Extended Methods to Handle Classification Biases

Emma Beauxis-Aussalet[1,2] and Lynda Hardman[1,2]

[1] CWI, Amsterdam, The Netherlands
[2] Utrecht University, Utrecht, The Netherlands
emalb@cwi.nl lynda.hardman@cwi.nl

Published in IEEE conference on Data Science and Advanced Analytics (DSAA)
DOI: 10.1109/DSAA.2017.52 (2017)

Abstract. Classifiers can be used to analyse class sizes, i.e., counts of items per class, but systematic classification errors yield biases (e.g., if a class is often misclassified as another, its size may be under-estimated). To handle classification biases, the statistics and epidemiology domains devised methods for estimating unbiased class sizes (or class probabilities) without identifying which individual items are misclassified. These bias correction methods are applicable to machine learning classifiers, but in some cases yield high result variance and increased biases. We present the applicability and drawbacks of existing methods and extend them with three novel methods. Our **Sample-to-Sample method** provides accurate confidence intervals for the bias correction results. Our **Maximum Determinant method** predicts which classifier yields the least result variance. Our **Ratio-to-TP method** provides detailed error estimations (i.e., how many items classified as class X truly belong to class Y, for all possible classes) and has properties of interest for applying the Maximum Determinant method. Our methods are demonstrated empirically, and we discuss the need for establishing theory and guidelines for choosing the methods to apply.

Keywords: Classification · Bias correction · Error estimation.

Methods for correcting biases due to systematic misclassifications have been researched in statistics and epidemiology [1,4], but seldom considered in machine learning besides land coverage estimation [2,3]. Yet a variety of use cases would benefit from applying bias correction methods, e.g., for analysing class sizes and distributions. For instance, let us consider ecologists classifying images of animals to analyse the species abundance. If species X is systematically misclassified as species Y, it yields under-estimations of species X and over-estimations of species Y. If species X increases over time while species Y is stable, the individuals from X misclassified as Y increase too. It yields a deceptive increase of species Y in the classification data. **Without applying bias correction methods, no scientific conclusion can be drawn from the classification data**.

Existing bias correction methods aim at estimating unbiased class sizes (i.e., numbers of items belonging to each class) or class proportions (i.e., class sizes divided by total number of items, also considered as class probabilities) without identifying which individual items are misclassified. The methods assume that

error rates measured in test sets are the same in the datasets to which classifiers are applied in practice (the *target sets*). Bias correction results are subject to potentially high variance due to random error rate deviations between test and target sets. For small datasets, the variance magnitude is critical and applying bias correction methods may even worsen the initial biases. **Two bias correction methods exist, one of which requires equal class proportions between test and target sets but has the least result variance**.

We extend bias correction methods to estimating the numbers of errors in a classifier output, i.e., within the items classified as class Y, how many would truly belong to class X. Such estimates describe the quality of classification data beyond accuracy or precision. In future research, they can also help identifying which individual items are misclassified [3]. We introduce a novel error estimation method, called Ratio-to-TP method, that has interesting properties to ensure the applicability of a bias correction method, and to predict its result variance.

The variance of bias correction results can be critical and is thus crucial to estimate. Existing variance estimation methods do not address the case of disjoint test and target sets, which is common in machine learning applications. Our Sample-to-Sample method addresses this issue. It estimates the variance at the level of the error rate estimator, using each class size in both test and target sets. It provides accurate confidence intervals for bias correction results in binary problems. Multiclass problems require bootstrapping techniques, or in simulations using Sample-to-Sample to specify error rates' variance.

Finally, we introduce the Maximum Determinant method for **predicting which classifier yields the least variance** when applying a bias correction method, without knowledge of the potential target sets. Initial results are promising but future research is needed to establish theory and investigate the applicability, e.g., depending on class sizes and proportions, number of classes, and error rate magnitudes. Future research is also needed if **feature distributions** (e.g., class models) differ between test and target sets (e.g., domain adaption). We illustrate such cases and their **critical impact on bias correction results**.

References

1. Buonaccorsi, J.P.: Measurement Error: Models, Methods and Applications. CRC Press, Taylor and Francis (2010)
2. Card, D.H.: Using known map category marginal frequencies to improve estimates of thematic map accuracy. Photogrammetric Engineering and Remote Sensing **48**, 431–439 (1982)
3. van Deusen, P.C.: Unbiased estimates of class proportions from thematic maps. Photogrammetric Engineering and Remote Sensing **62**(4), 409–412 (1996)
4. Tenenbein, A.: A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection. Technometrics **14**(1), 187–202 (1972)

[3] For instance, provided with class probabilities for each item and each class, and estimated number of error n_{xy} between classes X and Y, a method can be developed to select the n_{xy} items that are most likely classified as Y while belonging to X.

2017 International Conference on Data Science and Advanced Analytics

Extended Methods to Handle Classification Biases

Emma Beauxis-Aussalet
CWI and Utrecht University, The Netherlands
Email: emalb@cwi.nl

Lynda Hardman
CWI and Utrecht University, The Netherlands
Email: lynda.hardman@cwi.nl

***Abstract*—Classifiers can provide counts of items per class, but systematic classification errors yield biases (e.g., if a class is often misclassified as another, its size may be under-estimated). To handle classification biases, the statistics and epidemiology domains devised methods for estimating unbiased class sizes (or class probabilities) without identifying which individual items are misclassified. These bias correction methods are applicable to machine learning classifiers, but in some cases yield high result variance and increased biases. We present the applicability and drawbacks of existing methods and extend them with three novel methods. Our Sample-to-Sample method provides accurate confidence intervals for the bias correction results. Our Maximum Determinant method predicts which classifier yields the least result variance. Our Ratio-to-TP method details the error decomposition in classifier outputs (i.e., how many items classified as class C_y truly belong to C_x, for all possible classes) and has properties of interest for applying the Maximum Determinant method. Our methods are demonstrated empirically, and we discuss the need for establishing theory and guidelines for choosing the methods and classifier to apply.**

I. Introduction

Methods for correcting biases due to systematic misclassifications have been thoroughly researched in statistics and epidemiology [1]–[4], but seldom considered in machine learning besides land coverage estimation [5]–[8]. Yet a variety of use cases would benefit from applying bias correction methods, e.g., for analysing class sizes and distributions. For instance, let us consider ecologists classifying images of animals with computer vision software to analyse the species abundance. If species X is systematically misclassified as species Y, it yields under-estimations of species X and over-estimations of species Y. If species X increases over time while species Y is stable, the individuals from X misclassified as Y increase too. It yields a deceptive increase of species Y in the classification data. Without applying bias correction methods, no scientific conclusion can be drawn from the classification data.

Bias correction methods are based on error rates measured in a sample of items (the *test set*, also called groundtruth, gold standard, validation or calibration set). The error rates are assumed to be the same in the datasets to which classifiers are applied in practice (the *target sets*). Bias correction methods aim at estimating unbiased class sizes (i.e., numbers of items belonging to each class) or class proportions (i.e., class sizes divided by total number of items, also considered as class probabilities) without identifying which individual items are misclassified. Bias correction results are subject to potentially high variance due to random error rate deviations between test and target sets. The variance magnitude depends on bias correction methods. For small datasets, the variance magnitude is critical and applying bias correction methods may even worsen the initial biases. Two bias correction methods exist, one of which requires equal class proportions between test and target sets but has the least result variance (Section II).

The bias correction methods can be extended to detail the error composition in a classifier output, i.e., within the items classified as class Y, how many would truly belong to class X. Such estimates describe the quality of classification data beyond accuracy or precision. In future research, they can also help identifying which individual items are misclassified. For instance, provided with i) probabilities that an item belongs to a class, for all classes and items; and ii) estimated numbers n_{xy} of items classified as Y and truly belonging to class X; a method can be developed to select the n_{xy} items that are most likely classified as Y while belonging to X. We introduce a novel method for estimating the error decomposition, called Ratio-to-TP method. It provides exactly the same result as the extended state-of-the-art methods, but has interesting properties to ensure the applicability of a bias correction method, and to predict its result variance (Section III).

The variance of bias correction results can be critical and is thus crucial to estimate. Variance estimation methods exist for uses cases where the test set is randomly sampled within the target set, with the same class proportions [1]. For disjoint test and target sets, existing variance estimators describe the population from which target sets are sampled [2]–[4], [7]. If applied to describe the target set itself (i.e., class sizes, error composition), they provide inaccurate estimates. Our Sample-to-Sample method handles the latter issue. It estimates the variance at the level of the error rate estimator, using each class size in both test and target sets. It provides accurate confidence intervals for bias correction results in binary problems. Multiclass problems require bootstrapping techniques, or future research using Sample-to-Sample in simulations (Section IV).

Finally, we introduce the Maximum Determinant method for predicting which classifier yields the least variance when applying a bias correction method, without knowledge of the potential target sets. Initial results are promising but future research is needed to establish theory and investigate the applicability, e.g., depending on class sizes and proportions, number of classes, and error rate magnitudes (Section V).

Future research is also needed if feature distributions (e.g., class models) differ between test and target sets (e.g., domain adaption). We illustrate such cases and their critical impact on bias correction results (Section VI).

978-1-5090-5004-8/17 $31.00 © 2017 IEEE
DOI 10.1109/DSAA.2017.52

II. Existing Bias Correction Methods

Two bias correction methods exist: i) the *reclassification method* [4], also called double sampling [1], ratio method [6] or inverse calibration [9], which requires equal class proportions in test and target sets (Section II-A); ii) the *misclassification method* [4], also called matrix inversion method [10], classical calibration [9] or PERLE [11], which is robust to varying class proportions (Section II-B). The misclassification method yields a larger results variance than the reclassification method, as mentioned in [3] and shown in Fig. 1. It is preferable to use the reclassification method and a test set with class proportions similar to the target set. But this is often impossible as class proportions may vary over target sets, or are unknown when the test set is collected.

