Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Bayesian Prediction of Future Empirical ROC Curves and AUC

Olivier Caelen

olivier.caelen@uclouvain.be

UCLouvain

Gabriele Iannucci

Sapienza University of Rome

Research Article

Keywords: Receiver operating characteristic, Bayesian inference, Posterior predictive distribution, Dirichlet–multinomial model, Simultaneous predictive bands, Area under the curve

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11270016/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Bayesian Prediction of Future Empirical ROC Curves and AUC

Olivier Caelen[1*] and Gabriele Iannucci[2]

[1]LIDAM, UCLouvain, Voie du Roman Pays 20, Louvain-la-Neuve, 1348, Belgium (ORCID: 0000-0001-6970-9825).
[2]Department of Statistical Sciences, Sapienza University of Rome, Piazzale Aldo Moro 5, Rome, 00185, Italy.

*Corresponding author(s). E-mail(s): olivier.caelen@uclouvain.be;

Abstract

The measured performance of a classifier can change when it is evaluated on a new test sample, even when the classifier itself remains fixed. We propose DiP-ROC (Dirichlet predictive ROC), a Bayesian framework for predicting this variation from a single test sample. It predicts the receiver operating characteristic (ROC) curve and its area under the curve (AUC) that a new sample from the same populations could produce, with the same numbers of positive and negative observations. DiP-ROC groups prediction scores into bins defined by a fixed threshold grid and uses a Dirichlet–multinomial model for each class. Together, these models describe the joint uncertainty in the future ROC points and their corresponding grid-based AUC. They preserve the dependence between ROC points, which are calculated from the same observations. They also account for both uncertainty about the score distributions and the variation caused by drawing a new test sample. The resulting summaries are simultaneous predictive bands for all ROC points on the grid and predictive intervals for the AUC. Prior information can come from an earlier evaluation of the same classifier. This information may be a reported AUC value or two score histograms, one for each class. Experiments on synthetic data and nine real-world datasets show how prior strength and class counts affect the reliability and precision of predictions for a future evaluation. All code is available on GitHub to support reproducibility: https://github.com/ocaelen/bayesian_roc.

Keywords: Receiver operating characteristic, Bayesian inference, Posterior predictive distribution, Dirichlet–multinomial model, Simultaneous predictive bands, Area under the curve

1 Introduction

The receiver operating characteristic (ROC) curve describes how a binary classifier distinguishes positive from negative observations as its decision threshold changes. It shows the trade-off between the proportion of positive observations correctly identified and the proportion of negative observations incorrectly classified as positive. The area under the curve (AUC) summarizes this ability in a single number. When estimated from a test sample, both the ROC curve and the AUC depend on the observations used. A new evaluation can therefore produce different results, even when the trained classifier remains unchanged.

This variation raises a practical question: how much could the reported performance change if the same classifier were evaluated on a new sample? For example, suppose a classifier has been tested on 100 positive and 500 negative observations. An analyst plans a second evaluation on an independent sample from the same populations, with the same class counts and without retraining the classifier. Before collecting these new observations, the analyst would like to predict the range of ROC curves and AUC values that this second evaluation might produce.

We propose DiP-ROC (Dirichlet predictive ROC), a Bayesian framework that uses the scores and labels from one test sample to predict the results of an independent evaluation of the same fixed classifier. We assume that the future sample comes from the same populations and contains the same numbers of positive and negative observations. Both evaluations use the same threshold grid, chosen in advance.

DiP-ROC combines two sources of uncertainty: limited knowledge of the score distributions in each class, and variation caused by drawing a new test sample. It describes these sources through a joint posterior predictive distribution for the future ROC points and their corresponding AUC. This joint construction keeps the ROC points linked across thresholds, since they are all calculated from the same observations.

From this distribution, DiP-ROC provides simultaneous predictive bands for the ROC points and predictive intervals for the AUC. Each band describes uncertainty for all points on the chosen grid together. The AUC interval describes uncertainty in the area under the curve formed by those points. These summaries help analysts assess how much performance may change in a future evaluation before collecting the new observations.

2 Related Work

Methods for quantifying ROC uncertainty differ in what they describe and what they aim to predict or estimate. Some focus on the AUC alone, while others describe uncertainty across the ROC curve. They also differ in whether they estimate the classifier's underlying performance in the population or predict the empirical ROC curve and AUC obtained from a new test sample.

Frequentist methods commonly address uncertainty in population AUCs or ROC curves. DeLong et al. [1] account for dependence between AUC estimates when comparing classifiers evaluated on the same observations. Moving from AUCs to curves, Bertail et al. [2] construct confidence bands using a smoothed bootstrap, which repeatedly generates new samples from smoothed estimates of the score distributions and

measures how their ROC curves vary. Martínez-Camblor et al. [3] also develop nonparametric confidence bands for the population ROC curve, with widths that adapt to the estimated uncertainty along the curve. These methods estimate underlying performance from the available sample.

Macskassy et al. [4] explicitly consider a future evaluation of a fixed classifier. They compare simultaneous bands for the population ROC curve and wider "future-curve" bands intended to contain an empirical ROC curve from another test sample of the same total size. They obtain these future-curve bands by widening bootstrap-based fixed-width bands by a factor of $\sqrt{2}$, a heuristic adjustment for the additional sampling variation. DiP-ROC shares this predictive aim, deriving bands directly from a Bayesian model of future samples with fixed positive and negative sample sizes on a prespecified threshold grid. Other work focuses on individual thresholds. Urlus et al. [5] compare frequentist and Bayesian uncertainty regions for individual precision–recall and ROC points, each representing two performance measures at a single threshold; these remain pointwise rather than simultaneous across thresholds. Zheng et al. [6] construct conformal ROC bands that address differences between a fitted scoring function and the true one. Their reported coverage guarantee is pointwise and asymptotic, meaning that it applies at individual thresholds as sample sizes increase; it requires an estimate that approaches the true scoring function as the training sample grows, together with additional regularity conditions. Its target is oracle sensitivity, the true-positive rate that would be obtained on the test observations using the unknown true scoring function, with specificity (the true-negative rate) handled analogously. DiP-ROC instead predicts the empirical ROC curve of the fitted classifier on a new test sample.

Bayesian ROC methods describe uncertainty through posterior distributions, which combine prior information with observed data. Gu et al. [7] use the Bayesian bootstrap, assigning random Dirichlet weights that vary each observation's contribution to the class-specific score distributions, from which they obtain ROC curves and AUCs. They also construct simultaneous credible bands for the population ROC curve. Erkanli et al. [8] use flexible Dirichlet-process mixture models, which combine normal distributions to represent varied score patterns in each class, and derive posterior ROC summaries. For scores taking a finite set of ordered values, Al-Labadi et al. [9] use independent Dirichlet priors to model the probabilities of those values within each class. These summaries concern population ROC performance rather than empirical ROC curves from finite future test samples. DiP-ROC's fixed-grid count model also permits direct simulation of future curves, avoiding the iterative mixture fitting used by Erkanli et al.

A related line of work starts from the confusion matrix, which counts correct and incorrect predictions within each class at a fixed decision threshold. Caelen [10] uses a Dirichlet–multinomial model to simulate possible future confusion matrices and obtain predictive distributions for performance metrics computed from them. Tötsch and Hoffmann [11] distinguish uncertainty about underlying classifier performance from the additional variation in measured performance caused by drawing a new finite test sample. Shang et al. [12] study test samples in which predicted positives are deliberately overrepresented, correcting recall for this sampling design. They

derive approximate confidence intervals for precision and recall and posterior predictive intervals for these estimates on future test samples. Lovell et al. [13] model future confusion matrices and show how finite class counts shape the discrete distributions of the resulting performance estimates, which can take only finitely many values. These approaches quantify uncertainty at a fixed decision rule, rather than jointly across the thresholds of an ROC curve.

DiP-ROC connects these ideas to predict a future empirical ROC curve on a threshold grid chosen before observing the test data. It uses separate Dirichlet–multinomial models for the score counts in the two classes and simulates future samples from the same populations, with the same positive and negative counts as the observed test sample. Each simulated sample produces a complete curve on the grid, preserving dependence across thresholds. The maximum deviation of each sampled curve is used to construct a simultaneous posterior predictive band. The corresponding AUC values provide a separate predictive interval for the grid-based AUC.

3 Formalization of the learning problem

We consider a binary classification setting in which an observation is represented by $(\boldsymbol{X}, Y) \sim F_{\boldsymbol{X},Y}$, where $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^d$ is a random input vector and $Y \in \mathcal{Y} := \{0, 1\}$ is a binary random label. It is common to refer to $\{Y = 1\}$ as the positive case, and $\{Y = 0\}$ as the negative case. Let $\pi := \mathbb{P}(Y = 1)$ denote the prevalence of the positive class. We assume $0 < \pi < 1$, meaning that both classes are present in the population.

Note that throughout the paper, the typography distinguishes random objects from their realizations and scalar quantities from vector-valued ones. Scalar random variables are denoted by non-bold capital letters, such as Y, whereas their realizations, or deterministic scalar quantities, are denoted by non-bold lowercase letters, such as y. Vector- or matrix-valued random variables are denoted by bold capital letters, such as $\boldsymbol{X}$, and their realizations or deterministic counterparts by bold lowercase letters, such as $\boldsymbol{x}$. All vectors are treated as column vectors.

We define $\mathfrak{d}^{\mathrm{tr}} := ((\boldsymbol{x}_i^{\mathrm{tr}}, y_i^{\mathrm{tr}}))_{i=1}^{n^{\mathrm{tr}}}$ as a training dataset, where the pairs $(\boldsymbol{x}_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})$ are the observed realizations of an i.i.d. sample from $F_{\boldsymbol{X},Y}$, and n^{tr} denotes the training sample size. The dataset $\mathfrak{d}^{\mathrm{tr}}$ is used to learn the parameters of a probabilistic classifier $g : \mathcal{X} \to [0, 1]$ that returns a score $W := g(\boldsymbol{X}) \in [0, 1]$ which is intended to estimate the conditional probability $\mathbb{P}(Y = 1 \mid \boldsymbol{X})$. After training on $\mathfrak{d}^{\mathrm{tr}}$, we fix the parameters of the model, so that we can assume that the probabilistic classifier g is a deterministic function. For any decision threshold $\tau \in \mathbb{R}$, we define the binary classifier $h_\tau : \mathcal{X} \to \{0, 1\}$ as $h_\tau(\boldsymbol{x}) := \mathbb{1}\{g(\boldsymbol{x}) \geq \tau\}$, where $\mathbb{1}\{\cdot\}$ denotes the indicator function.

For a fixed threshold τ, we can define the class-conditional tail probabilities $p_\tau^+ := \mathbb{P}(W \geq \tau \mid Y = 1)$ and $p_\tau^- := \mathbb{P}(W \geq \tau \mid Y = 0)$. The quantity p_τ^+ is the population true positive rate $\mathrm{tpr}(\tau)$ and represents the probability that a positive observation is assigned by h_τ to the positive class. On the other hand, p_τ^- is the population false positive rate $\mathrm{fpr}(\tau)$ and represents the probability that a negative observation is assigned to the positive class.

Let $\mathcal{D}^{\mathrm{ts}} := ((\boldsymbol{X}_i^{\mathrm{ts}}, Y_i^{\mathrm{ts}}))_{i=1}^{n^{\mathrm{ts}}}$ be a random test sample, where $(\boldsymbol{X}_i^{\mathrm{ts}}, Y_i^{\mathrm{ts}}) \overset{\text{i.i.d.}}{\sim} F_{\boldsymbol{X},Y}$ and n^{ts} is the test sample size. We define $N_+^{\mathrm{ts}} := \sum_{i=1}^{n^{\mathrm{ts}}} \mathbb{1}\{Y_i^{\mathrm{ts}} = 1\}$ and

$N_-^{\text{ts}} := \sum_{i=1}^{n^{\text{ts}}} \mathbb{1}\{Y_i^{\text{ts}} = 0\}$ as the numbers of positive and negative test observations, respectively, where $N_+^{\text{ts}} + N_-^{\text{ts}} = n^{\text{ts}}$.

It is common to evaluate the performance of h_τ by computing a confusion matrix on the test dataset that can be represented as a random vector:

$$\boldsymbol{V}_\tau := (\#\text{TP}(\tau), \#\text{TN}(\tau), \#\text{FP}(\tau), \#\text{FN}(\tau))^\top,$$

where:

- $\#\text{TP}(\tau) := \sum_{i=1}^{n^{\text{ts}}} \mathbb{1}\{Y_i^{\text{ts}} = 1,\, W_i^{\text{ts}} \geq \tau\} \equiv$ number of true positives,
- $\#\text{TN}(\tau) := \sum_{i=1}^{n^{\text{ts}}} \mathbb{1}\{Y_i^{\text{ts}} = 0,\, W_i^{\text{ts}} < \tau\} \equiv$ number of true negatives,
- $\#\text{FP}(\tau) := \sum_{i=1}^{n^{\text{ts}}} \mathbb{1}\{Y_i^{\text{ts}} = 0,\, W_i^{\text{ts}} \geq \tau\} \equiv$ number of false positives,
- $\#\text{FN}(\tau) := \sum_{i=1}^{n^{\text{ts}}} \mathbb{1}\{Y_i^{\text{ts}} = 1,\, W_i^{\text{ts}} < \tau\} \equiv$ number of false negatives,

with $W_i^{\text{ts}} := g(\boldsymbol{X}_i^{\text{ts}})$. As all the elements in $\boldsymbol{V}_\tau$ are non-negative integers and as the sum of the entries is equal to n^{ts}, we have $\boldsymbol{V}_\tau \in \left\{\boldsymbol{v} \in \mathbb{N}^4 : \mathbf{1}_4^\top \boldsymbol{v} = n^{\text{ts}}\right\}$ where $\mathbb{N} = \{0, 1, 2, \ldots\}$ and $\mathbf{1}_q$ denotes a vector of q ones. This vector $\boldsymbol{V}_\tau$ can be modeled as a random vector following a multinomial distribution $\boldsymbol{V}_\tau \sim \text{Multinomial}\,(n^{\text{ts}}, \boldsymbol{\theta}_\tau)$ where the vector

$$\boldsymbol{\theta}_\tau = \left(\theta_\tau^{tp},\ \theta_\tau^{tn},\ \theta_\tau^{fp},\ \theta_\tau^{fn}\right)^\top := \left(\pi p_\tau^+,\ (1-\pi)(1-p_\tau^-),\ (1-\pi)p_\tau^-,\ \pi(1-p_\tau^+)\right)^\top$$

contains the probabilities of the four mutually exclusive outcomes $(\text{TP}, \text{TN}, \text{FP}, \text{FN})$ for a single test observation [10].

Many one-dimensional performance metrics can be computed from the confusion matrix (each ratio is defined when its denominator is positive):

- True positive rate (also known as recall): $\widehat{\text{TPR}}(\tau) := \frac{\#\text{TP}(\tau)}{\#\text{TP}(\tau)+\#\text{FN}(\tau)} = \frac{\#\text{TP}(\tau)}{N_+^{\text{ts}}}$.
- False positive rate: $\widehat{\text{FPR}}(\tau) := \frac{\#\text{FP}(\tau)}{\#\text{FP}(\tau)+\#\text{TN}(\tau)} = \frac{\#\text{FP}(\tau)}{N_-^{\text{ts}}}$.
- Accuracy: $\widehat{\text{ACC}}(\tau) := \frac{\#\text{TP}(\tau)+\#\text{TN}(\tau)}{n^{\text{ts}}}$.
- Precision: $\widehat{\text{PREC}}(\tau) := \frac{\#\text{TP}(\tau)}{\#\text{TP}(\tau)+\#\text{FP}(\tau)}$.
- F1-score: $\widehat{\text{F1}}(\tau) := \frac{2\cdot\#\text{TP}(\tau)}{2\cdot\#\text{TP}(\tau)+\#\text{FP}(\tau)+\#\text{FN}(\tau)}$.

All the performance metrics defined above depend on the choice of the decision threshold τ. Consequently, they characterize the performance of a specific threshold-based classifier h_τ rather than the overall performance of the probabilistic classifier g. To evaluate the predictive quality of g over the entire range of possible decision thresholds, it is common to study the evolution of one or several threshold-dependent performance metrics as functions of τ. Several evaluation tools follow this principle, including the empirical receiver operating characteristic (ROC) curve [14–16], which represents $\widehat{\text{TPR}}(\tau)$ as a function of $\widehat{\text{FPR}}(\tau)$, and the precision–recall curve [15], which represents $\widehat{\text{PREC}}(\tau)$ as a function of $\widehat{\text{RECALL}}(\tau)$ (where $\widehat{\text{RECALL}}(\tau) := \widehat{\text{TPR}}(\tau)$). In this paper, we focus on the ROC curve. A natural extension of this work could be the application of the same principles to other threshold-dependent performance curves such as the precision-recall curve. This is not treated in the current paper.

4 A Discretized Representation of the ROC Curve

If $\tau > 1$, then h_τ classifies every observation as negative. The model h_τ thus yields N_-^{ts} true negatives and N_+^{ts} false negatives and consequently $\boldsymbol{V}_\tau = \big(0,\, N_-^{\text{ts}},\, 0,\, N_+^{\text{ts}}\big)^\top$. If $\tau \leq 0$, then h_τ classifies every observation as positive and consequently $\boldsymbol{V}_\tau = \big(N_+^{\text{ts}},\, 0,\, N_-^{\text{ts}},\, 0\big)^\top$. As the threshold τ decreases from 1 to 0, progressively more observations are predicted by h_τ to be in the positive class. A decrease in the threshold causes some positive observations ($Y = 1$) to move from FN to TP, and some negative observations ($Y = 0$) to move from TN to FP. Intuitively, a good probabilistic classifier g should be characterized by a situation in which, as the threshold decreases, ground-truth positive observations ($Y = 1$) change their status from FN to TP more rapidly than ground-truth negative observations ($Y = 0$) change their status from TN to FP. Equivalently, the TPR should increase faster than the FPR. The empirical ROC curve captures this behavior by representing $\widehat{\text{TPR}}(\tau)$ as a function of $\widehat{\text{FPR}}(\tau)$ as the threshold τ varies.

The empirical ROC point at threshold τ is defined by

$$\widehat{\boldsymbol{R}}(\tau) := \Big(\widehat{\text{FPR}}(\tau), \widehat{\text{TPR}}(\tau)\Big)^\top, \qquad \tau \in \mathbb{R}. \tag{1}$$

Here, $\widehat{\boldsymbol{R}}(\tau)$ is a bivariate random variable in $[0,1] \times [0,1]$ for every τ.

As we have a finite test sample with n^{ts} elements, we are constructing a discretized representation of the empirical ROC curve. To do that, we discretize the score space $[0,1]$ into $\mathcal{K}$ bins using $\mathcal{K}+1$ grid points, where we assume that $\mathcal{K} \geq 2$. Let $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_{\mathcal{K}})^\top$ with $0 = \gamma_0 < \gamma_1 < \cdots < \gamma_{\mathcal{K}} = 1$ be a grid of bin boundaries, and use them to define the following intervals (see Figure 1):

$$\mathcal{I}_k(\boldsymbol{\gamma}) := \begin{cases} [\gamma_{k-1}, \gamma_k), & k = 1, \ldots, \mathcal{K}-1, \\ [\gamma_{\mathcal{K}-1}, \gamma_{\mathcal{K}}], & k = \mathcal{K}. \end{cases} \tag{2}$$

There are many ways to choose the grid of thresholds. It may be deterministic, for example uniformly spaced on $[0,1]$, or data-dependent. In this paper, we consider the deterministic case. Note that grids computed from the same test sample $\mathcal{D}^{\text{ts}}$ are not covered by the multinomial likelihood used below.

