ROC Curves for Methods of Evaluating Evidence:

Common Performance Measures Based on Similarity Scores

R. Bradley Patterson, Department of Statistics, George Mason University
John Miller, Department of Statistics, George Mason University
Chris Saunders, Department of Applied IT, George Mason University

August 11, 2011

Trace Evidence Symposium
Kansas City, MO

Acknowledgement and Disclaimer

The research detailed in this presentation was supported in part by Award No. 2009-DN-BX-K234 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice.

Outline

- Introduction
- Background
- Analysis
- Results
- Conclusion

INTRODUCTION

Context

- Two samples
  - Control
  - Recovered
- Two hypotheses
  - Same source
  - Different sources

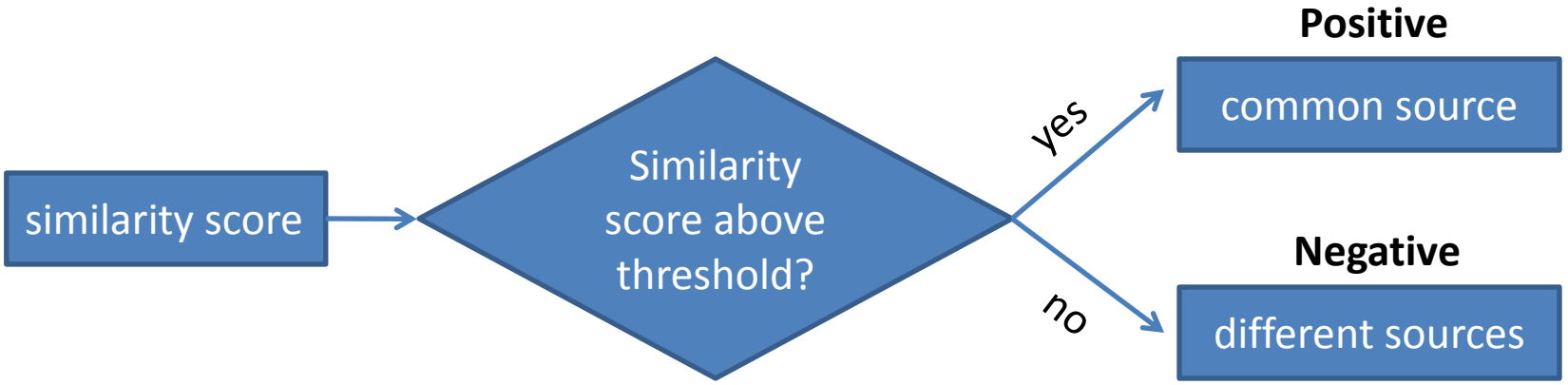
Evidence Evaluation

Similarity Score

- Numerical
- Indicative of association
- Higher values more suggestive of common source

Thresholds and Errors

- Threshold: fixed cutoff on similarity scores



- Method evaluation data: known sources
- Error types:
  - False positive
  - False negative

Outcomes for Fixed Threshold

		Truth	
		positive: pairs from same source	negative: pairs from different source
Evidence evaluation method's indication	positive	true positives	false positives
	negative	false negatives	true negatives

Error Rates

1. False positive rate = $$\frac{\text{number of false positives}}{\text{number of true negatives} + \text{number of false positives}}$$