We introduce the existing methods using the notation in Table I where n_{xy} are numbers of items belonging to class C_x and classified as C_y, $n_{x.}$ is the *true class size* for C_x, $n_{.x}$ the *output class size* from a classifier results, and $n_{..}$ the total number of items in the test set. The variables for the target set are denoted with the prime symbols, e.g., $n_{1.}$ is the true size of class C_1 in the test set, and $n'_{1.}$ the true class size in the target set. The bias correction methods estimate true class sizes $n'_{x.}$ in the target set, given the known output class sizes $n'_{.x}$ and numbers of error n_{xy} measured in the test set. We present bias correction results in terms of class size estimates $n'_{x.}$ rather than class proportion $n'_{x.}/n'_{..}$, the latter being easily derived from the former.

TABLE I
Confusion matrix and notation

		True Class				Output Count
		C_1	C_2	...	C_k	
Assigned Class	C_1	n_{11}	n_{21}	...	n_{k1}	$n_{.1}$
	C_2	n_{12}	n_{22}	...	n_{k2}	$n_{.2}$
	...	...	...	...	...	...
	C_k	n_{1k}	n_{2k}	...	n_{kk}	$n_{.k}$
True Count		$n_{1.}$	$n_{2.}$	...	$n_{k.}$	Total $n_{..}$

A. Reclassification Method

The reclassification method is based on error rates using the output class sizes $n_{.y}$ as denominators, e.g., precision in binary problems. Assuming equal error rates in test and target sets $\widehat{e'_{xy}} = e_{xy}$, true class sizes are estimated as (1). This assumption is violated, and the method is not applicable, if class proportions differ between test and target sets (see Subsection D).

$$e_{xy} = \frac{n_{xy}}{n_{.y}} \qquad \widehat{n'_{xy}} = e_{xy} n'_{.y} \qquad \widehat{n'_{x.}} = \sum_y e_{xy} n'_{.y} \tag{1}$$

Variance estimates $V(\widehat{n'_{x.}})$ are provided in [1] for test sets randomly sampled with target sets, using a weighted sum to account for both test and target set sizes.

B. Misclassification Method

The misclassification method is based on error rates with true class size $n_{x.}$ as denominator, e.g., recall in binary problems (2). Assuming equal error rates in test and target sets $\widehat{\theta'_{xy}} = \theta_{xy}$, true class sizes are estimated as (3), e.g., solving a system of linear equations as in [11].

$$\theta_{xy} = \frac{n_{xy}}{n_{x.}} \tag{2}$$

$$\begin{pmatrix} \widehat{n'_{1.}} \\ \widehat{n'_{2.}} \\ ... \\ \widehat{n'_{x.}} \end{pmatrix} = \begin{pmatrix} \theta_{11} & \theta_{21} & ... & \theta_{x1} \\ \theta_{12} & \theta_{22} & ... & \theta_{x2} \\ ... & ... & ... & ... \\ \theta_{1x} & \theta_{2x} & ... & \theta_{xx} \end{pmatrix}^{-1} \begin{pmatrix} n'_{.1} \\ n'_{.2} \\ ... \\ n'_{.x} \end{pmatrix} \tag{3}$$

Variance estimates $V(\widehat{n'_{x.}})$ are provided in [2] for test sets randomly sampled within target sets, and with similar class proportions $n_{x.}/n_{..} \simeq n'_{x.}/n'_{..}$. The case of disjoint test and target sets with different class proportions is addressed in [3], [4] and refined in Section IV.

C. Application

We apply reclassification and misclassification methods to open-source datasets from the UCI repository. We use a Naive Bayes classifier from the Weka platform with 10-fold cross validation. We randomly sample test sets of predefined sizes, and consider the remaining items as target sets. The predefined class sizes split the data into test and target sets with different class proportions (Table II). We draw 100 random splits to show the variance and bias in the initial classification results, and in the bias correction results (Fig. 1). The reclassification method yields biased results (median results are not on dashed line) as class proportions differ between test and target sets. The misclassification method is unbiased but yields larger variance than the reclassification method.

TABLE II
Datasets used for experiments in Fig. II

Dataset	Test Set Size $n_{x.}$	Target Set Size $n'_{x.}$
Iris	C_1:25 C_2:20 C_3:30	C_1:25 C_2:30 C_3:20
Ionosphere	C_1:63 C_0:150	C_1:63 C_0:75
Segment	$C_{1,3,5,7}$:210 $C_{2,4,6}$:110	$C_{1,3,5,7}$:120 $C_{2,4,6}$:220
Ohscal	C_0:471 C_1:433 C_2:124 C_3:125 C_4:275 C_5:205 C_6:738 C_7:339 C_8:490 C_9:613	C_0-C_{10}:400
Waveform	C_1:600 C_2:900 C_3:1200	C_1:1092 C_2:753 C_3:455
Chess	C_1:1000 C_0:500	C_1:669 C_0:1027

Data source: https://archive.ics.uci.edu/ml/datasets.html

D. Discussion

The misclassification method is unaffected by changes in class proportions because its error rates θ_{xy} involve items belonging to the same true class, unlike the error rates e_{xy} of the reclassification method, as shown in (4).

Class proportions: $\frac{n'_{x.}}{n'_{..}} = \alpha \frac{n_{x.}}{n_{..}}, \quad \frac{n'_{y.}}{n'_{..}} = \beta \frac{n_{y.}}{n_{..}}, \quad \alpha \neq \beta, \ \alpha, \beta \in \mathbb{R}_{<0}$

Assuming proportional errors $n'_{xy} = \alpha n_{xy}$ and $n'_{yx} = \beta n_{yx}$ then:

$$\theta'_{xy} = \frac{\alpha n_{xy}}{\alpha n_{x.}} = \theta_{xy} \qquad e'_{xy} = \frac{\alpha n_{xy}}{\beta n_{y.}} \neq e_{xy} \tag{4}$$

The misclassification method yields significantly higher variance than the reclassification method. The latter uses a simple linear sum of random variable $n'_{.y}\boldsymbol{e_{xy}}$ while the former uses a matrix inversion. Cramer's rule [12] shows that the random variables $\boldsymbol{\theta_{xy}}$ are involved several times in the denominator and numerator of a fraction, hence the higher variance (estimator is not linear).

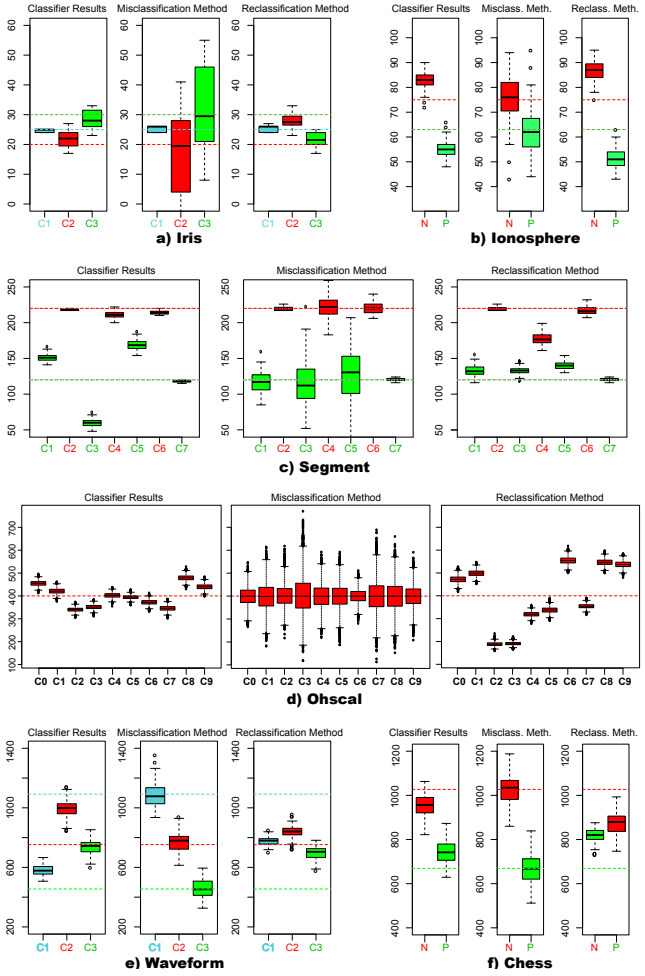

Fig. 1. Counts of items per class (median, 50%, 95% quartiles for 100 randomly sampled test and target sets) provided by the raw classifier output (left graphs), bias correction with the misclassification method (middle graphs) and reclassification method (right graphs). Horizontal dashed lines indicate true class sizes, and colors indicate the related class (e.g., green boxplots with median values on green dashed lines indicate unbiased results).

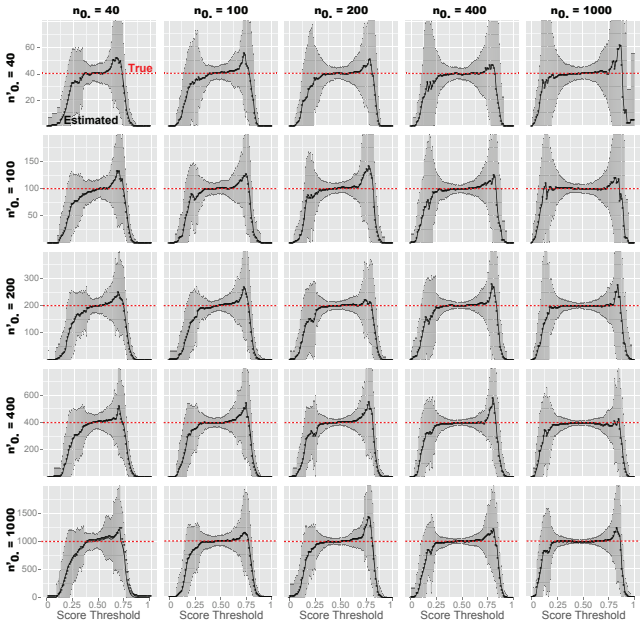

Fig. 2. Results of misclassification method for simulated data. Score thresholds (x axis) are used to assign classes C_0 or C_1 (Fig. 3), and simulate different magnitudes of error rate. Class sizes $\widehat{n'_{0.}}$ (y axis) are estimated for 10^4 pairs of test and target sets randomly sampled with score probability and class proportions in Fig. 3. Unbiased means $\widehat{n'_{0.}}$ (black line) are close to true $n'_{0.}$ (red line) unless test sets are too small and error rate too close to 0 or 1 (extreme thresholds yielding few observations, e.g., $n_{xy} \approx$ a few items).

