Pointwise sampling uncertainties on the Precision-Recall curve

R.E.Q. Urlus **M.A. Baak** **S. Collot** **I. Fridman Rojas**
ING Bank, Bijlmerdreef 106, 1102 CT Amsterdam, The Netherlands

Abstract

Quoting robust uncertainties on machine learning (ML) model metrics, such as f1-score, precision, recall, etc., from sources of uncertainty such as data sampling, parameter initialization, and target labelling, is typically not done in the field of data science, even though these are essential for the proper interpretation and comparison of ML models. This text shows how to calculate and visualize the impact of one dominant source of uncertainty – the sampling uncertainty of the test dataset – on each point of the Precision-Recall (PR) and Receiver Operating Characteristic (ROC) curves. This is particularly relevant for PR curves, where the joint uncertainty on recall and precision can be large and non-linear, especially at low recall. Four statistical methods to evaluate this uncertainty, both frequentist and Bayesian in origin, are compared in terms of coverage and speed. Of these, Wilks' method is the winner: it provides (near) correct coverage for samples as small as 10 records, works fine when the precision or recall are close to the edges of zero or one, and can be evaluated quickly for practical use. The presented algorithms are available through a public Python library. We recommend that showing uncertainty bands of PR or ROC curves becomes the norm, and believe our methodology forms a useful and necessary addition to any data scientist's toolbox.

1 INTRODUCTION

A meaningful comparison between any two ML methods is difficult without knowing the correct uncertainties on their corresponding model metrics. Knowledge of the uncertainties allows one to use best judgement in selecting and deploying a model (e.g. using heuristics such as those described in Cumming (2009)). Surprisingly perhaps, these uncertainties are typically not evaluated. Two reasons come to mind. First, their evaluation can be complex and time-consuming, e.g. most existing methods involve bootstrapping or cross-validation, which require retraining multiple times. Second, there are no out-of-the-box methods to calculate these, as far as we are aware.

The sources of uncertainty on model metrics are many, such as data sampling, model initialization, and hyper-parameter optimization (Bouthillier et al., 2021). The sampling uncertainty is often the dominant source (Bouthillier et al., 2021, Fig. 1). Priority is given here to the sampling uncertainty of the test set, which we refer to as the classifier uncertainty. Since the test set is usually smaller than the training set, its sampling uncertainty is generally the largest. This work calculates the confidence intervals from the classifier uncertainty on each point of a PR (or ROC) curve.

The classifier uncertainty is particularly impactful on PR curves in the low-recall, high-precision region, where the number of false and/or true positives can be small, and the dependency between recall and precision is strongly non-linear with a large joint uncertainty.

The following scenario is assumed, and often encountered in business settings. A trained binary classification model has been built, including a discrimination threshold at which to operate. A test set is available of limited size, several thousand data points or less.

This work compares four different statistical methods, based on the fact that any confusion matrix in the PR (or ROC) curve can be modelled with a multinomial distribution.

1. A frequentist approach using the profile likelihood ratio as a test statistic, which forms our baseline method.
2. Wilks' theorem (Wilks, 1938) states that – under certain conditions – the distribution of test statistic values follows the known χ^2 distribution.
3. A Bayesian approach, where for reasons of speed the Dirichlet conjugate prior is used, resulting in a closed-form posterior distribution.
4. The approximation of the precision-recall probability distribution as a bivariate normal distribution.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

The comparison focuses on: statistical coverage in case of small datasets, edge cases where the recall or precision are close to zero or one, evaluation speed, and extendability with other sources of uncertainty.

One could argue that running cross-validation and quoting the standard deviation over multiple folds is a form of reporting sampling uncertainties. Compared with an independent test set, however, this is not an unbiased uncertainty: a bias is introduced by overlapping training folds, introducing a correlation in the trained models (Bengio and Grandvalet, 2003). In addition, this uncertainty depends on the size of a fold, and is likely larger than on the test set.

Another standard method for computing uncertainties on precision and recall is the bootstrap, in this case applied to the test set exclusively. Three drawbacks of this approach are: it is (relatively) computationally expensive; when positives are scarce in the test set, the joint PR distribution will display artifacts in the form of banded, discrete structure; and that retrieving valid 2D confidence contours from the observed PR values is non-trivial (Kernel Density Estimates can be used, but these are CPU-intensive and not really fit for purpose). See Appendix A for further discussion on this.

Our contributions are as follows. The joint classifier uncertainty on recall and precision is derived (and on the true versus false positive rate), and a comparison of statistical methods to evaluate the classifier uncertainty in a PR (and ROC) curve is provided. To the best of our knowledge, Wilks' method – while used in other research fields, in particular in high energy physics (Cowan et al., 2011) – has never been used for model evaluation and comparison in the machine learning literature. In addition, we are not aware of functionality in the major scientific Python libraries that allows one to easily compute and visualise the uncertainty on the performance metrics of a classifier. (Though there is demand for this within the community, see Lemaitre (2021).)

Our recommendation is that, by default, every PR (or ROC) curve should show the classifier uncertainty, in order to provide a more complete view of the performance. Sampling uncertainties also have implications on the optimization of classifiers based on the area under the PR curve, which should itself be used with caution (Flach and Kull, 2015).

Relevant details of each statistical method are provided below, followed by their visualization (Sec. 8) and comparison (Sec. 9). The focus lies on PR curves; the procedure for ROC curves is illustrated in Appendix G.

2 RELATED WORK

The estimation of uncertainty as one of the central purposes of statistical learning – and the role of probabilistic modelling within that task – is perhaps most eloquently argued for by Lindley (2000). A modern review on the topic of uncertainty estimation as related to machine learning can be found in Hüllermeier and Waegeman (2021). A recent and comprehensive study to cover the topic of uncertainty estimation as it relates to model selection and accounting for multiple sources of variation in realistic setups, is Bouthillier et al. (2021).

The present work focuses on the uncertainty due to sampling variability in the test set, in contrast to previous seminal works Nadeau and Bengio (1999); Dietterich (1998) which consider uncertainty due to training set variability.

Focusing on probabilistic modelling of the confusion matrix, and thereby deducing uncertainty intervals on performance metrics derived from it, the most similar results to the present work are Caelen (2017) and Tötsch and Hoffmann (2021), where the former employs a multinomial distribution to model the confusion matrix, and the latter decomposes the two-class confusion matrix into three binomial processes.

For the particular case of the ROC curve, Hall et al. (2004) develops an asymptotic estimator for confidence intervals. Previous work also exists targeting the estimation of the area under the PR curve (AUCpr) and confidence intervals on it (Boyd et al., 2013), as well as functionality to use soft labels and visualize loose bounds on PR curves (Grau et al., 2015).

The present work differs from these existing approaches in that it focuses on the modelling the joint uncertainty on precision and recall (ROC). In addition, this work tackles the uncertainty estimation with a frequentist approach, making use of e.g. profile likelihood methods, next to a Bayesian approach.

Other works which have generated intervals on performance curves (i.e. not pointwise) using the bootstrap have been for example Everingham et al. (2015) on PR curves and Bertail et al. (2008) for ROC curves.

Yet other types of approaches for uncertainty estimation of performance metrics exist, such as tail bounds on the error probabilities (Langford, 2005). An example of the relevance of uncertainty estimation on real-world applications is the study of uncertainty in the context of prediction from electronic health records (Dusenberry et al., 2020).

3 MODELING THE CONFUSION MATRIX

In practice there is one confusion matrix per discrimination threshold along the PR curve. Assume a fixed, binary classification model and pick one discrimination threshold. Denote $x = (x_{TP}, x_{FP}, x_{TN}, x_{FN})$, where x_{TP}, x_{FP}, x_{TN}, and x_{FN} are the number of true positives, false positives, true negatives, and false negatives respectively, as obtained from the confusion matrix of the test set. Recall the usual

definitions for precision and recall:

$$\hat{P} = \frac{x_{TP}}{x_{TP} + x_{FP}} \quad ; \quad \hat{R} = \frac{x_{TP}}{x_{TP} + x_{FN}} \,, \tag{1}$$

which are best estimates of the true recall and precision parameters R and P.

The confusion matrix is described by a multinomial distribution with four (or more in the case of multi-class) mutually exclusive categories:

$$p_{MN}(x) = \frac{n!}{x_{TP}! \dots x_{FN}!} p_{TP}^{x_{TP}} \dots p_{FN}^{x_{FN}} \,, \tag{2}$$

with corresponding probabilities $p = (p_{TP}, p_{FP}, p_{TN}, p_{FN})$ subject to $\sum_i p_i = 1$ and $\sum_i x_i = n$. For our purpose we rewrite:

$$p_{FP} = \left(\frac{1-P}{P}\right) p_{TP} \quad ; \quad p_{FN} = \left(\frac{1-R}{R}\right) p_{TP} \,. \tag{3}$$

Given a confusion matrix x, the parameters of interest are R and P, and the auxiliary parameter is p_{TP}.

The assumption that the confusion matrix of an i.i.d. sample is described by a multinomial distribution can be easily verified using simulation studies, the results of one such simulation study are shown in Appendix B. See Caelen (2017) for an analogous Bayesian interpretation of the confusion matrix.

4 FREQUENTIST APPROACH

The frequentist method is CPU-intensive, using Monte Carlo simulations to determine the correct confidence intervals.

4.1 The Profile Log Likelihood Ratio

The aim is to find the uncertainty on the estimated Precision-Recall point of an observed confusion matrix x. To obtain this the profile likelihood method is used.

First, p is inferred from x using the multinomial distribution as likelihood function:

$$\hat{p} = \text{argmax}_p \, L(p) \,. \tag{4}$$

The maximum value $L(\hat{R}, \hat{P}, \hat{p}_{TP})$ is reached at $\forall i \; \hat{p}_i = x_i / n$, from which $\hat{R}$ and $\hat{P}$ follow as in Eqn. 1.

To set the uncertainty contours around $\hat{R}$, $\hat{P}$, hypothesis tests are performed for a square grid of R, P values in the surrounding region. Each hypothesis test is for exclusion, *i.e.* the aim is to reject the unlikely R, P values in the grid. The ranges of this grid are described further below.

Next, fixing R and P to each grid coordinate, the likelihood is maximized with respect to the auxiliary parameter p_{TP}:

$$\hat{\hat{p}}_{TP} = \text{argmax}_{p_{TP}} \, L(R, P, p_{TP}) \,. \tag{5}$$

resulting in the maximum likelihood value $L(R, P, \hat{\hat{p}}_{TP})$, where $\hat{\hat{p}}_{TP}$ can be derived as:

$$\hat{\hat{p}}_{TP} \quad = \quad \frac{x_{TP} + x_{FN} + x_{FP}}{\left(\frac{1}{P} + \frac{1}{R} - 1\right) n} \,. \tag{6}$$

(See Appendix C for details.) One can interpret $\hat{\hat{p}}_{TP}$ as the constrained fraction of true positives that best matches x.

The test statistic $q_{R,P}$ is a function of the likelihood values:

$$q_{R,P} = -2 \log \left(\frac{L(R, P, \hat{\hat{p}}_{TP})}{L(\hat{R}, \hat{P}, \hat{p}_{TP})} \right) , \tag{7}$$

known as the profile log likelihood ratio. This ratio is used in the frequentist approach as ordering variable.

4.2 Monte Carlo Simulation

In the frequentist approach the distribution of the test statistic $f(q_{R,P} | R, P, p_{TP})$ is determined using multinomial samples that are generated as in Baak et al. (2015), described in detail in Appendix H.

In summary, there is one confusion matrix per discrimination threshold, and one R, P grid per confusion matrix. For each R, P grid point the fraction of generated samples is determined with $q_{R,P}$ values smaller than the corresponding value of the test set's confusion matrix. This fraction signifies the exclusion p-value for that R, P point. Exclusion iso-contours are constructed based on these values in the R, P plane, *e.g.* the contour at $p = 0.9$ forms the 90% confidence interval.

