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

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

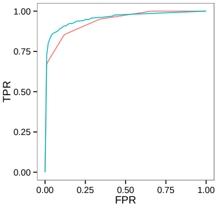
[1]University of Wisconsin-Madison, Madison, WI

[2]Roswell Park Cancer Institute, Buffalo, NY

September 26, 2013

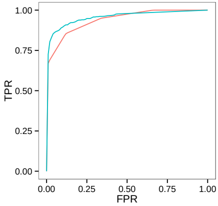
Binary Classification Evaluation

- Receiver-operating characteristic (ROC) curves
  - Preferred over accuracy alone [Provost et al., 1998]
  - Insensitive to skew (π = proportion of positives)
  - Area under ROC curve (AUCROC)

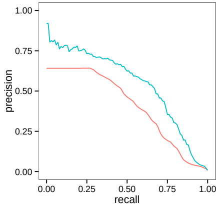


Binary Classification Evaluation

- Receiver-operating characteristic (ROC) curves
  - Preferred over accuracy alone [Provost et al., 1998]
  - Insensitive to skew (π = proportion of positives)
  - Area under ROC curve (AUCROC)

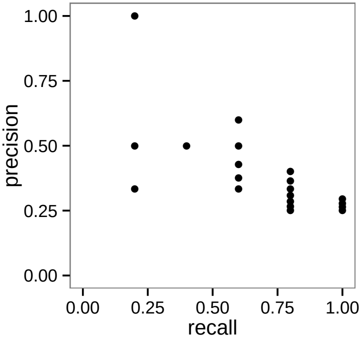


- Precision-recall (PR) curves
  - Alternative to ROC curves when π near 0 [Davis and Goadrich, 2006; Goadrich et al., 2006]
  - Sensitive to skew
  - Area under PR curve (AUCPR)



The Difficulty

Given



Do

- Estimate area: AUCPR $= 0.5$
- Obtain confidence interval: $[0.4, 0.6]$

Empirical PR Points

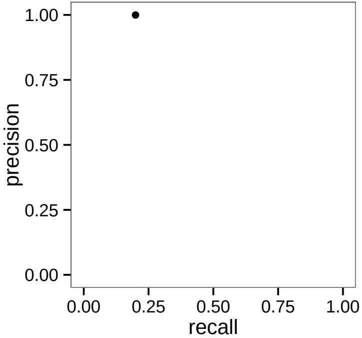
- 5 positives (n)
- 15 negatives (m)
- $\pi = 0.25$

score	true label
1.00	pos
0.95	neg
0.90	neg
0.85	pos
0.80	pos
0.75	neg
0.70	neg
0.65	neg
...	

Empirical PR Points

- 5 positives (n)
- 15 negatives (m)
- $\pi = 0.25$

score	true label
1.00	pos
0.95	neg
0.90	neg
0.85	pos
0.80	pos
0.75	neg
0.70	neg
0.65	neg
...	

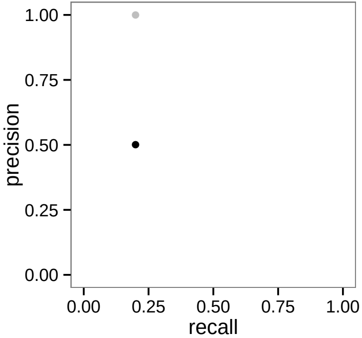


$$\text{Recall} = \frac{1}{5}$$
$$\text{Precision} = \frac{1}{1}$$

Empirical PR Points

- 5 positives (n)
- 15 negatives (m)
- $\pi = 0.25$

score	true label
1.00	pos
0.95	neg
0.90	neg
0.85	pos
0.80	pos
0.75	neg
0.70	neg
0.65	neg
...	

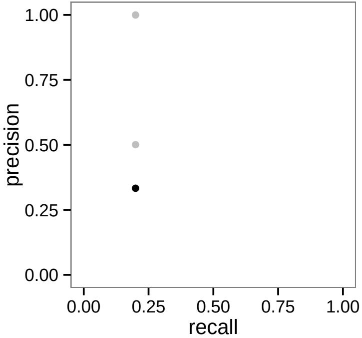


$$\text{Recall} = \frac{1}{5}$$
$$\text{Precision} = \frac{1}{2}$$

Empirical PR Points

- 5 positives (n)
- 15 negatives (m)
- $\pi = 0.25$

score	true label
1.00	pos
0.95	neg
0.90	neg
0.85	pos
0.80	pos
0.75	neg
0.70	neg
0.65	neg
...	

