Errors-In-Variables and Random Forests: Theory and Application to Eyewitness Identification Data

Alice Jia Liu
Charlottesville, VA

M.S., University of Virginia, Charlottesville, Virginia, 2017
B.A., University of Virginia, Charlottesville, Virginia, 2015

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Statistics

University of Virginia
May 2020

Abstract

Eyewitness identifications play a critical role in the investigation of crimes and the subsequent legal proceedings. However, law enforcement do not have the time and resources available to conduct the much-needed research for the development and validation of more reliable practices. Research in the effectiveness of law enforcement practices for eyewitness identification procedures remains incomplete. It is well known that eyewitnesses make errors, which often result in grievous consequences. Currently, there are a few options for eyewitness identification analysis, including receiver operating characteristic (ROC) curve analysis, Bayesian prior-posterior plots, and decision utility. All of these methods lack a fundamental way to include variability and the complex and interactive relationships of the variables affecting eyewitness identification accuracy.

We will also discuss new methods for eyewitness identification (EWID) data, which are borrowed from fields such as diagnostic medicine. The tools and procedures for analyzing the data in meaningful and utilitarian ways from these fields can provide thoughtful and valid conclusions. Such methodologies require ease of use and interpretation, flexibility, and efficient implementation. This compilation of chapters shows the thought process involved in considering what kinds of methods and approaches to thinking could help lead to better EWID procedures, with the intention of resulting in fewer errors, both in false convictions and false acquittals.

This research began with an interdisciplinary problem of understanding EWID data and existing statistical methodologies for the analysis of such data, as well as the consequences of an incomplete comprehension of the data. It was clear that there are latent variables to be estimated that are imperative to understand parts of the data, which resulted in the development of the proposed framework. This

framework allows researchers to estimate an individual's probability of accuracy, which is dependent on their individual probability of choosing a face from a lineup and the global probability of target presence in the lineup (i.e., base rate). The true value in the proposed method is how easily it is applied and interpreted, which could be helpful for law enforcement agents, lawyers, and jurors.

A component of the estimation relies on the algorithm of random forests. Since EWID data is susceptible to measurement error due to the human component, we discovered that the impact of measurement error on random forest models needs further study. This thesis addresses that problem. The literature provides a framework for the asymptotic behavior of random forests. This provides the groundwork to derive an estimator for the mean difference of two random forest models. In our case, the random forest models are developed with and without measurement error to simulate the behaviors of the differences. In the simulations, it was clear that there is an effect from measurement error. Since measurement error is usually assumed to be nonexistent or negligible, this is a valuable finding. The next steps should be to develop a methodology similar to those already in place for classical statistical models to account for these errors.

Keywords: statistics, eyewitness identification, random forests, confidence and accuracy, forensic evidence, ROC analysis, sensitivity, specificity, predictive value, choosing, classification, class probability estimation, measurement error, errors-in-variables

Acknowledgements

I would like to thank all of the people who have helped and inspired me throughout my doctoral study.

I would especially like to thank my adviser, Dr. Karen Kafadar, for all of her support, advice, and wisdom during my research and study at the University of Virginia. I remember the first day of the 2014 spring semester, sitting in her exploratory data analysis (EDA) class, and listening to her talk about John Tukey and emphasize the importance of EDA (she was absolutely right about that – it is one of my most used statistical tools). She has encouraged me throughout the proposal process, and provided guidance whenever I needed it. Her phenomenal advice and words of wisdom have provided waypoints on the winding path to the finish line. She has always been accessible and willing to help and meet with her students with their research, and I greatly appreciate that. Dr. Kafadar connected me with one of my first work experiences at the Defense Forensic Science Center (DFSC) through the Center for Statistics and Applications in Forensic Evidence (CSAFE), which broadened my perspective on the practical aspects of statistics in the industry.

I would also like to thank my other advising committee members – Dr. Chad Dodson, Dr. Daniel Keenan, and Dr. Xiwei Tang – for their continued support in my academic endeavors. Dr. Dodson was integral and absolutely invaluable to the process, as he provided the expertise in a completely different field, and was always willing to answer my questions and connect me with other researchers in his field. Dr. Keenan has been part of my journey since I was an undergraduate, and afforded me the opportunity for research at that level. He was also indispensable at the higher levels of doctoral classes (probability theory, measure theory, etc.). Dr.

Tang was always available to coordinate speakers for our graduate student meetings and colloquia, and is always enthusiastic about the work.

In addition, thank you to the faculty in the department who have taught and guided me throughout my undergraduate and graduate career, with special thanks to Dr. Tingting Zhang, Dr. Jianhui Zhou, Dr. Chao Du, and Dr. Jeff Holt.

Also, to my peers (Karen Pan, Justin Weinstock, Ye Lin, Yinge Sun, Tonghao Zhang, Brian Whitlow, Megan Yetman, David Bennett) and office mates (Evan Bagley, Taylor Brown, Maria Ferrara, Caitlin Steiner, Krista Varanyak) – thank you for being there as sounding boards when I was stuck on a homework question or theoretical concept. Your hard work and passion for the discipline is inspiring. Your support during mathematical statistics, measure theory, and through all of the steps of the dissertation process has been indispensable.

Thank you also to my family and friends who provided fundamental emotional support and words of encouragement. My father, Dr. Zhijun "George" Liu, was one of the primary reasons I sought a degree in statistics, which resulted in multiple degrees. He was always there, pushing me to do better, to better understand the concepts, to practice more, and to succeed. I could not have gotten this far without him. My mother, Rebecca Wu, was always there to encourage me and provide me with household and food necessities. My sister, Catherine Liu, was always free for conversation and to make me laugh.

My deepest gratitude goes to my boyfriend, Matthew West, who has put up with me and years of statistics jargon and has been my rock throughout my undergraduate and graduate studies (since 2012!). I am amazed that he is always there to give me feedback, even if he does not quite understand the concepts; to listen to me practice my poster talks, presentations, and practice lectures.

This research was supported by a grant from the Arnold Ventures. Their gen-

erous support is what has enabled this work. This report is an independent work product. The views expressed are those of the author and do not necessarily represent those of the funder.

Thank you to all of the people and organizations who have supported, molded, and aided me throughout this process!

Contents

Part I

Estimating Eyewitness Accuracy

Chapter 1

Introduction

1.1 Motivation

> "One of the main causes of wrongful convictions is eyewitness misidentifications. Despite a high rate of error (as many as 1 in 4 stranger eyewitness identifications are wrong), eyewitness identifications are considered some of the most powerful evidence against a suspect."
>
> *California Innocence Project*

In July 1984, Jennifer Thompson was sexually assaulted by an assailant, who, later that night, sexually assaulted a second woman. Thompson helped create the composite sketch that led to the assembly of a live lineup in which she positively identified Ronald Cotton as the perpetrator. "Yeah. This is the one... I think this is the guy," said Thompson at the live lineup (Garrett, 2012). A second lineup was assembled, with Cotton as the only repeated person. "This looks the most like him," Thompson confirms, stating that she was "absolutely sure" Cotton was the culprit. Cotton was convicted of sexual assault and burglary based on circumstantial evidence and Thompson's identification. He was sentenced to life in prison plus 54

years. In 1995, after 10 years in prison, Cotton was exonerated through DNA testing with help from the Innocence Project.[1] This is a particularly well-known example of a common problem. Approximately 71% of more than 360 post-conviction DNA exonerations documented by the Innocence Project since 1989 involved one or more mistaken eyewitness identifications.[2]

EWID plays a critical role in criminal cases, from the investigation to the prosecution of the crime. The core element of EWID is memory – remembering the suspect, the proceedings of the crime, and the emotions associated. Howe and Knott (2015) note that memory is first encoded, then consolidated with existing information in the brain, and then retrieved (i.e., reconstructed) at a later time. Each stage can cause memories to degrade or mutate over time, depending on the purpose for retrieving the information, to whom, and how it is recalled. In addition to internal factors, such as the person's own memory processes, external factors can distort one's information retrieval, such as length of time between the event and need for retrieval of the memory, intermediate events during that time, and identification procedures. We need experiments that faithfully represent EWID processes to assess which factors can be varied and set at levels that minimize the probabilities of grievous EWID errors.

Statistical methods, used to analyze datasets concerning eyewitness choices in experiments or in the field, give one a better understanding of what factors affect the likelihood that an eyewitness will choose correctly. Statisticians are working in conjunction with psychologists to conduct tests with high ecological validity to identify factors that improve the reliability of EWID evidence. We define "ecological validity" to mean that the study (including methods, materials, setting, etc.)

[1] `https://www.innocenceproject.org/cases/ronald-cotton/` accessed March 11, 2020.
[2] `https://www.innocenceproject.org/eyewitness-identification-reform/`, accessed March 11, 2020.

approximates the real-world, the generalize the study findings to real-world settings. Statistically designed experiments help identify factors that are more likely to lead to errors as well as those that are less likely to result in mistakes, by encouraging efficient experimental practices, integration of variability measures, and application of existing statistical models from other fields to EWID data.

The National Academy of Sciences (NAS) emphasized both needs in its important report on the subject issued in 2014: "The committee recommends a broad exploration of the merits of different statistical tools for use in the evaluation of eyewitness performance" (see National Research Council, 2014, pg. 108). With the creation of such experiments, relevant and appropriate methods of analysis need to identified and developed, which we are working towards.

1.2 Organization of the Dissertation

The motivation for this work stems from the interdisciplinary nature of statistics, since it lends itself well to supplement and enrich other fields. We motivate the development of a new framework for the analysis of EWID data, which in turn motivates a more general understanding of the asymptotic behavior of random forests.

1.2.1 The Main Contributions

(1) We provide an overview of the field of eyewitness identification and existing statistical methodologies.

(2) What statistical framework can be developed to provide an objective substantiation to the reliability of eyewitnesses? How flexible is this method in comparison to previously used methods in the field of psychology?

(3) Does the proposed framework satisfy the interpretational requirements of the field of psychology?

(4) Random forests are used as component of the proposed framework for the estimation process. How robust are the estimates from a random forest model? If the covariates are measured with error, how different can we expect the behavior to be? What is the asymptotic behavior of the distribution of the difference of two models (one built without any measurement error and another built with measurement error)?

This dissertation endeavors to address the above questions. In Chapter 2, we provide an overview of the field of eyewitness identification, as well as the details of data sets to be used throughout this thesis. Chapter 3 provides the current methods of statistical analysis for eyewitness identification data. In Chapter 4, we propose a new statistical framework for the estimation of the probability of accuracy of an eyewitness in a lineup procedure with at most one guilty target. In Chapter 5, we study the asymptotic behavior of random forest model. In Chapter 6, we derive the asymptotic behavior of the distribution of the difference of two models, based on covariates measured with and without error.

Chapter 2

The Field of Eyewitness Identification

2.1 Introduction

The major statistical contributions rely on the description of the motivational problem, which stems from the field of eyewitness identification. This requires an understanding of the eyewitness task and existing statistical methods used for the analysis of such data. This chapter serves to provide information on existing and viable statistical methods for analyzing EWID experiments. Whatever technique is used, proper characterization of the uncertainties associated with inferences must be calculated.

Background information of the eyewitness task and EWID data is provided in Section 2.2. A brief history of the development of analysis methodologies in the field of EWID and some issues present in the methodology currently are discussed in Section 2.3. In Section 2.4, current statistical methods in EWID research are reviewed. In Section 2.5, we present an example of analysis of variance to compare

EWID procedures.

Note: sections of this chapter appear as part of the *Handbook of Forensic Statistics* (see Liu et al., 2020, Chp. 21).

2.2 The Eyewitness Task

The task of the eyewitness is to attempt to identify the perpetrator of a crime that he or she witnessed. With a single suspect, the identification decision is binary: either the presented suspect is or is not the person whom he or she saw commit the crime. The binary choice results in a binary outcome: either the suspect was or was not the true perpetrator, and either the eyewitness does or does not implicate that suspect.

In the standard paradigm of EWID, the two correct outcomes are the conviction of the truly guilty (true positive) and the exoneration of the truly innocent (true negative). The two incorrect outcomes are the conviction of the truly innocent (false positive) and the exoneration of the truly guilty (false negative). Table 2.1 shows these outcomes from the eyewitness, who serves as the "binary classifier" for this task.

		Witness's Decision	
		"Guilty"	"Innocent"
Suspect's True Status	Guilty	True positive (TP)	False negative (FN)
	Innocent	False positive (FP)	True negative (TN)

Table 2.1: The eyewitness task shown visually as a two-by-two table, assuming the eyewitness serves as the "binary classifier."

The NAS report called attention to the consideration of the eyewitness as a

binary classifier for analysis purposes: "It is important that practitioners in this field broadly explore the large and rich field of statistical tools for evaluation of binary classifiers" (see National Research Council, 2014, pg. 91). For many people, minimizing false positives is the key priority, as the consequences to the wrongfully convicted are profound. Law enforcement personnel seek to minimize false negatives, to prevent perpetrators from committing further crimes.

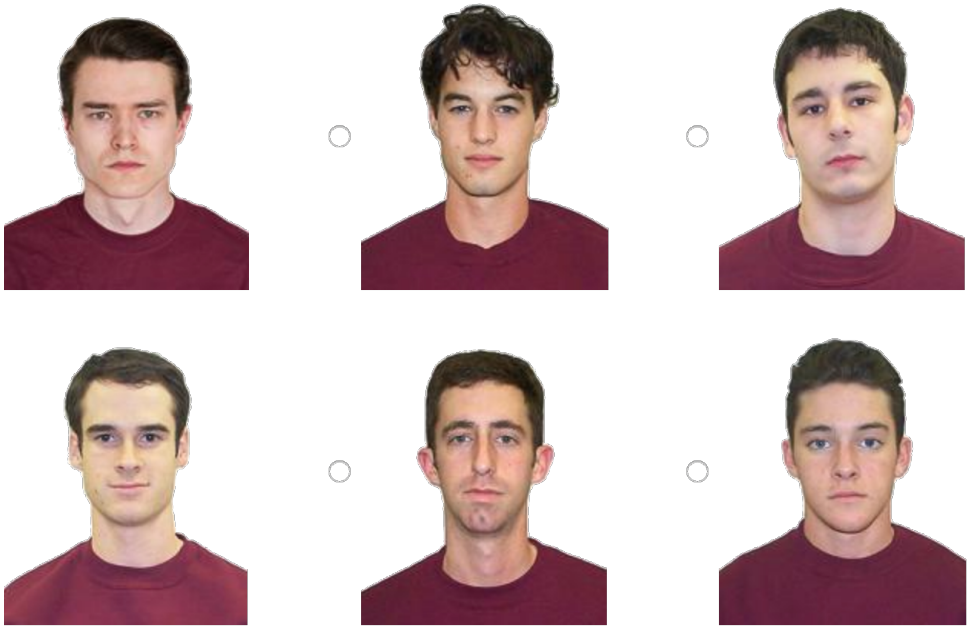


Figure 2.1: Example of a fair, target present simultaneous lineup in an experimental setting, target suspect (shown as the perpetrator in a video of the "crime") is in the top-left. This lineup was provided by Chad Dodson from the University of Virginia.

The perpetrator may not be in the lineup at all. Thus, the target is present (TIP) or target is absent (TIA) in the lineup. Law enforcement believes that the target is present in the lineup, but no data exist on the proportion of lineups that are TIP. Likely it varies substantially by jurisdiction or even from agency to agency.

Figure 2.1 shows an example of a simultaneous lineup with photos of six possible suspects that might be shown to an eyewitness; "Not Present" is also offered as an option. For more examples of simultaneous lineups used in such laboratory experiments, see Wells et al. (2011).

If the lineup is TIP, the eyewitness can make three possible decisions:

(P_1) Make the right decision and choose the guilty suspect;

(P_2) Make a wrong decision and choose an innocent foil[1];

(P_3) Make a wrong decision and state that the guilty suspect (i.e., target) is not present.

If the lineup is TIA, the eyewitness can make two possible decisions:

(A_1) Make the right decision and state that the guilty suspect is not present;

(A_2) Make a wrong decision and choose the innocent suspect or a foil.

Thus, five possible decision outcomes can occur, only two of which (P_1 and A_1) are correct; see Figure 2.2.

Researchers often include a designated "innocent suspect" to serve as the "target" in TIA lineups. Based on this set-up, the four categories of classification are:

(1) Correct suspect identification;

(2) Innocent suspect identification;

(3) Foil identification; and

(4) Lineup rejection (suspect not present).

[1]A "foil" is an innocent person in a police lineup. It is also sometimes referred to as "filler" in the literature.

Accurate	**Not Accurate**		
Choose Target (P_1)	Choose Foil (P_2)	Do Not Choose (P_3)	**Target Present**
Do Not Choose (A_1)	Choose Foil (A_2)		**Target Absent**

Figure 2.2: A display of the eyewitness decision outcome space, which takes into account the underlying status of the lineup.

Table 2.2 shows three different approaches to EWID data structure. The choice of structure will influence the analysis method. However, in general, the status of "innocent" suspect is unknown in a real lineup, making this structure highly unlikely. The concept of the five possible decision outcomes for the eyewitness task has been well-established in the field of psychology (Wells and Olson, 2002). These outcomes have been treated in a deterministic manner, with little regard for a generalizable statistical model that could move beyond a single input data set.

In the field of psychology, the traditional analysis dichotomizes the decision for a single suspect. Given an identification, memory theory from the paradigm of signal detection theory (SDT) indicates that the eyewitness applies "a simple rule to make an identification decision" (Clark et al., 2015). If the association between the suspect and the eyewitness's reconstruction (via memory) of the perpetrator exceeds a "threshold" of memory strength c, then the witness will identify that suspect as the perpetrator. If it falls below that individual's threshold, then the eyewitness will exclude the suspect as a perpetrator. This paradigm assumes that the decision is based on the individual's threshold for a single variable, "memory strength": a false identification occurs if the suspect is innocent but the individual's "memory

		Suspect's True Status Scenario	*"Guilty"*	*"Innocent"*
Eyewitness's Decision	Errors treated equally	*"Guilty" Suspect*	TP	FP
		Not the "Guilty" Suspect	FN	TN
	No designated innocent suspect	*"Guilty" Suspect*	TP	Forced 0
		Foil (Known Innocent)	Incorrect	FP
		Not Present	FN	TN
	Designated innocent suspect	*"Guilty" Suspect*	TP	Forced 0
		"Innocent" Suspect	Forced 0	FP
		Foil (Known Innocent)	Incorrect	Incorrect
		Not Present	FN	TN

Table 2.2: This table provides the three possible structures assumed for EWID data, from the two previously addressed structures in Table 2.1 and Figure 2.2 to the inclusion of an innocent suspect. The table provides the possible EWID outcomes based on the eyewitness's decision versus the true underlying status of the lineup, which could affect the analysis approaches used by researchers.

strength" falls above c, and false exclusions occur if the "memory strength" falls below c for an innocent suspect.

According to Gronlund and Benjamin (2018), SDT provides a cohesion for decision-making with ambiguous evidence, with a link to metacognition (i.e., awareness, understanding, analysis, and control of one's own cognitive (learning, thinking, reasoning, etc. processes). Figure 2.3 displays this memory theory paradigm: the eyewitness's decision comes from one of these two distributions, often conveniently assumed to be normal, and the memory "threshold" is flexible that can vary depending on factors, such as the cost of making a mistake. For example, in Figure 2.3, the memory threshold is set at the mean (median) of the "Guilty" distribution. The larger the separation between these two distributions, and the higher the quantile

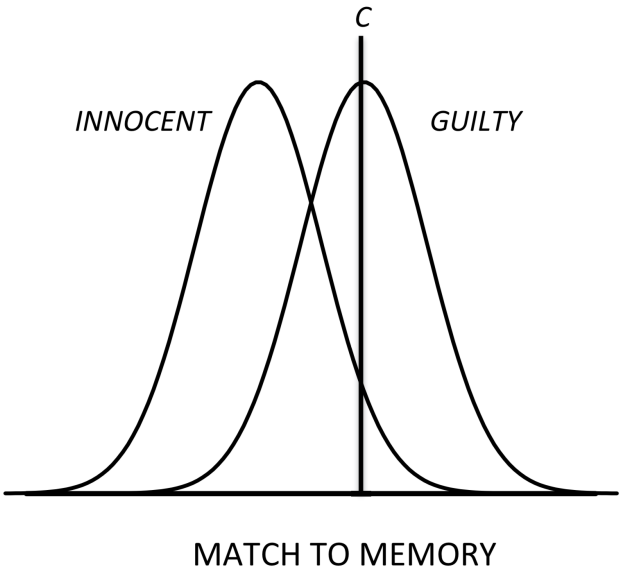


Figure 2.3: Distribution of "memory strength" for identification of guilty and innocent suspects (Clark et al., 2015). This illustration conforms with memory theory in assuming a normal distribution for "memory strength" and the individual's memory threshold c as the mean of the right-hand curve. Other models for the distribution of "memory strength" have be proposed, and individual thresholds may fall at different quantiles of the distribution other than 50%.

of the distribution of "Guilty" for the individual's threshold c, the lower the error rates (false negatives, false positives).

2.2.1 System and Estimator Variables

The accuracy of this eyewitness task depends on many factors. Some factors are under the control of law enforcement (e.g., type of lineup), while others arise by the circumstances (e.g., lighting). A summary of these factors is shown in Figure 2.4. Factors that can affect accuracy of eyewitness identification and are under the control of law enforcement have been called "system variables" in the eyewitness identification literature ("control" variables in experimental design literature); they include:

System Variables (Controllable by Law Enforcement)

- *Type of lineup* (or photo array, if it is a photos are used): typically "sequential" (suspects or photos show sequentially, one at at time), or "simultaneous" (shown together);

- *Size of lineup* (i.e., number of suspects shown): ideally chosen so that the probability of identifying an innocent suspect by chance is low (Brigham et al., 1999);

- *Fairness* (i.e., subjective similarity of appearance of people) of lineup: in a truly fair lineup, the probability that any one of the suspects is selected is equal; increasingly biased lineups are those for which the probabilities are not equal;

- *Delay* (i.e., retention interval): time between incident and eyewitness's identification task;

- *Lineup instructions*: degree of detail in guidance to the eyewitness in the identification procedure (i.e., instructing the eyewitness that the culprit may or may not be in the lineup); more details provided in Wixted and Wells (2017);

- *Blinding*: law enforcement officer (LEO) conducting the lineup either is, or is not, within view of the eyewitness and the photos (s)he is viewing, "unblinded" or ""blinded" lineup, respectively. (The concern is that the "unblinded" LEO may unconsciously deliver subtle cues that affect the eyewitness's selection of a suspect.)

Many other "environmental" factors affect eyewitness accuracy; they arise from the circumstances and are not under the control of LEOs conducting the lineup.

Factors that can affect accuracy of eyewitness identification and are *not* under the control of law enforcement have been called "estimator variables" in the eyewitness identification literature ("noise" variables in the experimental design literature); they include:

Estimator Variables (Not Controllable by Law Enforcement)

- *Weapon presence*: Presence or absence of weapon at time of incident (gun, knife, etc.);
- *Distinctive features*: presence or absence of a distinctive feature of the perpetrator;
- *Lighting*: this can affect visibility and recall of the incident;
- *Distance*: between eyewitness and perpetrator at time of incident;
- *Time elapsed*: length of exposure (seconds, minutes, etc.) to the suspect during the incident;
- *Stress*: for example, could be three levels (low, medium, high)
- *Race*: same- or cross-race (perpetrator and eyewitness are same or different races; studies suggest higher accuracy for the former. For references to these cross-race studies, please refer to Sporer (2001); Meissner and Brigham (2001); Wilson et al. (2013).

Datasets commonly used in EWID research proceed from designed experiments from the field of psychology, where some of the aforementioned variables are purposefully manipulated. Much EWID research has focused on comparing sequential versus simultaneous lineups (Amendola and Wixted, 2014; Lindsay and Wells, 1985;

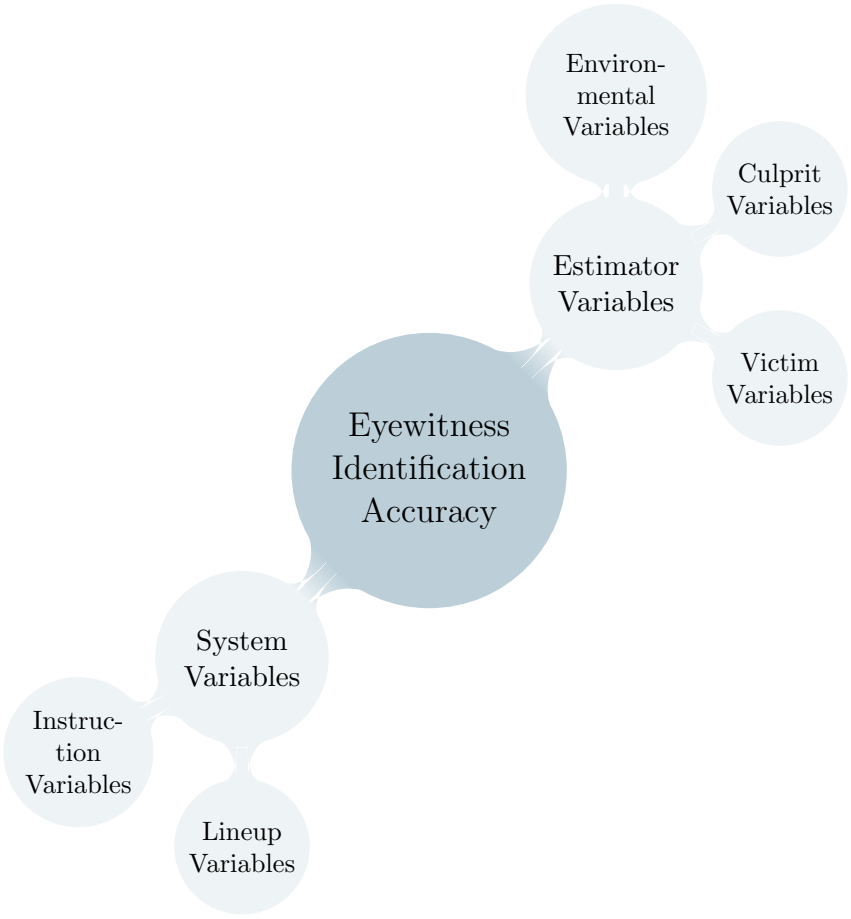


Figure 2.4: Diagram that shows the structure and relationships of examples of variables that could affect eyewitness identification accuracy

Carlson et al., 2008; Rotello et al., 2014). Relatively few studies considered the effects of several variables simultaneously in one experiment. Multi-factor experiments can be very informative in this context: if the effects of weapon presence, or cross-race, or delay (to identification) hugely dominate the effect of lineup type (sequential or simultaneous), then LEOs will know to focus their energy on, for example, minimizing the delay between incident and lineup, and less attention to the type of lineup. LEOs can also assess potential accuracy of an eyewitness in view of the conditions, such as presence or absence of a weapon or lighting that can affect visibility. Multi-factor experiments allow the estimation of jointly varying effects

from different sources. Studies that considered additional variables jointly include (Dodson and Dobolyi, 2016; Wixted et al., 2016b; Mickes et al., 2017; Sauerland et al., 2018; Clark, 2005; Sauer et al., 2010; Palmer et al., 2013; Humphries and Flowe, 2015; Carlson et al., 2016a,b; Colloff et al., 2016, 2017; Steblay, 1997). The National Research Council (2014) recommended the conduct of more multi-factor experiments, to better characterize the effect of presence (or absence) of weapon, relative to the choice of lineup (sequential or simultaneous).

Limitations of designed experimental data include lack of ecological validity, where some aspects of the reality of the EWID may not be reflected in the experimental situation. To address this issue, some researchers coordinated with law enforcement agencies to provide field datasets (Wixted et al., 2016a). In fact, multiple ongoing projects are seeking to collaborate with law enforcement agencies. However, field data lacks the underlying truth of suspect guilt. While the conviction and/or conclusions from evidence provide an estimated classification of guilt, using field data to train models to assess accuracy could lead to biased models. In this situation, the models would address the relationships of a particular law enforcement agency in identifying what they believe is a guilty suspect. Trade-offs exist for either forms of data, and the form used should be justified and consistent with the goal or research question.

Much literature exists on EWID from many perspectives (experimental, theory of memory, etc.); we give only a very brief background on that literature. Our main focus in this chapter and Chapter 3 expounds on the myriad of statistical tools that may offer powerful ways of identifying factors that affect EWID accuracy.

2.2.2 The Data

We include descriptions of data sets that can be found in the EWID field as representatives of the data available. These example data sets provide a snapshot of the types of information being collected in designed experiments, the size of such studies, and variables of particular interest to researchers. Some of these data sets will be used in later sections for evaluation purposes of the proposed framework in Chapter 4.

Most EWID experiments are conducted online using various survey platforms (such as Qualtrics©, Amazon© Mechanical Turk, SurveyMonkey©, etc.), but some are also conducted in-person. The participant views a video of a crime being committed, and then is asked to provide demographic information and answer questions concerning the video. The participant will make a decision in the lineup. Examples of lineups include six photos, used in an experimental setting by Chad Dodson of the Department of Psychology at the University of Virginia, as shown in Figure 2.1. The target suspect in this example photo lineup is in the top left, and represents the "true" perpetrator as shown in the video. If the researcher designates an innocent suspect, the innocent suspect will be placed in the same position as the true target, and is chosen as the filler that most resembles the true perpetrator. More example lineups can be found in Appendix A.

Dodson performed three different studies, obtained by way of online survey platforms. The three data sets will be referred to as the factor data set, repeated delay data set, and the delay data set. The factor data set focused on varying many conditions with a large sample size. The repeated delay and delay data sets focused on determining the effect of a delay from time of exposure to time of lineup identification. All three data sets share some common variables, shown in Table B.2.

Unique variables to each data set are shown in Table B.1. The factor data has 3233 respondents, the repeated delay data has 602 respondents with 12 lineup decisions per person (overall 7224 observations), and the delay data has 4301 respondents. Of primary interest are the counts for respondent decision: the target, the innocent suspect, a foil, or "target not present." From these counts, comparisons of the many other factors, such as weapon presence, lineup format, lineup bias, face recognizer ability, etc. can be created using the proportions of accuracies from various groupings of these factors. The accuracy rates of choosers versus non-choosers are also of specific interest.

The description of these data sets are examples of other such data collected in the field and in lab settings. Other data sets examined throughout this procedure include data from Mickes et al. (2017) and Seale-Carlisle et al. (2019). Table B.3 and Table B.4 provides the variable information for each data set from these sources, respectively.

The Mickes et al. (2017) data set has 5114 participants, after removing all participants that failed validation checks. Each participant was randomly assigned to different instructional environments:

(1) Provide a *confidence rating* for the decision made with unbiased instructions;

(2) *Liberal instruction*: instruct the participant to pick a person even if unsure;

(3) *Neutral instruction*: instruct the participant to pick a person if he or she sees the suspect from the video in the lineup or pick "not present";

(4) *Unbiased instruction*: instructs the participant the suspect may or may not be in the lineup, and to pick the suspect if he is in the lineup; and

(5) *Conservative instruction*: instruct to the person to pick "not present" if unsure.

Six experiments were run in total (experiment or expt. 1, 2, 3a, 3b, 4, and 5) in Seale-Carlisle et al. (2019). Each manipulated a different lineup condition:

(1) *Experiment 1*: manipulated the lineup format (simultaneous versus sequential) with 1993 participants;

(2) *Experiment 2*: manipulated stimulus format (photo or video) with 2271 participants;

(3) *Experiment 3a*: allowed different numbers of views for the lineup (1-lap vs. 2-lap vs. choice in video lineups) with 3096 participants;

(4) *Experiment 3b*: allowed different numbers of views for the lineup (1-lap vs. 2-lap vs. choice in photo lineups) with 3003 participants;

(5) *Experiment 4*: manipulated lineup size (six or nine total photos shown) with 2014 participants; and

(6) *Experiment 5*: manipulated the lineup format (simultaneous versus sequential) using a different set of stimuli from the previous five experiments with 2018 participants.

Unless otherwise specified, all lineups were shown using a simultaneous format. The data was cleaned of any discrepant or missing observations (i.e., if age was recorded as 0), which will account for any sample size differences from Seale-Carlisle et al. (2019).

These data sets will be used in Section 4.3 to assess the performance of the proposed modeling framework for EWID. They serve to represent the general forms of

experimentally-data obtained in psychology labs while trying to remain as ecologically valid as possible. A myriad of other EWID data sets may be available in full or summary form, but these work well as representatives.

2.3 Issues in Eyewitness Identification Research

EWID research began in the field of psychology, but has suffered from problems inherent to its historical development. The goal is to identify the EWID procedure that maximizes discriminability (i.e., the ability of eyewitnesses to discriminate between guilty and innocent suspects) (Gronlund et al., 2014). In effect, researchers seek to maximize eyewitness accuracy. The literature has conflicting results and conclusions about discriminability (Wixted and Wells, 2017). Due to these conflicting conclusions over the years of EWID research, the public has come to a consensus that eyewitness evidence is unreliable, and EWID researchers are seeking to reframe eyewitness evidence in an improved, more reliable light (Wixted et al., 2017a). They suggest eyewitness evidence collection procedures should follow certain conditions to ensure the evidence is not contaminated in a way that would render it unreliable, which are detailed in Section 2.2. Some of these issues, in addition to the questionable reliability of EWID evidence, include the relationship of confidence and accuracy, procedural decisions for eyewitness lineups, the choice of statistical methodology for the analysis of EWID experiments, etc.

2.3.1 Development of Eyewitness Identification Procedures

One of the underlying reasons for such discrepancies in the field is how EWID procedures are implemented, which have been historically applied in the field prior to being properly validated in a scientific setting (Wixted and Wells, 2017). Procedures

were primarily developed within the criminal justice system and used under the incorrect assumption that these were the best practices. For example, in 1999, the Department of Justice (DOJ) released a guide for law enforcement, developed by a technical working group, providing advisement on recommended lineup procedures (Department of Justice, 1999). The proposed guidelines discussed the collection and preservation of eyewitness evidence, and was expected to increase accuracy. It was "heralded as a 'successful application of eyewitness research,' 'from the lab to the police station'" (Gronlund et al., 2015). These reforms were meant to protect the innocent from wrongful conviction. At best, these reforms resulted in an increase in eyewitness conservatism (i.e., encouraging people to not choose a suspect if they are not sure). These researchers recommend that "future reforms are understood theoretically" so as to ensure "advocacy does not get ahead of the science."

In response to these missteps, a report from the National Research Council (2014) was released in 2014 as a more complete treatment of the problems, assessing the state of EWID research, shortcomings in the field, issues that should be considered or reconsidered, new methodologies that should or could be applied, etc. The report concludes that research in the effectiveness of law enforcement practices for implementing EWID procedures and the complex and interactive effects of system and estimator variables is incomplete. Yates (2017), the Deputy Attorney General at the time, released a memorandum for the heads of department of law enforcement and prosecutors promoting "sound professional practices and consistency" within the DOJ. There has been progress in the communication across different fields, but there is still progress yet to come. The consensus in the field is that the research "needs to be conducted in concert with the development and evaluation of theory" (Gronlund et al., 2015). The development of statistical methods should also work in conjunction with the development of eyewitness theory from psychology and policy

implementation from the judicial system.

2.3.2 Confidence-Accuracy Relationship

In addition to the conflict between real application and experimental settings, much eyewitness analysis is based on the confidence-accuracy (CA) relationship. Expressed confidence level (ECL) is considered a useful proxy for memory strength, which is said to be highly correlated with accuracy (Wixted and Mickes, 2010). Researchers have empirically identified evidence for the strong CA relationship in identifications made in a field study of police lineups from the Houston Police Department (Wixted et al., 2016a). The U.S. Supreme Court ruled in the case *Neil v. Biggers* (1972) that highly confident EWIDs are likely to be accurate, as long as these identifications meet certain criteria (III et al., 2012). Not all psychologists agree on this relationship, as their research has found confidence as a poor indicator of memory strength, and therefore of memory accuracy (Krug, 2007). Several issues accompany the usage of confidence as the sole predictor for accuracy. Confidence, as related to probability, depends on the status of the information of the subject who evaluates it. In 1947, Schrödinger said, "Since the knowledge may be different with different persons or with the same person at different times, they may anticipate the same event with more or less confidence, and thus different numerical probabilities may be attached to the same event."

One of the primary reasons for such high belief in the CA relationship is due to the calibration curve. Calibration is the agreement between objective (i.e., accuracy) and subjective variables (i.e., ECL) (Juslin et al., 1996). The calibration curve is a graph that plots accuracy on the x-axis and confidence on the y-axis. In an ideal situation, all participants with $c\%$ confidence should have $c\%$ accuracy, indi-

cating a well-calibrated respondent. These respondents would fall on a diagonal line where accuracy is equal to confidence. In Figure 2.5, well-calibrated participants would fall on a diagonal line where accuracy is equal to confidence (i.e., slope $b_1 = 1$ and intercept $b_0 = 0$). Overconfident participants would fall below this line, and under-confident participants would fall above this line. The over/under-confidence statistic ω is a supplementary statistic from the calibration curve, with a $\omega \in [-1, 1]$. Well-calibrated participants receive a score of 0 (i.e., perfect calibration). Under-confidence is indicated with a negative score, and overconfidence is indicated with a positive score.

These curves only provide information for the average captured CA relationship. While calibration curves may indicate "fair" and well-calibrated data, they may not clarify the impact of the various system and estimator variables on eyewitness choice and accuracy. The majority of calibration research in EWID attests that participants are usually over-confident in their assessment of their memory accuracy (Krug, 2007). Some psychologists view the calibration curve as a measure of the CA relationship, with good calibration as proof of a strong relationship (Gronlund et al., 2015). The CA relationship is not as strong for non-choosers as the relationship is for choosers (Clark et al., 2015; Brewer and Wells, 2006; Sporer et al., 1995). We define "non-choosers" as eyewitnesses who do not identify a suspect in a lineup, which is a "not present" decision. "Choosers" are defined as eyewitnesses who identify a suspect.

