Comments on Friedman's Method for Class Distribution Estimation

Dirk Tasche[0000−0002−2750−2970]

Unit for Data Science and Computing, North-West University, South Africa
55801447@nwu.ac.za

Abstract. The purpose of class distribution estimation (also known as quantification) is to determine the values of the prior class probabilities in a test dataset without class label observations. A variety of methods to achieve this have been proposed in the literature, most of them based on the assumption that the distributions of the training and test data are related through prior probability shift (also known as label shift). Among these methods, Friedman's method has recently been found to perform relatively well both for binary and multi-class quantification. We discuss the properties of Friedman's method and another approach mentioned by Friedman (called DeBias method in the literature) in the context of a general framework for designing linear equation systems for class distribution estimation.

Keywords: Prior probability shift · Label shift · Class prevalence · Quantification · Asymptotic variance

1 Introduction

The purpose of class distribution estimation (also known as quantification) is to determine the values of the prior class probabilities in a test dataset without class label observations. A variety of methods to achieve this have been proposed in the literature, most of them based on the assumption that the distributions of the training and test data are related through prior probability shift (also known as label shift). See González et al. [10] and Esuli et al. [6] for recent surveys of applications of and methods for quantification.

Friedman's [9] method has recently been found to perform relatively well both for binary and multi-class quantification (Schuhmacher et al. [18], Donyavi et al. [5]). On many real-world datasets, the performance of Friedman's method seems to exceed the performance of the EM algorithm (Saerens et al. [16]) which is an implementation of the maximum likelihood estimator for the test prior class probabilities (also called class prevalences). This observation is somewhat surprising because both Friedman's estimator and the EM algorithm involve estimates of the training posterior class probabilities which are notoriously hard to estimate. Hence one might expect that the performances of Friedman's method and the EM algorithm are at a more comparable level.

In order to find an explanation for the relatively good performance of Friedman's method, we study its properties and the properties of another approach mentioned by Friedman (called DeBias method by Castaño et al. [4]) in the context of a general framework for designing linear equation systems for class distribution estimation.

The outline of this paper and its contributions to the literature are as follows:

- Section 2 sets out the general assumptions and notation for the rest of the paper.
- In Section 3, we discuss the general framework for designing linear equation systems for class distribution estimation. (3b) and (3c) of Theorem 1 below appear to be novel, covariance-based versions of the basic equation (3a).
- In Section 4, we describe Friedman's method in detail and propose an alternative implementation that avoids direct estimation of the posterior class probabilities (Remark 1 below).
- In Section 5, we investigate conditions for the uniqueness of the solutions to linear equation systems for class distribution estimation. In Remark 3, we show that DeBias, the second method proposed for the binary case by Friedman [9] which involves the variance of one of the posterior class probabilities, is a special case of a covariance matrix-based approach to the multi-class case considered in Corollary 2. This provides an answer to the open research question "How to generalise the inequality of Corollary 6 of Tasche [21] to the multi-class case?" raised in Section 4.12 of Krempl et al. [13]. In addition, we show that the population versions of DeBias and 'Probabilistic adjusted count (PAC)' by Bella et al. [1] are identical (Remark 5 below).
- In Section 6, we compare the asymptotic variances of DeBias, Friedman's method and the maximum likelihood estimator in the binary case by means of a numerical example. The setting of the example is semi-asymptotic with an infinite training dataset and a finite large test dataset.
- Section 7 concludes the paper with a summary of the findings.

2 Setting

For this paper, we assume the following setting which is quite common in the study of dataset shift (see, for instance, Moreno-Torres et al. [15]):

- A class variable Y with values in $\mathcal{Y} = \{1, \ldots, \ell\}$ with $\ell \geq 2$ (multi-class case). A features vector X with values in $\mathcal{X}$.
- Each example (or instance) has a class label Y and features X.
- In the training dataset, for all examples their features X and labels Y are observed. P denotes the training (joint) distribution, also called source distribution, of (X, Y) of which the training dataset has been sampled.
- In the test dataset, only the features X of an example can immediately be observed. Its class label Y becomes known only with delay or not at all. Q denotes the test (joint) distribution, also called target distribution, of (X, Y) of which the test dataset has been sampled.

- We assume $0 < P[Y = y] < 1$, $0 < Q[Y = y] < 1$ for all $y \in \mathcal{Y}$.
- For the sake of a more concise notation, we define $p_y = P[Y = y]$ and $q_y = Q[Y = y]$ for $y \in \mathcal{Y}$.

We also use the notation $E_P[Z] = \int Z\,dP$ and $E_Q[Z] = \int Z\,dQ$ for integrable real-valued random variables Z.

The setting described above is called *dataset shift* or *distribution shift* in the literature if training and test distribution are not the same, i.e. $P \neq Q$. In the rest of the paper, we consider the following more specific type of dataset shift.

Definition 1. *The training distribution P and the test distribution Q are related through* prior probability shift *if for all $y \in \mathcal{Y}$ and all measurable sets $M \subset \mathcal{X}$ it holds that*[1]

$$P[X \in M|Y = y] \; = \; Q[X \in M|Y = y].$$

The term 'prior probability shift' appears to have been coined by Storkey [20]. In the literature, prior probability shift is also called *target shift* (Zhang et al. [27]), *label shift* (Lipton et al. [14]), or *global drift* (Hofer and Krempl [12]).

Prior probability shift implies dataset shift, i.e. $P \neq Q$, if $P[Y = y] \neq Q[Y = y]$ for at least one $y \in \mathcal{Y}$. Hence, as the class labels Y are not observed in the test dataset, the test prior probabilities $q_y = Q[Y = y]$ must be estimated from feature observations in the test dataset as well as feature and class label observations in the training dataset. Such an estimation procedure is called *quantification* or *class distribution estimation*.

