat{\theta}_{TR})$, defined in Corollary 2. However, the latter depends on unobservable quantities. An alternative is to minimize an estimate of $MSE(\widehat{\theta}_{TR})$. This estimate is presented in Definition 2.

Definition 2. Let $\hat{\theta}$ be an estimator of θ and, for each $i \in \{0,1\}$, let

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

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

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

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 $\mathscr{H}_K$ is the RKHS associated to K, then we choose g^* as

$$g^* := \arg\min_{g \in \mathscr{H}_K} \widehat{\mathrm{MSE}}(g) + \lambda ||g||^2_{\mathscr{H}_K} \tag{2.3}$$

In the following, Theorem 4 presents a characterization of the solution to Equation 2.3.

Theorem 4. Let K be a Mercer kernel and $\mathscr{H}_K$ the corresponding RKHS. Also define

- $\mathbb{K}$: 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 $N = \frac{\widehat{\theta}^2}{\widehat{p}_{1|L}} \widehat{\Sigma}_1 + \frac{(1-\widehat{\theta})^2}{\widehat{p}_{0|L}} \widehat{\Sigma}_0$.

For every $\lambda > 0$, if

$$\mathbf{w}^* = \arg\min_{w \in \mathbb{R}^{n_L}} \frac{\mathbf{w}^t N \mathbf{w}}{\mathbf{w}^t M \mathbf{w}} + \lambda \mathbf{w}^t \mathbb{K} \mathbf{w}. \tag{2.4}$$

then $g^*(\mathbf{x}) = \sum_{i \in A_1} w_i^* K(\mathbf{x}, \mathbf{x}_i)$ is such that

$$g^* = \arg\min_{g \in \mathscr{H}_K} \widehat{\mathrm{MSE}}(g) + \lambda ||g||^2_{\mathscr{H}_K}$$

Unfortunately, the minimization of Equation 2.4 is generally not trivial (ZHANG, 2013). A simpler optimization problem is obtained when choosing $\lambda = 0$. In this case, finding w^* in Equation 2.4 is equivalent to finding the vector, $\mathbf{w}^*$, associated to the largest eigenvalue, 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 2 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.

2.3 Extension: combined estimator

In this section we take advantage the fact that sometimes a few $S = 0$ samples can be labeled. For example, Company A may be starting to label its reviews, and hence wants to use these labeled samples to improve the estimate the proportion of positive reviews. The key idea is that we can obtain an estimate of θ using those samples and then combine it with the ratio estimator presented in Equation 2.1. Here, let $A_0^* \subset A_0$ be the subset of samples indexes which were labeled and define the *labeled* estimator as

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

It follows that in order to better estimate θ, one can combine the labeled estimator $\widehat{\theta}_L$ with the ratio estimator estimator $\widehat{\theta}_R$. We do this by using a convex combination of both,

$$\widehat{\theta}_C = w\widehat{\theta}_R + (1-w)\widehat{\theta}_L, \tag{2.5}$$

which we name the *combined* estimator. In order to choose the weight w , observe that, because $\widehat{\theta}_L$ and $\widehat{\theta}_R$ are not correlated,

$$\mathbb{V}[\widehat{\theta}_C] = w^2\mathbb{V}[\widehat{\theta}_R] + (1-w)^2\mathbb{V}[\widehat{\theta}_L],$$

which implies that the minimum variance is obtained by taking $w = \mathbb{V}[\widehat{\theta}_L]/(\mathbb{V}[\widehat{\theta}_L] + \mathbb{V}[\widehat{\theta}_R])$. We therefore use the weight

$$\widehat{w} = \widehat{\mathbb{V}}[\widehat{\theta}_L]/(\widehat{\mathbb{V}}[\widehat{\theta}_L] + \widehat{\mathbb{V}}[\widehat{\theta}_R]) \tag{2.6}$$

Remark 3. We have $\mathbb{V}[\widehat{\theta}_L] = \theta(1-\theta)/|A_0^*|$ and $\mathbb{V}[\widehat{\theta}_R]$ is given in Theorem 3. Therefore, we can estimate theses quantities using the respective empirical averages.

CHAPTER

3

EXTENSION: REGRESSION ESTIMATION

As a generalization of the quantification problem, one might be interested on how the prevalence of Y among the unlabeled data $(S = 0)$ varies according to a new set of covariates, $\mathbf{Z}$. For example, suppose that the Company A is implementing a program of continuous improvement in one of its products. Therefore, in order to measure the effects of its action, it is interested in how the proportion of positive reviews varies over time. Here, $\mathbf{Z}$ is the time which the review was posted. This problem is called sentiment analysis (WANG *et al.*, 2012) and is usually solved using a classify and count approach. This approach can criticized using the same arguments as the ones in Chapter 1.

Now we show how our framework allows the ratio estimator to be used in this regression setting. We assume that our sample is given by $(\mathbf{X}_1, \mathbf{Z}_1, Y_1, S_1), \ldots, (\mathbf{X}_n, \mathbf{Z}_n, Y_n, S_n)$. Again, not all samples are labeled (i.e., have Y available), and S_i indicates whether a sample is labeled or not. Our goal is to estimate

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

the proportion of positive labels at $\mathbf{Z} = \mathbf{z}$ on the unlabeled data $(S = 0)$.

Besides making Assumptions 1 and 3, we will also make two additional assumptions on how the covariates $\mathbf{z}$ related to the other quantities:

Assumption 5.

- $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent conditionally on $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$
- Let $\mathbf{z} \in \mathbb{R}^{d_z}$. Then $\mathbf{X}_1 | \mathbf{Z}_1 = \mathbf{z}, \ldots, \mathbf{X}_n | \mathbf{Z}_n = \mathbf{z}$ are identically distributed.

Assumption 6. [**Conditional Covariate Independence**] $\mathbf{X}$ is independent of $\mathbf{Z}$ conditionally on Y and S.

In our example, Assumption 5 states that the texts of the reviews have the same distribution and are independents at a fix time $\mathbf{z}$. Moreover, Assumption 6 states that, given that a review

is positive (or negative) and belongs to Company A (or B), the texts distribution do not change over the time.

3.1 The regression ratio estimator

We now show how the ratio estimator can be used in the regression context.

Using the same derivation of Section 2.1, one can show that, for every $\mathbf{z} \in \mathbb{R}^{d_z}$,

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

It then follows from Assumption 6 that

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

Moreover, from the prior probability shift assumption (Assumption 2), it holds that

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

This motivates the estimator

$$\widehat{\theta}(\mathbf{z}) = \frac{\widehat{\mathbb{E}}[g(\mathbf{X})|S=0,\mathbf{z}] - \widehat{\mathbb{E}}[g(\mathbf{X})|Y=0,S=1]}{\widehat{\mathbb{E}}[g(\mathbf{X})|Y=1,S=1] - \widehat{\mathbb{E}}[g(\mathbf{X})|Y=0,S=1]}. \tag{3.1}$$

The terms $\mathbb{E}[g(\mathbf{X})|Y=1,S=1]$ and $\mathbb{E}[g(\mathbf{X})|Y=0,S=1]$ may be estimated using the same estimator presented in Definition 1. Moreover, $\mathbb{E}[g(\mathbf{X})|S=0,\mathbf{z}]$ can be estimated using a non-parametric regression method. For instance, if $\mathbf{Z}$ is a continuous random vector, one may use a Nadaraya-Watson kernel estimator:

$$\widehat{\mathbb{E}}[g(\mathbf{X})|s=0,\mathbf{z}] = \sum_{i=1}^{n} w_i(\mathbf{z}) g(\mathbf{X}_i), \tag{3.2}$$

where $w_i(\mathbf{z}) = K(\mathbf{z}_i,\mathbf{z}) / \sum_{k=1}^{n} K(\mathbf{z}_k,\mathbf{z})$ and K is a kernel smoother.

