Comparing Correction Methods to Reduce Misclassification Bias

Kevin Kloos (s2370530)

Local Supervisor 1: Dr. Sander Scholtus

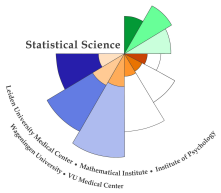
Local Supervisor 2: Quinten Meertens MSc

Statistical Science Supervisor: Dr. Julian Karch

MASTER THESIS

Defended on Month Day, Year

Specialization: Data Science



STATISTICAL SCIENCE
FOR THE LIFE AND BEHAVIOURAL SCIENCES

Foreword

First of all, I would like Statistics Netherlands for arranging my internship and this thesis project. During the first weeks before the COVID-crisis, I was able to work at their office and it has been a great experience. In particular, I would like to thank my internal supervisors Sander Scholtus and Quinten Meertens, and my statistical science supervisor Julian Karch for their help and contribution to this paper. It could not be achieved without their support and their interesting contributions. Furthermore, I would like to thank Maarten Kampert and the other contributors of the EMOS program for arranging interesting courses and a even more interesting internship. It has been a great addition to my masters degree. Moreover, my supervisors and I were able to write a paper for the BNAIC-BENELEARN 2020 conference*. It has been a great experience to write my first scientific paper next to my thesis and it is even greater that we were a nominee for the Best Paper Award. I would like to thank the anonymous referees and Arnout van Delden for reviewing this paper. Furthermore, I would like to thank the organisation of the BNAIC-BENELEARN conference for organizing the event and publishing the paper. Last of all, I would like to thank my family at home, and my friend and library-mate Maxine for enduring with me during the last few months. It has been a pleasure!

An important note is that this paper has overlapping content with the paper we wrote for the BNAIC-BENELEARN conference. With permission of the thesis committee, I was able to use the paper in chapters 1,2,3 and 6 of this thesis. Chapter 4 and 5 of this thesis is not discussed in the paper.

The figures in this paper are made with a dashboard. This dashboard can be found on `https://github.com/kevinkloos/MasterThesis` such that everyone is able to compare the estimators given a set of parameters. On this page, you can find the thesis itself and the dashboard.

*Conference website: `https://bnaic.liacs.leidenuniv.nl/`. Paper and additional content: `https://github.com/kevinkloos/Misclassification-Bias`

Contents

Abstract

When applying supervised machine learning algorithms, the classical goal is trying to estimate the true labels of the objects as accurately as possible. However, if the predictions of an accurate algorithm are aggregated, for example by counting the predicted labels per class, the result is often biased. The statistical bias that occurs when aggregating statistics is defined as misclassification bias. National statistical institutes try to avoid biased statistics as much as possible and therefore these algorithms cannot be used without adjustments. We propose five estimators to reduce the misclassification bias as much as possible. However, reducing misclassification bias can increase the variance and therefore the mean square error (MSE) will be used as test statistic to evaluate the estimators. Up to now, only asymptotic results existed for three of the five estimators and these results are extended in this paper to expressions for finite samples. Therefore, it is possible to compare the estimators and find the estimator with the lowest MSE. We concluded that the calibration estimator has in most of the cases the lowest RMSE and significantly reduces the misclassification bias of the uncorrected predictions. The baseline estimator performs well for bad classifiers, while the subtracted-bias estimator and the classify-and-count estimator perform well when the absolute amount of errors in both groups are almost equal.

The estimators perform differently when the distribution of the target population changes over time, i.e. the proportion of objects per class change over time. When the target population changes over time and the test set remains unchanged, the calibration estimation can be heavily biased. None of the five estimators perform consistently well over all possible combinations of the parameters and therefore a new estimator is proposed. The mixed estimator is a combination of the misclassification estimator and the calibration estimator which performs generally better than the five original estimators. However, this estimator has still some harmful properties like an extreme variance for bad classifiers.

This thesis addresses the dangers of misclassification bias and proposes new estimators in order to solve this misclassification bias. We know which estimators perform well in certain situations and can apply it to correct the outcomes of binary classifiers. Also, this thesis can be used as a starting point to explore multi-class classifiers or to explore what happens if you relax other assumptions. Last, this thesis can be used to compute new estimators and compare their performances with the five original estimators.

Chapter 1

Introduction

1.1 General Background

Over the last few years, National Statistical Institutes have made more and more use of new innovative techniques to generate their statistics. These techniques vary from big data to machine learning algorithms. An example where machine learning algorithms are used to generate statistics, is in estimating the proportion of households in the Netherlands that have a solar-panel installation on their roofs [2]. The algorithms are obviously not perfect and they will make errors in its classifications. These errors can lead to a systematic bias in the so called *base rate*, which is defined as the proportion of objects that belong to the class of interest [13, 14]. This systematic bias that occurs when aggregating classifications, is called *misclassification bias*.

The misclassification bias can be illustrated with an toy-example in Figure 1.1. First, we define our class of interest as the class that contains the blue boxes. The population contains 100 blue boxes and 900 orange boxes, where the algorithm classifies a blue box with a probability of 98% correctly and an orange box with a probability of 97% correctly. The expected output of the blue boxes contains the true positives of the blue boxes and the false negatives of the orange boxes, while the expected output of the orange boxes contains the true negatives of the orange boxes and the false positives of the blue boxes. The expected output of the amount of blue boxes in the population is therefore 125, instead of the true value of 100. Despite the high classification probabilities, there occurs a large amount misclassification bias.

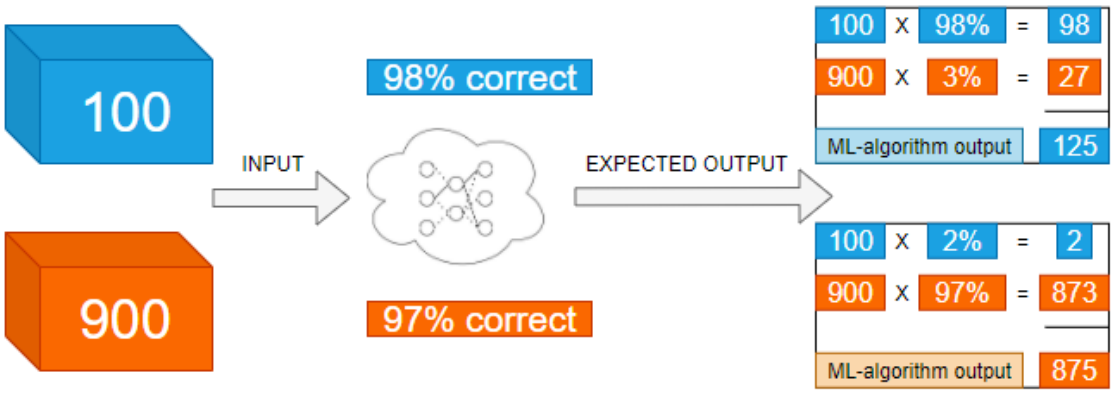


Figure 1.1: Illustration of misclassification bias with blue and orange boxes

In official statistics, there are more serious examples where misclassification bias occurs. An example is the classification of businesses per type in the Netherlands. The population size is large and the outcome of these statistics is important for Dutch policy makers. We already observed that a big population size does not remove bias, due to the imbalance between the errors of the classes. Therefore, we should apply corrections on these statistics, otherwise policy could be made on biased statistics.

Besides official statistics [10], misclassification bias occurs in a broad range of applications, like land cover mapping [9], political science [6, 17], and epidemiology [5]. The aim of these studies is to minimize the loss of the aggregated prediction, for example minimizing the base rate, instead of minimizing the loss of the individual predictions. Minimizing the loss of the aggregated prediction is in literature defined as *quantification learning* and there are different approaches to obtain this goal [3]. The approach used in this paper, is defined as *Classify, count and correct*. We first classify the sample, then count the estimates to obtain a first prediction and last apply correction methods to generate a new prediction of the base rate. With this approach, we try to minimize the misclassification bias as much as possible.

1.2 Research Aim

Minimizing the misclassification bias of the base rate however might lead to an increase in the variance of its predictions. Therefore, the quality of the estimators is measured in terms of the (root) mean square error. This leads to our research question: *"How can we reduce the mean squared error of the base rate for inaccurate data?"*. This research question can be divided to two subquestions.

The first subquestion is *"How can we reduce the mean squared error of the base rate for inaccurate data in case of a fixed base rate?"* We compare five estimators based on what is known in the available literature [1, 8]. For each of these estimators, we compute the bias, variance and so the mean square error. To the best of our knowledge, there only exists asymptotic expressions regarding to the MSE for three of the five estimators. This paper extends these expressions for finite datasets. We restrict ourselves to binary classifiers, but we believe that the expressions can be extended to multi-class models. The theoretical expressions make it possible to compare the estimators and give recommendations on which estimator performs well under which circumstances.

The second subquestion is formulated as *"How can we reduce the mean squared error of the base rate for inaccurate data in case of a changing base rate?"* In order to answer this research question, we will perform a simulation study in the case that the value of the base rate parameter changes over time. This phenomenon is defined as *concept drift* [16]. We will again apply the same estimators and try to compare the behaviour of the estimators under a fixed base rate with the behaviour of the estimators under concept drift. With these two subquestions, we should be able to answer the research question.

The thesis is therefore built on the following chapters. In chapter 2, we introduce the estimators and derive expressions for the bias, variance and MSE in case for a fixed value of the base rate. In chapter 3, we will draw conclusions from the derived expressions and can indicate which estimator performs well given the parameters. Chapter 4 discusses the impact of concept drift on the estimators, while chapter 5 proposes a new estimator under concept drift. Finally, chapter 6 presents a discussion and gives suggestions for future research.

Chapter 2

Computing the estimators for a fixed base rate

2.1 Notation and assumptions

In this section, we will define our model and its assumptions. We define a target population of size N. This target population consists of objects that belong to one of the two classes. Class 1 is our *class of interest* and the parameter of interest is the base rate of class 1, denoted as *alpha* (α). From our example in the Introduction, we define the blue boxes as our *class of interest* and we define the parameter of interest α as the proportion of blue boxes in the target population.

In the target population, only the predicted labels are observed, while the true labels are unobserved. In order to utilize our estimators, we need a test set. We assume that this test set is a simple random sample from our target population and contains both the predicted labels as the true labels, see figure 2.1. An important detail is that this test set is not used to train the classifier, otherwise we could not fairly use our estimators. Another assumption is that the size of the test set is much smaller than the size of the target population (i.e. $n \ll N$), such that we can apply our machine learning algorithm on the whole target population.

Target Population		
object	predicted class	true class
1	0	?
2	1	?
3	0	?
4	1	?
⋮	⋮	⋮
N-1	1	?
N	0	?

Simple random sample →

Test Set		
object	predicted class	true class
1	0	0
2	1	0
3	0	1
4	1	1
⋮	⋮	⋮
n-1	1	1
n	0	1

Figure 2.1: A simple illustration of the elements in the target population and the test set

The aim of the study is to use estimators that remove *misclassification bias* without adding too much variance. In order to apply the estimators, we need to define some parameters first. For instance, we need the classification probabilities for the objects of each class. We assume that the classification probabilities are equal for every object in the same class. Then p_{00} denotes the probability that an object in class 0 is correctly classified and p_{11} denotes the probability that an object in class 1, our class of interest, is correctly classified. In machine learning, p_{11} is

Table 2.1: Contingency tables for test set (a) and target population (b)

(a)

		Estimated class		
		0	1	Tot
True class	0	n_{00}	n_{01}	n_{0+}
	1	n_{10}	n_{11}	n_{1+}
	Tot	n_{+0}	n_{+1}	n

(b)

		Estimated class		
		0	1	Tot
True class	0	N_{00}	N_{01}	N_{0+}
	1	N_{10}	N_{11}	N_{1+}
	Tot	N_{+0}	N_{+1}	N

better known as the sensitivity and p_{00} is better known as the specificity. In the example of the boxes in the Introduction, the classification probability of an orange box (specificity) is 97% and the classification probability of an blue box (sensitivity) is 98%, i.e. $p_{00} = 0.97$ and $p_{11} = 0.98$. These probabilities can be noted in the *row-normalized confusion matrix* $\mathbf{P}$:

$$\mathbf{P} = \begin{pmatrix} p_{00} & 1 - p_{00} \\ 1 - p_{11} & p_{11} \end{pmatrix}. \tag{2.1}$$

These classification probabilities are unknown in practice. However, these classification probabilities can be estimated with help of the test set. Recall that this test set is a simple random sample from the target population and contains both the predicted labels and the true labels for all the objects. We can estimate the specificity, $\hat{p}_{00}$ by dividing the amount of objects correctly classified as class 0 in the test set by all the objects that belong to class 0 in the test set. In terms of the values in Table 2.1, we divide n_{00} by n_{0+} This works in a similar way for $\hat{p}_{11}$, where we divide n_{11} by n_{1+}. Furthermore, the base rate parameter α can be calculated by dividing N_{1+} by N. Similarly, we can apply calibration probabilities on our estimator. Instead of using a row-normalized confusion matrix $\mathbf{P}$, we can also define a *column-normalized confusion matrix* $\mathbf{C}$. Then, each entry c_{ij} is the probability that an object, classified as class j, belongs to class i. The calibration probabilities are, similar to the calibration probabilities, unknown and can be estimated from the test set. Each point c_{ij} can be estimated by dividing cell value n_{ij} by its column total n_{+j}.

$$\mathbf{C} = \begin{pmatrix} c_{00} & 1 - c_{11} \\ 1 - c_{00} & c_{11} \end{pmatrix}. \tag{2.2}$$

Finally, we can make some mathematical assumptions. The first assumption is that the size of the test set is much smaller than the size of the target population. This assumption makes it possible to assume, in combination with the fact that α is fixed, that n_{1+} follows a $Bin(n, \alpha)$-distribution, instead of a hyper-geometric distribution. Next, the predicted classes are independent random variables, whereby an object of class i has the same probability of being classified correctly as any other object of class i. Accordingly, we assume that the classifier has higher classification probabilities than just simply guessing, but is not error-free, i.e. $0.5 < p_{ii} < 1$. Last, we assume that the elements in the contingency tables all follow a binomial distribution conditioned on the row totals and with the classification probabilities as success probabilities. For example, $n_{00} \mid n_{0+} \sim Bin(n_{0+}, p_{00})$ and $N_{11} \mid N_{1+} \sim Bin(N_{1+}, p_{11})$. We should also notice that the expressions for the bias and variance of an estimator cannot always be evaluated exactly. In order to compute these expressions, we make use of Taylor Series approximations.

2.2 Computation of the five estimators

This chapter contains mathematical expressions of each estimator and expressions for their bias and variance. We assume a fixed value for the base rate α, wherefore we can say that the test set is drawn of the same population as the target population.

2.2.1 Baseline estimator

The baseline estimator ($\hat{\alpha}_a$) is the first of the five estimators we will discuss in this thesis. The baseline estimator can be calculated by counting the objects in the test set that belong to class 1 and divide it by the total amount of objects that occur in the test set.

$$\hat{\alpha}_a = \frac{n_{1+}}{n} \tag{2.3}$$

Under the assumptions discussed in section 2.1, we are able to say that the baseline estimator is unbiased, because we take a simple random sample from the target population.

$$B[\hat{\alpha}_a] = 0 \tag{2.4}$$

Under the mathematical assumptions, we have said that sampling a test set from the target population can be approximated as a simple random sample with replacement. Therefore, the row total n_{1+} can be approximated with a binomial distribution of size n and success parameter α. This approach results into a easier expression for the variance, and therefore MSE, than when we use a hypergeometric distribution.

$$V[\hat{\alpha}_a] = MSE[\hat{\alpha}_a] = \frac{\alpha(1-\alpha)}{n} \tag{2.5}$$

This estimator does not make use of the predicted labels of the machine learning algorithm. It would be preferable if the other estimators perform better than this baseline estimator, otherwise it would not make sense to use a machine learning algorithm. It would be then a lot cheaper to perform a survey instead of building an algorithm.

2.2.2 Classify-and-count estimator

The second estimator that we will discuss is the Classify-and-count estimator ($\hat{\alpha}^*$). When we apply the machine algorithm, we can just simply count the amount of labels that the algorithm classifies as class 1 and divide it by the size of the target population.

$$\hat{\alpha}^* = \frac{N_{+1}}{N} \tag{2.6}$$

Following the example in the Introduction, the estimate of $\hat{\alpha}^*$ can be highly biased. This bias can be shown by computing the expected value of this estimator first. The expected value can be calculated by adding the correctly classified objects in class 1 with the incorrectly classified objects in class 0. When we subtract base rate α from this expected value, we compute the bias.

$$E[\hat{\alpha}^*] = \alpha p_{11} + (1-\alpha)(1-p_{00}) \tag{2.7}$$
$$B[\hat{\alpha}^*] = E[\hat{\alpha}^*] - \alpha = \alpha(p_{11}-1) + (1-\alpha)(1-p_{00}) \tag{2.8}$$

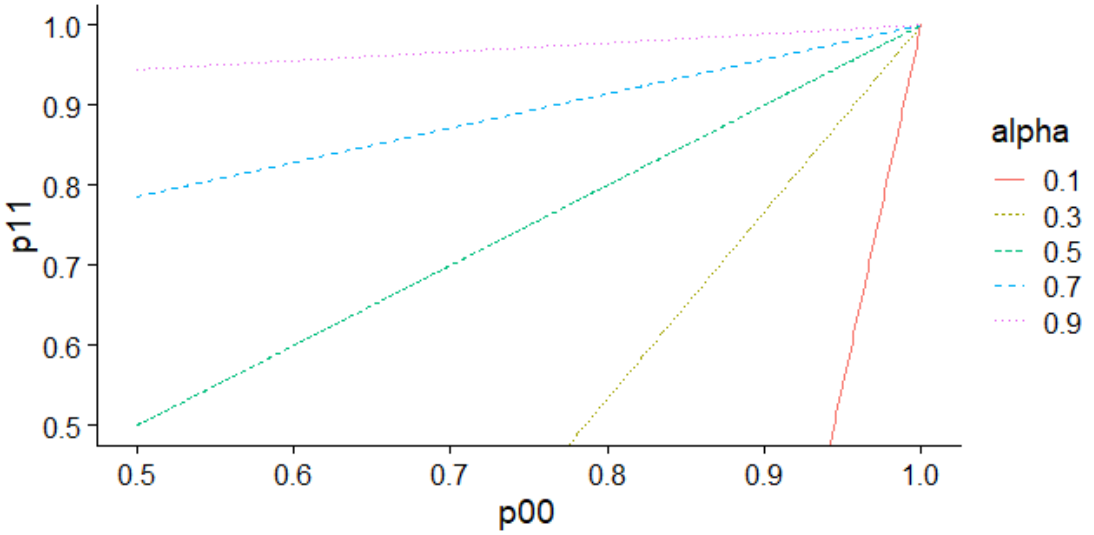


Figure 2.2: The parameter space where the Classify-and-count estimator is unbiased for different values of alpha.

The classify-and-count estimator is only unbiased when there are an equal amount of misclassified objects in both classes. This is the case when (p_{00}, p_{11}) lies on the line through $(1-\alpha, \alpha)$ and $(1, 1)$ [15], see Figure 2.2.

The variance of the classify-and-count estimator is computed in [13] and is equal to

$$V[\hat{\alpha}^*] = \frac{\alpha p_{11}(1-p_{11}) + (1-\alpha)p_{00}(1-p_{00})}{N} = O(\frac{1}{N}) \tag{2.9}$$

From (2.9), we notice inverse proportionality between the variance and the population size. Because we assume a large population size, we can neglect the variance as a higher order term. Note that the expected value does not dependent on the population size and will therefore not vanish when the population size increases. The final step is computing the MSE, which is in this case equal to the squared bias.

$$MSE[\hat{\alpha}^*] = (\alpha(p_{11}-1) + (1-\alpha)(1-p_{11}))^2 + O(\frac{1}{N}) \tag{2.10}$$

2.2.3 Subtracted-bias estimator

The third estimator that we will discuss is the Subtracted-bias estimator ($\hat{\alpha}_b$). This estimator takes the classify-and-count estimator, accordingly estimates the bias of the classify-and-count estimator and subtracts it to obtain a less biased estimator. From (2.8), we already know that the bias is dependent on α, p_{00} and p_{11}. However, these values are unknown in practice and therefore we need plug-in estimators to make an estimate of the bias. Instead of the true base rate α, we can use the classify-and-count estimator $\hat{\alpha}^*$ and instead of the true classification probabilities p_{00} and p_{11}, we use the estimated classification probabilities $\hat{p}_{00}$ and $\hat{p}_{11}$. The estimated bias for the classify-and-count estimator is then equal to

$$\begin{aligned}\hat{B}[\hat{\alpha}^*] &= \hat{\alpha}^*\hat{p}_{11} + (1-\hat{\alpha}^*)(1-\hat{p}_{00}) - \hat{\alpha}^* \\ &= \hat{\alpha}^*(\hat{p}_{00} + \hat{p}_{11} - 2) + (1-\hat{p}_{00})\end{aligned} \tag{2.11}$$

The formula for the estimator can be computed by taking the Classify-and-count estimator and subtract it with the estimated bias.

$$\hat{\alpha}_b = \hat{\alpha}^*(3 - \hat{p}_{00} - \hat{p}_{11}) - (1-\hat{p}_{00}) \tag{2.12}$$

To the best of our knowledge, the equations for the bias and variance of the Subtracted-bias method have not been published in scientific literature*. Therefore, we were able to derive both up to terms of order $1/n^2$, yielding the following result.

Theorem 1. *The bias of $\hat{\alpha}_b$ as estimator for α is given by*

$$B\left[\hat{\alpha}_b\right] = (1-p_{00})(2-p_{00}-p_{11}) - \alpha(p_{00}+p_{11}-2)^2. \tag{2.13}$$

The variance of $\hat{\alpha}_b$ equals

$$\begin{aligned}V\left[\hat{\alpha}_b\right] = &\frac{\left[\alpha(p_{00}+p_{11}-1)-p_{00}\right]^2 p_{00}(1-p_{00})}{n(1-\alpha)}\left(1+\frac{\alpha}{n(1-\alpha)}\right) \\ &+ \frac{\left[\alpha(p_{00}+p_{11}-1)+(1-p_{00})\right]^2 p_{11}(1-p_{11})}{n\alpha}\left(1+\frac{1-\alpha}{n\alpha}\right) \\ &+ O\left(\max\left[\frac{1}{n^3}, \frac{1}{N}\right]\right).\end{aligned} \tag{2.14}$$

Proof. See the Appendix. □

Note that (2.13) can be rewritten as $B\left[\hat{\alpha}_b\right] = (2-p_{00}-p_{11}) \times B\left[\hat{\alpha}^*\right]$. As a result of our assumption that the classification probabilities are always between 0.5 and 1, we can say that $1 \leq p_{00}+p_{11} \leq 2$. Hence, $|B\left[\hat{\alpha}_b\right]| \leq |B\left[\hat{\alpha}^*\right]|$. At the cost of reducing misclassification bias, is the increase in variance. The main reason of the increase in variance is the variance in estimating the misclassification probabilities from the test set. This variance can be reduced by taking a bigger sample from the target population.

One may notice that we can also use an unbiased variant of the subtracted-bias estimator by using the baseline estimator $\hat{\alpha}_a$, instead of the biased classify-and-count estimator. Obviously, $\hat{\alpha}_a$ has more variance than $\hat{\alpha}^*$, which will increase the total variance substantially. Therefore, we may conclude that this variant of the subtracted-bias estimator won't add new information and we will not use this estimator in this thesis.

2.2.4 Misclassification estimator

The fourth estimator that we will discuss in this paper is the misclassification estimator ($\hat{\alpha}_p$). As already shown in (2.1), $\mathbf{P}$ presents the expected values for the row-normalized confusion matrix in the target population, trained by the machine learning algorithm. In other words, each entry p_{ij} stands for the probability that an object of class i is labeled as class j by the algorithm. We can denote the base rate of each class as $\boldsymbol{\alpha}$, which is in the binary case equal to $(1-\alpha, \alpha)$.