By construction, when using this procedure the p-value obtained for the hypothesis test will not undercover. The procedure guarantees exact statistical coverage in the case where the fitted value of p_{TP} corresponds to its true value. When these are different it will over-cover, for the reason that any other value is less consistent with the test set: $q_{R,P}$ must be higher, resulting in a tighter level of exclusion.

5 WILKS' METHOD

Wilks' method is similar to the frequentist approach, but avoids the costly generation of Monte Carlo samples. Per confusion matrix of the PR curve the test statistic $q_{R,P}$ is evaluated for each R, P grid point. The test statistic $q_{R,P}$ has an important property. Wilks' theorem Wilks (1938) states that, asymptotically, $q_{R,P}$ is described by a χ^2 distribution with two degrees of freedom and is independent of the actual value of p_{TP}.

The χ^2 distribution has a well-known integral that serves as probability function, $\mathcal{P}$, to obtain the p-value of a hypothesis test. For example, with two degrees of freedom:

$$\mathcal{P}(q_{R,P} < 4.605) = 0.9 \,, \tag{8}$$

meaning the contour of R, P values at $q_{R,P} = 4.605$ forms the 90% confidence interval. (The 95% confidence interval corresponds to the value 5.991.) More precisely, 90% of the x vectors corresponding to a multinomial distribution with $p = (R, P, \hat{\hat{p}}_{TP})$ have $q_{R,P}$ values smaller than 4.605.

Wilks' theorem holds asymptotically, meaning for large statistics samples. The approximation of large statistics holds reasonably well in many cases, e.g. from as few as O(10) data points per confusion matrix cell (Baak et al., 2015; Cochran, 1952). Therefore, one often uses the asymptotic approximation to evaluate the p-value of a hypothesis test, avoiding the need for compute-intensive multinomial sample generation.

Wilks' theorem can break down for low-statistics samples, and when R or P reaches 0 or 1. These scenarios are tested explicitly in Sec. 9.2.1.

6 BAYESIAN METHOD

A known Bayesian approach Caelen (2017) for computing credible intervals on any performance metric, is to assume the multinomial likelihood model for the confusion matrix, as defined in Eqn 2, and choosing its conjugate prior – the Dirichlet distribution – for computational convenience.

For a given choice of prior, *i.e.* the values of the concentration parameters of the Dirichlet prior, the posterior distribution over the entries of the confusion matrix is then also a Dirichlet distribution, giving a nice closed-form posterior from which to sample or compute credible intervals for the probability parameters p.

Other forms of Bayesian models for the confusion matrix are possible, such as the Beta-Binomial model for a binary classification confusion matrix shown in Tötsch and Hoffmann (2021). More general Bayesian models with yet other likelihood and prior combinations could be devised, however these would likely not possess a closed-form posterior and would therefore require Markov Chain Monte Carlo (MCMC) sampling.

For simplicity we solely make use of the Dirichlet model:

$$\begin{aligned} \mathrm{Dir}(p|\alpha) &= \frac{\Gamma(\alpha_{TP}+\alpha_{FP}+\alpha_{TN}+\alpha_{FN})}{\Gamma(\alpha_{TP})\Gamma(\alpha_{FP})\Gamma(\alpha_{TN})\Gamma(\alpha_{FN})} \times \\ & p_{TP}^{\alpha_{TP}-1} p_{FP}^{\alpha_{FP}-1} p_{TN}^{\alpha_{TN}-1} p_{FN}^{\alpha_{FN}-1}. \end{aligned} \tag{9}$$

Note that p_{TN} can be replaced using $\sum_i p_i = 1$. For the posterior distribution $\alpha = x + \nu$, where x are the confusion matrix counts and ν are the parameters of the prior. The choices considered here are: Jeffrey's prior ($\nu_i = \frac{1}{2}$), the uniform prior ($\nu_i = 1$), and the Legendre prior ($\nu_i = 2$).

Using the transformations in Eqn. 3 and marginalizing over p_{TP}, the probability distribution for precision and recall can be obtained as (see Appendix D for derivation):

$$\begin{aligned} f(R,P) &= \frac{\Gamma(\alpha_{TP}+\alpha_{FP}+\alpha_{FN})}{\Gamma(\alpha_{TP})\Gamma(\alpha_{FP})\Gamma(\alpha_{FN})} \times \\ & \left(\frac{1-P}{P}\right)^{\alpha_{FP}-1} \left(\frac{1-R}{R}\right)^{\alpha_{FN}-1} \times \\ & \left(\frac{1}{\gamma}\right)^{\alpha_{TP}+\alpha_{FP}+\alpha_{FN}} \frac{1}{R^2 P^2}\,. \end{aligned} \tag{10}$$

where

$$\gamma = \frac{1}{R} + \frac{1}{P} - 1\,. \tag{11}$$

This is a complex shape to integrate over in the R,P plane. The value of f to use as iso-contour that defines a given credible interval is determined using MC integration.

7 BIVARIATE NORMAL APPROXIMATION

An alternative method to obtain the PR uncertainty is based on the PR covariance matrix. This approach can be used to combine various sources of uncertainty, of both the training and test sets. Extendibility is achieved by summing the covariance matrices of all uncertainties.

The marginal distribution of precision or recall is a binomial distribution. As such, given sufficient sample size, the joint PR probability distribution can be readily approximated as a bivariate normal distribution, which is described by a PR covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma_R^2 & \rho\,\sigma_R\sigma_P \\ \rho\,\sigma_R\sigma_P & \sigma_P^2 \end{bmatrix}, \tag{12}$$

where σ_R (σ_P) is the width of the probability distribution in R (P), and the correlation parameter ρ signifies the linear tilt between R and P.

The linear approximation results in elliptical uncertainty contours around $(\hat{R}, \hat{P})$, which holds well for medium to high statistics datasets, but not for a confusion matrix with low statistics cells, or when the precision or recall are close to the limiting values of 0 or 1, both resulting in non-linear behaviour.

The PR covariance matrix of the test set can be derived using linear error propagation (see Appendix E), giving the following variances and covariance terms:

$$\sigma_P^2 = \frac{x_{TP}\,x_{FP}}{(x_{TP}+x_{FP})^3}\,, \tag{13}$$

$$\sigma_R^2 = \frac{x_{TP}\,x_{FN}}{(x_{TP}+x_{FN})^3}\,, \tag{14}$$

$$\begin{aligned} \sigma_{P,R} &= \sigma_{R,P} \\ &= \frac{x_{TP}\,x_{FP}\,x_{FN}}{(x_{TP}+x_{FP})^2\,(x_{TP}+x_{FN})^2}\,. \end{aligned} \tag{15}$$

For completeness, the correlation ρ, which equals $\sigma_{P,R}/\sigma_R\sigma_P$, is positive in value. This makes sense intuitively: if x_{TP} increases, both the precision and recall go up, meaning a positive correlation.

R and P can be rotated, shifted and scaled into two independent random variables $Z_1, Z_2 \sim \mathcal{N}(0, 1)$:

$$Z_1 = \frac{R - \mu_R}{\sigma_R}\,, \tag{16}$$

$$Z_2 = \frac{\frac{P-\mu_P}{\sigma_P} - \rho Z_1}{\sqrt{1-\rho^2}}\,, \tag{17}$$

yielding the joint probability distribution:

$$f_{\text{b.n.}}(Z_1, Z_2) = \frac{1}{2\pi}\exp\left[-\frac{1}{2}\left(Z_1^2 + Z_2^2\right)\right]. \tag{18}$$

The score $Z^2 = Z_1^2 + Z_2^2$ follows the χ^2 distribution with two degrees of freedom, with probability function $\mathcal{P}$ (as in Sec. 5).

8 COMBINATION OF UNCERTAINTY CONTOURS

A PR curve with uncertainty band is composed of many individual contours: each point on the curve corresponds to a different discrimination threshold and uncertainty contour. With many contours present, together these blur into a single uncertainty band. In addition, drawing individual contours or ellipse is slow when the number of thresholds is high. Because of this, an alternative computation and visualisation is presented here, one that is much faster.

Subsequent PR points are not statistically independent and neither are their uncertainty contours. No statistical correction is applied for this effect, for the following reason: only one discrimination threshold is normally used in practice, typically determined by business requirements, and one is interested in the uncorrected uncertainty on that point.

The constructed uncertainty band represent a conservative view on the uncertainty over the complete PR curve. By default the precision-recall grid is divided into 1000 bins per axis. For each point in the R, P grid the highest probability is retained. Or, put differently, the minimum Z-score observed for a grid point given the confusion matrices. This means that for any threshold, the curve's CI will never be smaller than the CI of the corresponding threshold. Hence this method can over-cover the true confidence interval.

While this procedure reduces the need to draw many contours, it still has $\mathcal{O}(tn^2)$ complexity, where n is the number of bins per axis, and t the number of available thresholds. Therefore, only the bins contained by ± 6 marginal standard deviations around $\hat{P}$ and $\hat{R}$ (see Eqns. 33,34), with a minimum (maximum) value of $0 + \epsilon$ ($1 - \epsilon$), are evaluated, where ϵ guards against floating point errors (default is 10^{-12}).

We argue that, for the purpose of visualisation, the trade-off between feasibility of generating the plot and the over-coverage is worthwhile. However, the uncertainty over the curve should not be directly used for inference or for threshold selection. For this one should use the uncertainties estimated for a single discrimination threshold.

The PR curve in Fig. 1 is obtained for a simple binary classification problem, the separation of two blobs, evaluated on a test sample of just 500 data points with a class balance of 50%, where each point on the curve corresponds to a different discrimination threshold.

The combined uncertainty band, shown in blue, is evaluated using Wilks' method. The green and red contours correspond to individual discrimination thresholds. As seen from the green contours, the combined method can slightly over-cover the true confidence interval.

Note that the precision uncertainty is relatively constant over a large part of the curve, except at low and high recall. At low recall both x_{TP} and x_{FP} decrease in value, eventually towards zero, resulting in ever larger statistical uncertainties. For example, with just one false positive remaining at a high discrimination threshold, the precision can drop down to low values. Equivalently this results in a relatively small area under the curve. At full recall the uncertainty on the precision does not shrink to zero. This simply means that, at low discrimination thresholds, the observed confusion matrix can also be explained by a slightly lower, true recall value.

9 COMPARISON OF METHODS

The four approaches are compared in terms of statistical coverage and evaluation speed.

9.1 Coverage Study

A coverage comparison study has been performed for Wilks' method and the bivariate normal approximation. We omit the Bayesian method as the credible intervals obtained from it are not directly comparable to the frequentist confidence intervals considered here. The baseline frequentist approach is also skipped given it has correct coverage by construction.

If the statistical method estimates the correct confidence interval, in $x\%$ of the tests the true population values should be contained within the interval. Let Δ_{cov} be the difference between the observed and nominal coverage for a confidence interval. If the method is correct $\Delta_{\text{cov}} \sim Normal(0, \sigma)$, where σ is due to the sample noise inherent to simulation studies with finite sample sizes.