Remark 4. The method above does not require the labeled data to have the same $\mathbf{z}$ distribution as the unlabeled data.

CHAPTER

4

EXPERIMENTS

In order to evaluate and compare the methods presented in the previous sections, some experiments were performed. In this section we present the results found by these experiments.

4.1 Ratio Estimator

First, we compare the ratio estimator using various g's functions. We also evaluate the classify and count estimator when g is a classifier. We use five datasets: Candles Dataset (IZBICKI; STERN, 2013; FREEMAN *et al.*, 2013), Bank Marketing (MORO; LAUREANO; CORTEZ, 2011), SPAM e-mail Database (BLAKE, 1998), Wisconsin Breast Cancer Database (MANGASARIAN, 1990) and Blocks Classification (MALERBA; ESPOSITO; SEMERARO, 1996). For the all datasets we considered $n_L = 300$, $n_0 = 150$ and $n_1 = 150$; that is $\mathbb{P}(Y = 1|S = 1) = 0.5$. The size of unlabeled sample for each dataset is presented in Table 1.

	Dataset	Unlabeled size
1	cancer	100
2	candles	300
3	block	800
4	spam	2000
5	bank	10000

Table 1 – Size of the unlabeled sample for each data set.

For all datasets we consider $\theta \in \{0.1; 0.2; 0.3; 0.4; 0.5\}$, performing 100 repetitions for each one of 11 methods presented in Table 2.

Results are shown in Figure 1 and Table 3. In figure, we show the average of the squared error (horizontal and red bar) and its respective confidence interval (vertical and blue bar). Moreover, the average of the squared error of each method and their standard deviations are shown in the table.

Estimator type	Prior shift estimator	$g(x)$ estimator
CC	Classify and count (CC)	Logistic regression (LR), k-NN, random forest (RF).
Ratio	Forman	Logistic regression (LR), k-NN, random forest (RF).
Ratio	Bella	Logistic regression (LR), k-NN, random forest (RF).
Ratio	RKHS	Linear Kernel (Linear), Gaussian Kernel (Gauss).

Table 2 – Methods compared in the experiments.

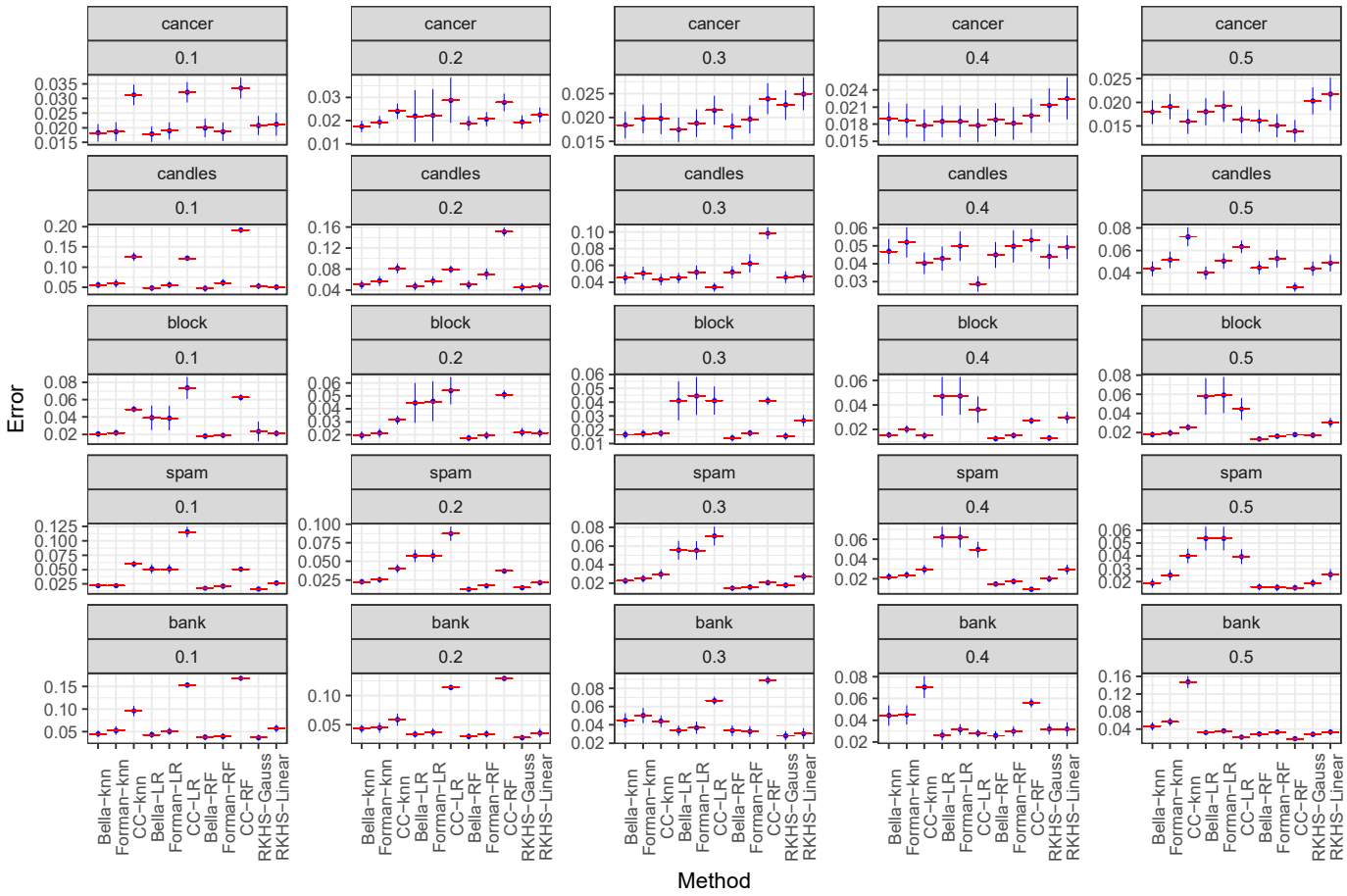

Figure 1 – Average and standard deviation of the mean squared error in each setting.

In summary, we can observe that:

- The ratio estimator lead to good results when compared with classify and count approach. Note that this does not occur in the setting where $\theta \approx 0.5$. In these cases, $P(Y = 1|S = 0) \approx P(Y = 1|S = 1)$, that is, there is no prior shift.

- The function g proposed by Bella *et al.* (2010) had a better performance than g proposed by Forman (2008) in essentially all settings. This means that, in these settings, it is better to use a probability instead of a hard-classification.

- The RKHS approach for choosing g is a competitive method. It had good results using a Gaussian Kernel particularly in the Block dataset.

Table 3 – Average and standard deviation (SD) of the mean squared error in each setting. Bold values indicate the best method in each setting.

