UNIVERSITY OF CALIFORNIA,
IRVINE

Statistical Methods for the Forensic Analysis of User-Event Data

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Christopher Galbraith

Dissertation Committee:
Chancellor's Professor Padhraic Smyth, Chair
Chancellor's Professor Hal S. Stern
Associate Professor Veronica Berrocal

2020

DEDICATION

To my parents, Lynn & Larry

TABLE OF CONTENTS

Page

LIST OF FIGURES **vi**

LIST OF TABLES **x**

LIST OF ALGORITHMS **xi**

ACKNOWLEDGMENTS **xii**

VITA **xiv**

ABSTRACT OF THE DISSERTATION **xvi**

1 Introduction **1**
1.1 Outline & Contributions 4

2 Computing Strength of Evidence with the Likelihood Ratio **8**
2.1 Evidence Types 8
2.1.1 Biological (DNA) Evidence 9
2.1.2 Trace Evidence 10
2.1.3 Pattern Evidence 11
2.2 Formal Problem Statement 12
2.2.1 Source Propositions 12
2.3 The Likelihood Ratio 15
2.3.1 The LR as a Bayesian Method 17
2.3.2 Estimation 18
2.3.3 Interpretation 21
2.4 Population Data 22
2.4.1 Reference Data 24
2.4.2 Validation Data 24
2.4.3 Leave-pairs-out Cross-validation 25
2.5 Empirical Classification Performance 25
2.6 Information-theoretic Evaluation 27
2.6.1 Uncertainty and Information 28
2.6.2 Choosing the Target Distribution 32
2.6.3 Strictly Proper Scoring Rules & Bayes Risk 33

2.6.4 Empirical Cross-Entropy (ECE) . . . 37
2.6.5 Log Likelihood Ratio Cost . . . 38
2.6.6 The ECE Plot . . . 39
2.7 Empirical Calibration . . . 41
2.7.1 Calibration in General . . . 42
2.7.2 Using LR Values to Calibrate Posterior Probabilities . . . 43
2.7.3 Isotonic Regression & the PAV Algorithm . . . 43
2.7.4 Obtaining Calibrated LR Values . . . 45
2.7.5 Assessing the Calibrated LR Values . . . 45
2.8 Discussion . . . 48
2.8.1 Contributions . . . 49

3 Score-Based Approaches for Computing the Strength of Evidence 50
3.1 The Score-based Likelihood Ratio . . . 51
3.1.1 Choosing an Appropriate Score Function . . . 53
3.1.2 Estimation . . . 54
3.2 The Coincidental Match Probability . . . 58
3.2.1 Estimation . . . 60
3.2.2 Interpretation . . . 62
3.3 Examples with Known Distributions . . . 63
3.3.1 Likelihood Ratio . . . 64
3.3.2 Score-based Likelihood Ratio . . . 65
3.3.3 Coincidental Match Probability . . . 66
3.3.4 Comparison . . . 67
3.4 Discussion . . . 71
3.4.1 Contributions . . . 72

4 Spatial Event Data 73
4.1 Motivating Example . . . 74
4.2 Related Work . . . 76
4.3 Forensic Question of Interest . . . 77
4.4 Computing the Likelihood Ratio . . . 80
4.4.1 Adaptive Bandwidth Kernel Density Estimators . . . 83
4.4.2 Choosing the Mixture Parameter . . . 84
4.5 Score Functions for Geolocation Data . . . 87
4.5.1 Nearest Neighbor Distances . . . 88
4.5.2 Earth Mover's Distance . . . 89
4.5.3 Geoparcel Data . . . 91
4.5.4 Weighting Events . . . 92
4.6 Score-based Techniques . . . 94
4.6.1 Score-based Likelihood Ratio . . . 94
4.6.2 Coincidental Match Probability . . . 95
4.7 Case Study—Twitter Data . . . 95
4.7.1 Event Data . . . 95
4.7.2 Geoparcel Data . . . 97

4.7.3 Results . . . 98
4.7.4 Error Analysis . . . 104
4.7.5 Discussion of Twitter Results . . . 107
4.8 Case Study—Gowalla Data . . . 108
4.8.1 Data . . . 108
4.8.2 Results . . . 111
4.8.3 Discussion of Gowalla Results . . . 114
4.9 Discussion . . . 115
4.9.1 Contributions . . . 117

5 Temporal Event Data **119**
5.1 Motivating Example . . . 120
5.2 Forensic Question of Interest . . . 121
5.3 Related Work . . . 124
5.4 Score Functions for Temporal Event Data . . . 126
5.4.1 Marked Point Process Indices . . . 127
5.4.2 Inter-Event Times . . . 129
5.5 Quantifying Strength of Evidence . . . 129
5.5.1 Population-based Approach . . . 130
5.5.2 Resampling Approach . . . 131
5.6 Case Study—Simulated Data . . . 135
5.6.1 Simulating Temporal Marked Point Processes . . . 135
5.6.2 Results . . . 137
5.6.3 Discussion of Simulation Results . . . 142
5.7 Case Study—Student Web Browsing Data . . . 143
5.7.1 Data . . . 143
5.7.2 Population-based Results . . . 144
5.7.3 Resampling Results . . . 148
5.7.4 Discussion of Student Web Browsing Results . . . 149
5.8 Case Study—LANL Authentication Data . . . 150
5.8.1 Data . . . 150
5.8.2 Results . . . 151
5.9 Discussion . . . 152
5.9.1 Contributions . . . 153

6 Discussion on Future Directions **154**

Bibliography **159**

Appendix A Spatial Results—Twitter Data **169**

Appendix B Spatial Results—Gowalla Data **176**

Appendix C Signal-to-Noise Ratio Calculation **179**

LIST OF FIGURES

Page

1.1 Example of temporal evidence from two individuals (i and j) taken from the case study of Chapter 5. A and B events generated by the same individual tend to cluster temporally, with less clustering in time for A and B events from different users. 4

2.1 Prior entropy $\mathcal{U}_P(H_s)$ (logarithm base 2) as a function of the prior probability $P(H_s)$. 29
2.2 Logarithmic scoring rule (base 2) as a loss function. 34
2.3 Empirical cross-entropy (ECE) plot for case study data from Chapter 4. C_{llr} values are the ECE evaluated at prior log odds of 0 and are given in the legend. Lower values are indicative of better performance on the validation data. . . 40
2.4 Empirical cross-entropy (ECE) plot of Figure 2.3 including the PAV calibrated LR values. C_{llr} values are the ECE evaluated at prior log odds of 0 and are given in the legend. Lower values are indicative of better performance on the validation data. 46

3.1 Hypothetical illustration of the conditional densities of the score function Δ under the hypotheses that the samples are from the same source (H_s, dashed line) and that the samples are from different sources (H_d, solid line). The score-based likelihood ratio SLR_Δ is the ratio of the conditional density functions g evaluated at δ. 52
3.2 Hypothetical illustration of the densities of the score function Δ under the hypotheses that the samples are from the same source (H_s, dashed line) and that the samples are from different sources (H_d, solid line). The coincidental match probability CMP_Δ is the shaded tail region of $g(\Delta(A,B)|H_d,I)$. . . . 61
3.3 Behavior of the various methods for evaluating evidence under known distributional forms. The value $B = b$ was fixed to the mean of the same-source distribution, $\mu_B = 0$, to eliminate one source of variability in the analysis. Columns represent a selection of values for the parameters μ_P, σ_ω, and σ_β. (Top row) Distribution of A under H_s and H_d; (middle row) behavior of the LR and SLR as a function of $A = a$; (bottom row) behavior of the CMP as a function of $A = a$. 68
3.4 Contour plots of the various evidence evaluation methods for the example with known distributions where $\mu_B = 0$, $\mu_P = -4$, $\sigma_\omega = 0.5$, and $\sigma_\beta = 1$ (corresponding to the third column of Figure 3.3). 70

4.1 Location data (taken from Section 4.7.1) in a 3.5 square mile region of Orange County, CA. Green boxes represent geofences with events in both sets. (a) Both the unknown and known source data were generated by the same individual; (b) the unknown and known source data were generated by different individuals. The unknown source data is the same in both panels. The geographic features of the map (i.e., street names and buildings) were removed to preserve the individuals' privacy. 75
4.2 Example of sets of locations for Twitter data from New York. The patterns correspond to geolocatons of tweets from the same account over two different months, with month 1 corresponding to A (red) and month 2 corresponding to B (black). 79
4.3 Example of the KDE models used to estimate the likelihood ratio for Twitter events in Orange County, CA, from the experimental results in Section 4.7. Overlaid on each panel are the set of points A from the motivating example in Section 4.1. (a) Population component used to estimate the denominator of the LR $f(B|H_d, I)$; (b) individual component built using the overlaid points; (c) mixture model with $\alpha = 0.8$ used to estimate the numerator of the LR $f(B|A, H_s, I)$. 81
4.4 Mixture weight α as a function of the number of events in the unknown source sample n_a. 87
4.5 Area around John Wayne Airport (SNA) in Orange County, California, highlighting the parcel corresponding to the airport and Twitter events in the region. Figure credit Lichman [2017]. 92
4.6 Adaptive bandwidth KDE for the population data $\mathcal{D}$ of Twitter visits. (a) Orange County, CA; (b) New York, NY. 96
4.7 Density estimate of the number of parcels versus (a) the number of visits in the parcel, and (b) the number of unique accounts in that parcel. Note that both figures are right-truncated due to the extremely long tails. 98
4.8 False positive rate of each method under different data regimes in (a) Orange County, and (b) New York. Low corresponds to 1 event in each of A and B, medium is between 2 and 19 events, and high is 20 or more events. Showing results for fixed α in the LR approach and the account weighted EMD for the SLR and CMP approaches. Thresholds are 1 for the LR and SLR and 0.05 for the CMP. Trends are similar for other choices of mixing weight and score function (see Figures A.2 and A.6 in Appendix A). 102
4.9 Empirical cross-entropy (ECE) plots for a selection of the LR and SLR methods applied on the Twitter data. C_{llr} values are provided in the legend. (a) Orange County likelihood ratio with the nonparametric mixing weight $\alpha(n_a)$; (b) Orange County score-based likelihood ratio using the earth mover's distance and account weighting scheme; (c) New York LR with $\alpha(n_a)$; (d) New York SLR using the EMD and account weights. 103

4.10 Empirical cross-entropy (ECE) plot of the likelihood ratio approach with the nonparametric weighting function $\alpha(n_a)$ for the Orange County Twitter data. (a) Standard ECE plot; (b) ECE plot contribution for each piece of same-source evidence. The black curve is from the evidence shown in Figure 4.11a; (c) ECE plot contribution for each piece of different-source evidence. The black curve is from the evidence shown in Figure 4.11b. Note the different scales on the y-axes. 104
4.11 Examples of misclassified evidence in the Orange County Twitter data corresponding to the highlighted individual ECE contributions in Figure 4.10. (a) A same-source pair with $\log(LR) \approx -33$. Due to overplotting, the point size of the locations in A was increased. (b) A different-source pair with $\log(LR) \approx 77$.105
4.12 Adaptive bandwidth KDE for the population data $\mathcal{D}$ of Gowalla check-in events in Southern California. 109
4.13 Density estimate of the number of locations versus the number of visits at the location (solid line), and the number of unique accounts in that location (dashed line). Note that both figures are right-truncated due to the extremely long tails. 110
4.14 False positive rate of each method under different data regimes for the Gowalla data. Low corresponds to less than 5 in each of A and B, medium is between 5 and 14 events, and high is 15 or more events. Showing results for fixed α in the LR approach and the account weighted EMD for the SLR and CMP approaches. Thresholds are 1 for the LR and SLR and 0.05 for the CMP. Trends are similar for other choices of mixing weight and score function (see Figure B.2 in Appendix B). 113
4.15 Empirical cross-entropy (ECE) plots for a selection of the LR and SLR methods applied on the Gowalla data. C_{llr} values are provided in the legend. (a) Likelihood ratio with the nonparametric mixing weight $\alpha(n_a)$; (b) Score-based likelihood ratio using the mean nearest neighbor distance and account weighting scheme. 114

5.1 Series of authentication events for logins to a unique computer (known source series B) and a shared compute resource (unknown source series A) taken from Section 5.8. Both series were generated by the same user. 121
5.2 Example of temporal marked point processes from two different individuals (i and k) taken from the case study of Section 5.7. Note that A and B events generated by the same individual tend to cluster temporally, with less clustering in time for A and B events from different users. 122
5.3 Example mean inter-event time calculation. 130
5.4 Example of sessionized resampling for a pair of event series (A^*, B^*) taken from the student web browsing data. Here we use $T = 10$ minutes, and the distribution of session start times $p(t_{ses})$ is the empirical distribution of session start times across all series B available in the data set. $B^{(\ell)}$ for $\ell = 1, \ldots, 5$ represent five event series simulated via Algorithm 2. 133

5.5 (Top) Boxplot of measures of association for $\Delta(A, B) = \overline{\mathcal{T}}_{BA}$ for simulated data with $p = 0.20$ and (Bottom) corresponding AUC values as a function of the signal-to-noise ratio. (a) Score-based likelihood ratio. Note the different scales of the SLR for SNR $\in \{7.3, 14.6\}$. (b) Coincidental match probability. Note the CMPs for independent pairs are uniformly distributed by definition and thus omitted. 138
5.6 ECE plots for simulated data with varying signal-to-noise ratios. In all cases, $p = 0.20$ and the SLR using the mean inter-event time score function. (a) SNR $= 0.073$; (b) SNR $= 0.73$; (c) SNR $= 3.65$. 139
5.7 AUC values for both the SLR and CMP as a function of SNR for simulated data with $p = 0.20$. 140
5.8 Generalized additive model (GAM) smoother of the score-based likelihood ratio for simulated associated pairs with $p = 0.20$ as a function of the number of events in series B. Smoother fit in black and 99% confidence interval in grey. Note the different scales on the y-axes. 141
5.9 Web browsing data observed over 7 days from a random sample of 10 users from the case study data. Each user has two rows corresponding to the two event series with the top row of grey bars representing non-Facebook events (A_i) and the bottom row of black bars representing Facebook web browsing events (B_i). Note that all events shown above are relative to the first day of observation for each student, and each tick mark on the x-axis represents midnight of the corresponding day. 144
5.10 Empirical distributions of the score functions from Section 5.4. Same source distributions (H_s, dashed line) and different source distributions (H_d, solid line) approximated via kernel density estimation with Gaussian kernels and Scott's rule of thumb bandwidth. Leave-pairs-out cross-validation was not used to produce these densities; instead all pairs were used for illustrative purposes. Note that for the mingling index in (b) same-source pairs typically have higher score values than different-source pairs, so the inequalities in the SLR and CMP of Equations 5.8 and 5.7 are reversed. 145
5.11 Empirical cross-entropy (ECE) plots for the SLR with each score function described in Section 5.4.1. C_{llr} values are provided in the legend. (a) Coefficient of segregation, $S(A, B)$; (b) mingling index, $\overline{M}_1(A, B)$; (c) mean inter-event time, $\overline{\mathcal{T}}_{BA}$; and (d) median inter-event time, $med(\mathcal{T}_{BA})$. 147
5.12 LANL authentication data. Unique machine refers to Target X and Target Y for Actor 1 and Actor 2, respectively, and shared machine refers to Target Z for both actors. 151

LIST OF TABLES

Page

2.1 Example verbal scale for presenting conclusions from the LR from Association of Forensic Science Providers [2009]. 21

4.1 Number of observed days, accounts, events and visits for the Twitter data sets. Average number per account denoted in parentheses. 97
4.2 Number of observed accounts and visits for the Twitter data sets used in the analysis. Average number per account denoted in parentheses. 97
4.3 Summary statistics for the distribution of number of visits (Type "Visits") and unique accounts with at least one visit (Type "Accounts") in each parcel computed from the full population data in Table 4.1. The minimum and 25th percentile are 1 for all cases. 98
4.4 Performance of a classifier based on LR for the Twitter data. 100
4.5 Performance of a classifier based on SLR_{Δ} for the Twitter data. 100
4.6 Performance of a classifier based on CMP_{Δ} for the Twitter data. 101
4.7 Number of observed days, accounts, events and visits for the Gowalla data. Average number per account denoted in parentheses. 109
4.8 Summary statistics for the distribution of number of visits (Type "Visits") and unique accounts with at least one visit (Type "Accounts") in each parcel computed from the full population data in Table 4.1. The minimum and 25th percentile are 1 for all cases. 110
4.9 Number of observed accounts and visits for the Gowalla data used in the analysis. Average number per account denoted in parentheses. 111
4.10 Performance of a classifier based on LR for the Gowalla data. 111
4.11 Performance of a classifier based on SLR_{Δ} for the Gowalla data. 112
4.12 Performance of a classifier based on CMP_{Δ} for the Gowalla data. 112

5.1 Performance of a classifier based on SLR_{Δ} for the student web browsing data. 146
5.2 Performance of a classifier based on CMP_{Δ} computed using population data for the student web browsing data. 148
5.3 Performance of a classifier based on CMP_{Δ} computed via resampling for the student web browsing data. 149
5.4 Number of login events for each user to each target computer on the first day of activity in the LANL authentication data. 151
5.5 Coincidental match probabilities for various score functions for the LANL authentication data. Lower scores are indicative of same source event series. 152

LIST OF ALGORITHMS

		Page
1	Pool Adjacent Violators (PAV) Algorithm for isotonic regression.	44
2	Sessionized Resampling	134
3	Simulation of associated marked point processes	136

ACKNOWLEDGMENTS

I would like to start by thanking my advisor, Padhraic Smyth. His guidance, support and mentorship have been invaluable throughout the duration of my time at UCI. Padhraic always pushed me to succeed in research, but not at the cost of coursework, personal goals or mental health. It has been an honor working with him.

I also thank rest of my committee, Hal Stern and Veronica Berrocal. The time they spent reading and providing feedback on this work helped to form it into the product it is today. I especially thank Hal for co-authoring the manuscripts that this dissertation is based upon and for his invaluable insights to both statistics and forensics. I thank the researchers that were a part of The Center for Statistics and Applications in Forensic Evidence (CSAFE) whom taught me that forensic analysis is a much more complicated process than I originally assumed. Statistics is an interdisciplinary endeavor, and I am grateful for the opportunities CSAFE presented to me for advancing such an impactful field. Finally, I thank Gloria Mark for providing the student web browsing data discussed in Chapter 5.

To the past and present members of the Smyth DataLab, you have challenged me to be a better researcher, and I thank you for that. In particular, Homer Strong challenged me in both coursework and research early on in my time at UCI. My understanding of statistical theory is owed in large part to our whiteboard discussions and studying sessions. Moshe Lichman also helped me a great deal in my first research projects at UCI by improving my coding skills and aiding in collection of the Twitter data analyzed in Chapter 4. Eric Nalisnick always asked excellent questions that challenged me to think about my projects differently and provided valuable feedback on this dissertation.

I have also learned much from the mentors, collaborators, and friends with whom I have crossed paths during my stints in industry, including Alexander Vandenberg-Rhodes, Matt Wolff, Naresh Chebolu, Michael Slawinski and Michael Wojnowicz. None of my accomplishments would have been possible without the guidance of my undergraduate advisor at South Dakota State University, Kurt Cogswell, whose door was always open for discussions.

On a personal note, my time at UCI would not have been as successful without the lasting friendships I have made here. Regular pub visits with Eric, Homer, Alexander, Lars Hertel, Brian Vegetabile, Andrew Holbrook, Maricela Cruz, Micah Jackson, Lingge Li and others provided some much needed relief from coursework and research. Moshe and Dimitrios Kotzias frequently dragged me away from my desk to the Eastern Sierra for snowboarding trips that are some of my fondest memories of UCI.

This dissertation would not have been successful without the support of my family. Knowing that my parents, Larry and Lynn, support me in all of my endeavors has made all of the difference throughout my life. And last but certainly not least, I want to thank Jordan Smith for all of the love and support she has given me over the years. You made the good times better, and the difficult times easier. I cannot imagine getting through this without you.

This research was partially funded through Cooperative Agreement #70NANB15H176 between the National Institute of Standards and Technology and Iowa State University, which includes activities carried out at Carnegie Mellon University, University of California, Irvine, and University of Virginia. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Institute of Science and Technology, nor of the Center for Statistics and Applications in Forensic Evidence.

VITA

Christopher Galbraith

EDUCATION

Doctor of Philosophy in Statistics **2020**
University of California, Irvine *Irvine, California*

Master of Science in Statistics **2016**
University of California, Irvine *Irvine, California*

Bachelor of Science in Mathematics **2014**
South Dakota State University *Brookings, South Dakota*

RESEARCH EXPERIENCE

Graduate Research Assistant **2014–2020**
University of California, Irvine *Irvine, California*

REU Scholar **2012**
SDSU Research Experience for Undergraduates *Brookings, South Dakota*

REFEREED JOURNAL PUBLICATIONS

Christopher Galbraith, Padhraic Smyth and Hal S. Stern. **Quantifying the association between discrete event time series with applications to digital forensics.** *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2020.

REFEREED CONFERENCE PUBLICATIONS

Christopher Galbraith, Padhraic Smyth and Hal S. Stern. **Statistical methods for the forensic analysis of geolocated event data.** *Digital Investigation*, in press, 2020.

Christopher Galbraith and Padhraic Smyth. **Analyzing user-event data using score-based likelihood ratios with marked point processes.** *Digital Investigation*, 22, S106 – S114, 2017.

SOFTWARE

assocr `https://github.com/UCIDataLab/assocr`
R implementation of SLR and CMP methods for temporal event data.

TEACHING EXPERIENCE

Instructor – Introduction to R **2018**
California State University, Long Beach *Long Beach, California*

Instructor – Introduction to Data Analysis with R **2016–2018**
UCI Data Science Initiative *Irvine, California*

Teaching Assistant – Basic Statistics **2016**
University of California, Irvine *Irvine, California*

PROFESSIONAL EXPERIENCE

Machine Learning Researcher **2018 – 2020**
Obsidian Security *Newport Beach, California*

Data Science Intern **2018**
Obsidian Security *Newport Beach, California*

Data Science Intern **2017**
Cylance Inc. *Irvine, California*

Analyst **2013–2014**
Wells Fargo & Company *Sioux Falls, South Dakota*

Analytics Intern **2013**
Wells Fargo & Company *Sioux Falls, South Dakota*

HONORS & AWARDS

Honorary Fellow, UCI Machine Learning and Physical Sciences Program **2017**
Honorable Mention, NSF Graduate Research Fellowships Program **2016**
Recipient, UCI Graduate Dean's Recruitment Fellowship **2014**
Recipient, SDSU Schultz-Werth Student Paper Award **2014**
Inductee, Pi Mu Epsilon Honor Society **2014**

PROFESSIONAL MEMBERSHIPS

Member, American Statistical Association **2018–present**

ABSTRACT OF THE DISSERTATION

Statistical Methods for the Forensic Analysis of User-Event Data

By

Christopher Galbraith

Doctor of Philosophy in Statistics

University of California, Irvine, 2020

Chancellor's Professor Padhraic Smyth, Chair

A common question in forensic analysis is whether two observed data sets originate from the same source or from different sources. Statistical approaches to addressing this question have been widely adopted within the forensics community, particularly for DNA evidence, providing forensic investigators with tools that allow them to make robust inferences from limited and noisy data. For other types of evidence, such as fingerprints, shoeprints, bullet casing impressions and glass fragments, the development of quantitative methodologies is more challenging. In particular, there are significant challenges in developing realistic statistical models, both for capturing the process by which the evidential data is produced and for modeling the inherent variability of such data from a relevant population.

In this context, the increased prevalence of *digital evidence* presents both opportunities and challenges from a statistical perspective. Digital evidence is typically defined as evidence obtained from a digital device, such as a mobile phone or computer. As the use of digital devices has increased, so too has the amount of user-generated event data collected by these devices. However, current research in digital forensics often focuses on addressing issues related to information extraction and reconstruction from devices and not on quantifying the strength of evidence as it relates to questions of source.

This dissertation begins with a survey of techniques for quantifying the strength of evidence

(the likelihood ratio, score-based likelihood ratio and coincidental match probability) and evaluating their performance. The evidence evaluation techniques are then adapted to digital evidence. First, the application of statistical approaches to same-source forensic questions for spatial event data, such as determining the likelihood that two sets of observed GPS locations were generated by the same individual, is investigated. The methods are applied to two geolocated event data sets obtained from social networks. Next, techniques are developed for quantifying the degree of association between pairs of discrete event time series, including a novel resampling technique when population data is not available. The methods are applied to simulated data and two real-world data sets consisting of logs of computer activity and achieve accurate results across all data sets. The dissertation concludes with suggestions for future work.

Chapter 1

Introduction

When a crime is committed, the subsequent investigation may identify a variety of evidence or information that can be used to help identify the perpetrator. This evidence can aid the investigation by helping law enforcement formulate hypotheses about the crime (e.g., how, when, who). It can also help the prosecution (or defense) in a court of law convince the judge or jury about a suspect's guilt (or innocence). The latter use of forensic evidence can be very powerful, but several recent events have raised questions about the scientific foundation of the analysis and interpretation of said evidence. For example, the Federal Bureau of Investigation mistakenly identified Brandon Mayfield as the source of a latent fingerprint found at the scene of a 2004 train bombing in Spain [Fine, 2006]. The FBI arrested and held the Portland, Oregon, based lawyer for over two weeks, although he had never been to Spain. Mayfield was later released and cleared of any association with the crime. A plethora of other examples can be found in the National Registry of Exonerations, which provides detailed information about every known wrongful conviction in the United States since 1989. As of early 2020, there have been over 2,600 exonerations, and unreliable or improper forensic science was found to be a contributing factor in roughly 25 percent of those cases [National Registry of Exonerations].

These controversies have elicited a reaction from a number of governmental agencies. For instance, the President's Council of Advisors on Science and Technology penned a 2016 report [PCAST, 2016] that identified a number of challenges associated with forensic science and issued recommendations on actions to take that could address them. While improving the practice of forensic science is clearly a multidisciplinary challenge (involving forensic subject matter experts and legal practitioners including judges, lawyers and law enforcement), the field of statistics has a significant role to play in the effort.

Forensic analysis involves examining evidence during a civil or criminal legal investigation. For this dissertation we focus on forensic analysis in criminal settings. Statistical techniques have played a key role in forensic analysis, providing forensic investigators with tools that allow them to make robust inferences from limited and noisy data. The best-known example in this context is the use of likelihood-ratio techniques for assessing the strength of the evidence that a DNA sample from a crime scene is a match to a suspect's DNA sample [Evett and Weir, 1998; Myers et al., 2011]. For other types of evidence, such as fingerprints, shoeprints, bullet casing impressions, glass fragments, and so on, the development of quantitative methodologies (such as likelihood ratio techniques) is more challenging [Stern, 2017]. In particular, there are significant challenges in developing realistic statistical models, both for capturing the process by which the evidential data is produced and for modeling the inherent variability of such data from a relevant population.

In this context, the increased prevalence of *digital evidence* presents both opportunities and challenges from a statistical perspective. Digital evidence is typically defined as evidence obtained from a digital device, such as a mobile phone or a computer, where the evidence is associated with a crime scene or with a suspect. As the use of digital devices has increased, so too has the amount of user-generated event data collected by these devices. Such data can be obtained from logs of timestamped events stored either directly on a device, such as a mobile phone or computer, or stored on a user's account in the cloud [Oh et al., 2011; Roussev and

McCulley, 2016]. Examples of such events include user actions within particular software, searching or browsing activities in a web browser, communicating via email or text messaging, and so on. This type of user-generated event data tends to be (i) inhomogeneous over time (often with circadian rhythms), (ii) bursty, with brief periods of high activity followed by periods of no activity [e.g., Radicchi, 2009; Barabasi, 2005], and (iii) heterogeneous across a across different users [e.g., Lichman and Smyth, 2014]. These general characteristics pose a number of challenges from the perspective of developing appropriate statistical models.

There is significant interest in the development of tools that can assist in the investigation of user-generated event logs from digital devices [e.g., Casey, 2011; Roussev, 2016]. Current research in digital forensics often focuses on addressing issues related to information extraction and information reconstruction from devices or from the cloud [e.g., SWGDE, 2019]. However, citing Casey [2018], there is "a growing expectation that forensic practitioners treat digital traces in a manner that is becoming widely accepted in forensic science: evaluating and expressing the relative probabilities of the forensic findings given at least two mutually exclusive hypotheses." As an example, the Organization of Scientific Area Committees (OSAC) for Forensic Science Task Group on Digital/Multimedia Science recently issued a recommendation to develop "systematic and coherent methods for studying the principles of digital/multimedia evidence to assess the causes and meaning of traces in the context of forensic questions, as well as any associated probabilities" [Pollitt et al., 2019].

This dissertation focuses on the problem of quantifying the degree of association between two sets of user-generated events in either the temporal (e.g., only timestamps of the events are available) or spatial (e.g., only the spatial locations of the events are available) setting. As an example, consider the case where one event series A consists of a log of timestamped events (such as logins, file access events, browsing, messaging) generated on a device associated with a crime (e.g., on a mobile phone found at a crime scene). A second event series B consists of a log of similar events associated with a suspect (e.g., user-generated events recorded on a

device owned by the suspect). The evidence consists of both event series of events A, B and the question of interest is to determine how likely it is that the two series were generated by the same individual. See Figure 1.1 for an example of such temporal event data.

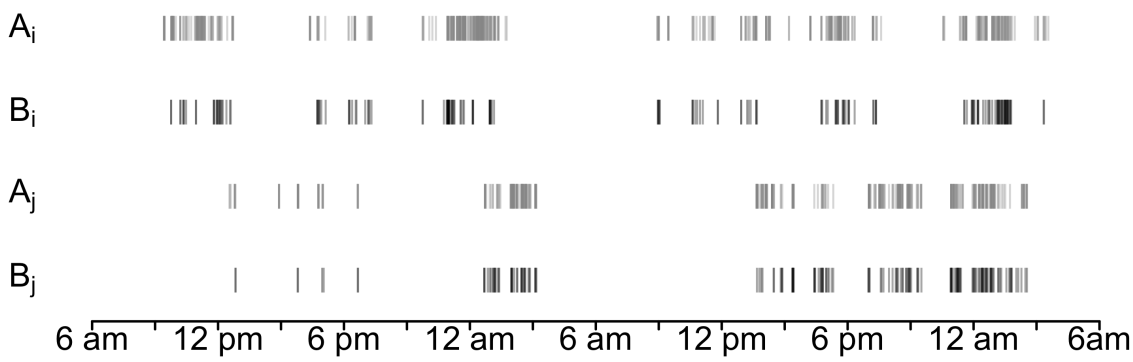


Figure 1.1: Example of temporal evidence from two individuals (i and j) taken from the case study of Chapter 5. A and B events generated by the same individual tend to cluster temporally, with less clustering in time for A and B events from different users.

While the forensic examiner may be able to determine if two sequences of events were in fact generated by the same individual (e.g., by comparison of the event series in Figure 1.1 by eye), it may take a significant amount of time and will result in a conclusion that does not meet the scientific standards for presentation in a court of law. The techniques presented in this dissertation solve both of these issues by providing investigators tools for the objective quantification of digital evidence that adhere to the standards for presentation in court.

1.1 Outline & Contributions

The structure of the dissertation, including novel contributions, is as follows:

Chapter 2 provides a brief review of a variety of types of forensic evidence and then describes how the likelihood ratio is used to quantify the strength of evidence by modeling the features

of said evidence directly. Various techniques to assess the performance of the likelihood ratio are then presented, including classification accuracy, information-theoretic value and calibration. The chapter finishes with a brief discussion of the key results and challenges from a statistical perspective.

The contributions of Chapter 2 include:

- A survey of likelihood ratio techniques for the analysis of forensic evidence that unifies the forensics, biometrics and statistics literature on the topic.
- A concise discussion of the methods to evaluate the LR via classification performance, information-theoretic value (and its relationship to frequentist decision theory), and calibration. This portion unifies concepts and notation from previous literature on the topics.

Chapter 3 presents an alternate approach to the likelihood ratio for quantifying strength of evidence. For many types of forensic evidence, it is rare that its underlying generative process is sufficiently understood to make the distributional assumptions required to compute the likelihood ratio. This chapter introduces the concept of score-based methods, which instead measure the similarity of the two observed data sets via a score function. That score function can then be used to assess the strength of evidence via the score-based likelihood ratio or coincidental match probability. A comparative evaluation using simulated data with known distributions is presented that illustrates the different behavior of the methods.

The contributions of Chapter 3 include:

- A survey of the score-based likelihood ratio for the analysis of forensic evidence that unifies the forensics, biometrics and statistics literature on the topic.
- A score-based adaptation of the coincidental match probability for the evaluation of

forensic evidence.

- A comparison of the direct modeling approach of the likelihood ratio to the score-based approaches of the score-based likelihood ratio and coincidental match probability for a theoretical example of evidence from a known distribution.

Chapter 4 presents an application of the methods in Chapters 2 and 3 to geolocated event data. Part of these results has been published in Galbraith et al. [2020b]. The evidence under consideration consists of sequences of GPS coordinates recorded from mobile devices. I show how mixtures of kernel density estimators can be used to estimate the likelihood ratio. I then investigate appropriate score functions for such data, including nearest neighbor distance and earth mover's distance, and weighting schemes for the score functions based on geoparcel data (i.e., disjoint polygons that partition a spatial region where each individual parcel represents a specific property). A comparative evaluation using geolocated Twitter event data and Gowalla check-in data from two large metropolitan areas shows the potential efficacy of such techniques.

The contributions of Chapter 4 include:

- A novel technique to quantifying the strength of evidence for geolocated event data via the likelihood ratio using mixtures of kernel density estimators.
- We developed a variety of appropriate score functions that can distinguish between same- and different-source series of geolocated events.
- Extensive experimental comparison of LR, SLR and CMP methods on two large real-world data sets in two different regions within the US.

Chapter 5 presents an application of the score-based methods of Chapter 3 to temporal event data. This chapter is more in-depth adaptation of my work in [Galbraith and Smyth,

2017] and [Galbraith et al., 2020a]. I focus in particular on the case where two associated event series exhibit temporal clustering such that the occurrence of one type of event at a particular time increases the likelihood that an event of the other type will also occur nearby in time. A non-parametric approach to the problem is pursued, and different score functions to quantify association, including characteristics of marked point processes and summary statistics of inter-event times, are investigated. Two techniques are proposed for assessing the significance of the measured degree of association: (i) a population-based approach to calculating score-based likelihood ratios when a sample from a relevant population is available, and (ii) a resampling approach to computing coincidental match probabilities when only a single pair of event series is available. The methods are applied to simulated data and to two real-world data sets consisting of logs of computer activity and achieve accurate results across all data sets.

The contributions of Chapter 5 include:

- A variety of appropriate score functions that can distinguish between same- and different-source series of temporal events.
- Novel techniques for quantifying the strength of evidence for temporal event data via the score-based likelihood ratio and coincidental match probability.
- A resampling technique that allows for the computation of strength of evidence when a reference population is not available.

Chapter 6 presents remaining open questions, providing a starting point for future work in the area of statistical modeling in digital forensics.

Chapter 2

Computing Strength of Evidence with the Likelihood Ratio

In this chapter, I provide a brief review of a variety of types of forensic evidence. I then describe how the likelihood ratio is used to quantify the strength of evidence by modeling the features of the evidence directly. Various techniques to assess the performance of the likelihood ratio are then presented, including classification accuracy, information-theoretical value, and calibration. The chapter finishes with a brief discussion of the key results and challenges from a statistical perspective.

2.1 Evidence Types

There are a variety of types of evidence that a forensic examiner may encounter in practice, e.g., DNA samples, fingerprints, handwriting and glass fragments. In general, the various types of evidence fall into one of three categories—biological evidence, trace evidence or pattern evidence—each of which are reviewed below. The applicable techniques for quan-

tifying the strength of evidence are contextually dependent on both the category and type of evidence under consideration. For example, appropriate statistical models for DNA evidence are not applicable to handwriting evidence and vice versa, as the measured features on each have vastly different properties and different levels of scientific understanding of their generative mechanisms.

In addition to modeling considerations, expert testimony in federal courts is governed by guidelines established in the 1993 case *Daubert v. Merrell Dow* known as the Daubert standard [Daubert v. Merrell Dow Pharmaceuticals, Inc., 1993]. The Daubert standard identifies many factors that should be considered by a judge when determining whether to allow expert scientific testimony about a given form of forensic evidence. These factors include whether the evaluation technique follows the scientific method, is generally accepted in the scientific literature, and provides information about the associated error rate. Legal precedent does not necessarily imply that an analysis method for a given evidence type satisfies the Daubert standard. For example, bite mark analysis has been admitted in court for over 30 years, although there has been little evidence that it can be reliably assessed by forensic examiners [Saks et al., 2016]. Therefore, the discussion below focuses on the scientific validity of methods for quantifying the strength of evidence and not on the legal precedents that currently exist in practice.

2.1.1 Biological (DNA) Evidence

DNA analysis of single-source samples is the gold standard in forensic science for evaluating evidence. There is a rich scientific literature in both biology and forensics that supports the analysis [e.g., Steele and Balding, 2014]. In this setting, a DNA profile from a known source (e.g., the suspect's DNA) is compared to another profile from an unknown source (e.g., a sample recovered from the crime scene). The profiles consist of a sequence of alleles

at a specific set of locations in the genome where there is known to be heterogeneity in the population. First, the examiner determines if the profiles "match," i.e., the known source profile and unknown source profile share the same sequence of alleles. A probability model is then formulated to assess the likelihood of observing these matching profiles by chance in the population—the lower this likelihood the more probable it is that the samples were from the same individual. Since the underlying biology of genetic inheritance is well understood, a precise model for the likelihood of matching DNA profiles can be formulated. Published databases are available that allow for accurate estimates of the unknown parameters of the model, i.e., the population frequencies of different alleles.

There are some settings, however, where considerable uncertainty about how to analyze DNA evidence exists. In particular, when the DNA sample recovered from the crime scene is a mixture of genetic material from multiple individuals, inter-laboratory studies have demonstrated variability in the conclusions reached for a given sample. Numerous statistical approaches can be applied in this setting [e.g., Kelly et al., 2014]. For this reason, when DNA evidence is referred to as the "gold standard" for forensic science it is assumed that single-source profiles are under consideration.

2.1.2 Trace Evidence

Trace evidence refers to materials that can be transferred during the commission of a crime, including hair, fibers, soil and glass. The analysis of such evidence varies depending on the specific type of trace evidence. In some cases, the evidence is comprised of measurements of the chemical concentrations of elements in the sample (e.g., glass, gasoline residue, gunshot residue, etc). For other types of trace evidence, including hair and fibers, the analysis more closely resembles that of pattern evidence (discussed later in Section 2.1.3).

The standard approach for analyzing trace evidence consisting of measurements of chemical

concentrations is to compare the population mean concentrations for the population of the crime scene evidence to the population mean for the population of the suspect evidence using a standard significance test or related procedure [e.g., Almirall and Trejos, 2006; Aitken and Lucy, 2004]. Failure to reject the hypothesis of equal population means is often said to indicate that the two samples are indistinguishable. This approach typically assumes a distributional form for the measurements [e.g., Aitken and Lucy, 2004; Vergeer et al., 2014].

2.1.3 Pattern Evidence

Pattern evidence refers to evidence produced when one object comes into contact with another and leaves an impression, including fingerprints, shoeprints, toolmarks, bullet casing impressions, and handwriting. The primary characteristic of pattern evidence is that a pattern left at a crime scene is compared to another pattern from a known source. This results in a broad category of evidence that can also include some types of digital evidence like voice recordings and event data, as these digital impressions are left by an individual on a device.

Pattern evidence is considered to be one of the most difficult types of evidence to work with in terms of producing appropriate statistical models for the observed features, which are typically high-dimensional and complex (usually in the form of an image). Unlike DNA evidence, it can also be difficult to obtain a relevant reference population for assessing the probability of coincidental matches. Often, simplifying assumptions must be made, e.g., the similarity of two impressions can be projected onto a lower dimensional space using a score function [e.g., Ali et al., 2014; Hepler et al., 2012; Ramos et al., 2017]. There is considerable research aimed at developing applicable methods for evaluating pattern evidence, and later in this thesis we will extend these methods to the domain of digital evidence.

2.2 Formal Problem Statement

A common problem in forensic science is that of determining the degree to which two samples of pattern evidence "match," or have the same generative mechanism. The evidence corresponds to observed data and can take different forms such as measurements related to DNA, fingerprints, or shoe prints [e.g., Aitken and Stoney, 1991]. Denote the evidence as $E = (A, B)$, where in general both A and B are a sets of observations (measured "features") for samples from the evidence type under question.

The goal of a forensic examination is to assess the likelihood of observing the evidence $E = (A, B)$ under two mutually exclusive hypotheses

H_s: (A, B) came from the same source;
H_d: (A, B) came from different sources.

These hypothesis are commonly referred to as the prosecution (H_s) and defense (H_d) propositions [Aitken and Taroni, 2004]. However, there can be ambiguity in the way that these hypotheses are currently stated that can result in the development of models that mismatch the needs of the criminal justice system. The following section briefly develops two formal scenarios that frame the inference for the source of forensic evidence: the common source proposition and the specific source proposition [Ommen and Saunders, 2018].

2.2.1 Source Propositions

The common source and specific source scenarios are often confused with one another. This can result in the development of models under one scenario to answer the question considered by the other one [Neumann and Ausdemore, 2019]. Thus, understanding their differences is important and helps assess the potential and limitations of the different inference frameworks

for forensic evidence.

Common Source

The common source scenario considers whether two samples of forensic evidence originate from the same source or from different sources without formally specifying which sources are considered. In this scenario, the goal is to determine if A and B originate from the same unknown source. For example, an investigator could be interested in determining if two fingerprint impressions or DNA samples recovered at one or more crime scenes are from the same individual, thus linking the crime scenes or helping form conclusions on the number or perpetrators. No specific individual must be identified in order for the analysis to proceed. Instead, the focus is solely on determining if a common source could be responsible for both samples.

Formally, the common source problem can be stated as follows:

H_s: A and B originate from the same, unknown source;
H_d: A and B originate from from two different, unknown sources.

Here the sources are assumed to be randomly selected from a relevant population of potential sources. Under the same-source hypothesis H_s, A and B have the same random source. Under different source hypothesis H_d, A and B originate from two different random sources.

For the digital event data under consideration in Chapters 4 and 5, the common source problem is generally not applicable. The heterogeneity of behavior across different individuals makes it difficult to build models that can reasonably assess a common source problem in practice.

Specific Source

The specific source scenario typically involves the comparison of a trace sample from an unknown source with a control sample from a known source. The goal is to determine if the trace sample was generated by the source that generated the control sample. Here, we define the sets of observations that compose the evidence as follows:

- A: set of observations for a sample from an unknown source (e.g., a sample recovered from a crime scene);
- B: set of observations for a reference sample from a known source (e.g., a sample from a suspect).

Formally, the specific source problem can be stated as follows:

- H_s: The unknown source evidence A and the specific source evidence B both originate from the specific source;
- H_d: The unknown source evidence A does not originate from the specific source, but from some other source in an alternative population.

This is the most common scenario in forensic investigations, with H_s and H_d above typically being referred to as the "same source hypothesis" and "different source hypothesis" respectively [Stern, 2017]. For instance, A could be a fingerprint found at the crime scene and B a suspect's fingerprint collected by law enforcement. The same source hypothesis is that both the crime scene fingerprint and suspect's fingerprint came from the same source (the suspect's finger). Under the different source hypothesis, the crime scene fingerprint was not generated by the suspect. Any observed similarities between A and B must be due to chance.

In the context of the user-generated event data under consideration in Chapters 4 and 5, "source" refers to a specific individual or user account, and "came from" can be interpreted

as "generated by." Thus, H_s is the proposition that the sample from the unknown source A was generated by the same individual or user account as the sample from the known source B. H_d is the proposition that the sample from the unknown source A was *not* generated by the specific source of B, but instead from another individual among an alternative source population.

We henceforth operate under the specific source problem assumptions, and all references to the evidence $E = (A, B)$ and hypotheses H_s and H_d will use the specific source definitions. The evidence will be referred to interchangeably as either E or (A, B).

2.3 The Likelihood Ratio

The *likelihood ratio (LR)* is widely accepted in the forensic science community as "a logically defensible way" to assess the strength of evidence [Willis et al., 2016] having been applied in a variety of forensic disciplines, including handwriting [Bozza et al., 2008], speech [Champod and Meuwly, 2000], fingerprints [Champod and Evett, 2001] and DNA [Aitken and Stoney, 1991; Evett and Weir, 1998]. See Stern [2017] for a thorough discussion of the likelihood ratio's application across a variety of forensic disciplines. The term "strength of evidence" refers to the amount of support that the LR provides to the same-source proposition H_s relative to the different-source proposition H_d. This term has a long history, with the first known mention of the synonymous "weight of evidence" occurring in Peirce [1878] in reference to the logarithm of the likelihood ratio.

The LR arises naturally in the application of Bayes' Theorem to updating the relative likelihoods (odds) of the two competing hypotheses given the evidence $E = (A, B)$. Bayes'

Theorem in the forensic context is

$$\underbrace{\frac{Pr(H_s|A,B,I)}{Pr(H_d|A,B,I)}}_{\text{posterior odds}} = \overbrace{\frac{Pr(A,B|H_s,I)}{Pr(A,B|H_d,I)}}^{\text{likelihood ratio}} \underbrace{\frac{Pr(H_s|I)}{Pr(H_d|I)}}_{\text{prior odds}} \tag{2.1}$$

where $Pr(\cdot)$ refers to the appropriate probability distribution, and I is all of the information available to the decision-maker prior to the introduction of the evidence (A, B). For the likelihood ratio term these are probability distributions for the evidence (i.e., either a probability mass function or probability density function) and for the prior and posterior odds these are probabilities assigned to the hypotheses.

The likelihood ratio measures the relative probability of obtaining the evidence (A, B) under the two hypotheses. A large likelihood ratio means the observed evidence is much more likely under the same-source hypothesis H_s than the different-source hypothesis H_d. A small LR means that the observed evidence is much less likely under the same-source hypothesis. Equation 2.1 tells the evaluator of the evidence (e.g., a member of the jury) how to modify their prior (pre-evidence) odds given the evidence in order to obtain posterior odds of the two hypotheses. One common view is that the goal of the forensic examination is to supply the LR to said evaluator.

In addition to the evidence, there may also be additional information I that should be considered during evaluation. This can include information about how the evidence itself was collected or information about the relevant population distribution of various characteristics. For instance, this could be population data relevant to (A, B) as discussed in Section 2.3.2. What should be included in I is a current topic of discussion in the forensic science community. It is generally understood that I should not include information about other evidence in the case or certain other local circumstances, as this could possibly result in task-irrelevant information forming cognitive biases for the examiner [National Commission on Forensic

Science, 2015].

2.3.1 The LR as a Bayesian Method

The likelihood ratio or Bayes factor arises in the application of Bayes' Theorem to updating the relative odds of the two competing hypotheses given the evidence (A, B). For this reason, the LR can be viewed as a Bayesian approach to the analysis of forensic evidence. This is somewhat of a misnomer, however, for three primary reasons: how the prior probabilities of each hypothesis are handled, the fact that the LR also arises in frequentist inference, and how nuisance parameters are handled in the LR.

Bayesian methodologies refer to inferential procedures in which Bayes' Theorem is used to combine prior information with observed data. Thus, the likelihood ratio can be reasonably justified as being a Bayesian approach to forensic inference. However, the LR itself does not address the role of the prior probabilities $Pr(H_s|I)$ and $Pr(H_d|I)$ in the inference, which implies the likelihood ratio is not a fully Bayesian approach. In fact, these prior probabilities can be difficult to specify, and differing priors can result in drastically different posterior conclusions being formed. For example, in the United States accused suspects are presumed innocent until proven guilty. One interpretation of this is that the prior probability of the same-source hypothesis should be zero, $Pr(H_s|I) = 0$, which would render the entire inferential procedure useless no matter the value of the LR. The role of the background information I plays a large role in formulating the priors, too. Take for instance, a case in which I restricts the relevant reference population from 100,000 suspects living in some region to 10 suspects known to be associated with the victim. This would result in a scenario in which the prior probability $Pr(H_d|I)$ increases from 1/100,000 to 1/10. Given the same observed evidence, hypotheses and procedure to calculate the LR the posteriors would differ dramatically, possibly resulting in different conclusions being made. For a more thorough

discussion of why fully Bayesian inference has not played a larger role in the legal system, including examples from real cases, see Fenton et al. [2016].