The Δ_{cov} distributions for both Wilks' and the Bivariate method have been estimated using the procedure described in Alg. 2a. The function f computes the χ^2 score given the population and the sample confusion matrix. If the χ^2

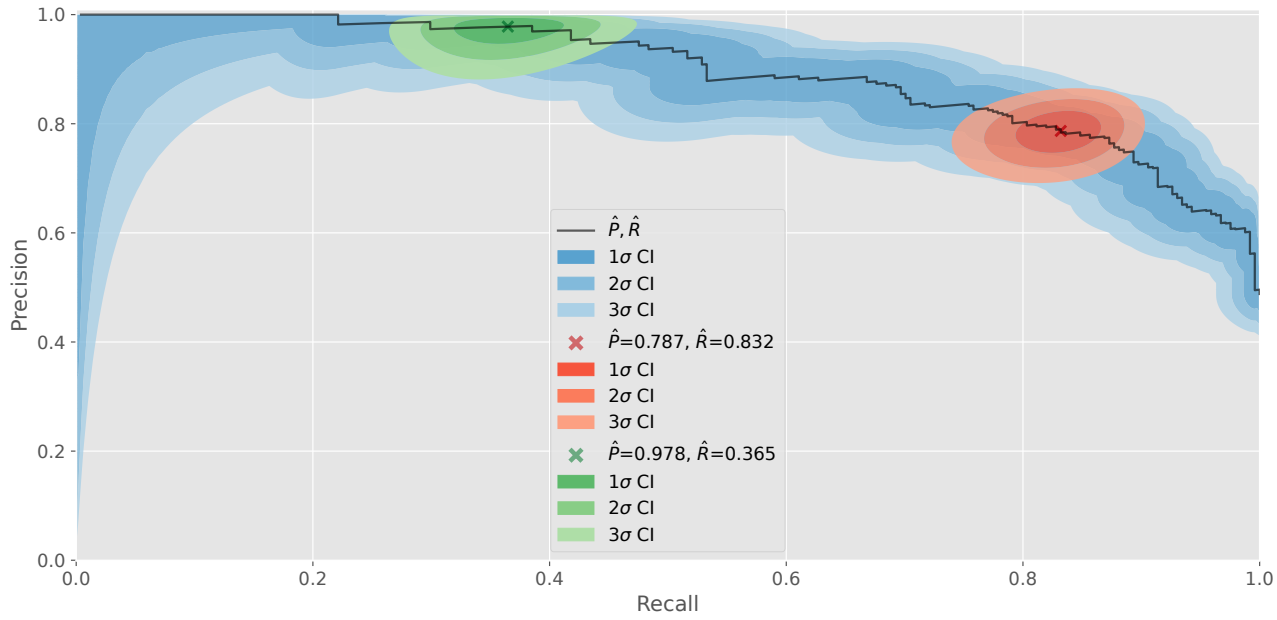


Figure 1: Shows the PR curve (black), the uncertainty over the complete curve (blue) and the uncertainties at discrimination thresholds of 0.5 (red) and 0.95 (green), as obtained with Wilks' method. The uncertainty contours are drawn at confidence levels of 68.3%, 95.4% and 99.7%. See the text for additional details.

score is smaller or equal to the corresponding critical value of a given confidence interval, the population parameters are contained within the confidence interval.

Fig. 2b shows the resulting mean Δ_{cov} over 5k (N) precision-recall scenarios with each 10k (K) simulations for test set sizes ranging from 10-100k data points. The green band reflects the sampling uncertainty inherent to the selected sample sizes, which is determined by drawing the same number of scenarios and samples from a $\chi^2(2)$ distribution that represents the null hypothesis.

For high statistics the two methods are very consistent in terms of coverage. Below 10k sample size the bivariate method starts to undercover more and more. The coverage of Wilks' method works well for test sets as small as 10 data points, which is the smallest dataset tested.

9.2 Statistical Coverage of Edge Cases

The focus lies on cases where one (or two) cells of the confusion matrix are close to zero, for which the differences in coverage are most pronounced. This happens naturally for edge cases where the recall or precision are close to zero or one. Three scenarios are explored, at low, mid and high recall, each based on the same test set as shown in Fig. 1. In Fig. 3 the uncertainty contours are drawn for each method at confidence intervals of 1, 2 and 3 standard deviations, *i.e.* at 68.3%, 95.4% and 99.7% confidence level. The results of this comparison are discussed step-by-step below.

9.2.1 Wilks' Method vs Frequentist Approach

Wilks' method is first compared with only the frequentist approach, as the former is an approximation of the latter.

The prediction of Wilks' theorem, *i.e.* asymptotically the distribution of the test statistic $q_{R,P}$ is the χ^2 distribution with two degrees of freedom, is illustrated in Appendix F. This approximation can break down for low-statistics samples or when R and P reach the edge values of 0 or 1.

Overall the the uncertainty contours obtained with Wilks' approximation (blue) and the frequentist (grey) agree very well, especially for the mid and high recall scenarios (see Fig. 3). Note that all uncertainty contours are non-elliptical in shape. The top-left graph explores the low recall, high precision scenario, and is based on a confusion matrix with just 1 false positive. Here the contours are highly non-elliptical. The two sets of contours still agree rather well, with some small deviations though in the corners of each contour.

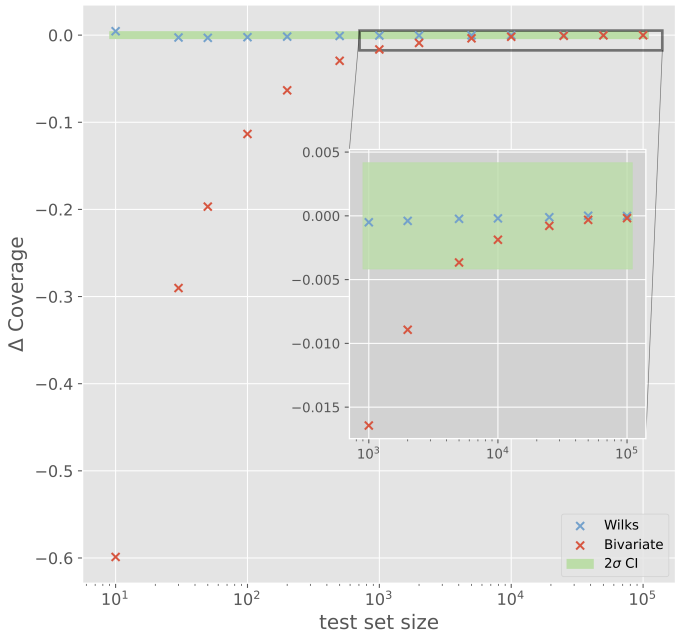
In practice, Wilks' method is not observed to break down for low-statistics samples or when R and P reach the edge values of 0 or 1. The agreement between the frequentist approach and Wilks' method is remarkably good; any differences in contours are hardly visible over the full recall and precision ranges.

$\alpha \leftarrow 0.95$; $q \leftarrow \chi^2(2).ppf(\alpha)$
$\mathbb{S} \leftarrow \{10, 30, 50, \dots, 100000\}$
$S \leftarrow |\mathbb{S}|$

$score \in \mathbb{R}$; $Y, \Delta \subset \mathbb{Z}^{S \times N}$; $\bar{\Delta} \subset \mathbb{R}^S$

for $i \in \{1, 2, \dots, S\}$ **do**
  $p \leftarrow Dirichlet(2, 1, 1, 2)$
  $prec \leftarrow p[4] \ / \ (p[4] + p[2])$
  $rec \leftarrow p[4] \ / \ (p[4] + p[3])$
  for $j \in \{1, 2, \dots, N\}$ **do**
    for $k \in \{1, 2, \dots, K\}$ **do**
      $cm \leftarrow Multinomial(\mathbb{S}[i], p)$
      $score \leftarrow f(cm, prec, rec)$
      if $score \leq q$ **then**
        $Y[i, j] \leftarrow Y[i, j] + 1$
      end if
    end for
    $\Delta[i, j] \leftarrow (Y[i, j] \ / \ K) - \alpha$
  end for
  $\bar{\Delta}[i] \leftarrow mean(\Delta[i, :])$
end for

(a) Coverage study procedure.



(b) Mean difference between observed and nominal coverage of 2σ confidence interval (95.4%) for Wilks' and the Bivariate method.

9.2.2 Bayesian Method

Compared with the frequentist methods, the Bayesian approach (green) is able to produce credible intervals for any performance metric that can be computed from the confusion matrices sampled from the posterior. The choice of prior can be used to encode pre-existing knowledge about the expected performance, whereas this is less straightforward to implement in frequentist methods. And whilst credible intervals are deemed more easily interpretable by some, the choice of prior does carry its own biases and issues requiring further choices from the user, as discussed in Caelen (2017); Tötsch and Hoffmann (2021). In the top and bottom-left scenarios of Fig. 3 Jeffrey's prior is used, which gives credible intervals that match the confidence intervals reasonably well. The edge cases are handled properly.

The bottom-right plot shows the low recall point evaluated with the Bayesian approach with Jeffrey's prior, the Legendre prior and the uniform prior. A significant dependency on prior is seen, in particular when one cell (or several) of the confusion matrix has a low number of counts.

9.2.3 Bivariate Normal Approximation

The bivariate approximation breaks down for low statistics cells or when the precision or recall care close to the limiting values of 0 or 1.

The effect of the linearity assumption is visible for all three scenarios in Fig. 3, particularly at low and high recall. In the top-left plot the confidence intervals (red) extend beyond the precision of 1, as it cannot be compressed, and it also underestimates the density below a precision of 0.95. In contrast, the other methods are capable of capturing the higher-order correlations in this region, for which the uncertainty contours are highly non-elliptical.

In fact the uncertainty contours of the bivariate approximation disappear completely when $\hat{P} = 1$, where $\sigma_P = 0$ (and the same happens for $\hat{R} = 1$, with $\sigma_R = 0$). In contrast, Wilks' method in Fig. 1 shows an increasing uncertainty for that region, which is expected given the low cell statistics.

As a rule of thumb, the bivariate normal approximation to the binomial is good when $np(1-p) \geq 10$, and improves as it becomes larger. For the mid recall scenario, with just 41 (55) false negatives (positives), the bivariate contours agree better with those of the other methods, though not yet perfectly. In the edge case of $\hat{P} \to 1$, and thus $p_{FP} \to 0$, the required number of test cases with $y = 1$ quickly approaches infinity.

In summary, the bivariate normal approximation has lower than expected coverage for low statistics confusion matrices, but this improves for high statistics samples.

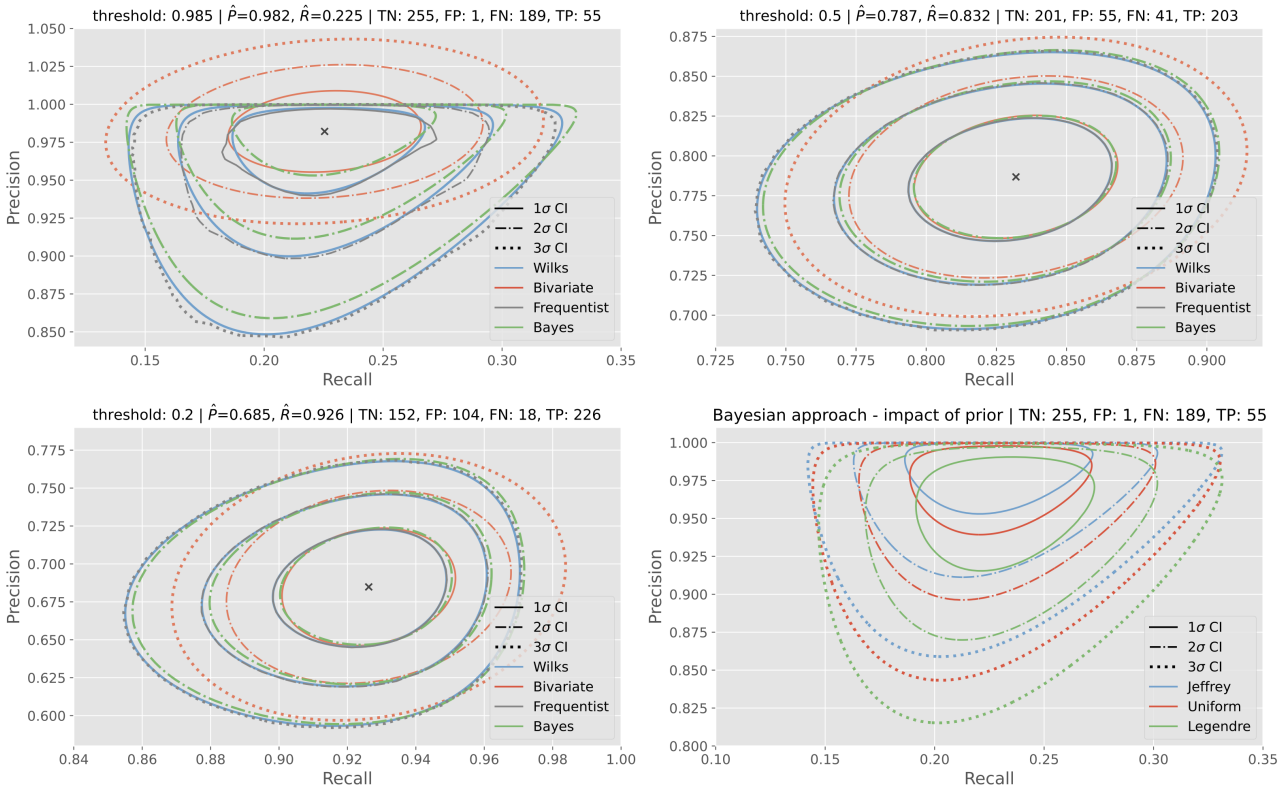


Figure 3: Example uncertainty contours at low (top left), mid (top right) and high (bottom left) recall values. The bottom right plot is the low recall scenario evaluated with the Bayesian approach using three different priors. For each method the uncertainty contours are drawn at confidence levels of 68.3%, 95.4% and 99.7%. For additional details see the text.