3 Linear equations for class distribution estimation

In the following, we treat class distribution estimation under prior probability shift as a parametric estimation problem in a family of mixture distributions:

- We consider the distributions Q_X on $\mathcal{X}$ that can be represented as

$$Q_X[M] \; = \; \sum\nolimits_{y=1}^{\ell} q_y\, P[X \in M|Y = y] \tag{1}$$

  for all measurable sets $M \subset \mathcal{X}$. The family of these distributions is parametrised through the test prior class probabilities $(q_1, \ldots, q_\ell) \in (0,1)^\ell$ with the additional constraint

$$\sum\nolimits_{y=1}^{\ell} q_y \; = \; 1. \tag{2}$$

- Unless stated otherwise, for the purposes of this paper we assume that the conditional feature distributions $P[X \in M|Y = y]$, $M \subset \mathcal{X}$, under the training distribution are known and do not contribute to the estimation uncertainty.

---

[1] Recall the notion of conditional probability for events A and B: $P[A|B] = \frac{P[A \cap B]}{P[B]}$ if $P[B] > 0$ and $P[A|B] = 0$ otherwise.

– The parametrised distribution family defined in (1) is identifiable in the sense of Definition 11.2.2 of Casella and Berger [3], i.e. Q_X and Q_X^* differ whenever the corresponding parametrisations $(q_1, \ldots, q_\ell)$ and $(q_1^*, \ldots, q_\ell^*)$ differ.

According to San Martín and Quintana [17], identifiability is necessary for the existence of both asymptotically unbiased estimates and consistent estimates. This observation leaves open the question of how to find such estimates. In the following, we strive to design estimators of the class prior probabilities q_y as unique solutions to systems of linear equations[2].

Calling the following result a theorem is an exaggeration as its proof is very short and basic. But it is fundamental for the study and estimation of prior probability shift and in that sense deserves being called a theorem. Of course, Theorem 1 is not novel. In particular (3a) was mentioned by Saerens et al. [16] (Eq. (2.5), with Z chosen as a hard classifier) and quite likely also in earlier works. Even so, linking the notion of prior probability shift to the training dataset covariances of functions of the features and the indicators of the classes or the posterior class probabilities might have some degree of novelty, at least in the multi-class case.

Theorem 1. *Let $p_y = P[Y = y]$ and $q_y = Q[Y = y]$ for $y \in \mathcal{Y}$. Suppose that P and Q are related through prior probability shift in the sense of Definition 1 and that the random variable Z is integrable both under P and Q. Then it holds that*[3]

$$E_Q[Z] = \sum\nolimits_{y=1}^{\ell} q_y \, E_P[Z|Y = y] \tag{3a}$$

$$= \sum\nolimits_{y=1}^{\ell} \frac{q_y}{p_y} \operatorname{cov}_P\big(Z, \, \mathbf{1}_{\{Y=y\}}\big) + E_P[Z]. \tag{3b}$$

If Z is X-measurable, i.e. if there is a function $f : \mathcal{X} \to \mathbb{R}$ such that $Z = f(X)$, then it also follows that[4]

$$E_Q[Z] = \sum\nolimits_{y=1}^{\ell} \frac{q_y}{p_y} \operatorname{cov}_P\big(Z, \, P[Y = y|X]\big) + E_P[Z]. \tag{3c}$$

Proof. The theorem follows from the law of total probability combined with the definitions of conditional expectation and covariance respectively. □

(3a) provides the theoretical basis for Firat's ([7], Section 3.2) constrained regression approach for quantification under prior probability shift. Firat's K

[2] Other popular approaches to designing estimators include distribution matching (Gonzàlez et al. [10] and the references therein), ensemble methods (Serapião et al. [19] and the references therein) and expectation maximisation as implementation of maximum likelihood estimation (Saerens et al. [16]).

[3] For sets S, define the indicator function $\mathbf{1}_S$ by $\mathbf{1}_S(s) = 1$ if $s \in S$ and $\mathbf{1}_S(s) = 0$ if $s \notin S$.

[4] $P[Y = y|X]$ denotes the posterior probability of $Y = y$ given X in the sense of general conditional probability as defined, for instance, in Section 33 of Billingsley [2].

classes correspond to the ℓ classes of this paper. The L rows of Firat's matrix $\mathbf{X}$ emerge when (3a) is applied to L different variables $Z_1, \ldots, Z_L$.

As noted by Firat, (3a), (3b) or (3c) can be the starting point for setting up a system of linear equations for estimating the class prior probabilities q_y under prior probability shift. For instance, the choice $f_y(X) = \mathbf{1}_{C_y}(X)$ as crisp (or hard) 'one vs. all' classifier for class y, learned on the training dataset only, leads to the 'Adjusted Count' estimation approach used in the popular paper by Lipton et al. [14] who described it as 'method of moments'. Observe that for this version of 'one vs. all', there is no problem with changing the type of dataset shift, in contrast to the issue for combined 'one vs. all' quantifiers noted by Friedman [9] and Donyavi et al. [5].

Some questions should be considered when designing a concrete instance of such a linear equation system for quantification.

How many equations should be used? If the number of classes in the model is $\ell = |\mathcal{Y}|$ one might conclude that at least ℓ equations are needed in order to obtain a unique solution. However, as another consequence of the law of total probability, the q_y must additionally fulfil the linear equation (2). Hence, in order to achieve uniqueness of the solution, at least ℓ equations must be set up if (2) is considered a constraint that is checked once a solution has been found. Alternatively, if (2) is to be taken into account at the same time as the other equations, for uniqueness as a minimum it suffices to set up $\ell - 1$ additional equations on the basis of Theorem 1. Sticking with $\ell - 1$ equations has the advantage of reducing the number of random variables Z that must be chosen for the equations in Theorem 1.

If more then ℓ equations are set up the resulting linear equation system for the q_y is overdetermined such that in its sample-based versions there might be no exact solution at all. Nonetheless, the overdetermined case is naturally encountered when distribution-matching algorithms are implemented via binning of the feature space $\mathcal{X}$ (DF_x methods) or of the range of a continuous scoring classifier (DF_y methods), see Firat [7], Castaño et al. [4] and the references in the latter paper. To work around the lack of exact solutions, typically approximate solutions are determined by jointly minimising the differences between the left-hand and right-hand sides of the equations with respect to some specific metric like the Euclidean norm or the Hellinger divergence (see for instance Castaño et al. [4]).