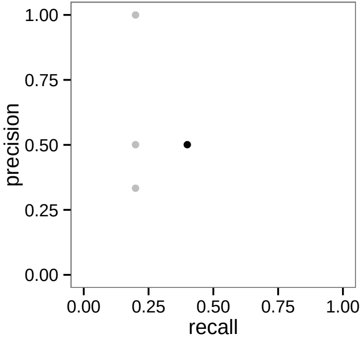


$$\text{Recall} = \frac{1}{5}$$
$$\text{Precision} = \frac{1}{3}$$

Empirical PR Points

- 5 positives (n)
- 15 negatives (m)
- $\pi = 0.25$

score	true label
1.00	pos
0.95	neg
0.90	neg
0.85	pos
0.80	pos
0.75	neg
0.70	neg
0.65	neg
...	

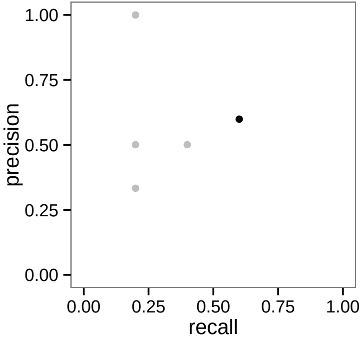


$$\text{Recall} = \frac{2}{5}$$
$$\text{Precision} = \frac{2}{4}$$

Empirical PR Points

- 5 positives (n)
- 15 negatives (m)
- $\pi = 0.25$

score	true label
1.00	pos
0.95	neg
0.90	neg
0.85	pos
0.80	pos
0.75	neg
0.70	neg
0.65	neg
...	

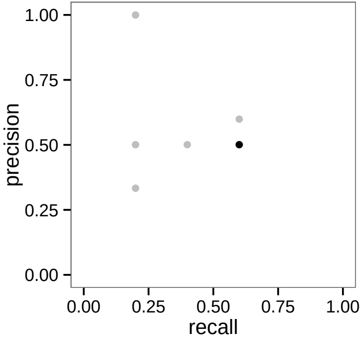


$$\text{Recall} = \frac{3}{5}$$
$$\text{Precision} = \frac{3}{5}$$

Empirical PR Points

- 5 positives (n)
- 15 negatives (m)
- $\pi = 0.25$

score	true label
1.00	pos
0.95	neg
0.90	neg
0.85	pos
0.80	pos
0.75	neg
0.70	neg
0.65	neg
...	



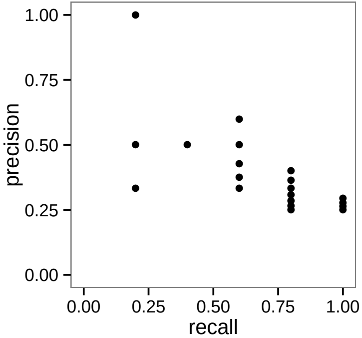
$$\text{Recall} = \frac{3}{5}$$
$$\text{Precision} = \frac{3}{6}$$

Empirical PR Points

- 5 positives (n)
- 15 negatives (m)
- $\pi = 0.25$

score	true label
1.00	pos
0.95	neg
0.90	neg
0.85	pos
0.80	pos
0.75	neg
0.70	neg
0.65	neg

...