9.3 Evaluation Speed

The four methods have significantly different runtimes, as shown in Table 1, which reports the times to evaluate and draw the $1, 2, 3\sigma$ confidence intervals of a single discrimination threshold and the full uncertainty band. The complexity per algorithm depends on: the number of discrimination thresholds (t), the number of bins per axis (n), by default 1000, and the number of multinomial samples per grid point (s), by default 10k.

The bivariate normal approximation is fastest, particularly for a single threshold as this solely needs a covariance matrix. This is closely followed by Wilks' method, where drawing the full uncertainty band takes just a fraction of a second. Bayesian methods with conjugate prior are not computationally competitive. The reason is the MC integration required to determine the isocontour values of Eqn. 10. The frequentist approach is extremely CPU-intensive, easily requiring 100 billion multinomial samples per PR curve. While doable, in practice this is only feasible for single confusion matrices, not for a full PR curve.

10 DISCUSSION

For low statistics test samples, with fewer than several thousand data points, or when enforcing a tight discrimination threshold, resulting in near-zero numbers of false and/or true positives, the classifier uncertainty is seen to be quite impactful.

The different techniques we have examined have been summarized in Table 2. Of the evaluated methods, the frequentist approach is too slow to evaluate in practice. The bivariate normal approximation breaks down for low statistics samples or when the precision or recall are close to their edge values of 0 or 1, but can be combined easily with other learning algorithm uncertainties. The Bayesian approach has a closed-form solution, but converting this to credible uncertainty bands on the PR curve still requires (slow) MC sampling from the posterior distribution. Wilks' method is our recommendation: its uncertainty contours can be accurately evaluated compared with the frequentist approach, even when some cells of the confusion matrix contain almost no entries. Some small discrepancies are observed in coverage, but these deviations are at an acceptable level, essentially invisible in the uncertainty bands of the full PR curve.

method	complexity	time (one threshold)	time (full band)	factor
Bivariate normal	$\mathcal{O}(tn^2)$	$41.7\,\mu s \pm 4.23\,\mu s$	$114\,ms \pm 3.1\,ms$	1.0
Wilks'	$\mathcal{O}(tn^2)$	$107\,\mu s \pm 5.26\,\mu s$	$242\,ms \pm 2.59\,ms$	2.7
Bayesian	$\mathcal{O}(t(s+n^2))$	$74.9\,ms \pm 2.47\,ms$	$39.5\,s \pm 556\,ms$	$1.8 \cdot 10^3$
Frequentist	$\mathcal{O}(tsn^2)$	$47\,s \pm 514\,ms$	$7.5\,h$	$1.1 \cdot 10^6$

Table 1: Average times to evaluate the confidence intervals of a single discrimination threshold and the full uncertainty band. See the text for a description of complexity. The speed tests have been performed with $n = 1000$, $t = 500$, $s = 10k$ on an Intel Core i9 CPU with 8 cores and a clock speed of 2.3Ghz. The factor column is with respect to the bivariate normal approximation for a single threshold.

method	coverage	edge effects	low statistics	impact of prior	extendability	speed
Bivariate normal	$+/-$	$-$	$-$	N.A.	$+$	$+$
Wilks'	$+$	$+$	$+$	N.A.	$-$	$+$
Bayesian	$+$	$+$	$+$	$-$	$-$	$-$
Frequentist	$+$	$+$	$+$	N.A.	$-$	$--$

Table 2: Behaviour of statistical methods, in terms of: statistical coverage, precision or recall close to the edges of zero or one, confusion matrix with low cell counts, dependency on choice of prior, extendability with other sources of uncertainty, and speed.

The following is noted regarding the area under the PR curve, known as the metric AUCpr, which is sometimes used as reward function to optimize classifiers. This metric depends strongly on the classifier uncertainty in the low-recall region. At low recall the uncertainty on the precision grows substantially, leading to a correspondingly large uncertainty on the AUCpr. As such, unless the user has interest in the low-recall region of the PR curve, it is advisable to exclude this region from the AUCpr metric during optimization.

11 CODE

As far as the authors are aware, there is no functionality in the major numerical/ML Python libraries that allows one to easily compute and or visualise the uncertainty on the performance metrics of a classifier. We have developed an open-source package named "Model Metric Uncertainty (MMU)" that implements the classifier uncertainty methods presented in this work.[1] The methods implemented are valid for any confusion matrix, regardless of the classification setup and/or dataset origin. In the future we hope to extend MMU and integrate it into one of the major libraries.

12 CONCLUSION

This work derives the joint test-set sampling uncertainty on recall and precision (and on true positive versus false positive rate, in Appendix G), and shows how to evaluate and plot the related uncertainty band on the points of the PR (or ROC) curve. Curves with this uncertainty band give a more realistic view of the performance of a classifier, and in our view should be the new standard. This is particularly relevant for low statistics test samples, with fewer than several thousand data points, and is most impactful in the low recall region, where the uncertainty can blow up. Of the four statistical technique that have been tested in terms of coverage and speed, Wilks' method is our recommendation: its uncertainty contours can be quickly and accurately evaluated, even when some cells of the confusion matrix contain (almost) no entries. The coverage of Wilks' method works for test sets as small as 10 data points. The methods described are easy to apply through a Python library that has been made publicly available. We believe this work fills a gap in the toolbox of any data scientist.

13 BROADER IMPACT

Improvements in the estimation of the uncertainty associated with the out-of-sample performance of a model should be a broadly positive influence on the usage of machine learning discriminators in the real world. Therefore we expect the present work to be a net-positive contribution to the community. Potential risks exist if the present methods are misused (and under-represent the uncertainty), or misinterpreted as representing the full uncertainty of a model, rather than just that due to sampling variability in the test set.

Acknowledgements

We are grateful to the reviewers for their valuable time and feedback, which led to improvements in this manuscript.

[1] See for code, examples and documentation: `https://github.com/RUrlus/ModelMetricUncertainty`

References

Baak, M., Besjes, G., Côte, D., Koutsman, A., Lorenz, J., and Short, D. (2015). Histfitter software framework for statistical data analysis. *The European Physical Journal C*, 75(4):1–20.

Bengio, Y. and Grandvalet, Y. (2003). No unbiased estimator of the variance of k-fold cross-validation. *Advances in Neural Information Processing Systems*, 16.

Bertail, P., Clémençcon, S., and Vayatis, N. (2008). On bootstrapping the roc curve. *Advances in Neural Information Processing Systems*, 21.

Bouthillier, X. et al. (2021). Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3:747–769.

Boyd, K., Eng, K. H., and Page, C. D. (2013). Area under the precision-recall curve: point estimates and confidence intervals. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pages 451–466. Springer.

Caelen, O. (2017). A bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, 81(3):429–450.

Cochran, W. (1952). The χ^2 Test of Goodness of Fit. *Annals of Mathematical Statistics*, 23(3):315–345.

Cowan, G., Cranmer, K., Gross, E., and Vitells, O. (2011). Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 71(2):1–19.

Cumming, G. (2009). Inference by eye: Reading the overlap of independent confidence intervals. *Statistics in medicine*, 28(2):205–220.

Diciccio, T. J. and Romano, J. P. (1988). A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(3):338–354.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

Dusenberry, M. W., Tran, D., Choi, E., Kemp, J., Nixon, J., Jerfel, G., Heller, K., and Dai, A. M. (2020). Analyzing the role of model uncertainty for electronic health records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 204–213.

Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136.

Flach, P. and Kull, M. (2015). Precision-recall-gain curves: Pr analysis done right. *Advances in neural information processing systems*, 28.

Grau, J., Grosse, I., and Keilwagen, J. (2015). Prroc: computing and visualizing precision-recall and receiver operating characteristic curves in r. *Bioinformatics*, 31(15):2595–2597.

Hall, P., Hyndman, R. J., and Fan, Y. (2004). Nonparametric confidence intervals for receiver operating characteristic curves. *Biometrika*, 91(3):743–750.

Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506.

Langford, J. (2005). Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6(10):273–306.

Lemaitre, G. (2021). Use cross-validation results in the different curve display to add confidence intervals.

Lindley, D. V. (2000). The philosophy of statistics. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3):293–337.

Nadeau, C. and Bengio, Y. (1999). Inference for the generalization error. In Solla, S., Leen, T., and Müller, K., editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Tötsch, N. and Hoffmann, D. (2021). Classifier uncertainty: evidence, potential impact, and probabilistic treatment. *PeerJ Computer Science*, 7:e398.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1):60–62.

A UNCERTAINTY ESTIMATION BY THE BOOTSTRAP

One of the most common existing techniques to estimate sampling error in the test set is the bootstrap (a vast amount of literature exists on the bootstrap, see Diciccio and Romano (1988) for a non-exhaustive review), which in the asymptotic limit is guaranteed to converge to sampling from the population distribution, and therefore to produce valid confidence intervals.

The bootstrap as it pertains to computing confidence intervals on ROC curves was studied in particular in Hall et al. (2004); Bertail et al. (2008). These papers allude to shortcomings in the bootstrap approach, related to the need for smoothing. We also faced a form of this issue, mentioned in the main text and discussed below.

In samples with low numbers of positive examples, the bootstrap results in discrete, banded structure in the precision-recall distribution, due to particular easy/difficult examples being sampled and contributing to the the true positive/false positive counts. This small sample effect – which would be smoothed out with increasing sample size – gives the resulting 2D precision-recall contour deduced from it spurious structure, see Fig. 4. In contrast, the procedures we use correctly smooth out these banded artifacts.