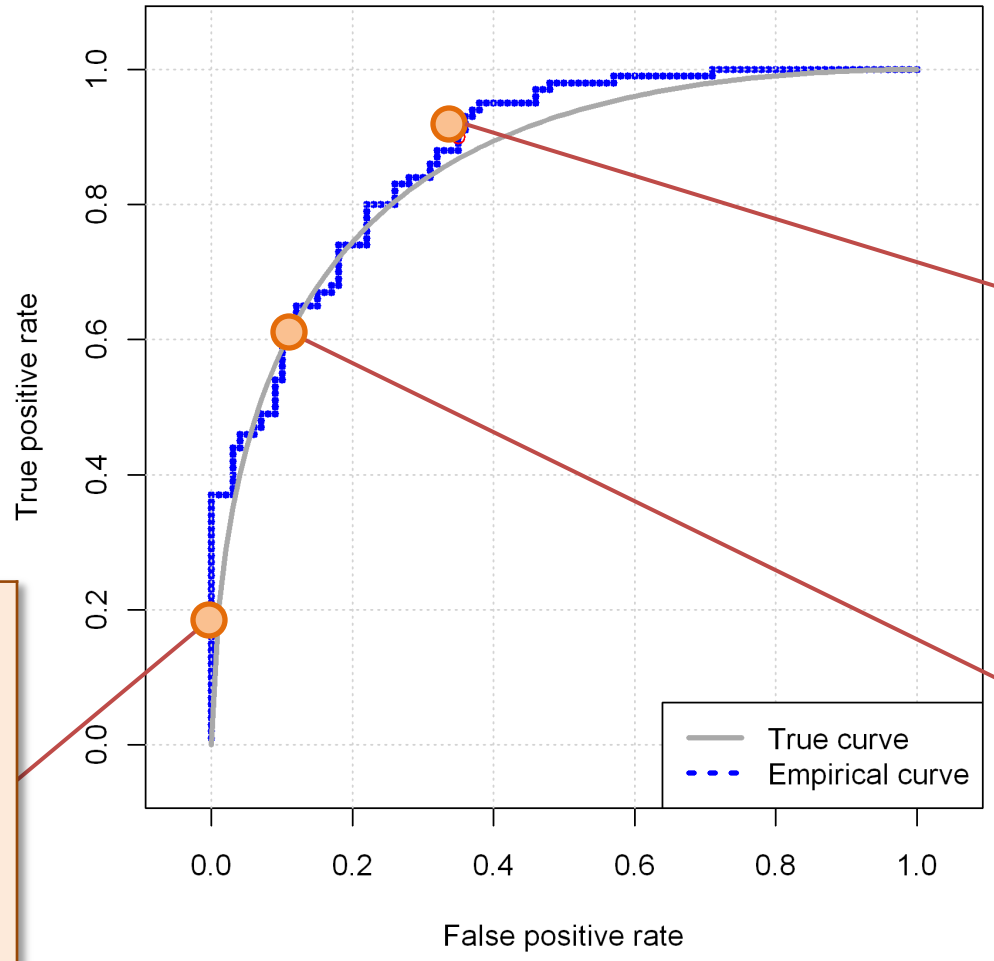
2. False negative rate = $$\frac{\text{number of false negatives}}{\text{number of true positives} + \text{number of false negatives}}$$