In the following, we focus on the cases of systems of ℓ and $\ell - 1$ equations, in the latter case together with constraint (2).

How should the random variables Z appearing in the equations of Theorem 1 be chosen? A very basic criterion for choosing the variables Z is that it must be possible to compute them from observations of the features X only. This follows from the fact that on the left-hand sides of the equations in Theorem 1, the variables Z are integrated under the test distribution Q but the class labels Y are not observed under Q. Hence one must make sure that $Z = f(X)$ for some function f.

Among others, the following criteria for selecting such functions f have been considered in the literature:

- Reducing the variances of the estimated q_y. See Friedman [9] and Vaz et al. [25] for approaches to the direct minimisation of the variance. Findings by Vaz et al. [24] and Tasche [22] suggest that deploying variables Z that are able to separate the classes with high accuracy also reduces the variances of the class prior estimates.
- Speed of computation. See for instance Hassan et al. [11].

With the exception of Hassan et al. [11], in the literature primarily the choices $Z = \mathbf{1}_{C_y}$ (hard classifier for one of the classes y in $\mathcal{Y}$) and $Z = P[Y = y|X]$ (posterior probability under P for class y) have been considered. Below, we consider Friedman's [9] choices of hard classifiers and $Z = P[Y = y|X]$ in more detail.

4 Friedman's method

Friedman [9] proposed two class distribution estimation methods:

- He discussed in detail one method (later called 'Friedman's method' by Schuhmacher et al. [18]) based on a specific choice of hard classifiers both for the binary and multi-class cases. We revisit Friedman's method in this section.
- Another method, specified only for the binary case, is based on the variance of the posterior positive class probability under the training distribution (later called 'DeBias' method by Castaño et al. [4]). This method, without being named, had been mentioned before by Tasche [21] as Corollary 6. We discuss this approach in Remark 3 below.

First, we consider Friedman's method in the binary case $\ell = 2$. As Friedman himself wrote he was not the first researcher to think about this method.

Method Max (Forman [8], Section 2.2). Forman wrote on page 173: "Considering the earlier discussion of small denominators, another likely policy is where the denominator *(tpr-fpr)* is maximized: *method Max*." Here, Forman referred to crisp binary classifiers (not necessary most accurate) which were derived from a 'raw classifier' (i.e. a real-valued scoring classifier).

Accordingly, Friedman's method in the binary case is the special case of Forman's method Max when the underlying scoring classifier is chosen as the Bayes classifier, i.e. the posterior probability of the positive class.

Derivation of Friedman's method. Firat [7] describes on p. 2 the rationale for Friedman's method as follows: "Friedman uses the optimum threshold that minimizes the variance of proportion estimates (Friedman, 2014)." This statement is somewhat misleading, as Friedman [9] actually does not maximise the variance of the estimator but only the denominator on the right-hand side of the following equation (in the notation of this paper)

$$q_1 = \frac{E_Q[Z] - E_P[Z|y=2]}{E_P[Z|y=1] - E_P[Z|y=2]}, \tag{4}$$

over all random variables $0 \leq Z = f(X) \leq 1$. Note that (4) is a special case of (3a) for $\ell = 2$.

It turns out that

$$\arg \max_{f:\mathcal{X}\to[0,1]} E_P[f(X)|y=1] - E_P[f(X)|y=2] \;=\; f^* \tag{5}$$

with $f^*(x) = 1$ if $P[Y=1|X=x] > p_1$, $f^*(x) = 0$ if $P[Y=1|X=x] < p_1$ and $f^*(x)$ arbitrary if $P[Y=1|X=x] = p_1$.

The solution to the problem of minimising the sample variance of the estimator defined by (4) is less obvious. It has been tackled numerically by Vaz et al. ([25], Section 2.3), and by Tian et al. [23] by involving influence functions.

Remark 1. Friedman [9] and subsequent users of his method appear to have implemented it by means of plugging-in an estimate of the posterior probability $P[Y=1|X]$ into the function f^* as defined in (5). However, as $P[Y=1|X]$ could be difficult to estimate with satisfactory accuracy, such an implementation might be suboptimal.

Note that (5) is equivalent to

$$\arg \min_{f:\mathcal{X}\to[0,1]} (1-p_1)\, E_P[f(X)\, \mathbf{1}_{\{Y=1\}}] + p_1\, E_P[(1-f(X))\, \mathbf{1}_{\{Y=2\}}] \;=\; 1-f^*, \tag{6}$$

with f^* as in (5). (6) can be read as the problem to minimise the expected cost-sensitive error for a binary classification problem. This problem can be dealt with directly through a variety of approaches, resulting in approximations of the optimal classifier which do not require the estimation of $P[Y=1|X]$. The cost-sensitive minimisation problem can also be translated into a standard classification problem by appropriate re-weighting (Zadrozny et al. [26]). □

Friedman's method for more than two classes. Friedman [9] suggested defining $Z_y = f_y^*(X)$ for $y \in \mathcal{Y}$ with $f_y^*(x) = 1$ if $P[Y=y|X=x] > p_y$, $f_y^*(x) = 0$ if $P[Y=y|X=x] \leq p_y$, and then using (3a) with Z_y, $y = 1, \ldots, \ell$, to obtain a system of ℓ linear equations for the test prior probabilities of the classes $y \in \mathcal{Y}$.

According to Schuhmacher et al. [18], Friedman's method works well in binary quantification problems and still achieves good performance in multi-class settings.

5 Uniqueness of solutions and covariance matrix-based approaches

As discussed in Section 3, uniqueness of the solutions is an important design criterion for setting up a system of linear equations for class distribution estimation under prior probability shift. In this section, we provide more detail regarding the number of equations needed and look closer at designs based on covariance matrices estimated in the training dataset.

