Recalibrating binary probabilistic classifiers

Dirk Tasche

Centre for Business Mathematics and Informatics, North-West University, South Africa
55801447@nwu.ac.za

Abstract. Recalibration of binary probabilistic classifiers to a target prior probability is an important task in areas like credit risk management. We analyse methods for recalibration from a distribution shift perspective. Distribution shift assumptions linked to the area under the curve (AUC) of a probabilistic classifier are found to be useful for the design of meaningful recalibration methods. Two new methods called parametric covariate shift with posterior drift (CSPD) and ROC-based quasi moment matching (QMM) are proposed and tested together with some other methods in an example setting. The outcomes of the test suggest that the QMM methods discussed in the paper can provide appropriately conservative results in evaluations with concave functions like for instance risk weights functions for credit risk.

Keywords: Probabilistic classifier · posterior probability · prior probability · calibration · distribution shift · dataset shift · credit risk.

1 Introduction

Occasionally binary probabilistic classifiers are learned on a training dataset and then are applied to a test dataset which reflects a joint distribution of features and labels different than the distribution of the training dataset. Actually, quite often it is unknown how much the training and test distributions differ because for the instances in the test dataset only the features but not the labels can be observed. Hence, the feature training and test distributions might be different while the posterior probabilities are identical for the training and test datasets. This kind of dataset shift is called covariate shift and is rather benign in principle as it would not require any change of the probabilistic classifier (Storkey [21]). If however, another type of dataset shift other than covariate shift is incurred, applying the classifier learned on the training dataset to the instances in the test dataset without any changes risks to generate unreliable predictions of the labels.

In this paper, we study the situation where indeed no labels are observed in the test dataset but where there exists an estimate of the proportion of positive labels, i.e. an estimate of the test prior probability of the positive class. The problem is then to *recalibrate* the probabilistic classifier learned on the training dataset such that the mean of the recalibrated classifier on the test dataset matches the estimate of the positive labels proportion.

This is a common situation in credit risk management, see for instance Chapter 4 of Bohn and Stein [3]. So called probabilities of default (PDs) are estimated

on a training sample with observations of solvent and defaulted borrowers and must be recalibrated before being evaluated for the borrowers in a live portfolio. The future solvency states of these borrowers in general are unknown but typically an estimate of the proportion of borrowers who are going to default is available. Sometimes, such estimates are conservative, i.e. they are likely to significantly overestimate the proportion of defaulters. Conservatism of the estimates can be a regulatory requirement or it could be part of stress testing exercises intended to assess the impact of unfavourable economic circumstances on the portfolio.

Recalibration of posterior probabilities learned on a training dataset to a target prior probability on a test dataset is not a well-defined problem because there is more than one way to transform the original posterior probabilities such that the target is matched. Therefore, we are going to study the impact of the recalibration method on the values of concave or nearly concave functions of the posterior probabilities as an additional criterion to identify meaningful solutions. The risk weight functions for the calculation of minimum required capital under the "internal ratings based (IRB) approach" of the Basel credit risk framework (BCBS [2]) are a primary example of such functions. The findings of this paper can inform the choice of the recalibration method in credit risk management and similar contexts.

This paper is organised as follows: Section 2 puts the paper into the context of related work. Section 3 describes the technical details of the setting of the paper and the recalibration problem. In addition, it introduces the evaluation of the solutions with a concave function as a criterion for assessing their appropriateness. Section 4 presents a number of methods for recalibration. Parametric CSPD (Section 4.4) and ROC-based QMM (Section 4.5) seem to be new. With an example in Section 5, we illustrate the dependence of the solutions to the recalibration problem on assumptions of distribution shift and identify some less reliable recalibration methods. Section 6 proposes a way forward to prudent recalibration and concludes the paper. Appendix A provides additional technical details needed for the implementation of some of the methods discussed in Section 4.

2 Related work

Calibration of probabilistic classifiers has often been treated in the literature, see the surveys by Ojeda et al. [14] and Silva Filho et al. [20]. In the typical calibration setting, a real-valued score is learned on a training dataset with joint observations of instances and labels. The score is then *calibrated* or *mapped* to become a probabilistic classifier on a test (or validation or calibration) dataset for which there are also joint observations of instances and labels. If the original score is already a probabilistic classifier then the term recalibration is sometimes used instead of calibration.

Cautious calibration (Allikivi et al. [1]) is a variant of binary calibration with the goal to avoid either overconfidence or underconfidence of probabilistic classifiers. Like for calibration, it is assumed that the labels of the instances in the calibration and test datasets are known. A common feature with recalibration

as defined in this paper is conservatism of the posterior estimates. In the setting of this paper, conservatism can be achieved by choosing an appropriate value for the target class-1 prior probability.

Recalibration of probabilities of default (PDs) on a dataset without observation of labels but with knowledge of the prior probability of positive labels has been a topic of research for twenty or more years in credit risk (Bohn and Stein [3] and the references therein). Such recalibration may be considered an extreme case of learning with label proportions (Quadrianto et al. [16]) where there are no individual label observations but label proportions for groups of instances are available. In general, 'learning with label proportions' in the binary case requires that there are at least two groups of instances with different proportions of positive labels such that results from the related research are not applicable to the recalibration problem as studied in this paper.

Quantification (or class distribution estimation, CDE) is another related problem. See Esuli et al. [8] for a recent survey. The goal of CDE in the binary case is to estimate the proportion of positive labels in a test dataset without any information on the labels. The primary common feature of the CDE and recalibration problems is the dependence of the solutions upon assumptions on the type of distribution shift between the training and test datasets. However, Remark 1 in Section 4.4 shows that the one-parameter version of the recalibration method 'parametric CSPD' may also be used for CDE.

Covariate shift with posterior drift (CSPD) as introduced by Scott [19] turns out to be a useful assumption on the type of distribution shift for tackling the recalibration problem. See Sections 4.4 and 5 below.