Method		Bella-KNN		Forman-KNN		CC-KNN		Bella-LR		Forman-LR		CC-LR		Bella-RF		Forman-RF		CC-RF		RKHS - Gauss		RKHS - Linear	
Dataset	θ	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
Cancer	0.1	0.0183	0.01439	0.01860	0.01577	0.03120	0.01577	**0.0178**	0.01304	0.01890	0.01439	0.03210	0.01439	0.0199	0.01526	0.0186	0.01540	0.03350	0.01540	0.0207	0.01580	0.02120	0.01895
	0.2	**0.0174**	0.01187	0.01930	0.01296	0.02390	0.01296	0.0219	0.05488	0.02220	0.05519	0.02870	0.05519	0.0188	0.01420	0.0206	0.01458	0.02780	0.01458	0.0193	0.01388	0.02240	0.01542
	0.3	0.0184	0.01372	0.01970	0.01466	0.01980	0.01466	**0.0174**	0.01252	0.01880	0.01409	0.02150	0.01409	0.0181	0.01331	0.0196	0.01436	0.02390	0.01436	0.0227	0.01517	0.02490	0.01668
	0.4	0.0189	0.01399	0.01970	0.01440	**0.0178**	0.01440	0.0185	0.01359	0.01850	0.01359	0.01780	0.01359	0.0188	0.01430	0.0182	0.01411	0.01950	0.01411	0.0213	0.01445	0.02250	0.01790
	0.5	0.0179	0.01226	0.01910	0.01310	0.01600	0.01310	0.018	0.01378	0.01920	0.01605	0.01640	0.01605	0.0161	0.01120	0.0151	0.01180	**0.0139**	0.01180	0.0203	0.01388	0.02180	0.01676
Candles	0.1	0.0557	0.03754	0.05930	0.04639	0.12570	0.04639	0.0481	0.03297	0.05580	0.03862	0.12180	0.03862	**0.0471**	0.03647	0.0611	0.04030	0.19150	0.04030	0.0531	0.03644	0.05040	0.03420
	0.2	0.0498	0.03935	0.05720	0.04392	0.08120	0.04392	0.0472	0.03695	0.05720	0.04124	0.07920	0.04124	0.0499	0.03991	0.0705	0.05140	0.15090	0.05140	**0.0449**	0.03759	0.04640	0.03495
	0.3	0.0451	0.03341	0.05060	0.03798	0.04310	0.03798	0.045	0.02906	0.05180	0.04001	**0.0339**	0.04001	0.0519	0.03597	0.0623	0.05269	0.09860	0.05269	0.0459	0.03352	0.04670	0.03385
	0.4	0.0469	0.03382	0.05190	0.04124	0.04020	0.04124	0.0428	0.03268	0.04980	0.04018	**0.0286**	0.04018	0.0449	0.03528	0.0496	0.04376	0.05310	0.04376	0.0441	0.03341	0.04910	0.03205
	0.5	0.0436	0.03179	0.05170	0.03564	0.07220	0.03564	0.04	0.02747	0.05050	0.03321	0.06330	0.03321	0.0447	0.02909	0.0528	0.03792	**0.0274**	0.03792	0.0439	0.02989	0.04880	0.03621
Block	0.1	0.0205	0.01454	0.02170	0.01656	0.04890	0.01656	0.0392	0.06901	0.03860	0.06911	0.07350	0.06911	**0.018**	0.01409	0.0188	0.01517	0.06240	0.01517	0.0233	0.05519	0.02100	0.01604
	0.2	0.0194	0.01542	0.02140	0.01642	0.03140	0.01642	0.0445	0.07573	0.04570	0.07565	0.05410	0.07565	**0.0175**	0.01171	0.0194	0.01406	0.05090	0.01406	0.0219	0.01531	0.02140	0.01511
	0.3	0.0164	0.01297	0.01710	0.01310	0.01730	0.01310	0.041	0.06930	0.04450	0.06757	0.04110	0.06757	**0.0141**	0.01119	0.0175	0.01162	0.04100	0.01162	0.0154	0.01213	0.02660	0.01949
	0.4	0.0154	0.01136	0.02000	0.01478	0.01480	0.01478	0.0472	0.07778	0.04740	0.07618	0.03620	0.07618	**0.0124**	0.00991	0.0149	0.01036	0.02710	0.01036	0.0129	0.01057	0.02960	0.02275
	0.5	0.0178	0.01299	0.01930	0.01527	0.02530	0.01527	0.0577	0.09396	0.05930	0.09286	0.04450	0.09286	**0.0129**	0.00974	0.0159	0.01083	0.01770	0.01083	0.0169	0.01224	0.03020	0.02350
Spam	0.1	0.0219	0.01565	0.02190	0.01839	0.05960	0.01839	0.0509	0.03701	0.05090	0.03697	0.11540	0.03697	0.0175	0.01419	0.0211	0.01824	0.05070	0.01824	**0.016**	0.01314	0.02660	0.02041
	0.2	0.0224	0.01578	0.02530	0.01824	0.04020	0.01824	0.0574	0.04058	0.05750	0.04040	0.08730	0.04040	**0.0126**	0.00943	0.0172	0.01420	0.03690	0.01420	0.0146	0.01286	0.02140	0.01560
	0.3	0.0227	0.01633	0.02500	0.01938	0.02980	0.01938	0.0555	0.04810	0.05520	0.04823	0.07070	0.04823	**0.0147**	0.01279	0.016	0.01186	0.02080	0.01186	0.0179	0.01291	0.02730	0.02045
	0.4	0.0221	0.01624	0.02390	0.01575	0.02940	0.01575	0.0622	0.05119	0.06210	0.05150	0.04950	0.05150	0.0148	0.01070	0.0175	0.01318	**0.0096**	0.01318	0.02	0.01592	0.02930	0.02292
	0.5	0.0187	0.01656	0.02500	0.01973	0.04000	0.01973	0.0536	0.04494	0.05360	0.04501	0.03940	0.04501	0.0158	0.01246	0.0154	0.01178	**0.0152**	0.01178	0.0188	0.01416	0.02540	0.02058
Bank	0.1	0.0454	0.03283	0.05280	0.03990	0.09610	0.03990	0.0434	0.03044	0.05080	0.03274	0.15300	0.03274	0.0381	0.02842	0.0396	0.03057	0.16830	0.03057	**0.037**	0.02698	0.05730	0.03627
	0.2	0.0432	0.02941	0.04510	0.03885	0.05870	0.03885	0.0339	0.02442	0.03680	0.03001	0.11380	0.03001	0.03	0.02147	0.0342	0.02615	0.12900	0.02615	**0.028**	0.02071	0.03590	0.03147
	0.3	0.0449	0.03536	0.05020	0.03739	0.04400	0.03739	0.0337	0.02490	0.03670	0.02961	0.06650	0.02961	0.0336	0.02555	0.0328	0.02600	0.08880	0.02600	**0.028**	0.02256	0.03040	0.02540
	0.4	0.0443	0.04200	0.04500	0.03990	0.07040	0.03990	0.0263	0.02190	0.03140	0.02252	0.02800	0.02252	**0.0258**	0.01860	0.0298	0.02005	0.05590	0.02005	0.0318	0.02185	0.03180	0.02860
	0.5	0.0461	0.03938	0.05620	0.04024	0.14730	0.04024	0.0318	0.02384	0.03550	0.02293	0.02160	0.02293	0.02860	0.02138	0.0329	0.02243	**0.0176**	0.02243	0.0277	0.02301	0.03320	0.02845

We also investigate the coverage of the confidence interval presented in Equation 2.1. The results are shown in Figure 2. Here, we consider all methods together. We can observe that the coverage is close to 0.95. Moreover, in general, the observed coverage is greater than the theoretical confidence level. That is, generally we overestimate the estimator variance making our interval more conservative.