5.1 How many equations are needed?

(3c) of Theorem 1 is interesting because the choice $Z = P[Y = y|X]$ for fixed $y = 1, \ldots, \ell$, implies the matrix identity

$$\begin{pmatrix} E_Q\big[P[Y=1|X]\big] - p_1 \\ \vdots \\ E_Q\big[P[Y=\ell|X]\big] - p_\ell \end{pmatrix} = \varSigma_P \times \begin{pmatrix} \frac{q_1}{p_1} \\ \vdots \\ \frac{q_\ell}{p_\ell} \end{pmatrix}, \tag{7}$$

$$\varSigma_P = \begin{pmatrix} \mathrm{cov}_P\big(P[Y=1|X],\, P[Y=1|X]\big) & \ldots & \mathrm{cov}_P\big(P[Y=1|X],\, P[Y=\ell|X]\big) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}_P\big(P[Y=\ell|X],\, P[Y=1|X]\big) & \ldots & \mathrm{cov}_P\big(P[Y=\ell|X],\, P[Y=\ell|X]\big) \end{pmatrix}.$$

(7) connects the prior class probabilities p_y under the training distribution, the prior class probabilities q_y under the test distribution, and the averages under the test distribution $E_Q\big[P[Y = y|X]\big]$ of the training posterior class probabilities through the covariance matrix $\varSigma_P$ of the training posterior probabilities under the training distribution. All quantities in (7) but the test class prior probabilities q_y can be estimated from the training dataset and the features in the test dataset in principle. Hence, if the square matrix $\varSigma_P$ were invertible, (7) could be solved for the q_y by matrix inversion.

Unfortunately, as follows from the following proposition, the covariance matrix $\varSigma_P$ is never invertible.

Proposition 1. *Let $Z_1, \ldots, Z_r$ be integrable random variables under the distribution P. Suppose that Y is a discrete random variable with values in $\mathcal{Y} = \{1, \ldots, \ell\}$ with $\ell \geq 2$ and X is a random vector with values in $\mathcal{X}$. Define the matrices $M = (m_{ij})_{\substack{i=1,\ldots,r \\ j=1,\ldots,\ell}}$ and $M^* = (m^*_{ij})_{\substack{i=1,\ldots,r \\ j=1,\ldots,\ell}}$ by*

$$m_{ij} = \mathrm{cov}(Z_i,\, \mathbf{1}_{\{Y=j\}}) \quad \textit{and} \quad m^*_{ij} = \mathrm{cov}\big(Z_i,\, P[Y=j|X]\big).$$

Then it follows that

$$\mathrm{rank}(M) \leq \ell - 1 \quad \textit{and} \quad \mathrm{rank}(M^*) \leq \ell - 1.$$

Proof. Due to the fact that $1 = \sum_{j=1}^{\ell} \mathbf{1}_{\{Y=j\}}$ and $1 = \sum_{j=1}^{\ell} P[Y = j|X]$, the vector $v = (1, 1, \ldots, 1)^T \in \mathbb{R}^\ell$ is an element of the kernels of M and M^*, i.e. it holds that $M \times v = 0 = M^* \times v$. This implies the assertion. □

As a consequence of Proposition 1, there is no possible choice of random variables $Z_1, \ldots, Z_\ell$ that could serve on the basis of (3b) or (3c) to create a system of ℓ linear equations with a unique solution for the ℓ unknowns $q_1, \ldots, q_\ell$. However, Proposition 1 leaves open the question if such an equation system can be constructed on the basis of (3a).

Remark 2. For integrable random variables $Z_1, \ldots, Z_r$, define the matrix $\widetilde{M} = (\widetilde{m}_{ij})_{\substack{i=1,\ldots,r \\ j=1,\ldots,\ell}}$ by $\widetilde{m}_{ij} = E_P[Z_i|Y = j]$.

$\widetilde{M}$ can be rewritten as

$$\widetilde{M} \;=\; L \times D, \tag{8a}$$

where

$$L \;=\; \begin{pmatrix} E_P[Z_1\,\mathbf{1}_{\{Y=1\}}] & \dots & E_P[Z_1\,\mathbf{1}_{\{Y=\ell\}}] \\ \vdots & \ddots & \vdots \\ E_P[Z_r\,\mathbf{1}_{\{Y=1\}}] & \dots & E_P[Z_r\,\mathbf{1}_{\{Y=\ell\}}] \end{pmatrix} \tag{8b}$$

and $D = (d_{ij})_{i,j=1,\dots,\ell}$ is the diagonal matrix with $d_{ij} = \frac{1}{p_i}$ if $i = j$ and $d_{ij} = 0$ if $i \neq j$. In particular, we have $\mathrm{rank}(D) = \ell$.

Define the vector $v = (1, 1, \dots, 1)^T$ as in the proof of Proposition 1. Then it follows that $L \times v = (E_P[Z_1], \dots, E_P[Z_r])^T$. If $Z_1, \dots, Z_r$ are chosen such that $(E_P[Z_1], \dots, E_P[Z_r]) \neq 0$, as a consequence $L \times v \neq 0$ results. Hence there is no obvious reason as in the case of Proposition 1 for the rank of L (and by (8a) also of $\widetilde{M}$) to be less than maximal, i.e. being equal to $\min(r, \ell)$. This observation suggests that (3a) can be used to obtain a system of ℓ linear equations with a unique solution for the test class prior probabilities $q_1, \dots, q_\ell$. □

5.2 Invertible covariance matrices

The fact that the covariance Σ_P of the posterior class probabilities $P[Y = y|X]$, $y \in \mathcal{Y}$ in (7) cannot be inverted is caused by the linear dependence between the posterior probabilities since $\sum_{y=1}^{\ell} P[Y = y|X] = 1$. This issue can be avoided by disregarding one of probabilities, say $P[Y = \ell|X]$. Indeed, making use of the identity $\mathbf{1}_{\{Y=\ell\}} = 1 - \sum_{y=1}^{\ell-1} \mathbf{1}_{\{Y=y\}}$ in (3b) produces the following corollary to Theorem 1.