*Except for the paper published in BNAIC (more formal)

Under the assumption that each point of class i has equal classification probabilities as all the other points in class i, we obtain the expression $E[\hat{\alpha}^*] = \mathbf{P}^T\boldsymbol{\alpha}$. The other way around, we obtain an expression for $\boldsymbol{\alpha}$ by inverting $\mathbf{P}$: $\boldsymbol{\alpha_p} = (\mathbf{P}^T)^{-1}E[\hat{\boldsymbol{\alpha}}^*]$. Note that we invert $\mathbf{P}$ and therefore the expression has no solution for $p_{00} = p_{11} = 0.5$, due to singularity. If the true classification probabilities were known, then $\hat{\boldsymbol{\alpha}}_{\boldsymbol{p}} = (\mathbf{P}^T)^{-1}\hat{\boldsymbol{\alpha}}^*$ would give an unbiased estimate for α. However, these values are unknown and we should use the plug-in estimator $\hat{\boldsymbol{P}}$, obtained from the test set, to compute the estimate. When we write out all the elements in the matrices, we obtain the following expression:

$$\hat{\alpha}_p = \frac{\hat{\alpha}^* + \hat{p}_{00} - 1}{\hat{p}_{00} + \hat{p}_{11} - 1} \tag{2.15}$$

It is already known that the estimator $\hat{\alpha}_p$ is asymptotically unbiased for α, see [1]. In [4], the asymptotic variance of this estimator is already analysed for an arbitrary number of classes. However, a simple analytic expression for the bias and variance of $\hat{\alpha}_p$ in the binary case does, as far as we know, not exist. Therefore, we have derived the bias and variance for finite datasets, yielding the following result.

Theorem 2. *The bias of $\hat{\alpha}_p$ as estimator for α is given by*

$$B\left[\hat{\alpha}_p\right] = \frac{p_{00} - p_{11}}{n(p_{00} + p_{11} - 1)} + O\left(\frac{1}{n^2}\right). \tag{2.16}$$

The variance of $\hat{\alpha}_p$ is given by

$$V[\hat{\alpha}_p] = \frac{(1-\alpha)p_{00}(1-p_{00})\left[1 + \frac{\alpha}{n(1-\alpha)}\right] + \alpha p_{11}(1-p_{11})\left[1 + \frac{1-\alpha}{n\alpha}\right]}{n(p_{00} + p_{11} - 1)^2} + O\left(\max\left[\frac{1}{n^2}, \frac{1}{N}\right]\right). \tag{2.17}$$

Proof. See the Appendix. □

2.2.5 Calibration estimator

The fifth, and final, estimator that we will discuss in this paper is the calibration estimator ($\hat{\alpha}_c$). In contrast to the row-normalized confusion matrix $\mathbf{P}$, let $\mathbf{C}$ be the column-normalized confusion matrix of the machine learning algorithm that we have trained. Then, each entry c_{ij} is the probability that an object, which is classified as class j, belongs to class i. We obtain an unbiased estimate for α by multiplying $\mathbf{C}$ with the vector of Classify-and-count estimators $\hat{\boldsymbol{\alpha}}^*$. However, the true calibration probabilities are unknown and therefore we need the plug-in estimator $\hat{\boldsymbol{C}}$ that we can obtain from the test set. Each entry in $\hat{\boldsymbol{C}}$, $\hat{c}_{ij}$, can be calculated by dividing each entry n_{ij} by n_{+j}. When we write out all the elements in the matrices, we obtain the following expression:

$$\hat{\alpha}_p = (1 - \hat{\alpha}^*)\frac{n_{10}}{n_{+0}} + \hat{\alpha}^*\frac{n_{11}}{n_{+1}}. \tag{2.18}$$

In literature, it already has been shown that $\hat{\alpha}_c$ is a consistent estimator for α [1]. Under the assumptions in this paper, we were able to proof that $\hat{\alpha}_c$ is also an unbiased estimator for finite datasets. Furthermore, we were also able to obtain an approximation of the variance of $\hat{\alpha}_c$. These expression were, as far as we know, not computed ever before in scientific literature.

Theorem 3. *The calibration estimator $\hat{\alpha}_c$ is an unbiased estimator for α:*

$$B\left[\hat{\alpha}_c\right] = 0. \tag{2.19}$$

The variance of $\hat{\alpha}_c$ is equal to the following expression:

$$\begin{aligned} V(\hat{\alpha}_c) = & \left[\frac{(1-\alpha)(1-p_{00}) + \alpha p_{11}}{n} + \frac{(1-\alpha)p_{00} + \alpha(1-p_{11})}{n^2}\right] \\ & \times \left[\frac{\alpha p_{11}}{(1-\alpha)(1-p_{00}) + \alpha p_{11}}\left(1 - \frac{\alpha p_{11}}{(1-\alpha)(1-p_{00}) + \alpha p_{11}}\right)\right] \\ & + \left[\frac{(1-\alpha)p_{00} + \alpha(1-p_{11})}{n} + \frac{(1-\alpha)(1-p_{00}) + \alpha p_{11}}{n^2}\right] \\ & \times \left[\frac{(1-\alpha)p_{00}}{(1-\alpha)p_{00} + \alpha(1-p_{11})}\left(1 - \frac{(1-\alpha)p_{00}}{(1-\alpha)p_{00} + \alpha(1-p_{11})}\right)\right] \\ & + O\left(\max\left[\frac{1}{n^3}, \frac{1}{Nn}\right]\right). \end{aligned} \tag{2.20}$$

Hereby, the overview of the five estimators for the base rate α is complete. Each estimator has a theoretical expression for the bias and variance and these expressions can be used to compare the (root) mean square error of the five estimators. This will be done in the next chapter amplified with a simulation study.

Chapter 3

Results of the model under a fixed base rate

3.1 Behaviour of the estimators

In this section, we describe the behaviour of the five estimators. In the previous chapter, we obtained mathematical expressions for the estimators and their bias and variance. We use these expressions, in addition with simulation studies, to describe the behaviour of the estimators.

3.1.1 Baseline estimator

The baseline estimator has the most predictable behaviour out of the five introduced estimators. This estimator is, under our assumptions, unbiased for each feasible combination of $(p_{00}, p_{11}, \alpha, n, N)$. Furthermore, its variance does not depend on the classification probabilities p_{00} and p_{11}. The estimator does not use the machine learning algorithm to estimate $\hat{\alpha}_a$ and performs therefore equally well for low classification probabilities as well as for high classification probabilities. However, the variance of the baseline estimator is heavily dependent on the sample size of the test set. Relatively small test sets lead to estimates with a high variance. Last, the variance of the baseline estimator is also dependent on the base rate, where a more balanced target population, leads to a higher variance. Both effects can be shown with Figure 3.1.

3.1.2 Classify-and-count estimator

The classify-and-count estimator is an estimator with high bias, but a negligible variance. As shown in Figure 2.2, this estimator is unbiased when (p_{00}, p_{11}) lies on the line through $(1-\alpha, \alpha)$ and $(1, 1)$ [15]. Furthermore, the bias is independent on the size of the target population. A larger target population does therefore not cancel the bias.

Furthermore, better classification probabilities do not lead directly to better aggregated statistics. In fact, better classification probabilities could even lead to a worse quality of the aggregated statistic. This phenomenon can be illustrated with the contourplots in Figure 3.2. Let C_1 be a classifier with classification probabilities $(p_{00} = 0.7, p_{11} = 0.8)$ and let C_2 be a classifier with classification probabilities $(p_{00} = 0.8, p_{11} = 0.8)$. The plots show that C_2 has a lower RMSE than C_1 when α is equal to 0.3 or 0.5, but on the other hand has C_1 a lower RMSE than C_2 when α is equal to 0.7. Therefore we can conclude that a classifier with high classification probabilities does not always have a lower RMSE for the base rate than a classifier with low classification probabilities. The reason why this can happen is the imbalance between the classification er-

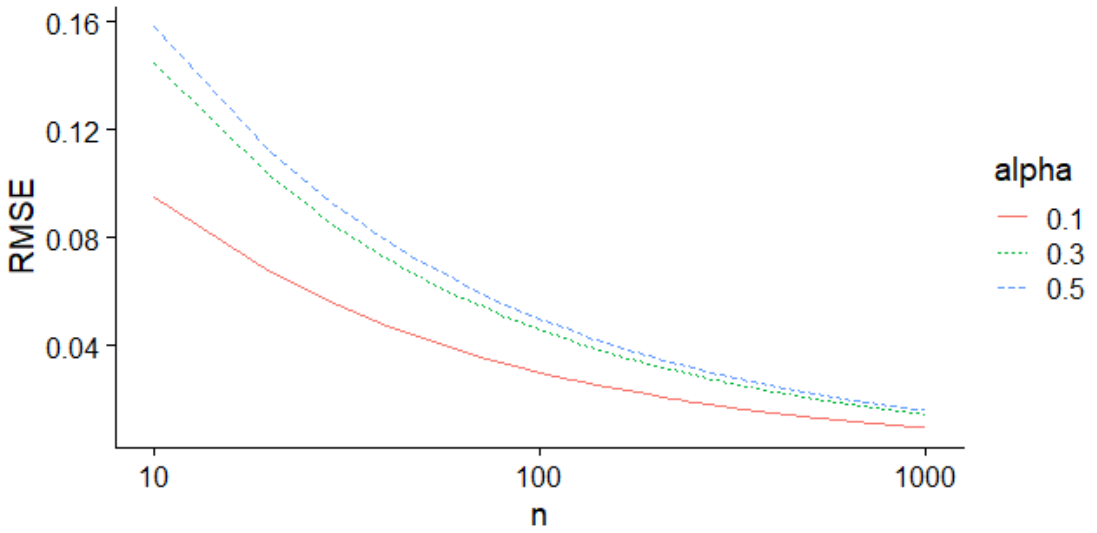


Figure 3.1: RMSE of the Baseline estimator against the size of the test set n for different values of *alpha*. An α of 0.1 has the same RMSE-curve as a α of 0.9, see Formula 2.5. Note that the x-axis is on a logarithmic scale.

rors. Improving classification probabilities for one class can lead to more imbalance of the errors between the classes, wherefore the bias of the estimator increases.

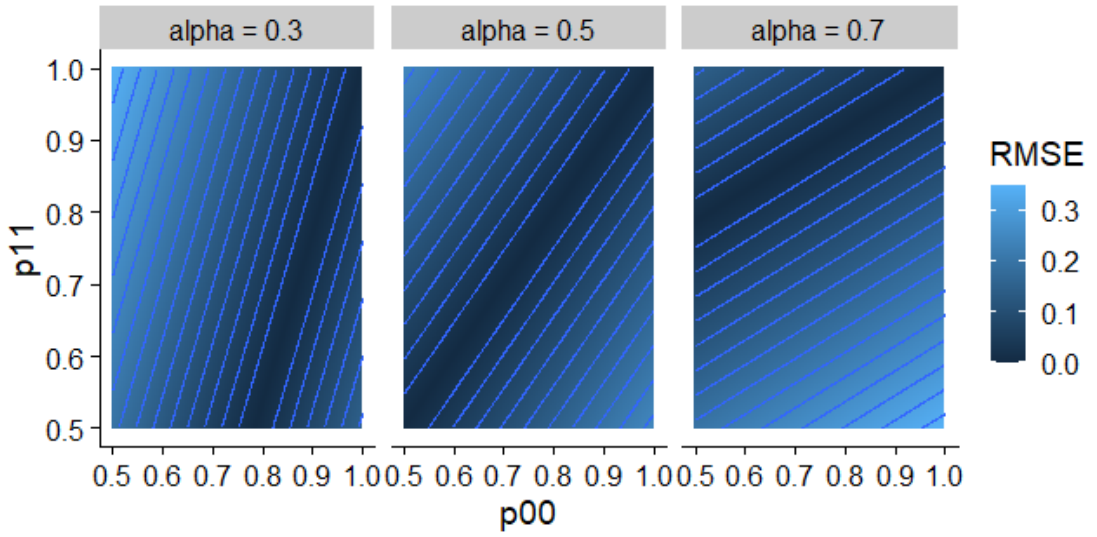


Figure 3.2: Contourplot of the RMSE for the Classify-and-count estimator against the classification probabilities for different values of α.

3.1.3 Subtracted-bias estimator

In contrast to the unbiased baseline estimator and the invariant classify-and-count estimator, the subtracted-bias estimator has a trade-off between bias and variance. As shown in section 2.2.3, the subtracted bias estimator has always an lower or equal bias than the classify-and-count estimator, but always has equal or more variance than the classify-and-count estimator. From the contourplots in Figure 3.3, we can extract some valuable information about the RMSE-curve. The shape of the RMSE-curve is on the first sight an upward opening parabola with the lowest values for high classification probabilities. Lower classification probabilities lead to a higher value for the RMSE and the slope of the curve is steeper in the direction of p_{00} when $\alpha < 0.5$ and steeper in the direction of p_{11} when $\alpha > 0.5$. It can also be noticed that the RMSE is low on the line where the classify-and-count estimator, and therefore the subtracted-bias estimator, is unbiased. However, for the subtracted-bias estimator, the RMSE is not equal to zero on this line, but increases when the classification probabilities decrease. Furthermore, a larger test set leads in general to less variance of this estimator, and thus a lower RMSE. The last insight from this plot is that the RMSE of the subtracted-bias estimator is less sensitive for changes in classification probabilities than the classify-and-count estimator. This is the result of the reduction in misclassification bias by this estimator.

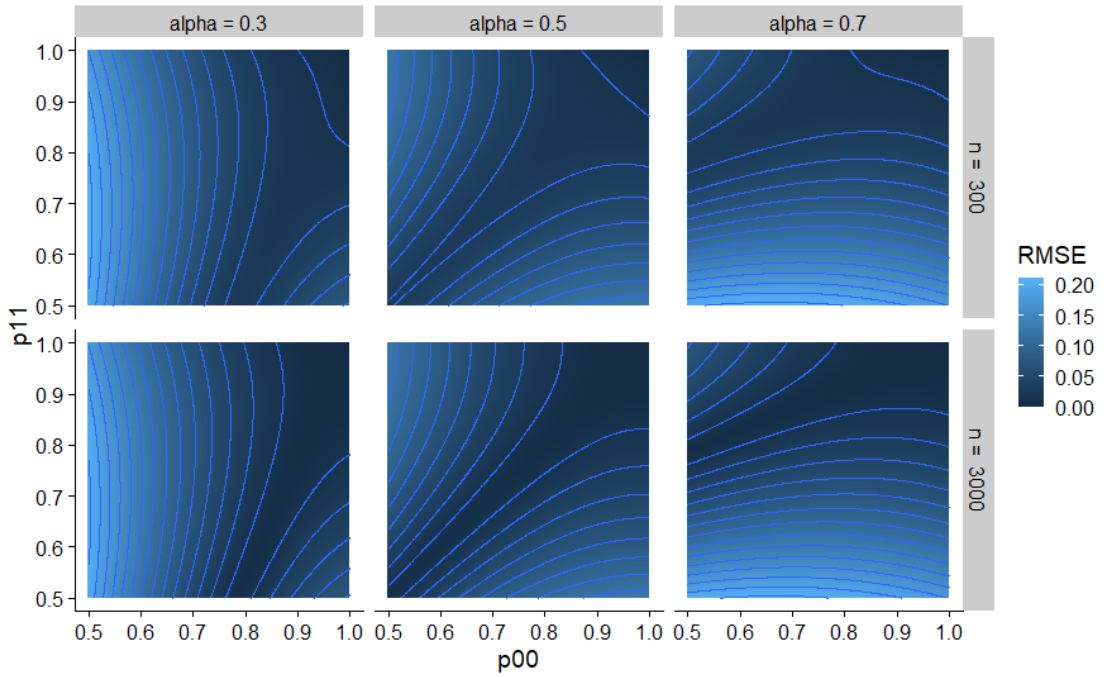


Figure 3.3: Contourplot of the RMSE for the Subtracted-bias estimator against the classification probabilities for different values of α and size of the test set.

3.1.4 Misclassification estimator

The misclassification estimator is an asymptotically unbiased estimator, which makes use of both the target population and the test set. We use the test set to approximate the classification probabilities and we use the target population to obtain the classify-and-count estimator. Because

the misclassification estimator is asymptotically unbiased, hence the biggest contributor to the RMSE is the variance. However, the bias is not negligibly small for test sets with a small sample size. The bias is relatively large for big differences between the classification probabilities and for low classification probabilities in general, see 2.16.

The variance of the misclassification estimator depends heavily on both the classification probabilities and the size of the test set. Very low classification probabilities can lead to extreme variances, see the denominator of (2.17) and the contourplot of Figure 3.4. When $p_{00} + p_{11}$ is close to 1, the variance of the misclassification estimator is large. The term $(\boldsymbol{P}^T)^{-1}$ causes this phenomenon, because the values of this matrix tend to take extreme values when the difference between p_{00} and $1 - p_{00}$, and the difference between p_{11} and $1 - p_{11}$, is small. It is even impossible to compute if $p_{00} = p_{11} = 0.5$, due to singularity. Thus, the value of the variance increases asymptotically when the classification probabilities decrease to $p_{00} = 0.5$ and $p_{11} = 0.5$. This can also be shown in the contourplot, where the RMSE tend to increase faster when the classification probabilities are lower. Note that the range of the classification probabilities only go to 0.6 instead of 0.5, because of readability reasons. Furthermore, a larger test set decreases the RMSE quite substantially, but it still performs badly for very low classification probabilities. The value of the base rate α determines the shape of the RMSE-curve: the RMSE increases slower in the direction of p_{00} than in the direction of p_{11} when $0 < \alpha < 0.5$ and increases faster in the direction of p_{00} than in the direction of p_{11} when $0.5 < \alpha < 1$.

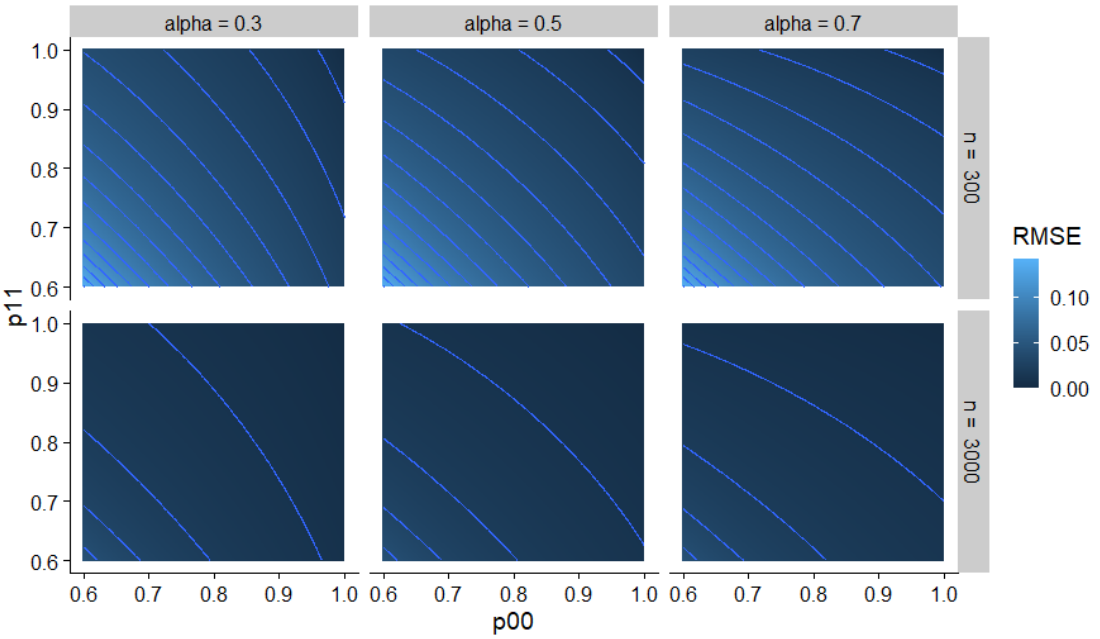


Figure 3.4: Contourplot of the RMSE for the Misclassification estimator against the classification probabilities for different values of α and size of the test set.

3.1.5 Calibration estimator

The calibration estimator is an unbiased estimator, which makes use of both the target population and the test set. The variance equation of this estimator is quite complex and depends on

many parameters, see (2.20). The reason why it is so complex compared to the misclassification estimator, is the dependency between the parameters of the calibration estimator. Recall that the misclassification estimator uses classification probabilities, which are calculated per row and that the calibration estimator uses calibration probabilities, which are calculated per column. The rows of the target population are independent from each other, but the columns are dependent from each other. We took this dependence into account, hence the variance equation of the calibration estimator becomes more complex.

Following the contourplot of Figure 3.5, we observe the following. First, the rate of the increase in the RMSE of the calibration estimator decelerates when the misclassification probabilities decrease, which is in contrast to the misclassification estimator where the rate of the increase in the RMSE accelerates when the misclassification probabilities decrease. The calibration estimator tends to perform well over the whole range of misclassification probabilities, even for bad classifiers. Furthermore, it can be shown from the plots that the RMSE is heavily dependent on the size of the test set. Larger test sets decrease the size of the RMSE drastically. Last, the value of the base rate α determines the shape of the RMSE-curve of the calibration estimator similarly to the RMSE-curve of the misclassification estimator: the RMSE increases slower in the direction of p_{00} than in the direction of p_{11} when $0 < \alpha < 0.5$ and increases faster in the direction of p_{00} than in the direction of p_{11} when $0.5 < \alpha < 1$.

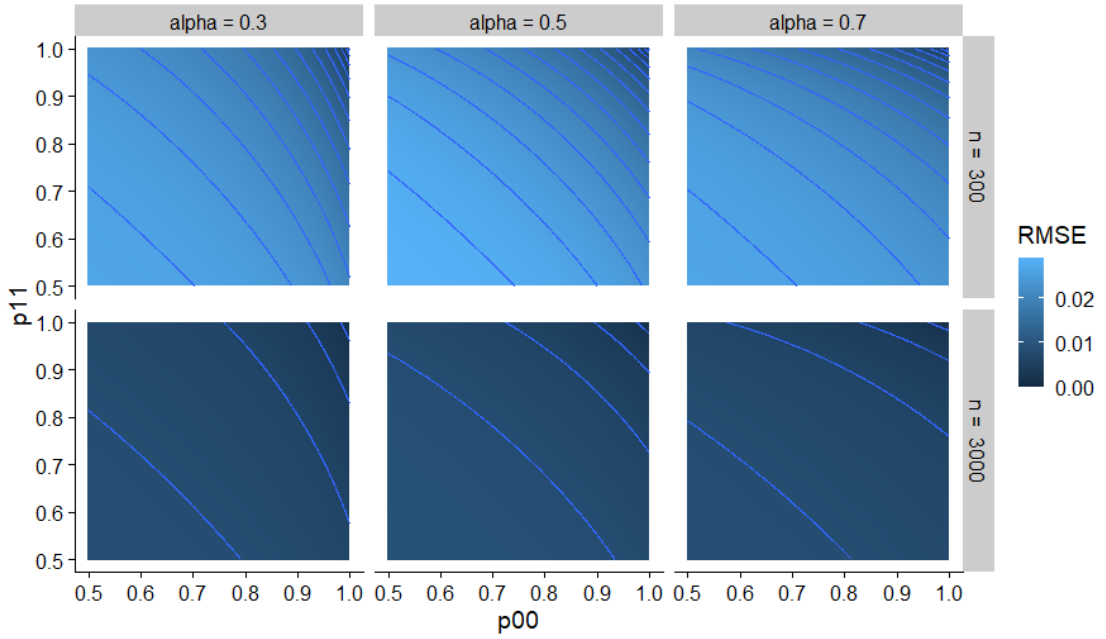


Figure 3.5: Contourplot of the RMSE for the Calibration estimator against the classification probabilities for different values of α and size of the test set.

3.2 Comparing the estimators

Now that we know the behaviour of the estimators separately, it is time to compare the sampling distribution of the estimators. Besides the mathematical equations, we perform a simulation study to check if the mathematical assumptions and approximations indeed model the correct

distribution and compare the distributions of the estimators with each other. We present two simple simulation studies that are computed according to algorithm 1.

Algorithm 1: Simulation study of the estimators

Result: Compute R estimates of each estimator
Set initial values for $\alpha, p_{00}, p_{11}, n$ and N
while *iter* <= R **do**
  Compute contingency table of target population with size N and base rate α.
  Sample test set without replacement from target population of size n
  Apply the estimators on the test set and target population
  $iter = iter + 1$
end
return *estimates*

In the first simulation study, we consider a class-balanced dataset ($\alpha = 0.5$), with a small test set of size $n = 1000$, a large population dataset of $N = 3 \times 10^5$ and a rather poor classifier having classification probabilities $p_{00} = 0.6$ and $p_{11} = 0.7$. We deliberately choose $p_{00} \neq p_{11}$, as otherwise the classify-and-count estimator would be unbiased: (p_{00}, p_{11}) would be on the line between $(1 - \alpha, \alpha)$ and $(1, 1)$, see also (2.8).