If an eyewitness decides that the guilty suspect is "not present" in the lineup, then the confidence statement may not indicate the accuracy of the decision. Furthermore, the strong relationship found by some researchers seems to only hold for confidence judgments made during the initial identification. Research has shown memory is malleable and can be distorted by other events (Clark et al., 2015; Lof-

tus, 2005; Wells and Bradfield, 1998). Initial confidence statements may adhere to the CA relationship more closely, but it does not necessarily translate for confidence statements made after the fact.

An eyewitness's ECL may vary at different times when presented with exactly the same circumstances, with similar variation demonstrated in other forensic sciences. A study done with fingerprint examiners shown the same evidence, but under different contextual circumstances, showed the examiners reversing their decisions (Dror and Charlton, 2006). Variance in and among eyewitness decisions should also be considered, such as within-eyewitness and between-eyewitnesses variation (Amendola and Wixted, 2015). Discrepancies may also exist in the laboratory setting versus the more stressful real-life situation. Confidence may be a good predictor for accuracy, but until these issues are addressed and studied more, it is not certain any conclusions made from confidence-based statistical methods are completely valid. We address this issue more in depth later in this section.

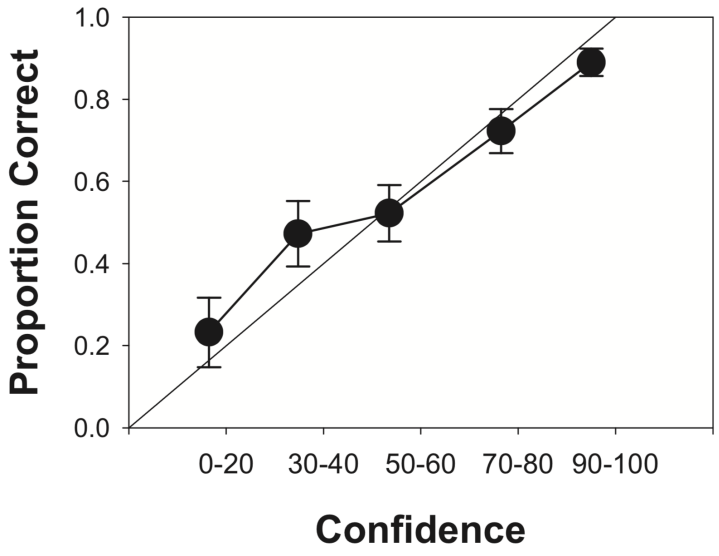


Figure 2.5: This plot shows the observed relationship between proportion of correct decisions and expressed confidence levels (Juslin et al., 1996; Wixted et al., 2015)

Additional variables likely work independently and/or interactively to predict for accuracy. Some researchers have started considering estimator variables in addition to the ones that are normally used. For example, the Cambridge Face Memory Test (CFMT), which provides a score for face recognition, could be a potential predictor for accuracy (Duchaine and Nakayama, 2006; McKone et al., 2012; Zhao et al., 2014; Andersen et al., 2014; Cho et al., 2015). Grabman et al. (2019) is the first study to relate CFMT to the predictive value of eyewitness confidence. Other face recognition tests exist such as the Glasgow Face Matching Test and the Recognition Memory Test (Burton et al., 2010; Warrington, 1984). The consideration of other variables such as the CFMT has not been well-studied in the past literature. As such, there is an obvious role here for generalized linear models (GLM), specifically logistic and multinomial logistic regressions, for the inclusion of such variables. The CFMT uses 72 sets of images to determine the face recognizer ability of the test-taker, which includes three sections:

(1) The *learning section*, where identical images are used;

(2) *Novel images*, where new images of the same face are used; and

(3) *Novel images with noise*, where additional noise is introduced.

Example images from the CFMT are shown in Appendix C.

Recently, Wixted et al. (2015) suggest the "combined weight of theory, empirical evidence, and revelations from DNA exoneration" lead to the conclusion that initial identifications are more reliable than have been previously thought. Simply because a coherent story has been generated does not necessarily mean this story is true. Confidence is the ubiquitous gathering of information and knowledge that supports some hypothesis, which is sometimes known as confirmation bias. It should be noted that most psychologist researchers in the field agree upon the CA relationship.

2.4 Current Statistical Methodologies

Surprisingly few statistical approaches have been used in analyzing data from EWID experiments. Some psychologists have historically used the diagnosticity ratio and the discriminability index (d') as measures of comparison across different eyewitness procedures (Wixted and Mickes, 2012; Mickes et al., 2014; Georgeson; Mickes et al., 2017). Other psychologists have been proponents of the point-biserial correlation coefficient (r_{pb}) and Goodman and Kruskal's gamma (G), which tend to result in misleading conclusions. Additional methods include: calibration curves (Juslin et al., 1996; Krug, 2007; Gronlund et al., 2015; Clark et al., 2015; Brewer and Wells, 2006; Sporer et al., 1995); ROC curve analysis based on the SDT paradigm using ECLs as the cut-points (Clark et al., 2015; Wixted and Mickes, 2010, 2012; Pepe, 2000; Wixted and Mickes, 2015b); partial area under the curve (pAUC) as an extension of ROC analysis (Walter, 2005; Mickes et al., 2014; Wixted et al., 2017b; Lampinen et al., 2019); estimation of posterior probability of guilt based on Bayes' Theorem (Wells and Lindsay, 1980; Wells et al., 2015a,b); expected utility (Lampinen et al., 2019; Smith et al., 2018); logistic regression (Wetmore et al., 2015; Andersen et al., 2014); and log-linear models (Luby, 2016, 2017). Overall, psychologists are exploring these methods to further the theory of eyewitness cognition, which consists of memory judgments (making a selection in a lineup) and accompanying metacognitive context (the associated confidence statement) (Gronlund and Benjamin, 2018). We describe these approaches in this section.

2.4.1 Diagnosticity Ratio and Discriminability Index

The diagnosticity ratio (DR), equivalently positive likelihood ratio LR_+, is the ratio of the odds of the suspect being guilty relative to the odds of the suspect being in-

nocent. It measures the probative value, which is how much information is available in the evidence, of a lineup procedure. The DR provides the posterior odds of guilt or the likelihood that a guilty suspect is identified in a lineup (Wixted and Mickes, 2012; Mickes et al., 2014).

$$\begin{aligned} \text{DR} &= \frac{\text{Correct ID Rate}}{\text{False ID Rate}} \\ &= \frac{\text{HR}}{\text{FAR}} \\ &= \frac{P(\text{suspect identified}|\text{suspect guilty})}{P(\text{suspect identified} \mid \text{suspect innocent})} \end{aligned} \tag{2.1}$$

The discriminability index d' (also known as the sensitivity index), which originates from signal detection theory, is a popular estimate signal strength (Georgeson). Discriminability is defined as the ability to perceive and respond to differences among stimuli. The discriminability or d' is defined as the separation between two means expressed in a common unit of their equal or unequal variances. A higher d' indicates a larger pAUC,[2]. Equation 2.2 shows the relationship of d' to area under the curve (AUC). Here, $z(\cdot)$ represents the normal score[3] associated with the function inputs:

$$\begin{aligned} d' &= z(\text{Correct ID Rate}) - z(\text{False ID Rate}) \\ &= \sqrt{2} \cdot z(\text{AUC}) \end{aligned} \tag{2.2}$$

Both the DR and d' are summaries that characterize EWID performance across all "levels" of system and estimator variables. However, any such measure oversimplifies

[2]More information on pAUC in Section 2.4.2.

[3]The normal score for some value x is found by normalizing the value such that $z = \frac{x-\mu}{\sigma/\sqrt{n}}$, where μ is the population mean, σ is the population standard deviation, and n is the number of observations available.

performance: a single index cannot capture all the information in a comparison between tow procedures. In experimental settings, maximizing the DR may lead to more conservative responding (i.e., more likely to choose a "not present" response) (Wixted and Mickes, 2012). For example, more extreme instructions designed to protect the innocent induced a higher DR, but did not necessarily lead to a better accuracy result. The DR has a tendency to naturally increase even if discriminability is constant (Mickes et al., 2017). This could result in misleading conclusions, since a different lineup instruction would not (and should not) change the witness's memory, which should be constant across conditions. The DR was a popular performance metric for comparing procedures (e.g., simultaneous versus sequential), until some researchers (e.g., Wixted and Mickes (2015a) observed that a third variable, ECL, can affect this ratio, and that DR could confound changes in accuracy with changes in "response bias."

The ROC curve is a plot of the hit rate (HR) versus the false alarm rate (FAR) for various levels of ECL (e.g., "at least 10% confident," ..., "at least 40% confident," ..., "at least 100% confident"); the slope of the ROC curve at one of these points corresponds to $\frac{\text{HR}}{\text{FAR}}$ (i.e., the DR) at that ECL. Hence, a straight line indicates the same DR for all ECLs (i.e., DR does not depend on ECL in this case). Because ROC curves incorporate additional information (e.g., ECL, they are viewed as more useful for comparing methods than the simple DR collapsed over all ECL categories. The DR allows the researcher to disregard suspect identifications that are categorized as "untrustworthy" (i.e., identifications made with low confidence) (Wixted and Mickes, 2015a). Both DR and d' should be accompanied by measures of variability (but often are not). The National Research Council (2014) report acknowledged advantages of ROC over DR in some circumstances, but emphasized that other statistical analyses of EWID experimental data are more powerful (e.g., logistic

regression, binary classifiers); see below.

2.4.2 ROC Curves

The ROC curve was originally developed in the 1950s and used with electronic SDT, with first applications in radar (Hajian-Tilaki, 2013). Since then, researchers in many other fields, including psychology, diagnostic radiology, medical diagnostics, and machine learning, use it to compare different techniques, often by its area under the curve (AUC) or pAUC. An ROC curve plots the HR or Se against the FAR or (1 - Sp). The curve is based on some decision variable, and the counts of good and bad results will vary based on the chosen threshold of that decision variable. It is a descriptive device that demonstrates the range of trade-offs between the true positive rate (TPR) and the false positive rate (FPR) within a particular test (Pepe, 2000). An ROC curve with better discriminant capacity will appear as a curve closer to the upper left-hand corner in the ROC space. A curve lying on a straight diagonal line with a slope $b_1 = 1$ indicates the test has a performance similar to that of chance. The slope of the tangent line at each point of the ROC curve is equal to the likelihood ratio, which is the ratio of the two density functions describing the two distributions of the decision variable in population one and population two. These distributions are usually assumed to be normal. Sensitivity (Se) and specificity (Sp) are defined as

$$\text{Se} = \frac{\text{Number of TP}}{\text{Number of TP} + \text{Number of FN}} \tag{2.3}$$

and

$$\text{Sp} = \frac{\text{Number of TN}}{\text{Number of TN} + \text{Number of FP}}. \tag{2.4}$$

Use of ROC analysis in EWID research was first proposed by Wixted and Mickes (2012) because a lineup procedure is characterized by a range of DRs, rather than a single DR. Wixted and Mickes state that the ROC can show which of two procedures is diagnostically preferable. Researchers disagree if ROC analysis is the best method to measure underlying discriminability (Wells et al., 2015a,b).

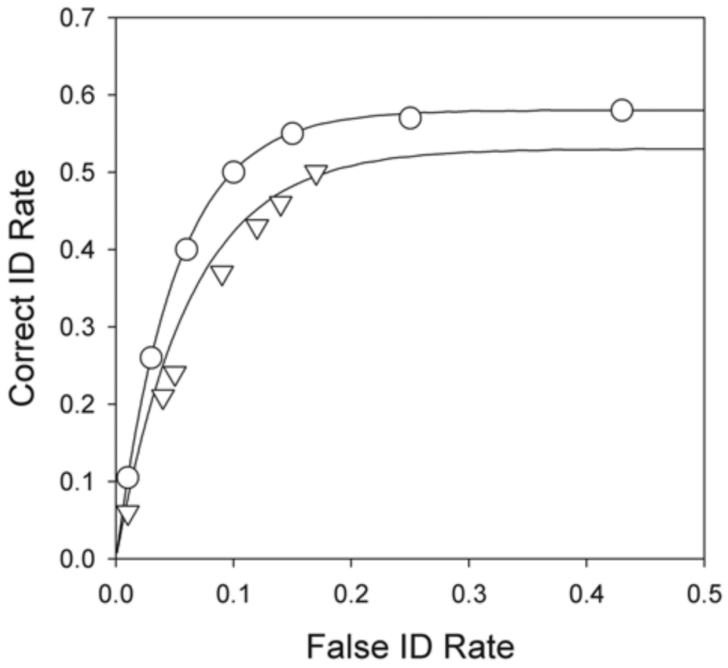


Figure 2.6: Hypothetical ROC curves for simultaneous (circles) and sequential (triangles) procedures. In this case, the plot concludes that simultaneous procedures are diagnostically superior (Gronlund et al., 2014).

Expressed Confidence Levels. The points on ROC curves (HR versus FAR) constructed from data in EWID experiments can be based on many "third" variables. A common "third" variable is the eyewitness's ECL at the time of the lineup. Researchers have stated that only confidence recorded immediately after the identification should be used (Sauer et al., 2019) In most lab experiments, the "mock eyewitness" reports an ECL often as numerical response along a scale (0 to 1) to

the question, "How confident are you in your identification?" with discrete choices (e.g., "0.0," "0.1," "0.2," ..., "1.0" (11 categories), or, more coarsely, "0.0," "0.2," "0.4," ..., "1.0" (six categories)). As with any scale, the difference in a respondent's reactions of, for example, "0.0" versus "0.2" may be more clear to a respondent than the difference in the respondent's reactions of "0.4" to "0.6," which the respondent may possibly view as less distinguishable. In real life, LEOs recognize that typical eyewitnesses are not comfortable with numerical scales, so they solicit their responses as verbal descriptors. The LEO's translation of those descriptors as a numeric value may depend on the LEO.

Sauer et al. (2019) state, "The extent of variation in the confidence-accuracy relation precludes us from making strong, generalized claims about the accuracy of high confidence identification decisions, even under pristine conditions[4], when evaluating individual identifications." They note that an individual identification differs from "aggregate level" confidence-accuracy relationship, which is equivalent to an ensemble of eyewitnesses. Either way, the ECL is likely subject to uncertainty, depending on many factors (such as high levels of stress) whose effects on ECL remain largely unstudied. These effects deserve further study so the uncertainty in ECL can be incorporated in the analysis of data from EWID experiments.

Confidence-based ROC analysis has some connection to ROC analysis in diagnostic medicine, used to compare the diagnostic superiority of different systems (e.g., magnetic resonance imaging (MRI) versus mammography). Target present [absent] lineups may be viewed as "condition present [absent]" (e.g., presence or absence of tumor) (Wixted and Mickes, 2015a). The analog of the ROC points in EWID (ECL) are ranges of assessment of condition (e.g., "definitely not malignant" to "definitely

[4]viz., only one suspect in the lineup, the suspect did not stand out, the witness was cautioned that the culprit may not be present, double-blind testing was used, and the confidence statement was obtained at the time of testing

malignant"); cf. (Park et al., 2004; Mickes et al., 2012). Note that radiologists are trained professionals, with their training based on medical standards, whereas eyewitnesses are rarely "trained" in face recognition and likely have no prior practice nor experience when identifying suspects in a lineup. Kantner and Dobbins (2019) suggest that a given confidence report is largely (if not completely) determined by individual differences when testing for memory for words. These differences are broadly defined as self-efficacy, use of the confidence scale, and/or other factors. While the study from Kantner and Dobbins (2019) tested memory for words, rather than memory for faces, there may possible extensions of their conclusions to memory for faces. Nonetheless, ROC analysis may have value in the analysis of EWID experiments, if sources of uncertainty are properly taken into account. In the subsequent sections, we discuss statistical methodology alternatives to ROC analysis.

Variability. Researchers realize that the decision criteria (in this case, ECL) may vary among participants, and use the term *criterial variance* to represent the variance in decision criteria (i.e., the differences among eyewitnesses in their criteria for making identification decisions). This is also known as criterial noise or criterion variance. Decision criteria refers to the cutoff that is used for making an identification or responding "not present." Since people use different criteria for their individual cutoffs, there is variability across people.

Researchers also assume variance in the underlying distributions for target and fillers in the SDT model. The variances assumed are the equal variance versus unequal variance model for the underlying normal distributions for memory strength; see Figure 2.7. These distributions are estimated for the latent variable of memory strength. For each ECL c, the DR d_c is theoretically calculated based on these

assumed normal distributions for target and foil decisions; see Equation 2.5.

$$d_c = \frac{\mu_{\text{Target}} - \mu_{\text{Foil}}}{\frac{1}{2}\left(\sigma^2_{\text{Target}} + \sigma^2_{\text{Foil}}\right)} \tag{2.5}$$

This variability aims to represent the between-participant, versus the within-participant, variability. Within-participant variability is a measure of a single participant's ECL across many lineups, of many or the same stimuli. The within-participant variability may be considered as a type of "measurement error." In this case, the measurer is not necessarily the experiment conductor or the LEO, but rather the eyewitness. Russ et al. (2018) examined the phenomenon, and reached the conclusion that a more realistic "field encounter" does not necessarily engender robust eyewitness identifications due to development of "limited cognitive representations of a target." A more controlled setting results in more consistent and correct identifications. They suggest that the degree of familiarity a participant has with a target could be a potential index for EWID accuracy. Kantner and Dobbins (2019) reiterate the point that ROC curves should be fitted to individual participants rather than in aggregate form across a large group. They found large inter-subject differences, and expect group ROC curves to be variable (i.e., noisy).

Both measures of variability differ from methods that provide intervals for point estimates. Each DR serves as a point estimate, which should have some measure of variability to capture the true DR. This point is explored further in Section 3.3.5.

Construction. A confidence-based ROC curve is constructed by plotting the number of correct identifications versus the number of false identifications, with each point of this curve within an ordinal category of expressed confidence level (ECL) from 0% to 100% (Gronlund et al., 2015). The number of categories of confi-

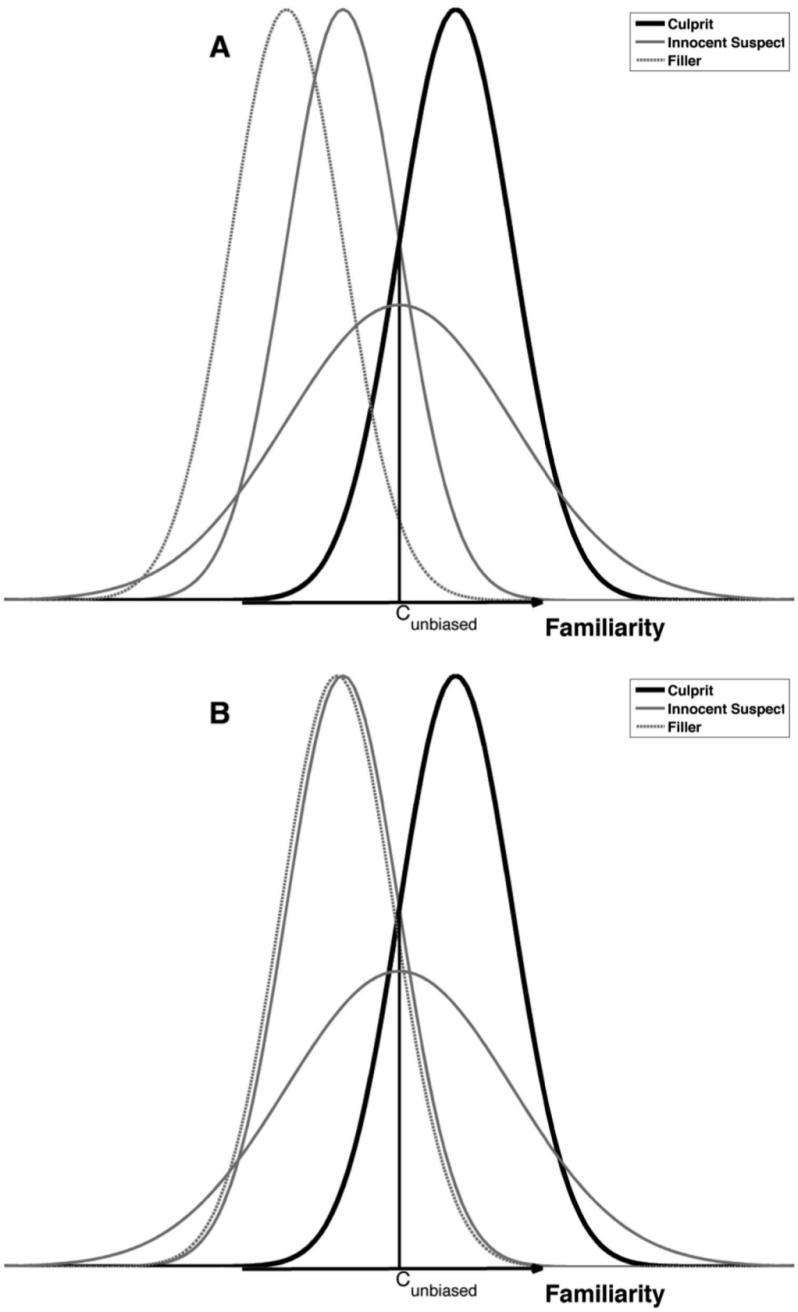


Figure 2.7: Plot A shows a lineup procedure under the assumption of equal variance for the culprit and innocent suspect normal distributions and plot B shows a lineup procedure under the assumption of unequal variance (i.e., assuming criterial variance). C represents the ECL (Smith et al., 2016).

dence varies among researchers. The correct identification rate[5] at a given ECL $c\%$ is estimated as the proportion of people who correctly chose the perpetrator in the "target present" condition and expressed a confidence level of at least $c\%$. The false identification rate[6] at a given ECL $c\%$ is estimated as the proportion of people who chose the "innocent" suspect incorrectly within the target absent population and expressed confidence of at least $c\%$. This is done for each ECL in $0\% \leq c \leq 100\%$. The slope (i.e., tangent at each plotted point) of the ROC curve is equal to the DR for that ECL.

Area Under the Curve. The AUC is a standard summary of an ROC curves for purposes of comparing procedures, with preference for the procedure with the larger AUC. Some authors (Mickes et al., 2014; Wixted et al., 2017b) prefer to summarize the method's performance via a pAUC. The AUC represents the average value of sensitivity over all possible FARs $\in [0, 1]$ (Walter, 2005), and is related to the Mann-Whitney U-statistic, which evaluates the significance of the difference between the sample distribution of positive and negative decisions (Pepe, 2000). Some authors (Mickes et al., 2014; Wixted et al., 2017b) prefer to summarize the method's performance via a pAUC, particularly in situations where the maximum value on the x-axis (here, false identification rate or FAR) is guaranteed to be less than 1. In a target present lineup, five possible false identifications and one correct identification exist, so the maximum possible false identification rate is $\frac{n-1}{n}$, where n is the number of people or photos in the lineup. The pAUC has limitations also (see Walter, 2005), and a comparison of procedures based on either AUC or pAUC may not be straightforward if one curve is not consistently higher than the other across the entire range of HR and FAR (Streiner and Cairney, 2007).

[5]HR or TPR
[6]FAR or FPR

2.4.3 Logistic Regression

The logistic regression model assumes a binary dependent variable and one or multiple continuous and/or categorical independent variables. In the EWID paradigm, the dependent variable is accuracy (correct or false identification, whose definition depends on the inclusion or exclusion of a designated innocent suspect) and the independent variables are relevant system and estimator variables (Wetmore et al., 2015; Andersen et al., 2014). An expansion of logistic regression use in the EWID paradigm is provided in Section 3.2.

An estimated coefficient in logistic regression provides the change in the odds ratio for a one unit increase (for a continuous variable) or the change in odds for one category versus the reference category (for a categorical variable). Variables selected in the model are deemed the informative variables of discriminability. The predicted accuracy from the fitted logistic regression model can be viewed in a contingency table with the observed accuracy, providing model performance. Cross-validation can be used as well.

As noted earlier, using a binary response variable may not be the most realistic choice. Researchers are interested in understanding what influences a witness to choose the true suspect, to choose an "innocent" suspect, to choose a foil, or to not choose at all. Multinomial logistic regression or some other multiple classification method may work better in this respect.

2.4.4 Expected Utility

Some researchers adopted the idea of expected utility and decision theory in the analysis of ROC curves (Lampinen et al., 2019). Since the dominating ROC curve is not always clear, some researchers have suggested an approach based on decision

theory and estimation of the base rate (BR) (equivalent to prior probabilities). The possible EWID task decisions result in some subjective benefit or cost, leading to the notion of a procedure's "utility," defined as the product of the probability of the outcome, BR, and cost or benefit. Other proposed measures of utility include terminal point utility (utility calculated at the right-most point on the ROC curve), high-confidence utility (based on only those participants expressing high utility), average utility (averaged over all confidence-level utilities), and maximum utility.

Smith et al. (2018) discuss a metric for ROC curve analysis based on expected utility, which distinguishes diagnostic utility and ECL. This metric, known as the deviation from perfect performance (DPP), claims to consistently indicate which of two lineup procedures has higher expected utility. DPP is based on the global measure of predictive performance r discussed by Shiu and Gatsonis (2008). The modified measure for ROC curve use is defined as

$$\text{DPP}(c) = [1 - \text{Suspect}(c)] + [\text{Innocent}(c)], \tag{2.6}$$

where $0 \leq \text{Suspect}(c),\ \text{Innocent}(c) \leq 1$. Suspect$(c)$ is the suspect identification rate at a given point and Innocent(c) is the "innocent" suspect identification rate at the same point. Perfect performance is achieved when Suspect$(c) = 1$ and Innocent$(c) = 0$ or DPP$(c) = 0$. The index is computed as the average DPP(c) of the entire ROC curve, providing a value of how much a lineup procedure deviates from "perfect performance." The DPP(c) is not tied to a specific region of the ROC space, which means it does not force researchers to make comparisons that are confounded by an eyewitness's ECL. In both of these expected utility approaches, the assumed cost and benefit of decisions is subjective and may not accurately portray the information. Both methods rely on a single value to summarize the entirety of a lineup, similar

to DR and d'.

2.5 Example

As noted in Section 2.4.1, the DR depends not only on an eyewitness's tendency towards "conservative" or "liberal" identification (as measured by "expressed confidence level"), but also on numerous other factors, including:

(A) *Type of lineup*: for example, two levels (simultaneous versus sequential);

(B) *Weapon presence*: usually has two levels (presence or absence of some weapon); more levels could be considered, for example, the presence or absence of multiple weapons, such as gun, knife, towel, none;

(C) *Stress*: for example, three levels such as low, medium, high;

(D) *Time elapsed*: between incident and exam (e.g., three levels: 30 minutes, two hours, one day, etc.);

(E) *Race*: for example, two levels (same or different race); or four levels (eyewitness and culprit are: white and white; white and non-white; non-white and white; non-white and non-white; non-white and white);

(F) *Subject*: N levels, corresponding to N subjects.

If a study is sufficiently large, one could construct an ROC for each participant corresponding to each of these conditions (i.e., plot HR versus FAR at different ECLs for each participant of the N participants). To avoid running such an enormous experiment, one would sensibly consider running a fraction of all possible combinations.[7] One can then summarize the information in the ROC via different

[7]See, for example, Box et al. (2005) on constructing fractional factorial designs.

measures, such as the logarithm of the AUC, or log(AUC). Consider the following approach:

The Model. Let $y_{ijk\ell mnr}$ denote the log(AUC) for the rth trial using participant n ($n = 1, ..., N$) for procedure i, weapon level j, stress level k, time condition ℓ, and cross-race effect m. Then we could write:

$$y_{ijk\ell mnr} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_\ell + \phi_m + (\alpha\beta)_{ij} + ...(\text{interactions})... + \epsilon_{ijk\ell mnr}$$

where μ represents the overall average log(AUC) across all conditions, the next six terms reflect the main effects of A (lineup procedure: $i = 1$ for sequential and $i = 2$ for simultaneous); B (weapon: $j = 1$ for presence and $j = 2$ for absence of weapon); C (stress level: $k = 1$ for low, $k = 2$ for medium, $k = 3$ for high); D (elapsed time between incident and report: $\ell = 1$ for 30 minutes, $\ell = 2$ for two hours, $\ell = 3$ for one day); E (cross-race effect: $m = 1$ for same race, $m = 2$ for different races); F (participant effect: $n = 1, 2, ..., N$ participants); "(interactions)" reflects the joint effect of two or more factors together, and the last term, $\epsilon_{ijk\ell mnr}$ represents any random error in the r^{th} trial that is not specified from the previous terms (e.g., measurement, ECL, multiple trials, etc.). This approach would allow separation of the effects of the different factors, enable one to assess which factors have the greatest influence on the outcome (here, logarithm of the area under the ROC curve: bigger is better), and especially to evaluate the importance of these factors relative to variation among "eyewitnesses." It may be that eyewitnesses are the greatest source of variability, dominating the effects of all other factors. Or it may be that, in spite of person-to-person variability, one or more factors still stand out as having strong influence on the outcome. Note that,

(i) Other covariates could be included, such as age and gender of participant; and

(ii) The ROC curve need not be defined in terms of "expressed confidence level" thresholds, if a more sensitive measure of "response bias" (tendency towards "liberal" versus "conservative" identifications) can be developed.

With Data. Carlson and Carlson (2014) use pAUC, as a summary measure of the information in an ROC curve (bigger is better), for each of 12 different conditions defined by three factors:

(A) *Procedure*: three levels (simultaneous [SIM: suspect in position four], sequential [SEQ2: suspect in position two], sequential [SEQ5: suspect in position five]);

(B) *Weapon focus*: two levels (present versus absent);

(C) *Distinctive feature*: two levels (present versus absent).

The data are provided in their Table 3, along with 95% confidence intervals. Because the length of a confidence interval is proportional to the standard error, pAUC values with shorter confidence intervals correspond to having smaller standard errors and hence should have higher weights. The logarithms of the reported pAUC values and weights (reciprocals of the lengths of the reported confidence intervals) are given in Table 2.3.

For this study, the data on all $n = 2675$ participants (720 undergraduates and 1955 SurveyMonkey© respondents) were combined, and ECLs were solicited on a seven-point scale. Variations in the 12 log(pAUC) values can be decomposed into three main effects (one each for procedure, weapon, and feature), and their two-way interactions. The raw data may permit a more detailed analysis. The data can be analyzed using a less complex model than that stated above (because we have fewer terms):

Condition	Procedure	Weapon	Feature	$5 + \log(\text{pAUC})$	Weight
1	Simultaneous	Yes	Yes	1.31	47.6
2	Simultaneous	Yes	No	1.73	33.3
3	Simultaneous	No	Yes	0.93	55.6
4	Simultaneous	No	No	1.88	45.5
5	Sequential 2	Yes	Yes	1.49	47.6
6	Sequential 2	Yes	No	1.23	47.6
7	Sequential 2	No	Yes	1.09	52.6
8	Sequential 2	No	No	1.59	41.7
9	Sequential 5	Yes	Yes	1.70	38.5
10	Sequential 5	Yes	No	0.98	58.8
11	Sequential 5	No	Yes	0.66	66.7
12	Sequential 5	No	No	1.49	55.6

Table 2.3: Logarithms of the reported pAUC values and weights (reciprocals of the lengths of the reported confidence intervals).

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \epsilon_{ijk}$$

where y_{ijk} denotes $[5 + \log(\text{pAUC})]$ for procedure $i = 1, 2, 3$; weapon condition $j = 1, 2$; feature $k = 1, 2$; μ represents the overall average log(pAUC) across all conditions; α_i represents the effect of procedure i; β_j represents the effect of weapon condition j; γ_k represents the effect of feature condition k; and the next three terms reflect the three two-factor interactions between the main factors. The analysis of variance, where log(pAUC) values are weighted according to the values in the last column of Table 2.4. None of the factors are significant.

We can decompose the two degrees of freedom in the sum of squares for *Procedure* (three levels), 8.04, into two single degree of freedom contrasts, *SEQ2* versus *SEQ5* (4.14) and *SIM* versus the average of *SEQ2* and *SEQ5* (3.90), and consider all pairwise interaction terms among the four "main effects." All single-degree-of-freedom effects remain non-significant, in either this weighted analysis or in an unweighted analysis.

The result is surprising, because all three factors in Table 2.4 (lineup type,

Source of Variation	**Degrees of Freedom**	**Sum of Squares**	**Mean Square**	**F-Statistic**	**p-value**
Procedure	2	8.04	4.02	1.129	0.470
Weapon	1	2.94	2.94	0.826	0.460
Feature	1	14.72	14.72	4.138	0.179
Procedure × Weapon	2	0.59	0.30	0.083	0.923
Procedure × Feature	2	10.41	5.21	1.463	0.406
Weapon × Feature	1	34.80	34.80	9.780	0.089
Residuals	2	7.12	3.56		

Table 2.4: Analysis of variance table for log(pAUC)[a]

[a]Data on pAUC from Table 3 in Carlson and Carlson (2014) (National Research Council, 2014, see Appendix C, pgs. 150-154).

presence/absence of weapon, and presence/absence of distinctive features) appear in the literature as having consequential effects on accuracy. The lack of significance could be due to low power in detecting small effect sizes, the use of ECL in the ROC, or the insensitivity of pAUC in characterizing a condition. For a discussion of the advantages and disadvantages of using AUC versus pAUC as a summary measure see Walter (2005). A complete set of raw data may yield a more powerful analysis with different results, as might a different summary measure of the ROC curve, such as the AUC.

Chapter 3

Statistical Models and Methods for Adaptation

Based on the amount of new research, the field of EWID is gaining much traction in terms of obtaining new sources of data, new methods for analysis, etc. We seek to expand on possible methods to be used with EWID data by incorporating and looking at existing statistical methodologies used in other fields, such as diagnostic medicine. We review potential statistical models to quantify the effects of factors influencing the accuracy of eyewitness identification in controlled experiments and to explore methods for analyzing the results from these experiments, using statistical models and intuitive displays of the effects of these factors.

For example, while the ROC curve has been used for decades in statistical quality control, diagnostic medicine, and many other fields where methods or techniques are being compared. The ROC curves using data from eyewitness identification experiments are constructed using the experimental participant's ECL in the identification, which is affected by error and variation. We present alternative statistical approaches, some of which have been used in similar scenarios (e.g., comparing

medical diagnostic imaging modalities) with the aim of developing more powerful analyses to better quantify the effects of variables (including or modifying the ECL) influencing the accuracy of EWID procedures. These statistical tools may offer powerful ways of identifying factors that affect EWID accuracy, beyond the conventional tools of diagnosticity ratios and ROC. Note that many of the proposed methods depend on treating eyewitnesses as binary classifiers, which, as previously discussed, is problematic.

We provide a literature review of potential avenues for adaptations of statistical methodologies from other fields. In Section 3.1, potential methods from the field of diagnostic medicine are reviewed. Alternative statistical methods to the conventional ROC curve are provided in Section 3.3. Finally, Section 3.4 provides a discussion of the methods in terms of their adaptability for EWID experiments as well as to suggest improved models. We recommend potential statistical approaches in the final section, depending on the data, experimental conditions, and concomitant information available.

Note: sections of this chapter appear as part of the *Handbook of Forensic Statistics* (see Liu et al., 2020, Chp. 21).

3.1 Statistical Models From Diagnostic Medicine

The tasks in diagnostic medicine, to identify abnormalities in an image, bear resemblance to the EWID task, to identify a perpetrator from a lineup. Accordingly, we discuss approaches that have been developed for comparing detection modalities and conducting meta-analyses in diagnostic medicine that may be suitable for comparing procedures (e.g., lineup format) in EWID. A successful model in diagnostic medicine is a bivariate random-effects statistical model for sensitivity and specificity

(HR and [1 – FAR] in EWID, respectively), which leads to models for positive predictive value (PPV) and negative predictive value (NPV). These methods apply to meta-analysis for combining data from similar studies.

Meta-analysis is used to provide synthesized statistics across similar studies, including multiple tests of diagnostic accuracy. Research synthesis, when done well, also provides a determination of study validity based on study design and execution of included studies, and is also used to test effects of patient and test characteristics and to identify areas for further research (Irwig et al., 1994). Likewise in EWID, several meta-analyses and other forms of research synthesis have been conducted, and a database identifying studies related to both single and combinations of variables (e.g., presence/absence of weapon or retention interval) is under development. This new database, which will be publicly accessible, is anticipated to identify gaps in the existing knowledge base and facilitate new research syntheses.[1]

A bivariate random-effects model (in the form of a hierarchical Bayesian model) was originally proposed by DuMouchel (1994) as a compromise between those who used the traditional fixed-effect meta-analytic methods and those who argued against meta-analysis (i.e., that data from across studies should never be combined) (Junaidi et al., 2012). Other researchers support Bayesian methods in meta-analysis for the study of fixed and random effects (Sutton and Abrams, 2001; Rutter and Gatsonis, 2001). Most of these methods compare test results to a "gold" reference standard, which does not necessarily exist in the EWID paradigm. Certain methods that overcome the lack of the reference standard could be adapted for the use in EWID data analysis, perhaps by simply comparing two experimental methods.

The framework of meta-analysis is natural for the EWID paradigm. Meta-analysis requires the combination of data from various sources (i.e., studies and

[1]Joanne Yaffe, personal communication.

experiments) that may have been performed using the same or similar settings with a common result, but were performed at different times. In the EWID field, each individual court case or eyewitness could be viewed as an individual "study." We are interested in combining the information obtained across many of these "studies" (i.e., persons) or court cases or for different experiments from various researchers in the EWID field. Should these models be adapted to EWID research, diagnostic test accuracy literature could provide a solid foundation for the work. For example, the Cochrane Collaboration, a non-profit organization formed to organize medical research findings, may provide guidelines to application through their Cochrane Handbook (Macaskill et al., 2010).

The sections below review statistical models for meta-analysis from the field of diagnostic medicine, which can be adapted to be used with EWID data.

Logitnormal Bivariate Random-Effects Model. A popular approach to assessing the impact of several variables on accuracy in diagnostic medicine, and hence also EWID experiments, is a bivariate model for the logit transformation[2] sensitivity and specificity proposed by Reitsma et al. (2005), and generalized by Chu and Cole (2006). For example, in comparing diagnostic technologies in a meta-analysis, the Reitsma model is a linear mixed effects model and assumes the logit-transformed sensitivity and specificity marginally follow a normal distribution, then combines the pair into a bivariate normal distribution.