Another distinction between the likelihood ratio and Bayesian methods is that the LR is not an exclusively Bayesian concept. The definition in Equation 2.1 is not the only way to arrive at the likelihood ratio. In fact, the likelihood ratio is a prominent feature in traditional frequentist inference. The most well-known example is the likelihood ratio test, which is used to assesses the goodness of fit of two competing statistical models based on the ratio of their likelihoods, specifically one found by maximization over the entire parameter space and another found after imposing some constraint (i.e., the null hypothesis). This test can be applied to both nested [Neyman and Pearson, 1933] and non-nested hypotheses [Vuong, 1989].

The final distinction relates to the handling of nuisance parameters. In many cases, the likelihood ratio will depend on unknown parameters (e.g., population frequencies of alleles in DNA analysis). These parameters can be handled by one of two methods: estimation using frequentist techniques such as the maximum likelihood estimator, yielding the frequentist version of the likelihood ratio, or by averaging over the distribution of the parameters in a Bayesian analysis, yielding the Bayes factor. This implies that either frequentist or Bayesian methods can be used to estimate the LR. We will use the term likelihood ratio to refer to both approaches, and the specific details as to how it is calculated for a particular type of evidence determine which method is used.

2.3.2 Estimation

The likelihood ratio of Equation 2.1 requires probabilistic generative models $Pr(\cdot)$ for the evidence $E = (A, B)$. Specifying such models can be extremely difficult. One would have to construct two models that not only specify the distribution of the features of A and B but also

the correlation between those features under the same- and different-source hypotheses. A well-known way [Stern, 2017] to simplify the likelihood ratio is to factor the joint distribution of (A, B) under each model such that

$$LR = \frac{Pr(A, B|H_s, I)}{Pr(A, B|H_d, I)} = \frac{Pr(B|A, H_s, I)Pr(A|H_s, I)}{Pr(B|A, H_d, I)Pr(A|H_d, I)} = \frac{Pr(B|A, H_s, I)}{Pr(B|H_d, I)}. \tag{2.2}$$

In this scenario $Pr(A|H_s, I) = Pr(A|H_d, I) = Pr(A|I)$ because the distribution of A itself does not depend on the same- or different-source hypothesis. For example, in fingerprint analysis the marginal distribution of the location and type of minutiae, or features of fingerprint ridges, for a single latent print A does not depend on propositions about its source. Furthermore it is natural to assume the distribution of B is independent of A under the different-source hypothesis, which results in $Pr(B|A, H_d, I) = Pr(B|H_d, I)$. For the fingerprint analysis example, this assumption implies the distribution of minutiae in latent print B is independent of the locations of the minutiae in another latent print A from a different individual, i.e., information about a fingerprint from some randomly selected alternate source in the population does not provide any additional information about the fingerprint from the known source. In contrast, if we condition on the same-source hypothesis H_s (i.e., that A is from the same source as B), then A is informative about B (i.e., features of a print A from an individual will be informative about the features of another print B from the same individual).

When the evidence A and B are comprised of multiple observations, one approach to compute the likelihood ratio is to assume that the observations are conditionally independent of one another given the appropriate hypothesis and background information. In this situation, $A = \{a_i : i = 1, \ldots, n_a\}$ is composed of n_a total observations $a_i \in \mathbb{R}^d$ where d is the dimension of each observation. A similar definition holds for B. Under the conditional

independence assumption, the likelihood ratio from Equation 2.2 can be expressed as

$$LR = \prod_{j=1}^{n_b} \frac{Pr(b_j|A, H_s, I)}{Pr(b_j|H_d, I)}. \tag{2.3}$$

The conditional probability distributions $Pr(\cdot)$ in the numerator and denominator of Equations 2.2 and 2.3 can be estimated via a variety of techniques, all of which require reference data from a relevant population. Assume that we have a reference sample of N_s same-source exemplars $\mathcal{D}_s = \{(A_i, B_i) : i = 1, \ldots, N_s\}$ that were generated under H_s. Similarly, define a reference sample of N_d different-source exemplars $\mathcal{D}_d = \{(A_j, B_k) : j \neq k\}$ that were generated under H_d. These reference data sets will be used in the estimation of methods that generate likelihood ratio values, and should be composed of exemplars that have similar characteristics as the evidence under question (e.g., similar number of observations, marginal distributions of features, etc). A thorough discussion of the requisite population data is presented in Section 2.4.

Due to the large variety of types of evidence and measurements taken, appropriate probability models for the likelihood ratio can take many different forms. Presenting all such models is out of the scope of this work, so the reader is encouraged to refer to [Stern, 2017] for a survey of techniques for a variety of types of evidence. For digital evidence, appropriate probability models (when possible) should take into account the properties of user-event data, which include inhomogeneity, burstiness, and user-specific deviations from group behavior. Such models are discussed in more detail in Chapter 4. In other situations, such probability models can be difficult or even impossible to specify (one such example is given in Chapter 5). Other techniques must be used to quantify the strength of evidence in such situations.

2.3.3 Interpretation

After computing the likelihood ratio, the forensic investigator can then come to a conclusion about the two propositions under consideration. This conclusion should express the degree of support provided by the evidence for the same-source hypothesis H_s versus the different-source hypothesis H_d depending on the magnitude of the LR. See Willis et al. [2016] for practical guidelines.

When the $LR = 1$ the conclusion should be that the evidence provides no assistance in distinguishing between the two hypotheses. For $LR > 1$ the conclusion should be that the evidence is more probable if the unknown-source sample and known-source sample were generated by the source of the known-source sample. For $LR < 1$ the conclusion should be that the evidence is more probable if the alternative is true, i.e., the evidence was generated by different sources.

An open area of research is how such analyses of forensic evidence should be presented in court. There have been studies demonstrating the difficulty of understanding likelihood ratios for jurors and the common misconceptions that arise in doing so [e.g., Martire et al., 2013; Thompson and Newman, 2015; Thompson et al., 2018]. Statisticians have an important role to play in developing techniques for presenting quantitative evaluations of evidence and in the design and analysis of juror studies.

LR Value	Verbal Expression
1-10	Weak or limited support
10-100	Moderate support
100-1,000	Moderately strong support
1,000-10,000	Strong support
10,000-100,000	Very strong support
$> 100,000$	Extremely strong support

Table 2.1: Example verbal scale for presenting conclusions from the LR from Association of Forensic Science Providers [2009].

One such technique for that can help jurors' understanding of the probative value of evidence is to express the likelihood ratio via a verbal equivalent according to a scale of conclusions [Nordgaard and Rasmusson, 2012]. Table 2.1 provides an example of such a verbal equivalent. For a more thorough discussion on expressing the probative value of forensic evidence in a clear and consistent manner, see Thompson [2017].

2.4 Population Data

The likelihood ratio relies on having samples from a population to estimate the likelihood of observing the evidence under each source hypothesis. It is a common belief in the forensics community that said population should be *relevant* to the case at hand [e.g., Hughes, 2014]. There is no ambiguity in the definition of a relevant population for the numerator of the likelihood ratio, or the probability of observing the evidence given the same-source hypothesis is true. However, the definition of relevance for the different-source hypothesis is more problematic. The defense typically offers a non-specific alternative proposition, e.g., it was not the defendant that was the source of the evidence but someone else. In this situation, the relevant population consists of anyone *except* the suspect. When using a broad population, the resulting LR values are typically small regardless of the ground truth of the source proposition [Aitken and Taroni, 2004]. Therefore, it is necessary to reduce the relevant population "to more manageable proportions" [Aitken and Taroni, 2004, page 206], unless "there is no evidence to separate the perpetrator from the ... population (at large)" or where the evidence is independent of variation within sub-populations [Robertson and Vignaux, 1995, page 36].

The concept of the relevant population was first defined by [Coleman and Walls, 1974, page 276] as:

> those persons who could have been involved (in the crime); sometimes it can be established that the crime must have been committed by a particular class of persons on the basis of age, sex, occupation or other sub-grouping, and it is then not necessary to consider the remainder of, say the United Kingdom.

Defining the relevant population by dividing the entire population into subgroups by factors is considered logically relevant if the factors affect the distribution of the evidence (or a parameter involved in its estimation) in the wider population [Kaye, 2004]. This approach is extensively used in the evaluation of DNA evidence, as allele frequencies are known to differ by race [Gill and Clayton, 2009]. DNA analysis is aided by having large databases that enable examiners to quickly and easily obtain a sample from the logically relevant population. For other forms of evidence, however, such databases do not exist, and the examiner must define and collect samples from the relevant population him- or herself. For example, in forensic speaker recognition, the use of case-specific data is common, with a common viewpoint that the examiner must "be prepared to go and get a suitable reference sample for each case" [Rose, 2007].

Clearly, values of the likelihood ratio depend on how relevant population is defined. For that reason, it is typically assumed that the background information I upon which the likelihood ratio and prior odds are conditioned on in Equation 2.1 includes the relevant population. In estimation of the LR, therefore, samples from the relevant population are necessary to estimate model parameters. These samples are referred to as the *reference data*. Once the parameters of the LR have been estimated, it is necessary to evaluate the performance of the LR as the strength of evidence using separate *validation data* that is also sampled from the relevant population. In the following sections, I describe these data sets in more detail and present a technique for evaluating out of sample performance when a limited amount of sample data is available.

2.4.1 Reference Data

The reference data consists of sample of N_s same-source exemplars $\mathcal{D}_s = \{(A_i, B_i)\}$ that were generated under H_s. Similarly, define a reference sample of N_d different-source exemplars $\mathcal{D}_d = \{(A_j, B_k) : j \neq k\}$ that were generated under H_d. These reference data sets are used in the estimation of methods that generate likelihood ratio values.

For the digital event data discussed in Chapters 4 and 5, the reference data sets are constructed from a sample of N individuals (e.g., users of mobile devices) from the relevant population. Let $E_i = (A_i, B_i)$ for $i = 1, \ldots, N$ be the sampled values of the evidence where A_i and B_i in each pair are from the same source and each pair E_i is from a different individual i. We construct a reference data set of all N^2 pairwise combinations of evidence, denoted $\mathcal{D} \equiv \{(A_j, B_k) : j, k \in \{1, \ldots, N\}\}$, where each of the N samples of A are paired up with each of the N samples of B.

Given a new observed evidence $E^* = (A^*, B^*)$ that is not from $\mathcal{D}$,[1] we can use the scores of all of the N same-source pairs, $\mathcal{D}_s = \{(A_i, B_i) : i = 1, \ldots, N\}$, to estimate the same-source likelihood, and the scores of all $N^2 - N$ pairs with different sources, $\mathcal{D}_d = \{(A_j, B_k) : j, k \in \{1, \ldots, N\}, j \neq k\}$, to estimate the different-source likelihood. Given that $N_d > N_s$, we may want to perform subsampling or stratification of the different-source reference data to improve class balance. Examples of this approach will be presented in the case studies of Chapters 4 and 5.

2.4.2 Validation Data

Similar to the reference data, the validation data consists of sample of N_s^* same-source exemplars $\mathcal{D}_s^* = \{(A_i, B_i)\}$ that were generated under H_s and a sample of N_d^* different-

[1] The * notation should be interpreted as referring to data that is not part of the reference data set used for estimation. This will be consistent throughout the remainder of the thesis.

source exemplars $\mathcal{D}_d^* = \{(A_j, B_k) : j \neq k\}$ that were generated under H_d. These validation data sets are used in the evaluation of methods that generate likelihood ratio values.

In many scenarios, including for digital event data, obtaining a large enough sample from the relevant population for creating both the reference and validation sets is difficult. If this is the case, the reference data must be partitioned (e.g., via cross-validation) in a way that allows simulation of in-sample data for estimation and out-of-sample data for validation. One such technique is presented in the following section.

2.4.3 Leave-pairs-out Cross-validation

To evaluate the out-of-sample performance of the evidence analysis method we use *leave-pairs-out cross-validation* to estimate the LR for every pairwise combination available in $\mathcal{D}$. Let $(A^*, B^*) = (A_\ell, B_m)$ be an arbitrary pair from $\mathcal{D}$, where ℓ and m may or may not be equal. Given (A_ℓ, B_m) let $\mathcal{D}_s = \{(A_j, B_j) : j \in \{1, \ldots, N\} \setminus \{\ell, m\}\}$ and $\mathcal{D}_d = \{(A_j, B_k) : j, k \in \{1, \ldots, N\} \setminus \{\ell, m\}, j \neq k\}$ compose the reference data. Essentially, we remove any pair from the reference data with a piece of evidence from either individual whose evidence is currently being evaluated. This process is then repeated for all pairs of evidence in $\mathcal{D}$, with each pair being treated as the validation data one time (i.e., the validation data is a single evidential sample). The resulting LR values can then be thought of as coming from a validation data set and can be used to assess the performance of the method using techniques discussed in the following sections.

2.5 Empirical Classification Performance

Measuring the performance of likelihood ratio methods is central to validate their forensic applicability prior to use in case work [Haraksim et al., 2015]. At the source level, this typi-

cally consists of using sample data to evaluate empirical performance. Therefore, assessing the classification performance of the LR relies on having evidence exemplars under both the same- and different-source hypotheses as described in Section 2.4. Here, we assume that each set of evidence in the validation data sets $\mathcal{D}_s^*$ and $\mathcal{D}_d^*$ have a LR value that has already been computed. Namely, the parameters of the conditional distributions in the LR have already been determined (e.g., via estimation on a reference set), and, given these parameters, the resulting conditional distributions can be used to compute LR values on new out-of-sample validation data $\mathcal{D}^*$.

The estimated likelihood ratio values can be ranked and thresholded to obtain binary decisions of same- or different-source, and these binary decisions can then be compared to the known ground truth to compute error rates. In this context a false positive (Type I error) occurs when $LR > 1$ for evidence known to be from different sources (i.e., $E \in \mathcal{D}_d^*$), and a false negative (Type II error) occurs when $LR < 1$ for evidence known to be from the same source (i.e., $E \in \mathcal{D}_s^*$).

The threshold can then be varied to achieve different trade-offs in terms of sensitivity and specificity. The area under the receiver operating characteristic (ROC) curve, abbreviated as AUC, can be used to summarize this trade-off. The ROC curve measures the true positive rate as a function of the false positive rate by varying the threshold used in classification. AUC is a measure of goodness of fit and can be thought of as the probability that the method will result in a larger LR for a randomly chosen same-source pair than for a randomly chosen different-source pair [e.g., Fawcett, 2006; Krzanowski and Hand, 2009]. Higher AUC values are indicative of better classification performance.

A benefit of using classification performance measures like error rates and AUC is that the likelihood ratio method can be compared to any arbitrary classifier, i.e., any other method that maps the observed features of the evidence to a decision on whether or not the evidence was from the same source or different sources (in Section 3.2 we will present one

such method). However, it should be noted that in the typical Bayesian decision theoretic framework decisions in favor of either hypothesis H_s or H_d are made using the posterior probabilities $Pr(H_s|E, I)$ and $Pr(H_d|E, I)$. This implies that the prior odds are known, which is not the case for the forensic examiner. Error rates and AUC only consider the LR values and ignore the prior odds and, therefore, the posterior probabilities. As a consequence, using these measures to assess performance do not represent the likelihood ratio's value in the Bayesian assessment of the evidence.

2.6 Information-theoretic Evaluation

Information theory studies the quantification, storage and communication of information. First proposed in the mid-twentieth century by Shannon in the area of communicating information over a noisy channel [Shannon, 1948], the field of information theory has since been applied in many disciplines, including physics, statistics, economics and computer science [Cover and Thomas, 2012]. In general, the concept of information is quite broad, but in the context of probability theory the *entropy* of a probability distribution has many properties that align with the intuitive notion of what a measure of information should be. Entropy is a measure of the uncertainty of a random variable. Additional knowledge about a random variable reduces its entropy and, therefore, the information about the variable increases.

Information-theoretic methods have been proposed to assess the performance of the likelihood ratio in the forensic analysis of evidence [e.g., Brümmer and du Preez, 2006; Ramos, 2007]. These approaches measure the reduction of the uncertainty about the same- and different-source propositions that the evidence evaluation (in the form of LRs) provides. Intuitively, the more information the evidence provides, the more the uncertainty about the source proposition is reduced and the less additional information is needed by the decision-maker to correctly determine which source proposition is true.

In this section, a measure of the information-theoretic value of the likelihood ratio based on entropy and divergence is presented. First introduced in Ramos [2007], the empirical cross-entropy (ECE) measures the accuracy of the LR for a given value of the prior probabilities of the source propositions in terms of average information loss. The notation and derivations are based on Ramos [2007].

2.6.1 Uncertainty and Information

The amount of information obtained from an inferential process is determined by the reduction in entropy of the variable of interest [Cover and Thomas, 2012]. In the forensic setting, the entropy represents the uncertainty about the true value of the source proposition. Let the binary random variable H_s represent the true value of the source hypotheses, such that $H_s = 1$ when the same-source hypothesis is true and $H_s = 0$ when the different-source hypothesis is true. Therefore, $H_s = 0$ is semantically equivalent to H_d. Further, let P and p represent probabilities and probability density functions, respectively, obtained from the evidence evaluation method.[2] Therefore, $P(H_s)$ is the prior probability that the same-source hypothesis is true. Here, we assume that the evidence E is in the form of a continuous measurement and that p denotes its density.

Prior to the introduction of the evidence, the uncertainty about the source hypotheses is only conditioned on the background information I.[3] This quantity is known as the *prior*

[2]The reasoning for the change in notation from Pr to P will become evident with the introduction of cross-entropy later in this section. Joint distributions will always be referred to with P, regardless of whether both variables are continuous, e.g., $P(E, H_s) = p(e|H_s)P(H_s)$.

[3]For the remainder of the derivations in this section, we drop the conditioning on I for simplicity, but every probability presented is conditioned on this background information.

entropy[4] [Cover and Thomas, 2012]

$$\begin{aligned}\mathcal{U}_P(H_s) &= -\mathbb{E}_{P(H_s)} \log P(H_s) \\ &= -P(H_s) \log P(H_s) - (1 - P(H_s)) \log(1 - P(H_s))\end{aligned} \tag{2.4}$$

where the base of the logarithm is arbitrary, with base 2 resulting in Shannon entropy (units in bits) [Shannon, 1948]. The graph of $\mathcal{U}_P(H_s)$ as a function of $P(H_s)$ is shown in Figure 2.1. The entropy is a concave function of the prior probability $P(H_s)$. The entropy equals 0 when the prior probability is either 0 or 1, which makes intuitive sense as the value of H_s is not random in that scenario and there is no uncertainty about its value. Similarly, the uncertainty, and therefore the entropy, is maximized when $P(H_s) = 0.5$.

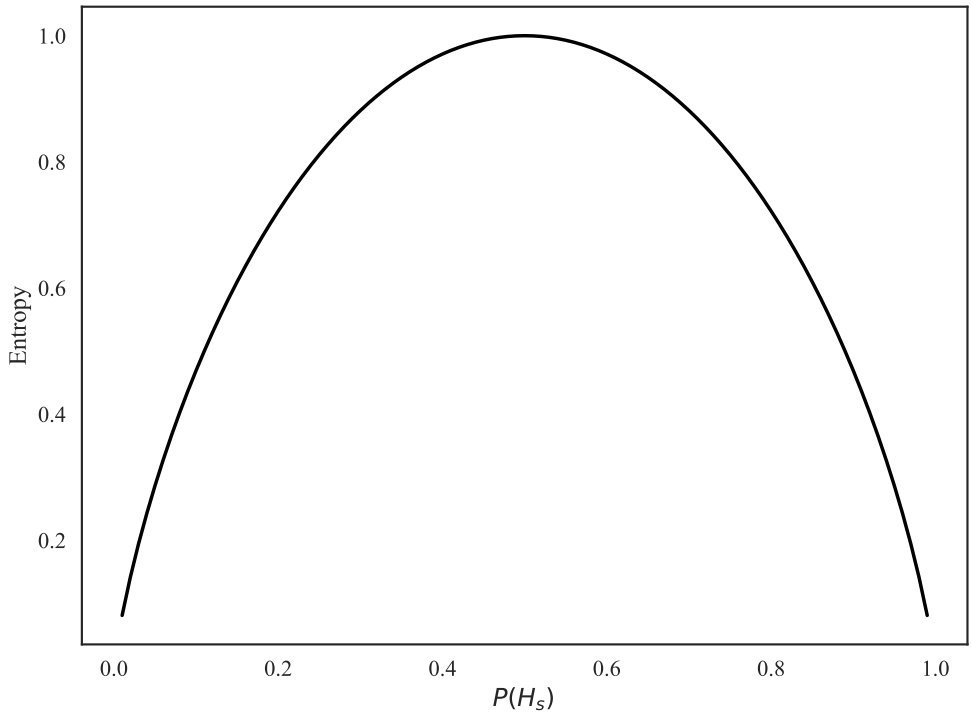


Figure 2.1: Prior entropy $\mathcal{U}_P(H_s)$ (logarithm base 2) as a function of the prior probability $P(H_s)$.

After analyzing the evidence E, the forensic examiner produces a likelihood ratio that can

[4]Entropy is typically denoted with H. Due to the source hypothesis being denoted by H_s and H_d, we will represent entropy with $\mathcal{U}$, which refers to uncertainty—the quantity that entropy measures.

be used to obtain a posterior probability for a given prior probability. The entropy of the posterior probability, or the *posterior entropy*, can then be computed via

$$\begin{aligned}\mathcal{U}_P(H_s|E) &= -\,\mathbb{E}_{P(E,H_s)} \log P(H_s|E) \\ &= -\sum_{k=0}^{1} P(H_s = k) \int p(e|H_s = k) \log P(H_s = k|e) de\end{aligned} \tag{2.5}$$

where the value of the evidence $E = e$ is integrated over its entire domain [Cover and Thomas, 2012, page 17]. The value of a specific piece of evidence e may or may not reduce uncertainty, but on average (with respect to the distribution of the evidence E) the posterior entropy will be lower than the marginal entropy (or equal to it if E is independent of H_s), implying that $\mathcal{U}_P(H_s|E) \leq \mathcal{U}_P(H_s)$ [Cover and Thomas, 2012, Theorem 2.6.5].

Computing the posterior entropy of Equation 2.5 is usually not practical, however, as it requires probability models $p(E|H_s = 1)$ and $p(E|H_s = 0)$ for the evidence conditional on the true value of the hypotheses. As discussed in Section 2.3.2, specifying such models can be difficult if not impossible. Even if the models could be specified, we would then need to integrate over every possible value of the evidence, which introduces a new set of challenges especially if the features of the evidence are high-dimensional.

A solution to this problem is to compare the posterior probabilities computed using the LR (and an assumed prior) with a *target posterior distribution* [Ramos, 2007; Brümmer, 2010]. Let Q and q represent probabilities and probability density functions, respectively, for the target distribution. The target probability $Q(H_s|E)$ represents a desired value of the posterior, compared to the actual posterior probability $P(H_s|E)$ computed from the LR and prior. For instance, one reasonable choice for the target posterior distribution is a Dirac delta function on the true value of the source hypothesis, i.e., $Q(H_s|E) = 1$ if H_s is true and 0 otherwise. A detailed discussion of the target distribution is presented in Section 2.6.2.

Instead of the posterior entropy of the forensic evaluation $\mathcal{U}_P(H_s|E)$, we can consider the

posterior cross-entropy

$$\begin{aligned}\mathcal{U}_{Q||P}(H_s|E) &= -\,\mathbb{E}_{Q(E,H_s)} \log P(H_s|E) \\ &= -\sum_{k=0}^{1} Q(H_s = k) \int q(e|H_s = k) \log P(H_s = k|e) de.\end{aligned} \tag{2.6}$$

Cover and Thomas [2012] showed that the cross-entropy can be decomposed in the following way

$$\mathcal{U}_{Q||P}(H_s|E) = \mathcal{U}_Q(H_s|E) + D_{Q||P}(H_s|E) \tag{2.7}$$

where $D_{Q||P}(H_s|E)$ is the Kullback-Leibler (KL) divergence between the target posterior distribution and the actual posterior distribution obtained from the forensic analysis of the evidence. The KL divergence is defined as

$$D_{Q||P}(H_s|E) = \sum_{k=0}^{1} Q(H_s = k) \int q(e|H_s = k) \log \frac{Q(H_s = k|e)}{P(H_s = k|e)} de. \tag{2.8}$$

It should be noted that the KL divergence is not symmetric in its arguments, i.e., $D_{Q||P}(H_s|E) \neq D_{P||Q}(H_s|E))$. However, the correct version is used in Equation 2.8 since we want to measure the information lost when approximating the target with the actual posterior, not the other way round. Thus, the cross-entropy is the sum of the posterior entropy of the target distribution, $\mathcal{U}_Q(H_s|E)$, and the deviation between the target posterior Q and the actual posterior P, $D_{Q||P}(H_s|E)$. This second term is additional information loss incurred from using Q and not P in the calculation. If we carefully select the target distribution many attractive properties of the decomposition in Equation 2.7 arise, as shown in the following section.

2.6.2 Choosing the Target Distribution

The target distribution Q must be carefully selected so that the information-theoretic value of the LR has an intuitive interpretation in the context of a forensic analysis [Ramos, 2007]. The prior probability of the hypothesis variable is assumed to be a parameter of the analysis because the forensic examiner does not supply this value (in the following sections we will show that the information-theoretic value of the evidence should be computed over a range of prior probabilities). Therefore, we are free to choose any prior probability for the target distribution, so let $Q(H_s) = P(H_s)$ and $\mathcal{U}_Q(H_s) = \mathcal{U}_P(H_s)$.

We now must choose a target posterior distribution of the hypothesis variable given the evidence. If the decision-maker already knew the true value of the hypothesis variable H_s, he or she would always obtain the following oracle posterior probabilities

$$Q(H_s|E) = \begin{cases} 1, & \text{if the same-source hypothesis } H_s \text{ is true} \\ 0, & \text{if the different-source hypothesis } H_d \text{ is true.} \end{cases} \tag{2.9}$$

Using the oracle distribution as the target posterior results in several attractive properties. First, the entropy of the oracle posterior is zero, $\mathcal{U}_Q(H_s|E) = 0$, and therefore the cross-entropy is equal to the KL-divergence between the posterior distribution produced by the forensic analysis with respect to the oracle posterior, i.e., $\mathcal{U}_{Q||P}(H_s|E) = D_{Q||P}(H_s|E)$. Second, the oracle posterior yields a simple interpretation. The larger the cross-entropy, the more information the decision-maker needs in order to know the true value of the hypotheses. If the LR is misleading (i.e., it favors the incorrect hypothesis), then the cross-entropy will grow as will the information needed to obtain the true value of the hypothesis.

Using the oracle distribution as the target posterior also provides another interpretation of the cross-entropy in terms of decision theory, as I show in the following section.

2.6.3 Strictly Proper Scoring Rules & Bayes Risk

The notion of assessing the quality of posterior probabilities is a familiar one in both the statistics and machine learning literature [e.g., DeGroot and Fienberg, 1983; Gneiting and Raftery, 2007; Niculescu-Mizil and Caruana, 2005]. A key concept in this area is that of *scoring rules*, which provide summary measures for the evaluation of probabilistic predictions by assigning numerical scores based on the predictive distribution and the observed value [Gneiting and Raftery, 2007]. One such strictly proper scoring rule that was first introduced by Good [1952], and that has been applied in the forensic context by Ramos et al. [2013] is the logarithmic scoring rule. For each observation $E = e$ in the forensic setting, the logarithmic scoring rule is defined as

$$\begin{aligned} &H_s \text{ true: } -\log P(H_s|E=e) \\ &H_d \text{ true: } -\log(1-P(H_s|E=e)). \end{aligned} \tag{2.10}$$

Note that the base of the logarithm is a scaling factor that does not impact any derivations that follow [Ramos et al., 2013].

Gneiting and Raftery [2007] showed that in estimation problems, strictly proper scoring rules provide loss functions that can be tailored to scientific problems (a special case of which is maximum likelihood estimation). Thus, the logarithmic scoring rule in Equation 2.10 can also be thought of as a loss function that assigns a penalty to the value of the posterior probability. For instance, if the posterior probability of the same-source hypothesis given the observed evidence $E = e$ is high, but the evidence was in fact generated by different sources, then the logarithmic scoring rule would assign a high penalty. See Figure 2.2 for the loss function defined by Equation 2.10.

The *loss function* implied by the logarithmic scoring rule of Equation 2.10 for a single piece

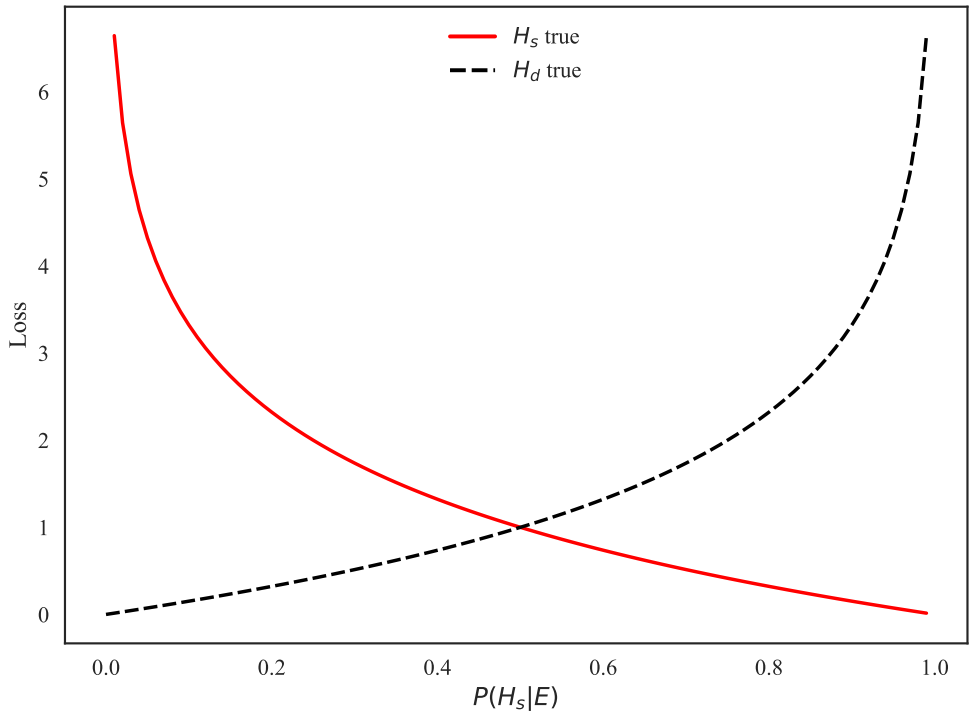


Figure 2.2: Logarithmic scoring rule (base 2) as a loss function.

of evidence $E = e$ can be expressed as

$$L\left[Q(H_s|e), P(H_s|e)\right] = -Q(H_s|e)\log P(H_s|e) - (1 - Q(H_s|e))\log(1 - P(H_s|e)) \quad (2.11)$$

where the target distribution Q is chosen as Equation 2.9, i.e., it assigns probability one to the correct hypothesis. The loss L measures the cost of estimating the true distribution $Q(H_s|E = e)$ with the estimator $P(H_s|E = e)$ for an observed sample of evidence e. We can then define the *expected loss*, or *risk*, by taking the expectation of the loss function with respect to the sampling distribution of the estimator, namely,

$$R(Q, P) = \mathbb{E}_{q(E|H_s)}\, L\left[Q(H_s|E), P(H_s|E)\right] = \int q(e|H_s) L\left[Q(H_s|e), P(H_s|e)\right] de. \quad (2.12)$$

From the frequentist perspective, the risk is averaged over e (thus ignoring the observed data) and is conditioned on the true value of the variable H_s (which is unknown). To choose

amongst estimators, therefore, we need to convert the risk $R(Q, P)$ into a measure of quality that does not depend on knowing H_s. One approach is to put a prior Q on H_s and integrate the risk with respect to said prior. This yields the *Bayes risk*, defined as

$$R_B(P) = \mathbb{E}_{Q(H_s)}\, R(Q, P) = \sum_{k=0}^{1} Q(H_s = k) R(Q, P). \tag{2.13}$$

Bayes risk is equivalent to the cross-entropy of Equation 2.6:

$$\begin{aligned} R_B(P) &= \sum_{k=0}^{1} Q(H_s = k)\, \mathbb{E}_{q(E|H_s=k)}\, L\left[Q(H_s|e), P(H_s|e)\right] \\ &= -\sum_{k=0}^{1} Q(H_s = k) \int q(e|H_s = k) \log P(H_s = k|e) de \\ &= \mathcal{U}_{Q||P}(\theta|E) \end{aligned} \tag{2.14}$$

The goal of frequentist approaches (e.g., classification, regression and parameter estimation) is to choose an estimator that minimizes Bayes risk (perhaps subject to some regularization constraints). Such an estimator is referred to as a *Bayes estimator* or *Bayes decision rule*.[5] However, frequentist decision theory suffers from the fundamental problem that one cannot actually compute the risk function, since it relies on knowing the true data distribution. The usual approach is to minimize the *empirical risk*, or the average value of the loss function, given a sample of n observed values of evidence e whose target distribution $Q(H_s|e)$ is known via

$$\widehat{R}_n(P) = \frac{1}{n} \sum_{i=1}^{n} L\left[Q(H_s|e_i), P(H_s|e_i)\right]. \tag{2.15}$$

Thus, the overall performance of the method used to generate a predictive distribution $P(H_s|E)$ is given by the average value of a strictly proper scoring rule over many different

[5]It has been shown that a Bayes estimator can be found by minimizing the *posterior expected loss* or *posterior risk* for each value of the evidence e, and, therefore, the frequentist and Bayesian decision-theoretic problem formulations are equivalent [Berger, 2013].

predictions for which the ground truth is known [Gneiting and Raftery, 2007].

In the forensic setting, the validation data $\mathcal{D}_s^*$ and $\mathcal{D}_d^*$ of known same- and different-source evidence can therefore be used to assess the utility of the LR for generating posterior probabilities via the empirical logarithmic loss[6]

$$\widehat{L} = -\frac{1}{N_s^* + N_d^*} \left(\sum_{i \in \mathcal{D}_s^*} \log P(H_s|E = e_i) + \sum_{j \in \mathcal{D}_d^*} \log(1 - P(H_s|E = e_j)) \right) \tag{2.16}$$

where e_i and e_j denote the observed values for same- and different-source evidence, respectively. In order to assess this loss function, however, one must know the prior probability of H_s. This becomes further evident when observing that the posterior probabilities can be expressed in terms of the likelihood ratio and prior probabilities via the following

$$\begin{aligned} P(H_s|E) &= \frac{LR \times \frac{P(H_s)}{P(H_d)}}{1 + LR \times \frac{P(H_s)}{P(H_d)}} \\ P(H_d|E) \equiv 1 - P(H_s|E)) &= \frac{1}{1 + LR \times \frac{P(H_s)}{P(H_d)}} \end{aligned} \tag{2.17}$$

where $P(H_d) \equiv 1 - P(H_s)$ for simplicity in notation. Thus, the empirical logarithmic loss can be expressed as

$$\widehat{L} = \frac{1}{N_s^* + N_d^*} \left(\sum_{i \in \mathcal{D}_s^*} \log \left(1 + \frac{1}{LR_i \times \frac{P(H_s)}{P(H_d)}} \right) + \sum_{j \in \mathcal{D}_d^*} \log \left(1 + LR_j \times \frac{P(H_s)}{P(H_d)} \right) \right) \tag{2.18}$$

where LR_i and LR_j denote the LR values for same- and different-source evidence, respectively. Ideally, we would like $\widehat{L}$ to be minimized on the validation data for some specified prior probability.

[6]The logarithmic loss L is in effect the same loss function used in training and evaluating binary classifiers in machine learning and statistics, and is equivalent to the conditional log-likelihood for predicting binary outcomes.

2.6.4 Empirical Cross-Entropy (ECE)

The cross-entropy (or Bayes risk) may be difficult to compute due to the integration over the evidence in the risk $\mathbb{E}_{q(E|H_s)}\, L\left[Q(H_s|e), P(H_s|e)\right]$. Therefore, an empirical procedure was presented to approximate its value by Ramos [2007]. Given validation samples of same- and different-source evidence $\mathcal{D}_s^*$ and $\mathcal{D}_d^*$, respectively, along with their corresponding LR values, we can obtain posterior probabilities using Equation 2.17 for an assumed value of the prior probabilities. The risk can then be estimated by averaging over the corresponding sample evidence, namely,

$$\begin{aligned}\mathbb{E}_{q(E|H_s=k)}\, L\left[Q(H_s|e), P(H_s|e)\right] &= \int q(e|H_s=k)\log P(H_s=k|e)de \\ &\approx \frac{1}{N_k^*}\sum_{i\in\mathcal{D}_k^*}\log P(H_s=k|e_i)\end{aligned} \tag{2.19}$$

where $k \in \{0,1\}$ represents the different- and same-source hypotheses are true, respectively, and $\mathcal{D}_k^*, N_k^*$ represent the corresponding validation data set and the number of samples in that set (this is a slight abuse of notation that is remedied below). The cross-entropy can then be estimated by substituting the approximations in Equation 2.19 into Equation 2.6, resulting in the *empirical cross-entropy* (ECE)

$$ECE = -\frac{P(H_s)}{N_s^*}\sum_{i\in\mathcal{D}_s^*}\log P(H_s|e_i) - \frac{1-P(H_s)}{N_d^*}\sum_{j\in\mathcal{D}_d^*}\log(1-P(H_s|e_j)) \tag{2.20}$$

where the target distribution presented in Section 2.6.2 is assumed to apply here so that $Q(H_s) = P(H_s)$. By the law of large numbers, $ECE \to \mathcal{U}_{Q||P}(H_s|E)$ as $N_s^*, N_d^* \to \infty$. Therefore, by the equivalence of the cross entropy and Bayes risk presented in Equation 2.14, $ECE \to R_B(P)$ as $N_s^*, N_d^* \to \infty$.

The empirical cross-entropy retains the same information-theoretic interpretation as the cross-entropy. It represents the mean additional information after consideration of the ev-

idence E that the decision maker (e.g., a juror) still needs to know the ground truth (i.e., which hypothesis, H_s or H_d, is true). The mean is computed over the LR values in the validation data $\mathcal{D}_s^*$ and $\mathcal{D}_d^*$. If the LR values for a given evidence evaluation method tend to support the correct hypothesis, then the amount of uncertainty regarding the ground truth—and therefore the ECE—decreases. Conversely, if the LR values are misleading, the amount of information needed to know the ground truth—and therefore the ECE—increases. Thus, small values of ECE are indicative of good performance of an LR method.

The empirical cross-entropy explicitly depends upon the prior probability $P(H_s)$. To make this relationship more explicit, the ECE can be expressed as a function of the likelihood ratio and prior probabilities by applying Equation 2.17 to the posterior probabilities in Equation 2.20

$$\begin{aligned} ECE = &\frac{P(H_s)}{N_s^*} \sum_{i \in \mathcal{D}_s^*} \log \left(1 + \frac{1}{LR_i \times \frac{P(H_s)}{P(H_d)}} \right) \\ &+ \frac{P(H_d)}{N_d^*} \sum_{j \in \mathcal{D}_d^*} \log \left(1 + LR_j \times \frac{P(H_s)}{P(H_d)} \right). \end{aligned} \tag{2.21}$$

Typically, one of two approaches for setting the prior probabilities is taken. The first approach is to choose the prior probability that yields the highest prior entropy and compute a single value for the ECE (see Section 2.6.5). The other approach is to compute the ECE over a range of prior probabilities (see Section 2.6.6).

2.6.5 Log Likelihood Ratio Cost

One solution for assigning prior probabilities that was first proposed in the context of forensic speaker recognition [Brümmer and du Preez, 2006] is to choose the values of the prior probabilities so that the prior entropy is maximized. This choice is well-motivated by the principle of maximum entropy, which states that the probability distribution which best

represents the current state of knowledge is the one with largest entropy [Jaynes, 1957].

Namely, evaluate the empirical cross-entropy for $P(H_s) = P(H_d) = 1/2$. This is referred to as the *log likelihood ratio cost* (C_{llr}) and has been applied in a variety of biometric and forensic settings including speaker recognition [e.g., Ramos et al., 2017; Morrison, 2009], physiochemical trace analysis [Zadora et al., 2013], glass fragment analysis [Ramos et al., 2013], and gasoline residue analysis [Vergeer et al., 2014]. The log likelihood ratio cost is defined as

$$C_{llr} = \frac{1}{2N_s^*} \sum_{i \in \mathcal{D}_s^*} \log\left(1 + \frac{1}{LR_i}\right) + \frac{1}{2N_d^*} \sum_{j \in \mathcal{D}_d^*} \log\left(1 + LR_j\right). \tag{2.22}$$

Lower values of C_{llr} are indicative of better performance. It has been shown that minimizing C_{llr} results in reduced Bayes decision costs for all possible decision costs and prior probabilities [Brümmer and du Preez, 2006].

2.6.6 The ECE Plot

The role of the forensic examiner is not to supply these probabilities but to give an objective numeric summary of the strength of the evidence. One way for the examiner to compare different methods for evidence evaluation is to compute the ECE over a range of prior probabilities and plot them. Figure 2.3 shows an example of such a plot for a case study described in detail in Chapter 4. This plot shows the ECE (using base 2 logarithms) versus the logarithm of the prior odds. The solid red curve is the ECE computed from LR values in the sample data. Each y-value is computed by evaluating Equation 2.21 for the prior odds implied by the corresponding x-value using the likelihood ratio values for the N_s^* same-source samples in $\mathcal{D}_s^*$ and N_d^* different-source samples in $\mathcal{D}_d^*$. The lower this curve, the less information is needed in order to know the ground truth on average for the evidence in the

sample, and, therefore, the better the method performs on the validation data.[7]

The dashed black curve depicts the performance of a neutral method for comparison. The neutral method assumes that the evidence has no value, so that $LR = 1$ and, therefore, the prior and posterior probabilities are equal. If the ECE of the LR values (solid red curve) is above that of the neutral method (dashed black curve), then the method performs worse in that range of prior odds than not using the evidence at all.

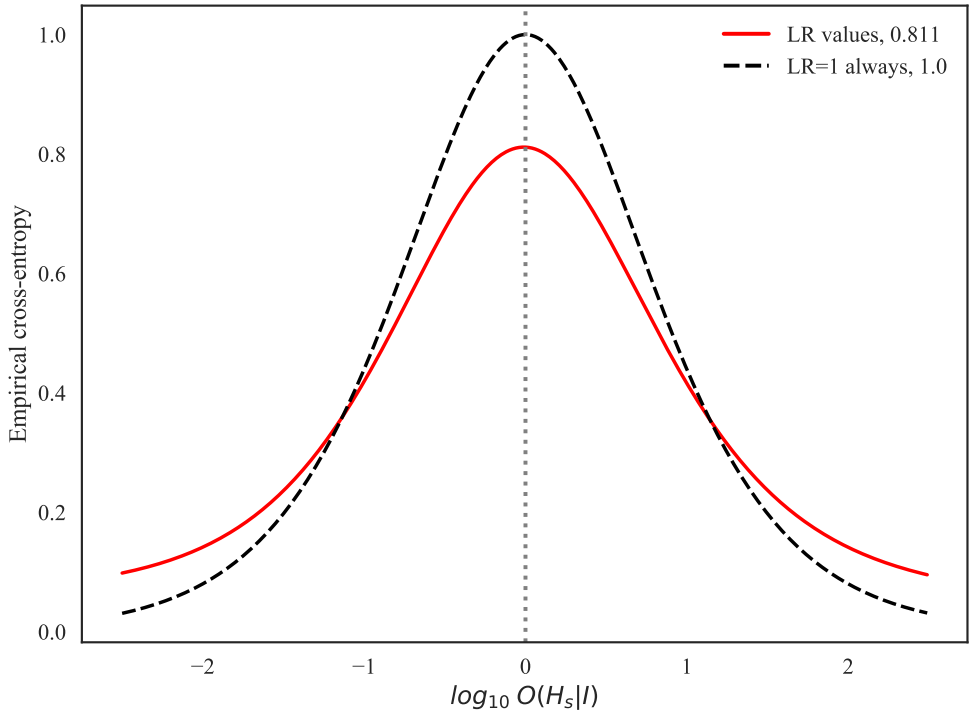


Figure 2.3: Empirical cross-entropy (ECE) plot for case study data from Chapter 4. C_{llr} values are the ECE evaluated at prior log odds of 0 and are given in the legend. Lower values are indicative of better performance on the validation data.

A thorough discussion on the use of ECE plots to compare different evidence evaluation methods and how to determine if a given method is adequate to evaluate the evidence is presented in Ramos et al. [2013]. In general, the most important aspect is in the choice of

[7]Note that the performance is related to both the method for evaluating the evidence and the feature values of the evidence itself. The method could be excellent at assessing the strength of evidence in general, but the validation sample under consideration could be "weak," e.g., the observed features are not distinguishable between same- and different-source pairs. Therefore, the selection of validation data used to evaluate the method is very important—poor selection could result in incorrectly dismissing the method as invalid.

the samples $\mathcal{D}_s^*$ and $\mathcal{D}_d^*$ used to validate the method. This validation data should come from conditions comparable to the materials in the case under review. Once that has been selected and LR values have been computed using one or more methods, ECE plots are made. If the ECE curve of the LR values falls under that of the neutral method, the method is informative for the decision maker (jury). If not, or if two different methods yield differing curves, then the methods that yields a lower ECE in a "reasonable" range of prior probabilities should be chosen. Since the examiner does not necessarily know what is a reasonable prior, Ramos et al. [2013] suggests presenting both methods and informing the court about the differing performance.

2.7 Empirical Calibration

The performance metrics from the previous sections provide an examiner a way to assess an evidence evaluation method in terms of its classification (or discriminative) ability and the behavior of posterior probabilities as a function of the prior probability. What is not evident, however, is how the method behaves in terms of calibration, which refers to the statistical consistency between the distributional predictions and the known ground truth. Calibration is important if we want the posterior probabilities (and indirectly the LR values for an assumed prior probability) to be interpretable as the chances of the same- or different-source hypothesis to be true.

The main idea in this section is to transform the LR values into posterior probabilities, then calibrate those posterior probabilities using a common technique such as isotonic regression. Given the calibrated posterior probabilities, we then transform them back into calibrated LR values. The results of the calibration can then be visualized by producing ECE plots of the calibrated LR values and comparing them to those obtained by the uncalibrated and neutral approaches. First, we begin by briefly introducing calibration in general before applying it

to the LR in the forensic setting.

2.7.1 Calibration in General

To introduce the notion of calibration, we will discuss it in the context of classification models. More specifically, assume we have a model (a binary classifier) that, given an input x, produces a real-valued score $s(x) \in [0, 1]$ that is an estimate of $Pr(c = 1|x)$ where $c \in \{0, 1\}$ are the possible values of the binary random variable C whose value is being predicted. This classifier is said to be well-calibrated if the empirical class membership probability $Pr(c|s(x) = s)$ converges to the score value $s(x) = s$ as the number of examples classified goes to infinity [Murphy and Winkler, 1977]. In other words, if we consider all observations to which the classifier assigns a score $s(x) = 0.9$, then 90% of these observations should be members of class c.