Table 3.1 summarizes the bias, variance and root mean square error (RMSE), computed using the analytic expression presented in Section 2.2. The classify-and-count estimator is an estimator with a high RMSE. Despite having the lowest variance, the estimator is highly biased and therefore has a high RMSE. The subtracted-bias estimator has a lower (absolute) value for the bias, but a higher variance. Nevertheless, the RMSE for the subtracted-bias estimator is lower than the RMSE for the classify-and-count estimator. The bias can almost be removed with the misclassification estimator. As discussed earlier in Section 3.1.4, the variance is very large for classifiers with low values for p_{00} and p_{11}, due to singularity of the inverted row-normalized confusion matrix $\mathbf{P}$. In this case, the large variance causes that the misclassification estimator has the highest RMSE of all estimators, despite having a very low bias. Last, the baseline estimator and the calibration estimator are both unbiased and have a similar variance. The calibration estimator has the lowest RMSE out of the two, so it performs better than the baseline estimator. This is in contrast to all the other estimators, who perform worse than the baseline estimator.

Table 3.1: A comparison of the bias, variance and RMSE of each of the five estimators for α, where $\alpha = 0.5$, $p_{00} = 0.6$, $p_{11} = 0.7$, $n = 1000$ and $N = 3 \times 10^5$.

Estimator	*Symbol*	Bias $\times 10^{-2}$	Variance $\times 10^{-4}$	RMSE $\times 10^{-2}$
Baseline	$\hat{\alpha}_a$	0.000	2.500	1.581
Classify-and-count	$\hat{\alpha}^*$	5.000	0.000	5.000
Subtracted-bias	$\hat{\alpha}_b$	3.500	2.365	3.823
Misclassification	$\hat{\alpha}_p$	-0.033	25.025	5.003
Calibration	$\hat{\alpha}_c$	0.000	2.275	1.508

To obtain additional insights of the sampling distributions, we perform a simulation study according to Algorithm 1. We create $R = 10000$ target populations and we sample from each target population a test set. Following, we apply the estimators on the available data and visualise it with a boxplot, see Figure 3.6. The theoretical results are similar to the simulation

study. The classify-and-count estimates are highly biased with a low variance; the subtracted-bias estimates are less biased, but with a higher variance. The misclassification estimator has a very high variance, with extreme outliers on both sides. The baseline estimator and the calibration estimator both produce stable estimates around the true α.

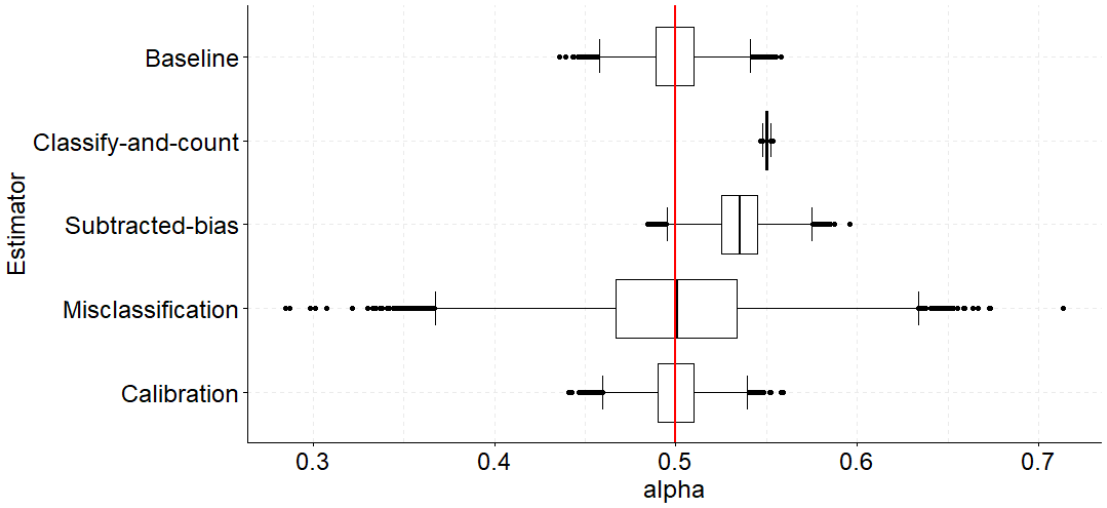


Figure 3.6: The boxplots show the sampling distribution of the estimators for α, where $\alpha = 0.5$, $p_{00} = 0.6$, $p_{11} = 0.7$, $n = 1000$ and $N = 3 \times 10^5$. The true value of α is highlighted by a vertical line.

In the second simulation study, we consider a highly imbalanced dataset with $\alpha = 0.98$. We again assume a test set of size $n = 1000$ and a target population of size $N = 3 \times 10^5$. On the other hand, we assume that the classifier has classification probabilities $p_{00} = 0.94$ and $p_{11} = 0.97$. Similarly to the first simulation study, Table 3.2 summarizes the bias, variance and RMSE of each of the estimators and Figure 3.7 shows the sampling distribution of each of the estimators. Again, the classify-and-count estimator is highly biased with low variance, while the subtracted-bias estimator is less biased, but has more variance. Even with high classification probabilities, both estimators are biased. The misclassification estimator has now less variance than in the previous simulation study, because of the higher classification probabilities, but performs still worse than the baseline estimator. The calibration estimator again performs the best out of the five estimators and is still the only estimator that performs better than the baseline. Last, it can be noticed that both the misclassification estimator and the subtracted-bias estimator sometimes produce estimates that exceed 1. It is obvious that these values cannot occur in the target population. For the misclassification estimator, this effect gets bigger when $p_{00} + p_{11}$ gets closer to 1.

3.3 Finding the optimal estimator

Now that we know the properties of each of the estimators, we can find the optimal estimator, i.e., the estimator with the lowest RMSE, for every combination of (α, p_{00}, p_{11} and n). In the first situation, we suppose that (p_{00}, p_{11}) is close to the line in the plane through the points $(1 - \alpha, \alpha)$ and $(1, 1)$. As noted before, it implies that the classify-and-count estimator has then low bias. Consequently, the subtracted-bias estimator has low bias as well. We can visualize

Table 3.2: A comparison of the bias, variance and RMSE of each of the five estimators for α, where $\alpha = 0.98$, $p_{00} = 0.94$, $p_{11} = 0.97$, $n = 1000$ and $N = 3 \times 10^5$.

Method	*Symbol*	Bias $\times 10^{-2}$	Variance $\times 10^{-4}$	RMSE $\times 10^{-2}$
Baseline	$\hat{\alpha}_a$	0.000	0.196	0.443
Classify-and-count	$\hat{\alpha}^*$	-2.820	0.000	2.820
Subtracted-bias	$\hat{\alpha}_b$	-0.254	0.307	0.609
Misclassification	$\hat{\alpha}_p$	-0.003	0.359	0.599
Calibration	$\hat{\alpha}_c$	0.000	0.127	0.356

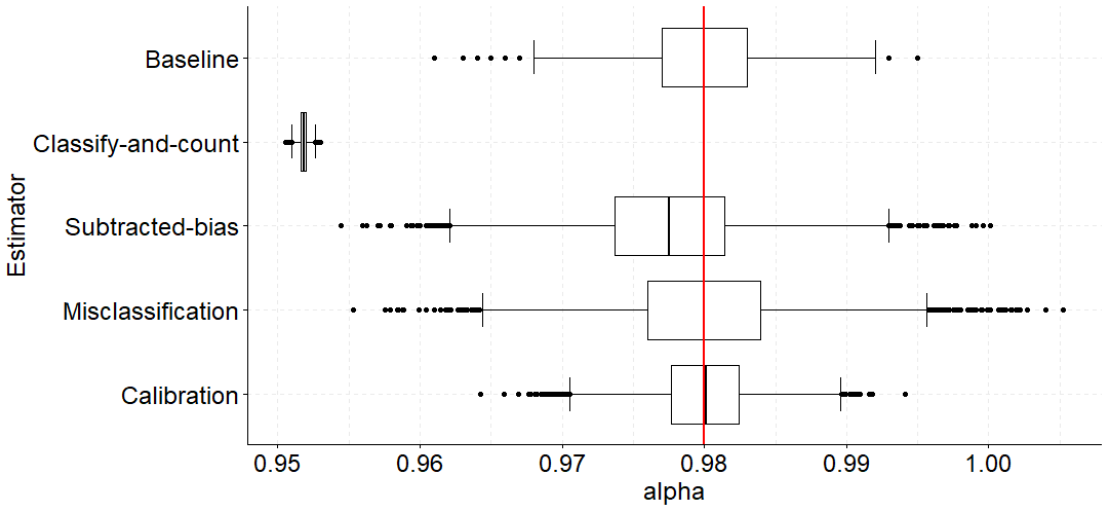


Figure 3.7: The boxplots show the sampling distribution of the estimators for α, where $\alpha = 0.98$, $p_{00} = 0.94$, $p_{11} = 0.97$, $n = 1000$ and $N = 3 \times 10^5$. The true value of α is highlighted by a vertical line.

which estimator has the lowest RMSE with Figure 3.8, where we compare the RMSE of the consistent calibration estimator with the two biased methods. The two biased estimators have both a low RMSE around the specific line, but they perform relatively worse when the size of the test set increases. In general, the biased methods perform relatively well when (1) the classification probabilities are close to the line through the points $(1 - \alpha, \alpha)$ and $(1, 1)$ and (2) when there is a large class-imbalance, i.e., an α close to 0 or 1, and the classification probability for the largest group is high.

As we have seen in Table 3.1 and Table 3.2, the calibration estimator competes with the baseline estimator as the estimator with the lowest RMSE. The baseline estimator is independent of the classification probabilities, while the calibration estimator has a higher RMSE when the classification probabilities decrease. For finite test sets, the baseline estimator has always a lower RMSE than the calibration estimator when $p_{00} = p_{11} = 0.5$. However, for every α and n, there must exist a curve in the $(p_{00}.p_{11})$-plane beyond which the calibration estimator performs better than the baseline estimator. The left-hand panels in Figure 3.9 show this curve for $\alpha = 0.2$

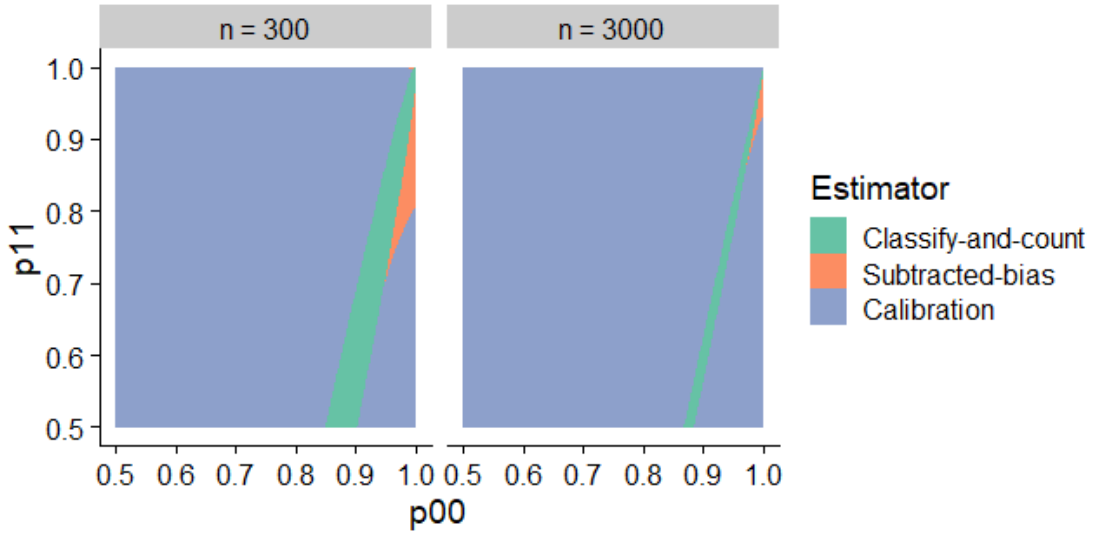


Figure 3.8: For each coordinate (p_{00}, p_{11}), the depicted color indicates which estimator has the lowest RMSE, considering only the classify-and-count estimator (green), the subtracted-bias estimator (orange) and the calibration estimator (purple). In the left panel , we have set $\alpha = 0.2$ and $n = 300$, whereas $\alpha = 0.2$ and $n = 3000$ in the right panel. The red and green regions are smaller in the right panel, as the variance of the calibration estimator is decreasing in n, while the bias of the classify-and-count estimator and of the subtracted-bias estimator do not depend on n.

and two different values for n. For larger values of n, the curve where the calibration estimator performs better than the baseline estimator gets closer to $p_{00} = p_{11} = 0.5$ and therefore covers a larger area in the (p_{00}, p_{11})-plane.

Furthermore, we could have seen in Table 3.1 and Table 3.2 that the misclassification estimator only performs well if the classification probabilities are high. This can be confirmed by the formulas for the bias and the variance, which both show singularity at $p_{00} + p_{11} = 1$ (see (2.16) and (2.17)). The right-hand panels in Figure 3.9 show, for $\alpha = 0.2$ and two different values for n, the curve in the (p_{00}, p_{11})-plane that the misclassification estimator has a lower RMSE than the baseline estimator. Observe that increasing the size of the test set, does not have much impact on the position of the curve. This is in contrast to the calibration estimator where the test set has more impact on the position on the curve. The area where the baseline estimator performs better than the misclassification estimator is substantially large. Only for high classification probabilities in general or high classification probabilities in the largest group, the misclassification estimator performs better than the baseline estimator.

The final analysis of the thesis is to compare the calibration estimator and the misclassification estimator for high values of p_{00} and p_{11}. In Theorem 4 it is proven that, for all possible combinations of α and sufficiently large n, the MSE of the calibration estimator is consistently lower than that of the misclassification estimator.

Theorem 4. *Let $\widetilde{MSE}[\hat{\alpha}_p]$ and $\widetilde{MSE}[\hat{\alpha}_c]$ denote the approximate mean squared errors, up to terms of order $1/n$, of the misclassification estimator and the calibration estimator, respectively.*

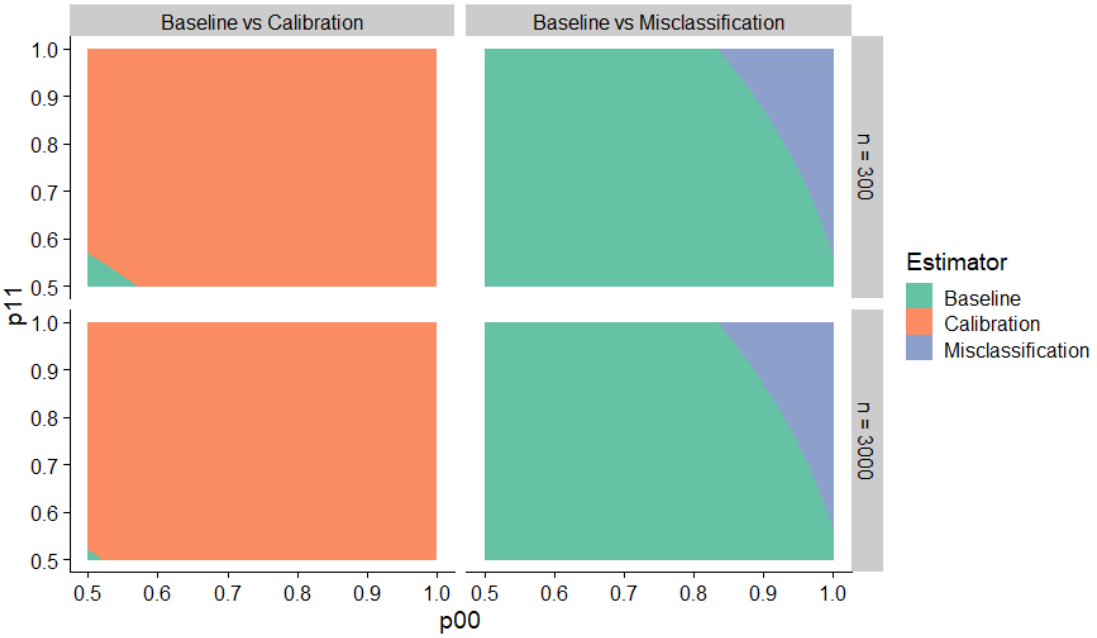


Figure 3.9: For each coordinate (p_{00}, p_{11}), the depicted color indicates which estimate has the lowest RMSE, considering only the baseline estimator (green), the calibration estimator (orange) and the misclassification estimator (purple). The top-row panels consider $\alpha = 0.2$ and $n = 300$, while the bottom-row panels consider $\alpha = 0.2$ and $n = 3000$.

It holds that:

$$\widetilde{MSE}[\hat{\alpha}_p] - \widetilde{MSE}[\hat{\alpha}_c] = \frac{\Big[(1-\alpha)p_{00}(1-p_{00}) + \alpha p_{11}(1-p_{11})\Big]^2}{(p_{00}+p_{11}-1)^2\beta(1-\beta)}, \tag{3.1}$$

in which $\beta \coloneqq (1-\alpha)(1-p_{00}) + \alpha p_{11}$.

Proof. See the Appendix. □

Thus, neglecting terms of order $1/n^2$ and higher, the result implies that the calibration estimator has a lower mean squared error than the misclassification estimator, except that both are equal if and only if $p_{00} = p_{11} = 1$. (Note that $0 < \beta < 1$.)

We do remark that the difference in MSE is large in particular for values of p_{00} and p_{11} close to $\frac{1}{2}$. More specifically, it diverges when $p_{00} + p_{11} \to 1$. It is the result of the misclassification estimator having a singularity at $p_{00} + p_{11} = 1$ (see Equation (2.17)), while the variance of the calibration estimator is bounded. An unpleasant consequence of the singularity at $p_{00} + p_{11} = 1$ is that, for fixed n and α, the probability that $\hat{\alpha}_p$ takes values outside the interval $[0, 1]$ increases as $p_{00} + p_{11} \to 1$; see [12] for a discussion and a possible solution.

Chapter 4

Concept Drift

4.1 Theory

In the previous chapters, we assumed that our model contains a fixed base rate. In terms of official statistics, we create a test set and apply an machine learning algorithm to estimate a base rate at time 0. In an ideal situation, we would create a new test set every time we apply the algorithm to estimate the base rate. At National Statistical Institutes, many statistics are produced on a weekly, monthly or annual basis. It is expensive to create a new test set over and over again and therefore we prefer to recycle the test set. We cannot assume that the test set at time t is still a simple random sample from the target population, because the underlying distribution has been changed. In literature, this is called *concept drift* [16].

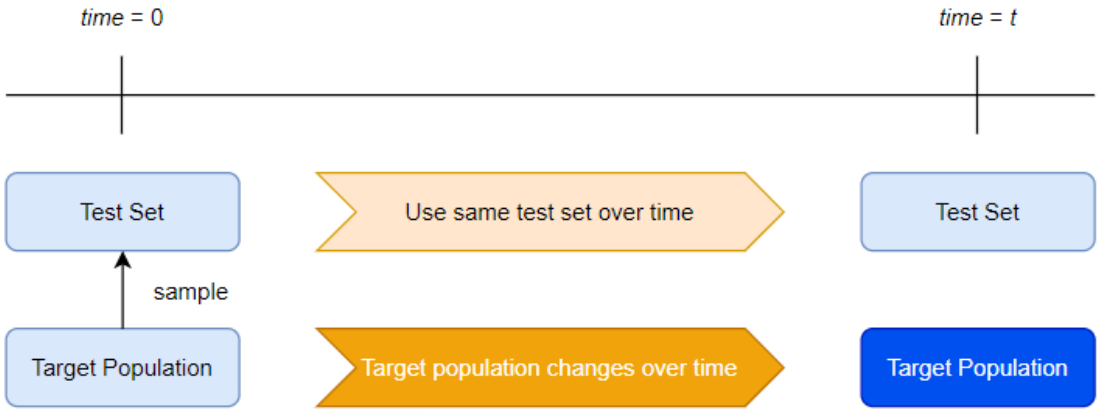


Figure 4.1: Conceptualisation of concept drift. The test set is a simple random sample from the target population at time 0, but is not fully representative on the target population on time t.

The term concept drift is widely used in literature and authors can interpret it differently. In our study, we assume that objects can disappear from the target population and new objects can be included to the target population. Moreover, we assume that the classification probabilities are equal over time, i.e., the distribution conditioned on the true class of the model would not change. However, the base rate of the target population at time 0, α can be different than the base rate of the target population at time t, α'. To be consistent, we define U as the target population with base rate α. The test set is sampled from U at time 0. We define U' as the

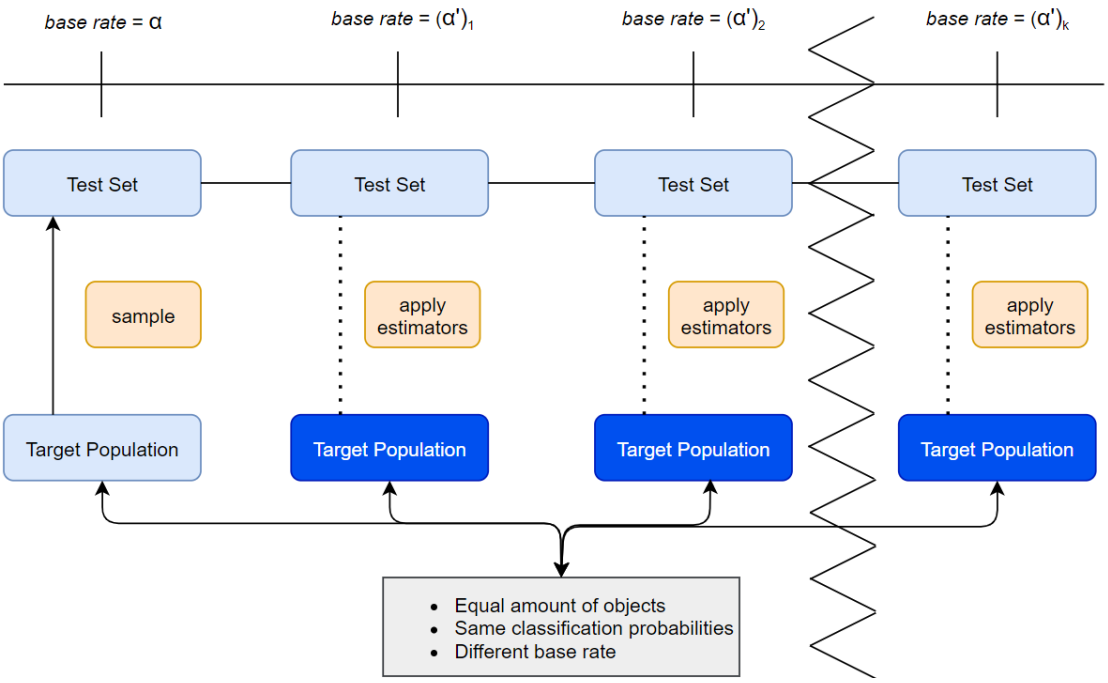


Figure 4.2: Conceptualisation under concept drift while using new target populations. The test set is sampled once and used in combinations with the different target populations to compute estimates of the base rate.

target population with base rate α'. To estimate α', we use both target population U' and the test set sample from target population U. The aim of this chapter is comparing the behaviour of the estimators under concept drift and to see if the properties of the estimators under concept drift change compared to the study under a fixed base rate.

In the simulation study, our approach consists of sampling complete new target populations with different base rates, see Figure 4.2. The other parameters, like the size of the target population and the classification probabilities, remain the same in this simulation study. We can use the sampled test set in combination with the new target populations to compute new estimates of the base rate. This makes it possible to compare the effects of the estimators under concept drift in an efficient way.

More precisely, the algorithm works as follows. First, sample a test set from the target population U in the same way as in Algorithm (1). Accordingly, compute k new target populations U' with a different base rate α', but with an equal amount of objects and the same classification probabilities. We can apply our estimators by combining the new target populations with the original test set. This process will be replicated R times, in order to approximate the RMSE of the estimators.

In the next subsections, we will discuss the theoretical effect of concept drift on the estimators, based on the expressions in section (2.2). In section (4.2), we will apply the algorithm and compare the estimators based on a simulation study.