For $k = 1, \ldots, \mathcal{K}$, we can define the class-specific bin counts

$$D_k^+ := \sum_{i=1}^{n^{\text{ts}}} \mathbb{1}\{Y_i^{\text{ts}} = 1,\ W_i^{\text{ts}} \in \mathcal{I}_k(\boldsymbol{\gamma})\}, \qquad D_k^- := \sum_{i=1}^{n^{\text{ts}}} \mathbb{1}\{Y_i^{\text{ts}} = 0,\ W_i^{\text{ts}} \in \mathcal{I}_k(\boldsymbol{\gamma})\}$$

where D_k^+ counts the positive test observations whose scores fall in $\mathcal{I}_k(\boldsymbol{\gamma})$, whereas D_k^- counts the corresponding negative observations. At the population level, for a single test observation, we can define

$$\eta_k^+ := \mathbb{P}(Y = 1, W \in \mathcal{I}_k(\boldsymbol{\gamma})) \quad \text{and} \quad \eta_k^- := \mathbb{P}(Y = 0, W \in \mathcal{I}_k(\boldsymbol{\gamma})) \tag{3}$$

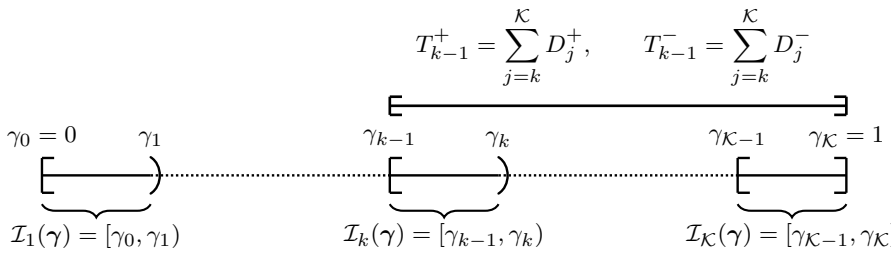


Fig. 1 Discretization of the score space $[0, 1]$ into $\mathcal{K}$ bins. The first $\mathcal{K} - 1$ bins are left-closed and right-open, whereas the last bin is closed at both endpoints. For each bin $\mathcal{I}_j(\boldsymbol{\gamma})$, D_j^+ and D_j^- denote the numbers of positive and negative observations, respectively, whose scores fall in $\mathcal{I}_j(\boldsymbol{\gamma})$. The upper segment represents the right-tail region associated with the threshold γ_{k-1}; all observations with scores in $[\gamma_{k-1}, 1]$ contribute to the cumulative counts $T_{k-1}^+ = \sum_{j=k}^{\mathcal{K}} D_j^+$ and $T_{k-1}^- = \sum_{j=k}^{\mathcal{K}} D_j^-$

where D_k^+/n^{ts} and D_k^-/n^{ts} are the empirical estimators of η_k^+ and η_k^-, respectively.

For compactness, in the following we will use $\ell \in \{+, -\}$ as the class-specific sign index with $y_+ = 1$ and $y_- = 0$. Let us define the histogram vectors

$$\boldsymbol{D}^\ell := (D_1^\ell, \ldots, D_{\mathcal{K}}^\ell)^\top$$

where, by construction, we have $\mathbf{1}_{\mathcal{K}}^\top \boldsymbol{D}^\ell = N_\ell^{\mathrm{ts}}$. We can now reconstruct the empirical ROC curve on the score grid $\boldsymbol{\gamma}$ from these histograms $\boldsymbol{D}^+$ and $\boldsymbol{D}^-$. For $k = 0, \ldots, \mathcal{K} - 1$, we define (see Figure 1):

$$T_k^\ell := \sum_{j=k+1}^{\mathcal{K}} D_j^\ell$$

with the convention that $T_{\mathcal{K}}^+ = T_{\mathcal{K}}^- = 0$. Thus, for $k = 0, \ldots, \mathcal{K} - 1$, T_k^+ and T_k^- are the numbers of ground-truth positive and negative test observations, respectively, that will be classified as positive by h_{γ_k}. Conditionally on the class counts

$$\begin{aligned} T_k^+ \mid N_+^{\mathrm{ts}} &\sim \mathrm{Binomial}\big(N_+^{\mathrm{ts}}, p_{\gamma_k}^+\big)\,, \\ T_k^- \mid N_-^{\mathrm{ts}} &\sim \mathrm{Binomial}\big(N_-^{\mathrm{ts}}, p_{\gamma_k}^-\big)\,, \qquad k = 0, \ldots, \mathcal{K} - 1. \end{aligned}$$

Moreover, for each k, T_k^+ and T_k^- are independent conditionally on N_+^{ts} and N_-^{ts}. But note that for a given ℓ, the tail counts $T_1^\ell, T_2^\ell, \ldots, T_{\mathcal{K}}^\ell$ are not independent; they are indeed nested in the sense that $T_{k-1}^\ell = T_k^\ell + D_k^\ell$.

For given T_k^+ and T_k^-, the corresponding empirical ROC coordinates on the score grid are

$$\widehat{\mathrm{TPR}}_k := \frac{T_k^+}{N_+^{\mathrm{ts}}}, \qquad \widehat{\mathrm{FPR}}_k := \frac{T_k^-}{N_-^{\mathrm{ts}}}, \qquad k = 0, \ldots, \mathcal{K}. \tag{4}$$

Note that here we assume that both random variables N_+^{ts} and N_-^{ts} take positive values, which is a natural assumption in practice ensuring that the test sample contains at least one observation from each class.

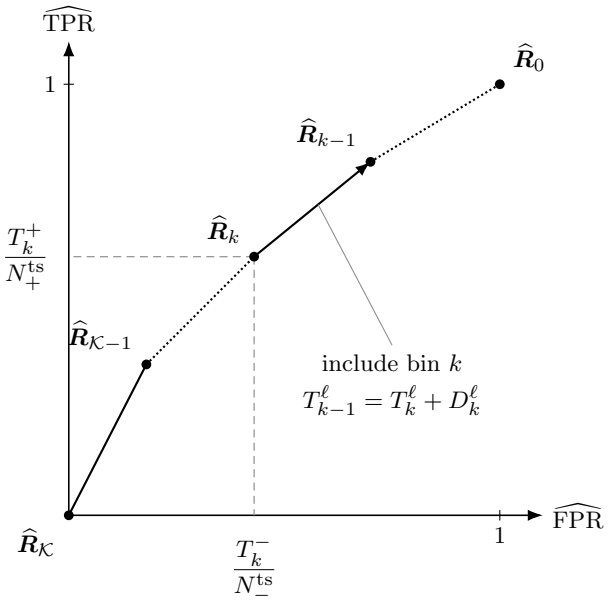


Fig. 2 Evolution of the empirical ROC curve on the score grid as the decision threshold decreases

Hence, the empirical ROC points on the score grid are the random vectors

$$\widehat{\boldsymbol{R}}_k := \left(\widehat{\mathrm{FPR}}_k, \widehat{\mathrm{TPR}}_k\right)^\top = \left(\frac{\sum_{j=k+1}^{\mathcal{K}} D_j^-}{N_-^{\mathrm{ts}}}, \frac{\sum_{j=k+1}^{\mathcal{K}} D_j^+}{N_+^{\mathrm{ts}}}\right)^\top, \qquad k = 0, \ldots, \mathcal{K}. \quad (5)$$

Figure 2 illustrates these empirical ROC points on the score grid and their evolution as the decision threshold decreases.

For $k = 0, \ldots, \mathcal{K}-1$, the empirical ROC point $\widehat{\boldsymbol{R}}_k$ on the score grid coincides with the empirical ROC point at threshold γ_k defined by (1), namely

$$\widehat{\boldsymbol{R}}_k = \widehat{\boldsymbol{R}}(\gamma_k).$$

The last point $\widehat{\boldsymbol{R}}_{\mathcal{K}}$ is the appended endpoint corresponding to any threshold $\tau > 1$; that is, $\widehat{\boldsymbol{R}}_{\mathcal{K}} := \widehat{\boldsymbol{R}}(\tau)$ for any $\tau > 1$. Therefore,

$$\widehat{\boldsymbol{R}}_0 = (1, 1)^\top, \qquad \widehat{\boldsymbol{R}}_{\mathcal{K}} = (0, 0)^\top.$$

5 A Multinomial Representation of the ROC Curve

The test dataset $\mathcal{D}^{\mathrm{ts}}$ contains n^{ts} i.i.d. random observations. As the probabilistic classifier is deterministic, the scores $W_i^{\mathrm{ts}} = g(\boldsymbol{X}_i^{\mathrm{ts}})$ are also i.i.d. random variables and consequently we have n^{ts} pairs of random variables $(W_i^{\mathrm{ts}}, Y_i^{\mathrm{ts}})$.

Each pair $(W_i^{\text{ts}}, Y_i^{\text{ts}})$ can be used to define the outcome of a random experiment with $2\mathcal{K}$ mutually exclusive categories. The label Y_i^{ts} first determines whether the observation contributes to the positive histogram $\boldsymbol{D}^+$ or to the negative histogram $\boldsymbol{D}^-$. Conditional on this choice, the partition $\{\mathcal{I}_1(\boldsymbol{\gamma}), \ldots, \mathcal{I}_{\mathcal{K}}(\boldsymbol{\gamma})\}$ ensures that the score W_i^{ts} belongs to one bin of the selected histogram.

Let $\boldsymbol{D} := \left(D_1^+, \ldots, D_{\mathcal{K}}^+, D_1^-, \ldots, D_{\mathcal{K}}^-\right)^\top$ be the vector of all the $2\mathcal{K}$ bin counts. Each pair $(W_i^{\text{ts}}, Y_i^{\text{ts}})$ contributes to exactly one of the $2\mathcal{K}$ bins of $\boldsymbol{D}$. This vector $\boldsymbol{D}$ can be modeled as a random vector following a multinomial distribution with n^{ts} trials and $2\mathcal{K}$ categories, that is

$$\boldsymbol{D} \sim \text{Multinomial}\left(n^{\text{ts}}, \boldsymbol{\eta}\right) \tag{6}$$

where the vector

$$\boldsymbol{\eta} = (\eta_1^+, \ldots, \eta_{\mathcal{K}}^+, \eta_1^-, \ldots, \eta_{\mathcal{K}}^-)^\top \in \Delta_{2\mathcal{K}-1} := \left\{\boldsymbol{u} \in [0,1]^{2\mathcal{K}} : \mathbf{1}_{2\mathcal{K}}^\top \boldsymbol{u} = 1\right\}$$

contains the probabilities of the $2\mathcal{K}$ mutually exclusive outcomes for a single test observation where η_k^+ and η_k^- are defined in (3).

The vector $\boldsymbol{D}$ records the allocation of test observations across class labels and score bins. For the ROC curve, the coordinates are class-conditional quantities. Thus, the relevant information is obtained by splitting $\boldsymbol{D}$ according to the class label into the two histograms $\boldsymbol{D}^+$ and $\boldsymbol{D}^-$.

Once we condition on N_+^{ts} and N_-^{ts}, the only remaining randomness concerns how the positive observations are distributed among the $\mathcal{K}$ score bins and how the negative observations are distributed among the same $\mathcal{K}$ score bins. At the population level, we define

$$\lambda_k^\ell := \mathbb{P}(W \in \mathcal{I}_k(\boldsymbol{\gamma}) \mid Y = y_\ell), \qquad k = 1, \ldots, \mathcal{K},$$

where

$$\eta_k^+ = \pi \lambda_k^+, \qquad \eta_k^- = (1-\pi)\lambda_k^-.$$

Let us also define the vectors

$$\boldsymbol{\lambda}^\ell = (\lambda_1^\ell, \ldots, \lambda_{\mathcal{K}}^\ell)^\top \in \Delta_{\mathcal{K}-1} := \left\{\boldsymbol{u} \in [0,1]^{\mathcal{K}} : \mathbf{1}_{\mathcal{K}}^\top \boldsymbol{u} = 1\right\}.$$

These vectors $\boldsymbol{\lambda}^+$ and $\boldsymbol{\lambda}^-$ determine the discretized population ROC curve on the chosen grid. Indeed, at threshold γ_k, the class-conditional right-tail probability is

$$p_{\gamma_k}^\ell = \mathbb{P}(W \geq \gamma_k \mid Y = y_\ell) = \sum_{j=k+1}^{\mathcal{K}} \lambda_j^\ell, \qquad \ell \in \{+,-\}, \quad k = 0, \ldots, \mathcal{K}-1$$

and thus the corresponding population ROC point is

$$\boldsymbol{r}_k := \begin{pmatrix} p_{\gamma_k}^- \\ p_{\gamma_k}^+ \end{pmatrix} = \begin{pmatrix} \text{fpr}(\gamma_k) \\ \text{tpr}(\gamma_k) \end{pmatrix} = \begin{pmatrix} \sum_{j=k+1}^{\mathcal{K}} \lambda_j^- \\ \sum_{j=k+1}^{\mathcal{K}} \lambda_j^+ \end{pmatrix}, \qquad k = 0, \ldots, \mathcal{K}-1 \tag{7}$$

where $\boldsymbol{r}_{\mathcal{K}} = (0,0)^\top$. Let us define $\mathfrak{d}^{\text{ts}} := ((\boldsymbol{x}_i^{\text{ts}}, y_i^{\text{ts}}))_{i=1}^{n^{\text{ts}}}$ as a realization of the test dataset $\mathcal{D}^{\text{ts}}$. After observing $\mathfrak{d}^{\text{ts}}$, we obtain one realization $\boldsymbol{d}^\ell$. This observed histogram

$\boldsymbol{d}^\ell$ is one realization of $\boldsymbol{D}^\ell$ that follows a multinomial distribution governed by the vector $\boldsymbol{\lambda}^\ell$, that is

$$\boldsymbol{D}^\ell \mid N_\ell^{\text{ts}} \sim \text{Multinomial}\left(N_\ell^{\text{ts}}, \boldsymbol{\lambda}^\ell\right), \qquad \ell \in \{+,-\} \tag{8}$$

with the following probability mass functions

$$\mathbb{P}_{n_\ell^{\text{ts}}}\left(\boldsymbol{D}^\ell = \boldsymbol{d}^\ell\right) = \frac{n_\ell^{\text{ts}}!}{\prod_{k=1}^{\mathcal{K}} d_k^\ell!} \prod_{k=1}^{\mathcal{K}} \left(\lambda_k^\ell\right)^{d_k^\ell} \mathbb{1}\left\{\sum_{k=1}^{\mathcal{K}} d_k^\ell = n_\ell^{\text{ts}}\right\}. \tag{9}$$

Here, for simplicity, we define $\mathbb{P}_{n_\ell^{\text{ts}}}(\cdot) := \mathbb{P}(\cdot \mid N_\ell^{\text{ts}} = n_\ell^{\text{ts}})$.

Note that, conditionally on the class counts, these two multinomial allocations concern disjoint subsets of observations and are independent:

$$\boldsymbol{D}^+ \perp\!\!\!\perp \boldsymbol{D}^- \mid N_+^{\text{ts}}, N_-^{\text{ts}}.$$

Thus, the original $2\mathcal{K}$-category multinomial model defined in (6) decomposes, after conditioning on the class counts, into two independent $\mathcal{K}$-category multinomial models given by (8).

6 A Bayesian Representation of the ROC Curve

In the probabilistic model (8), both parameters $\boldsymbol{\lambda}^+$ and $\boldsymbol{\lambda}^-$ are assumed to be fixed and unknown. By adopting a Bayesian formulation, we can consider them as realizations of random vectors $\boldsymbol{\Lambda}^+$ and $\boldsymbol{\Lambda}^-$. In this Bayesian setting, the left-hand side of (9) is written as

$$\mathbb{P}_{n_\ell^{\text{ts}}}\left(\boldsymbol{D}^\ell = \boldsymbol{d}^\ell \mid \boldsymbol{\Lambda}^\ell = \boldsymbol{\lambda}^\ell\right).$$

Bayes' rule and the law of total probability tell us that

$$\begin{aligned} f_{\boldsymbol{\Lambda}^\ell \mid \boldsymbol{D}^\ell}(\boldsymbol{\lambda}^\ell \mid \boldsymbol{d}^\ell) &= \frac{\mathbb{P}_{n_\ell^{\text{ts}}}\left(\boldsymbol{D}^\ell = \boldsymbol{d}^\ell \mid \boldsymbol{\Lambda}^\ell = \boldsymbol{\lambda}^\ell\right) f_{\boldsymbol{\Lambda}^\ell}(\boldsymbol{\lambda}^\ell)}{\mathbb{P}_{n_\ell^{\text{ts}}}\left(\boldsymbol{D}^\ell = \boldsymbol{d}^\ell\right)} \\ &= \frac{\mathbb{P}_{n_\ell^{\text{ts}}}\left(\boldsymbol{D}^\ell = \boldsymbol{d}^\ell \mid \boldsymbol{\Lambda}^\ell = \boldsymbol{\lambda}^\ell\right) f_{\boldsymbol{\Lambda}^\ell}(\boldsymbol{\lambda}^\ell)}{\int_{\Delta_{\mathcal{K}-1}} \mathbb{P}_{n_\ell^{\text{ts}}}\left(\boldsymbol{D}^\ell = \boldsymbol{d}^\ell \mid \boldsymbol{\Lambda}^\ell = \boldsymbol{u}\right) f_{\boldsymbol{\Lambda}^\ell}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}} \end{aligned} \tag{10}$$

where $f_{\boldsymbol{\Lambda}^\ell \mid \boldsymbol{D}^\ell}(\boldsymbol{\lambda}^\ell \mid \boldsymbol{d}^\ell)$ is the conditional density of the parameter vector $\boldsymbol{\Lambda}^\ell$ given an observed $\boldsymbol{d}^\ell$ computed from a realization $\mathfrak{d}^{\text{ts}}$. In (10), the denominator is a normalizing constant to ensure that $\int_{\Delta_{\mathcal{K}-1}} f_{\boldsymbol{\Lambda}^\ell \mid \boldsymbol{D}^\ell}(\boldsymbol{u} \mid \boldsymbol{d}^\ell)\mathrm{d}\boldsymbol{u} = 1$ and it does not depend on $\boldsymbol{\lambda}^\ell$. Consequently, we only have to evaluate the numerator and then normalize it so that its integral equals 1. Thus,

$$\underbrace{f_{\boldsymbol{\Lambda}^\ell \mid \boldsymbol{D}^\ell}(\boldsymbol{\lambda}^\ell \mid \boldsymbol{d}^\ell)}_{\text{Posterior}} \propto \underbrace{\mathbb{P}_{n_\ell^{\text{ts}}}\left(\boldsymbol{D}^\ell = \boldsymbol{d}^\ell \mid \boldsymbol{\Lambda}^\ell = \boldsymbol{\lambda}^\ell\right)}_{\text{Likelihood}} \underbrace{f_{\boldsymbol{\Lambda}^\ell}(\boldsymbol{\lambda}^\ell)}_{\text{Prior}}. \tag{11}$$

Thanks to Bayes' rule, which allows us to interpret the parameters $\boldsymbol{\lambda}^+$ and $\boldsymbol{\lambda}^-$ as realizations of random vectors, (11) gives the conditional distribution of the parameters $\boldsymbol{\Lambda}^\ell$. This conditional density is proportional to the product of two terms. The likelihood returns the plausibility score of an observed $\boldsymbol{d}^\ell$ given a value for the parameters $\boldsymbol{\lambda}^\ell$. The second term is the prior distribution and, as we will see, it allows us to incorporate prior knowledge about the parameters $\boldsymbol{\lambda}^\ell$. The posterior distribution $f_{\boldsymbol{\Lambda}^\ell \mid \boldsymbol{D}^\ell}(\boldsymbol{\lambda}^\ell \mid \boldsymbol{d}^\ell)$ is a compromise between the prior and the likelihood.