Corollary 1. *Let $p_y = P[Y = y]$ and $q_y = Q[Y = y]$ for $y \in \mathcal{Y}$. Suppose that P and Q are related through prior probability shift in the sense of Definition 1 and that the random variable Z is integrable both under P and Q. Then it holds that*

$$E_Q[Z] = \sum_{y=1}^{\ell-1} \left(\frac{q_y}{p_y} - \frac{q_\ell}{p_\ell}\right) \mathrm{cov}_P\big(Z,\, \mathbf{1}_{\{Y=y\}}\big) + E_P[Z]. \tag{9a}$$

If Z is X-measurable, i.e. if there is a function $f : \mathcal{X} \to \mathbb{R}$ such that $Z = f(X)$, then it also follows that

$$E_Q[Z] = \sum_{y=1}^{\ell-1} \left(\frac{q_y}{p_y} - \frac{q_\ell}{p_\ell}\right) \mathrm{cov}_P\big(Z,\, P[Y = y|X]\big) + E_P[Z]. \tag{9b}$$

Corollary 1 suggests the following approach to estimating the test class prior probabilities $q_1, \dots, q_\ell$.

Corollary 2. *Assume that the functions $f_1, \dots, f_{\ell-1} : X \to \mathbb{R}$ are such that the matrix*

$$C \;=\; \begin{pmatrix} \mathrm{cov}(f_1(X),\, \mathbf{1}_{\{Y=1\}}) & \dots & \mathrm{cov}(f_1(X),\, \mathbf{1}_{\{Y=\ell-1\}}) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(f_{\ell-1}(X),\, \mathbf{1}_{\{Y=1\}}) & \dots & \mathrm{cov}(f_{\ell-1}(X),\, \mathbf{1}_{\{Y=\ell-1\}}) \end{pmatrix} \tag{10a}$$

has rank $\ell - 1$, i.e. it is invertible.

Let $\big(E_Q[f_1(X)] - E_P[f_1(X)], \ldots, E_Q[f_{\ell-1}(X)] - E_P[f_{\ell-1}(X)]\big)^T = z$ *and* $C^{-1} \times z = (s_1, \ldots, s_{\ell-1})^T$.

Then it follows that

$$q_y = p_y \left(s_y + 1 - \sum\nolimits_{i=1}^{\ell-1} p_i\, s_i\right),\ y = 1, \ldots, \ell-1, \quad q_\ell = p_\ell \left(1 - \sum\nolimits_{i=1}^{\ell-1} p_i\, s_i\right). \tag{10b}$$

Observe that as a consequence of the general properties of conditional expectation[5] matrix C of (10a) can be represented as

$$C = \begin{pmatrix} \mathrm{cov}(f_1(X),\, P[Y=1|X]) & \ldots & \mathrm{cov}(f_1(X),\, P[Y=\ell-1|X]) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(f_{\ell-1}(X),\, P[Y=1|X]) & \ldots & \mathrm{cov}(f_{\ell-1}(X),\, P[Y=\ell-1|X]) \end{pmatrix}. \tag{11}$$

With the special choice $f_y(X) = P[Y = y|X]$ for $y = 1, \ldots, \ell-1$ matrix C as represented in (11) becomes the covariance matrix of Σ_P^* of $P[Y = 1|X], \ldots, P[Y = \ell - 1|X]$.

Remark 3 (DeBias method). Suppose we are in the binary case $\ell = 2$ and apply Corollary 2 with C as given in (11) and $f_1(X) = P[Y = 1|X]$. This implies $C = \Sigma_P^* = \mathrm{var}\big[P[Y = 1|X]\big]$. We then obtain by means of (10b)

$$q_1 = \frac{p_1\,(1-p_1)}{\mathrm{var}_P\big[P[Y=1|X]\big]} \big(E_Q\big[P[Y=1|X]\big] - p_1\big) + p_1, \tag{12a}$$

which is equivalent to

$$E_Q\big[P[Y=1|X]\big] = q_1\, \frac{\mathrm{var}_P\big[P[Y=1|X]\big]}{p_1\,(1-p_1)} + p_1 \left(1 - \frac{\mathrm{var}_P\big[P[Y=1|X]\big]}{p_1\,(1-p_1)}\right). \tag{12b}$$

(12b) appears to have been first published by Tasche [21] (Corollary 6) and then to have been presented at a conference by Friedman [9]. This approach to estimating q_1 has been called ‘DeBias’ method by Castaño et al. [4].

Hence, Corollary 2 with $C = \Sigma_P^*$ may be interpreted as multi-class extension of the DeBias approach. □

Remark 4 (Probabilistic Adjusted Count (PAC)). Suppose again we are in the binary case $\ell = 2$ and apply Corollary 2, this time with C as represented in (10a) and $f_1(X) = P[Y = 1|X]$. This implies $C = E_P[\big[P[Y = 1|X]\, \mathbf{1}_{\{Y=1\}}\big] - p_1^2$. We then obtain by means of (10b)

$$q_1 = p_1\,(1-p_1) \frac{E_Q\big[P[Y=1|X]\big] - p_1}{E_P[\big[P[Y=1|X]\,\mathbf{1}_{\{Y=1\}}\big] - p_1^2} + p_1 \tag{13a}$$

[5] See, for instance, Problem 34.6 of Billingsley [2]).

which is equivalent to

$$q_1 = \frac{E_Q\big[P[Y=1|X]\big] - E_P\big[P[Y=1|X]\,\big|\,Y=2\big]}{E_P\big[P[Y=1|X]\,\big|\,Y=1\big] - E_P\big[P[Y=1|X]\,\big|\,Y=2\big]}. \tag{13b}$$

(13b) was called 'probability estimation & average (P&A)' method by Bella et al. [1] but is now commonly referred to as 'probabilistic adjusted count (PAC)' (González et al. [10]). Its multi-class extension is sometimes called 'generalized probabilistic adjusted count (GPAC)' (see, for instance, Schuhmacher et al. [18]) and also covered by Corollary 2 with the choice $f_y(X) = P[Y = y|X]$ in (10a). □

Remark 5. Observe that in (13a) it holds that

$$E_P\big[\big[P[Y=1|X]\,\mathbf{1}_{\{Y=1\}}\big] - p_1^2 \;=\; \mathrm{var}_P\big[P[Y=1|X]\big].$$

By (12a), therefore in the binary case the DeBias and PAC methods for class distribution estimation are identical at population level, i.e. with infinite training and test datasets. This observation is not necessarily true in practice when DeBias and PAC estimates respectively are calculated based on sample versions of (12a) and (13a). □

6 Comparing asymptotic variances

As mentioned in Section 2, we consider class distribution estimation as a two-sample problem:

- A training sample for estimating certain quantities (e.g. the true positive and false negative rates of a classifier) under the training distribution because the quantities are needed for estimating the class prior probabilities under the test distribution.
- A test sample for estimating the class prior probabilities under the test distribution, based on the quantities estimated on the training sample.