4.1.1 Baseline estimator and Classify-and-count estimator

First, we will discuss the effects of concept drift under the baseline estimator and the classify-and-count estimator. The expression of the baseline estimator remains the same under concept drift, see (2.3).

$$\hat{\alpha}'_a = \frac{n_{1+}}{n} \tag{4.1}$$

The effect of concept drift on the baseline estimator is large. Under a fixed base rate, the baseline estimator is an unbiased estimator. However, this estimator can, obviously, be biased under concept drift. If the difference between the base rate of the new target population U' and the target population U that is used to sample the test set is large, the baseline estimator $\hat{\alpha}_a$ is highly biased. It is straight forward to see that the bias of this estimator is equal to the difference between α and α', see (4.2). It depends therefore on the variability of the base rate over time whether this estimator is usable. The variance of this estimator stays equal over time, because the same test is the only source of information that is used for this estimator, see (4.3).

$$B[\hat{\alpha}'_a] = \alpha - \alpha' \tag{4.2}$$

$$V[\hat{\alpha}'_a] = \frac{\alpha(1-\alpha)}{n} \tag{4.3}$$

On the other hand, concept drift does not change the RMSE, up to $O(\frac{1}{N})$, of the classify-and-count estimator. This makes sense, because the classify-and-count estimator does only rely on the target population and does therefore not rely on the test set, see (2.6). Similar to the study with the fixed base rate, we use this classify-and-count estimator to compute the estimates of the final three estimators, see (4.4).

$$(\hat{\alpha}')^* = \frac{(N_{+1})'}{N} \tag{4.4}$$

$$B[(\hat{\alpha}')^*] = \alpha'(p_{11} - 1) + (1 - \alpha')(1 - p_{00}) \tag{4.5}$$

$$V[(\hat{\alpha}')^*] = O(\frac{1}{N}) \tag{4.6}$$

4.1.2 Subtracted-bias estimator

The change between the RMSE under U and U' for the subtracted-bias estimator is somewhat harder to analyse, see (4.7). In the model assumptions, we have noticed that estimates of the classification probabilities are always unbiased, given every proper value of the base rate α. However, the variance of the classification probabilities does depend on the base rate. Therefore, the variance equation of $\hat{\alpha}'_b$ is different from the variance equation of $\hat{\alpha}_b$, see (2.13) and (4.9).

$$\hat{\alpha}'_b = (\hat{\alpha}')^*(3 - \hat{p}_{00} + \hat{p}_{11}) - (1 - \hat{p}_{00}) \tag{4.7}$$

In (4.8), we observe that the change in bias is dependent on the new base rate α'. When α' is bigger than α, the bias under target population U' is smaller than the bias under target population U and when α' is smaller than α, the bias target population U' is larger than the bias under target population U. The difference of the variance between α_b and α'_b is dependent on the change in the classify-and-count estimator and the change of the new base rate α'. Substituting

α' in (2.13) gives

$$B[\hat{\alpha}'_b] = (1-p_{00})(2-p_{00}-p_{11}) - \alpha'(p_{00}+p_{11}-2)^2 \tag{4.8}$$

$$\begin{aligned} V[\hat{\alpha}'_b] = &\frac{[\alpha'(p_{00}+p_{11}-1)-p_{00}]^2 p_{00}(1-p_{00})}{n(1-\alpha)}\left(1+\frac{\alpha}{n(1-\alpha)}\right) \\ &+ \frac{[\alpha'(p_{00}+p_{11}-1)+(1-p_{00})]^2 p_{11}(1-p_{11})}{n\alpha}\left(1+\frac{1-\alpha}{n\alpha}\right) \\ &+ O\left(\max\left[\frac{1}{n^3},\frac{1}{N}\right]\right). \end{aligned} \tag{4.9}$$

From 4.9, we can see that $V[\hat{\alpha}'_b] > V[\hat{\alpha}_b]$ when $\alpha' > \alpha$ and vice versa. We clearly see that the variance of $\hat{\alpha}'_b$ depends on both α and α'.

4.1.3 Misclassification estimator

The equation of the misclassification estimator under concept drift is almost the same as the equation of the misclassification estimator under a fixed base rate, see (4.10). The derivation of the bias and variance of $\hat{\alpha}_p$ under concept drift are more complex and are derived in [11].

$$\hat{\alpha}'_p = \frac{(\hat{\alpha}')^* + \hat{p}_{00} - 1}{\hat{p}_{00} + \hat{p}_{11} - 1} \tag{4.10}$$

$$B[\hat{\alpha}'_p] = \frac{p_{00}-p_{11}}{n(p_{00}+p_{11}-1)} + \frac{\alpha'-\alpha}{n(p_{00}+p_{11}-1)^2}\cdot\left(\frac{p_{11}(1-p_{11})}{\alpha}+\frac{p_{00}(1-p_{00})}{1-\alpha}\right) + O(\frac{1}{n^2}) \tag{4.11}$$

$$\begin{aligned} V[\hat{\alpha}'_p] = &\frac{1}{n(p_{00}+p_{11}-1)^2}\cdot\Big[(1-\alpha)p_{00}(1-p_{00}) + \alpha p_{11}(1-p_{11}) \\ &+ 2(\alpha'-\alpha)(p_{00}-p_{11})(p_{00}+p_{11}-1) + (\alpha'-\alpha)^2\cdot\left(\frac{p_{11}(1-p_{11})}{\alpha}+\frac{p_{00}(1-p_{00})}{1-\alpha}\right)\Big] \\ &+ O(\frac{1}{n^2}) \end{aligned} \tag{4.12}$$

From (4.10), we see that the bias of $\hat{\alpha}'_p$ is, up to $O(\frac{1}{n})$, equal to the bias of $\hat{\alpha}'_p$. Therefore, the change in bias is very small and vanishes when the size of the test set n is large. We cannot say that the variance of $\hat{\alpha}'_p$ is bigger than the variance of $\hat{\alpha}_p$ when $\hat{\alpha}'_p > \hat{\alpha}_p$ and vice versa. This is dependent on the classification probabilities p_{00} and p_{11} and base rate α. The other assumptions are still valid: also under concept drift, the variance is large when $p_{00}+p_{11}$ gets close to 1.

4.1.4 Calibration estimator

Like the misclassification estimator, the equation of the calibration estimator under concept drift is almost the same as the equation of the calibration estimator under a fixed base rate, see equation (4.13). Moreover, the derivation of the bias and variance of $\hat{\alpha}_c$ under concept drift are more complex and are derived in [11]. We can define $\beta := (1-\alpha)(1-p_{00}) + \alpha p_{11}$ and

$T := (1-\alpha)p_{00}(1-p_{00}) + \alpha p_{11}(1-p_{11})$.

$$\hat{\alpha}'_c = (1-(\hat{\alpha}')^*)\frac{n_{10}}{n_{+0}} + (\hat{\alpha}')^*\frac{n_{11}}{n_{+1}} \tag{4.13}$$

$$B[\hat{\alpha}'_c] = (\alpha-\alpha')\frac{T}{\beta(1-\beta)} + O(\frac{1}{n^2}) \tag{4.14}$$

$$V[\hat{\alpha}'_c] = \frac{\alpha(1-\alpha)}{n}\left[\frac{T}{\beta(1-\beta)} + 2(\alpha'-\alpha)(p_{00}+p_{11}-1)\left(\frac{p_{11}(1-p_{00})}{\beta^2} - \frac{p_{00}(1-p_{11})}{(1-\beta)^2}\right) + (\alpha'-\alpha)^2(p_{00}+p_{11}-1)^2\left(\frac{p_{11}(1-p_{00})}{\beta^3} - \frac{p_{00}(1-p_{11})}{(1-\beta)^3}\right)\right] + O(\frac{1}{n^2}) \tag{4.15}$$

While the RMSE of the misclassification estimator changes mildly under concept drift, the RMSE of the calibration estimator changes a lot under concept drift. The estimator is biased under concept drift, in contrast to the situation under a fixed base rate where this estimator is unbiased. Equation (4.14) shows that the variance increases linear with respect to the difference between α' and α. Furthermore, this bias is dependent on the base rate α and classification probabilities p_{00} and p_{11}. We cannot say that the variance of $\hat{\alpha}'_p$ is bigger than the variance of $\hat{\alpha}_c$ when $\hat{\alpha}'_c > \hat{\alpha}_c$ and vice versa. This is dependent on the classification probabilities p_{00} and p_{11} and base rate α.

4.2 Comparing the estimators

Now that we know the general properties of the estimators, we can compare the estimators with each other. We will perform two new simulation studies similar to the simulation studies in section 3.2. The starting values remain the same, but we will also show the effect of concept drift for values α' around the original base rate α. For both simulation studies, we present tables for the bias, variance and RMSE and we show boxplots that visualise the distributions for each combination of α' and estimators.

The first simulation study starts with a balanced data set, i.e. $\alpha = 0.5$. The algorithm has fairly low classification probabilities of $p_{00} = 0.6$ and $p_{11} = 0.7$. The test set of size $n = 1000$ is small compared to the size of the target population $N = 3 \times 10^5$. The difference from the prior simulation study, is that we add concept drift. Besides plotting the boxplot given these parameters and α, we also compute the boxplots for other values for the base rate α', see Figure 4.3. Recall that we only compute a test set for $\alpha = 0.5$ and carry this test set over to the other target populations with different base rates. We can also estimate the bias, variance and RMSE with these simulations, presented in Table 4.1.

First, we observe that the baseline estimator is highly biased when the base rate changes. We can clearly see that the median value of each boxplot can be computed by subtracting α' from α. Also, in Table 4.1 is shown that the bias increases with the same rate as the change in α'. The variance remains the same over each base rate. Next, the classify-and-count estimator is not affected by concept drift, because it does not use the values from the test set. The bias could be very high, while the variance is negligible small. Accordingly, we see that the subtracted-bias estimators cannot solve a lot of bias, because the classification probabilities are low. It seems that the variance remains fairly constant over the values of α'. Moreover, we can say from the plots that the misclassification estimators has a low amount of bias, also under concept drift. However, due to the low classification probabilities, these estimators have a high amount of variance. Finally, the calibration estimator seems to show the same behaviour as the baseline estimator. A small shift in α' leads to a big amount of bias, nearly the same amount as the change in α'. The variance of this estimator remains stable over time. We observe that not a single estimator is able to have a low RMSE for every value of α', which is a severe problem.

Table 4.1: A comparison of the bias, variance and RMSE of each of the five estimators for α', where $\alpha = 0.5$, $p_{00} = 0.6$, $p_{11} = 0.7$, $n = 1000$ and $N = 3 \times 10^5$. The test set is drawn for $\alpha = 0.5$, values in italic.

		Bias $\times 10^{-2}$ for values of α'						
Method	*Symbol*	0.41	0.44	0.47	*0.5*	0.53	0.56	0.59
Baseline	$\hat{\alpha}_a$	9.000	6.000	3.000	*0.000*	-3.000	-6.000	-9.000
Classify-and-count	$\hat{\alpha}^*$	11.300	9.200	7.100	*5.000*	2.900	0.800	-1.300
Subtracted-bias	$\hat{\alpha}_b$	7.910	6.440	4.970	*3.500*	2.030	0.560	-0.910
Misclassification	$\hat{\alpha}_p$	-0.123	-0.093	-0.063	*-0.033*	-0.003	0.027	0.057
Calibration	$\hat{\alpha}_c$	8.182	5.455	2.727	*0.000*	-2.727	-5.455	-8.182

		Variance $\times 10^{-4}$ for values of α'						
Method	*Symbol*	0.41	0.44	0.47	*0.5*	0.53	0.56	0.59
Baseline	$\hat{\alpha}_a$	2.500	2.500	2.500	*2.500*	2.500	2.500	2.500
Classify-and-count	$\hat{\alpha}^*$	0.000	0.000	0.000	*0.000*	0.000	0.000	0.000
Subtracted-bias	$\hat{\alpha}_b$	2.365	2.365	2.365	*2.365*	2.365	2.365	2.365
Misclassification	$\hat{\alpha}_p$	26.436	25.786	25.315	*25.025*	24.915	24.985	25.235
Calibration	$\hat{\alpha}_c$	2.274	2.272	2.272	*2.273*	2.275	2.279	2.284

		RMSE $\times 10^{-2}$ for values of α'						
Method	*Symbol*	0.41	0.44	0.47	*0.5*	0.53	0.56	0.59
Baseline	$\hat{\alpha}_a$	9.138	6.205	3.391	*1.581*	3.391	6.205	9.138
Classify-and-count	$\hat{\alpha}^*$	11.300	9.200	7.100	*5.000*	2.900	0.800	1.300
Subtracted-bias	$\hat{\alpha}_b$	8.058	6.621	5.202	*3.823*	2.547	1.637	1.787
Misclassification	$\hat{\alpha}_p$	5.143	5.079	5.032	*5.003*	4.991	4.999	5.024
Calibration	$\hat{\alpha}_c$	8.320	5.659	3.116	*1.508*	3.117	5.660	8.320

The second simulation study starts with a highly imbalanced data set, i.e. $\alpha = 0.98$. The algorithm has high classification probabilities of $p_{00} = 0.94$ and $p_{11} = 0.97$. The test set of size $n = 1000$ is small compared to the size of the target population $N = 3 \times 10^5$. In this simulation study, the shift in α' is smaller than in the previous simulation study, because we observe the shift around an already high base rate α. The estimates of the bias, variance and RMSE are shown in Table 4.2 and the boxplots are shown in Figure 4.4.

It is important to note that the big outliers of the misclassification estimator are removed from the plot. A handful of the estimates had a difference between 1 and 3 from the true base rate. However, we can still use the plots to interpret the general behaviour of the estimators. We observe that the baseline and the calibration estimator still perform well if the new base rate α' has extreme values and does not differ much from the original base rate α. Even if there is some bias, is still performs better than the misclassification estimator. Even though the misclassification estimator has a low bias when the base rate changes, it has a lot of variance. A lot of estimates are outside the interval $[0, 1]$, which are impossible to obtain in practice. The classify-and-count estimator and the subtracted-bias estimator remain highly biased.

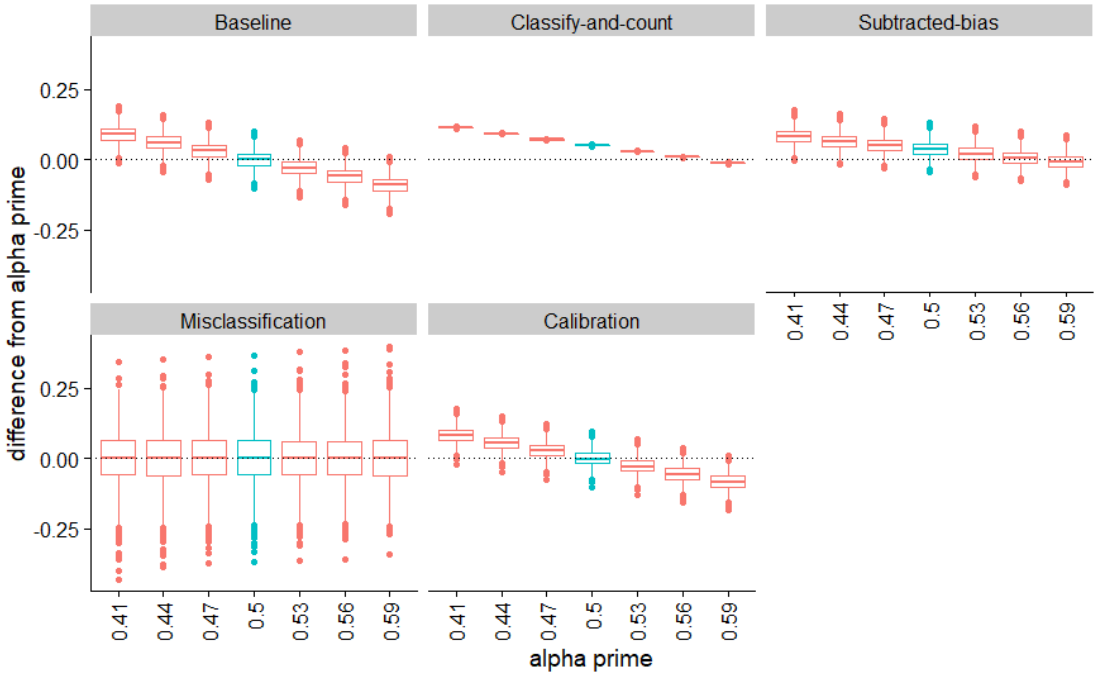


Figure 4.3: Simulation study to observe the change in prediction error under concept drift using boxplots. Five estimators are compared given a initial base rate $\alpha = 0.5$ (blue) and different values of α' (red). The x-axis shows the different base rates and the y-axis shows the distribution of the difference from α'. All the parameters: $p_{00} = 0.6$, $p_{11} = 0.7$, $n = 1000$ and $N = 3 \times 10^5$.

From the simulation study, we can conclude that none of the estimators has consistently a low RMSE in estimating the base rate α' under concept drift. In the next chapter, we provide a solution in the form of a new estimator: the *mixed estimator*.

Table 4.2: A comparison of the bias, variance and RMSE of each of the five estimators for α', where $\alpha = 0.98$, $p_{00} = 0.94$, $p_{11} = 0.97$, $n = 1000$ and $N = 3 \times 10^5$. The test set is drawn for $\alpha = 0.98$, values in italic.

		Bias $\times 10^{-2}$ for values of α'						
Method	*Symbol*	0.93	0.94	0.95	0.96	0.97	*0.98*	0.99
Baseline	$\hat{\alpha}_a$	5.000	4.000	3.000	2.000	1.000	*0.000*	-1.000
Classify-and-count	$\hat{\alpha}^*$	-2.370	-2.460	-2.550	-2.640	-2.730	*-2.820*	-2.910
Subtracted-bias	$\hat{\alpha}_b$	-0.213	-0.221	-0.230	-0.238	-0.246	*-0.254*	-0.262
Misclassification	$\hat{\alpha}_p$	-0.021	-0.017	-0.014	-0.010	-0.007	*-0.003*	0.000
Calibration	$\hat{\alpha}_c$	3.231	2.585	1.939	1.292	0.646	*0.000*	-0.646

		Variance $\times 10^{-4}$ for values of α'						
Method	*Symbol*	0.93	0.94	0.95	0.96	0.97	*0.98*	0.99
Baseline	$\hat{\alpha}_a$	0.196	0.196	0.196	0.196	0.196	*0.196*	0.196
Classify-and-count	$\hat{\alpha}^*$	0.000	0.000	0.000	0.000	0.000	*0.000*	0.000
Subtracted-bias	$\hat{\alpha}_b$	0.307	0.307	0.307	0.307	0.307	*0.307*	0.307
Misclassification	$\hat{\alpha}_p$	0.485	0.445	0.413	0.388	0.370	*0.359*	0.355
Calibration	$\hat{\alpha}_c$	0.444	0.364	0.293	0.229	0.174	*0.127*	0.088

		RMSE $\times 10^{-2}$ for values of α'						
Method	*Symbol*	0.93	0.94	0.95	0.96	0.97	*0.98*	0.99
Baseline	$\hat{\alpha}_a$	5.020	4.024	3.032	2.048	1.094	*0.443*	1.094
Classify-and-count	$\hat{\alpha}^*$	2.370	2.460	2.550	2.640	2.730	*2.820*	2.910
Subtracted-bias	$\hat{\alpha}_b$	0.594	0.596	0.600	0.603	0.606	*0.609*	0.613
Misclassification	$\hat{\alpha}_p$	0.697	0.668	0.643	0.623	0.608	*0.599*	0.596
Calibration	$\hat{\alpha}_c$	3.299	2.654	2.013	1.378	0.769	*0.356*	0.711

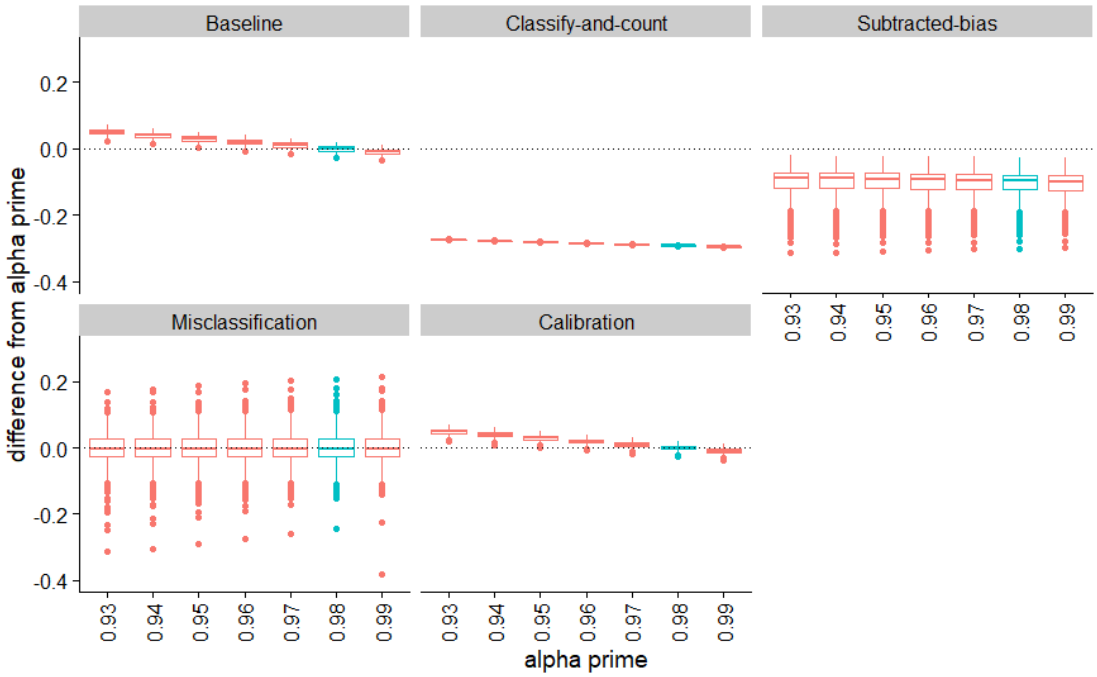


Figure 4.4: Simulation study to observe the change in prediction error under concept drift using boxplots. Five estimators are compared given a initial base rate $\alpha = 0.98$ (blue) and different values of α' (red). The x-axis shows the different base rates and the y-axis shows the distribution of the difference from α'. All the parameters: $p_{00} = 0.94$, $p_{11} = 0.97$, $n = 1000$ and $N = 3 \times 10^5$.

Chapter 5

Mixed estimator

In Chapter 4, we have discussed the effects of concept drift on our estimators. We concluded that none of the five estimators has a consistently low RMSE under concept drift. In this chapter, we propose a new mixed estimator. This mixed estimator combines properties of the misclassification estimator and the calibration estimator, such that we can obtain an estimator with a low bias and a low variance. We will write down expressions for the estimator and its bias, variance and RMSE. Furthermore, we will compare the mixed estimator with the misclassification estimator and the calibration estimator with a simulation study. Finally, we will draw conclusions about the properties of the mixed estimator.

5.1 Theory

The mixed estimator is a combination of the misclassification estimator and the calibration estimator. In Figure 4.3 and 4.4, we could see that the calibration estimator has a stable RMSE for $\alpha' = \alpha$. In other words, the calibration has a stable RMSE when the test set is drawn from the same distribution as the target population. Furthermore, the RMSE of the misclassification estimator does not change much over different values of α'. Even though that the overall RMSE can be high, the difference of the RMSE between α and α' is small. The mixed estimator starts with the values of the calibration estimator, but adds the change between the misclassification estimators of α and α', i.e. $\hat{\alpha}'_p - \hat{\alpha}_p$.

$$\begin{aligned}\hat{\alpha}'_m &= \hat{\alpha}_c + \left[\hat{\alpha}'_p - \hat{\alpha}_p\right] \\ &= \frac{n_{10}}{n_{+0}}(1-\hat{\alpha}^*) + \frac{n_{11}}{n_{+1}}\hat{\alpha}^* + \frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00} + \hat{p}_{11} - 1}\end{aligned} \tag{5.1}$$

As far as we know, the mixed estimator does not exist in scientific literature. Therefore, closed expression for the bias and variance of this estimator do not exist. Although the computations were heavy, we were able to calculate the bias and the variance of this mixed estimator. The mixed estimator is biased when $\alpha \neq \alpha'$.

Theorem 5. *The mixed estimator $\hat{\alpha}'_m$ is an biased estimator for $\alpha \neq \alpha'$:*

$$B\left[\hat{\alpha}'_m\right] = \frac{(\alpha' - \alpha)(V(\hat{p}_{00}) + V(\hat{p}_{11}))}{(p_{00} + p_{11} - 1)^2} + O(n^{-2}). \tag{5.2}$$

The variance of $\hat{\alpha}'_m$ is equal to the following expression:

$$\begin{aligned}
V(\hat{\alpha}'_m) = &\left[\frac{(1-\alpha)(1-p_{00})+\alpha p_{11}}{n} + \frac{(1-\alpha)p_{00}+\alpha(1-p_{11})}{n^2}\right] \\
&\times \left[\frac{\alpha p_{11}}{(1-\alpha)(1-p_{00})+\alpha p_{11}}\left(1 - \frac{\alpha p_{11}}{(1-\alpha)(1-p_{00})+\alpha p_{11}}\right)\right] \\
&+ \left[\frac{(1-\alpha)p_{00}+\alpha(1-p_{11})}{n} + \frac{(1-\alpha)(1-p_{00})+\alpha p_{11}}{n^2}\right] \\
&\times \left[\frac{(1-\alpha)p_{00}}{(1-\alpha)p_{00}+\alpha(1-p_{11})}\left(1 - \frac{(1-\alpha)p_{00}}{(1-\alpha)p_{00}+\alpha(1-p_{11})}\right)\right] \\
&+ (\alpha'-\alpha)^2 \times \frac{V(\hat{p}_{00})+V(\hat{p}_{11})}{(p_{00}+p_{11}-1)^2} \\
&+ (\alpha'-\alpha) \times \left[\frac{\alpha p_{00}(1-p_{00})(1-p_{11}) + p_{00}p_{11}(1-p_{11})}{n(p_{00}-\alpha(p_{00}+p_{11}-1))(p_{00}+p_{11}-1)}\right. \\
&\left. - \frac{\alpha p_{00}(1-p_{00})p_{11} + (1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00})+\alpha p_{11})(p_{00}+p_{11}-1)}\right] + O(n^{-2}). \qquad (5.3)
\end{aligned}$$

The bias will be large when the variance of the estimated classification probabilities are high and/or the classification probabilities are low. Low classification probabilities lead in fact to more variance in the estimated classification probabilities, but this effect will be even stronger when the size of the test set n is low. The variance of the mixed estimator is low when the difference between $\alpha \neq \alpha'$ is small, when the size of the test set n is large and the classification probabilities are high.