However, we only have a finite number of observations in practice. To assess the calibration of the generic binary classifier, we can use the scores $s(x)$ to rank the observations from the least to the most probable member of class c. Namely, for two observations x and y, if $s(x) < s(y)$ then $Pr(c|x) < Pr(c|y)$. What is needed to assess the calibration is not simply the ranking, but an accurate estimate of the probability that each observation is a member of class c given the observed value. Suppose we have a set of examples for which we know the ground truth (i.e., the class membership for each observation). Thus, we would like for the resulting probability estimates to be $P(c|x) = 1$ for positive examples, and $P(c|x) = 0$ for negative examples. After applying the classifier to those examples to obtain scores $s(x)$, we can learn a function mapping scores to probability estimates $\widehat{Pr}(c|x)$. These new probability estimates are calibrated to the data set used to generate them. Many methods have been explored to perform this calibration, including parametric techniques like logistic regression

and non-parametric techniques like binning and isotonic regression[8] [Zadrozny and Elkan, 2002].

2.7.2 Using LR Values to Calibrate Posterior Probabilities

We now formulate the calibration problem defined above for the forensic setting. The likelihood ratio value computed for each evidential sample in the validation data can be thought of as a score for that evidence. Similar to the generic calibration problem presented above, we can rank the LR values for each sample in ascending order. Furthermore, we know which samples were from the same source (those in $\mathcal{D}_s^*$) and different sources (those in $\mathcal{D}_d^*$), so that we can assign target values for the posterior probability of H_s given the evidence for each sample. Namely, $Pr(H_s|E \in \mathcal{D}_s^*, I) = 1$ and $Pr(H_s|E \in \mathcal{D}_d^*, I) = 0$. If the likelihood ratio values are well-calibrated, then the ranking will be such that $LR_j < LR_i$ for all $j \in \mathcal{D}_d^*$ and $i \in \mathcal{D}_s^*$. If not, then we wish to obtain new posterior probability estimates $\widehat{Pr}(H_s|E, I)$ that are calibrated to the validation data and use these estimates to calculate calibrated LR values. A common approach to calibrating posterior probabilities for a given set of LR values is the pool adjacent violators (PAV) algorithm for nonparametric isotonic regression [Brümmer and du Preez, 2006; Ramos et al., 2013].

2.7.3 Isotonic Regression & the PAV Algorithm

We begin by formulating the isotonic regression problem in general and then apply it to the forensic setting. In isotonic regression a monotone increasing function is fit to a set of points in the plane [Barlow et al., 1972]. Suppose that $X = \{x_1, x_2, \ldots, x_n\}$ is a finite set of ordered predictors such that $x_1 < x_2 < \cdots < x_n$. Let $Y = \{y_1, y_2, \ldots, y_n\}$ be the observed responses

[8] In general, parametric formulations of isotonic regression exist. Here, we assume that isotonic regression consists of fitting a non-parametric step function to the data.

Algorithm 1 Pool Adjacent Violators (PAV) Algorithm for isotonic regression.

Inputs:

Ordered predictors $X = \{x_1, \ldots, x_n\}$ such that $x_1 < x_2 < \cdots < x_n$
Observed responses $Y = \{y_1, \ldots, y_n\}$
Weights $\{\omega_1, \ldots, \omega_n\}$

Output:

Fitted responses $\widehat{Y} = \{\widehat{y}_1, \ldots, \widehat{y}_n\}$ such that $\widehat{y}_1 \leq \widehat{y}_2 \leq \cdots \leq \widehat{y}_n$

Initialize iteration counter $\ell = 0$
Initialize solution $a_i^{(0)} = y_i$ and $\omega_i^{(0)} = \omega_i$ for all i
Initialize block index $r = 1, \ldots, B$ such that for the first iteration $\ell = 0$, $B = n$ so each observation $a_r^{(0)}$ forms a block
while $a_{r+1}^{(\ell)} \leq a_r^{(\ell)}$ for all r **do**
    Adjacent pooling: merge $\{a_{r+1}^{(\ell)}, a_r^{(\ell)}\}$ values into blocks if $a_{r+1}^{(\ell)} < a_r^{(\ell)}$ and update B
    for $r = 1$ to B **do**
        Update $a_r^{(\ell+1)} = \frac{\omega_r^{(\ell)} a_r^{(\ell)} + \omega_{r+1}^{(\ell)} a_{r+1}^{(\ell)}}{\omega_r^{(\ell)} + \omega_{r+1}^{(\ell)}}$
        Update $\omega_r^{(\ell+1)} = \omega_r^{(\ell)} + \omega_{r+1}^{(\ell)}$
    end for
    Increment $\ell = \ell + 1$
end while
Expand the block values to all observations within the block to obtain $\widehat{Y}$

that preserve the ordering determined by X, and $\widehat{Y}$ be the unknown responses to be fitted. The least squares problem for nonparametric isotonic regression can be expressed as

$$\widehat{Y} = \underset{a \in \mathbb{R}^n}{\arg\min} \sum_{i=1}^{n} \omega_i (y_i - a_i)^2 \quad \text{subject to} \quad \widehat{y}_1 \leq \widehat{y}_2 \leq \cdots \leq \widehat{y}_n \tag{2.23}$$

where ω_i for $i = 1, \ldots, n$ are optional weights. The monotone increasing restriction on both the predictors X and fitted values $\widehat{Y}$ restrict us from obtaining the best fit obtainable from traditional least squares regression. A simple iterative approach for solving the least squares problem in Equation 2.23 is the pool adjacent violators (PAV) algorithm, which fits a step function to the data [Barlow et al., 1972]. The PAV algorithm can be considered as a dual active set method whose computational complexity is $\mathcal{O}(n)$ [Best and Chakravarti, 1990]. See Algorithm 1 for the general representation of the PAV algorithm.

2.7.4 Obtaining Calibrated LR Values

To calibrate likelihood ratio values, first apply the PAV algorithm to the validation LR values ranked in ascending order (i.e., X in Algorithm 1) and the known ground truth of the same-source hypothesis as response values (i.e., Y in Algorithm 1). The resulting posterior probabilities, $\widehat{Pr}(H_s|E,I)$, are well-calibrated. These posterior probabilities can then be transformed back to calibrated log likelihood ratio values, denoted as $\log(\widetilde{LR})$, via

$$\begin{aligned}\log(\widetilde{LR}) &= \log(\widehat{O}(H_s|E,I)) - \log(O(H_s|I)) \\ &= \log\left(\frac{\widehat{Pr}(H_s|E,I)}{1-\widehat{Pr}(H_s|E,I)}\right) - \log\left(\frac{N_s^*}{N_d^*}\right)\end{aligned} \tag{2.24}$$

where N_s^* and N_d^* are the number of same- and different-source samples of evidence in the validation data set. Calibrated likelihood ratios $\widetilde{LR}$ can then be computed for new cases by using the isotonic regression model to map the original LR values to calibrated posterior probabilities and then using Equation 2.24 to obtain a calibrated LR value for the new case.

2.7.5 Assessing the Calibrated LR Values

The calibrated LR values have the same discriminative power (in terms of classification performance) as their non-calibrated counterparts because the calibrated probabilities, and consequently the calibrated LRs, are monotonic functions of the original probabilities (and LRs) by construction [Ahuja and Orlin, 2001]. One benefit of the calibrated LR is that it allows the examiner to compare the performance of the evidence evaluation method in a manner that decomposes the performance in terms of discriminative power and calibration loss.

Given the calibrated LR values, we can then recompute the ECE (and therefore C_{llr}) over a range of prior values. Brümmer and du Preez [2006] showed that the PAV algorithm

decomposes the empirical cross-entropy to components due to discrimination and calibration costs:

$$ECE = ECE^{min} + ECE^{cal}. \tag{2.25}$$

ECE^{min} represents the discrimination loss of the method, and is obtained from the calibrated LR values (i.e., the LR after applying PAV optimization). ECE^{min} is the lowest ECE obtainable while preserving the discriminative power of the LR values. On the other hand, ECE^{cal} represents the calibration loss of the method. The lower the calibration loss, the better the method performs. Ideally, an evidence evaluation method that produces LR values will have the smallest ECE^{cal} possible.

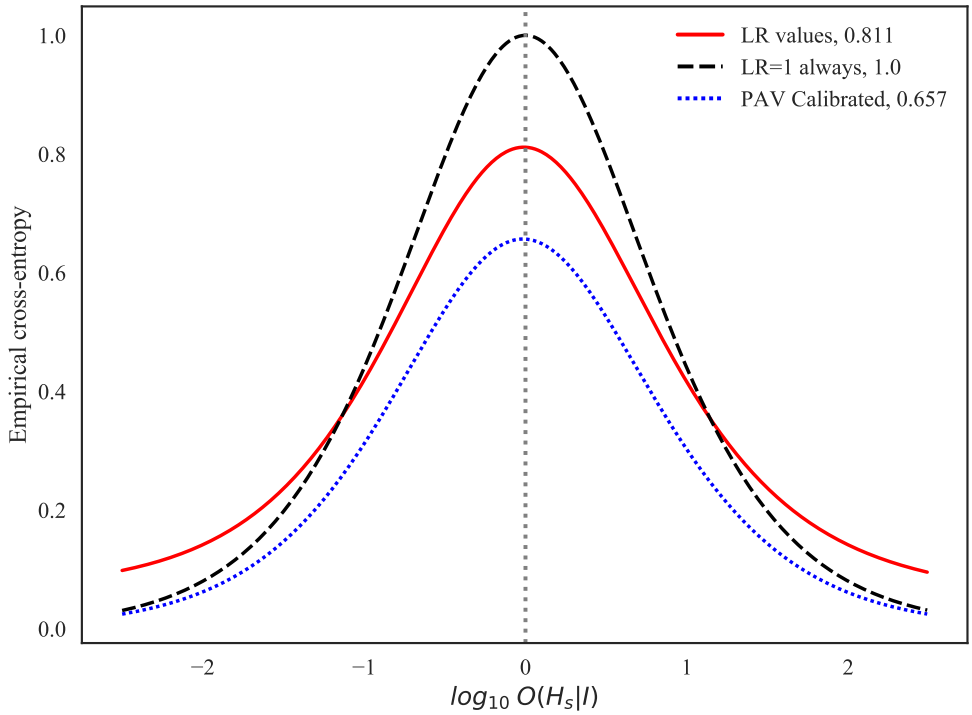


Figure 2.4: Empirical cross-entropy (ECE) plot of Figure 2.3 including the PAV calibrated LR values. C_{llr} values are the ECE evaluated at prior log odds of 0 and are given in the legend. Lower values are indicative of better performance on the validation data.

It is now possible to introduce the full ECE plot in Figure 2.4. This plot is the same as that of Figure 2.3 but with the ECE derived from the PAV calibrated LR values shown as a

blue dotted curve that shows the performance of a method that has the same discriminating power as the original one (red curve), but optimizes its calibration to the validation data. For any given value of the prior odds (x-axis), the height of the PAV calibrated ECE curve (corresponding y-value) represents ECE^{min}, i.e., the component of the ECE arising from the nonperfect discriminating power of the validation set of LR values being analyzed, because the component of ECE due to calibration has been minimized. It can only be obtained if the ground truth is known, and therefore, it represents a ceiling of performance rarely achievable in practice. The difference between the ECE of original LR values (red curve) and the calibrated LR values (blue dotted curve) represents ECE^{cal}, i.e., the calibration loss of the method.

In general, calibrating likelihood ratio values using methods like isotonic regression provides desirable performance measures (i.e., discrimination versus calibration loss) for the method used to generate the LR in a given sample of evidence. However, it should be noted that the calibration property of these updated LR values applies to the sample data used to perform the calibration and may not generalize to new evidence. This is especially true if the new evidence is an outlier of the distribution of samples used for calibration (i.e., drawn from the tails of the distribution of evidence implied under $\mathcal{D}_s$ or $\mathcal{D}_d$), is out of distribution (i.e., the new evidence is from a completely different distribution that than samples used for calibration), or is anomalous (i.e., the new evidence has an entirely different support than the samples used for calibration). Assessing if any of these scenarios apply can be difficult in practice. Therefore, it is typical for the examiner to report the non-calibrated LR value. The decomposition of the loss of information due to discrimination versus calibration is moreso helpful for the comparison of competing methods.

2.8 Discussion

When a crime is committed, the subsequent investigation may identify a variety of evidence that can be used to help identify the perpetrator. In general, this evidence falls into one of three categories—biological evidence, trace evidence, or pattern evidence. In any case, the goal of the forensic evaluation is to identify the source of the evidence discovered at the crime scene. The direct approach to addressing the goal of the forensic analysis involves modeling the observed features of the evidence under two competing propositions—that the evidence from the scene of the crime was from the same source as that of a sample from a suspect, or that it was generated by some alternate source from a relevant population. The likelihood ratio is widely accepted in the forensic science community as "a logically defensible way" to assess the strength of evidence in this setting [Willis et al., 2016].

Methods for estimating the likelihood ratio from the observed evidence using reference samples of evidence from a relevant population were presented along with a variety of approaches for evaluating its performance for a particular type of evidence—including classification accuracy, information-theoretical value, and calibration. These evaluation techniques can help a forensic investigator determine which method of computing the likelihood ratio offers the most value to the decision-maker (i.e., the judge or jury) so they can come to the correct conclusion regarding the source of the evidence in question.

The applicability of the LR for a particular type of evidence, however, varies greatly. While biological DNA evidence is the gold standard for the LR, for other types of evidence such as pattern evidence (including digital evidence) obtaining a LR can be difficult if not impossible. In the following chapter, an alternative approach to quantifying the strength of evidence is presented. Instead of modeling the observed features of the evidence directly, this alternative approach relies on modeling the similarity between the crime scene and suspect samples and is applicable in many more settings than the LR approach.

2.8.1 Contributions

The following contributions were made in this chapter:

- A survey of likelihood ratio techniques for the analysis of forensic evidence that unifies the forensics, biometrics and statistics literature on the topic.
- A concise discussion of the methods to evaluate the LR via classification performance, information-theoretic value (and its relationship to frequentist decision theory), and calibration. This portion unifies concepts and notation from previous literature on the topics.

Chapter 3

Score-Based Approaches for Computing the Strength of Evidence

For many types of forensic evidence, obtaining the likelihood ratio has proven difficult if not impossible. It is rare that the underlying process that generates A and B is sufficiently understood to make distributional assumptions. In some cases, the evidence is assumed to follow a distribution from a common family of distributions [e.g., Aitken and Lucy, 2004; Bolck et al., 2015; Vergeer et al., 2014]. For example, in some forensic evaluations of the elemental composition of glass fragments an assumption that A and B are normally distributed is often made [e.g., Aitken and Lucy, 2004], but this is not always reasonable Garvin and Koons [2011]. Even for the most basic forms of DNA evidence—often considered the gold standard for likelihood ratio computation—there were many years of research before reasonable distributional assumptions could be made to an adequate degree of certainty that is required in the courts Evett and Weir [1998]. For digital evidence, such simple distributional assumptions are not even plausible [Galbraith et al., 2020a,b]. Knowing (or reasonably assuming) the distribution of the evidence is only the first challenge, however. True distributional parameters are rarely known and must be estimated, which can be difficult for complex and

often high dimensional measurements arising in pattern evidence including digital evidence and handwriting evidence [e.g., Hepler et al., 2012].

3.1 The Score-based Likelihood Ratio

Instead of specifying a generative model for the observed data, an alternative approach is to instead measure the similarity between unknown source data A (i.e., evidence recovered from the crime scene) and known source data B (i.e., evidence taken from the suspect) via a *score function* $\Delta(A, B)$ that is usually univariate and continuous. Typically, low scores indicate the samples are similar, while high scores indicate considerable differences. If the score function can capture similarities (or differences) between the two samples, then the dimensionality of the problem is reduced from the dimensionality of observed features of the evidence to the dimensionality of the score (typically 1-dimensional). However, the examiner must still transform the scores into an assessment of the strength of evidence.

A natural approach to assess the strength of evidence via score functions is the *score-based likelihood ratio (SLR)*, which has been gaining popularity in forensic science [e.g., Bolck et al., 2015; Meuwly et al., 2017; Galbraith et al., 2020a]. Given an observed set of evidence (A, B) related to a forensic investigation and a score function to assess the similarity of that type of evidence $\Delta(A, B)$, the SLR arises naturally in the application of Bayes' Theorem to updating the relative likelihoods (odds) of the two competing hypotheses given the observed score $\Delta(A, B) = \delta$. Bayes' Theorem in this context is

$$\underbrace{\frac{Pr(H_s|\Delta(A,B)=\delta, I)}{Pr(H_d|\Delta(A,B)=\delta, I)}}_{\text{posterior odds}} = \overbrace{\frac{Pr(\Delta(A,B)=\delta|H_s, I)}{Pr(\Delta(A,B)=\delta|H_d, I)}}^{\text{score-based likelihood ratio}} \underbrace{\frac{Pr(H_s|I)}{Pr(H_d|I)}}_{\text{prior odds}}. \tag{3.1}$$

Since the score function Δ under consideration is typically assumed to be a continuous,

1-dimensional summary of the similarity between the observed evidence, the SLR can be expressed as

$$SLR_{\Delta} = \frac{g(\Delta(A,B) = \delta | H_s, I)}{g(\Delta(A,B) = \delta | H_d, I)} \tag{3.2}$$

where g denotes the conditional probability density function of $\Delta(A, B)$ given one of the two propositions (H_s for same-source or H_d for different-source). The subscript Δ in SLR_{Δ} is present to distinguish which score function was used to produce the SLR. The definition of the SLR parallels that of the likelihood ratio of Equation 2.1, but with the relevant distributions over the value of the score function instead of the evidence. See Figure 3.1 for an example of how the SLR approach might be applied.

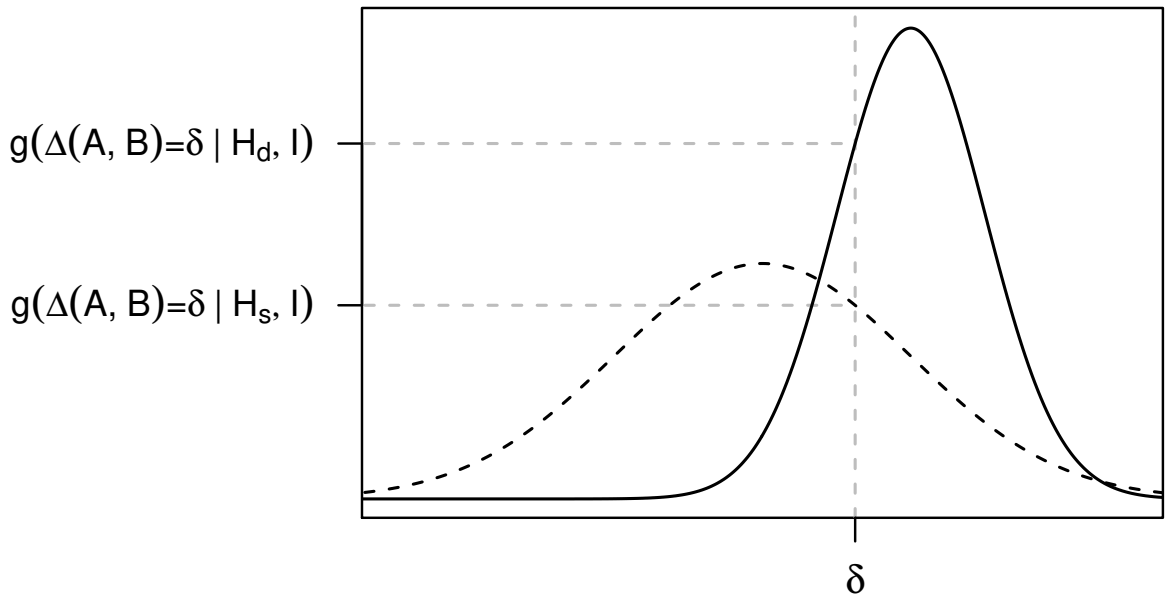


Figure 3.1: Hypothetical illustration of the conditional densities of the score function Δ under the hypotheses that the samples are from the same source (H_s, dashed line) and that the samples are from different sources (H_d, solid line). The score-based likelihood ratio SLR_{Δ} is the ratio of the conditional density functions g evaluated at δ.

Following the specific source proposition discussed in Section 2.2.1, the numerator of the SLR in Equation 3.2 can be interpreted as the likelihood of observing the score $\Delta(A, B) = \delta$

if A and B came from the same source. The interpretation of the denominator is the likelihood of observing the score $\Delta(A, B) = \delta$ if the unknown source sample A is taken from a randomly selected individual from an alternative population and paired with the known source sample B. This approach is often used in biometrics for example [Ross et al., 2006]. The interpretation of the SLR is similar to that of the LR, with values greater than 1 favoring the same-source proposition.

3.1.1 Choosing an Appropriate Score Function

Scores quantify the degree of similarity between measurements made on pairs of objects, e.g., an unknown source sample A and a known source sample B.[1] Here we will assume that pairs from the same source tend to have smaller scores than pairs from different sources, although all arguments hold if the reverse is true. Determining what score function is appropriate for a particular case largely depends on what type of evidence is under consideration and what features were measured from said evidence. The score function should be able to discriminate between same- and different-source pairs of objects, and the better it is at doing so the better the resulting score-based likelihood ratio will perform in terms of both classification accuracy and calibration.

In some settings, only one measurement is made on each sample A and B. This measurement can be either univariate or multivariate, and the score function chosen is typically some type of distance between the measurements [e.g., Nordgaard and Höglund, 2011; Baechler et al., 2013; Vergeer et al., 2014; Bolck et al., 2015]. In other settings, such as biometric fingerprint or face recognition, output from a commercial software product can be used as scores [e.g., Hepler et al., 2012; Ali et al., 2014; Ramos et al., 2017]. Such commercial software is often treated as a black box in the sense that the techniques used to produce the scores are

[1]Score functions are defined on the *measurements* $x_a, x_b \in \mathbb{R}^m$ from samples A and B, respectively, but for notational simplicity the score functions will be defined on the samples $\Delta(A, B)$ rather than on the corresponding measurements $\Delta(x_a, x_b)$.

unknown and may be protected trade secrets. Finally, measurements made on the samples A and B may be multidimensional with potentially complex multimodal distributions [e.g. Tang and Srihari, 2014; Kong et al., 2019; Galbraith et al., 2020a,b]. In this setting, scores can be considered a projection of the complex distribution in the original feature space onto a simple (often univariate) distribution in the score space. This necessarily implies a loss of information, but simple parametric models with few parameters or nonparametric models in the lower-dimensional space tend to produce much better calibrated likelihood ratio results than complex models requiring many parameter values to be fitted in the original feature space [Morrison and Enzinger, 2018].

It is up to the examiner to choose an appropriate score function for the evidence at hand. Whenever possible, a method that has been peer reviewed should be used. It is also possible to use more than one score function and compare SLR results (via the methods discussed for the LR in Section 2.5) across the different methods on a reference data set of cases similar to the one under consideration.

3.1.2 Estimation

Suppose we are given a particular piece of evidence $E^* = (A^*, B^*)$ and a score function Δ. We wish to assess the strength of evidence between A^* and B^*. To do so, we must consider the likelihood of observing $\Delta(A^*, B^*) = \delta^*$ under two competing hypotheses, namely, that A^* and B^* were generated by the same source, or that they were generated by different sources.[2]

In order to compute the SLR for a particular piece of evidence (A^*, B^*), we need a sample from a reference population as defined in Section 2.4. The reference population data, denoted

[2]Note the slight change in notation for the evidence under consideration. This is to avoid any confusion between the score function evaluated for evidence in question, $\Delta(A^*, B^*) = \delta^*$, and the score function for evidence of the same type, $\Delta(A, B)$, which is a random variable whose distribution arises in the definition of the SLR in Equation 3.2.

$\mathcal{D}_s$ and $\mathcal{D}_d$ for same- and different-source exemplars, is used for parameter estimation for $g(\cdot)$ or directly as the basis for the non-parametric estimate of $g(\cdot)$ as shown below.

Generative Models

The most straightforward way to estimate the score-based likelihood ratio is by directly estimating the conditional distributions of the scores given the same- and different-source hypotheses. Given the observed score $\Delta(A^*, B^*) = \delta^*$, the SLR is estimated via

$$\widehat{SLR}_{\Delta} = \frac{\widehat{g}(\Delta(A, B) = \delta^* | H_s, I)}{\widehat{g}(\Delta(A, B) = \delta^* | H_d, I)} \tag{3.3}$$

where $\widehat{g}$ is a generative model fit to the respective conditional distribution. Many techniques can be used to fit $\widehat{g}$, including maximum likelihood estimation under an assumed distributional form such as the Gaussian distribution [e.g., Hepler et al., 2012], kernel density estimation [e.g., Gonzalez-Rodriguez et al., 2005; Baechler et al., 2013; Bolck et al., 2015], Gaussian mixture models [e.g., Gonzalez-Rodriguez et al., 2005], and neural networks [e.g., Kong et al., 2019].

The most prevalent choice for estimating the score-based likelihood ratio is kernel density estimation (KDE) of the conditional distributions in the numerator and denominator. KDE is a common choice for the non-parametric estimation of a probability density function g. The kernel function K is usually defined as any symmetric density function that satisfies the following conditions

1. K integrates to unity: $\int K(x)dx = 1$
2. K has mean zero: $\int xK(x)dx = 0$
3. K has finite variance: $0 < \int x^2K(x)dx < \infty$.

Common examples of kernel functions include the Gaussian (or normal) and Epanechnikov kernels.

To model the conditional distribution in the numerator (denominator) of the SLR, we use the collection of N_s (N_d) scores $\{\delta_1, \ldots, \delta_{N_s}\}$ obtained from applying the score function $\Delta(A, B)$ on the pairs in $\mathcal{D}_s$ ($\mathcal{D}_d$). Given a kernel function K and a bandwidth $h > 0$, the kernel density estimator of the same-source score distribution is defined as

$$\widehat{g}(\delta^*|H_s, I) = \frac{1}{N_s} \sum_{i=1}^{N_s} \frac{1}{h} K\left(\frac{\delta^* - \delta_i}{h}\right). \tag{3.4}$$

A similar definition holds for $\widehat{g}(\delta^*|H_d, I)$. In either case, the estimated density is the average of the kernel centered at the observed score δ_i and scaled by h across all N_s (N_d) observations. KDEs are essentially a local smoothing method.

The choice of the kernel itself is not as important as the selection of the bandwidth h. As h decreases, the height of the peak at each observation increases resulting in less smoothing. As h increases, the height of the peak at each observation decreases and probability mass is pushed away from the observation resulting in more smoothing. Multiple techniques for bandwidth selection exist, one of the most popular of which is automatic bandwidth selection via the "rule of thumb" from [Scott, 1992].

Logistic Regression

Logistic regression is a common technique for mapping score values to score-based likelihood ratios in forensic biometrics [e.g., Brümmer and du Preez, 2006; Ramos et al., 2017]. Logistic regression models in the forensic setting estimate the posterior log odds of the competing hypotheses directly, so to obtain SLR values one must apply a post-hoc transformation. The

logistic model in this setting is as follows

$$\begin{aligned} \log O(H_s|\delta, I) &= \log\left(\frac{Pr(H_s|\delta, I)}{Pr(H_d|\delta, I)}\right) = \beta_0 + \sum_{j=1}^{d} \beta_j \delta_j \\ Pr(H_s|\delta, I) &= \frac{1}{1 + \exp\{-\log O(H_s|\delta, I)\}} \\ &= \frac{1}{1 + \exp\{-\log SLR_\Delta - \log O(H_s|I)\}} \end{aligned} \tag{3.5}$$

where the score values δ are assumed to be a d-dimensional vector. The coefficients $\boldsymbol{\beta}$ are estimated via numerical optimization for a given value of the prior odds $O(H_s|I)$. Estimates of the SLR are then computed by removing the influence of the choice of prior odds on $\widehat{O}(H_s|\delta, I)$, i.e., the odds computed by plugging in the coefficient estimates $\widehat{\boldsymbol{\beta}}$ into Equation 3.5. Namely, for the evidence (A^*, B^*) under question, the score-based likelihood ratio is approximated by

$$\begin{aligned} \widehat{SLR}_\Delta &= \exp\left(\log \widehat{O}(H_s|\delta^*, I) - \log O(H_s|I)\right) \\ &= \exp\left(\widehat{\beta}_0 + \sum_{j=1}^{d} \widehat{\beta}_j \delta_j^* - \log O(H_s|I)\right). \end{aligned} \tag{3.6}$$

The logistic regression model above can be used to estimate SLRs in a variety of settings. When $d = 1$, (i.e., only a single univariate score is under consideration) this method is simply an alternative to the generative approach because it assumes a linear form for the log-odds, which is equivalent to a generative approach with a Gaussian model under equal variance assumptions. In general, logistic regression is more flexible in that the coefficients are optimized rather than based on a Gaussian assumption. When $d > 1$, logistic regression can be used to "fuse" different scores together to compute the SLR, which is common in biometric systems [Pigeon et al., 2000; Brümmer et al., 2007]. More recently, multiple logistic regression models have been used to produce multiple scores and then fuse them together to produce likelihood ratios [Bosma et al., 2020].

Isotonic Regression

In Section 2.7, we showed how the PAV algorithm for isotonic regression could be used to calibrate likelihood ratio values. However, it can also be used to directly compute score-based likelihood ratios [Ramos et al., 2017] for 1-dimensional scores. To estimate SLR values, first apply the PAV algorithm to the score values ranked in ascending order and the known ground truth of the same-source hypothesis as response values (as determined by $\mathcal{D}_s$ and $\mathcal{D}_d$). The resulting posterior probabilities, $\widehat{Pr}(H_s|\delta, I)$ can then be transformed back to calibrated log likelihood ratio values via Equation 2.24. These SLRs can then be computed for new cases by using the isotonic regression model to map the observed score to calibrated posterior probabilities and then using Equation 2.24 to obtain a calibrated LR value for the new case. Note that PAV leads to a non-invertible transformation, as it results in a step function. Smoothing techniques can be applied to mitigate this issue. For instance, adding a small slope or interpolating with spline approaches result in invertible functions.

3.2 The Coincidental Match Probability

Likelihood ratio techniques (i.e., the LR and SLR) are not the only way to evaluate forensic evidence. In fact, a recent—albeit controversial—report from the National Institute of Standards and Technology stated "we hope the forensic science community comes to view the LR as one *possible*, not normative or necessarily optimum, tool for communicating to DMs (decision makers)" [Lund and Iyer, 2017, authors' emphasis]. While there is a significant opposition to this position [e.g., Aitken et al., 2018], there are certain situations in which computing the likelihood ratio or its score-based variant is not possible. In this section, I introduce one such method, the coincidental match probability, that is a score-based extension of the two-stage inference framework [Parker, 1966, 1967].

Parker [1966, 1967] views the forensic inference process as occurring in two stages, the *similarity stage* and the *identification stage*. In the similarity stage, the goal is to compare the characteristics of the samples A and B and determine whether they are distinguishable. As the difference between the sets of characteristics increases, the hypothesis that the samples originate from the same source is weakened to the point that it can be rejected. However, establishing that A and B are indistinguishable is not sufficient in itself to conclude they originated from the same source, as it could occur by chance. Intuitively, the value of finding that A and B are indistinguishable is a function of the number of sources whose characteristics would also be deemed indistinguishable from the unknown source sample A (i.e., fewer potential sources that are indistinguishable from A results in the evidence being more valuable). Thus, the goal of the identification stage is to determine the rarity of the characteristics observed in the first stage among a population of potential sources (assuming that A and B were determined to be indistinguishable in the first stage). This level of rarity is often called a *probability of coincidence*. In the remainder of this section, a score-based extension of the probability of coincidence is presented.

The *coincidental match probability (CMP)* is the probability that two samples of evidence A^* and B^* exhibit the characteristics of a same source pair, i.e., a small value of $\Delta(A^*, B^*)$, by chance given that they are from different sources. Coincidental match probabilities are intrinsically related to the denominator of the score-based likelihood ratio, i.e., the conditional likelihood of observing the value $\Delta(A^*, B^*)$ given that the series have different sources. Further, CMPs share conceptual similarities with random match probabilities (RMPs) that are frequently used in forensics, particularly in DNA analysis [Thompson and Newman, 2015]. When computing RMPs, forensic scientists first determine if two samples match, and, if so, they then compute the probability that the samples match by chance. When calculating CMPs, however, we do not first attempt to determine if the series A^* and B^* are from the same source, but instead calculate the probability that they exhibit the observed degree of association by chance. Thus, CMPs and RMPs are related but have different interpretations.

The coincidental match probability relies on fixing the unknown source sample A^* and having multiple realizations of the known-source sample B^* under a null model for different-source data (H_d). These realizations are then used to induce a distribution of scores under this model. Specifically, given an observed pair (A^*, B^*) we define the CMP as the probability that a randomly sampled pair (A^*, B') under the different source model has a more extreme score $\Delta(A^*, B')$ (indicating greater similarity) than the observed score $\Delta(A^*, B^*) = \delta^*$,

$$CMP_\Delta = g(\Delta(A^*, B') < \delta^* | H_d, I). \tag{3.7}$$

"More extreme" here is related to the notion of hypothesis tests and can be defined as either one-sided or two-sided. The definition of coincidental match probability in Equation 3.7 assumes a one-sided test where the observed score for same source pairs tends to be less than that of different source pairs. See Figure 3.2 for an illustration.

3.2.1 Estimation

We propose the following natural estimator for the coincidental match probability

$$\widehat{CMP}_\Delta = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \mathbb{I}[\Delta(A^*, B^{(i)}) < \delta^*] \tag{3.8}$$

where $B^{(i)}$ for $i = 1, \ldots, n_{sim}$ are randomly sampled under H_d using the null model for different-source data, and $\mathbb{I}$ is the indicator function. The smaller this empirical probability, the less likely it is that the pair (A^*, B^*) was generated by different sources.

When reference data of known different-source pairs $\mathcal{D}_d$ are available, the estimator defined in Equation 3.8 is straightforward to compute, e.g., by averaging the indicator function over the N_d different-source pairs in the reference set $\mathcal{D}_d$. However, there are many situations in practice where data from a reference population is not readily available. Furthermore,

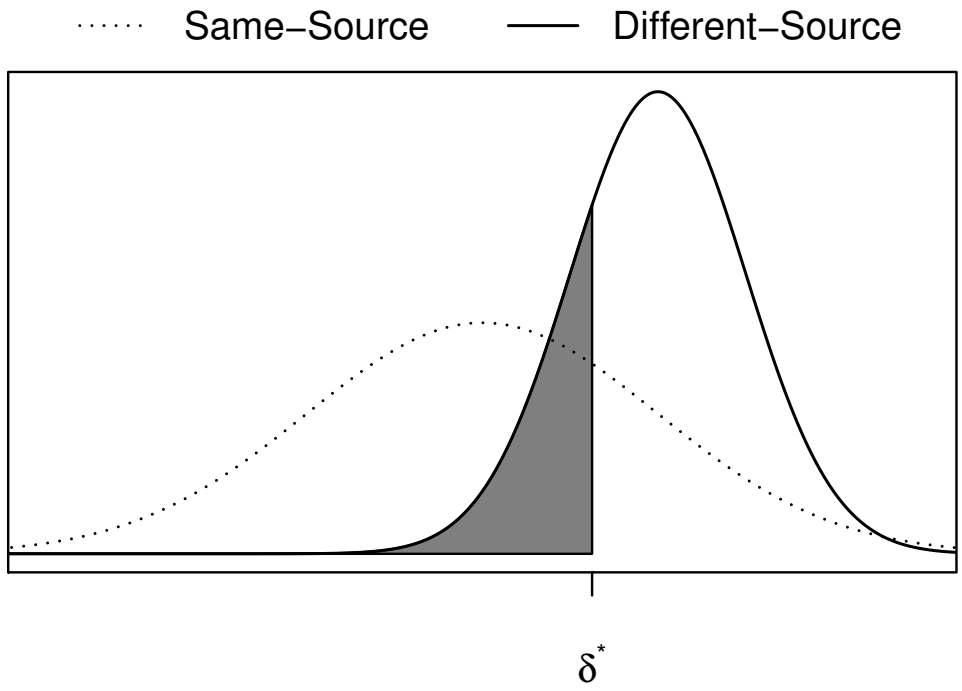


Figure 3.2: Hypothetical illustration of the densities of the score function Δ under the hypotheses that the samples are from the same source (H_s, dashed line) and that the samples are from different sources (H_d, solid line). The coincidental match probability CMP_Δ is the shaded tail region of $g(\Delta(A, B)|H_d, I)$.

even when a population is available, it is often quite difficult to define the *relevant* reference population of interest in a forensic setting. Should the relevant population be a sample from all individuals in general, or from everyone who matches the description of a suspect in a given region, or from some other group? To address these potential problems we propose a resampling approach that computes coincidental match probabilities using only a single sample of evidential data.

With only a single sample of evidence, the likelihood ratio and score-based likelihood ratio are not possible to compute, as reference data is necessary to estimate model parameters

or build non-parametric models like kernel density estimators. The CMP can be estimated by simulating data under the different-source model. Typically, this distribution is easier to make reasonable assumptions about as it models the two sources of information as conditionally independent given the background information. How exactly these samples are simulated is application specific. One such technique for temporal event data is presented in Chapter 5. It is not always possible to simulate realistic data, however, so this method should be applied cautiously. For that reason, the coincidental match probability is more of an investigative tool than one that is ready for presentation in court.

This approach is similar in spirit to the use of simulation envelopes for computing confidence intervals for a spatial assocation function in spatial point process models (e.g., Baddeley et al. [2014]) where resampling techniques are used to estimate confidence intervals under a null model (such as complete spatial randomness). In our approach the null model assumes that the two samples of evidence were generated by different sources.

3.2.2 Interpretation

Similar to the likelihood ratio methods, we can view the CMP as a discriminant function for a binary classification decision (e.g., pairs with CMP values less than some threshold are considered same source) and compare the true and false positive rates. The thresholds can be varied to compute the AUC of the ROC curve for direct comparison to the LR methods.

For instance, assume that we are interested in the classification performance of the method at two particular thresholds: 0.05 and 0.001. A pair of evidential samples (A^*, B^*) whose CMP value is equal to one of these thresholds has the following interpretation: the probability that a score comparable to or lower than that obtained from the pair (A^*, B^*) was generated by different source data is 5% (or 0.1%).

The thresholds have a special interpretation when using simulated data. The probability model is known in this scenario, so the threshold controls the false positive rate, i.e., incorrectly classifying a different-source pair as having come from the same source. This is an attractive property as it allows the examiner to limit the rate of false identifications in practice.

3.3 Examples with Known Distributions

Having introduced the LR, SLR and CMP for measuring strength of evidence, we now compare the three methods on simulated data. To illustrate each method, we begin with simple examples where the generative mechanism of the evidence is known. This is a common approach to analyze and compare the behavior of methods for assessing the strength of evidence, particularly in comparing the likelihood ratio to its score-based equivalent [Lindley, 1977; Grove, 1980; Hepler et al., 2012; Ommen and Saunders, 2018; Morrison and Enzinger, 2018; Garton et al., 2020]. In general, we follow the notation of Hepler et al. [2012] in the problem formulation.

Let A and B denote real-valued, univariate random variables corresponding to samples of evidence from an unknown source and from a specific known source, respectively, and let a and b denote single realizations (or observations) of these random variables. Assume that it is known that samples of this type follow a normal distribution with some known mean parameter and a known within-source variance σ_ω^2. Therefore, the known source evidence B follows a univariate normal distribution with mean μ_B variance σ_ω^2, denoted $B|\mu_B, \sigma_\omega^2 \sim N(\mu_B, \sigma_\omega^2)$. Similarly, assuming that the mean μ_A of the unknown source evidence A is known, A also follows a normal distribution, denoted $A|\mu_A, \sigma_\omega^2 \sim N(\mu_A, \sigma_\omega^2)$. Here, we make the simplyfing assumption that the variance of a given sample is the same within each source.

Under the same source hypothesis, assume that A and B are conditionally independent random variables given the model parameters and that $\mu_A = \mu_B$. Thus, the unknown source variable A and known source variable B have the same distribution.

Under the different source hypothesis, a common assumption is that the source of A is a randomly selected individual from a relevant population of alternate sources. Thus, an additional variance term is needed because we do not know the value of μ_A. There exist two sources of variance—within-source variance σ_ω^2 and between-source variance σ_β^2—that are both assumed to be known. These sources of variance play a key role in defining the relevant reference population's distribution. Denote this reference population as P, and assume that it is normally distributed with known mean μ_P and variance $\sigma_P^2 = \sigma_\omega^2 + \sigma_\beta^2$. Under H_d, A has the same distribution as P so that $A \sim N(\mu_P, \sigma_P^2)$. We can then state the specific source hypotheses as follows:

$$
\begin{aligned}
H_s &: A \sim N(\mu_B, \sigma_\omega^2), \quad B \sim N(\mu_B, \sigma_\omega^2) \\
H_d &: A \sim N(\mu_P, \sigma_P^2 = \sigma_\omega^2 + \sigma_\beta^2), \quad B \sim N(\mu_B, \sigma_\omega^2).
\end{aligned}
$$

We now explore the behavior of the LR, SLR and CMP under these hypotheses.

3.3.1 Likelihood Ratio

Under the same source hypothesis, A and B are conditionally independent random variables given the model parameters. Therefore the joint density can be expressed as the product of their respective densities, which yields

$$
f(A = a, B = b | H_s, \mu_B, \sigma_\omega) = \frac{1}{\sigma_\omega^2} \phi\left(\frac{a - \mu_B}{\sigma_\omega}\right) \phi\left(\frac{b - \mu_B}{\sigma_\omega}\right) \tag{3.9}
$$

where ϕ is the standard normal probability density function.

Under the different source hypothesis A and B are assumed to be independent—information about the known source provides no additional information about the unknown source. Thus, we can again express their joint density as the product of their respective densities resulting in

$$f(A = a, B = b|H_d) = \frac{1}{\sigma_P \sigma_\omega} \phi\left(\frac{a - \mu_P}{\sigma_P}\right) \phi\left(\frac{b - \mu_B}{\sigma_\omega}\right). \tag{3.10}$$

Taking the ratio of Equations 3.9 and 3.10 and eliminating common terms yields the likelihood ratio

$$LR = \quad = \frac{\sqrt{\sigma_\omega^2 + \sigma_\beta^2}\phi\left(\frac{a-\mu_B}{\sigma_\omega}\right)}{\sigma_\omega \phi\left(\frac{a-\mu_P}{\sqrt{\sigma_\omega^2+\sigma_\beta^2}}\right)}. \tag{3.11}$$

Note that the LR above does not rely on the value of the observation b of the known source sample B, but only on the value a of the unknown source sample A. This is strictly because, under both H_s and H_d, we assumed that A and B were conditionally independent. The independence assumption under the same source hypothesis was done for convenience, as we will show that the distributional forms of the SLR and CMP are known.

3.3.2 Score-based Likelihood Ratio

To compare the behavior of the likelihood ratio with that of the score-based methods we must define a suitable score function Δ for the data under consideration. The score should be small for samples from the same source and large for samples from different sources. Define the random variable $\Delta(A, B) = (A - B)^2$. Under the assumptions made under each hypothesis, this score function satisfies the desired properties. Additionally, since A and B are normally distributed exact expressions for the SLR and CMP can be obtained

by leveraging the relationship between squared normal distributions and the chi-squared distribution.

Under the same source hypothesis, it is straightforward to show that

$$\frac{(A-B)^2}{2\sigma_\omega^2} \sim \chi_1^2. \tag{3.12}$$

Under the different source hypothesis,

$$\frac{(A-B)^2}{2\sigma_\omega^2 + \sigma_\beta^2} \sim \chi_{1,\lambda}^2 \tag{3.13}$$

where $\lambda = \frac{(\mu_P - \mu_B)^2}{2\sigma_\omega^2 + \sigma_\beta^2}$ is the noncentrality parameter.

Taking the ratio of the probability density functions in Equations 3.12 and 3.13 yields the score-based likelihood ratio

$$SLR = \frac{g(\Delta(A,B) = \delta | H_s)}{g(\Delta(A,B) = \delta | H_d)} = \frac{(2\sigma_\omega^2 + \sigma_\beta^2)\chi_1^2\left(\frac{\delta}{2\sigma_\omega^2}\right)}{2\sigma_\omega^2 \chi_{1,\lambda}^2\left(\frac{\delta}{2\sigma_\omega^2 + \sigma_\beta^2}\right)} \tag{3.14}$$

where $\chi_{1,\lambda}^2(\cdot)$ is the probability density function of the noncentral chi-squared distribution with noncentrality parameter $\lambda = \frac{(\mu_B - \mu_P)^2}{2\sigma_\omega^2 + \sigma_\beta^2}$.

3.3.3 Coincidental Match Probability

The coincidental match probability is simply the tail probability of the score $\Delta(A,B)$ under the different source hypothesis. Given the probability density function of the random variable specified in Equation 3.13, it can be formally stated as

$$CMP = Pr(\Delta(A,B) < \delta | H_d) = \int_0^\delta \frac{1}{2\sigma_\omega^2 + \sigma_\beta^2} \chi_{1,\lambda}^2 \left(\frac{x}{2\sigma_\omega^2 + \sigma_\beta^2}\right) dx. \tag{3.15}$$

We approximate the above integral via an alternative series representation [Ding, 1992].

3.3.4 Comparison

Instead of comparing the functional forms of the evidence evaluation methods, we focus on analyzing their behavior graphically [Hepler et al., 2012; Garton et al., 2020]. Figure 3.3 depicts the distributions of the evidence under the same- and different-source hypotheses along with the corresponding values of the LR, SLR, and CMP for various settings for the parameters μ_P, σ_ω, and σ_β. For ease of comparison, the measurement b taken from the known source sample B was assumed to be fixed to the mean of its distribution, i.e., $b = \mu_B = 0$. This was done to eliminate one source of variability in the analysis as was done in [Hepler et al., 2012].

The columns of Figure 3.3 represent different settings of the parameter μ_P, σ_ω, and σ_β (in every example $\mu_B = 0$). The top row of plots in Figure 3.3 shows the distribution of the unknown source sample A under both H_s (solid curve) and H_d (dashed curve) for the given values of the parameters. The middle row of plots depicts the behavior (in log scale) of the likelihood ratio of Equation 3.11 (black curve) and score-based likelihood ratio of Equation 3.14 (dashed curve) as a function of the observed value $A = a$. As a reference, a horizontal line is plotted at 0 for the log ratio value. Values above this threshold support H_s, while values below it support H_d. Finally, the third row of plots shows the behavior of the coincidental match probability of Equation 3.15 as a function of a. Again, a horizontal reference line was plotted at 0.05 as this is a common choice of threshold in hypothesis testing. CMP values below this threshold support H_s.