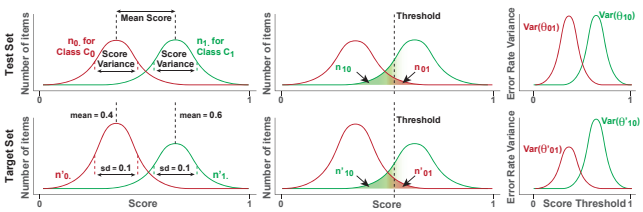

Fig. 3. Specification of classification problem in Fig. 2. Left: score distribution with means μ_0=μ'_0=0.4 for C_0, μ_1=μ'_1=0.6 for C_1, and σ_x=σ'_x=0.1. Middle: example of score threshold and related errors n_{01}, n_{10}. Right: error rate variance over thresholds. $V(\theta'_{01})$<$V(\theta'_{10})$ because we use $n'_{0.}$=$2n'_{1.}$ and $n_{0.}$=$n_{1.}$ to simulate different class proportions in test and target sets.

If the test or target sets are small, or the change in class proportions not significant, the variance of the misclassification method may introduce more bias than the reclassification method or the initial classification results (Fig. 1-a to -d). Combining both methods does not reduce the variance (e.g., estimate $n'_{x.}$ with the misclassification method, subsample the test set with similar class proportions $n_{x.}$=$\alpha n'_{x.}$ $\forall x$, and apply reclassification method using the resampled test set)[1].

To address the challenge of high result variance for the misclassification method, we introduce novel methods for estimating result variance for specific target sets (Section IV), and for minimising the variance without knowledge of the potential target sets (Section V). These methods are also able to deal with the variance of n'_{xy} estimates describing the error decomposition of $n'_{x.}$ (Section III).

[1]Demonstration omitted for brevity but reproducible with code in Appendix.

III. Error Decomposition

We extend the bias correction methods to detail the class-to-class errors in a classifier output, i.e., in an output of $n'_{.y}$ items classified as C_y, we estimate the n'_{xy} items that truly belong to C_x. Such estimates are of interest for describing the quality of classification data, and potentially for identifying which items are misclassified. For instance, if a classifier provides class probabilities for each item, a method can infer which items are most likely to be classified as C_x while belonging to C_y.

If class proportions do not differ between test and target sets, the reclassification method is a trivial solution, i.e., $\widehat{n'_{xy}}$=$e_{xy}n'_{.y}$ (1). We discuss two methods addressing varying class proportions (Subsections A and B). Both result in the exact same estimates, impacted by the same variance[1], but use different error rates. We discuss both methods because their error rate matrices have specific properties of interest.

A. Extension of Misclassification Method

The misclassification method is easily extended to estimate n'_{xy} as (5), after estimating the class size $\widehat{n'_{x.}}$.

$$\widehat{n'_{xy}} = \theta_{xy}\widehat{n'_{x.}} \tag{5}$$

B. Ratio-to-TP Method

The Ratio-to-TP method is based on atypical error ratios using True Positives n_{xx} as denominators (6), with r_{xx}=1 (assuming $n_{xx}\neq 0$). The method assumes equal error rates in test and target sets, i.e., $\widehat{r'_{xy}} = r_{xy}$. The true positives in the target set n'_{xx} are estimated by solving the linear system (7) in (8). The number of errors n'_{xy} are derived using $\widehat{n'_{xx}}$ in (9).

$$r_{xy} = \frac{n_{xy}}{n_{xx}} \qquad n'_{.y} = \sum_x n'_{xy} = \sum_x n'_{xx} r'_{xy} \tag{6}$$

$$\begin{Bmatrix} n'_{.1} = & n'_{11} & + & n'_{22}\, r'_{21} & + & \ldots + & n'_{xx}\, r'_{x1} \\ n'_{.2} = & n'_{11}\, r'_{12} & + & n'_{22} & + & \ldots + & n'_{xx}\, r'_{x2} \\ \ldots & = & \ldots & + & \ldots & + & \ldots + & \ldots \\ n'_{.x} = & n'_{11}\, r'_{1x} & + & n'_{22}\, r'_{2x} & + & \ldots + & n'_{xx} \end{Bmatrix} \tag{7}$$

$$\begin{pmatrix} \widehat{n'_{11}} \\ \widehat{n'_{22}} \\ \ldots \\ \widehat{n'_{xx}} \end{pmatrix} = \begin{pmatrix} 1 & r_{21} & \ldots & r_{x1} \\ r_{12} & 1 & \ldots & r_{x2} \\ \ldots & \ldots & \ldots & \ldots \\ r_{1x} & r_{2x} & \ldots & 1 \end{pmatrix}^{-1} \begin{pmatrix} n'_{.1} \\ n'_{.2} \\ \ldots \\ n'_{.x} \end{pmatrix} \tag{8}$$

$$\widehat{n'_{xy}} = \widehat{n'_{xx}}\; r_{xy} \tag{9}$$

C. Application

We apply the Ratio-to-TP and extension of misclassification methods, using the same experimental setup as in Section II-C. Both methods result in the same estimates[2], which are unbiased but with potentially high variance due to random differences between θ_{xy} and θ'_{xy} (Fig. 4). The potentially high variance is a challenge for estimating both $\widehat{n'_{x.}}$ and $\widehat{n'_{xy}}$.

D. Discussion

The error rate matrix $\mathbf{M_r}$=$\begin{pmatrix} 1 & r_{2x} & \ldots \\ r_{x2} & 1 & \ldots \\ \ldots & \ldots & \ldots \end{pmatrix}$ of the Ratio-to-TP method has all diagonal values equal to 1. It offers a simple condition to ensure its invertibility (i.e., that its determinant $|\mathbf{M_r}|\neq 0$), needed for the Ratio-to-TP method to be applicable. Under condition (10) $\mathbf{M_r^\top}$ is diagonally dominant, thus invertible, and since $|\mathbf{M_r}| = |\mathbf{M_r^\top}|$, $\mathbf{M_r}$ is also invertible. Setting a threshold t for all error rates $r_{xy,x\neq y} < t$ can ensure that (10) is satisfied. $\mathbf{M_r}$ is always invertible under condition (11) where c is the number of classes (e.g., for 3-class problems t=0.5, 4-class t=0.33, 5-class t=0.25). It is also possible that $\mathbf{M_r}$ is invertible even if the condition is not met.

$$|\mathbf{M_r}| \neq 0 \;\text{ if for all } C_x \sum_{y,y\neq x} r_{xy} < 1 \tag{10}$$

$$\text{If all } r_{xy} < \frac{1}{c-1} \text{ then } \sum_{y,y\neq x} r_{xy} < (c-1)\frac{1}{c-1} = 1$$

$$\text{Thus } |\mathbf{M_r}| \neq 0 \;\text{ if all }\; r_{xy,y\neq x} < \frac{1}{c-1} \tag{11}$$

[2]Demonstration is omitted but reproducible with code in Appendix.

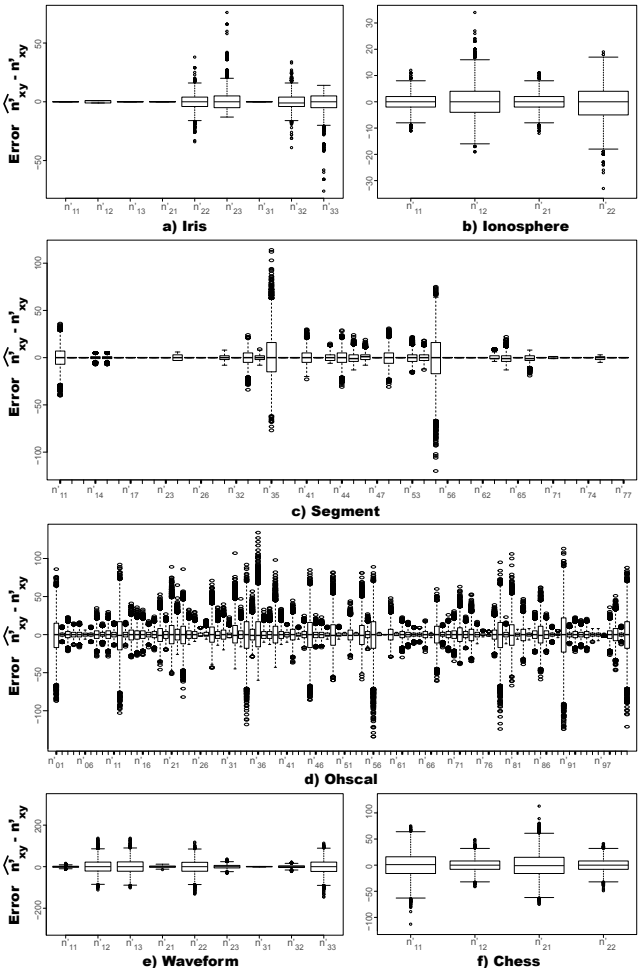

Fig. 4. Evaluation of estimated $\widehat{n'_{xy}}$, showing the absolute error $n'_{xy} - \widehat{n'_{xy}}$ for 10^4 pairs test and target sets sampled as in Section II-C.

The misclassification method also requires its error rate matrix $\mathbf{M}_\theta$ = $\begin{pmatrix} \theta_{11} & \theta_{21} & \ldots \\ \theta_{12} & \theta_{22} & \ldots \\ \ldots & \ldots & \ldots \end{pmatrix}$ to be invertible, but the Ratio-to-TP method offers a simple threshold condition to guarantee its matrix invertibility. We empirically observed that error rate matrices $\mathbf{M_r}$ and $\mathbf{M}_\theta$ drawn from the same test set were either both invertible, or both non-invertible. Future work is needed to establish if the threshold condition (11) ensuring the invertibility $\mathbf{M_r}$ also ensures the invertibility of $\mathbf{M}_\theta$.