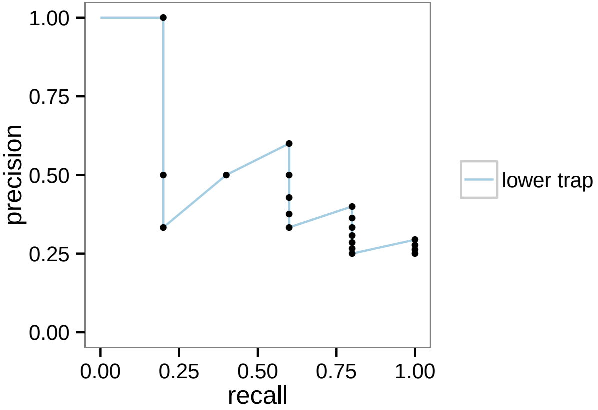


AUCPR Estimators

Many existing methods to estimate AUCPR

AUCPR Estimators

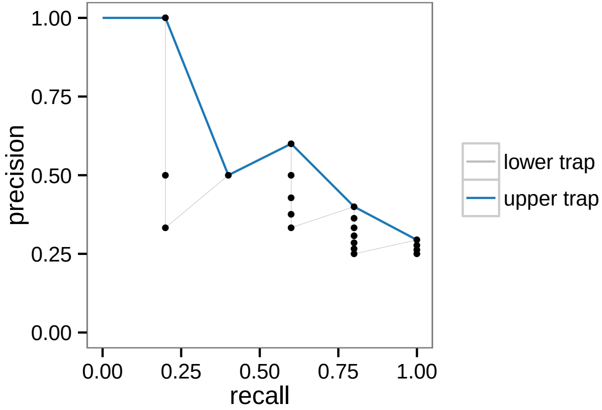
Many existing methods to estimate AUCPR



[Abeel et al., 2009]

AUCPR Estimators

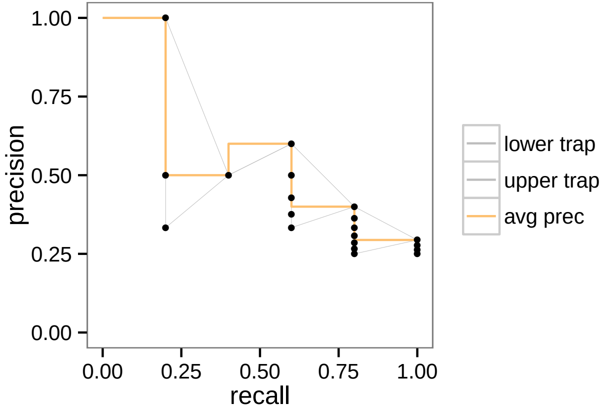
Many existing methods to estimate AUCPR



[Abeel et al., 2009; Davis and Goadrich, 2006]

AUCPR Estimators

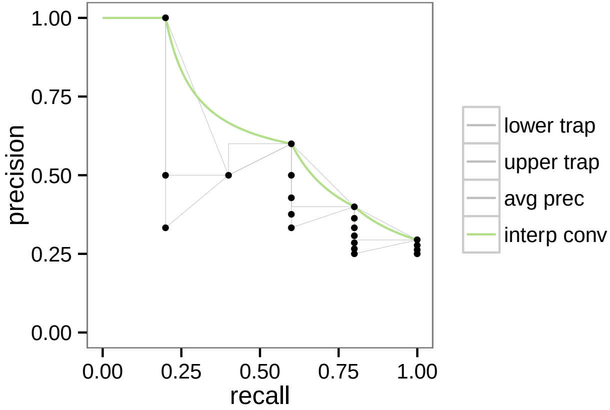
Many existing methods to estimate AUCPR



[Manning et al., 2008]

AUCPR Estimators

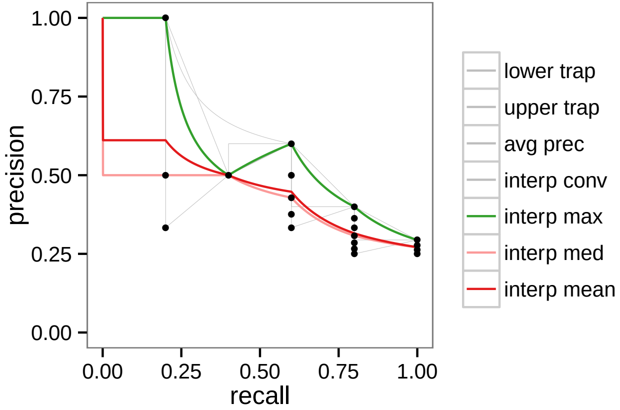
Many existing methods to estimate AUCPR



[Davis and Goadrich, 2006]

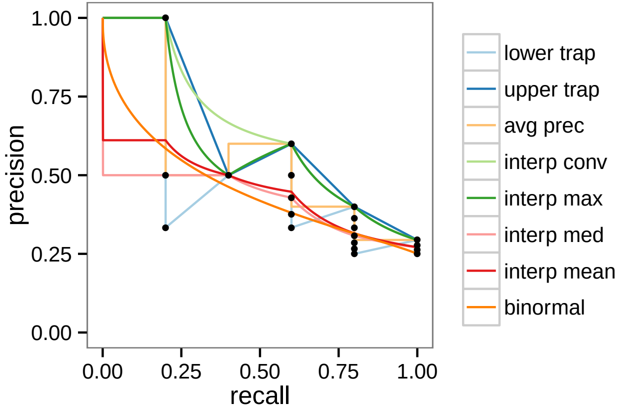
AUCPR Estimators

Many existing methods to estimate AUCPR



AUCPR Estimators

Many existing methods to estimate AUCPR



Estimator Evaluation Goals

Estimator Desiderata

- Unbiased: average estimate is equal to true AUCPR

Estimator Evaluation Goals

Estimator Desiderata

- Unbiased: average estimate is equal to true AUCPR
- Robust to different output distributions