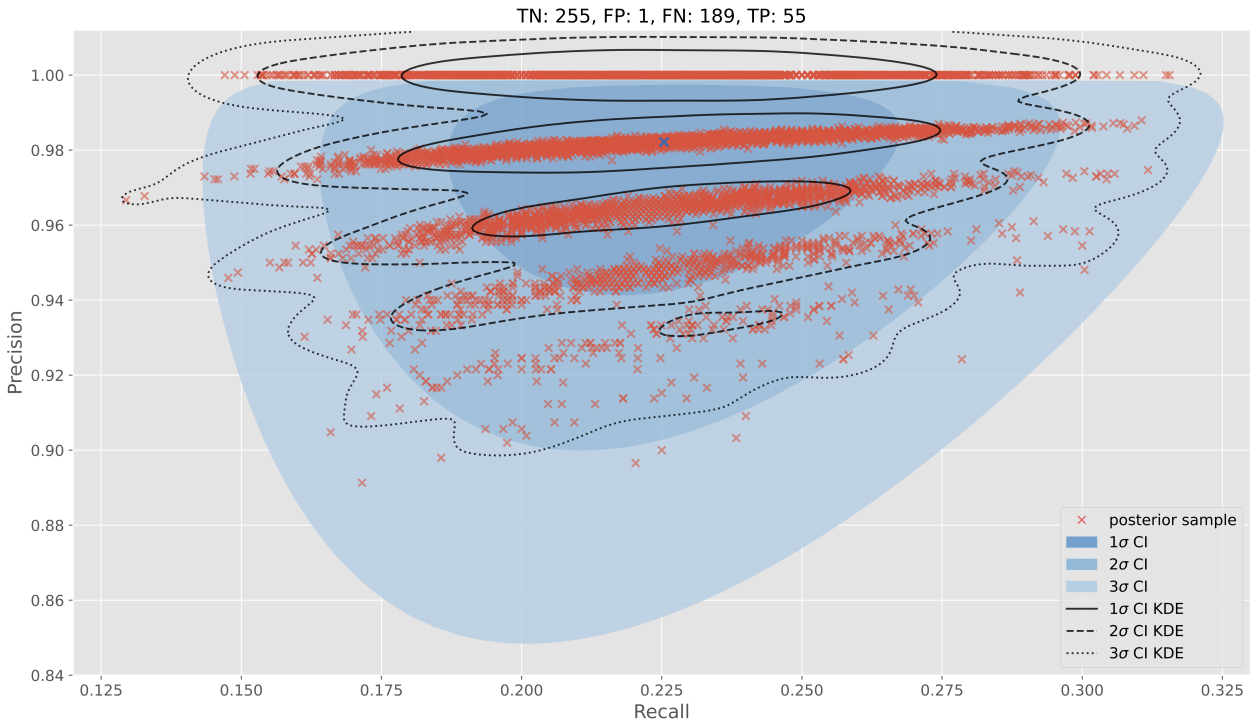


Figure 4: Example of PR distribution obtained using the bootstrap (red) on a test set with a single false positives, compared to the confidence contours (blue) obtained from Wilks' method. The uncertainty contours from both KDE and Wilks' method are drawn at confidence levels of 68.3%, 95.4% and 99.7%.

A second reason is that though deducing confidence intervals for single variables from bootstrap samples is straightforward (e.g. using percentiles), for 2D distributions determining the desired confidence contours is less unambiguous since it requires fitting a density-estimating model (such as a kernel density estimator, which is slow) to the bootstrapped precision-recall data points, and then deducing the confidence contours from the fitted density. In contrast, the procedures we propose in the main text naturally produce the required confidence contours.

B ASSUMPTION OF MULTINOMIAL DISTRIBUTION

Here the assumption is tested that the confusion matrix of an independent and identically distributed test sample can be described by a multinomial distribution. If the confusion matrix can be described by a multinomial distribution, each of the elements of the confusion matrix follow binomial distributions. Hence, the statistical uncertainty of the elements of the

confusion matrix are thus described by the second moment of a binomial.

This assumption can be easily verified using simulation studies. For a range of test set sizes, 100–100K, one computes the standard deviation of the theoretical distribution given the observed probabilities of the test set. These are then compared to the standard deviation across many hold-out sets that are identically distributed and of equal size to the test set. From Fig. 5 it is evident that the individual entries of the confusion matrix follow a binomial distribution, as the observed standard deviation over the hold-out sets closely matches the standard deviation of a binomial.

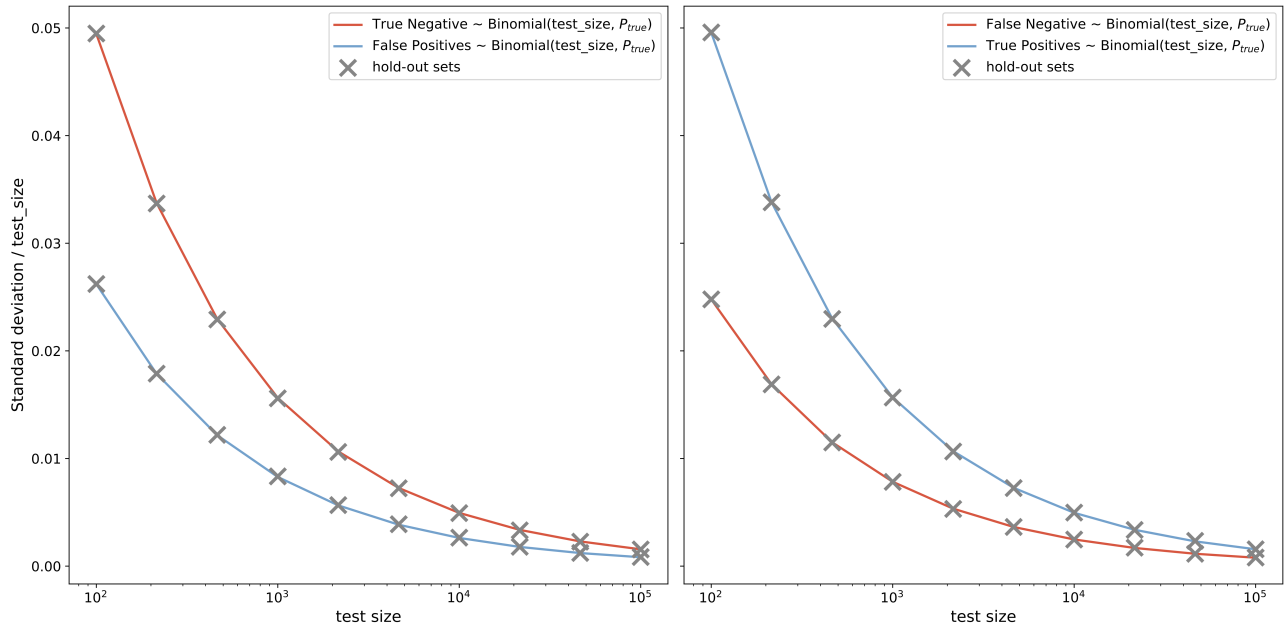


Figure 5: left: shows the σ_{tn}/n and σ_{fp}/n of the true negative and false positive elements of the confusion matrix for a sequence of test sizes. right: contains the same for the false-negative and true-positive entries. These confusion matrices are generated by fitting a logistic regression to synthetic data generated from the make_blobs dataset generator from sklearn (Pedregosa et al., 2011).

Using linear error propagation, the first-order statistical uncertainties on both P and R can be derived, as detailed in Eqns. 33 and 34 in Sec. E:

$$\sigma_P = \sqrt{\frac{x_{TP}\, x_{FP}}{(x_{TP}+x_{FP})^3}} \quad ; \quad \sigma_R = \sqrt{\frac{x_{TP}\, x_{FN}}{(x_{TP}+x_{FN})^3}}\,.$$

The same experiment is performed for the marginal standard deviation of precision and recall. As seen in Fig. 6 the observed and theoretical values are quite close even for small test set sizes. Note that the statistical uncertainties σ_P and σ_R scale roughly with $1/\sqrt{n}$, which can be seen when inserting $x_i \approx p_i n$. So the common rule of thumb applies: if the test dataset quadruples the corresponding uncertainties reduce by a factor of two.

C ANALYTICAL CALCULATION OF THE MAXIMUM LIKELIHOOD FOR FIXED RECALL AND PRECISION

Once fixing R and P, the likelihood is maximized with respect to p_{TP} only:

$$\hat{p}_{TP} = \text{argmax}_{p_{TP}}\, L(R, P, p_{TP})\,,$$

which results in the maximum likelihood value $L(R, P, \hat{\hat{p}}_{TP})$. $\hat{\hat{p}}_{TP}$ can be derived analytically by finding the maximum value of the log likelihood with respect to p_{TP}:

$$\begin{aligned}
\log(p_{MN}(p_{TP})) &= \log\left(\frac{n!}{x_{TP}!x_{FP}!x_{TN}!x_{FN}!}\, p_{TP}^{x_{TP}}\, p_{FP}^{x_{FP}}\, p_{TN}^{x_{TN}}\, p_{FN}^{x_{FN}}\right) \\
&= \log\left(c_0\right) + x_{TP}\log(p_{TP}) + x_{FP}\log(c_1 p_{TP}) \\
&\quad + x_{TN}\log\left(1 - c_3 p_{TP}\right) + x_{FN}\log(c_2 p_{TP})\,,
\end{aligned}$$

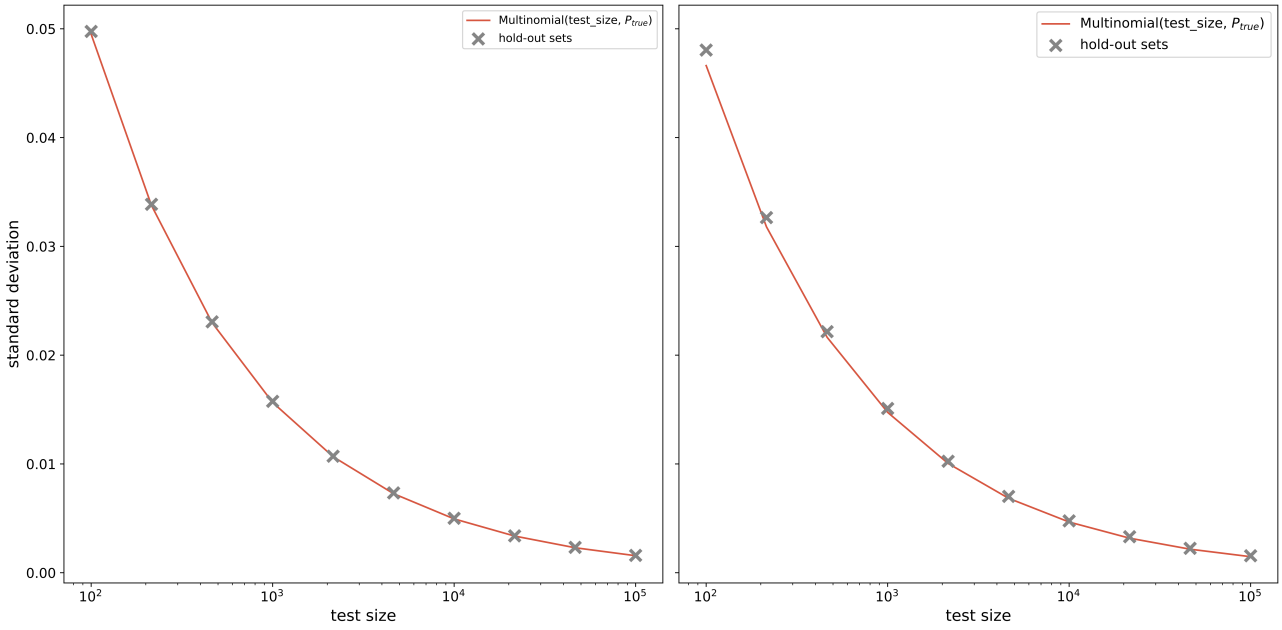


Figure 6: Shows σ_P (left) and σ_R (right) for a sequence of test sizes. See text for a description.

where c_0 is a constant, $c_1 = \frac{1-P}{P}$ and $c_2 = \frac{1-R}{R}$ and $c_3 = 1 + c_1 + c_2$. Taking the derivative to find the maximum:

$$\frac{\partial \log(p_{MN}(p_{TP}))}{\partial p_{TP}} = 0 \;\Rightarrow\; \frac{x_{TP}}{p_{TP}} + \frac{x_{FP}}{p_{TP}} + \frac{c_3 x_{TN}}{c_3 p_{TP} - 1} + \frac{x_{FN}}{p_{TP}} = 0\,,$$

and with some juggling this gives the analytical solution:

$$\hat{\hat{p}}_{TP} = \frac{x_{TP} + x_{FN} + x_{FP}}{\left(\frac{1}{P} + \frac{1}{R} - 1\right)n}\,. \tag{19}$$

D POSTERIOR PROBABILITY DISTRIBUTION OF PRECISION AND RECALL

The Dirichlet distribution is given by:

$$\mathrm{Dir}(p_1, p_2, p_3 | \bar{\alpha}) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)\Gamma(\alpha_4)}\, p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} p_3^{\alpha_3 - 1} p_4^{\alpha_4 - 1}\,, \tag{20}$$

where

$$p_1 + p_2 + p_3 + p_4 = 1\,. \tag{21}$$

Because the Dirichlet function is a conjugate prior of the multinomial distribution, the posterior takes on the same form. The probability distribution of precision and recall are are derived below from the Dirichlet function.

