

UniLeiden at LeQua 2022: The first step in understanding the behaviour of the median sweep quantifier using continuous sweep^{*}

Kevin Kloos^{1,2,*}, Quinten A. Meertens^{2,3} and Julian D. Karch¹

¹Leiden University, Faculty of Social Sciences, Institute of Psychology, Department Methodology and Statistics, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands

²Statistics Netherlands, Henri Faasdreef 312, 2492 JP Den Haag, The Netherlands

³University of Amsterdam, Amsterdam School of Economics, Center for Nonlinear Dynamics in Economics and Finance, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands

Abstract

This paper presents the continuous sweep quantifier, a smoothed adaptation of the median sweep quantifier. Previous research has shown that the median sweep quantifier is a good quantifier. However, it is not well understood why it performs well because it is hard to derive its theoretical properties. The continuous sweep quantifier is a modification of the median sweep quantifier that enables computing theoretical results. The continuous sweep quantifier 1) uses kernel estimates instead of the empirical distribution, 2) constructs decision boundaries instead of applying discrete decision rules, and 3) uses the mean instead of the median. We show that a simplified adaptation of the continuous sweep quantifier performs similarly to the median sweep quantifier in terms of bias and variance on the LeQua 2022 dataset. The continuous sweep quantifier can therefore be used to provide insights into the median sweep quantifier by computing theoretical expressions for bias and variance.

Keywords

quantification learning, learning to quantify, classification, machine learning, median sweep, continuous sweep, LeQua 2022

1. Introduction

Quantification Learning, also known as *learning to quantify* or *quantification*, is a machine learning task with the aim to compute the class prevalences from an unlabeled test set [1]. Quantification used to be seen as a side product of classification: a good classifier should also produce good prevalence estimates. However, Forman has objected against this statement and showed that simply *classifying and counting* the estimated labels from a classifier may lead to a severe bias [2]. Therefore, more advanced techniques are needed.

Over the past decades, specific techniques for quantification learning called *quantifiers* have been developed. Binary quantifiers can be categorized into three groups [1]: the group based


CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

✉ k.kloos@fsw.leidenuniv.nl (K. Kloos); q.a.meertens@uva.nl (Q. A. Meertens); j.d.karch@fsw.leidenuniv.nl (J. D. Karch)

🌐 <https://github.com/kevinkloos> (K. Kloos)

🆔 0000-0001-6980-4259 (K. Kloos); 0000-0002-3485-8895 (Q. A. Meertens); 0000-0002-1625-2822 (J. D. Karch)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

on *Classify, Count and Correct*, the group based on *direct learners*, and the group based on *distribution matching* [3, 4].

Currently, there is no consensus in the academic literature about which group of techniques performs best. According to Vapnik’s principle [5], a problem should be solved directly without solving a more general problem as an intermediate step. Quantification is a more generalized task than classification. Therefore, Vapnik’s principle implies that quantifiers should be created without the intermediate step of constructing a classifier [5, 6]. Schumacher compared quantification techniques empirically using an extensive simulation study [7]. They conclude that some techniques based on *Classify, Count and Correct*, that is, quantifiers that construct a classifier as an intermediate step, performed best. Especially the *median sweep* method from Forman performs well among all popular quantifiers [3]. These two approaches are rather different. An open question is when and why median sweep is such a good quantifier [7].

In this paper, we take the first step in understanding why median sweep is a good quantifier. We propose to perform a theoretical analysis. More specifically, we aim to derive the mean squared error of the median sweep method as a quantifier for the prevalence of the positive class (α) in a binary classification setting. Fortunately, theoretical results of several threshold-based quantifiers have already been derived [8, 9, 10, 11, 12]. We aim to extend these results to median sweep, which is, in fact, an *ensemble* of threshold-based quantifiers. The key challenge in the theoretical analysis is the discrete nature of median sweep.

Therefore, this paper introduces the new *continuous sweep* quantifier. Continuous sweep is a quantifier that is empirically similar to median sweep. It is constructed to have similar empirical performance as median sweep and to allow for easier analytical derivations. Since continuous sweep and median sweep are closely related, we anticipate that thoroughly understanding the theoretical properties of continuous sweep will also provide insight into the properties of median sweep. In this paper, we construct the continuous sweep quantifier, study its empirical performance, and specify a research agenda for the theoretical analysis of this new quantifier.

The remainder of the paper is organized as follows. In Section 2, we will introduce the mathematical notation and reiterate the mathematical expressions for the common quantifiers from the group *Classify, Count and Correct*, including median sweep. Moreover, we will introduce the continuous sweep quantifier and we will show how it is related to the median sweep quantifier. In Section 3, we will evaluate and compare the performance of median sweep and continuous sweep using data of the LeQua2022 Task [6]. In Section 4, we will discuss our new continuous sweep quantifier and provide suggestions for future research.

2. Methods

In this section, we introduce the continuous sweep quantifier and explain the differences between the continuous sweep quantifier and the median sweep quantifier. First, we introduce the notation and reiterate the definition of the median sweep. Second, we present three theoretical difficulties in analyzing the median sweep quantifier and introduce the continuous sweep quantifier.

2.1. Notation and median sweep

Consider a population of observations where each observation consists of a feature vector $x \in \mathcal{X} = \mathbb{R}^p$ and class label $y \in \mathcal{Y} = \{+, -\}$. The feature vector x consists of p (numeric) covariate values. Denote a training set of size n_{train} by D_{train} where the feature vectors are independent and identically distributed (i.i.d.) with density f_{train} . Moreover, we denote a validation set of size n_{val} by D_{val} with corresponding density f_{val} . Last, denote the test set of size n_{test} by D_{test} with density f_{test} .

Importantly, the class label y is only observed in D_{train} and D_{val} . The class label y is unobserved in D_{test} . The aim of quantification in a binary setting is to estimate the proportion of observations with a positive label in D_{test} using the available data and machine learning techniques.