Hence minimising the error of a method for class distribution estimation means minimising the estimation errors on the two samples.

In the following, we look at the semi-asymptotic binary case ($\ell = 2$) where

- the training distribution P is known (infinite sample) such that the prior class probabilities p_y and the posterior class probabilities $P[Y = y|X]$ can be exactly determined in the sense that the estimation error on the training sample vanishes.
- From the test distribution a finite but large sample of size n is given, and we focus upon unbiased estimators of the class prior probabilities.

For unbiased estimators the Cramér-Rao lower bound specifies a minimum value for the variance that cannot be undercut. Denote by $\widehat{q}_n^{\mathrm{ML}}$ the maximum-likelihood

(ML) estimator of the test prior probability q_1 of class 1 and by σ^2_{ML} its so-called asymptotic variance. Then $\frac{\sigma^2_{\mathrm{ML}}}{n}$ is the Cramér-Rao lower bound for the variances of the unbiased estimators of q_1 on test samples of size n when the training distribution is known (called here 'asymptotic setting'), see Section 5 of Tasche [22].

We compare σ^2_{ML} with the asymptotic variances in the sense of Definition 10.1.9 of Casella and Berger [3] of the Friedman estimator $\widehat{q}_n^{\mathrm{Fried}}$ and the DeBias estimator $\widehat{q}_n^{\mathrm{DeBias}}$ of the test prior probability q_1 of class 1.

We assume that both the conditional distribution of X given $Y = 1$ and the conditional distribution of X given $Y = 2$ have densities $g_1 > 0$ and $g_2 > 0$ with respect to some measure[6] μ. In particular, then the posterior probability $P[Y = 1|X = x]$ can be represented as

$$P[Y=1|X=x] \;=\; \frac{p_1\,g_1(x)}{p_1\,g_1(x)+(1-p_1)\,g_2(x)}, \tag{14}$$

and the density of the feature vector X under the test distribution Q is given by

$$g_Q \;=\; q_1\,g_1 + (1-q_1)\,g_2. \tag{15}$$

Since the training distribution P is assumed to be known, in the following all expected values $E_P[Z]$ are deterministic values that need not be estimated. In particular, also the prior probabilities p_1 and $p_2 = 1 - p_1$ are known constants. In contrast, the test distribution Q is not known but an i.i.d. sample $X_1, \ldots, X_n$ of the feature vector X drawn from its distribution under Q is observed.

ML estimator. For a detailed description of the ML estimator $\widehat{q}_n^{\mathrm{ML}}(X_1, \ldots, X_n) = \widehat{q}_n^{\mathrm{ML}}$ we refer to Section 4 of Tasche [22], as there is no closed-form representation of the ML estimator. However, its asymptotic variance σ_{ML} under Q is known:

$$\sigma^2_{\mathrm{ML}} = E_Q\left[\Big(\frac{g_1(X)-g_2(X)}{g_Q(X)}\Big)^2\right]^{-1} \;=\; \frac{q_1^2\,(1-q_1)^2}{\mathrm{var}_Q\big[E_Q[Y=1|X]\big]}. \tag{16}$$

σ^2_{ML} is characterised through the property that $\sqrt{n}\,\big(\widehat{q}_n^{\mathrm{ML}} - q_1\big)$ converges in distribution toward the normal distribution with mean 0 and variance σ^2_{ML}. Observe that σ^2_{ML} is a function of q_1 but not of p_1.

Friedman estimator. In the binary case, under the assumption on semi-asymptotics made for this section, the Friedman estimator $\widehat{q}_n^{\mathrm{Fried}}(X_1, \ldots, X_n) = \widehat{q}_n^{\mathrm{Fried}}$ based on the homonymous method presented in Section 4 can be written as

$$\widehat{q}_n^{\mathrm{Fried}} \;=\; \frac{\frac{1}{n}\sum_{i=1}^n f^\ast(X_i) - E_P[f^\ast(X)|y=2]}{E_P[f^\ast(X)|y=1] - E_P[f^\ast(X)|y=2]}, \tag{17a}$$

with $f^\ast$ defined through (5). Friedman [9] observed that $f^\ast$ can also be represented as

$$f^\ast(x) \;=\; \begin{cases} 1, & \text{if } g_1(x) > g_2(x),\\ 0, & \text{if } g_1(x) \le g_2(x). \end{cases} \tag{17b}$$

[6] In Example 1 below μ is the Lebesgue measure on $\mathbb{R}$.