The Ratio-to-TP method uses error ratios r_{xy} that follow a Cauchy distribution, in contrast to θ_{xy} which follows a binomial distribution. Estimating the variance $V(r_{xy})$ is more complex. Hence in the next section we focus on error rates θ_{xy} to estimate the variance of $\widehat{n'_{x.}}$ and $\widehat{n'_{xy}}$ derived from the misclassification method.

IV. Sample-to-Sample Method

The Sample-to-Sample method aims at estimating the variance estimates $\widehat{\theta'_{xy}}$, $\widehat{n'_{x.}}$, $\widehat{n'_{xy}}$ for a target set $\boldsymbol{S'}$, using measurements from a disjoint test set $\boldsymbol{S}$ (i.e., $\boldsymbol{S}\cap\boldsymbol{S'}$=$\varnothing$). We first approximate the variance of the $\widehat{\theta'_{xy}}$ estimator, and validate our approach using known $n'_{x.}$. The method is then evaluated

in practice with unknown $n'_{x.}$, using estimated $\widehat{n'_{x.}}$ instead. The method performs well for estimating the variance of $\widehat{n'_{x.}}$ and $\widehat{n'_{xy}}$ in binary problems. Multiclass problems require bootstrapping techniques, or future work on simulations using Sample-to-Sample estimates of $\widehat{V}\left(\widehat{\theta'_{xy}}\right)$ (see Subsection E).

A. Error Rate Estimator

We focus on the estimator $\widehat{\theta'_{xy}}=\theta_{xy}$ for the unknown target set error rate θ'_{xy} based on the known error rate θ_{xy} in a disjoint test set. Test and target sets are assumed to be randomly sampled from the same population $n^*_{x.}\to\infty$ with error rate θ^*_{xy}. For test and target sets sampled with $n_{x.}$ and $n'_{x.}$ items, the expected value and variance of $\boldsymbol{\theta_{xy}}$ and $\boldsymbol{\theta'_{xy}}$ are given in (12) [13].

$$E[\boldsymbol{\theta_{xy}}] = E[\boldsymbol{\theta'_{xy}}] = \theta^*_{xy}$$
$$V(\boldsymbol{\theta_{xy}}) = \frac{\theta^*_{xy}(1-\theta^*_{xy})}{n_{x.}} \qquad V(\boldsymbol{\theta'_{xy}}) = \frac{\theta^*_{xy}(1-\theta^*_{xy})}{n'_{x.}} \tag{12}$$

The estimator $\widehat{\boldsymbol{\theta'_{xy}}}=\boldsymbol{\theta_{xy}}$ yields the mean squared error in (13). The notation below omits the subscripts, e.g, $\boldsymbol{\theta}=\boldsymbol{\theta_{xy}}$.

$$MSE(\widehat{\boldsymbol{\theta}'}) = E\left[(\boldsymbol{\theta}-\boldsymbol{\theta}')^2\right] = E\left[(\boldsymbol{\theta}-E[\boldsymbol{\theta}]+E[\boldsymbol{\theta}]-\boldsymbol{\theta}')^2\right]$$
$$= E\Big[\big(\boldsymbol{\theta}-E[\boldsymbol{\theta}]\big)^2 + 2\big(\boldsymbol{\theta}-E[\boldsymbol{\theta}]\big)\big(E[\boldsymbol{\theta}]-\boldsymbol{\theta}'\big) + \big(E[\boldsymbol{\theta}]-\boldsymbol{\theta}'\big)^2\Big]$$
$$= E\Big[\big(\boldsymbol{\theta}-E[\boldsymbol{\theta}]\big)^2\Big] - 2E\Big[\big(\boldsymbol{\theta}-E[\boldsymbol{\theta}]\big)\big(\boldsymbol{\theta}'-E[\boldsymbol{\theta}']\big)\Big] + E\Big[\big(\boldsymbol{\theta}'-E[\boldsymbol{\theta}']\big)^2\Big]$$
$$= V(\boldsymbol{\theta}) - 2Cov(\boldsymbol{\theta},\boldsymbol{\theta}') + V(\boldsymbol{\theta}')$$
$$Cov(\boldsymbol{\theta},\boldsymbol{\theta}')=0 \text{ since } \boldsymbol{S}\cap\boldsymbol{S}'=\varnothing \text{ and } \boldsymbol{\theta},\boldsymbol{\theta}' \text{ i.i.d., thus}$$
$$MSE(\widehat{\boldsymbol{\theta'_{xy}}}) = V(\boldsymbol{\theta_{xy}}) + V(\boldsymbol{\theta'_{xy}}) \tag{13}$$

Hence the Sample-to-Sample method considers that the estimator $\widehat{\boldsymbol{\theta'_{xy}}} = \boldsymbol{\theta_{xy}}$ is approximately distributed as (14).

$$\widehat{\boldsymbol{\theta'_{xy}}} \sim N\Big(\theta_{xy}, V(\boldsymbol{\theta_{xy}}) + V(\boldsymbol{\theta'_{xy}})\Big) \tag{14}$$

B. Evaluation of Error Rate Estimator

We evaluate the Sample-to-Sample estimates in (14) by simulating binary datasets and drawing confidence intervals for $\widehat{\boldsymbol{\theta'_{01}}}$. We focus on a single class C_0 and ignore C_1, i.e., we simulate only n_{0y} and n'_{0y}. We draw 68% rather than 95% confidence level for a better verification of over-estimated intervals (e.g., coverage may be slightly higher than 95% but significantly higher than 68%). To estimate $V(\boldsymbol{\theta'_{01}})$ we use the known $n'_{0.}$ and apply (12) as (15). Further evaluations address realistic cases where $n'_{x.}$ is unknown (Subsections C and D).

$$V\left(\widehat{\boldsymbol{\theta'_{01}}}\right) = \frac{\theta_{01}(1-\theta_{01})}{n_{0.}} + \frac{\theta_{01}(1-\theta_{01})}{n'_{0.}} \tag{15}$$

We sample 100 test sets of sizes $n_{0.}\in\{20,...,50\,000\}$ randomly drawn from an infinite population with $\theta^*_{01}\in\{0.01,0.5\}$. For each test set, we measured θ_{01} and use (14-15) to draw confidence intervals for $\widehat{\theta'_{01}}$ in target sets of sizes $n'_{0.}\in\{20,...,50\,000\}$. For each interval, we randomly sample 100 target sets with the same population rate θ^*_{01}.

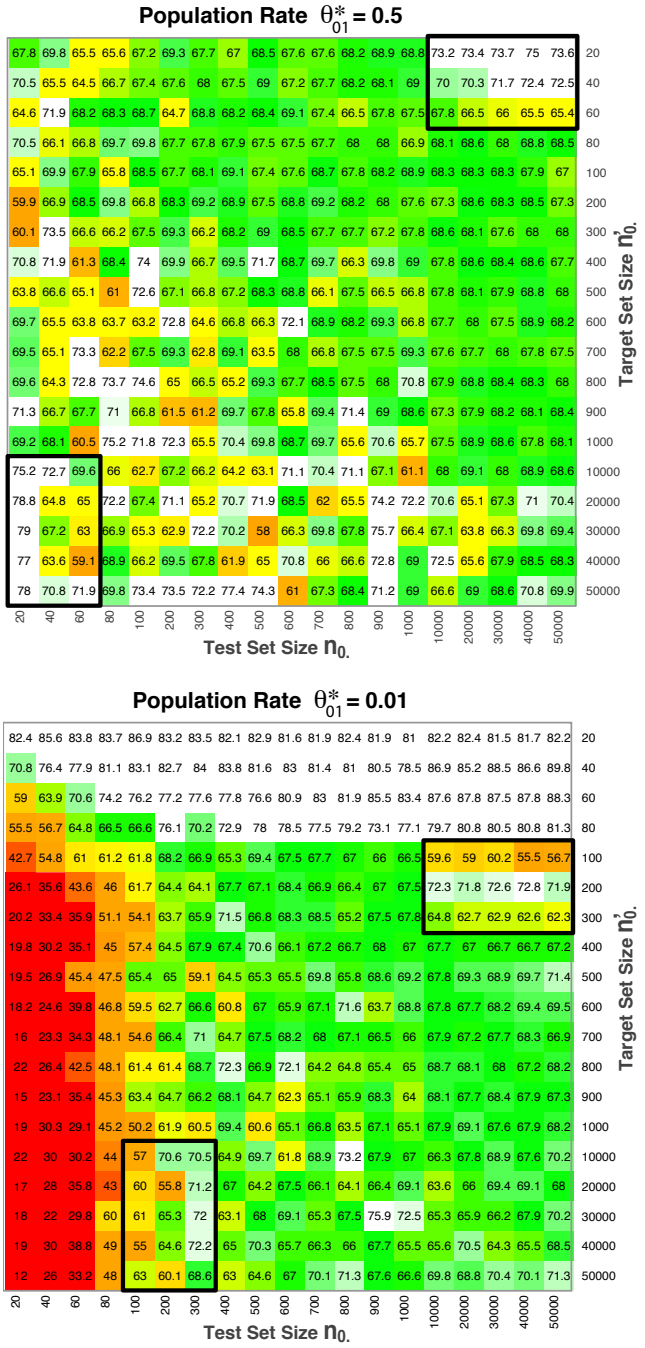

Fig. 5. Evaluation of Sample-to-Sample using known $n'_{x.}$ to derive $V(\widehat{\theta'_{xy}})$ and draw 68% confidence intervals for $\widehat{\theta'_{xy}}$. The cells show the % of intervals containing true θ'_{01} for a total of 10^4 tests. Green cells have correct coverages ≈68%, red indicates too small coverages, white indicates too large coverages.

The graph cells in Fig. 5 show the percentage of θ'_{xy} contained in confidence intervals derived using the Sample-to-Sample method. The confidence intervals achieve the desired confidence level, except when sample sizes $n_{x.}$ and $n'_{x.}$ are too small w.r.t. error rates θ^*_{xy} (in bottom graph only, e.g., $n_{xy}\approx 1$ item, same as the biases observed in Fig. 2), or w.r.t. each other ($n_{x.} \ll$ or $\gg n'_{x.}$, black contours). The interval

coverage varies more if $n_{x.}$<$n'_{x.}$ (lower left triangle of graphs) but mean coverage is correct (e.g., for θ^*_{xy}=0.5, in lower left triangle μ=68.1% and σ=4, otherwise μ=68.3% and σ=1.5).