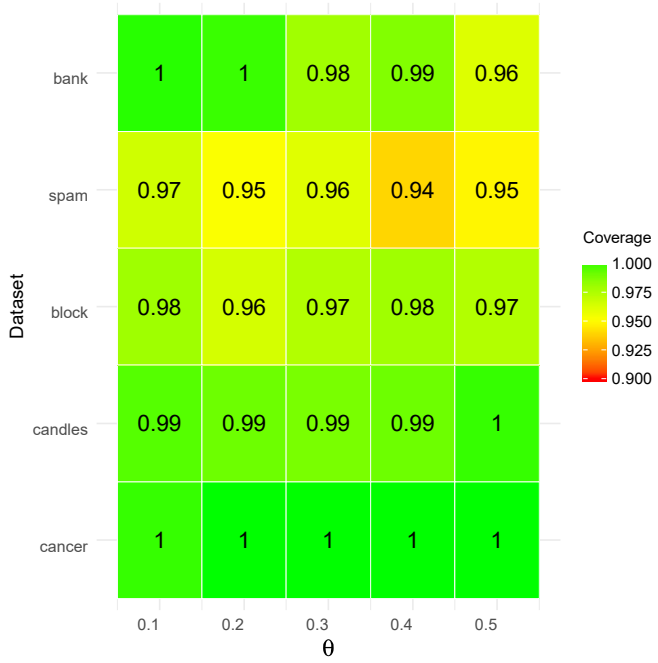

Figure 2 – Empirical coverage of the confidence interval.

4.2 Combined estimator

In order to evaluate the combined estimator developed in Section 2.3, we performed an experiment under the same settings as that of Section 4.1. For each repetition we to label 10, 20, 30 and 50 samples. Figure 3 presents the squared errors for each setting. We consider all the θ's values ($\theta \in \{0.1; 0.2; 0.3; 0.4; 0.5\}$) together. We can notice that the combined estimator lead to good results when compare to labeled and ratio estimators. Even when one of these methods is much better than the other, the combined estimator essentially assumes the same behavior as it. This fact may be observed in the block dataset when the $g(x)$ is based on random forest method (RF). These results also indicate that labeling some samples is a good way to improve the quantification results.

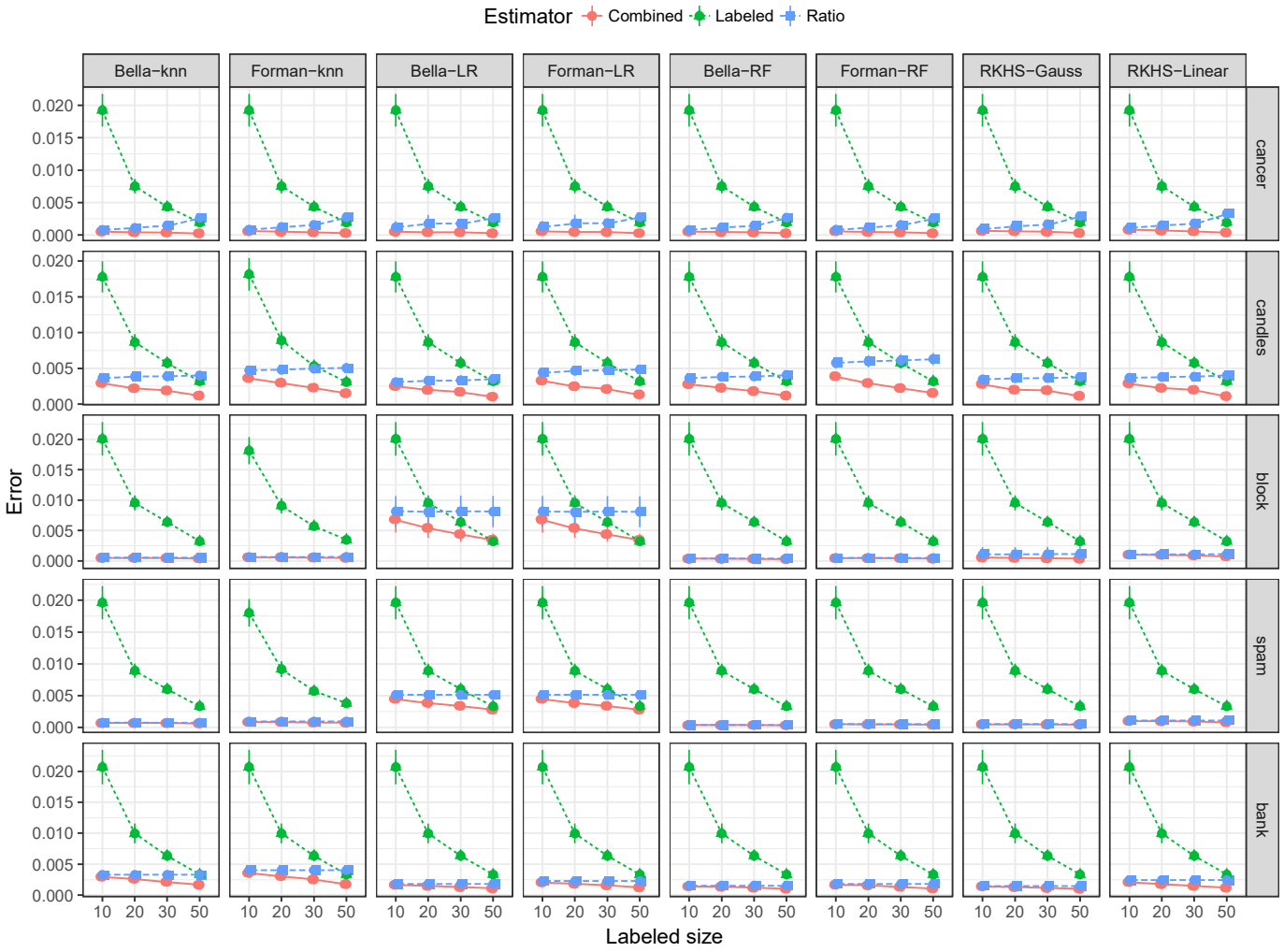

Figure 3 – Empirical error for each setting and each size of labeled sample.

Figure 4 shows the behavior of the weight w under each setting. We can observe that w decreases as the labeled size increases (as expected). Moreover, when $\theta = 0.1$, w decreases more fast. This behavior may be related to the fact that in this setting the labeled estimator has less variance. The opposite can be seen as theta approximate to 0.5. In almost all cases, w is greater than 0.5. This means that the estimated variability of the ratio estimator is less than the estimated variability of the labeled estimator.

