Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals

Kendrick Boyd[1], Kevin H. Eng[2], and C. David Page[1]

[1] University of Wisconsin-Madison, Madison, WI
boyd@cs.wisc.edu,page@biostat.wisc.edu
[2] Roswell Park Cancer Institute, Buffalo, NY
Kevin.Eng@RoswellPark.org

Abstract. The area under the precision-recall curve (AUCPR) is a single number summary of the information in the precision-recall (PR) curve. Similar to the receiver operating characteristic curve, the PR curve has its own unique properties that make estimating its enclosed area challenging. Besides a point estimate of the area, an interval estimate is often required to express magnitude and uncertainty. In this paper we perform a computational analysis of common AUCPR estimators and their confidence intervals. We find both satisfactory estimates and invalid procedures and we recommend two simple intervals that are robust to a variety of assumptions.

1 Introduction

Precision-recall (PR) curves, like the closely-related receiver operating characteristic (ROC) curves, are an evaluation tool for binary classification that allows the visualization of performance at a range of thresholds. PR curves are increasingly used in the machine learning community, particularly for imbalanced data sets where one class is observed more frequently than the other class. On these imbalanced or skewed data sets, PR curves are a useful alternative to ROC curves that can highlight performance differences that are lost in ROC curves [1]. Besides visual inspection of a PR curve, algorithm assessment often uses the area under a PR curve (AUCPR) as a general measure of performance irrespective of any particular threshold or operating point (e.g., [2,3,4,5]).

Machine learning researchers build a PR curve by first plotting precision-recall pairs, or points, that are obtained using different thresholds on a probabilistic or other continuous-output classifier, in the same way an ROC curve is built by plotting true/false positive rate pairs obtained using different thresholds. Davis and Goadrich [6] showed that for any fixed data set, and hence fixed numbers of actual positive and negative examples, points can be translated between the two spaces. After plotting the points in PR space, we next seek to construct a curve and compute the AUCPR and to construct 95% (or other) confidence intervals (CIs) around the curve and the AUCPR.

However, the best method to construct the curve and calculate area is not readily apparent. The PR points from a small data set are shown in Fig. 1. Questions

H. Blockeel et al. (Eds.): ECML PKDD 2013, Part III, LNAI 8190, pp. 451–466, 2013.

immediately arise about what to do with multiple points with the same x-value (recall), whether linear interpolation is appropriate, whether the maximum precision for each recall are representative, if convex hulls should be used as in ROC curves, and so on. There are at least four distinct methods (with several variations) that have been used in machine learning, statistics, and related areas to compute AUCPR, and four methods that have been used to construct CIs. The contribution of this paper is to discuss and analyze eight estimators and four CIs empirically. We provide evidence in favor of computing AUCPR using the *lower trapezoid*, *average precision*, or *interpolated median* estimators and using *binomial* or *logit* CIs rather than other methods that include the more widely-used (in machine learning) ten-fold *cross-validation*. The differences in results using these approaches are most striking when data are highly skewed, which is exactly the case when PR curves are most preferred over ROC curves.

Section 2 contains a review of PR curves, Section 3 describes the estimators and CIs we evaluate, and Section 4 presents case studies of the estimators and CIs in action.

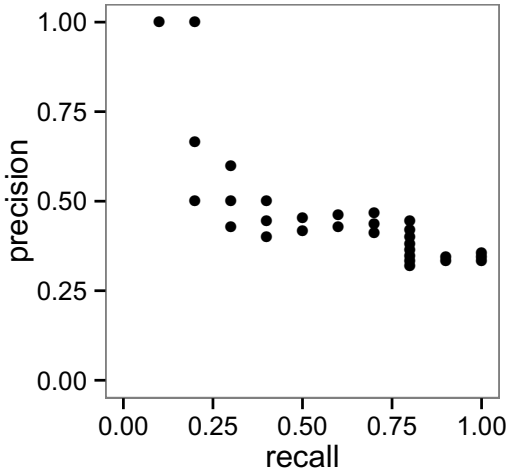

Fig. 1. Empirical PR points obtained from a small data set with 10 positive examples and 20 negative examples

2 Area Under the Precision-Recall Curve

Consider a binary classification task where models produce continuous outputs, denoted Z, for each example. Diverse applications are subsumed by this setup, e.g., a medical test to identify diseased and disease-free patients, a document ranker to distinguish relevant and non-relevant documents to a query, and generally any binary classification task. The two categories are often naturally labelled as positive (e.g., diseased, relevant) or negative (e.g., disease-free, non-relevant). Following the literature on ROC curves [7,8], we denote the output values for the negative examples by X and the output values for the positive examples by Y (Z is a mixture of X and Y). These populations are assumed to be independent when the class is known. Larger output values are associated with positive examples, so for a given threshold c, an example is predicted positive if its value is greater than c. We represent the category (or class) with the indicator variable D where $D = 1$ corresponds to positive examples and $D = 0$ to negative examples. An important aspect of a task or data set is the class skew $\pi = P(D = 1)$. Skew is also known as prevalence or a prior class distribution.

Several techniques exist to assess the performance of binary classification across a range of thresholds. While ROC analysis is the most common, we are

interested in the related PR curves. A PR curve may be defined as the set of points:

$$PR(\cdot) = \{(Recall(c), Prec(c)), -\infty < c < \infty\}$$

where $Recall(c) = P(Y > c)$ and $Prec(c) = P(D = 1|Z > c)$. Recall is equivalent to true positive rate or sensitivity (the y-axis in ROC curves), while precision is the same as positive predictive value. Since larger output values are assumed to be associated with positive examples, as c decreases, $Recall(c)$ increases to one and $Prec(c)$ decreases to π. As c increases, $Prec(c)$ reaches one as $Recall(c)$ approaches zero under the condition that "the first document retrieved is relevant" [9]. In other words, whether the example with the largest output value is positive or negative greatly changes the PR curve (approaching $(0, 1)$ if positive and $(0, 0)$ if negative). Similarly, estimates of precision for recall near 0 tend to have high variance, and this is a major difficulty in constructing PR curves.