As a consequence of (17b), the right-hand side of (17a) does not depend on p_1 or p_2 for $f^*(X)$ or any of the $f^*(X_i)$. Therefore, also $\widehat{q}_n^{\text{Fried}}$ as defined in (17a) does not change if p_1 or p_2 are changed. From the central limit theorem, it follows that $\sqrt{n}\,(\widehat{q}_n^{\text{Fried}} - q_1)$ under Q converges toward a normal distribution with mean 0 and variance σ^2_{Fried}. More precisely, the asymptotic variance of $\widehat{q}_n^{\text{Fried}}$ is

$$\sigma^2_{\text{Fried}} \;=\; \frac{E_Q[f^*(X)]\,\bigl(1 - E_Q[f^*(X)]\bigr)}{(E_P[f^*(X)|y=1] - E_P[f^*(X)|y=2])^2}. \tag{17c}$$

DeBias estimator. In the binary case, under the assumption on semi-asymptotics made for this section, the DeBias estimator $\widehat{q}_n^{\text{DeBias}} = \widehat{q}_n^{\text{DeBias}}(X_1, \ldots, X_n)$ based on the method presented in Remark 3 can be written as

$$\widehat{q}_n^{\text{DeBias}} \;=\; \frac{p_1\,(1-p_1)}{\text{var}_P\bigl[P[Y=1|X]\bigr]}\,\left(\frac{1}{n}\sum\nolimits_{i=1}^n P[Y=1|X=X_i] - p_1\right) + p_1. \tag{18a}$$

From the central limit theorem, it follows that $\sqrt{n}\,(\widehat{q}_n^{\text{DeBias}} - q_1)$ under Q converges toward the normal distribution with mean 0 and variance σ^2_{DeBias}, or more precisely, the asymptotic variance of $\widehat{q}_n^{\text{DeBias}}$ is

$$\sigma^2_{\text{DeBias}} \;=\; \left(\frac{p_1\,(1-p_1)}{\text{var}_P\bigl[P[Y=1|X]\bigr]}\right)^2 \text{var}_Q\bigl[P[Y=1|X]\bigr]. \tag{18b}$$

Note that it follows from (16) and (18b) that $\sigma^2_{\text{ML}} = \sigma^2_{\text{DeBias}}$ in the case of no shift, i.e. $p_1 = q_1$. As all quantities derived from P are assumed to be constant in the setting of this section, it follows as in Remark 5 that the asymptotic variance σ^2_{PAC} of the PAC estimator discussed in Remark 4 is identical with σ^2_{DeBias}, i.e. $\sigma^2_{\text{DeBias}} = \sigma^2_{\text{PAC}}$. For this reason, PAC is omitted from the following numerical example.

Example 1. We consider the same univariate binormal model with equal variances of the class-conditional distributions as in Section 7 of Tasche [22]: The two normal class-conditional distributions of the feature variable X are given by

$$X\,|\,Y=i \;\sim\; \mathcal{N}(\mu_i, \sigma^2), \qquad i = 1, 2 \tag{19a}$$

for conditional means $\mu_2 < \mu_1$ and some $\sigma > 0$. We choose

$$\mu_1 = 1.5, \quad \mu_2 = 0, \quad \text{and} \quad \sigma = 1. \tag{19b}$$

The model is then completely specified by choosing $p_1 = 0.15$ for the training prior probability of class 1. The test prior probability q_1 of class 1 is not fixed as we calculate asymptotic variances of the three above-mentioned prior distribution estimators for the whole range $(0, 1)$ of q_1. The results are shown in Figure 1. □

The following observations can be made from Figure 1:

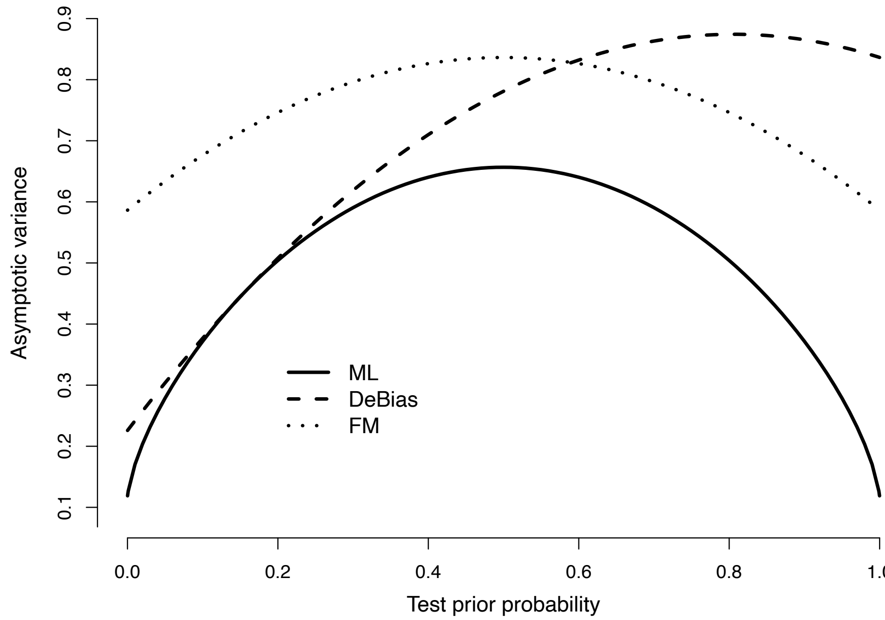

Fig. 1. Asymptotic variances of maximum likelihood estimator, DeBias estimator and Friedman estimator in a binormal model. See Example 1 for the specification of the underlying model.

- The asymptotic variance of the ML estimator is uniformly lower than the asymptotic variances of the other estimators for the whole possible range of the test prior probability of class 1 as is to be expected as a consequence of the Cramér-Rao inequality.
- The asymptotic variance of the Friedman estimator is not uniformly lower than the asymptotic variance of the DeBias estimator and vice versa.
- The DeBias estimator is almost optimal in the vicinity of the training prior probability ($p_1 = 0.15$) of class 1, as a consequence of (16) and (18b).
- In contrast, the asymptotic variance of the DeBias estimator is much larger than the asymptotic variance of the Friedman estimator in the $(0.8, 1)$ range of the test prior probability that is far away from the training prior probability 0.15.

7 Conclusions

We have considered Friedman's [9] method in the context of a general framework for designing linear equation systems for class distribution estimation and compared its binary version with DeBias which is another method proposed by Friedman, and the maximum likelihood estimator. The main findings of this paper are the following:

- The population versions of DeBias and Probability Adjusted Count (PAC, Bella et al. [1]) are identical and the binary special case of a new estimation approach based on inverting the covariance matrix of the training posterior class probabilities (see Section 5.2).
- Although the definition of Friedman's method appears to involve evaluations of the posterior probabilities under the training distribution, the method is potentially less sensitive to inaccuracies of the posterior estimates on smaller training datasets than the maximum likelihood estimator. This is a consequence of the fact that Friedman's methode can be implemented without a need to estimate the training posterior class probabilities (see Section 4).
- As shown in Example 1, Friedman's method may be locally outperformed in terms of asymptotic variance by DeBias. But thanks to its independence of the training prior class probabilities its performance is relatively uniform over the full range of possible values of the test prior probability of the positive class (class 1 in Example 1), in contrast to DeBias' poor performance for test prior probabilities which are very different to the corresponding training prior probability.