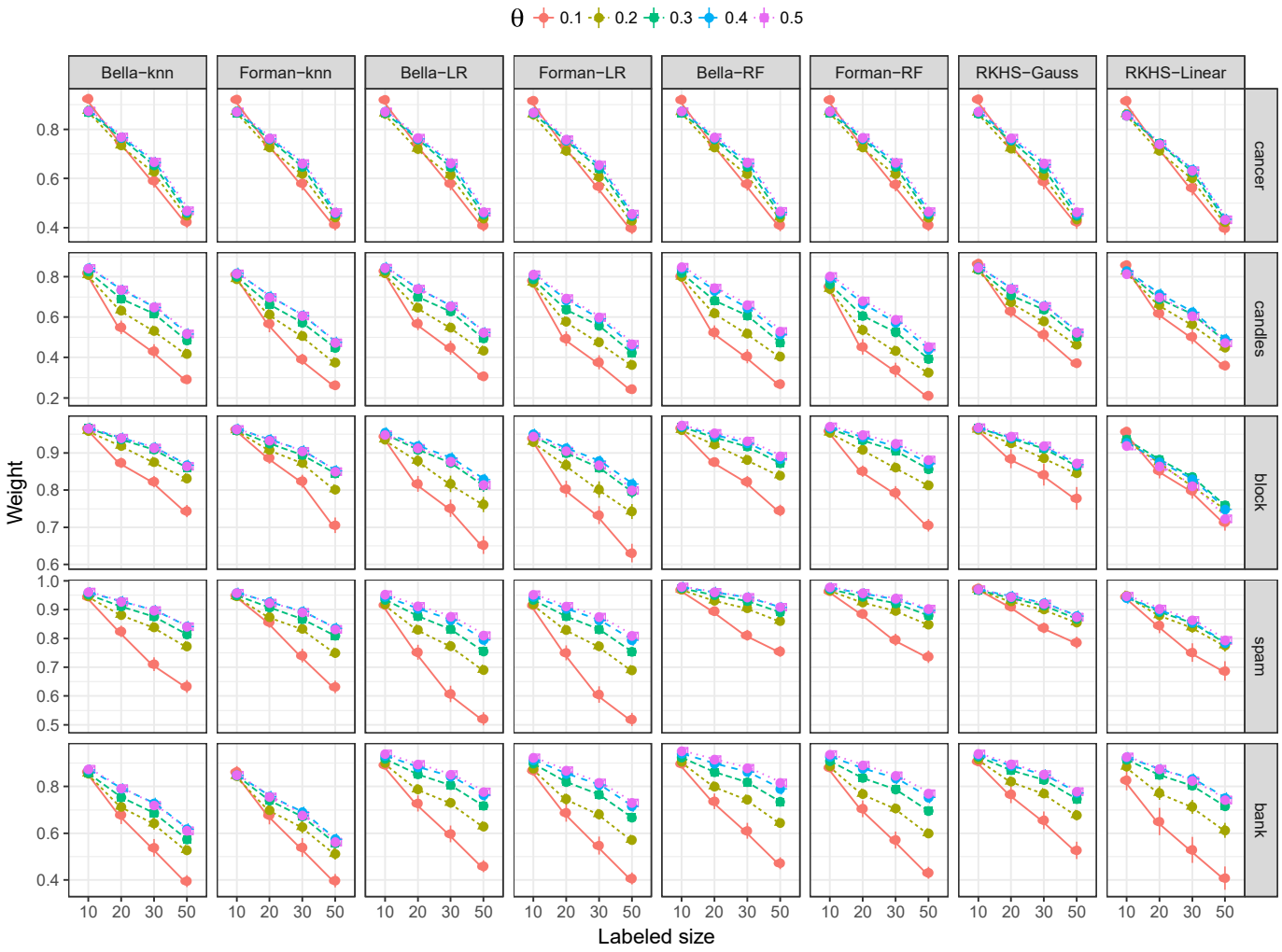

Figure 4 – Empirical weight for each setting and each size of labeled sample.

4.3 Regression Estimation

Now, we evaluate the regression ratio estimator against the classify and count method. In order to do so, an artificial dataset was generate under the following settings:

- $Z \sim U(0,1)$
- $Y = 1|S = 1 \sim \text{Ber}(1/2)$ independent of Z
- $\theta(z) = P(Y = 1|S = 0, Z = z) = (\sin(2zk\pi) + 1)/2$ for $k \in \{1,2\}$
- $X|Y = 0 \sim N(\mu,1)$ and $X|Y = 0 \sim N(-\mu,1)$ for $\mu \in \{0.5,1,1.5,2\}$
- $n_L = n_U = 1000$
- $g(X) = I(X > 0)$ (i.e., the Bayes classifier)

For each combination of the parameters k and μ were generate 400 samples. In Figure 5 we evaluate the boxplot of the error in each setting. Moreover, in Figure 6 we evaluate the average regression of the 400 samples. In both Figures, we can observe that the performance of the ratio estimator is aways better than the classify and count method. In Figure 5 we see that the errors of the ratio estimator tend to be smaller than those of the classify and count estimator, especially when μ is small. This may be associated to fact that the classification problem is harder in these case. Note that both classify and count and the ratio estimator improve their performance as the problem becomes easier (i.e., as μ gets larger), but the ratio estimator always lead to better results. Moreover, both estimators appear to have the same sensitivity to the lack of smoothness of $\theta(z)$, since both increase their errors when we go from $k = 1$ to $k = 2$. Finally, in Figure 6 we can observe that the average regression obtained by the ratio estimator fits better the real curve in all settings. In general, the classify and count underestimates $\theta(z)$ when $\theta(z) > 1/2$ and overestimates it when $\theta(z) < 1/2$.

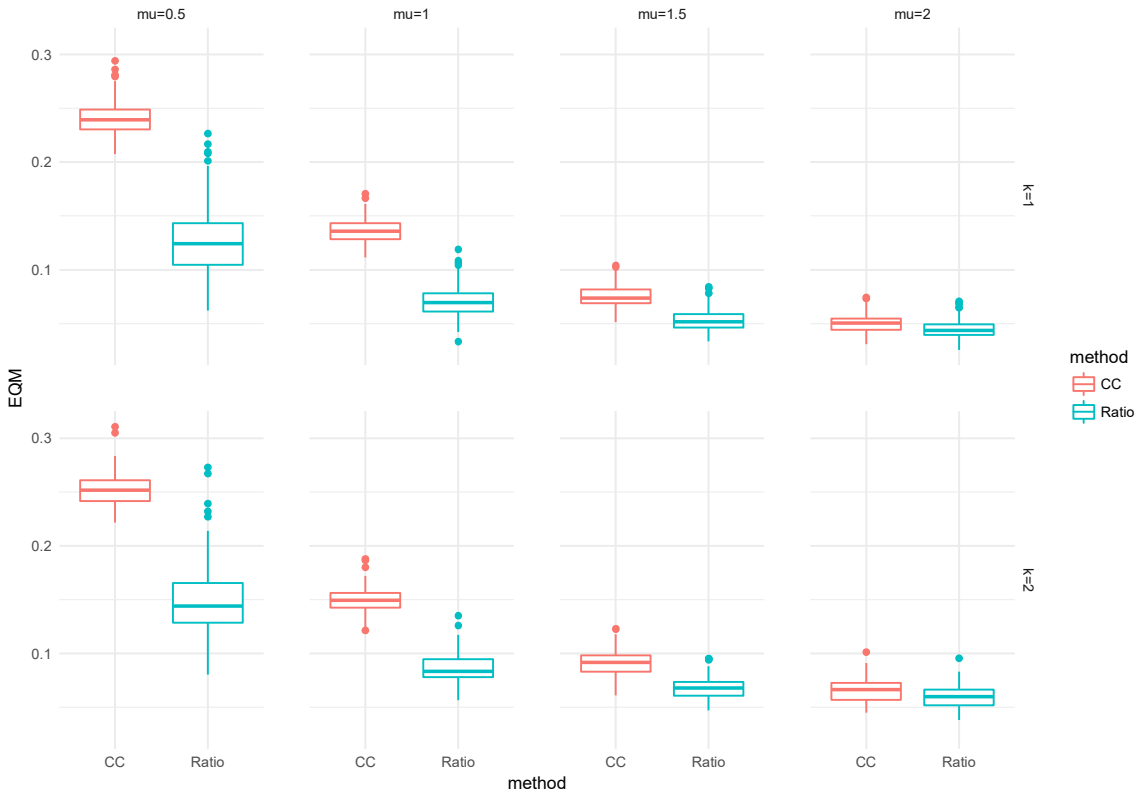

Figure 5 – Boxplot of the error in each setting.