It is often desirable to summarize the PR curve with a single scalar value. One summary is the area under the PR curve (AUCPR), which we will denote θ. Following the work of Bamber [7] on ROC curves, AUCPR is an average of the precision weighted by the probability of a given threshold.

$$\theta = \int_{-\infty}^{\infty} Prec(c) dP(Y \leq c) \tag{1}$$

$$= \int_{-\infty}^{\infty} P(D = 1|Z > c) dP(Y \leq c). \tag{2}$$

By Bayes' rule and using that Z is a mixture of X and Y,

$$P(D = 1|Z > c) = \frac{\pi P(Y > c)}{\pi P(Y > c) + (1 - \pi) P(X > c)}$$

and we note that $0 \leq \theta \leq 1$ since $Prec(c)$ and $P(Y \leq c)$ are bounded on the unit square. Therefore, θ might be viewed as a probability. If we consider Eq. (2) as an importance-sampled Monte Carlo integral, we may interpret θ as the fraction of positive examples among those examples whose output values exceed a randomly selected $c \sim Y$ threshold.

3 AUCPR Estimators

In this section we summarize point estimators for θ and then introduce CI methods.

3.1 Point Estimators

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ represent observed output values from negative and positive examples, respectively. The skew π is assumed to be given or is set

to $n/(n+m)$. An empirical estimate of the PR curve, $\widehat{PR}(\cdot)$, can be derived by the empirical estimates of each coordinate:

$$\widehat{Recall}(c) = n^{-1} \sum_{i=1}^{n} I(Y_i > c)$$

$$\widehat{Prec}(c) = \frac{\pi \widehat{Recall}(c)}{\pi \widehat{Recall}(c) + (1-\pi) m^{-1} \sum_{j=1}^{m} I(X_j > c)}$$

where $I(A)$ is the indicator function for event A.

We review a number of possible estimators for θ.

Trapezoidal Estimators. For fixed $\widehat{Recall}(t)$, the estimated precision may not be constant (so $\widehat{PR}(\cdot)$ is often not one-to-one). This corresponds to cases where an observed $Y_{(i)} < X_j < Y_{(i+1)}$ for some i and j where $Y_{(i)}$ denotes the ith order statistic (ith largest value among the Y_i's). As the threshold is increased from $Y_{(i)}$ to X_j, recall remains constant while precision decreases. Let $r_i = \widehat{Recall}(Y_{(n-i)})$, so that $r_1 \leq r_2 \leq \cdots \leq r_n$, and p_i^{max} be the largest sample precision value corresponding to r_i. Likewise, let p_i^{min} be the smallest sample precision value corresponding to r_i. This leads immediately to a few choices for estimators based on the empirical curve using trapezoidal estimation [10].

$$\widehat{\theta}_{LT} = \sum_{i=1}^{n-1} \frac{p_i^{min} + p_{i+1}^{max}}{2} (r_{i+1} - r_i) \tag{3}$$

$$\widehat{\theta}_{UT} = \sum_{i=1}^{n-1} \frac{p_i^{max} + p_{i+1}^{max}}{2} (r_{i+1} - r_i) \tag{4}$$

corresponding to a *lower trapezoid* (Eq. (3)) and *upper trapezoid* (Eq. (4)) approximation. Note the *upper trapezoid* method uses an overly optimistic linear interpolation [6]; we include it for comparison as it is one of the first methods a non-expert is likely to use due to its similarity to estimating area under the ROC curve.

Interpolation Estimators. As suggested by Davis and Goadrich [6] and Goadrich et al. [1], we use PR space interpolation as the basis for several estimators. These methods use the non-linear interpolation between known points in PR space derived from a linear interpolation in ROC space.

Davis and Goadrich [6] and Goadrich et al. [1] examine the interpolation in terms of the number of true positives and false positives corresponding to each PR point. Here we perform the same interpolation, but use the recall and precision of the PR points directly, which leads to the surprising observation that the interpolation (from the same PR points) does not depend on π.

Theorem 1. *For two points in PR space* (r_1, p_1) *and* (r_2, p_2) *(assume WLOG* $r_1 < r_2$*), the interpolation for recall* r' *with* $r_1 \leq r' \leq r_2$ *is*

$$p' = \frac{r'}{ar' + b} \tag{5}$$

where

$$a = 1 + \frac{(1-p_2)r_2}{p_2(r_2 - r_1)} - \frac{(1-p_1)r_1}{p_1(r_2 - r_1)}$$
$$b = \frac{(1-p_1)r_1}{p_1} - \frac{(1-p_2)r_1 r_2}{p_2(r_2 - r_1)} + \frac{(1-p_1)r_1^2}{p_1(r_2 - r_1)}$$

Proof. First, we convert the points to ROC space. Let s_1, s_2 be the false positive rates for the points (r_1, p_1) and (r_2, p_2), respectively. By definition of false positive rate,

$$s_i = \frac{(1-p_i)\pi r_i}{p_i(1-\pi)}. \tag{6}$$

A linear interpolation in ROC space for $r_1 \leq r' \leq r_2$ has a false positive rate of

$$s' = s_1 + \frac{r' - r_1}{r_2 - r_1}(s_2 - s_1). \tag{7}$$

Then convert back to PR space using

$$p' = \frac{\pi r'}{\pi r + (1-\pi)s'}. \tag{8}$$

Substituting Eq. (7) into Eq. (8) and using Eq. (6) for s_1 and s_2, we have

$$\begin{aligned} p' &= \pi r' \left[\pi r' + \frac{\pi(1-p_1)r_1}{p_1} + \frac{\pi(r' - r_1)}{r_2 - r_1}\left(\frac{(1-p_2)r_2}{p_2} - \frac{(1-p_1)r_1}{p_1}\right)\right]^{-1} \\ &= r' \left[r'\left(1 + \frac{(1-p_2)r_2}{p_2(r_2 - r_1)} - \frac{(1-p_1)r_1}{p_1(r_2 - r_1)}\right) + \right. \\ &\quad \left. \frac{(1-p_1)r_1}{p_1} - \frac{(1-p_2)r_1 r_2}{p_2(r_2 - r_1)} + \frac{(1-p_1)r_1^2}{p_1(r_2 - r_1)}\right]^{-1} \end{aligned}$$

□

Thus, despite PR space being sensitive to π and the translation to and from ROC space depending on π, the interpolation in PR space *does not* depend on π. One explanation is that each particular PR space point inherently contains the information about π, primarily in the precision value, and no extra knowledge of π is required to perform the interpolation.

The area under the interpolated PR curve between these two points can be calculated analytically using the definite integral:

$$\begin{aligned} \int_{r_1}^{r_2} \frac{r'}{ar' + b}\, dr' &= \left.\frac{br' - a\log(a + br')}{b^2}\right|_{r'=r_1}^{r'=r_2} \\ &= \frac{br_2 - a\log(a + br_2) - br_1 + a\log(a + br_1)}{b^2} \end{aligned}$$

where a and b are defined as in Theorem 1.