In the first column of Figure 3.3, the known source sample is normally distributed with mean $\mu_B = 0$ and standard deviation $\sigma_\omega = 0.5$. Under H_s, the unknown source sample A also follows this distribution. Under H_d, A arises from the population of different sources, which

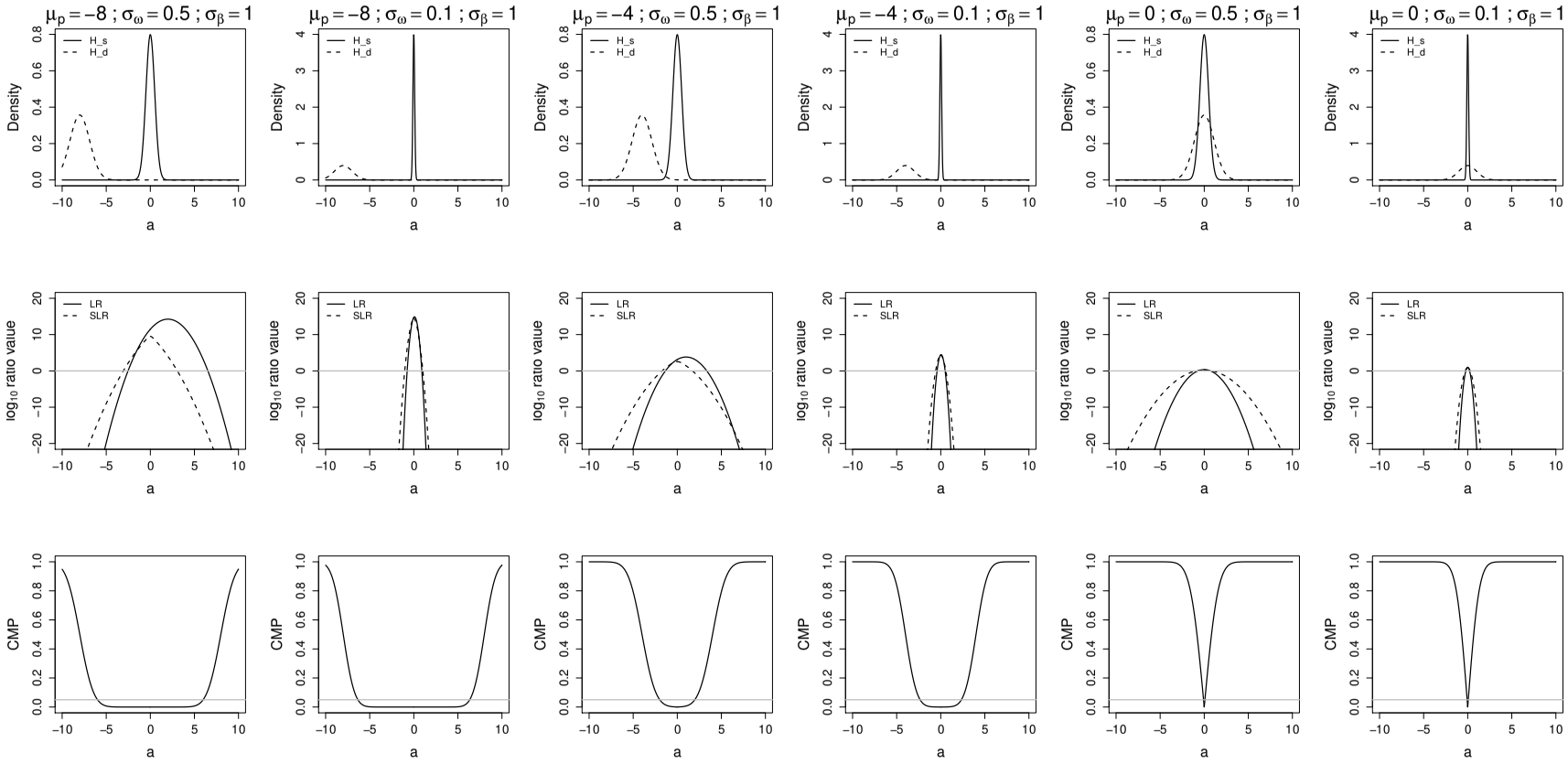


Figure 3.3: Behavior of the various methods for evaluating evidence under known distributional forms. The value $B = b$ was fixed to the mean of the same-source distribution, $\mu_B = 0$, to eliminate one source of variability in the analysis. Columns represent a selection of values for the parameters μ_P, σ_ω, and σ_β. (Top row) Distribution of A under H_s and H_d; (middle row) behavior of the LR and SLR as a function of $A = a$; (bottom row) behavior of the CMP as a function of $A = a$.

is a normal distribution with mean $\mu_P = -8$ and standard deviation $\sigma_p = \sqrt{0.5^2 + 1}$. The log likelihood ratio takes on positive values as the measurement from the unknown source sample a approaches the mean of the known source. It continues to increase as the value of a increases until it reaches the far right-tail of the known-source distribution. Once far in the tail, it becomes less and less likely to have come from the known source because the density values in the tail of that distribution are lower than that of the population distribution (due to the higher variance $\sigma_P^2 > \sigma_\beta^2$).

The log score-based likelihood ratio, on the other hand, is symmetrical around the observed value of b (the mean of the known source distribution). This behavior is expected, as the score function used was the squared difference between the two observations. Values a near the observed value of the known source sample result in SLR values that support H_s, with decreasing support further from 0 in either direction. When $a < -1.5$ the log SLR is greater than the log LR, implying that the SLR is overstating the value of the evidence in favor of the same-source hypothesis. This overstatement even results in different conclusions being drawn when $-3 < a < -2.5$, as log SLR is above zero while log LR is below zero. When $a > -1.5$ the SLR is underestimating the value of the evidence, in some cases drastically so. This shows that some amount of evidential value is not being adequately captured by the score-based method, albeit in a contrived hypothetical example.

Similar to the SLR, the coincidental match probability is also symmetric around the observed value of b. Depending on the choice of threshold for favoring H_s, the CMP performs worse than both the LR and SLR. If using a threshold of 5%, the region that favors H_s is $[-6, 6]$. As the threshold is lowered, this interval shrinks. Choosing a threshold of 0.001% results in this region about half of the size $[-2.75, 2.75]$, which is similar to that of the SLR.

The properties displayed in the first column of figure 3.3 are the most extreme for the parameter settings chosen. When the between-source variability σ_β is much larger than the within-source variability σ_ω, the SLR is well-behaved in comparison to the LR (e.g.,

comparing the second column to the first, fourth to third, and sixth to fifth). The CMP is unaffected by changing the ratio of these variances, as the noncentrality parameter of Equation 3.15 is more impacted by the value of the score function than it is by the relatively small variance terms considered.

As the mean of the population distribution approaches that of the known source distribution, the magnitudes of the SLR and LR values are reduced, but the overall shapes of the curves remain similar. An exception is when $\mu_P = \mu_B = 0$, where the LR values are also symmetric around 0 (fifth column of Figure 3.3). The SLR also tends to overestimate the LR in this setting (though both the LR and SLR values are extremely small and support the different-source hypothesis). Furthermore, the rejection region produced by the CMP shrinks as the means of the two distributions approach one another.

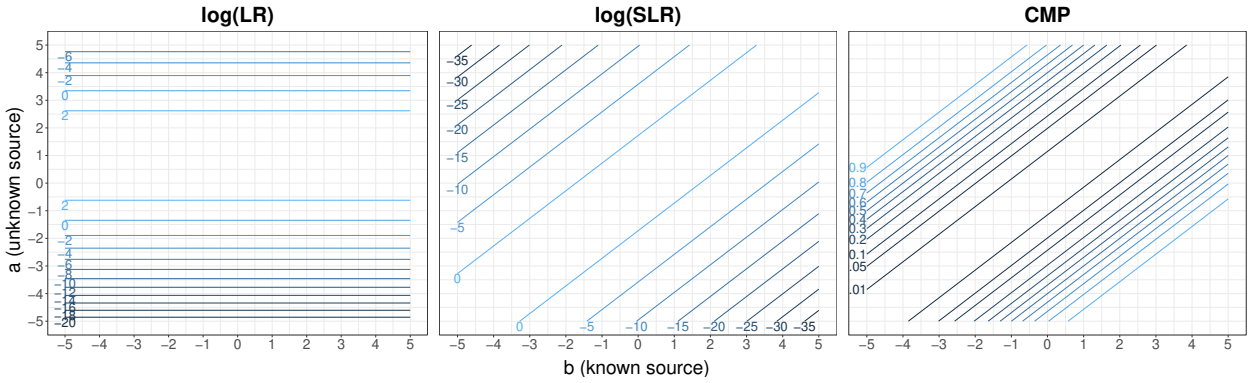


Figure 3.4: Contour plots of the various evidence evaluation methods for the example with known distributions where $\mu_B = 0$, $\mu_P = -4$, $\sigma_\omega = 0.5$, and $\sigma_\beta = 1$ (corresponding to the third column of Figure 3.3).

Another way to compare the behaviors of the evidence evaluation methods in the known distribution setting is to compute their values over a grid of values a and b for the random variables A and B, as done in [Garton et al., 2020]. Figure 3.4 depicts contour plots of the LR, SLR and CMP when $\mu_B = 0$, $\mu_P = -4$, $\sigma_\omega = 0.5$, and $\sigma_\beta = 1$ (corresponding to the third column of Figure 3.3). Under H_s we assumed that A and B are conditionally independent of one another given the parameter values, so the resulting LR does not depend on the value of B. However, the SLR and CMP depend on the values of A and B through

the score function. This results in horizontal contours for the LR, and diagonal contours for the SLR and CMP.

The comparison presented between the methods in this section shows considerable differences in the behavior of the LR, SLR and CMP for a simple example where the distributions are known. In practice, one would expect greater differences in their behavior as real data requires estimation of unknown parameters and is typically from much more complex generative distributions. In practical situations, however, the SLR or CMP might be the only option available for the forensic examiner. For such scenarios, the performance of the SLR and CMP should be evaluated using held out data prior to use in court to determine their validity for the problem at hand. The contrasts presented here are meant to illustrate differing behaviors of the methods, not to discredit the use of score-based approaches altogether.

3.4 Discussion

A common question in forensic analysis is whether two observed data sets originated from the same source or from different sources. Statistical approaches to addressing this question have been widely adopted within the forensics community, particularly for DNA evidence. For other types of evidence, especially pattern evidence, appropriate statistical approaches are a current area of research.

In general, there are two primary methods for quantifying the strength of forensic evidence. The first is a likelihood ratio approach based on modeling the observed features of the evidence directly as discussed in Chapter 2. The second approach, as discussed in this chapter, is to instead measure the similarity of the two observed data sets via a score function. That score function can then be used to assess the strength of evidence via the score-based likelihood ratio or coincidental match probability. A comparative evaluation using simulated data

with known distributions was presented. The comparison showed considerable differences in the approaches, even for a relatively simple example. However, the SLR and CMP may be useful options in situations when the LR is not practical, as we will see in Chapters 4 and 5.

In the following chapters, I will present extensions of the evidence evaluation framework presented here for user-generated event data, which can be viewed as a type of pattern evidence.

3.4.1 Contributions

The following contributions were made in this chapter:

- A survey of the score-based likelihood ratio for the analysis of forensic evidence that unifies the forensics, biometrics and statistics literature on the topic.
- A score-based adaptation of the coincidental match probability for the evaluation of forensic evidence.
- A comparison of the direct modeling approach of the likelihood ratio to the score-based approaches of the score-based likelihood ratio and coincidental match probability for a theoretical example of evidence from a known distribution.

Chapter 4

Spatial Event Data

Logs of geolocation data are now routinely available on modern mobile devices. This type of data is typically associated with events generated on the device, such as actions taken by a user in a software application. Such data can be collected in a variety of ways—from the device itself, from servers that store the locations based on IP addresses, from cellular towers, and so on. Given the general prevalence of mobile devices, this type of spatial event data is now encountered with increasing regularity during forensic investigations. For instance, an investigator might wish to determine if two sets of events with geolocations, corresponding to different accounts or devices, were in fact generated by the same individual.

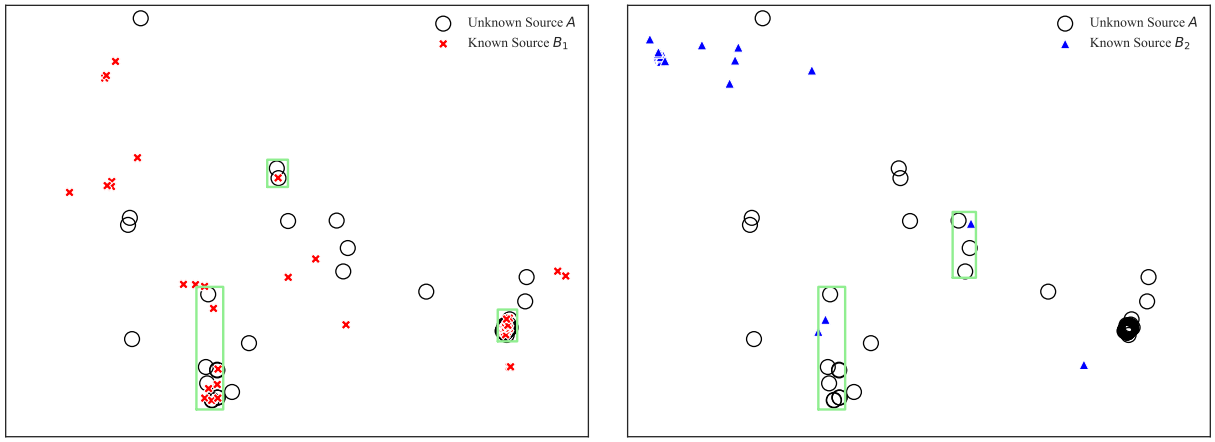
The primary contribution of this chapter is the development and evaluation of quantitative techniques for the forensic analysis of geolocated event data. In particular two types of approaches to obtain strength of evidence are proposed: a likelihood ratio approach based on modeling the evidential data directly and a score-based approach that instead models a summary measure of the similarity of the evidence and then quantifies that measure with either the score-based likelihood ratio or coincidental match probability.

4.1 Motivating Example

Suppose that a forensic investigator is given a set of GPS coordinates associated with criminal activity and is tasked with finding the most likely suspect from a set of individuals for whom reference location data is available. The GPS coordinates could be the locations of crime scenes (e.g., in the case of serial crime) or data gathered from a device of unknown origin (e.g., a burner phone recovered from a crime scene). In either case, we do not know who generated this location data and will refer to it as the *unknown source data*.

One investigative approach in this context is to gather location data for a set of potential suspects via a geofence warrant (e.g., Valentino-DeVries [2019]). A geofence warrant refers to a situation where a "fence" or bounding box is constructed around a set of locations, such as locations associated with a crime. A law enforcement agency then requests data from a service provider (such as Google or Twitter) for any individuals whose devices were within the geofence during a time-period of interest (e.g., in a window of time around which the criminal activity occurred). For individuals who match the geofence (i.e., potential suspects), their geolocation data is given an anonymous identifier and their data is sent to law enforcement to aid in the investigation. We will refer to the location data for these individuals as the *known source data* because, once persons of interest have been identified, the service provider can reveal their identities.

Figure 4.1 provides an illustrative example of such a geofence situation. The data points are geolocated events, with colors and shapes indicating different accounts. Here we treat A (black circles) as the unknown source data, where each point has an associated geofence surrounding it whose size and shape was determined based on the land parcel data described in 4.5.3. Figures 4.1a and 4.1b show geolocation events from two different known source accounts B_1 (red crosses) and B_2 (blue triangles) with at least one GPS coordinate inside a geofence (highlighted by green boxes).



(a) Same source (b) Different sources

Figure 4.1: Location data (taken from Section 4.7.1) in a 3.5 square mile region of Orange County, CA. Green boxes represent geofences with events in both sets. (a) Both the unknown and known source data were generated by the same individual; (b) the unknown and known source data were generated by different individuals. The unknown source data is the same in both panels. The geographic features of the map (i.e., street names and buildings) were removed to preserve the individuals' privacy.

An investigator looking at this data would need to infer how likely it is that the locations in each panel of the figure match to the same source (e.g., were generated by the same individual). In this example, we have selected the data so that the points in Figure 4.1a are from the same account (over different time-periods) and the points in 4.1b are from different accounts. Determining if sets of locations "match" can be a difficult task due to many factors including variability in human behavior and the typicality of locations of interest (e.g., how common they are in the population in general). To address this problem we propose a technique that, given two sets of locations, produces an objective measure of their probative evidential value. Using our method for the data in Figure 4.1, an investigator would be able to conclude that there is strong support for the hypothesis that the two sets of locations in Figure 4.1a were generated by the same individual. She would also be able to conclude that the individual that generated the known source data (blue triangles) in Figure 4.1b can likely be excluded as the source of the unknown source data (black circles). In the remainder of this chapter, we will show how to produce such conclusions given this type of location

data.

4.2 Related Work

Evaluating location-related mobile device evidence and expressing probative conclusions in the forensic setting is challenging due to both technological and circumstantial subtleties that can be present in the data. Casey et al. [2020] discuss these challenges and present a structured framework for the evaluation of geolocation data. However, the hypotheses considered in that work are focused on specific locations of interest rather than comparing sets of spatial patterns (which is the focus here).

The recent work of Bosma et al. [2020] is similar in spirit to our work in that they address same-source problems using mobile geolocation data. They develop a method that uses the location and time of cellular tower registrations of mobile phones to assess the strength of evidence that a pair of phones were used by the same person. Their approach creates features from the cell tower data and makes parametric modeling assumptions via logistic regression in how those features indicate same- and different-source phone usage patterns. The methods that we propose differ in that we make no such parametric assumptions, and no data has to be held out to estimate model parameters (although we do require a reference set of data in order to estimate the typicality of locations, e.g., how frequently-visited they are by the population in general).

For spatial event data there is a long history in the development of statistical analysis and modeling methods (e.g, see Ripley [1977], Isham [1981], Diggle [2013] and Illian et al. [2008]) motivated by a variety of problems in application areas such as astronomy, ecology, geology, and criminology. Of potential relevance to digital forensics is work on modeling and evaluation of dependence between spatial point patterns [e.g., Berman, 1986; Schlather et al.,

2004]. Measuring the relationship between spatial point patterns is typically done by performing inference (either analytical or numerical) under a null model that typically assumes some form of independence between the patterns. One of the most well-known techniques is Ripley's cross-K function [Dixon, 2014], which measures the number of occurrences of one type of point within a given radius r of the other type of point as a function of r. Under certain assumptions on the processes themselves (e.g., stationarity) and the relationship between the processes (e.g., complete spatial randomness), significance tests can be performed to determine if the observed function is consistent with the assumed relationship [e.g., Diggle and Chetwynd, 1991; Gaines et al., 2000]. Much of this type of work relies on assumptions such as spatial homogeneity that are not well-suited to the type of bursty and non-stationary human-generated event data that is often of interest in a forensics setting. Nonetheless this prior work in spatial point processes can provide a useful starting point for analyzing spatial event data in a systematic manner.

4.3 Forensic Question of Interest

Geolocated event data can be viewed as a realization of a spatial point pattern, which consists of a set of positions $\{s_1, s_2, \ldots\}$ in a defined region R at which events have been recorded [e.g., Diggle, 2013]. Each point, $s_i = (s_{ix}, s_{iy})$, refers to the position of the ith observed event and is simply a shorthand way of identifying the x and y coordinates of an event. The region R limits where events can be observed, i.e., only events whose location is within R are considered. The region is assumed to be predetermined by the forensic investigator (sensitivity to the choice of the region is beyond the scope of the current work).

In this chapter we consider bivariate point patterns, where we observe some additional feature of the events that can be used to dichotomize them into one of two classes (a classic example of this is in epidemiology, where individuals with a disease are marked as "cases" and those

without as "controls"). Bivariate point patterns can be thought of as a superposition of two point patterns corresponding to the set of positions for events in each class, e.g., pattern A and pattern B with n_a and n_b events of type A and type B, respectively. We base our notation for bivariate point patterns on that of Illian et al. [2008], with

$$E = (A, B) = \{(s_i, m(s_i)) : i = 1, \ldots, n_a + n_b\} \tag{4.1}$$

where $m(s_i) \in \{\mathrm{A}, \mathrm{B}\}$ is the type (or mark) of the ith event at position s_i.

To formally define the question of interest, we adopt notation and terminology from the specific source problem posed in Section 2.2.1. For geolocated event data, the unknown source data A and known source data B could consist of locations at which actions were taken on two different devices, e.g., locations where phone calls were made. The forensic question of interest in this scenario would be to determine how likely it is that the events on the different devices were generated by the same individual. Alternatively, sets A and B could consist of locations at which events were generated from a single account (e.g., accounts on a social media platform such as Twitter) or locations from the same device but over two different time periods. The forensic question of interest would be to determine if the same individual was responsible for generating both sets of events. This scenario is relevant for example when the person of interest invokes the "it wasn't me" defense, with A corresponding to events for which the individual claims they are not responsible and B corresponding to a sample of his or her typical activity.

In the scenarios above, A and B refer to sets of (longitude, latitude) coordinates at which events occurred. Figure 4.2 provides an example of such geolocated event data. In this specific example $A = \{(-73.984, 40.754), (-73.977, 40.761), \ldots, (-73.987, 40.727)\}$ for a total of $n_a = 42$ events in A, and $B = \{(-73.988, 40.742), (-74.009, 40.711), \ldots, (-73.995, 40.718)\}$

for a total of $n_b = 39$ events in B.[1]

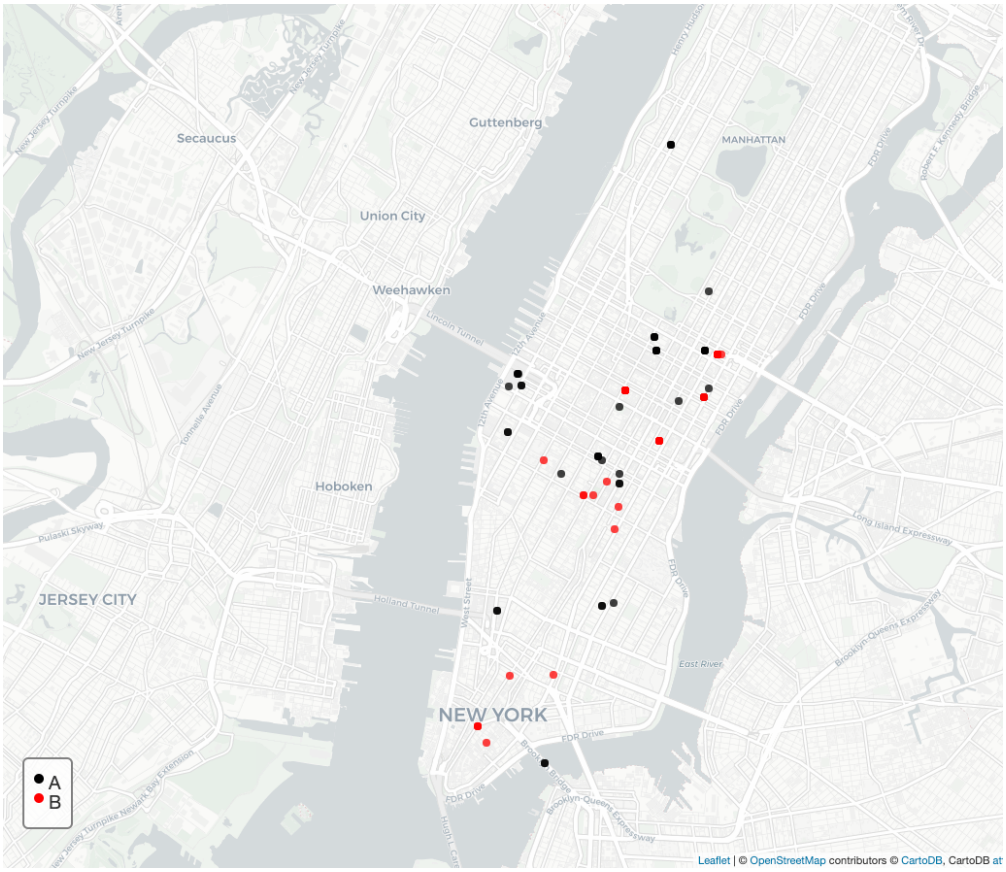


Figure 4.2: Example of sets of locations for Twitter data from New York. The patterns correspond to geolocatons of tweets from the same account over two different months, with month 1 corresponding to A (red) and month 2 corresponding to B (black).

The goal of a forensic examination is to assess the likelihood of observing the evidence (A, B) under two the same- and different-source hypotheses H_s and H_d, respectively. In the context of the geolocation data we will be focusing on in this chapter, the term "source" refers to a specific individual or user account, and the term "came from" can be interpreted as meaning "generated by." Thus, H_s is the proposition that the sample from the unknown source A was generated by the same individual or user account as the sample from the known source B. H_d is the proposition that the sample from the unknown source A was *not* generated by the specific source of B, but instead from another individual among an alternative source

[1]The latitude and longitude values presented in the text were rounded. Generally much higher precision is available, e.g., for coordinates provided by GPS.

population. See Section 2.2.1 for a more detailed discussion of the competing hypotheses.

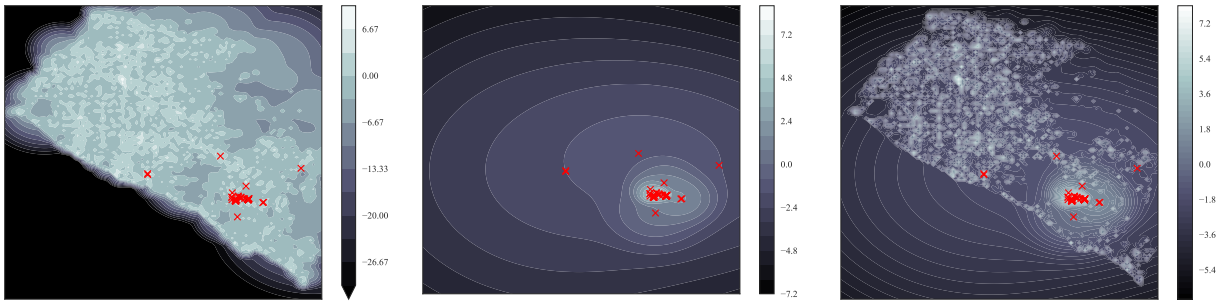
We propose and investigate two approaches for assessing the strength of evidence in this context. The first is a likelihood ratio approach that uses kernel density estimation techniques to estimate the relative likelihood of the observed location evidence under each proposition, H_s and H_d. The second approach is to instead measure the similarity of the two sets of locations via a score function and then assess the strength of the observed score resulting in either the score-based likelihood ratio or coincidental match probability.

4.4 Computing the Likelihood Ratio

We treat the event locations in A and B as real-valued numbers in the two-dimensional plane, and thus the numerator and denominator terms in the likelihood ratio of Equation 2.2 are modeled via probability densities, henceforth referred to by $f(\cdot)$. Therefore, the likelihood ratio for spatial event data can be expressed as

$$LR = \frac{f(B|A, H_s, I)}{f(B|H_d, I)} = \frac{\prod_{j=1}^{n_b} f(s_j^b|A, H_s, I)}{\prod_{j=1}^{n_b} f(s_j^b|H_d, I)} \tag{4.2}$$

where s_j^b is the location of the jth event in B. An additional assumption that the event locations in B are conditionally independent given the model allows us to express the terms in the numerator and denominator of the LR as products over the appropriate conditional density evaluated for each event in B. The support of the conditional probability densities is determined by the spatial region R defined by the investigation. Thus, the integral of f within any subregion is the probability that an event is in that subregion given that it occurred somewhere in R. Traditional parametric models for $f(B|A, H_s, I)$ and $f(B|H_d, I)$ above, such as spatial Poisson point process models, are often insufficient to capture the typical characteristics of user-generated geolocated event data that tends to be bursty and



(a) Population Component (b) Individual Component (c) Mixture

Figure 4.3: Example of the KDE models used to estimate the likelihood ratio for Twitter events in Orange County, CA, from the experimental results in Section 4.7. Overlaid on each panel are the set of points A from the motivating example in Section 4.1. (a) Population component used to estimate the denominator of the LR $f(B|H_d, I)$; (b) individual component built using the overlaid points; (c) mixture model with $\alpha = 0.8$ used to estimate the numerator of the LR $f(B|A, H_s, I)$.

inhomogeneous. For that reason, we focus on non-parametric kernel density estimation techniques for modeling sets of locations A and B.

To estimate the likelihood ratio, we first define a sample from a relevant reference population $\mathcal{D}$ of geolocated events, denoted $\mathcal{D} = \{s_k : k = 1, \ldots, n_p\}$ where s_k is the (longitude, latitude) coordinate of the kth event and n_p is the total number of population events in $\mathcal{D}$.[2] For a particular set of geolocated events B, the probability density function in the denominator of Equation 4.2, $f(B|H_d, I)$, can be estimated by

$$\widehat{f}(B|H_d, I) = \prod_{j=1}^{n_b} f_{KD}(s_j^b | \mathcal{D}) \tag{4.3}$$

where $f_{KD}(\cdot | \mathcal{D})$ is an adaptive bandwidth kernel density estimator built on the population data $\mathcal{D}$. See 4.4.1 for a more detailed discussion of adaptive bandwidth kernel density estimation and a formal definition of f_{KD}. The term $f_{KD}(s | \mathcal{D})$ is the likelihood that a randomly selected event from the population will occur at some particular location s, conditioned on s

[2]The definition of the sample from the reference population $\mathcal{D}$ is slightly different than that of Section 2.4, but it is equivalent. In this chapter, $\mathcal{D}$ refers directly to the observed locations and not to the evidence exemplars (A, B) that the locations come from. Essentially, $\mathcal{D}$ defines where individuals in the population are generating events regardless of the source of A and B.

lying within region R. Thus, Equation 4.3 is the likelihood of observing the set of locations B in the reference population, under the assumption of conditional independence of events given the model. Locations that are often-visited in the population (e.g., airports, shopping malls, etc) receive high probability in this model, while rare locations (e.g., individual homes, areas without cellular service, etc) receive low probability. Figure 4.3a provides an illustration of a population model of this type, using the Twitter geolocation event data that we describe in more detail in Section 4.7.

The numerator of Equation 4.2, $f(B|A, H_s, I)$, is the probability of observing new location data B given that we have already seen location data A and under the hypothesis that A and B came from the same source. Effectively, it is a predictive density for geolocated events from A. We model this as a mixture of two densities where the first density corresponds to an individual component based on the locations of events in A, and the second density corresponds to the population component defined in Equation 4.3. See Figure 4.3 for an example of such a mixture model using the data presented in the motivating example of Section 4.1. This addresses two potential problems. First, if A has very little data this model will appear similar to the population model resulting in LR values near 1 and proper calibration. Second, it allows for the possibility that an individual would visit new locations in a second sample.

We use a non-parametric kernel density approach for the mixture model components in $f(B|A, H_s, I)$, defined as

$$\widehat{f}(B|A, H_s, I) = \prod_{j=1}^{n_b} f_{MKD}(s_j^b|A, \mathcal{D}, \alpha). \tag{4.4}$$

Here $f_{MKD}(\cdot|A, \mathcal{D}, \alpha)$ refers to a mixture of adaptive bandwith kernel density estimators [e.g., see Lichman and Smyth, 2014], defined as

$$f_{MKD}(s_j^b|A, \mathcal{D}, \alpha) = \alpha f_{KD}(s_j^b|A) + (1 - \alpha) f_{KD}(s_j^b|\,\mathcal{D}) \tag{4.5}$$

where $f_{KD}(\cdot|A)$ is a kernel density built on the unknown source data A, which we refer to as the individual component. The parameter $\alpha \in [0, 1]$ determines how much weight to put on the individual component $f_{KD}(\cdot|A)$ of the model relative to the population component $f_{KD}(\cdot|\,\mathcal{D})$. The choice of α is presented in Section 4.4.2. If the set of events B contains locations nearby to those in A, $\widehat{f}(B|A, H_s, I)$ will be large relative to $\widehat{f}(B|H_d, I)$ and the LR will have a value greater than 1 which indicates that A and B are likely to have been generated by the same individual.

4.4.1 Adaptive Bandwidth Kernel Density Estimators

In general, we follow the notation of Lichman and Smyth [2014] for our definition of kernel densities and mixtures of kernel densities. Assume that we are given a set of 2-dimensional points $s = (x, y)$ that represent the location of an event, denoted $\mathcal{D} = \{s_i : i = 1, \ldots, n_p\}$. Kernel density estimation (KDE) is a common choice for the non-parametric estimation of a bivariate probability density function f using this data. Given the bivariate Gaussian kernel function K and a bandwidth parameter h, we get the following bivariate KDE

$$f_{KD}(s|\,\mathcal{D}) = \frac{1}{n_p}\sum_{i=1}^{n_p} K(s, s_i|h) \tag{4.6}$$

$$K(s, s_i|h) = \frac{1}{2\pi h}\exp\left(-\frac{1}{2}(s - s_i)^T\Sigma_h^{-1}(s - s_i)\right) \tag{4.7}$$

$$\Sigma_h = \begin{pmatrix} h & 0 \\ 0 & h \end{pmatrix} \tag{4.8}$$

Thus the estimated density at s is the average of the kernels centered at the observations s_i and scaled by h across all n observations. KDEs are essentially a local smoothing method.

The choice of the kernel itself is not as important as the selection of the bandwidth h. As h decreases, the height of the peak at each observation increases resulting in undersmoothing.

As h increases, the height of the peak at each observation decreases and probability mass is pushed away from the observation resulting in oversmoothing. Geolocated event data is hard to model via a homogeneous bandwidth given the high density of events in urban areas and low density in sparsely populated areas. More appropriate for this data is an adaptive bandwidth method [Breiman et al., 1977] where h is replaced with a bandwidth that depends on the observation s_i

$$f_{KD}(s|\,\mathcal{D}) = \frac{1}{n_p}\sum_{i=1}^{n_p} K(s, s_i|h = h(s_i)). \tag{4.9}$$

A similar definition holds for $f_{KD}(s|A)$, where the KDE is computed over the n_a events in A. In Lichman and Smyth [2014], an adaptive bandwidth $h(s_i)$ determined from the geodesic distance from each event s_i to its 5th nearest neighbor was used, and the minimum bandwidth was set to 50 meters (i.e., $h(s_i) \geq 50$m for all i) to prevent overfitting that arises due to events occurring at the exact same location. Here we adopt the same values, but note that for other types of geolocation data other values may be more appropriate.

4.4.2 Choosing the Mixture Parameter

Given the adaptive bandwidth kernel density estimators $f_{KD}(\cdot|A)$ and $f_{KD}(\cdot|\,\mathcal{D})$ of Equation 4.5, the only remaining free parameter in the mixture KDE model for the numerator of the likelihood ratio is the mixture weight α. In general, α should be estimated by maximizing the likelihood of observed events in a validation data set using, for example, the Expectation-Maximization (EM) algorithm as done in Lichman and Smyth [2014]. However, as discussed in Section 2.4, in forensic applications obtaining a large enough sample from the relevant population for creating both the reference and validation sets can be difficult. Cross-validation can be used for assessing the out of sample performance of the LR using only reference data, but then we are left without validation data for fitting model parameters

such as the mixture weight. Therefore, we used hand-designed values for mixture weight that were chosen to have intuitive interpretations.

The first mixture weight that we investigated was a constant $\alpha = 0.80$ for all evidential samples (A, B). Under this choice, we assume that 80% of the time new events (e.g., events in B) are generated from the individual model (i.e., the adaptive bandwidth KDE built using events in A) and 20% of the time from the population model. Intuitively, a constant mixture weight implies that individuals are generally self-consistent over time (i.e., repeatedly visiting locations that they have previously been to) and occasionally generate new events in locations that are prevalent in the relevant population. However, this does not take into account the number of observed events for any particular individual. Namely, an individual with few observed events and another with many events will have the same amount of weight placed on the individual model. If a particular individual has very few events in both A and B that happen to occur close to one another spatially, the fixed mixing weight $\alpha = 0.80$ would result in a very large likelihood ratio. In this scenario, we would like the LR to show the uncertainty in the source propositions due to having few observations, and, therefore, be shrunk toward a neutral value of 1 indicating that both source hypotheses are equally likely. In order to achieve this behavior, mixture weights that are a function of the number of observed events in A were also considered.

Two forms of the mixture weight α as a function of the number of events in A were considered.

The first is a step function, $\alpha(n_a)$, defined by the following

$$\alpha(n_a) = \begin{cases} 0.05, & \text{for } n_a \leq 5 \\ 0.15, & \text{for } n_a \in (5, 10] \\ 0.40, & \text{for } n_a \in (10, 20] \\ 0.55, & \text{for } n_a \in (20, 50] \\ 0.70, & \text{for } n_a \in (50, 100] \\ 0.85, & \text{for } n_a > 100. \end{cases} \tag{4.10}$$

The values in Equation 4.10 were selected heuristically based on some familiarity with the data. The second is a parameterized function that was chosen to mirror the behavior of the step function, defined by

$$\alpha(n_a|\gamma, \rho, \phi) = \frac{1}{1 + \exp(-\gamma n_a)} - \frac{\rho}{\sqrt{n_a}} - \phi \tag{4.11}$$

where $\gamma = 0.02$, $\rho = 0.35$ and $\phi = 0.1$. As the number of events in A increases, the weight on the individual component increases as shown in Figure 4.4. The parametric function places more weight on the individual component than the step function when n_a is small (i.e., $n_a < 10$). The parametric function asymptotes at 0.9, while the step function reaches its maximum of 0.85 for $n_a \geq 100$. In between these extremes, the parametric function is essentially a smoothed version of the step function.

Alternative choices are also possible for the function defining the mixture weights. The settings for α in this section tend to work well for the Twitter and Gowalla event data sets presented in Sections 4.7 and 4.8.

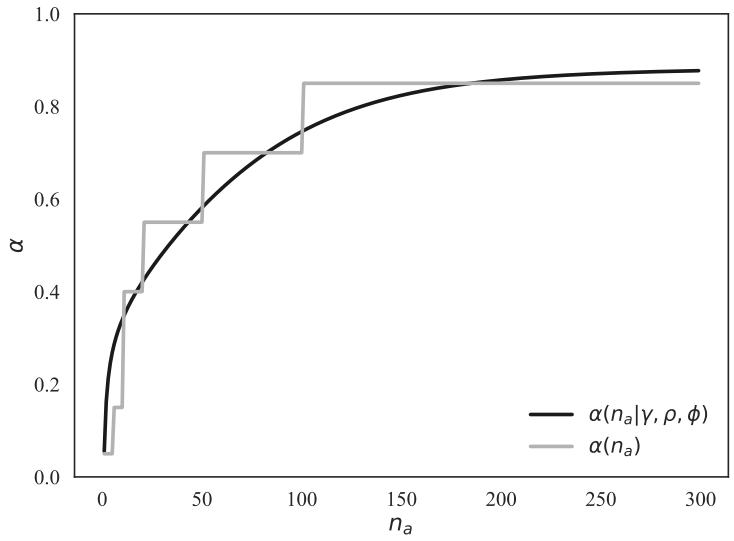


Figure 4.4: Mixture weight α as a function of the number of events in the unknown source sample n_a.

4.5 Score Functions for Geolocation Data

As an alternative to specifying the likelihood directly (as we did in Section 4.4), we also investigate score-based approaches based on similarities of sets of geolocated points. In terms of defining a suitable score $\Delta(A, B)$ for sets of locations A and B, there are a number of techniques that can be borrowed from the statistics literature on spatial point patterns. In general, they fall into two categories: distance-based and area-based techniques [Haggett et al., 1977]. Distance-based techniques use information on the spacing of points to characterize the pattern (typically, mean distance to the nearest neighboring point). Area-based techniques rely on characteristics of the frequency of observed points in subregions of the region under consideration. We investigate two different distance-based score functions $\Delta(A, B)$ to quantify the similarity of the points within the sets A and B and incorporate area-based information via various event-weighting strategies.

The two score functions we use are the mean nearest neighbor distance (denoted $\overline{D}_{min}$) and the earth mover's distance (denoted EMD), which both rely on computing the distance from

each event in B to the nearest neighboring event in A. Intuitively, we expect same-source pairs to contain events at locations near each other in the spatial region as individuals tend to be self-consistent (repeatedly generating events from the same locations over time). If events in B are spatially clustered among (i.e., "close to") events in A, then the score functions considered tend to be smaller than if the A and B events are generated independently and do not spatially cluster together.

To define the score functions, we first construct an inter-event distance matrix by measuring the geodesic distance [Karney, 2013] from each event in B to each event in A. Let $\boldsymbol{D} = [d_{jk}]$ represent the $n_b \times n_a$ distance matrix where each element $d_{jk} = d(s_j^b, s_k^a)$ denotes the geodesic distance between the position of the jth event in set B and the position of the kth event in set A.

4.5.1 Nearest Neighbor Distances

Treating each point in set B as the focus, we can compute the inter-event distance to its nearest neighbor in A. Let D_{min} represent the collection of the n_b nearest neighbor distances from B to A, and define it as follows

$$\begin{aligned} D_{min} &\equiv \left\{ d_j^{min} : j = 1, \ldots, n_b \right\} \\ \text{where } d_j^{min} &= \min_{k \in \{1,\ldots,n_a\}} d_{jk} \end{aligned} \tag{4.12}$$

If events of type B are spatially clustered among events of type A, then the nearest neighbor distances D_{min} tend to be smaller than if A and B events are generated independently and do not cluster together. A variety of characteristics of the distribution of nearest neighbor distances can be used as score functions $\Delta(A, B)$. We consider variants of the weighted

arithmetic average nearest neighbor distance from B to A, defined in general as

$$\overline{D}_{min}(B, A|\Omega^b) = \frac{\sum_{j=1}^{n_b} \omega_j^b d_j^{min}}{\sum_{j=1}^{n_b} \omega_j^b} \tag{4.13}$$

where $\Omega^b = \{\omega_j^b : j = 1, \ldots, n_b\}$ are weights assigned to each of the events in B. A discussion of the motivation for using weights and definitions of the various weighting strategies used here are provided in Sections 4.5.3 and 4.5.4.

Note that it is also possible to define a nearest neighbor distance from A to B. That distance would compute the nearest neighbor for each event of type A and weight these according to weights Ω^a. The asymmetry of the nearest neighbor distance is one motivation for seeking an alternative.

4.5.2 Earth Mover's Distance

The *earth mover's distance (EMD)*, or Wasserstein metric, is a measure of the distance between two probability distributions. To gain an intuition for the EMD, consider the problem of having multiple piles of earth of different sizes spread over some region that you wish to move into a collection of holes of different volumes in that same region. The EMD measures the least amount of "work" it takes to fill the holes with earth, where a unit of work consists of transporting a unit of earth by a unit of ground distance. For the problem at hand, we can think of the piles of earth as one point pattern (B) and the holes as the other (A). EMD has been widely used as a general approach for measuring distances between two sets as a function of the distance between elements of the sets [e.g., Rubner et al., 1998; Cohen, 1999]. We develop the use of EMD in the context of measuring the similarity of spatial point patterns.

Computing the EMD is based on a solution to the transportation problem [Hitchcock, 1941].

The first step is to find a flow $\boldsymbol{F'} = [f'_{jk}]$, where f'_{jk} is the flow (or amount of mass) moved from s^b_j to s^a_k, that minimizes the overall cost

$$\boldsymbol{F'} = \arg\min_{[f_{jk}]} \sum_{j=1}^{n_b} \sum_{k=1}^{n_a} f_{jk} d_{jk} \tag{4.14}$$

subject to the following constraints

$$f_{jk} \geq 0 \quad j \in \{1, \ldots, n_b\}, k \in \{1, \ldots, n_a\} \tag{4.15}$$

$$\sum_{k=1}^{n_a} f_{jk} \leq \omega^b_j \quad j \in \{1, \ldots, n_b\} \tag{4.16}$$

$$\sum_{j=1}^{n_b} f_{jk} \leq \omega^a_k \quad k \in \{1, \ldots, n_a\} \tag{4.17}$$

$$\sum_{j=1}^{n_b} \sum_{k=1}^{n_a} f_{jk} = \min \left(\sum_{j=1}^{n_b} \omega^b_j, \sum_{k=1}^{n_a} \omega^a_k \right). \tag{4.18}$$

where in principle the weights Ω^a and Ω^b are the same as those used in Equation 4.13. The first constraint (4.15) restricts the flow of mass from B to A and not vice versa. The next two constraints (4.16, 4.17) limit the amount of mass that can be sent from points in B to their weights, and the points in A receive no more mass than their corresponding weights. The last constraint (4.18) ensures the total amount of mass moved is equal to that of the lighter distribution, and is referred to as the total flow. Given the solution $\boldsymbol{F'}$ that minimizes (4.14), define the score function $\Delta(A, B)$ based on the earth mover's distance as the cost normalized by the total flow

$$EMD(B, A|\Omega) = \frac{\sum_{j=1}^{n_b} \sum_{k=1}^{n_a} f'_{jk} d_{jk}}{\sum_{j=1}^{n_b} \sum_{k=1}^{n_a} f'_{jk}} \tag{4.19}$$

where $\Omega = \{\Omega^a, \Omega^b\}$.

Note that the earth mover's distance is a metric when the distance between the points is a metric and the total weights of the point patterns are equal [Cohen, 1999, page 63]. Since

geodesic distance is a metric, the first property is satisfied. We enforce that the weights sum to 1 for both sets A and B. Therefore, the earth mover's distance considered is a metric which implies that $EMD(B, A|\Omega) = EMD(A, B|\Omega)$. This simplifies computation and results in the same conclusions being drawn regardless of which pattern you consider as the focus of analysis.

4.5.3 Geoparcel Data

Geolocated event data is quite useful, but additional information can be incorporated in the score functions (i.e., via the weights in the definitions of $\overline{D}_{min}$ and EMD) if we also consider spatial properties of locations at which the events occur. High-traffic locations, such as shopping malls, theme parks and stadiums, will have a high likelihood of appearing in any randomly selected point pattern and thus make patterns generated by different individuals look alike. Conversely, less common locations such as homes are highly unlikely to appear in multiple point patterns unless those patterns were generated by the same individual or someone close to him or her.

One option for incorporating spatial information is to partition the spatial region into a regular grid of disjoint cells, and compute population frequencies of events in each grid cell. However, defining the grid is a difficult problem as the result can be highly arbitrary since locations very rarely fall perfectly into a grid. Further, the spatial resolution of the grid is proportional to the amount of events in each cell—too small of a grid size results in highly sparse data. Given these limitations, we chose to use geoparcel information. Geoparcels are disjoint polygons (or parcels) that partition a spatial region where each individual parcel represents a specific property. The parcels vary in size and shape depending on the function of the property, solving the issues posed by using a grid. Within each parcel, we can use the frequency of visits to that particular location as a way to weight events as discussed in the

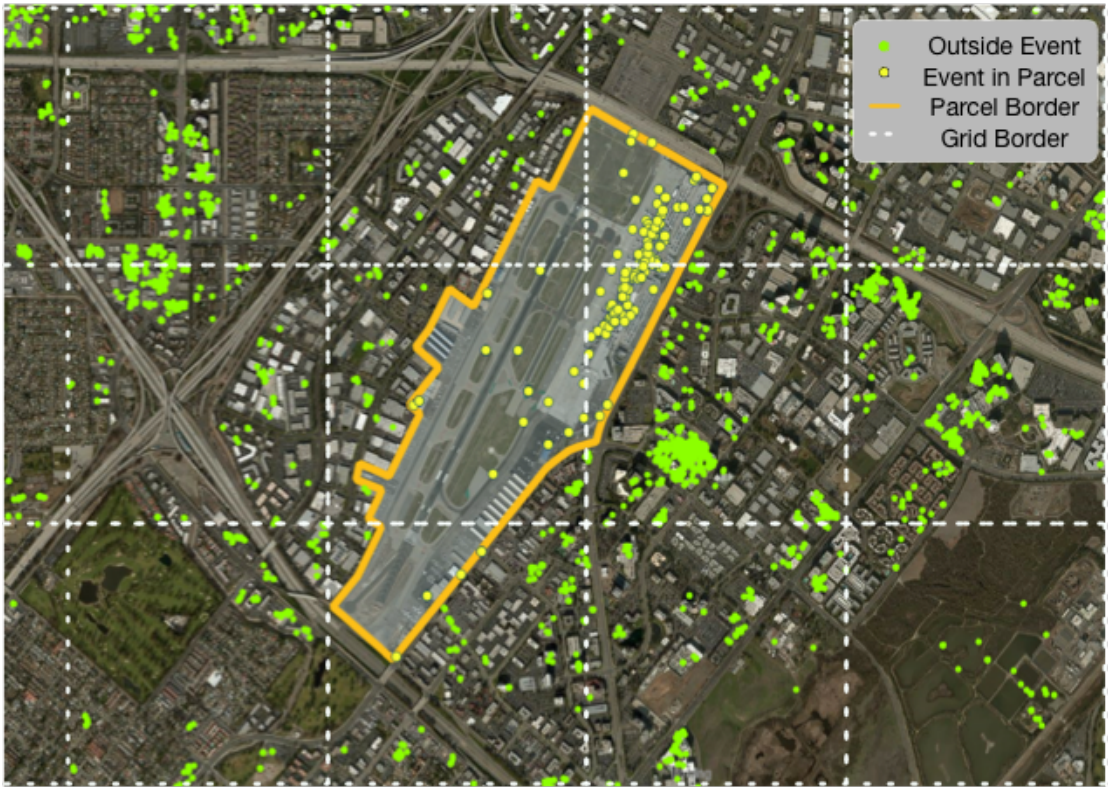


Figure 4.5: Area around John Wayne Airport (SNA) in Orange County, California, highlighting the parcel corresponding to the airport and Twitter events in the region. Figure credit Lichman [2017].

following section. See Figure 4.5 for an example of a parcel and a comparison to a grid-based approach.