3 Setting

In this paper, we consider binary, i.e. two-class classification problems for which we assume the following setting:

- There are a class variable Y with values in $\mathcal{Y} = \{0, 1\}$ and a features (also called covariates) vector X with values in $\mathcal{X}$. Each example (or instance) to be classified 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) from which the training dataset has been sampled.
- In the test dataset, only the features X of an example can immediately be observed. Its label Y becomes known only with delay or not at all. Q denotes the test joint distribution, also called target distribution, of (X, Y) from which the test dataset has been sampled.
- For the sake of a more concise notation, we write for short $p = P[Y = 1]$ and $q = Q[Y = 1]$ and assume $0 < p < 1$ and $0 < q < 1$.

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* (Storkey [21]) or *distribution shift* (Lipton et al. [12]) if source and target distribution of (X, Y) are not the same, i.e. in case of $P(X, Y) \neq Q(X, Y)$.

3.1 Recalibration

We assume that since the joint source distribution $P(X, Y)$ of features $X \in \mathcal{X}$ and class $Y \in \{0, 1\} = \mathcal{Y}$ is given, also the posterior probability $\eta_P(X) = P[Y = 1 \,|\, X]$ is known or can be estimated from the training dataset. $\eta_P(X)$ is a special case of probabilistic classifiers η which are real-valued statistics with $0 \le \eta \le 1$ and typically intended to approximate $\eta_P(X)$. Probabilistic classifiers for their part are special cases of scores (or scoring classifiers) which are real-valued statistics intended to provide a ranking of the instances in the dataset in the sense that a high score suggests a high likelihood that the instance has class label 1 (positive label).

By assumption only the target marginal feature distribution $Q(X)$ is observed while the target joint distribution $Q(X, Y)$ is unknown. Nonetheless, for the recalibration problem, we also assume the target marginal label distribution, specified by $q = Q[Y = 1]$ to be known. This is not in contradiction to $Q(X, Y)$ being unknown because in general the ensemble of marginal distributions does not uniquely determine the joint distribution. We will encounter an example for this phenomenon in Section 5 below.

The goal is to fit a posterior probability $\eta_Q(x) = Q[Y = 1 \,|\, X = x]$ as some transformation T of $\eta_P(x)$ such that in particular it holds that

$$q \;=\; E_Q[\eta_Q(X)] \;=\; E_Q\big[T(\eta_P(X))\big]. \tag{1a}$$

In the following, we call this problem *recalibration* of the posterior probability $\eta_P(X)$ to a new prior probability q of class 1 under the target distribution.

Note that the assumption

$$\eta_Q(X) \;=\; T\big(\eta_P(X)\big) \tag{1b}$$

appears quite natural but actually is rather strong. Indeed, by Theorems 32.5 and 32.6 of Devroye et al. [7], (1b) is equivalent to $\eta_P(X)$ being sufficient for X with respect to Y under Q, i.e. $Q[Y = 1 \,|\, X] \;=\; Q[Y = 1 \,|\, \eta_P(X)]$. This sufficiency property may be interpreted as 'the information provided by $\eta_P(X)$ about Y is as good as the information by the whole set of features X under the target distribution Q'.

3.2 Non-uniqueness of recalibration

The recalibration problem is not well-posed in the sense that its solution is not unique. Therefore, we study it in a context where underestimating $E_Q\big[C(\eta_Q(X))\big]$ for some fixed concave function $C : [0, 1] \to \mathbb{R}$ ought to be avoided. In general, it holds that (by Jensen's inequality and Lemma 1.2 of Lalley [11])

$$(1 - q)\, C(0) + q\, C(1) \;\le\; E_Q\big[C(Z)\big] \;\le\; C(q) \tag{2}$$

for any random variable $0 \le Z \le 1$ with $E_Q[Z] = q$, and $Z = \eta_Q(X)$ in particular. The maximum value $C(q)$ is taken for constant $Z = q$, the minimum value $(1 - q)\, C(0) + q\, C(1)$ is realised for Z with $Q[Z = 1] = q = 1 - Q[Z = 0]$.

However, assuming a distribution shift between source P and target Q which results in a constant posterior probability $\eta_Q(x) = q$ appears too restrictive in most real world environments. In Section 5 below, we demonstrate that assuming preservation of classification performance as measured by AUC (Area Under the Curve, see next section) between source P and target Q strikes a sensible note between too restrictive and too tolerant assumptions for the estimation of $E_Q\big[C(\eta_Q(X))\big]$ under the target features distribution.

3.3 Area Under the Curve (AUC)

AUC (Area Under the Curve[1]) is a popular measure of performance of a binary classifier, i.e. AUC is considered an appropriate measure of the classifier's ability to predict the true class label of an instance. See Chen et al. [5] for related comments. Since AUC plays an important role in some of the recalibration methods discussed in the following sections, we present here the population-level (in contrast to sample-based) representations of AUC that are used in this paper.

Let $S = h(X)$ be a score which is a function of the features with values in an ordered set $\boldsymbol{\mathcal{S}}$. Define the class-conditional distributions P_y, $y \in \mathcal{Y} = \{0, 1\}$, of S by $P_y[S \in M] = P[S \in M \,|\, Y = y]$, for all measurable $M \subset \boldsymbol{\mathcal{S}}$.

If the score S is assumed to be large for instances with high likelihood to have class 1 and small for instances with high likelihood to have class 0, then the AUC for S is defined as

$$AUC_S \;=\; P^*[S_1 > S_0] + \frac{1}{2}\,P^*[S_1 = S_0], \tag{3a}$$

where P^* denotes the product measure of P_1 and P_0, and S_1 and S_0 are the coordinate projections of the space $\boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{S}}$ on which P^* is defined. As a consequence, S_1 and S_0 are independent and $P^*[S_y \le s] = P_y[S \le s]$ for $y \in \{0, 1\}$ and all $s \in \boldsymbol{\mathcal{S}}$.

By Definition (3a), on the one hand AUC_S is "equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance" (p. 868 of Fawcett [9]) if both of the class-conditional distribution functions of the score S are continuous. On the other hand, it holds that $AUC_S = \frac{1}{2}$ for any uninformative score S – i.e. in the case $P_0 = P_1$ – even if both of the class-conditional score distribution functions have discontinuities. Note that computing AUC_S by means of (3a) at first glance requires knowledge of the joint distribution $P(S, Y)$ of S and Y which would have to be inferred from a sample of paired (S, Y) observations.