Given the definitions of precision P and recall R, apply the parameter transformation $(p_{TP}, p_{FP}, p_{FN}) \to (t, P, R)$:

$$\begin{aligned} p_{TP} &= t \\ p_{FP} &= \frac{1-P}{P}\, t \\ p_{FN} &= \frac{1-R}{R}\, t \\ p_{TN} &= (1 - \gamma\, t) \end{aligned} \tag{22}$$

where

$$\gamma = \frac{1}{R} + \frac{1}{P} - 1\,. \tag{23}$$

Note that $0 \leq t \leq \frac{1}{\gamma}$, as $0 \leq p_{TN} \leq 1$. The determinant of the corresponding Jacobian J is:

$$|\det(J)| = \frac{t^2}{R^2 P^2}\,. \tag{24}$$

Substituting the formulas above into Eq. 20 and integrating over the components containing t gives:

$$\begin{aligned} I &= \int_0^{\frac{1}{\gamma}} t^{\alpha_1+\alpha_2+\alpha_3-1}(1-\gamma\,t)^{\alpha_4-1}dt \\ &= \left(\frac{1}{\gamma}\right)^{\alpha_1+\alpha_2+\alpha_3} \int_0^1 x^{\alpha_1+\alpha_2+\alpha_3-1}(1-x)^{\alpha_4-1}dx \\ &= \left(\frac{1}{\gamma}\right)^{\alpha_1+\alpha_2+\alpha_3} \frac{\Gamma(\alpha_1+\alpha_2+\alpha_3)\Gamma(\alpha_4)}{\Gamma(\alpha_1+\alpha_2+\alpha_3+\alpha_4)}\,. \end{aligned} \tag{25}$$

Collecting all components results in the following probability density function for R and P:

$$f(R,P) = \frac{\Gamma(\alpha_1+\alpha_2+\alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}\,\frac{1}{R^2P^2}\left(\frac{1-P}{P}\right)^{\alpha_1-1}\left(\frac{1-R}{R}\right)^{\alpha_3-1}\left(\frac{1}{\gamma}\right)^{\alpha_1+\alpha_2+\alpha_3}\,. \tag{26}$$

Note that the parameter corresponding to the true negatives (α_4) has dropped out completely.

E BIVARIATE NORMAL APPROXIMATION

For the test set, the PR uncertainty contour can be approximated using linear error propagation, resulting in an uncertainty ellipse. To do so a linear Taylor expansion is made in both P and R. The assumption of local linearity of P and R does not generally hold, in particular when dealing with a confusion matrix with low statistics cells, or when the precision or recall are close to their limiting values of 0 or 1. Evidence of such non-linear behaviour can be seen in Fig. 3, in the differences between the bivariate normal approximation and the other methods. The approximation results in useful first-order approximations of the (co-)variances of P and R, which work well for medium to high statistics samples.

The covariance matrix of the multinomial distribution is given by:

$$\Sigma^x = \begin{bmatrix} \sigma^2_{x_1} & \dots & \sigma_{x_k,x_1} \\ \vdots & \ddots & \vdots \\ \sigma_{x_1,x_k} & \dots & \sigma^2_{x_k} \end{bmatrix} = \begin{bmatrix} np_1(1-p_1) & \dots & -np_kp_1 \\ \vdots & \ddots & \vdots \\ -np_1p_k & \dots & np_k(1-p_k) \end{bmatrix}\,, \tag{27}$$

where the variance per category is on-diagonal, and the covariance between any two categories is off-diagonal. Note that the variance formula $\sigma^2_{x_i}$ is the same as for a binomial distribution. The covariance terms σ_{x_i,x_j} tend to be smaller because they are second order in p_i. The negative covariance is understood as follows: for a fixed-size dataset, if the number of data points in one category goes up, then the (sum of) numbers in the other categories must go down.

For any test dataset, x is now defined as the counts in the confusion matrix: $x = (x_{TP}, x_{FP}, x_{TN}, x_{FN})$. Evaluate the covariance matrix by using the plug-in estimators for the probabilities: $\hat{p}_i = x_i/n$. When $x_i = 0$ (n) note that this results in values of zero variance for $\sigma^2_{x_i}$ and σ_{x_i,x_j}.

The covariance matrix of any PR point is then obtained using uncertainty propagation applied to the precision and recall formulas. Define f as the vector of model metrics to be plotted against each other: $f = (P, R)$. One wishes to evaluate the covariance matrix of f:

$$\Sigma = E[(f - E[f]) \otimes (f - E[f])]\,. \tag{28}$$

To do so one uses:

$$P = \frac{x_{TP}}{x_{TP}+x_{FP}} \quad ; \quad R = \frac{x_{TP}}{x_{TP}+x_{FN}}\,,$$

and makes a first-order Taylor expansion of f with respect to x:

$$f \approx f^0 + J \cdot x \quad ; \quad E[f] \approx f^0 + J \cdot E[x]\,, \tag{29}$$

where J is the Jacobian:

$$J = \begin{bmatrix} \frac{\partial P}{\partial x_{TP}} & \cdots & \frac{\partial P}{\partial x_{FN}} \\ \frac{\partial R}{\partial x_{TP}} & \cdots & \frac{\partial R}{\partial x_{FN}} \end{bmatrix}, \tag{30}$$

containing the partial derivatives of P and R to x.

For completeness here are the relevant partial derivatives in J:

$$\begin{aligned} \frac{\partial P}{\partial x_{TP}} &= \frac{x_{FP}}{(x_{TP}+x_{FP})^2}, \\ \frac{\partial P}{\partial x_{FP}} &= \frac{-x_{TP}}{(x_{TP}+x_{FP})^2}, \\ \frac{\partial R}{\partial x_{TP}} &= \frac{x_{FN}}{(x_{TP}+x_{FN})^2}, \\ \frac{\partial R}{\partial x_{FN}} &= \frac{-x_{TP}}{(x_{TP}+x_{FN})^2}. \end{aligned} \tag{31}$$

Using this formulation, the PR covariance matrix can linearly approximated as:

$$\begin{aligned} \Sigma &= E[(f - E[f]) \otimes (f - E[f])] \\ &\approx E[J(x - E[x]) \otimes J(x - E[x])] \\ &\approx J\, E[(x - E[x]) \otimes (x - E[x])]\, J^T \\ &\approx J\, \Sigma^x J^T, \end{aligned} \tag{32}$$

demonstrating that, for each PR point, Σ can be obtained directly from the corresponding covariance matrix of x and the Jacobian J.

This gives the following variances and covariance terms:

$$\begin{aligned} \sigma_P^2 &= \Big(\frac{\partial P}{\partial x_{TP}}\Big)^2 \sigma^2_{x_{TP}} + \Big(\frac{\partial P}{\partial x_{FP}}\Big)^2 \sigma^2_{x_{FP}} + 2\,\frac{\partial P}{\partial x_{TP}}\frac{\partial P}{\partial x_{FP}}\,\sigma_{x_{TP,FP}} \\ &= \frac{x_{TP}\,x_{FP}}{(x_{TP}+x_{FP})^3}, \end{aligned} \tag{33}$$

$$\begin{aligned} \sigma_R^2 &= \Big(\frac{\partial R}{\partial x_{TP}}\Big)^2 \sigma^2_{x_{TP}} + \Big(\frac{\partial R}{\partial x_{FN}}\Big)^2 \sigma^2_{x_{FN}} + 2\,\frac{\partial R}{\partial x_{TP}}\frac{\partial R}{\partial x_{FN}}\,\sigma_{x_{TP,FN}} \\ &= \frac{x_{TP}\,x_{FN}}{(x_{TP}+x_{FN})^3}, \end{aligned} \tag{34}$$

$$\begin{aligned} \sigma_{P,R} &= \sigma_{R,P} \\ &= \frac{\partial P}{\partial x_{TP}}\frac{\partial R}{\partial x_{TP}}\,\sigma^2_{x_{TP}} + \frac{\partial P}{\partial x_{TP}}\frac{\partial R}{\partial x_{FN}}\,\sigma_{x_{TP,FN}} + \frac{\partial P}{\partial x_{FP}}\frac{\partial R}{\partial x_{TP}}\,\sigma_{x_{FP,TP}} + \frac{\partial P}{\partial x_{FP}}\frac{\partial R}{\partial x_{FN}}\,\sigma_{x_{FP,FN}} \\ &= \frac{x_{TP}\,x_{FP}\,x_{FN}}{(x_{TP}+x_{FP})^2\,(x_{TP}+x_{FN})^2}. \end{aligned} \tag{35}$$

The partial derivatives of P and R to x_{TN} are both zero, therefore in Eqn. 32 the column and row of Σ^x corresponding to the true negatives drop out. There is still an indirect dependency of σ_P, σ_R and $\sigma_{P,R}$ on x_{TN}, albeit weakly, namely via the (co-)variance terms $\sigma_{x_i^2}$ and σ_{x_i,x_j}, in particular through n and its impact on the values of $p_{i\neq TN}$.

The statistical uncertainties σ_P and σ_R scale with $1/\sqrt{n}$, which can be seen using $x_i \approx p_i n$, so a quadratic increase in test set size results in a linear reduction of their values.

An example of the PR curve using these uncertainties is shown in Fig. 7 (in red). Note that the uncertainties σ_P and σ_R become zero when $\hat{P} = 0$ and $\hat{R} = 0$ respectively. This is an underestimation of the correct statistical uncertainty, as is visible when comparing with the uncertainty band from Wilks' approximation.

F ILLUSTRATION OF WILKS' APPROXIMATION

The prediction of Wilks' theorem – *i.e.* asymptotically the distribution of the test statistic $q_{R,P}$ is the χ^2 distribution with two degrees of freedom – can break down for low-statistics samples or when R and P approach the edge values of 0 or 1.

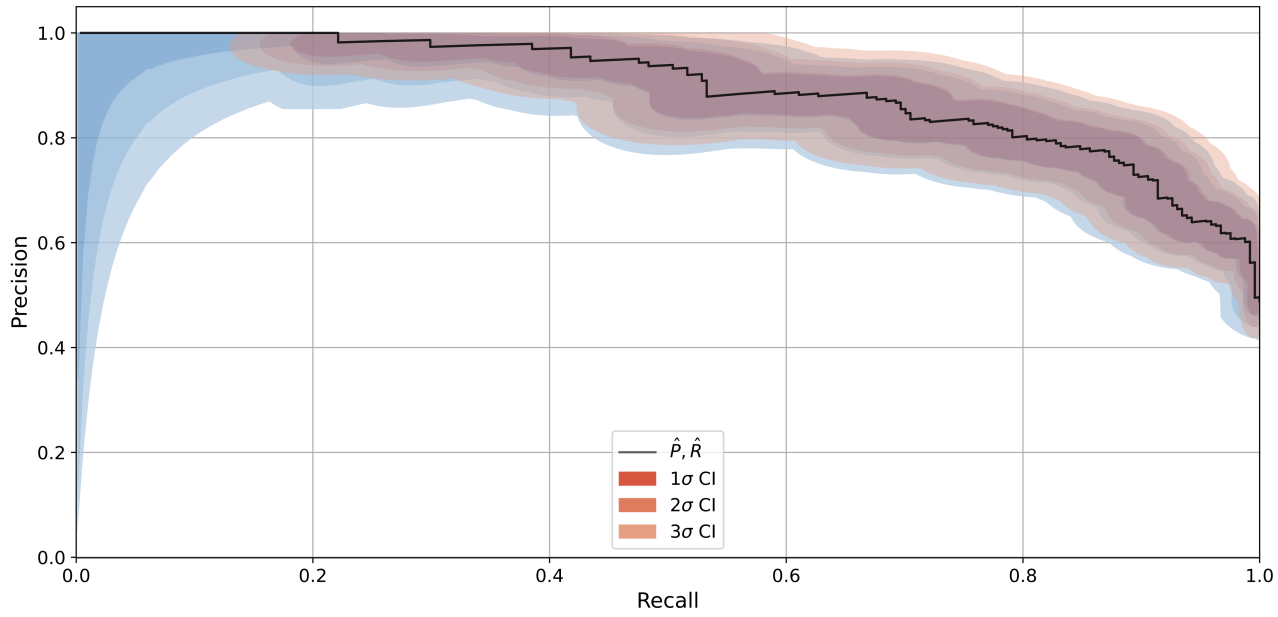


Figure 7: Shows the uncertainty over the complete curve, $|y_{\text{test}}| = 500$, using Wilks' approximation (blue) and linearly propagated errors modelled as a bivariate normal distribution (red).