The Dirichlet distribution is the conjugate distribution of the multinomial distribution [17]. It ensures that if the likelihood follows a multinomial distribution and the prior follows a Dirichlet distribution, then the posterior distribution also follows a Dirichlet distribution. The Dirichlet distribution is commonly used in Bayesian statistics to model the parameters of a multinomial distribution. In this context, we say that the random vector $\boldsymbol{\Lambda}^\ell$ follows a Dirichlet distribution

$$\boldsymbol{\Lambda}^\ell \sim \text{Dirichlet}(\boldsymbol{\alpha}^\ell)$$

with the following density function

$$\begin{aligned} f_{\boldsymbol{\Lambda}^\ell}(\boldsymbol{\lambda}^\ell) &= \frac{\Gamma\left(\sum_{k=1}^{\mathcal{K}} \alpha_k^\ell\right)}{\prod_{k=1}^{\mathcal{K}} \Gamma(\alpha_k^\ell)} \prod_{k=1}^{\mathcal{K}} \left(\lambda_k^\ell\right)^{\alpha_k^\ell - 1} \mathbb{1}\left\{\boldsymbol{\lambda}^\ell \in \Delta_{\mathcal{K}-1}\right\} \\ &\propto \prod_{k=1}^{\mathcal{K}} \left(\lambda_k^\ell\right)^{\alpha_k^\ell - 1} \mathbb{1}\left\{\boldsymbol{\lambda}^\ell \in \Delta_{\mathcal{K}-1}\right\}. \end{aligned} \tag{12}$$

Here, $\boldsymbol{\alpha}^\ell = (\alpha_1^\ell, \ldots, \alpha_{\mathcal{K}}^\ell)^\top \in \mathbb{R}_{>0}^{\mathcal{K}}$ is a vector of positive hyperparameters and $\Gamma(\cdot)$ is the gamma function. The first term $\Gamma\left(\sum_{k=1}^{\mathcal{K}} \alpha_k^\ell\right) / \prod_{k=1}^{\mathcal{K}} \Gamma(\alpha_k^\ell)$ in the previous equation can be omitted because it is only there to ensure the integral of $f_{\boldsymbol{\Lambda}^\ell}(\boldsymbol{\lambda}^\ell)$ on $\Delta_{\mathcal{K}-1}$ is equal to 1. The two class-conditional score distributions are modeled separately; therefore, we assume that $\boldsymbol{\Lambda}^+ \perp\!\!\!\perp \boldsymbol{\Lambda}^-$.

The Bayesian counterpart of the fixed population ROC point $\boldsymbol{r}_k$ from (7) is the population ROC point under the Bayesian model, represented by the random vector

$$\boldsymbol{R}_k := (\text{FPR}_k, \text{TPR}_k)^\top = \left(\sum_{j=k+1}^{\mathcal{K}} \Lambda_j^-, \sum_{j=k+1}^{\mathcal{K}} \Lambda_j^+\right)^\top \tag{13}$$

with $\boldsymbol{R}_{\mathcal{K}} := (0, 0)^\top$. The discretized population ROC curve under the Bayesian model is represented by the random matrix

$$\boldsymbol{R}_{\mathcal{K}:0} := [\boldsymbol{R}_{\mathcal{K}} \ \cdots \ \boldsymbol{R}_0] \in [0, 1]^{2 \times (\mathcal{K}+1)},$$

whose columns are the successive population ROC points under the Bayesian model.

Equation (12) defines the prior distribution of the parameters $\boldsymbol{\Lambda}^\ell$ and prior knowledge can be incorporated into the prior distribution via $\boldsymbol{\alpha}^\ell$. This topic will be covered in Section 7.

If in (11) we replace the likelihood and the prior by their respective expressions in (9) and (12), we obtain

$$f_{\boldsymbol{\Lambda}^\ell|\boldsymbol{D}^\ell}(\boldsymbol{\lambda}^\ell \mid \boldsymbol{d}^\ell) \propto \prod_{k=1}^{\mathcal{K}} \left(\lambda_k^\ell\right)^{d_k^\ell} \prod_{k=1}^{\mathcal{K}} \left(\lambda_k^\ell\right)^{\alpha_k^\ell - 1} \mathbb{1}\left\{\boldsymbol{\lambda}^\ell \in \Delta_{\mathcal{K}-1}\right\}$$
$$= \prod_{k=1}^{\mathcal{K}} \left(\lambda_k^\ell\right)^{d_k^\ell + \alpha_k^\ell - 1} \mathbb{1}\left\{\boldsymbol{\lambda}^\ell \in \Delta_{\mathcal{K}-1}\right\}. \tag{14}$$

This is the posterior distribution of the unknown parameters $\boldsymbol{\Lambda}^\ell$ given an observed $\boldsymbol{d}^\ell$. Given an observed $\boldsymbol{d}^\ell$ and the prior knowledge encoded in $\boldsymbol{\alpha}^\ell$, the posterior distribution is a Dirichlet distribution

$$\boldsymbol{\Lambda}^\ell \mid \boldsymbol{D}^\ell = \boldsymbol{d}^\ell \sim \text{Dirichlet}(\boldsymbol{\alpha}^\ell + \boldsymbol{d}^\ell) = \text{Dirichlet}(\boldsymbol{\omega}^\ell) \tag{15}$$

where $\boldsymbol{\omega}^\ell := \boldsymbol{\alpha}^\ell + \boldsymbol{d}^\ell$ is the parameter vector of the posterior density function. Note that the same class-specific Dirichlet update is used for ordered discrete diagnostics in Section 3.2 of [9].

Let $\widetilde{\boldsymbol{D}}^\ell$ be a new random vector representing a histogram of counts following the same multinomial distribution as $\boldsymbol{D}^\ell$ but independent, so that

$$\widetilde{\boldsymbol{D}}^\ell \mid N_\ell^{\text{ts}}, \boldsymbol{\Lambda}^\ell \sim \text{Multinomial}\left(N_\ell^{\text{ts}}, \boldsymbol{\Lambda}^\ell\right), \qquad \widetilde{\boldsymbol{D}}^\ell \perp\!\!\!\perp \boldsymbol{D}^\ell \mid N_\ell^{\text{ts}}, \boldsymbol{\Lambda}^\ell.$$

Let $\tilde{\boldsymbol{d}}^\ell$ be an arbitrary vector of counts with $\sum_{k=1}^{\mathcal{K}} \tilde{d}_k^\ell = n_\ell^{\text{ts}}$. Conditional on the observed class count $N_\ell^{\text{ts}} = n_\ell^{\text{ts}}$, but before using the observed allocation $\boldsymbol{d}^\ell$ across score bins, the probability of observing $\tilde{\boldsymbol{d}}^\ell$ is given by

$$\mathbb{P}_{n_\ell^{\text{ts}}}\left(\widetilde{\boldsymbol{D}}^\ell = \tilde{\boldsymbol{d}}^\ell\right) = \int_{\Delta_{\mathcal{K}-1}} \mathbb{P}_{n_\ell^{\text{ts}}}\left(\widetilde{\boldsymbol{D}}^\ell = \tilde{\boldsymbol{d}}^\ell \mid \boldsymbol{\Lambda}^\ell = \boldsymbol{u}\right) f_{\boldsymbol{\Lambda}^\ell}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}$$
$$= \mathbb{E}_{f_{\boldsymbol{\Lambda}^\ell}}\left[\mathbb{P}_{n_\ell^{\text{ts}}}\left(\widetilde{\boldsymbol{D}}^\ell = \tilde{\boldsymbol{d}}^\ell \mid \boldsymbol{\Lambda}^\ell\right)\right]$$

where $f_{\boldsymbol{\Lambda}^\ell}(\cdot)$ is only parameterized by $\boldsymbol{\alpha}^\ell$.

After observing $\boldsymbol{d}^\ell$, we can use (15) to compute the new predictive distribution of any vector of counts $\tilde{\boldsymbol{d}}^\ell$ via

$$\mathbb{P}_{n_\ell^{\text{ts}}}\left(\widetilde{\boldsymbol{D}}^\ell = \tilde{\boldsymbol{d}}^\ell \mid \boldsymbol{D}^\ell = \boldsymbol{d}^\ell\right) = \int_{\Delta_{\mathcal{K}-1}} \mathbb{P}_{n_\ell^{\text{ts}}}\left(\widetilde{\boldsymbol{D}}^\ell = \tilde{\boldsymbol{d}}^\ell \mid \boldsymbol{\Lambda}^\ell = \boldsymbol{u}\right) f_{\boldsymbol{\Lambda}^\ell|\boldsymbol{D}^\ell}(\boldsymbol{u} \mid \boldsymbol{d}^\ell) \mathrm{d}\boldsymbol{u} \tag{16}$$
$$= \mathbb{E}_{f_{\boldsymbol{\Lambda}^\ell|\boldsymbol{D}^\ell = \boldsymbol{d}^\ell}}\left[\mathbb{P}_{n_\ell^{\text{ts}}}\left(\widetilde{\boldsymbol{D}}^\ell = \tilde{\boldsymbol{d}}^\ell \mid \boldsymbol{\Lambda}^\ell\right)\right]. \tag{17}$$

For $\ell \in \{+,-\}$ and $k = 0, \ldots, \mathcal{K}$, we can define the future right-tail counts

$$\widetilde{T}_k^\ell := \sum_{j=k+1}^{\mathcal{K}} \widetilde{D}_j^\ell$$

where $\widetilde{T}_0^\ell = n_\ell^{\mathrm{ts}}$ and by convention we set $\widetilde{T}_{\mathcal{K}}^\ell = 0$.

Following the same reasoning as in (4), the future empirical ROC point associated with the future histograms is

$$\widetilde{\boldsymbol{R}}_k := \left(\widetilde{\mathrm{FPR}}_k, \widetilde{\mathrm{TPR}}_k\right)^\top := \left(\widetilde{T}_k^- / n_-^{\mathrm{ts}},\ \widetilde{T}_k^+ / n_+^{\mathrm{ts}}\right)^\top, \qquad k = 0, \ldots, \mathcal{K}.$$

Consequently,

$$\widetilde{\boldsymbol{R}}_{\mathcal{K}} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \widetilde{\boldsymbol{R}}_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The corresponding future empirical ROC curve on the score grid $\boldsymbol{\gamma}$ is represented by the random matrix

$$\widetilde{\boldsymbol{R}}_{\mathcal{K}:0} := \left[\widetilde{\boldsymbol{R}}_{\mathcal{K}} \ \cdots \ \widetilde{\boldsymbol{R}}_0\right] \in [0,1]^{2\times(\mathcal{K}+1)}.$$

7 Adding prior knowledge to the Bayesian model

As already mentioned, the Bayesian model allows us to incorporate prior knowledge into the prior distribution of the parameters $\boldsymbol{\Lambda}^\ell$. There are many ways to encode prior knowledge. We consider three prior specifications: a symmetric prior, a linear prior calibrated to a target expected AUC, and a prior informed by historical score histograms.

But before presenting these specifications, we will derive a general expression for the prior expected discretized AUC in terms of the prior mean bin probabilities. This expression will simplify the subsequent derivations. Let us denote by $A_{\mathcal{K}}$ the population AUC associated with $\boldsymbol{R}_{\mathcal{K}:0}$. In the following, we derive a general expression for the prior expected AUC defined as $a_{\mathcal{K}} := \mathbb{E}[A_{\mathcal{K}}]$.

Lemma 1 *Assume that*

$$\boldsymbol{\Lambda}^+ \perp\!\!\!\perp \boldsymbol{\Lambda}^-,$$

and define

$$q_k^\ell := \mathbb{E}[\Lambda_k^\ell], \qquad \ell \in \{+,-\}, \quad k = 1, \ldots, \mathcal{K}. \tag{18}$$

Then the prior expected discretized AUC is

$$a_{\mathcal{K}} = \mathbb{E}[A_{\mathcal{K}}] = \sum_{k=1}^{\mathcal{K}} q_k^- \left(\sum_{j=k+1}^{\mathcal{K}} q_j^+ + \frac{1}{2} q_k^+ \right), \tag{19}$$

where an empty sum is equal to zero.

The proof of Lemma 1 is given in Appendix A.

7.1 Symmetric uniform Dirichlet prior

A simple default choice is to use a symmetric uniform Dirichlet prior,

$$\boldsymbol{\alpha}^\ell = \left(\frac{m}{\mathcal{K}}, \ldots, \frac{m}{\mathcal{K}}\right)^\top, \qquad m > 0 \tag{20}$$

where $\sum_{k=1}^{\mathcal{K}} \alpha_k^\ell = m$. As $\boldsymbol{\Lambda}^\ell \sim \text{Dirichlet}(\boldsymbol{\alpha}^\ell)$, we have the following prior mean

$$q_k^\ell := \mathbb{E}[\Lambda_k^\ell] = \frac{\alpha_k^\ell}{\sum_{j=1}^{\mathcal{K}} \alpha_j^\ell} = \frac{m/\mathcal{K}}{m} = \frac{1}{\mathcal{K}}.$$

In this case, this probability is the same for all bins and it treats all score bins equally. The scalar m controls the total amount of prior information: larger values of m induce stronger prior information. In the posterior distribution, this prior acts as if $m/\mathcal{K}$ pseudo-counts had been added to each bin.

Note that this symmetric uniform prior does not represent an absence of prior information. As

$$\mathbb{E}[\text{TPR}_k] = \sum_{j=k+1}^{\mathcal{K}} \mathbb{E}[\Lambda_j^+] = \frac{\mathcal{K}-k}{\mathcal{K}} \qquad \text{and} \qquad \mathbb{E}[\text{FPR}_k] = \frac{\mathcal{K}-k}{\mathcal{K}}$$

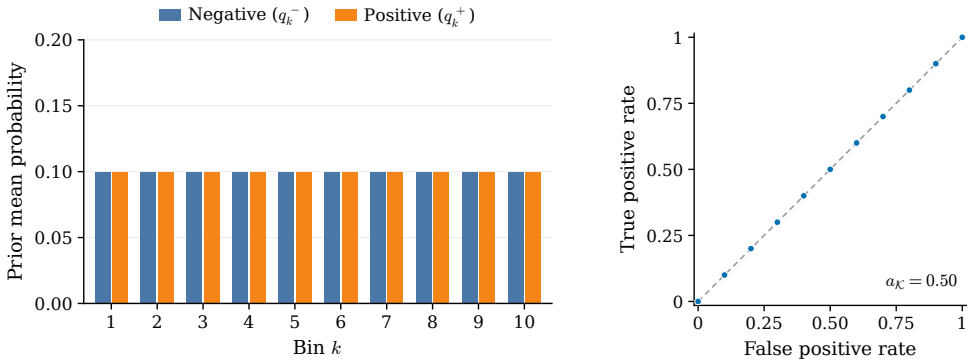
we have

$$\mathbb{E}[\boldsymbol{R}_k] = \frac{\mathcal{K}-k}{\mathcal{K}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{21}$$

This symmetric uniform prior sets both class-conditional prior mean profiles to the same uniform histogram, so that the prior mean population ROC curve lies on the diagonal.

From Lemma 1, we know that the prior expected AUC equals 0.5:

$$\begin{aligned}
a_{\mathcal{K}} &= \sum_{k=1}^{\mathcal{K}} \frac{1}{\mathcal{K}} \left(\sum_{j=k+1}^{\mathcal{K}} \frac{1}{\mathcal{K}} + \frac{1}{2\mathcal{K}} \right) \\
&= \sum_{k=1}^{\mathcal{K}} \frac{1}{\mathcal{K}} \left(\frac{\mathcal{K}-k}{\mathcal{K}} + \frac{1}{2\mathcal{K}} \right) \\
&= \frac{1}{\mathcal{K}^2} \sum_{k=1}^{\mathcal{K}} \left(\mathcal{K} - k + \frac{1}{2} \right) \\
&= \frac{1}{\mathcal{K}^2} \left[\sum_{k=1}^{\mathcal{K}} (\mathcal{K}-k) + \sum_{k=1}^{\mathcal{K}} \frac{1}{2} \right] \\
&= \frac{1}{\mathcal{K}^2} \left[\frac{\mathcal{K}(\mathcal{K}-1)}{2} + \frac{\mathcal{K}}{2} \right]
\end{aligned}$$



(a) Prior mean bin probabilities

(b) Prior mean discretized population ROC curve

Fig. 3 Effect of the symmetric uniform Dirichlet prior for $\mathcal{K} = 10$. (a) Both classes share the uniform prior mean $q_k^+ = q_k^- = 1/\mathcal{K} = 0.1$. (b) This symmetry places the prior mean population ROC curve on the diagonal $\mathbb{E}[\mathrm{TPR}_k] = \mathbb{E}[\mathrm{FPR}_k]$, yielding $a_{\mathcal{K}} := \mathbb{E}[A_{\mathcal{K}}] = 1/2$

$$= \frac{1}{\mathcal{K}^2}\frac{\mathcal{K}^2}{2}$$
$$= \frac{1}{2}.$$

Figure 3 illustrates the evolution for $\mathcal{K} = 10$, showing the common uniform prior mean over the score bins and the resulting prior mean population ROC curve on the diagonal.

7.2 A linear prior calibrated to a target expected AUC

In some applications, a previous study provides an AUC summary, but neither individual scores nor class-specific score histograms are available. It is therefore useful to develop a method that allows this prior AUC information to be incorporated directly into our model. Let a_0 denote this target prior expected AUC value. For a given pair of priors $(\boldsymbol{\alpha}^+, \boldsymbol{\alpha}^-)$, $a_{\mathcal{K}} := \mathbb{E}[A_{\mathcal{K}}]$ denotes the induced prior expected discretized AUC. Our goal is to choose the priors so that $a_{\mathcal{K}} = a_0$.