With the definite integral to calculate the area between two PR points, the question is: which points should be used. The achievable PR curve of Davis of Goadrich [6] uses only those points (translated into PR space) that are on the ROC convex hull. We also use three methods of summarizing from multiple PR points at the same recall to a single PR point to interpolate through. The summaries we use are the max, mean, and median of all p_i for a particular r_i. So we have four estimators using interpolation: *convex*, *max*, *mean*, and *median*.

Average Precision. Avoiding the empirical curve altogether, a plug-in estimate of θ, known in information retrieval as *average precision* [11], is:

$$\widehat{\theta}_A = \frac{1}{n}\sum_{i=1}^{n} \widehat{Prec}(Y_i) \tag{9}$$

which replaces the distribution function $P(Y \leq c)$ in Eq. (2) with its empirical cumulative distribution function.

Binormal Estimator. Conversely, a fully parametric estimator may be constructed by assuming that $X_j \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y_j \sim \mathcal{N}(\mu_y, \sigma_y^2)$. In this *binormal* model [12], the MLE of θ is

$$\widehat{\theta}_B = \int_0^1 \frac{\pi t}{\pi t + (1-\pi)\Phi\left(\frac{\hat{\mu}_y - \hat{\mu}_x}{\sigma_x} + \frac{\hat{\sigma}_y}{\hat{\sigma}_x}\Phi^{-1}(t)\right)}\, dt \tag{10}$$

where $\hat{\mu}_x, \hat{\sigma}_x, \hat{\mu}_y, \hat{\sigma}_y$ are sample means and variances of X and Y and $\Phi(t)$ is the standard normal cumulative distribution function.

3.2 Confidence Interval Estimation

Having discussed AUCPR estimators, we now turn our attention to computing confidence intervals (CIs) for these estimators. Our goal is to determine a simple, accurate interval estimate that is logistically easy to implement. We will compare two computationally intensive methods against two simple statistical intervals.

Bootstrap Procedure. A common approach is to use a *bootstrap* procedure to estimate the variation in the data and to either extend a symmetric, normal-based interval about the point estimate or to take the empirical quantiles from resampled data as interval bounds [13]. Because the relationship between the number of positive examples n and negative examples m is crucial for estimating PR points and hence curves, we recommend using stratified bootstrap so π is preserved exactly in all replicates. In our simulations we chose to use empirical quantiles for the interval bounds and perform 1000 bootstrap replicates.

Cross-Validation Procedure. Similarly, a *cross-validation* approach is a wholly data driven method for simultaneously producing the train/test splits required for unbiased estimation of future performance and producing variance estimates. In k-fold cross-validation, the available data are partitioned into k folds. $k-1$ folds are used for training while the remaining fold is used for testing. By performing evaluation on the results of each fold separately, k estimates of performance are obtained. A normal approximation of the interval can be constructed using the mean and variance of the k estimates. For more details and discussion of k-fold cross-validation, see Dietterich [14]. For our case studies we use the standard $k = 10$.

Binomial Interval. Recalling that $0 \leq \theta \leq 1$, we may interpret $\hat{\theta}$ as a probability associated with some binomial$(1, \theta)$ variable. If so, a CI for θ can be constructed through the standard normal approximation:

$$\hat{\theta} \pm \Phi_{1-\alpha/2} \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

We use n for the sample size as opposed to $n+m$ because n specifies the (maximum) number of unique recall values in $\widehat{PR}(\cdot)$. The *binomial* method can be applied to any $\hat{\theta}$ estimate once it is derived. A weakness of this estimate is that it may produce bounds outside of $[0, 1]$, even though $0 \leq \theta \leq 1$.

Logit Interval. To obtain an interval which is guaranteed to produce endpoints in $[0, 1]$, we may use the logistic transformation $\hat{\eta} = \log \frac{\hat{\theta}}{(1-\hat{\theta})}$ where $\hat{\tau} = s.e.(\hat{\eta}) = (n\hat{\theta}(1-\hat{\theta}))^{-1/2}$ by the delta method [15].

On the logistic scale, an interval for η is $\hat{\eta} \pm \Phi_{1-a/2}\hat{\tau}$. This can be converted pointwise to produce an asymmetric *logit* interval bounded in $[0, 1]$:

$$\left[\frac{e^{\hat{\eta} - \Phi(1-\alpha/2)\hat{\tau}}}{1 + e^{\hat{\eta} - \Phi(1-\alpha/2)\hat{\tau}}}, \frac{e^{\hat{\eta} + \Phi(1-\alpha/2)\hat{\tau}}}{1 + e^{\hat{\eta} + \Phi(1-\alpha/2)\hat{\tau}}} \right].$$

4 Case Studies

We use simulated data to evaluate the merits of the candidate point and interval estimates discussed in Section 3 with the goal of selecting a subset of desirable procedures.[1] The ideal point estimate would be unbiased, robust to various distributional assumptions on X and Y, and have good convergence as $n+m$ increases. A CI should have appropriate coverage, and smaller widths of the interval are preferred over larger widths.

[1] R code for the estimators and simulations may be found at `http://pages.cs.wisc.edu/~boyd/projects/2013ecml_aucprestimation/`

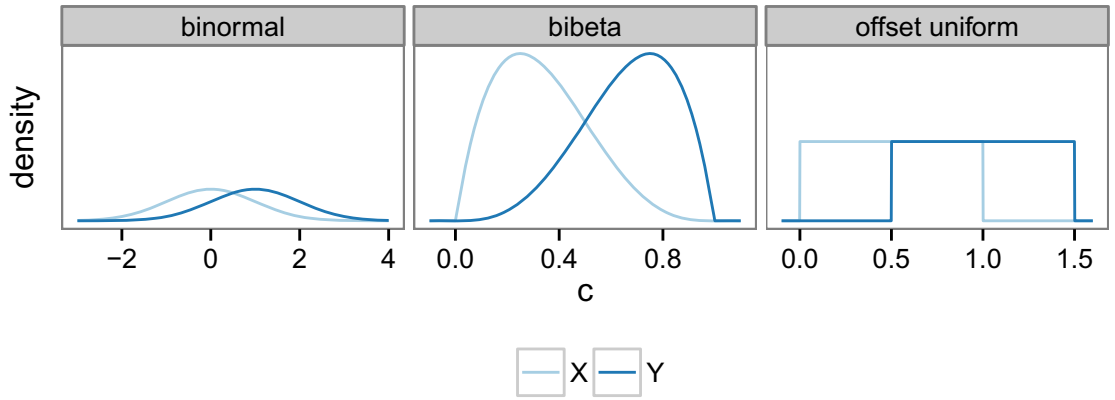

Fig. 2. Probability density functions for X (negative) and Y (positive) output values for binormal ($X \sim N(0,1), Y \sim N(1,1)$), bibeta ($X \sim B(2,5), Y \sim B(5,2)$), and offset uniform ($X \sim U(0,1), Y \sim U(0.5,1.5)$) case studies