However, AUC_S can also be determined if the distribution $P(S)$ of S and the posterior probabilities $\eta_P(S) = P[Y = 1 \,|\, S]$ are known. Then with $p =$

[1] 'Curve' refers to *ROC* (Receiver Operating Characteristic). See Fawcett [9] and the references therein for the most common definitions of ROC and AUC.

$E_P[\eta_P(S)]$ it holds that[2]

$$\begin{aligned} P_1[S \in M] &= \frac{E_P\big[\eta_P(S)\,\mathbf{1}_M(S)\big]}{p} \quad \text{and} \\ P_0[S \in M] &= \frac{E_P\big[(1-\eta_P(S))\,\mathbf{1}_M(S)\big]}{1-p} \end{aligned} \tag{3b}$$

for all measurable sets $M \subset \boldsymbol{\mathcal{S}}$. Plug P_0 and P_1 from (3b) in (3a) to compute AUC_S.

To indicate the way AUC_S is computed, in this paper we refer simply to AUC if AUC_S is assumed to be computed or inferred by means of (3a) irrespectively of the origin of P_0 and P_1. We refer to *implied* AUC if AUC_S is assumed to be computed by means of (3b) in combination with (3a). See Appendix A.1 for a more detailed formula for implied AUC in the case of a discrete-valued score S.

4 Approaches to recalibration

Any solution as in (1b) to the recalibration problem together with the marginal feature distribution $Q(X)$ completely determines the target distribution $Q(X,Y)$. Since $Q(X,Y) \neq P(X,Y)$ in case $q \neq p$, any given solution specifies some distribution shift between the source and target distributions. Hence, in order to better understand the consequences of selecting a particular solution transformation T, it is natural to explore the solution approaches through the lens of distribution shift. In the following, we assume $0 < \eta_P(x) < 1$ for all $x \in \mathcal{X}$.

4.1 Recalibration under assumption of stretched or compressed covariate shift

At first glance recalibration of $\eta_P(X)$ to a fixed target prior probability q might appear to be straightforward: Just define $\eta_Q(x) = \frac{q}{E_Q[\eta_P(X)]}\,\eta_P(x)$ for $x \in \mathcal{X}$, then $E_Q[\eta_Q(X)] = q$ immediately follows. This approach is called *scaling* (Section 3.1 of Ptak-Chmielewska and Kopciuszewski [15]).

Unfortunately, in some cases there is a problem with this approach: $1 < \eta_Q(x)$ may be incurred in the case $q > E_Q[\eta_P(X)]$. To avoid this issue, one can modify the approach to become

$$\eta_Q(x) = \min\big(t\,\eta_P(x), 1\big), \tag{4a}$$

with $t > 0$ being determined by

$$q \;=\; E_Q\big[\min\big(t\,\eta_P(x), 1\big)\big]. \tag{4b}$$

(4a) could be called *capped scaling* of the source posterior probabilities. Obviously, (4a) implies that (1b) is satisfied with $T(\eta) = \min(t\,\eta, 1)$.

Recall that source distribution $P(X,Y)$ and target distribution $Q(X,Y)$ are related through *covariate shift* (in the sense of Storkey [21]) if it holds that

[2] The *indicator function* $\mathbf{1}_A$ is defined by $\mathbf{1}_A(a) = 1$ if $a \in A$ and $\mathbf{1}_A(a) = 0$ if $a \notin A$.

$P[Y = y \,|\, X] = Q[Y = y \,|\, X]$ for all $y \in \mathcal{Y}$ with probability 1 both under P and Q. Hence if $t > 1$ in (4a), one might call the implied distribution shift *stretched covariate shift*.

In case $t < 1$ the induced distribution shift could be considered *compressed covariate shift*. In case $q > E_Q[\eta_P(X)]$, (4a) and (4b) imply $t > 1$ and $\eta_Q(x) = 1$ for all x with $t\,\eta_P(x) \geq 1$. This may have the consequence of an unjustified increase of $AUC_{\eta_Q(X)}$ compared to $AUC_{\eta_P(X)}$ for the area under the curve AUC defined as in Section 3.3 below.

4.2 Recalibration under assumption of label shift

Source distribution $P(X, Y)$ and target distribution $Q(X, Y)$ are related through *label shift* (Lipton et al. [12]), previously called *prior probability shift* in the literature (Storkey [21]), if $P[X \in M \,|\, Y = y] = Q[X \in M \,|\, Y = y]$ for all measurable sets $M \subset \mathcal{X}$ and $y \in \mathcal{Y}$.

Under the assumption of label shift, the target feature distribution $Q(X)$ can be represented as

$$Q[X \in M] \;=\; q\,P[X \in M \,|\, Y = 1] + (1 - q)\,P[X \in M \,|\, Y = 0], \tag{5}$$

for all measurable sets $M \subset \mathcal{X}$. If a prior probability $q = Q[Y = 1]$ is given, then according to the *posterior correction formula* (Eq. (2.4) of Saerens et al. [18]), η_Q is determined through (recall $p = P[Y = 1]$)

$$\eta_Q(x) \;=\; \frac{\frac{q}{p}\,\eta_P(x)}{\frac{q}{p}\,\eta_P(x) + \frac{1-q}{1-p}\,(1 - \eta_P(x))}, \tag{6}$$

for all $x \in \mathcal{X}$ with probability 1 under Q. In the credit risk community, the use of (6) for recalibration is popular (Section "Calibrating to PDs" of Bohn and Stein [3], Section 3.1 of Ptak-Chmielewska and Kopciuszewski [15]) because it avoids the problem of $\eta_Q(X)$ potentially taking the value 1 for large $\eta_P(X)$ which may be encountered with capped scaling as in (4a). Cramer [6] (Sections 6.2 and 6.3) pointed out that in the context of logistic regression the related special case of (6) was known at least since 1979. Note that (6) implies (1b) with $T(\eta) = \frac{\frac{p}{q}\,\eta}{\frac{p}{q}\,\eta + \frac{1-p}{1-q}\,(1-\eta)}$ strictly increasing in η.