The proposed bivariate model is a logitnormal bivariate random-effects model that relies on a two-level structure, which estimates the between-study variation and the correlation between sensitivity and specificity. The correlation provides information on the heterogeneity of the studies. Let $\theta_{A,i}$ be the true logit sensitivity

[2]The logit transformation is defined $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$.

of individual study i, with common mean value θ_A and within-study variance σ_A^2. Similar notation is used for the true logit specificity using $\theta_{B,i}$. Let σ_{AB} represent the covariance between logit sensitivity and logit specificity. Then the model is:

$$\begin{pmatrix} \theta_{A,i} \\ \theta_{B,i} \end{pmatrix} \sim N \left[\begin{pmatrix} \theta_A \\ \theta_B \end{pmatrix}, \Sigma \right], \text{ where } \Sigma = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix}.$$

The Chu and Cole (2006) extension reduces some Reitsma et al. (2005) model assumptions. First, they assume the number of true negatives n_{00} and the number of true positives n_{11} follow binomial distributions,

$$\begin{aligned} n_{11,i}|\theta_{A,i} &\sim \text{Binomial}(N_{A,i}, p_{A,i}) \\ n_{00,i}|\theta_{B,i} &\sim \text{Binomial}(N_{B,i}, p_{B,i}). \end{aligned} \tag{3.1}$$

Let $p_{A,i}$ and $p_{B,i}$ be the observed proportions for sensitivity and specificity, respectively. The logit-transformation is

$$\begin{aligned} \text{logit}(p_{A,i}) &= \boldsymbol{X_i}\boldsymbol{\alpha} + \theta_{A,i} \\ \text{logit}(p_{B,i}) &= \boldsymbol{Z_i}\boldsymbol{\beta} + \theta_{B,i}. \end{aligned} \tag{3.2}$$

Here, $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ are vectors of covariates that are related to sensitivity and specificity, which are possibly overlapping. The Chu and Cole (2006) extension assumes the following structure

$$\begin{pmatrix} \theta_{A,i} \\ \theta_{B,i} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right], \text{ where } \Sigma = \begin{pmatrix} \sigma_A^2 & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \sigma_B^2 \end{pmatrix}.$$

This model for sensitivity and specificity was adapted by Leeflang et al. (2012) for PPV and NPV; it is identical to the model found in Chu and Cole (2006), except PPV and NPV are used in place of sensitivity and specificity, respectively. PPV and NPV take account of prevalence, so Leeflang et al. (2012) chose to incorporate prevalence in their model by allowing it to vary, thereby avoiding its estimation.

Chu et al. (2009) proposed a trivariate model that jointly models PPV, NPV, and prevalence. Ma et al. (2014) modified the trivariate model to handle a missing reference test outcome (i.e., missing disease status). The model extends the Reitsma et al. (2005) model and Chu and Cole (2006) by adding prevalence as an additional random variable, assuming a trivariate normal distribution. The latent class bivariate model is another way to evaluate the accuracy of diagnostic tests in the absence of a "gold standard" reference (Eusebi et al., 2014); this approach models the between-study heterogeneity by assuming each study in the meta-analysis belongs to one of K latent classes.

The logitnormal bivariate random-effects model performs well in characterizing the performance of different diagnostic modalities (EWID procedures), in part because it models the logits of the probabilities; models for dependent outcomes restricted to a range such as $[0, 1]$ must incorporate constraints in the parameter estimation.

The model does involve only one correlation parameter, although extensions to incorporate additional correlation structures are straightforward. Finally, parameter estimation via maximum likelihood estimation (MLE) may require computational methods, such as numerical integration or Markov chain Monte Carlo (MCMC) techniques.

Due to the occasional non-convergence with the standard likelihood method, Chen et al. (2017) proposed a composite likelihood (CL) function that uses an

independent working assumption between sensitivity and specificity. The method specifies a pseudo-likelihood for sensitivity and specificity based on the marginal distributions. Equation 3.3 defines the pseudo-likelihood, and $\log L_B(\theta_B)$ is defined similarly, shown below

$$\log L_p(\theta_A, \theta_B) = \log L_A(\theta_A) + \log L_B(\theta_B), \tag{3.3}$$

where

$$\begin{aligned} \log L_A(\theta_A) &= \sum_{i=1}^{m} \log P(n_{i,11}|n_{i,1};\theta_A) \\ &= \sum_{i=1}^{m} \left\{ \log \int \text{Bin}(n_{i,11}|n_{i,1}, \text{Se}_i)\ \phi(\text{Se}_i;\theta_A)\ d\,\text{Se}_i \right\}. \end{aligned} \tag{3.4}$$

The authors note that approximation errors decrease in this method as only one-dimensional integrals are involved in the calculation. This method also relies on the marginal normality of the logit sensitivity and logit specificity, allowing the estimation to be more robust to the misspecification of the joint distribution assumption. Nikoloulopoulos (2018) compared CL versus MLE methods, and found that the CL method is nearly as efficient as the MLE method. Neither estimation method is robust to marginal distribution misspecification. The CL method proposed by Chen et al. (2017) will always converge because the proposed pseudolikelihood has a closed form.

Nonparametric Meta-Analysis for Diagnostic Accuracy Studies. Zapf et al. (2015) proposed a non-parametric method for meta-analysis. The authors assume fixed effects only, using a vector of individual test results that is a multivariate

Bernoulli distribution

$$(\mathbf{X}'_{i0}, \mathbf{X}'_{i1}) = (X_{i01}, \ldots, X_{i0n_{i0}}, X_{i11}, \ldots, X_{i1n_{i1}}). \tag{3.5}$$

This format is based on the unified, nonparametric approach for sensitivity, specificity, and ROC curves from Lange and Brunner (2012). Overall sensitivity and specificity are given as

$$\begin{aligned} \widehat{\text{Se}} &= \frac{1}{n_1} \sum_{i=1}^{I} \sum_{s=1}^{n_{i1}} X_{i1s} \\ \widehat{\text{Sp}} &= \frac{1}{n_0} \sum_{i=1}^{I} \sum_{s=1}^{n_{i0}} (1 - X_{i0s}), \end{aligned} \tag{3.6}$$

where 1 indicates "diseased" and 0 indicates "non-diseased." Then, a multivariate normal (MVN) distribution is defined from the overall Se and Sp, using asymptotic theory, as shown below

$$\sqrt{I} \left[\begin{pmatrix} \widehat{\text{Se}} \\ \widehat{\text{Sp}} \end{pmatrix} - \begin{pmatrix} \text{Se} \\ \text{Sp} \end{pmatrix} \right] \sim \text{MVN}(\mathbf{0}, \mathbf{V}), \tag{3.7}$$

where

$$\mathbf{V} = \text{Cov} \left(\sqrt{I} \left[\begin{pmatrix} \widehat{\text{Se}} \\ \widehat{\text{Sp}} \end{pmatrix} - \begin{pmatrix} \text{Se} \\ \text{Sp} \end{pmatrix} \right] \right).$$

The covariance matrix is estimated by the following unbiased estimator

$$\widehat{\mathbf{V}} = \frac{I^2}{I-1} \sum_{i=1}^{I} (\mathbf{Y}_i - \mathbf{S}_i) \cdot (\mathbf{Y}_i - \mathbf{S}_i)', \tag{3.8}$$

where

$$\mathbf{Y}_i = \left(\frac{\mathrm{TP}_i}{n_1}, \frac{\mathrm{TN}_i}{n_0} \right) \tag{3.9}$$
$$\mathbf{S}_i = \left(\frac{n_{i1}}{n_1^2} \cdot \mathrm{TP}, \frac{n_{i0}}{n_0^2} \cdot \mathrm{TN} \right).$$

TP and TN are the total counts across individual tests of TPs and TNs. No assumptions are made regarding the distribution of the data or the correlation structure. But the model assumes homogeneity of sensitivities and specificities across studies, and the method does not yet have a way to include covariates.

Quadrivariate Logistic Regression Model. Hoyer and Kuss (2016) proposed the quadrivariate logistic regression model to compare different diagnostic tests via meta-analysis. For EWID data, researchers seek to compare different lineup procedures to determine the diagnostically superior one. This methodology could work well in the EWID paradigm. Each study reports two four-fold tables with TP_{ij}, TN_{ij}, FP_{ij}, and FN_{ij} for the i-th study and the j-th diagnostic test, $j = 1, 2$. The TPs and TNs are still assumed to binomially distributed,

$$\mathrm{TP}_{ij} \mid \mathrm{Se}_{ij} \sim \mathrm{Bin}(\mathrm{TP}_{ij} + \mathrm{FN}_{ij}, \mathrm{Se}_{ij}) \tag{3.10}$$
$$\mathrm{TN}_{ij} \mid \mathrm{Sp}_{ij} \sim \mathrm{Bin}(\mathrm{TN}_{ij} + \mathrm{FP}_{ij}, \mathrm{Sp}_{ij}).$$

The models for the logit transformations of sensitivity and specificity are additive in two effects: an effect, μ_j and ν_j, respectively, for the method j, and a random effect, ϕ_{ij} and ψ_{ij}, respectively; viz., $\mathrm{logit}(\mathrm{Se}_{ij}) = \mu_j + \phi_{ij}$ and $\mathrm{logit}(\mathrm{Sp}_{ij}) = \nu_j + \psi_{ij}$. Four random effects ϕ_{i1}, ψ_{i1}, ϕ_{i2}, and ψ_{i2} are assumed to follow a quadrivariate normal

distribution, such as

$$\begin{pmatrix} \phi_{i1} \\ \psi_{i1} \\ \phi_{i2} \\ \psi_{i2} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\phi_{i1}} & \rho_{\phi_1\psi_1}\sigma_{\phi_1}\sigma_{\psi_1} & \rho_{\phi_1\phi_2}\sigma_{\phi_1}\sigma_{\phi_2} & \rho_{\phi_1\psi_2}\sigma_{\phi_1}\sigma_{\psi_2} \\ & \sigma^2_{\psi_1} & \rho_{\psi_1\phi_2}\sigma_{\psi_1}\sigma_{\phi_2} & \rho_{\psi_1\psi_2}\sigma_{\psi_1}\sigma_{\psi_2} \\ & & \sigma^2_{\phi_2} & \rho_{\phi_2\psi_2}\sigma_{\phi_2}\sigma_{\psi_2} \\ & & & \sigma^2_{\psi_2} \end{pmatrix} \right]. \quad (3.11)$$

The model captures the potential between-study heterogeneity of sensitivities and specificities, as well as the corresponding correlation among the random effects. The main parameters of interest are the differences of sensitivities and specificities between the meta-analyses. The difference in the logistic transformations (i.e., inverse logit) of sensitivity and specificity between the two studies provide the following formula for the parameter of interest,

$$\begin{aligned} \Delta\text{Se} &= \frac{\exp(\hat{\mu}_1)}{1+\exp(\hat{\mu}_1)} - \frac{\exp(\hat{\mu}_2)}{1+\exp(\hat{\mu}_2)} \\ \Delta\text{Sp} &= \frac{\exp(\hat{\nu}_1)}{1+\exp(\hat{\nu}_1)} - \frac{\exp(\hat{\nu}_2)}{1+\exp(\hat{\nu}_2)}. \end{aligned} \quad (3.12)$$

Similarly, Dimou et al. (2016) proposed a multivariate method for the meta-analytic comparison of diagnostic tests. It is an extension of the bivariate model for the comparison of two or more tests.

3.2 Supervised Learning Classification Methods

The classification problem has a long history in the statistical literature; it has reappeared in the machine learning field as "supervised learning" but the goal is the same: create "rules" by which to categorize new observations into groups. We provide a brief overview of some common classification methods as potential models for

eyewitness identification accuracy. The algorithms result in predicted decisions that is compared to the underlying truth and the influential predictors for the decisions. The goal is to minimize all types of errors. The resulting model can be adjusted by changing the thresholds of errors, depending on which error is considered more grievous. Once the model has been trained properly under the supervised learning framework, and validated with representative test data, it can be applied to real world data.

The difference between the methods mentioned in this section and the methods mentioned in the previous sections is the lack of a meta-analytic framework. At this time, the described methods cannot accommodate the meta-analytic framework. Some researchers are exploring methods of integrating machine learning algorithms to aid in study selection and data extraction for systematic reviews and meta-analysis. Methods have not yet been developed for computational purposes.

In classification methods, point estimates, which can be characterized by finding variance estimates using simulation and/or repetition, are obtained per data set. The true value in classification methods is how easily they are applied, which could be helpful for law enforcement agents, lawyers, and jurors. How well these models work in practice is yet unknown, but can be determined through simulation or application to real data sets.

Classification Models

Some common classification methods include linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), boosted logistic regression (in addition the standard logistic regression), decision trees, random forests, graphical models via Bayesian networks, support vector machines (SVM), and neural networks. Brief

descriptions of these methods, as well as graphical approaches, are provided in the following sections; see also *The Elements of Statistical Machine Learning, 2nd Edition* by Hastie et al. (2013) for in-depth discussions on all methods. Some of these methods (SVMs, random forests, and neural networks) suffer from "black box" syndrome, where the the results are not necessarily interpretable due to injected randomness, etc. Machine learning researchers have developed methodologies to mitigate this issue, which is beyond the scope of this chapter. These methodologies include the partial dependence plots (PDP) from Friedman (1991), local interpretable model-agnostic explanations (LIME) from Ribeiro et al. (2016), and Shapley additive explanations (SHAP) from Lundberg and Lee (2017).

Discriminant Analysis. LDA and QDA are conventional classification methods proposed by Fisher (1936) that use linear and quadratic decision boundaries, respectively, in the space spanned by the covariates that influence the outcome. In the framework of EWID, the outcome is "accuracy," using vectors of covariates to predict eyewitness's decisions. The choice between LDA and QDA depends heavily on the structure and amount of data, and the assumption of normally-distributed covariates; QDA for an underlying linear model results in highly biased predictions.

(Boosted) Logistic Regression. Logistic regression has been considered in EWID research (see Section 2.4.3), but not to a great extent (Wetmore et al., 2015; Andersen et al., 2014). Logistic regression assumes a binary response variable and one or multiple continuous and/or categorical independent variables x, which lends itself pretty easily to the EWID paradigm. However, logistic regression does not show discriminability from the contribution of response bias, since the "correct and false identifications are analyzed separately" (Gronlund and Neuschatz, 2014). Let

$\pi(\boldsymbol{x}) = P(Y = 1)$ for the binary response Y and $\boldsymbol{x} = (x_1, \ldots, x_k)$ of k predictors β. The logistic regression model is defined as

$$g(\boldsymbol{x} \mid \beta) = \log \frac{\mu(\boldsymbol{x} \mid \beta)}{1 - \mu(\boldsymbol{x} \mid \beta)} = \beta^T \boldsymbol{x} \tag{3.13}$$

$$\pi(\boldsymbol{x}) = \mu(\boldsymbol{x} \mid \beta) = \frac{\exp(\beta^T \boldsymbol{x})}{1 + \exp(\beta^T \boldsymbol{x})}. \tag{3.14}$$

The estimated coefficients in a logistic regression provide the effect of x_j on the log odds that $Y = 1$, adjusting for the other x_i. The change in odds would predict for accuracy and the variables selected in the model are the informative variables of discriminability. The predicted accuracy from the fitted model is put into a contingency table with the observed accuracy, providing a measure for model performance. Beyond the relationship of response and predictor variables, predictive modeling such as logistic regression would allow researchers to generalize models to new cases via extrapolation. For EWID analysis, covariates will be added in to account for differences in probability for a correct or incorrect identification. Logistic regression requires little to no multicollinearity among the covariates, which means it requires independent covariates. Given the framework of EWID, it does not seem possible to perform a one-level logistic regression, but might require a a more complicated model. As discussed, the nature of EWID data does not lend itself well to such binarization. Researchers are interested in understanding what causes a witness to choose the true perpetrator, to choose an innocent suspect, to choose a filler, or to not choose at all. Multinomial logistic regression or some other multi-class classification method may work better in this respect.

Multinomial logistic regression models have been well-studied in general, but have not seen much (if any) use in the EWID field. In general, polytomous classification models are not frequently used or even considered in the analysis of EWID

data. This stems as a result of how the data is perceived, which is primarily due to the extensive use of ROC curve analysis. Multinomial logistic regression models can be viewed as a set of $K-1$ independent binary regressions, shown in Equation 3.16. One outcome is chosen as a baseline, and the other outcomes are separately regressed on this baseline. Let $1 \leq J \leq K-1$ and Y_i, $\boldsymbol{x}_i$ represent the i-th set of response and predictor observations. The multinomial logistic regression model is defined as

$$g(\boldsymbol{x}_i \mid \beta_J) = \log \frac{P(Y_i = J)}{P(Y_i = K)} = \beta_J^T \boldsymbol{x}_i \tag{3.15}$$

$$P(Y_i = J) = \frac{\exp(\beta_J^T \boldsymbol{x}_i)}{1 + \sum_{k=1}^{J} \exp(\beta_k^T \boldsymbol{x}_i)}. \tag{3.16}$$

"Boosting," which was originally proposed by Schapire (1990), makes logistic regression more powerful. The idea was further adapted to gradient boosting machines by Friedman et al. (2000). Boosting combines the performance of many "weak" classifiers to produce a more powerful "committee." For EWID analysis, covariates are added to account for differences in probability for a correct or incorrect identification. For more on boosting, see Hastie et al. (2013).

Decision Trees and Random Forests. Decision or classification trees provide the foundation for random forests. The goal of decision trees is to create a model that predicts a value of a target variable based on several covariates. Nodes on the tree are the decision points that provide the path for the particular datum considered. Decision trees are simple to understand and easy to interpret. Classification trees are the individual units of random forests. Given the data, the covariates will be used as splitting variables to branch the data into sorted clusters. The splits are determined based on the homogeneity of observations in the resulting child nodes from the parent node. The resulting terminal nodes will be the decision determined

by classification and regression tree (CART) algorithm (Breiman et al., 1984).

Random forests are an ensemble classifier based on decision trees. Votes arise from groups of decision trees. Tree bagging (bootstrap aggregating) draws repeated samples from the original data. Each sample is drawn randomly with replacement, and creates a classification tree. One generates M such trees. When one wants to classify a new observation, one uses each of the M trees in the "forest" (collection of de-correlated trees) and uses majority (or plurality) rule to assign the classification. This decreases the variance in the model. Random forests are also generated using feature bagging, where random samples of covariates are used for each tree rather than the entire set of covariates. For each candidate (observation), a random subset of features is obtained. An observation is classified by majority vote from all the trees. Explaining the concept of a random forest can be done using visualizations. Further exploration of random forests is pursued in Chapter 5.

Support Vector Machines. Similar to other supervised learning algorithms, SVMs take as input the covariates for EWID to build the model based on training data. SVMs construct a hyperplane that is used to separate the data. A high-dimensional divider classifies the data into groups based on the interaction of several covariates. SVMs rely to classifying using hyperplanes (i.e., some sort of separator) in high dimensions, depending on the number of included covariates. Conveying this concept of high-dimensionality to laypeople may be difficult, which may affect its use in EWID and law enforcement settings. While SVMs can be effective and accurate in prediction in some circumstances, both the SVM algorithm and the output are difficult to interpret, making SVMs possibly problematic for a court setting.

Neural Networks. Neural networks is a black box method that uses layers or neurons $p_j(t)$, which receive input. These neurons then change their internal state (activation) $a_j(t)$ based on that input, and produces output. Some threshold θ_j determines activation, which is an input to some activation function

$$a_j(t+1) = f[a_j(t), p_j(t), \theta_j]. \tag{3.17}$$

The output function is expressed as

$$o_j(t) = f_{out}[a_j(t)]. \tag{3.18}$$

The network is formed by the connection of several of these neurons. Neural networks are flexible and can model a variety of functional forms, making it useful for complex and/or abstract problems. Like other machine learning algorithms, neural networks require training and computational resources. The covariates in an EWID experiment are used to determine the hidden units of the neural network, which are processed by the output function, resulting in a decision for each person. The decision from the algorithm for each person is then compared to the person's actual outcome.

Graphical Models. Graphical models, used in other forensic analysis, are also useful for the EWID paradigm (Dawid and Mortera, 2017). Luby (2016) explored this approach with log-linear analysis. In this model, the data are in the form of a multi-way table with TIP/TIA (two levels) $\times$ eyewitness (EW) Decision (two levels) $\times$ ECL (five or more levels) $\times$ Witness instructions (two or more levels); additional variables can be included without changing the theoretical foundation for the analysis. The model is fit iteratively to find the expected counts for each cell

using a training set of data. Based on the experiment and corresponding data, we generate different graphical models as follows. Let α represent the main effects, β represent the two-way interactions, subscript wc represent witness choice, subscript t represent target absence or presence, i represent witness instructions, and c represent ECL. Equation 3.19 shows an example of a fitted model. The model includes system and estimator variables, previously discussed in Section 2.2,

$$\log m_{wc,t,i,c} = \alpha_{wc} + \alpha_t + \alpha_i + \alpha_c + \alpha_e + \beta_{wc,t} + \beta_{wc,i} + \beta_{wc,c} + \beta_{e,i} + \beta_{c,i}. \quad (3.19)$$

Garbolino (2016) discusses the use of Bayesian networks for evaluating testimony; Garbolino's model is actually very general, and applies to testimony of any kind, not just from an eyewitness. The proposed model assumes that the witness is accurate, objective, and truthful. Each of these characteristics corresponds to an inference about the witness' personality:

(1) Senses give evidence of what is seen;

(2) Belief in the evidence from the senses; and

(3) Belief in what is said.

In the end, Garbolino (2016) proposes an object-oriented Bayesian network class for the analysis of the reliability of human witnesses. D'Agostini (2016) notes that Bayesian networks are a technical tool, but their true value is as a very powerful conceptual tool that can handle complex problems with variables related by both probabilistic and causal links. Even with subjective probability (i.e., eyewitness testimony), the intuitive idea of probability is recovered.

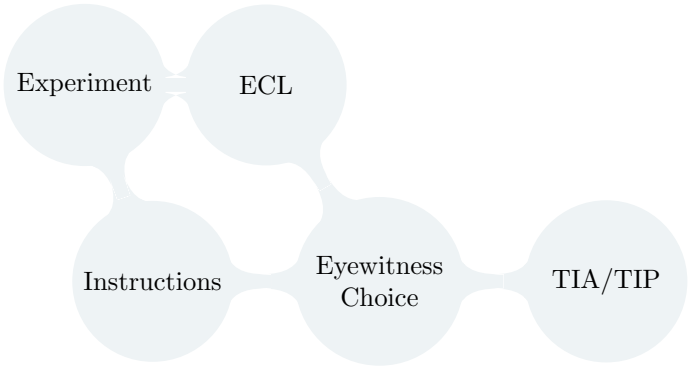


Figure 3.1: This is the graphical model corresponding to log-linear model in Equation 3.19.

3.3 Tools Based on ROC Methods

The popularity of the ECL-based ROC curve to compare lineup procedures, together with its limitations (see Section 2.4.2), leads us to consider other methods that augment and improve upon ROC curves for a more complete comparison between methods.

We discuss the predictive receiver operating characteristic (PROC) curve (which utilizes PPV and NPV in a similar way that HR and FAR are used in ROC curves), multivariate ROC curves, and AUC estimation for these curves. We also discuss the inclusion of variability measures for ROC curves that could also be adapted for the PROC curve and multivariate ROC curves.

3.3.1 Methodology Development

The National Research Council (2014) report called for a broader "exploration of the merits of different statistical tools for use in the evaluation of eyewitness performance" as an important area of research.

The analysis of EWID experimental data should consider three aspects:

(1) *Sensitivity*: the probability that an eyewitness correctly identifies the true

perpetrator given that the perpetrator is present in the lineup;

(2) *Specificity*: the probability that an eyewitness correctly chooses "Not Present" given that the perpetrator is not in the lineup; and

(3) *Prevalence*: the proportion of individuals who might be the culprit.

PPV and NPV are functions of all three factors, so accurate estimation of all three quantities is essential (Kafadar, 2015).

PPV is the probability that, when the eyewitness makes an identification, the identified person is truly the perpetrator. Similarly, NPV is the probability that, when the eyewitness fails to identify a person as a perpetrator, that person was truly not the perpetrator. In general, and in real life, we do not know if the eyewitness's decision is correct, but we can estimate the probability (PPV, NPV). PPV can be rewritten in terms of the odds ratio (OR) and positive likelihood ratio (LR_+) as follows. Let Se and Sp denote sensitivity and specificity, respectively, and p the probability that the suspect is truly the perpetrator. For example, in a lineup with six photos, p could be equal to $1/6$. Define PPV as the conditional probability that the identification was correct, given that the eyewitness selected a person from the lineup:

$$\begin{aligned} \mathrm{PPV} &= \frac{\#\ \mathrm{TP}}{\#\ \mathrm{TP} + \#\ \mathrm{FP}} \\ &= \frac{\mathrm{Se} \cdot p}{\mathrm{Se} \cdot p + (1 - \mathrm{Sp}) \cdot (1 - p)} \\ &= \frac{1}{1 + \frac{1}{\mathrm{OR} \cdot \mathrm{LR}_+}}, \end{aligned} \tag{3.20}$$

where $\mathrm{OR} = \frac{p}{1-p}$ denotes the ratio of probabilities that a suspect P is guilty versus

is innocent, and LR_+ is the likelihood ratio of a positive call:

$$\begin{aligned}\mathrm{LR}_+ &= \frac{P\{\text{eyewitness selects } P \mid P \text{ is perpetrator}\}}{P\{\text{eyewitness selects } P \mid P \text{ is innocent}\}} \\ &= \frac{\mathrm{Se}}{1-\mathrm{Sp}} \\ &= \frac{\mathrm{HR}}{\mathrm{FAR}} \\ &= \mathrm{DR}.\end{aligned} \tag{3.21}$$

Since LR_+ can be written as $\frac{\mathrm{Se}}{1-\mathrm{Sp}} = \frac{\mathrm{HR}}{\mathrm{FAR}}$, it is equivalent to the DR; as discussed in Section 2.4.1. Note also that the slope of the ECL-based ROC curve at ECL level c is the DR (LR_+) for those persons who expressed confidence of at least c. A higher LR_+ leads to a higher PPV for the same prevalence. Thus, if the probability that the guilty suspect is in the lineup (i.e., population under consideration), then the lineup procedure with the higher DR yields a higher PPV. PPV is more affected by specificity.

NPV is the probability that the excluded person is truly not the perpetrator. NPV can also be written in terms of OR and the negative likelihood ratio LR_-, or the likelihood ratio of a negative call. Similar to how a higher LR_+ results in a higher PPV, a lower LR_- would result in a higher NPV, given the same prevalence. We define NPV as

$$\begin{aligned}\mathrm{NPV} &= \frac{\#\ \mathrm{TN}}{\#\ \mathrm{TN} + \#\ \mathrm{FN}} \\ &= \frac{\mathrm{Sp}\cdot(1-p)}{(1-\mathrm{Se})\cdot p + \mathrm{Sp}\cdot(1-p)} \\ &= \frac{1}{1+(\mathrm{OR}\cdot\mathrm{LR}_-)}.\end{aligned} \tag{3.22}$$

NPV is normally not considered, as most EWID researchers, practitioners, and poli-

cymakers are less concerned with the probabilities associated with choosing innocent foils (Amendola and Wixted, 2014). We define LR_- as

$$\mathrm{LR}_- = \frac{1 - \mathrm{Se}}{\mathrm{Sp}} = \frac{1 - \mathrm{HR}}{1 - \mathrm{FAR}}. \tag{3.23}$$

Provided the conditions for comparing two lineup procedures are the same (e.g., OR is the same), then procedure one is preferred over procedure two if PPV_1 (PPV for procedure one) is greater than PPV_2 (PPV for procedure two). This is true if and only if

$$\frac{\mathrm{OR}_1}{\mathrm{DR}_1} < \frac{\mathrm{OR}_2}{\mathrm{DR}_2} \equiv \frac{\mathrm{DR}_1}{\mathrm{OR}_1} > \frac{\mathrm{DR}_2}{\mathrm{OR}_2}. \tag{3.24}$$

Using PPV (i.e., LR_+) as the criterion, procedure one is preferred over procedure two if

$$\mathrm{DR}_1 > \mathrm{DR}_2 \equiv (\mathrm{LR}_+)_1 > (\mathrm{LR}_+)_2. \tag{3.25}$$

Similarly, for NPV, method one is preferred if

$$\frac{1}{(\mathrm{LR}_-)_1} > \frac{1}{(\mathrm{LR}_-)_2}. \tag{3.26}$$

Thus, both LR_+ and LR_- need to be be considered when choosing "optimal" procedures. In the EWID paradigm, a vector of match-to-the-witness's-memory values (i.e., memory strength) for $n - 1$ alternatives with a lineup size of n between the eyewitness's memory of the perpetrator and the lineup member could be used in conjunction with the NPV (Clark, 2005). This emulates the framework for ROC curves in the SDT model. Since the PROC curve is an extension of the ROC curve, we can use some of the same ideas.

Some EWID researchers state that as responding becomes more conservative,

both LR_+ and LR_- increase, suggesting these values depict the tradeoff related to liberal versus conservative responding, not discriminability (Mickes et al., 2017). LR_+ and LR_- are functions of only sensitivity and specificity. The targeted values in EWID accuracy experiments are PPV and NPV; hence they need to be jointly considered also in the analysis of identification accuracy.

3.3.2 Predictive ROC Curve

Shiu and Gatsonis (2008) proposed a way of displaying PPV and NPV jointly via a PROC curve. The PROC curve is defined as $\{1 - \text{NPV}(c),\ \text{PPV}(c)\}$ for $c \in R$, where R is the set of all possible thresholds for test positivity. This curve is affected by prevalence p. Specifically speaking, PPV increases and NPV decreases when prevalence increases. Thus, with increasing prevalence, a point on the PROC curve will move towards the upper-right direction.

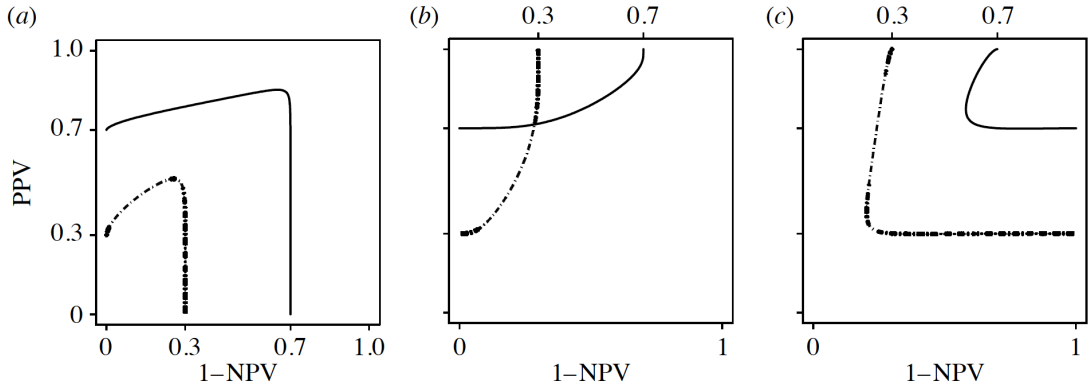


Figure 3.2: This plot shows the predictive curves with $a = 0.8$: (a) $b = 0.7$, (b) $b = 1$, (c) $b = 1.5$. The solid line represents high prevalence ($p = 0.7$) and the dot-dashed line represents low prevalence ($p = 0.3$) (Shiu and Gatsonis, 2008).

The PROC curve lacks monotonicity, which occurs if a one-to-one correspondence between PPV and NPV exists. The criteria for monotonicity is established using hazard rate order, reverse hazard rate order, and likelihood ratio order. The

likelihood ratio order says the ratio $\frac{f(c)}{g(c)}$ is a monotonic function of c; this is a sufficient condition for monotonicity of the PROC curve. But the monotonicity properties are complex in certain cases. It seems that in the binormal case, if the scaling parameter $b = 1$ for the binormal model, then there is an obvious trade-off between PPV and NPV, and the PROC curve is monotone. For $b \neq 1$, monotonicity is guaranteed for only certain segments along the curve.

Figure 3.2 demonstrates this complicated pattern of monotonicity. The middle plot shows the clear monotonicity – an increase in PPV has a corresponding decrease in NPV. For the other two plots, overlap is visible depending on the location along the curve. Figure 3.3 shows this phenomenon, where monotonicity is defined on $(-\infty,\ c^*_{\text{PPV}})\,, [c^*_{\text{PPV}},\ c^*_{\text{NPV}}]\,,$ and $(c^*_{\text{NPV}},\ \infty)$. The visually vertical and horizontal lines in this figure result from either PPV or NPV converging faster than the other.

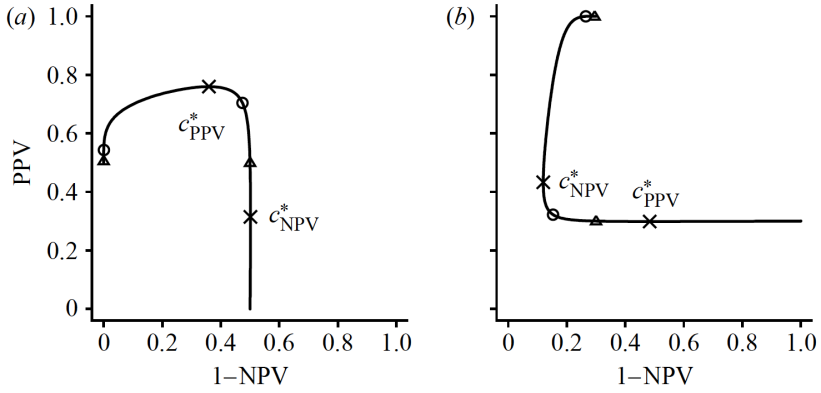


Figure 3.3: Predictive curves with (a) $a = 1$, $b = 0.5$, $p = 0.5$; (b) $a = 2$, $b = 2$, $p = 0.3$. Circles denote points corresponding to operating thresholds at -1 and $a+b$, triangles denote points corresponding to operating thresholds at -2 and $a + 2b$, and crosses denote c^*_{PPV} and c^*_{NPV} (Shiu and Gatsonis, 2008)

3.3.3 Multivariate ROC Curves

The standard ROC curve cannot account for more than two covariates. In 2009, Jin and Lu (2009) proposed the ROC region, which plots the TPR over the FPR for all possible choices of the decision thresholds for two continuous covariates. The thresholds arise from a tree-based nonlinear combination of multiple predictors. Wang and Li (2012) proposed a bivariate ROC and a bivariate weighted receiver operating characteristic (WROC) for biomarker measurements. The authors defined a bivariate ROC as a conditional expectation of TP as a function of the two continuous biomarkers given the FP as a function of the two biomarkers. Let $0 \leq q \leq 1$, (U_0, V_0) be a pair of bivariate markers from a non-diseased group,

$$\mathrm{ROC}(q) = E\big[\mathrm{TP}(U_0, V_0) \mid \mathrm{FP}(U_0, V_0) = q\big]. \tag{3.27}$$

The authors further defined WROC as the unconditional expectation of TP as a function of the two continuous biomarkers given the FP as a function of the two biomarkers. This idea was extended in 2013 to multivariate biomarkers (Wang and Li, 2013). In the multivariate markers extension, the decision thresholds for the continuous biomarkers were decided by classification tree-based methods. Another similar method, proposed by Pundir and Amala (2015), is the bivariate lognormal ROC curve for detecting the accuracy of two biomarkers. The WROC is a plot of the TPRs and FPRs as functions of the two thresholds from the biomarkers. We can possibly adapt these multivariate ROC methods to both continuous and categorical predictors, and apply them to EWID contexts.

Hong (2012) proposed a bivariate ROC model, which assumes a sliced bivariate normal distribution function for two predictors, X_1 and X_2. In the method, $X_2 = h(X_1)$ by using some linear function that passes through the mean vector of the

X_2 pseudo-random variable. The points for the ROC come from the cumulative distribution function defined for the ROC curve. Hong and Jeong (2012) proposed an optimal classification function for this bivariate ROC curve.[3]

3.3.4 AUC Estimation

The AUC for any one of the bivariate ROC curves can be modeled as a function of eyewitness, and lineup procedure, and any other variables at play, using a hierarchical model, similar to the model proposed by Wang and Gatsonis (2008). In that article, the authors propose a hierarchical model for multi-reader, multi-modality[4] studies in diagnostic medicine. Heterogeneity is introduced at the first level of the hierarchy. Effects for some covariates, such as reader variability, are treated as random (not fixed), and MCMC can be used to estimate model parameters. In the model, three levels (i.e., types) of correlation exist:

(1) Within-reader variability;

(2) Between-reader variability; and

(3) Variation of hyperparameters.

The within-reader variability represents the correlation due to readers between AUC estimates for a reader in two modalities. Let $\boldsymbol{Z}_j$ represent the reader level covariates and β represent the reader random effects. The authors assume correlation r_{1j} is common across all readers. Letting γ be a vector of regression coefficients, following independent normal prior distributions with mean zero and large variances,

[3]Both of these papers, Hong (2012) and Hong and Jeong (2012), are in Korean; their abstracts are in English.

[4]A modality is a particular diagnostic procedure.

the authors model the within-reader variability as

$$\begin{pmatrix} y_{1,j} \\ y_{2,j} \end{pmatrix} \Bigg| \, \boldsymbol{\beta}_j \sim N \left[\begin{pmatrix} \mu_{1,j} \\ \mu_{2,j} \end{pmatrix}, \Sigma_j \right], \tag{3.28}$$

where

$$\Sigma_j = \begin{bmatrix} f(\mu_{1j}) & r_{1j}\sqrt{f(\mu_{1j})f(\mu_{2j})} \\ r_{1j}\sqrt{f(\mu_{1j})f(\mu_{2j})} & f(\mu_{2j}) \end{bmatrix} \tag{3.29}$$

and

$$\boldsymbol{\mu}_j = \begin{pmatrix} \mu_{1j} \\ \mu_{2j} \end{pmatrix} = g\left(\boldsymbol{Z}_j^T \boldsymbol{\gamma} + \boldsymbol{\beta}_j\right). \tag{3.30}$$

The between-reader variability represents the correlation from two different readers in the same modality. This model assumes that the random effects follow normal distributions, where $\beta_{1j} \mid \sigma_1^2 \sim N(0, \sigma_1^2)$, $\beta_{2j} \mid \sigma_2^2 \sim N(0, \sigma_2^2)$, and $\boldsymbol{\beta}_j = (\beta_{1j}, \beta_{2j})^T$. The variation of hyperparameters represents the correlation due to cases between any two AUC estimates since the estimates arise from the same set of subjects. We also assume proper prior distributions on the hyperparameters, such that σ_1^2 has an inverse gamma distribution with parameters a_1 and b_1 and, similarly, σ_2^2 has an inverse gamma distribution with parameters a_2 and b_2. Additionally, r_{1j} has a beta distribution with parameters c_j and d_j.