Wilks' theorem is shown in action in Fig. 8. For two sets of multinomial parameters, with known true recall and precision values, 10M multinomial samples are generated of size 500, where the test statistic $q_{R,P}$ is determined for each sample. The test statistic distributions are overlaid with 10M entries drawn from a χ^2 distribution with 2 degrees of freedom.

In the top plot, with on average 40 false positives per generated sample (the smallest confusion matrix element), the two distributions overlap very well, over 6 orders of magnitude. In this asymptotic regime, the distribution of $f(q_{R,P}|R, P, p_{TP})$ has become independent of the values of the auxiliary measurements used to generate the multinomial samples, consistent with Wilks' theorem.

In the bottom plot, with on average just 5 false positives per sample, the same exponentially dropping behaviour is visible in both distributions. Bumps are visible in the test statistic distribution of the MC simulation (in blue). Each bump corresponds to fixed, low numbers of false positives in the generated samples. With a higher number of expected false positives, these bumps blur together into one smooth distribution. However the overlap between the two distributions is still nearly perfect, and enough for Wilks' to serve as a good approximation.

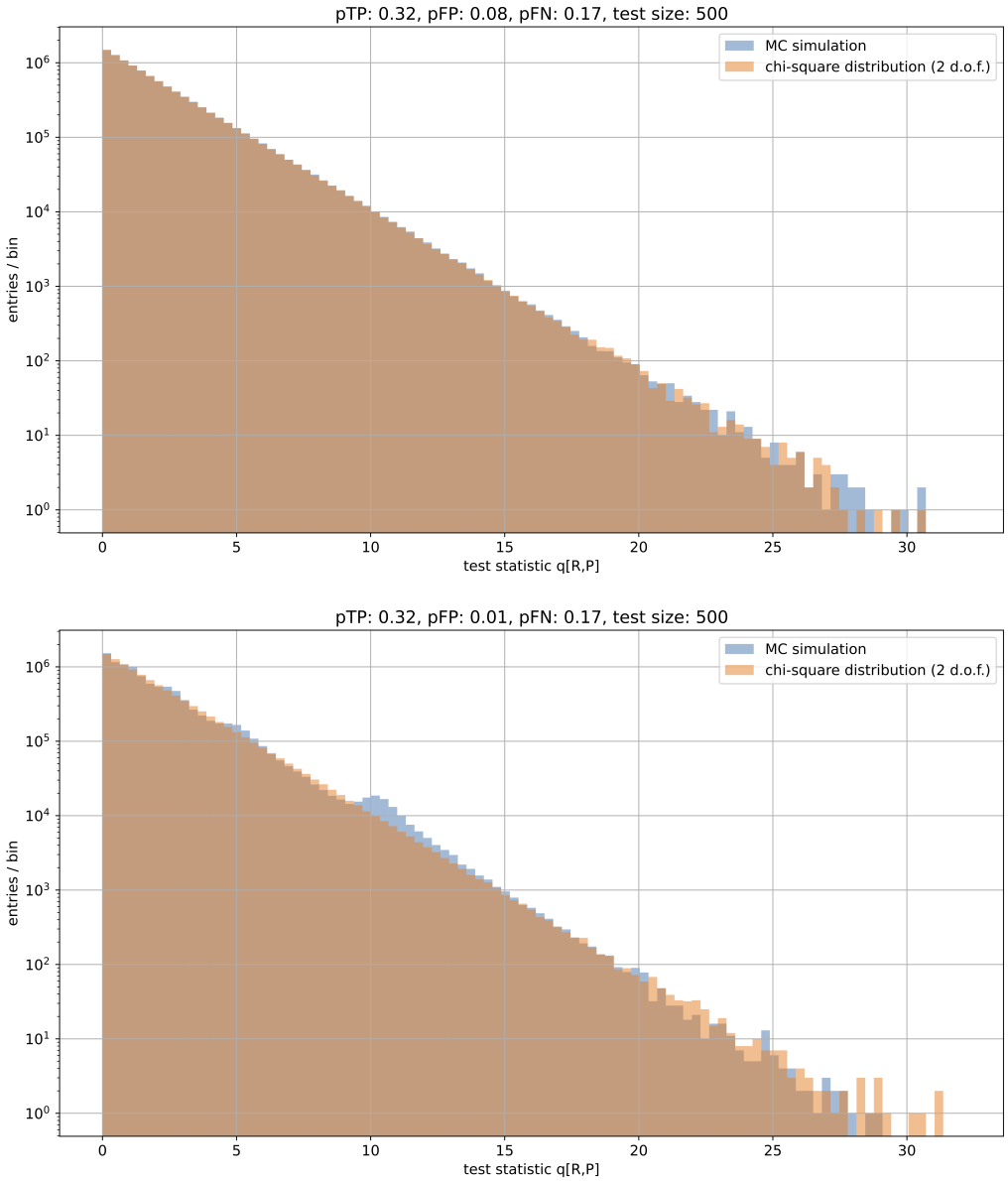


Figure 8: Wilk's theorem in action for two sets of multinomial parameter settings. See the text for a description.

G PROCEDURE FOR ROC CURVES

The Receiver Operating Characteristic (ROC) curve is an important graph that shows the True Positive Rate (T) on the y-axis as a function of the False Positive Rate (F) on the x-axis. The best values for T and F are defined as:

$$\begin{aligned} \hat{T} &= \frac{TP}{TP + FN}, \\ \hat{F} &= \frac{FP}{FP + TN}. \end{aligned} \tag{36}$$

The following subsections are similar to Appendices C, D and E, but discuss the uncertainties for the ROC curve.

G.1 Approach Based On Profile Log-Likelihood

One can compute the analytical solution of p_{TP} when the multinomial likelihood function is maximized for fixed values of T and F. In the first step one expresses p_{FN}, p_{FP} and p_{TN} as:

$$\begin{aligned} p_{FN} &= \left(\frac{1-T}{T}\right) p_{TP} \\ &= a_1 p_{TP}\,, \\ p_{FP} &= \left(\frac{F}{1-F}\right) p_{TN}\,, \\ p_{TN} &= 1 - p_{TP} - p_{FP} - p_{FN}\,, \end{aligned}$$

such that p_{FP} and p_{TN} can be rewritten as functions of T, F and p_{TP}:

$$\begin{aligned} p_{FP} &= F - \left(\frac{F}{T}\right) p_{TP} \\ &= b_0 + b_1 p_{TP}\,, \\ p_{TN} &= 1 - F + \left(\frac{F-1}{T}\right) p_{TP} \\ &= c_0 + c_1 p_{TP}\,, \end{aligned}$$

where the constants are functions of F and T. Once fixing T and F, the likelihood is maximized with respect to p_{TP}:

$$\hat{\hat{p}}_{TP} = \text{argmax}_{p_{TP}}\, L(T, F, p_{TP})\,.$$

resulting in the maximum likelihood value $L(R, P, \hat{\hat{p}}_{TP})$. The expression for $\hat{\hat{p}}_{TP}$ is found by solving for the maximum value of the log likelihood with respect to p_{TP}:

$$\begin{aligned} \log(p_{MN}(p_{TP})) &= \log(c) + x_{TP}\log(p_{TP}) + x_{FP}\log(b_0 + b_1 p_{TP}) \\ &\quad + x_{TN}\log(c_0 + c_1 p_{TP}) + x_{FN}\log(a_1 p_{TP})\,, \end{aligned}$$

where c is another constant. Taking the derivative to find the maximum:

$$\frac{\partial \log(p_{MN}(p_{TP}))}{\partial p_{TP}} = 0 \;\Rightarrow\; \frac{x_{TP}}{p_{TP}} + \frac{x_{FP} b_1}{b_0 + b_1 p_{TP}} + \frac{x_{TN} c_1}{c_0 + c_1 p_{TP}} + \frac{x_{FN}}{p_{TP}} = 0\,,$$

where after some arithmetic $\hat{\hat{p}}_{TP}$ is found to be:

$$\hat{\hat{p}}_{TP} = \frac{T\,(x_{FN} + x_{TP})}{n}\,. \tag{37}$$

This makes sense intuitively: when $T = \hat{T}$ then $\hat{\hat{p}}_{TP} = x_{TP}/n$.

G.2 Posterior Probability Distribution Of True Positive And Negative Fractions

The probability distribution of the true positive and true negative fraction are are derived below from the Dirichlet function. Given the true positive fraction T and true negavite fraction F, apply the transformation $(p_{TP}, p_{FP}, p_{FN}) \to (t, T, F)$:

$$\begin{aligned} p_{TP} &= t \\ p_{FP} &= \frac{1-T}{T}\, t \\ p_{FN} &= \frac{1-F}{F}\, s \\ &= (1-F)\left(1 - \frac{t}{T}\right) \\ p_{TN} &= s \\ &= F\left(1 - \frac{t}{T}\right), \end{aligned} \tag{38}$$

where from Eq. 21:

$$\begin{aligned} \frac{t}{T} + \frac{s}{F} &= 1 \\ s &= F\left(1 - \frac{t}{T}\right). \end{aligned} \tag{39}$$

As $0 < p_{TN} < 1$, this implies $0 < t < T$. The determinant of Jacobian J reads:

$$|\det(J)| = \frac{t}{T^2}\left(1 - \frac{t}{T}\right). \tag{40}$$

Substituting these formulas into Eq. 20 and integrating only over the pieces containing t results in:

$$\begin{aligned} I &= \int_0^T t^{\alpha_1+\alpha_2-1}\left(1 - \frac{t}{T}\right)^{\alpha_3+\alpha_4-1} dt \\ &= T^{\alpha_1+\alpha_2}\int_0^1 x^{\alpha_1+\alpha_2-1}(1-x)^{\alpha_3+\alpha_4-1} dx \\ &= T^{\alpha_1+\alpha_2}\frac{\Gamma(\alpha_1+\alpha_2)\Gamma(\alpha_3+\alpha_4)}{\Gamma(\alpha_1+\alpha_2+\alpha_3+\alpha_4)}. \end{aligned} \tag{41}$$

Collecting all components results in the probability density function for T and F:

$$f(T,F) = \frac{\Gamma(\alpha_1+\alpha_2)\Gamma(\alpha_3+\alpha_4)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)\Gamma(\alpha_4)}\,(1-T)^{\alpha_1-1}\,T^{\alpha_2-1}\,(1-F)^{\alpha_3-1}\,F^{\alpha_4-1}, \tag{42}$$

which is simply the product of two binomial distributions.

G.3 Approach Based On Bivariate Normal Approximation

Similar as in Sec. E one can compute with the linear error propagation, the (first-order) statistical uncertainties for both T and F. For completeness here are the relevant partial derivatives in J:

$$\begin{aligned} \frac{\partial T}{\partial x_{TP}} &= \frac{x_{FN}}{(x_{TP}+x_{FN})^2}, \\ \frac{\partial T}{\partial x_{FN}} &= \frac{-x_{TP}}{(x_{TP}+x_{FN})^2}, \\ \frac{\partial F}{\partial x_{FP}} &= \frac{x_{TN}}{(x_{FP}+x_{TN})^2}, \\ \frac{\partial F}{\partial x_{TN}} &= \frac{-x_{FP}}{(x_{FP}+x_{TN})^2}. \end{aligned} \tag{43}$$

The ROC covariance matrix can approximated similarly as in Eqn. 32 giving the following variances and covariance term:

$$\sigma_T^2 = \frac{x_{TP}\,x_{FN}}{(x_{TP}+x_{FN})^3}, \tag{44}$$

$$\sigma_F^2 = \frac{x_{FP}\,x_{TN}}{(x_{FP}+x_{TN})^3}, \tag{45}$$

$$\sigma_{T,F} = 0. \tag{46}$$

The covariance term $\sigma_{T,F}$ is null because T and F are independent.