4.5.4 Weighting Events

In our definitions of score functions $\Delta(A, B)$ for spatial point patterns in Equations 4.13 and 4.19, we require weights for each event. We consider three different weighting schemes that rely upon the geoparcel data discussed in the previous section.[3] The weights are defined for events in point pattern B, but similar definitions hold for events in A. All weights are normalized for each point pattern, i.e., $\sum_{j=1}^{n_b} \omega_j^b = 1$.

[3]If geoparcel data is not available, one could use grids or another data-adaptive partition of the space.

1. *Uniform.* Let $\omega_j^b = n_b^{-1}$ for $j = 1, \ldots, n_b$. Under uniform weighting, Equation 4.13 simplifies to the unweighted mean nearest neighbor distance. Furthermore, uniform weights result in the empirical distribution for each point pattern being used as the relevant distribution in the earth mover's distance calculation. The uniform weighting scheme is the most naive method, and requires no geoparcel data.

2. *Location Visits.* Define the weight for each event as a function of the number of visits occurring at the location (geoparcel) of that event across the reference population. Namely,

$$\omega_j^b \propto \left[n_{vis}(\ell(s_j^b))\right]^{-1} \tag{4.20}$$

   where $n_{vis}(\ell)$ is the number of visits at location ℓ, in this case the geoparcel in which the jth event in B occurred.

3. *Location Accounts.* Define the weight for each event as a function of the number of unique accounts in the reference population with at least one visit at the location of that event. Namely,

$$\omega_j^b \propto \left[n_{acc}(\ell(s_j^b))\right]^{-1} \tag{4.21}$$

   where $n_{acc}(\ell)$ is the number of unique accounts with at least one visit at location ℓ, in this case the geoparcel in which the jth event in B occurred.

Both location-based weighting schemes attempt to address high traffic locations by down-weighting their contribution to the score function. Specifically, in some spatial regions, a small subset of parcels can be responsible for a large fraction of activity. At such locations, it is highly likely that any randomly-selected account will generate an event there. The location-based weighting schemes above down-weights events from such parcels, placing more weight

on events at rarer locations such as homes.

4.6 Score-based Techniques

Given one of the score functions discussed in the previous section, the strength of evidence can be quantified via the techniques discussed in Chapter 3. In this section, I briefly demonstrate how the score-based likelihood ratio and coincidental match probability are computed for the mean nearest neighbor distance $\overline{D}_{min}$ score function. The same arguments hold for the earth mover's distance and any of the various weighting schemes.

The reference data sets $\mathcal{D}_s$ and $\mathcal{D}_d$ used below are constructed via leave-pairs-out cross-validation as described in Section 2.4.3.

4.6.1 Score-based Likelihood Ratio

Given the observed score $\delta^* = \overline{D}_{min}(B^*, A^*|\Omega^b)$, the estimator of the score-based likelihood ratio in Equation 3.3 becomes

$$\widehat{SLR}_{\overline{D}_{min}} = \frac{\widehat{g}(\overline{D}_{min}(B, A|\Omega^b) = \overline{D}_{min}(B^*, A^*|\Omega^b)|\{\overline{D}_{min}(B, A|\Omega^b) : (A, B) \in \mathcal{D}_s\})}{\widehat{g}(\overline{D}_{min}(B, A|\Omega^b) = \overline{D}_{min}(B^*, A^*|\Omega^b)|\{\overline{D}_{min}(B, A|\Omega^b) : (A, B) \in \mathcal{D}_d\})} \quad (4.22)$$

where $\widehat{g}$ is a kernel density estimator with a Gaussian kernel and rule-of-thumb bandwidth [Scott, 1992].[4] We explicitly condition on the reference sets $\mathcal{D}_s$ and $\mathcal{D}_d$ because the score values for the point patterns in these sets along with the kernel density parameters fully specify the estimated density. Further, the reference data sets fully encompass the relevant source hypothesis and background information.

[4] All of the scores we work with in this chapter are bounded (either above, below, or both), and an unconstrained KDE will push probability mass outside these bounds (e.g., below 0 for inter-event time score functions). More sophisticated methods could be used to estimate these densities, but for simplicity and computational efficiency we used a generic KDE method.

4.6.2 Coincidental Match Probability

Given the observed score $\delta^* = \overline{D}_{min}(B^*, A^*|\Omega^b)$, the estimator of the coincidental match probability in Equation 3.8 becomes

$$\widehat{CMP}_{\overline{D}_{min}} = \frac{1}{N_d} \sum_{(A,B)\in\mathcal{D}_d} \mathbb{I}\left(\overline{D}_{min}(B, A|\Omega^b) < \overline{D}_{min}(B^*, A^*|\Omega^b)\right) \tag{4.23}$$

where N_d is the number of evidential samples in the different-source reference data set $\mathcal{D}_d$.

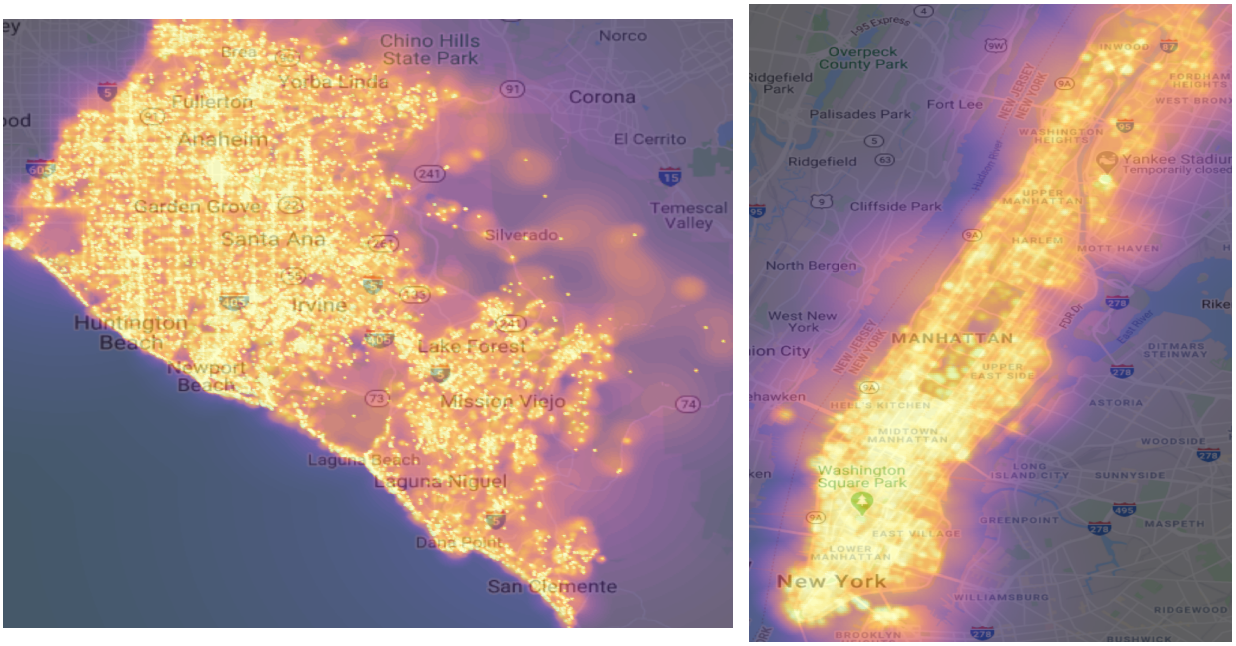
4.7 Case Study—Twitter Data

We used geolocation datasets of Twitter events to evaluate our proposed approaches. Twitter, a popular social media and microblogging service, provides a useful publicly accessible[5] source of user-event data that, given certain account configurations, exposes the geolocation of each event generated by that account. This data can be thought of as a subset of data collected from a given mobile device during a forensic investigation and is sufficient for illustrating our methods.

4.7.1 Event Data

We consider two spatial regions: Orange County, California, and the Manhattan borough of New York City. The data was collected from May 2015 to February 2016, selecting only events (tweets) with GPS coordinates from public accounts. Each event is composed of tuples of the following form:

[5]Note that while the data is publicly available via Twitter's API (`https://developer.twitter.com/en/docs/tweets/filter-realtime/overview`), the terms of use require that collected data sets cannot be shared amongst researchers.



(a) Orange County

(b) New York

Figure 4.6: Adaptive bandwidth KDE for the population data $\mathcal{D}$ of Twitter visits. (a) Orange County, CA; (b) New York, NY.

`<account_id, longitude, latitude, timestamp>`.

Thus, for any given account we have a set of geolocated events occurring in some bounded region. Twitter data from an account often consists of bursts of events from the same location within a short period of time. To handle this burstiness we replace such repeated events with a single effective event, which we refer to as a *visit*. Visits are defined as a set of events occurring within the same hour and within 50 meters of each other, and the first event from each visit is kept. Table 4.1 provides summary statistics for the Twitter data before and after filtering for visits. The visit data in this table is referred to as the population data, and was used for constructing the reference population $\mathcal{D}$ discussed in Section 4.4. Figure 4.6 depicts the adaptive bandwidth KDE for the population component of both spatial regions used in the construction of the LR.

To generate the spatial event data for our experiments we filtered the data based on sequential time periods of activity. Accounts with at least 1 visit per month in each of the first 2

Region	Days	Accounts	Events	Visits
OC	240	103,271	655,917 (6.4)	545,697 (5.3)
NY	239	194,224	1,162,871 (6.0)	989,494 (5.1)

Table 4.1: Number of observed days, accounts, events and visits for the Twitter data sets. Average number per account denoted in parentheses.

months were considered. For each account, we defined the sets of locations A and B to be all geolocated events in the first or second month, respectively (i.e., A and B correspond to the events in two consecutive time windows). Table 4.2 contains summary statistics for the Twitter data used in the analysis.

Region	Accounts	Visits in A	Visits in B
OC	6,714	44,310 (6.6)	38,697 (5.8)
NY	13,523	72,799 (5.4)	65,852 (4.9)

Table 4.2: Number of observed accounts and visits for the Twitter data sets used in the analysis. Average number per account denoted in parentheses.

4.7.2 Geoparcel Data

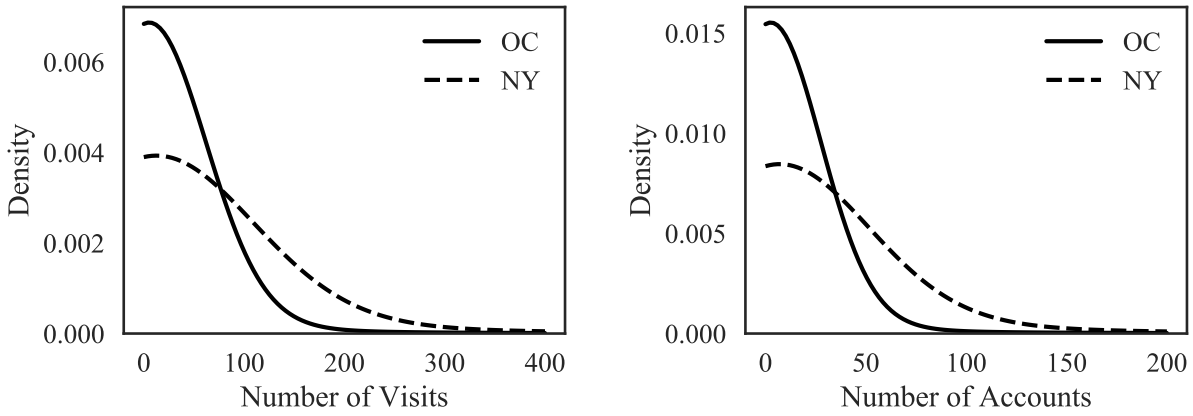
We use the same publicly available geoparcel data as Kotzias et al. [2018] for defining the weights used in the score functions. The 32,978 parcels for Orange County were collected from the Southern California Association of Government website.[6] The 21,312 parcels for New York were collected via the OpenStreetMap API.[7] Both the OC and NY data sets exhibit long-tailed distributions for the number of visits and number of unique accounts with at least one visit in each parcel, as shown in Table 4.3 and Figure 4.7. On average, parcels in New York have more visits and unique accounts than parcels in Orange County.

[6] `https://www.scag.ca.gov/`

[7] `https://wiki.openstreetmap.org/wiki/API`

Region	Type	Mean	Med.	75^{th}%ile	Max
OC	Visits	16.5	2	5	72,290
OC	Accounts	7.9	1	2	30,874
NY	Visits	46.4	4	16	77,760
NY	Accounts	26.8	3	10	25,775

Table 4.3: Summary statistics for the distribution of number of visits (Type "Visits") and unique accounts with at least one visit (Type "Accounts") in each parcel computed from the full population data in Table 4.1. The minimum and 25th percentile are 1 for all cases.



(a) Visits

(b) Accounts

Figure 4.7: Density estimate of the number of parcels versus (a) the number of visits in the parcel, and (b) the number of unique accounts in that parcel. Note that both figures are right-truncated due to the extremely long tails.

4.7.3 Results

For both the Orange County and New York regions, we compared the likelihood-ratio and the score-based likelihood-ratio techniques in terms of their effectiveness in quantifying the strength of evidence for pairs of sets of locations A and B. For computational efficiency we included all same-source pairs (6,714 in OC and 13,523 in NY) and a stratified random sample of different-source pairs. The sampling was stratified by the number of visits in each pattern, n_a and n_b, because the data is highly skewed towards a small number of visit events per individual and we wanted to assess performance of the methods under varying amounts of data. The strata correspond to all $3 \times 3 = 9$ combinations of 1 visit, between

2 and 19 visits, and 20 or more visits for n_a and n_b. 1,000 different-source pairs in each strata were randomly sampled, resulting in 9,000 total different-source pairs in each region. Leave-pairs-out cross-validation was used to estimate the LR, SLR and CMP for each piece of evidence.

Motivating Example

We begin the exploration of the results by revisiting the motivating example in Figure 4.1 of Section 4.1. Recall that the investigator was given one set of locations from an unknown source, A, as well as sets of locations from two known sources, B_1 and B_2. She was tasked with assessing the probative value of each pair of evidence—(A, B_1) and (A, B_2)—in order to determine the likelihood that either pair was generated by the same source. Using the likelihood ratio approach with fixed mixing weights, the LR for (A, B_1) was approximately 1,137. Following the verbal equivalents provided in Section 2.3.3, the investigator would conclude that there is strong support that A and B_1 were generated by the same individual. For the second pair, (A, B_2), the LR was approximately 2.8e-28 which would lead the investigator to conclude that the individual that generated B_2 could be excluded as the source of A.

Overall Results

As discussed in Section 2.5, one way to assess the performance of evidence evaluation techniques is to threshold the values to obtain binary decisions of same- or different-source and compare these binary decisions to the known ground truth to compute true and false positive rates. Using likelihood ratios with a threshold of 1, corresponding to the data being equally likely to have been generated under either hypothesis, we classify pairs with LR greater than 1 as same-source and those with LR less than 1 as different-source. We can then compare the true and false positive rates for each choice of the mixing weight α. Table 4.4 provides these

rates (listed as TP@1 and FP@1, respectively) along with the AUC. In both spatial regions the LR had similar performance, with the highest true positive rate and AUC belonging to the varying mixing weight approach and the lowest false positive rate for fixed α.

Region	Weight	TP@1	FP@1	AUC
OC	0.80	0.340	**0.026**	0.787
	$\alpha(n_a)$	**0.380**	0.038	**0.845**
	$\alpha(n_a\|\gamma,\rho,\phi)$	0.375	0.037	0.817
NY	0.80	0.251	**0.067**	0.711
	$\alpha(n_a)$	**0.285**	0.089	**0.768**
	$\alpha(n_a\|\gamma,\rho,\phi)$	0.282	0.088	0.734

Table 4.4: Performance of a classifier based on LR for the Twitter data.

Similarly, using SLRs with a threshold of 1 we can compare the true and false positive rates for each score function. Table 4.5 provides these rates along with the AUC. In both spatial regions, the SLR built on the EMD score function tends to outperform that using $\overline{D}_{min}$ within a given weighting scheme across TP, FP and AUC. Uniform weights tend to out-perform both the account and visit weighting schemes in terms of TP and AUC, but not FP. In Orange County account weights yield the lowest FP rate, while in NY both account and visit weights yield similarly low FP rates within a given score function.

Region	Δ	Weights	TP@1	FP@1	AUC
OC	$\overline{D}_{min}$	Uniform	0.628	0.202	0.768
	$\overline{D}_{min}$	Account	0.610	0.171	0.774
	$\overline{D}_{min}$	Visit	0.611	0.180	0.768
	EMD	Uniform	**0.654**	0.197	**0.790**
	EMD	Account	0.614	**0.162**	0.783
	EMD	Visit	0.602	0.169	0.774
NY	$\overline{D}_{min}$	Uniform	0.508	0.287	0.656
	$\overline{D}_{min}$	Account	0.494	0.254	0.666
	$\overline{D}_{min}$	Visit	0.493	0.257	0.663
	EMD	Uniform	**0.530**	0.253	**0.686**
	EMD	Account	0.511	0.235	0.685
	EMD	Visit	0.504	**0.234**	0.679

Table 4.5: Performance of a classifier based on SLR_Δ for the Twitter data.

For the coincidental match probability, the false positive rate is fixed at the choice of thresh-

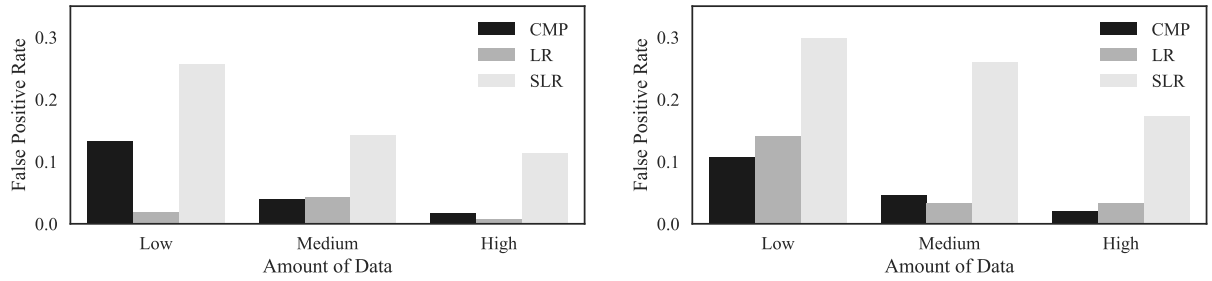
old as discussed in Section 3.2.2. Thus, only the true positive rate and AUC are indicative of its performance as a classifier. Table 4.6 contains these values for thresholds of 0.05 and 0.01 (listed as TP@0.05 and TP@0.01, respectively). The CMP performs similarly in both OC and NY, with the uniform weighted EMD having the highest AUC. Further, the account weighted score functions have the highest true positive rates for the choices of thresholds presented.

Region	Δ	Weights	TP@0.05	TP@0.01	AUC
OC	$\overline{D}_{min}$	Uniform	0.389	0.187	0.771
	$\overline{D}_{min}$	Account	0.441	**0.236**	0.776
	$\overline{D}_{min}$	Visit	0.415	0.209	0.771
	EMD	Uniform	0.397	0.154	**0.791**
	EMD	Account	**0.448**	0.208	0.784
	EMD	Visit	0.425	0.182	0.775
NY	$\overline{D}_{min}$	Uniform	0.242	0.153	0.656
	$\overline{D}_{min}$	Account	0.269	**0.186**	0.667
	$\overline{D}_{min}$	Visit	0.264	0.179	0.665
	EMD	Uniform	0.265	0.139	**0.687**
	EMD	Account	**0.283**	0.161	0.686
	EMD	Visit	0.276	0.156	0.681

Table 4.6: Performance of a classifier based on CMP_{Δ} for the Twitter data.

Regardless of the region considered and choice of α, Δ and weighting scheme used, the likelihood ratio approach outperforms the score-based approaches in terms of AUC (e.g., for the Orange County region the best AUC for the LR in Table 4.4 is 0.845, the SLR in Table 4.5 is 0.790, and the CMP in Table 4.6 is 0.791) and false positive rate (e.g., for the OC the best FPR for the LR is 0.026, the SLR is 0.162, and the CMP is 0.236). While the SLR and CMP have larger true positive rates than the LR (e.g., for OC the best TPR for the SLR is 0.654, the CMP is 0.448, and the LR is 0.380), the cost is FP rates that are significantly larger. This phenomenon not only appears in the overall results, but also when considering performance of the techniques within the strata. Figure 4.8 depicts the FP rate of the approaches versus the amount of data in the sets A and B (corresponding to a selection of 3 of the 9 strata used in sampling) for both spatial regions. Regardless of

technique, as the amount of data increases the false positive rate decreases. The SLR has much higher FP rate than the LR and CMP across all data regimes.[8]



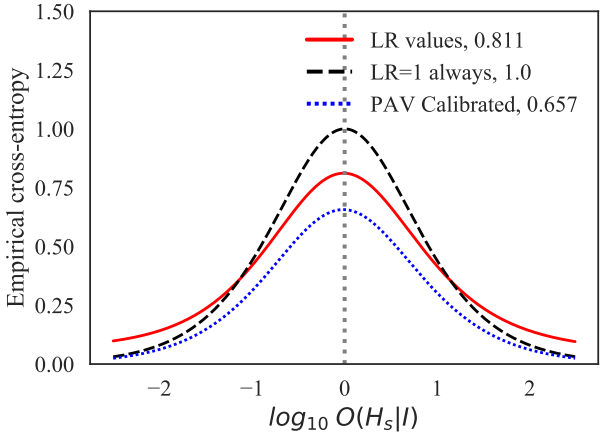
(a) Orange County

(b) New York

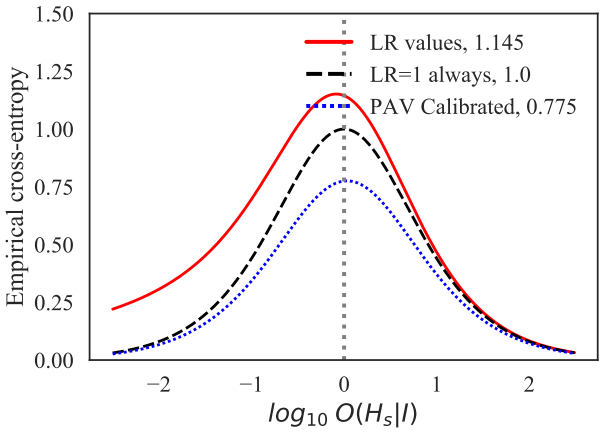
Figure 4.8: False positive rate of each method under different data regimes in (a) Orange County, and (b) New York. Low corresponds to 1 event in each of A and B, medium is between 2 and 19 events, and high is 20 or more events. Showing results for fixed α in the LR approach and the account weighted EMD for the SLR and CMP approaches. Thresholds are 1 for the LR and SLR and 0.05 for the CMP. Trends are similar for other choices of mixing weight and score function (see Figures A.2 and A.6 in Appendix A).

In addition to classification performance, we can also consider the information-theoretic value of the methods as discussed in Section 2.6. Figure 4.9 contains ECE plots for a selection of LR and SLR methods in both regions. For all ECE plots, see Appendix A. For the Orange County region, the best performing method is the LR with nonparametric mixture weight $\alpha(n_A)$ since it has a lower value of ECE (solid red curve in Figure 4.9a) for most prior log odds compared to other LR approaches or score-based approaches (e.g., solid red curve of Figure 4.9b). However, the LR with $\alpha(n_A)$ performs poorly for large magnitudes of the prior logs odds as shown in the tails of the ECE curve being larger than the neutral method that treats the $LR = 1$ for all evidence. The LR has better discriminating power than the SLR, which is shown by the lower C_{llr} and PAV calibrated ECE (blue dotted curve) in Figure 4.9a compared to the same in 4.9b. Overall, both the LR and SLR reduce the uncertainty in the source propositions after calibration and, therefore, are useful tools to aid the decision maker.

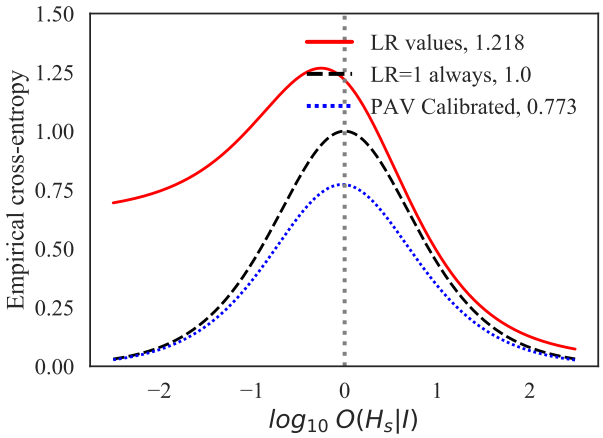
[8]While the overall false positive rate of the CMP is set by the choice of threshold, the FP rate within each strata can vary.



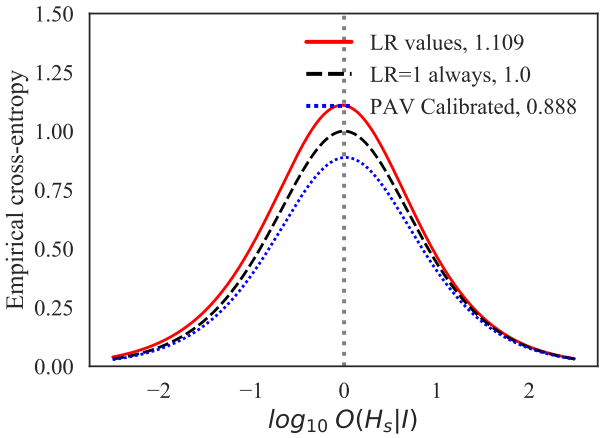
(a) OC: LR with $\alpha(n_a)$



(b) OC: SLR_{EMD} with Account Weights



(c) NY: LR with $\alpha(n_a)$



(d) NY: SLR_{EMD} with Account Weights

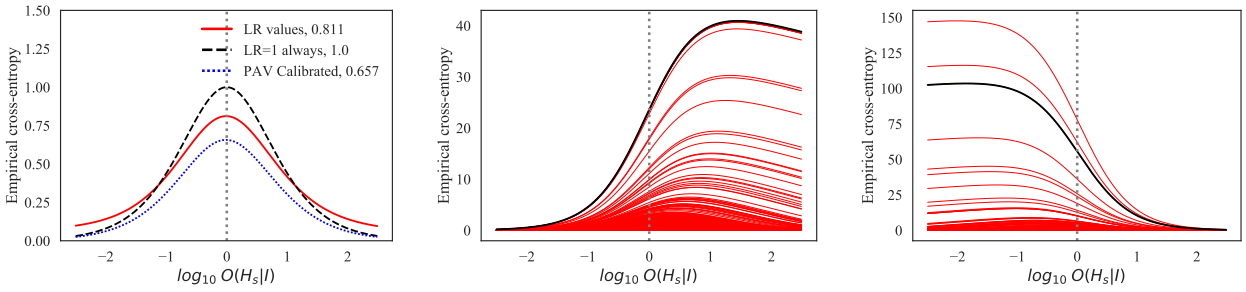
Figure 4.9: Empirical cross-entropy (ECE) plots for a selection of the LR and SLR methods applied on the Twitter data. C_{llr} values are provided in the legend. (a) Orange County likelihood ratio with the nonparametric mixing weight $\alpha(n_a)$; (b) Orange County score-based likelihood ratio using the earth mover's distance and account weighting scheme; (c) New York LR with $\alpha(n_a)$; (d) New York SLR using the EMD and account weights.

For the New York region, the best performing method is the SLR using the earth mover's distance and account weights since it has a lower value of ECE (solid red curve in Figure 4.9d) for all prior log odds compared to other SLR approaches and LR approaches (e.g., solid red curve of Figure 4.9c). However, the SLR has worse discriminating power, as shown by the PAV calibrated ECE (and therefore the C_{llr}) being larger for the SLR than that of the LR. This points to a larger phenomenon regarding the trade off between discrimination power

and calibration. Namely, the lower-dimensional SLR is better calibrated than the LR, but the cost is worse discrimination power.

The classification and information-theoretic performance of the evidence evaluation approaches in both regions together show the strengths and weaknesses of the techniques. In terms of classification performance, the LR approach is clearly the favorite with higher AUC and lower calibrated ECE values than the SLR regardless of the region or choice of mixture weight and score function. Furthermore, the LR is better behaved for "weak" evidence, or exemplars whose strength of evidence is near 1. This trend is apparent in the better false positive rate for the LR compared to both the SLR and CMP for natural threshold choices (1 for likelihood ratios and 0.05 or 0.01 for the CMP). However, the LR suffers from more calibration issues than the SLR, as shown by the tail performance in Figure 4.9a and the entirety of Figure 4.9c. This is due to a small subset of the evaluated same-source (or different-source) evidence being incorrectly classified with very small (or large) LR values, which I discuss in more detail in the following section.

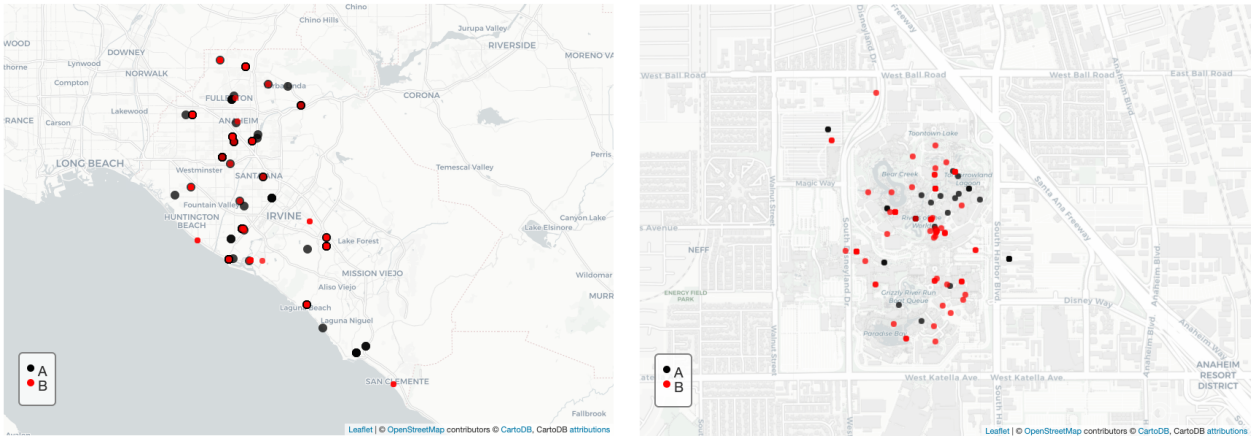
4.7.4 Error Analysis



(a) ECE Plot (b) H_s true (c) H_d true

Figure 4.10: Empirical cross-entropy (ECE) plot of the likelihood ratio approach with the nonparametric weighting function $\alpha(n_a)$ for the Orange County Twitter data. (a) Standard ECE plot; (b) ECE plot contribution for each piece of same-source evidence. The black curve is from the evidence shown in Figure 4.11a; (c) ECE plot contribution for each piece of different-source evidence. The black curve is from the evidence shown in Figure 4.11b. Note the different scales on the y-axes.

To illustrate the calibration issues of the likelihood ratio approach, I focus on an error analysis of the LR with nonparametric mixture weight $\alpha(n_A)$ in the Orange County data. Similar arguments apply for other LR and SLR techniques and to the New York region. Of primary interest is determining the cause of poor calibration of the LR for large magnitudes of the prior odds (i.e., the tail regions of Figure 4.9a). Ramos et al. [2013] showed that one can plot the individual contributions of LR values to the ECE for each piece of evidence to identify potential outliers in the validation set and to investigate potential issues with the evidence evaluation method. Figure 4.10 shows such plots for the LR approach with $\alpha(n_a)$ in the Orange County region. Figure 4.10a shows the ECE plot, and Figures 4.10b,c show the individual contributions to the ECE plot for known same-source and different-source evidence, respectively. Relatively few pieces of evidence are contributing to the poor tail performance of the technique (i.e., the curves with very large ECE values), as made apparent by these figures. These correspond to misclassified evidence, i.e., same-source (or different-source) evidence with very small (or large) LR values.



(a) Misclassified same-source pair (b) Misclassified different-source pair

Figure 4.11: Examples of misclassified evidence in the Orange County Twitter data corresponding to the highlighted individual ECE contributions in Figure 4.10. (a) A same-source pair with $\log(LR) \approx -33$. Due to overplotting, the point size of the locations in A was increased. (b) A different-source pair with $\log(LR) \approx 77$.

We now explore an example of a same-source pair and different-source pair to better understand why they result in such large ECE values. Figure 4.11 shows the location data

for these examples corresponding to the highlighted individual ECE curves (black curves) in Figure 4.10. The same-source pair in Figure 4.11a has a corresponding log likelihood ratio of approximately -33, which strongly supports the different-source hypothesis. This is somewhat surprising, as the locations of events in A and B have a good amount of overlap. Further investigation, however, showed that 5 other accounts in the reference data used to build the population component (i.e., the denominator and part of the mixture model in the numerator of the LR) had the same event data as this example (i.e., the event times and locations were exactly the same). The activity corresponds to multiple Twitter "bot" accounts that reproduce the same events. Thus, each of these events contributes 5 times to the population component (in both the numerator and denominator) compared to just once for the individual component in the numerator. The mixture weight α for this example places 70% of the weight on the individual component in the numerator of the LR, resulting in the different-source model being much more likely than the same-source model.

Other misclassified same-source evidence is typically caused by the two location patterns not being consistent between time periods, i.e., the event locations in A and B do not occur near each other spatially. Such scenarios arise due to natural variability in human behavior, e.g., when an individual's underlying generative model for event locations shifts or encounters a changepoint. It would be unreasonable if the evidence evaluation method were to quantify the strength of evidence in a manner that results in the investigator concluding such location patterns were generated by the same individual. Therefore, the proposed techniques handle this misleading evidence in a reasonable manner by producing LR values that suggest the generative mechanisms of A and B are different.

The different-source pair in Figure 4.11b has a corresponding log likelihood ratio of approximately 77, which strongly supports the same-source hypothesis. Further investigation of this example shows all events in B occurred very close spatially to events in A (upon which the individual component of the same-source model is built), and therefore the evidence

appears to be from the same source. In geolocated event data from social networks like Twitter, this is a somewhat common occurrence at very popular locations. In this example, all events were in or near Disneyland Resort. Thus, the LR is not able to properly assess the strength of evidence when *all* events occur at these common locations. Furthermore, the SLR for this evidence also failed to accurately assess this example (i.e., all SLR values were greater than 1 regardless of score function). However, the CMP with either location weighting scheme (account or visits) resulted in values of approximately 0.015, which would result in the investigator concluding that this evidence was from different sources using a threshold of 0.01. This shows that the same-source models in the numerator of either the LR or SLR over-value such examples. More investigation is needed to determine suitable models for overcoming this issue.

4.7.5 Discussion of Twitter Results

Overall, the likelihood ratio approach and score-based approaches worked well in quantifying the strength of evidence in the Twitter data. In terms of classification performance, the LR outperforms both the SLR and CMP in both regions. However, it suffers from calibration issues when the LR values are either very small or very large. The score-based approaches mitigate the calibration issues by reducing the dimensionality of the evidence. Investigators need to take these factors into account when analyzing geolocation evidence. As shown in the error analysis, the investigator must use their intuition in order to determine whether or not the resulting strength of evidence makes intuitive sense for the evidence at hand.

4.8 Case Study—Gowalla Data

Another source of geolocated event data was used to evaluate our proposed approaches. Gowalla, a location-based social networking application, allowed users to share check-ins at various locations around the world. The check-ins provide not only the physical location of the device when the check-in was recorded, but also what was at that particular location allowing for the distinction between a check-in to an office on the second floor and a check-in to a coffee shop on the 1st floor of the same building. Check-ins to location-based social networks are usually sporadic [Noulas et al., 2011], but work well to illustrate our approaches for quantifying the strength of geolocation evidence.

4.8.1 Data

The Gowalla data used was the same data that was studied by Cho et al. [2011], but restricted to southern California. The data was collected from March 2009 to October 2010, where each check-in event was composed of tuples of the form:

`<account_id, longitude, latitude, timestamp, location_id>`.

Unlike the Twitter data discussed in the previous section, the Gowalla data has categorical location information available so that no additional source of parcel data is required. The locations, however, are more specific than parcels and represent individual businesses or public locations, i.e., there could be many locations in any particular parcel. For consistency, we will refer to the categorical locations as parcels.

Similar to the Twitter data, we focused on visits and not events themselves. For the Gowalla data, visits are defined as a set of events occurring within the same hour, parcel (as determined by Gowalla's `location_id` feature) and within 50 meters of each other. Each visit is

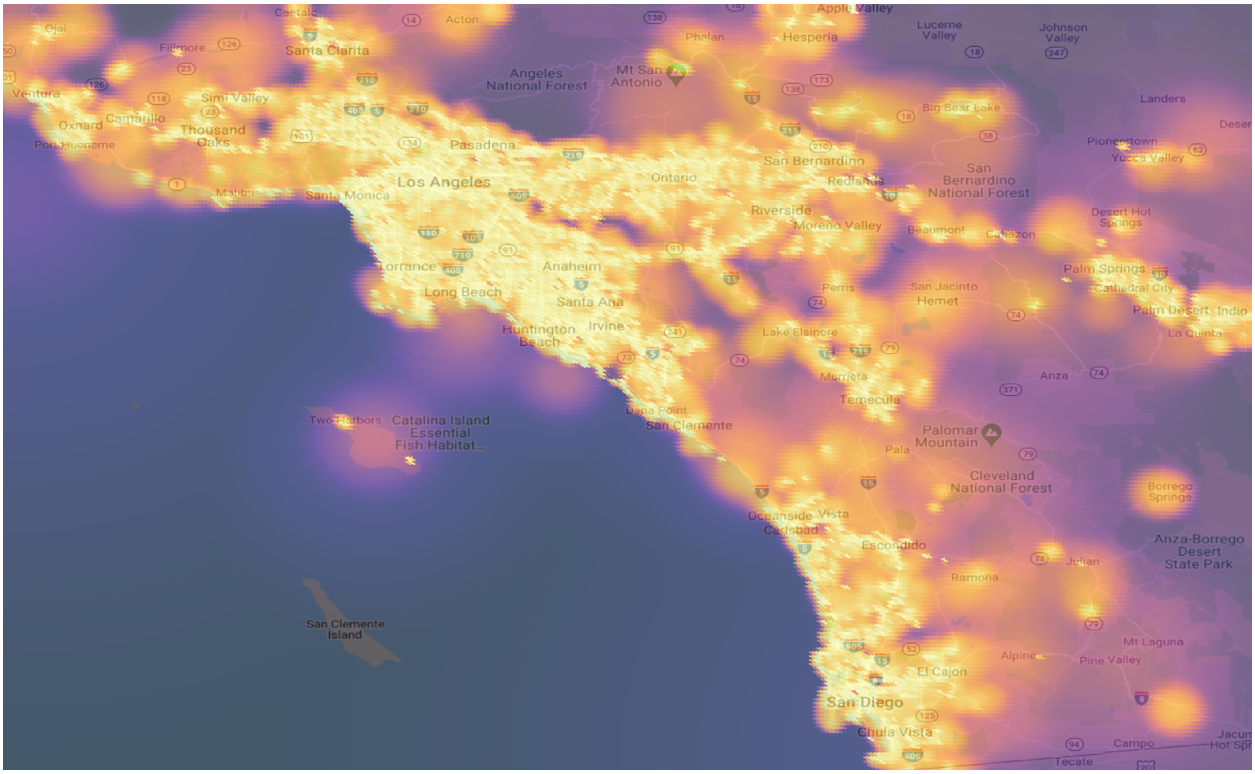


Figure 4.12: Adaptive bandwidth KDE for the population data $\mathcal{D}$ of Gowalla check-in events in Southern California.

treated as a single effective event (the first event from each visit is kept). Table 4.7 provides summary statistics for the Gowalla data before and after filtering for visits. Again, the visit data in this table is referred to as the population data, and was used for constructing the reference population $\mathcal{D}$ discussed in Section 4.4. Figure 4.12 depicts the adaptive bandwidth KDE for the population component used in the construction of the LR. The Gowalla check-in data is much less bursty than the Twitter event data, as is shown by the 0.35% reduction from events to visits compared to 16.8% and 14.9% reductions for the Orange County and New York regions of Twitter data, respectively.

Days	Accounts	Events	Visits
451	8,459	276,924 (32.7)	275,956 (32.6)

Table 4.7: Number of observed days, accounts, events and visits for the Gowalla data. Average number per account denoted in parentheses.

There were 52,097 parcels with at least one check-in in the population data. Similar to the

Twitter data, the Gowalla parcel data exhibits long-tailed distributions for the number of visits and number of unique accounts with at least one visit in each parcel, as shown in Table 4.8 and Figure 4.13. However, the Gowalla locations tend to be more sparsely visited than the Twitter parcels, with 75% of the locations having 5 or fewer total visits spread across 4 or fewer accounts. This is an attractive attribute of the Gowalla data, as fewer unique accounts visiting the same location should increase the discriminative power of the score functions used in constructing the score-based likelihood ratio and coincidental match probability.

Type	Mean	Med.	75^{th}%ile	Max
Visits	5.3	2	5	3,345
Accounts	3.5	2	4	1,898

Table 4.8: Summary statistics for the distribution of number of visits (Type "Visits") and unique accounts with at least one visit (Type "Accounts") in each parcel computed from the full population data in Table 4.1. The minimum and 25th percentile are 1 for all cases.

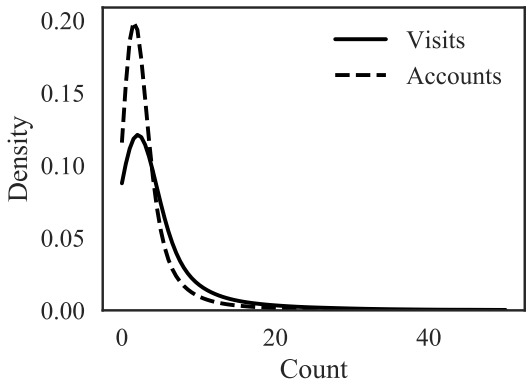


Figure 4.13: Density estimate of the number of locations versus the number of visits at the location (solid line), and the number of unique accounts in that location (dashed line). Note that both figures are right-truncated due to the extremely long tails.

To generate the spatial event data for our experiments we filtered the data based on sequential time periods of activity. Users with at least 1 visit per month in each of the last 2 months were considered.[9] For a given user, we define the sets of locations A and B to be all geolocated

[9]For the Twitter data, the first two months were considered. The Gowalla data, however, has very few events from March to November 2009, so we chose to focus on the last months of observed data, September and October 2010.

events in the October or September 2010, respectively. Table 4.9 contains summary statistics for the Gowalla data used in the analysis.

Accounts	Visits in A	Visits in B
1,214	16,682 (13.7)	26,448 (21.8)

Table 4.9: Number of observed accounts and visits for the Gowalla data used in the analysis. Average number per account denoted in parentheses.

4.8.2 Results

We compared the LR, SLR and CMP techniques in terms of their effectiveness in quantifying the strength of evidence for pairs of sets of locations A and B. For computational efficiency we included all 1,214 same-source pairs and a stratified random sample of different-source pairs. The sampling was stratified by the number of visits in each pattern, n_a and n_b, because the data is highly skewed towards a small number of visit events per individual and we wanted to assess performance of the methods under varying amounts of data. The strata correspond to all $3 \times 3 = 9$ combinations of less than 5 visits, between 5 and 14 visits, and 15 or more visits for n_a and n_b. 1,000 different-source pairs in each strata were randomly sampled, resulting in 9,000 total different-source pairs in each region.

Weight	TP@1	FP@1	AUC
0.80	0.301	**0.012**	0.717
$\alpha(n_a)$	**0.432**	0.037	**0.786**
$\alpha(n_a \vert \gamma, \rho, \phi)$	0.417	0.030	0.771

Table 4.10: Performance of a classifier based on LR for the Gowalla data.

Table 4.10 shows the classification performance of the likelihood ratio approach for each choice of mixing weight. The LR performed similarly on the Gowalla data as it did for the Twitter data, with the highest true positive rate and AUC belonging to the varying mixing weight approach and the lowest false positive rate for fixed α.

Δ	Weights	TP@1	FP@1	AUC
$\overline{D}_{min}$	Uniform	0.823	0.236	0.875
$\overline{D}_{min}$	Account	**0.827**	0.218	**0.884**
$\overline{D}_{min}$	Visit	0.821	0.232	0.875
EMD	Uniform	0.801	0.219	0.865
EMD	Account	0.797	**0.202**	0.873
EMD	Visit	0.790	0.226	0.858

Table 4.11: Performance of a classifier based on SLR_{Δ} for the Gowalla data.

Table 4.11 provides the true and false positive rates along with the AUC for the SLR using the various score functions considered. The mean nearest neighbor distance with account weighting scheme performed the best in terms of true positive rate and AUC, while the earth mover's distance with account weights had the lowest false positive rate.

Δ	Weights	TP@0.05	TP@0.01	AUC
$\overline{D}_{min}$	Uniform	0.572	0.273	0.877
$\overline{D}_{min}$	Account	**0.600**	**0.316**	**0.886**
$\overline{D}_{min}$	Visit	0.561	0.259	0.876
EMD	Uniform	0.490	0.301	0.868
EMD	Account	0.529	0.306	0.876
EMD	Visit	0.481	0.242	0.860

Table 4.12: Performance of a classifier based on CMP_{Δ} for the Gowalla data.

For the coincidental match probability, the false positive rate is fixed at the choice of threshold as discussed in Section 3.2.2. Thus, only the true positive rate and AUC are indicative of its performance as a classifier. Table 4.6 contains these values for thresholds of 0.05 and 0.01 (listed as TP@0.05 and TP@0.01, respectively). The CMP using the account weighted mean nearest neighbor distance had the highest true positive rate for both thresholds considered and the largest AUC.