We can easily extend the model to include additional covariates. The approach may be a more efficient alternative to stratified analyses, and it also is sufficiently flexible to accommodate complex correlation structures. Finally, the model fitting can be done on publicly available software.

Lange and Brunner (2012) proposed a unified, nonparametric approach to multi-reader diagnostic trials based on ranks, which allows them to estimate the AUC as a vector for different modalities. The authors suggest the approach is equivalent to a factorial chi-squared test on repeated measures. In the factorial design, the reader (i.e., eyewitness) and modality (i.e., lineup procedure) are the two factors. They use a nonparametric approach to show that sensitivity and specificity are areas under particular ROC curves. Additional nonparametric estimations for AUC could also be considered.

3.3.5 Confidence Intervals for ROC

Based on the considerations of previous work done with ROC curves in the EWID paradigm, variability is an increasingly important consideration for more robust conclusions. We continue the discussion of variability in EWID data and for confidence-based ROC curves. These methods acknowledge that the calculated DR values for ROC curves are point estimates at the center of an interval that captures the true population DR value. The methods presented below are possible ways to incorporate measures of variability.

Point-wise confidence intervals for ROC curves are the intervals of sensitivity at a given value of specificity. We can construct confidence bands for a range of specificity or for the entire ROC curve (Yin, 2014). Some EWID researchers have used bootstrap resampling to estimate standard errors for their ROC curves (Mickes et al., 2017). Luby (2017) used confidence boxes for the HR and FAR. Confidence bands are routinely calculated for medical applications, and should be used for EWID applications also.[5]

Macskassy and Provost (2008) provide an empirical study of methods for the

[5]See examples in Appendix D of the National Research Council (2014) report.

estimation of confidence bands. These methods include vertical averaging (VA), threshold averaging (THA), simultaneous joint confidence regions (SJR), Working-Hotelling based bands (WHB), and fixed-width confidence bands (FWB). These rely on sweep methodology, which samples the observed ROC point and the confidence boundary around it to generate upper and lower confidence bands.

Method	**Description**
VA	This is a sweep method that looks at successive FP rates and averages TPs for multiple bootstrapped ROC curves at a specific FP rate.
THA	This is a sweep method that freezes the threshold of the test rather than the FP rate, by identifying the set of ROC points that would be generated using a particular threshold on each of multiple ROC curves.
SJR	This method utilizes the Kolmogorov-Smirnov (KS) one-sample test statistic to find the global confidence interval (i.e., simultaneous confidence rectangles) for TP and FP, which are generated by freezing FP to identify the respective TP,
WHB	This method fits a a regression line $y = a - b \cdot x$, of the form $\ell(x, \pm k) = a - b \cdot x \pm k \cdot \sigma(x)$ for $k \geq 0$ and $\sigma(x) = \sqrt{\sigma_a^2 - 2\rho\sigma_a\sigma_b \cdot x + \sigma_b^2 \cdot x^2}$. The line is estimated using MLE. Other estimation methods include coaxial ellipses based on an envelop of a system of ellipses.
FWB	This method identifies a slope $b = -\sqrt{m/n} < 0$, where m and n are sample sizes, along which to displace the original ROC curve to generate confidence bands, and sweeps along the FP axis to identify the TP value at that FP.

Table 3.1: This table provides a summary of ROC curve confidence interval/band estimation methods.

The VA, THA, and point-wise WHB did not translate well to confidence bands, and failed to perform robustly for varied parameters used for the generation of ROC

curves. The authors attribute the failure of VA and THA to the naive methodology. Both the SJR and FWB worked well, and quite robustly, given the data. The SJR does not require any samples to generate the confidence bands, but has a higher variance. The FWB uses the bootstrap to empirically determine the proper displacement for the confidence band generation, but was found to be stable and consistent. Demidenko (2012) introduced parametric confidence bands for the binormal ROC curve, and the ellipse-envelope (EE) confidence band construction based on the Working-Hotelling approach, with variation calculated via the delta method. The EE confidence band has a shorter width than the WHB, under the assumption of a binormal curve. More details of these methods are described in Table 3.1.

3.4 Discussion

Long-standing conventional statistical methodologies, including logistic regression and, more generally, generalized linear models, particularly for bivariate outcomes (sensitivity and specificity), remain valuable and appropriate tools for analyzing EWID experiments, especially when the experiment includes concomitant information, such as environmental variables of the experiment and demographic characteristics of the "eyewitness." In the absence of such information, ROC curves remain a useful comparison of two methods in diagnostic medicine, statistical process control, and eyewitness experiments. Newer approaches from statistical machine learning may be useful with very large experiments, though the impact of specific variables on the outcome may not always be as interpretable as with conventional linear models. Whichever technique is used, proper characterization of the uncertainties associated with the inferences must be calculated.

As recommended by the National Research Council (NRC) report, a broader

"exploration of the merits of different statistical tools for use in the evaluation of eyewitness performance" is an important area of research (National Research Council, 2014). The goal is to encourage the use of more discerning statistical models and analytical methods for assessing EWID procedures used to increase the accuracy of identifications, decrease the number of false convictions, and ensure guilty perpetrators are properly convicted.

The methodologies that we discussed in this chapter provide a foundation for future work, and raise several issue that remain for future research. We have indicated statistical approaches to evaluating current methods that LEOs use in the field (e.g., sequential versus simultaneous lineups, presence or absence of an officer during the eyewitness's deliberations, etc.). We note that procedures from other fields, such as diagnostic medicine, can apply to EWID experimental data, with some modifications as needed. adapted from other fields, such as diagnostic medicine.

Due to the popularity of the ECL-based ROC curve to compare accuracies of EWID procedures, we offered several alternatives related to the quantities used in ROC curves, namely sensitivity and specificity, with a focus on the real targets of interest, PPV and NPV. This exploration led to the following questions:

First, what kind of curve can we use to describe PPV and NPV in an intuitive manner that will also hold up theoretically? Are there multivariate ROC curves that will display comparisons among several procedures simultaneously? How informative is a plot of LR_+ versus LR_- as "proxies" for PPV and NPV. For the classification methods, the determination of effectiveness and accuracy depend on useable and good data that replicate real world scenarios so that our proper assessment of the method's efficacy is valid. Confidence intervals or confidence bands should also always be included with any point estimates.

Second, what role can supervised learning classification methods play in pre-

dicting the accuracy of an eyewitness's decision? In a meta-analytic framework, how can we adapt the established bivariate and/or hierarchical modeling methods to the EWID framework? The answers to these questions require simulations and experimental data with the underlying truth known.

Alternative ways of examining the data could also lead to new modeling procedures or algorithms that would be useful in practice. We proposed a method for estimating the probability of accuracy for eyewitnesses that takes proper account of individuals' probabilities of choosing or not choosing a suspect from a lineup. This method is a potential tool that could provide an in-field assessment of eyewitness reliability, which can be explained to and understood by juries, judges, lawyers, law enforcement officers, and any other non-statisticians working in EWID. Further methods depend on the available types of EWID data, which could include recordings of eyewitness proceedings by working in conjunction with police departments.

Larger, more "ecologically valid" studies, may more properly reflect real-world scenarios by encompassing the realism of the stress, timelines, etc., than lab-based experiments where subjects know they are part of studies. Some ideas include staging a minor crime such as a robbery in a convenience store, and then asking participants if they recall the target. Another idea might be to have participants walk around with a camera on their heads, and ask them to recall targets or faces they may have seen. The camera would provide an objective source of comparison to eyewitness reports.

Researchers who conduct more varied and complex types of experiments will produce sets of observational data (National Research Council, 2014), leading to the development of novel modeling procedures and statistical methods.

We include this chapter as a window for new research for new methodologies and new applications of existing methodologies. Chapter 4 will approach the analysis

of EWID in a completely novel direction by looking at the eyewitness as a multiple classifier.

Chapter 4

Probability of Accuracy: Rethinking the Framework

We propose a new approach to examining EWID data in this chapter. This new approach reconsiders the perceived structure of EWID data, and results in a tool that could potentially be used as an in-field assessment of eyewitness reliability. The tool also proves to be generalizable to other types of data following a similar underlying structure.

4.1 Modeling Eyewitness Accuracy

The eyewitness will see a set number of faces, with a designated target who the police believe is guilty. The eyewitness will be asked to choose the person who they believe is guilty. This probability of choosing has been previously treated as fixed (e.g., 50% probability of choosing versus not choosing). In general, the psychologists who treat this probability of choosing as how conservative or liberal a person is in their response (i.e., response bias). That is, the eyewitness can be influenced to

be more likely or less likely to choose depending on the construction of the lineup (i.e., creating a biased lineup encourages more liberal responses). The choosing rate is more likely also affected by an individual's cost/benefit analysis of making a misidentification. In fact, possibly a more realistic model is that each person has his or her own probability of choosing that may fluctuate depending on the circumstances, regardless of the artificially induced response bias. A probabilistic approach to estimating an individual's predilection to choose has never been used before in EWID research.

We seek to determine the probability of a participant's accuracy given that they choose. The probability of being accurate is a random variable unique to each eyewitness, since each participant has a potentially different rate of accuracy based on various factors, which follows some frequency distribution. The probability of choosing may depend on various covariates (e.g., age or gender of the eyewitness, etc.). We also recognize the possibility that the non-chooser is accurate (i.e., the police failed to include the true perpetrator in the lineup); hence, "no choice" was the correct response. In many current analyses of EWID data, non-chooser responses are ignored. The proposed model uses all of the data, in that it considers the "accuracy" of both choosers and non-choosers. The end goal is to better model EWID accuracy.

To estimate the probability of accuracy, we need to return to the foundation of eyewitness identification – the structure of the data. Thus far, the statistical methodologies have focused on a strictly binary approach to an eyewitness's decision in a lineup. As highlighted in Section 2.2, the data structure consists of five possible decision outcomes that are divided into a group of two and a group of three based on the unknown status of target presence or absence. Recall, there are five possible decision outcomes that occur, only two of which (P_1 and A_1) are correct; see Figure 2.2. The entire probability space for eyewitness decision outcomes clearly

sums to one. The conditional probabilities on target presence or absence for the corresponding eyewitness decision outcomes also sum to one. It is possible to look at any of the possible conditional and/or marginal probability combinations in any order to achieve a probability of one.

This view of the data affords a new understanding and new possible framework for analysis. The data structure clearly lends itself to multiple classification, where the possible decision outcomes serve as the response variable for classification.

Very few multiple classification models have been seen in EWID literature. One popular model is the log-linear (Poisson regression) model, which are GLMs that use a log link in place of a logit link (Luby, 2016). In this model, the data is formatted as a contingency table Luby (2017). For example, the table could be TIP/TIA (two levels) $\times$ EW Decision (two levels) $\times$ ECL (five or more levels) $\times$ Witness instructions (two or more levels). The model can be fit iteratively to find the expected counts for each cell using a training set of data. The log-linear model allows insight into the data that logistic or multinomial regression does not. Log-linear models allow estimation of interactions among variables. Multinomial logistic regression models are more flexible in terms of addition of covariates, continuous and discrete, and are easier to use. Both models are closely related. Other polytomous classification methods can be considered, such as SVMs and random forests. In fact, psychologists working in the field of EWID have begun embracing the integration of machine learning into their analyses. Price et al. (2020) utilized SVMs to predict suspect guilt with predictors based on eye-tracking. The framework being proposed in this chapter extend beyond the possible uses of these methods.

The true value in classification methods is how easily they are applied and interpreted, which could be helpful for law enforcement agents, lawyers, and jurors. How well these models work in practice has not been seen much in publication, but

can be determined through simulation or application to real data sets. We present the proposed framework in a way that enables a non-statistician to follow the logic and understand the model proposed in this thesis.

4.2 Probability of Accuracy

The goal is to determine the probability that a participant responds accurately given that (s)he makes a choice (i.e., is a "chooser"). We treat the probability of being accurate as a random variable, because each participant has a potentially different rate of accuracy based on individual-level factors; a common distribution for this random variable is a beta distribution (restricted to $[0, 1]$ as a probability is), whose mean will depend on the levels of the factors. We have also the probability of choosing, which also is random, because each participant will have a different rate of choosing whose mean also is based on factor levels, which can vary due to random circumstances (e.g., how the person feels that day). For example, the rate that a statistician may choose might be fairly low, around 50%, since the statistician may be less sure of absolution due to statistics knowledge. The rate that a politician who is used to making highly confident statements and choosing with absolution may have a much higher choosing rate of, perhaps, 90%. Then, the combination of accuracy and choosing form a bivariate random variable (X, Y), with some probability distribution.

The probability of choosing can be thought of as a latent random variable, since it is likely unique to each eyewitness. The key is charting a path to the estimation of such value. To do this, we need to decompose the probability of eyewitness accuracy (i.e., the possible eyewitness decision outcomes) into quantifiable components. The probability of eyewitness accuracy will depend on the unknown probability of target

presence or absence, which can be thought of as a global latent variable that is the same for all eyewitnesses under the "same" conditions (e.g., at the same police station).

For the decomposition of this probability, we need to revisit the eyewitness task decision outcome space shown in Figure 2.2. We refer to it as we develop the model for the EWID task. We use the system and estimator variables to estimate an individual's *probability of choosing* dependent upon target presence, which form the components for estimating the probability of eyewitness accuracy.

Suppose $Y_i \in \{0, 1\}$ is the random variable of inaccurate or accurate, respectively, and $T \in \{0, 1\}$ is the random variable of TIP or TIA, respectively. We can decompose the probability of accuracy $P(\text{Accurate})$ into two components, which will depend upon the rate of target presence and the probability of making certain decisions. Let

$$\begin{aligned} P_1 &= P(\text{Choose Target} \mid \text{TIP }) \\ P_2 &= P(\text{Choose Foil} \mid \text{TIP}) \\ P_3 &= P(\text{Don't Choose} \mid \text{TIP}), \end{aligned} \tag{4.1}$$

where $P_1 + P_2 + P_3 = 1$. Further, let

$$\begin{aligned} A_1 &= P(\text{Don't Choose} \mid \text{TIA}) \\ A_2 &= P(\text{Choose Foil} \mid \text{TIA}), \end{aligned} \tag{4.2}$$

where $A_1 + A_2 = 1$. Also, let

$$\theta = P(\text{TIP}) = P(T = 1) \tag{4.3}$$

$$1 - \theta = P(\text{TIA}) = P(T = 0).$$

Then, for the i-th participant, the decompositions are given below,

$$\begin{aligned} P(\text{Accurate}) &= P(Y_i = 1) \\ &= P(Y_i = 1 \ \cap \ T = 1) + P(Y_i = 1 \ \cap \ T = 0) \\ &= P(\text{Choose Target} \ \cap \ T = 1) + P(\text{Don't Choose} \ \cap \ T = 0) \\ &= P(\text{Choose Target} \mid T = 1) \cdot P(T = 1) \\ &\quad + P(\text{Don't Choose} \mid T = 0) \cdot P(T = 0) \\ &= P_{1i} \cdot \theta + A_{1i} \cdot (1 - \theta). \end{aligned} \tag{4.4}$$

From a visual standpoint, we can think of this decomposition as a portion of our entire eyewitness decision outcome space, as shown in Figure 4.1.

Accurate	**Not Accurate**		
Choose Target (P_1)	Choose Foil (P_2)	Do Not Choose (P_3)	**Target Present**
Do Not Choose (A_1)	Choose Foil (A_2)		**Target Absent**

Figure 4.1: A display of the eyewitness decision outcome space, which takes into account the underlying status of the lineup, where we are looking at the decomposition of $P(\text{Accurate})$.

Similarly, we can decompose for $P(\text{Not Accurate})$ using the law of total probability,

$$P(\text{Not Accurate}) = P(Y_i = 0) \tag{4.5}$$

$$
\begin{aligned}
&= P(Y_i = 0 \ \cap \ T = 1) + P(Y_i = 0 \ \cap \ T = 0) \\
&= P(\text{Choose Foil} \ \cap \ T = 1) + P(\text{Don't Choose} \ \cap \ T = 1) \\
&\quad + P(\text{Choose Foil} \ \cap \ T = 0) \\
&= P(\text{Choose Foil} \,|T = 1) \cdot P(T = 1) \\
&\quad + P(\text{Don't Choose} \mid T = 1) \cdot P(T = 1) \\
&\quad + P(\text{Choose Foil} \mid T = 0) \cdot P(T = 0) \\
&= P_{2i} \cdot \theta + P_{3i} \cdot \theta + A_{2i} \cdot (1 - \theta).
\end{aligned}
$$

From these equations, we can see that $P(Y_i = 1)$ depends on P_1, A_1, and θ. We can estimate P_1 and A_1 by fitting any kind of multiple classification model to the subsets of target present and target absent data, respectively, from an ecologically valid, designed experiment. Given that the target is present, we have a multiple classification problem: (1) choose target; (2) choose foil; and (3) don't choose. Given that the target is absent, we have a binary classification problem: (1) choose foil; and (2) don't choose.

4.2.1 Random Forests

We propose to use random forests as the classification model of choice. We chose this model for many reasons, though we would like to note that it is certainly not the only classification model that can be used here.

Comprehensibility. We have found that, in practice, the concept of obtaining predictions in a random forest is easy for laypeople (i.e., people who are not familiar with the random forests or have little to no exposure to random forests) to understand. For example, it easy to output one of the trees fit with the random forest

model, and explain that the splits along the tree to the corresponding nodes are achieved by maximizing class homogeneity in the child nodes. Decision trees place observations to the most commonly occurring class of training observations in the region to which it belongs based on input variables. Random forests create *votes* in many decision trees, creating an ensemble model. New observations are obtained by following these rules established using data with known answers. An example tree is shown in Figure 4.2.

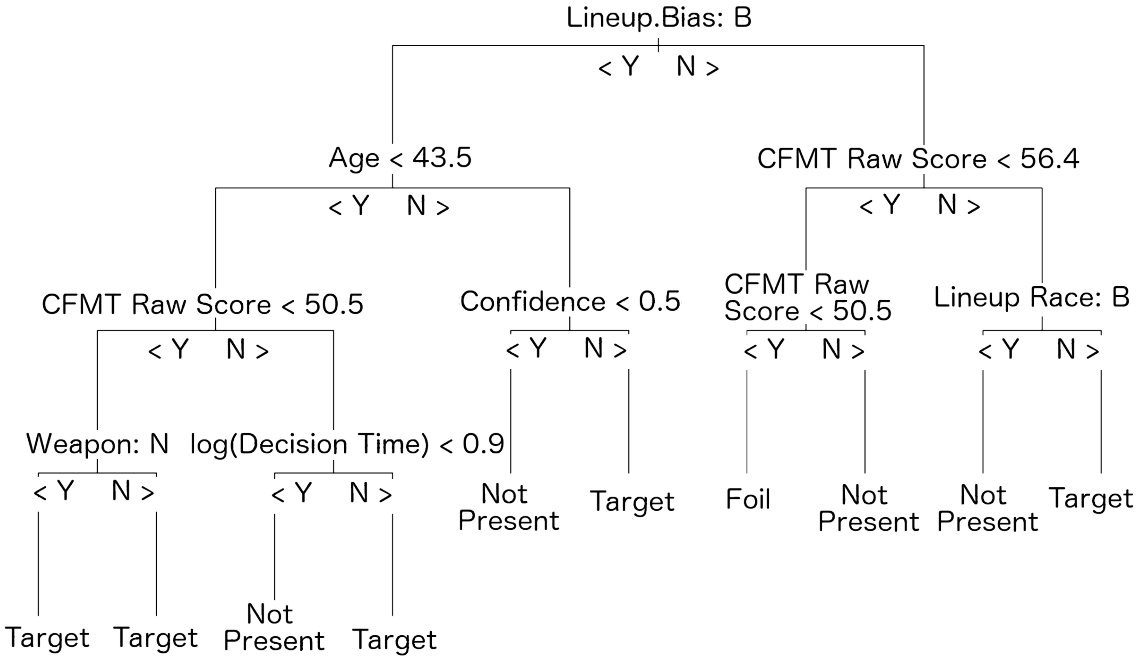


Figure 4.2: Example of a single tree in a random forest based on EWID data.

Classification. Additionally, random forests have good classification performance in general. Depending on the performance criteria (e.g., classification accuracy, mean squared error (MSE), etc.), random forests work as well or better than other classification methods. We can see the relative performance of random forest in terms of classification in comparison to logistic regression and support vector machines in Table 4.1. Random forests outperform the other two methods. Neural

Method	P(Choose Target | TIP)	P(Don't Choose | TIA)
Logistic regression	61.72% (4.07%)	70.89% (4.17%)
Random forests	96.18% (0.9%)	96.58% (1.47%)
Support vector machines	65.61% (3.48%)	58.42% (13.59%)

Table 4.1: Relative accuracy performance using study 1 data from Dodson (unpublished) from logistic regression, random forests, and support vector machines. The mean accuracy is given as the percentage on top and the standard deviation is given as the percentage in parentheses. These values were obtained using 10-fold cross-validation, where 80% of the data was reserved for training and 20% was used for testing. The reported accuracy values are from the testing sets.

networks are not considered, since the complexity of the data did not require the flexibility of a deep learning model and since the logistic regression itself could already be considered a "shallow neural network."

Good classification performance does not necessarily translate to good probability estimates for the relative classes. Class probability estimation requires every quantile to be estimated well, which contrasts with classification in which only the median quantile needs to be estimated well. Random forests are used to obtain probabilities in the literature (Li, 2013). Some researchers use the majority votes across all decision trees in the random forest as the probabilities, while other researchers grow the tree to some node size $k > 1$ to obtain the proportion of observations with class i, which is averaged across all trees via probability machines (Kruppa et al., 2013, 2014; Malley et al., 2012). Researchers are also seeking alternative ways of understanding the class probability estimation process by linking the process to kernel regression methods (Scornet, 2016b; Olson, 2018; Olson and Wyner, 2018). Some of these methods include proximity weighting, by using the concept of a nearest

neighbors to find a weighted average of nodes depending on the distance from some target node.

In this case, we do not seek to necessarily understand the class probability estimation process, but rather would like to use random forests as a means of class probability estimation. In fact, from Section 3.7 in Li (2013), the performance in terms of mean squared loss of the majority vote method versus the other methods of proximity weighting, kernel methods, and probability machines show similar or better performance. Given the occasional marginal increase in performance of other methods, we choose the simplest approach of a majority vote. However, probability estimates obtained using majority vote are not necessarily consistent. Thus, it may be of interest to implement other methods of random forest probability estimation. The performance of the random forests using majority vote seems to work well for the EWID data presented in this chapter, and we expect similar performance for EWID data from experiments of similar construct and size.

Variable Importance. Random forests have also have a built-in method to identify important covariates using the variable importance values such as mean decreased accuracy (MDA) or the Gini score for homogeneity. The larger the value of the metric, the more important the covariate is when looking at either MDA or Gini score. These metrics could identify covariates that more strongly affect or are more influential on the eyewitness outcomes. An example of a variable importance plot showing both MDA and Gini score is provided in Figure 4.3.

In general, we found that covariates such as decision time and lineup bias were routinely the most important covariates of the several that were included for the factor dataset in the target present models across the 50 subsamples. Other covariates that occasionally appeared included ECL. For target absent models, decision time

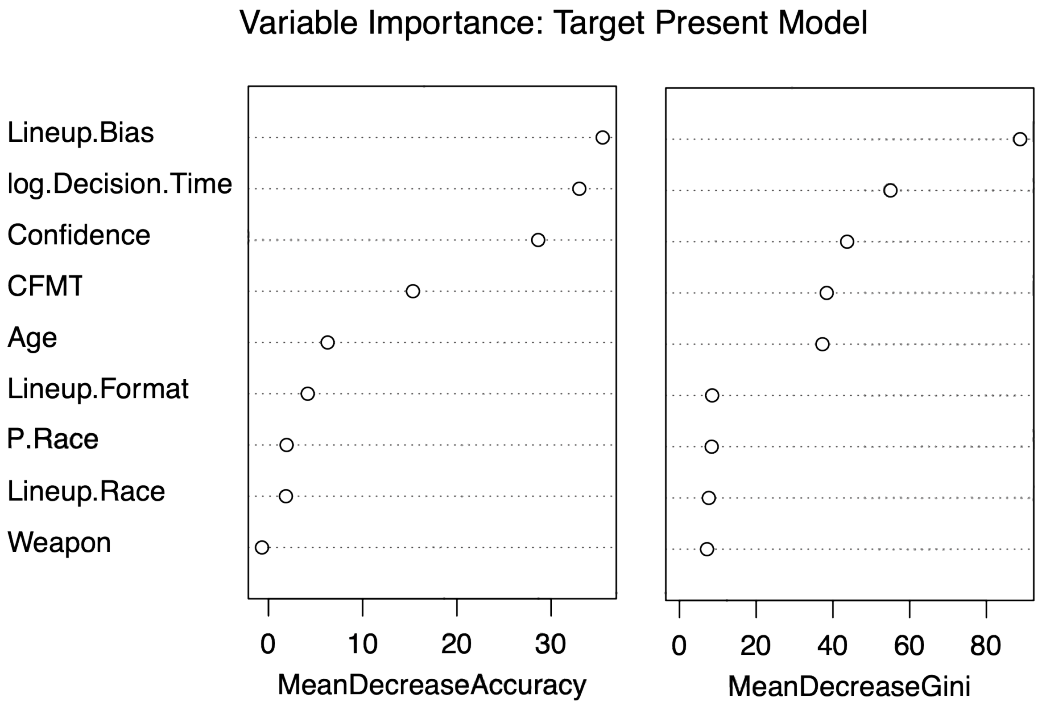


Figure 4.3: Example of a variable importance plot from a target present model using a subset of the factor data set. The more influential covariates are at the top of the plot, with decreasing influence as the plot progresses from top to bottom. Note here that "P.Race" refers to participant race.

was the most important factor, almost universally. The lineup instructions were the most influential in the Mickes et al. (2017) dataset in both the target present and target absent models. On occasion, age and ethnicity appeared as important factors, depending on the subsample. The datasets from Seale-Carlisle et al. (2019) showed similar behavior.

4.2.2 Rate of Target Presence

The rate of target presence θ is then estimated using a decomposition of a subset of the decision outcome space. In this case, we need to decompose something that will always be observed, which is the probability of choosing $P(\text{Choose})$. Let probability

of choosing $X_i \sim \text{Bern}\left[p_i(\theta)\right]$ for each person $i = 1, \ldots, n$ where $p_i(\theta)$ is the probability of "success" (i.e., choosing). Let

$$a_i = (1 - A_{1i}) \tag{4.6}$$

and

$$b_i = P1_{1i} + P_{2i} + A_{1i} - 1. \tag{4.7}$$

Then, we expand the probability of choosing using the law of total probability to write it in terms of a_i and b_i,

$$\begin{aligned}
P(\text{Choose}) &= P(X_i = 1) \\
&= P(X_i = 1 \mid T = 1) \cdot P(T = 1) \\
&\quad + P(X_i = 1 \mid T = 0) \cdot P(T = 0) \\
&= P(X_i = 1 \cap\ Y_i = 1 \mid T = 1) \cdot P(T = 1) \\
&\quad + P(X_i = 1\ \cap\ Y_i = 0 \mid T = 1) \cdot P(T = 1) \\
&\quad + P(X_i = 1\ \cap\ Y_i = 1 \mid T = 0) \cdot P(T = 0) \\
&\quad + P(X_i = 1\ \cap\ Y_i = 1 \mid T = 0) \cdot P(T = 0) \\
&= P_{1i} \cdot \theta + P_{2i} \cdot \theta + 0 \cdot (1 - \theta) + A_{2i} \cdot (1 - \theta) \\
&= (1 - A_{1i}) + (P_{1i} + P_{2i} + A_{1i} - 1) \cdot \theta \\
&= a_i + b_i \cdot \theta \\
&= p_i(\theta).
\end{aligned} \tag{4.8}$$

This decomposition simply reframes the eyewitness decision outcome space by looking at a different set of conditional probabilities, as illustrated in Figure 4.4.

We can find a point estimate for θ using MLE, such that $\theta \in [0, 1]$. The joint

likelihood $L\big[p_i(\theta)\big]$ of X_i for n observations is,

$$\begin{aligned} L\big[p_i(\theta)\big] &= \prod_{i=1}^{n} \big[p_i(\theta)\big]^{x_i} \cdot \big[1 - p_i(\theta)\big]^{1-x_i} \\ &= \prod_{i=1}^{n} (a_i + b_i \cdot \theta)^{x_i} \cdot (1 - a_i - b_i \cdot \theta)^{1-x_i}, \end{aligned} \tag{4.9}$$

which is used to find the joint log-likelihood $\ell\big[p_i(\theta)\big]$,

$$\ell\big[p_i(\theta)\big] = \sum_{i=1}^{n} \big[x_i \log(a_i + b_i \cdot \theta) + (1 - x_i) \log(1 - a_i - b_i \cdot \theta)\big]. \tag{4.10}$$

Now, take the first partial derivative of the joint log-likelihood with respect to θ and set equal to zero to solve for the local maximum, since we are assuming $0 < \theta < 1$,

$$\frac{\partial\ \ell\big[p_i(\theta)\big]}{\partial\theta} = \sum_{i=1}^{n} \left[\frac{x_i \cdot b_i}{a_i + b_i \cdot \theta} - \frac{(1 - x_i) \cdot b_i}{1 - a_i - b_i \cdot \theta}\right] \triangleq 0. \tag{4.11}$$

We solve Equation 4.11 numerically for $\hat{\theta}_{\text{MLE}}$, using (for example) `uniroot()` (Team, 2019), constrained to $[0, 1]$. Once the probability of target presence $\hat{\theta}_{\text{MLE}}$ is estimated, we have all of the requisite components to estimate the underlying proba-

Accurate	**Not Accurate**		
Choose Target (P_1)	Choose Foil (P_2)	Do Not Choose (P_3)	**Target Present**
Do Not Choose (A_1)	Choose Foil (A_2)		**Target Absent**

Figure 4.4: A display of the eyewitness decision outcome space, which takes into account the underlying status of the lineup, where considering P(Choose).

bility of accuracy $\hat{\rho}_i$ for each participant,

$$\begin{aligned} \widehat{P(i\text{-th participant accurate})} &= \widehat{P(Y_i = 1)} \\ &= \hat{P}_{1i} \cdot \hat{\theta}_{\text{MLE}} + \hat{A}_{1i} \cdot (1 - \hat{\theta}_{\text{MLE}}) \\ &= \hat{\rho}_i. \end{aligned} \tag{4.12}$$

Theoretically, the point estimate should follow the asymptotic properties of any other MLE. It should be consistent, efficient, and functionally invariant. That is the estimator will converge to some value, which means that the variance and bias of the estimator will converge to zero as sample size n tends to infinity. The estimator will also be asymptotically efficient (i.e., it achieves the Cramér-Rao lower bound, or lowest MSE, as n tends to infinity) The distribution assumptions made here are based more on convenience and hope, rather than any demonstrable evidence for its validity.

To ensure the existence of estimators using MLE with the constraint $\hat{\theta}_{\text{MLE}}$, as the parameter space may not necessarily be convex and the likelihood function may not necessarily be concave, we recommend running multiple subsamples to train multiple models. Then, to average the estimates across the models. In our application, we ran $m = 50$ models with subsamples $s = 0.7n$, where n is the total number of observations. That is, we randomly selected subsamples without replacement that were 70% of the original sample size. Of course, we reserved the remaining 30% of the samples as testing samples.

Other Methods of Estimating P(TP)

Other researchers have looked into estimating the probability of target presence. The primary method is based on the idea of minimization of some loss function, and

utilizes the assumptions of the SDT model (Wixted et al., 2016a, 2018).[1] Semmler et al. (2018) provides the code to estimate this value, which is referred to as maximum likelihood estimation for signal detection theory. This method assumes two normal distributions: one for the memory strength corresponding to the suspect $N\left(\mu_{\text{target}}, \sigma^2_{\text{target}}\right)$ and one for memory strength corresponding to the lure (i.e., innocent suspect) $N\left(\mu_{\text{lure}}, \sigma^2_{\text{lure}}\right)$. These distributions could be parsed into finer intervals based on some confidence levels c_1, c_2, and c_3. Thus, there could be up to five parameters (the means, variances, and confidence levels) to estimate based on these assumptions. Only up to five parameters could be estimated, as only six degrees of freedom exist. Often times the model is assumed to follow an equal variance model, where $\sigma^2_{\text{target}} = \sigma^2_{\text{lure}}$.

The data is a table of counts, separated into suspect identification (guilty or innocent suspect), foil identification, and not present decisions. The estimation procedure depends on the number of foil identification and suspect identification, which are used in the derived formulas for estimation. In some applications, the estimation is performed using iterative proportional fitting for a two-dimensional table, by iteratively generating estimates until the difference of the known truth and the estimate are below some threshold. The estimates are iteratively generated with the goal of minimizing the observed frequencies or cell counts for the data and the cell frequencies for the model. Semmler et al. (2018) uses the likelihood ratio statistic G^2, for observed counts o_i and expected counts e_i, whereas Wixted et al. (2018) use the chi-squared goodness-of-fit statistics of the actual and estimated tables. Wixted et al. (2018) showcase the performance of their procedure by simulating various rates of target presence for a chosen data set in Figure 4.5. Both the G^2 statistic and

[1] Andrew Cohen, Jeffrey Starns, Caren Rotello, and Andrea Cataldo from University of Massachusetts at Amherst are working on extending this method; at the time of this writing, they have not published their work. Knowledge of their work comes from personal communication.

the chi-squared statistics approximately follow chi-squared distributions, with the G^2 approximating the distribution more closely for small sample sizes.

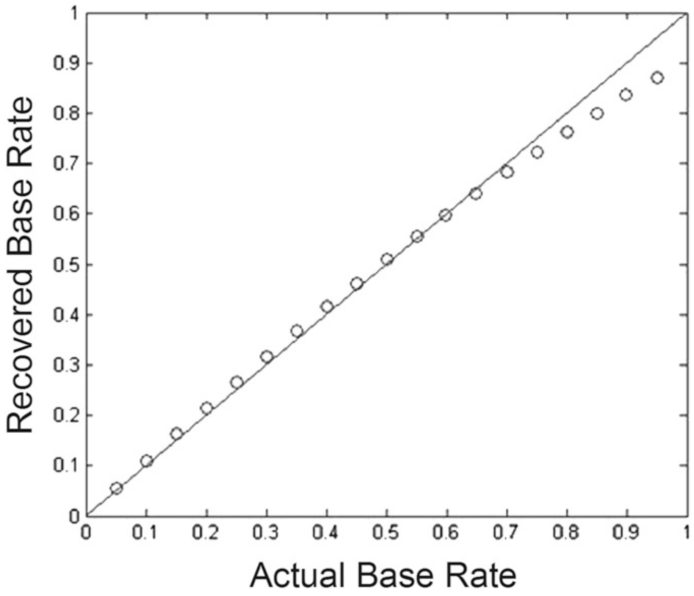


Figure 4.5: Estimated probability of target presence (named base rate in this plot) from different underlying rates of actual target presence in the data. This plot was taken from Wixted et al. (2018) to show the performance of their model fitting procedure. The underlying true base rate is on the x-axis and the estimated (i.e., recovered) base rate is on the y-axis.

An advantage to this estimation procedure is the simplicity of the data structure required. The estimation procedure needs only summary counts, whereas for our proposed framework, we require the entire data set. However, this method relies solely on minimizing the loss function, which results in very similar behavior across different data sets. Both procedures are generalizable to other data sets, but our proposed framework enables the estimation of an additional (primary) variable of individual probability of choosing.

We will use the code from Semmler et al. (2018) for a rough comparison of performance. It seems with the MLE SDT method, the behavior will be similar for similar groups of people or similar experimental types. While the results from

Cohen's group show a mixture of behaviors, we found that the MLE SDT code from Semmler et al. (2018) exhibits similar behavior for estimation for similar types of data. Using the experimental data from Mickes et al. (2017), we fit the proposed framework and the MLE SDT estimation method for each of the experiments.

4.3 Application

Empirically, it seems that this framework performs best given an optimal combination of sample size and covariate information. While the factor data set was not the largest data set fit, it did include the most number of covariates from demographic information and characterizing the lineup. The performance on the delay data set was almost, but not quite, as good, primarily due to the availability of covariates and the relatively large number of observations.

4.3.1 Factor Data Set

Now that the procedure has been established, we seek to show proof-of-concept of performance and usability. We return to one of the data sets described in Section 2.2.2 obtained by Dodson and his lab at the University of Virginia. This is the data, denoted as the *factor data set*. The data set has 3160 observations with 16 variables. Recall the data set arises from a 2^4 full-factorial designed study. Figure 4.6 shows only the first six rows of this data set; each row corresponds to a different participant, recruited via Qualtrics©, SurveyGizmo©, or Amazon© Mechanical Turk. Additional covariate information and counts for the factor data set can be found in Table B.1 and Table B.2.