We denote the probability density functions of the feature vector for observations in the positive and negative class by $f^{(+)}(x)$ and $f^{(-)}(x)$, respectively. The probability density functions of the feature vector for the training, validation and test set are each a mixture of $f^{(+)}(x)$ and $f^{(-)}(x)$, but with different mixture parameters α_{train} , α_{val} and α_{test} , respectively. So, we assume $f_{\text{train}}(x) = \alpha_{\text{train}} \cdot f^{(+)}(x) + (1 - \alpha_{\text{train}}) \cdot f^{(-)}(x)$, $f_{\text{val}}(x) = \alpha_{\text{val}} \cdot f^{(+)}(x) + (1 - \alpha_{\text{val}}) \cdot f^{(-)}(x)$, and $f_{\text{test}}(x) = \alpha_{\text{test}} \cdot f^{(+)}(x) + (1 - \alpha_{\text{test}}) \cdot f^{(-)}(x)$. In other words, we assume that the distributions of the positive class in the training, validation, and test set are identical (and we make the same assumption for the negative class). Moreover, we assume that the mixture parameters differ across the data sets. The combination of these assumptions is referred to as prior-probability shift [13].

We consider a soft-classifier $\hat{\delta}$ that maps each feature vector x to an estimate of $P(Y = + | X = x)$. The soft-classifier $\hat{\delta}$ can be obtained from a machine learning algorithm which is trained using the training data D_{train} . Then, we compute probability estimates $\hat{\delta}(x)$ for all feature vectors in the validation set D_{val} . Note that these values can only be interpreted as classification probabilities if the classifier is properly calibrated. Otherwise, we interpret these values as scores. With those scores, we can estimate marginal densities for $\hat{\delta}(x)$ for both classes. We define $\hat{f}^{(i)}$ as the estimated marginal probability density function for $\hat{\delta}(x)$ given that $y = i$. The true positive rate and false positive rate can be computed by integrating $\hat{f}^{(i)}$. Hence, $F^{(+)}(x)$ denotes the true positive rate and $F^{(-)}(x)$ denotes the false positive rate.

Quantifiers of type *Classify*, *Count* and *Correct* use a threshold to make an initial guess of the prevalence. The threshold value is based on the estimated score that an observation in D_{test} has a positive label. Usually, classifiers use a threshold with a score/probability of $\frac{1}{2}$ to classify an observation. Observations with an estimated score larger than or equal to $\frac{1}{2}$ are labeled as positive and observations with an estimated score smaller than $\frac{1}{2}$ are labeled as negative. Other score values could also be chosen as the threshold value. We will define the threshold value by θ , where we assume $\theta \in [0, 1]$ for convenience. Then, observations with an estimated score larger than θ are positively labelled and observations with an estimated score smaller than θ are negatively labelled.

There are several ways to estimate the prevalence of D_{test} using D_{train} and D_{val} , which we will discuss in the next subsections.

Classify-and-count ($\hat{\alpha}_{CC}$)

The most straightforward technique to estimate the prevalence α is by simply counting the number of observations that have a score larger than a certain threshold $\theta \in [0, 1]$ in D_{test} and dividing it by the total number of observations in D_{test} . This technique is more commonly known as the classify-and-count quantifier $\hat{\alpha}_{CC}$. The classify-and-count quantifier $\hat{\alpha}_{CC}$ is not a good quantifier for α , even when the underlying classifier performs well. Good classification performance is not sufficient enough for reliable quantification [1]. The most common threshold for θ is $\frac{1}{2}$, which makes sense for classification but is, in general, suboptimal for quantification. For a biased soft-classifier $\hat{\delta}(x)$ and/or when the prevalences differ across the training, validation and test set, a threshold of $\theta = 0.5$ is suboptimal for quantification. Given the notation from the previous paragraphs, we define the classify-and-count quantifier as

$$\hat{\alpha}_{CC}(D_{\text{test}}, \theta) = \frac{1}{n_{\text{test}}} \sum_{x \in D_{\text{test}}} \mathbb{1}_{\{\hat{\delta}(x) \geq \theta\}}. \quad (1)$$

In the next subsection, we use the classify-and-count quantifier to define the adjusted-count quantifier.

Adjusted count ($\hat{\alpha}_{AC}$)

The adjusted-count quantifier ($\hat{\alpha}_{AC}$) corrects the classify-and-count quantifier $\hat{\alpha}_{CC}$ using estimated classification rates. The adjusted-count quantifier uses the true positive rate and false positive rate for the classifier $\hat{\delta}(x) \geq \theta$ to adjust the classify-and-count estimate. The two classification rates are estimated from the validation set. The classification rates of class i are computed by counting the proportion of observations in D_{val} with a label $y = i$ that have a score $\hat{\delta}(x)$ larger than θ . Then, the classification rates are defined as

$$\hat{F}^{(i)}(D_{\text{val}}, \theta) = \frac{\sum_{(x,y) \in D_{\text{val}} : y=i} \mathbb{1}_{\{\hat{\delta}(x) \geq \theta\}}}{\sum_{y \in D_{\text{val}}} \mathbb{1}_{\{y=i\}}}. \quad (2)$$

The adjusted-count quantifier is then derived as

$$\hat{\alpha}_{AC}(D_{\text{test}}, D_{\text{val}}, \theta) = \frac{\hat{\alpha}_{CC}(D_{\text{test}}, \theta) - \hat{F}^{(-)}(D_{\text{val}}, \theta)}{\hat{F}^{(+)}(D_{\text{val}}, \theta) - \hat{F}^{(-)}(D_{\text{val}}, \theta)}. \quad (3)$$

In contrast to classify-and-count, the adjusted-count quantifier has been proven to be asymptotically unbiased [8, 10, 12]. The adjusted-count quantifier does not compute reliable prevalence estimates for each threshold value θ . If θ is such that the difference between true positive rate $\hat{F}^{(+)}(D_{\text{val}}, \theta)$ and false positive rate $\hat{F}^{(-)}(D_{\text{val}}, \theta)$ is small, then the numerator of Eq. (3) is small, which, in turn, leads to a large variance of the quantifier [8, 12].