Under the label shift assumption, the following observation is well-known. Define AUC (area under the curve) as in Section 3.3.

Proposition 1. *Define the score S by $S = \eta_P(X)$ and the score S^* by $S^* = \eta_Q(X)$. If P and Q are related through label shift then it follows that $AUC_S = AUC_{S^*}$.*

Proposition 1 is a consequence of (3a) in Section 3.3 as well as (1b) and (6). According to Proposition 1, recalibration under the assumption of label shift leaves the implied performance under the target distribution – sometimes called *discriminatory power* in the credit risk management literature (e.g. Bohn and Stein [3]) – unchanged when compared to the implied performance under the source distribution. In particular, Proposition 1 implies a necessary criterion for distribution shift to be label shift. Accordingly, in case $AUC_S \neq AUC_{S^*}$ source distribution P and target distribution Q cannot be related through label shift.

4.3 Recalibration under assumption of factorizable joint shift (FJS)

According to He at al. [10] and Tasche [25], the source distribution $P(X,Y)$ and the target distribution $Q(X,Y)$ are related through *factorizable joint shift* (FJS) if there are functions $g : \mathcal{X} \to [0,\infty)$ and $b : \mathcal{Y} \to [0,\infty)$ such that $(x,y) \mapsto g(x)\,b(y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ is a density of $Q(X,Y)$ with respect to $P(X,Y)$, i.e. it holds that $Q[(X,Y) \in M] = E_P\big[\mathbf{1}_M(X,Y)\,g(X)\,b(Y)\big]$ for all measurable sets $M \subset \mathcal{X} \times \mathcal{Y}$. See footnote 2 for the definition of the indicator $\mathbf{1}_M$.

By Corollary 4 of Tasche [25] and subject to mild technical conditions, under FJS the joint target distribution $Q(X,Y)$ is given by the target feature distribtuion $Q(X)$ and the class 1 posterior probability

$$\eta_Q(X) \;=\; \frac{\frac{q}{p}\,\eta_P(X)}{\frac{q}{p}\,\eta_P(X) + \frac{1}{\varrho}\,\frac{1-q}{1-p}\,(1-\eta_P(X))}, \tag{7a}$$

where $0 < \frac{p}{(1-p)\,E_Q\left[\frac{\eta_P(X)}{1-\eta_P(X)}\right]} \le \varrho \le \frac{p}{(1-p)}\,E_Q\left[\frac{1-\eta_P(X)}{\eta_P(X)}\right]$ and ϱ is the unique solution to the equation

$$q \;=\; E_Q\left[\frac{\frac{q}{p}\,\eta_P(X)}{\frac{q}{p}\,\eta_P(X) + \frac{1}{\varrho}\,\frac{1-q}{1-p}\,(1-\eta_P(X))}\right]. \tag{7b}$$

Note that recalibration of $\eta_P(X)$ to a target prior probability q under the assumption of FJS works for arbitrary target feature distributions $Q(X)$. This is in stark contrast to recalibration under the assumption of label shift which only works when assuming that $Q(X)$ is given by (5). However, by Proposition 1 recalibration under the label shift assumption entails $AUC_{\eta_Q(X)} = AUC_{\eta_P(X)}$. The example in Section 5 below shows that AUC preservation is not in general true for recalibration under the FJS assumption.

(7a) implies (1b) with $T(\eta) = T_\varrho(\eta) = \frac{\frac{p}{q}\,\eta}{\frac{p}{q}\,\eta + \frac{1}{\varrho}\,\frac{1-p}{1-q}\,(1-\eta)}$ strictly increasing in η. If ϱ happens to take the value 1, at first glance we are back in the context of label shift as in Section 4.2 above. However, as (5) need not hold true under the FJS assumption, in such cases there cannot be label shift. The type of shift modelled instead is called 'invariant density ratio' shift (Tasche [23]), defined as the special case of FJS with $\varrho = 1$.

4.4 Recalibration under assumption of covariate shift with posterior drift (CSPD)

In Sections 4.1, 4.2 and 4.3, we identified the transformation T of (1b) after we had described a recalibration method designed under assumption of one of three types of distribution shift, namely slightly modified covariate shift, label shift, and FJS. In contrast, in this section, we define T and use it to characterise the type of distribution shift implied by the recalibration method.

According to Scott [19], the source distribution $P(X,Y)$ and the target distribution $Q(X,Y)$ are related through *covariate shift with posterior drift* (CSPD) if there is a *strictly increasing transformation* T such that (1b) holds true.

As mentioned in Section 3.1, (1b) implies that $\eta_P(X)$ is sufficient for X with respect to Y under the target distribution Q. Since under CSPD the transformation T is strictly increasing, (1b) also implies that $\eta_Q(X)$ and $\eta_P(X)$ are strongly comonotonic (Tasche [24]). As a consequence, Kendall's τ and Spearman's rank correlation both take the maximum value 1 when applied to $(\eta_Q(X), \eta_P(X))$. This suggests that CSPD is a strong assumption which might be less often true than one would hope for. Nonetheless, comonotonicity of score and posterior probability is a common assumption in the literature. Chen et al. [5] call this assumption the *rationality assumption*.

If labels are available in the test dataset, under the CSPD assumption the transformation T of (1b) can be approximately determined by means of isotonic regression. However, the general assumption for this paper is that there are no label observations for the instances in the test dataset. Therefore, we are going to apply a moment matching approach instead (quasi moment matching, QMM).

To be able to do so, we consider *parametric CSPD* where the transformation T of (1b) is specified as follows through some strictly increasing and continuous distribution function F on the real line and parameters $a, b \in \mathbb{R}$:

$$T_{a,b}(u) \;=\; F\big(a\,F^{-1}(u) + b\big), \qquad \text{for } 0 < u < 1. \tag{8a}$$

This is not a radically new approach but rather a variation of what was called regression-based calibration by Ojeda et al. [14]. Here are two examples for natural choices of F in (8a):

- Logistic distribution function (inverse logit): $F(x) = \frac{1}{1+\exp(-x)}$, $x \in \mathbb{R}$. The parametric CSPD approach with inverse logit is often called 'Platt scaling' in the literature. However, as pointed out by Ojeda et al. [14], 'Platt scaling' sometimes also refers to the specification of T as

$$T_{a,b}(u) \;=\; \frac{1}{1 + \exp(-(a\,u + b))}, \qquad \text{for } 0 < u < 1. \tag{8b}$$

  In the following, we use the term *Platt scaling* to refer to (8b) and *logistic CSPD* to refer to (8a) with inverse logit.
- Standard normal distribution function (inverse probit): $F(x) = \Phi(x)$, $x \in \mathbb{R}$. We refer to this choice of F as *normal CSPD*.