C. Application to Class Size Estimates

We evaluate the Sample-to-Sample method applied to estimating confidence intervals for the target class sizes $\widehat{n'_{x.}}$ estimated with the misclassification method in binary problems. As in Section IV-B, we simulate 100 test sets and 100 target sets for each test set, with sizes $n_{x.}, n'_{x.} \in \{300, 500, 1000, 2000\}$, drawn from populations with θ^*_{xy} specified in (16).

$$\begin{pmatrix}\theta^*_{00} & \theta^*_{10}\\ \theta^*_{01} & \theta^*_{11}\end{pmatrix} \in \left\{ \begin{pmatrix}.9 & 0\\ .1 & 1\end{pmatrix}, \begin{pmatrix}.9 & .1\\ .1 & .9\end{pmatrix}, \begin{pmatrix}.9 & .2\\ .1 & .8\end{pmatrix}, \begin{pmatrix}.8 & .2\\ .2 & .8\end{pmatrix} \right\} \quad (16)$$

Confidence intervals are estimated using Fieller's theorem, as in [3]. We express the results of the misclassification method as ratios in (17), assuming $1-\widehat{\theta'_{01}}-\widehat{\theta'_{10}} \neq 0$. Fieller's theorem applies to ratios of correlated random variables A/B, e.g., $A = n'_{.0} - \widehat{\boldsymbol{\theta'_{10}}}\, n'_{..}$ and $B = 1 - \widehat{\boldsymbol{\theta'_{01}}} - \widehat{\boldsymbol{\theta'_{10}}}$. The variance and covariance of A and B are detailed in Appendix. We use estimator $\widehat{\theta'_{xy}}$=θ_{xy} with variance (18) derived using the Sample-to-Sample method, and the results of the misclassification method $\widehat{n'_{x.}}$ as estimates of the unknown $n'_{x.}$.

$$\widehat{n'_{0.}} = \frac{n'_{.0} - \widehat{\theta'_{10}}\, n'_{..}}{1 - \widehat{\theta'_{01}} - \widehat{\theta'_{10}}} \qquad \widehat{n'_{1.}} = \frac{n'_{.1} - \widehat{\theta'_{01}}\, n'_{..}}{1 - \widehat{\theta'_{01}} - \widehat{\theta'_{10}}} \quad (17)$$

$$\widehat{V}\left(\widehat{\boldsymbol{\theta'_{xy}}}\right) = \frac{\theta_{xy}(1-\theta_{xy})}{n_{x.}} + \frac{\theta_{xy}(1-\theta_{xy})}{\widehat{n'_{x.}}} \quad (18)$$

The results in Fig. 6 show that the Sample-to-Sample method provides accurate confidence intervals for $\widehat{n'_{x.}}$. For each model in (16) respectively, the mean and variance of intervals' coverage are respectively: μ=68.1% σ=0.7, μ=68.2% σ=0.7, μ=68.2% σ=0.7, μ=68.2% σ=0.7.

These results are obtained without rounding the estimated $\widehat{n'_{x.}}$ nor the confidence limits. If these are rounded, the intervals are slightly biased and over-estimated, e.g., in our experiments the coverage aproximatively varied by ±3% for 68% intervals with μ=69.1%, and ±1% for 95% intervals with μ=95.6%.

D. Application to Error Decomposition

We evaluate the Sample-to-Sample method applied to estimating confidence intervals for the results $\widehat{n'_{xy}}$ of the extended misclassification method (Section III-A). As in Section IV-C, Fieller's theorem is applied with the same experimental setup, to derive confidence intervals for $\widehat{n'_{01}}$ instead of $\widehat{n'_{0.}}$. In this case A=$\widehat{\boldsymbol{\theta'_{01}}}\left(n'_{.0} - \widehat{\boldsymbol{\theta'_{10}}}\, n'_{..}\right)$ (5), (17). The variance and covariance of A and B are detailed in the Appendix.

Instead of drawing a graph as Fig. 6, we report the mean and variance of interval coverage for each model in (16), respectively: μ=68.0% σ=0.7, μ=68.1% σ=0.8, μ=68.2% σ=0.7, μ=68.3% σ=0.7. It shows that the Sample-to-Sample method provides accurate confidence intervals for $\widehat{n'_{xy}}$.

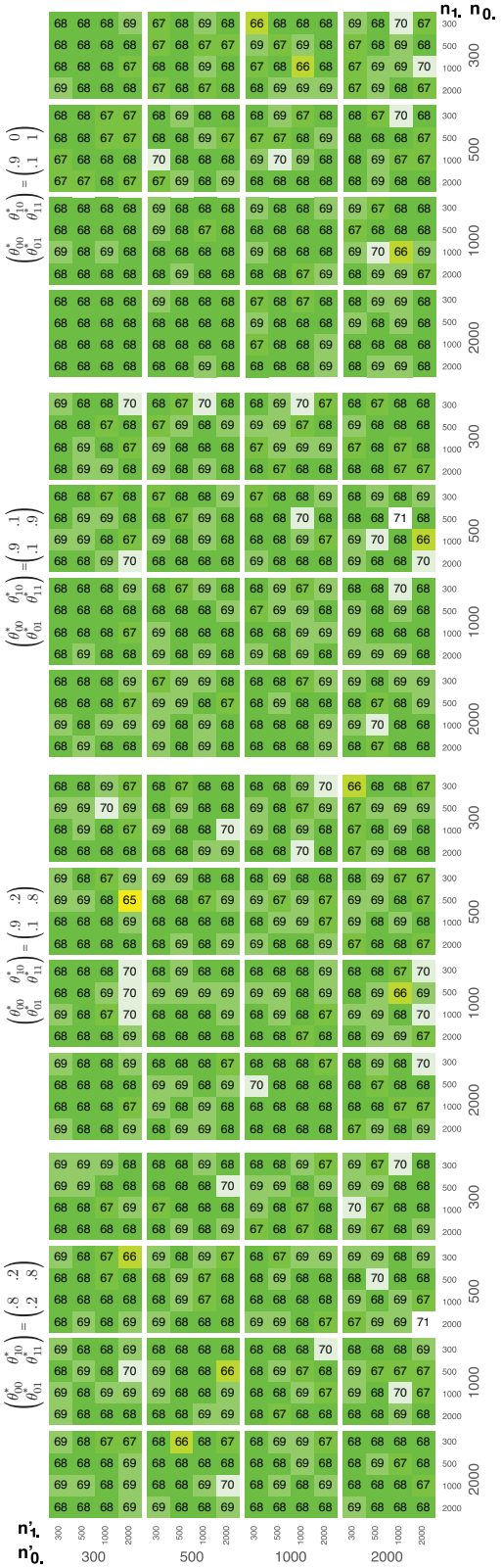

Fig. 6. Results of Sample-to-Sample applied to estimating confidence intervals for $\widehat{n'_{x.}}$. Test and target datasets are randomly sampled with sizes on columns and rows. The cells show the % of intervals that contained $n'_{0.}$ for a total of 10^4 tests (the % are rounded for clarity).

E. Discussion

1) Prior Work: Fieller's theorem is also used in [3], [4] but focuses on class proportions $\pi'_x{=}n'_{x.}/n'_{..}$. The prior method is restated for class C_0 in (19), using our notation. The main difference with our approach is how the test and target sizes $n_{x.}$, $n'_{x.}$ are considered for variance estimation. The Sample-to-Sample approach accounts for test and target sizes $n_{x.}$, $n'_{x.}$ in the estimation of error rate variance $V\big(\widehat{\theta'_{xy}}\big)$. In [3], [4] the error rate variance $V\big(\widehat{\theta'_{xy}}\big)$ considers only the test sizes $n_{x.}$. The target sizes $n'_{x.}$ are considered only for estimating the variance of class proportions $n'_{.y}/n'_{..}$ in the initial classifier output prior to applying bias correction (20).

$$\widehat{\pi_0} = \frac{\widehat{n'_{0.}}}{n'_{..}} = \frac{n'_{.0}/n'_{..} - \theta_{10}}{1 - \theta_{01} - \theta_{10}} \tag{19}$$

$$V\big(n'_{.0}/n'_{..} - \theta_{10}\big) = \frac{n'_{.0}/n'_{..}\big(1 - n'_{.0}/n'_{..}\big)}{n'_{..}} + \frac{\theta_{10}(1-\theta_{10})}{n_{1.}} \tag{20}$$

The results of [3], [4] are applied in Fig. 7 using variables similar to the evaluation in [3]: $n_{x.}{\in}\{25, 50, 125, 250\}$, $n'_{x.}{\in}\{50, 125, 250, 500\}$, θ_{01}=0.1, θ_{10}=0.2. The result are biased for some values of $n_{x.}$ and $n'_{x.}$ because the method originally aims at estimating class proportions in the population $\widehat{\pi^*_x}{=}n'_{x.}/n'_{..}$. This prior work is not applicable for estimating the class sizes or proportions of the target set itself.

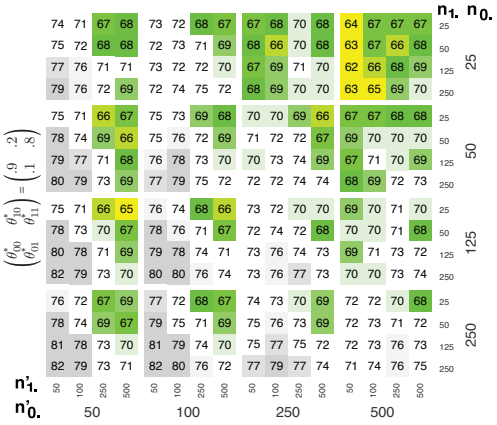

Fig. 7. Confidence intervals drawn using prior work in [3], [4]. Test and target datasets are randomly sampled with sizes on columns and rows. The cells show the % of intervals that contained $\pi_0{=}n'_{0.}/n'_{..}$ for a total of 10^4 tests per cell (the % are rounded for clarity).