Median sweep ($\hat{\alpha}_{\text{MS}}$)

The median sweep quantifier uses the adjusted-count quantifier to compute prevalence estimates for a range of threshold values. Then, it takes the median value of the computed range of prevalence estimates as the final estimate [3]. As a remedy for the large variance of the adjusted-count quantifier, Forman advised to only compute the adjusted-count quantifier for those threshold values θ for which the difference between $\hat{F}^{(+)}(D_{\text{val}}, \theta)$ and $\hat{F}^{(-)}(D_{\text{val}}, \theta)$ is bigger than $\frac{1}{4}$ [3]. In notation, the median sweep is

$$\hat{\alpha}_{\text{MS}}(D_{\text{test}}, D_{\text{val}}) = \text{med} \left(\left\{ \hat{\alpha}_{\text{AC}}(D_{\text{test}}, D_{\text{val}}, \theta) : \hat{F}^{(+)}(D_{\text{val}}, \theta) - \hat{F}^{(-)}(D_{\text{val}}, \theta) > \frac{1}{4} \right\} \right). \quad (4)$$

This can be simplified by only considering thresholds $\theta \in \{\hat{\delta}(x) : x \in D_{\text{test}}\}$, which can be computed easily. Now, we have a finite set of prevalence estimates, which makes it easy to compute the median.

We implement median sweep by fitting the estimated probabilities/scores of the validation set D_{val} to an empirical cumulative density function (*ecdf*). The empirical cumulative density function is computed similarly to the classify-and-count quantifier $\hat{\alpha}_{\text{CC}}$, but then using the validation data D_{val} conditional on the labels. Hence, $\hat{F}_{\text{MS}}^{(+)}(x)$ defines the function of the true positive rate using the median sweep paradigm and $\hat{F}_{\text{MS}}^{(-)}(x)$ defines the function of the false positive rate using the median sweep paradigm.

2.2. Continuous sweep

In this section, we first explain why it is difficult to derive the mean square error for the median sweep quantifier. Second, we introduce the continuous sweep quantifier. We introduce two variants of the continuous sweep quantifier: the original continuous sweep quantifier and the simplified continuous sweep quantifier.

Difficulties median sweep

In Section 2.1, we explained how the median sweep quantifier works. The median sweep quantifier has a few properties that make it difficult to derive the mean square error. We present the three most important reasons.

First, the classify-and-count quantifier $\hat{\alpha}_{\text{CC}}$ and classification rates $\hat{F}^{(-)}$ and $\hat{F}^{(+)}$, are interpreted as step functions in θ . Step functions are not differentiable, so are therefore difficult to study analytically. Second, outliers are removed using a complicated data-dependent function, see Eq. (4). Every test set has a different number of observations that pass the data-dependent function and therefore computations grow fast since we need to compute the variance for every number of observations that pass the data-dependent function. Third, it is in general difficult to compute the mean and variance of the median as a function, especially for complex algorithms and distributions. Even for proper densities, we need to invert the cumulative density function to compute the median analytically, which is often unavailable.

In the next subsection, we propose solutions to the problems that occur with median sweep and we introduce the continuous sweep quantifier.

Continuous sweep quantifier

The continuous sweep quantifier is a smoothed adaptation of the median sweep quantifier. The continuous sweep quantifier provides solutions for the problems that occur for the median sweep regarding computing theoretical results.

Instead of using step functions for the classify-and-count quantifier $\hat{\alpha}_{CC}$ and the classification rates $\hat{F}^{(-)}$ and $\hat{F}^{(+)}$, the continuous sweep quantifier uses continuous functions. If we know the type of distribution, estimating the classify-and-count quantifier $\hat{\alpha}_{CC}$ and the classification rates $\hat{F}^{(-)}$ and $\hat{F}^{(+)}$ can be done parametrically with maximum likelihood estimation. If we do not know the type of distribution, we use kernel methods to estimate the marginal densities. In this paper, we use kernel estimates to compute the continuous functions for the classify-and-count quantifier $\hat{\alpha}_{CC}$ and the classification rates $\hat{F}^{(-)}$ and $\hat{F}^{(+)}$. These functions are now continuous instead of discrete. Then, classification rates $\hat{F}^{(-)}$ and $\hat{F}^{(+)}$ are kernel cumulative density functions given a soft-classifier $\hat{\delta}(x)$ and validation data D_{val} , and where the classify-and-count quantifier $\hat{\alpha}_{CC}$ is a kernel cumulative density function given a soft-classifier $\hat{\delta}(x)$ and test data D_{test} . Figures 1a, 1b and 1c show some examples. The black dots in Figures 1a and 1b show the observations in D_{val} where from we construct the empirical density functions for the true positive rate and the false positive rate. The red lines show the continuous function of the classification rates using a kernel. The black dots in Figure 1c show the classify-and-count estimate for each observation in D_{test} and the red line shows the continuous function of the classify-and-count quantifier for each threshold value θ . Figure 1d shows two things: the continuous function of the adjusted-count quantifier using the functions in Figures 1a, 1b, and 1c, and the prevalence estimates of each observation in D_{test} that we need to compute median sweep. All continuous functions seem to resemble their discrete equivalent.

With continuous sweep, we should still consider that prevalence estimates for extreme values of θ have large variances. With median sweep, we discard every prevalence estimate where the difference between the classification rates is smaller than $\frac{1}{4}$. In order to keep the differences between continuous sweep and median sweep as small as possible, we propose to apply the same decision rule to continuous sweep as to median sweep. Consider two decision boundaries θ_l and θ_r , where θ_l is the lower (left) threshold value where $\hat{F}^{(+)}(D_{val}, \theta) - \hat{F}^{(-)}(D_{val}, \theta) = \frac{1}{4}$ and θ_r is the upper (right) threshold value where $\hat{F}^{(+)}(D_{val}, \theta) - \hat{F}^{(-)}(D_{val}, \theta) = \frac{1}{4}$. In Figure 1d, the decision boundaries θ_l and θ_r are showed with a vertical orange line. Then we integrate to compute the area between θ_l and θ_r , where $\hat{F}^{(+)}(D_{val}, \theta) - \hat{F}^{(-)}(D_{val}, \theta) \geq \frac{1}{4}$, and divide it by the difference between θ_l and θ_r . In Figure 1d, we see a slight difference between the decision boundaries of the continuous sweep quantifier and the decision rule of the median sweep quantifier. In this example, the median sweep quantifier allows observations with more extreme threshold values θ than the continuous sweep quantifier in their calculations. This can be seen by the blue dots that lay at the outside of the orange decision boundaries. Therefore, the kernels do not exactly match the discrete observations.