Estimator Evaluation Goals

Estimator Desiderata

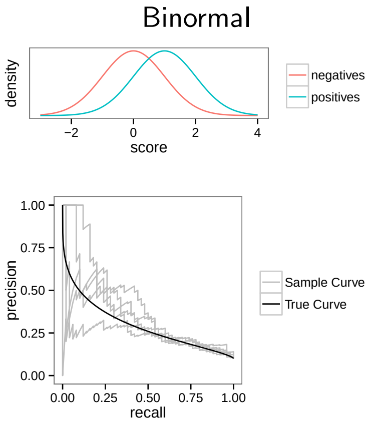
- Unbiased: average estimate is equal to true AUCPR
- Robust to different output distributions
- Robust to various skews (π) and data sizes ($n + m$)

Simulation Setup

- Assume example scores are drawn from known distributions
  - Similar to binormal analysis on ROC curves [Pepe, 2004; Bamber, 1975]
  - Allows calculation of true PR curve and AUCPR

Simulation Setup

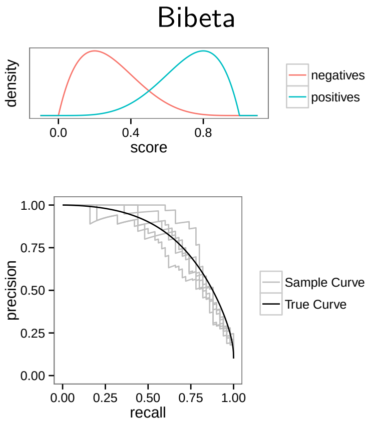
- Assume example scores are drawn from known distributions
  - Similar to binormal analysis on ROC curves [Pepe, 2004; Bamber, 1975]
  - Allows calculation of true PR curve and AUCPR
- Analyzed distributions
  - Binormal



$$\text{negatives} \sim N(0,1)$$
$$\text{positives} \sim N(1,1)$$

Simulation Setup

- Assume example scores are drawn from known distributions
  - Similar to binormal analysis on ROC curves [Pepe, 2004; Bamber, 1975]
  - Allows calculation of true PR curve and AUCPR
- Analyzed distributions
  - Binormal
  - Bibeta

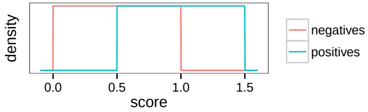


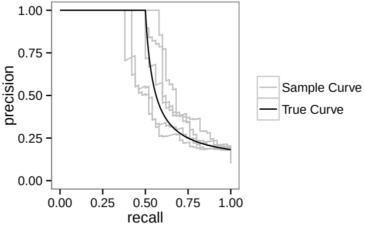
$$\text{negatives} \sim \mathit{Beta}(2,5)$$
$$\text{positives} \sim \mathit{Beta}(5,2)$$

Simulation Setup

- Assume example scores are drawn from known distributions
  - Similar to binormal analysis on ROC curves [Pepe, 2004; Bamber, 1975]
  - Allows calculation of true PR curve and AUCPR
- Analyzed distributions
  - Binormal
  - Bibeta
  - Offset uniform