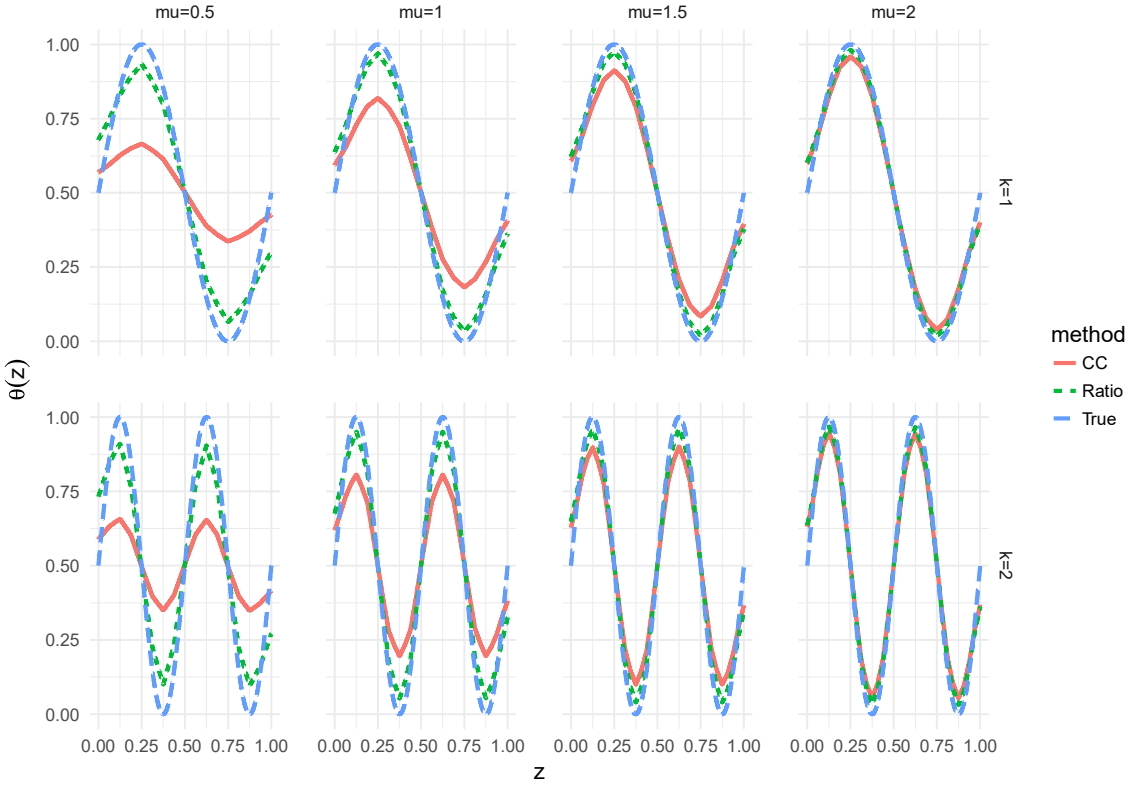

Figure 6 – Averege of the fitted regression in each setting.

CHAPTER

5

FINAL DISCUSSION

In this work, we have shown that the ratio estimator is a promising method to solve the quantification problem under the prior probability shift assumption once we proved it is consistent and show how one can build confidence intervals based on it (the coverage of these intervals was satisfactory). We have also developed a novel way to construct the function g based on RKHS which led to good results. We extend these estimator of two different way: combining it with a estimator based on a few labeled observation from target population and associating it with another covariate in order to solve regression problems. In all of these points, the ratio estimator lead to good results when compared to the classify and count (usual) approach. In the future works, novels methods to construct g's function may be propose. Moreover, still not clear under what settings each g function work better. Because of this, intensives study of simulations can be performed.

BIBLIOGRAPHY

BARRANQUERO, J.; DÍEZ, J.; COZ, J. J. del. Quantification-oriented learning based on reliable classifiers. **Pattern Recognition**, Elsevier, v. 48, n. 2, p. 591–604, 2015. Citation on page 22.

BELLA, A.; FERRI, C.; HERNÁNDEZ-ORALLO, J.; RAMIREZ-QUINTANA, M. J. Quantification via probability estimators. In: IEEE. **Data Mining (ICDM), 2010 IEEE 10th International Conference on**. [S.l.], 2010. p. 737–742. Citations on pages 22, 25, 27, and 34.

BLAKE, C. Uci repository of machine learning databases. **http://www. ics. uci. edu/˜ mlearn/MLRepository. html**, University of California, Department of Information and Computer Science, 1998. Citation on page 33.

CASELLA, G.; BERGER, R. L. **Statistical inference**. [S.l.]: Duxbury Pacific Grove, CA, 2002. Citation on page 47.

FAWCETT, T.; FLACH, P. A. A response to webb and ting's on the application of roc analysis to predict classification performance under varying class distributions. **Machine Learning**, Springer, v. 58, n. 1, p. 33–38, 2005. Citation on page 23.

FORMAN, G. Quantifying trends accurately despite classifier error and class imbalance. In: ACM. **Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.], 2006. p. 157–166. Citation on page 22.

____. Quantifying counts and costs via classification. **Data Mining and Knowledge Discovery**, Springer, v. 17, n. 2, p. 164–206, 2008. Citations on pages 21, 22, 25, 27, and 34.

FREEMAN, P.; IZBICKI, R.; LEE, A.; NEWMAN, J.; CONSELICE, C.; KOEKEMOER, A.; LOTZ, J.; MOZENA, M. New image statistics for detecting disturbed galaxy morphologies at high redshift. **Monthly Notices of the Royal Astronomical Society**, The Royal Astronomical Society, v. 434, n. 1, p. 282–295, 2013. Citation on page 33.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. **The elements of statistical learning**. [S.l.]: Springer series in statistics New York, 2001. Citation on page 22.

GART, J. J.; BUCK, A. A. 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**, v. 83, n. 3, p. 593–602, 1966. Citations on pages 22 and 25.

IZBICKI, R.; STERN, R. B. Learning with many experts: model selection and sparsity. **Statistical Analysis and Data Mining: The ASA Data Science Journal**, Wiley Online Library, v. 6, n. 6, p. 565–577, 2013. Citation on page 33.

MALERBA, D.; ESPOSITO, F.; SEMERARO, G. A further comparison of simplification methods for decision-tree induction. In: **Learning from data**. [S.l.]: Springer, 1996. p. 365–374. Citation on page 33.