Using the estimated continuous distributions, we can estimate the adjusted-count quantifier for any threshold. Hence, instead of computing the median of discrete data points, we propose to use integration across the whole probability range to compute the expected value of $(\hat{\alpha}_{CS})$. Finding the median is more complex since we need to find the quantile function of $\hat{\alpha}_{AC}$. In Figure 1d, we see that the function of the adjusted-count quantifier against the threshold values

is not bijective. This property makes it hard to find the inverse function, which enables to compute the median. Therefore, we propose to compute the (weighted) mean instead of the median. Even though the median is a more robust estimator, we think that the mean should give similar estimates because the outliers are discarded using the decision rule. The mean can be computed by computing the area under the curve using integrals of the continuous functions.

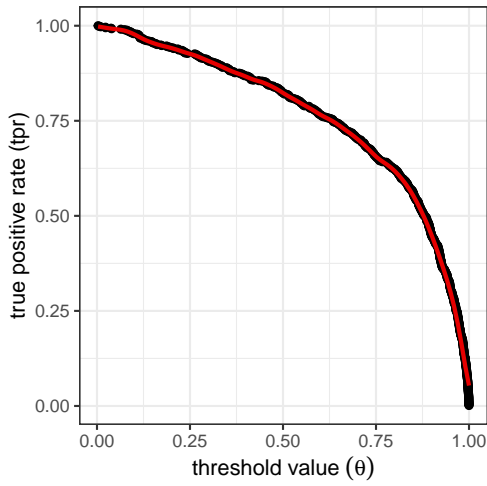
In order to make the continuous sweep quantifier as similar as the median sweep quantifier, we should weight areas with many observations in D_{test} more than areas with little observations in D_{test} . The probability density function of the observations' threshold values in D_{test} ($\hat{f}_{\delta(x)}(\theta)$) defines the weights of the continuous sweep quantifier. In fact, this is the (negative value of the) derivative of the classify-and-count quantifier with respect to θ . We have already computed the function of the classify-and-count quantifier and can use its derivative with respect to θ to compute the weights. Taking into account the decision boundaries, the expected value of the continuous sweep quantifier $\hat{\alpha}_{\text{CS}}$ is given by

$$\begin{aligned}\hat{\alpha}_{\text{CS}}(D_{\text{test}}, D_{\text{val}}, \theta_l, \theta_r) &= \frac{1}{\hat{F}(\theta_r) - \hat{F}(\theta_l)} \int_{\theta=\theta_l}^{\theta_r} \hat{f}_{\delta(x)}(\theta) \cdot \hat{\alpha}_{\text{AC}}(D_{\text{test}}, D_{\text{val}}, \theta) d\theta \\ &= \frac{1}{\hat{\alpha}_{\text{CC}}(D_{\text{test}}, \theta_r) - \hat{\alpha}_{\text{AC}}(D_{\text{test}}, \theta_l)} \int_{\theta=\theta_l}^{\theta_r} -\left(\frac{d}{d\theta} \hat{\alpha}_{\text{CC}}(D_{\text{test}}, \theta)\right) \hat{\alpha}_{\text{CC}}(D_{\text{test}}, D_{\text{val}}, \theta) d\theta \\ &= \frac{1}{\hat{\alpha}_{\text{AC}}(D_{\text{test}}, \theta_r) - \hat{\alpha}_{\text{CC}}(D_{\text{test}}, \theta_l)} \int_{\theta=\theta_l}^{\theta_r} -\left(\frac{d}{d\theta} \hat{\alpha}_{\text{CC}}(D_{\text{test}}, \theta)\right) \frac{\hat{\alpha}_{\text{CC}}(D_{\text{test}}, \theta) - \hat{F}^{(-)}(D_{\text{val}}, \theta)}{\hat{F}^{(+)}(D_{\text{val}}, \theta) - \hat{F}^{(-)}(D_{\text{val}}, \theta)} d\theta.\end{aligned}\tag{5}$$

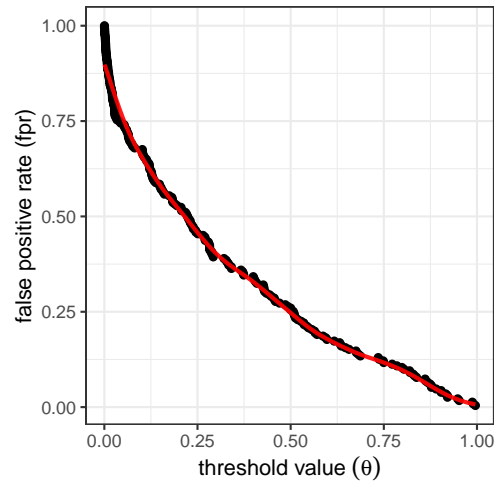
The integral of Eq. (5) is numerically tedious because it contains many estimates from the data. In order to reduce numerical complexity, we introduce the *simplified continuous sweep quantifier* $\hat{\alpha}_{\text{SCS}}$. The simplified continuous sweep quantifier does not contain $\hat{f}_{\delta(x)}(\theta)$ in the integral. The interpretation of leaving out this density is that we no longer weight areas with many observations in D_{test} more than areas with little observations in D_{test} . We believe that the impact of this omission on the theoretical properties of the quantifier are limited. We include a brief explanation, as an elaborate theoretical analysis is out of scope of this paper. First, we note that the adjusted count estimator is asymptotically unbiased for *every threshold value* θ [8, 10, 12]. Hence, the continuous sweep quantifier can be interpreted as a weighted average of asymptotically unbiased estimators and the simplified continuous sweep quantifier can be interpreted as an unweighted average of asymptotically unbiased estimators. Both quantifiers are therefore asymptotically unbiased estimators. The difference between the two is the asymptotic variance. A more detailed theoretical comparison between median sweep, continuous sweep, and simplified continuous will be included in a future paper. The key take home message is that the simplified continuous sweep quantifier is theoretically similar to the continuous sweep quantifier and has more appealing numerical properties. The simplified continuous sweep quantifier $\hat{\alpha}_{\text{SCS}}$ that can be computed as