Overall, the score-based approaches have better true positive rates and AUC than the LR approach regardless of score function or mixture weight used. This is in contrast to the Twitter data, where the LR had a better AUC than either the SLR or CMP. The worse performance of the LR in terms of AUC can be attributed to the sparsity of locations in

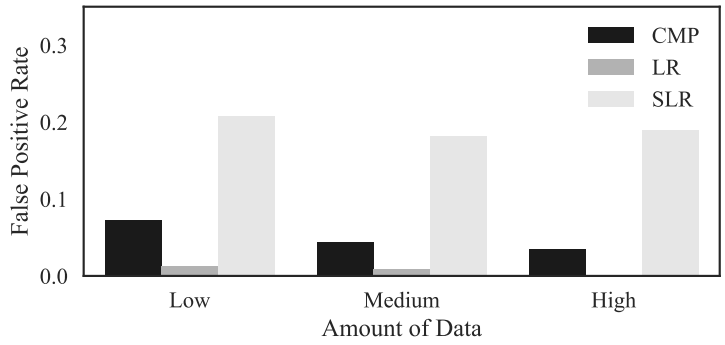
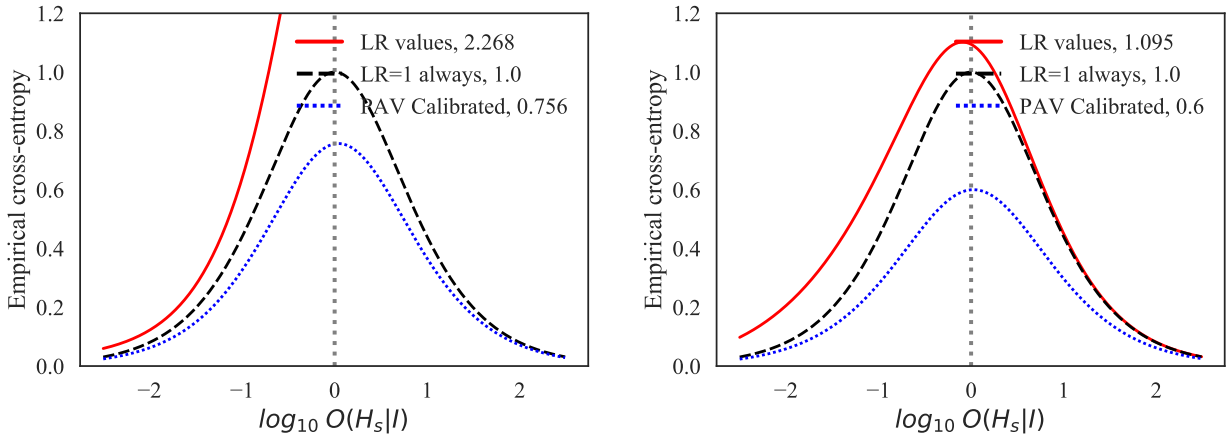


Figure 4.14: False positive rate of each method under different data regimes for the Gowalla data. Low corresponds to less than 5 in each of A and B, medium is between 5 and 14 events, and high is 15 or more events. Showing results for fixed α in the LR approach and the account weighted EMD for the SLR and CMP approaches. Thresholds are 1 for the LR and SLR and 0.05 for the CMP. Trends are similar for other choices of mixing weight and score function (see Figure B.2 in Appendix B).

the Gowalla data that makes the distinction of same-source pairs more difficult. Gowalla limits users to check-ins at public places only, which removes the possibility of events being generated at rare locations such as homes that are helpful in distinguishing unique behavior patterns across the population of users. This intuition is furthered by the fact that the false positive rate of the LR with a threshold of 1 is significantly lower than the FP of the SLR with the same threshold or CMP with any reasonably chosen threshold, which implies that the LR is accurately quantifying the strength of evidence for different-source pairs better than the score-based approaches. This phenomenon not only appears in the overall results, but also when considering performance of the techniques within the strata. Figure 4.14 depicts the FP rate of the approaches versus the amount of data in the sets A and B (corresponding to a selection of 3 of the 9 strata used in sampling). As the amount of data increases, the false positive rate decreases for the LR and CMP, while the SLR's FP rate remains fairly constant. The SLR has much higher FP rate than the LR and CMP across all data regimes.

In addition to classification performance, we can also consider the information-theoretic value of the methods as discussed in Section 2.6. Figure 4.15 contains ECE plots for the best performing LR and SLR method on the Gowalla data. For all ECE plots, see Appendix B.



(a) LR with $\alpha(n_a)$

(b) $SLR_{\overline{D}_{min}}$ with Account Weights

Figure 4.15: Empirical cross-entropy (ECE) plots for a selection of the LR and SLR methods applied on the Gowalla data. C_{llr} values are provided in the legend. (a) Likelihood ratio with the nonparametric mixing weight $\alpha(n_a)$; (b) Score-based likelihood ratio using the mean nearest neighbor distance and account weighting scheme.

The best performing method is the SLR using the mean nearest neighbor distance and account weights since it has a lower value of ECE (solid red curve in Figure 4.15b) for most regions of the prior log odds compared to other SLR approaches and LR approaches (e.g., solid red curve of Figure 4.15a). Unlike the Twitter data, the SLR also has better discriminating power than the LR, as shown by the PAV calibrated ECE (and therefore the C_{llr}) being lower for the SLR than that of the LR.

4.8.3 Discussion of Gowalla Results

The classification and information-theoretic performance of the evidence evaluation approaches show mixed results for the Gowalla data. While the score-based approaches have better true positive rates, AUC values, and lower calibrated ECE values (SLR only) than the LR, their false positive rates for natural threshold choices (1 for likelihood ratios and 0.05 or 0.01 for the CMP) are significantly higher. Overall, the account weighting strategy performed the best for the score-based approaches, which is not surprising given the increased

resolution of locations in the Gowalla data compared to the parcels in the Twitter data. Down-weighting very common locations (as determined by the number of accounts seen at those locations) is a natural choice for check-in data that is limited in terms of spatial resolution, i.e., events can only be generated at pre-defined locations. While the score-based approaches with account weights have many intuitive advantages and are better calibrated, the likelihood ratio approach is still preferred given the significantly lower false positive rate.

4.9 Discussion

Analysis of user-generated spatial event data is likely to become increasingly important in the forensic investigation of digital evidence. However, few methods have been developed to date that use statistical techniques for analysis of such data. In this chapter we have taken a step towards the development of such techniques, focusing on the problem of investigating whether two sets of user-generated geolocated events were generated by the same source or by different sources. Given a reference population, we proposed two approaches to quantify the strength of evidence in this setting. The first is a likelihood ratio approach based on modeling the location data directly. The second is to instead measure the similarity of the two sets of locations via a score function and then assess the strength of the score resulting in the score-based likelihood ratio or coincidental match probability. Experimental results, based on analysis of Twitter data in two spatial regions and Gowalla data, indicate the proposed methodology provides a useful starting point for forensic investigation of geolocated event data. However, many factors must be taken into account in regards to both the applicability of the methods and their performance.

It is worth noting that the manner in which we defined the sets A and B for the Twitter and Gowalla data (via time) is just one approach and the techniques we propose are not dependent on how the events in A and B are defined. For example, other ways of defining

the sets of locations could include events from two different devices (e.g., mobile phones) collected over the same time period where an investigator is interested if they are associated with the same individual. The data sets investigated were simply chosen for convenience.

For the data sets investigated, we found that the methods showed promise in terms of being able to separate same-source pairs of spatial patterns from different-source pairs. This observation leads us to believe these methods could be useful for discovery, e.g., as a method to rank the similarity of multiple different sets of locations from known sources to a single set of locations from an unknown source (similar to the motivating example in Section 4.1). Applicability of these methods for presentation in court should proceed with extreme caution as human event data is quite variable and, therefore, the methods can be ill-calibrated as shown in both case studies. Investigators must use their best judgement to determine whether or not the result should be trusted to prevent errors like those presented in Section 4.7.4.

There are two main areas that impact the behavior of the techniques: the characteristics of the spatial region under consideration and the amount of evidential data available.

Region Characteristics

The spatial regions considered have very different characteristics. Southern California (Gowalla data) and Orange County (Twitter data) are largely suburban, while New York is the most densely populated city in the United States. As a result, the characteristics of the locations and how they are used tend to be quite different in each of these regions. In Southern California, land parcels are typically single-use with one business or home at each location (this is especially true for the more granular Gowalla locations). However, in New York, the parcels are mostly high-rise buildings that contain many residences and businesses. We found that the different characteristics of the spatial regions manifest in different performance of the LR, SLR, and CMP. In general the classification problems for Southern

California (both the Gowalla data and Orange County Twitter data) are easier than for New York. The AUC illustrates this phenomenon, with each method having a larger AUC for Southern California and OC than NY. This suggests that an analyst may need to take into account his or her knowledge of the region under consideration when presenting error rates of the method.

Amount of Evidential Data

Varying the number of events in A and B can significantly impact the behavior of our approaches. The score-based methods tend to be sensitive to the amount of evidential data available because the variance of the underlying score functions is high when the number of events is low. The high variance in the score function would be expected under both the same- and different-source distributions, making them more similar and generally leading to smaller SLR values for same-source pairs and larger values for different-source pairs. There is no natural way to alter behavior of the score functions when the number of observations is low. The LR approach is less sensitive to the amount of data, which makes intuitive sense as the likelihoods in both the numerator and denominator have no explicit reliance on the number of observed events.

4.9.1 Contributions

The following contributions were made in this chapter:

- A novel technique to quantifying the strength of evidence for geolocated event data via the likelihood ratio using mixtures of kernel density estimators.
- We developed a variety of appropriate score functions that can distinguish between same- and different-source series of geolocated events.

- Extensive experimental comparison of LR, SLR and CMP methods on two large real-world data sets in two different regions within the US.

Chapter 5

Temporal Event Data

The most common feature of user-generated event data—regardless of how the data was collected or what device it was collected from—is the time at which events occurred. Thus, the most universally applicable investigatory tool for digital evidence should consider only the event times. As an example, consider the case where one event series A consists of a log of timestamped events (such as logins, file access events, browsing, messaging) generated on a device associated with a crime (e.g., on a mobile phone found at a crime scene). A second event series B consists of a log of similar events associated with a suspect (e.g., user-generated events recorded on a device owned by the suspect). The evidence consists of both event series A, B and the question of interest is to determine how likely it is that the two series were generated by the same individual.

The primary contribution of this chapter is the development and evaluation of quantitative techniques for the forensic analysis of temporal event data. We begin by proposing non-parametric score functions for quantifying the association between pairs of potentially related discrete event time series, focusing on a particular (and common) situation where event dependence arises because one type of event tends to occur within bursts of the other type.

We then assess the strength of evidence (also referred to here as the degree of the association) in two scenarios: (i) using score-based likelihood ratios when multiple pairs of discrete event series are available to serve as reference data and (ii) using a resampling approach to compute the probability of a coincidental match when only a single pair of event series is available.

All techniques described below are implemented in the open source R package `assocr` available at `https://github.com/UCIDataLab/assocr`. The case study data sets are also contained in this repository. This chapter expands upon previously published work in Galbraith et al. [2020a], which contains additional resources including the scripts run to perform the experiments and documentation describing their usage.

5.1 Motivating Example

Suppose that a forensic investigator is given two authentication (or login) sequences for two different compute resources in a network. The sequences contain only the timestamps at which the logins occurred and the resource that was accessed (i.e., the computer that was logged into). The first series consists of logins to a shared resource, i.e., a computer that many users in the system have access to. The source of this first series of authentications to the shared resource is under question, so it is referred to as the *unknown source data A*. The second series consists of logins to a personal machine, i.e., a computer that is only accessed by one user in the system. The source of this second series of authentications to the unique computer is known, so it is referred to as the *known source data B*. See Figure 5.1 for an example of this scenario.

The investigator has reason to believe the user who generated the known source data is also responsible for the logins to the shared resource, i.e., they are also responsible for generating the unknown source data. From a visual inspection of the two series, it is apparent that the

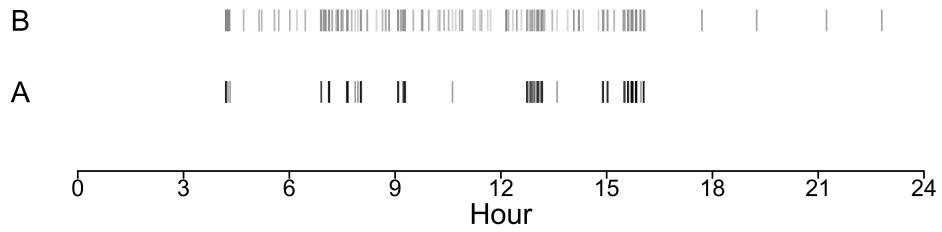


Figure 5.1: Series of authentication events for logins to a unique computer (known source series B) and a shared compute resource (unknown source series A) taken from Section 5.8. Both series were generated by the same user.

logins to the two different resources tend to occur near each other in time. The investigator needs to quantify this relationship (i.e., the association between the series), however, in an objective manner. To address this problem, we propose techniques in this chapter that, given two temporal event series, produces a measure of their probative evidential value without needing a sample from a relevant reference population. Using our method on the data in Figure 5.1, the investigator would be able to conclude that it is highly unlikely that the two series A and B were generated by different sources. In the remainder of this chapter, we will show how to produce such conclusions given this type of temporal event data.

5.2 Forensic Question of Interest

Consider a pair of user-generated *event series*, where each event series is defined by a set of times when events of the appropriate type occurred, e.g., series A and series B with n_a and n_b events of type A and type B, respectively. Equivalently, the pair of event series (A, B) can be thought of as a *temporal marked point process* E [e.g., Daley and Vere-Jones, 2007], with

$$E = (A, B) = \{(t_j, m(t_j)) : j = 1, \ldots, n_a + n_b\} \tag{5.1}$$

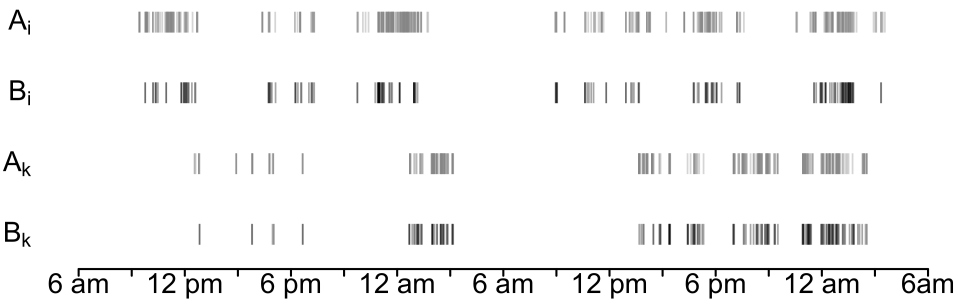


Figure 5.2: Example of temporal marked point processes from two different individuals (i and k) taken from the case study of Section 5.7. Note that A and B events generated by the same individual tend to cluster temporally, with less clustering in time for A and B events from different users.

where $t_j \in \mathbb{R}^+$ and $m(t_j) \in \{\mathrm{A}, \mathrm{B}\}$ are the time and type (or mark) of the j^{th} event, respectively. For example, series A and B could consist of timestamped events corresponding to activity from two user accounts (e.g., accounts on a social media platform such as Facebook) where we may be interested in determining if the two accounts belong to the same individual.

We focus on the case where the two event series exhibit temporal clustering such that the occurrence of one type of event at a particular time increases the likelihood that an event of the other type will also occur nearby in time. In contrast, one can also have "negative association," where one type of event tends to repel the other type (e.g., when one individual uses two devices or accounts at distinct and clearly separated times). This alternative is not pursued in this work, although it may be possible to adapt the present framework to such situations.

Figure 5.2 provides an example of the types of pairs of temporally clustered event series that we focus on. The data consists of two pairs of event series of browser actions, (A_i, B_i) and (A_k, B_k) generated by users i and k, respectively, taken from the case study discussed later in Section 5.7. From the figure it is visually apparent that A_i and B_i are associated with each other, as are A_k and B_k.

In this general context we address the problem of developing methods to quantify the likelihood of observing the pair of event series (A, B) under different hypotheses regarding their source. While it is possible to model the event series directly via Hawkes processes (for example) specifying the relationship between two event series under the same- and different-source hypotheses is difficult. For this reason we only consider score-based approaches for temporal event data in this work. In particular, we focus on two specific aspects of this problem:

1. Investigating suitable measures $\Delta(A, B)$ to quantify the association between two event series A and B, and

2. Quantifying the likelihoods of observing the pair (A, B)—or more precisely the likelihood of observing the relevant score $\Delta(A, B)$—under the hypotheses that the series were generated by the same source or by different sources. We will refer to this (second) aspect of the problem as *assessing the degree of association* (or strength of evidence) between the two event series.

We address the first question by leveraging ideas from the marked point process literature where a variety of techniques have been developed for measuring association between marks, particularly for spatial point processes [Illian et al., 2008; Baddeley et al., 2015]. Real-world pairs of event series $E = (A, B)$ of user-generated event data can exhibit significant burstiness and inhomogeneity over time (e.g., as in Figure 5.2), making it challenging to develop robust parametric models of association between A and B. For this reason we pursue non-parametric measures of association between temporal processes, particularly measures based on near-neighbor and inter-event time characteristics.

To address the second question, the quantification of the likelihood of observing A and B (or more precisely $\Delta(A, B)$) under competing hypotheses about the source(s) of the series, we investigate two different methods. The first is a population-based approach where we have

many realizations of same- and different-source pairs of processes $\mathcal{D}_s$ and $\mathcal{D}_d$, respectively, that are used to compute the score-based likelihood ratio. The second is a randomization technique when only a single pair E is available (i.e., we don't have access to a sample from a relevant population of realizations) and the coincidental match probability is computed. The resampling approach in particular opens up the proposed methodology to a much broader range of applications in practice, given that it relaxes the need for data from a reference population.

5.3 Related Work

Measuring the association between event series is an issue that arises in various applications. For example, a common problem in spatial statistics is determining the relationship between point patterns (i.e., marked point processes) by performing inference (either analytical or numerical) under a null model that typically assumes some form of independence between the patterns. One of the most well-known techniques is Ripley's cross-K function [Dixon, 2014], which measures the number of occurrences of one type of point within a given radius r of the other type of point as a function of r. Under certain assumptions on the processes themselves (e.g., stationarity) and the relationship between the processes (e.g., complete spatial randomness), significance tests can be performed to determine if the observed function is consistent with the assumed relationship [e.g., Diggle and Chetwynd, 1991; Gaines et al., 2000].

A popular alternative to the analytical significance test uses *simulation envelope* techniques which compute a summary function of the observed point patterns (such as Ripley's cross-K function) and compare the observed function with the envelope of a set of functions obtained from simulations of the null model [Baddeley et al., 2014]. Numerous methods exist for computing simulation envelopes, but the most relevant to the present work use

some form of a bootstrap to perform the simulation. Loh [2008] fixed the spatial locations and resampled the marks of the points to obtain confidence envelopes of spatial correlation functions. Niehof and Morley [2012] kept the marks fixed, but resampled the (temporal) locations of the points by incorporating a moving block bootstrap [Synowiecki, 2007]. Our method is conceptually related to simulation envelopes obtained by bootstrapping. However, we focus on a single point (e.g., a single score) rather than a function and the focus of our analysis is on data that is both inhomogeneous and bursty.

Another related line of work has been developed in the neuroscience literature. One approach in that context is the spike train reliability statistic R introduced by Hunter and Milton [2003] to measure the association of neural spike trains. R is taken to be the mean normalized exponentiated time between events in one series and the corresponding nearest neighbors in another series. This is related to the inter-event time score functions that we introduce in Section 5.4.2. However, no method to assess the statistical significance of R has been developed in prior work. Another method in this context is event synchronization (ES), as proposed by Quiroga et al. [2002] for measuring correlation in left and right EEG channels. More recently, the technique has also been used in climatological applications [Boers et al., 2016; Malik et al., 2010]. Event synchronization is similar to the reliability statistic in that it is a function of the time between events in one series to nearest neighbors in the other, but it also takes into account the marginal inter-event times (i.e., the time between events for the subprocess restricted to events of a single type). In this manner, ES is similar to the signal-to-noise ratio that we utilize to determine the detectability of association (see Section 5.6).

Building on work from both of the aforementioned domains, Donges et al. [2016] proposed a framework called event coincidence analysis (ECA) with an open source software package for quantifying the strength, directionality, and time lag of relationships between event series [Siegmund et al., 2017]. ECA focuses on coincidences, which are defined by the Donges et al.

[2016] as the occurrence of at least one event in each series in some (τ lagged) time window ΔT. Under the assumptions that both series are independent Poisson processes and that events are rare (i.e., the number of events multiplied by ΔT is sufficiently smaller than the period spanned by the series), analytical significance tests for the number of observed coincidences have been derived. Donges et al. [2016] relaxed these assumptions by proposing two surrogate- data-based tests that rely on simulating realizations of both processes via either random event times or from some prescribed inter-event time distribution. Conceptually this method is the most similar to our proposed technique but requires the specification of both a time lag τ and coincidence window ΔT before performing any analysis. One could perform the significance tests for multiple values of τ and ΔT, but then the analysis would suffer from multiple-testing issues.

5.4 Score Functions for Temporal Event Data

We investigate a number of different score functions $\Delta(A, B)$ that characterize the association between a pair of event time series (focusing solely on the times at which events occur, so the series are 1-dimensional). Two different types of score functions are considered, based on either nearest-neighbor characteristics or summary statistics of inter-event times. Note the score functions rely on the notion of a *reference point*, which is a term related to the Palm distribution [Hanisch, 1984]. For practical purposes, we define the reference point as an arbitrarily selected event in the pair of event series $E = (A, B)$. The reference point may be of either type.

5.4.1 Marked Point Process Indices

Coefficient of Segregation

The *coefficient of segregation* [Pielou, 1977] is a function of (i) the ratio of the probability that a (randomly chosen) reference point and its nearest neighbor have different marks to (ii) the same probability for independent marks, defined as

$$S(A,B) = 1 - \frac{p_{ab} + p_{ba}}{p_a p_{\cdot b} + p_a p_{\cdot b}}. \tag{5.2}$$

Here p_{ab} (or p_{ba}) is the joint probability that the reference point is type A and its nearest neighbor in time is type B (or vice-versa), p_a and p_b are the relative frequencies of the two types of points, and $p_{\cdot a}$ (or $p_{\cdot b}$) is the probability that the nearest neighbor is type A (or B) irrespective of the type of the reference point. These probabilities are naturally estimated by the empirical relative frequencies of the appropriate events as observed in the data

$$\begin{aligned}
\hat{p}_a &= \frac{n_a}{n_a + n_b} = \frac{n_a}{n} \\
\hat{p}_{ab} &= \frac{1}{n_a} \sum_{j=1}^{n} \mathbb{I}\left[m(t_j) = \mathrm{A}\right] \mathbb{I}\left[m(z_1(t_j)) = \mathrm{B}\right] \\
\hat{p}_{\cdot b} &= \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}\left[m(z_1(t_j)) = \mathrm{B}\right]
\end{aligned} \tag{5.3}$$

where $z_1(t_j)$ denotes the nearest neighbor of the point t_j, $m(\cdot)$ the mark of the given point, and $\mathbb{I}(\cdot)$ the indicator function. Similar definitions hold for $\hat{p}_b$, $\hat{p}_{ba}$, and $\hat{p}_{\cdot a}$.

Note that $S(A,B) \in [-1,1]$. If the reference point and its nearest neighbor always are the same type, then $p_{ab} = p_{ba} = 0$ and $S(A,B) = 1$. This corresponds to repulsion or segregation of points by their mark (i.e., points of type A always occur near each other and never near points of type B and vice-versa). If the reference point and its nearest neighbor always have different marks, then $p_{aa} = p_{bb} = 0$ which implies that $p_{\cdot a} = p_{ba}$ and $p_{\cdot b} = p_{ab}$ so $S(A,B) < 0$

with a minimum of $S(A,B) = -1$ if $p_a = p_b = 1/2$. This is the opposite of segregation, indicating that points of different marks are attracted to one another. If the marks are independent then $S(A,B)$ will tend to 0 as the size of the observed data set grows since $p_{ab} \approx p_a p_{\cdot b}$ and $p_{ba} \approx p_b p_{\cdot a}$.

Mingling Index

The *mingling index* is also based on local neighborhoods of a reference point. It compares the type of the reference point to those of its k nearest neighbors, and is calculated by an average of terms in which each point is treated as the reference point:

$$\begin{aligned} \overline{M}_k(A,B) &= \frac{1}{n_a+n_b} \sum_{j=1}^{n_a+n_b} M_{k,j}(A,B) \\ &= \frac{1}{k(n_a+n_b)} \sum_{j=1}^{n_a+n_b} \sum_{\ell=1}^{k} \mathbb{I}\left[m(t_j) \neq m(z_\ell(t_j))\right] \end{aligned} \tag{5.4}$$

where $M_{k,j}$ is the mingling index when the j^{th} event, t_j, is treated as the reference point and $z_\ell(t_j)$ denotes its ℓ^{th} nearest neighbor. Thus, $\overline{M}_k(A,B) \in [0,1]$ describes the mean fraction of points among the k nearest neighbors whose type is different than that of the reference point. In the results in this chapter, we use the single nearest neighbor with $k = 1$ and, therefore, the mingling index simplifies to $\overline{M}_1(A,B) = p_{ab} + p_{ba}$.

The mingling index can be thought of as a characterization of the mixture of points of different type. If the reference point and its k nearest neighbors tend to be of the same type, then $\overline{M}_k(A,B)$ has a small value and the process can be viewed as segregated (repulsion between points of different types). In the opposite case, the mingling index has a large value and the process can be viewed as mixed (attraction).

5.4.2 Inter-Event Times

We also investigate score functions based inter-event times for a pair of event time series. In principle, we expect the inter-event times to carry more information than the nearest neighbor characteristics used in the coefficient of segregation. We construct distributions of inter-event times by measuring the time from each event in B to the closest event in A. Here, we define "closest in time" to mean the closest event either forward or backward in time, but a directional definition (e.g., forward or backward only) could also be used.

Let $\mathcal{T}_{BA}$ represent the set of n_b inter-event times from B to A, and define it as follows

$$\begin{aligned} \mathcal{T}_{BA} &\equiv \left\{ \tau_{BA,j} : j = 1, \ldots, n_b \right\} \\ \text{where } \tau_{BA,j} &= \min_{k \in \{1,\ldots,n_a\}} |t_{b,j} - t_{a,k}| \end{aligned} \tag{5.5}$$

and $t_{b,j}$ denotes the j^{th} event time of series B, and similarly for series A. See Figure 5.3 for an illustration. If events of type B are clustered in time with events of type A, then the inter-event times $\mathcal{T}_{BA}$ tend to be smaller than if A and B events are generated independently. A variety of characteristics of the distribution of inter-event times could be used as score functions. We consider the mean inter-event time from B to A

$$\overline{\mathcal{T}}_{BA} = \frac{1}{n_b} \sum_{j=1}^{n_b} \tau_{BA,j} \tag{5.6}$$

and the median inter-event time from B to A, $med(\mathcal{T}_{BA})$.

5.5 Quantifying Strength of Evidence

Suppose we are given a pair of event time series $E^* = (A^*, B^*)$ and a particular score function Δ, such as one of those defined in the previous section. We wish to assess the degree of

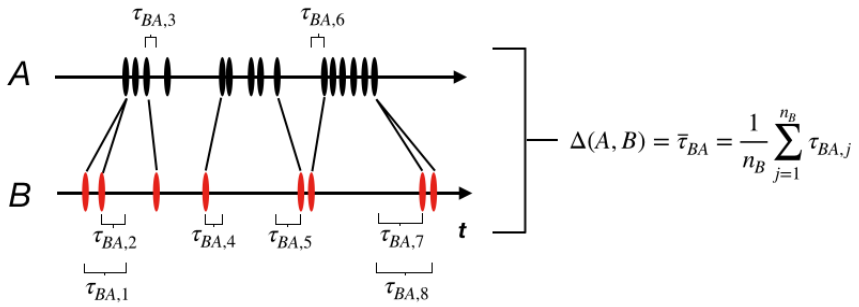


Figure 5.3: Example mean inter-event time calculation.

association between A^* and B^*. To do so, we must consider the likelihood of observing $\Delta(A^*, B^*)$ under two competing hypotheses, namely, that A^* and B^* were generated by the same source, or that they were generated by two different sources. We investigate two different score-based methods in this context: a population-based approach in which we have a sample from a relevant population, and a resampling approach when only a single pair E^* is available.

5.5.1 Population-based Approach

We begin by considering a situation in which we have a sample from a relevant reference population as discussed in Chapter 3. The reference data sets $\mathcal{D}_s$ and $\mathcal{D}_d$ used below are constructed via leave-pairs-out cross-validation as described in Section 2.4.3.

To illustrate this approach, consider the score function for the mean inter-event time proposed in Equation 5.6 and let $\Delta(A, B) = \overline{\mathcal{T}}_{BA}$ for any given pair of event time series (A, B). Thus, $\delta^* = \overline{\mathcal{T}}_{B^*A^*}$ is the observed value that is plugged into the estimator of the score-based likelihood ratio in Equation 3.3:

$$\widehat{SLR}_{\overline{\mathcal{T}}_{BA}} = \frac{\widehat{g}(\overline{\mathcal{T}}_{BA} = \overline{\mathcal{T}}_{B^*A^*} | \{\overline{\mathcal{T}}_{BA} : (A, B) \in \mathcal{D}_s\})}{\widehat{g}(\overline{\mathcal{T}}_{BA} = \overline{\mathcal{T}}_{B^*A^*} | \{\overline{\mathcal{T}}_{BA} : (A, B) \in \mathcal{D}_d\})}. \tag{5.7}$$

where $\widehat{g}$ is a kernel density estimator with a Gaussian kernel and rule-of-thumb bandwidth

[Scott, 1992].[1] We explicitly condition on the reference sets $\mathcal{D}_s$ and $\mathcal{D}_d$ because the score values for the point patterns in these sets along with the kernel density parameters fully specify the estimated density. Further, the reference data sets fully encompass the relevant source hypothesis and background information.

In addition to the score-based likelihood ratio, the coincidental match probability can also be estimated using the reference data. Given the observed score $\delta^* = \overline{\mathcal{T}}_{B^*A^*}$, the estimator of the coincidental match probability in Equation 3.8 becomes

$$\widehat{CMP}_{\overline{\mathcal{T}}_{BA}} = \frac{1}{N_d} \sum_{(A,B)\in\mathcal{D}_d} \mathbb{I}\left(\overline{\mathcal{T}}_{BA} < \overline{\mathcal{T}}_{B^*A^*}\right) \tag{5.8}$$

where N_d is the number of evidential samples in the different-source reference data set $\mathcal{D}_d$.

5.5.2 Resampling Approach

The population-based approach above is useful when a reference population of pairs of event series is available, e.g., user-generated data from a relevant population of users. However, there are many situations in practice where data from a population of users is not readily available. Furthermore, even when a population is available, it is often quite difficult to define the *relevant* reference population of interest in a forensic setting. Should the relevant population be a sample from all individuals in general, or from everyone who matches the description of a suspect in a given region, or from some other group? (See Section 2.4 for additional discussion of this issue.) To address these potential problems we propose below a resampling approach that computes coincidental match probabilities using only a single pair of event series.

[1] All of the scores we work with in this chapter are bounded (either above, below, or both), and an unconstrained KDE will push probability mass outside these bounds (e.g., below 0 for inter-event time score functions). More sophisticated methods could be used to estimate these densities, but for simplicity and computational efficiency we used a generic KDE method.

Coincidental Match Probability

In order to estimate the coincidental match probability we use resampling in time to simulate new realizations of the event series B^* under a null model for different source data and thus induce a distribution of scores under this model (see Section 3.2 for more details). Specifically, assuming that $\Delta(A, B) = \overline{\mathcal{T}}_{BA}$, we propose the following natural estimator for the coincidental match probability

$$\widehat{CMP}_{\overline{\mathcal{T}}_{BA}} = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \mathbb{I}[\overline{\mathcal{T}}_{B^{(i)}A^*} < \overline{\mathcal{T}}_{B^*A^*}] \tag{5.9}$$

where $B^{(i)}$ for $i = 1, \ldots, n_{sim}$ are randomly sampled under the null model for different source data. The smaller this empirical probability, the less likely it is that the pair (A^*, B^*) was generated by different sources. In the following section a method for resampling an event series is presented.

Sessionized Resampling

For the different source hypothesis we use a null model that assumes the known-source event series B^* is conditionally independent of the unknown source series A^* given an inhomogeneous background intensity process (e.g., that varies with time-of-day for user activity). In particular, we generate simulated series B' that depend on the background intensity and that have similar marginal characteristics to the observed series B^*. The particular details of how the simulation is carried out can be domain-specific. Since the user-generated event data of interest to this chapter are typically inhomogeneous and bursty, we pursue an approach that we call *sessionized resampling*.

Specifically, we keep the event times in A^* fixed and generate multiple random realizations B' of B^* by randomly perturbing the event times in B^*. This process can be thought of as

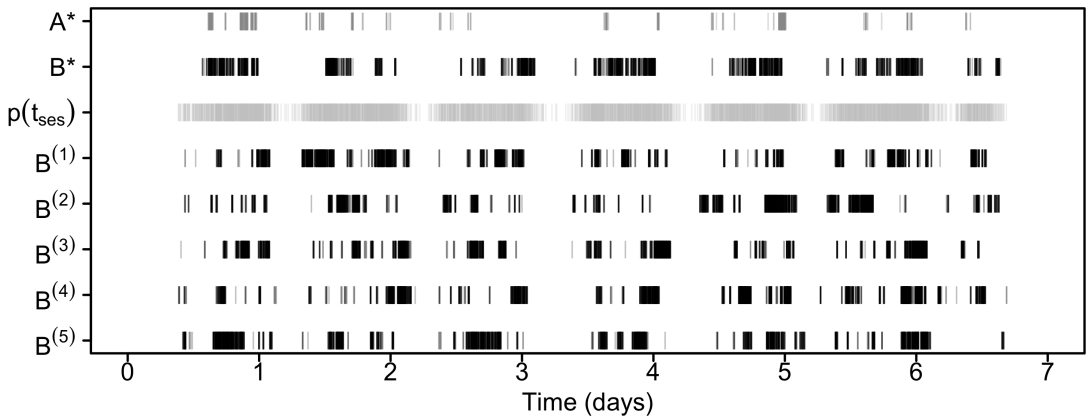


Figure 5.4: Example of sessionized resampling for a pair of event series (A^*, B^*) taken from the student web browsing data. Here we use $T = 10$ minutes, and the distribution of session start times $p(t_{ses})$ is the empirical distribution of session start times across all series B available in the data set. $B^{(\ell)}$ for $\ell = 1, \ldots, 5$ represent five event series simulated via Algorithm 2.

generating many event series B' from other suspects. In particular, in order to preserve the bursty and inhomogeneous nature of the data, we work with sessions (collections of event times) rather than individual event times. Sessions are defined formally below. We sample new times for the starts of sessions (rather than new times for individual events). Each session is then shifted to the corresponding sampled (perturbed) start time. The new session start times are sampled from an inhomogeneous background distribution over time. We next describe the steps in this approach in more detail.

To sessionize the data we proceed by defining the first event in a session to be any event that occurs after a period of T or more time units of inactivity [e.g., Spiliopoulou et al., 2003]. We define the set of session start times for series B^* as

$$\begin{aligned} B^*_{ses} &= \{t_j : j = 1 \text{ or } t_j - t_{j-1} \geq T \text{ for } j = 2, \ldots, n_{b^*}\} \\ &\equiv \{t_{ses,k} : k = 1, \ldots, r_{b^*}\}. \end{aligned} \tag{5.10}$$

Thus B^* has $r_{b^*} \leq n_{b^*}$ sessions, where a session is defined as all of the events after one

session start and before the next. We then define the sessionized series B^* to be composed of the process of session start times B^*_{ses} and the event times in each corresponding session. Namely, $B^* = \bigcup_{k=1}^{r_{b^*}} S_k$ where

$$S_k = \begin{cases} \{t_j : t_{ses,k} \leq t_j < t_{ses,k+1}\} & \text{if } k = 1, \ldots, r_{b^*} - 1 \\ \{t_j : t_{ses,k} \leq t_j \leq t_{n_{b^*}}\} & \text{if } k = r_{b^*}. \end{cases} \tag{5.11}$$

This definition leaves the event series unchanged but groups activity according to bursts of activity.

A replicate of B^* can be generated as follows. Sample r_{b^*} new session start times from a distribution of session start times $p(t_{ses})$. (In general any distributional form for $p(t_{ses})$ that reflects the inhomogeneous nature of event series could be used.) Then shift all events in each of the sessions S_k for $k = 1, \ldots, r_{b^*}$ so that the first event $t_{ses,k}$ occurs at the k^{th} newly sampled session start time. This process preserves the total number of events in B^* as well as the number of and spacing of events in each session. See Algorithm 2 for the pseudo-code to generate resampled series and Figure 5.4 for an illustration of how the approach works.

Algorithm 2 Sessionized Resampling

Input: Pair of event series (A, B); resampling time distribution $p(t_{ses})$
Output: Set of n_{sim} resampled pairs $\mathcal{D}_{sim}$

Derive the r_b sessions $S_k, k = 1, \ldots, r_b$ of B as defined in the text
for $\ell = 1$ to n_{sim} **do**
  for $k = 1$ to r_b **do**
    Draw $t_{new} \sim p(t_{ses})$
    Elementwise, set $S_k^{(\ell)} = S_k - t_{ses,k} + t_{new}$
  end for
  Set $B^{(\ell)} = \bigcup_{k=1}^{r_b} S_k^{(\ell)}$
end for
return $\mathcal{D}_{sim} = \{(A, B^{(\ell)}) : \ell = 1, \ldots, n_{sim}\}$

5.6 Case Study—Simulated Data

We simulated event series to form pairs of temporal processes, both independent pairs and pairs with varying degrees of association, in order to assess the behavior of our proposed methods for computing score-based likelihood ratios and coincidental match probabilities. The simulated series have similar marginal characteristics (in terms of overall rates of event generation) to the data from one of our case studies (the student web browsing data described in Section 5.7) where individuals generate two types of events, at different rates per individual, over a period of one week.

5.6.1 Simulating Temporal Marked Point Processes

The simulation process was as follows. We generate A events over a window of one week from a Poisson process with rate λ_A, where λ_A is sampled from a kernel density estimate of observed event rates across different users from the case study with web browsing data (Section 5.7). The rate of events for series B is proportional to λ_A, with $\lambda_B = p\lambda_A$ where $p \in [0, 1]$ is the relative frequency of type B events to type A events.[2]

For independent series we simulate events independently from two Poisson processes with rates λ_A and λ_B. For dependent processes we again simulate A events from a Poisson process with rate λ_A, but now generate B events according to Algorithm 3. For every simulated A event, a B event is generated with probability p, and (if generated) the time of the B event is distributed via a Gaussian density with standard deviation σ centered at the time of the A event[3]. The degree of association between two simulated processes is controlled by (i) the relative frequency p of events of type B to events of type A, (ii) the standard deviation σ,

[2]Here, we assume that the expected number of events in B is less than the expected number of events in A, i.e., $\mathbb{E}(n_b) = wp\lambda_A < w\lambda_A = \mathbb{E}(n_a)$ where w is the length of the observation window, to simplify the discussion of the results.

[3]We also experimented with sampling from non-Gaussian distributions such as the exponential and obtained similar results to those described here.

Algorithm 3 Simulation of associated marked point processes

Input: Intensity λ_A, relative frequency of B events to A events p, standard deviation σ
Output: Simulated pair of processes (A, B)

Simulate $A = \{t_j : j = 1, \ldots, n_a\}$ from a Poisson point process with rate λ_A
Set $k = 0$
for $j = 1$ to n_a **do**
    Draw $d_j \sim Bernoulli(p)$
    if $d_j = 1$ **then**
        Increment $k = k + 1$
        Draw $t_k \sim Normal(\mu = t_j, \sigma^2)$
    end if
end for
return $A = \{t_j : j = 1, \ldots, n_a\}, B = \{t_k : k = 1, \ldots, n_b = \sum_{j=1}^{n_a} d_j\}$

and (iii) the intensity λ_A of process A which also controls the number of events in both A and B. Our ability to detect an association is expected to decrease as (i) p decreases, and (ii) σ increases. The relationship between detectability and the number of events in each process is more complex, as we discuss later.

Simulations were performed with different combinations of parameter settings to investigate the sensitivity of our detection methods across a variety of scenarios. To ensure sufficient variation in the event counts we sampled the rates for process A (λ_A) in Algorithm 3 from a kernel density estimate as described earlier and multiplied the sampled intensity by a rate multiplier $r \in \{1, 10\}$ in order to asses the impact of dense event series on the SLR and CMP. For a given setting of parameter values, (r, p, σ), we simulated both independent and dependent series. The relative frequency of type B events to type A events was one of $p \in \{0.01, 0.10, 0.20, 0.50, 0.75, 0.95\}$. Finally, the standard deviation of the Gaussian distribution used to generate timestamps for events of type B was $\sigma \in \{0.5, 1, 2, 5, 10\}$ minutes.

For each combination of parameters (r, p, σ), we generated 10,000 independent event series pairs and 10,000 event series pairs with association, and computed SLR and CMP values for each pair. For both the SLR and CMP, we utilized the leave-pairs-out cross-validation

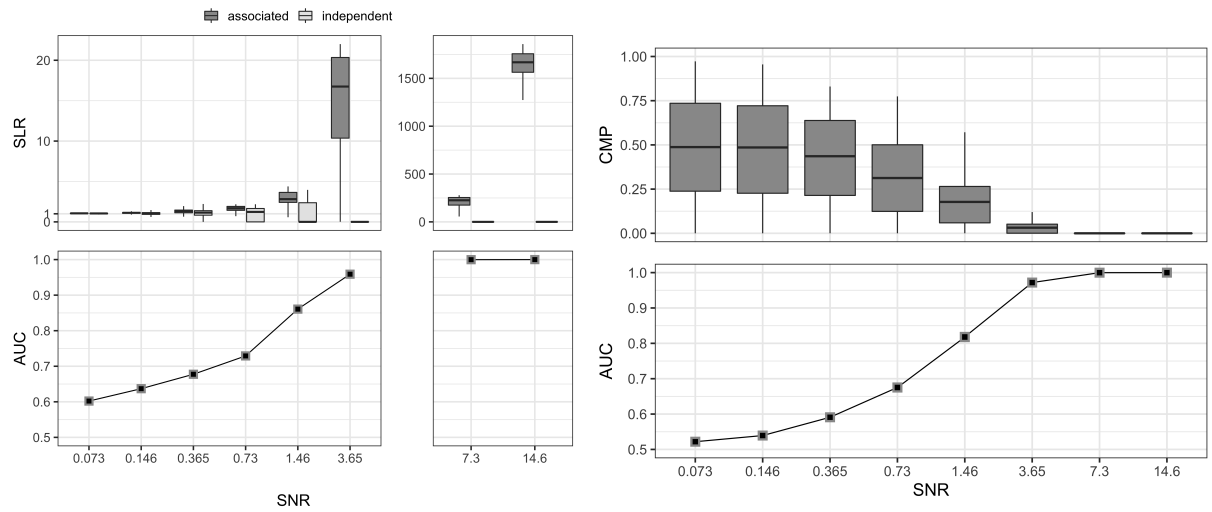
methodology. Therefore, the CMP was computed using the population data (and not via resampling) in the simulation results discussed below.

5.6.2 Results

In general we found that we could detect associated event series pairs over a wide variety of parameter settings, for both the marked point process indices and inter-event time statistics and for both SLR and CMP methods. Here we focus on results for the mean inter-event time score function (the other scores were not as accurate in detection). We found that the two most important factors in assessing the performance of our methods are the number of events in process B and the signal-to-noise ratio (SNR), defined as

$$SNR \equiv \frac{(\overline{\lambda}_A)^{-1}}{\sigma}. \tag{5.12}$$

Here, the "signal" is inversely proportional to the standard deviation σ of the Gaussian distribution used to generate event times in B. Smaller values of σ correspond to smaller inter-event times from events in B to events in A and, therefore, higher signal. The "noise" is the reciprocal of the mean intensity across the simulated realizations of process A, denoted $(\overline{\lambda}_A)^{-1}$. As this value decreases, the noise (or the density of events in realizations of A) increases. As an extreme case, consider a single process A' with $\lambda_{A'}^{-1} \to \infty$, which implies that $\mathbb{E}(\overline{\mathcal{T}}_{A'A'}) \to 0$. Regardless of the strength of the signal, events in B will occur close in time to events in A', and therefore any other series will appear to be associated with A'. The SNR controls for this phenomenon. As the SNR increases we expect the association of two event series to be more easily detected via methods such as the SLR or CMP. For the simulation study, the signal to noise ratio is known. In practice, a natural estimator of the



(a) (b)

Figure 5.5: (Top) Boxplot of measures of association for $\Delta(A, B) = \overline{\mathcal{T}}_{BA}$ for simulated data with $p = 0.20$ and (Bottom) corresponding AUC values as a function of the signal-to-noise ratio. (a) Score-based likelihood ratio. Note the different scales of the SLR for SNR $\in \{7.3, 14.6\}$. (b) Coincidental match probability. Note the CMPs for independent pairs are uniformly distributed by definition and thus omitted.

SNR for a single pair of event series (A, B) is given by

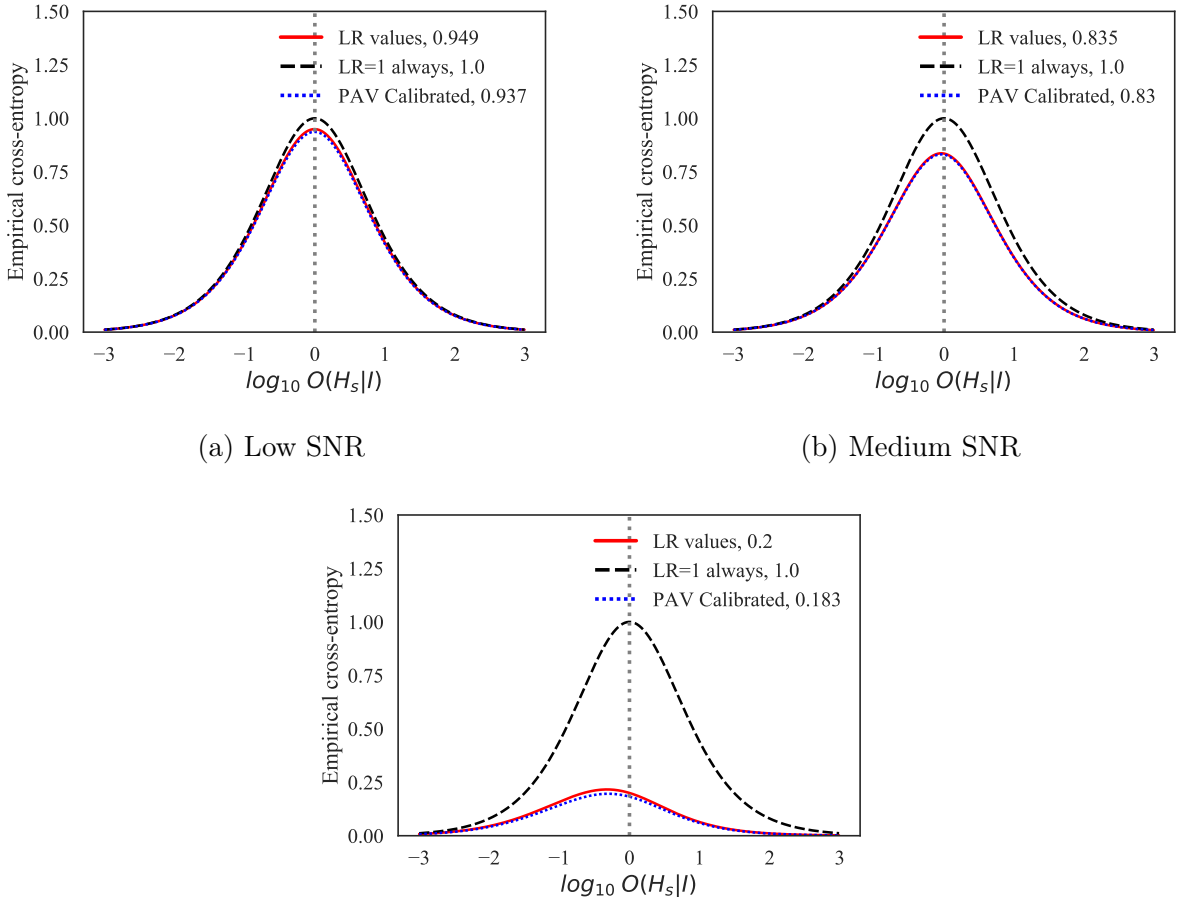
$$\widehat{SNR} = \frac{\overline{\mathcal{T}}_{AA}}{\overline{\mathcal{T}}_{BA}}. \tag{5.13}$$

See Appendix C for the reasoning behind this choice of estimator.