Offset Uniform



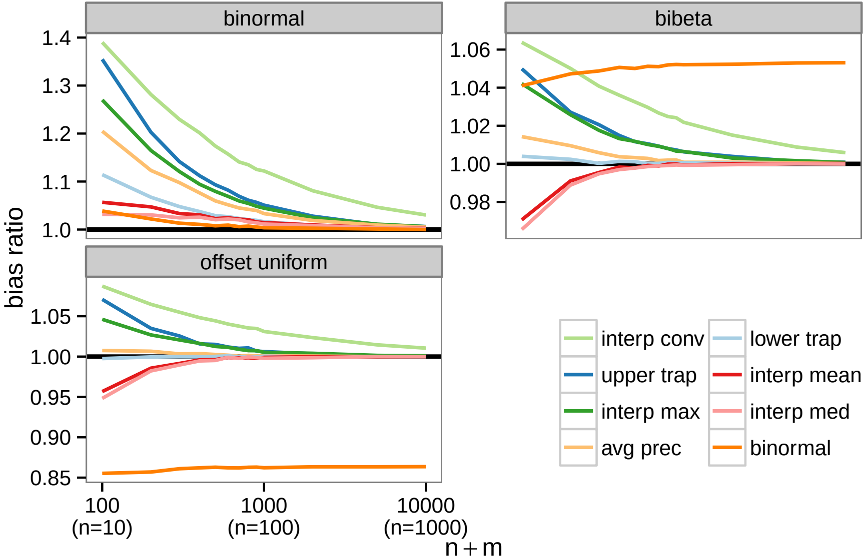


negatives $\sim U(0, 1)$
positives $\sim U(0.5, 1.5)$

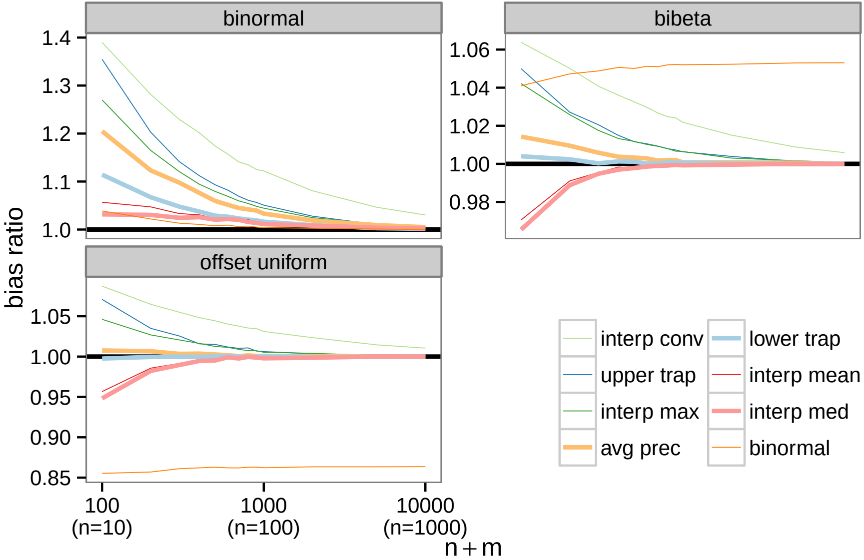
Simulation Setup

- Assume example scores are drawn from known distributions
  - Similar to binormal analysis on ROC curves [Pepe, 2004; Bamber, 1975]
  - Allows calculation of true PR curve and AUCPR
- Analyzed distributions
  - Binormal
  - Bibeta
  - Offset uniform
- Additional parameters
  - # of examples ($n + m$)
  - skew ($\pi = 0.1$)

AUCPR Estimator Results



AUCPR Estimator Results



AUCPR Confidence Intervals

Definition

A $1 - \alpha$ confidence interval is an interval that contains the true value with probability at least $(1 - \alpha)$. [Wasserman, 2004]

AUCPR Confidence Intervals

Definition

A $1-\alpha$ confidence interval is an interval that contains the true value with probability at least $(1-\alpha)$. [Wasserman, 2004]

- Empirical
  - Cross-validation: compute interval using mean and variance of K estimates from K-fold cross-validation

AUCPR Confidence Intervals

Definition

A $1 - \alpha$ confidence interval is an interval that contains the true value with probability at least $(1 - \alpha)$. [Wasserman, 2004]

- Empirical
  - Cross-validation: compute interval using mean and variance of K estimates from K-fold cross-validation
  - Bootstrap: choose interval that contains $(1 - \alpha)\%$ of empirical distribution of AUCPR estimates

AUCPR Confidence Intervals

Definition

A $1-\alpha$ confidence interval is an interval that contains the true value with probability at least $(1-\alpha)$. [Wasserman, 2004]

- Empirical
  - Cross-validation: compute interval using mean and variance of K estimates from K-fold cross-validation
  - Bootstrap: choose interval that contains $(1-\alpha)$% of empirical distribution of AUCPR estimates
- Parametric
  - Binomial: $\hat{\theta} \pm \Phi_{1-\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$

    $\hat{\theta}$ is the estimated AUCPR

AUCPR Confidence Intervals

Definition

A $1-\alpha$ confidence interval is an interval that contains the true value with probability at least $(1-\alpha)$. [Wasserman, 2004]