We ran the estimation procedure on the factor data set to visualize the empirical density of the probabilities of accuracy for the factor data set as shown in

```
     Platform Participant.Race    Sex Age Group
1 SurveyGizmo            White   Male  37    31
2   TurkPrime            White   Male  25    22
3 SurveyGizmo            White   Male  38    24
4 SurveyGizmo            White   Male  32    10
5 SurveyGizmo            White Female  37    14
6   Qualtrics            Black   Male  21    24

  Lineup.Race Weapon Lineup.Format Lineup.Bias Target.Present
1       Black     No           Sim        Fair             TA
2       White    Yes           Seq        Fair             TA
3       Black    Yes           Seq        Fair             TA
4       White     No           Seq        Bias             TP
5       White     No           Seq        Fair             TP
6       Black    Yes           Seq        Fair             TA

  CFMT Chooser    Category Accuracy Confidence Decision.Time
1   56     Yes        Foil        0        0.6        0.1920
2   24     Yes        Foil        0        0.4        0.6670
3   61     Yes        Foil        0        1.0        0.8700
4   37     Yes        Foil        0        0.4        1.0130
5   65     Yes        Foil        0        1.0        1.2350
6   40      No Not Present        1        0.6        1.2285
```

Figure 4.6: First six observations of the factor data set with all 16 variables

Figure 4.7. The following covariates are included in the random forests: weapon presence, lineup race, lineup format, lineup bias, CFMT score, participant race, confidence in response, logarithm of decision time, and age.

The first measure of performance is to gauge the ability of the method to "recover" (i.e., estimate) the rate of target presence. The factor data set has a fixed rate of 50% target present observations, so data with other rates need to be simulated. Since it is known that MLE may suffer at the edges of the response surface, we chose to simulate rates ranging from 5% to 95%. This is realistic, since it is highly unlikely that a law enforcement agency would have near perfect nor near imperfect accuracy. The rates are simulated at each percentage point between 5% and 95%. For each simulated rate, 50 random subsamples without replacement were taken from the entire data set by downsampling within each subset of target present and target absent observations such that the requisite proportions of observations

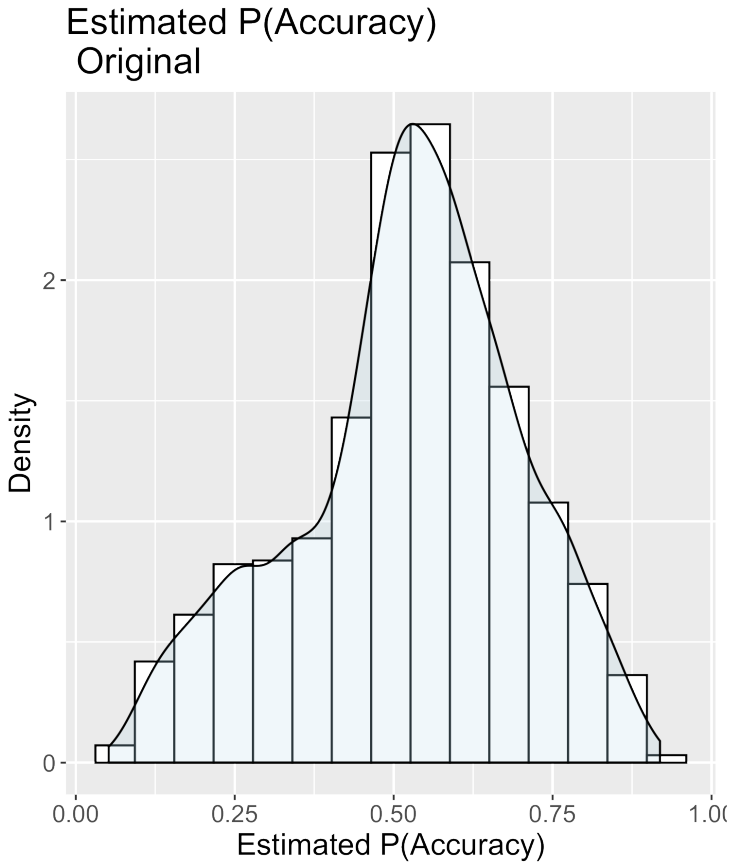


Figure 4.7: Empirical density of estimate probabilities of accuracy from the factor data set

are met. Each training set subsample consists of approximately 60% to 70% of the original, full data set. The remaining 30% to 40% of the data were denoted the testing set. Among the 50 models fit, we found the mean and median estimated values of the rate of target presence. The results are shown in Figure 4.8, and have extraordinarily good results, with very little bias for the majority of rates and minimal and constant variance across the span of all rates. Overall, the method seems to be biased for more extreme values of true $P(\text{TIP})$, by slightly over-predicting for low true $P(\text{TIP})$ and under-predicting for high true $P(\text{TIP})$.

The second measure of performance is a bit of fine-tuning in terms of node size and number of trees for the random forests. The default node size is 1. We also allow

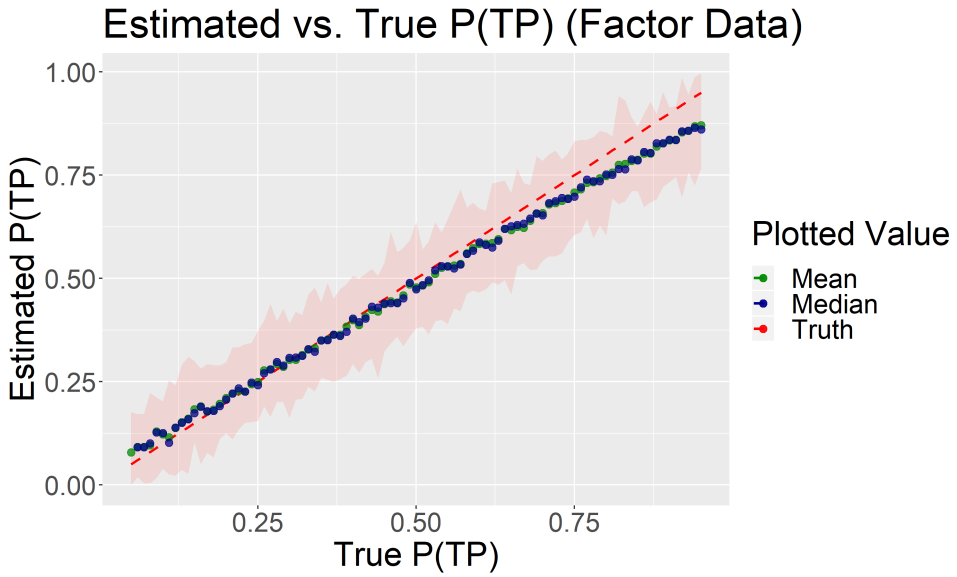


Figure 4.8: Factor data set: estimated probabilities of target presence (base rate) versus underlying truth for rates from 5% to 95% at increments of 1% with 50 subsamples without replacement for each rate. The green points are the mean estimated values and the blue points are the median estimated values. The underlying truth is on the x-axis, while the estimated rates are on the y-axis. The shaded red area represents the full range of estimates (minimum to maximum estimates).

the node size to be 1% to 10% of the training sample size, with increments of 1%. The number of trees range from 100 trees to 1000 trees, with increments of 100. We evaluate the performance of the random forest by looking at all combinations of the node size and number of trees with 50 iterations (i.e., 50 subsamples of $n = 1000$ taken without replacement); i.e., node sizes $[1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100]$. From the 50 iterations, we look at the range of the values (maximum minus minimum) for each iteration, and find the median value, as shown in Figure 4.9, to assess the variability of the estimates based on different subsamples. The median is established based on testing values run through the model, across the 50 iterations for all combinations. Based on Figure 4.9, variability of the estimates decreases as the node size increases, but only up to a certain node size. The largest decrease of variability is from node size of 1 to node size of 10 (1% of the sample size). Vari-

ability decreases at smaller steps for node sizes of 20 and larger. This suggests an optimal node size is potentially 1% of the sample size. Of course there seems to be less variability for node size of 40 (4% of the sample size). Additionally, 500 trees appears to be the point where the variability flattens.

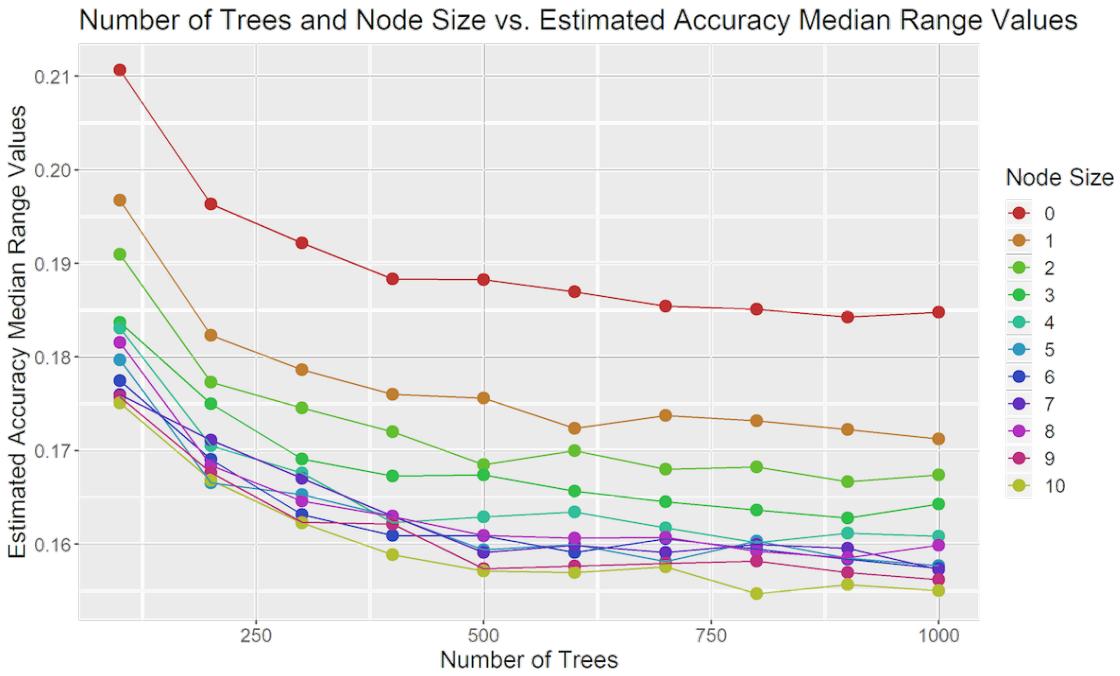


Figure 4.9: Factor data set: estimated probabilities of accuracy showing the effect of node size and number of trees. The x-axis shows the number of trees, ranging from 100 to 1000 at increments of 100. The y-axis represents the median range of estimated probabilities across all samples. The colored lines represent the node sizes, where "0" is the a node size of 1 and the node sizes labeled "1" to "10" are the relative proportions of the sample size.

To further assess the performance of various combinations of node sizes and numbers of trees, we look at the bias of the estimate of the rate of target presence. We would look at the bias of the estimated probabilities of accuracy, but no obtainable true values exist for comparison. Figure 4.10 shows the increase of bias node size increases. The bias increases minimally between node size of 1 and 10, but more significantly increases for larger node sizes. Thus, the suggestion of optimal node

size of 1% of sample size is justified in this case and, potentially, for other data sets. Number of trees minimally affects the bias.

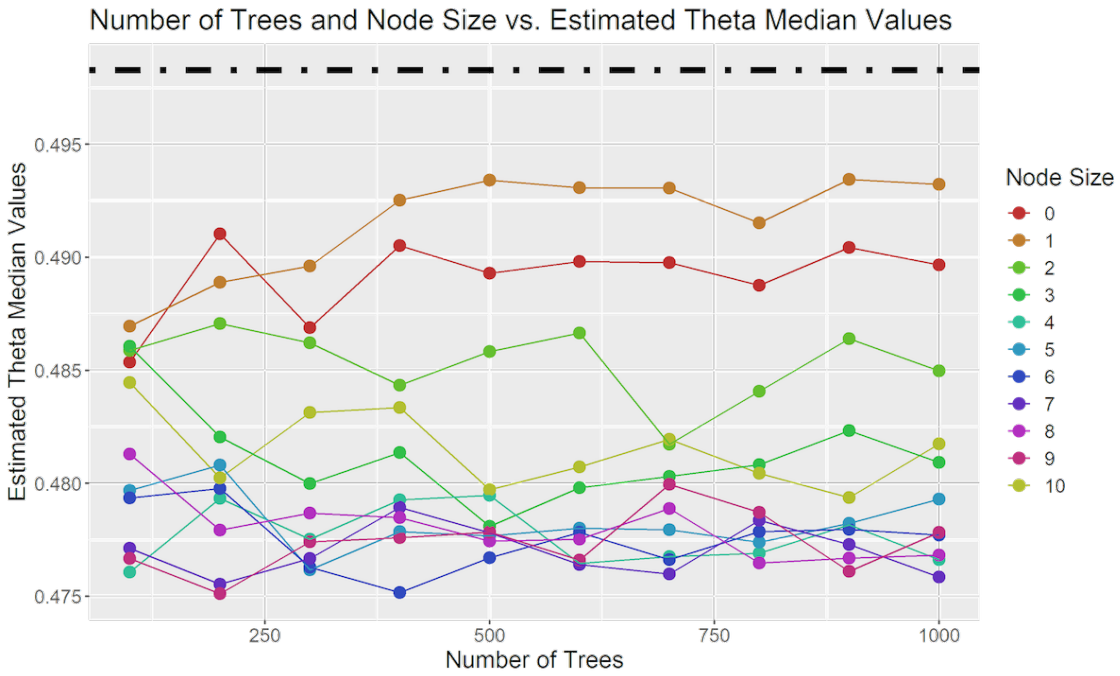


Figure 4.10: Factor data set: estimated probabilities of target presence showing the effect of node size and number of trees. The x-axis shows the number of trees, ranging from 100 to 1000 at increments of 100. The y-axis represents the median estimated rate of target presence across all samples. The colored lines represent the node sizes, where "0" is the a node size of 1 and the node sizes labeled "1" to "10" are the relative proportions of the sample size. The black dotted line represents the underlying truth.

The final measure of performance is the comparison of interpretability of the estimated probabilities of accuracy to existing methods. Specifically, we are interested to see if universally accepted or agreed-upon conclusions of lineup characteristics remain true for the estimated probabilities in the case of the factor data set. From the literature, researchers universally agree that fair lineups are superior to biased lineups, as biased lineups cause more high confidence misidentifications than do fair lineups.[2] Researchers also strongly agree that for choosers:

[2]Chad Dodson, personal communication

(1) Fair lineups have better accuracy than biased lineups;

(2) Same-race lineups have better accuracy than cross-race lineups;

(3) Crimes without weapons result in lineups that have better accuracy than crimes with weapons;

(4) Simultaneous lineups are superior to sequential lineups; and

(5) Lineups of size one (i.e., a showup) is inferior to a lineup size of six.

Note that these statements are not necessarily terminal nor absolutely conclusive, but serve as a snapshot of the current beliefs in the literature.

We can assess four of the five items from above by looking at the estimated probabilities of accuracy. We subset the probabilities by the appropriate factor, which are lineup fairness, cross-race effect, weapon presence, and type of lineup in this case. We assess this via notched box plots in Figure 4.11 to view the distribution of estimated probabilities by the levels of the factors. The box plots suggest that fair lineups have higher accuracy than biased lineups by a subjectively marginal amount, but have a larger variability. Simultaneous lineups have much better performance in terms of accuracy and lower variability. The cross-race and weapon effects are less obvious, with cross-race and weapon presence lineups showing a slightly higher accuracy. These boxplots also collapse across choosers and non-choosers.

From an applied standpoint, it may be important to separate chooser and non-choosers responders, as there are different consequences when an eyewitness has responded "not present" versus selecting a face from a lineup. Thus, we can examine the boxplots for the subsetted data (choosers separate from non-choosers), as shown in Figure 4.12 and Figure 4.13, respectively.

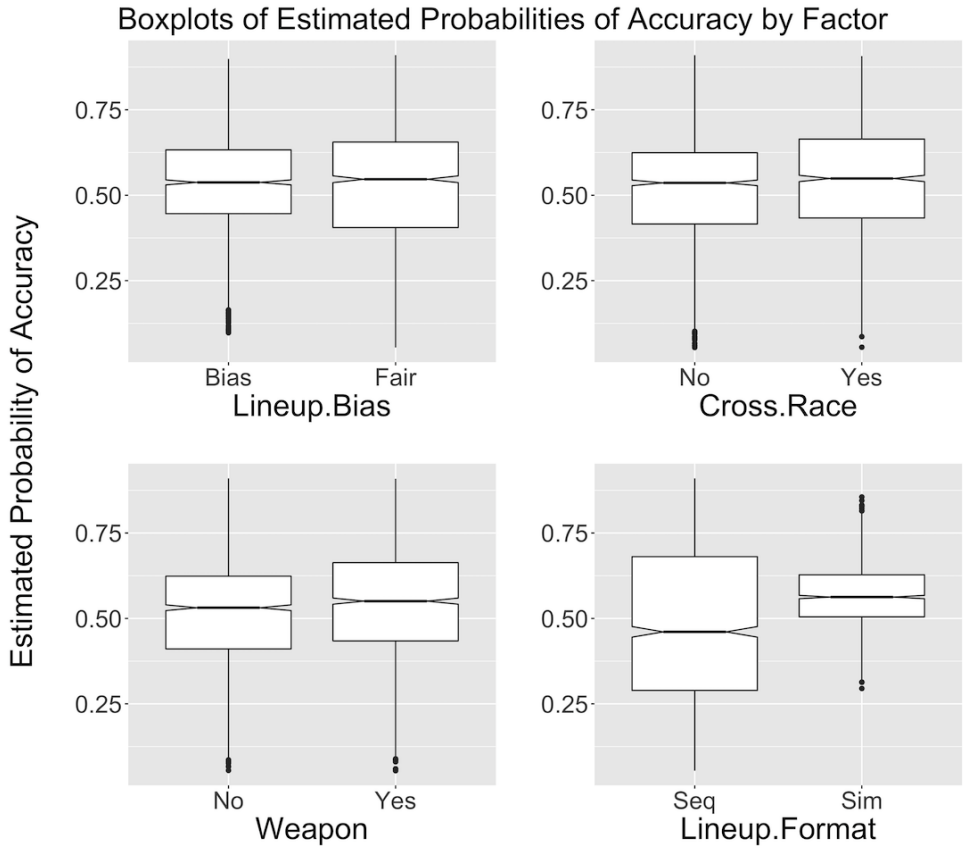


Figure 4.11: Factor data set, comparison of estimated probabilities of accuracy by four factors: (1) lineup fairness (top left); (2) cross-race effect (top right); (3) weapon presence (bottom left); and (4) type of lineup (bottom right). The estimated probabilities are on the y-axis, while the two levels for the four factors are on the appropriate x-axes.

For choosers, fair lineups do appear to be more accurate than biased lineups. However, it seems that there is little difference from the collapsed data regarding the accuracy of same- versus cross-race lineups and weapon presence. The cross-race result may be a result of an extremely memorable actor for the white perpetrator. There is a switch in the sequential versus simultaneous lineup conclusion, where there is less accuracy. However, simultaneous lineups do seem to have much less variability.

For non-choosers, biased lineups appear to be more accurate than fair lineups.

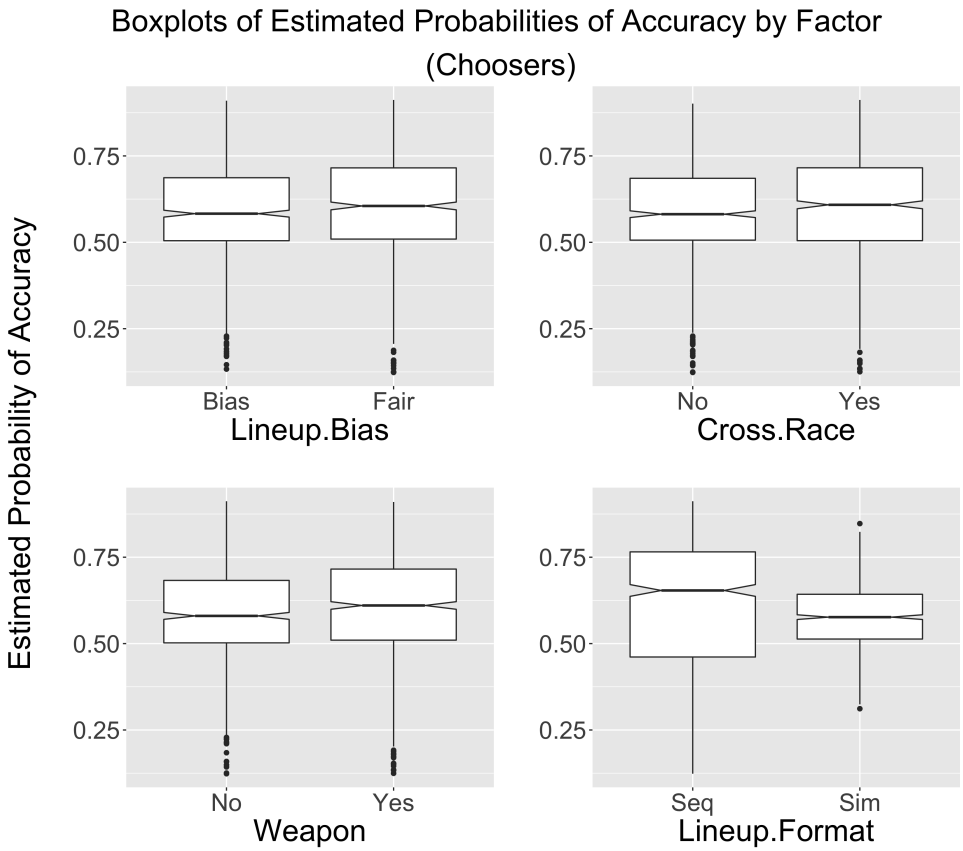


Figure 4.12: Factor data set, comparison of estimated probabilities of accuracy by four factors for **choosers**: (1) lineup fairness (top left); (2) cross-race effect (top right); (3) weapon presence (bottom left); and (4) type of lineup (bottom right).

However, again there seems to be little difference in the results for the same- versus cross-race and weapon presence conclusions. Simultaneous lineups are more accurate than sequential lineups.

These conclusions are ambiguous, as the notches do appear to overlap for the majority of the cases, regardless of the collapse or separation of choosers versus non-choosers. Additionally, these box plots are only based on a singular experiment, which may or may not be 100% ecologically valid. We would like to explore more data sets similar to this one in terms of complexity level and number of observations for further analysis.

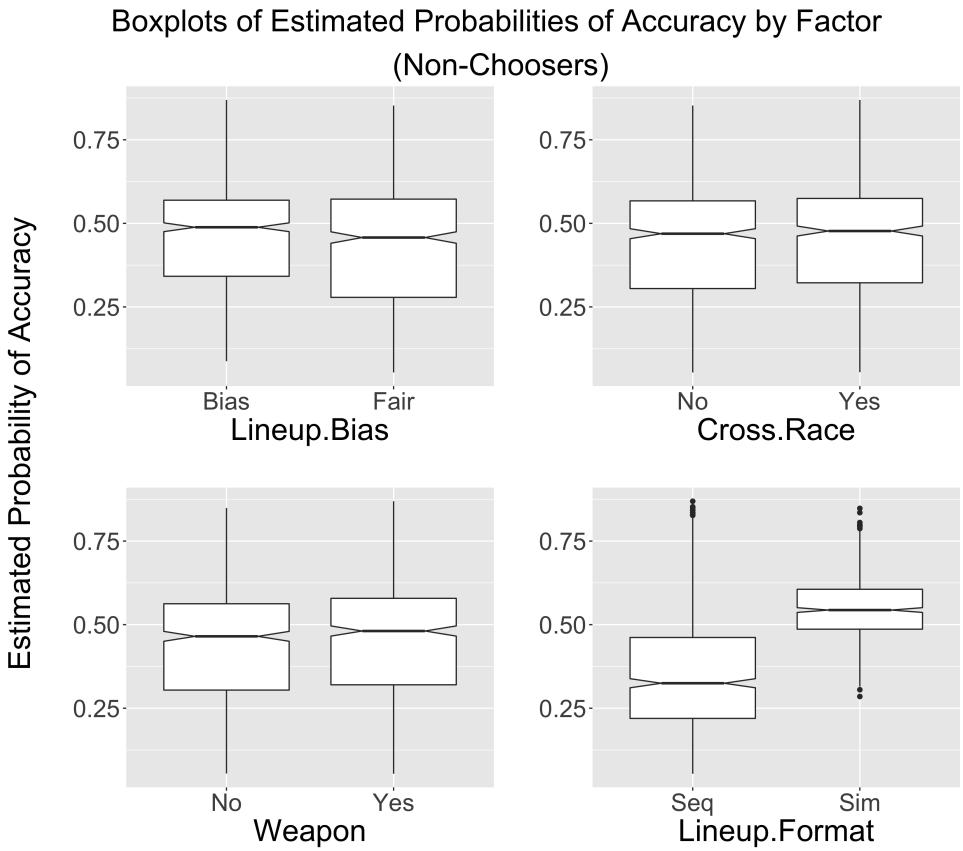


Figure 4.13: Factor data set, comparison of estimated probabilities of accuracy by four factors for **non-choosers**: (1) lineup fairness (top left); (2) cross-race effect (top right); (3) weapon presence (bottom left); and (4) type of lineup (bottom right).

In fact, this framework may not be the most appropriate method for the separate analysis of choosers versus non-choosers, as we are looking at the probability of accuracy, which averages a single eyewitness's. It may be possible to consider the components separately as $P(\text{Choose Target} \mid T = 1) \cdot P(T = 1)$ for choosers and $P(\text{Don't Choose} \mid T = 0) \cdot P(T = 0)$ for non-choosers.

Another option for the completely separate analysis of choosers versus non-choosers is to consider a slightly different framework than the one proposed here. It may be possible to decompose accuracy in terms of the probability of choosing $P(\text{Accurate} \mid \text{Choose})$ and the probability of not choosing $P(\text{Accurate} \mid \text{Do Not Choose})$.

The details of this framework may more complicated, as there is overlap in the estimates of these two probabilities.

4.3.2 Other Data Sets

Similar results were found for other data sets, such as the delay and repeated delay data sets described in Section 2.2.2. The performance of the estimator for TIP observations declined with less complicated data sets with fewer potential covariates. This is due to the limited information to capture the true behavior of the global variable of target presence. Clearly, we need an ecologically valid data set, with as many variables measured as possible, to better generalize the fitted model to new data sets. Nonetheless, the performance of the estimation procedure appears to be more informative than the procedure from Wixted et al. (2018) and the Cohen's team at UMass Amherst.[3]

The MLE SDT method from Semmler et al. (2018) was fit to a table where the ratio of target present and target absent lineups was known, giving it an advantage in terms of estimation performance. It was done this way for two reasons. First, this enables an "unfair" comparison, giving a "best case scenario" result for the MLE SDT method and a "realistic" scenario for our methodology. A comparison of metrics (squared bias, variance, and MSE) is provided in Table 4.2 for multiple data sets, including the factor data set, delay data set, Mickes et al. (2017) data set, and all data from all experiments in Seale-Carlisle et al. (2019). Figure 4.14 shows the performance from four data sets: the factor data set, the delay data set, the Mickes et al. (2017) data set, and experiment 1 from Seale-Carlisle et al. (2019).

Covariates included in the delay data set random forest models are delay, weapon

[3]Unfortunately, there was difficulty obtaining the full data sets to run a full performance comparison.

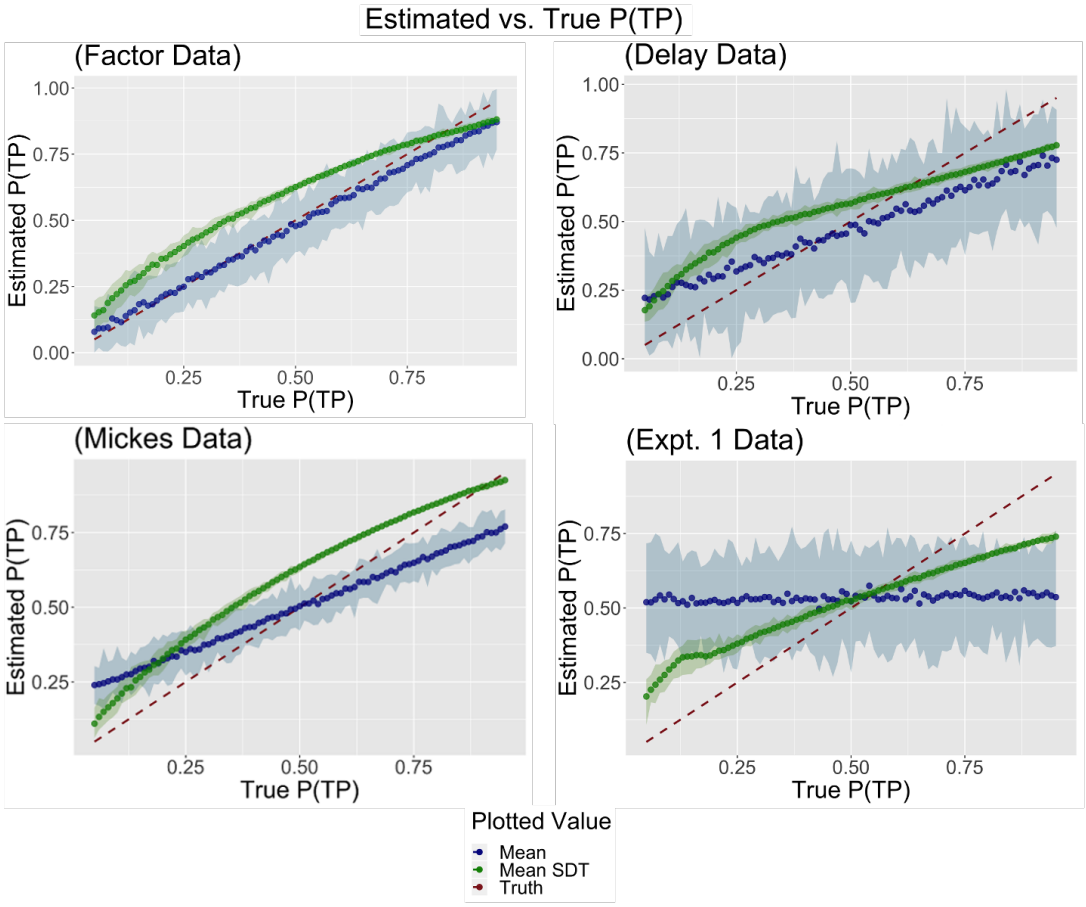


Figure 4.14: Estimated probabilities of target presence (base rate) versus underlying truth for rates from 5% to 95% at increments of 1% with 50 subsamples without replacement for each rate. The green points are the mean estimated values from the MLE SDT estimation method from Semmler et al. (2018) and the blue points are the mean estimated values from our framework. The underlying truth is on the x-axis, while the estimated rates are on the y-axis. The shaded areas with their respective colors represent the full range of estimates (minimum to maximum estimates).

presence, confidence rating, logarithm of reaction time to making a confidence rating, CFMT score, sex, age, and logarithm of decision time. Covariates included in the Mickes et al. (2017) data set random forest models are age, ethnicity, education level, sex, and instructional biasing. We did not include confidence, since only 1/6 of the participants were asked to report a confidence rating, which means that 5/6 of the data set would have a missing value for that covariate. Covariates included

in the experiment 1 from Seale-Carlisle et al. (2019) include experimental condition (simultaneous versus sequential), confidence, age, ethnicity, and sex.

We evaluated the methods via squared bias, variance, and MSE (the sum of squared bias and variance), where Bias^2 and Var are defined in Equation 4.13. Suppose there are T known truths for probability of target presence.

$$\begin{aligned} \text{Bias}^2 &= \frac{1}{T}\sum_{t=1}^{T}\left[\left(\frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_{mt}\right) - \theta_t\right]^2 \\ \text{Var} &= \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_{mt} - \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_{mt}\right)^2\right] \end{aligned} \tag{4.13}$$

Overall, the MLE SDT method appears more robust based on the metrics in Table 4.2, but the visual comparison provides a different story. In Figure 4.14, the MLE SDT performance is nearly identical across the four data sets due to the maximization of the goodness-of-fit criteria for the table counts. The method overestimates for the majority of the true rates. However, our framework provides unique fits for the available information in the data.

It seems that the more information available that can characterize more of the variability in the response, the better the estimation is, as shown by the factor data set. The extreme deterioration in performance from the experiment 1 data is due to the potential lack of information in the manipulated experimental conditions. The experiment 1 data from Seale-Carlisle et al. (2019) has arguably the least amount of information available, which explains the poor performance. Some information exists in the demographic covariates, which enables a slightly increasing estimate as the true value of $P(\text{TIP})$ increases. It is clear that the information is not nearly sufficient. We also like to note the performance from Seale-Carlisle et al. (2019) experiment 1 is the worst across all data sets fit using our framework, including the

other experiments 2 through 5 from Seale-Carlisle et al. (2019).

Additionally, the small variability in estimates using the MLE SDT method suggests that if (and when) an incorrect estimate is made, it is unlikely to lie close to the truth. This issue could potentially be corrected with a bias correction, which has not been addressed in the literature.

Data Set	**Method**	**Bias^2**	**Variance**	**MSE**
Factor data set	SDT	0.013	1.62E-04	0.013
	Framework	0.004	0.002	0.006
Delay data set	SDT	0.017	1.78E-04	0.017
	Framework	0.022	0.009	0.031
Mickes et al. (2014) data set	SDT	0.012	1.10E-04	0.012
	Framework	0.012	9.44E-09	0.013
Expt. 1	SDT	0.015	1.67E-04	0.015
	Framework	0.073	0.006	0.079
Expt. 2	SDT	0.018	1.73E-04	0.018
	Framework	0.068	0.006	0.074
Expt. 3a	SDT	0.014	1.21E-04	0.014
	Framework	0.056	0.006	0.062
Expt. 3b	SDT	0.015	1.03E-04	0.015
	Framework	0.060	0.007	0.067
Expt. 4	SDT	0.017	1.55E-04	0.017
	Framework	0.051	0.007	0.058
Expt. 5	SDT	0.013	1.89E-04	0.014
	Framework	0.040	0.007	0.047

Table 4.2: Comparison of squared bias, variance, and MSE from the Dodson data sets (factor and delay data sets), Mickes et al. (2014) data set, and Seale-Carlisle et al. (2019) data sets. Here, "SDT" is the MLE SDT method and "Framework" is our proposed modeling framework.

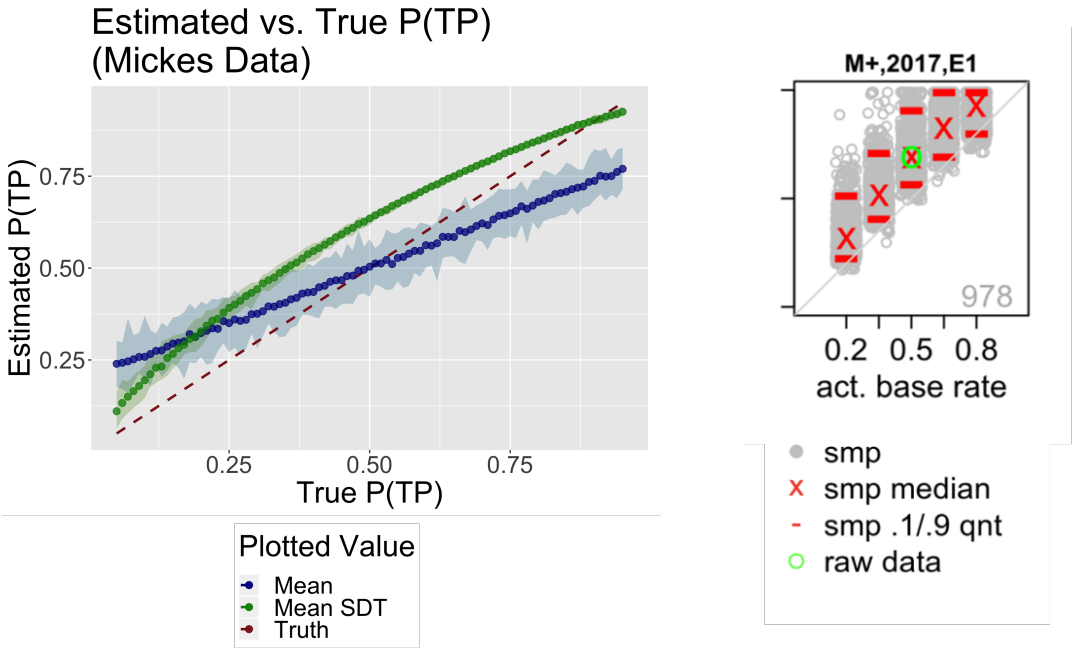


Figure 4.15: Comparison of estimated base rates to true base rates for our framework (on the left) and the MLE SDT method implemented by Cohen's group (on the right)

The method implemented from Semmler et al. (2018) is not the exact method implemented by Wixted et al. (2018), but rely on the same underlying concepts. The described methodology from Wixted et al. (2018) was unable to be replicated, and the upcoming contributions from Cohen's group at UMass Amherst has not yet been published. In the future, once their proposed R package `sdtlu` is available, a more in-depth and thorough comparison could be performed. Cohen's group did run their estimation method on the data from Mickes et al. (2017), to which we had access.[4] Cohen's group seemed to use a subsample size of approximately 1000 observations, whereas, we used approximately 600 observations. The performance of Cohen's group's method appears visually different from the Semmler et al. (2018) method, but this is due to the difference in step sizes between simulated base rates. The MLE

[4]Travis Seale-Carlisle, personal communication.

SDT method overestimates the actual base rate with obvious regularity, whereas our framework simultaneously overestimates for lower base rates and underestimates for higher base rates, with exact estimates for the middle range of base rates.

4.4 Framework

We have provided the details and construction of the framework. Overall, the framework provides an estimation procedure for a specific data structure that contains two types of latent variables: (1) a per unit latent variable and (2) a global latent variable. In the case of EWID data, the per unit latent variable is the probability of choosing for each person, which is dependent upon the global latent variable of rate of target presence. Neither variable is known a priori, and, in fact, neither is usually ever known. The general algorithm detailed in this chapter is provided in Algorithm 1.

We could extend this framework to other types of data that follow the same inherent idea of a per unit latent variable and a global latent variable. The two latent variables need to be related in some way by some possible outcome. In the case of EWID data, the outcomes are the eyewitness decisions within a lineup. Figure 4.16 shows the general data structure required, where "Var. 1" represents the per unit latent variable and "Var. 2" represents the global latent variable. The outcomes connecting the two latent variables are listed as cross-items within the table. There can be as many categories within each latent variable as necessary for the data, with as many outcomes as necessary for the combinations of latent variable categories, as denoted by the triple dots.