Figure 5.5 shows the detectability of association of simulated event series via the SLR (Figure 5.5a) and CMP (Figure 5.5b) as a function of the SNR of the simulated series.[4] The upper plots show box plots of the values of SLR and CMP, and the points below show the corresponding AUC values, each as a function of SNR. Here we present results for simulations with a relative frequency of $p = 0.2$. Results for other values of p are qualitatively similar (the magnitude of the SLR increases slightly for small SNR as p increases, but the CMPs are indistinguishable for varying p). As the SNR increases, the SLR increases, the

[4]Note that we considered 5 values of σ and 2 values of r, but when computing the SNR there are only 8 unique values due to the overlap of $\sigma \in \{5, 10\}$ with $r = 1$ and $\sigma \in \{0.5, 1\}$ with $r = 10$, e.g., $(1\overline{\lambda}_A 5)^{-1} = (10\overline{\lambda}_A 0.5)^{-1}$. Furthermore, $(\overline{\lambda}_A)^{-1} = 7.3$ minutes, which is the corresponding mean inter-event time in the first case study.

CMP decreases, and the AUC increases, all of which are indicative of being better able to discriminate associated and independent pairs of event series. Note that as the SNR gets large (e.g., SNR > 7.3) the AUC values of both classifiers have an asymptote near 1 and CMP values near 0 for associated pairs, but the SLR continues to increase. This implies that both methods perform similarly for classification, but that the SLR is better calibrated with values increasing indefinitely as the signal-to-noise ratio increases.



(a) Low SNR

(b) Medium SNR

(c) High SNR

Figure 5.6: ECE plots for simulated data with varying signal-to-noise ratios. In all cases, $p = 0.20$ and the SLR using the mean inter-event time score function. (a) SNR $= 0.073$; (b) SNR $= 0.73$; (c) SNR $= 3.65$.

To further illustrate the calibration and discrimination performance of the SLR, ECE plots were created for the various signal-to-noise ratio values considered. Figure 5.6 contains such plots for simulations with a relative frequency of $p = 0.2$ with low, medium and high SNR

values. Under all SNR regimes, the SLR is well-calibrated as shown by the ECE (solid red curves) being almost identical to the PAV calibrated ECE (blue dotted curves). When the signal-to-noise ratio is low (e.g., Figure 5.6a), the SLR provides little in terms of reduction of uncertainty about the source hypotheses since the ECE curve is similar to that of the neutral method. As the SNR increases the discrimination performance of the SLR increases, as shown by the ECE curves shifting down in Figures 5.6b and 5.6c. For simulations with high SNR values (e.g., Figure 5.6c), the SLR greatly reduced the uncertainty regarding the source hypotheses and, therefore, its discrimination power is high.

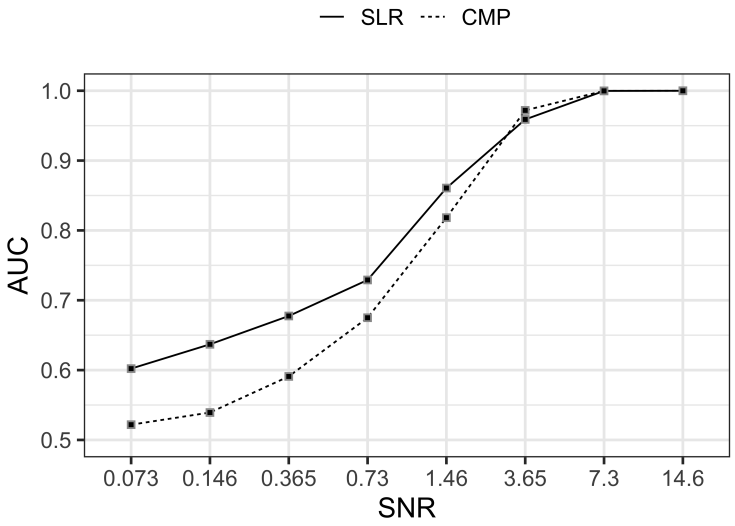
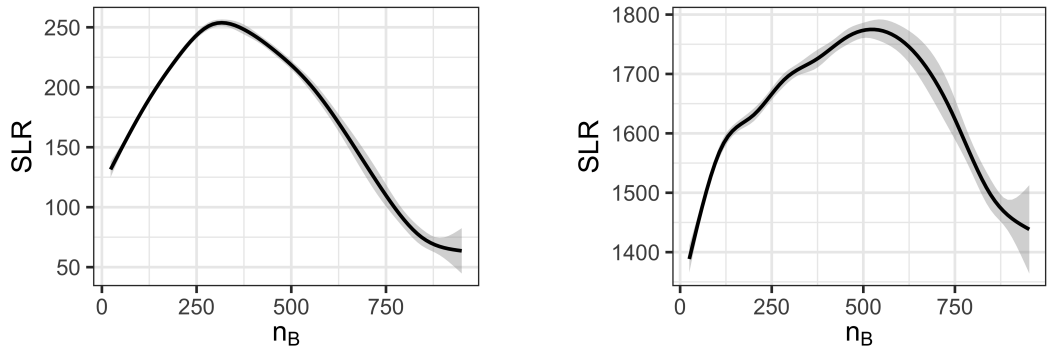


Figure 5.7: AUC values for both the SLR and CMP as a function of SNR for simulated data with $p = 0.20$.

Figure 5.7 overlays AUC curves for SLR and CMP from Figure 5.5 on the same plot. The SLR performs better than the CMP for low values of the SNR, indicating that the same source score distribution aids in quantification of degree of association when the signal-to-noise ratio is low, but the both techniques perform similarly when SNR ≥ 3.65. Therefore, the CMP results in no information loss compared to the SLR for pairs of processes exhibiting high degrees of association.

In addition to the signal-to-noise ratio, the number of events in series B also influenced

the detectability of association. We found little sensitivity in the SLR to p for any given intensity λ_A, but varying both together resulted in dramatically different behavior. We use n_b as a proxy for the combination of relative frequency and rate because it is a stochastic function of the two such that $\mathbb{E}(n_b) = wp\lambda_A$ where w is the length of the observation window. Figure 5.8 depicts a nonparametric regression of SLR on n_b for associated pairs of processes with $p = 0.20$. If the observed number of events is small, then the score function (the mean inter-event time) will have higher variance. The high variance in the score function would be expected under both the same source and different source distributions, generally leading to smaller SLR values. Conversely, if this number is large the processes become so dense that the inter-event times for independent pairs decrease (i.e., inter-event times under the different source hypothesis) and, therefore, behave more like the inter-event times of associated pairs. In either extreme, we observe lower SLRs relative to the peak value that occurs in the middle of this range. For these data the nature of the conclusion doesn't change (i.e., the SLR favors the same source hypothesis across the entire range of n_b values) but the observed patterns suggest relationships that may be important with other data.



(a) SNR = 7.3 (b) SNR = 14.6

Figure 5.8: Generalized additive model (GAM) smoother of the score-based likelihood ratio for simulated associated pairs with $p = 0.20$ as a function of the number of events in series B. Smoother fit in black and 99% confidence interval in grey. Note the different scales on the y-axes.

Note the trend exhibited in Figure 5.8 is consistent across all values of the relative frequency p and SNR. However, the magnitudes differ with smaller SNR yielding a lower peak SLR. This

makes sense intuitively because as the signal-to-noise ratio decreases, the nearest neighbors of events in B are no longer the generating events[5] in A (i.e., they can be closer to another unassociated event). Thus, process B behaves similarly to an independently generated Poisson process when SNR is small.

5.6.3 Discussion of Simulation Results

The results of the simulation study illustrate the promise of score-based approaches for quantifying strength of evidence for temporal event series. In terms of classification performance, the SLR outperformed the CMP as shown by larger AUC values across all simulation regimes. Furthermore, the SLR was well-calibrated for the simulated data sets considered as shown in ECE plots. Therefore, when population data is available, the SLR is preferred over the CMP, given its better performance for pairs exhibiting weak association (low signal-to-noise ratio) and the fact that it performs similarly to the CMP for strongly associated pairs (high SNR).

One limitation of the simulation study, however, was that the generated data sets are not realistic patterns of human behavior. We purposely chose to analyze Poisson processes (that were neither homogeneous nor bursty, as real-world event data tends to be) in order to make the simulations tractable and the results more interpretable. The simplifying assumptions increased our understanding of the performance of the evidence evaluation methods. In the following case studies, we investigate their performance on user-generated event data.

[5]Here "generating event" refers to the time t_j of the simulated event in A which was used as the mean of the Gaussian distribution to generate a given event in B in Algorithm 3.

5.7 Case Study—Student Web Browsing Data

The data in this case study attempts to mimic a situation in which an investigator obtains an event series from a known-source account stored by a cloud service provider (e.g., events from a suspect's account stored by a company such as Google or Facebook) and another event series from a device of unknown origin (obtained from a crime scene). The forensic question of interest is whether or not the events stored by the cloud service provider are associated with the events obtained from the device, i.e., if the two event series were generated by the same source.

5.7.1 Data

The data considered in this section come from an *in situ* observational study of student activity over time on digital devices, conducted at a large US university [Wang et al., 2015]. 124 undergraduate students with Windows computers voluntarily participated in the study for one week and browser activity was automatically logged. Participants were instructed to continue using their devices as normal.

The event logs were dichotomized by the type of web browsing event to create pairs of event time series (A, B) for each student. Series B (known-source data) corresponds to Facebook events (i.e., any web browser activity occurring on *facebook.com*), and series A (unknown-source) corresponds to non-Facebook events (i.e., any web browser activity not occurring on *facebook.com*). Students were included in our analysis if they had at least 50 events of each type. Of the 124 students originally recorded, 55 met the inclusion criteria. These students generated 90,340 log records, with 13,995 (15.5%) Facebook and 76,345 (84.5%) non-Facebook browser events. A graphical illustration of a subset of the data is shown in Figure 5.9.

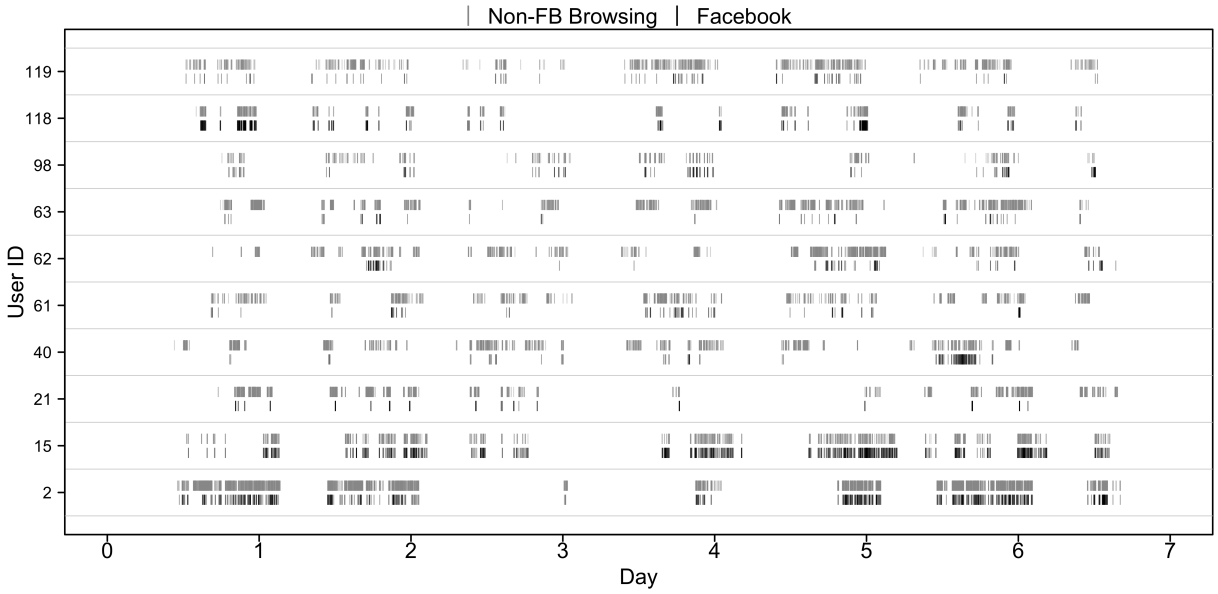
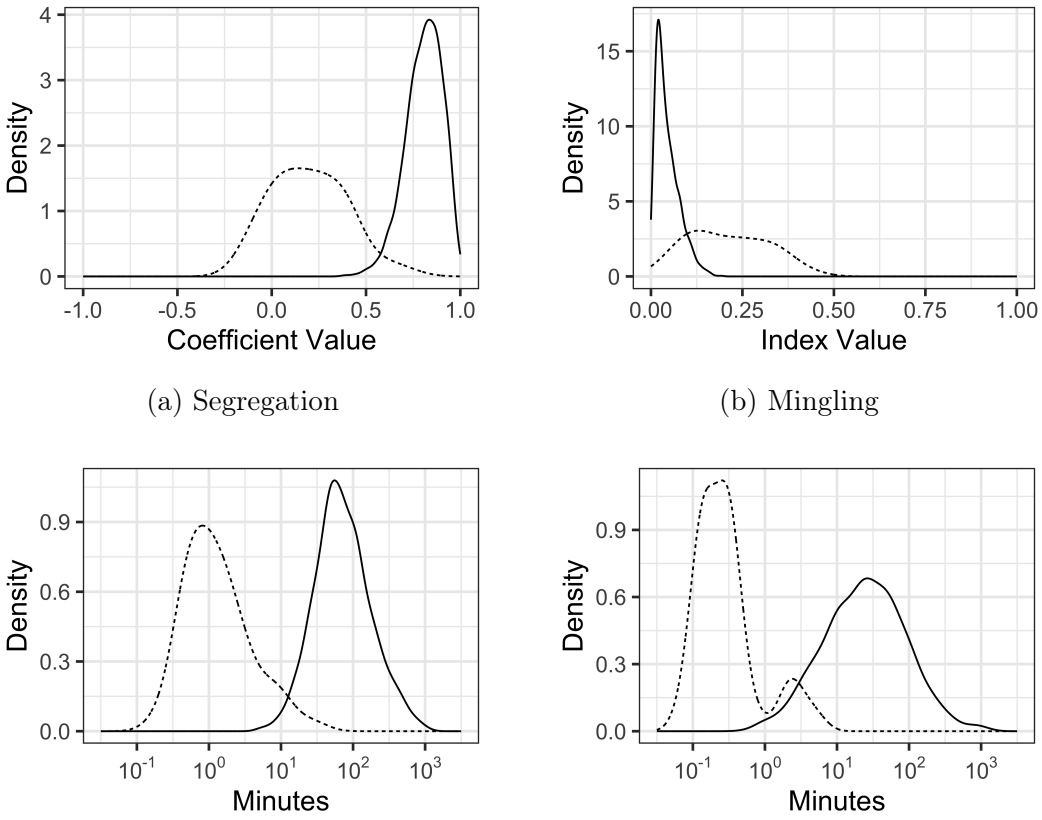


Figure 5.9: Web browsing data observed over 7 days from a random sample of 10 users from the case study data. Each user has two rows corresponding to the two event series with the top row of grey bars representing non-Facebook events (A_i) and the bottom row of black bars representing Facebook web browsing events (B_i). Note that all events shown above are relative to the first day of observation for each student, and each tick mark on the x-axis represents midnight of the corresponding day.

5.7.2 Population-based Results

Figure 5.10 shows the empirical distributions of each of the score functions for same and different source pairs as discussed in Section 5.4. Note that all pairwise combinations of the data were included in the reference data set $\mathcal{D}$ used to create these densities for illustrative purposes (leave-pairs-out cross-validation was used for the rest of the results in this section). While there is some overlap in the same and different source densities for all score functions, it is clear that the majority of the probability mass does not occur in the same region. This suggests that both the SLR and CMP should be able to accurately assess the strength of association.

Using the score-based likelihood ratios estimated via leave-pairs-out cross-validation and a threshold of 1, which corresponds to the data being equally likely to have been generated



(a) Segregation

(b) Mingling

(c) Mean IET

(d) Median IET

Figure 5.10: Empirical distributions of the score functions from Section 5.4. Same source distributions (H_s, dashed line) and different source distributions (H_d, solid line) approximated via kernel density estimation with Gaussian kernels and Scott's rule of thumb bandwidth. Leave-pairs-out cross-validation was not used to produce these densities; instead all pairs were used for illustrative purposes. Note that for the mingling index in (b) same-source pairs typically have higher score values than different-source pairs, so the inequalities in the SLR and CMP of Equations 5.8 and 5.7 are reversed.

under either hypothesis, we can compare the true- and false-positive rates for each score function. Table 5.1 provides these rates (listed as TP@1 and FP@1, respectively). Note that the SLR based on the mean inter-event time yields the highest true-positive rate and lowest false-positive rate for this particular choice of threshold. The area under the receiver (AUC) operating characteristic (ROC) curve is also given in Table 5.1. The choice of score function seems to have little impact on AUC, since all score functions yield a value greater that 0.99.

Perhaps of most interest to forensic scientists is the threshold of SLR values that gives

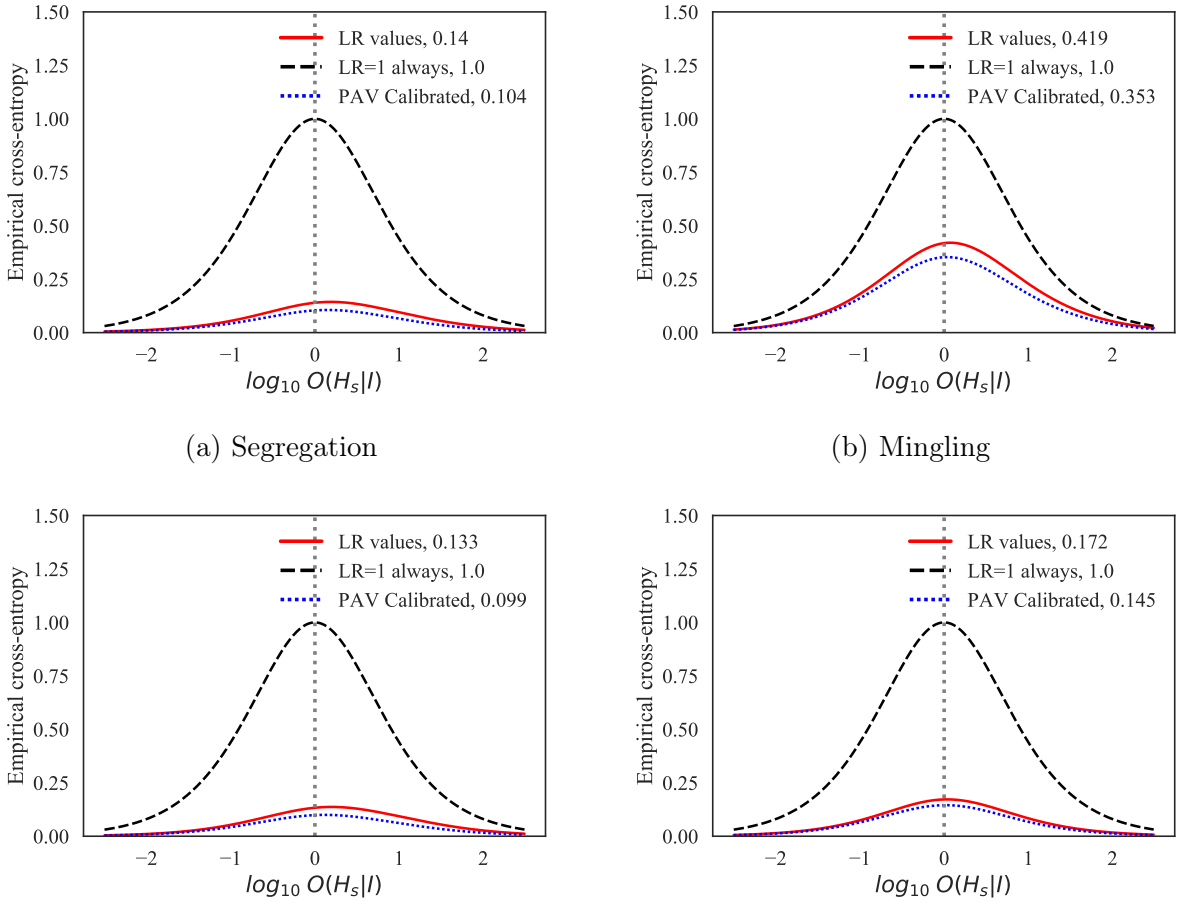
Δ	TP@1	FP@1	FP=0 Threshold	TP@FP=0	AUC
S	0.945	0.022	1871	0.745	**0.992**
$\overline{M}_1$	0.855	0.116	218	0.473	0.946
$\overline{\mathcal{T}}_{BA}$	**0.964**	**0.021**	**122**	**0.873**	**0.992**
$med(\mathcal{T}_{BA})$	0.945	0.060	279	0.818	0.990

Table 5.1: Performance of a classifier based on SLR_Δ for the student web browsing data.

a 0% empirical false-positive rate, since wrongfully convicting the innocent has extremely negative societal consequences. Table 5.1 lists this threshold and its corresponding true positive rate (TP@FP=0) for all of the score functions considered. The inter-event time summary statistics require a lower SLR than the marked point process characteristics and yield a higher true positive rate, which is evidence that they are better calibrated. Note that a pair of event series (A, B) whose SLR value is equal to 122 for the mean inter-event time, $\overline{\mathcal{T}}_{BA}$, has the following interpretation: the observed mean inter-event time for pair (A, B) was 122 times more likely to have been generated by a same source series than by different source series.

To further illustrate the calibration and discrimination performance of the SLR, ECE plots are given in Figure 5.11 for each score function considered. Regardless of the score function used, the ECE plots show that the score-based likelihood ratio is both well-calibrated (shown by the ECE for the SLR values being almost identical to the PAV calibrated ECE) and has excellent discrimination power (shown by the low ECE values regardless of prior log odds) for the student web browsing data. The best-performing scores were the coefficient of segregation (Figure 5.11a) and mean inter-event time (Figure 5.11c), which is not surprising given that these methods also performed well in terms of AUC and false positive rates (see Table 5.1). The ECE plots for the student web browsing data look similar to the ECE plot for the simulated data with a high signal-to-noise ratio (Figure 5.6c), which shows that there is very strong signal in the web browsing data.

The coincidental match probability was also computed using leave-pairs-out cross-validation,



(a) Segregation

(b) Mingling

(c) Mean IET

(d) Median IET

Figure 5.11: Empirical cross-entropy (ECE) plots for the SLR with each score function described in Section 5.4.1. C_{llr} values are provided in the legend. (a) Coefficient of segregation, $S(A, B)$; (b) mingling index, $\overline{M}_1(A, B)$; (c) mean inter-event time, $\overline{\mathcal{T}}_{BA}$; and (d) median inter-event time, $med(\mathcal{T}_{BA})$.

and Table 5.2 contains the true positive rates for thresholds of 0.05 and 0.001 along with the AUC for each score function considered. (Recall that for the CMP computed using the population data the false positive rate is fixed at the choice of the threshold.) In terms of AUC, the best performing CMP method had a slightly larger AUC (0.996 for $\overline{\mathcal{T}}_{BA}$ in Table 5.2) than the best performing SLR method (0.992 for $\overline{\mathcal{T}}_{BA}$ in Table 5.1). This phenomenon is not too surprising, however, given the separation between same- and different-source score distributions in Figure 5.10. Computing tail probabilities of the different-source score distribution is adequate for distinguishing between the source hypotheses for this data. The same-source

Δ	TP@0.05	TP@0.001	AUC
S	0.982	0.945	0.995
$\overline{M}_1$	0.800	0.618	0.950
$\overline{\mathcal{T}}_{BA}$	**0.982**	**0.891**	**0.996**
$med(\mathcal{T}_{BA})$	0.945	0.855	0.992

Table 5.2: Performance of a classifier based on CMP_Δ computed using population data for the student web browsing data.

score distribution used in the numerator of the SLR, however, provides additional information resulting in increasing SLRs as one moves further into the tail of the different-source score distribution while the CMP is constrained.

5.7.3 Resampling Results

We now consider the case where we only have a single pair of event series (A^*, B^*) available for analysis—for example, for any pair of event series in our case study data we would like to assess the degree of association between that pair only using the data for the pair. We used sessionized resampling with B^* fixed to estimate the coincidental match probability (CMP) for each pair of series using both the mean and median inter-event times. In order to more directly compare to the results of the population-based approach, $p(t_{ses})$ was chosen to be the empirical distribution of all session start times in the data excluding those from the particular pair (A^*, B^*) being analyzed in a fashion similar to leave-pairs-out cross-validation. We define the estimated CMP as the fraction of simulated pairs whose score function is less than that of the observed pair.

For each of the $55^2 = 3025$ pairwise combinations of the event series, 1000 samples were generated via sessionized resampling (Algorithm 2). Similar to our population-based approach, we can view the CMP as a discriminant function for a binary classification decision (e.g., pairs with CMP values less than some threshold are considered same source) and compare the true and false positive rates for each score function. Table 5.3 provides these rates for

thresholds of 0.05 and 0.001. Note that a pair of event series (A^*, B^*) whose CMP value is equal to one of these thresholds has the following interpretation: the probability that a score comparable to or lower than that obtained from the pair (A^*, B^*) was generated by different source event series is 5% (or 0.1%).

Δ	TP@.05	FP@.05	TP@.001	FP@.001	AUC
S	1	**0.091**	0.800	0.008	0.998
$\overline{M}_1$	1	0.098	0.800	0.009	0.997
$\overline{\mathcal{T}}_{BA}$	1	0.146	0.800	0.023	0.992
$med(\mathcal{T}_{BA})$	1	0.098	**0.982**	**0.004**	**0.999**

Table 5.3: Performance of a classifier based on CMP_Δ computed via resampling for the student web browsing data.

The resampling approach to computing the coincidental match probability worked quite well in quantifying the degree of association for this data, as shown in Table 5.3. Across all score functions, the CMPs computed with sessionized resampling resulted in higher true positive rates than their population-based counterparts (the results in Table 5.2). However, the false positive rate tends to be much higher, e.g., the best FP rate for the resampling approach with a threshold of 0.05 was 0.091 (the population-based FP is 0.05). When using AUC as the evaluation metric, the CMP (e.g., the best AUC in Table 5.2 was 0.996 and in Table 5.3 was 0.999) performed better than the SLR (0.992 in Table 5.1). We also observed this phenomenon in the simulation study for pairs of processes exhibiting a high signal-to-noise ratio, which implies that the student web browsing data is well-suited for the resampling approach because it is comprised of highly associated pairs of same source event series.

5.7.4 Discussion of Student Web Browsing Results

Score-based approaches worked well in quantifying the strength of evidence in the student web browsing data. The score-based likelihood ratio showed excellent discrimination performance and calibration properties, which were similar to those of the simulation studies with

high signal-to-noise ratio values. The coincidental match probability also worked well in this data set regardless of whether it was computed from population data or by resampling. Overall, the results show that the techniques presented in this chapter work well for web browsing data, albeit in a relatively small, homogeneous sample of individuals.

5.8 Case Study—LANL Authentication Data

Suppose that an examiner is given two sets of timestamped authentication, or login, events. The two series of events are composed of logins to a user's personal computer and a shared computer, respectively. The examiner is tasked with quantifying the association between these event series in order to determine if both were generated by the same user (i.e., the user whose personal computer authentication event series was collected) or not. Further assume that the examiner is only presented with the event series in question and does not have access to a population of similar event series.

5.8.1 Data

We show a proof of concept of the efficacy of the coincidental match probability on this task with real authentication data. The data represent successful authentication, or login, events from users to computers on the Los Alamos National Laboratory (LANL) enterprise network [Kent, 2014; Hagberg et al., 2014]. Each authentication event is composed of a timestamp (represented by the number of seconds from some unknown origin time), the user account that generated the event (the *actor*), and the computer that was logged into (the *target*). We focus on two users—Actor 1 and Actor 2—and their authentication events to three particular target computers—Target X, Target Y and Target Z—during the first day of available data.[6]

[6]For simplicity, the users and computers were renamed. In the original LANL data, Actor 1 is U4116, Actor 2 is U7250, Target X is C4751, Target Y is C8268, and Target Z is C248.

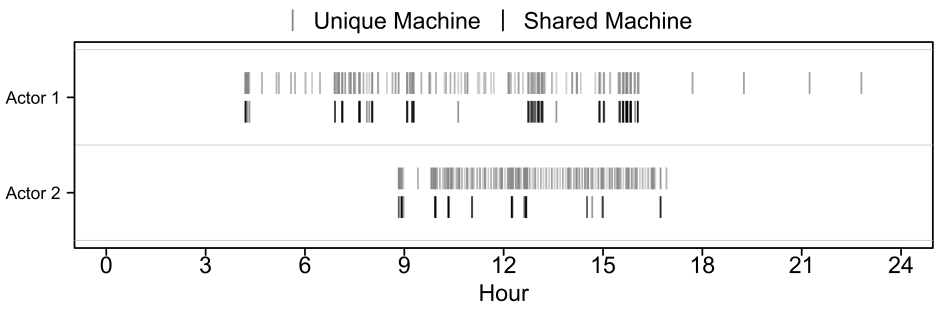


Figure 5.12: LANL authentication data. Unique machine refers to Target X and Target Y for Actor 1 and Actor 2, respectively, and shared machine refers to Target Z for both actors.

Target X and Target Y are only logged into by Actor 1 and Actor 2, respectively, and not by any other users in the entirety of the LANL authentication data (and are referred to as the unique machines). Both users authenticate to Target Z (the shared machine). Table 5.4 gives the number of authentication events for each user-computer pair, and Figure 5.12 shows the event series themselves.

User	Target X	Target Y	Target Z
Actor 1	318	0	120
Actor 2	0	362	53

Table 5.4: Number of login events for each user to each target computer on the first day of activity in the LANL authentication data.

5.8.2 Results

We computed the coincidental match probability for each pairwise combination of event series available. Thus, there were two same-source pairs (Actor 1's authentications to machines Target X and Target Z, and Actor 2's authentications to machines Target Y and Target Z) and two different-source pairs (Actor 1-Target Z, Actor 2-Target Y and Actor 2-Target Z, Actor 1-Target X). In each case, authentication events to the shared machine (series A) were fixed and the event times of authentications to the unique machine (series B) were resampled $10,000$ times via sessionized resampling. Session start times were sampled from

Same-Source	Shared Machine (A)	Unique Machine (B)	$\overline{\mathcal{T}}_{BA}$	$med(\mathcal{T}_{BA})$	S	$\overline{M}_1$
Yes	Actor 1, Target Z	Actor 1, Target X	0.000	0.000	0.000	0.000
	Actor 2, Target Z	Actor 2, Target Y	0.000	0.000	0.000	0.000
No	Actor 1, Target Z	Actor 2, Target Y	0.200	0.325	0.267	0.240
	Actor 2, Target Z	Actor 1, Target X	0.607	0.799	0.947	0.930

Table 5.5: Coincidental match probabilities for various score functions for the LANL authentication data. Lower scores are indicative of same source event series.

a uniform distribution over the range of observed event times in series B (e.g., logins to the unique machine). Note that other distributions over session start times, including Gaussian and empirical distributions, yield similar conclusions (see Galbraith et al. [2020a] for these results). The resulting CMPs for each score function discussed in Section 5.4 are provided in Table 5.5.

Across all score functions, the same-source authentication event series for both actors exhibit CMPs numerically equivalent to 0, which is strongly indicative that they were in fact generated by the same user. Conversely, the different-source event series exhibit CMPs ranging from 0.20 to 0.95, indicating that the association of these series was at a level that would be typical for different-source series and unlikely for same-source series. Overall, the resampling-based approach proved effective for this particular data set.

5.9 Discussion

Drawing on previous work from the forensics and statistics literature, we explored a variety of measures for quantifying the association between two discrete event time series. Multiple score functions were used to determine the similarity between the series, including characteristics of marked point processes (coefficient of segregation and mingling index) and inter-event time summary statistics (mean and median). These score functions were shown

to be discriminative for same and different source pairs of event series.

We then proposed two methods for assessing the strength of association for a given pair of event series. The population-based approach uses a sample from the relevant population to construct score-based likelihood ratios that assess the relative likelihood of observing a given degree of association when the series came from the same or different sources. The resampling approach considers only a single pair of event series, simulates a different source score distribution via sessionized resampling, and uses that distribution to calculate coincidental match probabilities.

While the population-based approach with score-based likelihood ratios remains the preferred technique in terms of accuracy and interpretability, our proposed resampling technique with coincidental match probabilities shows considerable promise for assessing the degree of association of pairs of user-generated event series. It should be noted, however, that both techniques require more extensive study and testing before being used in practice by forensic examiners. Overall, work in the area of quantifying association of user-event data shows considerable promise and potential efficacy in both forensic and cybersecurity settings.

5.9.1 Contributions

The following contributions were made in this chapter:

- A variety of appropriate score functions that can distinguish between same- and different-source series of temporal events.
- Novel techniques for quantifying the strength of evidence for temporal event data via the score-based likelihood ratio and coincidental match probability.
- A resampling technique that allows for the computation of strength of evidence when a reference population is not available.

Chapter 6

Discussion on Future Directions

This thesis has presented the current state of the application of statistical techniques to same-source forensic questions for digital evidence. Chapters 2 and 3 provided an overview of common techniques used to quantify and evaluate the strength of evidence across a variety of forensic disciplines. Chapters 4 and 5 adapted these techniques to user-generated event data in the spatial and temporal domains, respectively. Numerous case studies illustrate the potential efficacy of the methods and introduce a number of new research questions that should be explored in the future, which are discussed in further detail below.

- **Reference Data**

  The data sets used in the case studies attempt to mimic real-world scenarios that an investigator may encounter. They were chosen for the features present and convenience in collection (i.e., either public source or easy to collect via an API). Therefore, they are not entirely representative of what one would find in practice. More studies must be performed with realistic case data before use in the courtroom. Obtaining such data sets is non-trivial, however, due to the sensitive nature of digital event data. This issue is common to all forensic disciplines, but, being that the field of digital

evidence is relatively young, databases do not exist like they do for DNA (e.g., the FBI's Combined DNA Index System (CODIS)) or fingerprints [e.g., Garris and McCabe, 2000; Tackett, 2020]. An effort to create such databases for digital evidence exists, but so far has been limited to tampering activities such as mass deletion, reformatting and backdating [Casey, 2013, 2020]. To facilitate the development, evaluation and use of digital evidence evaluation techniques, robust databases of digital evidence must be created and shared among researchers and law enforcement agencies. Unlike DNA and fingerprint databases, however, digital evidence databases will need to be diverse in the types of information they store. For instance, multiple databases for geolocated event data must exist for regions around the United States, since events occurring in Los Angeles are of no use for a case in South Dakota. Companies that offer location-based services (e.g., Google, Apple, telecom companies, etc.) have rich datasets that could be leveraged for development and application of statistical techniques for quantifying strength of evidence in this setting, but there are complex legal issues in using such data for forensic purposes. Therefore, practitioners and researchers must work together in the effort to create a wide variety of digital evidence databases, potentially beginning by sharing case data. A large part of this effort will be in developing protocols for how the information is shared and assuring the privacy and compliance of all individuals involved.

- **Assessment Techniques**

  Numerous methods of assessing the performance of an evidence evaluation technique on a validation data set were presented in this dissertation. In traditional forensic science disciplines, assessing the classification performance and calibration of a method is relatively straightforward. We do not expect, for instance, two fingerprints from the same individual to look completely different when observed at two different times (assuming there is not significant degradation or occlusion of the print). That is not the case for digital evidence, especially so for user-generated event data. Natural vari-

ability in human behavior manifests when an individual's underlying generative model for event locations/times shifts or encounters a changepoint. It would be unreasonable if the evidence evaluation method were to quantify the strength of evidence in a manner that results in the investigator concluding such shifted patterns were generated by the same individual. This phenomenon makes assessing evidence evaluation techniques difficult and time-consuming. Developing extensions of the discrimination and calibration assessments that systematically account for these "misclassified" exemplars will help to speed up the validation of digital evidence evaluation methods and make the process uniform across different researchers.

- **Discovery**

  The techniques proposed in this dissertation are well-suited for use as an investigatory tool for "discovery," e.g., the task of finding the most similar known source in a database given an unknown source sample. Such tools exist for other evidence types, such as automatic fingerprint identification systems (AFIS) that allow law enforcement agents to search for suspects by uploading a fingerprint image from a crime scene [Komarinski, 2005]. Assuming that relevant databases for digital evidence exist, the remaining open problem becomes speeding up the computation of the likelihood ratio, score-based likelihood ratio or coincidental match probability to reduce the time it takes to return the most likely suspects for a given digital trace. Furthermore, the assessment issues mentioned above would be alleviated since the human in the loop (e.g., the investigator looking at the most similar suspects' data) could reduce the false positive rate and aid in calibration issues.

- **Modeling Extensions**

  The applications to spatial (Chapter 4) and temporal (Chapter 5) event data represent a baseline for the forensic analysis of such data. Many other modeling choices should be explored in the future. For instance, models that take into account both

the spatial locations and times at which events occur are a natural next step. The models considered in the applications only considered two types of events (represented by A and B), but higher-dimensional extensions that take into account an arbitrary number of event types are also possible. Other extensions, including incorporating event metadata would also aid in discriminating between same- and different-source evidence. One example of this metadata is the user-agent string, which encodes information about the device that the event was generated on (e.g., the operating system and version). Other metadata is context-dependent, such as incorporating natural language processing models for email or SMS text, or taking into account the target of the event (e.g., the recipient of the message) in a graphical model.

- **Missing Data**

  In many cases digital evidence has been tampered with, so only part of the evidence is observed by the examiner [e.g., Casey, 2020]. A missing data problem can be framed from such scenarios. Using the score-based likelihood ratio for spatial event data as an example, one would only observe a fraction of the locations at which an individual visited during the observation window. Thus, there would be uncertainty in the value of the score function due to these unobserved events that would propagate into uncertainty in the value of the SLR. Numerous approaches could be taken to estimate this uncertainty, e.g., a fully Bayesian approach in which a posterior distribution of the SLR is approximated via data augmentation [e.g., Tanner and Wong, 1987; Little and Rubin, 2002, chapter 10]. Alternative approaches, such as marginalizing out the uncertainty in the score value to compute a point estimate of the SLR are also possible.

Digital evidence is being encountered with increasing regularity in forensic investigations, but current tools are ill-suited for answering same-source questions that meet the scientific standards for presentation in a court of law. The work developed within this dissertation provides a starting point for the statistical modeling of event data in the context of forensic

analysis of digital evidence, as well as motivating many research topics for future investigation.

Bibliography

R. K. Ahuja and J. B. Orlin. A fast scaling algorithm for minimizing separable convex functions subject to chain constraints. *Operations Research*, 49(5):784–789, 2001.

C. Aitken and D. Stoney. *The Use of Statistics in Forensic Science*. Ellis Horwood Limited, Chichester, England, 1991.

C. Aitken and F. Taroni. *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley & Sons, 2nd edition, 2004.

C. Aitken, A. Nordgaard, F. Taroni, and A. Biedermann. Commentary: Likelihood ratio as weight of forensic evidence: A closer look. *Frontiers in Genetics*, 9:224, 2018.

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

T. Ali, L. Spreeuwers, R. Veldhuis, and D. Meuwly. Effect of calibration data on forensic likelihood ratio from a face recognition system. *IET Biometrics*, 3:335–346, 2014.

J. Almirall and T. Trejos. Advanced in the forensic analysis of glass fragments with a focus on refractive index and elemental analysis. *Forensic Science Review*, 18(2):73–96, 2006.

Association of Forensic Science Providers. Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice*, 49:161–164, 2009.

A. Baddeley, P. J. Diggle, A. Hardegen, T. Lawrence, R. K. Milne, and G. Nair. On tests of spatial pattern based on simulation envelopes. *Ecological Monographs*, 84(3):477–489, 2014.

A. Baddeley, E. Rubak, and R. Turner. *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC, 2015.

S. Baechler, V. Terrasse, J.-P. Pujol, T. Fritz, O. Ribaux, and P. Margot. The systematic profiling of false identity documents: Method validation and performance evaluation using seizures known to originate from common and different sources. *Forensic Science International*, 232(1):180–190, 2013.

A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039): 207–211, 2005.

R. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical Inference Under Order Restrictions*. John Wiley & Sons, 1972.

J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2013.

M. Berman. Testing for spatial association between a point process and another stochastic process. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 35(1): 54–62, 1986.

M. J. Best and N. Chakravarti. Active set algorithms for isotonic regression; a unifying framework. *Mathematical Programming*, 47(1-3):425–439, 1990.

N. Boers, B. Bookhagen, N. Marwan, and J. Kurths. Spatiotemporal characteristics and synchronization of extreme rainfall in South America with focus on the Andes Mountain range. *Climate Dynamics*, 46(1-2):601–617, 2016.

A. Bolck, H. Ni, and M. Lopatka. Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: Applied to forensic MDMA comparison. *Law, Probability and Risk*, 14(3):243–266, 2015.

W. Bosma, S. Dalm, E. van Eijk, R. el Harchaoui, E. Rijgersberg, H. T. Tops, A. Veenstra, and R. Ypma. Establishing phone-pair co-usage by comparing mobility patterns. *Science & Justice*, 60:180–190, 2020.

S. Bozza, F. Taroni, R. Marquis, and M. Schmittbuhl. Probabilistic evaluation of handwriting evidence: likelihood ratio for authorship. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(3):329–341, 2008.

L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19(2):135–144, 1977.

N. Brümmer. *Measuring, Refining and Calibrating Speaker and Language Information Extracted from Speech*. PhD thesis, University of Stellenbosch, 2010.

N. Brümmer and J. du Preez. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2-3):230–275, 2006.

N. Brümmer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim. Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST Speaker Recognition Evaluation 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2072–2084, 2007.

E. Casey. *Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet*. Academic Press, 2011.

E. Casey. *Reinforcing the scientific method in digital investigations using a case-based reasoning (CBR) system*. PhD thesis, University College Dublin, 2013.

E. Casey. Clearly conveying digital forensic results. *Digital Investigation*, 24:1–3, 2018.

E. Casey. Standardization of forming and expressing preliminary evaluative opinions on digital evidence. *Forensic Science International: Digital Investigation*, 2020.

E. Casey, D.-O. Jaquet-Chiffelle, H. Spichiger, E. Ryser, and T. Souvignet. Structuring the evaluation of location-related mobile device evidence. *Forensic Science International: Digital Investigation*, 2020.

C. Champod and I. W. Evett. A probabilistic approach to fingerprint evidence. *Journal of Forensic Identification*, 51(2):101–122, 2001.

C. Champod and D. Meuwly. The inference of identity in forensic speaker recognition. *Speech Communication*, 31(2-3):193–203, 2000.

E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1082–1090, 2011.

S. Cohen. *Finding Color and Shape Patterns in Images*. PhD thesis, Stanford University, 1999.

R. Coleman and H. Walls. Evaluation of scientific evidence. *Criminal Law Review*, pages 276–288, 1974.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*. Springer, 2nd edition, 2007.

Daubert v. Merrell Dow Pharmaceuticals, Inc. 509 U.S. 579. 1993.

M. H. DeGroot and S. E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

P. J. Diggle. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC press, 2013.

P. J. Diggle and A. G. Chetwynd. Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, pages 1155–1163, 1991.

C. G. Ding. Algorithm AS 275: Computing the non-central χ^2 distribution function. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):478–482, 1992.

P. M. Dixon. Ripley's K function. *Wiley StatsRef: Statistics Reference Online*, 2014.

J. F. Donges, C.-F. Schleussner, J. F. Siegmund, and R. V. Donner. Event coincidence analysis for quantifying statistical interrelationships between event time series. *The European Physical Journal Special Topics*, 225(3):471–487, 2016.

I. W. Evett and B. S. Weir. *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sinauer, 1998.

T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

N. Fenton, M. Neil, and D. Berger. Bayes and the law. *Annual Review of Statistics and Its Application*, 3:51–77, 2016.

G. A. Fine. Review of the FBI's handling of the Brandon Mayfield case. *Office of the Inspector General, Oversight and Review Division, US Department of Justice*, 2006.

K. F. Gaines, A. L. Bryan Jr, and P. M. Dixon. The effects of drought on foraging habitat selection of breeding wood storks in coastal Georgia. *Waterbirds*, pages 64–73, 2000.

C. Galbraith and P. Smyth. Analyzing user-event data using score-based likelihood ratios with marked point processes. *Digital Investigation*, 22:S106–S114, 2017.

C. Galbraith, P. Smyth, and H. S. Stern. Quantifying the association between discrete event time series with applications to digital forensics. *Journal of the Royal Statistical Society: Series A (Statistics in Society), in press*, 2020a.

C. Galbraith, P. Smyth, and H. S. Stern. Statistical methods for the forensic analysis of geolocated event data. *Digital Investigation, in press*, 2020b.

M. D. Garris and R. M. McCabe. NIST special database 27: Fingerprint minutiae from latent and matching tenprint images. Technical report NISTIR 6534, National Institute of Standards and Technology, 2000.

N. Garton, D. Ommen, J. Niemi, and A. Carriquiry. Score-based likelihood ratios to evaluate forensic pattern evidence. *arXiv preprint arXiv:2002.09470*, 2020.

E. J. Garvin and R. D. Koons. Evaluation of match criteria used for the comparison of refractive index of glass fragments. *Journal of Forensic Sciences*, 56(2):491–500, 2011.

P. Gill and T. Clayton. The current status of DNA profiling in the UK. *Handbook of Forensic Science*, pages 29–56, 2009.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, D. Ramos-Castro, and J. Ortega-Garcia. Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems. *Forensic Science International*, 155(2-3):126–140, 2005.

I. J. Good. Rational decisions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 107–114, 1952.

D. Grove. The interpretation of forensic evidence using a likelihood ratio. *Biometrika*, 67 (1):243–246, 1980.

A. Hagberg, A. Kent, N. Lemons, and J. Neil. Credential hopping in authentication graphs. In *2014 International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*. IEEE Computer Society, Nov. 2014.

P. Haggett, A. D. Cliff, and A. Frey. *Locational analysis in human geography*, volume 2. Arnold London, 1977.

K. Hanisch. Some remarks on estimators of the distribution function of nearest neighbour distance in stationary spatial point processes. *Series Statistics*, 15(3):409–412, 1984.

R. Haraksim, D. Ramos, D. Meuwly, and C. E. Berger. Measuring coherence of computer-assisted likelihood ratio methods. *Forensic Science International*, 249:123–132, 2015.

A. B. Hepler, C. P. Saunders, L. J. Davis, and J. Buscaglia. Score-based likelihood ratios for handwriting evidence. *Forensic Science International*, 219(1):129–140, 2012.

F. L. Hitchcock. The distribution of a product from several sources to numerous localities. *Journal of Mathematics and Physics*, 20(1-4):224–230, 1941.

V. Hughes. *The Definition of the Relevant Population and the Collection of Data for Likelihood Ratio-based Forensic Voice Comparison*. PhD thesis, University of York, 2014.

J. D. Hunter and J. G. Milton. Amplitude and frequency dependence of spike timing: implications for dynamic regulation. *Journal of Neurophysiology*, 90(1):387–394, 2003.

J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan. *Statistical Analysis and Modelling of Spatial Point Patterns*. John Wiley & Sons Ltd, West Sussex, England, 2008.

V. Isham. An introduction to spatial point processes and markov random fields. *International Statistical Review*, pages 21–43, 1981.

E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620, 1957.

C. F. F. Karney. Algorithms for geodesics. *Journal of Geodesy*, 87(1):43–55, 2013.

D. H. Kaye. Logical relevance: Problems with the reference population and DNA mixtures in People v. Pizarro. *Law, Probability and Risk*, 3(3-4):211–220, 2004.

H. Kelly, J.-A. Bright, J. S. Buckleton, and J. M. Curran. A comparison of statistical models for the analysis of complex forensic DNA profiles. *Science & Justice*, 54(1):66–70, 2014.

A. D. Kent. User-computer authentication associations in time. Los Alamos National Laboratory, 2014.

P. Komarinski. *Automated Fingerprint Identification Systems (AFIS)*. Elsevier, 2005.

B. Kong, J. Supancic, D. Ramanan, and C. C. Fowlkes. Cross-domain image matching with deep feature maps. *International Journal of Computer Vision*, 127(11-12):1738–1750, 2019.