The variance is dependent on many parameters, so it is complex to analyse the behaviour on the formula only. In the next section, we will perform the same simulation studies as before, but with the addition of the new mixed estimator.

5.2 Properties of the Mixed Estimator

In this section, we compare the distribution of the mixed estimator with the misclassification estimator and the calibration estimator. We will use the same two simulation studies as in the previous chapter, but we add the mixed estimator to the simulation study. Combining the insights from the simulation study with the mathematical expressions, we are able to explain when the mixed estimator works.

In the first simulation study, we consider a class-balanced dataset ($\alpha = 0.5$), with a small test set of size $n = 1000$, a large population dataset of $N = 3 \times 10^5$ and a rather poor classifier having classification probabilities $p_{00} = 0.6$ and $p_{11} = 0.7$. From Figure 5.1, we can see that the mixed estimator is in general a stable estimator with a low amount of bias and way less variance than the misclassification estimator. However, the variance of the mixed estimator tends to increase when the difference between α' and α gets larger, which is in line with the observations in the previous section.

A situation where the mixed estimator does not work as well as expected, can be found in Figure 5.2. We again have the same parameters as the second simulation study in the previous chapter: $p_{00} = 0.94$, $p_{11} = 0.97$, $\alpha = 0.98$, $n = 1000$ and $N = 3 \times 10^5$. The misclassification estimator tend to have more extreme outliers when the difference between α' and α increases. This affects the mixed estimator in terms of variance. Furthermore, the mixed estimator can predict values outside the $[0, 1]$-interval. Obviously, we cannot encounter these values in practice and it is therefore a problem that we obtain these estimates. Finally, we can observe that

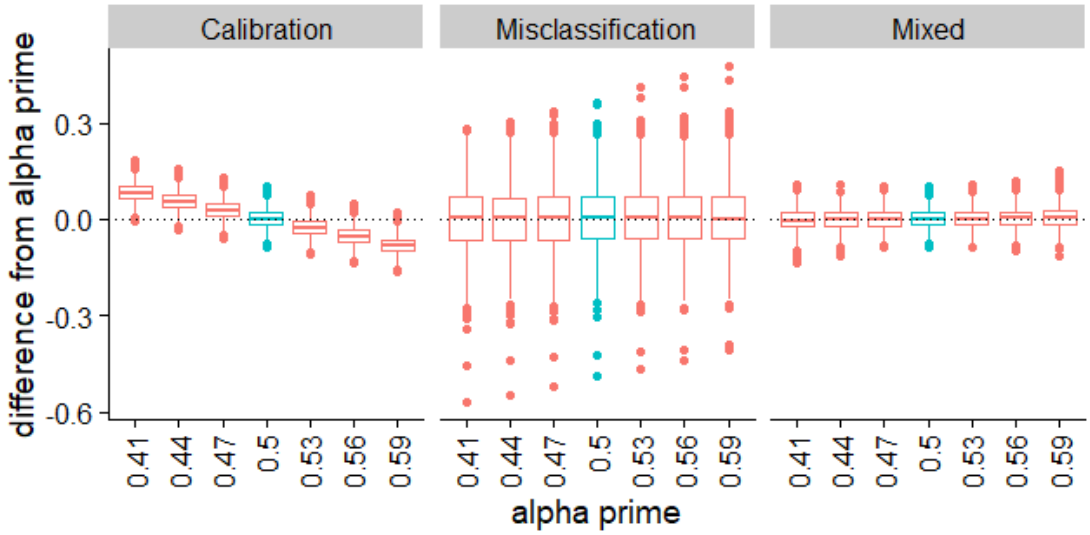


Figure 5.1: Simulation study to observe the change in prediction error under concept drift using boxplots. The calibration, misclassification and mixed estimator are compared given a initial base rate $\alpha = 0.5$ (blue) and different values of α' (red). The x-axis shows the different base rates and the y-axis shows the distribution of the difference from α'. All the parameters: $p_{00} = 0.6$, $p_{11} = 0.7$, $n = 1000$ and $N = 3 \times 10^5$.

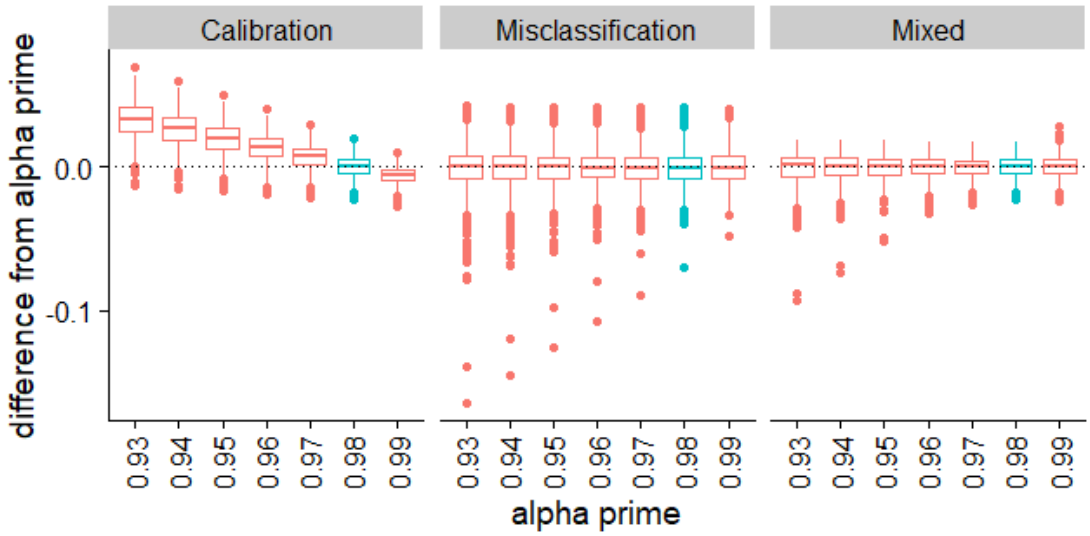


Figure 5.2: Simulation study to observe the change in prediction error under concept drift using boxplots. The calibration, misclassification and mixed estimator are compared given a initial base rate $\alpha = 0.98$ (blue) and different values of α' (red). The x-axis shows the different base rates and the y-axis shows the distribution of the difference from α'. All the parameters: $p_{00} = 0.94$, $p_{11} = 0.97$, $n = 1000$ and $N = 3 \times 10^5$.

the variance of the mixed estimator is always lower than the variance of the misclassification estimator.

In both the two simulation studies, the misclassification estimator did not work properly, and we showed values of α' that are close to α. It is also interesting to see what happens when the misclassification estimator has a low RMSE for α and what happens when α' differs substantially from α. We perform a simulation study with $\alpha = 0.75$, $p_{00} = 0.85$, $p_{11} = 0.9$, $n = 1000$ and $N = 3 \times 10^5$, shown in Figure 5.3. We observe that the distribution of the mixed estimator is similar to the distribution of the misclassification estimator. However, it seems that the mixed estimator still performs consistently better than the misclassification estimator.

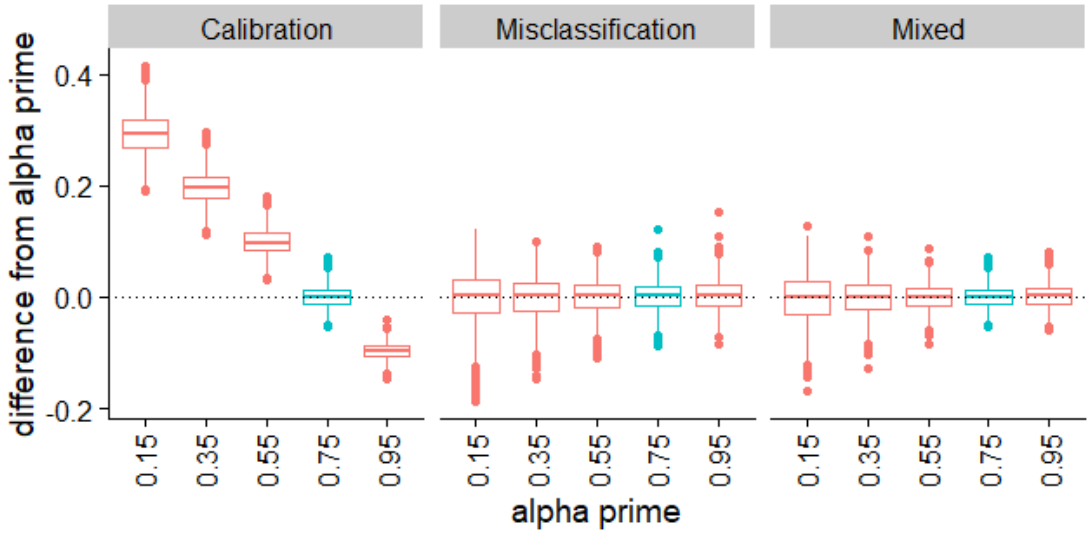


Figure 5.3: Simulation study to observe the change in prediction error under concept drift using boxplots. The calibration, misclassification and mixed estimator are compared given a initial base rate $\alpha = 0.75$ (blue) and different values of α' (red). The x-axis shows the different base rates and the y-axis shows the distribution of the difference from α'. All the parameters: $p_{00} = 0.85$, $p_{11} = 0.90$, $n = 1000$ and $N = 3 \times 10^5$.

All with all, we observe that the mixed estimator performs in general better than the calibration estimator and the mixed estimator. However, the RMSE of the mixed estimator can be large when there is a big difference between α' and α and when there are small classification probabilities p_{00} and p_{11}. Possible solutions will be discussed in the next, and final, chapter of this thesis.

Chapter 6

Discussion and Conclusion

In this thesis, we have compared five correction methods to reduce misclassification bias. We obtained our results by computing mathematical expressions for the bias, variance and mean square error and by performing simulation studies. These results make it possible to answer our research question that is proposed in the introduction: *"How can we reduce the mean squared error of the base rate for inaccurate data?"* This research question is supported by two subquestions.

The first subquestion, *"How can we reduce the mean squared error of the base rate for inaccurate data in case of a fixed base rate?"*, can be answered with our findings in chapters 2 and 3. First, we have seen that the baseline estimator $\hat{\alpha}_a$ has the lowest RMSE for small test sets and bad classifiers, i.e., n is small and (p_{00}, p_{11}) are close to 0.5. Second, we have seen that the classify-and-count estimator $\hat{\alpha}^*$ and the subtracted-bias estimator $\hat{\alpha}_b$ have the lowest RMSE when the classification probabilities p_{00} and p_{11} are close to the line closed by $(1-\alpha, \alpha)$ and $(1, 1)$. The area around that line gets smaller when the size of the test set n increases. Last, the calibration estimator $\hat{\alpha}_c$ has the smallest RMSE in all the other cases and is in general an estimator with a low RMSE. The calibration estimator $\hat{\alpha}_c$ has always a lower or equal RMSE than the misclassification estimator $\hat{\alpha}_p$.

The second subquestion, *"How can we reduce the mean squared error of the base rate for inaccurate data in case of a changing base rate?"*, can be answered with our findings in chapters 4 and 5. First, we have seen that the baseline estimator $\hat{\alpha}_a$ and the calibration estimator $\hat{\alpha}_a$ can be highly biased under concept drift. A large shift in the base rate parameter α' leads to a large bias of these estimators. These estimators are only useful in a small area around α. Second, the classify-and-count estimator and the subtracted-bias estimator are hardly affected by the concept drift. The change in variance under concept drift of the classify-and-count estimator is only visible in higher order terms, while the change in variance under concept drift of the subtracted-bias estimator is dependent on the difference between α and α'. Third, the misclassification estimator remains as the only estimator that is asymptotically unbiased under concept drift. The disadvantage of the misclassification estimator is the high variance under low classification probabilities or an extreme base rate parameter. We proposed a solution in the mixed estimator, which performs better than the misclassification estimator under almost all circumstances. However, the mixed estimator has still the same awkward properties as the misclassification estimator like high variance when there is a large difference between α and α' and singularity when $p_{00} + p_{11} \to 1$.

We can apply this knowledge to improve the quality of the official statistics. However, this thesis is not complete and we would like to point out some directions for future research. First, we can generalize this problem to multi-class problems. A binary classification algorithm is a good starting point, but many classifiers in official statistics have to do with three or more

classes. The expressions would become more complex, but can be partially solved with matrix notation. Second, we could relax the assumptions even more. We can question what will happen for small test sets and what will happen if the classification probabilities are estimated in other ways. Third and last, we could combine the estimators in a new estimator. A first step is made in introducing the mixed estimator, but this estimator has its disadvantages. Including biased estimators could potentially lead to more stable results. All with all, we made awareness for the fact that misclassification bias can lead to inaccurate statistics, even for good classifiers and we introduced new estimators that can reduce this misclassification bias without adding much variance.

Bibliography

[1] John P Buonaccorsi. *Measurement Error: Models, Methods, and Applications*. en. Boca Raton, FL: Chapman & Hall/CRC, 2010.

[2] R.L. Curier et al. "Monitoring spatial sustainable development: Semi-automated analysis of satellite and aerial images for energy transition and sustainability indicators". In: *arXiv preprint arXiv:1810.04881* (2018).

[3] Pablo González et al. "A Review on Quantification Learning". In: *ACM Computing Surveys* 50.5 (2017), 74:1–74:40. DOI: 10.1145/3117807.

[4] A. Grassia and R. Sundberg. "Statistical Precision in the Calibration and Use of Sorting Machines and Other Classifiers". en. In: *Technometrics* 24.2 (1982), pp. 117–121.

[5] Sander Greenland. "Sensitivity Analysis and Bias Analysis". en. In: *Handbook of Epidemiology*. Ed. by Wolfgang Ahrens and Iris Pigeot. New York, NY: Springer, 2014.

[6] Daniel J. Hopkins and Gary King. "A Method of Automated Nonparametric Content Analysis for Social Science". en. In: *American Journal of Political Science* 54.1 (2010), pp. 229–247. ISSN: 1540-5907.

[7] Paul Knottnerus. *Sample survey theory: some Pythagorean perspectives*. Springer Science & Business Media, 2003.

[8] J. Kuha and C. J. Skinner. "Categorical data analysis and misclassification". In: *Survey Measurement and Process Quality*. Ed. by L.E. Lyberg et al. Wiley, Mar. 1997, pp. 633–670.

[9] Fabian Löw, Patrick Knöfel, and Christopher Conrad. "Analysis of uncertainty in multi-temporal object-based classification". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 105 (2015), pp. 91–106. ISSN: 0924-2716.

[10] Q. A. Meertens et al. "A data-driven supply-side approach for estimating cross-border Internet purchases within the European Union". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 183.1 (2020), pp. 61–90. DOI: 10.1111/rssa.12487.

[11] Q.A. Meertens. "Understanding the Output Quality of Official Statistics that are Based on Machine Learning Algorithms". In: ().

[12] Q.A. Meertens et al. *A Bayesian Approach for Accurate Classification-Based Aggregates*. en. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2019.

[13] Sander Scholtus and Arnout van Delden. "On the Accuracy Of estimators based on a binary classifier". en. In: (Feb. 2020). Discussion Paper, Statistics Netherlands, The Hague.

[14] Joseph E. Schwartz. "The Neglected Problem of Measurement Error in Categorical Data". In: *Sociological Methods & Research* 13.4 (1985), pp. 435–466. DOI: 10.1177/0049124185013004001.

[15] A. Van Delden, S. Scholtus, and J. Burger. "Accuracy of Mixed-Source Statistics as Affected by Classification Errors". In: *Journal of Official Statistics* 32.3 (2016), pp. 619–642.

[16] Geoffrey I Webb et al. "Characterizing concept drift". In: *Data Mining and Knowledge Discovery* 30.4 (2016), pp. 964–994.

[17] Gregor Wiedemann. "Proportional Classification Revisited: Automatic Content Analysis of Political Manifestos Using Active Learning". en. In: *Social Science Computer Review* 37.2 (2019), pp. 135–159. ISSN: 0894-4393.

Chapter 7

Appendix

This appendix contains the proofs of the theorems presented in the thesis entitled "Comparing Correction Methods to Reduce Misclassification Bias". Recall that we have assumed a population of size N in which a fraction $\alpha := N_{1+}/N$ belongs to the class of interest, referred to as the class labelled as 1. We assume that a binary classification algorithm has been trained that correctly classifies a data point that belongs to class $i \in \{0, 1\}$ with probability $p_{ii} > 0.5$, independently across all data points. In addition, we assume that a test set of size $n \ll N$ is available and that it can be considered a simple random sample from the population. The classification probabilities p_{00} and p_{11} are estimated on that test set as described in Section 2.2. Finally, we assume that the classify-and-count estimator $\hat{\alpha}^*$ is distributed independently of $\hat{p}_{00}$ and $\hat{p}_{11}$, which is reasonable (at least as an approximation) when $n \ll N$.

It may be noted that the estimated probabilities $\hat{p}_{11}$ and $\hat{p}_{00}$ defined in Section 2.2 cannot be computed if $n_{1+} = 0$ or $n_{0+} = 0$. Similarly, the calibration probabilities c_{11} and c_{00} cannot be estimated if $n_{+1} = 0$ or $n_{+0} = 0$. We assume here that these events occur with negligible probability. This will be true when n is sufficiently large so that $n\alpha \gg 1$ and $n(1-\alpha) \gg 1$.

Preliminaries

Many of the proofs presented in this appendix rely on the following two mathematical results. First, we will use univariate and bivariate Taylor series to approximate the expectation of non-linear functions of random variables. That is, to estimate $E[f(X)]$ and $E[g(X, Y)]$ for sufficiently differentiable functions f and g, we will insert the Taylor series for f and g at $x_0 = E[X]$ and $y_0 = E[Y]$ up to terms of order 2 and utilize the linearity of the expectation. Second, we will use the following conditional variance decomposition for the variance of a random variable X:

$$V(X) = E[V(X \mid Y)] + V(E[X \mid Y]). \tag{7.1}$$

The conditional variance decomposition follows from the tower property of conditional expectations [7]. Before we prove the theorems presented in the paper, we begin by proving the following lemma.

Lemma 1. *The variance of the estimator $\hat{p}_{11}$ for p_{11} estimated on the test set is given by*

$$V(\hat{p}_{11}) = \frac{p_{11}(1-p_{11})}{n\alpha}\left[1 + \frac{1-\alpha}{n\alpha}\right] + O\left(\frac{1}{n^3}\right). \tag{7.2}$$

Similarly, the variance of $\hat{p}_{00}$ is given by

$$V(\hat{p}_{00}) = \frac{p_{00}(1-p_{00})}{n(1-\alpha)}\left[1 + \frac{\alpha}{n(1-\alpha)}\right] + O\left(\frac{1}{n^3}\right). \tag{7.3}$$

Moreover, $\hat{p}_{11}$ and $\hat{p}_{00}$ are uncorrelated: $C(\hat{p}_{11}, \hat{p}_{00}) = 0$.

Proof of Lemma 1. We approximate the variance of $\hat{p}_{00}$ using the conditional variance decomposition and a second-order Taylor series, as follows:

$$\begin{aligned}
V(\hat{p}_{00}) &= V\left(\frac{n_{00}}{n_{0+}}\right) \\
&= E_{n_{0+}}\left[V\left(\frac{n_{00}}{n_{0+}} \mid n_{0+}\right)\right] + V_{n_{0+}}\left[E\left(\frac{n_{00}}{n_{0+}} \mid n_{0+}\right)\right] \\
&= E_{n_{0+}}\left[\frac{1}{n_{0+}^2} V(n_{00} \mid n_{0+})\right] + V_{n_{0+}}\left[\frac{1}{n_{0+}} E(n_{00} \mid n_{0+})\right] \\
&= E_{n_{0+}}\left[\frac{n_{0+} p_{00}(1-p_{00})}{n_{0+}^2}\right] + V_{n_{0+}}\left[\frac{n_{0+} p_{00}}{n_{0+}}\right] \\
&= E_{n_{0+}}\left[\frac{1}{n_{0+}}\right] p_{00}(1-p_{00}) \\
&= \left[\frac{1}{E[n_{0+}]} + \frac{1}{2}\frac{2}{E[n_{0+}]^3} \times V[n_{0+}]\right] p_{00}(1-p_{00}) + O\left(\frac{1}{n^3}\right) \\
&= \frac{p_{00}(1-p_{00})}{E[n_{0+}]}\left[1 + \frac{V[n_{0+}]}{E[n_{0+}]^2}\right] + O\left(\frac{1}{n^3}\right) \\
&= \frac{p_{00}(1-p_{00})}{n(1-\alpha)}\left[1 + \frac{\alpha}{n(1-\alpha)}\right] + O\left(\frac{1}{n^3}\right).
\end{aligned}$$

The variance of $\hat{p}_{11}$ is approximated in the exact same way.

Finally, to evaluate $C(\hat{p}_{11}, \hat{p}_{00})$ we use the analogue of (7.1) for covariances:

$$\begin{aligned}
C(\hat{p}_{11}, \hat{p}_{00}) &= C\left(\frac{n_{11}}{n_{1+}}, \frac{n_{00}}{n_{0+}}\right) \\
&= E_{n_{1+},n_{0+}}\left[C\left(\frac{n_{11}}{n_{1+}}, \frac{n_{00}}{n_{0+}} \mid n_{1+}, n_{0+}\right)\right] \\
&\qquad + C_{n_{1+},n_{0+}}\left[E\left(\frac{n_{11}}{n_{1+}} \mid n_{1+}, n_{0+}\right), E\left(\frac{n_{00}}{n_{0+}} \mid n_{1+}, n_{0+}\right)\right] \\
&= E_{n_{1+},n_{0+}}\left[\frac{1}{n_{1+}n_{0+}} C(n_{11}, n_{00} \mid n_{1+}, n_{0+})\right] \\
&\qquad + C_{n_{1+},n_{0+}}\left[\frac{1}{n_{1+}} E(n_{11} \mid n_{1+}), \frac{1}{n_{0+}} E(n_{00} \mid n_{0+})\right].
\end{aligned}$$

The second term is zero as before. The first term also vanishes because, conditional on the row totals n_{1+} and n_{0+}, the counts n_{11} and n_{00} follow independent binomial distributions, so $C(n_{11}, n_{00} \mid n_{1+}, n_{0+}) = 0$. □

Note: in the remainder of this appendix, we will not add explicit subscripts to expectations and variances when their meaning is unambiguous.

Subtracted-bias estimator

We will now prove the bias and variance approximations for the subtracted-bias estimator $\hat{\alpha}_b$ that was defined in Equation 2.12.