We consider three scenarios for generating output values X and Y. Our intention is to cover representative but not exhaustive cases whose conclusions will be relevant generally. The densities for these scenarios are plotted in Fig. 2. The true PR curves (calculated using the cumulative distribution functions of X and Y) for $\pi = 0.1$ are shown in Fig. 3. Fig. 3 also contains sample empirical PR curves that result from drawing data from X and Y. These are the curves the estimators work from, attempting to recover the area under the true curve as accurately as possible.

For unbounded continuous outputs, the binormal scenario assumes that $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(\mu,1)$ where $\mu > 0$. The distance between the two normal distributions, μ, controls the discriminative ability of the assumed model. For test values bounded on $[0,1]$ (such as probability outputs), we replace the normal distribution with a beta distribution. So the bibeta scenario has $X \sim B(a,b)$ and $Y \sim B(b,a)$ where $0 < a < b$. The larger the ratio between a and b, the better able to distinguish between positive and negative examples. Finally, we model an extreme scenario where the support of X and Y is not the same. This offset uniform scenario is given by $X \sim U(0,1)$ and $Y \sim U(\gamma, 1+\gamma)$ for $\gamma \geq 0$: that is X lies uniformly on $(0,1)$ while Y is bounded on $(\gamma, \gamma+1)$. If $\gamma = 0$ there is no ability to discriminate, while $\gamma > 1$ leads to perfect classification of positive and negative examples with a threshold of $c = 1$. All results in this paper use $\mu = 1, a = 2, b = 5$, and $\gamma = 0.5$. These were chosen as representative examples of the distributions that produce reasonable PR curves.

This paper exclusively uses simulated data drawn from specific, known distributions because this allows calculation of the true PR curve (shown in Fig. 3) and the true AUCPR. Thus, we have a target value to compare the estimates against and are able to evaluate the bias of an estimator and the coverage of a CI. This would be difficult to impossible if we used a model's predictions on real data because the true PR curve and AUCPR are unknown.

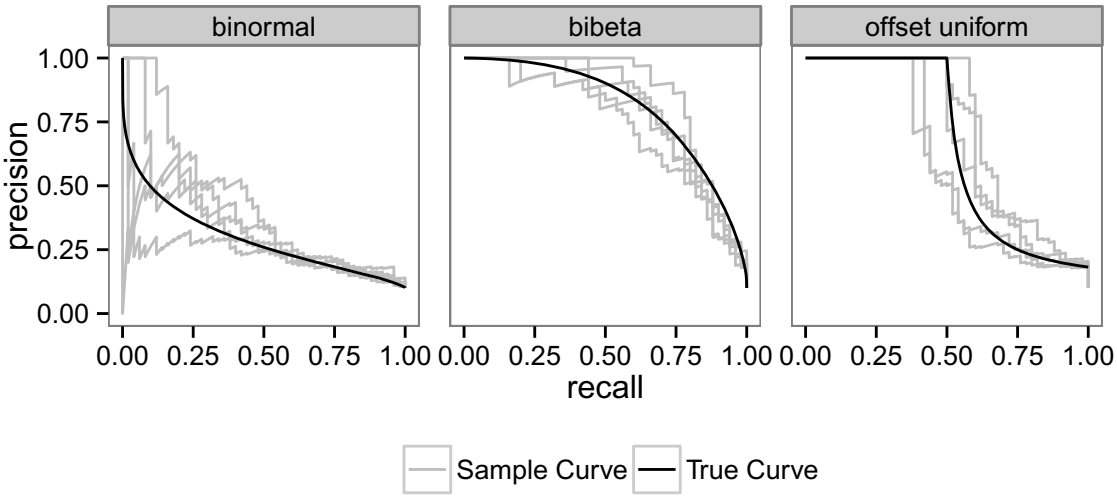

Fig. 3. True PR curves (calculated from the theoretical density functions) and sampled empirical PR curves, both at $\pi = 0.1$. Sampled PR curves use $n + m = 500$. The sampled PR curves were generated by connecting PR points corresponding to adjacent thresholds values.

4.1 Bias and Robustness in Point Estimates

For each scenario, we evaluate eight estimators: the non-parametric *average precision*, the parametric *binormal*, two trapezoidal estimates, and four interpolated estimates. Fig. 4 shows the bias ratio versus $n + m$ where $\pi = 0.1$ over 10,000 simulations, and Fig. 5 shows the bias ratio versus π where $n + m = 1000$. The bias ratio is the mean estimated AUCPR divided by the true AUCPR, so an unbiased estimator has a bias ratio of 1.0. Good point estimates of AUCPR should be unbiased as $n + m$ and π increase. That is, an estimator should have an expected value equal to the true AUCPR (calculated by numerically integrating Eq. 2).

As $n + m$ grows large, most estimators converge to the true AUCPR in every case. However, the *binormal* estimator shows the effect of model misspecification. When the data are truly binormal, it shows excellent performance but when the data are bibeta or offset uniform, the *binormal* estimator converges to the wrong value. Interestingly, the bias due to misspecification that we observe for the *binormal* estimate is lessened as the data become more balanced (π increases).

The *interpolated convex* estimate consistently overestimates AUCPR and appears far from the true value even at $n + m = 10000$. The poor performance of the *interpolated convex* estimator seems surprising given how it uses the popular convex hull ROC curve and then converts to PR space. Because the other interpolated estimators perform adequately, the problem may lie in evaluating the convex hull in ROC space. The convex hull chooses those particular points that give the best performance on the *test* set. Analogous to using the test set during training, the convex hull procedure may be overly optimistic and lead to the observed overestimation of AUCPR.