The bias in Fig. 7 is corrected by using the Sample-to-Sample method, and considering no variance for the initial class proportion $n'_{.y}/n'_{..}$ (21). The corrected results in Fig. 8 have a small bias, which is explained by the small sample sizes $n_{x.}$, $n'_{x.}$. Estimates drawn using the larger sample sizes in Fig. 6 are unbiased with mean coverage μ=68.2%, σ=0.7. Hence the Sample-to-Sample method is also suitable for estimating the class proportions $\pi'_x{=}n'_{x.}/n'_{..}$.

$$\widehat{V}_{corrected}\big(n'_{.0}/n'_{..} - \theta_{10}\big) = \frac{\theta_{10}(1-\theta_{10})}{n_{1.}} + \frac{\theta_{10}(1-\theta_{10})}{\widehat{n'_{1.}}} \tag{21}$$

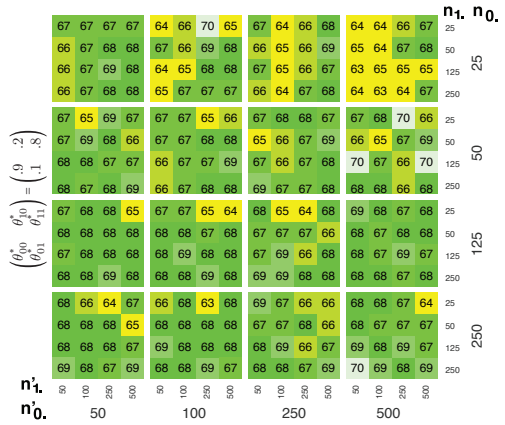

Fig. 8. Results of Sample-to-Sample method used to correct the bias in Fig. 7.

2) Multiclass Problems: Fieller's theorem does not apply to multiclass problems, which are not easily solved as fractions of random variable using Cramer's rule as in (17). Sarrus rule applies to 3-class problems, but the resulting ratio under Cramer's rule remains complex. Thus bootstrapping methods are recommended for multiclass problems [4]. Monte Carlo simulations are also of interest. Datasets can be simulated using error rates θ_{xy} with the Sample-to-Sample method, i.e., with variance (18). Future work should investigate such simulation, and compare its results to bootstrapping methods.

V. Maximum Determinant Method

Prior to applying the method, several classifiers may be available, with no knowledge of the potential target sets. To choose an optimal classifier, the Maximum Determinant method aims at predicting which classifier yields the smallest variance when applying the misclassification method. The approach is agnostic of the potential target sets, hence providing *a priori* results.

The method focuses on the determinant of the error rate matrix, i.e., $|\mathbf{M}_\theta| = \begin{vmatrix} \theta_{11} & \theta_{21} & \dots \\ \theta_{12} & \theta_{22} & \dots \\ \dots & \dots & \dots \end{vmatrix}$ for the misclassification method, or $|\mathbf{M_r}| = \begin{vmatrix} 1 & r_{21} & \dots \\ r_{12} & 1 & \dots \\ \dots & \dots & \dots \end{vmatrix}$ for the Ratio-to-TP method. According to Cramer's rule, the results of the misclassification and Ratio-to-TP methods are fractions of two matrix determinants [12]. The fraction's denominator is the determinant of the error rate matrix $|\mathbf{M}_\theta|$ or $|\mathbf{M_r}|$. If the determinant $|\mathbf{M}|{\to}0$ then $\widehat{n'_{x.}}{\to}\infty$. For a small determinant $|\mathbf{M}|{\to}0$, a variation $|\mathbf{M}^+|{=}|\mathbf{M}|{+}\delta$ can yield a high variation in $\widehat{n'_{x.}}$ as $\widehat{n'_{x.}}{\to}\infty$. For a larger determinant $|\mathbf{M}|{\gg}0$, the same variation $|\mathbf{M}^+|{=}|\mathbf{M}|{+}\delta$ yields a smaller variation in $\widehat{n'_{x.}}$. Hence the Maximum Determinant method assumes that the larger the difference $|\mathbf{M}| - 0$ the smaller the variance $V(\widehat{n'_{x.}})$.

An initial evaluation is provided in Fig. 9 and Table III, using the same datasets as Section II-C. To sample several target sets for the same test set, we use smaller sample sizes (i.e., in Table III, $n_{x.}{+}n'_{x.}{<}n^*_{x.}$ where $n^*_{x.}$ is the total number of items available for class C_x). We sample 1000 test sets and measure their matrix determinants $|\mathbf{M}_\theta|$ and $|\mathbf{M_r}|$. For each test set, we sample 100 distinct target sets and compute the variance $V(\widehat{n'_{x.}})$ over the target sets.

We find that $V(\widehat{n'_{x.}})$ seems to be a linear function of $|\mathbf{M}|$ (Fig. 9). The negative correlation between $|\mathbf{M}_\theta|$ and $\sum_x V(\widehat{n'_{x.}})$ or $\sum_x \sum_y V(\widehat{n'_{xy}})$ (Table III) supports the Maximum Determinant assumption that high determinants indicate lower variance $V(\widehat{n'_{x.}})$ and $V(\widehat{n'_{xy}})$. The correlation is significant for the multiclass datasets, and less significant but consistent for the binary datasets (i.e., negative or null). Hence the method may not be relevant for some binary problems.

TABLE III
RESULTS OF MAXIMUM DETERMINANT METHOD

				Correlation $\|\mathbf{M}_\theta\|$ and $\sum$Var		Correlation $\|\mathbf{M_r}\|$ and $\sum$Var	
	Data	**Test Set $n_{x.}$**	**Target Set $n'_{x.}$**	**V($\widehat{n'_{x.}}$)**	**V($\widehat{n'_{xy}}$)**	**V($\widehat{n'_{x.}}$)**	**V($\widehat{n'_{xy}}$)**
Tests T1	Iris	C_{1-2}:20 C_3:15	C_{1-3}:25	-0.81	-0.79	**-0.91**	**-0.89**
	Iono.	C_1:50 C_0:50	C_1:50 C_0:100	**-0.35**	**-0.13**	-0.21	-0.01
	Segm.	C_{1-7}:100	$C_{1,3,5,7}$:100 $C_{2,4,6}$:200	**-0.83**	**-0.81**	-0.79	-0.76
	Ohsc.	C_{0-9}:400	C_{0-4}:100 C_{5-10}:200	-0.72	-0.52	**-0.75**	**-0.64**
	Wave.	C_{1-3}:300	C_1:300 C_2:600 C_3:900	**-0.53**	**-0.40**	-0.16	-0.08
	Chess	C_1:300 C_0:500	C_1:1000 C_0:500	-0.01	0.08	0	0.08
Tests T2	Iris	C_{1-2}:10 C_3:15	C_{1-3}:25	-0.79	-0.77	**-0.89**	**-0.87**
	Iono.	C_1:30 C_0:30	C_1:50 C_0:100	**-0.36**	**-0.12**	-0.23	0.01
	Segm.	C_{1-7}:50	$C_{1,3,5,7}$:100 $C_{2,4,6}$:200	**-0.83**	**-0.81**	-0.78	-0.75
	Ohsc.	C_{0-9}:200	C_{0-4}:100 C_{5-10}:200	-0.71	-0.53	**-0.75**	**-0.65**
	Wave.	C_{1-3}:200	C_1:300 C_2:600 C_3:900	**-0.49**	**-0.35**	-0.18	-0.10
	Chess	C_1:200 C_0:300	C_1:1000 C_0:500	-0.01	0.08	0	0.09
Tests T3	Iris	C_{1-3}:25	C_{1-2}:10 C_3:15	-0.24	-0.24	**-0.35**	**-0.34**
	Iono.	C_1:50 C_0:100	C_1:30 C_0:30	**-0.80**	**-0.64**	-0.75	-0.58
	Segm.	$C_{1,3,5,7}$:100 $C_{2,4,6}$:200	C_{1-7}:50	-0.72	-0.71	**-0.77**	**-0.74**
	Ohsc.	C_{0-4}:100 C_{5-10}:200	C_{0-9}:200	-0.68	-0.49	**-0.72**	**-0.59**
	Wave.	C_1:300 C_2:600 C_3:900	C_{1-3}:200	**-0.61**	**-0.46**	-0.16	-0.08
	Chess	C_1:1000 C_0:500	C_1:200 C_0:300	-0.33	-0.16	**-0.34**	**-0.17**

The initial results are promising, but future work is required for establishing theory (e.g., parameters of function $f(|\mathbf{M}|)$=$V(\widehat{n'_{x.}})$; binary problems for which the method is not relevant; in which cases $|\mathbf{M}_\theta|$ or $|\mathbf{M_r}|$ is a better predictor). Future work should also investigate the potential of resampling the test set (e.g., do smaller test sets with a higher matrix determinant perform better than larger test sets with a lower determinant?), and the consistency with Sample-to-Sample estimates (e.g., do higher $|\mathbf{M}|$ have smaller Sample-to-Sample variance estimates?).

VI. DISCUSSION AND FUTURE WORK

The choice of a bias correction method, and a variance estimation method, depends on the characteristics of both test and target sets. If class proportions differ between test and target set, the reclassification method is not applicable. If the test set is not sampled from the target set, and the bias correction method is applied to describe the target set (rather than a general population from which the target set is sampled), then prior variance estimation methods are not applicable and the Sample-to-Sample method must be used.