For example, suppose a geophysicist is choosing locations to obtain samples to measure the presence of two compounds. The geophysicist would be interested in the

Algorithm 1: Estimation procedure framework for estimating a per unit latent variable and global latent variable for EWID data

Input: Data set with n samples, pre-determined number of models to fit m
Output: Predicted probability of choosing for each participant ρ_i and estimated probability of target presence θ

for $j = 1, \ldots, m$ **do**
- Randomly select subsample without replacement such that total number of training samples is equal to $0.7n$.
- Subset subsampled data into target present and target absent participants.
- Fit a random forest to each subset of data using the eyewitness decisions as the response variable.
- **for** $i = 1, \ldots, n$ **do**
  - Determine the probability estimates P_1, P_2, P_3, A_1, and A_2 for each participant.
  - Calculate a_i and b_i.
  - Solve numerically for θ_{MLE} using the provided formula in Equation 4.11.
  - Plug in estimated values to find ρ_i.
- **end for**

end for

probability of finding useful samples in a given location. Suppose a unit is a sample in this case. The per sample (i.e., column) variable is the indicator function of useful sample (yes or no). The global (i.e., row) variable are the possible locations (location one, location two, etc.). The outcomes (i.e., within the boxes) are the status: zero, one, or two compounds.

Another example is determining the probability of payment for a credit card company, given a general propensity for charge-off. A charge-off occurs when a credit card user fails to pay the full balance on a card within some specified period of time, which requires the credit card company to assume full responsibility of the revenue loss. Of course, it is in the best interest of the credit card company to minimize this loss to maximize profit. In this case, the per user variable is the probability of payment. The global variable is the probability of charge-off, perhaps

Var. 1a	**Var. 1b**	**⋯**	
Outcome 1	Outcome 3	⋯	**Var. 2a**
Outcome 2	Outcome 4	⋯	**Var. 2b**
...	...	...	**...**

Figure 4.16: General data structure required for utilizing the proposed estimation framework

averaged over types of credit cards or credit score brackets (i.e., subprime, near prime, prime, and super prime). The outcomes are identifying when the user last made a payment (i.e., on-time payment in the last cycle, two cycles, three cycles, etc.)

Now suppose an agricultural science company runs an experiment to determine the performance of fungicides. They are looking at the probability of resistance of fruit to some fungus. The per fruit (e.g., apples, oranges, lemons, etc.) variable is the probability of resisting the fungus, inherent to the fruit. The global variable is the average effectiveness (as measured in some appropriate metric) of fungicide A, fungicide B, etc. The outcomes are the types of harvest, whether it is a usable harvest, unusable harvest, or no harvest at all.

4.4.1 Limitations

This method has some limitations. First, the method requires the entire data set, rather than a set of summary counts. It may be more difficult to obtain full data

sets rather than those summary counts. A more immediate problem is the lack of available, replicated full data sets. In fact, it was not feasible to fit models to existing "training" data and then fit new "testing" data even at this stage. The available data sets did not replicate each other in terms of manipulated and/or recorded covariates.

Second, the data needs to be ecologically valid for use in the field. The end goal is to fit a model that could be used at a law enforcement agency for day-to-day use, which will not be possible unless the training data set is representative of the future test points. From what is available in the literature, the current data sets are quite limited in terms of ecological validity, primarily due to resource constraints. Groups are working to fix these deficits.

Finally, the current use of random forests for probability estimation could be problematic in terms of asymptotic behavior, as noted in Section 4.2.1. Other probability estimation methods could be employed, but from the performance we have seen, the current majority vote method seems sufficient.

4.4.2 Discussion

Overall, our method provides a substantial contribution to the EWID field, because it enables the estimation of two latent variables: (1) the probability of accuracy for each eyewitness and (2) the probability of target presence or base rate for a given data set. It is not only applicable to EWID data, but to data from other fields that follow a similar structure. In comparison to an existing method of estimating base rate, it performs much more variably, but with increasing accuracy as the complexity of the data set increases.

A key component to the success of implementation is to obtain as much infor-

mation as possible in terms of both the system and estimator variables relevant to eyewitness lineups. Additionally, in order for implementation to take place for real use, ecologically valid data sets need to be collected to train a suitable model. This is a tremendous step to advancing the analysis of EWID data that needs to be supplemented by the psychology experts in the field, which is core to the original interdisciplinary motivation.

Part II

Errors-In-Variables and Random Forests

Chapter 5

Asymptotic Theory for Random Forests

5.1 Introduction

Random forests are generally thought of as "black box" methods, in which the predictions from such models are validated empirically through some hold-out validation data set. Users consider these models as black box since the model is built based on randomized inner components that are chosen without guidance from the user, resulting in an "opaque" implementation. Such empirical validation can cause issues when attempting to utilize more classical inference procedures such as hypothesis tests and confidence intervals on the predicted responses. The key to to evaluating results from random forest models in such a manner is by establishing an asymptotic framework. The investigation of the asymptotic behavior of random forests is a relatively new field and offers useful insights. Biau and Scornet (2015) provide a "guided tour" of the recent literature in random forests, covering topics such as the connection of random forests to nearest neighbors and kernels; resam-

pling mechanisms commonly found in random forests including CART; consistency; asymptotic normality; and variable importance measures. We review some of these topics below to set the foundation for the novel contributions to be discussed, by giving an abbreviated version of the tour from Biau and Scornet (2015). We further review the framework for asymptotic normality as established by Wager and Athey (2018), which will provide the foundation for the novel contributions in Chapter 6.

5.2 Random Forests

Random forests were originally developed by Breiman (2001) based on the CART algorithm (see Breiman et al., 1984). Although for some users, "random forests" simply means some aggregation of random decision trees, regardless of the algorithm, we will take the view that random forests refer to the original algorithm proposed by Breiman (2001).

5.2.1 Principles

Regression

The random forest follows the framework of nonparametric regression (i.e., regression tree) estimation with an observed input random vector $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^p$. That is, we observe p variables $\boldsymbol{X}$ that are a subset of real numbers in a p-dimensional vector space. The goal is to predict the square integrable random response $Y \in \mathbb{R}$. Assume we have some training sample $\mathcal{Z}_n = \big[(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\big]$ of independent random variables. Let $Z_i = (\boldsymbol{X}_i, Y_i)$. The model uses $\mathcal{Z}_n$ to create an estimate

$$\mu_n(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R} \tag{5.1}$$

of the function $\mu(\boldsymbol{x})$ for some test point $\boldsymbol{x}$. We are estimating the true conditional mean response function

$$\mu(x) = \mathrm{E}\left(Y \mid X = x\right) \tag{5.2}$$

The random forest predictor consists of a collection of M randomized regression trees. For the j-th tree in the collection, the prediction at test point x is denoted as $T(x; \theta_j; \mathcal{Z}_n)$ where $\theta_j \sim \Theta$ for $j = 1, \ldots, M$ is a variable to encompass randomness from resampling and the splitting procedure. This randomness is known as *auxiliary randomness* Wager and Athey (2018). Here, T is the regression tree used to estimate the conditional mean response function at x. Mathematically speaking, the j-th tree estimate is of the form

$$T(\boldsymbol{x}; \theta_j; \mathcal{Z}_n) = \sum_{i \in \mathcal{Z}_n^*(\theta_j)} \frac{\mathbb{1}_{\boldsymbol{X}_{i \in A_n(\boldsymbol{x}; \theta_j, \mathcal{Z}_n)}} Y_i}{N_n(\boldsymbol{x}; \theta_j, \mathcal{Z}_n)} \tag{5.3}$$

where $\mathcal{Z}_n^*(\theta_j)$ represents the set of data points selected before the construction of the tree, $A_n(\boldsymbol{x}; \theta_j, \mathcal{Z}_n)$ is the node containing x, and $N_n(\boldsymbol{x}; \theta_j, \mathcal{Z}_n)$ is the number of points that fall into $A_n(\boldsymbol{x}; \theta_j, \mathcal{Z}_n)$. This estimator is counting the number of times the observed value and the fitted value fall within the same node among all points that fall within that node. Some authors may also refer to the "node" of a tree as a "cell."

The M regression trees are combined to give the finite forest estimate

$$\mu_{M,n}(\boldsymbol{x}; \theta_j, \mathcal{Z}_n) = \frac{1}{M} \sum_{j=1}^{M} T(\boldsymbol{x}; \theta_j, \mathcal{Z}_n). \tag{5.4}$$

The R package `randomForest` (Liaw and Wiener, 2002) sets the default value for M at `ntree` $= 500$. In practice, M is only limited by computing resources. Allowing

M to tend to infinity gives the infinite forest estimate

$$\mu_{\infty,n}(\boldsymbol{x}; \mathcal{Z}_n) = \mathrm{E}_{\Theta}\big[\mu_n(\boldsymbol{x}; \theta; \mathcal{Z}_n)\big]. \tag{5.5}$$

Here, $\mathbb{E}_\Theta$ denotes the expectation with respect to θ given $\mathcal{Z}_N$. We marginalize over the auxiliary randomness. Scornet (2016a) showed that, in general, $M \to \infty$ is justified via the law of large numbers. Conditional on $\mathcal{Z}_n$, the following converges almost surely

$$\lim_{M\to\infty} \mu_{M,n}(\boldsymbol{x}; \theta_j, \mathcal{Z}_n) = \mu_{\infty,n}(\boldsymbol{x}; \mathcal{Z}_n). \tag{5.6}$$

Classification

The framework for regression random forests extends to supervised classification problems as well. For simplicity, let us assume the binary classification problem. The framework can inherently model multi-class problems as well. In the binary classification problem, the random response is $Y \in 0, 1$ and, given $\boldsymbol{X}$, the model classifies the responses Y. The classifier T is a Borel-measurable function of $\boldsymbol{X}$ and $\mathcal{Z}_n$ that labels Y and $\mathcal{Z}_n$. T is consistent if the conditional probability of error satisfies

$$L(T) = P\big[T(\boldsymbol{X}) \neq Y\big] \underset{n\to\infty}{\to} L^*, \tag{5.7}$$

where L^* is the error of the unknown, optimal Bayes classifier

$$T^*(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \mathrm{P}\left(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}\right) > \mathrm{P}\left(Y = 0 \mid \boldsymbol{X} = \boldsymbol{x}\right) \\ 0 & \text{otherwise} \end{cases}. \tag{5.8}$$

The random forest classifier is then obtained through a majority vote among the

M classification trees,

$$\mu_{M,n}(\boldsymbol{x};\theta_1,\ldots,\theta_M,\mathcal{Z}_n) = \begin{cases} 1 & \text{if } \frac{1}{M}\sum_{j=1}^{M} T(\boldsymbol{x};\theta_j,\mathcal{Z}_n) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}. \tag{5.9}$$

Suppose a node represents region A, then the randomized tree classifier becomes

$$T(\boldsymbol{x};\theta_j,\mathcal{Z}_n) = \begin{cases} 1 & \text{if } \sum_{i\in\mathcal{Z}_n^*(\theta)} \mathbf{1}_{\boldsymbol{X}_i\in A, Y_i=1} > \sum_{i\in\mathcal{Z}_n^I(\theta)} \mathbf{1}_{\boldsymbol{X}_i\in A, Y_i=0}, \boldsymbol{x}\in A, \\ 0 & \text{otherwise} \end{cases} \tag{5.10}$$

where $\mathcal{Z}_n^*(\theta)$ contains all of the data points chosen in the resampling step. In each node, a majority vote is taken over all points $Z_i = (\boldsymbol{X}_i, Y_i)$ for which $\boldsymbol{X}_i$ is in the same region A. By convention, any ties are broken in favor of class 0. Algorithm 2 easily adapts to classification by utilizing a slightly different CART-split criterion.

5.2.2 Algorithm

The random forest algorithm grows M different, randomized trees. The trees are grown from s_n observations drawn at random with[out] replacement from the full, original data set. If the samples are drawn with replacement, then there may be repeated observations. These s_n observations are used to construct each tree, and are redrawn for each tree. At each cell or node of the each tree, the sample is split by maximizing the CART-split criterion (described below) over `mtry` directions that are chosen uniformly from the full set of p covariates. Let $\mathcal{M}_{\text{try}}$ denote the resulting subset of chosen coordinates. Each tree completes its building process once each node contains fewer than $k =$ `nodesize` points. For any test point $\boldsymbol{x} \in \mathcal{X}$, each regression tree predicts the average of the Y_i that were in the drawn s_n points for

which the corresponding $\boldsymbol{X}_i$ falls into the node of $\boldsymbol{x}$. This process is summarized in Algorithm 2, which is adapted from Biau and Scornet (2015). In this algorithm, $\mathcal{P}$ is the list of chosen partitions to determine the resulting nodes. A similar algorithm can be constructed for supervised classification.

The three key parameters in the algorithm are:

1. $s_n \in 1, \ldots, n$: the number of sampled data points per tree;
2. $m = \texttt{mtry} \in 1, \ldots, p$: the number of possible directions (i.e., covariates) to split the sample at each node of each tree;
3. $k = \texttt{nodesize} \in 1, \ldots, s_n$: the number of observations in each node that below which node the splitting is terminated (i.e., the *terminal node*).

The terminology is given in terms of the `randomForest` R package (Liaw and Wiener, 2002). The default size for `mtry` is $\lceil p/3 \rceil$, s_n is set to n or the full size of the original data set, and `nodesize` is set to five for regression trees and one for classification trees. Here, $\lceil \cdot \rceil$ is the ceiling function.

The CART-split criterion maximizes the homogeneity in each child node from the parent node at each split-point. Let A be a generic node and $N_n(A)$ as the number of the data points that fall in A. A pair (c, d) is a cut in A, where $c \in 1, \ldots, p$ and d is the position of the cut along the c-th coordinate within the bounds of A. Define $\mathcal{C}_A$ be the set of all possible cuts in A. Let $\boldsymbol{X}_i = \left(\boldsymbol{X}_i^{(1)}, \ldots, \boldsymbol{X}_i^{(p)}\right)$ for any $(c, d) \in \mathcal{C}_A$, the CART-split criterion is

$$\begin{aligned} L_{\text{reg},n}(c, d) =& \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_A\right)^2 \mathbb{1}_{\boldsymbol{X}_i \in A} \\ & - \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_{\text{left}}} \mathbb{1}_{\boldsymbol{X}_i^{(c)} < d} - \bar{Y}_{A_{\text{right}}} \mathbb{1}_{\boldsymbol{X}_i^{(j)} \geq d}\right)^2 \mathbb{1}_{\boldsymbol{X}_i \in A} \end{aligned} \tag{5.11}$$

Algorithm 2: Breiman's random forest, predicted value at $\boldsymbol{x}$

Input: Training set $\mathcal{Z}_n$; number of trees $M > 0$; $s_n \in 1, \ldots, n$; `mtry` $\in 1, \ldots, p$; `nodesize` $\in 1, \ldots, s_n$; and $\boldsymbol{x} \in \mathcal{X}$

Output: Prediction of the random forest model at $\boldsymbol{x}$

for $j = 1, \ldots, M$ **do**

- Pick s_n points with or without replacement, uniformly in $\mathcal{Z}_n$. Use only s_n for the following steps.
- Set $\mathcal{P} = (\mathcal{X})$ as the list containing the node associated with the root of the tree.
- Set $\mathcal{P}_{\text{final}} = \varnothing$, an empty list.
- **while** $\mathcal{P} \neq \varnothing$ **do**
  - Let A be the first element in $\mathcal{P}$.
  - **if** *A contains fewer elements than* ***`nodesize`*** *points or if all $\boldsymbol{X}_i \in A$ are equal* **then**
    - Remove the node A from the list $\mathcal{P}$
    - $\mathcal{P}_{\text{final}} \leftarrow \text{Concatenate}(\mathcal{P}_{\text{final}},\ A)$
  - **else**
    - Select uniformly, without replacement, a subset of $\mathcal{M}_{\text{try}} \subset 1, \ldots, p$ of cardinality $m =$ `mtry`.
    - Select the best split in A by optimizing the CART-split criterion (described in the text) along the coordinates in $\mathcal{M}_{\text{try}}$.
    - Cut the parent node A according to the best split.
    - Call A_{left} and A_{right} the two resulting child nodes.
    - Remove node A from the list $\mathcal{P}$.
    - $\mathcal{P} \leftarrow \text{Concatenate}(\mathcal{P},\ A_{\text{left}},\ A_{\text{right}})$.
  - **end if**
- **end while**
- Compute the predicted value $T(\boldsymbol{x}; \theta_j, \mathcal{Z}_n)$ at $\boldsymbol{x}$ equal to the average of the Y_i falling in the node of $\boldsymbol{x}$ in partition $\mathcal{P}_{\text{final}}$.

end for

Compute the random forest estimate $\mu_{M,n}(\boldsymbol{x}; \theta_1, \ldots, \theta_M, \mathcal{Z}_n)$ at the test point $\boldsymbol{x}$ according to Equation 5.4.

where $A_{\text{left}} = \boldsymbol{x} \in A : \boldsymbol{x}^{(j)} < d$ for the left child node, $A_{\text{right}} = \boldsymbol{x} \in A : \boldsymbol{x}^{(j)} \geq d$ for the right child node, and $\bar{Y}_A$ is the average of the Y_i belonging to A. Let $\bar{Y}_A = 0$ if

no point $\boldsymbol{X}_i$ belongs to A.

$L_{\text{reg},n}(c,\ d)$ determines the total number of data points within a given node and averages the deviations between the observed value Y_i and the fitted value from the average of the node $\bar{Y}_A$ for three different nodes: the parent node and the two child (left and right) nodes. This is the normalized difference in empirical variance between the parent nodes and the children nodes before and after a cut is made. The chosen cut is made uniformly across the possible directions in $\mathcal{M}_{\text{try}}$, returning the best one. The best cut (c_n^*, d_n^*) is chosen by maximizing $L_{\text{reg},n}(c, d)$ over $\mathcal{M}_{\text{try}}$ and $\mathcal{C}_A$,

$$(c_n^*, d_n^*) \in \underset{\substack{j \in \mathcal{M}_{\text{try}} \\ (c,d) \in \mathcal{C}_A}}{\arg\max}\, L_{\text{reg},n}(c, d). \tag{5.12}$$

The criterion for classification follows a similar framework. Define $p_{0,n}(A)$ and $p_{1,n}(A)$ as the empirical probability of a data point in node A having label 0 and 1, respectively. Then, for any $(c, d) \in \mathcal{C}_A$, the classification CART-split criterion (Breiman et al., 1984) is

$$L_{\text{class},n}(c, d) = p_{0,n}(A) \cdot p_{1,n}(A) - \frac{N_n(A_L)}{N_n(A)} \cdot p_{0,n}(A_L) \cdot p_{1,n}(A_L) \tag{5.13}$$

$$- \frac{N_n(A_R)}{N_n(A)} \cdot p_{0,n}(A_R) \cdot p_{1,n}(A_R). \tag{5.14}$$

This criterion $L_{\text{class},n}$ is based on the Gini impurity measure $2 \cdot p_{0,n}(A) \cdot p_{1,n}(A)$. In `randomForest` (Liaw and Wiener, 2002), the default values are `nodesize` $= 1$ and `mtry` $= \sqrt{p}$.

5.3 Asymptotic Normality

Two separate groups showed that random forests are asymptotically normal, first by Mentch and Hooker (2016); then by Wager and Athey (2018). The underlying approach of the two follow similar structures, both based on Hájek projections (Hájek, 1968), but use different assumptions resulting in different rates of convergence. Both rely on a simplified version of Breiman's random forest procedure (Breiman, 2001). Instead of assuming bootstrap resampling for each tree, these authors assume proper subsampling for each tree, where each observation may not be chosen more than once. From this point forward, the random forests will be assumed to be by training trees T on subsamples of size s out of n possible observations, drawn without replacement. Note that this is implemented practically in `R` by setting the option `replace = FALSE` in `randomForest` (Liaw and Wiener, 2002).

The bootstrap resampling scheme is often replaced by a subsampling one due to the resistant nature of the bootstrap to classical statistical methods. While the bootstrap is simplistic in its practical implementation, the asymptotic behavior of the bootstrap is often unpredictable. For example, Friedman and Hall (2007) derived some properties by decomposing bagged predictors. However, in application, given a linear model, the behavior of the variance was incongruous with what was expected from the theoretical derivations. In particular, the distribution of the bootstrap sample $\mathcal{B}_n^*$ is different from the original distribution $\mathcal{B}_n$ (Biau and Scornet, 2015). For example, suppose there exists some random variable $\boldsymbol{X}$ with some density f. Whenever observations are sampled with replacement with some probability $p > 0$, at least one observation will be chosen more than once. Thus, with positive probability, there will be two identical data points in $\mathcal{B}_n^*$ and the distribution of $\mathcal{B}_n^*$ cannot be absolutely continuous. Many of the asymptotic properties for random

forests rely on Lipschitz-continuity, which is not guaranteed for bootstrap samples.

Wager and Athey (2018) assume a random forest that averages trees trained over all possible size-s subsamples of the training data $\mathcal{Z}_n$, marginalizing over the the noise (i.e., auxiliary randomness) θ. The forest is computed by Monte Carlo averaging,

$$\mu_{M,n}(\boldsymbol{x}; \mathcal{Z}_n) \approx \frac{1}{M} \sum_{j=1}^{M} T(\boldsymbol{x}; \theta_j^*, Z_{j1}^*, \ldots, Z_{js}^*) \tag{5.15}$$

where $Z_{j1}^*, \ldots, Z_{js}^*$ is drawn without replacement from $Z_1, \ldots, Z_s$ and θ_j^* is a random draw from Θ. The usual application of random forests use the auxiliary randomness θ is used to randomly regulate the number of covariates on which the trees can split at any of the training steps. For each step, m features are randomly chosen from the full set of p possible covariates, and the tree predictor splits on one of these m features. If $m = p$, then the tree will always be able to split on any feature, and the random forest converts to a bagged tree. If $m = 1$, then the tree is completely restricted to one covariate on which to split.

Recall that Mentch and Hooker (2016) also establish asymptotic normality, which assumes stronger conditions than that of Wager and Athey (2018). Mentch and Hooker (2016) require the subsample size to grow slower than $\sqrt{n}$ (i.e., $\frac{s_n}{\sqrt{n}} \to 0$). Wager and Athey (2018) note that the random forests are not generally asymptotically biased. Assuming the number of covariates $p = 2$ and $\mu(x) = \|x\|_1$, evaluate this forest at $x = 0$. It can be shown that the bias of the forest decays as $\frac{1}{\sqrt{s_n}}$, while the variance decays as $\frac{s_n}{n}$. If $\frac{s_n}{\sqrt{n}} \to 0$, the squared bias decays slower than the variance. Thus, any confidence interval built using the asymptotic normality framework of Mentch and Hooker (2016) will not provide coverage for $\mu(x)$.

We follow the framework from Wager and Athey (2018) for establishing asymptotic normality.

5.3.1 Assumptions and Definitions

We begin by providing a formal definition for a random forest. The random forest defined in Definition 5.3.1 is a random kernel U-statistic (Mentch and Hooker, 2016; Wager and Athey, 2018).

Definition 5.3.1. The *random forest* with tree T and subsample size s is

$$\mu(\boldsymbol{x}; \mathcal{Z}_n) = \binom{n}{s}^{-1} \sum_{1 \le i_1 < \ldots < i_s \le n} E_{\theta \sim \Theta}\big[T(\boldsymbol{x}; \theta_j, Z_{i_1}, \ldots, Z_{i_s})\big] \tag{5.16}$$

The trees T in the forest must be *honest*, *random-split*, α*-regular*, and *symmetric*, as defined in Definition 5.3.2 to Definition 5.3.5, respectively.

Definition 5.3.2. A tree grown using training sample $Z_i = (\boldsymbol{X}_i, Y_i)$ for $i = 1, \ldots, s$ is *honest* if the tree, conditionally on $\boldsymbol{X}_i$, does not use the responses Y_i when deciding where to place its splits.

The concept of honesty stems from the similarity between random forests and adaptive nearest neighbors (ANN), as detailed by Lin and Jeon (2006). An honest tree does not reuse the training response values Y_i for both choosing split-points of the tree and for prediction. In other words, the tree is grown using one subsample, while the predictions at the nodes of the tree are estimated using a different subsample. If the condition is not required, then arbitrarily biased trees can be easily constructed. This ensures that the split criterion used to identify the selection variable S_i is independent of Y_i conditional on $\boldsymbol{X}_i$. Wager (2016) notes that an easy way to enforce honesty is to divide the training points into a set of structure points $\mathcal{S}$ that are only used to pick the split-points and a set of prediction points $\mathcal{P}$ that are only used to make predictions once the splits are chosen.

Empirically, honest trees are unbiased regardless of the number of observations n. Note, that CART trees are not honest by Definition 5.3.2, and result in biased output. The bias seems to increase with n, since CART trees are assertively separating outliers from the other data, thereby pushing outliers into far corners of the feature space and increasing bias. In general, it seems that the bias of CART trees is not necessarily that intrusive in the performance of random forests. As long as the minimum `nodesize` grows with n, the phenomenon of growing bias is avoided. Thus, even with the bias, Wager (2016) notes that the asymptotic results still provide valuable insight into understanding the behavior of CART random forests. The results still work well in practice. We move forward with the framework.

Definition 5.3.3. A tree is *random-split* if at every step of the tree-growing process, marginalizing over θ, the probability that the next split occurs along the j-th feature is bounded from below by $\frac{\pi}{p}$ for some $0 < \pi \leq 1$ for all $j = 1, \ldots, p$.

The random-split conditions guarantees consistency, by ensuring that the nodes of the trees becomes small across all dimensions of the covariate space as the number of observations $n \to \infty$. The condition forces each variable to be chosen with some randomness during the tree building process. The condition is also mentioned by Meinshausen (2006) in his development of quantile regression trees.

Definition 5.3.4. A tree grown via recursive partitioning is α-*regular* for some $\alpha > 0$ if each split apportions at least a fraction α of the available training observations on either side of the split (i.e., for each child node). Trees are fully grown to depth k for $k \in \mathbb{N}$. Thus, each terminal node of the grown tree has between k and $2k - 1$ observations.

The concept of α-regularity provides control over the shape of the nodes in T,

which ensures that predictions from random forests are local as the depth of the tree increases.

Definition 5.3.5. A tree is *symmetric* if the (possibly randomized) output of the predictor does not depend on the order in which the training examples are indexed for $i = 1, 2, \ldots, n$.

Symmetry of trees allows the establishment of asymptotic normality for random forests.

5.3.2 Central Limit Theorem

The seminal result from Wager and Athey (2018) is Theorem 5.3.1 establishing asymptotic normality for random forests (see Wager and Athey, 2018, theorem 1).

Theorem 5.3.1. *Suppose there are n independent and identically distributed training observations $Z_i = (\boldsymbol{X}_i,\ Y_i) \in [0,1]^p \times \mathbb{R}$. Moreover, suppose that the covariates are independent and uniformly distributed $\boldsymbol{X}_i \sim Uniform\,([0,1]^p)$. Suppose also that $\mu(\boldsymbol{x}) = E(Y \mid \boldsymbol{X} = \boldsymbol{x})$ and $\mu_2(\boldsymbol{x}) = E(Y^2 \mid \boldsymbol{X} = \boldsymbol{x})$ are Lipschitz-continuous, that*

$$Var\,(Y \mid \boldsymbol{X} = \boldsymbol{x}) > 0, \tag{5.17}$$

and that

$$E\left[\left|Y - E(Y \mid \boldsymbol{X} = \boldsymbol{x})\right]\right|^{2+\delta} \Big|\ \boldsymbol{X} = \boldsymbol{x}\Big] \leq Q \tag{5.18}$$

for some constants $\delta,\ Q > 0$ uniformly over all $\boldsymbol{x} \in [0,1]^p$. Given this process for generating data, let T be an honest, α-regular, and symmetric random-split tree as defined in Definition 5.3.2 to Definition 5.3.5 with $\alpha \leq 0.2$. Let $\hat{\mu}(\boldsymbol{x})$ be the estimate for $\mu(\boldsymbol{x})$ given by a random forest with trees T and a subsample size s_n.

Finally, suppose that the subsample size s_n scales as

$$s_n \asymp n^{\beta} \text{ for some } \beta_{\min} \coloneqq 1 - \left(1 + \frac{p}{\pi} \cdot \frac{\log(\alpha^{-1})}{\log[(1-\alpha)^{-1}]}\right) \tag{5.19}$$

Then, random forest predictions are asymptotically normal,

$$\frac{\hat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x})}{\sigma_n(\boldsymbol{x})} \to N(0,1) \text{ for a sequence } \sigma_n(\boldsymbol{x}) \to 0 \tag{5.20}$$

Here, $N(0,1)$ represents the standard normal distribution. Moreover, the asymptotic variance σ_n can be consistently estimated using the infinitesimal jackknife (IFJ),

$$\frac{\hat{V}_{IFJ}(\boldsymbol{x})}{\sigma_n^2(\boldsymbol{x})} \xrightarrow{p} 1 \tag{5.21}$$

where

$$\hat{V}_{IFJ}(\boldsymbol{x}) = \frac{n-1}{n}\left(\frac{n}{n-s}\right)^2 \sum_{i=1}^{n} Cov_*\left[\hat{\mu}_m^*(\boldsymbol{x}), Y_{im}^*\right]^2. \tag{5.22}$$

The covariance is taken with respect to the set of all the trees $m = 1, \ldots, M$. In this case, Y_{im}^* represents the number of times the i-th training observation appears in the m resamples (i.e., m trees).

By assuming independent and identical distributions, the observations are assumed to be random, independent draws from some underlying distribution. Independent and uniformly distributed covariates allows for the bias of the terminal nodes to be bounded. Lipschitz-continuity ensures that the differentiable expectations have a bounded derivative, and thus, are differentiable. The IFJ estimate $\hat{V}_{\text{IFJ}}$ is based on the idea of the nonparametric delta-method estimate of standard deviation derived by Efron (2014).

The authors remark that Theorem 5.3.1 holds for binary classification random

forests with `nodesize` = 1. In this case, the output $\mu(\boldsymbol{x})$ of the random forest is the estimate for the probability $\mathrm{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})$. Theorem 5.3.1 allows for the construction of confidence intervals about this probability. If `nodesize` > 1, then Theorem 5.3.1 holds if the trees are built by averaging observations within a node, but not if the predictions are made via majority vote, as is the case for `randomForest` (Liaw and Wiener, 2002). They note that future work could focus on establishing a central limit theorem (CLT) for classification random forests by majority vote.

We follow, at a high-level, the proof of the results in Theorem 5.3.1, as shown by Wager and Athey (2018). Technical details are fully provided in their work.

Bounding the Bias

Wager and Athey (2018) begin by bounding the bias of regression trees

$$RF_{\text{bias}} = E\big[\hat{\mu}_n(x) - \mu(x)\big] \tag{5.23}$$

by showing that as the subsample size s becomes large, the node sizes become small. They utilize the diameter of the node, which acts as the pathway for a particular observation down a tree. The diameter with respect to the j-th axis is the pathway for nodes containing the j-th covariate, to derive an upper bound for the bias of a single tree. The diameter $\text{diam}[L(x)]$ of a node $L(x)$ is the length of the longest segment within $L(x)$. Similarly, $\text{diam}_j[L(x)]$ is the length of the longest such segment that is parallel to the j-th axis. The derivation for this bound employs the α-regularity condition from Definition 5.3.4 and Chernoff's inequality.

They establish that, for $\alpha \leq 0.2$, the bias of the random forest is bounded by

$$\left|\mathrm{E}\left[\hat{\mu}(\boldsymbol{x})\right] - \mu(\boldsymbol{x})\right| = \mathcal{O}\left(s^{\frac{1}{2}\cdot\frac{\log[(1-\alpha)^{-1}]}{\log(\alpha^{-1})}\cdot\frac{\pi}{p}}\right) \tag{5.24}$$

Bounding the bias ensures the predictions will converge at the rate established in Equation 5.24, which is necessary to establish the consistency of the estimator $\hat{\mu}(\boldsymbol{x})$.

U-Statistics and Hájek Projections

Wager and Athey (2018) continue to lay the foundation for asymptotic normality of random forests by building upon the idea of asymptotically normal U-statistics from Hoeffding (1948). Mentch and Hooker (2016) provides a connection of U-statistics to random forests, which is also utilized by Wager and Athey (2018). The authors reference Lee (2019) for a more complete treatment of U-statistics. See Appendix D for a brief introduction to U-statistics.

Specifically, Wager and Athey (2018) and Mentch and Hooker (2016), use the Hájek projection of a random variable, which is the projection of the random variable onto the set of sums $\sum_{i=1}^{n} g_i(X_i)$ of measurable functions satisfying $\mathrm{E}\left[g_i(X_i)\right]^2 < \infty$. The Hájek projection guarantees asymptotic normality if the ratio of the projection $\mathring{T}$ and the original predictor T tends to 1 since $\mathring{T}$ is the sum of independent random variables. The Hájek projection is used as a theoretical tool to establish normality of some statistic T_n by comparing it with another statistic $\mathring{T}_n$ that is known to be asymptotically normal by showing that

$$\mathrm{E}\left(T_n - \mathring{T}_n\right)^2 \to 0. \tag{5.25}$$

Suppose there exists a predictor T and a set of independent training observations

$Z_1 \ldots, Z_n$. The Hájek projection of T onto $\mathring{T}$ is defined as

$$\begin{aligned} \mathring{T} &= \sum_{i=1}^{n} \Big[\mathrm{E}\left(T|Z_i\right) \Big] - (n-1)\mathrm{E}\left(T\right) \\ &= \mathrm{E}\left(T\right) + \sum_{i=1}^{n} \Big[\mathrm{E}\left(T|Z_i\right) - \mathrm{E}\left(T\right) \Big], \end{aligned} \tag{5.26}$$

given that T has a finite second moment. Classically, $\mathrm{Var}\left(\mathring{T}\right) \leq \mathrm{Var}\left(T\right)$ and,

$$\lim_{n\to\infty} \frac{\mathrm{Var}\left(\mathring{T}\right)}{\mathrm{Var}\left(T\right)} = 1 \tag{5.27}$$

implies that

$$\lim_{n\to\infty} \frac{E\left(\|\mathring{T} - T\|_2^2 \right)}{\mathrm{Var}\left(T\right)} = 0. \tag{5.28}$$

The condition from Equation 5.26 and Equation 5.27 do not apply directly to regression trees, and must be modified. Wager and Athey (2018) argue that if the tree T is 1-incremental, as defined in Definition 5.3.6, then the condition in Equation 5.26 and Equation 5.27 can be used. The idea of 1-incremental is a specific case of ν-incremental where $\nu = 1$, and is a weaker version of Equation 5.27. Establishing ν-incrementality shows that the Hájek projection $\mathring{T}$ of T retains some of the variation of T.

Definition 5.3.6. The predictor T is $\nu(s)$-incremental at $\boldsymbol{x}$ if

$$\liminf_{s\to\infty} \frac{\mathrm{Var}\left[\mathring{T}(\boldsymbol{x}; Z_1, \ldots, Z_s)\right] \;/\; \mathrm{Var}\left[T(\boldsymbol{x}; Z_1, \ldots, Z_s)\right]}{\nu(s)} \geq 1. \tag{5.29}$$

They establish the condition of ν-incrementality under the framework of predictive nearest neighbors (PNN), following the lead of Lin and Jeon (2006) and Biau and Devroye (2010). A predictor T is a k-PNN predictor over training observations

$\mathcal{Z}$ if T outputs the average of the responses Y_i over a k-PNN set of x. This states that any decision tree T that makes axis-aligned splits (i.e., the previously described splitting criterion) and has nodes of size between k and $2k-1$ (i.e., α-regular trees) is a k-PNN predictor. The original CART trees from Breiman et al. (1984) are k-PNN predictors. All k-PNN predictors can be written as

$$T(\boldsymbol{x}; \theta; Z_1, \ldots, Z_s) = \sum_{i=1}^{s} S_i Y_i \tag{5.30}$$

where the selection variable S_i is

$$S_i = \begin{cases} \frac{1}{\left|\{i : \boldsymbol{X}_i \in L(\boldsymbol{x})\}\right|} & i \in L(\boldsymbol{x}) \\ 0 & \text{otherwise} \end{cases}. \tag{5.31}$$

For honest trees T, S_i is independent of Y_i, conditional on $\boldsymbol{X}_i$ for each i. Wager and Athey (2018) establish a lower bound for $s \cdot \text{Var}\left[\text{E}\left(S_1 \mid Z_1\right)\right]$, which gives an idea on if S_i is non-zero even if only Z_i is observed. The bound is given as

$$\liminf_{s \to \infty} \frac{s \cdot \text{Var}\left[\text{E}\left(S_1 \mid Z_1\right)\right]}{\frac{1}{k} \cdot \frac{C_{f,p}}{\log(s)^p}} \geq 1 \tag{5.32}$$

where f is a density bounded away from infinity and p is the covariates in the p-dimensional space. When f is uniform over $[0,1]^p$, then the bound holds with $C_{f,d} = \frac{(d-1)!}{2^{d+1}}$. This expression provides a lower bound on how much information about S_1 is contained in Z_1. A tree T is ν-incremental at $\boldsymbol{x}$ with

$$\nu(s) = \frac{C_{f,p}}{\log(s)^p} \tag{5.33}$$

where T is an honest, α-regular, symmetric tree with Lipschitz-continuous $\mu(\boldsymbol{x})$ and

$\mu_2(\boldsymbol{x})$ and $\text{Var}\left(Y \mid \boldsymbol{X} = \boldsymbol{x}\right) > 0$.