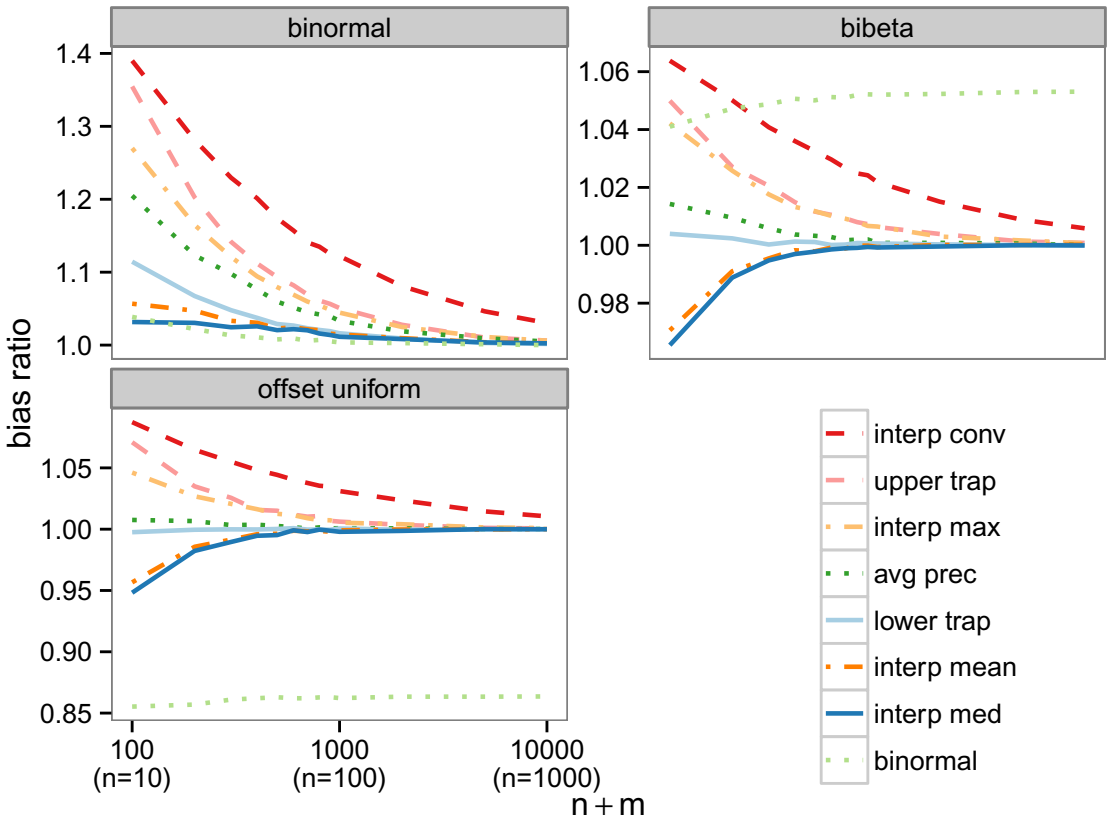

Fig. 4. Ratio of estimated AUCPR to true AUCPR (bias ratio) versus total number of examples ($n+m$). $\pi = 0.1$ for all cases.

It is important to note that since $\pi = 0.1$ in Fig. 4, data are sparse at $n+m = 100$: there are $n = 10$ values of Y to evaluate the estimate. In these situations there is no clear winner across all three scenarios and estimators tend to overestimate AUCPR when n is small with a few exceptions. Among related estimators, *lower trapezoid* appears more accurate than the *upper trapezoid* method and the *mean* or *median interpolation* outperform the *convex* and *max interpolation*. Consequently, we will only consider the *average precision*, *interpolated median*, and *lower trapezoid* estimators since they are unbiased in the limit, less biased for small sample sizes, and robust to model misspecification.

4.2 Confidence Interval Evaluation

We use a two-step approach to evaluate confidence intervals (CIs) based on Chapter 7 of Shao [16]. In practice, interval estimates must come with a confidence guarantee: if we say an interval is an $(1-\alpha)\%$ CI, we should be assured that it covers the true value in at least $(1-\alpha)\%$ of datasets [16,17,18]. It may be surprising to non-statisticians that an interval with slightly low coverage is ruled inadmissible, but this would invalidate the guarantee. Additionally, targeting an exact $(1-\alpha)\%$ interval is often impractical for technical reasons, hence the *at least* $(1-\alpha)\%$. When an interval provides at least $(1-\alpha)\%$ coverage, it is considered a valid interval and this is the first criteria a potential interval must satisfy.

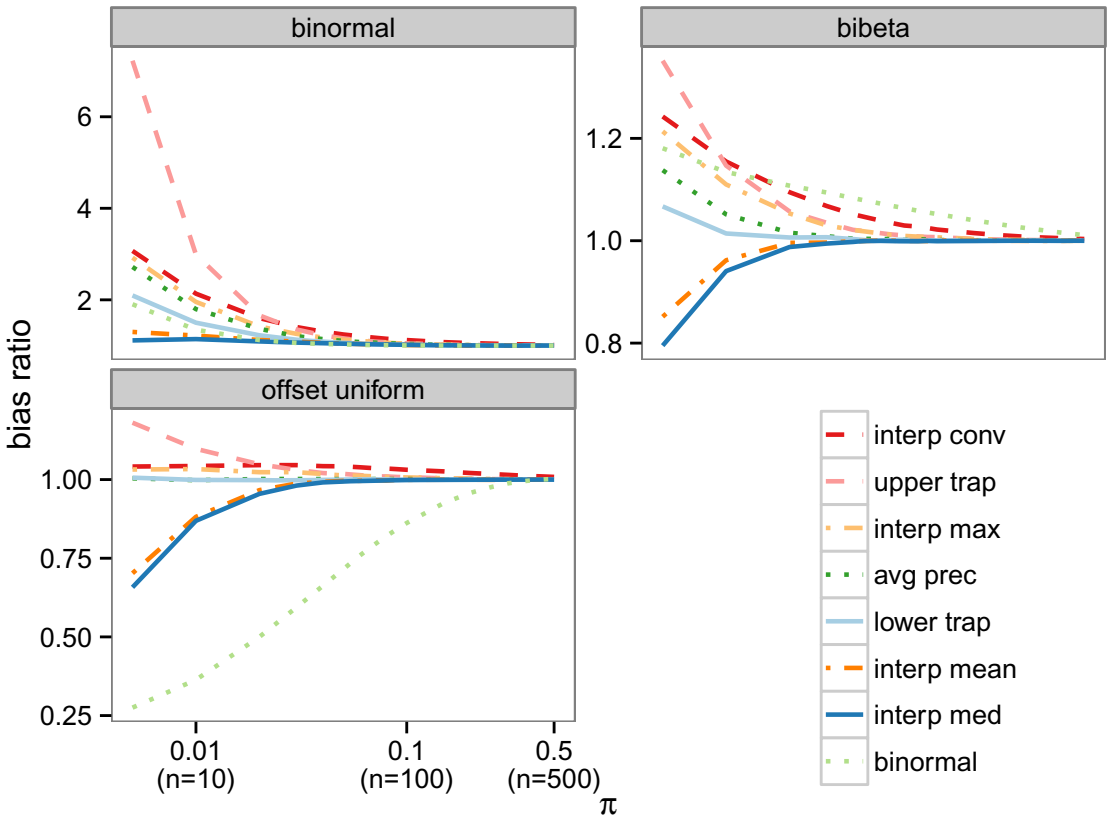

Fig. 5. Ratio of estimated AUCPR to true AUCPR (bias ratio) versus π. In all cases $n + m = 1000$.