Or,

True positive rate = $$\frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

Outcomes for Varied Threshold

Receiver operating characteristic (ROC) plot



Threshold = 0.37

Indication		Truth p	Truth n
	p	90	35
	n	10	65

Threshold = 2.37

Indication		Truth p	Truth n
	p	20	0
	n	80	100

Threshold = 1.16

Indication		Truth p	Truth n
	p	60	10
	n	40	90

Application to Forensics

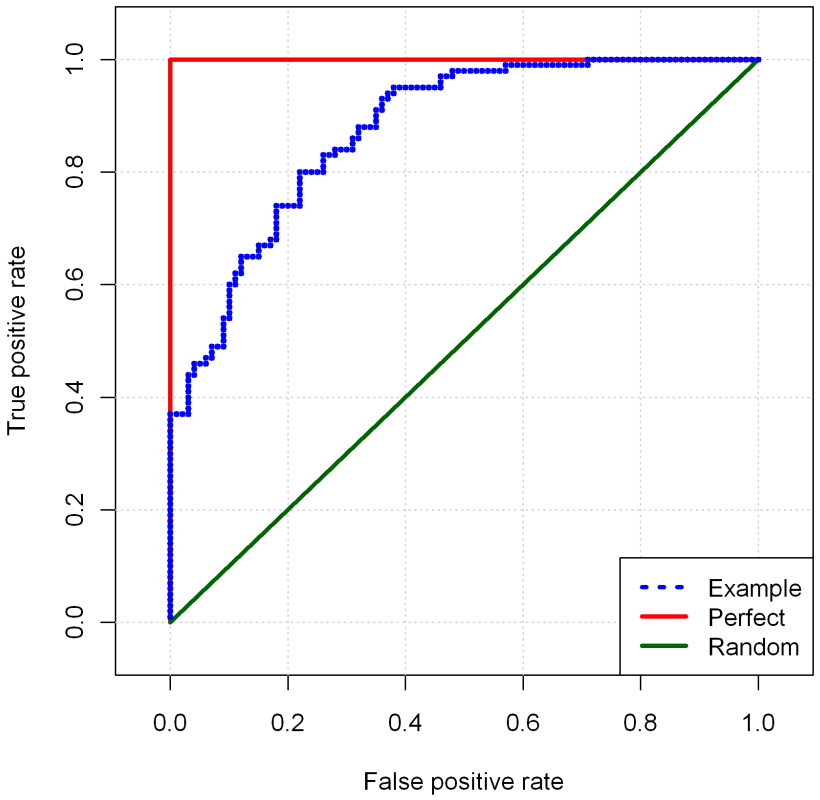
- Glass fragments
- Statistical methods of evaluating evidence

BACKGROUND

History and Uses of ROC Curves

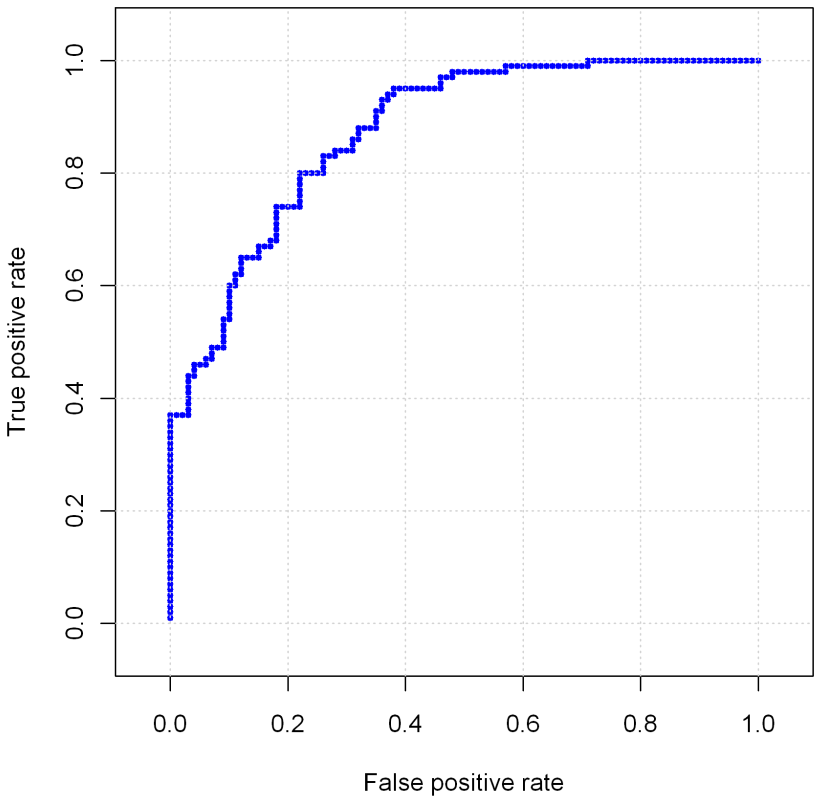
- 1940s: Radar hits and misses
- 1950s and 1960s: Signal detection theory (Green and Swets, 1966)
- 1980s: Diagnostic systems (Swets and Pickett, 1982)
- Today: Medicine, machine learning, astronomy and more