Following the analysis of variance (ANOVA) decomposition from Efron and Stein (1981), Wager and Athey (2018) provide a bound for the variance of tree T given the Hájek projection,

$$E\left[\left(\hat{\mu}(\boldsymbol{x}) - \mathring{\hat{\mu}}(\boldsymbol{x})\right)^2\right] \leq \left(\frac{s}{n}\right)^2 \cdot \text{Var}\left[T(\boldsymbol{x}; \theta, Z_1, \ldots, Z_s)\right] \tag{5.34}$$

Thus, as long as the subsample size s_n satisfies

$$\lim_{n\to\infty} s_n = \infty \tag{5.35}$$

and

$$\lim_{n\to\infty} s_n \cdot \frac{\log(n)^p}{n} = 0 \tag{5.36}$$

and

$$\text{E}\left[\left|Y - \text{E}\left(Y|\boldsymbol{X} = \boldsymbol{x}\right)\right|^{2+\delta} \;\middle|\; \boldsymbol{X} = \boldsymbol{x}\right] \leq Q \tag{5.37}$$

for some constants $\delta, Q > 0$, uniformly over all $\boldsymbol{x} \in [0,1]^p$. Then, there exists some sequence $\sigma_n(\boldsymbol{x}) \to 0$ such that

$$\frac{\hat{\mu}_n(\boldsymbol{x}) - \text{E}\left[\hat{\mu}_n(\boldsymbol{x})\right]}{\sigma_n(\boldsymbol{x})} \xrightarrow{d} N\left(0, 1\right) \tag{5.38}$$

and

$$\frac{\hat{V}_{\text{IFJ}}(\boldsymbol{x}; Z_1, \ldots, Z_n)}{\sigma_n^2(\boldsymbol{x})} \xrightarrow{p} 1. \tag{5.39}$$

Thus, this result confirms Theorem 5.3.1. Wager and Athey (2018) check Lyapunov-style[1] conditions for the CLT established above. The Lyapunov-style conditions

[1] Also spelled Liapounov.

assume for a set of random variables X_i for $i = 1, \ldots, n$ with means $E(X_i) = \xi_i$, variances σ_i^2, and finite third moments that

$$Y_n = \frac{\bar{X} - E(\bar{X})}{\sqrt{\text{Var}\,(\bar{X})}} \xrightarrow{d} N\,(0, 1) \tag{5.40}$$

provided

$$\left[E\left(\sum_{i=1}^{n} |X_i - \xi_i|^3\right)\right]^2 = o\left[\left(\sum_{i=1}^{n} \sigma_i^2\right)^3\right]. \tag{5.41}$$

Bias Correction

A bias correction (i.e., finite sample correction) for a finite number of trees is given as $\frac{n-1}{n}\left(\frac{n}{n-2}\right)^2$. This is motivated by looking the case of trivial trees that do not make any splits $T(\boldsymbol{x}; \theta; Z_{i_1}, \ldots, Z_{i_s}) = \frac{1}{s}\sum_{j=1}^{s} Y_{i_j}$. The full random forest in this case is simply

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{5.42}$$

with variance estimator

$$\hat{V}_{\text{trivial}} = \frac{1}{n(n-1)}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \tag{5.43}$$

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. $\hat{V}_{\text{trivial}}$ is established to be unbiased for Var $(\hat{\mu})$. In the case of trivial trees, we can show E $\left(\hat{V}_{\text{IFJ}}\right) =$ Var $(\hat{\mu})$ (see Wager and Athey, 2018, Proposition 10).

5.3.3 Infinitesimal Jackknife

The IFJ is a modification of the standard Quenouille-Tukey jackknife introduced by Jaeckel (1972). In the standard jackknife procedure, the sample follows a leave-one-

out (LOO) procedure, where the i-th observation is removed and the procedure is run on the remaining $n-1$ observations. All n observations are subjected to the same treatment, resulting in n estimates, which provides a sampling distribution for a specified statistic. The removed observation has no weight. The IFJ gives the removed observation slightly less weight than the kept observations rather than no weight at all. The removed observation is given a weight of $\frac{n-1}{n}$ instead of the usual weight of 0.

Define a functional statistic $\hat{\eta} = \eta(\hat{F})$ with empirical distribution $\hat{F}$. Let $\boldsymbol{p}^* = (p_1^*, \ldots, p_M^*)$ be a resampling vector, which can be any probability vector. Each $\boldsymbol{p}^*$ has some re-weighted empirical probability distribution $\hat{F}^*$, which is the mass p_i^* on some observations Z. Then, $\boldsymbol{p}^*$ is any vector on an M-dimensional simplex,

$$\mathcal{L}_n = \left\{ \boldsymbol{p}^* : p_i^* \geq 0, \sum_{i=1}^{n} p_i^* = 1 \right\}. \tag{5.44}$$

A simplex is the generalization of a tetrahedral region of space to M dimensions. It is also known as a hypertetrahedron.

Let $\boldsymbol{z} = (z_1, \ldots, z_M)$. Under $\hat{F}$, z can take on n distinct values z_i, each with probability $\frac{1}{M}$. For each $i \in \{1, \ldots, M\}$, define a random variable p_i^*

$$\boldsymbol{Y}_i^* = \frac{1}{n}\{\text{number of times } z_i \text{ occurs in the sample}\} \tag{5.45}$$

as a random sample of size n from $\hat{F}$. The vector $(nY_1^*, \ldots, nY_M^*)$ has a multinomial distribution and $\sum_{i=1}^{M} Y_i^* = 1$. Thus, in general, the resampling vectors are selected using a rescaled multinomial distribution,

$$\boldsymbol{Y}^* \sim \frac{\text{Multinomial}_n(M, \boldsymbol{p}^\circ)}{n}, \tag{5.46}$$

with n independent draws on M categories that each has probability $\frac{1}{M}$. Here, $\boldsymbol{p}^\circ = \left(\frac{1}{M}, \ldots, \frac{1}{M}\right)$, such that $\hat{\eta}^* = \eta\big[\hat{F}(\boldsymbol{p}^\circ)\big] = \hat{\eta}(\boldsymbol{p}^\circ)$. That is, $\hat{F}(\boldsymbol{p})$ is the distribution as specified in Equation 5.46 evaluated at probability vector $\boldsymbol{p}^\circ$.

We can verify that

$$\begin{aligned} \mathrm{E}\left(Y_i^*\right) &= \frac{1}{M}, \\ \mathrm{Var}\left(Y_i^*\right) &= \frac{1}{nM}\left(1 - \frac{1}{M}\right), \text{ and} \\ \mathrm{Cov}\left(Y_i^*, Y_j^*\right) &= -\frac{1}{nM^2}. \end{aligned} \tag{5.47}$$

Then, as $n \to \infty$, the vector

$$\sqrt{n}(\hat{\boldsymbol{Y}}^* - \boldsymbol{Y}^*) = \left[\sqrt{n}\left(Y_1^* - \frac{1}{n}\right), \ldots, \sqrt{n}\left(Y_n^* - \frac{1}{n}\right)\right] \tag{5.48}$$

has a multivariate normal limiting distribution, for which all means are zero, all variances are $\frac{1}{M}\left(1 - \frac{1}{M}\right)$, and all covariances are $-\frac{1}{M^2}$. Assume η is defined for discrete distributions, which assign arbitrary non-negative weights $\boldsymbol{B}_i^*$ to the z_i. We can write the estimate as

$$\hat{\eta} = \eta\big[\hat{F}(\boldsymbol{p}^*)\big] = \eta(\boldsymbol{z}, \hat{\boldsymbol{Y}}^*) \tag{5.49}$$

and

$$\eta = \eta\big[F(\boldsymbol{p}^\circ)\big] = \eta(\boldsymbol{z}, \boldsymbol{Y}^*). \tag{5.50}$$

Since the z_i are fixed, $\hat{\eta}(\boldsymbol{z}, \hat{\boldsymbol{Y}}^*)$ is a function of the M variables $\hat{Y}_i^*$. In Equation 5.49, we assume that $\hat{\eta}$ is differentiable with respect to $\hat{Y}_i^*$, so that

$$D_i = \left.\frac{\partial \hat{\eta}}{\partial Y_i^*}\right|_{\boldsymbol{Y}^* = \hat{\boldsymbol{Y}}^*} \tag{5.51}$$

and

$$D_{ij} = \left. \frac{\partial^2 \hat{\eta}}{\partial Y_i^* \, \partial Y_j^*} \right|_{\boldsymbol{Y}^* = \hat{\boldsymbol{Y}}^*}. \tag{5.52}$$

Also, define

$$\hat{D}_i = \left. \frac{\partial \hat{\eta}}{\partial Y_i^*} \right|_{\boldsymbol{Y} = \hat{\boldsymbol{Y}}^*}. \tag{5.53}$$

Now, by Jaeckel (1972), the IFJ variance is

$$V = \frac{1}{M} \sum_{j=1}^{M} D_j^2, \tag{5.54}$$

which, by Theorem 1 in Jaeckel (1972), can be estimated by $\hat{V}$

$$n\hat{V} = \frac{1}{n} \sum_{i=1}^{n} \hat{D}_i^2 \tag{5.55}$$

since $n\hat{V} \xrightarrow{p} V$, providing a consistent estimator for V.

Efron (2014) provides a theorem to estimate Equation 5.55 by deriving a non-parametric delta-method estimate of standard deviation for the ideal smoothed bootstrap statistic $s(\boldsymbol{z}) = \sum_{i=1}^{R} \frac{\hat{\eta}}{R}$ for R bootstrap replicates. He shows that there exists a relationship between the bootstrap and the IFJ. We approximate $\hat{\eta}$ with the hyperplane tangent $\hat{\eta}_{\text{tan}}$ to the surface $\hat{\eta}$ at the point $\boldsymbol{Y}^* = \boldsymbol{Y}^\circ$ instead of using $\hat{\eta}_{\text{lin}}$,

$$\hat{\eta}_{\,\text{lin}}(\hat{F}) = \hat{\eta}_{(\cdot)} + \left(\hat{F} - F\right) \boldsymbol{U}, \tag{5.56}$$

where

$$U_i = (n-1)\left(\hat{\eta}_{(\cdot)} - \hat{\eta}_{(i)}\right), \tag{5.57}$$

for $i = 1, \ldots, n$. The estimate for the standard deviation is

$$\widehat{\mathrm{SD}} = \mathrm{Var}\,(\hat{F}) \cdot \hat{\eta}_{\tan}(\hat{F}) \tag{5.58}$$

where $\hat{\eta}_{\tan}(\cdot)$ is

$$\hat{\eta}_{\tan}(\hat{F}) = \hat{\eta}(F) + (\hat{F} - F)\boldsymbol{U} \tag{5.59}$$

and

$$\begin{aligned} \dot{U}_i &= \lim_{\varepsilon \to 0} \frac{\hat{\eta}\Big[F + \varepsilon\big(\boldsymbol{\delta}_i - F\big)\Big] - \hat{\eta}\,(F)}{\varepsilon} \\ &= \frac{d}{d\varepsilon}\ \hat{\eta}\Big[F + \varepsilon\big(\boldsymbol{\delta}_i - F\big)\Big]\ \Big|_{\varepsilon=0} \\ &= \frac{\partial}{\partial \varepsilon} \hat{\eta}\Big[\hat{F}_i(\varepsilon)\Big]. \end{aligned} \tag{5.60}$$

The defined $\dot{U}_i$ is equivalent to the partial derivatives $\hat{D}_i$ from Equation 5.53 as defined by Jaeckel (1972).

Here, $\boldsymbol{\delta}_i$ is the i-th coordinate vector of the probability mass on the i-th coordinate. The U_i are directional derivatives, and are also known as influence function (IF). The IF measures the rate at which the functional $\hat{\eta}$ changes when F is contaminated with a small probability of picking up an observation i. This provides a measure of the influence of the contamination. It measures the influence of a small proportion of observations at i that are not a "part of" F, which Lehmann (2004) calls "gross errors." Thus, the IFJ resamples $\hat{\eta}$ at $\hat{F}$ values infinitesimally close to F, rather than the $O\left(\frac{1}{n}\right)$ that the ordinary jackknife uses.

The next step is to connect the results from Jaeckel (1972) to the estimation procedure from Wager et al. (2014), which was based on an estimation procedure

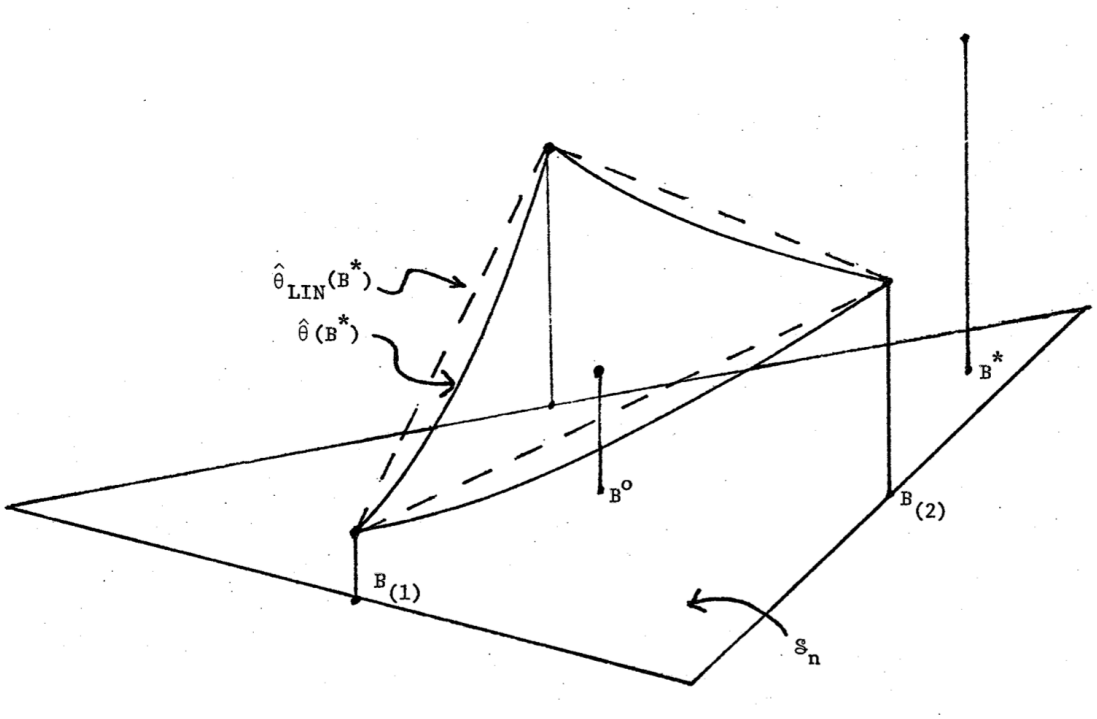


Figure 5.1: Illustrative representation of $\hat{\eta}(\boldsymbol{B}^*)$ as a function on the simplex $\boldsymbol{\mathcal{L}}_n$, where $\boldsymbol{B}^*$ are resamples. In this specific case, $\boldsymbol{B}^*$ are bootstrap resamples, but Efron (2014) shows that bootstrap resamples are directly related to IFJ resamples. The curved surface $\hat{\eta}(\cdot)$ is approximated by the linear function $\hat{\eta}_{\text{lin}}(\cdot)$. Visualization is taken from Efron (2014).

from Efron (2014). Efron (1982) derives the IFJ estimate of standard deviation as

$$\widehat{\text{SD}}_{\text{IFJ}}(\hat{\eta}) = \frac{1}{M}\sqrt{\sum_{i=1}^{M} \dot{U}_i^2}, \tag{5.61}$$

which is equivalent to the derivation Equation 5.54 from Jaeckel (1972). Efron (2014) estimates Equation 5.61 as

$$\hat{V} = \sum_{i=1}^{M} \text{Cov}_*\left[Y_i^*, \hat{\eta}\right]^2 \tag{5.62}$$

where $\text{Cov}_*\left[Y_i^*, \hat{\eta}\right]$ is the covariance between the estimate $\hat{\eta}$ and the number of

times Y_i^* the i-th training observation appears in a resample. We need to show Equation 5.61 is equal to $\text{Cov}_*\big[Y_i^*, \hat{\eta}\big]$ to show Equation 5.62. Define $w_i(\boldsymbol{p})$ as the ratio of probabilities under Y_i^* under Equation 5.46, for probability vectors $\boldsymbol{p}^*$ and $\boldsymbol{p}^\circ$,

$$w_i(\boldsymbol{p}) = \prod_{k=1}^{M} (np_k)^{Y_k^*}, \tag{5.63}$$

so that

$$E(\hat{\eta}) = \sum_{i=1}^{R} \frac{w_i(\boldsymbol{p}) \cdot \hat{\eta}}{R}. \tag{5.64}$$

We include the factor $\frac{1}{R}$ to express that under $\boldsymbol{p}^\circ$, all of the Y_i^* random variables have probability $\frac{1}{R} = \frac{1}{M^n}$. Following Equation 5.60, we have $F(\varepsilon) = F + \varepsilon(\boldsymbol{\delta}_i - F)$, then

$$w_i(\boldsymbol{p}) = \big[1 + (n-1)\varepsilon\big]^{Y_i^*} \cdot (1-\varepsilon)^{\sum_{i \neq k} Y_{ik}^*}. \tag{5.65}$$

Let $\varepsilon \to 0$, then

$$w_i(\boldsymbol{p}) \doteq 1 + n\varepsilon(Y_i^* - 1), \tag{5.66}$$

since $\sum_{i=1}^{M} Y_i^* = 1$.[2] Combining Equation 5.64 and Equation 5.66 yields

$$E(\hat{\eta}) \doteq \sum_{i=1}^{R} \frac{\big[1 + \varepsilon(Y_i^* - 1)\big] \cdot \hat{\eta}}{R} \tag{5.67}$$

$$= E\Big\{\eta + n \cdot \varepsilon \cdot \text{Cov}_*\big[Y_i^*, \hat{\eta}\big]\Big\}. \tag{5.68}$$

From the definition of the directional derivative from Equation 5.60, we have

$$\dot{U}_i = n \cdot \text{Cov}_*\big[Y_i^*, \hat{\eta}\big], \tag{5.69}$$

[2] The equal sign with a dot above $\doteq$ means that the series converges in mean to some function $f(\cdot)$.

which establishes Equation 5.62. Wager et al. (2014) provides the connection between the IJK variance estimator and the Hájek projections.

Practically, we can transform samples from the empirical distribution $\hat{F}$ into samples from $\hat{F}_i(\varepsilon)$ in the following way:

1. Let $z_1^*, \ldots, z_n^*$ be a sample from $\hat{F}$;
2. Progressing through the entire sample and, independently for each j, take z_j^*;
3. With probability ε, replace it with z_i. The sample can now be considered a sample from $\hat{F}_i(\varepsilon)$.

When ε tends to zero, the probability of replacing two of the z_i^* with this procedure becomes trivial. We can equivalently transform our sample into a sample from $\hat{F}_i(\varepsilon)$ by transforming a single random element from $\{z_j^*\}$ into z_i with probability $n \cdot \varepsilon$. Without loss of generality, assuming this element is the first one, we can rewrite Equation 5.60 as

$$\begin{aligned} &\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \Big[\mathrm{E}_{\hat{F}_i(\varepsilon)}\big[\hat{\eta}\big] - \mathrm{E}_{\hat{F}}\big[\hat{\eta}\big] \Big] \\ &\quad = n \Big[\mathrm{E}_{\hat{F}}\big[\hat{\eta} \mid z_1^* = z_i\big] - \mathrm{E}_{\hat{F}}\big[\hat{\eta}\big] \Big]. \end{aligned} \tag{5.70}$$

Accordingly,

$$\frac{1}{n}\hat{\eta}\big[\hat{F}_i(\varepsilon)\big] = \mathrm{E}_{\hat{F}}(\hat{\eta} \mid z_1^* = z_i) - \mathrm{E}_{\hat{F}}(\hat{\eta}), \tag{5.71}$$

and so, using the Hájek projection of $\hat{\eta}$,

$$\begin{aligned} \hat{V} &= \sum_{i=1}^{n} \Big[\mathrm{E}_{\hat{F}}(\hat{\eta} \mid z_1^* = z_i) - \mathrm{E}_{\hat{F}}(\hat{\eta}) \Big]^2 \\ &\approx \sum_{i=1}^{n} \Big[\mathrm{E}_{F}(\eta \mid z_1^* = z_i) - \mathrm{E}_{F}(\eta) \Big]^2, \end{aligned} \tag{5.72}$$

where we replace the empirical approximation $\hat{F}$ with its true value F. At this point, we have established the connection among the IFJ, Hájek projections, and the asymptotic variance estimate for the predictions of random forests.

We will extend the asymptotic theory established here to the difference in predictions of two random forest models to evaluate the behavior of the inclusion of measurement error for covariates in random forests.

Chapter 6

Random Forest Models and Measurement Error

6.1 Introduction

The EWID paradigm invites error in measurements, from the recording of data to the performance on some benchmark test such as the CFMT to the measurement error inherent to machines. This motivates the question: how different could the predictions be in terms of bias and variance if we assume the presence of measurement error? This chapter seeks to explore the behavior of the presence of measurement error in random forest models.

6.1.1 Introduction to Error

All data is subject to error, both in the sense of traditional measurement error from instruments to recording errors from humans. In a sense, humans can be thought as the "instrument" or "machine" measuring some value. In general, no data is safe from the general sense of [measurement] error.

Generally speaking, machine learning models are fit without measurement error, because it is assumed that the data provided is representative of the population and modeling goals. Machine learning models tend to be fit without regard for causal inferences nor for attribution (i.e., significance). The primary goals of machine learning models are prediction and, on occasion, estimation, since these models are used for their high prediction accuracy. With increased use of machine learning models, many modelers seek to go beyond just prediction, into estimation and attribution. One such consequence is understanding the limitations and key assumptions of the data.

Almost all data is assumed to be measured without error, or that the error is small enough to be contained in the assumed standard normal noise variable. For example, if we ask for a person's height, weight, blood pressure, etc., we assume the measurement received is exactly the same as whatever we are trying to measure. In some cases, this assumption is perfectly reasonable and feasible (e.g., age, biological sex, if a person owns a cat, etc.). In other cases, the measurement may be "close enough" or "good enough" to ignore any measurement error. However, these cases tend to more often be exceptions rather than the standard. Making such a strong assumption that measurements are taken without error may have serious statistical consequences, and could lead to spurious conclusions. Failing to account for measurement error could potentially invalidate findings.

The idea of modeling measurement error is not new or novel, and has large presence in modeling observational data from economics and nutritional studies. In fact, it is well-studied in those fields, where the measurement error is widely prevalent in the types of data collected. Wallace (2020) addresses many of these issues in the February 2020 issue of *Significance* magazine to raise awareness of the dire consequences of ignoring measurement error, possible sources of measurement

error, and methods to address these issues.

For example, patients often notice elevated blood pressure readings while at the doctor's office due to "white coat hypertension" from a higher-stress situation. Errors are assumed to follow various structures. Classical measurement error assumes the value recorded is the truth plus some "random" noise structure. Berkson measurement error assumes each unit is exposed to some condition, and the observed exposure varies from unit to unit, with added noise. Carroll et al. (2006) provides more detailed information on measurement error and methodologies developed in classical statistics to account for such errors.

This chapter serves to explore the behavior of measurement error in random forest models to see if we need to account for measurement error or if the random forest model can account for the error by itself. Can measurement error in these cases be ignored because its impact is "not that bad?" We first establish the asymptotic framework to provide an estimator for the mean difference and variance of the two distributions, one of which is measured with error and the second is measured without error.

6.2 Measurement Error and Random Forests

Assume we have some covariates W that are measured exactly with no error. Now, further assume that we have some other covariates U that are measured with error. In fact, let $U = W + \varepsilon$, where ε represents the measurement error. Finally, let us assume we have response variable Y. From Wager and Athey (2018), we know that for test point x

$$\frac{\hat{\mu}_n(x) - \mathrm{E}\left[\hat{\mu}_n(x)\right]}{\sigma_n(x)} \xrightarrow{d} N(0,1). \tag{6.1}$$

For any random forest, the results Wager and Athey (2018) show that the forest is asymptotically normal, which means that a random forest built on either W or U will be asymptotically normal with a different mean and a different variance. We know that U is in fact W with additional variability due to the presence of the measurement error term ε. How different can we expect that predictions of the two asymptotically normal distributions to be?

6.2.1 Asymptotic Behavior

We follow the derivation process from Wager and Athey (2018) as detailed in Section 5.3. Suppose we have two random forests $\mu\left(x; \mathcal{Z}_{1,n}\right)$ and $\mu\left(x; \mathcal{Z}_{2,n}\right)$ as defined in Definition 5.3.1, where

$$\begin{aligned} \mathcal{Z}_{1,n} &= \left(Z_{u,1}, \ldots, Z_{u,n}\right) = \left[\left(U_1, Y_1\right), \ldots, \left(U_n, Y_n\right) \right] \\ \mathcal{Z}_{2,n} &= \left(Z_{w,1}, \ldots, Z_{w,n}\right) = \left[\left(W_1, Y_1\right), \ldots, \left(W_n, Y_n\right) \right]. \end{aligned} \tag{6.2}$$

From Equation 5.15, we can show that the asymptotic means are

$$\begin{aligned} \hat{\mu}_{u,n}(x) &= \frac{1}{M} \sum_{m=1}^{M} T\left(x; Z^*_{u,m_1}, \ldots, Z^*_{u,m_s}\right) \\ \hat{\mu}_{w,n}(x) &= \frac{1}{M} \sum_{m=1}^{M} T\left(x; Z^*_{w,m_1}, \ldots, Z^*_{w,m_s}\right), \end{aligned} \tag{6.3}$$

respectively, for M trees, subsample size s, and test point x. Here, Z^*_u and Z^*_w are drawn without replacement from their respective sets U and W. We can also show that the asymptotic variances are

$$\hat{\sigma}^2_{u,n}(x) = \frac{n-1}{n} \left(\frac{n}{n-s}\right)^2 \sum_{i=1}^{n} \mathrm{Cov}_*\left[\hat{\mu}_{u,m}(x), Y^*_{u,im}\right]^2 \tag{6.4}$$

$$\hat{\sigma}^2_{w,n}(x) = \frac{n-1}{n}\left(\frac{n}{n-s}\right)^2 \sum_{i=1}^{n} \mathrm{Cov}_*\left[\hat{\mu}_{w,m}(x), Y^*_{w,im}\right]^2$$

respectively, from Equation 5.22. Here, $\hat{\mu}_{u,n}(x)$ and $\hat{\mu}_{w,n}(x)$ are the the estimates for $\mu_u(x)$ and μ_w(x) from a single regression tree m, respectively, and $Y^*_{u,im}$ and $Y^*_{w,im}$ are the number of times $Z_{u,i}$ and $Z_{w,i}$ appears in the subsamples used by T, respectively. The covariances are taken with respect to the set of all trees $m = 1, \ldots, M$. Asymptotically, the respective distributions for $\hat{\mu}_u(x)$ and $\hat{\mu}_w(x)$ are normal with means $\mathrm{E}\left[\hat{\mu}_u(x)\right]$ and $\mathrm{E}\left[\hat{\mu}_w(x)\right]$ and variances $\hat{\sigma}^2_u(x)$ and $\hat{\sigma}^2_w(x)$ according to Equation 6.1.

6.2.2 Framework for a Difference

Mentch and Hooker (2016) provide a general framework for a test of significance for the predictive influence of covariates. In particular, they consider a "full" model with p features $\{X_1, \ldots, X_p\}$ versus a "reduced" model with $\boldsymbol{X}^{(R)} \subset \{X_1, \ldots, X_p\}$. Their goal was to determine if $\mu(x) = \mu^{(R)}(x)$, which provides information on whether or not a covariate not included in the reduced model makes a significant contribution to the prediction of the test point x. Let $x_{\text{test}} = \{x_1, \ldots, x_j\}$. Formally, the proposed hypothesis test

$$\begin{aligned} H_0 &: \mu(x_i) = \mu^{(R)}(x_i) \;\forall\; x_i \in x_{\text{test}} \\ H_A &: \mu(x_i) \neq \mu^{(R)}(x_i) \text{ for some } x_i \in x_{\text{test}}. \end{aligned} \tag{6.5}$$

They define a difference function

$$\begin{aligned} \hat{D}(x) &= \hat{\mu}(x) - \hat{\mu}^{(R)}(x) \\ &= \frac{1}{M}\sum_{m=1}^{M} T\left(x; Z^*_{m_1}, \ldots, Z^*_{m_s}\right) - \frac{1}{M}\sum_{m=1}^{M} T^{(R)}\left(x; Z^*_{m_1}, \ldots, Z^*_{m_s}\right) \end{aligned} \tag{6.6}$$

$$= \frac{1}{M} \sum_{m=1}^{M} \left[T\left(x; Z^*_{m_1}, \ldots, Z^*_{m_s}\right) - T^{(R)}(x; Z^*_{m_1}, \ldots Z^*_{m_s}) \right]$$

as the difference between the full and reduced models (i.e., the fitted random forest models based on the relevant sets of covariates). The difference function is a U-statistic, which means it can be subjected to the same treatment as any other random forest. $\hat{D}$ is asymptotically normal given one test point x.

For multiple j test points $\{x_1, \ldots, x_j\}$, define $\hat{\boldsymbol{D}}$ as the vector of observed differences in the predictions $\hat{\boldsymbol{D}} = \left[\hat{D}(x_1), \ldots, \hat{D}(x_j)\right]$, which has an MVN distribution with mean vector

$$\boldsymbol{M} = \left[\mu(x_1) - \mu^{(R)}(x_1), \ldots, \mu(x_j) - \mu^{(R)}(x_j)\right]^T \tag{6.7}$$

and covariance matrix Σ, provided that a joint distributions exists with respect to Lebesgue measure. For a single test point, the predictions will have a normal distribution. We will assume a single test point for the derivations below.

We can adapt this framework to study the difference in predictions between two random forests built from covariates measured with and without error, U and W, respectively. The key difference is the exclusion of the actual hypothesis test, since we are not interested in determining the significance of importance for the measurement error nor is it possible in practice to determine this difference.

Define a difference function $\mathbb{D}$ for a single test point x, similar to the one defined in Equation 6.6,

$$\begin{aligned} \mathbb{D} &= \hat{\mu}_u(x) - \hat{\mu}_w(x) \\ &= \frac{1}{M} \sum_{m=1}^{M} T(x; Z^*_{u,m_1}, \ldots, Z^*_{u,m_s}) - \frac{1}{M} \sum_{m=1}^{M} T(x; Z^*_{w,m_1}, \ldots, Z^*_{w,m_s}) \end{aligned} \tag{6.8}$$

$$= \frac{1}{M} \sum_{m=1}^{M} \Big[T(x; Z^*_{u,m_1}, \ldots, Z^*_{u,m_s}) - T(x; Z^*_{w,m_1}, \ldots, Z^*_{w,m_s}) \Big].$$

The above is still a U-statistic, which means that $\mathbb{D} \sim N(\mu_{\text{diff}}, \sigma^2_{\text{diff}})$, where

$$\mu_{\text{diff}} = \mu_u(x) - \mu_w(x) \tag{6.9}$$

and σ^2_{diff} is similar to $\hat{\sigma}^2_u$ and $\hat{\sigma}^2_w$ from Equation 6.4. The estimation procedure in this case would require the construction of both random forests.

As stated earlier in Equation 6.3 and Equation 6.4, we know asymptotically that $\mu_u(x) \sim N\left[\hat{\mu}_u(x), \hat{\sigma}^2_u(x)\right]$ and $\mu_w(x) \sim N\left[\hat{\mu}_w(x), \hat{\sigma}^2_w(x)\right]$. Thus, we treat $\mu_u(x)$ and $\mu_w(x)$ as we would any other random variable to determine the asymptotic distribution of $\mathbb{D}$, as defined in Equation 6.8. To estimate μ_{diff} and σ^2_{diff}, we will rely on the results from Wager and Athey (2018). We already know that $\mathbb{D}$ is asymptotically normal, with a theoretical mean of μ_{diff} and σ^2_{diff}. For μ_{diff}, we can show

$$\begin{aligned} \mathrm{E}\left(\mathbb{D}\right) &= \mathrm{E}\left[\mu_u(x) - \mu_w(x)\right] \\ &= \mathrm{E}\left[\mu_{u,n}(x)\right] - \mathrm{E}\left[\mu_{w,n}(x)\right] \\ &\xrightarrow{p} \hat{\mu}_{u,n}(x) - \hat{\mu}_{w,n}(x) \end{aligned} \tag{6.10}$$

due to linearity of expectation. Thus, the estimation process for μ_{diff} depends wholly on the construction of the two random forests based on U and W, and taking the difference of the predictions. In terms of the variance, we can show

$$\begin{aligned} \mathrm{Var}\left(\mathbb{D}\right) &= \mathrm{Var}\left[\mu_u(x) - \mu_w(x)\right] \\ &= \mathrm{Var}\left[\mu_{u,n}(x)\right] + \mathrm{Var}\left[\mu_{w,n}(x)\right] - 2 \cdot \mathrm{Cov}\left[\mu_{u,n}(x), \mu_{w,n}(x)\right] \end{aligned} \tag{6.11}$$

$$\xrightarrow{p} \hat{\sigma}^2_{u,n}(x) + \hat{\sigma}^2_{w,n}(x) - 2\sum_{i=1}^{n} \text{Cov}_* \left[\hat{\mu}_{u,m}(x), Y^*_{u,im}\right] \cdot \text{Cov}_* \left[\hat{\mu}_{w,m}(x), Y^*_{w,im}\right]$$

for m trees. We can derive the expression for $\text{Cov}\left[\mu_{u,n}(x), \mu_{w,n}(x)\right]$ in Equation 6.11 by using a combination of the original IFJ results from Jaeckel (1972) and the relationship of of the IFJ and Hájek projection from Wager et al. (2014).

6.2.3 Variance Estimate

We can show that the IFJ estimation procedure still holds for difference distribution defined in Equation 6.8. Jaeckel (1972) provides a bivariate extension for two statistics Q and R, evaluated at some empirical distribution $\hat{F}$, such that

$$Q(\hat{F}) = Q(F) + \sum_i \left(\hat{W}_i - \frac{1}{M}\right) D^Q_i + \dots \tag{6.12}$$

and

$$R(\hat{F}) = U(F) + \sum_j \left(\hat{W}_j - \frac{1}{M}\right) D^R_j + \dots$$

are the Taylor series expansion of Q and R. Jaeckel (1972) then derives the expression for the covariance of Q and R,

$$\begin{aligned} \text{Cov}\left[Q(\hat{F}), R(\hat{F})\right] &= \text{E}\left\{\left[Q(\hat{F}) - Q(F)\right]\left[R(\hat{F}) - R(F)\right]\right\} \\ &\approx \text{E}\left[\sum_i \left(\hat{W}_i - \frac{1}{M}\right) D^Q_i \cdot \sum_j \left(\hat{W}_j - \frac{1}{M}\right) D^R_j\right] \\ &= \text{E}\Big[\sum_j \left(\hat{W}_j - \frac{1}{M}\right)^2 D^Q_j D^R_j \\ &\qquad + \sum_{i \neq j}\sum \left(\hat{W}_i - \frac{1}{M}\right)\left(\hat{W}_j - \frac{1}{M}\right) D^Q_i D^R_j\Big] \end{aligned} \tag{6.13}$$

$$= \frac{1}{nM} \sum_j D_j^Q D_j^R,$$

where D_j is defined in a similar manner as in Equation 5.51. The above quantity can be estimated by

$$\hat{V}_{\text{Cov}} = \frac{1}{n^2} \sum_{i=1}^{n} \hat{D}_i^Q \hat{D}_i^R \tag{6.14}$$

if Q and R are well-behaved, which means they both have finite variances and differentiable expectations with respect to the parameter being estimated. The random variables are assumed to have a distribution that exists with respect to Lebesgue measure.

We can easily extend these results to covariance $\hat{V}_{\text{Cov}}$ of Q and R from Equation 6.14, such that

$$\hat{V}_{\text{Cov}} = \sum_{i=1}^{n} \text{Cov}_*^Q \left[Y_i^*, \hat{\eta} \right] \cdot \text{Cov}_*^R \left[Y_i^*, \hat{\eta} \right] \tag{6.15}$$

as shown previously from Equation 5.62. We previously established that the directional derivatives (denoted as U_i from Efron (1982) or D_j from Jaeckel (1972)) are equivalent to the covariance of the prediction from the functional statistic and the random variable designating the number of times the i-th observation is in the resample. This then returns us to the variance derivation of $\text{Var}(\mathbb{D})$ from Equation 6.4, which allows us to estimate $\text{Var}(\mathbb{D})$ using the point estimate shown in Equation 6.11.