To do that, we first assume that the two Dirichlet priors have the same total concentration,

$$\sum_{r=1}^{\mathcal{K}} \alpha_r^+ = \sum_{r=1}^{\mathcal{K}} \alpha_r^- = m, \qquad m > 0.$$

We can then define

$$q_k^\ell := \mathbb{E}[\Lambda_k^\ell] = \frac{\alpha_k^\ell}{m}, \qquad \ell \in \{+, -\} \tag{22}$$

so that

$$\sum_{k=1}^{\mathcal{K}} q_k^\ell = 1. \tag{23}$$

The target value a_0 does not uniquely determine the two vectors $\boldsymbol{\alpha}^+$ and $\boldsymbol{\alpha}^-$. Since different pairs of vectors may produce the same expected AUC, we have to consider a structured family of priors defined by some assumptions. We make four assumptions about the prior mean bin probabilities q_k^+ and q_k^-. First, for the positive class, the prior mean probability q_k^+ increases linearly as the score bin index k increases. Second, for the negative class, the prior mean probability q_k^- decreases linearly as the score bin index k increases. Third, the two class-specific profiles are mirror images of one another, so that bins favored by the positive class are disfavored by the negative class, and vice versa. Finally, when $a_0 = 1/2$, both classes share the same uniform distribution over the $\mathcal{K}$ bins.

Let us define the centered bin indices

$$x_k := k - \frac{\mathcal{K}+1}{2}, \qquad k = 1, \ldots, \mathcal{K},$$

which measure the position of each bin relative to the center of the discretized score range and satisfy $\sum_{k=1}^{\mathcal{K}} x_k = 0$. Centering the bin indices simplifies the construction of symmetric linear prior profiles. For example, if $\mathcal{K} = 4$, then

$$(x_1, x_2, x_3, x_4)^\top = \left(-\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}\right)^\top.$$

Under the assumptions stated above, we consider a one-parameter family for the prior mean bin probabilities. For $k = 1, \ldots, \mathcal{K}$, we define

$$q_k^+ = \frac{1}{\mathcal{K}} + s x_k, \qquad q_k^- = \frac{1}{\mathcal{K}} - s x_k. \tag{24}$$

The parameter $s \geq 0$ controls the separation between the two class-conditional prior mean distributions.

Because the x_k sum to zero,

$$\sum_{k=1}^{\mathcal{K}} q_k^+ = \sum_{k=1}^{\mathcal{K}} \left(\frac{1}{\mathcal{K}} + s x_k\right) = 1 + s \sum_{k=1}^{\mathcal{K}} x_k = 1,$$

and similarly $\sum_{k=1}^{\mathcal{K}} q_k^- = 1$. We recover the constraint in (23) for both classes.

Moreover,

$$q_{\mathcal{K}-k+1}^+ = \frac{1}{\mathcal{K}} + s x_{\mathcal{K}-k+1} = \frac{1}{\mathcal{K}} + s\left(\mathcal{K} - k + 1 - \frac{\mathcal{K}+1}{2}\right) = \frac{1}{\mathcal{K}} - s x_k = q_k^-.$$

Thus, the positive and negative profiles are exact mirror images. We recover the symmetry assumption made above.

It is also easy to see that when $s = 0$, then all the q_k^ℓ are equal to $1/\mathcal{K}$ and we recover the symmetric uniform Dirichlet prior defined in (20). When $s > 0$, we have

$$q_1^+ < q_2^+ < \cdots < q_{\mathcal{K}}^+, \qquad q_1^- > q_2^- > \cdots > q_{\mathcal{K}}^-.$$

The profiles in (24) thus satisfy the four assumptions stated above.

We will now use Lemma 1 to derive an explicit expression for the expected discretized AUC of these linear prior mean profiles. The goal is then to invert this expression to find the value of s that produces a target expected AUC a_0.

Corollary 2 *Consider the linear prior mean profiles*

$$q_k^+ = \frac{1}{\mathcal{K}} + sx_k, \qquad q_k^- = \frac{1}{\mathcal{K}} - sx_k,$$

where

$$x_k = k - \frac{\mathcal{K}+1}{2}, \qquad k = 1, \ldots, \mathcal{K}.$$

Then the induced expected discretized AUC is

$$a_{\mathcal{K}} = \frac{1}{2} + \frac{s(\mathcal{K}^2-1)}{6}. \tag{25}$$

The proof is provided in Appendix C.

Given Corollary 2, we can now calibrate the linear prior profiles in (24) to a target expected AUC a_0. The goal is to choose s_0 so that $a_{\mathcal{K}} = a_0$:

$$\begin{aligned} a_0 &= \frac{1}{2} + s_0 \frac{\mathcal{K}^2-1}{6} \\ \iff \quad s_0 &= \frac{6(a_0 - 1/2)}{\mathcal{K}^2-1}. \end{aligned} \tag{26}$$

It is easy to see that $a_0 = 1/2 \iff s_0 = 0$ and that s_0 increases linearly with the target expected AUC a_0.

It is now possible to obtain the prior mean bin probabilities by substituting (26) into (24)

$$q_k^+ = \frac{1}{\mathcal{K}} + \frac{3(2a_0-1)}{\mathcal{K}^2-1}\left(k - \frac{\mathcal{K}+1}{2}\right) \qquad q_k^- = \frac{1}{\mathcal{K}} - \frac{3(2a_0-1)}{\mathcal{K}^2-1}\left(k - \frac{\mathcal{K}+1}{2}\right).$$

Since the total concentration is m, from (22) we can obtain the Dirichlet hyperparameters

$$\begin{aligned} \alpha_k^+ &= m\left[\frac{1}{\mathcal{K}} + \frac{3(2a_0-1)}{\mathcal{K}^2-1}\left(k - \frac{\mathcal{K}+1}{2}\right)\right], \\ \alpha_k^- &= m\left[\frac{1}{\mathcal{K}} - \frac{3(2a_0-1)}{\mathcal{K}^2-1}\left(k - \frac{\mathcal{K}+1}{2}\right)\right]. \end{aligned} \tag{27}$$

7.2.1 Feasible target AUC values

We need to check that all the probabilities q_k^ℓ are strictly positive. This is necessary because the Dirichlet parameters $\alpha_k^\ell = mq_k^\ell$ must be strictly positive. As $s \geq 0$, the minimum value of q_k^+ occurs at $k = 1$ and the minimum value of q_k^- occurs at $k = \mathcal{K}$. As we have symmetry, we can consider only the positive class and ensure that $q_1^+ > 0$. When $k = 1$, we have

$$x_1 = -\frac{\mathcal{K}-1}{2}$$

and therefore

$$q_1^+ = \frac{1}{\mathcal{K}} - s\frac{\mathcal{K}-1}{2} > 0 \quad \iff \quad s < \frac{2}{\mathcal{K}(\mathcal{K}-1)}. \tag{28}$$

By symmetry, the same constraint holds for the negative class. Therefore, to ensure that both q_k^+ and q_k^- are positive, we must impose an upper bound on s.

The target expected AUC a_0 determines the slope s_0 through (26):

$$s_0 = \frac{6(a_0 - 1/2)}{\mathcal{K}^2 - 1}.$$

However, s_0 cannot be arbitrarily large. To ensure that all prior mean bin probabilities remain strictly positive, (28) requires

$$s_0 < \frac{2}{\mathcal{K}(\mathcal{K}-1)}.$$

Combining these two conditions gives

$$\frac{6(a_0 - 1/2)}{\mathcal{K}^2 - 1} < \frac{2}{\mathcal{K}(\mathcal{K}-1)}$$

and this simplifies to

$$a_0 < \frac{5\mathcal{K}+2}{6\mathcal{K}}. \tag{29}$$

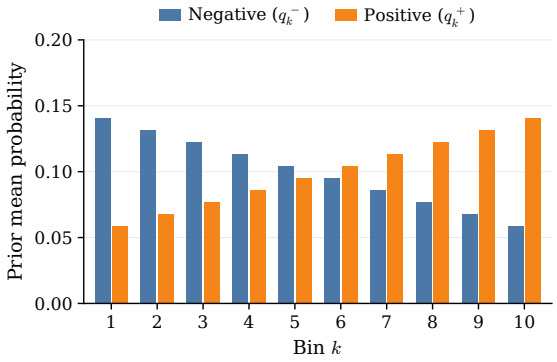
For example, when $\mathcal{K} = 10$, the target expected AUC must satisfy

$$a_0 < \frac{52}{60} \approx 0.867$$

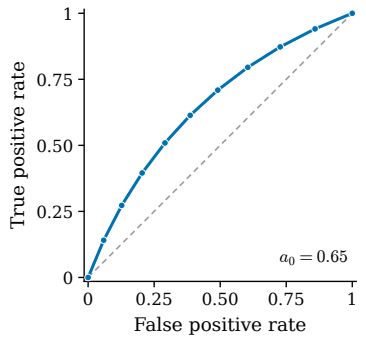
with

$$\lim_{\mathcal{K}\to\infty} \frac{5\mathcal{K}+2}{6\mathcal{K}} = \frac{5}{6} = 0.8333\ldots$$

This upper bound is a consequence of the linear prior profiles in (24). As the slope increases, the probabilities assigned to the least favored bins decrease linearly and eventually reach zero. More flexible nonlinear profiles could preserve positivity while allowing target expected AUC values closer to one. Constructing such families from an AUC target alone is left for future work.



(a) Calibrated prior mean bin probabilities



(b) Calibrated prior mean discretized population ROC curve

Fig. 4 Linear prior calibrated to $a_0 = 0.65$ for $\mathcal{K} = 10$. (a) The mirrored prior means satisfy $q_k^{\pm} = 1/\mathcal{K} \pm s_0\big(k - (\mathcal{K}+1)/2\big)$, where $s_0 = 6(a_0 - 1/2)/(\mathcal{K}^2 - 1) = 1/110$. (b) The resulting prior mean population ROC curve lies above the diagonal and satisfies $a_{\mathcal{K}} = \mathbb{E}[A_{\mathcal{K}}] = a_0 = 0.65$

7.2.2 Prior mean ROC curve under the linear prior

Having calibrated the linear prior to a target expected AUC a_0, we now derive the corresponding prior mean ROC curve.

Corollary 3 *Let $\mathcal{K} \geq 2$ and consider the Dirichlet priors with hyperparameters defined in (27), where $m > 0$ and*

$$\frac{1}{2} \leq a_0 < \frac{5\mathcal{K}+2}{6\mathcal{K}}.$$

Then, for every $k = 0, \ldots, \mathcal{K}$, the prior mean discretized population ROC point is

$$\mathbb{E}[\boldsymbol{R}_k] = \frac{\mathcal{K}-k}{\mathcal{K}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{3(2a_0-1)k(\mathcal{K}-k)}{2(\mathcal{K}^2-1)} \begin{pmatrix} -1 \\ 1 \end{pmatrix}. \tag{30}$$

The proof is provided in Appendix D.

This expression has a useful interpretation. The first term is the prior mean point in (21) for the symmetric uniform prior. The second term is a correction that shifts the interior prior mean points above the diagonal when $a_0 > 1/2$. When $a_0 = 1/2$, the second term vanishes and the prior mean discretized population ROC curve lies on the diagonal.

Using (30), Figure 4 illustrates the calibration for $\mathcal{K} = 10$ and $a_0 = 0.65$, showing how the target expected AUC determines both the mirrored prior mean profiles and the resulting prior mean population ROC curve.

7.3 A prior informed by historical score histograms

This section provides a third way to inject prior information into the model. A previous study may provide class-specific score histograms. They must refer to the same fixed

classifier g and score grid $\boldsymbol{\gamma}$, and come from observations collected independently of the current test sample. For each class $\ell \in \{+, -\}$, let

$$\boldsymbol{d}^{\mathrm{pr},\ell} := \left(d_1^{\mathrm{pr},\ell}, \ldots, d_{\mathcal{K}}^{\mathrm{pr},\ell}\right)^\top, \qquad n_\ell^{\mathrm{pr}} := \sum_{k=1}^{\mathcal{K}} d_k^{\mathrm{pr},\ell} > 0$$

denote the observed historical histogram and its sample size. We define the historical mean profile and the Dirichlet parameters by

$$\widehat{q}_k^{\mathrm{pr},\ell} := \frac{d_k^{\mathrm{pr},\ell} + 1/\mathcal{K}}{n_\ell^{\mathrm{pr}} + 1}, \qquad \alpha_k^\ell := m\,\widehat{q}_k^{\mathrm{pr},\ell}, \qquad k = 1, \ldots, \mathcal{K}, \quad m > 0. \tag{31}$$

The term $1/\mathcal{K}$ distributes one smoothing pseudo-count equally across all bins, ensuring positive probabilities even when a historical bin is empty. The denominator normalizes these counts, so that

$$\widehat{q}_k^{\mathrm{pr},\ell} > 0, \qquad \sum_{k=1}^{\mathcal{K}} \widehat{q}_k^{\mathrm{pr},\ell} = 1, \qquad \sum_{k=1}^{\mathcal{K}} \alpha_k^\ell = m, \qquad \mathbb{E}[\Lambda_k^\ell] = \widehat{q}_k^{\mathrm{pr},\ell}.$$

Thus, the historical sample $\boldsymbol{d}^{\mathrm{pr},\ell}$ determines the prior mean profile $\widehat{q}_k^{\mathrm{pr},\ell}$, while m separately controls its concentration. Note that in our construction, increasing the historical sample size does not increase the chosen concentration m.

The posterior update in (15) remains unchanged:

$$\omega_k^\ell = d_k^\ell + m\,\widehat{q}_k^{\mathrm{pr},\ell}.$$

8 Monte Carlo Sampling of Future Empirical ROC Curves

Assume that we have observed test data $\mathfrak{d}^{\mathrm{ts}}$ and that, for a fixed probabilistic classifier g and a fixed grid $\boldsymbol{\gamma}$, we have obtained the observed histograms $\boldsymbol{d}^+$ and $\boldsymbol{d}^-$. Given the prior parameters $\boldsymbol{\alpha}^+$ and $\boldsymbol{\alpha}^-$, Algorithm 1 describes a Monte Carlo procedure for generating $\mathcal{S}$ realizations of the future empirical ROC curve where $\mathfrak{s}$ is used to index the Monte Carlo samples.

The inputs of this algorithm are the observed histograms $\boldsymbol{d}^+$ and $\boldsymbol{d}^-$, the prior parameters $\boldsymbol{\alpha}^+$ and $\boldsymbol{\alpha}^-$, the number of bins $\mathcal{K}$, and the desired number of Monte Carlo samples $\mathcal{S}$.

In lines (2–5), the algorithm first computes the numbers of ground-truth positive and negative observations n_+^{ts} and n_-^{ts}, and the posterior Dirichlet parameters $\boldsymbol{\omega}^+$ and $\boldsymbol{\omega}^-$. It assumes $n_+^{\mathrm{ts}} > 0$ and $n_-^{\mathrm{ts}} > 0$.

For each Monte Carlo sample $\mathfrak{s}$, the algorithm draws class-specific probability vectors from the posterior Dirichlet distributions in (15) (lines 7–8), and then draws posterior predictive histograms from the corresponding multinomial distributions

Algorithm 1 Posterior predictive sampling of future empirical ROC curves on the score grid

1: **function** ROC-SAMPLER($\boldsymbol{d}^+, \boldsymbol{d}^-, \boldsymbol{\alpha}^+, \boldsymbol{\alpha}^-, \mathcal{K}, \mathcal{S}$)
2: $\quad n_+^{\text{ts}} \leftarrow \mathbf{1}_{\mathcal{K}}^\top \boldsymbol{d}^+$
3: $\quad n_-^{\text{ts}} \leftarrow \mathbf{1}_{\mathcal{K}}^\top \boldsymbol{d}^-$
4: $\quad \boldsymbol{\omega}^+ \leftarrow \boldsymbol{\alpha}^+ + \boldsymbol{d}^+$
5: $\quad \boldsymbol{\omega}^- \leftarrow \boldsymbol{\alpha}^- + \boldsymbol{d}^-$
6: $\quad$ **for** $\mathfrak{s} \leftarrow 1$ to $\mathcal{S}$ **do**
7: $\quad\quad$ Draw $\boldsymbol{\lambda}^{+,(\mathfrak{s})}$ from Dirichlet($\boldsymbol{\omega}^+$)
8: $\quad\quad$ Draw $\boldsymbol{\lambda}^{-,(\mathfrak{s})}$ from Dirichlet($\boldsymbol{\omega}^-$)
9: $\quad\quad$ Draw $\widetilde{\boldsymbol{d}}^{+,(\mathfrak{s})}$ from Multinomial $\left(n_+^{\text{ts}}, \boldsymbol{\lambda}^{+,(\mathfrak{s})}\right)$
10: $\quad\quad$ Draw $\widetilde{\boldsymbol{d}}^{-,(\mathfrak{s})}$ from Multinomial $\left(n_-^{\text{ts}}, \boldsymbol{\lambda}^{-,(\mathfrak{s})}\right)$
11: $\quad\quad$ **for** $k \leftarrow 0$ to $\mathcal{K}-1$ **do**
12: $\quad\quad\quad \widetilde{t}_k^{+,(\mathfrak{s})} \leftarrow \sum_{j=k+1}^{\mathcal{K}} \widetilde{d}_j^{+,(\mathfrak{s})}$
13: $\quad\quad\quad \widetilde{t}_k^{-,(\mathfrak{s})} \leftarrow \sum_{j=k+1}^{\mathcal{K}} \widetilde{d}_j^{-,(\mathfrak{s})}$
14: $\quad\quad\quad \widetilde{\boldsymbol{r}}_k^{(\mathfrak{s})} \leftarrow \left(\widetilde{t}_k^{-,(\mathfrak{s})}/n_-^{\text{ts}}, \widetilde{t}_k^{+,(\mathfrak{s})}/n_+^{\text{ts}}\right)^\top$
15: $\quad\quad$ **end for**
16: $\quad\quad \widetilde{\boldsymbol{r}}_{\mathcal{K}}^{(\mathfrak{s})} \leftarrow (0,0)^\top$
17: $\quad\quad \widetilde{\boldsymbol{r}}_{\mathcal{K}:0}^{(\mathfrak{s})} \leftarrow \left[\widetilde{\boldsymbol{r}}_{\mathcal{K}}^{(\mathfrak{s})}, \widetilde{\boldsymbol{r}}_{\mathcal{K}-1}^{(\mathfrak{s})}, \ldots, \widetilde{\boldsymbol{r}}_0^{(\mathfrak{s})}\right]$
18: $\quad$ **end for**
19: $\quad$ **return** $\widetilde{\boldsymbol{r}}_{\mathcal{K}:0}^{(1)}, \ldots, \widetilde{\boldsymbol{r}}_{\mathcal{K}:0}^{(\mathcal{S})}$
20: **end function**

(lines 9–10). From these histograms, it computes (lines 11–16) realizations of the future empirical ROC points $\widetilde{\boldsymbol{r}}_0^{(\mathfrak{s})}, \ldots, \widetilde{\boldsymbol{r}}_{\mathcal{K}}^{(\mathfrak{s})}$ where $\widetilde{\boldsymbol{r}}_0^{(\mathfrak{s})} = (1,1)^\top$ and $\widetilde{\boldsymbol{r}}_{\mathcal{K}}^{(\mathfrak{s})} = (0,0)^\top$, and where

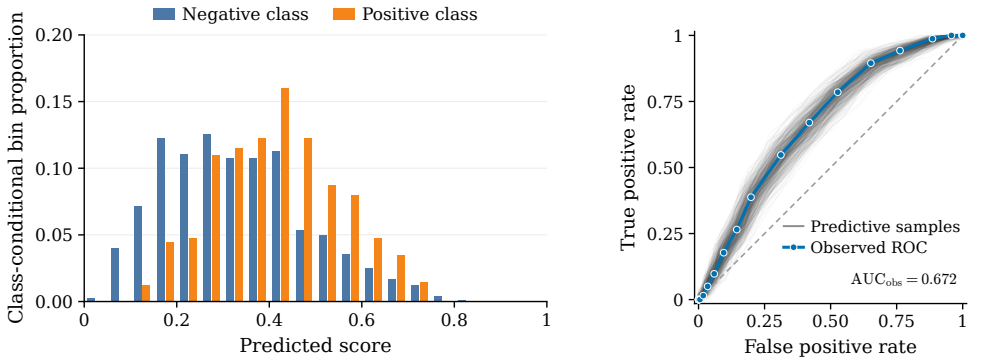
$$\widetilde{\boldsymbol{r}}_k^{(\mathfrak{s})} = \left(\widetilde{\text{fpr}}_k^{(\mathfrak{s})}, \widetilde{\text{tpr}}_k^{(\mathfrak{s})}\right)^\top.$$

The output for the $\mathfrak{s}$th Monte Carlo sample is a matrix $\widetilde{\boldsymbol{r}}_{\mathcal{K}:0}^{(\mathfrak{s})} \in [0,1]^{2\times(\mathcal{K}+1)}$ where the columns are ordered from $(0,0)^\top$ to $(1,1)^\top$, corresponding to decreasing the threshold values (line 17). Finally, the algorithm returns the collection of $\mathcal{S}$ sampled curves (line 19).