Proof of Theorem 1. The bias of $\hat{\alpha}_b$ is given by

$$\begin{aligned} B(\hat{\alpha}_b) &= E\left[\hat{\alpha}^\star - \hat{B}[\hat{\alpha}^\star]\right] - \alpha \\ &= E[\hat{\alpha}^\star - \alpha] - E\left[\hat{B}[\hat{\alpha}^\star]\right] \\ &= B[\hat{\alpha}^\star] - E\left[\hat{B}[\hat{\alpha}^\star]\right] \\ &= [\alpha(p_{00}+p_{11}-2)+(1-p_{00})] - E\left[\hat{\alpha}^\star(\hat{p}_{00}+\hat{p}_{11}-2)+(1-\hat{p}_{00})\right]. \end{aligned}$$

Because $\hat{\alpha}^*$ and $(\hat{p}_{00}+\hat{p}_{11}-2)$ are assumed to be independent, the expectation of their product equals the product of their expectations:

$$\begin{aligned} B(\hat{\alpha}_b) &= \alpha(p_{00}+p_{11}-2)+(1-p_{00}) - E[\hat{\alpha}^\star](p_{00}+p_{11}-2) - (1-p_{00}) \\ &= (\alpha - E[\hat{\alpha}^\star])(p_{00}+p_{11}-2) \\ &= B[\hat{\alpha}^\star](2-p_{00}-p_{11}) \\ &= (1-p_{00})(2-p_{00}-p_{11}) - \alpha(p_{00}+p_{11}-2)^2. \end{aligned}$$

This proves the formula for the bias of $\hat{\alpha}_b$ as estimator for α. To approximate the variance of $\hat{\alpha}_b$, we apply the conditional variance decomposition (7.1) conditional on $\hat{\alpha}^*$ and look at the two resulting terms separately. First, consider the expectation of the conditional variance:

$$\begin{aligned} E\left[V(\hat{\alpha}_b \mid \hat{\alpha}^*)\right] &= E\left[V(\hat{\alpha}^*(3-\hat{p}_{00}-\hat{p}_{11})-(1-\hat{p}_{00}) \mid \hat{\alpha}^*)\right] \\ &= E\big[V(\hat{\alpha}^*(3-\hat{p}_{00}-\hat{p}_{11}) \mid \hat{\alpha}^*) + V(1-\hat{p}_{00} \mid \hat{\alpha}^*) \\ &\qquad - 2C(\hat{\alpha}^*(3-\hat{p}_{00}-\hat{p}_{11}), 1-\hat{p}_{00} \mid \hat{\alpha}^*)\big] \\ &= E\big[(\hat{\alpha}^*)^2 V(3-\hat{p}_{00}-\hat{p}_{11} \mid \hat{\alpha}^*) + V(1-\hat{p}_{00} \mid \hat{\alpha}^*) \\ &\qquad - 2\hat{\alpha}^* C(3-\hat{p}_{00}-\hat{p}_{11}, 1-\hat{p}_{00} \mid \hat{\alpha}^*)\big] \\ &= E\big[(\hat{\alpha}^*)^2\left[V(\hat{p}_{00})+V(\hat{p}_{11})\right] + V(\hat{p}_{00}) - 2\hat{\alpha}^* V(\hat{p}_{00})\big] \\ &= E\left[(\hat{\alpha}^*)^2\right]\left[V(\hat{p}_{00})+V(\hat{p}_{11})\right] + V(\hat{p}_{00}) - 2E\left[\hat{\alpha}^*\right]V(\hat{p}_{00}). \end{aligned}$$

In the penultimate line, we used that $C(\hat{p}_{11}, \hat{p}_{00}) = 0$. The second moment $E\left[(\hat{\alpha}^*)^2\right]$ can be written as $E\left[\hat{\alpha}^*\right]^2 + V(\hat{\alpha}^*)$. Because $V(\hat{\alpha}^*)$ is of order $1/N$, it can be neglected compared to $E\left[\hat{\alpha}^*\right]^2$, which is of order 1. In particular, we find that the expectation of the conditional variance equals:

$$\begin{aligned} E\left[V(\hat{\alpha}_b \mid \hat{\alpha}^*)\right] &= E\left[(\hat{\alpha}^*)\right]^2\left[V(\hat{p}_{00})+V(\hat{p}_{11})\right] + V(\hat{p}_{00}) - 2E\left[\hat{\alpha}^*\right]V(\hat{p}_{00}) + O\left(\frac{1}{N}\right) \\ &= V(\hat{p}_{00})\left[E\left[\hat{\alpha}^*\right]-1\right]^2 + V(\hat{p}_{11})E\left[\hat{\alpha}^*\right]^2 + O\left(\frac{1}{N}\right). \end{aligned}$$

Next, the variance of the conditional expectation can be seen to be equal the following:

$$\begin{aligned} V\left[E(\hat{\alpha}_b \mid \hat{\alpha}^*)\right] &= V\left[E(\hat{\alpha}^*(3-\hat{p}_{00}-\hat{p}_{11})-(1-\hat{p}_{00}) \mid \hat{\alpha}^*)\right] \\ &= V\left[\hat{\alpha}^* E(3-\hat{p}_{00}-\hat{p}_{11} \mid \hat{\alpha}^*) - E(1-\hat{p}_{00} \mid \hat{\alpha}^*)\right] \\ &= V(\hat{\alpha}^*)(3-p_{00}-p_{11})^2. \end{aligned}$$

Because $V(\hat{\alpha}^*)$ is of order $1/N$, it can be neglected in the final formula. Furthermore, the variances of $\hat{p}_{00}$ and $\hat{p}_{11}$ can be written out using the result from Lemma 1:

$$\begin{aligned} V(\hat{\alpha}_b) = &\frac{\left[\alpha(p_{00}+p_{11}-1)-p_{00}\right]^2 p_{00}(1-p_{00})}{n(1-\alpha)}\left[1+\frac{\alpha}{n(1-\alpha)}\right] \\ &+\frac{\left[\alpha(p_{00}+p_{11}-1)+(1-p_{00})\right]^2 p_{11}(1-p_{11})}{n\alpha}\left[1+\frac{1-\alpha}{n\alpha}\right] \\ &+O\left(\max\left[\frac{1}{n^3},\frac{1}{N}\right]\right). \end{aligned}$$

This concludes the proof of Theorem 1. □

Misclassification estimator

We will now prove the bias and variance approximations for the misclassification estimator $\hat{\alpha}_p$ as defined in Equation (2.15).

Proof of Theorem 2. Under the assumption that $\hat{\alpha}^*$ is distributed independently of $(\hat{p}_{00}, \hat{p}_{11})$, it holds that

$$\begin{aligned} E(\hat{\alpha}_p) &= E\left(\frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1}\right) + E\left[E\left(\frac{\hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\;\middle|\;\hat{\alpha}^*\right)\right] \\ &= E\left(\frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1}\right) + E(\hat{\alpha}^*)E\left(\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right). \end{aligned} \tag{7.4}$$

$E(\hat{\alpha}^*)$ is known from (2.7). To evaluate the other two expectations, we use a second-order Taylor series approximation. The first- and second-order partial derivatives of $f(x,y) = 1/(x+y-1)$ and $g(x,y) = (x-1)/(x+y-1) = 1-[y/(x+y-1)]$ are given by:

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} = \frac{-1}{(x+y-1)^2}, \tag{7.5}$$

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial^2 f}{\partial y^2} = \frac{2}{(x+y-1)^3},$$

$$\frac{\partial g}{\partial x} = \frac{y}{(x+y-1)^2}, \tag{7.6}$$

$$\frac{\partial g}{\partial y} = \frac{-(x-1)}{(x+y-1)^2}, \tag{7.7}$$

$$\frac{\partial^2 g}{\partial x^2} = \frac{-2y}{(x+y-1)^3},$$

$$\frac{\partial^2 g}{\partial y^2} = \frac{2(x-1)}{(x+y-1)^3}.$$

Now also using that $C(\hat{p}_{11}, \hat{p}_{00}) = 0$, we obtain for the first expectation:

$$\begin{aligned} E\left(\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right) &= \frac{1}{p_{00}+p_{11}-1} + \frac{V(\hat{p}_{00})+V(\hat{p}_{11})}{(p_{00}+p_{11}-1)^3} + O(n^{-2}) \\ &= \frac{1}{p_{00}+p_{11}-1}\left[1+\frac{\frac{p_{00}(1-p_{00})}{n(1-\alpha)}+\frac{p_{11}(1-p_{11})}{n\alpha}}{(p_{00}+p_{11}-1)^2}\right] + O(n^{-2}). \end{aligned} \tag{7.8}$$

Here, we have included only the first term of the approximations to $V(\hat{p}_{00})$ and $V(\hat{p}_{11})$ from Lemma 1, since this suffices to approximate the bias up to terms of order $O(1/n)$. Similarly, for the second expectation we obtain:

$$\begin{aligned} E\left(\frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1}\right) &= \frac{p_{00}-1}{p_{00}+p_{11}-1} + \frac{(p_{00}-1)V(\hat{p}_{11}) - p_{11}V(\hat{p}_{00})}{(p_{00}+p_{11}-1)^3} + O(n^{-2}) \\ &= \frac{p_{00}-1}{p_{00}+p_{11}-1}\left[1 + p_{11}\frac{\frac{1-p_{11}}{n\alpha} + \frac{p_{00}}{n(1-\alpha)}}{(p_{00}+p_{11}-1)^2}\right] + O(n^{-2}). \end{aligned} \tag{7.9}$$

Using (7.4), (2.7), (7.8), and (7.9), we conclude that:

$$\begin{aligned} E(\hat{\alpha}_p) = {} & \frac{\alpha(p_{00}+p_{11}-1)-(p_{00}-1)}{p_{00}+p_{11}-1}\left[1 + \frac{\frac{p_{00}(1-p_{00})}{n(1-\alpha)} + \frac{p_{11}(1-p_{11})}{n\alpha}}{(p_{00}+p_{11}-1)^2}\right] \\ & + \frac{p_{00}-1}{p_{00}+p_{11}-1}\left[1 + p_{11}\frac{\frac{1-p_{11}}{n\alpha} + \frac{p_{00}}{n(1-\alpha)}}{(p_{00}+p_{11}-1)^2}\right] + O\left(\frac{1}{n^2}\right). \end{aligned}$$

From this, it follows that an approximation to the bias of $\hat{\alpha}_p$ that is correct up to terms of order $O(1/n)$ is given by:

$$\begin{aligned} B(\hat{\alpha}_p) = {} & \frac{\alpha(p_{00}+p_{11}-1)-(p_{00}-1)}{n(p_{00}+p_{11}-1)^3}\left[\frac{p_{00}(1-p_{00})}{1-\alpha} + \frac{p_{11}(1-p_{11})}{\alpha}\right] \\ & + \frac{(p_{00}-1)p_{11}}{n(p_{00}+p_{11}-1)^3}\left[\frac{1-p_{11}}{\alpha} + \frac{p_{00}}{1-\alpha}\right] + O\left(\frac{1}{n^2}\right). \end{aligned}$$

By expanding the products in this expression and combining similar terms, the expression can be simplified to:

$$B(\hat{\alpha}_p) = \frac{p_{11}(1-p_{11}) - p_{00}(1-p_{00})}{n(p_{00}+p_{11}-1)^2} + O\left(\frac{1}{n^2}\right).$$

Finally, using the identity $p_{11}(1-p_{11}) - p_{00}(1-p_{00}) = (p_{00}+p_{11}-1)(p_{00}-p_{11})$, we obtain the required result for $B(\hat{\alpha}_p)$.

To approximate the variance of $\hat{\alpha}_p$, we apply the conditional variance decomposition conditional on $\hat{\alpha}^*$ and look at the two resulting terms separately. First, consider the variance of the conditional expectation:

$$\begin{aligned} V\left[E(\hat{\alpha}_p \mid \hat{\alpha}^*)\right] &= V\left[E\left(\hat{\alpha}^*\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1} + \frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1} \mid \hat{\alpha}^*\right)\right] \\ &= V\left[\hat{\alpha}^*\frac{1}{p_{00}+p_{11}-1}\right] \\ &= \frac{1}{(p_{00}+p_{11}-1)^2}V\left[\hat{\alpha}^*\right] = O\left(\frac{1}{N}\right), \end{aligned} \tag{7.10}$$

where in the last line we used (2.9). Note: the factor $1/(p_{00}+p_{11}-1)^2$ can become arbitrarily large in the limit $p_{00}+p_{11} \to 1$. It will be seen below that this same factor also occurs in the lower-order terms of $V(\hat{\alpha}_p)$; hence, the relative contribution of (7.10) remains negligible even in the limit $p_{00}+p_{11} \to 1$.

Next, we compute the expectation of the conditional variance.

$$\begin{aligned}
E\left[V(\hat{\alpha}_p \mid \hat{\alpha}^*)\right] &= E\left[V\left(\hat{\alpha}^* \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1} + \frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1} \mid \hat{\alpha}^\star\right)\right] \\
&= E\Bigg[V\left(\hat{\alpha}^* \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1} \mid \alpha^\star\right) + V\left(\frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1} \mid \hat{\alpha}^\star\right) \\
&\qquad + 2C\left(\hat{\alpha}^* \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}, \frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1} \mid \hat{\alpha}^\star\right)\Bigg] \\
&= E\left[(\hat{\alpha}^*)^2\right] V\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] + V\left[\frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] \\
&\qquad + 2E\left[\hat{\alpha}^\star\right] C\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}, \frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] \\
&= E\left[\hat{\alpha}^\star\right]^2 \left[1+O\left(\frac{1}{N}\right)\right] V\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] + V\left[\frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] \\
&\qquad + 2E\left[\hat{\alpha}^\star\right] C\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}, \frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1}\right]. \qquad (7.11)
\end{aligned}$$

To approximate the variance and covariance terms, we use a first-order Taylor series. Using the partial derivatives in (7.5), (7.6) and (7.7), we obtain:

$$\begin{aligned}
V\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] &= \frac{V(\hat{p}_{00})+V(\hat{p}_{11})}{(p_{00}+p_{11}-1)^4} + O(n^{-2}) \\
V\left[\frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] &= \frac{V(\hat{p}_{00})(p_{11})^2}{(p_{00}+p_{11}-1)^4} + \frac{V(\hat{p}_{11})(1-p_{00})^2}{(p_{00}+p_{11}-1)^4} + O(n^{-2}) \\
C\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}, \frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] &= \frac{V(\hat{p}_{00})(-p_{11})}{(p_{00}+p_{11}-1)^4} + \frac{V(\hat{p}_{11})(p_{00}-1)}{(p_{00}+p_{11}-1)^4} + O(n^{-2}).
\end{aligned}$$

Substituting these terms into Formula (7.11) and accounting for Formula (7.10) yields:

$$\begin{aligned}
V(\hat{\alpha}_p) &= \frac{V(\hat{p}_{00})\left[E\left[\hat{\alpha}^\star\right]^2 - 2p_{11}E\left[\hat{\alpha}^\star\right] + p_{11}^2\right]}{(p_{00}+p_{11}-1)^4} \\
&\qquad + \frac{V(\hat{p}_{11})\left[E\left[\hat{\alpha}^\star\right]^2 - 2(1-p_{00})E\left[\hat{\alpha}^\star\right] + (1-p_{00})^2\right]}{(p_{00}+p_{11}-1)^4} + O\left(\max\left[\frac{1}{n^2}, \frac{1}{N}\right]\right) \\
&= \frac{V(\hat{p}_{00})\left[E\left[\hat{\alpha}^\star\right] - p_{11}\right]^2}{(p_{00}+p_{11}-1)^4} + \frac{V(\hat{p}_{11})\left[E\left[\hat{\alpha}^\star\right] - (1-p_{00})\right]^2}{(p_{00}+p_{11}-1)^4} + O\left(\max\left[\frac{1}{n^2}, \frac{1}{N}\right]\right) \\
&= \frac{V(\hat{p}_{00})(1-\alpha)^2}{(p_{00}+p_{11}-1)^2} + \frac{V(\hat{p}_{11})\alpha^2}{(p_{00}+p_{11}-1)^2} + O\left(\max\left[\frac{1}{n^2}, \frac{1}{N}\right]\right).
\end{aligned}$$

Finally, inserting the expressions for $V(\hat{p}_{00})$ and $V(\hat{p}_{11})$ from Lemma 1 yields:

$$\begin{aligned}
V(\hat{\alpha}_p) = &\frac{\frac{p_{00}(1-p_{00})}{n(1-\alpha)}\left[1+\frac{\alpha}{n(1-\alpha)}\right](1-\alpha)^2}{(p_{00}+p_{11}-1)^2} + \frac{\frac{p_{11}(1-p_{11})}{n\alpha}\left[1+\frac{1-\alpha}{n\alpha}\right]\alpha^2}{(p_{00}+p_{11}-1)^2} \\
&+ O\left(\max\left[\frac{1}{n^2}, \frac{1}{N}\right]\right),
\end{aligned}$$

from which expression (4.12) follows. This concludes the proof of Theorem 2. □

Calibration estimator

We will now prove the bias and variance approximations for the calibration estimator $\hat{\alpha}_c$ that was defined in Equation (2.18).

Proof of Theorem 3. To compute the expected value of $\hat{\alpha}_c$, we first compute its expectation conditional on the 4-vector $\boldsymbol{N} = (N_{00}, N_{01}, N_{10}, N_{11})$:

$$\begin{aligned}
E(\hat{\alpha}_c \mid \boldsymbol{N}) &= E\left[\hat{\alpha}^* \frac{n_{11}}{n_{+1}} + (1-\hat{\alpha}^*)\frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}\right] \\
&= \hat{\alpha}^* E\left[\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}\right] + (1-\hat{\alpha}^*) E\left[\frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}\right] \\
&= \hat{\alpha}^* E\left[E\left(\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}, n_{+1}\right) \mid \boldsymbol{N}\right] \\
&\qquad + (1-\hat{\alpha}^*) E\left[E\left(\frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}, n_{+0}\right) \mid \boldsymbol{N}\right] \\
&= \frac{N_{+1}}{N} E\left[\frac{1}{n_{+1}} n_{+1} \frac{N_{11}}{N_{+1}} \mid \boldsymbol{N}\right] + \frac{N_{+0}}{N} E\left[\frac{1}{n_{+0}} n_{+0} \frac{N_{10}}{N_{+0}} \mid \boldsymbol{N}\right] \\
&= \frac{N_{11}}{N} + \frac{N_{10}}{N} \\
&= \frac{N_{1+}}{N} = \alpha.
\end{aligned} \tag{7.12}$$

By the tower property of conditional expectations, it follows that $E[\hat{\alpha}_c] = E\left[E(\hat{\alpha}_c \mid \boldsymbol{N})\right] = \alpha$. This proves that $\hat{\alpha}_c$ is an unbiased estimator for α.

To compute the variance of $\hat{\alpha}_c$, we use the conditional variance decomposition, again conditioning on the 4-vector $\boldsymbol{N}$. We remark that N_{0+} and N_{1+} are deterministic values, but that N_{+0} and N_{+1} are random variables. As shown above in Equation (7.12), the conditional expectation is deterministic, hence it has no variance: $V(E[\hat{\alpha}_c \mid \boldsymbol{N}]) = 0$. The conditional variance decomposition then simplifies to the following:

$$V(\hat{\alpha}_c) = E\left[V(\hat{\alpha}_c \mid \boldsymbol{N})\right]. \tag{7.13}$$

The conditional variance $V(\hat{\alpha}_c \mid \boldsymbol{N})$ can be written as follows:

$$\begin{aligned}
V[\hat{\alpha}_c \mid \boldsymbol{N}] &= V\left[\hat{\alpha}^* \frac{n_{11}}{n_{+1}} + (1-\hat{\alpha}^*)\frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}\right] \\
&= (\hat{\alpha}^*)^2 V\left[\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}\right] + (1-\hat{\alpha}^*)^2 V\left[\frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}\right] \\
&\qquad + 2\hat{\alpha}^*(1-\hat{\alpha}^*) C\left[\frac{n_{11}}{n_{+1}}, \frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}\right].
\end{aligned} \tag{7.14}$$

We will consider these terms separately. First, the variance of n_{11}/n_{+1} can be computed by applying an additional conditional variance decomposition:

$$V\left[\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}\right] = V\left[E\left(\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}, n_{+1}\right) \mid \boldsymbol{N}\right] + E\left[V\left(\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}, n_{+1}\right) \mid \boldsymbol{N}\right].$$

The first term is zero, which can be shown as follows:

$$\begin{aligned} V\left[E\left(\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}, n_{+1}\right)\right] &= V\left[\frac{1}{n_{+1}} E(n_{11} \mid \boldsymbol{N}, n_{+1}) \mid \boldsymbol{N}\right] \\ &= V\left[\frac{1}{n_{+1}} n_{+1} \frac{N_{11}}{N_{+1}} \mid \boldsymbol{N}\right] \\ &= V\left[\frac{N_{11}}{N_{+1}} \mid \boldsymbol{N}\right] = 0. \end{aligned}$$

For the second term, we find under the assumption that $n \ll N$:

$$\begin{aligned} E\left[V\left(\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}, n_{+1}\right) \mid \boldsymbol{N}\right] &= E\left[\frac{1}{n_{+1}^2} V(n_{11} \mid \boldsymbol{N}, n_{+1}) \mid \boldsymbol{N}\right] \\ &= E\left[\frac{1}{n_{+1}^2} n_{+1} \frac{N_{11}}{N_{+1}} (1 - \frac{N_{11}}{N_{+1}}) \mid \boldsymbol{N}\right] \\ &= E\left[\frac{1}{n_{+1}} \mid \boldsymbol{N}\right] \frac{N_{11} N_{01}}{N_{+1}^2}. \end{aligned}$$

The expectation of $\frac{1}{n_{+1}}$ can be approximated with a second-order Taylor series:

$$\begin{aligned} V\left[\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}\right] &= \left[\frac{1}{E[n_{+1} \mid \boldsymbol{N}]} + \frac{1}{2} \frac{2}{E[n_{+1} \mid \boldsymbol{N}]^3} V\left[n_{+1} \mid \boldsymbol{N}\right]\right] \frac{N_{11} N_{01}}{N_{+1}^2} + O(n^{-3}) \\ &= \frac{1}{E[n_{+1} \mid \boldsymbol{N}]} \left[1 + \frac{V\left[n_{+1} \mid \boldsymbol{N}\right]}{E[n_{+1} \mid \boldsymbol{N}]^2}\right] \frac{N_{11} N_{01}}{N_{+1}^2} + O(n^{-3}) \\ &= \frac{1}{n\hat{\alpha}^*} \left[1 + \frac{1 - \hat{\alpha}^*}{n\hat{\alpha}^*}\right] \frac{N_{11} N_{01}}{N_{+1}^2} + O(n^{-3}). \end{aligned} \tag{7.15}$$

The variance of n_{10}/n_{+0} can be approximated in the same way, which yields the following expression:

$$V\left[\frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}\right] = \frac{1}{n(1 - \hat{\alpha}^*)} \left[1 + \frac{\hat{\alpha}^*}{n(1 - \hat{\alpha}^*)}\right] \frac{N_{00} N_{10}}{N_{+0}^2} + O(n^{-3}). \tag{7.16}$$

Finally, it can be shown that the covariance in the final term is equal to zero:

$$\begin{aligned} C\left[\frac{n_{11}}{n_{+1}}, \frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}\right] &= E\left[C\left(\frac{n_{11}}{n_{+1}}, \frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}, n_{+0}, n_{+1}\right) \mid \boldsymbol{N}\right] \\ &\quad + C\left[E\left(\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}, n_{+0}, n_{+1}\right), E\left(\frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}, n_{+0}, n_{+1}\right) \mid \boldsymbol{N}\right] \\ &= E\left[\frac{1}{n_{+0} n_{+1}} C\left(n_{11}, n_{10} \mid \boldsymbol{N}, n_{+0}, n_{+1}\right) \mid \boldsymbol{N}\right] \\ &\quad + C\left[\frac{1}{n_{+1}} E\left(n_{11} \mid \boldsymbol{N}, n_{+0}, n_{+1}\right), \frac{1}{n_{+0}} E\left(n_{10} \mid \boldsymbol{N}, n_{+0}, n_{+1}\right) \mid \boldsymbol{N}\right] \\ &= 0 + C\left[\frac{1}{n_{+1}} n_{+1} \frac{N_{11}}{N_{+1}}, \frac{1}{n_{+0}} n_{+0} \frac{N_{10}}{N_{+0}} \mid \boldsymbol{N}\right] = 0. \end{aligned} \tag{7.17}$$

Combining Formulas (7.15), (7.16) and (7.17) with (7.14) gives:

$$\begin{aligned} V[\hat{\alpha}_c \mid \boldsymbol{N}] &= \frac{N_{+1}^2}{N^2}\frac{1}{n\hat{\alpha}^*}\left[1+\frac{1-\hat{\alpha}^*}{n\hat{\alpha}^*}\right]\frac{N_{11}N_{01}}{N_{+1}^2} \\ &\quad + \frac{N_{+0}^2}{N^2}\frac{1}{n(1-\hat{\alpha}^*)}\left[1+\frac{\hat{\alpha}^*}{n(1-\hat{\alpha}^*)}\right]\frac{N_{00}N_{10}}{N_{+0}^2} + O(n^{-3}) \\ &= \frac{1}{n\hat{\alpha}^*}\left[1+\frac{1-\hat{\alpha}^*}{n\hat{\alpha}^*}\right]\frac{N_{11}N_{01}}{N^2} \\ &\quad + \frac{1}{n(1-\hat{\alpha}^*)}\left[1+\frac{\hat{\alpha}^*}{n(1-\hat{\alpha}^*)}\right]\frac{N_{00}N_{10}}{N^2} + O(n^{-3}). \end{aligned}$$