- Empirical
  - Cross-validation: compute interval using mean and variance of K estimates from K-fold cross-validation
  - Bootstrap: choose interval that contains $(1-\alpha)\%$ of empirical distribution of AUCPR estimates
- Parametric
  - Binomial: $\hat{\theta} \pm \Phi_{1-\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$

    $\hat{\theta}$ is the estimated AUCPR
  - Logit: $\left[\frac{e^{\hat{\eta}-\Phi(1-\alpha/2)\hat{\tau}}}{1+e^{\hat{\eta}-\Phi(1-\alpha/2)\hat{\tau}}}, \frac{e^{\hat{\eta}+\Phi(1-\alpha/2)\hat{\tau}}}{1+e^{\hat{\eta}+\Phi(1-\alpha/2)\hat{\tau}}}\right]$

    $\hat{\eta} = \log\frac{\hat{\theta}}{1-\hat{\theta}}, \hat{\tau} = (n\hat{\theta}(1-\hat{\theta}))^{-1/2}$

Confidence Interval Evaluation Goals

Confidence Interval Desiderata

- Valid - at least $(1 - \alpha)\%$ coverage

Confidence Interval Evaluation Goals

Confidence Interval Desiderata

- Valid - at least $(1-\alpha)\%$ coverage
- Prefer narrower (but still valid)

Confidence Interval Evaluation Goals

Confidence Interval Desiderata

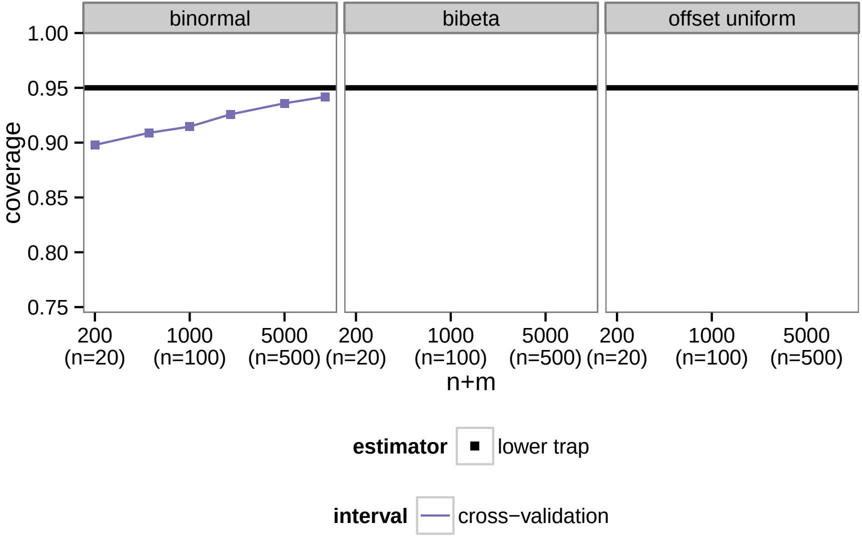
- Valid - at least $(1-\alpha)\%$ coverage
- Prefer narrower (but still valid)
- Robust to different output distributions

Confidence Interval Evaluation Goals

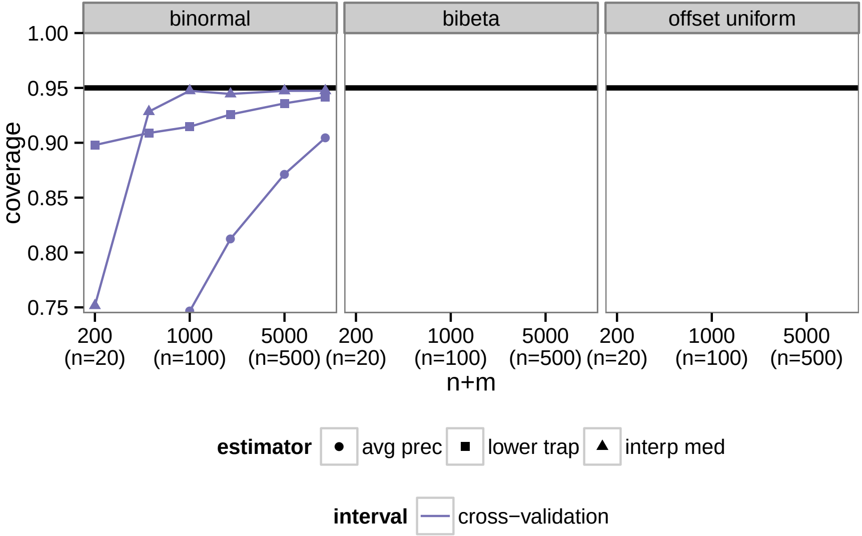
Confidence Interval Desiderata

- Valid - at least $(1-\alpha)\%$ coverage
- Prefer narrower (but still valid)
- Robust to different output distributions
- Robust to various skews (π) and data sizes ($n+m$)