After identifying valid methods for CIs, the second step is that we prefer the narrowest (or optimal) intervals among the valid methods. The trivial $[-\infty, +\infty]$ interval is a valid 95% CI because it always has at least 95% coverage (indeed, it has 100% coverage), but it conveys no useful information about the estimate. Thus we seek methods that produce the narrowest, valid intervals.

CI Coverage. The first step in CI evaluation is to identify valid CIs with coverage at least $(1-\alpha)$%. In Fig. 6, we show results over 10,000 simulations for the coverage of the four CI methods described in 3.2. These are 95% CIs, so the target coverage of 0.95 is denoted by the thick black line. As mentioned at the end of Section 4.1, we only consider the *average precision*, *interpolated median*, and *lower trapezoid* estimators for our CI evaluation.

A strong pattern emerges from Fig. 6 where the *bootstrap* and *cross-validation* intervals tend to have coverage below 0.95, though asymptotically approaching 0.95. Since the coverage is below 0.95, this makes the computational intervals technically invalid. The two formula-based intervals are consistently above the requisite 0.95 level. So *binomial* and *logit* produce valid confidence intervals.

Given the widespread use of *cross-validation* within machine learning, it is troubling that the CIs produced from that method fail to maintain the confidence guarantee. This is not an argument against *cross-validation* in general, only a

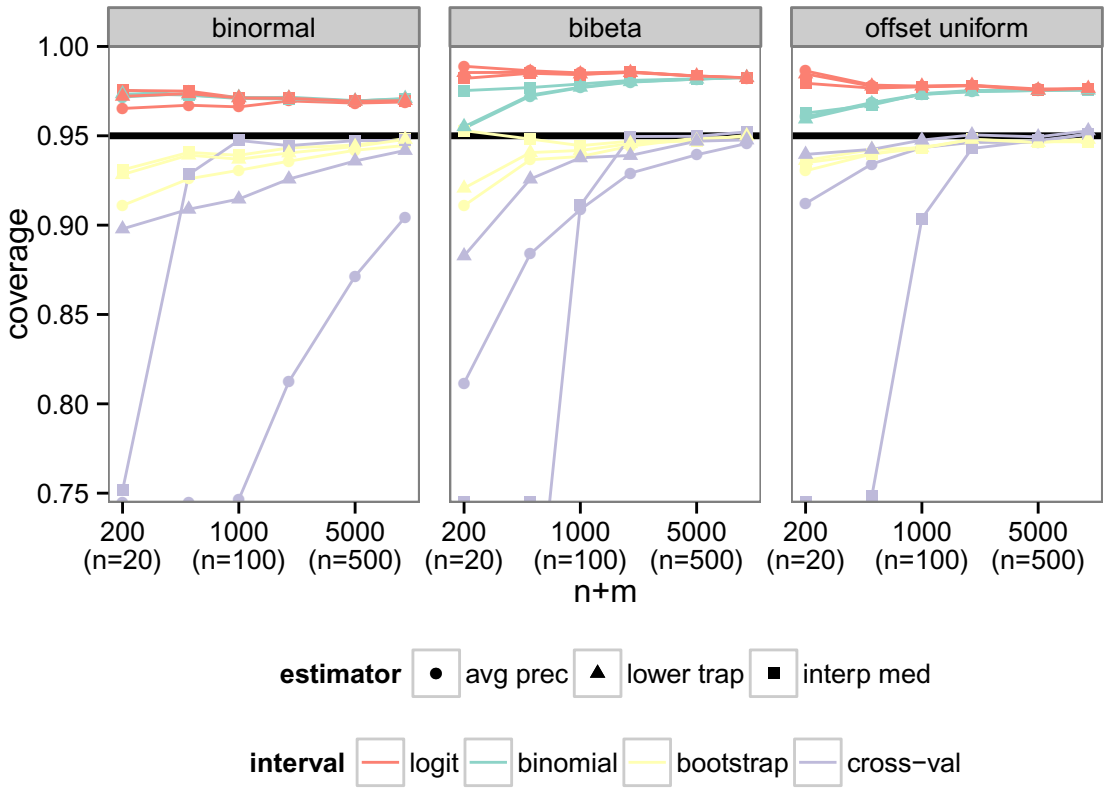

Fig. 6. Coverage for selected estimators and 95% CIs calculated using the four interval methods. Results for selected $n + m$ are shown for $\pi = 0.1$. To be valid 95% CIs, the coverage should be at least 0.95. Note that the coverage for a few of the *cross-validation* intervals is below 0.75. These points are represented as half-points along the bottom border.

caution against using it for AUCPR inference. Similarly, *bootstrap* is considered a rigorous (though computationally intensive) fall-back for non-parametrically evaluating variance, yet Fig. 6 shows it is only successful assymptotically as data size increases (and the data size needs to be fairly large before it nears 95% coverage).