Note that lines 7–10 of Algorithm 1 implement a hierarchical model. Equivalently, it can be seen as a one-step procedure that draws posterior predictive histograms directly from the Dirichlet–Multinomial distribution [18]

$$\widetilde{\boldsymbol{D}}^\ell \mid \boldsymbol{D}^\ell = \boldsymbol{d}^\ell \sim \text{Dirichlet-Multinomial}\left(n_\ell^{\text{ts}}, \boldsymbol{\omega}^\ell\right), \qquad \ell \in \{+,-\}. \tag{32}$$

Figure 5 illustrates Algorithm 1 on a simulated binary classification example. A logistic regression model is fitted on synthetic data, and its predicted scores are evaluated on an independent test sample. We use $\mathcal{K} = 20$ fixed uniform bins with a



(a) Observed class-conditional score histograms

(b) Sampled future empirical ROC curves

Fig. 5 Posterior predictive sampling of future empirical ROC curves on the score grid. The left panel shows the observed class-conditional score histograms computed from the observed test dataset $\mathfrak{d}^{\mathrm{ts}}$. It corresponds to the observed $\boldsymbol{d}^{+}/n_{+}^{\mathrm{ts}}$ and $\boldsymbol{d}^{-}/n_{-}^{\mathrm{ts}}$. The right panel shows the sampled posterior predictive curves generated by Algorithm 1

symmetric uniform Dirichlet prior with total prior mass $m = \mathcal{K}$, so that $\alpha_k^{+} = \alpha_k^{-} = 1$ for every bin (see (20)). The algorithm generates $\mathcal{S} = 500$ posterior predictive samples of the future empirical ROC curve. The left panel of Figure 5 shows the observed class-conditional score histograms computed from $\mathfrak{d}^{\mathrm{ts}}$. The right panel shows the sampled curves, together with the observed empirical ROC curve on the score grid $\boldsymbol{\gamma}$.

There are two ways to generate ROC curves under the Bayesian model from the posterior distribution: the latent approach and the predictive approach. In the latent approach, we sample the vectors $\boldsymbol{\Lambda}^{+}$ and $\boldsymbol{\Lambda}^{-}$ from their posterior distribution and construct realizations of the discretized population ROC curve under the Bayesian model directly from these samples. This is a way to model the uncertainty about the population ROC curve only. In the posterior predictive approach, one additionally draws new class-specific histograms from the multinomial model and reconstructs realizations of the future empirical ROC curve on the score grid from these simulated counts; this incorporates both parameter uncertainty and the sampling variability that would be observed on a future test sample. This distinction is consistent with the interpretation of posterior predictive confusion matrices discussed in [11]. In this paper, we focus on the posterior predictive approach, because our objective is to quantify the variability of future empirical ROC curves. This posterior predictive construction is the one implemented in Algorithm 1.

9 Posterior Predictive Credible Bands for the Future Empirical ROC Curve

Macskassy et al. [4] study "future-curve" bands: regions intended to contain the empirical ROC curve obtained when the same trained classifier is evaluated on a new test

sample. We pursue this predictive aim using a Bayesian model. Specifically, we construct posterior predictive credible bands that describe uncertainty in all ROC points on the chosen threshold grid.

The future empirical ROC curve is represented by

$$\widetilde{\boldsymbol{R}}_{\mathcal{K}:0} := \left[\widetilde{\boldsymbol{R}}_{\mathcal{K}}, \ldots, \widetilde{\boldsymbol{R}}_0\right] \in [0,1]^{2\times(\mathcal{K}+1)}.$$

A simultaneous posterior predictive credible band with credible level $1-\delta$ is a collection of sets $\mathcal{C}_0, \ldots, \mathcal{C}_{\mathcal{K}} \subseteq [0,1]^2$ such that

$$\mathbb{P}\Big(\widetilde{\boldsymbol{R}}_k \in \mathcal{C}_k \text{ for all } k = 0, \ldots, \mathcal{K} \;\Big|\; \boldsymbol{D}^+ = \boldsymbol{d}^+, \boldsymbol{D}^- = \boldsymbol{d}^-\Big) \geq 1 - \delta$$

where $\mathcal{C}_k$ is the credible region assigned to the kth future empirical ROC point and $\delta \in (0,1)$ is the error probability.

Budig et al. [19] consider a related prediction problem for overdispersed multinomial counts, whose variability exceeds that predicted by a standard multinomial model. One of their Bayesian methods compares simulated future counts with their predictive means, expressing each difference in units of its predictive standard deviation. For each simulated sample, they retain the largest absolute standardized difference across categories. They use the distribution of these maxima to set the widths of predictive intervals intended to cover all category counts together. Our construction below uses a related idea for ROC curves, based on the largest Euclidean distance between a future ROC point and its predictive mean.

For a Euclidean construction, let us define

$$Q := \max_{0\leq k\leq \mathcal{K}} \left\| \widetilde{\boldsymbol{R}}_k - \boldsymbol{\mu}_k \right\|_2, \qquad \boldsymbol{\mu}_k := \mathbb{E}\Big[\widetilde{\boldsymbol{R}}_k \mid \boldsymbol{D}^+ = \boldsymbol{d}^+, \boldsymbol{D}^- = \boldsymbol{d}^-\Big],$$

so that the event $Q \leq \rho$ is equivalent to the simultaneous inclusion of all future empirical ROC points in Euclidean balls of common radius ρ centered at $\boldsymbol{\mu}_0, \ldots, \boldsymbol{\mu}_{\mathcal{K}}$ where (see Appendix E)

$$\boldsymbol{\mu}_k = \left(\sum_{j=k+1}^{\mathcal{K}} \frac{\omega_j^-}{\omega_\bullet^-}, \sum_{j=k+1}^{\mathcal{K}} \frac{\omega_j^+}{\omega_\bullet^+} \right)^\top \tag{33}$$

with $\omega_\bullet^\ell := \sum_{j=1}^{\mathcal{K}} \omega_j^\ell$. The point $\boldsymbol{\mu}_k$ gives the average false-positive and true-positive rates that our model predicts for the kth point of a future empirical ROC curve. Considering all points on the chosen grid, Q measures the largest distance between a future empirical ROC point and its corresponding point $\boldsymbol{\mu}_k$.

To construct a simultaneous Euclidean credible band, we seek a common radius that is sufficiently large so that all future empirical ROC points are contained in the band with posterior predictive probability at least $1-\delta$. The smallest radius satisfying

Algorithm 2 Constructing a Euclidean posterior predictive credible band

function CREDIBLE-BAND($\boldsymbol{\omega}^+, \boldsymbol{\omega}^-, \widetilde{\boldsymbol{r}}^{(1)}_{\mathcal{K}:0}, \ldots, \widetilde{\boldsymbol{r}}^{(\mathcal{S})}_{\mathcal{K}:0}, \mathcal{K}, \mathcal{S}, \delta$)

Compute $\omega^\ell_\bullet \leftarrow \sum_{j=1}^{\mathcal{K}} \omega^\ell_j$, for $\ell \in \{+, -\}$.

Compute the posterior predictive means $\boldsymbol{\mu}_k$ from (33), for $k = 0, \ldots, \mathcal{K}$,

$$\boldsymbol{\mu}_k \leftarrow \begin{pmatrix} \sum_{j=k+1}^{\mathcal{K}} \frac{\omega^-_j}{\omega^-_\bullet} \\ \sum_{j=k+1}^{\mathcal{K}} \frac{\omega^+_j}{\omega^+_\bullet} \end{pmatrix}.$$

Compute the maximum Euclidean deviation of each sampled curve, for $\mathfrak{s} = 1, \ldots, \mathcal{S}$,

$$q^{(\mathfrak{s})} \leftarrow \max_{0 \le k \le \mathcal{K}} \left\| \widetilde{\boldsymbol{r}}^{(\mathfrak{s})}_k - \boldsymbol{\mu}_k \right\|_2.$$

Let $q_{[1]} \le \cdots \le q_{[\mathcal{S}]}$ be the ordered values of $q^{(1)}, \ldots, q^{(\mathcal{S})}$ and set

$$\widehat{\rho}_{1-\delta} \leftarrow q_{[\lceil (1-\delta)\mathcal{S} \rceil]}.$$

Define the credible regions for $k = 0, \ldots, \mathcal{K}$,

$$\widehat{\mathcal{C}}^{\mathrm{ROC}}_{k,1-\delta} \leftarrow \left\{ \boldsymbol{z} \in [0,1]^2 : \|\boldsymbol{z} - \boldsymbol{\mu}_k\|_2 \le \widehat{\rho}_{1-\delta} \right\}.$$

return $\widehat{\mathcal{C}}^{\mathrm{ROC}}_{0,1-\delta}, \ldots, \widehat{\mathcal{C}}^{\mathrm{ROC}}_{\mathcal{K},1-\delta}$.

end function

this requirement is

$$\rho_{1-\delta} := \inf \left\{ \rho \ge 0 : \mathbb{P}\big(Q \le \rho \,\big|\, \boldsymbol{D}^+ = \boldsymbol{d}^+, \boldsymbol{D}^- = \boldsymbol{d}^-\big) \ge 1 - \delta \right\}.$$

We can now assign to each grid point k the ball centered at $\boldsymbol{\mu}_k$ with common radius $\rho_{1-\delta}$

$$\mathcal{C}^{\mathrm{ROC}}_{k,1-\delta} := \left\{ \boldsymbol{z} \in [0,1]^2 : \|\boldsymbol{z} - \boldsymbol{\mu}_k\|_2 \le \rho_{1-\delta} \right\}, \qquad k = 0, \ldots, \mathcal{K}.$$

After running Algorithm 1, we have a collection $\widetilde{\boldsymbol{r}}^{(1)}_{\mathcal{K}:0}, \ldots, \widetilde{\boldsymbol{r}}^{(\mathcal{S})}_{\mathcal{K}:0}$ where each $\widetilde{\boldsymbol{r}}^{(\mathfrak{s})}_{\mathcal{K}:0}$ is a posterior predictive realization of the future empirical ROC curve on the score grid $\boldsymbol{\gamma}$. In Algorithm 2, we will use these samples to construct a simultaneous Euclidean credible band $\widehat{\mathcal{C}}^{\mathrm{ROC}}_{k,1-\delta}$ around the posterior predictive means $\boldsymbol{\mu}_k$.

To do that, for each simulated curve, we define the maximum pointwise Euclidean deviation from this center curve by

$$q^{(\mathfrak{s})} := \max_{0 \le k \le \mathcal{K}} \left\| \widetilde{\boldsymbol{r}}^{(\mathfrak{s})}_k - \boldsymbol{\mu}_k \right\|_2, \qquad \mathfrak{s} = 1, \ldots, \mathcal{S}.$$

Let $q_{[1]} \le \cdots \le q_{[\mathcal{S}]}$ denote the ordered values of $q^{(1)}, \ldots, q^{(\mathcal{S})}$, and define the $(1-\delta)$-quantile of the Monte Carlo sample by

$$\widehat{\rho}_{1-\delta} := q_{[\lceil (1-\delta)\mathcal{S} \rceil]}.$$

The resulting Monte Carlo approximation is the collection

$$\widehat{\mathcal{C}}^{\mathrm{ROC}}_{k,1-\delta} := \left\{ \boldsymbol{z} \in [0,1]^2 : \|\boldsymbol{z} - \boldsymbol{\mu}_k\|_2 \le \widehat{\rho}_{1-\delta} \right\}, \qquad k = 0, \ldots, \mathcal{K}. \tag{34}$$

10 Posterior Predictive Distribution of the Discretized AUC

The area under the ROC curve (AUC) is defined as the area enclosed between the ROC curve and the horizontal axis. It quantifies in one scalar the overall discriminative performance of the probabilistic classifier g. We approximate this by applying the trapezoidal rule to each sampled future empirical ROC curve on the score grid generated by Algorithm 1. From the $\mathfrak{s}$th posterior predictive draw $\widetilde{\boldsymbol{r}}^{(\mathfrak{s})}_{\mathcal{K}:0}$, the corresponding grid-based posterior predictive AUC realization is

$$\widetilde{a}^{(\mathfrak{s})}_{\boldsymbol{\gamma}} := \sum_{k=0}^{\mathcal{K}-1} \frac{\widetilde{\mathrm{tpr}}^{(\mathfrak{s})}_k + \widetilde{\mathrm{tpr}}^{(\mathfrak{s})}_{k+1}}{2} \left(\widetilde{\mathrm{fpr}}^{(\mathfrak{s})}_k - \widetilde{\mathrm{fpr}}^{(\mathfrak{s})}_{k+1} \right), \qquad \mathfrak{s} = 1, \ldots, \mathcal{S}.$$

Hence, $\widetilde{a}^{(1)}_{\boldsymbol{\gamma}}, \ldots, \widetilde{a}^{(\mathcal{S})}_{\boldsymbol{\gamma}}$ form an i.i.d. Monte Carlo sample. Let $\widetilde{a}_{[1]} \le \cdots \le \widetilde{a}_{[\mathcal{S}]}$ denote the ordered values of $\widetilde{a}^{(1)}_{\boldsymbol{\gamma}}, \ldots, \widetilde{a}^{(\mathcal{S})}_{\boldsymbol{\gamma}}$. For a credibility level $1-\delta$, we define the Monte Carlo posterior predictive credible interval

$$\widehat{\mathcal{C}}^{\mathrm{AUC}}_{1-\delta} := \left[\widetilde{a}_{[\lceil (\delta/2)\mathcal{S} \rceil]},\ \widetilde{a}_{[\lceil (1-\delta/2)\mathcal{S} \rceil]} \right]. \tag{35}$$

11 Experiments

We evaluate how well DiP-ROC predicts the ROC curve and AUC in another evaluation of a fixed classifier, using synthetic and real-world datasets. The main results use a uniform ten-bin grid $\boldsymbol{\gamma}$ and nominal levels of 80% and 95%. Each configuration is repeated $\mathcal{B}$ times with newly generated samples or a new data split, indexed by $\mathfrak{b} = 1, \ldots, \mathcal{B}$. Within each replication, $\mathcal{S} = 4000$ predictive draws represent possible future evaluations given the test sample and prior parameters. Table 1 summarizes the protocols. All code is available on GitHub to support reproducibility: https://github.com/ocaelen/bayesian_roc.

11.1 Evaluation protocol

Each replication $\mathfrak{b}$ uses a test sample $\mathfrak{d}^{\mathrm{ts}}$ and a separate external sample $\mathfrak{d}^{\mathrm{ext}}$, both with class counts n^{ts}_{+} and n^{ts}_{-}. Table 2 distinguishes these samples from the training and historical samples used in the respective experiments. For real data, the training,

Table 1 Overview of the two experimental protocols. Test and external class counts match in both experiments.

	Synthetic data	**Real-world datasets**
Data	Three score scenarios and five class-count pairs	Nine binary classification datasets
Classifier	Scores generated directly	Logistic regression fitted on training data
DiP-ROC prior	Uniform, linear, and histograms, each at $m = 1, 5, 20$	Uniform, $m = 1$
Grid bins $\mathcal{K}$	10	$5, 10, 20$
Replications $\mathcal{B}$	10000 per configuration	300 per dataset
Predictive draws $\mathcal{S}$	4000 per setting and replication	4000 per grid and replication
Nominal levels $1 - \delta$	80% and 95%	80% and 95%

Table 2 Role of the observations at each stage.

Sample	**Purpose**	**Used in**
Training, $\mathfrak{d}^{\mathrm{tr}}$	Fit the classifier	Real data only
Historical, $\mathfrak{d}^{\mathrm{pr}}$	Estimate the AUC summary or class-specific histograms used to construct informed priors	Synthetic data only
Test, $\mathfrak{d}^{\mathrm{ts}}$	Construct the predictive ROC band and AUC interval	Both experiments
External, $\mathfrak{d}^{\mathrm{ext}}$	Check whether the predictions contain the new empirical ROC curve and AUC	Both experiments

test, and external samples are disjoint; the classifier g is fitted on the training sample and then held fixed.

For each setting and grid $\boldsymbol{\gamma}$, the predictive draws $\widetilde{\boldsymbol{r}}^{(\mathfrak{s})}_{\mathcal{K}:0}$, $\mathfrak{s} = 1, \ldots, \mathcal{S}$, are used to construct both the ROC bands in (34) and the AUC intervals in (35) at both nominal levels. Coverage is then assessed on the external sample $\mathfrak{d}^{\mathrm{ext}}$, which is not used to construct predictions or tune their parameters.

11.1.1 Coverage.

In replication $\mathfrak{b}$, let $\widehat{\boldsymbol{r}}^{(\mathfrak{b})}_{\mathrm{ext},k}$ and $\widehat{a}^{(\mathfrak{b})}_{\mathrm{ext}}$ denote the external ROC points and AUC on the same grid $\boldsymbol{\gamma}$. ROC coverage requires all external grid points to lie within their corresponding predictive regions. AUC coverage requires the external AUC to lie within its predictive interval. The coverage estimates over $\mathcal{B}$ replications are

$$\widehat{\mathrm{COV}}^{\mathrm{ROC}}_{\mathrm{ext}} = \frac{1}{\mathcal{B}} \sum_{\mathfrak{b}=1}^{\mathcal{B}} \mathbb{1}\left\{ \widehat{\boldsymbol{r}}^{(\mathfrak{b})}_{\mathrm{ext},k} \in \widehat{\mathcal{C}}^{\mathrm{ROC},(\mathfrak{b})}_{k,1-\delta} \text{ for all } k = 0, \ldots, \mathcal{K} \right\}$$

and

$$\widehat{\mathrm{cov}}_{\mathrm{ext}}^{\mathrm{AUC}} = \frac{1}{\mathcal{B}} \sum_{\mathfrak{b}=1}^{\mathcal{B}} \mathbb{1}\left\{ \widehat{a}_{\mathrm{ext}}^{(\mathfrak{b})} \in \widehat{\mathcal{C}}_{1-\delta}^{\mathrm{AUC},(\mathfrak{b})} \right\}.$$

We report these proportions and compare them with the corresponding target nominal level.