The idea for *quasi-moment matching (QMM)* with parametric CSPD is to determine the parameters $a, b \in \mathbb{R}$ by solving the following equation system:

$$q = E_Q\big[T_{a,b}(\eta_P(X))\big] \quad \text{and} \quad AUC_{\eta_P(X)} = AUC_{T_{a,b}(\eta_P(X))}, \tag{8c}$$

with $T_{a,b}$ as in (8a) or (8b) and AUC defined in Section 3.3. More precisely, in (8c), $AUC_{\eta_P(X)}$ is computed with respect to $P(X, Y)$ while $AUC_{T_{a,b}(\eta_P(X))}$ is computed as implied AUC with respect to $Q(X)$. See Section 3.3 for the different ways to compute AUC.

In (8c), AUC works like a second moment of the posterior probabilities. But in contrast to its effect on the second moment or the variance, the prior probability of the positive class has no effect on AUC. This makes the assumption of AUC invariance between source and target distributions more plausible (Tasche [22]).

Remark 1. One-parameter CSPD as in (8a) with $b = 0$ can be used for class distribution estimation (also called quantification), i.e. to determine an unknown class prior probability $q = Q[Y = 1]$ under the target distribution. In this case, instead of solving (8c) for two parameters a and b, the following equation is solved for parameter a only:

$$AUC_{\eta_P(X)} \;=\; AUC_{T_{a,0}(\eta_P(X))}. \tag{9a}$$

Then the mean of the resulting posterior probability $\eta_Q(X) = T_{a,0}(\eta_P(X))$ under the target distribution feature distribution $Q(X)$ is computed to obtain an estimate $\widehat{q}$ of q:

$$\widehat{q} \;=\; E_Q\big[T_{a,0}(\eta_P(X))\big]. \tag{9b}$$

One-parameter CSPD may be interpreted as a modification of quantification under the assumption of covariate shift. In contrast to assuming covariate shift when AUC can differ between source distribution and target distribution, with one-parameter CSPD by construction source AUC and target AUC are equal.

4.5 Quasi moment matching based on parametrised receiver operating characteristics

There is no guarantee that the QMM approach presented in Section 4.4 is feasible in the sense that there exists a solution to equation system (8c) or that the solution is unique. This reservation motivates the following alternative approach where instead of beginning with representation (8a) of the posterior probabilities, the starting point is a parametrised representation of the receiver operating characteristic (ROC) curve associated with the target distribution $Q(X, Y)$.

Tasche ([22], Section 5.2) modified an idea of van der Burgt [4] by assuming the ROC curve of a real-valued score S to be

$$ROC_S(u) \;=\; \Phi\big(c + \Phi^{-1}(u)\big), \qquad u \in (0, 1), \tag{10a}$$

for some fixed parameter $c \in \mathbb{R}$, with Φ denoting the standard normal distribution function. The ROC curve of (10a) emerges when the class-conditional score distributions in a binary classification problem are both univariate normal distributions with equal variances. But ROC curves like in (10a) may also be incurred in circumstances where the class-conditional distributions are not normal (Proposition 5.3 of Tasche [22]).

Tasche [22] showed that (10a) implies the following representation for the posterior probability given the score S under the target distribution Q:

$$Q[Y = 1 \,|\, S] \;=\; \frac{1}{1 + \frac{1-q}{q}\,\exp\big(c^2/2 - c\,\Phi^{-1}(F_0(S))\big)}, \tag{10b}$$

where $F_0(s) = Q[S \le s \,|\, Y = 0]$ stands for the class-conditional distribution function of S given $Y = 0$.

For AUC_S as defined in Section 3.3, (10a) implies $AUC_S = \Phi\big(\frac{c}{\sqrt{2}}\big)$. Hence, if AUC_S is known the parameter c of (10b) is determined by

$$c \;=\; \sqrt{2}\,\Phi^{-1}(AUC_S). \tag{10c}$$

The score S in (10b) can be chosen as $\eta_P(X)$ or any probabilistic classifier which approximates $\eta_P(X)$. The target prior probability q is known by general assumption for this paper. Similarly to Section 4.4, for QMM to work one has to make the assumption that a prudent choice of AUC_S under the target distribution (defined as implied AUC by (3b) and (3a) above) is informed by AUC_S observed in the training dataset such that by (10c) also parameter c is known.

However, the distribution function F_0 of the score S conditional on $Y = 0$ appearing in (10b) is assumed not to be known under the target distribution since by general assumption for this paper, only the features but not the labels can be observed in the test dataset. Making use of (10b) therefore requires an iterative approach where in each step a refined estimate of the negative class-conditional score distribution function F_0 is calculated.

Denote by f an unconditional density of S under Q and by f_0 a density of F_0. Then the posterior probability $Q[Y = 1 \,|\, S = s]$ can be represented as

$$Q[Y = 1 \,|\, S = s] \;=\; 1 - \frac{(1-q)\, f_0(s)}{f(s)} \tag{11a}$$

for s in the range $\mathcal{S}$ of S. By (10b), (11a) implies

$$f_0(s) \;=\; \frac{f(s)}{1-q} \left(1 - \frac{1}{1 + \frac{1-q}{q}\, \exp\bigl(c^2/2 - c\, \Phi^{-1}(F_0(s))\bigr)} \right), \quad s \in \mathcal{S}. \tag{11b}$$

In the case of scores S under a discrete or empirical distribution, as described in Appendices A.1 and A.2, (11b) can be treated as a fixed point equation for the probabilities $Q[S = s \,|\, Y = 0] = f_0(s)$ and be solved by a straightforward fixed-point iteration with intial values $Q[S = s]$, $s \in \mathcal{S}$. The numerical example of Section 5 below suggests that such an iterative approach converges as long as the prior probability q is small and, hence, f and f_0 are close to each other.