Overview of ROC Curves



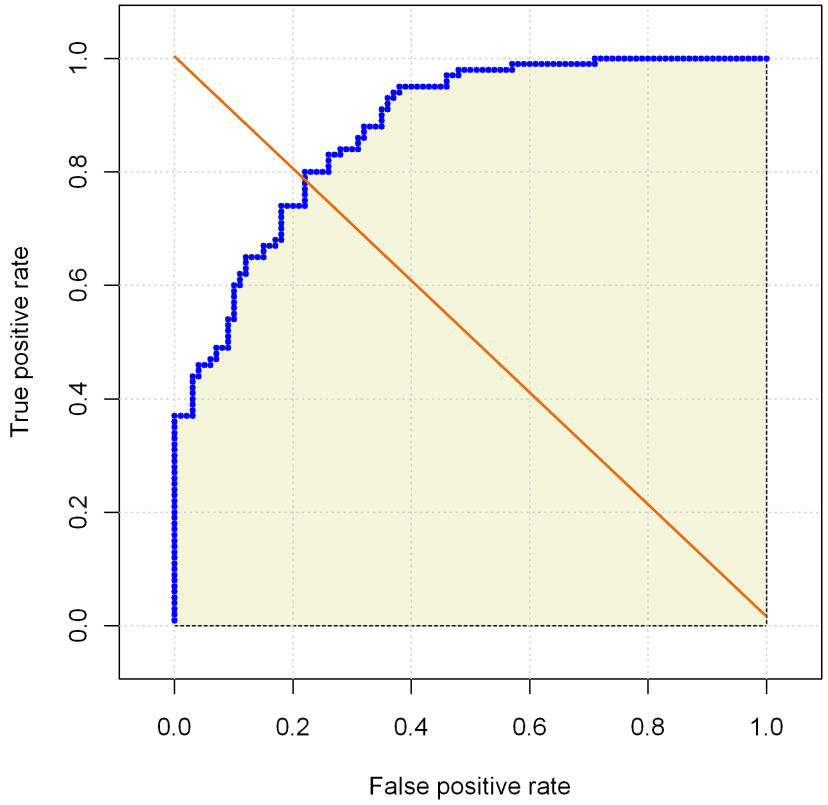
- Axes
  - True positive rate vs false positive rate
  - Range from 0 to 1
- Possible ROC curves
  - Perfect from (0,0) to (0,1) to (1,1).
  - Random along 45 degree line
  - In practice, usually somewhere between

Properties of ROC Curves



- Complete range of error rates
- Independent of scale of similarity scores
  - Order of similarity scores determines curve
  - Invariant under non-decreasing monotone transformation

ROC Performance Measures



- Error rates (upper left region)
- Equal error rate
- Equal likelihood
- Area under the curve (AUC)

$$= \Pr(\text{score}_{\text{same}} > \text{score}_{\text{diff}})$$

ANALYSIS

Glass data

- 62 windows
  - Three types
- Five fragments from each window
- Measurements of Si, K, Ca, and Fe
- Variables: log(Ca/K), log(Ca/Si), log(Ca/Fe)

Aitken, C. G. G., and Lucy, D. (January 2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53 (1), 109-122.

Data publicly available at
http://www.blackwellpublishing.com/rss/

Statistical Methods of Evaluating Evidence

- Methods (Aitken and Lucy, 2004):
  - Multiple *t*-statistics
  - Hotelling's T^2-statistic
  - Normal-based likelihood ratio
  - Density-based likelihood ratio
- Similarity scores:

  We can treat all methods as mappings from two samples to a similarity score.