$$\begin{aligned}\hat{\alpha}_{\text{SCS}}(D_{\text{test}}, D_{\text{val}}, \theta_l, \theta_r) &= \frac{1}{\theta_r - \theta_l} \int_{\theta=\theta_l}^{\theta_r} \hat{\alpha}_{\text{AC}}(D_{\text{test}}, D_{\text{val}}, \theta) d\theta \\ &= \frac{1}{\theta_r - \theta_l} \int_{\theta=\theta_l}^{\theta_r} \frac{\hat{\alpha}_{\text{CC}}(D_{\text{test}}, \theta) - \hat{F}^{(-)}(D_{\text{val}}, \theta)}{\hat{F}^{(+)}(D_{\text{val}}, \theta) - \hat{F}^{(-)}(D_{\text{val}}, \theta)} d\theta.\end{aligned}\tag{6}$$

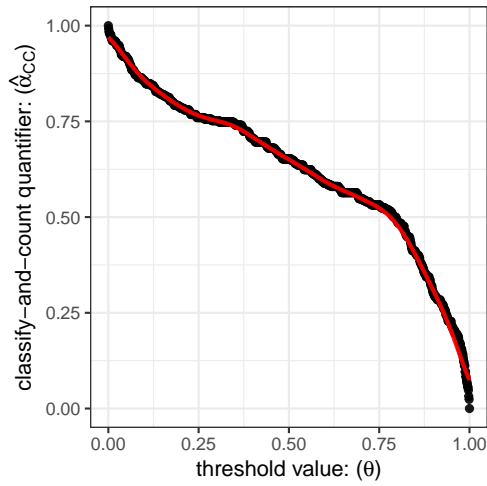
Concluding, the continuous sweep quantifiers are continuous adaptations of median sweep, but makes it easier to compute theoretical results. In the next section, we compare the continuous sweep quantifiers with the median sweep quantifier with the data provided by the LeQua2022 Task.



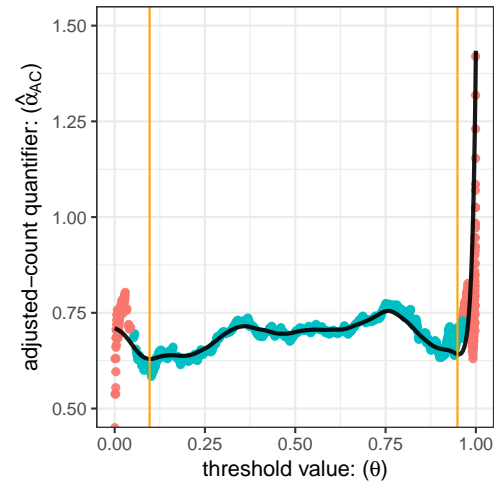
(a) True positive rate



(b) False positive rate



(c) Classify-and-count



(d) Adjusted-count

Figure 1: This figure shows the strong numerical similarity between median sweep as in Eq. 4 and our continuous sweep method as in Eqs. 5 and 6. In subfigures (a)-(c), the red curves are the continuous version of the discrete median sweep estimates (black dots). In subfigure (d), the black line shows the estimated adjusted-count value for every threshold value θ using the curve from subfigures (a)-(c). The vertical, orange lines show the decision boundaries θ_l and θ_r . The blue dots shows the adjusted-count estimates from median sweep that pass the criterion that the difference between the true positive rate and the false positive rate is larger than $\frac{1}{4}$, the red dots are the estimates that fail the criterion. The median sweep quantifier is computed by taking the median of the blue dots in subfigure (d). The simplified continuous sweep quantifier can be computed by integrating the area between the decision boundaries of subfigure (d) and divide it by the distance between the decision boundaries. The original continuous sweep quantifier can be computed by integrating the weighted area between the decision boundaries of subfigure (d) and divide it by the weighted distance between the decision boundaries. These weights are based on the classify-and-count quantifier.

3. Evaluation

In this section, we evaluate the continuous sweep quantifiers and the median sweep quantifier. In short, the objective is to quantify the prevalence α of positive product reviews (from a webshop) as accurate as possible across 5,000 test sets. For more information on the quantification task, we refer to the paper of the LeQua 2022 Task [6]. First, we explain the technical details of our study. Second, we show the results of the quantifiers on the test datasets. Third, we explain the similarities and differences between the continuous sweep quantifiers and the median sweep quantifier regarding the quantification task.

3.1. Technical setup

The analysis is performed using statistical software *R* version 4.1.3 [14]. Besides the core packages, we used *tidyverse* and *tidymodels* [15, 16]. The training data consists of 5,000 observations, each with 300 covariates and a label on whether the review is positive or negative. The training set is imbalanced: 3,870 reviews are positive and 1,130 reviews are negative. We randomly split this dataset in two parts: a training set D_{train} containing 4,000 observations and validation set D_{val} containing 1,000 observations from the complete training data. The training data D_{train} was balanced, which means that some of the negatively labelled observations are replicated to match the number of positively labelled observations.

Our classification model was a support vector machine (SVM) [17], denoted by $\hat{\delta}$. The SVM is trained with the training data D_{train} . The model had a linear kernel boundary and a regularisation parameter $C = 1$. We converted the decision values of the SVM to probabilities/scores using Platt scaling [18], such that we could use the theory of the previous section.

We computed the classify-and-count estimator $\hat{\alpha}_{\text{CC}}$ and classification rates $F_{\text{MS}}^{(+)}(x)$ and $F_{\text{MS}}^{(-)}(x)$ for the median sweep quantifier using the *ecdf* function. The *ecdf* function fits an empirical step function from the input data.