A further issue when making (10b) operational may occur when the class-conditional distribution function F_0 or any function approximating it takes the value 1. In particular, this will happen if F_0 is approximated by an empirical distribution function. Then the term $\Phi^{-1}\bigl(F_0(s)\bigr)$ is ill-defined. See Appendix A.2 for a workaround to deal with this issue.

Deploying a discrete distribution F_0 in (10b) as for instance an empirical approximation, makes it unlikely if not impossible to exactly match a pre-defined AUC_S when using (10b). To control for the unavoidable deviation from the AUC_S objective in this case we take recourse to the original QMM approach as proposed by Tasche [22]. Define

$$T^*_{a,b}(S) \;=\; \frac{1}{1 + \exp\bigl(b + a\, \Phi^{-1}(F_0(S))\bigr)}, \tag{12a}$$

with F_0 as in (10b) and parameters $a, b \in \mathbb{R}$ to be determined by quasi-moment matching (QMM) like in (8c):

$$q = E_Q\bigl[T^*_{a,b}(S)\bigr] \quad \text{and} \quad AUC_S = AUC_{T^*_{a,b}(S)}, \tag{12b}$$

To distinguish the two recalibration approaches presented in this section, we call the approach based on (10b) and (10c) *ROC-based QMM* and refer to the approach based on (12a) and (12b) as *2-parameter QMM*.

5 Example

The example presented in this section illustrates the impact of the recalibration methods and assumptions discussed in Section 4 on the following characteristics of the target distribution:

- The class 1 posterior probabilities.
- The mean of the class 1 posterior probabilities (which ought to equal the class 1 prior probability).
- The AUC implied by the class 1 posterior probabilities.
- The mean of the square root of the class 1 posterior probabilities as an expample involving a concave function of the posterior probabilities.

For the example we chose discrete source and target distributions of a feature (called score in the following) with values in an ordered set with 17 elements. These distributions may be interpreted as empirical distributions of samples with many ties or as the genuine distributions of discrete scores or ratings. For instance, the major credit rating agencies Standard & Poor's, Moody's and Fitch use rating scales with 17 to 19 different grades.

The source distribution is specified as follows: (1) The conditional feature distribution for class 0 is a binomial distribution with success probability 0.4, the conditional feature distribution for class 1 is a binomial distribution with success probability 0.55. The number of trials for both binomial distributions is 16 such that the support of the distribution is the set $\{0, 1, \ldots, 16\}$. (2) The class 1 prior probability is $p = 0.01$.

The target distribution is incompletely specified as follows: (1) The unconditional feature distribution is a binomial distribution with number of trials 16 whose success probability is a Vasicek-distributed random variable (Section 2.2 of Meyer [13]) with mean 0.3 and correlation parameter 0.3. (2) The target class 1 prior probability is $q = 0.05$.

Figure 1 shows[3] the source and target unconditional score distributions. They were intentionally chosen to be quite different such that any distribution shift assumed for the recalibration must be significant.

Figure 2 presents the source posterior probabilities and eight different sets of target posterior probabilities that have been calculated with the methods and assumptions discussed in Section 4:

- Capped scaling: Section 4.1
- Label shift: Section 4.2
- FJS: Factorizable joint shift, Section 4.3
- Platt scaling: Section 4.4, Eq. (8b)
- ROC QMM: ROC-based QMM, Section 4.5, Eq. (10b) and Eq. (10c)
- 2-param QMM: 2-parameter QMM, Section 4.5, Eq. (12a) and Eq. (12b)
- Logistic CSPD: Section 4.4, Eq. (8a) with $F(x) = \frac{1}{1+\exp(-x)}$
- Normal CSPD: Section 4.4, Eq. (8a) with $F(x) = \Phi(x)$, the standard normal distribution function

[3] Calculations were performed with R (R Core Team [17]). Details of some more involved calculations are described in Appendices A.1 and A.2. The R-scripts can be downloaded from https://www.researchgate.net/profile/Dirk-Tasche.

Fig. 1. Source and target score distributions for the example in Section 5.

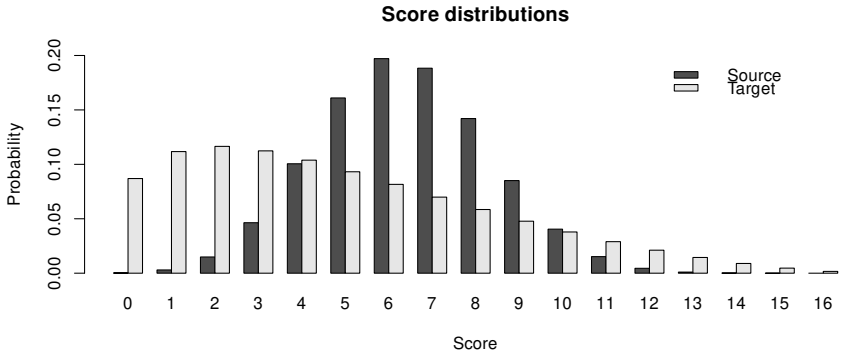

Fig. 2. Class 1 posterior probabilities for the example in Section 5. The scale of the vertical axis is logarithmic. The dots for the probabilities have been connected with straight lines for better readability. 'Source' refers to the posterior probabilities in the source distribution without recalibration.