11.1.2 Prediction size (sharpness).

Sharpness describes how concentrated the predictions are; following [20], we assess it alongside coverage. We report averages across replications of the ROC-band radii $\widehat{\rho}_{1-\delta}^{(\mathfrak{b})}$ and AUC-interval widths $u_{1-\delta}^{(\mathfrak{b})} - l_{1-\delta}^{(\mathfrak{b})}$, where l and u are the interval endpoints. Smaller values are useful only alongside adequate coverage in the same configuration.

11.2 Synthetic-data experiment

We generate scores directly, without fitting a classifier, and check whether the resulting predictions cover an independent evaluation. We vary how well the score distributions separate the two classes, the number of observations in each class, and the prior settings. Table 3 defines the fifteen configurations; Figure 6 shows the score distributions.

Table 3 The fifteen synthetic configurations: combine each score scenario in panel A with each class-count pair in panel B. External samples use the same class counts as test samples.

A. Class-conditional score distributions

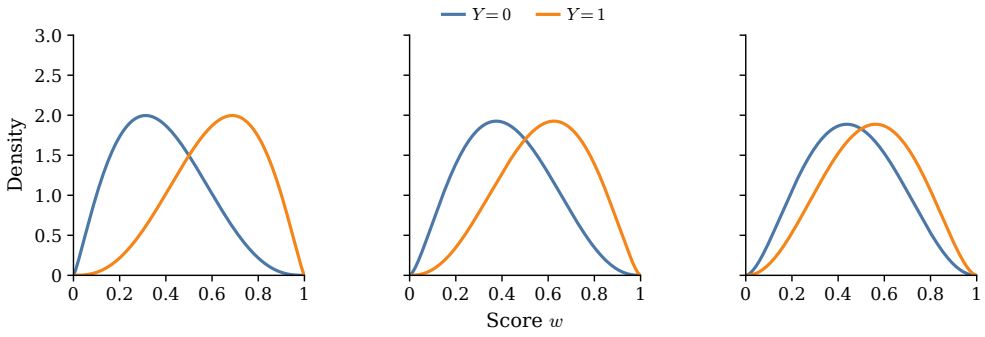
Scenario	**Positive scores** $(Y = 1)$	**Negative scores** $(Y = 0)$
Easy	Beta$(3.75, 2.25)$	Beta$(2.25, 3.75)$
Medium	Beta$(3.5, 2.5)$	Beta$(2.5, 3.5)$
Hard	Beta$(3.25, 2.75)$	Beta$(2.75, 3.25)$

B. Observations in each test or external sample

Class-count setting	**Positive** n_+^{ts}	**Negative** n_-^{ts}
Small, balanced	20	20
Medium, balanced	100	100
Large, balanced	500	500
Fewer positive observations	20	500
Fewer negative observations	500	20

11.2.1 Shared samples and predictive analysis.

We use $\mathcal{K} = 10$, with $\gamma_k = k/10$ for $k = 0, \ldots, 10$, and the bins $\mathcal{I}_k(\boldsymbol{\gamma})$ defined in (2). In each of $\mathcal{B} = 10000$ replications per configuration, we generate test and external



(a) Easy (b) Medium (c) Hard

Fig. 6 Class-conditional score densities for the three simulation scenarios. The overlap between the distributions increases from Easy to Hard

samples $\mathfrak{d}^{\mathrm{ts},(\mathfrak{b})}$ and $\mathfrak{d}^{\mathrm{ext},(\mathfrak{b})}$ with the displayed class counts n_+^{ts} and n_-^{ts}, together with one historical sample $\mathfrak{d}^{\mathrm{pr},(\mathfrak{b})}$ containing $n_+^{\mathrm{pr}} = 500$ positive and $n_-^{\mathrm{pr}} = 500$ negative observations. The dataset $\mathfrak{d}^{\mathrm{pr},(\mathfrak{b})}$ is only used to construct the informed prior. All three samples use the same class-conditional distributions from Table 3. They are mutually independent, with independent scores within and between classes; new samples are drawn independently in every replication. Table 4 describes the settings for the method DiP-ROC. We vary the prior family and its concentration $m \in \{1, 5, 20\}$. All nine settings share the test observations $\mathfrak{d}^{\mathrm{ts},(\mathfrak{b})}$ and external observations $\mathfrak{d}^{\mathrm{ext},(\mathfrak{b})}$, and the six informed settings share the historical sample $\mathfrak{d}^{\mathrm{pr},(\mathfrak{b})}$.

Table 4 The nine DiP-ROC settings evaluated in every synthetic configuration.

	Setting	m	**Historical information received**
1	DiP-ROC uniform	1	None
2	DiP-ROC uniform	5	None
3	DiP-ROC uniform	20	None
4	DiP-ROC linear	1	One empirical grid-AUC summary
5	DiP-ROC linear	5	One empirical grid-AUC summary
6	DiP-ROC linear	20	One empirical grid-AUC summary
7	DiP-ROC histograms	1	Two class-specific score histograms
8	DiP-ROC histograms	5	Two class-specific score histograms
9	DiP-ROC histograms	20	Two class-specific score histograms

The *uniform* settings use the symmetric Dirichlet prior in (20), $\alpha_k^\ell = m/\mathcal{K}$. The *linear* settings use (27) with the same target for all three concentrations $m \in \{1, 5, 20\}$:

$$a_0^{(\mathfrak{b})} = \min\Big\{0.86, \max\Big\{0.5, \widehat{a}_{\mathrm{pr}}^{(\mathfrak{b})}\Big\}\Big\},$$

where $\widehat{a}_{\mathrm{pr}}^{(\mathfrak{b})}$ is the historical sample's AUC computed from its empirical histograms on the same grid $\boldsymbol{\gamma}$ (using (19) with $q_k^\ell = d_k^{\mathrm{pr},\ell}/n_\ell^{\mathrm{pr}}$). For $\mathcal{K} = 10$, the cap $0.86 < 13/15$ ensures feasibility by (29). The *histogram* settings use (31) to incorporate the two class-specific score histograms $\boldsymbol{d}^{\mathrm{pr},+}$ and $\boldsymbol{d}^{\mathrm{pr},-}$ from the historical sample.

11.2.2 Results

Table 5 gives an overview of coverage and prediction size across the nine prior settings. Table 6 then shows how prediction size changes with class counts under the weak uniform prior. Table 7 turns to coverage in the Easy, Medium, and Hard scenarios. Finally, Table 8 brings coverage and prediction size together for each class-count setting, covering both balanced and imbalanced samples.

Table 5 summarizes results averaged equally over the fifteen configurations. With weak priors ($m = 1$), the three prior families give similar results. The uniform prior provides a useful baseline, with average coverage close to nominal. At the 95% level, simultaneous ROC coverage is 94.70% and AUC coverage is 95.01%. Note that these averages can conceal shortfalls: in the Easy scenario with $(20, 500)$, the histogram prior at $m = 1$ gives 92.69% AUC coverage at the 95% level, compared with 94.96% for the uniform prior at the same concentration.

Historical information improves precision with the histogram prior at $m = 5$. At the 95% level, its mean ROC radius is 3.8% smaller than that of the weak uniform prior. Its mean AUC interval width is 5.1% smaller. Average coverage remains close to nominal, at 95.32% for ROC and 95.19% for AUC. These gains use independent historical data drawn from the same populations as the test and external samples.

Increasing the concentration to $m = 20$ further reduces mean ROC radii and AUC widths for both informed priors. The histogram prior gives the smallest prediction sizes at both nominal levels. However, average coverage exceeds nominal for both informed priors at this concentration. For example, histogram ROC coverage reaches 83.74% at the 80% level and 96.43% at the 95% level. Stronger informed priors therefore give smaller predictions, but do not necessarily bring coverage closer to its target.

The strong uniform prior shows why the choice of prior also matters. At $m = 20$, its average AUC coverage falls to 64.12% and 86.46% at the 80% and 95% levels. This occurs despite its AUC intervals having the largest mean widths in the table.

Table 5 DiP-ROC synthetic results, averaged equally over the fifteen configurations. Column headings give nominal levels. Mean radii and AUC widths describe prediction size and should be read alongside coverage.

		ROC coverage (%)		AUC coverage (%)		Mean ROC radius		Mean AUC width	
Prior	m	80%	95%	80%	95%	80%	95%	80%	95%
Uniform	1	79.16	94.70	80.00	95.01	0.2049	0.2703	0.1685	0.2558
Uniform	5	80.66	95.30	80.34	95.13	0.2029	0.2630	0.1768	0.2681
Uniform	20	76.69	94.35	64.12	86.46	0.1936	0.2492	0.1814	0.2747
Linear	1	79.42	94.75	79.47	94.58	0.2050	0.2704	0.1649	0.2503
Linear	5	80.76	95.35	81.82	95.94	0.2016	0.2616	0.1645	0.2499
Linear	20	81.37	95.94	84.95	97.08	0.1915	0.2469	0.1600	0.2431
Histograms	1	80.07	94.93	79.08	94.31	0.2051	0.2708	0.1637	0.2485
Histograms	5	81.01	95.32	80.48	95.19	0.2002	0.2601	0.1599	0.2428
Histograms	20	83.74	96.43	82.33	96.10	0.1874	0.2431	0.1505	0.2288

Table 6 shows how class counts affect prediction size under the weak uniform prior ($m = 1$). For balanced test samples, the number of observations per class increases from 20 to 500. At the 95% level, the mean AUC interval width falls from 0.416 to 0.084. The corresponding mean ROC radius falls from 0.398 to 0.083. Both sizes decrease by about a factor of five for a twenty-fivefold increase in each class count. This pattern is consistent with inverse-square-root scaling.

Increasing only one class count gives a smaller gain in precision. With 20 observations in one class and 500 in the other, the mean 95% AUC interval width remains 0.295. These imbalanced samples contain 520 observations each, compared with 200 for the balanced $(100, 100)$ samples. Yet their ROC bands and AUC intervals are wider. The smaller class therefore limits precision, even when the total sample is large. DiP-ROC reflects this limitation by accounting for the information available in each class.

Table 6 Synthetic uncertainty size for the weak uniform reference ($m = 1$), averaged equally over the three score scenarios. Counts apply to each test sample.

		Mean ROC radius		Mean AUC width	
n_+^{ts}	n_-^{ts}	80%	95%	80%	95%
20	20	0.312	0.398	0.277	0.416
100	100	0.146	0.184	0.124	0.188
500	500	0.066	0.083	0.055	0.084
20	500	0.250	0.343	0.193	0.295
500	20	0.250	0.344	0.194	0.295

Table 7 separates coverage by score scenario. The strong uniform prior performs poorly in the Easy scenario. For the individual Easy $(20, 20)$ configuration, $m = 20$ yields only 58.04% AUC coverage at the 95% level. The mean interval width is nevertheless larger than with $m = 1$. Scenario averages can therefore hide substantial shortfalls. This is consistent with the prior properties described in Section 7.1. The uniform prior has an expected AUC of 0.5. The population AUC in the Easy scenario is much higher. With small samples, a strong uniform prior can pull the predicted AUC downwards. This helps explain why even wide intervals can miss future AUC values.

The scenario averages also show that DiP-ROC can maintain coverage close to nominal across Easy, Medium, and Hard. With the weak uniform prior ($m = 1$), mean 95% ROC coverage ranges from 94.67% to 94.72%. Mean 95% AUC coverage ranges from 94.93% to 95.17%. With histogram priors at $m = 5$, both 95% coverage measures stay within half a percentage point of nominal in every scenario. This complements the smaller overall mean prediction sizes reported in Table 5. The Easy scenario also shows the benefit of historical information when the prior receives more weight. At $m = 20$, average 95% AUC coverage increases from 75.74% with the uniform prior to 96.06% with the histogram prior. This is a substantial improvement, although coverage is then slightly above nominal.

Table 7 Synthetic coverage by score scenario, averaged equally over the five class-count configurations. Column headings give nominal levels. All nine prior settings are shown, with $\mathcal{B} = 10000$, $\mathcal{S} = 4000$, and $\mathcal{K} = 10$. In the Setting column, DiP-U, DiP-L, and DiP-H denote uniform, linear, and histogram priors; the suffix gives m.

Scenario	Setting	ROC 80%	ROC 95%	AUC 80%	AUC 95%
Easy	DiP-U1	78.88	94.67	79.73	95.17
Easy	DiP-U5	80.44	95.28	78.11	94.20
Easy	DiP-U20	71.29	92.41	46.30	75.74
Easy	DiP-L1	79.42	94.83	78.62	94.21
Easy	DiP-L5	80.72	95.36	81.00	95.79
Easy	DiP-L20	80.96	95.77	84.25	96.96
Easy	DiP-H1	80.23	95.03	78.23	93.96
Easy	DiP-H5	81.09	95.49	79.76	95.00
Easy	DiP-H20	83.85	96.59	81.86	96.06
Medium	DiP-U1	78.97	94.72	79.95	94.94
Medium	DiP-U5	80.52	95.28	80.58	95.30
Medium	DiP-U20	77.43	94.65	65.44	88.36
Medium	DiP-L1	79.21	94.73	79.57	94.62
Medium	DiP-L5	80.63	95.34	81.91	95.97
Medium	DiP-L20	81.15	95.78	85.00	97.11
Medium	DiP-H1	79.93	94.93	79.19	94.38
Medium	DiP-H5	80.89	95.25	80.54	95.17
Medium	DiP-H20	83.50	96.36	82.26	96.05
Hard	DiP-U1	79.62	94.70	80.33	94.93
Hard	DiP-U5	81.02	95.34	82.33	95.90
Hard	DiP-U20	81.35	95.99	80.62	95.27
Hard	DiP-L1	79.65	94.70	80.22	94.91
Hard	DiP-L5	80.95	95.36	82.55	96.06
Hard	DiP-L20	81.99	96.26	85.60	97.18
Hard	DiP-H1	80.04	94.84	79.83	94.60
Hard	DiP-H5	81.06	95.22	81.13	95.41
Hard	DiP-H20	83.87	96.34	82.88	96.19

Table 8 gives the class-count results. Prior effects generally diminish with larger balanced samples. With unequal counts, the strong linear prior can give coverage well above nominal. For $(20, 500)$, the linear prior at $m = 20$ gives 98.41% AUC coverage at the 95% level and mean width 0.292. The histogram prior at the same concentration gives 96.55% coverage and width 0.259. These values are averages over the three scenarios.

For an evaluation with only 20 positive observations, relevant historical data can make predictions more informative. In the same $(20, 500)$ setting, the histogram prior at $m = 5$ gives average 95% AUC coverage of 94.96%. Its mean interval width is 0.277, compared with 0.295 under the weak uniform prior, a reduction of about 6%.

The analyst therefore obtains a narrower range of AUC values for a future evaluation while retaining average coverage close to the requested level. This is useful when historical data are available but collecting additional positive observations is difficult. The stronger histogram prior gives narrower intervals still, but $m = 5$ matches the requested coverage more closely in these experiments.

For the configurations studied, these results suggest using the weak uniform prior ($m = 1$) when relevant historical information is unavailable. When class counts are small, compare this baseline with a histogram prior at $m = 5$ if suitable historical data are available. The historical histograms should use the same classifier and score grid, with independent observations from the same populations. With 20 observations per class, this histogram prior reduces the mean 95% ROC radius by about 4% relative to the weak uniform prior. The mean 95% AUC interval width decreases by about 6%. Average coverage remains close to nominal, at 95.66% for ROC and 95.98% for AUC. With 500 observations per class, the two priors give the same mean prediction sizes to the reported precision. The additional precision from historical information is therefore more useful when observations are scarce.

Table 8 Synthetic coverage and prediction size by class counts, averaged equally over Easy, Medium, and Hard. Column headings give nominal levels; coverage is in percent. Counts apply to each test and external sample. Balanced configurations; imbalanced configurations follow in the continuation.

	ROC coverage (%)		AUC coverage (%)		Mean ROC radius		Mean AUC width	
Setting	80%	95%	80%	95%	80%	95%	80%	95%
$(n_+^{\mathrm{ts}}, n_-^{\mathrm{ts}}) = (20, 20)$								
DiP-U1	78.15	94.43	81.47	95.69	0.312	0.398	0.277	0.416
DiP-U5	81.55	95.66	78.70	94.66	0.310	0.389	0.287	0.430
DiP-U20	75.53	93.95	51.74	77.74	0.295	0.367	0.281	0.423
DiP-L1	77.82	94.28	80.73	95.17	0.312	0.397	0.269	0.405
DiP-L5	81.89	95.71	83.42	96.61	0.306	0.384	0.262	0.394
DiP-L20	83.43	96.80	85.89	97.45	0.287	0.359	0.243	0.366
DiP-H1	77.93	94.44	80.23	94.93	0.311	0.398	0.268	0.404
DiP-H5	81.66	95.66	82.13	95.98	0.303	0.382	0.260	0.391
DiP-H20	86.20	97.28	84.30	96.88	0.281	0.353	0.240	0.362
$(n_+^{\mathrm{ts}}, n_-^{\mathrm{ts}}) = (100, 100)$								
DiP-U1	79.58	94.95	80.22	95.03	0.146	0.184	0.124	0.188
DiP-U5	80.24	95.26	79.91	94.82	0.145	0.183	0.125	0.191
DiP-U20	76.81	94.09	61.81	84.46	0.143	0.179	0.128	0.195
DiP-L1	79.50	94.84	80.12	94.99	0.146	0.184	0.123	0.187
DiP-L5	80.31	95.16	80.87	95.33	0.145	0.182	0.122	0.186
DiP-L20	80.68	95.66	82.77	96.17	0.141	0.178	0.119	0.181
DiP-H1	79.51	94.88	80.11	94.93	0.146	0.184	0.123	0.187
DiP-H5	80.38	95.26	80.53	95.15	0.144	0.182	0.122	0.185
DiP-H20	82.72	96.23	81.61	95.67	0.140	0.176	0.118	0.180
$(n_+^{\mathrm{ts}}, n_-^{\mathrm{ts}}) = (500, 500)$								
DiP-U1	79.37	94.70	79.71	94.84	0.066	0.083	0.055	0.084
DiP-U5	79.55	94.74	79.63	94.79	0.066	0.083	0.055	0.085
DiP-U20	78.69	94.51	74.10	91.78	0.065	0.082	0.056	0.085
DiP-L1	79.33	94.66	79.76	94.81	0.066	0.083	0.055	0.084
DiP-L5	79.60	94.78	79.91	94.89	0.066	0.083	0.055	0.084
DiP-L20	79.75	94.97	80.32	95.12	0.065	0.082	0.055	0.084
DiP-H1	79.40	94.74	79.75	94.81	0.066	0.083	0.055	0.084
DiP-H5	79.60	94.76	79.86	94.81	0.066	0.083	0.055	0.084
DiP-H20	80.25	95.01	80.13	94.98	0.065	0.082	0.055	0.083

Table 8 Synthetic coverage and prediction size by class counts, averaged equally over Easy, Medium, and Hard. Column headings give nominal levels; coverage is in percent. Counts apply to each test and external sample. Imbalanced configurations (continued).