We computed the classify-and-count estimator $\hat{\alpha}_{\text{CC}}$ and classification rates $F_{\text{CS}}^{(+)}(x)$ and $F_{\text{CS}}^{(-)}(x)$ for the continuous sweep quantifiers using the *kde* function from the *ks* package [19]. Moreover, we computed $\hat{f}_{\delta(x)}(\theta)$ using the *kde* function from the same *ks* package. We added no additional arguments for both functions, except the boundaries for the estimated probabilities, which are set to 0 and 1.

3.2. Results

In this section, we evaluated the median sweep quantifier and the continuous sweep quantifiers on the test sets of the LeQua2022 task. First, we compared the median sweep quantifier and the continuous sweep quantifiers with the true prevalences. Second, we compared the median sweep quantifier with the continuous sweep quantifiers.

First, we evaluated the median sweep quantifier on the test sets. Figures 2a, 2b plot the estimated prevalence by the median sweep quantifier against the true prevalence, and the residuals. Obviously, the error of very small estimated prevalences is positive and the error of the very large estimated prevalence is negative. Moreover, it seems that there is a small positive bias among the estimated prevalences.

Table 1

Comparing summary statistics between the median sweep and continuous sweep quantifiers with the test sets.

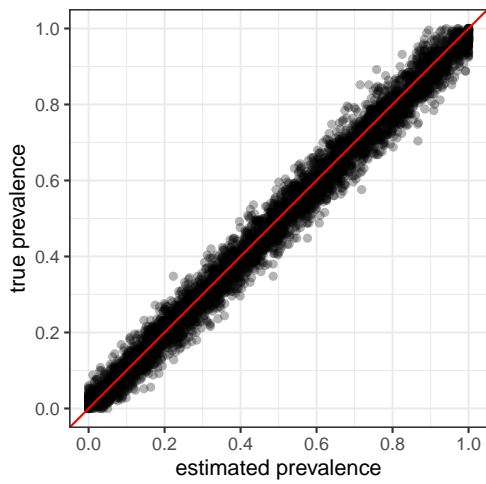
Quantifier	Bias	Variance	MAE
Continuous sweep	0.02565	0.00302	0.0473
Simplified continuous sweep	-0.00916	0.00151	0.0317
Median sweep	0.00650	0.00129	0.0289

Second, we evaluate the continuous sweep quantifiers on the test sets. Figure 2c and 2e plots the estimated prevalence by the continuous sweep quantifiers against the true prevalence, Figure 2d and 2f plot the estimated prevalence by the continuous sweep quantifiers against the residuals. We see different results between the continuous sweep quantifiers. The continuous sweep quantifier performs worse than the simplified continuous sweep quantifier: the continuous sweep quantifier has a large bias for large prevalence values and it has more variance than the simplified continuous sweep quantifier. It is clear that the simplified continuous sweep quantifier performs better than the original continuous sweep quantifier and therefore, we will now only compare the simplified continuous sweep quantifier with the median sweep quantifier.

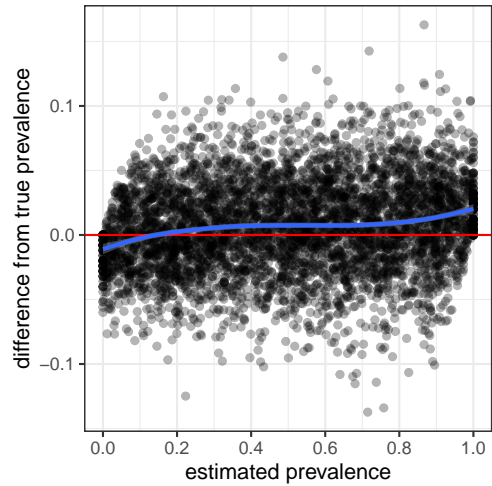
When we compare the median sweep quantifier with the simplified continuous sweep quantifier, we see some similarities and differences. The two quantifiers seem to have only little bias across the range of prevalences, however, the direction of the bias is different. Moreover, the pattern of the bias is different. The bias of the median sweep quantifier is monotonically increasing (Figure 2b) while the bias of the simplified continuous sweep quantifier seems to have a local minimum and a local maximum (Figure 2f). The variance of the simplified continuous sweep quantifier is slightly larger than the variance of the median sweep quantifier, see Table 1, hence the mean absolute error (MAE) of the simplified continuous sweep quantifier is slightly larger than the MAE of the median sweep quantifier.

The simplified continuous sweep quantifier has more variance than the median sweep quantifier. A reason could be that the mean is more sensitive to extreme values than the median. Figure 3 shows nine examples of the adjusted-count integral and the median sweep estimates. Remarkable is that the continuous sweep function is close to the discrete estimates over the whole range of θ , except around the value of θ_r . This difference can be a possible cause of the small difference between the continuous sweep quantifier and the median sweep quantifier.

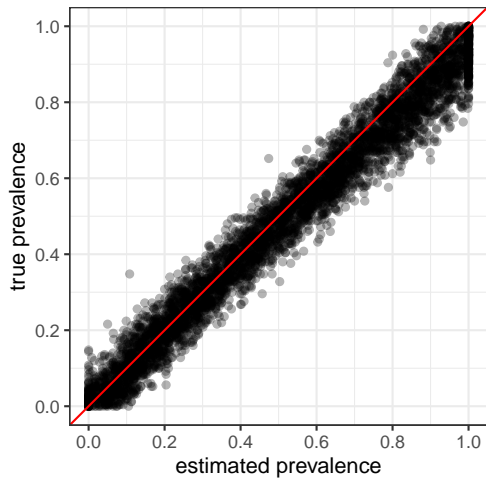
Concluding, the simplified continuous sweep quantifier is a quantifier that performs slightly worse than the median sweep quantifier using the procedure described in this section. The original continuous sweep quantifier performs much worse than the other two quantifiers. The results for the simplified continuous sweep quantifier and the median sweep quantifier are similar and we believe that we can use the (simplified) continuous sweep quantifier to compute theoretical results that are related to the median sweep quantifier.