Acknowledgments. The author would like to thank three anonymous reviewers for their useful comments and suggestions.

References

1. Bella, A., Ferri, C., Hernandez-Orallo, J., Ramírez-Quintana, M.: Quantification via probability estimators. In: Data Mining (ICDM), 2010 IEEE International Conference on Data Mining. pp. 737–742. IEEE (2010)
2. Billingsley, P.: Probability and measure. John Wiley & Sons, second edn. (1986)
3. Casella, G., Berger, R.: Statistical Inference. Duxbury Press, second edn. (2002)
4. Castaño, A., Alonso, J., González, P., Pérez, P., del Coz, J.: QuantificationLib: A Python library for quantification and prevalence estimation. SoftwareX **26** (2024). https://doi.org/10.1016/j.softx.2024.101728
5. Donyavi, Z., Serapião, A., Batista, G.: MC-SQ: A Highly Accurate Ensemble for Multi-class Quantification. In: Proceedings of the 2023 SIAM International Conference on Data Mining (SDM). pp. 622–630. SIAM (2023)
6. Esuli, A., Fabris, A., Moreo, A., Sebastiani, F.: Learning to Quantify. Springer Cham (2023). https://doi.org/https://doi.org/10.1007/978-3-031-20467-8
7. Firat, A.: Unified framework for quantification. arXiv preprint arXiv:1606.00868 (2016)
8. Forman, G.: Quantifying counts and costs via classification. Data Mining and Knowledge Discovery **17**(2), 164–206 (2008)
9. Friedman, J.: Class counts in future unlabeled samples. Presentation at MIT CSAIL Big Data Event (2014)
10. González, P., Castaño, A., Chawla, N., Coz, J.D.: A Review on Quantification Learning. ACM Comput. Surv. **50**(5), 74:1–74:40 (2017)
11. Hassan, W., Maletzke, A., Batista, G.: Accurately Quantifying a Billion Instances per Second. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). pp. 1–10 (2020)

12. Hofer, V., Krempl, G.: Drift mining in data: A framework for addressing drift in classification. Computational Statistics & Data Analysis **57**(1), 377–391 (2013)
13. Krempl, G., Hofer, V., Webb, G., Hüllermeier, E.: Beyond Adaptation: Understanding Distributional Changes (Dagstuhl Seminar 20372). Dagstuhl Reports **10**(4), 1–36 (2021). https://doi.org/10.4230/DagRep.10.4.1
14. Lipton, Z., Wang, Y.X., Smola, A.: Detecting and Correcting for Label Shift with Black Box Predictors. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 3122–3130. PMLR (10–15 Jul 2018)
15. Moreno-Torres, J., Raeder, T., Alaiz-Rodriguez, R., Chawla, N., Herrera, F.: A unifying view on dataset shift in classification. Pattern Recognition **45**(1), 521–530 (2012)
16. Saerens, M., Latinne, P., Decaestecker, C.: Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. Neural Computation **14**(1), 21–41 (2002). https://doi.org/10.1162/089976602753284446
17. San Martín, E., Quintana, F.: Consistency and identifiability revisited. Brazilian Journal of Probability and Statistics pp. 99–106 (2002)
18. Schumacher, T., Strohmaier, M., Lemmerich, F.: A comparative evaluation of quantification methods. arXiv preprint arXiv:2103.03223 (2021)
19. Serapião, A., Donyavi, Z., Batista, G.: Ensembles of Classifiers and Quantifiers with Data Fusion for Quantification Learning. In: Bifet, A., Lorena, A., Ribeiro, R., Gama, J., Abreu, P. (eds.) Discovery Science. pp. 3–17. Springer Nature Switzerland, Cham (2023)
20. Storkey, A.: When Training and Test Sets Are Different: Characterizing Learning Transfer. In: Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N. (eds.) Dataset Shift in Machine Learning, chap. 1, pp. 3–28. The MIT Press, Cambridge, Massachusetts (2009)
21. Tasche, D.: Exact fit of simple finite mixture models. Journal of Risk and Financial Management **7**(4), 150–164 (2014)
22. Tasche, D.: Minimising quantifier variance under prior probability shift. In: Cong, G., Ramanath, M. (eds.) Proceedings of the CIKM 2021 Workshops (2021), first International Workshop on Learning to Quantify: Methods and Applications (LQ 2021)
23. Tian, Q., Zhang, X., Zhao, J.: ELSA: Efficient Label Shift Adaptation through the Lens of Semiparametric Models. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23 (2023), https://proceedings.mlr.press/v202/
24. Vaz, A., Izbicki, R., Stern, R.: Prior Shift Using the Ratio Estimator. In: Polpo, A., Stern, J., Louzada, F., Izbicki, R., Takada, H. (eds.) International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. pp. 25–35. Springer (2017)
25. Vaz, A., Izbicki, R., Stern, R.: Quantification Under Prior Probability Shift: the Ratio Estimator and its Extensions. Journal of Machine Learning Research **20**(79), 1–33 (2019), http://jmlr.org/papers/v20/18-456.html
26. Zadrozny, B., Langford, J., Abe, N.: Cost-sensitive learning by cost-proportionate example weighting. In: Third IEEE International Conference on Data Mining. pp. 435–442 (2003). https://doi.org/10.1109/ICDM.2003.1250950
27. Zhang, K., Schölkopf, B., Muandet, K., Wang, Z.: Domain Adaptation Under Target and Conditional Shift. In: Proceedings of the 30th International Conference on International Conference on Machine Learning – Volume 28. pp. III–819–III–827. ICML'13, JMLR.org (2013)