	ROC coverage (%)		AUC coverage (%)		Mean ROC radius		Mean AUC width	
Setting	80%	95%	80%	95%	80%	95%	80%	95%
$(n_+^{\mathrm{ts}}, n_-^{\mathrm{ts}}) = (20, 500)$								
DiP-U1	79.46	94.77	79.38	94.67	0.250	0.343	0.193	0.295
DiP-U5	81.06	95.51	81.89	95.74	0.247	0.330	0.208	0.318
DiP-U20	76.24	94.63	66.60	89.37	0.233	0.309	0.221	0.335
DiP-L1	80.43	95.00	78.51	93.83	0.251	0.344	0.189	0.288
DiP-L5	81.19	95.53	82.54	96.47	0.246	0.329	0.192	0.293
DiP-L20	81.67	96.19	87.86	98.41	0.232	0.308	0.192	0.292
DiP-H1	81.89	95.27	77.78	93.31	0.251	0.345	0.186	0.283
DiP-H5	81.90	95.48	80.04	94.96	0.244	0.327	0.181	0.277
DiP-H20	84.95	96.89	82.82	96.55	0.226	0.302	0.170	0.259
$(n_+^{\mathrm{ts}}, n_-^{\mathrm{ts}}) = (500, 20)$								
DiP-U1	79.22	94.64	79.25	94.84	0.250	0.344	0.194	0.295
DiP-U5	80.90	95.33	81.57	95.66	0.247	0.330	0.209	0.318
DiP-U20	76.19	94.58	66.36	88.92	0.233	0.309	0.221	0.335
DiP-L1	80.04	94.99	78.22	94.07	0.251	0.344	0.189	0.288
DiP-L5	80.83	95.57	82.37	96.41	0.246	0.329	0.192	0.293
DiP-L20	81.31	96.06	87.90	98.26	0.232	0.308	0.192	0.292
DiP-H1	81.59	95.33	77.55	93.59	0.251	0.345	0.186	0.284
DiP-H5	81.51	95.46	79.82	95.07	0.244	0.327	0.182	0.277
DiP-H20	84.59	96.73	82.80	96.41	0.226	0.302	0.170	0.259

11.3 Real-world-data experiment

This experiment assesses predictive coverage and prediction size across nine binary classification datasets. We use the weak uniform prior ($m = 1$) without historical information. We use three grid resolutions: $\mathcal{K} = 5, 10, 20$. Table 9 specifies the nine binary classification tasks, their sources, and the class counts in each held-out sample. For each dataset, the original target values satisfying the rule in Table 9 are encoded as $Y = 1$, and all other target values are encoded as $Y = 0$.

Table 9 Real-data sources and fixed binary encodings. Rules refer to original target labels, except for Abalone, where the target is the number of rings. The last two columns apply to each of the test and external samples; all remaining cases form training data.

Dataset	OpenML ID (version)	Source rule for $Y = 1$	Complete cases	n_+^{ts}	n_-^{ts}
Abalone	183 (1)	≥ 10	4177	416	419
Adult	1590 (2)	`>50K`	45222	2241	6802
Banknote Authentication	1462 (1)	`2`	1372	122	152
Diabetes	37 (1)	`tested_positive`	768	53	100
Haberman	43 (1)	`2`	306	16	45
Mammographic	45557 (2)	`1`	830	80	85
Titanic	40704 (2)	`1`	2201	142	298
Transfusion	1464 (1)	`2`	748	35	114
Contraceptive Method Choice	23 (1)	`2` or `3`	1473	168	125

11.3.1 Sampling and fitting.

We remove rows with missing values in any source column before excluding predictors or splitting the data. For each dataset, we independently repeat the random split into training, test, and external samples $\mathcal{B} = 300$ times. For each class, the test and external samples each receive one fifth of the observations, rounded down, and the remaining observations form the training sample. Preprocessing and an L_2-regularized logistic regression are fitted anew on training observations only. Each numerical variable is centered and scaled. For categorical variables, each category observed in training is represented by a one-hot encoded column. A category encountered only in a test or external sample receives zeros in all indicator columns for that variable. The same fitted transformation is applied to the test and external samples.

The classifier uses `C=1`, an intercept, no class weighting, `lbfgs`, tolerance 10^{-8}, and at most 10000 iterations, with no hyperparameter tuning. The fitted preprocessing and classifier produce positive-class probability scores for both held-out samples. Following Section 11.1, the test histograms are used to generate $\mathcal{S} = 4000$ posterior predictive ROC curves. We use these draws to construct both the ROC bands and AUC intervals at the 80% and 95% nominal levels. For the main ten-bin analysis, we then check coverage using the external ROC and trapezoidal AUC on the same ten-bin grid.

11.3.2 Results.

Table 10 summarizes performance across the nine datasets, giving each dataset equal weight. At $\mathcal{K} = 10$, mean AUC coverage is 81.37% and 95.11%; simultaneous ROC coverage is 79.85% and 95.00%. The mean coverages are closer to their nominal targets at 95% than at 80%. The higher level requires wider bands and intervals: the mean ROC radius increases from 0.135 to 0.175, and the mean AUC width increases from 0.104 to 0.158. Table 11 reports each dataset separately; Table 12 gives the corresponding mean prediction sizes. Note that conditional on each available dataset, coverage estimates from 300 random splits have Monte Carlo standard errors of approximately 2.3

and 1.3 percentage points near 80% and 95%, respectively. Therefore, small deviations from the nominal coverage levels should be interpreted with caution, especially for individual datasets.

Table 10 Coverage and prediction size, averaged equally over the nine datasets. Minimum–maximum ranges describe variation between datasets, not confidence intervals.

	80% nominal level		95% nominal level	
Measure	Mean	Minimum–maximum	Mean	Minimum–maximum
ROC coverage (%)	79.85	75.00–82.67	95.00	93.33–97.00
AUC coverage (%)	81.37	77.00–91.67	95.11	93.33–99.33
Mean ROC radius	0.135	0.027–0.291	0.175	0.035–0.378
Mean AUC width	0.104	0.005–0.279	0.158	0.009–0.419

Table 11 Class counts and coverage for uniform DiP-ROC with $m = 1$. Counts apply to each test and external sample; coverage headings give nominal levels.

	Each held-out sample		ROC coverage (%)		AUC coverage (%)	
Dataset	n_+^{ts}	n_-^{ts}	80%	95%	80%	95%
Abalone	416	419	80.67	94.00	79.33	94.00
Adult	2241	6802	82.33	96.00	83.67	95.33
Banknote Authentication	122	152	82.67	97.00	91.67	99.33
Diabetes	53	100	81.33	95.67	78.00	96.00
Haberman	16	45	77.33	94.00	80.33	94.67
Mammographic	80	85	78.00	93.33	82.33	94.33
Titanic	142	298	80.33	96.67	78.67	95.67
Transfusion	35	114	75.00	95.00	77.00	93.33
Contraceptive Method Choice	168	125	81.00	93.33	81.33	93.33

As in the synthetic experiment, average coverage can conceal shortfalls. For Transfusion, ROC and AUC coverage at the 80% level are 75.00% and 77.00%, respectively. For Banknote Authentication, AUC coverage is above the nominal 95% level, at 99.33%, even though the mean interval width is only 0.00945. The differences between datasets do not establish a single explanation for these results.

Read alongside Table 11, Table 12 also provides examples of precise predictions with coverage close to nominal. For Adult, the mean 95% AUC interval width is 0.021, with coverage of 95.33%. The mean ROC radius is 0.035, with simultaneous coverage of 96.00%. For Abalone, the mean 95% AUC interval width is 0.066 and the mean ROC radius is 0.090. Both predictions attain 94.00% coverage. These examples show that DiP-ROC can provide narrow predictions while retaining coverage close to the requested level.

Table 12 Mean prediction sizes for the predictions evaluated in Table 11. Smaller values indicate greater precision only when coverage is adequate.

	Mean ROC radius		Mean AUC width	
Dataset	80%	95%	80%	95%
Abalone	0.072	0.090	0.043	0.066
Adult	0.027	0.035	0.014	0.021
Banknote Authentication	0.066	0.093	0.005	0.009
Diabetes	0.178	0.229	0.124	0.189
Haberman	0.291	0.378	0.279	0.419
Mammographic	0.156	0.197	0.103	0.157
Titanic	0.105	0.138	0.092	0.141
Transfusion	0.195	0.257	0.165	0.252
Contraceptive Method Choice	0.123	0.156	0.112	0.171

We evaluate $\mathcal{K} \in \{5, 10, 20\}$ with $\gamma_k = k/\mathcal{K}$, reusing each replication's split, fitted preprocessing and classifier, and held-out scores across grids. Each grid uses the uniform prior $\alpha_k^\ell = 1/\mathcal{K}$ and its own 4000 predictive draws for both nominal levels, with coverage assessed against the external ROC and AUC on that grid. Table 13 summarizes coverage and prediction size across the nine datasets for each grid. Changing the grid changes the predicted ROC points and trapezoidal AUC; these results do not identify a grid resolution preferable for every dataset.

Table 13 also shows that average AUC coverage remains close to nominal across all three grids. At the 80% level, it ranges from 80.93% to 81.44%. At the 95% level, it ranges from 95.11% to 95.48%. Increasing $\mathcal{K}$ from 5 to 20 raises the mean 95% AUC interval width only from 0.157 to 0.160, an increase of about 1.4%. This is useful when an application requires more score thresholds. In these experiments, the finer grid retains near-nominal average coverage for its AUC target with little increase in interval width. Average 95% ROC coverage also remains within 0.9 percentage points of nominal across the grids.

Table 13 Real-data summary by grid resolution. Entries average the nine dataset-specific replication means with equal weight. These descriptive averages do not guarantee coverage on individual datasets.

		Coverage (%)		Mean uncertainty size	
$\mathcal{K}$	Nominal level (%)	ROC	AUC	ROC radius	AUC width
5	80	80.04	80.93	0.121	0.103
5	95	95.81	95.37	0.161	0.157
10	80	79.85	81.37	0.135	0.104
10	95	95.00	95.11	0.175	0.158
20	80	78.85	81.44	0.144	0.105
20	95	94.78	95.48	0.185	0.160

12 Conclusions

DiP-ROC uses a Bayesian framework to anticipate how the empirical ROC curve and its grid-based AUC may vary in a new evaluation of a fixed classifier. It provides one coherent predictive distribution for both summaries on a predefined threshold grid, assuming new observations from the same populations with matching class counts. It combines uncertainty about the score distributions with variation from a new sample to produce simultaneous bands for the ROC points and predictive intervals for the AUC.

The synthetic experiments show that a weak uniform prior gives average coverage close to the requested levels, while historical score histograms from the same populations can improve precision. Across nine real-world datasets, the weak uniform prior also achieved average coverage close to the requested levels, although coverage varied between datasets. A small minority class can leave predictions wide even when the total sample is large. Stronger priors and narrower bands or intervals do not necessarily make predictions more reliable.

Future work could explore more flexible ways to incorporate prior information, including priors calibrated solely from an AUC summary over a wider range of target values and approaches that account for uncertainty in historical summaries. Extending this predictive framework to other performance curves, particularly precision–recall curves, would also broaden its use, with explicit attention to class balance in the future evaluation.

Statements and Declarations

Corresponding author

Olivier Caelen (olivier.caelen@uclouvain.be).

Funding

No funding received for this work.

Competing interests

No financial or non-financial interests to disclose.

Data availability

The real-world datasets are available from OpenML (https://www.openml.org/); Table 9 gives their identifiers, versions, and binary encodings. The synthetic-data generation procedure is described in Section 11.2.

Code availability

The code repository is https://github.com/ocaelen/bayesian_roc.

Appendix A Proof of Lemma 1: Prior Expected Discretized AUC

This is the proof of Lemma 1 on page 13, which gives the prior expected discretized AUC in terms of the Dirichlet parameters.

Proof Let B^+ denote the bin containing the score of a randomly selected positive observation, and let B^- denote the bin containing the score of a randomly selected negative observation. Thus, B^+ and B^- take values in $\{1, \ldots, \mathcal{K}\}$. Conditionally on $(\boldsymbol{\Lambda}^+, \boldsymbol{\Lambda}^-)$, we assume that these two bin indices are independent and satisfy

$$B^+ \perp\!\!\!\perp B^- \mid (\boldsymbol{\Lambda}^+, \boldsymbol{\Lambda}^-),$$

with

$$\mathbb{P}\Big(B^+ = j \mid \boldsymbol{\Lambda}^+, \boldsymbol{\Lambda}^-\Big) = \Lambda_j^+, \qquad \mathbb{P}\Big(B^- = j \mid \boldsymbol{\Lambda}^+, \boldsymbol{\Lambda}^-\Big) = \Lambda_j^-, \qquad j = 1, \ldots, \mathcal{K}.$$

Following the standard probabilistic interpretation of the AUC [21], we define the discretized population AUC by

$$A_{\mathcal{K}} := \mathbb{P}\left(B^+ > B^- \mid \boldsymbol{\Lambda}^+, \boldsymbol{\Lambda}^-\right) + \frac{1}{2}\mathbb{P}\left(B^+ = B^- \mid \boldsymbol{\Lambda}^+, \boldsymbol{\Lambda}^-\right). \tag{A1}$$

The quantity $A_{\mathcal{K}}$ is random because it is a function of the random vectors $\boldsymbol{\Lambda}^+$ and $\boldsymbol{\Lambda}^-$.

The event $B^+ > B^-$ occurs when the negative observation belongs to bin k and the positive observation belongs to any bin $j \in \{k+1, \ldots, \mathcal{K}\}$. For a fixed k,

$$\mathbb{P}(B^- = k \mid \boldsymbol{\Lambda}^-) = \Lambda_k^-, \qquad \text{and} \qquad \mathbb{P}(B^+ > k \mid \boldsymbol{\Lambda}^+) = \sum_{j=k+1}^{\mathcal{K}} \Lambda_j^+.$$

By the conditional independence of B^+ and B^- given $(\boldsymbol{\Lambda}^+, \boldsymbol{\Lambda}^-)$, we have

$$\mathbb{P}(B^- = k,\ B^+ > k \mid \boldsymbol{\Lambda}^+, \boldsymbol{\Lambda}^-) = \Lambda_k^- \sum_{j=k+1}^{\mathcal{K}} \Lambda_j^+.$$

Summing over all possible k gives

$$\mathbb{P}(B^+ > B^- \mid \boldsymbol{\Lambda}^+, \boldsymbol{\Lambda}^-) = \sum_{k=1}^{\mathcal{K}} \Lambda_k^- \sum_{j=k+1}^{\mathcal{K}} \Lambda_j^+, \tag{A2}$$

where, for $k = \mathcal{K}$, the inner sum is empty and equals zero.

The event $B^+ = B^-$ occurs when both observations belong to the same bin k:

$$\mathbb{P}(B^+ = k,\ B^- = k \mid \boldsymbol{\Lambda}^+, \boldsymbol{\Lambda}^-) = \Lambda_k^+ \Lambda_k^-.$$

Summing over k, we have

$$\mathbb{P}(B^+ = B^- \mid \boldsymbol{\Lambda}^+, \boldsymbol{\Lambda}^-) = \sum_{k=1}^{\mathcal{K}} \Lambda_k^+ \Lambda_k^-. \tag{A3}$$

Therefore, by substituting (A2) and (A3) into (A1), we obtain

$$A_{\mathcal{K}} = \sum_{k=1}^{\mathcal{K}} \Lambda_k^- \sum_{j=k+1}^{\mathcal{K}} \Lambda_j^+ + \frac{1}{2} \sum_{k=1}^{\mathcal{K}} \Lambda_k^+ \Lambda_k^-$$

$$= \sum_{k=1}^{\mathcal{K}} \left(\sum_{j=k+1}^{\mathcal{K}} \Lambda_k^- \Lambda_j^+ + \frac{1}{2} \Lambda_k^- \Lambda_k^+ \right).$$

Taking expectations with respect to the joint prior distribution of $(\boldsymbol{\Lambda}^+, \boldsymbol{\Lambda}^-)$ gives the prior expected discretized AUC

$$a_{\mathcal{K}} = \mathbb{E}[A_{\mathcal{K}}] \tag{A4}$$

$$= \sum_{k=1}^{\mathcal{K}} \left(\sum_{j=k+1}^{\mathcal{K}} \mathbb{E}[\Lambda_k^- \Lambda_j^+] + \frac{1}{2} \mathbb{E}[\Lambda_k^- \Lambda_k^+] \right)$$

$$= \sum_{k=1}^{\mathcal{K}} \left[\sum_{j=k+1}^{\mathcal{K}} \mathbb{E}[\Lambda_k^-] \mathbb{E}[\Lambda_j^+] + \frac{1}{2} \mathbb{E}[\Lambda_k^-] \mathbb{E}[\Lambda_k^+] \right] \tag{A5}$$

$$= \sum_{k=1}^{\mathcal{K}} \mathbb{E}[\Lambda_k^-] \left(\sum_{j=k+1}^{\mathcal{K}} \mathbb{E}[\Lambda_j^+] + \frac{1}{2} \mathbb{E}[\Lambda_k^+] \right), \tag{A6}$$

where the third equality (A5) follows from the independence assumption $\boldsymbol{\Lambda}^+ \perp\!\!\!\perp \boldsymbol{\Lambda}^-$. Substituting the definitions of q_k^+ and q_k^- from (18) into (A6), we obtain

$$a_{\mathcal{K}} = \sum_{k=1}^{\mathcal{K}} q_k^- \left(\sum_{j=k+1}^{\mathcal{K}} q_j^+ + \frac{1}{2} q_k^+ \right).$$

□

Appendix B An Auxiliary Identity for the Expected Discretized AUC

This appendix establishes an identity for the deviation of the expected discretized AUC from 1/2. This identity is used in the proof of Corollary 2.

From (19), we have

$$a_{\mathcal{K}} = \sum_{k=1}^{\mathcal{K}} \left[\sum_{j=k+1}^{\mathcal{K}} q_k^- q_j^+ + \frac{1}{2} q_k^- q_k^+ \right]$$

$$= \sum_{k=1}^{\mathcal{K}} \sum_{j=k+1}^{\mathcal{K}} q_j^+ q_k^- + \frac{1}{2} \sum_{k=1}^{\mathcal{K}} q_k^+ q_k^-.$$

Lemma 4 *Let* $\boldsymbol{q}^+ = (q_1^+, \ldots, q_{\mathcal{K}}^+)^\top \in \Delta_{\mathcal{K}-1}$ *and* $\boldsymbol{q}^- = (q_1^-, \ldots, q_{\mathcal{K}}^-)^\top \in \Delta_{\mathcal{K}-1}$ *be two probability vectors. For the expected discretized AUC*

$$a_{\mathcal{K}} = \sum_{k=1}^{\mathcal{K}} \sum_{j=k+1}^{\mathcal{K}} q_j^+ q_k^- + \frac{1}{2} \sum_{k=1}^{\mathcal{K}} q_k^+ q_k^-,$$

we have

$$a_{\mathcal{K}} - \frac{1}{2} = \frac{1}{2} \sum_{j>k} \left(q_j^+ q_k^- - q_k^+ q_j^- \right) \tag{B7}$$

where $\sum_{j>k} := \sum_{k=1}^{\mathcal{K}} \sum_{j=k+1}^{\mathcal{K}}$.