CI Width. To better understand why *bootstrap* and *cross-validation* are failing, an initial question is: are the intervals *too* narrow? Since we have simulated 10,000 data sets and obtained AUCPR estimates on each using the various estimators, we have an empirical distribution from which we can calculate an ideal empirical width for the CIs. When creating a CI, only 1 data set is available, thus this empirical width is not available, but we can use it as a baseline to compare the mean width obtained by the various interval estimators. Fig. 7 shows coverage versus the ratio of mean width to empirically ideal width. As expected there is a positive correlation between coverage and the width of the intervals: wider intervals tend to provide higher coverage. For *cross-validation*, the widths tend to be slightly smaller than the *logit* and *binomial* intervals but still larger than the empirically ideal width. Coverage is frequently much lower though,

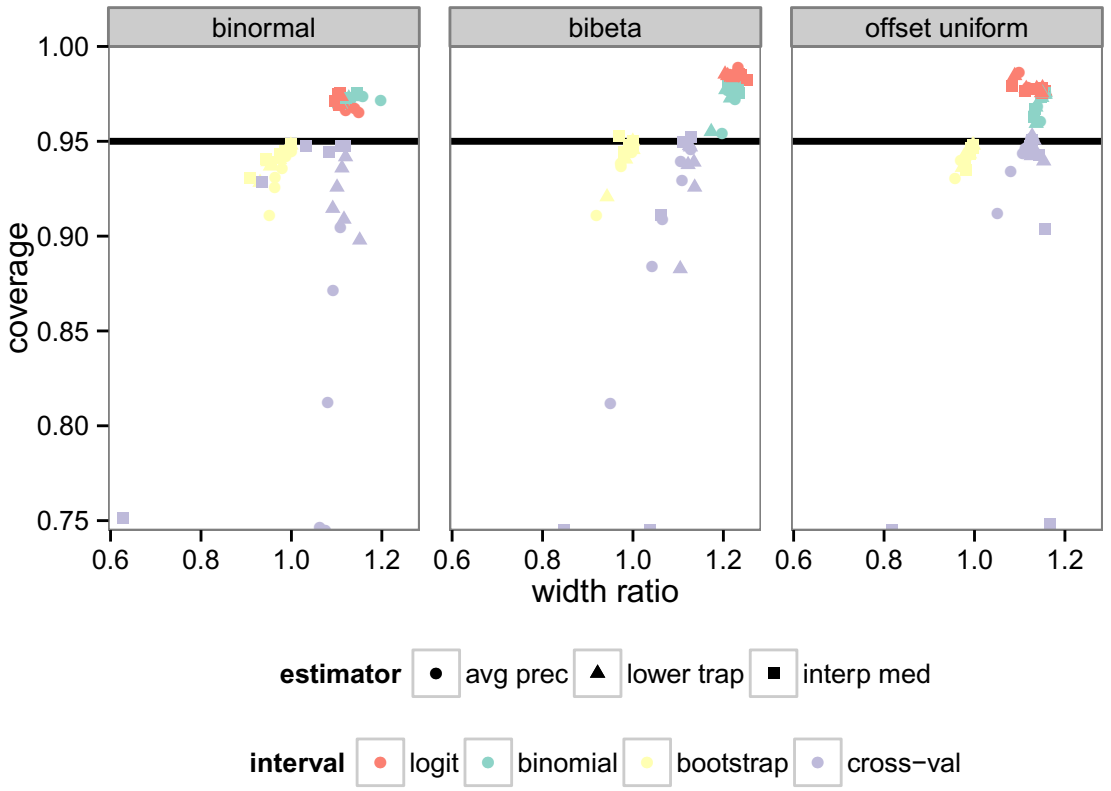

Fig. 7. Mean normalized width ratio versus coverage for *binomial*, *logit*, *cross-validation*, and *bootstrap* methods. Normalized width is the ratio of the CI width to the empirically ideal width. Width ratios below 1 suggest the intervals are overoptimistic. Results shown for $n + m \in 200, 500, 1000, 5000, 10000$ and $\pi = 0.1$. Note that the coverage for some of the *cross-validation* intervals is below 0.75. These points are represented as half-points along the bottom border.

suggesting the width of the interval is not the reason for the poor performance of *cross-validation*. However, interval width may be part of the issue with *bootstrap*. The *bootstrap* widths are either right at the empirically ideal width or even smaller.

CI Location. Another possible cause for poor coverage is that the intervals are for the wrong target value (i.e., the intervals are biased). To investigate this, we analyze the mean location of the intervals. We use the original estimate on the full data set as the location for the *binomial* and *logit* intervals since both are constructed around that estimate, the mid-point of the interval from *cross-validation*, and the median of the *bootstrap* replicates since we use the quantiles to calculate the interval. The ratio of the mean location to the true value (similar to Fig. 4) is presented in Fig. 8. The location of the *cross-validation* intervals is much farther from the true estimate than either the *bootstrap* or *binomial* locations, with *bootstrap* being a bit worse than *binomial*. This targeting of the wrong value for small $n + m$ is the primary explanation for the low coverages seen in Fig. 6.