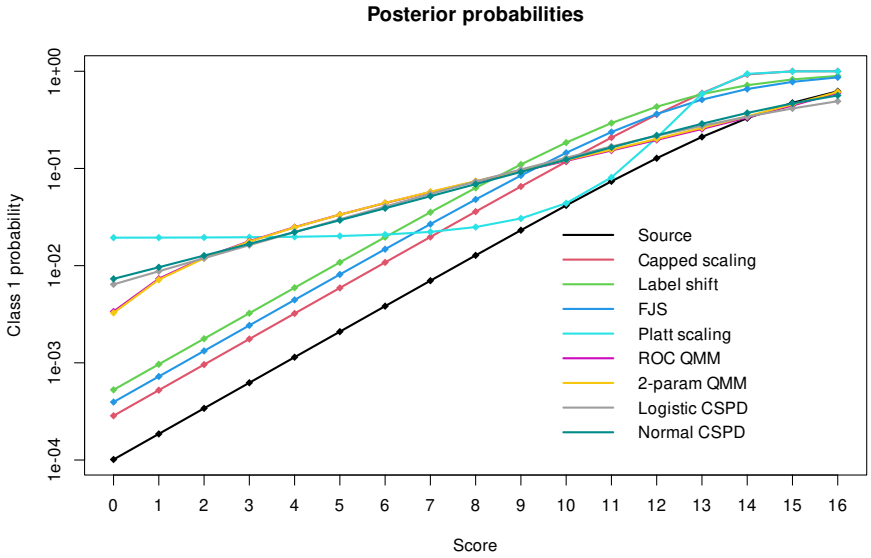

Table 1. Results for the recalibration methods presented in Section 4. The row 'Source' shows the values for the source distribution without recalibration. The numbers in all other rows refer to the target distribution. 'mean(probs)' is the class 1 prior probability calculated as mean of the recalibrated posterior probabilities. 'AUC' is the area under the ROC curve implied by the recalibrated posterior probabilities. 'mean(sqrt(probs))' is the mean of the square root of the recalibrated posterior probabilities.

Method	mean(probs)	AUC	mean(sqrt(probs))
Source	0.010	0.802	0.084
Capped scaling	0.050	0.950	0.132
Label shift	0.060	0.930	0.160
FJS	0.050	0.932	0.142
Platt scaling	0.050	0.802	0.179
ROC QMM	0.049	0.799	0.191
2-param QMM	0.050	0.802	0.191
Logistic CSPD	0.050	0.803	0.192
Normal CSPD	0.050	0.802	0.192

All target posterior probabilities are well above the source posterior probabilities. This does not come as a surprise given that the target class 1 prior probability of 5% is much higher than the source class 1 prior probability of 1%. Otherwise, there are two groups of target posterior probability curves: The curves based on capped scaling, label shift and FJS on the one hand, and the curves based on the five QMM methods described in Sections 4.4 and 4.5 on the other hand. The curves of the former group are rather steep compared to the curves of the latter group.

Table 1 displays three characteristics for the source distribution as well as for the eight different target distributions that result from the recalibration methods discussed in Section 4. Concluding from the entries in the column 'mean(probs)', only the recalibration method 'label shift' – introduced in Section 4.2 – is unreliable in so far as it does not achieve the required target class 1 prior probability $p = 0.05$. With all seven other recalibration methods the target prior probability is matched.

Column 'AUC' of Table 1 is more varied than column 'mean(probs)'. For the five QMM methods from Sections 4.4 and 4.5, AUC is essentially the same as the AUC of 0.802 under the source distribution. This is a consequence of the QMM design since one of the moment matching objectives is hitting the source AUC. Thanks to Proposition 1, one might expect also the 'label shift' AUC to match the source AUC. However, Proposition 1 is not applicable to the example of this section because by design the target feature distribution is not a mixture of the source class-conditional distributions. In any case, the high AUC values for methods 'capped scaling', 'label shift' and 'FJS' cause the greater slopes of their posterior probability curves in Figure 2 compared to the slopes of the QMM posterior probability curves.

As an example for the application of a concave function to the target posterior probabilities, we chose the function $C(u) = \sqrt{u}$, $u \in [0, 1]$, and the related

expected value $E_Q\left[C(\eta_Q(X))\right]$ under the target distribution. Note that (2) implies $0.05 = q \leq E_Q\left[\sqrt{\eta_Q(X)}\right] \leq \sqrt{q} \approx 0.2236$.

It is clear from column 'mean(sqrt(probs))' of Table 1 that AUC is a driver of the mean of the concave function of the posterior probabilities. The lower the value of AUC, the higher is the mean of the concave function of the probabilities. Hence it makes sense to have AUC as a second objective to be matched, in addition to requiring that the target class 1 prior probability is reached. But even if both of these conditions are met there can still be variation in the mean of the concave function. This is demonstrated by the 'Platt scaling' posterior probabilities for which $E_Q\left[\sqrt{\eta_Q(X)}\right]$ takes a notedly lower value than for the other four QMM posterior probabilities with almost identical values.

6 Conclusions

Recalibration of binary probabilistic classifiers to a target prior probability is an important task in areas like credit risk management. This paper presents analyses and methods for recalibration from a distribution shift perspective. It has turned out that distribution shift assumptions linked to the performance in terms of area under the curve (AUC) of a probabilistic classifier under the source distribution are useful for the design of meaningful recalibration methods. Two new methods called parametric CSPD (covariate shift with posterior drift) and ROC-based QMM (quasi moment matching) have been proposed and have been tested together with some other methods in an example setting. The outcomes of this testing exercise suggest that the QMM methods discussed in the paper can provide appropriately conservative results in evaluations with concave functions like for instance risk weights functions for credit risk.

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

References

1. Allikivi, M.L., Järve, J., Kull, M.: Cautious Calibration in Binary Classification. Frontiers in Artificial Intelligence and Applications, vol. 392, pp. 1503–1510 (2024). https://doi.org/10.3233/FAIA240654
2. BCBS: CRE – Calculation of RWA for credit risk. Basel Committee on Banking Supervision, https://www.bis.org/basel_framework/standard/CRE.htm, Regulatory Standard
3. Bohn, J., Stein, R.: Active Credit Portfolio Management in Practice. John Wiley & Sons, Inc. (2009)
4. van der Burgt, M.: Calibrating low-default portfolios, using the cumulative accuracy profile. Journal of Risk Model Validation **1**(4), 17–33 (2008)
5. Chen, W., Sahiner, B., Samuelson, F., Pezeshk, A., Petrick, N.: Calibration of medical diagnostic classifier scores to the probability of disease. Statistical methods in medical research **27**(5) (2018). https://doi.org/10.1177/0962280216661371
6. Cramer, J.: Logit Models From Economics and Other Fields. Cambridge University Press (2003)

7. Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Springer (1996)
8. Esuli, A., Fabris, A., Moreo, A., Sebastiani, F.: Learning to Quantify. Springer Cham (2023). https://doi.org/10.1007/978-3-031-20467-8
9. Fawcett, T.: An introduction to ROC analysis. Pattern Recognit. Lett. **27**(8), 861–874 (2006). https://doi.org/10.1016/J.PATREC.2005.10.010
10. He, H., Yang, Y., Wang, H.: Domain Adaptation with Factorizable Joint Shift. Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning (2021). https://doi.org/10.48550/ARXIV.2203.02902
11. Lalley, S.: Concentration inequalities (2013), https://galton.uchicago.edu/~lalley/Courses/386/index.html, lecture notes
12. 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)
13. Meyer, C.: Estimation of intra-sector asset correlations. The Journal of Risk Model Validation **3**(3), 47–79 (2009)
14. Ojeda, F., Jansen, M., Thiéry, A., Blankenberg, S., Weimar, C., Schmid, M., Ziegler, A.: Calibrating machine learning approaches for probability estimation: A comprehensive comparison. Statistics in Medicine **42**(29), 5451–5478 (2023). https://doi.org/10.1002/sim.9921
15. Ptak-Chmielewska, A., Kopciuszewski, P.: New Definition of Default Recalibration of Credit Risk Models Using Bayesian Approach. Risks **10**(1) (2022), https://www.mdpi.com/2227-9091/10/1/16
16. Quadrianto, N., Smola, A., Caetano, T., Le, Q.: Estimating Labels from Label Proportions. Journal of Machine Learning Research **10**(82), 2349–2374 (2009), http://jmlr.org/papers/v10/quadrianto09a.html
17. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2024), https://www.R-project.org/
18. 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
19. Scott, C.: A Generalized Neyman-Pearson Criterion for Optimal Domain Adaptation. In: Proceedings of Machine Learning Research, 30th International Conference on Algorithmic Learning Theory. vol. 98, pp. 1–24 (2019)
20. Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., Flach, P.: Classifier calibration: a survey on how to assess and improve predicted class probabilities. Machine Learning **112**(9), 3211–3260 (2023). https://doi.org/10.1007/s10994-023-06336-7
21. 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)
22. Tasche, D.: Estimating discriminatory power and PD curves when the number of defaults is small. Working paper (2009), http://arxiv.org/abs/0905.3928
23. Tasche, D.: Fisher Consistency for Prior Probability Shift. Journal of Machine Learning Research **18**(95), 1–32 (2017)
24. Tasche, D.: Calibrating sufficiently. Statistics **55**(6), 1356–1386 (2021). https://doi.org/10.1080/02331888.2021.2016767
25. Tasche, D.: Factorizable Joint Shift in Multinomial Classification. Machine Learning and Knowledge Extraction **4**(3), 779–802 (2022). https://doi.org/10.3390/make4030038

26. Tsyplakov, A.: Evaluation of Probabilistic Forecasts: Proper Scoring Rules and Moments (2013), https://mpra.ub.uni-muenchen.de/45186/
27. Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., Schön, T.: Evaluating model calibration in classification. In: The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. vol. 89, pp. 3459–3467. PMLR (2019)

A Appendix

A.1 The AUC of discrete auto-calibrated probabilitistic classifiers

How to calculate implied AUC for an auto-calibrated[4] probabilistic classifier S under a joint distribution $\mu(S, Y)$ of S and $Y \in \mathcal{Y} = \{0, 1\}$? We assume here that the distribution of S under distribution μ is discrete, i.e. it is is given by pairs of score values s_i and their probabilities $\mu[S = s_i] = \pi_i > 0$, $i = 1, \ldots, n$. Then it follows from the AUC-definition and properties in Section 3.3 that

$$AUC_S \;=\; \frac{\sum_{i=2}^{n} \pi_i \, s_i \Big(\frac{1}{2}\, \pi_i (1 - s_i) + \sum_{k=1}^{i-1} \pi_k \, (1 - s_k) \Big)}{\Big(1 - \sum_{j=1}^{n} \pi_j \, s_j \Big) \sum_{j=1}^{n} \pi_j \, s_j} \tag{13}$$

We use (13) to compute imlied AUC both under the source distribution for $S = \eta_P(X)$ with $\mu = P$ and under the target distribtion for $S = \eta_Q(X) = T\big(\eta_P(X)\big)$ with $\mu = Q$. Note that $\eta_P(X)$ under P and $\eta_Q(X)$ under Q are both auto-calibrated probabilistic classifiers (Proposition 1 of Vaicenavicius et al. [27]).

A.2 Adapting discrete distributions for QMM

Consider for the real-valued score or probabilistic classifier S with discrete distribution $\mu[S = s_i] = \pi_i$, $i = 1, \ldots, n$, its distribution function $G(s) = \mu[S \le s]$ for $s \in \mathbb{R}$. If the s_i are increasingly ordered with $s_1 < \ldots < s_n$, then we obtain $G(s_i) = \sum_{j=1}^{i} \pi_j$, for $i = 1, \ldots, n$. In particular, it follows $G(s_n) = 1$ such that $\Phi^{-1}\big(G(s_n)\big) = \infty$ for the standard normal distribution function Φ and for the logistic distribution function.

For the purpose of this paper, to avoid this problem we apply the workaround proposed by van der Burgt [4]. He suggested replacing the distribution function G with the mean of itself and its left-continuous version, i.e. with G^* defined by $G^*(s) = \frac{1}{2}\big(\mu[S \le s] + \mu[S < s]\big)$, for $s \in \mathbb{R}$.

This implies for the s_i which represent the support of μ that $G^*(s_i) = \Big(\sum_{j=1}^{i} \pi_j\Big) - \pi_i/2$, for $i = 1, \ldots, n$.

[4] The probabilistic classifier S is *auto-calibrated* if for all s in the range of S it holds that $\mu[Y = 1 \,|\, S = s] = s$ (Tsyplakov [26], Section 2.2).