Applications of bias correction methods face issues with potentially high result variance. Random error rate variations between test and target sets can worsen the initial classification bias when applying bias correction methods. This issue is addressed in [3] with a method balancing the uncorrected classifier output $n'_{.y}$ and estimated $\widehat{n'_{x.}}$ in a linear combination, e.g., fitting the α parameter in (22).

$$\widehat{n'_{x.,combined}} = \alpha\, \widehat{n'_{x.}} + (1-\alpha)\, n'_{.y} \qquad (22)$$

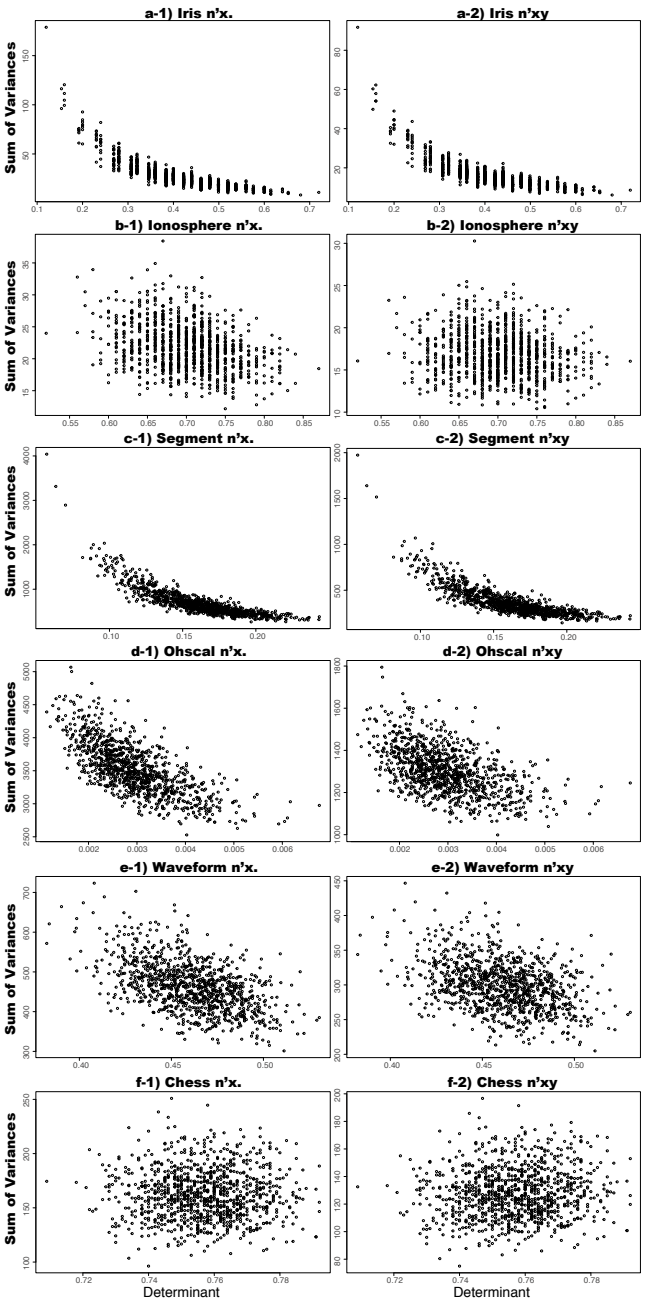

Fig. 9. Results of Maximum Determinant method, applied using $|\mathbf{M}_\theta|$ (misclassification method) and the datasets of test T1 in Table III. Each dot represents a test set for which 10^2 target sets are sampled. The x-axis shows $|\mathbf{M}_\theta|$, and the y-axis $\sum_x V(\widehat{n'_{x.}})$ (left graphs) and $\sum_x \sum_y V(\widehat{n'_{xy}})$ (right graphs). The summation may explain the exponentiality in graph -a) -c).

A. Small Datasets

High variance is particularly critical for small datasets, e.g., if $n_{x.}, n'_{x.}, n_{xy}$ or n'_{xy} are less than a few items. Further research is needed to identify the data sizes for which bias correction methods are not recommended, or linear combinations (22) are preferable (e.g., depending on error rate magnitudes). Cases where small n_{xy} yield error rate $\theta_{xy} \to 0$ or 1 should also be investigated (e.g., higher error rates may be preferable).

B. Negative Estimates of $n'_{x.}$ and n'_{xy}

The misclassification method can yield negative estimates $\widehat{n'_{x.}} < 0$ (although it rarely happened in our experiments). Negative estimates are easily handled for binary problems, i.e., if $\widehat{n'_{0.}} < 0$, set $\widehat{n'_{1.}}$ to $\widehat{n'_{1.}} + \widehat{n'_{0.}}$, and $\widehat{n'_{0.}}$ to 0. Future research is required to handle negative estimates in multiclass problems.

C. Applicability of Fieller's theorem

Fieller's theorem is not applicable when its denominator is null, i.e., $\theta_{01} + \theta_{10} = 1$. Such impractical cases occur with random classifiers (i.e., $\theta_{01} = \theta_{10} = 0.5$), or with classifiers performing worse than random for one class and inversely proportional for the other class (e.g., $\theta_{01} = 0.8$ and $\theta_{10} = 0.2$).

D. Varying Feature Distributions

Classifiers typically use feature distributions to build models of each class (e.g., describing the characteristics of the classes). If the feature distributions differ between test and target sets, the error rates may differ too (e.g., if a target set has more low-contrast images, more images may be misclassified). This may worsen the classification biases when applying bias correction methods. Fig. 10 shows examples where a single feature is used, a score as in Fig. 2. Small variations of the feature distribution have created significant biases.

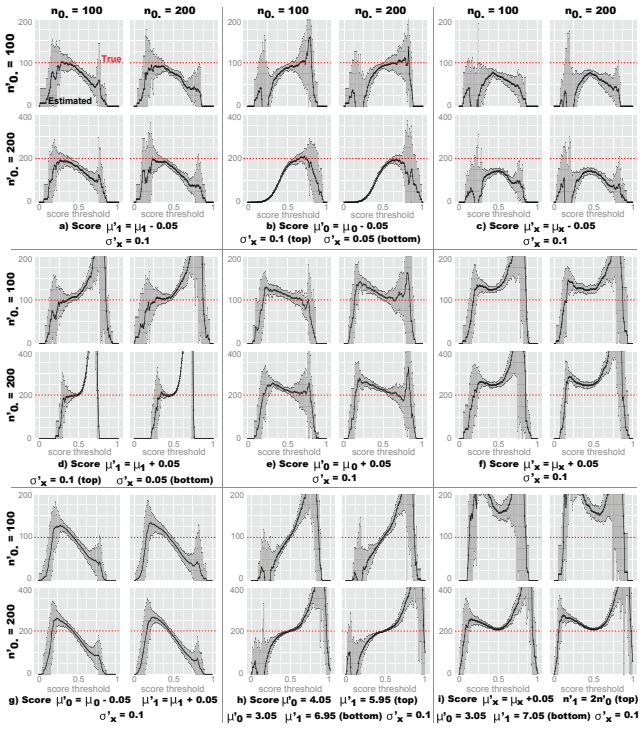

Fig. 10. Results of misclassification method for simulated data with varying feature distribution. As in Fig. 2, a score threshold (x axis) is used to assign classes C_0 or C_1. Class sizes $\widehat{n'_{0.}}$ (y axis) are estimated for 10^4 pairs of test and target sets. Test sets are randomly sampled with class proportions $n_{0.}$=$n_{1.}$, mean scores μ_0=0.4 for C_0, μ_1=0.6 for C_1, and score variance σ_x=0.1. Target sets are sampled from score distributions that differ from the test sets, with μ'_x=$\mu_x \pm 0.5$ and variance $\sigma'_x \in \{0.05, 0.1\}$, and with class proportions $n'_{0.}$=$2n'_{1.}$. Additional variations are illustrated in graphs -h) -i).

Further research is required to handle varying feature distributions. Prior work in [14] addresses this issue with a logistic regression fitted on feature distributions, but requires equal class proportions between test and target sets, and items with a single feature per class (e.g., a dimension reduction, or a measure of the similarity with class models). For binary problems using tuning parameters, Fig. 10 suggests that parameters yielding equal error rates $\theta_{01} = \theta_{10}$ (e.g., thresholds averaging the mean scores $(\mu_0 + \mu_1)/2$ in our simulations) may minimise the biases due to varying feature distributions (and the variance in any case, as suggested in Fig. 2).

VII. Conclusion

We demonstrated the characteristics of bias correction methods regarding their variance magnitude, and their applicability if class proportions or feature distributions differ between test and target sets. It informs the choice of a method depending on the use case at hand.

We investigated methods for estimating the error composition in a classifier output. Given the $n'_{.y}$ items classified as class C_y, they estimate how many n'_{xy} items truly belong to class C_x. Such methods describe the quality of classification data beyond accuracy or precision.

We introduced a novel variance estimation method, called Sample-to-Sample. It applies to specific target sets rather than general populations from which target sets are sampled, the latter being addressed in prior work. It provides accurate confidence intervals for class sizes $\widehat{n'_{x.}}$ estimated from bias correction methods, and for error composition estimates $\widehat{n'_{xy}}$.

Finally, we introduced a promising method for predicting the variance of bias correction results without prior knowledge of the potential target sets. It shows a correlation between the determinant of error rate matrices and the variance of two bias correction methods. Such method can inform the choice of a classifier, or its test set if several of them are available, before applying a bias correction method.

The methods we introduced support uncertainty-aware analyses of classification data, e.g., to investigate class sizes and distributions.

Acknowledgement

We are grateful to Arjen P. de Vries, Nishant Mehta, Erik Quaeghebeur and Rebecca Holman for their invaluable remarks and suggestions. Part of this research was funded by the Fish4Knowledge project (EU FP7 Grant 257024).