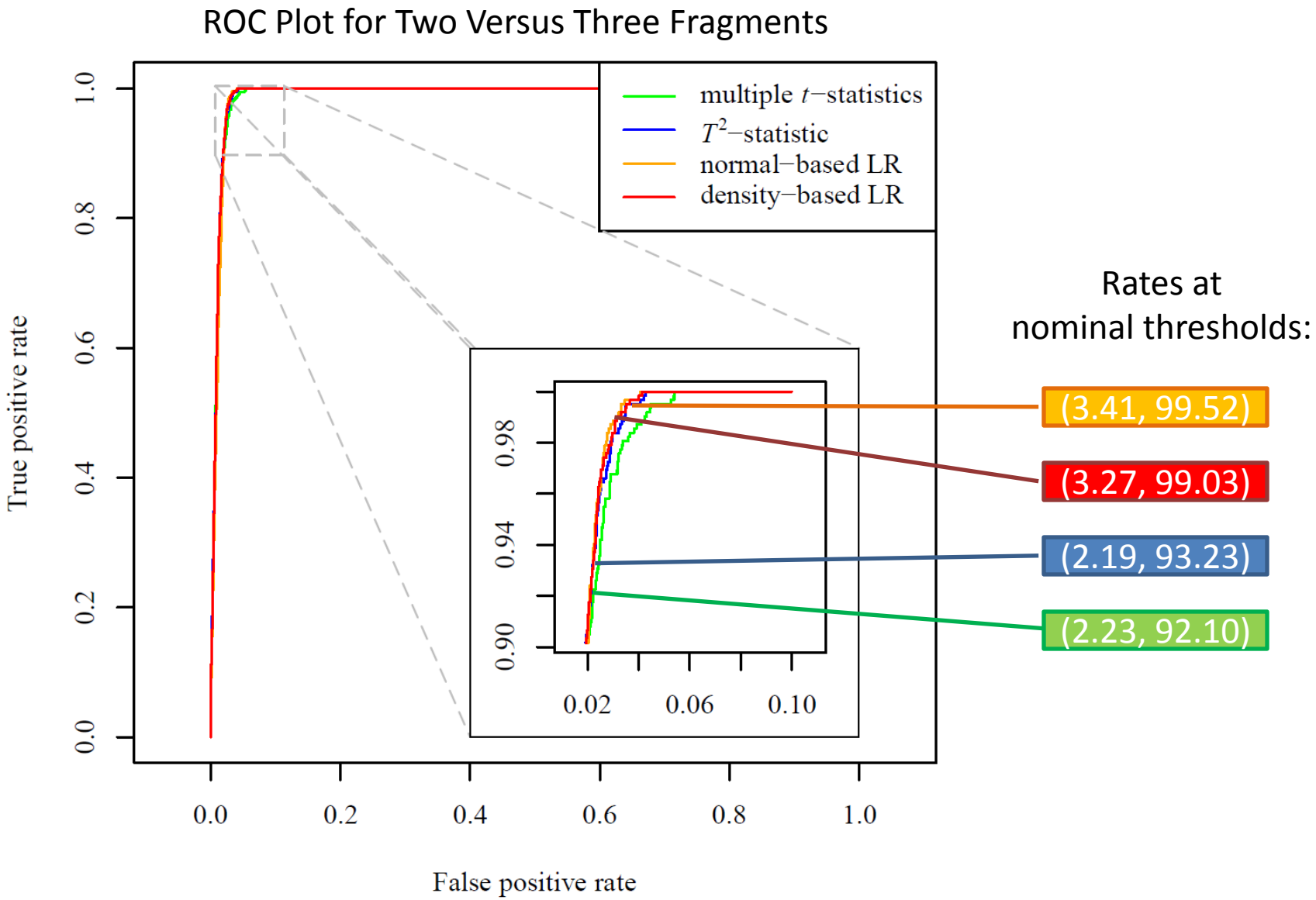
RESULTS

Nominal Error Rates

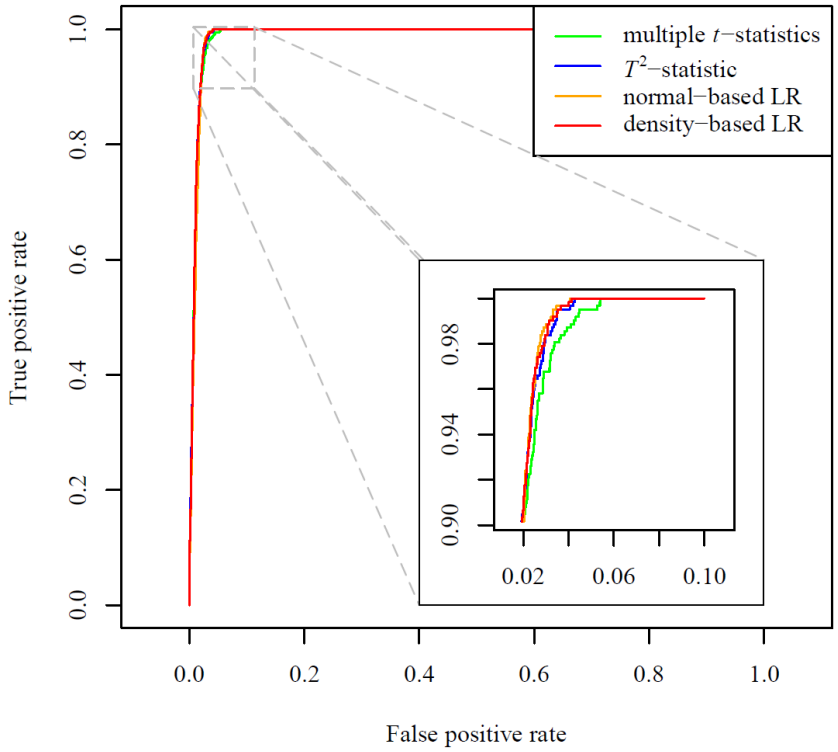
Error rates at nominal thresholds for two versus three fragments:

Error type	Multiple *t*-statistics	T^2-statistic	Normal-based LR	Density-based LR
false negative	7.90	6.77	0.48	0.97
false positive	2.23	2.19	3.41	3.27

ROC Plots


ROC Plot for Two Versus Three Fragments
multiple t−statistics