MANGASARIAN, O. L. Cancer diagnosis via linear programming. **SIAM news**, v. 23, n. 5, p. 18, 1990. Citation on page 33.

MORO, S.; LAUREANO, R.; CORTEZ, P. Using data mining for bank direct marketing: An application of the crisp-dm methodology. In: EUROSIS. **Proceedings of European Simulation and Modelling Conference-ESM'2011**. [S.l.], 2011. p. 117–121. Citation on page 33.

SAERENS, M.; LATINNE, P.; DECAESTECKER, C. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. **Neural computation**, MIT Press, v. 14, n. 1, p. 21–41, 2002. Citations on pages 22 and 25.

TASCHE, D. Does quantification without adjustments work? **arXiv preprint arXiv:1602.08780**, 2016. Citation on page 22.

____. Fisher consistency for prior probability shift. **Journal of Machine Learning Research**, v. 18, n. 95, p. 1–32, 2017. Available: <http://jmlr.org/papers/v18/17-048.html>. Citations on pages 22, 23, and 25.

WAHBA, G. **Spline models for observational data**. [S.l.]: Siam, 1990. Citations on pages 28 and 48.

WANG, H.; CAN, D.; KAZEMZADEH, A.; BAR, F.; NARAYANAN, S. A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In: **Proceedings of the ACL 2012 System Demonstrations**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. (ACL '12), p. 115–120. Citations on pages 22 and 31.

ZHANG, L. H. On optimizing the sum of the rayleigh quotient and the generalized rayleigh quotient on the unit sphere. **Computational Optimization and Applications**, Springer, v. 54, n. 1, p. 111–139, 2013. Citation on page 29.

APPENDIX

A

PROOFS AND ADDITIONAL LEMMAS

Theorem 1. Follows directly from the definition of $\widehat{\theta}_R$ and $\widehat{\theta}_{TR}$, and Lemma 1. □

Lemma 2. 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 $\varepsilon_1, \varepsilon_2 \in (0,1)$

$$\mathbb{E}\left[\left(T - \frac{\mathbb{E}[Z_1|S]}{\mathbb{E}[Z_2|S]}\right)^2 \middle| S\right] \leq \frac{4(|\mathbb{E}[Z_1|S]| + \varepsilon_1)\max\left(\mathbb{V}[Z_1|S], \mathbb{V}[Z_2|S]\right)}{\min(1, (1-\varepsilon_2)^4 \mathbb{E}[Z_2|S]^4)} + \varepsilon_1^{-2}\mathbb{V}[Z_1|S] + (\varepsilon_2 \mathbb{E}[Z_2|S])^{-2}\mathbb{V}[Z_2|S].$$

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 $A = \{|Z_1 - \mathbb{E}[Z_1|S]| \leq \varepsilon_1, |Z_2 - \mathbb{E}[Z_2|S]| \leq \varepsilon_2 \mathbb{E}[Z_2|S]\}$, obtain

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{Z_1}{Z_2} - \frac{\mathbb{E}[Z_1|S]}{\mathbb{E}[Z_2|S]}\right)^2 \middle| A, S\right] \mathbb{P}(A|S) \\
=&\mathbb{E}\left[\left(\frac{1}{Z_{2,*}}(Z_1 - \mathbb{E}[Z_1|S]) - \frac{Z_{1,*}}{Z_{2,*}^2}(Z_2 - \mathbb{E}[Z_2|S])\right)^2 \middle| A, S\right] \mathbb{P}(A|S) \\
\leq& 4\max\left(\mathbb{E}\left[\frac{1}{Z_{2,*}^2}(Z_1 - \mathbb{E}[Z_1|S])^2 \middle| A, S\right], \mathbb{E}\left[\frac{Z_{1,*}^2}{Z_{2,*}^4}(Z_2 - \mathbb{E}[Z_2|S])^2 \middle| A, S\right]\right) \mathbb{P}(A|S) \\
\leq& \frac{4(|\mathbb{E}[Z_1|S]| + \varepsilon_1)\max\left(E\left[(Z_1 - \mathbb{E}[Z_1|S])^2 | A, S\right], E\left[(Z_2 - \mathbb{E}[Z_2|S])^2 \middle| A, S\right]\right) \mathbb{P}(A|S)}{\min(1, (1-\varepsilon_2)^4 \mathbb{E}[Z_2|S]^4)} \\
\leq& \frac{4(|\mathbb{E}[Z_1|S]| + \varepsilon_1)\max\left(\mathbb{V}[Z_1|S], \mathbb{V}[Z_2|S]\right)}{\min(1, (1-\varepsilon_2)^4 \mathbb{E}[Z_2|S]^4)}
\end{aligned} \tag{A.1}$$

Finally, obtain that

$$
\begin{aligned}
E\left[\left(T-\frac{\mathbb{E}[Z_1|S]}{\mathbb{E}[Z_2|S]}\right)^2\middle|S\right] &= \mathbb{E}\left[\mathbb{E}\left[\left(T-\frac{\mathbb{E}[Z_1|S]}{\mathbb{E}[Z_2|S]}\right)^2\middle|\mathbb{I}_A,S\right]\middle|S\right] \\
&\leq \mathbb{E}\left[\left(\frac{Z_1}{Z_2}-\frac{\mathbb{E}[Z_1|S]}{\mathbb{E}[Z_2|S]}\right)^2\middle|A,S\right]\mathbb{P}(A|S)+\mathbb{P}(A^c|S) \qquad T,\frac{\mathbb{E}[Z_1]}{\mathbb{E}[Z_2]}\in[0,1] \\
&\leq \frac{4(|\mathbb{E}[Z_1|S]|+\varepsilon_1)\max\left(\mathbb{V}[Z_1|S],\mathbb{V}[Z_2|S]\right)}{\min(1,(1-\varepsilon_2)^4\mathbb{E}[Z_2|S]^4)}+\mathbb{P}(A^c|S) \qquad \text{eq. A.1}
\end{aligned}
$$

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

$$
\begin{aligned}
\mathbb{P}(A^c|S) &\leq \mathbb{P}(|Z_1-\mathbb{E}[Z_1|S]|>\varepsilon_1|S)+\mathbb{P}(|Z_2-\mathbb{E}[Z_2|S]|>\varepsilon_2\mathbb{E}[Z_2|S]|S) \\
&\leq \varepsilon_1^{-2}\mathbb{V}[Z_1|S]+(\varepsilon_2\mathbb{E}[Z_2|S])^{-2}\mathbb{V}[Z_2|S]
\end{aligned}
$$

□

Theorem 2. Define $Z_1=\widehat{\mathbb{E}}[g(\mathbf{X})|S=0]-\widehat{\mathbb{E}}[g(\mathbf{X})|Y=0,S=1]$ and also $Z_2=\widehat{\mathbb{E}}[g(\mathbf{X})|Y=1,S=1]-\widehat{\mathbb{E}}[g(\mathbf{X})|Y=0,S=1]$ (Definition 1). Note that