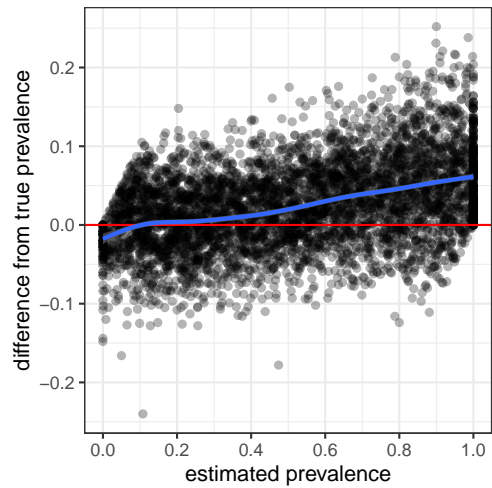
(a) Median sweep against true prevalences



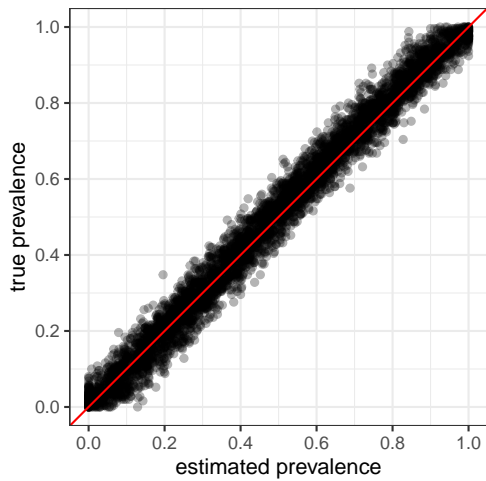
(b) Fitted residuals median sweep



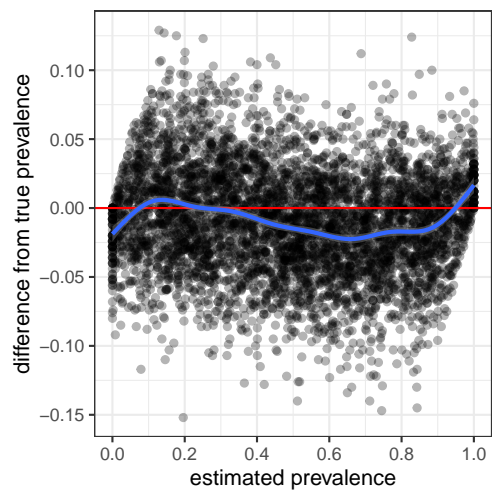
(c) Continuous sweep against true prevalences



(d) Fitted residuals continuous sweep



(e) Simplified continuous sweep against true prevalences



(f) Fitted residuals simplified continuous sweep

Figure 2: Quantifiers against true prevalence among 5,000 test sets. The fitted red lines plot the line where the estimated prevalence is equal to the true prevalence. The blue lines plot a fitted GAM-model representing the bias among the prevalences.

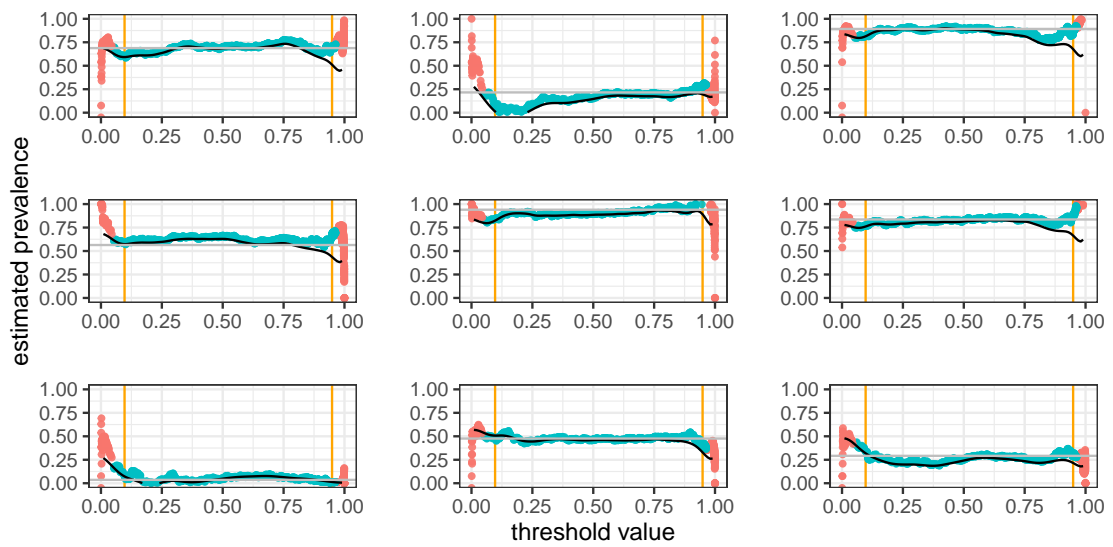


Figure 3: Nine examples of the adjusted-count integral. The black line denotes the estimated adjusted count quantifier at threshold θ for a development set. The orange vertical lines are the two decision boundaries θ_l and θ_r , and the grey horizontal lines denote the prevalence of each development set. The blue dots show the adjusted-count estimates from the median sweep that pass the criterion that the difference between the true positive rate and the false positive rate is larger than $\frac{1}{4}$, and the red dots are the estimates that fail the criterion.

4. Conclusion and Discussion

The goal of this paper was to design the continuous sweep quantifier, study its empirical performance, and specify a research agenda for the theoretical analysis of this new quantifier.

In this paper, we constructed the continuous sweep quantifier. We provided two versions of the continuous sweep quantifier: the original continuous sweep quantifier where every threshold is weighted with the classify-and-count quantifier and the simplified continuous sweep quantifier without weights. The continuous sweep quantifiers are based on the well-known median sweep quantifier. Previous research has shown that the median sweep quantifier is a good quantifier. However, it is not well understood why it performs well because it is hard to derive its theoretical properties. The median sweep quantifier uses empirical distributions for the classify-and-count quantifier $\hat{\alpha}_{AC}$ and the classification rates $F^{(+)}(x)$ and $F^{(-)}(x)$ which makes it hard to do proper calculations on, like differentiating and integrating. Moreover, median sweep uses discrete decision rules to remove outliers, which makes the calculations more complicated. Last, the median is hard to compute analytically since the functions of the prevalence α against threshold θ is non-bijective. Therefore, we proposed a new quantifier named the continuous sweep. The continuous sweep quantifier is a modification of the median sweep quantifier that enables computing theoretical results. The continuous sweep quantifier 1) used kernel estimates instead of the empirical distribution, 2) constructed decision boundaries instead of applying discrete decision rules, and 3) used the mean instead of the median. Figure 1 showed that the continuous functions are closely related to the empirical functions.