T^2−statistic
normal−based LR
density−based LR
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.90
0.94
0.98
0.02
0.06
0.10
Rates at nominal thresholds:
(3.41, 99.52)
(3.27, 99.03)
(2.19, 93.23)
(2.23, 92.10)

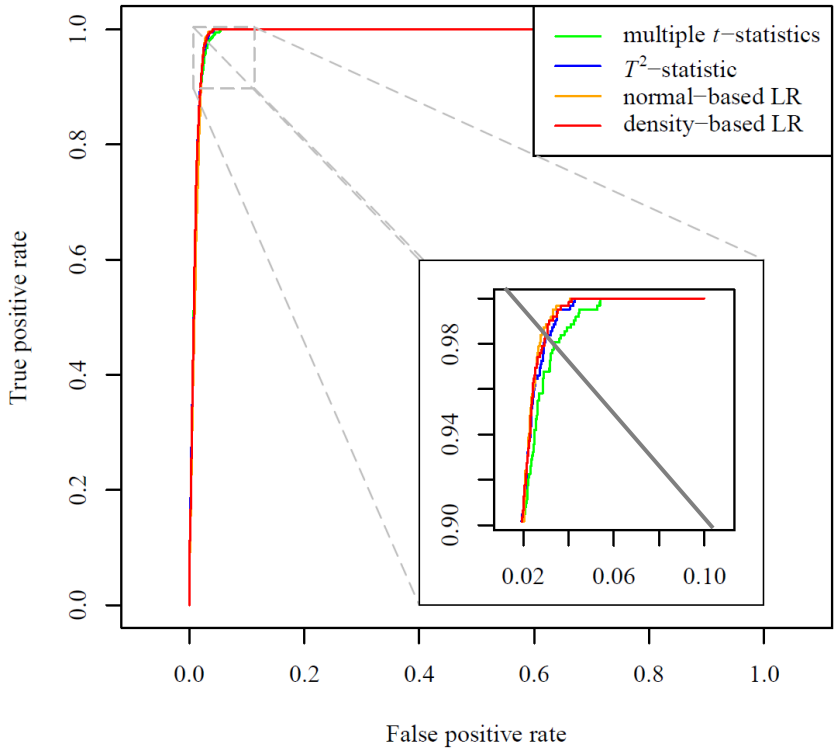
AUC Values



AUC values

Method	AUC
Multiple *t*-statistics	99.03
T^2-statistic	99.08
Normal-based LR	98.98
Density-based LR	99.09