G.4 Example ROC Curve

An example ROC curve using elliptical uncertainties is shown in Fig. 9.

One interesting difference with PR curves is that the uncertainty band does not blow up (as much) at low recall, as the number of true negatives generally does not go down to zero, so the uncertainty σ_F stays relatively small.

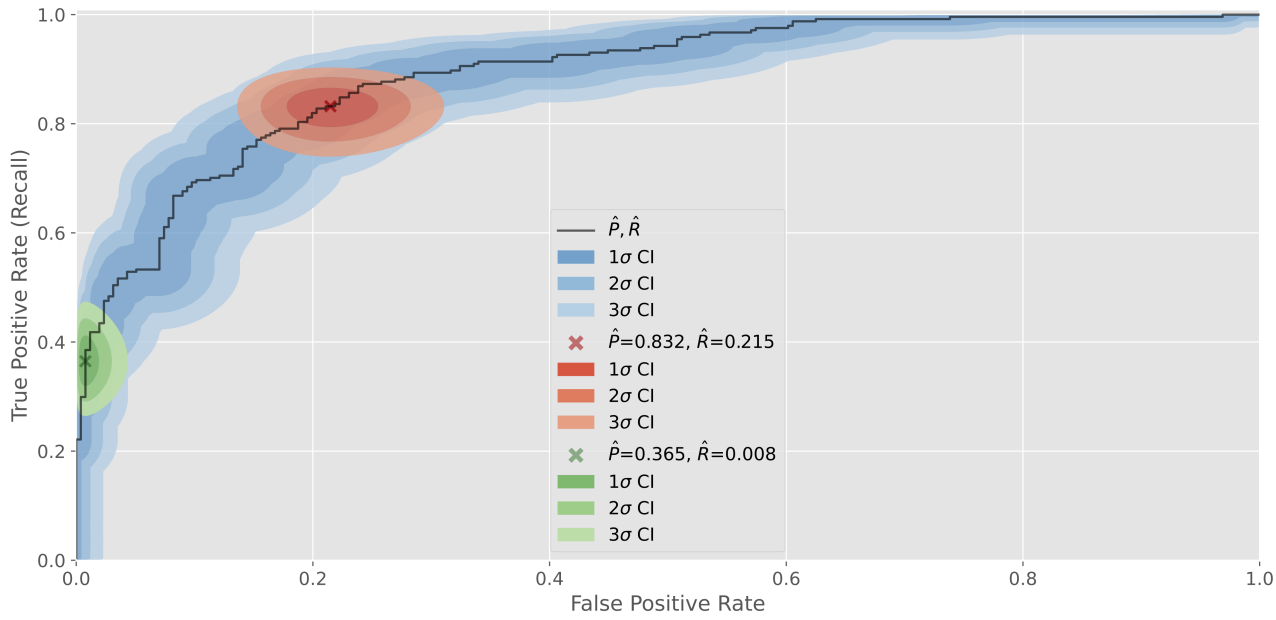


Figure 9: Example of ROC curve (black), the uncertainty over the complete curve (blue) and the uncertainties at discrimination thresholds of 0.5 (red) and 0.95 (green), as obtained with Wilks' method. The uncertainty contours are drawn at confidence levels of 68.3%, 95.4% and 99.7%.

H ALGORITHMS

This section gives the pseudo code of each of the four statistical methods. For each PR curve there is one confusion matrix per discrimination threshold, and one R, P grid per confusion matrix. The same, configurable R, P grid is used for each confusion matrix. The subsection below describes how for each statistical method the exclusion p-value or Z-score is determined for each R, P grid point. Extra explanation is first provided for the baseline frequentist method. Exclusion iso-contours are then constructed based on these values in the full R, P plane, *e.g.* the contour at $p = 0.9$ forms the 90% confidence interval. For the Bayesian approach the algorithm is presented to draw credible contours.

H.1 Frequentist Method

The frequentist approach generates multinomial samples following the procedure of Baak et al. (2015).

Given a confusion matrix, the parameters of interest are R and P, and the auxiliary parameter is p_{TP}. Since the true value of p_{TP} is unknown, ideally one scans p_{TP} for each R, P point and generates a sufficiently high number of multinomial samples for each set. In this way one can find the p_{TP} value that gives the most conservative exclusion p-value for each R, P point. For example, one cannot exclude an R, P point if there is a p_{TP} value where the exclusion p-value is greater than 10%.

This is an inefficient procedure when there is a large set of R, P values to consider. However, it turns out a good estimate can be made of which p_{TP} value maximizes the p-value per R, P point. As the p-value is based on the observed data, the largest value essentially corresponds to the scenario that is most compatible with the observed confusion matrix. Therefore one fits p_{TP} based on the confusion matrix and the hypothesized values of R and P. Based on the p_{TP} value thus found (Eqn. 5) one generates the multinomial samples that are expected to maximize the p-value for each R, P point. The observed p-value is evaluated as below. This procedure is called "the profile construction", where p_{TP} has been "profiled" on the observed data.

There is one confusion matrix per discrimination threshold, and one R, P grid per confusion matrix. Each R, P grid has a configurable number of scan points (default is 100 steps per axis). For each R, P grid point 100.000 multinomial samples are generated, and the fraction of samples is determined with $q_{R,P}$ values smaller than the value on the confusion matrix of the test set.

H.2 Pseudo Code

Algorithm 1 Frequentist Method

$x \subset \mathbb{W}^4$ ▷ the confusion matrix
$R_{grid} \in (0,1)^N \subset \mathbb{R}^N$ ▷ recall vector
$P_{grid} \in (0,1)^M \subset \mathbb{R}^M$ ▷ precision vector
$grid \leftarrow R_{grid} \otimes P_{grid}$ ▷ precision and recall grid
$p_{grid} \in [0,1]^{N \times M} \subset \mathbb{R}^{N \times M}$ ▷ p-values
procedure FREQUENTISTMETHOD($x, grid, p_{grid}$)
  $x_{TN} \leftarrow x[0]$ ▷ true negative count
  $x_{FP} \leftarrow x[1]$ ▷ false positive count
  $x_{FN} \leftarrow x[2]$ ▷ false negative count
  $x_{TP} \leftarrow x[3]$ ▷ true positive count
  $y \subset \mathbb{W}^4$
  $q_{R,P}, q_{sim} \in \mathbb{R}_+$
  for all $i \in \{0, \dots, N-1\}$ **do**
    for all $j \in \{0, \dots, M-1\}$ **do**
      $c \leftarrow 0$
      $q_{R,P} \leftarrow$ PROFILELIKELIHOOD($x_{TN}, x_{FP}, x_{FN}, x_{TP}, grid[i,j]$) ▷ See Equation 7
      $\hat{\hat{p}}_{TP} \leftarrow \mathrm{F}(x)$ ▷ See Equation 6
      for all $i \in \{0, \dots, s-1\}$ **do**
        $y \leftarrow$ MULTINOMIAL($grid[i,j], \hat{\hat{p}}_{TP}$)) ▷ See Equation 3
        $q_{sim} \leftarrow$ PROFILELIKELIHOOD($y_{TN}, y_{FP}, y_{FN}, y_{TP}, grid[i,j]$)
        if $q_{sim} < q_{R,P}$ **then**
          $c \leftarrow c + 1$
        end if
      end for
      $p_{grid}[i,j] \leftarrow c/s$
    end for
  end for
end procedure

Algorithm 2 Wilks' method

$x \subset \mathbb{W}^4$ ▷ the confusion matrix
$R_{grid} \in (0,1)^N \subset \mathbb{R}^N$ ▷ recall vector
$P_{grid} \in (0,1)^M \subset \mathbb{R}^M$ ▷ precision vector
$grid \leftarrow R_{grid} \otimes P_{grid}$ ▷ precision and recall grid
$f_{grid} \subset \mathbb{R}^{N \times M}$ ▷ likelihood scores
procedure WILKSMETHOD($x, grid, f_{grid}$)
  $x_{TN} \leftarrow x[0]$ ▷ true negative count
  $x_{FP} \leftarrow x[1]$ ▷ false positive count
  $x_{FN} \leftarrow x[2]$ ▷ false negative count
  $x_{TP} \leftarrow x[3]$ ▷ true positive count
  for all $i \in \{0, \dots, N-1\}$ **do**
    for all $j \in \{0, \dots, M-1\}$ **do**
      $f_{grid}[i,j] \leftarrow$ PROFILELIKELIHOOD($x_{TN}, x_{FP}, x_{FN}, x_{TP}, grid[i,j]$) ▷ See Equation 7
    end for
  end for
end procedure

Algorithm 3 Bayesian approach

$CL \leftarrow 0.954$ ▷ confidence level, by default 2 standard deviations
$x \subset \mathbb{W}^4$ ▷ the confusion matrix
$v \subset \mathbb{R}^4$ ▷ prior, by default Uniform prior
$R_{grid} \in (0,1)^N \subset \mathbb{R}^N$ ▷ recall vector
$P_{grid} \in (0,1)^M \subset \mathbb{R}^M$ ▷ precision vector
$grid \leftarrow R_{grid} \otimes P_{grid}$ ▷ precision and recall grid
$f_{grid} \subset \mathbb{R}_+^{N \times M}$ ▷ likelihood scores
procedure BAYESIANMETHOD($x, v, CL, f_{grid}, grid$)
  $\alpha \subset \mathbb{R}^4$
  $\alpha \leftarrow v + x$
  $p \in (0,1)^4 \subset \mathbb{R}^4$
  $q \subset \mathbb{R}_+^s$
  for $i \in \{0, \ldots, s-1\}$ **do**
    $p \leftarrow$ DIRICHLET(α)
    $R \leftarrow$ RECALL(p)
    $P \leftarrow$ PRECISION(p)
    $q[i] \leftarrow$ F(R, P, α) ▷ See Equation 10
  end for
  SORT(q)
  $q_\sigma \leftarrow$ QUANTILE(q, CL)
  for $i \in \{0, \ldots, N-1\}$ **do**
    for $j \in \{0, \ldots, M-1\}$ **do**
      $f_{grid}[i,j] \leftarrow$ F($grid[i,j], \alpha$) ▷ See Equation 10
    end for
  end for
  CONTOUR($grid, f_{grid}, q_\sigma$)
end procedure

Algorithm 4 Bivariate normal method

$x \subset \mathbb{W}^4$ ▷ the confusion matrix
$R_{grid} \in (0,1)^N \subset \mathbb{R}^N$ ▷ recall vector
$P_{grid} \in (0,1)^M \subset \mathbb{R}^M$ ▷ precision vector
$grid \leftarrow R_{grid} \otimes P_{grid}$ ▷ precision and recall grid
$f_{grid} \subset \mathbb{R}^{N \times M}$ ▷ likelihood scores
procedure BIVARIATENORMALMETHOD($x, grid, f_{grid}$)
  $x_{TN} \leftarrow x[0]$ ▷ true negative count
  $x_{FP} \leftarrow x[1]$ ▷ false positive count
  $x_{FN} \leftarrow x[2]$ ▷ false negative count
  $x_{TP} \leftarrow x[3]$ ▷ true positive count
  for all $i \in \{0, \ldots, N-1\}$ **do**
    for all $j \in \{0, \ldots, M-1\}$ **do**
      $f_{grid}[i,j] \leftarrow$ ZSCORE($x_{TN}, x_{FP}, x_{FN}, x_{TP}, grid[i,j]$) ▷ See Equations 16, 17
    end for
  end for
end procedure