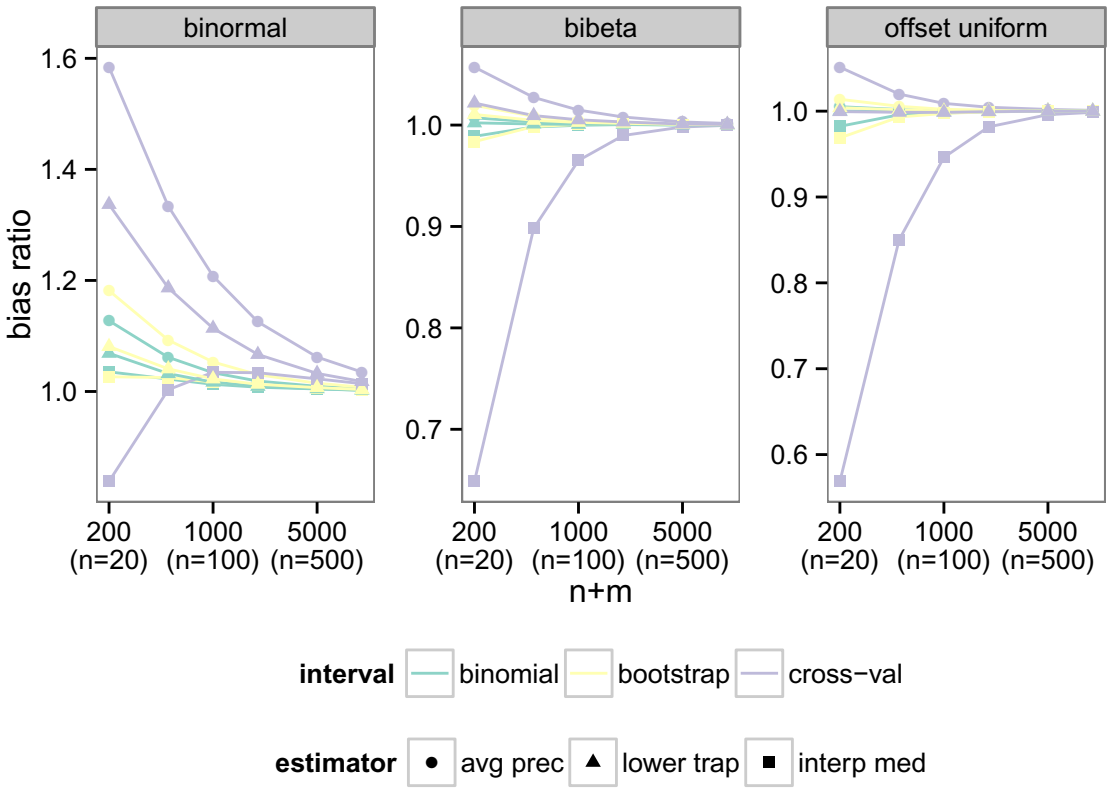

Fig. 8. Mean location of the intervals produced by the *binomial*, *bootstrap*, and *cross-validation* methods (*logit* is identical to *binomial*). As in Fig. 4, the y-axis is the bias ratio, the ratio of the location (essentially a point estimate based on the interval) to the true AUCPR. *Cross-validation* is considerably more biased than the other methods and *bootstrap* is slightly more biased than *binomial*.

Comments on *Bootstrap* and *Cross-validation* Intervals. The increased bias in the intervals produced by *bootstrap* and *cross-validation* occurs because these methods use many smaller data sets to produce a variance estimate. K-fold *cross-validation* reduces the effective data sets by a factor of k while *bootstrap* is less extreme but still reduces the effective data sets by a factor of 1.5. Since the estimators become more biased with smaller data sets (demonstrated in Fig. 4), the point estimates used to construct the *bootstrap* and *cross-validation* intervals are more biased, leading to the misplaced intervals and less than $(1 - \alpha)\%$ coverage.

Additionally, the *bootstrap* has no small sample theoretical justification and it is acknowledged it tends to break down for very small sample sizes [19]. When estimating AUCPR with skewed data, the critical number for this is the number of positive examples n, not the size of the data set $n + m$. Even when the data set itself seems reasonably large with $n + m = 200$, at $\pi = 0.1$ there are only $n = 20$ positive examples. With just 20 samples, it is difficult to get representative samples during the *bootstrap*. This also contributes to the lower than expected 95% coverage and is a possible explanation for the *bootstrap* widths being even smaller than the empirically ideal widths seen in Fig. 7.

We emphasize that both the *binomial* and *logit* intervals are valid and do not require the additional computation of *cross-validation* and *bootstrap*. For large sample sizes *bootstrap* approaches $(1-\alpha)\%$ coverage, but it approaches from below, so care should be taken. *Cross-validation* is even more problematic, with proper coverage not obtained even at $n+m = 10,000$ for some of our case studies.

5 Conclusion

Our computational study has determined that simple estimators can achieve nearly ideal width intervals while maintaining valid coverage for AUCPR estimation. A key point is that these simple estimates are easily evaluated and do not require resampling or add to computational workload. Conversely, computationally expensive, empirical procedures (*bootstrap* and *cross-validation*) yield interval estimates that do not provide adequate coverage for small sample sizes and only asymptotically approach $(1-\alpha)\%$ coverage.

We have also tested a variety of point estimates for AUCPR and determined that the parametric *binormal* estimate is extremely poor when the true generating distribution is not normal. Practically, data may be re-scaled (e.g., the Box-Cox transformation) to make this assumption fit better, but, with easily accessible nonparametric estimates that we have shown to be robust, this seems unnecessary.

The scenarios we studied are by no means exhaustive, but they are representative, and the conclusions can be further tested in specific cases if necessary. In summary, our investigation concludes that the *lower trapezoid*, *average precision*, and *interpolated median* point estimates are the most robust estimators and recommends the *binomial* and *logit* methods for constructing interval estimates.

Acknowledgments. We thank the anonymous reviewers for their detailed comments and suggestions. We gratefully acknowledge support from NIGMS grant R01GM097618, NLM grant R01LM011028, UW Carbone Cancer Center, ICTR NIH NCA TS grant UL1TR000427, CIBM Training Program grant 5T15LM007359, Roswell Park Cancer Institute, and NCI grant P30 CA016056.

References

1. Goadrich, M., Oliphant, L., Shavlik, J.: Gleaner: Creating ensembles of first-order clauses to improve recall-precision curves. Machine Learning 64, 231–262 (2006)
2. Richardson, M., Domingos, P.: Markov logic networks. Machine Learning 62(1-2), 107–136 (2006)
3. Liu, Y., Shriberg, E.: Comparing evaluation metrics for sentence boundary detection. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, vol. 4, pp. IV–185. IEEE (2007)

4. Yue, Y., Finley, T., Radlinski, F., Joachims, T.: A support vector method for optimizing average precision. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 271–278. ACM (2007)
5. Natarajan, S., Khot, T., Kersting, K., Gutmann, B., Shavlik, J.: Gradient-based boosting for statistical relational learning: The relational dependency network case. Machine Learning 86(1), 25–56 (2012)
6. Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine learning, ICML 2006, pp. 233–240. ACM, New York (2006)
7. Bamber, D.: The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology 12(4), 387–415 (1975)
8. Pepe, M.S.: The statistical evaluation of medical tests for classification and prediction. Oxford University Press, USA (2004)
9. Gordon, M., Kochen, M.: Recall-precision trade-off: A derivation. Journal of the American Society for Information Science 40(3), 145–151 (1989)
10. Abeel, T., Van de Peer, Y., Saeys, Y.: Toward a gold standard for promoter prediction evaluation. Bioinformatics 25(12), i313–i320 (2009)
11. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
12. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The binormal assumption on precision-recall curves. In: 2010 20th International Conference on Pattern Recognition (ICPR), pp. 4263–4266. IEEE (2010)
13. Efron, B.: Bootstrap methods: Another look at the jackknife. The Annals of Statistics 7(1), 1–26 (1979)
14. Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation 10, 1895–1923 (1998)
15. DeGroot, M.H., Schervish, M.J.: Probability and Statistics. Addison-Wesley (2001)
16. Shao, J.: Mathematical Statistics, 2nd edn. Springer (2003)
17. Wasserman, L.: All of statistics: A concise course in statistical inference. Springer (2004)
18. Lehmann, E.L., Casella, G.: Theory of point estimation, vol. 31. Springer (1998)
19. Efron, B.: Bootstrap confidence intervals: Good or bad? Psychological Bulletin 104(2), 293–296 (1988)