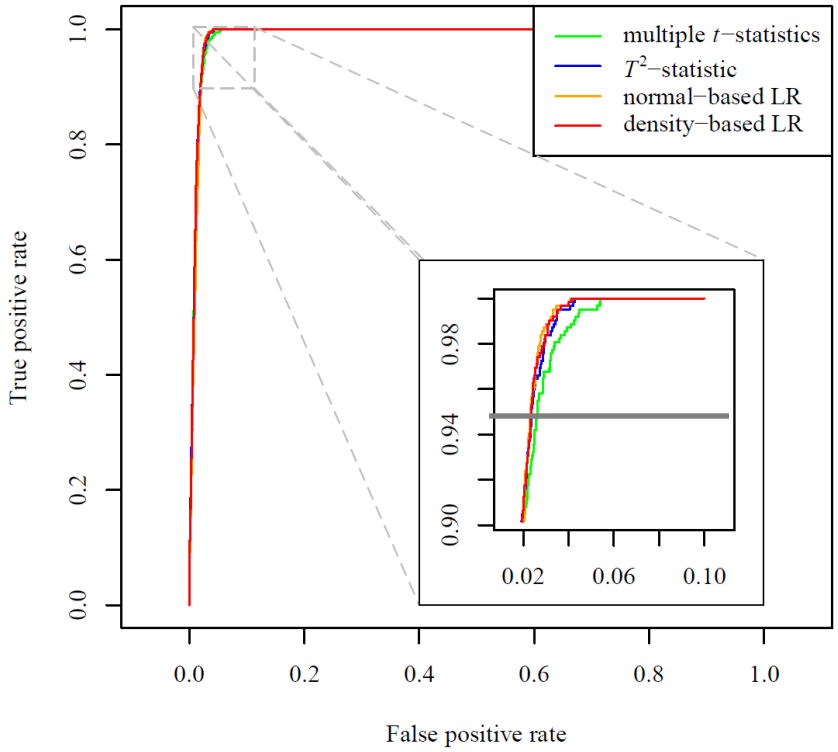
EER Values



EER values

Method	EER	Threshold
Multiple *t*-statistics	3.23	3.29
T^2-statistic	2.81	12.44
Normal-based LR	2.61	38.32
Density-based LR	2.62	17.98

5% False Negative Rate



False Positive Rates

Method	False Positive Rate	Threshold
Multiple *t*-statistics	2.61	2.72
T^2-statistic	2.36	8.96
Normal-based LR	2.30	117.33
Density-based LR	2.35	44.71

CONCLUSION

Conclusions

- ROC curves from similarity scores are
  - Comprehensive
    - Full range of error rates
  - Comparable
    - Independent of scale of scores
    - Objective performance measures

- Application to trace evidence and statistical methods (glass data) showed high performance from all methods.

References

- Aitken, C. G. G. and D. Lucy (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society Series C* 53 (1), 109–122.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters* 27 (8), 861–874.
- Green, D. M. and J. A. Swets (1966). *Signal detection theory and psychophysics*. New York: John Wiley and Sons, Inc.
- Krzanowski, W. J. and D. J. Hand (2009). *ROC curves for continuous data*. Boca Raton: Chapman & Hall.
- Pepe, M. S. (2004). *The statistical evaluation of medical tests for classification and prediction*. Oxford: Oxford University Press.
- Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer (2005). ROCR: Visualizing classifier performance in R. *Bioinformatics* 21 (20), 3940–3941.
- Swets, J. A., R. M. Dawes, and J. Monahan (2000). Better decisions through science. *Scientific American* 283 (4), 82–87.
- Swets, J. A. and R. M. Pickett (1982). *Evaluation of diagnostic systems: Methods from signal detection theory.* New York: Academic Press.
- Zhou, X., D. K. McClish, and N. A. Obuchowski (2002). *Statistical methods in diagnostic medicine* (1 ed.). New York: Wiley-Interscience.
- Zou, K. H., A. J. O'Malley, and L. Mauri (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation* 115 (5), 654–657.