D. Kotzias, M. Lichman, and P. Smyth. Predicting consumption patterns with repeated and novel events. *IEEE Transactions on Knowledge and Data Engineering*, 31(2):371–384, 2018.

W. J. Krzanowski and D. J. Hand. *ROC Curves for Continuous Data*. Chapman and Hall/CRC, 2009.

M. Lichman. *Context-Based Smoothing for Personalized Prediction Models*. PhD thesis, UC Irvine, 2017.

M. Lichman and P. Smyth. Modeling human location data with mixtures of kernel densities. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 35–44, 2014.

D. V. Lindley. A problem in forensic science. *Biometrika*, 64(2):207–213, 1977.

R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2nd edition, 2002.

J. M. Loh. A valid and fast spatial bootstrap for correlation functions. *The Astrophysical Journal*, 681(1):726, 2008.

S. P. Lund and H. Iyer. Likelihood ratio as weight of forensic evidence: A closer look. *Journal of research of National Institute of Standards and Technology*, 122(27), 2017.

N. Malik, N. Marwan, and J. Kurths. Spatial structures and directionalities in monsoonal precipitation over south asia. *Nonlinear Processes in Geophysics*, 17(5):371–381, 2010.

K. A. Martire, R. I. Kemp, I. Watkins, M. A. Sayle, and B. R. Newell. The expression and interpretation of uncertain forensic science evidence: Verbal equivalence, evidence strength, and the weak evidence effect. *Law and Human Behavior*, 37(3):197, 2013.

D. Meuwly, D. Ramos, and R. Haraksim. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International*, 276:142–153, 2017.

G. S. Morrison. Likelihood-ratio-based forensic speaker comparison using parametric representations of vowel formant trajectories. *Journal of the Acoustical Society of America*, 125 (4), 2009.

G. S. Morrison and E. Enzinger. Score based procedures for the calculation of forensic likelihood ratios—Scores should take account of both similarity and typicality. *Science & Justice*, 58(1):47–58, 2018.

A. H. Murphy and R. L. Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 26(1):41–47, 1977.

S. P. Myers, M. D. Timken, M. L. Piucci, G. A. Sims, M. A. Greenwald, J. J. Weigand, K. C. Konzak, and M. R. Buoncristiani. Searching for first-degree familial relationships in California's offender DNA database: Validation of a likelihood ratio-based approach. *Forensic Science International: Genetics*, 5(5):493–500, 2011.

National Commission on Forensic Science. Ensuring that forensic analysis is based upon task-relevant information. Technical report, National Institute of Standards & Technology, 2015.

National Registry of Exonerations. Exonerations by contributing factor. `https://www.law.umich.edu/special/exoneration/Pages/ExonerationsContribFactorsByCrime.aspx`. Accessed: 2020-04-20.

C. Neumann and M. A. Ausdemore. Defence against the modern arts: the curse of statistics "score-based likelihood ratios". *arXiv preprint arXiv:1910.05240*, 2019.

J. Neyman and E. S. Pearson. IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632, 2005.

J. T. Niehof and S. K. Morley. Determining the significance of associations between two series of discrete events: Bootstrap methods. Report LA-14453, Los Alamos National Laboratory, 2012.

A. Nordgaard and T. Höglund. Assessment of approximate likelihood ratios from continuous distributions: A case study of digital camera identification. *Journal of Forensic Sciences*, 56(2):390–402, 2011.

A. Nordgaard and B. Rasmusson. The likelihood ratio as value of evidence—More than a question of numbers. *Law, Probability and Risk*, 11(4):303–315, 2012.

A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in Foursquare. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

J. Oh, S. Lee, and S. Lee. Advanced evidence collection and analysis of web browser activity. *Digital Investigation*, 8:S62–S70, 2011.

D. M. Ommen and C. P. Saunders. Building a unified statistical framework for the forensic identification of source problems. *Law, Probability and Risk*, 17(2):179–197, 2018.

J. B. Parker. A statistical treatment of identification problems. *Journal of the Forensic Science Society*, 6(1):33–39, 1966.

J. B. Parker. The mathematical evaluation of numerical evidence. *Journal of the Forensic Science Society*, 7(3):134–144, 1967.

PCAST. Report to the president: Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods. Technical report, Executive Office of the President's Council of Advisors on Science and Technology, 2016.

C. S. Peirce. The probability of induction, 1878. Reprinted (1956) in The World of Mathematics, Vol. 2, Ed. J. R. Newman, pp. 1341–1354. New York: Simon and Schuster.

E. Pielou. *Mathematical Ecology*. John Wiley & Sons, Inc., 1977.

S. Pigeon, P. Druyts, and P. Verlinde. Applying logistic regression to the fusion of the NIST'99 1-speaker submissions. *Digital Signal Processing*, 10(1-3):237–248, 2000.

M. Pollitt, E. Casey, D.-O. Jaquet-Chiffelle, and P. Gladyshev. A framework for harmonizing forensic science practices and digital/multimedia evidence. OSAC Task Group on Digital/Multimedia Science, OSAC Technical Series 0002R1, NIST Organization of Scientific Area Committees for Forensic Science, 2019.

R. Q. Quiroga, A. Kraskov, T. Kreuz, and P. Grassberger. Performance of different synchronization measures in real data: A case study on electroencephalographic signals. *Physical Review E*, 65(4):041903, 2002.

F. Radicchi. Human activity in the web. *Physical Review E*, 80(2):026118, 2009.

D. Ramos. *Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems*. PhD thesis, Universidad Autonoma de Madrid, 2007.

D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, and C. Aitken. Information-theoretical assessment of the performance of likelihood ratio computation methods. *Journal of Forensic Sciences*, 58(6):1503–1518, 2013.

D. Ramos, R. P. Krish, J. Fierrez, and D. Meuwly. From biometric scores to forensic likelihood ratios. In *Handbook of Biometrics for Forensic Science*, pages 305–327. Springer, 2017.

B. D. Ripley. Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(2):172–192, 1977.

B. Robertson and G. A. Vignaux. *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. Oxford University Press, 1995.

P. Rose. Going and getting it—Forensic speaker recognition from the perspective of a traditional practitioner-researcher. In *Australian Research Council Network in Human Communicative Science Workshop: FSI not CSI—Perspectives in State-of-the-Art Forensic Speaker Recognition, Sydney*, 2007.

A. Ross, K. Nandakumar, and A. Jain. *Handbook of Multibiometrics*. Springer, New York, NY, 2006.

V. Roussev. Digital forensic science: Issues, methods, and challenges. *Synthesis Lectures on Information Security, Privacy, & Trust*, 8(5):1–155, 2016.

V. Roussev and S. McCulley. Forensic analysis of cloud-native artifacts. *Digital Investigation*, 16:S104–S113, 2016.

Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pages 59–66. IEEE, 1998.

M. J. Saks, T. Albright, T. L. Bohan, B. E. Bierer, C. M. Bowers, M. A. Bush, P. J. Bush, A. Casadevall, S. A. Cole, M. B. Denton, S. S. Diamond, R. Dioso-Villa, J. Epstein, D. Faigman, L. Faigman, S. E. Fienberg, B. L. Garrett, P. C. Giannelli, H. T. Greely, E. Imwinkelried, A. Jamieson, K. Kafadar, J. P. Kassirer, J. Koehler, D. Korn, J. Mnookin, A. B. Morrison, E. Murphy, N. Peerwani, J. L. Peterson, D. M. Risinger, G. F. Sensabaugh, C. Spiegelman, H. S. Stern, W. C. Thompson, J. L. Wayman, S. Zabell, and R. E. Zumwalt. Forensic bitemark identification: Weak foundations, exaggerated claims. *Journal of Law and the Biosciences*, 3(3):538–575, 2016.

M. Schlather, P. J. Ribeiro Jr, and P. J. Diggle. Detecting dependence between marks and locations of marked point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):79–93, 2004.

D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992.

C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27 (3):379–423, 1948.

J. F. Siegmund, N. Siegmund, and R. V. Donner. Coincalc—A new R package for quantifying simultaneities of event series. *Computers & Geosciences*, 98:64–72, 2017.

M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa. A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *INFORMS Journal on Computing*, 15(2):171–190, 2003.

C. D. Steele and D. J. Balding. Statistical evaluation of forensic DNA profile evidence. *Annual Review of Statistics and Its Application*, 1:361–384, 2014.

H. S. Stern. Statistical issues in forensic science. *Annual Review of Statistics and Its Application*, 4(1):225–244, 2017.

SWGDE. Best practices for mobile device evidence collection & preservation, handling, and acquisition. Technical report, Scientific Working Group on Digital Evidence, 2019.

R. Synowiecki. Consistency and application of moving block bootstrap for non-stationary time series with periodic and almost periodic structure. *Bernoulli*, 13(4):1151–1178, 2007.

M. Tackett. *Creating Fingerprint Databases and a Bayesian Approach to Quantify Dependencies in Evidence*. PhD thesis, University of Virginia, 2020.

Y. Tang and S. N. Srihari. Likelihood ratio estimation in forensic identification using similarity and rarity. *Pattern Recognition*, 47(3):945–958, 2014.

M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

W. C. Thompson. How should forensic scientists present source conclusions? *Seton Hall Law Review*, 48:773, 2017.

W. C. Thompson and E. J. Newman. Lay understanding of forensic statistics: Evaluation of random match probabilities, likelihood ratios, and verbal equivalents. *Law and Human Behavior*, 39(4):332–349, 2015.

W. C. Thompson, R. H. Grady, E. Lai, and H. S. Stern. Perceived strength of forensic scientists' reporting statements about source conclusions. *Law, Probability and Risk*, 17 (2):133–155, 2018.

J. Valentino-DeVries. Tracking phones, Google is a dragnet for the police. *The New York Times*, 2019. URL `https://www.nytimes.com/interactive/2019/04/13/us/google-location-tracking-police.html`.

P. Vergeer, A. Bolck, L. J. Peschier, C. E. Berger, and J. N. Hendrikse. Likelihood ratio methods for forensic comparison of evaporated gasoline residues. *Science & Justice*, 54(6): 401–411, 2014.

Q. H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pages 307–333, 1989.

Y. Wang, M. Niiya, G. Mark, S. Reich, and M. Warschauer. Coming of age (digitally): An ecological view of social media use among college students. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 571–582, 2015.

S. Willis, L. McKenna, S. McDermott, G. O'Donell, A. Barrett, B. Rasmusson, A. Nordgaard, C. Berger, M. Sjerps, J. L. Molina, G. Zadora, C. Aitken, L. Lunt, C. Champod, A. Biedermann, T. Hicks, and F. Taroni. ENFSI guideline for evaluative reporting in forensic science. *European Network of Forensic Science Institutes*, 2016.

G. Zadora, A. Martyna, D. Ramos, and C. Aitken. *Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data*. John Wiley & Sons, 2013.

B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699, 2002.

Appendix A

Spatial Results—Twitter Data

A.1 Orange County Region

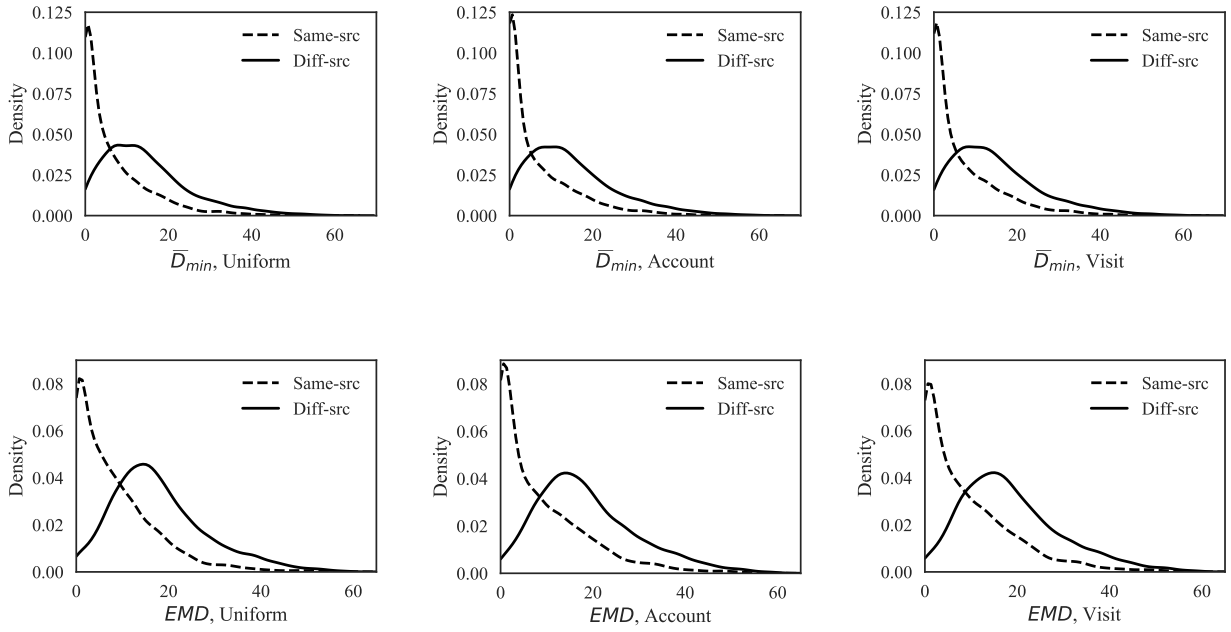


Figure A.1: Empirical distributions of the score functions from Section 4.5 for Twitter data in Orange County, CA. Same source distributions (H_s, dashed line) and different source distributions (H_d, solid line) approximated via kernel density estimation with Gaussian kernels and Scott's rule of thumb bandwidth. Leave-pairs-out cross-validation was not used to produce these densities; instead all pairs were used for illustrative purposes.

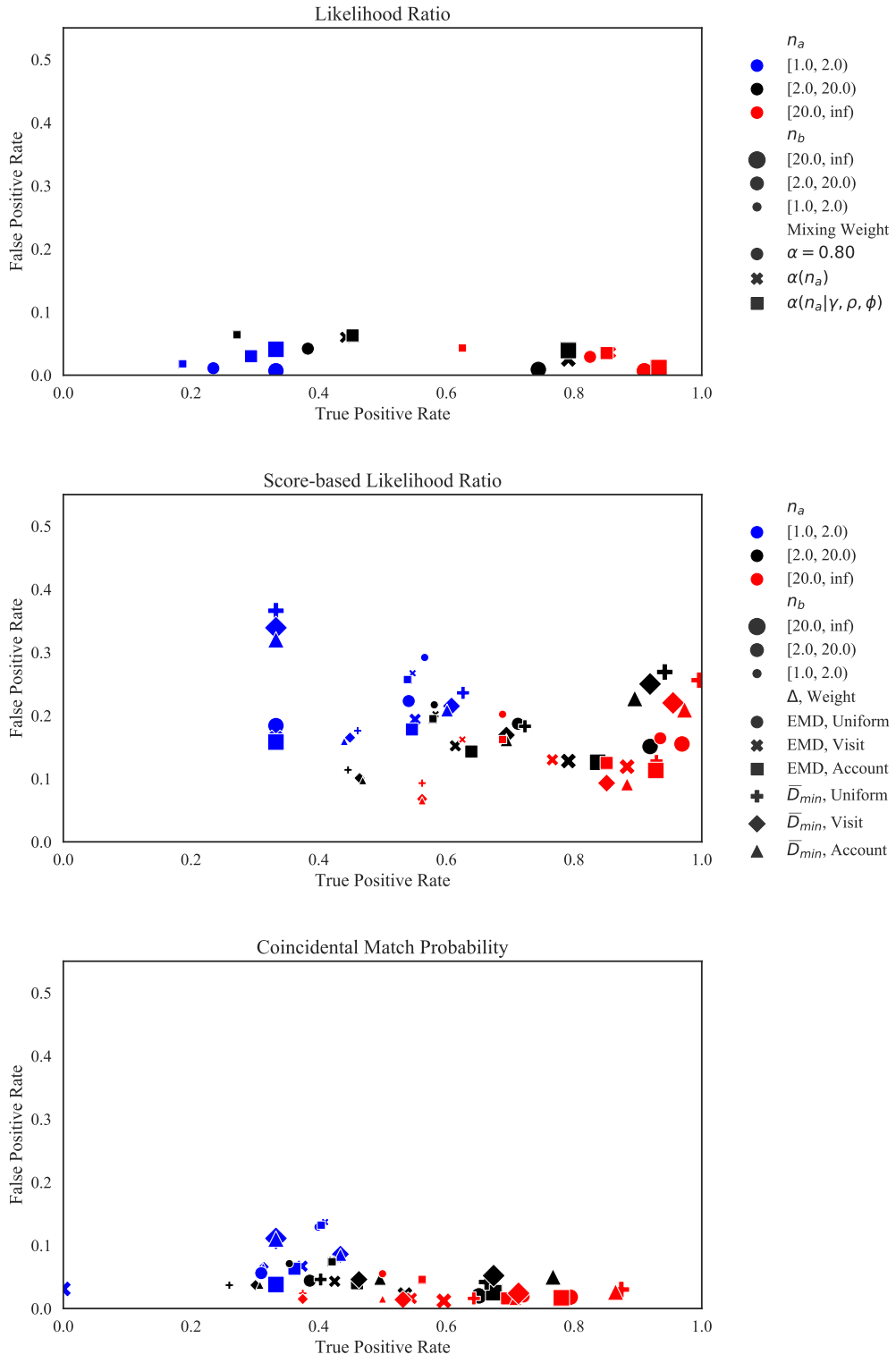
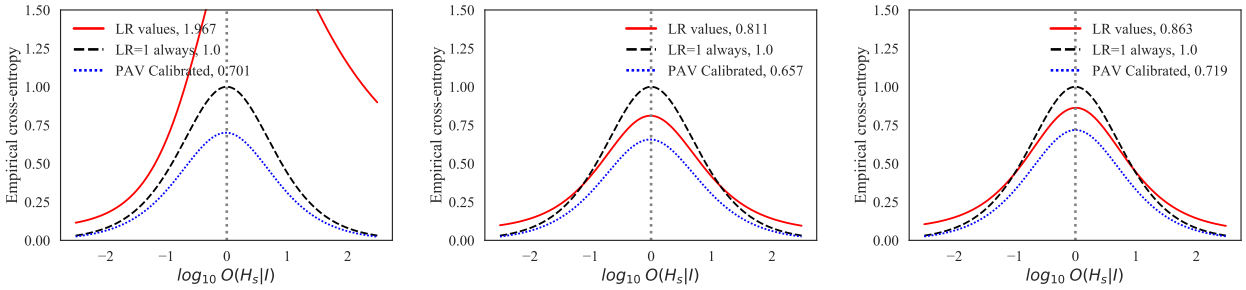
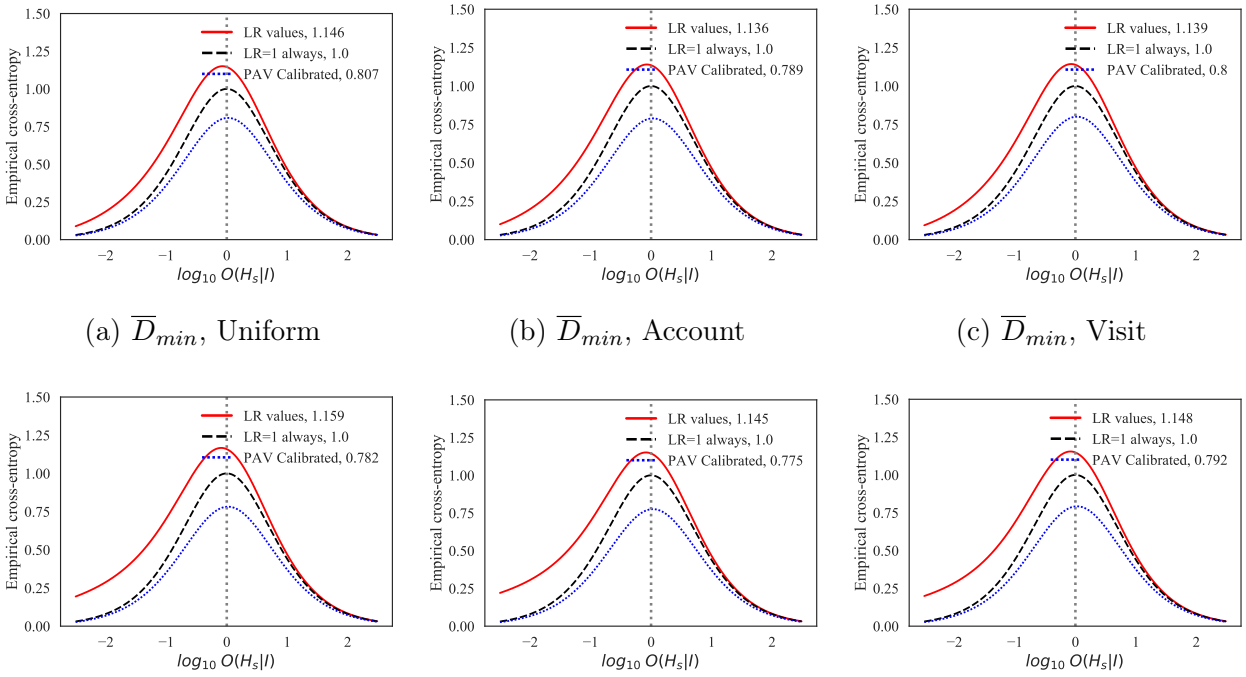


Figure A.2: False positive rate versus true positive rate for each evidence evaluation technique and corresponding weighting scheme or score function for Twitter data in Orange County, CA. The LR and SLR use a threshold of 1, while the CMP uses a threshold of 0.05. Point size and color correspond to the strata used for sampling as discussed in Section 4.7.3. The barplot in Figure 4.8a is a subset of the y-axis of this plot.



(a) $\alpha = 0.80$ (b) $\alpha(n_a)$ (c) $\alpha(n_a|\gamma, \rho, \phi)$

Figure A.3: Empirical cross-entropy (ECE) of the likelihood ratio (LR) approaches on the Twitter data in Orange County, CA, using different mixing weights α.



(a) $\overline{D}_{min}$, Uniform (b) $\overline{D}_{min}$, Account (c) $\overline{D}_{min}$, Visit

(d) EMD, Uniform (e) EMD, Account (f) EMD, Visit

Figure A.4: Empirical cross-entropy (ECE) of the score-based likelihood ratio (SLR) approaches on the Twitter data in Orange County, CA, using different score functions and weighting schemes.

A.2 New York Region

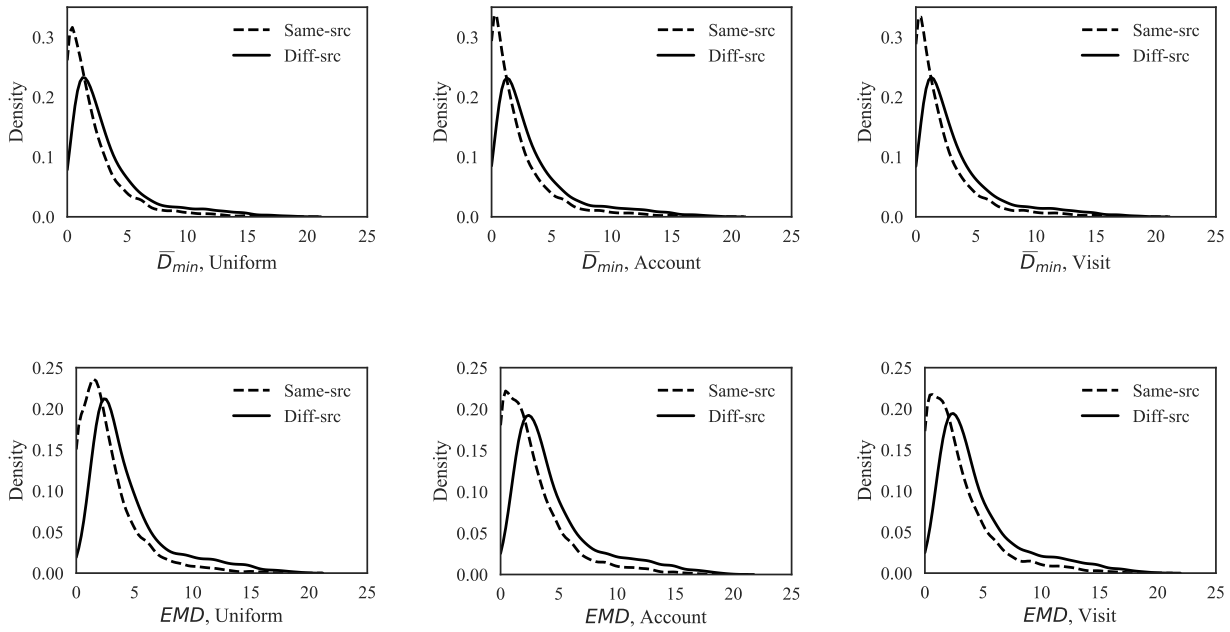


Figure A.5: Empirical distributions of the score functions from Section 4.5 for Twitter data in New York, NY. Same source distributions (H_s, dashed line) and different source distributions (H_d, solid line) approximated via kernel density estimation with Gaussian kernels and Scott's rule of thumb bandwidth. Leave-pairs-out cross-validation was not used to produce these densities; instead all pairs were used for illustrative purposes.

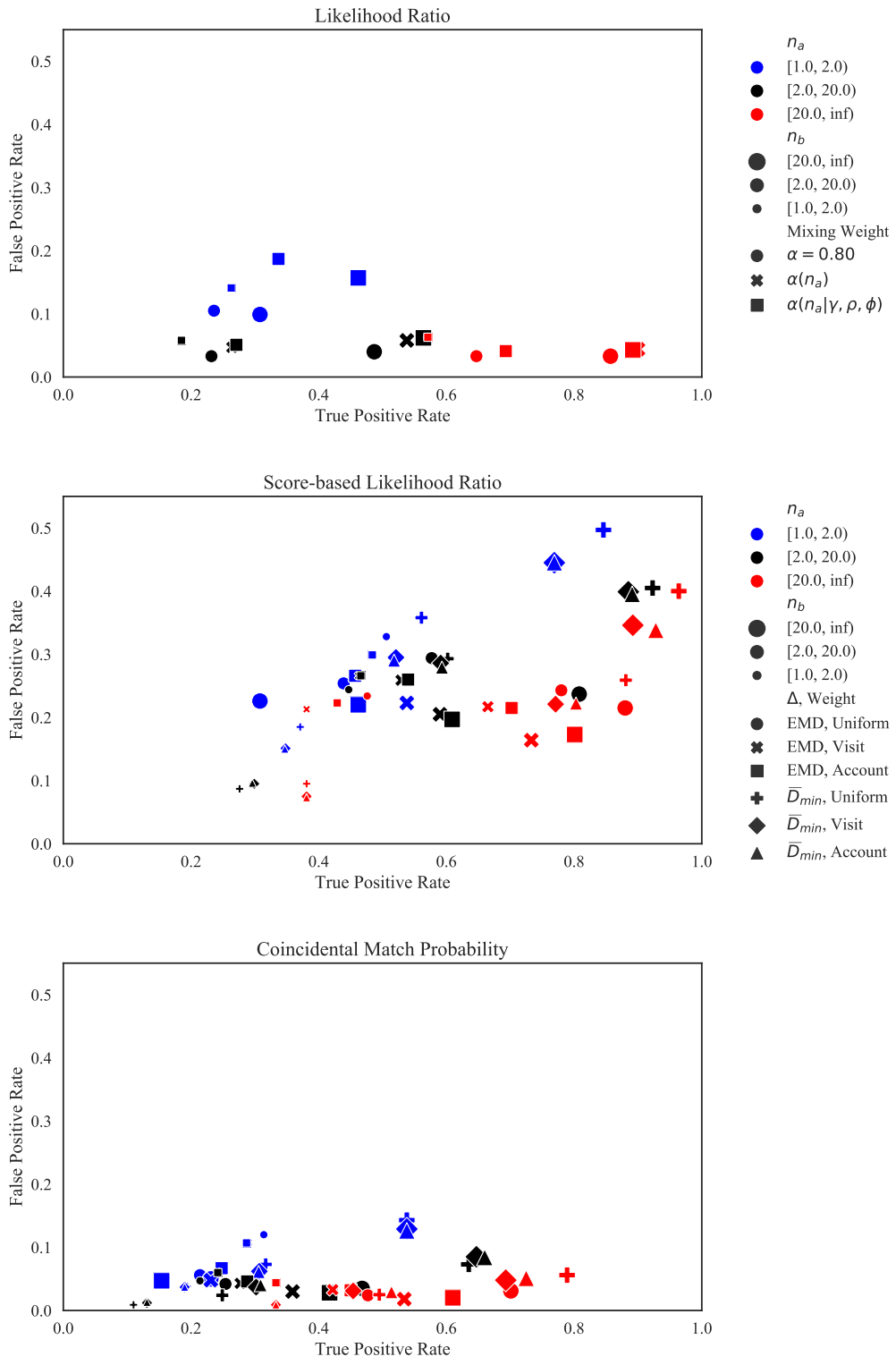
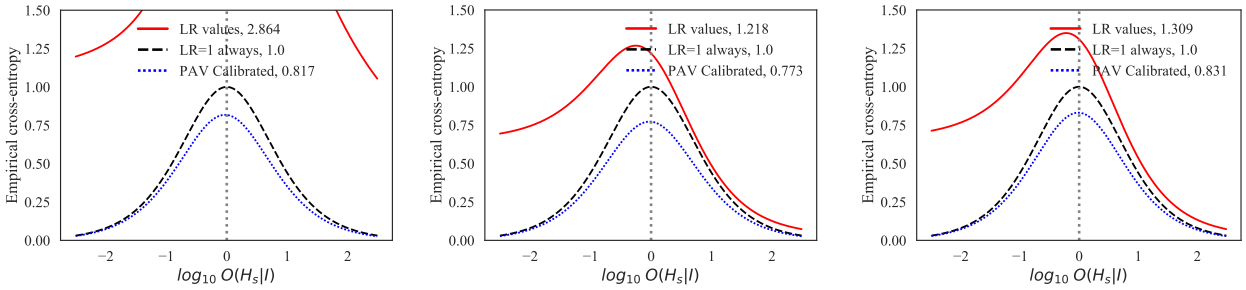
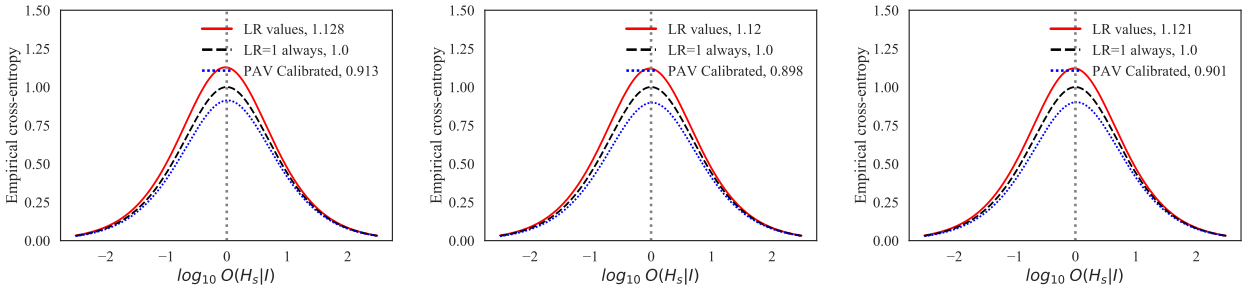


Figure A.6: False positive rate versus true positive rate for each evidence evaluation technique and corresponding weighting scheme or score function for Twitter data in Orange County, CA. The LR and SLR use a threshold of 1, while the CMP uses a threshold of 0.05. Point size and color correspond to the strata used for sampling as discussed in Section 4.7.3. The barplot in Figure 4.8b is a subset of the y-axis of this plot.

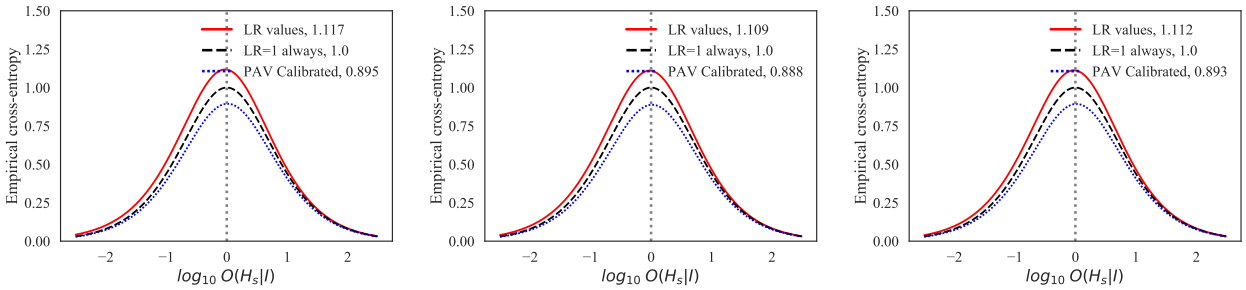


(a) $\alpha = 0.80$ (b) $\alpha(n_a)$ (c) $\alpha(n_a|\gamma, \rho, \phi)$

Figure A.7: Empirical cross-entropy (ECE) of the likelihood ratio (LR) approaches on the Twitter data in New York, NY, using different mixing weights α.



(a) $\overline{D}_{min}$, Uniform (b) $\overline{D}_{min}$, Account (c) $\overline{D}_{min}$, Visit



(d) EMD, Uniform (e) EMD, Account (f) EMD, Visit

Figure A.8: Empirical cross-entropy (ECE) of the score-based likelihood ratio (SLR) approaches on the Twitter data in New York, NY, using different score functions and weighting schemes.

Appendix B

Spatial Results—Gowalla Data

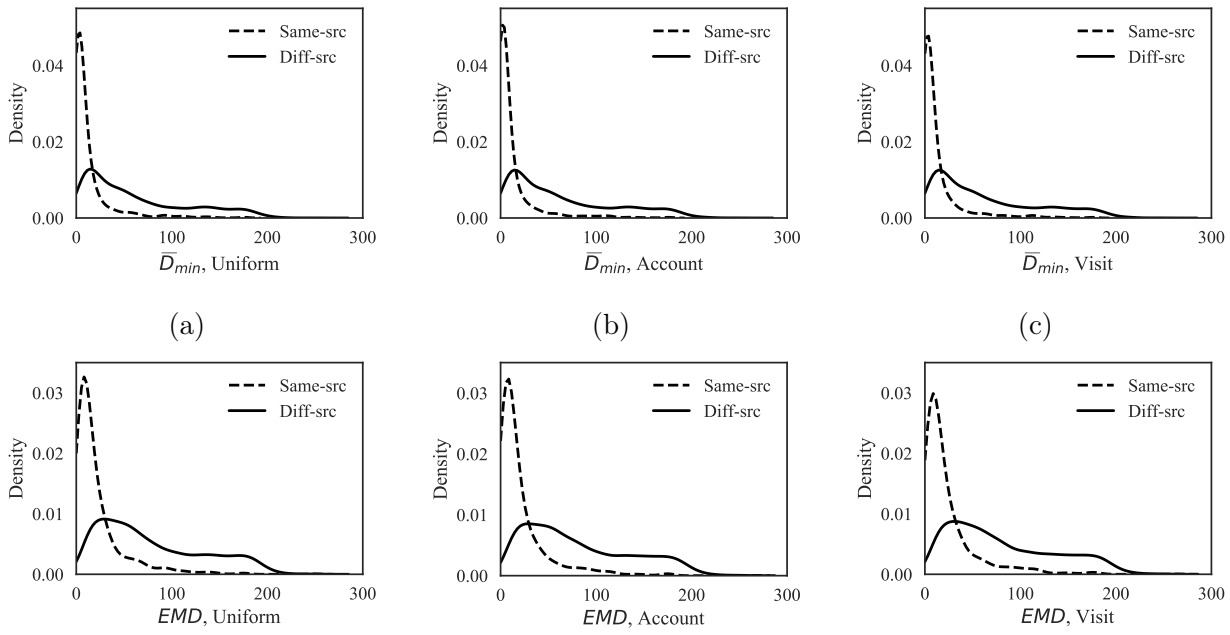


Figure B.1: Empirical distributions of the score functions from Section 4.5 for Gowalla data in Southern California. Same source distributions (H_s, dashed line) and different source distributions (H_d, solid line) approximated via kernel density estimation with Gaussian kernels and Scott's rule of thumb bandwidth. Leave-pairs-out cross-validation was not used to produce these densities; instead all pairs were used for illustrative purposes.

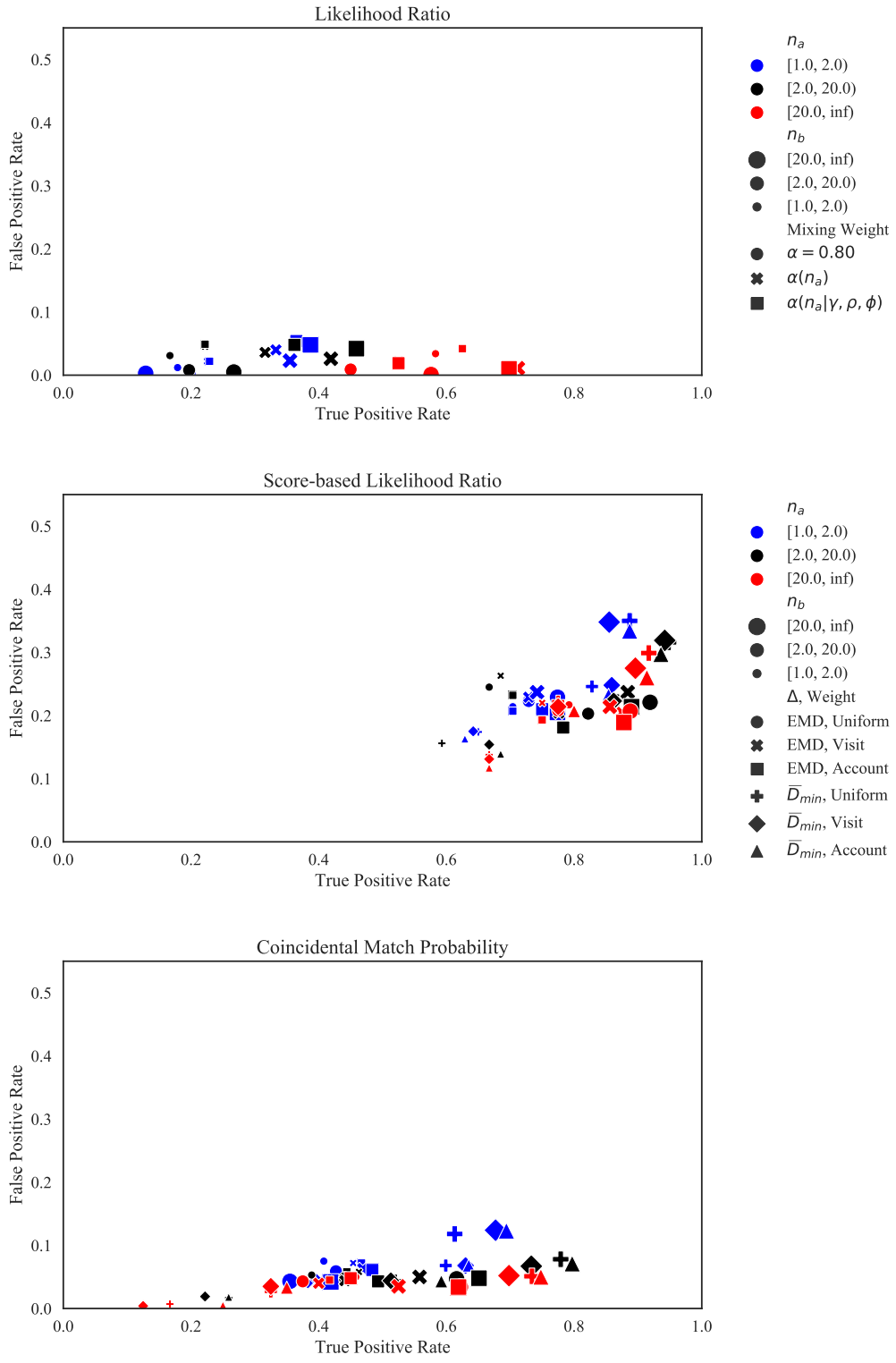
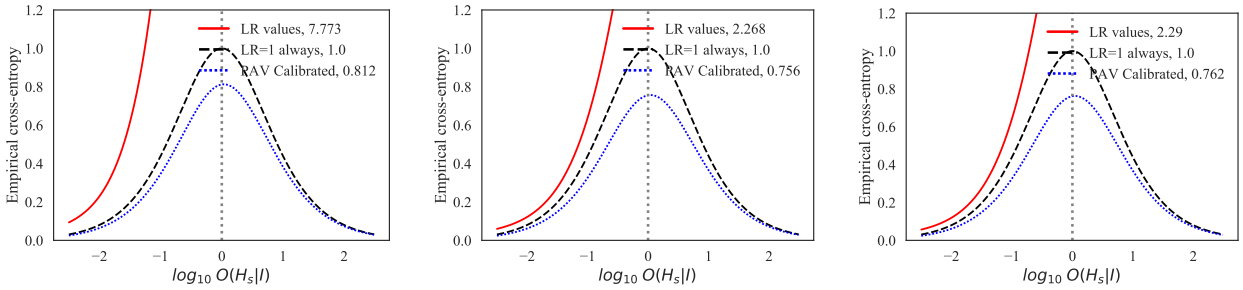
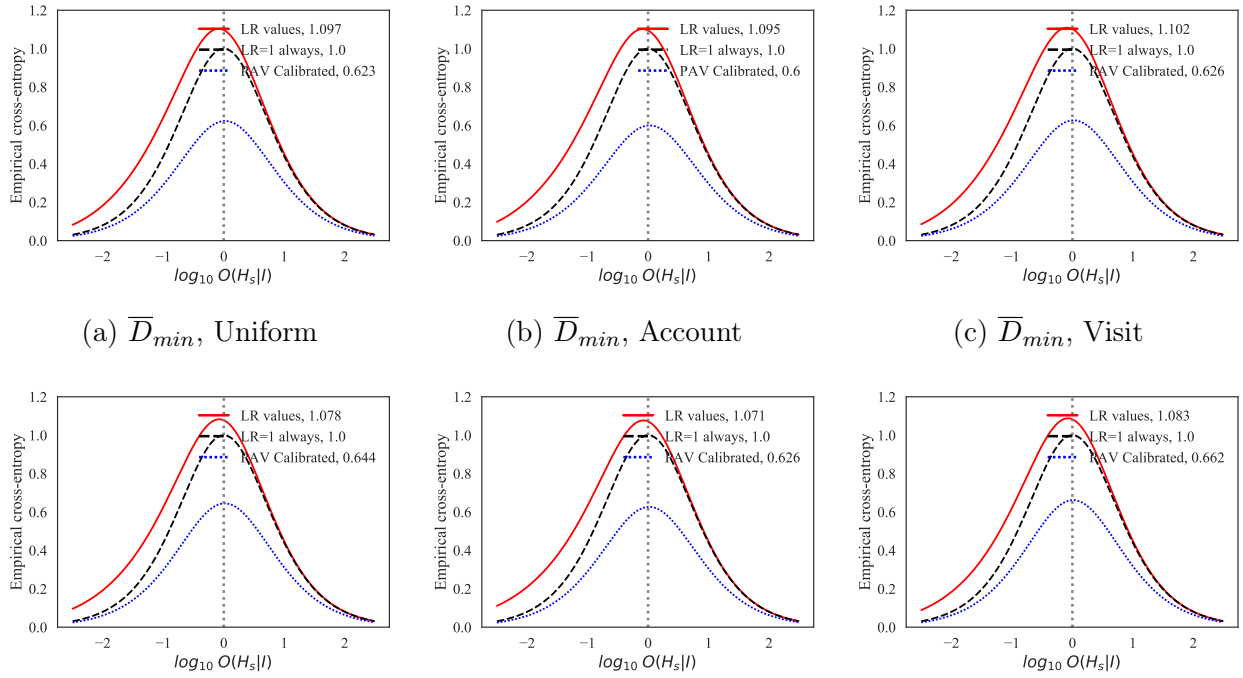


Figure B.2: False positive rate versus true positive rate for each evidence evaluation technique and corresponding weighting scheme or score function for Gowalla data. The LR and SLR use a threshold of 1, while the CMP uses a threshold of 0.05. Point size and color correspond to the strata used for sampling as discussed in Section 4.8.2. The barplot in Figure 4.14 is a subset of the y-axis of this plot.



(a) $\alpha = 0.80$ (b) $\alpha(n_a)$ (c) $\alpha(n_a|\gamma, \rho, \phi)$

Figure B.3: Empirical cross-entropy (ECE) of the likelihood ratio (LR) approaches on the Gowalla data in Southern California, using different mixing weights α.



(a) $\overline{D}_{min}$, Uniform (b) $\overline{D}_{min}$, Account (c) $\overline{D}_{min}$, Visit

(d) EMD, Uniform (e) EMD, Account (f) EMD, Visit

Figure B.4: Empirical cross-entropy (ECE) of the score-based likelihood ratio (SLR) approaches on the Gowalla data in Southern California, using different score functions and weighting schemes.

Appendix C

Signal-to-Noise Ratio Calculation

Recall that the numerator of the signal-to-noise ratio (SNR) in Equation 5.12 is the reciprocal of the mean intensity of the simulated realizations of process A, i.e.,

$$\overline{\lambda}_A^{-1} = \left[n^{-1} \sum_{i=1}^{n} \lambda_A^{(i)} \right]^{-1} . \tag{C.1}$$

where n is the number of simulated processes and $\lambda_A^{(i)}$ is the intensity of the i^{th} realization of process A. Since each realization of A is a Poisson process, the inter-event times $\tau_{AA}^{(i,j)}$ for $j = 1, \ldots, n_A^{(i)}$ are distributed *i.i.d.* Exponential($\lambda_A^{(i)}$), and their expectation is

$$\mathbb{E}_\tau \left(\tau_{AA}^{(i,j)} \right) = \left(\lambda_A^{(i)} \right)^{-1} \quad \forall j. \tag{C.2}$$

Note that each realization of A is independent of the other $n-1$ realizations. Thus the expected inter-event time across the realizations of A is

$$\mathbb{E}_\tau\left(\overline{\tau}_{AA}^{(\cdot,\cdot)}\right) = \mathbb{E}_\tau\left(n^{-1}\sum_{i=1}^{n}\overline{\tau}_{AA}^{(i,\cdot)}\right) \tag{C.3}$$

$$= n^{-1}\sum_{i=1}^{n}\mathbb{E}_\tau\left(\tau_{AA}^{(i,j)}\right) \tag{C.4}$$

$$= n^{-1}\sum_{i=1}^{n}\left(\lambda_A^{(i)}\right)^{-1} \tag{C.5}$$

$$\to \mathbb{E}_\lambda\left(\frac{1}{\lambda_A}\right) \quad \text{as } n\to\infty. \tag{C.6}$$

Since λ_A^{-1} is a convex function, we can apply Jensen's inequality to (C.6) to obtain

$$\frac{1}{\overline{\lambda}_A} \to \frac{1}{\mathbb{E}_\lambda(\lambda_A)} \leq \mathbb{E}_\lambda\left(\frac{1}{\lambda_A}\right). \tag{C.7}$$

Therefore, $\overline{\lambda}_A^{-1}$ is a lower bound on the expected inter-event time across the simulated realizations of process A, denoted $\mathbb{E}_\lambda(\lambda_A^{-1})$. As stated in Section 5.6.2, as the expected inter-event time $\mathbb{E}_\lambda(\lambda_A^{-1})$ decreases the noise (or the density of events in realizations of A) increases. The inequality (C.7) shows that it is more conservative to use $\overline{\lambda}_A^{-1}$ than $\mathbb{E}_\lambda(\lambda_A^{-1})$ for estimating the SNR since it results in an over-estimate of the amount of noise present in the processes (and thus a lower SNR).