Proof Since $\boldsymbol{q}^+$ and $\boldsymbol{q}^-$ are probability vectors, we have $\sum_{j=1}^{\mathcal{K}} q_j^\ell = 1$ and therefore

$$\begin{aligned}
1 &= \left(\sum_{j=1}^{\mathcal{K}} q_j^+\right)\left(\sum_{k=1}^{\mathcal{K}} q_k^-\right) \\
&= \sum_{k=1}^{\mathcal{K}}\sum_{j=1}^{\mathcal{K}} q_j^+ q_k^- \\
&= \sum_{k=1}^{\mathcal{K}}\left[\sum_{j=1}^{k-1} q_j^+ q_k^- + q_k^+ q_k^- + \sum_{j=k+1}^{\mathcal{K}} q_j^+ q_k^-\right] \\
&= \sum_{k=1}^{\mathcal{K}}\sum_{j=1}^{k-1} q_j^+ q_k^- + \sum_{k=1}^{\mathcal{K}} q_k^+ q_k^- + \sum_{k=1}^{\mathcal{K}}\sum_{j=k+1}^{\mathcal{K}} q_j^+ q_k^- \\
&= \sum_{j>k} q_j^+ q_k^- + \sum_{j<k} q_j^+ q_k^- + \sum_{k=1}^{\mathcal{K}} q_k^+ q_k^- \\
&= \sum_{j>k} q_j^+ q_k^- + \sum_{j>k} q_k^+ q_j^- + \sum_{k=1}^{\mathcal{K}} q_k^+ q_k^- \qquad \text{(B8)} \\
&= \sum_{j>k}\left(q_j^+ q_k^- + q_k^+ q_j^-\right) + \sum_{k=1}^{\mathcal{K}} q_k^+ q_k^-
\end{aligned}$$

where, for compactness, we use the notation

$$\sum_{j>k} := \sum_{k=1}^{\mathcal{K}}\sum_{j=k+1}^{\mathcal{K}} \qquad \text{and} \qquad \sum_{j<k} := \sum_{k=1}^{\mathcal{K}}\sum_{j=1}^{k-1}$$

with empty sums equal to zero. Dividing by two gives

$$\frac{1}{2} = \frac{1}{2}\sum_{j>k}\left(q_j^+ q_k^- + q_k^+ q_j^-\right) + \frac{1}{2}\sum_{k=1}^{\mathcal{K}} q_k^+ q_k^- . \qquad \text{(B9)}$$

We also have from (19) that

$$a_{\mathcal{K}} = \sum_{j>k} q_j^+ q_k^- + \frac{1}{2}\sum_{k=1}^{\mathcal{K}} q_k^+ q_k^- . \qquad \text{(B10)}$$

Subtracting (B9) from (B10) gives

$$\begin{aligned}
a_{\mathcal{K}} - \frac{1}{2} &= \sum_{j>k} q_j^+ q_k^- + \frac{1}{2}\sum_{k=1}^{\mathcal{K}} q_k^+ q_k^- - \frac{1}{2}\sum_{j>k}\left(q_j^+ q_k^- + q_k^+ q_j^-\right) - \frac{1}{2}\sum_{k=1}^{\mathcal{K}} q_k^+ q_k^- \\
&= \sum_{j>k} q_j^+ q_k^- - \frac{1}{2}\sum_{j>k}\left(q_j^+ q_k^- + q_k^+ q_j^-\right) \\
&= \sum_{j>k} q_j^+ q_k^- - \frac{1}{2}\sum_{j>k} q_j^+ q_k^- - \frac{1}{2}\sum_{j>k} q_k^+ q_j^-
\end{aligned}$$

$$= \frac{1}{2} \sum_{j>k} \left(q_j^+ q_k^- - q_k^+ q_j^- \right).$$

Assuming $j > k$, $q_j^+ q_k^-$ can be interpreted as the probability that a randomly selected positive observation falls in the higher-score bin j while a randomly selected negative observation falls in the lower-score bin k. Similarly, $q_k^+ q_j^-$ is the probability of the reverse configuration. Hence, $q_j^+ q_k^- - q_k^+ q_j^-$ measures the excess probability of correct over incorrect ordering for this pair of bins. Summing these excess probabilities over all pairs of distinct bins quantifies the overall tendency of the positive class to be ranked above the negative class. Since $a_{\mathcal{K}} - \frac{1}{2}$ is the signed area between the prior mean discretized population ROC curve and the diagonal, the factor $1/2$ converts this cumulative ordering advantage into area: each pairwise excess contributes twice the corresponding elementary triangular area, so dividing by two yields exactly the total signed area between this prior mean curve and the diagonal.

□

Appendix C Proof of Corollary 2: Expected Discretized AUC under Linear Prior Profiles

This is the proof of Corollary 2 on page 17, which expresses the prior expected discretized AUC under the linear prior mean profiles.

Proof By (B7), we have

$$a_{\mathcal{K}} - \frac{1}{2} = \frac{1}{2} \sum_{j>k} \left(q_j^+ q_k^- - q_k^+ q_j^- \right).$$

For $j > k$, using the definitions of q_j^+ and q_k^- in (24),

$$q_j^+ q_k^- = \left(\frac{1}{\mathcal{K}} + s x_j \right) \left(\frac{1}{\mathcal{K}} - s x_k \right),$$

and similarly,

$$q_k^+ q_j^- = \left(\frac{1}{\mathcal{K}} + s x_k \right) \left(\frac{1}{\mathcal{K}} - s x_j \right).$$

Subtracting the two expressions gives

$$q_j^+ q_k^- - q_k^+ q_j^- = \frac{2s}{\mathcal{K}} (x_j - x_k) = \frac{2s}{\mathcal{K}} (j - k).$$

Therefore,

$$a_{\mathcal{K}} - \frac{1}{2} = \frac{s}{\mathcal{K}} \sum_{k=1}^{\mathcal{K}} \sum_{j=k+1}^{\mathcal{K}} (j - k).$$

For a fixed k, let $r = j - k$. Then

$$\sum_{j=k+1}^{\mathcal{K}} (j - k) = \sum_{r=1}^{\mathcal{K}-k} r = \frac{(\mathcal{K} - k)(\mathcal{K} - k + 1)}{2}.$$

Hence,

$$a_{\mathcal{K}} - \frac{1}{2} = \frac{s}{2\mathcal{K}} \sum_{k=1}^{\mathcal{K}} (\mathcal{K} - k)(\mathcal{K} - k + 1).$$

Changing variables to $r = \mathcal{K} - k + 1$ and reversing the order of summation, we have

$$\sum_{k=1}^{\mathcal{K}} (\mathcal{K} - k)(\mathcal{K} - k + 1) = \sum_{r=1}^{\mathcal{K}} (r-1)r.$$

Consequently,

$$\begin{aligned} a_{\mathcal{K}} - \frac{1}{2} &= \frac{s}{2\mathcal{K}} \sum_{r=1}^{\mathcal{K}} (r^2 - r) \\ &= \frac{s}{2\mathcal{K}} \left(\frac{\mathcal{K}(\mathcal{K}+1)(2\mathcal{K}+1)}{6} - \frac{\mathcal{K}(\mathcal{K}+1)}{2} \right) \\ &= \frac{s(\mathcal{K}^2 - 1)}{6}. \end{aligned}$$

Therefore,

$$a_{\mathcal{K}} = \frac{1}{2} + \frac{s(\mathcal{K}^2 - 1)}{6}.$$

□

Appendix D Proof of Corollary 3: Prior Mean Discretized Population ROC Curve under Linear Prior Profiles

This is the proof of Corollary 3 on page 19. It gives an explicit expression for the prior mean discretized population ROC points under the linear prior profiles calibrated to a target expected AUC a_0.

Proof Fix $k \in \{0, \ldots, \mathcal{K}\}$, with empty sums interpreted as zero. By (13), the true positive rate is the sum of the positive-class probabilities over bins $k+1, \ldots, \mathcal{K}$. Applying linearity of expectation and substituting the prior means gives

$$\begin{aligned} \mathbb{E}[\mathrm{TPR}_k] &= \mathbb{E}\left[\sum_{j=k+1}^{\mathcal{K}} \Lambda_j^+ \right] \\ &= \sum_{j=k+1}^{\mathcal{K}} \mathbb{E}[\Lambda_j^+] \\ &= \sum_{j=k+1}^{\mathcal{K}} \left(\frac{1}{\mathcal{K}} + s_0 x_j \right) \\ &= \frac{\mathcal{K} - k}{\mathcal{K}} + s_0 \sum_{j=k+1}^{\mathcal{K}} x_j. \end{aligned}$$

The last equality follows because the sum contains $\mathcal{K} - k$ terms equal to $1/\mathcal{K}$, and s_0 is deterministic. Linearity of expectation applies regardless of the dependence between the Dirichlet components.

To evaluate the remaining sum, substitute the definition of x_j:

$$\sum_{j=k+1}^{\mathcal{K}} x_j = \sum_{j=k+1}^{\mathcal{K}} \left(j - \frac{\mathcal{K}+1}{2} \right)$$

$$
\begin{aligned}
&= \sum_{j=k+1}^{\mathcal{K}} j - (\mathcal{K}-k)\frac{\mathcal{K}+1}{2} \\
&= \left(\sum_{j=1}^{\mathcal{K}} j - \sum_{j=1}^{k} j\right) - (\mathcal{K}-k)\frac{\mathcal{K}+1}{2} \\
&= \frac{\mathcal{K}(\mathcal{K}+1) - k(k+1) - (\mathcal{K}-k)(\mathcal{K}+1)}{2} \\
&= \frac{k(\mathcal{K}+1) - k(k+1)}{2} \\
&= \frac{k(\mathcal{K}-k)}{2}.
\end{aligned}
$$

Here, the second equality counts the $\mathcal{K}-k$ occurrences of the centering constant $(\mathcal{K}+1)/2$. The third equality expresses the sum from $k+1$ to $\mathcal{K}$ as the difference of two sums starting at 1. The fourth equality uses $\sum_{j=1}^{n} j = n(n+1)/2$. Expanding $(\mathcal{K}-k)(\mathcal{K}+1)$ cancels the term $\mathcal{K}(\mathcal{K}+1)$, and factoring out k gives the final expression.

Consequently,

$$
\mathbb{E}[\mathrm{TPR}_k] = \frac{\mathcal{K}-k}{\mathcal{K}} + \frac{s_0 k(\mathcal{K}-k)}{2}.
$$

For the false positive rate, the same argument applies to the negative-class probabilities. Their prior means have the opposite sign in front of $s_0 x_j$, so

$$
\begin{aligned}
\mathbb{E}[\mathrm{FPR}_k] &= \mathbb{E}\left[\sum_{j=k+1}^{\mathcal{K}} \Lambda_j^-\right] \\
&= \sum_{j=k+1}^{\mathcal{K}} \mathbb{E}[\Lambda_j^-] \\
&= \sum_{j=k+1}^{\mathcal{K}} \left(\frac{1}{\mathcal{K}} - s_0 x_j\right) \\
&= \frac{\mathcal{K}-k}{\mathcal{K}} - s_0 \sum_{j=k+1}^{\mathcal{K}} x_j \\
&= \frac{\mathcal{K}-k}{\mathcal{K}} - \frac{s_0 k(\mathcal{K}-k)}{2}.
\end{aligned}
$$

Since $\boldsymbol{R}_k = (\mathrm{FPR}_k, \mathrm{TPR}_k)^\top$, combining the two expected coordinates yields

$$
\mathbb{E}[\boldsymbol{R}_k] = \frac{\mathcal{K}-k}{\mathcal{K}}\begin{pmatrix}1\\1\end{pmatrix} + \frac{s_0 k(\mathcal{K}-k)}{2}\begin{pmatrix}-1\\1\end{pmatrix}.
$$

Finally, substituting the calibrated slope from (26),

$$
s_0 = \frac{6(a_0 - 1/2)}{\mathcal{K}^2 - 1} = \frac{3(2a_0 - 1)}{\mathcal{K}^2 - 1}
$$

gives (30). □

Appendix E Derivation of the Posterior Predictive Mean ROC Points

This appendix derives equation (33). By the Dirichlet-multinomial distribution in (32) and its mean formula, $\mathbb{E}\left[\widetilde{D}_j^{\ell} \mid \boldsymbol{D}^{\ell} = \boldsymbol{d}^{\ell}\right] = n_{\ell}^{\mathrm{ts}} \frac{\omega_j^{\ell}}{\omega_{\bullet}^{\ell}}$, where $\omega_{\bullet}^{\ell} := \sum_{j=1}^{\mathcal{K}} \omega_j^{\ell}$. Since n_{+}^{ts} is fixed after conditioning on $\boldsymbol{D}^{+} = \boldsymbol{d}^{+}$, we have

$$\begin{aligned}
\mathbb{E}\left[\widetilde{\mathrm{TPR}}_k \mid \boldsymbol{D}^{+} = \boldsymbol{d}^{+}\right] &= \mathbb{E}\left[\frac{\widetilde{T}_k^{+}}{n_{+}^{\mathrm{ts}}} \mid \boldsymbol{D}^{+} = \boldsymbol{d}^{+}\right] \\
&= \frac{1}{n_{+}^{\mathrm{ts}}} \sum_{j=k+1}^{\mathcal{K}} \mathbb{E}\left[\widetilde{D}_j^{+} \mid \boldsymbol{D}^{+} = \boldsymbol{d}^{+}\right] \\
&= \frac{1}{n_{+}^{\mathrm{ts}}} \sum_{j=k+1}^{\mathcal{K}} n_{+}^{\mathrm{ts}} \frac{\omega_j^{+}}{\omega_{\bullet}^{+}} \\
&= \sum_{j=k+1}^{\mathcal{K}} \frac{\omega_j^{+}}{\omega_{\bullet}^{+}}.
\end{aligned}$$

Similarly,

$$\mathbb{E}\left[\widetilde{\mathrm{FPR}}_k \mid \boldsymbol{D}^{-} = \boldsymbol{d}^{-}\right] = \sum_{j=k+1}^{\mathcal{K}} \frac{\omega_j^{-}}{\omega_{\bullet}^{-}}.$$

Therefore, the exact posterior predictive mean of the future empirical ROC point is

$$\boldsymbol{\mu}_k = \begin{pmatrix} \sum_{j=k+1}^{\mathcal{K}} \frac{\omega_j^{-}}{\omega_{\bullet}^{-}} \\ \sum_{j=k+1}^{\mathcal{K}} \frac{\omega_j^{+}}{\omega_{\bullet}^{+}} \end{pmatrix}.$$

References

[1] DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics **44**(3), 837–845 (1988) https://doi.org/10.2307/2531595

[2] Bertail, P., Clémençon, S.J., Vayatis, N.: On bootstrapping the ROC curve. In: Advances in Neural Information Processing Systems 21, pp. 137–144. Curran Associates, Inc. (2008)

[3] Martínez-Camblor, P., Pérez-Fernández, S., Corral, N.: Efficient nonparametric confidence bands for receiver operating-characteristic curves. Statistical Methods in Medical Research **27**(6), 1892–1908 (2018) https://doi.org/10.1177/0962280216672490

[4] Macskassy, S.A., Provost, F., Rosset, S.: ROC confidence bands: An empirical evaluation. In: Proceedings of the 22nd International Conference on Machine Learning, pp. 537–544. Association for Computing Machinery (2005). https://doi.org/10.1145/1102351.1102419

[5] Urlus, R.E.Q., Baak, M., Collot, S., Fridman Rojas, I.: Pointwise sampling uncertainties on the Precision–Recall curve. In: Ruiz, F., Dy, J., van de Meent, J.-W. (eds.) Proceedings of the 26th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 206, pp. 8211–8232. PMLR (2023). https://proceedings.mlr.press/v206/urlus23a.html

[6] Zheng, Z., Yang, B., Song, P.X.-K.: Classification Uncertainty Quantification: A Comparison Between Bootstrap and Conformal ROC Confidence Bands. Statistica Sinica, Volume 37, Issue 4. forthcoming. https://doi.org/10.5705/ss.202025.0127 . https://www3.stat.sinica.edu.tw/statistica/fp/SS-2025-0127.html

[7] Gu, J., Ghosal, S., Roy, A.: Bayesian bootstrap estimation of ROC curve. Statistics in Medicine **27**(26), 5407–5420 (2008) https://doi.org/10.1002/sim.3366

[8] Erkanli, A., Sung, M., Costello, E.J., Angold, A.: Bayesian semi-parametric ROC analysis. Statistics in Medicine **25**(22), 3905–3928 (2006) https://doi.org/10.1002/sim.2496

[9] Al-Labadi, L., Evans, M., Liang, Q.: ROC analyses based on measuring evidence using the relative belief ratio. Entropy **24**(12), 1710 (2022) https://doi.org/10.3390/e24121710

[10] Caelen, O.: A Bayesian interpretation of the confusion matrix. Annals of Mathematics and Artificial Intelligence **81**(3–4), 429–450 (2017) https://doi.org/10.1007/s10472-017-9564-8

[11] Tötsch, N., Hoffmann, D.: Classifier uncertainty: evidence, potential impact, and probabilistic treatment. PeerJ Computer Science **7**, e398 (2021) https://doi.org/10.7717/peerj-cs.398

[12] Shang, H., Langlois, J.-M., Tsioutsiouliklis, K., Kang, C.: Precision/Recall on imbalanced test data. In: Ruiz, F., Dy, J., van de Meent, J.-W. (eds.) Proceedings of the 26th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 206, pp. 9879–9891. PMLR (2023). https://proceedings.mlr.press/v206/shang23a.html

[13] Lovell, D., Miller, D., Capra, J., Bradley, A.P.: Never mind the metrics—what about the uncertainty? visualising binary confusion matrix metric distributions to put performance in perspective. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 22702–22757. PMLR (2023). https://proceedings.mlr.press/v202/lovell23a.html

[14] Fawcett, T.: An introduction to ROC analysis. Pattern recognition letters **27**(8), 861–874 (2006) https://doi.org/10.1016/j.patrec.2005.10.010

[15] Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240 (2006). https://doi.org/10.1145/1143844.1143874

[16] Nahm, F.S.: Receiver operating characteristic curve: overview and practical use for clinicians. Korean journal of anesthesiology **75**(1), 25–36 (2022) https://doi.org/10.4097/kja.21209

[17] Murphy, K.P.: Probabilistic Machine Learning: an Introduction. MIT press (2022)

[18] Mosimann, J.E.: On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. Biometrika **49**(1/2), 65–82 (1962) https://doi.org/10.1093/biomet/49.1-2.65

[19] Budig, S., Schaarschmidt, F., Menssen, M.: Prediction intervals for overdispersed multinomial data with application to historical controls. arXiv preprint arXiv:2603.20349 (2026) https://doi.org/10.48550/arXiv.2603.20349

[20] Gneiting, T., Balabdaoui, F., Raftery, A.E.: Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society Series B: Statistical Methodology **69**(2), 243–268 (2007) https://doi.org/10.1111/j.1467-9868.2007.00587.x

[21] Muschelli III, J.: ROC and AUC with a binary predictor: a potentially misleading metric. Journal of classification **37**(3), 696–708 (2020) https://doi.org/10.1007/s00357-019-09345-1