The simplified continuous sweep quantifier performed similar to the median sweep quantifier in terms of bias and variance. The original continuous sweep quantifier performed much worse than the simplified continuous sweep quantifier. Both continuous sweep quantifiers can be further optimized by choosing better kernels and other hyper-parameters.

The outline for the theoretical agenda is separated into two parts: defining the assumptions of the continuous distributions, and second we discuss how to compute the theoretical results.

First, we make assumptions about the continuous distributions. In this paper, the continuous distributions are kernels with default parameters estimated from the training and validation data. Deriving theoretical results from default kernels is still a cumbersome task. Therefore, we can make assumptions on the distributions. We start using basic distributional distributions such as the uniform. Later on, we can extend it to more complex distributional distributions such as the beta.

Second, we discuss how to compute the theoretical results. In the first step, we assume that the classification rates follow a uniform distribution where the limits are given. Then, we can compute the expected value of the classify-and-count quantifier for each prevalence α over each threshold value θ . Adding the information of the distributions of the classification rates, we can compute the expected value of the adjusted-count quantifier using [8] and iterate over the whole range of θ to compute the expected value of the continuous sweep quantifier. We can apply a similar strategy for the variance. Combining the expected value and the variance results in a value for the mean square error for the continuous sweep quantifier. Having the mean square error of the continuous sweep quantifier, we can compare it with the mean square error for other quantifiers like the adjusted count, calibration or mixed quantifier [8, 9, 10].

After computing theoretical results for the continuous sweep quantifier, we can further im-

prove the continuous sweep quantifier. The continuous sweep quantifier has been constructed to compute theoretical results for the median sweep quantifier. With innovative techniques regarding kernel estimates and handling large variances, we can improve the predictive performance of the continuous sweep quantifier.

In conclusion, the continuous sweep quantifier can be used to understand median sweep more thoroughly. It enables us to compute theoretical results for bias and variance in future papers.

References

- [1] P. González, A. Castaño, N. V. Chawla, J. Coz, A review on quantification learning, *ACM Computing Surveys* 50 (2017) 74:1–74:40.
- [2] G. Forman, Counting Positives Accurately Despite Inaccurate Classification, volume 3720, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 564–575. URL: http://link.springer.com/10.1007/11564096_55, series Title: Lecture Notes in Computer Science DOI: 10.1007/11564096_55.
- [3] G. Forman, Quantifying counts and costs via classification, *Data Mining and Knowledge Discovery* 17 (2008) 164–206. URL: <http://link.springer.com/10.1007/s10618-008-0097-y>. doi:10.1007/s10618-008-0097-y.
- [4] L. Milli, A. Monreale, G. Rossetti, F. Giannotti, D. Pedreschi, F. Sebastiani, Quantification trees, *IEEE*, 2013, p. 528–536.
- [5] V. N. Vapnik, *Statistical learning theory*, 1998.
- [6] A. Esuli, A. Moreo, F. Sebastiani, Lequa@clef2022: Learning to quantify, 2021. URL: <https://arxiv.org/abs/2111.11249>. doi:10.48550/ARXIV.2111.11249.
- [7] T. Schumacher, M. Strohmaier, F. Lemmerich, A comparative evaluation of quantification methods, *arXiv:2103.03223 [cs]* (2021). URL: <http://arxiv.org/abs/2103.03223>, arXiv: 2103.03223.
- [8] K. Kloos, Q. Meertens, S. Scholtus, J. Karch, Comparing correction methods to reduce misclassification bias, Springer International Publishing, Cham, 2021, pp. 64–90.
- [9] K. Kloos, A new generic method to improve machine learning applications in official statistics, *Statistical Journal of the IAOS* 37 (2021) 1181–1196. URL: <http://dx.doi.org/10.3233/SJI-210885>. doi:10.3233/sji-210885.
- [10] Q. A. Meertens, C. G. H. Diks, H. J. Van Den Herik, F. W. Takes, Understanding the output quality of official statistics that are based on machine learning algorithms, 2021.
- [11] D. Tasche, Fisher consistency for prior probability shift, *The Journal of Machine Learning Research* 18 (2017) 3338–3369.
- [12] D. Tasche, Minimising quantifier variance under prior probability shift, 2021. URL: <https://arxiv.org/abs/2107.08209>. doi:10.48550/ARXIV.2107.08209.
- [13] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, F. Herrera, A unifying view on dataset shift in classification, *Pattern recognition* 45 (2012) 521–530.
- [14] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. URL: <https://www.R-project.org/>.
- [15] H. Wickham, M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Golemund,

- A. Hayes, L. Henry, J. Hester, M. Kuhn, T. Pedersen, E. Miller, S. Bache, K. Müller, J. Ooms, D. Robinson, D. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, H. Yutani, Welcome to the tidyverse, *J. Open Source Softw.* 4 (2019) 1686. URL: <http://dx.doi.org/10.21105/joss.01686>. doi:10.21105/joss.01686.
- [16] M. Kuhn, H. Wickham, *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.*, 2020. URL: <https://www.tidymodels.org>.
- [17] J. H. Friedman, T. Hastie, R. Tibshirani, et al., *The elements of statistical learning*, Springer, New York, 2001. doi:10.1007/978-0-387-84858-7.
- [18] A. Karatzoglou, A. Smola, K. Hornik, A. Zeileis, *kernlab – an S4 package for kernel methods in R*, 2004. URL: <http://www.jstatsoft.org/v11/i09/>.
- [19] T. Duong, *ks: Kernel Smoothing*, 2022. URL: <https://CRAN.R-project.org/package=ks>, r package version 1.13.4.