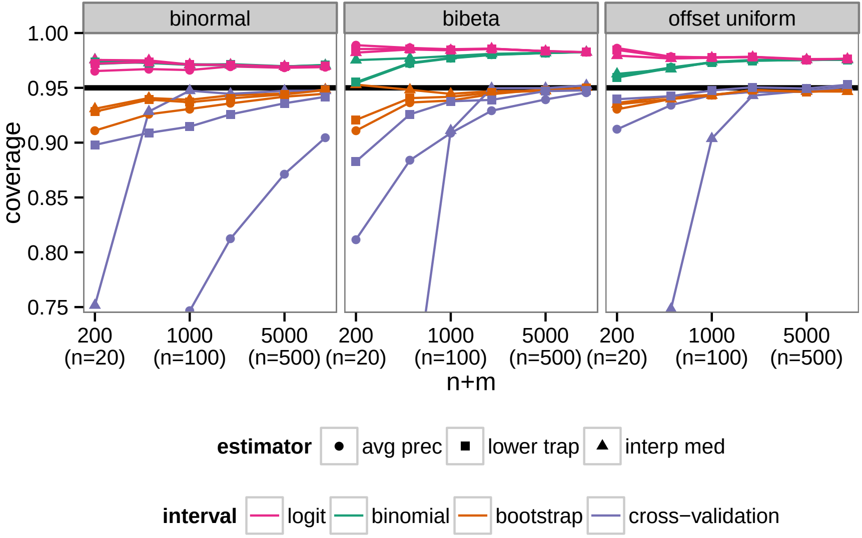
AUCPR Confidence Interval Results



AUCPR Confidence Interval Results



AUCPR Confidence Interval Results



Conclusion

- Choice of AUCPR estimator and confidence interval is important
  - Particularly for small data sets

Conclusion

- Choice of AUCPR estimator and confidence interval is important
  - Particularly for small data sets
- Recommended estimators
  - Lower trapezoid
  - Average precision
  - Interpolated median

Conclusion

- Choice of AUCPR estimator and confidence interval is important
  - Particularly for small data sets
- Recommended estimators
  - Lower trapezoid
  - Average precision
  - Interpolated median
- Recommended confidence intervals
  - Binomial
  - Logit

Conclusion

- Choice of AUCPR estimator and confidence interval is important
  - Particularly for small data sets
- Recommended estimators
  - Lower trapezoid
  - Average precision
  - Interpolated median
- Recommended confidence intervals
  - Binomial
  - Logit
  - What about cross-validation and bootstrap?
    - Converge to proper coverage, but from below
    - Problematic for small data sets and low numbers of positive examples

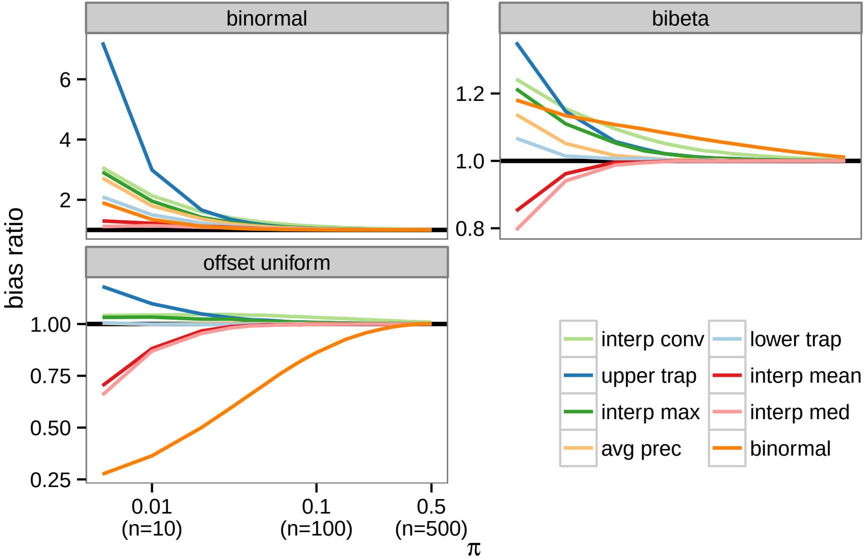
Thank You

Questions?