Recall from Formula (7.13) that $\mathrm{V}\,[\hat{\alpha}_c] = \mathrm{E}\,[V[\hat{\alpha}_c \mid \boldsymbol{N}]] = \mathrm{E}\,[\mathrm{E}\,[V[\hat{\alpha}_c \mid \boldsymbol{N}] \mid N_{+1}]]$. Hence,

$$\begin{aligned} V[\hat{\alpha}_c] = E\left[\frac{1}{n\hat{\alpha}^*}\left(1+\frac{1-\hat{\alpha}^*}{n\hat{\alpha}^*}\right)E\left(\frac{N_{11}N_{01}}{N^2} \mid N_{+1}\right)\right. \\ \left. + \frac{1}{n(1-\hat{\alpha}^*)}\left(1+\frac{\hat{\alpha}^*}{n(1-\hat{\alpha}^*)}\right)E\left(\frac{N_{00}N_{10}}{N^2} \mid N_{+1}\right)\right] + O(n^{-3}). \end{aligned} \tag{7.18}$$

To evaluate the expectations in this expression, we observe that, conditional on the column total N_{+1}, N_{11} is distributed as $Bin(N_{+1}, c_{11})$, where c_{11} is a calibration probability as defined in Section 2.2.5. Hence,

$$\begin{aligned} E\,[N_{11} \mid N_{+1}] &= N_{+1}c_{11} = \frac{N_{+1}\alpha p_{11}}{(1-\alpha)(1-p_{00}) + \alpha p_{11}} \\ V\,[N_{11} \mid N_{+1}] &= N_{+1}c_{11}(1-c_{11}). \end{aligned} \tag{7.19}$$

Similarly, since $N = N_{+1} + N_{+0}$ is fixed,

$$\begin{aligned} E\,[N_{00} \mid N_{+1}] &= N_{+0}c_{00} = \frac{N_{+0}(1-\alpha)p_{00}}{(1-\alpha)p_{00} + \alpha(1-p_{11})} \\ V\,[N_{00} \mid N_{+1}] &= N_{+0}c_{00}(1-c_{00}). \end{aligned} \tag{7.20}$$

Using these results, we obtain:

$$\begin{aligned} E\left[\frac{N_{11}N_{01}}{N^2} \mid N_{+1}\right] &= \frac{1}{N^2}E\,[N_{11}N_{01} \mid N_{+1}] \\ &= \frac{1}{N^2}E\,[N_{11}(N_{+1}-N_{11}) \mid N_{+1}] \\ &= \frac{1}{N^2}\left[N_{+1}E\,[N_{11} \mid N_{+1}] - E\left[N_{11}^2 \mid N_{+1}\right]\right] \\ &= \frac{1}{N^2}\left[N_{+1}E\,[N_{11} \mid N_{+1}] - V\,[N_{11} \mid N_{+1}] - E\,[N_{11} \mid N_{+1}]^2\right] \\ &= \frac{1}{N^2}\left[N_{+1}^2c_{11} - N_{+1}c_{11}(1-c_{11}) - N_{+1}^2c_{11}^2\right] \\ &= \frac{N_{+1}^2}{N^2}c_{11}(1-c_{11}) + O\left(\frac{1}{N}\right), \end{aligned} \tag{7.21}$$

and similarly

$$E\left[\frac{N_{00}N_{10}}{N^2} \mid N_{+1}\right] = \frac{N_{+0}^2}{N^2}c_{00}(1-c_{00}) + O\left(\frac{1}{N}\right). \tag{7.22}$$

Substituting expressions (7.21) and (7.22) into (7.18) and noting that $N_{+1}^2/N^2 = (\hat{\alpha}^*)^2$ and $N_{+0}^2/N^2 = (1-\hat{\alpha}^*)^2$, we obtain:

$$\begin{aligned} V[\hat{\alpha}_c] &= E\left[\frac{\hat{\alpha}^*}{n}\left(1+\frac{1-\hat{\alpha}^*}{n\hat{\alpha}^*}\right)c_{11}(1-c_{11})\right. \\ &\qquad \left. +\frac{1-\hat{\alpha}^*}{n}\left(1+\frac{\hat{\alpha}^*}{n(1-\hat{\alpha}^*)}\right)c_{00}(1-c_{00})\right] + O\left(\max\left[\frac{1}{n^3},\frac{1}{Nn}\right]\right) \\ &= \left[\frac{E(\hat{\alpha}^*)}{n}+\frac{1-E(\hat{\alpha}^*)}{n^2}\right]c_{11}(1-c_{11}) \\ &\qquad +\left[\frac{1-E(\hat{\alpha}^*)}{n}+\frac{E(\hat{\alpha}^*)}{n^2}\right]c_{00}(1-c_{00}) + O\left(\max\left[\frac{1}{n^3},\frac{1}{Nn}\right]\right). \end{aligned}$$

Finally, substituting the expressions for $E(\hat{\alpha}^*)$ from (2.7) and the expressions for c_{11} and c_{00} from (7.19) and (7.20), the desired expression (2.20) is obtained. This concludes the proof of Theorem 3. □

Comparing mean squared errors

To conclude, we present the proof of Theorem 4, which essentially shows that the mean squared error (up to and including terms of order $1/n$) of the calibration estimator is lower than that of the misclassification estimator.

Proof of Theorem 4. Recall that the bias of $\hat{\alpha}_p$ as an estimator for α is given by

$$B\left[\hat{\alpha}_p\right] = \frac{p_{00}-p_{11}}{n(p_{00}+p_{11}-1)} + O\left(\frac{1}{n^2}\right).$$

Hence, $(B\left[\hat{\alpha}_p\right])^2 = O(1/n^2)$ is not relevant for $\widetilde{MSE}[\hat{\alpha}_p]$. It follows that $\widetilde{MSE}[\hat{\alpha}_p]$ is equal to the variance of $\hat{\alpha}_p$ up to order $1/n$. From (2.17) we obtain:

$$\widetilde{MSE}[\hat{\alpha}_p] = \frac{1}{n}\left[\frac{(1-\alpha)p_{00}(1-p_{00})+\alpha p_{11}(1-p_{11})}{(p_{00}+p_{11}-1)^2}\right]. \tag{7.23}$$

Recall that $\hat{\alpha}_c$ is an unbiased estimator for α, i.e., $B[\hat{\alpha}_c] = 0$. Also recall the notation $\beta = (1-\alpha)(1-p_{00})+\alpha p_{11}$. It follows from (2.20) that the variance, and hence the MSE, of $\hat{\alpha}_c$ up to terms of order $1/n$ can be written as:

$$\begin{aligned} \widetilde{MSE}[\hat{\alpha}_c] &= \frac{1}{n}\left[\beta\frac{\alpha p_{11}}{\beta}\left(1-\frac{\alpha p_{11}}{\beta}\right)+(1-\beta)\frac{(1-\alpha)p_{00}}{1-\beta}\left(1-\frac{(1-\alpha)p_{00}}{1-\beta}\right)\right] \\ &= \frac{\alpha(1-\alpha)}{n}\left[\frac{(1-p_{00})p_{11}}{\beta}+\frac{p_{00}(1-p_{11})}{1-\beta}\right]. \end{aligned} \tag{7.24}$$

To prove Expression (3.1), first note that

$$\frac{(1-p_{00})p_{11}}{\beta}+\frac{p_{00}(1-p_{11})}{1-\beta} = \frac{(1-p_{00})p_{11}+\beta(p_{00}-p_{11})}{\beta(1-\beta)}. \tag{7.25}$$

The numerator of this equation can be rewritten as follows:

$$\begin{aligned} &(1-p_{00})p_{11}+\beta(p_{00}-p_{11}) \\ &= (1-p_{00})p_{11}+(1-\alpha)p_{00}(1-p_{00})+\alpha p_{00}p_{11}-(1-\alpha)(1-p_{00})p_{11}-\alpha p_{11}^2 \\ &= (1-\alpha)p_{00}(1-p_{00})+\alpha p_{00}p_{11}+\alpha(1-p_{00})p_{11}-\alpha p_{11}^2 \\ &= (1-\alpha)p_{00}(1-p_{00})+\alpha p_{11}(1-p_{11}). \end{aligned}$$

Note that the obtained expression is equal to the numerator of Expression (7.23). Write $T = (1-\alpha)p_{00}(1-p_{00}) + \alpha p_{11}(1-p_{11})$ for that expression. It follows that

$$\begin{aligned}
&\widetilde{MSE}[\hat{\alpha}_p] - \widetilde{MSE}[\hat{\alpha}_c] \\
&= \frac{T}{n(p_{00}+p_{11}-1)^2} - \frac{T\alpha(1-\alpha)}{n\beta(1-\beta)} \\
&= \frac{T}{n(p_{00}+p_{11}-1)^2\beta(1-\beta)}\Big[\beta(1-\beta) - \alpha(1-\alpha)(p_{00}+p_{11}-1)^2\Big].
\end{aligned}$$

Writing out the second factor in the last expression gives the following:

$$\begin{aligned}
&\beta(1-\beta) - \alpha(1-\alpha)(p_{00}+p_{11}-1)^2 \\
&= (1-\alpha)^2 p_{00}(1-p_{00}) + \alpha(1-\alpha)\Big((1-p_{00})(1-p_{11}) + p_{00}p_{11}\Big) + \alpha^2 p_{11}(1-p_{11}) \\
&\qquad - \alpha(1-\alpha)(p_{00}+p_{11}-1)^2 \\
&= (1-\alpha)^2 p_{00}(1-p_{00}) + \alpha(1-\alpha)\Big(p_{00}(1-p_{00}) + p_{11}(1-p_{11})\Big) + \alpha^2 p_{11}(1-p_{11}) \\
&= (1-\alpha)p_{00}(1-p_{00}) + \alpha p_{11}(1-p_{11}) \\
&= T.
\end{aligned}$$

This concludes the proof of Theorem 4. □

Mixed estimator

In this section, we will prove the bias and the variance of the mixed estimator under concept drift. The mixed estimator is dependent on the calibration estimator at time 0, the misclassification estimator on time 0 and the misclassification estimator on time t.

Proof of Theorem 5. First, we will make a proof for the bias of the Mixed Estimator. The expression for the Mixed Estimator is:

$$\begin{aligned}
\hat{\alpha}'_m &= \hat{\alpha}_c + (\hat{\alpha}'_p - \hat{\alpha}_p) \\
&= \hat{\alpha}_c + [(\hat{\alpha}')^* - \hat{\alpha}^*] \times \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}.
\end{aligned} \tag{7.26}$$

The bias is defined as the difference between the expected value of the estimator minus the true value of the target variable:

$$B[\hat{\alpha}'_m] = E[\hat{\alpha}'_m] - \alpha' \tag{7.27}$$

Using Equation 7.26, we can write out the expected value of the Mixed estimator.

$$\begin{aligned}
E[\hat{\alpha}'_m] &= E\left[\hat{\alpha}_c + [(\hat{\alpha}')^* - \hat{\alpha}^*] \times \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] \\
&= E\left[\hat{\alpha}_c\right] + E\left[[(\hat{\alpha}')^* - \hat{\alpha}^*] \times \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right]
\end{aligned} \tag{7.28}$$

From Theorem 3, we already know that:

$$E\left[\hat{\alpha}_c\right] = E\left[E\left[\hat{\alpha}_c \mid \boldsymbol{N}\right]\right] = \alpha \tag{7.29}$$

$E\left[[(\hat{\alpha}')^* - \hat{\alpha}^*] \times \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right]$ can be computed by conditioning on the Classify-and-count estimators $(\hat{\alpha}')^*$ and $\hat{\alpha}^*$.

$$\begin{aligned} E\left[[(\hat{\alpha}')^* - \hat{\alpha}^*] \times \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] &= E\left[E\left[[(\hat{\alpha}')^* - \hat{\alpha}^*] \times \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1} \mid (\hat{\alpha}')^*, \hat{\alpha}^*\right]\right] \\ &= E\left[((\hat{\alpha}')^* - \hat{\alpha}^*) \times E\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1} \mid (\hat{\alpha}')^*, \hat{\alpha}^*\right]\right] \\ &= E\left[((\hat{\alpha}')^* - \hat{\alpha}^*) \times E\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right]\right] \end{aligned} \tag{7.30}$$

From Theorem 2, we used Taylor Series to approximate the expected value of $\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}$.

$$E\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] = \frac{1}{p_{00}+p_{11}-1} + \frac{V(\hat{p}_{00}) + V\hat{p}_{11})}{(p_{00}+p_{11}-1)^3} + O(n^{-2}) \tag{7.31}$$

Now it only remains to calculate the expected values of the Classify-and-count estimators.

$$E\left[(\hat{\alpha}')^* - \hat{\alpha}^*\right] = E\left[(\hat{\alpha}')^*\right] - E\left[\hat{\alpha}^*\right] \tag{7.32}$$

$$E\left[(\hat{\alpha}')^*\right] = \alpha' p_{11} + (1-\alpha')(1-p_{00}) = \alpha'(p_{00}+p_{11}-1) + (1-p_{00}) \tag{7.33}$$

$$E\left[\hat{\alpha}^*\right] = \alpha p_{11} + (1-\alpha)(1-p_{00}) = \alpha(p_{00}+p_{11}-1) + (1-p_{00}) \tag{7.34}$$

Combining these expressions, $E\left[(\hat{\alpha}')^* - \hat{\alpha}^*\right]$ can be simplified towards the following expression.

$$E\left[(\hat{\alpha}')^* - \hat{\alpha}^*\right] = (\alpha' - \alpha)(p_{00}+p_{11}-1) \tag{7.35}$$

Combining (7.31) and (7.35) gives the expression what should be in the big expectation of (7.30).

$$\begin{aligned} E\left[\frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\right] &= E\left[((\hat{\alpha}')^* - \hat{\alpha}^*) \times E\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right]\right] \\ &= E\left[(\hat{\alpha}')^* - \hat{\alpha}^*\right] \times E\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] \\ &= (\alpha' - \alpha)(p_{00}+p_{11}-1) \times \left[\frac{1}{p_{00}+p_{11}-1} + \frac{V(\hat{p}_{00}) + V(\hat{p}_{11})}{(p_{00}+p_{11}-1)^3}\right] + O(n^{-2}) \\ &= \alpha' - \alpha + \frac{(\alpha' - \alpha)(V(\hat{p}_{00}) + V(\hat{p}_{11}))}{(p_{00}+p_{11}-1)^2} + O(n^{-2}) \end{aligned} \tag{7.36}$$

Finalizing the proof given the equations (7.27), (7.29) and (7.36).

$$\begin{aligned} B[\hat{\alpha}'_m] &= E[\hat{\alpha}'_m] - \alpha' \\ &= \alpha + \alpha' - \alpha + \frac{(\alpha' - \alpha)(V(\hat{p}_{00}) + V(\hat{p}_{11}))}{(p_{00}+p_{11}-1)^2} - \alpha' + O(n^{-2}) \\ &= \frac{(\alpha' - \alpha)(V(\hat{p}_{00}) + V(\hat{p}_{11}))}{(p_{00}+p_{11}-1)^2} + O(n^{-2}) \end{aligned} \tag{7.37}$$

Now it only remains to proof the variance of the mixed estimator. Recall that the mixed estimator can be written as

$$\hat{\alpha}'_m = \hat{\alpha}_c + [(\hat{\alpha}')^* - \hat{\alpha}^*] \times \frac{1}{\hat{p}_{00} + \hat{p}_{11} - 1}. \tag{7.38}$$

It clearly follows from (7.38) that the variance of this mixed estimator can be written as

$$V\left[\alpha'_m\right] = V\left[\hat{\alpha}_c\right] + V\left[\frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00} + \hat{p}_{11} - 1}\right] + 2C\left[\hat{\alpha}_c, \frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00} + \hat{p}_{11} - 1}\right]. \tag{7.39}$$

From Theorem 3, we already know that the variance of the calibration estimator is equal to

$$\begin{aligned} V(\hat{\alpha}_c) = &\left[\frac{(1-\alpha)(1-p_{00}) + \alpha p_{11}}{n} + \frac{(1-\alpha)p_{00} + \alpha(1-p_{11})}{n^2}\right] \\ &\times \left[\frac{\alpha p_{11}}{(1-\alpha)(1-p_{00}) + \alpha p_{11}}\left(1 - \frac{\alpha p_{11}}{(1-\alpha)(1-p_{00}) + \alpha p_{11}}\right)\right] \\ &+ \left[\frac{(1-\alpha)p_{00} + \alpha(1-p_{11})}{n} + \frac{(1-\alpha)(1-p_{00}) + \alpha p_{11}}{n^2}\right] \\ &\times \left[\frac{(1-\alpha)p_{00}}{(1-\alpha)p_{00} + \alpha(1-p_{11})}\left(1 - \frac{(1-\alpha)p_{00}}{(1-\alpha)p_{00} + \alpha(1-p_{11})}\right)\right] \\ &+ O\left(\max\left[\frac{1}{n^3}, \frac{1}{Nn}\right]\right). \end{aligned} \tag{7.40}$$

The second term in equation (7.39) makes use of previous assumptions in this paper. We can say that $\hat{p}_{00}$ and $\hat{p}_{11}$ are independent of our Classify-and-count estimators $\hat{\alpha}^*$ and $(\hat{\alpha}')^*$. Furthermore, a well-known result on variances states that for two independent random variables A and B, it holds that $V(AB) = E[A]^2V(B) + E[B]^2V(A) + V(A)V(B)$. Combining these statements gives

$$\begin{aligned} V\left[\frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00} + \hat{p}_{11} - 1}\right] = &\left[E\left((\hat{\alpha}')^* - \hat{\alpha}^*\right)\right]^2 V\left[\frac{1}{\hat{p}_{00} + \hat{p}_{11} - 1}\right] \\ &+ \left[E\left(\frac{1}{\hat{p}_{00} + \hat{p}_{11} - 1}\right)\right]^2 V\left[(\hat{\alpha}')^* - \hat{\alpha}^*\right] \\ &+ V\left[(\hat{\alpha}')^* - \hat{\alpha}^*\right] V\left[\frac{1}{\hat{p}_{00} + \hat{p}_{11} - 1}\right]. \end{aligned} \tag{7.41}$$

Assuming that $N \gg n$, we can make the statement that $V\left[(\hat{\alpha}')^* - \hat{\alpha}^*\right]$ is of $O(\frac{1}{N})$.

$$V\left[\frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00} + \hat{p}_{11} - 1}\right] = \left[E\left((\hat{\alpha}')^* - \hat{\alpha}^*\right)\right]^2 V\left[\frac{1}{\hat{p}_{00} + \hat{p}_{11} - 1}\right] + O(\frac{1}{N}) \tag{7.42}$$

The expected value of the differences between the classify-and-count estimators is already computed in (7.35) and the variance term in (7.42) is already proven in Theorem 2. This eases the derivation of the second term in (7.39).

$$V\left[\frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00} + \hat{p}_{11} - 1}\right] = (\alpha' - \alpha)^2 \times \frac{V(\hat{p}_{00}) + V(\hat{p}_{11})}{(p_{00} + p_{11} - 1)^2} + O\left(\max\left[\frac{1}{N}, \frac{1}{n^2}\right]\right) \tag{7.43}$$

Thus it remains to evaluate the covariance term in (7.39). By conditioning on the classify-and-count estimators $\hat{\alpha}^*$ and $(\hat{\alpha}')^*$, we obtain:

$$\begin{aligned} C\left[\hat{\alpha}_c, \frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00} + \hat{p}_{11} - 1}\right] &= E\left[C\left[\hat{\alpha}_c, \frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00} + \hat{p}_{11} - 1} \mid (\hat{\alpha}')^*, \hat{\alpha}^*\right]\right] \\ &\quad + C\left[E\left[\hat{\alpha}_c \mid (\hat{\alpha}')^*, \hat{\alpha}^*\right], E\left[\frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00} + \hat{p}_{11} - 1} \mid (\hat{\alpha}')^*, \hat{\alpha}^*\right]\right] \end{aligned} \tag{7.44}$$

It can be proven that the second term of (7.44) is equal to zero. The expectation of the calibration estimator, given Classify-and-count estimators, is equal to α. This is a constant and the covariance with a constant is equal to zero.

$$\begin{aligned} E\left[\hat{\alpha}_c \mid (\hat{\alpha}')^*, \hat{\alpha}^*\right] &= E\left[\frac{n_{10}}{n_{+0}}(1 - \hat{\alpha}^*) + \frac{n_{11}}{n_{+1}}\hat{\alpha}^* \mid (\hat{\alpha}')^*, \hat{\alpha}^*\right] \\ &= E\left[\frac{n_{10}}{n_{+0}}(1 - \hat{\alpha}^*) \mid \hat{\alpha}^*\right] + E\left[\frac{n_{11}}{n_{+1}}\hat{\alpha}^* \mid \hat{\alpha}^*\right] \\ &= (1 - \hat{\alpha}^*)E\left[\frac{n_{10}}{n_{+0}} \mid \hat{\alpha}^*\right] + \hat{\alpha}^* E\left[\frac{n_{11}}{n_{+1}} \mid \hat{\alpha}^*\right] \\ &= (1 - \hat{\alpha}^*)E\left[E\left[\frac{n_{10}}{n_{+0}} \mid n_{+0}, \boldsymbol{N}\right] \mid \hat{\alpha}^*\right] + \hat{\alpha}^* E\left[E\left[\frac{n_{11}}{n_{+1}} \mid n_{+1}, \boldsymbol{N}\right] \mid \hat{\alpha}^*\right] \\ &= (1 - \hat{\alpha}^*)E\left[\frac{1}{n_{+0}}E\left[n_{10} \mid n_{+0}, \boldsymbol{N}\right] \mid \hat{\alpha}^*\right] + \alpha^* E\left[\frac{1}{n_{+1}}E\left[n_{11} \mid n_{+1}, \boldsymbol{N}\right] \mid \hat{\alpha}^*\right] \\ &= \frac{N_{+0}}{N}E\left[\frac{1}{n_{+0}}n_{+0}\frac{N_{10}}{N_{+0}} \mid \hat{\alpha}^*\right] + \frac{N_{+1}}{N}E\left[\frac{1}{n_{+1}}n_{+1}\frac{N_{11}}{N_{+1}} \mid \hat{\alpha}^*\right] \\ &= E\left[\frac{N_{10}}{N} \mid \hat{\alpha}^*\right] + E\left[\frac{N_{11}}{N} \mid \hat{\alpha}^*\right] = E\left[\frac{N_{10}}{N} + \frac{N_{11}}{N} \mid \hat{\alpha}^*\right] = E\left[\frac{N_{1+}}{N} \mid \hat{\alpha}^*\right] = \alpha \end{aligned} \tag{7.45}$$

Therefore, the covariance term can also be written as:

$$C\left[\hat{\alpha}_c, \frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00} + \hat{p}_{11} - 1}\right] = E\left[C\left[\hat{\alpha}_c, \frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00} + \hat{p}_{11} - 1} \mid (\hat{\alpha}')^*, \hat{\alpha}^*\right]\right]. \tag{7.46}$$

We can derive an expression for the inner covariance, which is written as

$$C\left[\hat{\alpha}_c, \frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00} + \hat{p}_{11} - 1} \mid (\hat{\alpha}')^*, \hat{\alpha}^*\right] = \left[(\hat{\alpha}')^* - \hat{\alpha}^*\right] C\left[\hat{\alpha}_c, \frac{1}{\hat{p}_{00} + \hat{p}_{11} - 1} \mid \hat{\alpha}^*\right]. \tag{7.47}$$

The terms in (7.47) can be written in terms of the test set $(n_{00}, n_{01}, n_{10}, n_{11})$. This eases the computation further on. Note that the elements of this test set do not depend on the Classify-and-count estimator $\hat{\alpha}^*$.

$$
\begin{aligned}
C\left[\hat{\alpha}_c, \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1} \mid (\hat{\alpha}')^*, \hat{\alpha}^*\right] &= C\left[\frac{n_{10}}{n_{+0}}(1-\hat{\alpha}^*)+\frac{n_{11}}{n_{+1}}\hat{\alpha}^*, \frac{1}{\frac{n_{00}}{n_{0+}}+\frac{n_{11}}{n_{1+}}-1} \mid \hat{\alpha}^*\right] \\
&= C\left[\frac{n_{10}}{n_{+0}}(1-\hat{\alpha}^*)+\frac{n_{11}}{n_{+1}}\hat{\alpha}^*, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}} \mid \hat{\alpha}^*\right] \\
&= (1-\hat{\alpha}^*)C\left[\frac{n_{10}}{n_{+0}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\right] \\
&\quad + \hat{\alpha}^* C\left[\frac{n_{11}}{n_{+1}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\right] \qquad (7.48)
\end{aligned}
$$