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

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

$$
\begin{aligned}
\mathbb{E}\left[\left(\widehat{\theta}_{TR}-\theta\right)^2\middle|\mathbf{S}_1^n\right] &= \mathbb{E}\left[\left(T-\frac{\mathbb{E}[Z_1|\mathbf{S}_1^n]}{\mathbb{E}[Z_2|\mathbf{S}_1^n]}\right)^2\middle|\mathbf{S}_1^n\right] \\
&\leq \frac{4(|\mathbb{E}[Z_1|\mathbf{S}_1^n]|+\varepsilon_1)\max\left(\mathbb{V}[Z_1|\mathbf{S}_1^n],\mathbb{V}[Z_2|\mathbf{S}_1^n]\right)}{\min(1,(1-\varepsilon_2)^4\mathbb{E}[Z_2|\mathbf{S}_1^n]^4)} \\
&\quad+\varepsilon_1^{-2}\mathbb{V}[Z_1|\mathbf{S}_1^n]+(\varepsilon_2\mathbb{E}[Z_2|\mathbf{S}_1^n])^{-2}\mathbb{V}[Z_2|\mathbf{S}_1^n] \qquad \text{Lemma 2}
\end{aligned}
$$

The result follows from observing that $\mathbb{E}[Z_1|\mathbf{S}_1^n]$ and $\mathbb{E}[Z_2|\mathbf{S}_1^n]$ are constants $\mathbf{S}_1^n$, which includes n_L and n_U, $\mathbb{V}[Z_1|\mathbf{S}_1^n]=O(\max(n_L^{-1},n_U^{-1}))$, and $\mathbb{V}[Z_2|\mathbf{S}_1^n]=O(n_L^{-1})$. □

Theorem 3. Define $\mu_U:=\mathbb{E}[g(\mathbf{X}_i)|S_i=0]$, $\sigma_U^2=\mathbb{V}[g(\mathbf{X}_i)|S_i=0]$, and

- $Z_{U,n}:=\frac{\sqrt{n_U}}{\sigma_U}(\widehat{\mathbb{E}}[g(\mathbf{X})|S=0]-\mu_U)=\frac{\sqrt{n_U}}{\sigma_U}\left(\frac{\sum_{i=1}^n g(\mathbf{X}_i)\mathbb{I}(S_i=0)}{n_U}-\mu_U\right)$,
- $Z_{0,n}:=\frac{\sqrt{n_0}}{\sigma_0}\left(\widehat{\mathbb{E}}[g(\mathbf{X})|S=1,Y=0]-\mu_0\right)=\frac{\sqrt{n_0}}{\sigma_0}\left(\frac{\sum_{i=1}^n g(\mathbf{X}_i)\mathbb{I}(S_i=0,Y_i=0)}{n_0}-\mu_0\right)$,
- $Z_{1,n}:=\frac{\sqrt{n_1}}{\sigma_1}\left(\widehat{\mathbb{E}}[g(\mathbf{X})|S=1,Y=1]-\mu_1\right)=\frac{\sqrt{n_1}}{\sigma_1}\left(\frac{\sum_{i=1}^n g(\mathbf{X}_i)\mathbb{I}(S_i=0,Y_i=1)}{n_1}-\mu_1\right)$,
- $F_i=\mathbb{I}(S_i=1)(Y_i+1)$, $A_U=\{F_1=0\}$, $A_0=\{F_1=1\}$, and $A_1=\{F_1=2\}$.

$$\begin{aligned}\lim_{n\to\infty}\phi_{Z_{U,n},Z_{0,n},Z_{1,n}}(t_U,t_0,t_1) &= \lim_{n\to\infty}\mathbb{E}\left[\mathbb{E}\left[\exp\left(\sum_{j\in\{U,0,1\}} it_jZ_{j,n}\right)\middle|F_1,\ldots,F_n\right]\right]\\ &= \lim_{n\to\infty}\mathbb{E}\left[\prod_{j\in\{U,0,1\}}\mathbb{E}\left[\exp\left(it_jZ_{j,n}\right)\middle|F_1,\ldots,F_n\right]\right]\\ &= \lim_{n\to\infty}\mathbb{E}\left[\prod_{j\in\{U,0,1\}}\left(\phi_{\frac{g(\mathbf{X}_1)-\mu_j}{\sigma_j}|A_j}(t_jn_j^{-0.5})\right)^{n_j}\right] \qquad (A.2)\end{aligned}$$

It follows from the Central Limit Theorem for i.i.d. random variables that, for every $j\in\{U,0,1\}$, $\phi_{\frac{g(\mathbf{X}_1)-\mu_j}{\sigma_j}|A_j}(t_jn_j^{-0.5})\to\exp(-0.5t_j^2)$ as $n_j\to\infty$. Since $n_j\xrightarrow{a.s.}\infty$, conclude from eq. A.2 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

$$(Z_{U,n},Z_{0,n},Z_{1,n})\xrightarrow{\mathscr{L}}N(0,\mathbb{I}) \qquad (A.3)$$

Assume that $p_L\neq 0$. In this case, since $\frac{n_L}{n}\xrightarrow{\mathbb{P}}p_L$, it follows from eq. A.3 that

$$\sqrt{n}\left(\widehat{\mathbb{E}}[g(\mathbf{X})|S=0]-\mu_U,\widehat{\mathbb{E}}[g(\mathbf{X})|S=1,Y=0]-\mu_0,\widehat{\mathbb{E}}[g(\mathbf{X})|S=1,Y=1]-\mu_1\right)$$

converges in distribution to $N\left(0,diag\left(\frac{\sigma_U^2}{1-p_L},\frac{\sigma_0^2}{p_Lp_{0|L}},\frac{\sigma_1^2}{p_Lp_{1|L}}\right)\right)$. Since $\theta=\frac{\mu_U-\mu_0}{\mu_1-\mu_0}$ (Lemma **??**) and $\widehat{\theta}_R=\frac{\widehat{\mathbb{E}}[g(\mathbf{X})|S=0]-\widehat{\mathbb{E}}[g(\mathbf{X})|S=1,Y=0]}{\widehat{\mathbb{E}}[g(\mathbf{X})|S=1,Y=1]-\widehat{\mathbb{E}}[g(\mathbf{X})|S=1,Y=0]}$, it follows from the delta method (CASELLA; BERGER, 2002) that

$$\sqrt{n}(\widehat{\theta}_R-\theta)\xrightarrow{\mathscr{L}}N\left(0,\frac{\sigma_U^2(1-p_L)^{-1}}{(\mu_1-\mu_0)^2}+\frac{(\mu_U-\mu_1)^2\sigma_0^2(p_Lp_{0|L})^{-1}}{(\mu_1-\mu_0)^4}+\frac{(\mu_U-\mu_0)^2\sigma_1^2(p_Lp_{1|L})^{-1}}{(\mu_1-\mu_0)^4}\right)$$

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

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

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

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

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

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

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

Theorem 4. It follows from the Representer Theorem (WAHBA, 1990) that, for every $g \in \mathscr{H}_K$, $g(\mathbf{x}) = \sum_{k \in A_1} w_k K(\mathbf{x}, \mathbf{x}_k)$. 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 $g \in \mathscr{H}_K$,

$$\widehat{\mathrm{MSE}}(g) + \lambda ||g||_{\mathscr{H}_K} = \frac{\mathbf{w}^t N \mathbf{w}}{\mathbf{w}^t M \mathbf{w}} + \lambda \mathbf{w}^t \mathbb{K} \mathbf{w}.$$

□

ANNEX

A

R CODES

<https://github.com/afonsofvaz/PriorShiftQuantification> R codes used in the experiments.


ICMC
USP
SÃO CARLOS