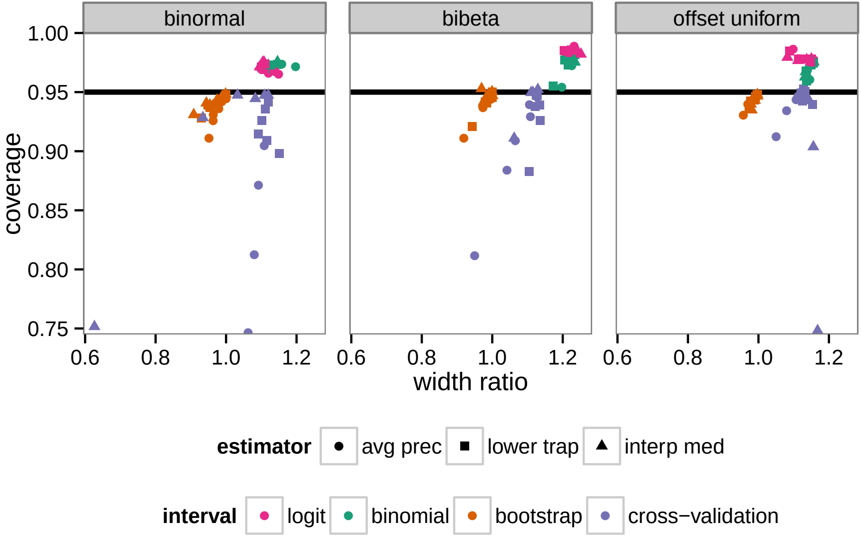
Acknowledgments

- 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
- NCI grant P30 CA016056

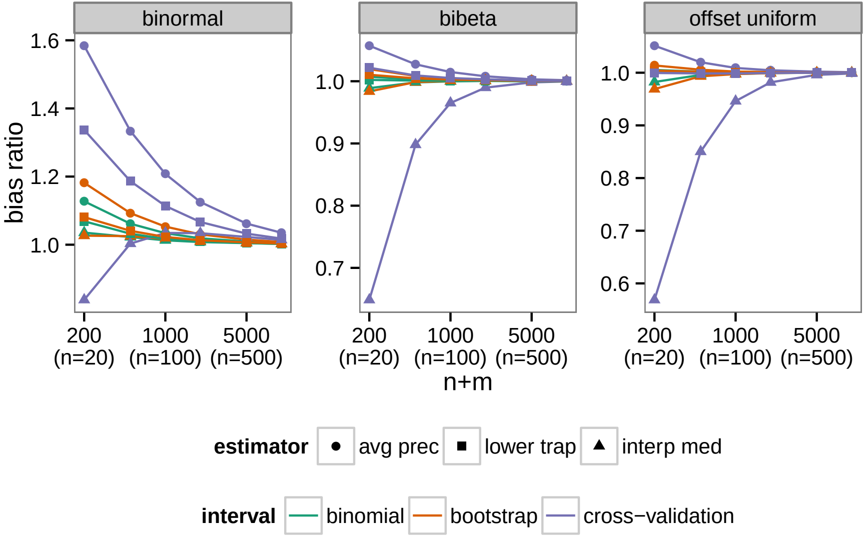
AUCPR Estimators Results by π



AUCPR Confidence Interval Widths



AUCPR Confidence Interval Locations



References I

Thomas Abeel, Yves Van de Peer, and Yvan Saeys. Toward a gold standard for promoter prediction evaluation. *Bioinformatics*, 25(12):i313–i320, 2009.

Donald Bamber. 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.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine learning*, ICML '06, pages 233–240, New York, NY, USA, 2006. ACM.

Mark Goadrich, Louis Oliphant, and Jude Shavlik. Gleaner: Creating ensembles of first-order clauses to improve recall-precision curves. *Machine Learning*, 64: 231–262, 2006.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

Margaret Sullivan Pepe. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA, 2004.

References II

Foster J Provost, Tom Fawcett, and Ron Kohavi. The case against accuracy estimation for comparing induction algorithms. In *ICML*, volume 98, pages 445–453, 1998.

Larry Wasserman. *All of statistics: A concise course in statistical inference*. Springer Verlag, 2004.