We are able to evaluate both covariance terms with the same methods. We can condition on one of the row totals. Note that the other row total is also fixed, because we work with binary classifiers ($n_{1+} = n - n_{0+}$). Furthermore, we are able to write as many variables as possible in terms of n_{0+} and n_{1+}. This helps with the Taylor Series that we apply to approximate the covariances.

$$
\begin{aligned}
&C\left[\frac{n_{10}}{n_{+0}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\right] \\
&= E\left[C\left[\frac{n_{10}}{n_{+0}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}} \mid n_{1+}\right]\right] + C\left[E\left[\frac{n_{10}}{n_{+0}} \mid n_{1+}\right], E\left[\frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}} \mid n_{1+}\right]\right] \\
&= E\left[C\left[\frac{n_{1+}-n_{11}}{n_{1+}+n_{00}-n_{11}}, \frac{n_{0+}n_{1+}}{n_{0+}n_{11}+n_{1+}n_{00}-n_{0+}n_{1+}} \mid n_{1+}\right]\right] \\
&\quad + C\left[E\left[\frac{n_{1+}-n_{11}}{n_{1+}+n_{00}-n_{11}} \mid n_{1+}\right], E\left[\frac{n_{0+}n_{1+}}{n_{0+}n_{11}+n_{1+}n_{00}-n_{0+}n_{1+}} \mid n_{1+}\right]\right] \qquad (7.49)
\end{aligned}
$$

While we condition on the row totals, the other variables in the covariance functions are n_{00} and n_{11}. Say $\frac{n_{10}}{n_{+0}} = f(n_{00}, n_{11})$ and $\frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}} = g(n_{00}, n_{11})$, with

$$
f(x,y) = \frac{n_{1+}-y}{n_{1+}+x-y} \qquad (7.50)
$$

$$
g(x,y) = \frac{n_{0+}n_{1+}}{n_{1+}x+n_{0+}y-n_{0+}n_{1+}} \qquad (7.51)
$$

we are able to compute first-order Taylor series approximations for these terms to obtain an approximation for $C\left[\frac{n_{10}}{n_{+0}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\right]$.

$$
\frac{\partial f}{\partial x} = \frac{(n_{1+}+x-y)\cdot 0-(n_{1+}-y)\cdot 1}{(n_{1+}+x-y)^2} = \frac{y-n_{1+}}{(n_{1+}+x-y)^2} \qquad (7.52)
$$

$$
\frac{\partial f}{\partial y} = \frac{(n_{1+}+x-y)\cdot -1-(n_{1+}-y)\cdot -1}{(n_{1+}+x-y)^2} = \frac{-x}{(n_{1+}+x-y)^2} \qquad (7.53)
$$

$$
\frac{\partial g}{\partial x} = \frac{-(n_{0+}n_{1+})n_{1+}}{(n_{0+}y+n_{1+}x-n_{0+}n_{1+})^2} = \frac{-n_{1+}^2 n_{0+}}{(n_{0+}y+n_{1+}x-n_{0+}n_{1+})^2} \qquad (7.54)
$$

$$
\frac{\partial g}{\partial y} = \frac{-(n_{0+}n_{1+})n_{0+}}{(n_{0+}y+n_{1+}x-n_{0+}n_{1+})^2} = \frac{-n_{0+}^2 n_{1+}}{(n_{0+}y+n_{1+}x-n_{0+}n_{1+})^2} \qquad (7.55)
$$

The approximation can be made with substituting $x = E[n_{00} \mid n_{1+}]$ and $y = E[n_{11} \mid n_{1+}]$ and applying the approximation rules for covariance. Given that n_{00} and n_{11} are independent from each other given the row totals, we can cross out $C(n_{00}, n_{11})$.

$$\begin{aligned} C\left[\frac{n_{10}}{n_{+0}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11} - n_{01}n_{10}} \mid n_{1+}\right] \approx & \frac{E[n_{11} \mid n_{1+}] - n_{1+}}{(n_{1+} + E[n_{00} \mid n_{1+}] - E[n_{11} \mid n_{1+}])^2} \\ & \times \frac{-n_{1+}^2 n_{0+}}{(n_{0+}E[n_{11} \mid n_{1+}] + n_{1+}E[n_{00} \mid n_{1+}] - n_{0+}n_{1+})^2} V(n_{00} \mid n_{1+}) \\ & + \frac{-E[n_{00} \mid n_{1+}]}{(n_{1+} + E[n_{00} \mid n_{1+}] - E[n_{11} \mid n_{1+}])^2} \\ & \times \frac{-n_{0+}^2 n_{1+}}{(n_{0+}E[n_{11} \mid n_{1+}] + n_{1+}E[n_{00} \mid n_{1+}] - n_{0+}n_{1+})^2} V(n_{11} \mid n_{1+}) \end{aligned} \tag{7.56}$$

In order to use this approximation, we can use the following properties:

$$\begin{aligned} E(n_{00} \mid n_{1+}) &= n_{0+}p_{00} \\ V(n_{00} \mid n_{1+}) &= n_{0+}p_{00}(1 - p_{00}) \\ E(n_{11} \mid n_{1+}) &= n_{1+}p_{11} \\ V(n_{11} \mid n_{1+}) &= n_{1+}p_{11}(1 - p_{11}) \end{aligned}$$

Substituting these elements gives

$$\begin{aligned} C\left[\frac{n_{10}}{n_{+0}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11} - n_{01}n_{10}} \mid n_{1+}\right] \approx & \frac{(n_{1+}p_{11}) - n_{1+}}{(n_{1+} + n_{0+}p_{00} - n_{1+}p_{11})^2} \\ & \times \frac{-n_{1+}^2 n_{0+}}{(n_{0+}(n_{1+}p_{11}) + n_{1+}(n_{0+}p_{00}) - n_{0+}n_{1+})^2} n_{0+}p_{00}(1 - p_{00}) \\ & + \frac{-n_{0+}p_{00}}{(n_{1+} + n_{0+}p_{00} - n_{1+}p_{11})^2} \\ & \times \frac{-n_{0+}^2 n_{1+}}{(n_{0+}(n_{1+}p_{11}) + n_{1+}(n_{0+}p_{00}) - n_{0+}n_{1+})^2} n_{1+}p_{11}(1 - p_{11}). \end{aligned} \tag{7.57}$$

This expression simplifies to

$$C\left[\frac{n_{10}}{n_{+0}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11} - n_{01}n_{10}} \mid n_{1+}\right] \approx \frac{n_{1+}p_{00}(1 - p_{00})(1 - p_{11}) + n_{0+}p_{00}p_{11}(1 - p_{11})}{(n_{1+} + n_{0+}p_{00} - n_{1+}p_{11})^2(p_{00} + p_{11} - 1)^2} \tag{7.58}$$

Now that the inner covariance of (7.49) is computed, we can move on and calculate the inner expectations of (7.49). This can be done with a second order Taylor series approximation.

$$\frac{\partial^2 f}{\partial x^2} = 2 \times \frac{n_{1+} - y}{(n_{1+} + x - y)^3} \tag{7.59}$$

$$\frac{\partial^2 f}{\partial y^2} = 2 \times \frac{-x}{(n_{1+} + x - y)^3} \tag{7.60}$$

$$\frac{\partial^2 g}{\partial x^2} = 2 \times \frac{n_{1+}^3 n_{0+}}{(n_{0+} y + n_{1+} x - n_{0+} n_{1+})^3} \tag{7.61}$$

$$\frac{\partial^2 g}{\partial y^2} = 2 \times \frac{n_{0+}^3 n_{1+}}{(n_{0+} y + n_{1+} x - n_{0+} n_{1+})^3} \tag{7.62}$$

Applying the Taylor rules for approximating an expected value and substituting $x = E[n_{00} \mid n_{1+}]$ and $y = E[n_{11} \mid n_{1+}]$ into the formulas gives:

$$\begin{aligned} E\left[\frac{n_{10}}{n_{+0}} \mid n_{1+}\right] \approx & \frac{n_{1+} - E[n_{11} \mid n_{1+}]}{n_{1+} + E[n_{00} \mid n_{1+}] - E[n_{11} \mid n_{1+}]} \\ & + \frac{n_{1+} - E[n_{11} \mid n_{1+}]}{(n_{1+} + E[n_{00} \mid n_{1+}] - E[n_{11} \mid n_{1+}])^3} V[n_{00} \mid n_{1+}] \\ & - \frac{E[n_{00} \mid n_{1+}]}{(n_{1+} + E[n_{00} \mid n_{1+}] - E[n_{11} \mid n_{1+}])^3} V[n_{11} \mid n_{1+}] \end{aligned} \tag{7.63}$$

$$\begin{aligned} = & \frac{n_{1+} - n_{1+} p_{11}}{n_{1+} + n_{0+} p_{00} - n_{1+} p_{11}} \\ & + \frac{n_{1+} - n_{1+} p_{11}}{(n_{1+} + n_{0+} p_{00} - n_{1+} p_{11})^3} n_{0+} p_{00} (1 - p_{00}) \\ & - \frac{n_{0+} p_{00}}{(n_{1+} + n_{0+} p_{00} - n_{1+} p_{11})^3} n_{1+} p_{11} (1 - p_{11}) \end{aligned} \tag{7.64}$$

$$= \frac{n_{1+}(1 - p_{11})}{n_{1+} + n_{0+} p_{00} - n_{1+} p_{11}} + \frac{n_{0+} n_{1+} p_{00} (p_{11} - 1)(p_{00} + p_{11} - 1)}{(n_{1+} + n_{0+} p_{00} - n_{1+} p_{11})^3} \tag{7.65}$$

$$\begin{aligned} E\left[\frac{n_{0+} n_{1+}}{n_{00} n_{11} - n_{01} n_{10}} \mid n_{1+}\right] \approx & \frac{n_{0+} n_{1+}}{n_{0+} E[n_{11} \mid n_{1+}] + n_{1+} E[n_{00} \mid n_{1+}] - n_{0+} n_{1+}} \\ & + \frac{n_{1+}^3 n_{0+}}{(n_{0+} E[n_{11} \mid n_{1+}] + n_{1+} E[n_{00} \mid n_{1+}] - n_{0+} n_{1+})^3} V[n_{00} \mid n_{1+}] \\ & + \frac{n_{0+}^3 n_{1+}}{(n_{0+} E[n_{11} \mid n_{1+}] + n_{1+} E[n_{00} \mid n_{1+}] - n_{0+} n_{1+})^3} V[n_{11} \mid n_{1+}] \end{aligned} \tag{7.66}$$

$$\begin{aligned} = & \frac{n_{0+} n_{1+}}{n_{0+} n_{1+} p_{11} + n_{1+} n_{0+} p_{00} - n_{0+} n_{1+}} \\ & + \frac{n_{1+}^3 n_{0+}}{(n_{0+} n_{1+} p_{11} + n_{1+} n_{0+} p_{00} - n_{0+} n_{1+})^3} n_{0+} p_{00} (1 - p_{00}) \\ & + \frac{n_{0+}^3 n_{1+}}{(n_{0+} n_{1+} p_{11} + n_{1+} n_{0+} p_{00} - n_{0+} n_{1+})^3} n_{1+} p_{11} (1 - p_{11}) \end{aligned} \tag{7.67}$$

$$= \frac{1}{p_{00} + p_{11} - 1} + \frac{n_{1+} p_{00} (1 - p_{00}) + n_{0+} p_{11} (1 - p_{11})}{(n_{0+} n_{1+})(p_{00} + p_{11} - 1)^3} \tag{7.68}$$

The next step is computing the outer expectation and the outer covariance of (7.49). The outer expectation can be approximated with a zero-order Taylor series.

$$E\left[C\left[\frac{n_{10}}{n_{+0}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}} \mid n_{1+}\right]\right] \approx \frac{n\alpha p_{00}(1-p_{00})(1-p_{11}) + n(1-\alpha)p_{00}p_{11}(1-p_{11})}{(n\alpha + n(1-\alpha)p_{00} - n\alpha p_{11})^2(p_{00}+p_{11}-1)^2}$$
$$= \frac{\alpha p_{00}(1-p_{00})(1-p_{11}) + (1-\alpha)p_{00}p_{11}(1-p_{11})}{n(p_{00}-\alpha(p_{00}+p_{11}-1))^2(p_{00}+p_{11}-1)^2} \tag{7.69}$$

Furthermore, it can be proven that the outer covariance of the two expectations is of $O(n^{-2})$ and can therefore be neglected in (7.49). In general, $C[f(X), g(X)] \approx f'(E[X]) \times g'(E[X]) \times V(X)$ holds. If we derive the derivations correctly, we can prove that the bigger covariance term is negligible with respect to the other terms.

Let $f(x)$ and $g(x)$ be the expectations of equation (7.66) and (7.68), with $x = n_{1+}$. Deriving them with respect to x gives.

$$f(x) = \frac{x(1-p_{11})}{x+(n-x)p_{00}-xp_{11}} + \frac{(n-x)xp_{00}(p_{11}-1)(p_{00}+p_{11}-1)}{(x+(n-x)p_{00}-xp_{11})^3}$$
$$f'(x) = \frac{np_{00}(p_{11}-1)}{(np_{00}-x(p_{00}+p_{11}-1))^2}$$
$$+ \frac{[p_{00}(1-p_{11})(p_{00}+p_{11}-1)][(2x-n)+3(x^2-nx)(np_{00}-x(p_{00}+p_{11}-1))^2(p_{00}+p_{11}-1)}{(np_{00}-x(p_{00}+p_{11}-1)^6} \tag{7.70}$$

$$g(x) = \frac{1}{p_{00}+p_{11}-1} + \frac{xp_{00}(1-p_{00})+(n-x)p_{11}(1-p_{11})}{((n-x)x)(p_{00}+p_{11}-1)^3}$$
$$g'(x) = \frac{(nx-x^2)(p_{00}(1-p_{00})-p_{11}(1-p_{11})) + (2x-n)[xp_{00}(1-p_{00})+(n-x)p_{11}(1-p_{11})])}{(nx-x^2)^2(p_{00}+p_{11}-1)^3} \tag{7.71}$$

If we substitute $x = E[n_{1+}] = n\alpha$ in the derivatives, we obtain the following expressions:

$$f'(E[n_{1+}]) = \frac{p_{00}(p_{11}-1)}{n((1-\alpha)p_{00}+\alpha(1-p_{11}))^2}$$
$$+ \frac{p_{00}(1-p_{11})(p_{00}+p_{11}-1)}{n^6((1-\alpha)p_{00}+\alpha(1-p_{11}))^6}$$
$$\times \frac{n(2\alpha-1)+3n^4(1-\alpha)((1-\alpha)p_{00}+\alpha(1-p_{11}))^2(p_{00}+p_{11}-1)}{n^6((1-\alpha)p_{00}+\alpha(1-p_{11}))^6} \tag{7.72}$$

$$g'(E[n_{1+}]) = \frac{(\alpha-\alpha^2)(p_{00}(1-p_{00})-p_{11}(1-p_{11})}{n^2(\alpha-\alpha^2)(p_{00}+p_{11}-1)^3 + (2\alpha-1)(\alpha p_{00}(1-p_{00})+(1-\alpha)p_{11}(1-p_{11}))} \tag{7.73}$$

It can be clearly seen that $f'(E[n_{1+}]) = O(\frac{1}{n})$, $g'(E[n_{1+}]) = O(\frac{1}{n^2})$ and that $V(E[n_{1+}]) = O(\frac{1}{n})$. Therefore, the whole covariance term is small enough to be negligible and that the covariance term can be written as:

$$C\left[\frac{n_{10}}{n_{+0}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\right] \approx \frac{\alpha p_{00}(1-p_{00})(1-p_{11}) + (1-\alpha)p_{00}p_{11}(1-p_{11})}{n(p_{00}-\alpha(p_{00}+p_{11}-1))^2(p_{00}+p_{11}-1)^2}. \tag{7.74}$$

Similarly, $C\left[\frac{n_{11}}{n_{+1}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\right]$ can be computed. First, $C\left[\frac{n_{11}}{n_{+1}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}} \mid n_{+1}\right]$ can be computed with a first-order Taylor series approximation. Because we condition on the row-totals, we rewrite $\frac{n_{11}}{n_{+1}}$ as

$$\frac{n_{11}}{n_{+1}} = \frac{n_{11}}{n - n_{00} - n_{10}} = \frac{n_{11}}{n - n_{00} - (n_{1+} - n_{11})} = \frac{n_{11}}{n_{0+} - n_{00} + n_{11}}$$

and make a function dependent on $x = n_{00}$ and $y = n_{11}$, which we can derive.

$$\begin{aligned} h(x,y) &= \frac{y}{n_{0+} - x + y} \\ \frac{\partial h}{\partial x} &= \frac{y}{(n_{0+} - x + y)^2} \qquad (7.75) \\ \frac{\partial h}{\partial y} &= \frac{n_{0+} - x}{(n_{0+} - x + y)^2} \qquad (7.76) \end{aligned}$$

Accordingly, we can borrow the expectations from the previous covariance term. Therefore we end up with the following term:

$$\begin{aligned} C\left[\frac{n_{11}}{n_{+1}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11} - n_{01}n_{10}} \mid n_{1+}\right] \approx & \frac{n_{1+}p_{11}}{(n_{0+}(1-p_{00}) + n_{1+}p_{11})^2} \\ & \times \frac{-n_{1+}^2 n_{0+}}{(n_{0+}(n_{1+}p_{11}) + n_{1+}(n_{0+}p_{00}) - n_{0+}n_{1+})^2} n_{0+}p_{00}(1-p_{00}) \\ & + \frac{n_{0+}(1-p_{00})}{(n_{0+}(1-p_{00}) + n_{1+}p_{11})^2} \\ & \times \frac{-n_{0+}^2 n_{1+}}{(n_{0+}(n_{1+}p_{11}) + n_{1+}(n_{0+}p_{00}) - n_{0+}n_{1+})^2} n_{1+}p_{11}(1-p_{11}). \end{aligned} \qquad (7.77)$$

This simplifies to:

$$C\left[\frac{n_{11}}{n_{+1}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11} - n_{01}n_{10}} \mid n_{1+}\right] \approx -\frac{n_{1+}p_{00}(1-p_{00})p_{11} + n_{0+}(1-p_{00})p_{11}(1-p_{11})}{(n_{0+}(1-p_{00}) + n_{1+}p_{11})^2(p_{00} + p_{11} - 1)^2} \qquad (7.78)$$

The next step is computing the expected value of this expression.

$$\begin{aligned} E\left[C\left[\frac{n_{11}}{n_{+1}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11} - n_{01}n_{10}} \mid n_{1+}\right]\right] &\approx -\frac{n\alpha p_{00}(1-p_{00})p_{11} + n(1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{(n(1-\alpha)(1-p_{00}) + n\alpha p_{11})^2(p_{00} + p_{11} - 1)^2} \\ &= -\frac{\alpha p_{00}(1-p_{00})p_{11} + (1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00}) + \alpha p_{11})^2(p_{00} + p_{11} - 1)^2} \end{aligned} \qquad (7.79)$$

The covariance between the expectation is again of a negligible low order, so the covariance term can be written as:

$$C\left[\frac{n_{11}}{n_{+1}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11} - n_{01}n_{10}}\right] \approx -\frac{\alpha p_{00}(1-p_{00})p_{11} + (1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00}) + \alpha p_{11})^2(p_{00} + p_{11} - 1)^2} \qquad (7.80)$$

Now that we have obtained the two conditional covariance in (7.74) and (7.80), we can substitute these terms in (7.48).

$$C\left[\hat{\alpha}_c, \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1} \mid (\hat{\alpha}')^*, \hat{\alpha}^*\right] \approx (1-\hat{\alpha}^*) \times \frac{\alpha p_{00}(1-p_{00})(1-p_{11}) + (1-\alpha)p_{00}p_{11}(1-p_{11})}{n((1-\alpha)p_{00}+\alpha(1-p_{11}))^2(p_{00}+p_{11}-1)^2} - \hat{\alpha}^* \times \frac{\alpha p_{00}(1-p_{00})p_{11} + (1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00})+\alpha p_{11})^2(p_{00}+p_{11}-1)^2} \tag{7.81}$$

Combining (7.46), (7.47) and (7.81), we can compute $C\left[\hat{\alpha}_c, \frac{(\hat{\alpha}')^*-\hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\right]$ by taking the expected value of the difference between the Classify-and-count estimators multiplied by the expected value of (7.81). Note that the first part of both denominators are equal to respectively the expected value of $(1-\hat{\alpha}^*)$ and $\hat{\alpha}^*$ squared.

$$\begin{aligned}
C\left[\hat{\alpha}_c, \frac{(\hat{\alpha}')^*-\hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\right] &= E\left[[(\hat{\alpha}')^*-\hat{\alpha}^*]\, C\left[\hat{\alpha}_c, \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1} \mid \hat{\alpha}^*\right]\right] \\
&\approx E\Bigg[[(\hat{\alpha}')^*-\hat{\alpha}^*]\Bigg[(1-\hat{\alpha}^*)\frac{\alpha p_{00}(1-p_{00})(1-p_{11}) + (1-\alpha)p_{00}p_{11}(1-p_{11})}{n(p_{00}-\alpha(p_{00}+p_{11}-1))^2(p_{00}+p_{11}-1)^2} \\
&\quad - \hat{\alpha}^* \times \frac{\alpha p_{00}(1-p_{00})p_{11} + (1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00})+\alpha p_{11})^2(p_{00}+p_{11}-1)^2}\Bigg]\Bigg] \\
&= E\Bigg[[(\hat{\alpha}')^*-\hat{\alpha}^*]\left[\frac{\alpha p_{00}(1-p_{00})(1-p_{11}) + p_{00}p_{11}(1-p_{11})}{n(p_{00}-\alpha(p_{00}+p_{11}-1))(p_{00}+p_{11}-1)^2}\right. \\
&\quad \left. - \frac{\alpha p_{00}(1-p_{00})p_{11} + (1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00})+\alpha p_{11})(p_{00}+p_{11}-1)^2}\right]\Bigg] \\
&= [(\alpha')^*-\alpha^*]\,(p_{00}+p_{11}-1)\left[\frac{\alpha p_{00}(1-p_{00})(1-p_{11}) + p_{00}p_{11}(1-p_{11})}{n(p_{00}-\alpha(p_{00}+p_{11}-1))(p_{00}+p_{11}-1)^2}\right. \\
&\quad \left. - \frac{\alpha p_{00}(1-p_{00})p_{11} + (1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00})+\alpha p_{11})(p_{00}+p_{11}-1)^2}\right] \\
&= [(\alpha')-\alpha]\left[\frac{\alpha p_{00}(1-p_{00})(1-p_{11}) + p_{00}p_{11}(1-p_{11})}{n(p_{00}-\alpha(p_{00}+p_{11}-1))(p_{00}+p_{11}-1)}\right. \\
&\quad \left. - \frac{\alpha p_{00}(1-p_{00})p_{11} + (1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00})+\alpha p_{11})(p_{00}+p_{11}-1)}\right]
\end{aligned} \tag{7.82}$$

Combining all elements gives the total variance of the mixed estimator.

$$
\begin{aligned}
V(\hat{\alpha}'_m) = &\left[\frac{(1-\alpha)(1-p_{00})+\alpha p_{11}}{n} + \frac{(1-\alpha)p_{00}+\alpha(1-p_{11})}{n^2}\right] \\
&\times \left[\frac{\alpha p_{11}}{(1-\alpha)(1-p_{00})+\alpha p_{11}}\left(1 - \frac{\alpha p_{11}}{(1-\alpha)(1-p_{00})+\alpha p_{11}}\right)\right] \\
&+ \left[\frac{(1-\alpha)p_{00}+\alpha(1-p_{11})}{n} + \frac{(1-\alpha)(1-p_{00})+\alpha p_{11}}{n^2}\right] \\
&\times \left[\frac{(1-\alpha)p_{00}}{(1-\alpha)p_{00}+\alpha(1-p_{11})}\left(1 - \frac{(1-\alpha)p_{00}}{(1-\alpha)p_{00}+\alpha(1-p_{11})}\right)\right] \\
&+ (\alpha'-\alpha)^2 \times \frac{V(\hat{p}_{00})+V(\hat{p}_{11})}{(p_{00}+p_{11}-1)^2} \\
&+ (\alpha'-\alpha) \times \left[\frac{\alpha p_{00}(1-p_{00})(1-p_{11})+p_{00}p_{11}(1-p_{11})}{n(p_{00}-\alpha(p_{00}+p_{11}-1))(p_{00}+p_{11}-1)}\right. \\
&\left. - \frac{\alpha p_{00}(1-p_{00})p_{11}+(1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00})+\alpha p_{11})(p_{00}+p_{11}-1)}\right] + O(n^{-2}). \qquad (7.83)
\end{aligned}
$$

□

This concludes the proof of the bias and variance of the mixed estimator.