Appendix

A. Code

The R code used to apply and evaluate the methods described in this paper is available online, free of use: *https://github.com/emma-cwi/classification_error*

B. Application of Fieller's theorem

Fieller's theorem [15] defines the confidence intervals limits $[\ell^-, \ell^+]$ for a ratio of correlated random variables A/B as (23), with z=1 for 68% confidence level.

$$\ell^{\pm} = \frac{(\mu_A\mu_B - z^2\sigma_{A,B}) \pm \sqrt{(\mu_A\mu_B - z^2\sigma_{A,B})^2 - (\mu_A^2 - z^2\sigma_A^2)(\mu_B^2 - z^2\sigma_B^2)}}{\mu_B^2 - z^2\sigma_B^2} \tag{23}$$

1) Section IV-C: For $\widehat{n'_{0.}}$, $A = n'_{.0} - \widehat{\boldsymbol{\theta'_{10}}}(n'_{.1} + n'_{.0})$ and $B = 1 - \widehat{\boldsymbol{\theta'_{01}}} - \widehat{\boldsymbol{\theta'_{10}}}$ (17). The mean, variance, covariance of A, B are detailed below, knowing that $\widehat{\boldsymbol{\theta'_{01}}}$, $\widehat{\boldsymbol{\theta'_{10}}}$ are independent with null covariance.

$$\mu_B = E\left[1 - \widehat{\boldsymbol{\theta'_{01}}} - \widehat{\boldsymbol{\theta'_{10}}}\right] \qquad \widehat{\mu_B} = 1 - \theta_{01} - \theta_{10}$$

$$\sigma_B^2 = V\left(1 - \widehat{\boldsymbol{\theta'_{01}}} - \widehat{\boldsymbol{\theta'_{10}}}\right) = V\left(\widehat{\boldsymbol{\theta'_{01}}}\right) + V\left(\widehat{\boldsymbol{\theta'_{10}}}\right)$$
$$\widehat{\sigma_B^2} = \frac{\theta_{01}(1-\theta_{01})}{n_{0.}} + \frac{\theta_{01}(1-\theta_{01})}{\widehat{n'_{0.}}} + \frac{\theta_{10}(1-\theta_{10})}{n_{1.}} + \frac{\theta_{10}(1-\theta_{10})}{\widehat{n'_{1.}}}$$

$$\mu_A = E\left[n'_{.0} - \widehat{\boldsymbol{\theta'_{10}}}\, n'_{..}\right] \qquad \widehat{\mu_A} = n'_{.0} - \theta_{10}\, n'_{..}$$
$$\sigma_A^2 = V\left(n'_{.0} - \widehat{\boldsymbol{\theta'_{10}}}\, n'_{..}\right) = {n'_{..}}^2\, V\left(\widehat{\boldsymbol{\theta'_{10}}}\right)$$
$$\widehat{\sigma_A^2} = {n'_{..}}^2\left(\frac{\theta_{10}(1-\theta_{10})}{n_{1.}} + \frac{\theta_{10}(1-\theta_{10})}{\widehat{n'_{1.}}}\right)$$

$$\sigma_{A,B} = Cov\left(n'_{.0} - \widehat{\boldsymbol{\theta'_{10}}}\, n'_{..},\ 1 - \widehat{\boldsymbol{\theta'_{01}}} - \widehat{\boldsymbol{\theta'_{10}}}\right) = n'_{..} V\left(\widehat{\boldsymbol{\theta'_{10}}}\right)$$

$$\widehat{\sigma_{A,B}} = n'_{..}\left(\frac{\theta_{10}(1-\theta_{10})}{n_{1.}} + \frac{\theta_{10}(1-\theta_{10})}{\widehat{n'_{1.}}}\right)$$

2) Section IV-D: For $\widehat{n'_{01}}$, $A = \widehat{\boldsymbol{\theta'_{01}}}\left(n'_{.0} - \widehat{\boldsymbol{\theta'_{10}}}\, n'_{..}\right)$. B remains unchanged. Their mean, variance, covariance are detailed below, using [16].

$$\mu_A = E\left[\widehat{\boldsymbol{\theta'_{01}}}\left(n'_{.0} - \widehat{\boldsymbol{\theta'_{10}}}\, n'_{..}\right)\right] \quad \widehat{\mu_A} = \theta_{01}\left(n'_{.0} - \theta_{10}\, n'_{..}\right)$$
$$\sigma_A^2 = E\left[\widehat{\boldsymbol{\theta'_{01}}}\right]^2 V\left(n'_{.0} - \widehat{\boldsymbol{\theta'_{10}}}\, n'_{..}\right) + E\left[n'_{.0} - \widehat{\boldsymbol{\theta'_{10}}}\, n'_{..}\right]^2 V\left(\widehat{\boldsymbol{\theta'_{01}}}\right) + V\left(\widehat{\boldsymbol{\theta'_{01}}}\right) V\left(n'_{.0} - \widehat{\boldsymbol{\theta'_{10}}}\, n'_{..}\right)$$
$$\widehat{\sigma_A^2} = \theta_{01}^2 {n'_{..}}^2\, \widehat{V}\left(\widehat{\boldsymbol{\theta'_{10}}}\right) + \left(n'_{.0} - \theta_{10}\, n'_{..}\right)^2 \widehat{V}\left(\widehat{\boldsymbol{\theta'_{01}}}\right) + {n'_{..}}^2 \widehat{V}\left(\widehat{\boldsymbol{\theta'_{01}}}\right)\widehat{V}\left(\widehat{\boldsymbol{\theta'_{10}}}\right)$$

$$\sigma_{A,B} = n'_{..}\left(Cov\left(\widehat{\boldsymbol{\theta'_{01}}}\widehat{\boldsymbol{\theta'_{10}}}, \widehat{\boldsymbol{\theta'_{01}}}\right) + Cov\left(\widehat{\boldsymbol{\theta'_{01}}}\widehat{\boldsymbol{\theta'_{10}}}, \widehat{\boldsymbol{\theta'_{10}}}\right)\right) - n'_{.0} V\left(\widehat{\boldsymbol{\theta'_{10}}}\right)$$

$$Cov\left(\widehat{\boldsymbol{\theta'_{xy}}}\widehat{\boldsymbol{\theta'_{yx}}}, \widehat{\boldsymbol{\theta'_{xy}}}\right) = E\left[\widehat{\boldsymbol{\theta'_{xy}}}\right] Cov\left(\widehat{\boldsymbol{\theta'_{yx}}}, \widehat{\boldsymbol{\theta'_{xy}}}\right) + E\left[\widehat{\boldsymbol{\theta'_{yx}}}\right] Cov\left(\widehat{\boldsymbol{\theta'_{xy}}}, \widehat{\boldsymbol{\theta'_{xy}}}\right) = E\left[\widehat{\boldsymbol{\theta'_{yx}}}\right] V\left(\widehat{\boldsymbol{\theta'_{xy}}}\right)$$

$$\widehat{\sigma_{A,B}} = n'_{..}\left(\theta_{10}\widehat{V}\left(\widehat{\boldsymbol{\theta'_{01}}}\right) + \theta_{01}\widehat{V}\left(\widehat{\boldsymbol{\theta'_{10}}}\right)\right) - n'_{.0}\widehat{V}\left(\widehat{\boldsymbol{\theta'_{01}}}\right)$$
With $\widehat{V}\left(\widehat{\boldsymbol{\theta'_{xy}}}\right) = \frac{\theta_{xy}(1-\theta_{xy})}{n_{x.}} + \frac{\theta_{xy}(1-\theta_{xy})}{\widehat{n'_{x.}}}$ (Sample-to-Sample)

3) Section IV-E: For $\widehat{\pi_0} = \widehat{n'_{0.}}/n'_{..}$ in [3], [4], $A = n'_{.0}/n'_{..} - \widehat{\boldsymbol{\theta'_{10}}}$ (19). B remains unchanged. The mean, variance, covariance used in [3], [4] are restated below.

$$\mu_A = n'_{.0}/n'_{..} - \theta_{10}$$
$$\sigma_A^2 = \frac{n'_{.0}/n'_{..}\left(1 - n'_{.0}/n'_{..}\right)}{n'_{..}} + \frac{\theta_{10}(1-\theta_{10})}{n_{1.}}$$
$$\sigma_B^2 = \frac{\theta_{01}(1-\theta_{01})}{n_{0.}} + \frac{\theta_{10}(1-\theta_{10})}{n_{1.}}$$
$$\sigma_{A,B} = \frac{\theta_{10}(1-\theta_{10})}{n_{1.}}$$

REFERENCES

[1] A. Tenenbein, "A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection," *Technometrics*, vol. 14, no. 1, pp. 187–202, 1972.

[2] A. Grassia and R. Sundberg, "Statistical precision in the calibration and use of sorting machines and other classifiers," *Technometrics*, vol. 24, no. 2, pp. 117–121, 1982.

[3] M. S. Shieh, "Correction methods, approximate biases, and inference for misclassified data," Ph.D. dissertation, Univ. of Massachusetts, 2009.

[4] J. P. Buonaccorsi, *Measurement Error: Models, Methods and Applications*. CRC Press, Taylor and Francis, 2010.

[5] D. H. Card, "Using known map category marginal frequencies to improve estimates of thematic map accuracy." *Photogrammetric Engineering and Remote Sensing*, vol. 48, pp. 431–439, 1982.

[6] A. M. Hay, "The derivation of global estimates from a confusion matrix," *International Journal of Remote Sensing*, vol. 9, no. 8, 1988.

[7] P. C. van Deusen, "Unbiased estimates of class proportions from thematic maps," *Photogrammetric Engineering and Remote Sensing*, vol. 62, no. 4, pp. 409–412, 1996.

[8] G. M. Foody, "Status of land cover classification accuracy assessment," *Remote Sensing of Environment*, vol. 80, no. 1, pp. 185–201, 2002.

[9] M. Katila, *Forest Inventory: Methodology and Applications*. Springer, 2006, no. 13, ch. Correcting map errors in forest inventory estimates for small areas, pp. 225–233.

[10] A. M. Hay, "Global estimates from a confusion matrix, a reply to Jupp," *International Journal of Remote Sensing*, vol. 10, no. 9, 1989.

[11] E. Beauxis-Aussalet and L. Hardman, "Multifactorial uncertainty assessment for monitoring population abundance using computer vision," in *IEEE Conference on Data Science and Advanced Analytics (DSAA)*, 2015.

[12] A. Kosinski, "Cramer's rule is due to Cramer," *Mathematics Magazine*, vol. 74, pp. 310–312, 2001.

[13] W. G. Cochran, *Sampling techniques*. John Wiley & Sons, 2007.

[14] B. J. Boom, E. Beauxis-Aussalet, L. Hardman, and R. B. Fisher, "Uncertainty-aware estimation of population abundance using machine learning," *Multimedia System Journal*, vol. 22, no. 6, 2016.

[15] E. C. Fieller, "Some problems in interval estimation," *Journal of the Royal Statistical Society. Series B*, 1954.

[16] G. Bohrnstedt and A. Goldberger, "On the exact covariance of products of random variables." *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1439–1442, 1969.