Bias Correction

We also derive the finite bias correction for the estimate of covariance. It was established in Section 5.3.2 that the bias correction for the variance estimate is

$\left(\frac{n-1}{n}\right)\left(\frac{n}{n-s}\right)^2$. Now, assume we have two sets of trivial trees that do not make any splits

$$T_1\left(x;\theta,Z_{1i_1},\ldots,Z_{1i_s}\right)=\frac{1}{s}\sum_{j=1}^{s}Y_{i_j} \tag{6.16}$$
$$T_2\left(x;\theta,Z_{2i_1},\ldots,Z_{2i_s}\right)=\frac{1}{s}\sum_{j=1}^{s}X_{i_j}$$

Then, the full random forests for these sets of trees are $\hat{\mu}_1=\frac{1}{n}\sum_{i=1}^{n}Y_i$ and $\hat{\mu}_2=\frac{1}{n}\sum_{i=1}^{n}X_i$, respectively, with standard variance estimators of

$$\hat{V}_{1,\text{trivial}}=\frac{1}{n(n-1)}\sum_{i=1}^{n}(Y_i-\bar{Y})^2 \tag{6.17}$$
$$\hat{V}_{2,\text{trivial}}=\frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

Further, we can show

$$\widehat{\text{Cov}}_{\text{trivial}}=\frac{1}{n(n-1)}\sum_{i=1}^{n}(Y_i-\bar{Y})(X_i-\bar{X}) \tag{6.18}$$

The idea here is similar to that of sample variance, where it is the average of the "squared" deviations. Let Y_i^* denote if the i-th observation was in the subsample, which has a multinomial distribution with s observations and probability $\frac{1}{n}$ for each observation to be chosen. For any $i=1,\ldots,n$, we know $\text{E}_*(\hat{\mu}^*)$ and $E(Y_i^*)=\frac{s}{n}\cdot\bar{Y}$. Further,

$$\begin{aligned}\text{E}_*(\hat{\mu}_1^*\cdot Y_i^*)&=\frac{s}{n}\left[\frac{Y_i}{s}+\left(\frac{s-1}{s}\right)\left(\frac{n\bar{Y}-Y_i}{n-1}\right)\right] \\ &=\left(\frac{1}{n}\right)\left(\frac{n-s}{n-1}\right)\cdot Y_i+\left(\frac{s-1}{n-1}\right)\cdot\bar{Y}\end{aligned} \tag{6.19}$$

and

$$\begin{aligned}\mathrm{Cov}_*(\hat{\mu}_1^*, Y_i^*) &= \left(\frac{1}{n}\right)\left(\frac{n-s}{n-1}\right)\cdot Y_i + \left(\frac{s-1}{n-1} - \frac{s}{n}\right)\cdot \bar{Y} \\ &= \left(\frac{1}{n-1}\right)\left(\frac{n-s}{n}\right)(Y_i - \bar{Y}).\end{aligned} \tag{6.20}$$

This can be similarly shown for the second set of trivial trees $T_2(\cdot)$, where

$$\mathrm{E}_*(\hat{\mu}_2^* \cdot Y_i^*) = \left(\frac{1}{n}\right)\left(\frac{n-s}{n-1}\right)\cdot X_i + \left(\frac{s-1}{n-1}\right)\cdot \bar{X} \tag{6.21}$$

and

$$\mathrm{Cov}_*(\hat{\mu}_2^*, Y_i^*) = \left(\frac{1}{n-1}\right)\left(\frac{n-s}{n}\right)(X_i - \bar{X}). \tag{6.22}$$

Thus,

$$\begin{aligned}\widehat{\mathrm{Cov}}_{\mathrm{IFJ}} &= \frac{n-1}{n}\left(\frac{n}{n-s}\right)^2 \sum_{i=1}^{n} \mathrm{Cov}_*(\hat{\mu}_1^*, Y_i^*)\cdot \mathrm{Cov}_*(\hat{\mu}_2^*, Y_i^*) \\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \\ &= \widehat{\mathrm{Cov}}_{\mathrm{trivial}}.\end{aligned} \tag{6.23}$$

6.2.4 Consistency

We can show that the derived variance estimator from Equation 6.11 is consistent with the population σ^2_{diff}. We already know that $\hat{\sigma}^2_{u,n}(x)$ and $\hat{\sigma}^2_{w,n}(x)$ are consistent estimators for σ_u^2 and σ_w^2, as shown by Wager and Athey (2018). Following a similar proof, we can also show that $\hat{V}_{\mathrm{Cov}}$ is consistent for $\mathrm{Cov}\left[\mu_{u,n}(x), \mu_{x,n}(x)\right]$. Let F be the distribution from which $Z_1, \ldots, Z_n$ are drawn, then the variance $\sigma^2_{\mathrm{diff},n}$ of the

Hájek projection of $\hat{\mu}_{u,n}(x) - \hat{\mu}_{w,n}(x)$ is

$$\sigma^2_{\text{diff},n} = \sum_{i=1}^{n} \left(\mathrm{E}\left[\hat{\mu}_u(x)|Z_i\right] - \mathrm{E}\left[\hat{\mu}_u(x)\right]\right)\left(\mathrm{E}\left[\hat{\mu}_w(x)|Z_i\right] - \mathrm{E}\left[\hat{\mu}_w(x)\right]\right) \tag{6.24}$$

$$= \frac{s^2}{n^2}\sum_{i=1}^{n} \left[\mathrm{E}\left(T_u|Z_i\right) - \mathrm{E}\left(T_u\right)\right]\left[\mathrm{E}\left(T_w|Z_i\right) - \mathrm{E}\left(T_w\right)\right] \tag{6.25}$$

and the IFJ estimated defined in Equation 6.15 is equal to

$$\hat{V}_{\text{Cov}} = \frac{n-1}{n}\left(\frac{n}{n-2}\right)^2 \sum_{i=1}^{n} \left[\mathrm{E}\left(T_u|Z_1^* = Z_i\right)\right]\left[\mathrm{E}\left(T_w|Z_1^* = Z_i\right)\right] \tag{6.26}$$

for Z^* drawn from empirical distribution $\hat{F}$ on $\{Z_1, \ldots, Z_n\}$. Recall that Z^* is drawn without replacement from $\hat{F}$. We can rewrite $\hat{V}_{\text{Cov}}$ using the Hájek projection $\mathring{T}$ of T,

$$\hat{V}_{\text{Cov}} = \frac{n-1}{n}\left(\frac{n}{n-2}\right)^2 \sum_{i=1}^{n} (A_{i,u} + R_{i,u})(A_{i,w} + R_{i,w}) \tag{6.27}$$

where

$$\begin{aligned}
A_{i,u} &= E(\mathring{T}_u|Z_1^* = Z_i) - E(\mathring{T}_u), \\
A_{i,w} &= E(\mathring{T}_w|Z_1^* = Z_i) - E(\mathring{T}_w), \\
R_{i,u} &= E(T_u - \mathring{T}_u|Z_1^* = Z_i) - E(T_u - \mathring{T}_u), \text{ and} \\
R_{i,w} &= E(T_u - \mathring{T}_w|Z_1^* = Z_i) - E(T_u - \mathring{T}_w).
\end{aligned} \tag{6.28}$$

Here, A_i are the main effects and R_i are the secondary effects. Wager and Athey (2018) showed in their Lemma 12 that the main effects A_i will give the covariance $\text{Cov}\left[\mu_{u,n}(x), \mu_{w,n}(x)\right]$ such that

$$\left(\frac{1}{\text{Cov}\left[\mu_{u,n}(x), \mu_{w,n}(x)\right]}\right)\left(\frac{s^2}{n^2}\right)\sum_{i=1}^{n} A_{i,u}A_{i,w} \xrightarrow{p} 1. \tag{6.29}$$

In their Lemma 13, they show that

$$\left(\frac{1}{\text{Cov}\left[\mu_{u,n}(x), \mu_{w,n}(x)\right]}\right)\left(\frac{s^2}{n^2}\sum_{i=1}^{n} R_{i,u}R_{i,w}\right) \xrightarrow{p} 0 \tag{6.30}$$

The Cauchy-Schwarz inequality can be used to bound the cross terms and that

$$\lim_{n\to\infty} \frac{n(n-1)}{(n-s)^2} = 1. \tag{6.31}$$

Thus, $\hat{V}_{\text{Cov}}\,/\text{Cov}\left[\mu_{u,n}(x), \mu_{w,n}(x)\right]$ converges in probability to 1. Since we know that $\hat{\sigma}_u^2$, $\hat{\sigma}_w^2$, and $\hat{V}_{\text{Cov}}$ converge in probability to constants, then the sum of these will converge also converge in probability to a constant $\text{Var}\left(\mathbb{D}\right)$.

6.3 Simulations

Now that we have derived estimators for expectation and variance for the difference of two random forest models, we can simulate what could be expected in practice. Since asymptotic theory was established from two different parties, and to investigate the behavior under a variety of conditions, we will borrow the synthetic distributions from Wager and Athey (2018) and Mentch and Hooker (2016). Additionally, we will consider a more realistic model that more resembles the original motivation of this problem.

6.3.1 Synthetic Distributions

First, we will establish the models we will be using for the simulations. These distributions were used to test the performance of the IFJ estimate for variance, as well as establish some behavioral observations for the difference in distributions

for data measured with and without error. The measurement errors assume the classical measurement error structure, where the observed value is equal to the true value plus some random error structure.

From Wager and Athey (2018), we will use one model. For p variables, let $X \sim \text{Unif}([0,1]^p)$. Further, assume noise $\varepsilon_{\cos} \sim N(0,1)$. Define the cosine model as

$$Y = 3 \cdot \cos[\pi \cdot (X_1 + X_2)] + \varepsilon_{\cos}. \tag{6.32}$$

In this case, only two of the p covariates are ever used to generate Y, but the remaining $p-2$ covariates are included as predictors in the random forest models.

Moving on to the two models used by Mentch and Hooker (2016). Define the simple linear regression (SLR) model as

$$Y = 2X_1 + \varepsilon_{\text{SLR}}, \tag{6.33}$$

where $X_1 \sim \text{Uniform}(0, 20)$.

Define the multivariate adaptive regression splines (MARS) model used by Friedman (1991) as

$$Y = 10\sin(\pi \cdot X_1 X_2) \ + \ 20(X_3 \ - \ 0.05)^2 \ + 10X_4 \ + \ 5X_5 \ + \ \varepsilon_{\text{MARS}} \tag{6.34}$$

where $\boldsymbol{X} = [X_1, \ldots, X_5] \sim \text{Uniform}\big([0,1]^5\big)$.

In simulations, $\varepsilon_{\text{MARS}}$ and ε_{SLR} could follow one of two distributions: $N(0,10)$ or $N(0,1)$.

For the above cases, two distributions are created using the described framework: one with measurement error and the other without measurement error. For distributions generated with measurement error, an additional error term $\varepsilon_{\text{error}} \sim$

$N\ (\mu_{\text{error}}, \sigma^2_{\text{error}})$ was added to each simulated X_i.

Finally, define the beta model to simulate a model where the CFMT is the covariate. This model is a "proof of concept." By construction, the CFMT has a range of 0 to 72, which makes either the beta or gamma distribution excellent candidates. For the sake of this simulation, we will use the beta distribution with parameters α and β. From the factor data set, we estimated that CFMT $\sim$ Beta$(3, 1.05)$. Assume a measurement error variable ε_3 and noise variable $\varepsilon_{\text{CFMT}}$ that follow a $N(0, 0.1)$ distribution. If we assume a scaled standard deviation of 0.1, this provides a "real" standard deviation of ± 7 CFMT score. For more realism, we estimated the parameters of the beta distribution of the normalized decision times from the factor data set, which was estimated as Beta$(0.319, 15.626)$. Define the CFMT 1 model as

$$Y^{\text{CFMT 1}}_{\text{no error}} = \frac{1}{1 + \exp(\text{CFMT} + \text{Decision Time} + \varepsilon_{\text{CFMT}})}. \tag{6.35}$$

We will also assume a secondary error structure of $\varepsilon_{\text{CFMT}} \sim \text{Unif}(-0.1, +0.1)$, which will be the CFMT 2 model.

We would like to note that this is likely the "worst case scenario." In a realistic situation, measurement errors like those found in the CFMT score or confidence rating would be likely diluted by other covariates not subject to measurement error. We use the sigmoid function[1] to map the response values to the range $[0, 1]$.

The response value Y will be the same for the pairs of error and error-free generated distributions. That is, the error distributions will consist of the response Y generated by the error free distributions, but with covariates that include the error terms.

[1]The sigmoid function is $f(x) = \frac{1}{1+\exp[h(x)]}$, where $h(x)$ is the function to create the responses from the simulated covariates.

6.3.2 Results

Since we use only $M = \mathcal{O}(n)$ replicates, $\hat{V}_{\text{IFJ}}$ can experience substantial Monte Carlo bias. To mitigate this issue, Wager (2016) proposed a Monte Carlo bias correction, which we have adapted for our purposes. For subsample size s and M replicates, the bias correction is,

$$\hat{V}_{\text{IFJ}}^{M} = \sum_{i=1}^{n} C_i^2 - \frac{s(n-s)}{n}\left(\frac{\hat{v}}{M}\right), \text{ where} \tag{6.36}$$

$$C_i = \frac{1}{M}\sum_{m=1}^{M}\left(Y_{mi}^{*} - \frac{s}{n}\right)\left(T_m^{*} - \bar{T}^{*}\right) \text{ and} \tag{6.37}$$

$$\hat{v} = \frac{1}{M}\sum_{m=1}^{M}(T_m^{*} - \bar{T}^{*})^2. \tag{6.38}$$

First, we can assess the performance of the derive estimators based on their bias, variance, and mean squared error. For evaluation purposes, we consider the MSE, as a combination of the squared bias and variance. We first draw $K = 100$ random test points $\{x_{(k)}\}_{k=1}^{K}$ from the data-generating distributions describe previously. Then, for each test point, we construct $R = 100$ random training sets $\{Z_{(r)}\}_{r=1}^{R}$. Evaluate both the prediction $\text{RF}_s(x_{(k)}; Z_{(r)})$ for subsample size s and the variance estimate $\hat{V}_{\text{IFJ}}(x_{(k)}; Z_{(r)})$. The numbers shown in Table 6.1 are averaged over k test points:

$$\text{Bias}^2 = \frac{1}{k}\sum_{k=1}^{K}\left(\frac{1}{R}\sum_{r=1}^{R}\hat{V}_{\text{IFJ}}^{B}(x_{(k)}; Z_{(r)}) - \text{Var}_r\left[\text{RF}_s(x_{(k)}; Z_{(r)})\right]\right)^2, \tag{6.39}$$

$$\text{Var} = \frac{1}{k}\sum_{k=1}^{K}\frac{1}{R-1}\sum_{r=1}^{R}\left[\hat{V}_{\text{IFJ}}^{B}(x_{(k)}; Z_{(r)}) - \frac{1}{R}\sum_{r=1}^{R}\hat{V}_{\text{IFJ}}\left(x_{(k)}; Z_{(r)}\right)\right]^2. \tag{6.40}$$

We take the variance of the predictions from the forest across the R training sets and compare that variance to the IFJ estimate of the variance to find the squared bias. We will take the average of the K squared bias values. To find variance, we

Distribution	p	n	IFJ Estimate	Bias^2	Absolute Variance	MSE
SLR[a]	1	200	4.550	27.804	7.606	35.409
		1000	2.807	9.533	1.721	11.254
SLR[b]	1	200	14.782	256.982	43.420	300.402
		1000	11.716	149.061	12.801	161.86
MARS[a]	4	200	1.509	2.320	0.225	2.545
		1000	0.685	0.479	0.025	0.504
MARS[b]	4	200	5.276	27.481	1.563	29.044
		1000	2.928	8.629	0.277	8.906
Cosine	2	200	0.278	0.086	0.012	0.098
		1000	0.132	0.018	0.001	0.020
	10	200	0.138	0.019	0.001	0.020
		1000	0.062	0.004	1.95E-04	0.004
CFMT 1	2	200	8.47E-05	8.86E-09	1.87E-10	1.07E-09
		1000	2.04E-05	4.98E-10	9.45E-11	5.93E-10
CFMT 2	2	200	6.87E-05	5.90E-09	1.32E-10	7.21E-09
		1000	1.39E-05	2.25E-10	3.73E-11	2.63E-10

Table 6.1: Performance of the IFJ for random forests for a set of synthetic distributions. The "absolute" metrics describe the accuracy of $\hat{V}_{\text{IFJ}}$. All metrics are obtained using a training sample size of n with p features, subsample size of $s = \lfloor n^{0.7} \rfloor$, and use $M = 5n$ trees grown for each forest.

[a] Assuming a noise structure of $N(0, 1)$.
[b] Assuming a noise structure of $N(0, 10)$.

take the variance of the IFJ estimates, which is simply the sample variance formula. We previously used a similar evaluation method when comparing methods for the estimation of probability of target presence in Section 4.3.2.

The "absolute" metrics describe the accuracy of $\hat{V}_{\text{IFJ}}$. All metrics are obtained using a training sample size of n with p features, subsample size of $s = \lfloor n^{0.7} \rfloor$, and use $M = 5n$ trees grown for each forest. All synthetic distributions were run with $n = 200, 1000$ observations each, with the subsample sizes and number of trees per forest modified accordingly.

Table 6.1 serves to show the behavior of the estimator derived in Equation 6.11, which should follow similar behavior in terms of metrics to the behavior seen from the results in Wager and Athey (2018). In Table 6.1, we can see that as the simulated sample size increases, the estimate becomes less variable and less biased, resulting in a smaller MSE. This behavior is indeed consistent with what was previously seen from the simpler case of one distribution. A more variable noise structure results results in a much larger IFJ estimate, which is logical from a noise variance of ten versus one. The bias, absolute variance, and MSE are much larger. The range for MSE is quite variable for different data-generating procedures; however, the values are relative to the possible outputs from the procedures. The calculated metrics are consistent with what was originally seen in Wager and Athey (2018) and consistent with the generated data.

We do seem to observe some interesting behavior with the comparison of simpler models (i.e., the SLR model) versus more complex models (i.e., the MARS model), in which the noise structure very heavily impacts the IFJ estimate. Given the estimated MSE for these models, larger samples may be required for certain types models for more accurate estimates. The requisite sample size depends on the model used, even with the derived finite bias correction.

Distribution	p	n	Mean Difference	IFJ Estimate
SLR[a]	1	200	-0.912	4.550
		1000	-0.077	2.807
SLR[b]	1	200	3.615	14.782
		1000	1.134	11.716
MARS[a]	4	200	-0.400	1.509
		1000	-0.148	0.685
MARS[b]	4	200	0.738	5.276
		1000	0.123	2.928
Cosine	2	200	0.105	0.278
		1000	0.107	0.132
	10	200	0.241	0.138
		1000	0.217	0.062
CFMT 1	2	200	2.51E-04	8.47E-05
		1000	-3.89E-04	2.036E-05
CFMT 2	2	200	-8.65E-04	6.87E-05
		1000	-1.33E-04	1.40E-05

Table 6.2: Distributions of the differences of sets of random forests for a set of synthetic distributions, generated with and without measurement error. The "absolute" metrics describe the accuracy of $\hat{V}_{\text{IFJ}}$. All metrics are obtained using a training sample size of n with p features, subsample size of $s = \lfloor n^{0.7} \rfloor$, and use $M = 5n$ trees grown for each forest.

[a] Assuming a noise structure of $N(0, 1)$.
[b] Assuming a noise structure of $N(0, 10)$.

In Table 6.2, we record the point estimates for mean difference and variance of the distribution of differences for the previously defined data generation structures. The mean difference and IFJ estimates are proportional to the relative ranges of possible outputs from the data-generating distributions.

It seems that the measurement error structure needs to be quite egregious to make an impact on the predictions of the random forest, which is consistent with the idea of the "robustness" and flexibility of random forests. Error structures consisting of higher variance simply flattens the density curve, while error structures with non-zero means shift the peaks of the curves to the relative directions.

6.3.3 Different Measurement Error Structures

In Section 6.3.1, the goal was to assess the behavior of predictions during the presence of measurement error. Here, we assume the same models as before, but vary the measurement error distributions. The measurement error structures are assumed with some minor structure to vary the biases and variances. The measurement error distributions are all assumed to be normal: $N(-2, 2)$, $N(-1, 2)$, $N(0, 1)$, $N(0, 5)$, $N(1, 2)$, and $N(2, 2)$. We aim to show the relative impacts of the measurement errors on the predictions, primarily to see if the model could indeed capture the measurement error structure. Figure 6.1 shows the behavior of the estimated means and variances of the measurement error structure for the MARS model, with noise structure $N(0, 1)$ and $N(0, 10)$.

For $n = 200$, we randomly generated 100 distributions following the MARS data-generating procedure for 100 different models. The average estimates for mean difference and variance to find a point estimate to generate theoretical normal densities. These densities are plotted in Figure 6.1. We found similar behavior with

models fit using $n = 1000$, but for efficiency, we chose to use $n = 200$. The other data-generating models produced similar behavior, but it is most evident in the MARS model.

6.3.4 Summary

It is clear that measurement error does affect the predictions from the underlying truth. While this chapter does not seek to provide a solution to measurement error, it does illustrate the impact of measurement error by deriving an asymptotic estimator to quantify that impact. Moving forward, it is important to characterize the size and structure of this measurement error when fitting future models. The different assumed measurement error structures in Section 6.3.3 noticeably impact the predictions when the structures assume zero versus non-zero mean and the relative sizes of the variance.

What is typically done in classical statistical models (such as GLMs) in fields that encounter much measurement error (such as medicine, nutrition, economics, etc.) is to generate more data to better understand the measurement error. This can be done by taking repeated measurements (i.e., obtaining replicates) on a single unit. Large discrepancies in the measurements from a single unit may indicate large measurement error. Another method is to obtain validation data, where some measurements are known to be perfect. A less preferred method is to use external data, outside of a designed study, to understand the measurement data.

In terms of EWID data, it is unclear what may be the best course of action to understand the measurement error structure. Due to the nature of the data, it is quite difficult to obtain replication data, since there may be a learning curve of identifying suspects in lineups. Additionally, it may be quite expensive and resource-

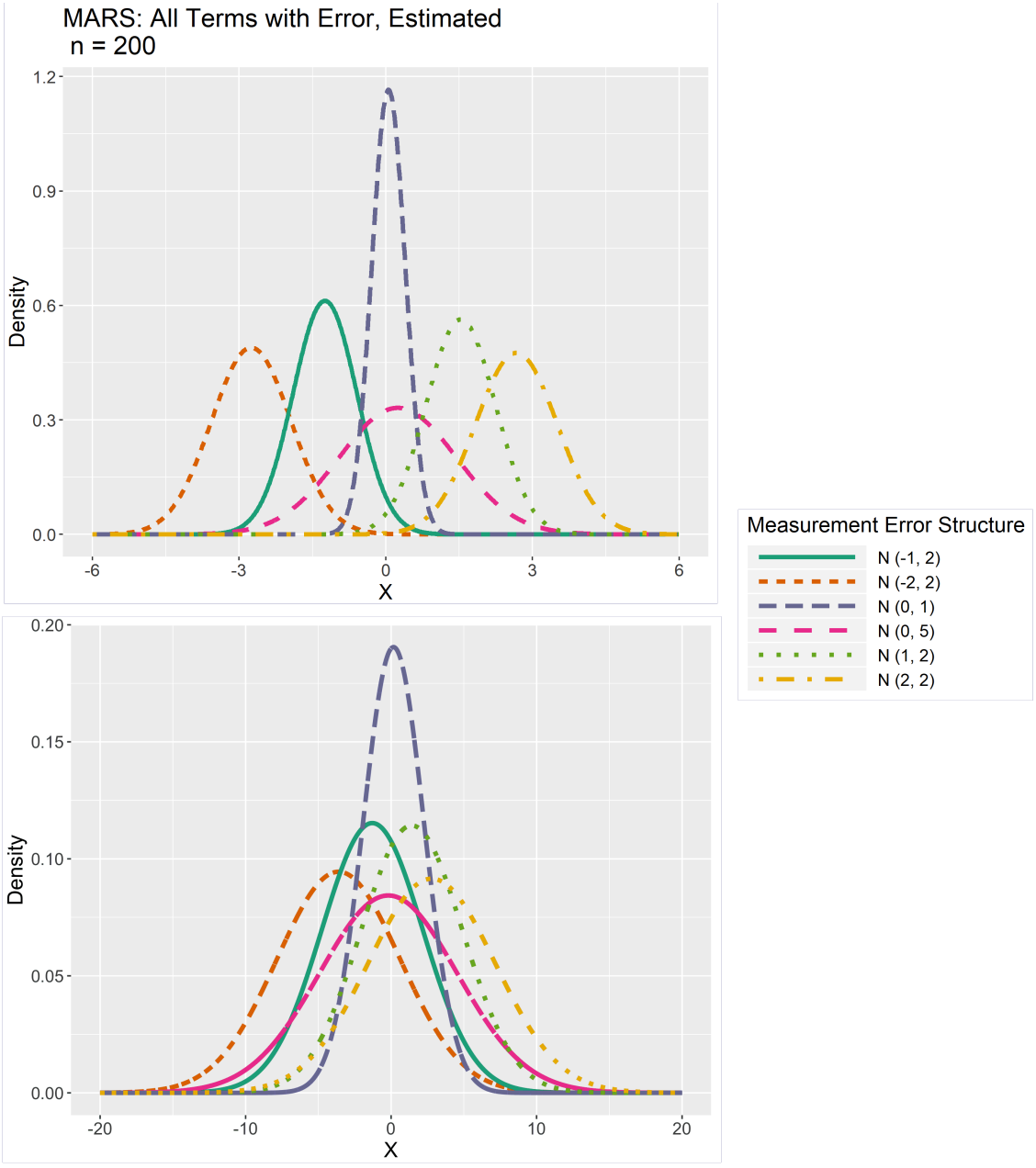


Figure 6.1: Captured behavior of different measurement error structures. Plot on the top shows a model fit with a noise distribution of $N(0, 1)$. Plot on the bottom shows a model fit with a noise distribution of $N(0, 10)$. The measurement error distributions added to each generated covariate are $N(-2, 2)$, $N(-1, 2)$, $N(0, 1)$, $N(0, 5)$, $N(1, 2)$, and $N(2, 2)$.

intensive to generate an adequate number of different lineups with different suspect and foils. Validation data may also be difficult to obtain, since no "gold standard" or method of obtaining "perfect" data without measurement error exists. If the data were affected or manipulated in any way to engineer "perfect" data, the data would no longer be ecologically valid. Finally, external data has many of the same issues as discussed with replication and validation data, with the additional uncertainty of how accurately the information in the external data can translate to the relevant setting.

Part III

Summary and Future Work

Throughout this dissertation, we provided an overview of the motivation for this body of work: eyewitness identification. We reviewed existing statistical methodologies, and introduced a new statistical framework to estimate an individual eyewitness' probability of accuracy. Since EWID data could fluctuate based on individual factor levels, we sought to understand the impact of measurement error in variables in random forest models.

Long-standing conventional statistical methodologies, including logistic regression and, more generally, generalized linear models, particularly for bivariate outcomes (sensitivity and specificity), remain valuable and appropriate tools for analyzing EWID experiments, especially when the experiment includes concomitant information, such as environmental variables of the experiment and demographic characteristics of the "eyewitness." In the absence of such information, ROC curves remain a useful comparison of two methods in diagnostic medicine, statistical process control, and eyewitness experiments. Newer approaches from statistical machine learning may be useful with very large experiments, though the impact of specific variables on the outcome may not always be as interpretable as with conventional linear models. Whichever technique is used, proper characterization of the uncertainties associated with the inferences must be calculated.

Alternative ways of examining the data could also lead to new modeling procedures or algorithms that would be useful in practice. We proposed a method for estimating the probability of accuracy for eyewitnesses that takes proper account of individuals' probabilities of choosing or not choosing a suspect from a lineup. This new framework estimates eyewitness identification accuracy by estimating, as intermediate steps, an individual's probability of choosing and the global probabilities of target presence using random forests.

This method is a potential tool that could provide an in-field assessment of

eyewitness reliability, which can be explained to and understood by juries, judges, lawyers, LEOs, and any other non-statisticians working in EWID. Further methods depend on the available types of EWID data, which could include recordings of eyewitness proceedings by working in conjunction with police departments. Researchers who conduct more varied and complex types of experiments will produce sets of observational data (National Research Council, 2014), leading to the development of novel modeling procedures and statistical methods.

Overall, our method provides a substantial contribution to the EWID field, because it enables the estimation of two latent variables: (1) the probability of accuracy for each eyewitness and (2) the probability of target presence or base rate for a given data set. It is not only applicable to EWID data, but to data from other fields that follow a similar structure. In comparison to an existing method of estimating base rate, it performs much more variably, but with increasing accuracy as the complexity of the data set increases. The problem of eyewitness accuracy is treated probabilistically in our framework, and provides a way to assess accuracy even if the underlying truth has not (and cannot) be observed directly.

A key component to the success of implementation is to obtain as much information as possible in terms of both the system and estimator variables relevant to eyewitness lineups. The framework inherently incorporates the interactive effects of the system and estimator variables present in the data sets. Additionally, in order for implementation to take place for real use, ecologically valid data sets need to be collected to train a suitable model. This is a tremendous step to advancing the analysis of EWID data that needs to be supplemented by the psychology experts in the field, which is core to the original interdisciplinary motivation.

The framework established in Chapter 4 also lends itself well to fields with similarly structured data sets. Some of these fields were identified, such as geophysics,

financial services, and agricultural sciences.

Since we use random forests for estimation purposes, we were interested in assessing the behavior of random forests in the presence of variables that are very highly likely to have measurement error. Thus, we derived asymptotic estimators for the difference of two random forests models in order to illustrate and quantify the impact of the presence of measurement error. The asymptotic theory reviewed and established in Chapter 5 and Chapter 6 only applies to regression and binary classification, since the "majority votes" are essentially averages. The "majority vote" framework for multiple classification would require a different asymptotic framework that has not yet been pursued. Thus, the asymptotic theory could be expanded for multiple classification.

It is clear that measurement error does affect the predictions from the underlying truth. While this chapter does not seek to provide a solution to measurement error, it does illustrate the impact of measurement error by deriving an asymptotic estimator to quantify that impact. Moving forward, it is important to characterize the size and structure of this measurement error when fitting future models. The different assumed measurement error structures in Section 6.3.3 noticeably impact the predictions when the structures assume zero versus non-zero mean and the relative sizes of the variance.

What is typically done in classical statistical models (such as GLMs) in fields that encounter much measurement error (such as medicine, nutrition, economics, etc.) is to generate more data to better understand the measurement error. This can be done by taking repeated measurements (i.e., obtaining replicates) on a single unit. Large discrepancies in the measurements from a single unit may indicate large measurement error. Another method is to obtain validation data, where some measurements are known to be perfect. A less preferred method is to use external

data, outside of a designed study, to understand the measurement data.

In terms of EWID data, it is unclear what may be the best course of action to understand the measurement error structure. Due to the nature of the data, it is quite difficult to obtain replication data, since there may be a learning curve of identifying suspects in lineups. Additionally, it may be quite expensive and resource-intensive to generate an adequate number of different lineups with different suspect and foils. Validation data may also be difficult to obtain, since no "gold standard" or method of obtaining "perfect" data without measurement error exists. If the data were affected or manipulated in any way to engineer "perfect" data, the data would no longer be ecologically valid. Finally, external data has many of the same issues as discussed with replication and validation data, with the additional uncertainty of how accurately the information in the external data can translate to the relevant setting.

This thesis seeks to expand the methodologies available for EWID data analysis, and emphasize the necessary cooperation of statisticians and psychologists to work to the ideal of ecologically valid data. We have introduced a framework for the estimation of eyewitness identification accuracy, but have discovered a new issue with the necessity to quantify measurement error in EWID system and/or estimator variables. While we do not provide a solution for measurement error, we do emphasize the importance to work towards characterizing the size and structure of measurement error in future EWID models.

Part IV

Appendices

Appendix A

Example Lineups

The four lineups used in the factor data set from Chad Dodson at the University of Virginia.

(Top right) Fair lineup, with the target present;

(Top left) Biased lineup, with the target present;

(Bottom left) Fair lineup, with the target absent, innocent suspect in place of true target; and

(Bottom right) Biased lineup, with the target absent, innocent suspect in place of true target.

Not Present

Not Present

Not Present

Not Present

Appendix B

Data Sets

The tables included here describe the data sets introduced in Section 2.2.2.

Data Set	Variable	Possible Values	Description
Factor data	Lineup bias	Yes, no	Was the lineup biased?
Repeated delay data	Categorization	Category of recognition[a]	Basis for respondent's choice
	Delay	5 minutes, 1 day	Assigned delay condition
	Lineup race	White, black	What was the race in lineup?
Delay data	Actor	Actor A, actor B	Actor for each race
	Delay	Immediate, 2 day, 4 day, 8 day	Assigned delay condition
	Video	Black or white man	Race/sex of actor in video

Table B.1: Unique variables to the factor data set, delay data set, and repeated delay data set

[a]Unobserved features, observed features, familiarity, recognition

Variable	Possible Values	Description
Accuracy	0, 1	Was the eyewitness correct in his or her choice?
Age	>18 years old	Age of the eyewitness
CFMT score	0 to 72	What was the eyewitness's CFMT score?
Chooser	Yes, no	Did the eyewitness choose a person?
Confidence	0, 0.2, 0.4, 0.6, 0.8, 1.0	How confident was the eyewitness in their decision?
Cross-race	Yes, no	Was the lineup cross-race?
Decision	Target, innocent suspect[a], foil, not present	What was the eyewitness's decision?
Decision time	>0 seconds	How long did the eyewitness's decision take?
Lineup format	Simultaneous, sequential	What was the format of the lineup?
Participant race	White, Black	Race of eyewitness
Sex	Male, female	Sex of eyewitness
Target present	Present, absent	Was the target present or absent?
Weapon	Yes, no	Was a weapon present?

Table B.2: Common variables to the factor data set, delay data set, and repeated delay data set

[a]Innocent suspect choice was only available for the factor data set, as no innocent suspect was designated in either of the other two data sets.

Variable	**Possible Values**	**Description**
Age	>18 years old	Age of the participant
Ethnicity	Asian, Black, Hispanic, Native American, White, Other, Did Not State	Ethnicity of the participant
Education	Less than high school, high school/GED, currently attending college, 2-year college degree, 4-year college degree, graduate degree, did not state	Education level of the participant
Sex	Male, female	Sex of the participant
Biasing	Confidence rating, liberal, neutral, unbiased, or conservative	Assigned instruction condition

Table B.3: Variables included in the Mickes et al. (2017) data set

Variable	Possible Values	Description
Expt. 1	Simultaneous, sequential	Expt. 1 manipulated conditions
Expt. 2	Photo stimulus (lineup), video stimulus	Expt. 2 manipulated conditions
Expt. 3a	1-lap, 2-lap, choice in video lineups	Expt. 3a manipulated conditions
Expt. 3b	1-lap, 2-lap, choice in photo lineups	Expt. 3b manipulated conditions
Expt. 4	Six or nine photos	Expt. 4 manipulated conditions
Expt. 5	Simultaneous, sequential	Expt. 5 manipulated conditions
Age	>18 years old	Age of the participant
rating	0 to 1.0, intervals of 0.1	Confidence rating from participant's decision
Ethnicity	White, Asian, Black, Hispanic, Native American, Other, Did Not State	Ethnicity of the participant
Sex	Male, female	Sex of the participant

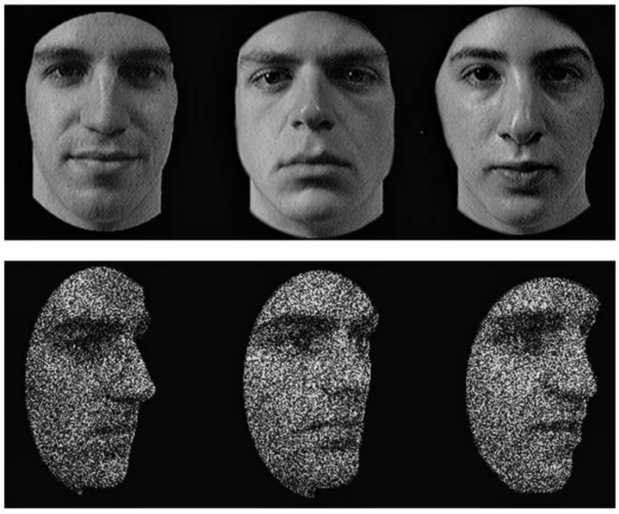
Table B.4: Variables included in the Seale-Carlisle et al. (2019) data sets

Appendix C

CFMT

Below are example images from the CFMT (Duchaine and Nakayama, 2006). There are six total faces that the test taker needs to remember. This set represents a portion of one of the six faces. Each face has three sets of images in the the learning phase, five sets of images in the novel images phase, and four sets of images in the novel images with noise phase. First, we have the set of learning images. The top set of images belong are the facades of target to be remembered, and the bottom set of images is the "question" asked by the test.

Next, we have the novel images (top) and novel images with noise (bottom):



Appendix D

U-Statistics

Overview. A U-statistic is a special case of statistical functionals. Any such population quantity is a function of some distribution F of the independent and identically distributed X_i, and can be written as $g(F)$, where g is some real-valued function that is defined over the collection $\mathcal{F}$ of distributions F (for more information, see Lehmann, 2004, Chapter 6).

Definition. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F_{X,\eta}$, where η is the parameter to be estimated. Further, suppose there exists an unbiased estimator g of η as a function of $k \leq n$ arguments. Define

$$\eta = \mathrm{E}\left[g(X_1, \ldots, X_k)\right], \tag{D.1}$$

The function g is of k arguments and is known as η's expectation functionals. Without loss of generality, we can assume that g is permutation symmetric in its arguments, since any given g can be replaced by another permutation symmetric version. That is, the arguments in any given g can be substituted by an equivalent permutation symmetric version. Permutation symmetry is a concept where some object is invariant under the action of some operator. It remains unchanged when operated

upon, even by elements that can be "exchanged" (see French and Rickles, 2003). In other words, $g(X_{i_1}, \ldots, X_{i_k})$ satisfies η for any permutation $(i_1, \ldots, i_k)$ of $(1, \ldots, k)$. So, therefore does the symmetric function

$$g^*(X_1, \ldots, X_k) = \frac{1}{k!} \sum_{(i_1, \ldots, i_k)} \cdots \sum g(X_{i_1}, \ldots, X_{i_k}), \tag{D.2}$$

where the sum extends over all $k!$ such permutations.

The minimum variance unbiased estimator (MVUE) U_n for η is given by taking the sum over all possible $\binom{n}{k}$ subsamples of size k. U_n is known as a U-statistic with kernel g of rank k,

$$U_n = \binom{n}{k}^{-1} \sum_{(i_1,\ \ldots,\ i_k)} g(X_{i_1}, \ldots, X_{i_k}), \tag{D.3}$$

Thus, a U-statistic U_n is the average of the kernel $g(X_{i_1}, \ldots, X_{i_k})$ over all possible k-tuples of observations in the sample. It is al an unbiased estimator of η, where the sum extends over all k-tuples such that $1 \leq i_1 < \cdots < i_k \leq n$.

Hoeffding (1948) shows for $n - 1 \geq k$, the smallest possible sample size, U-statistics satisfy

$$\frac{n-1}{n} \operatorname{Var}_{n-1} \geq \operatorname{Var}_n \tag{D.4}$$

and

$$\mathrm{E}_F(\widehat{\mathrm{Var}}) = \frac{n-1}{n} \mathrm{E}_F(\widetilde{\mathrm{Var}}) \geq \frac{n-1}{n} \operatorname{Var}_{n-1} \geq \operatorname{Var}_n. \tag{D.5}$$

Asymptotic Theory Hoeffding (1948) also shows that U_n is asymptotically normal with limiting variance $\frac{s^2}{n} \cdot \xi_{1,s}$ when both the kernel and rank are fixed. Here,

$$\xi_{1,k} = \operatorname{Cov}\left[g(X_1, \ldots, X_s),\ g(Z_1, Z_2', \ldots, Z_s')\right], \tag{D.6}$$

where

$$Z'_2, \ldots, Z'_s \overset{\text{i.i.d.}}{\sim} F_{X,\eta}. \tag{D.7}$$

In this example, the "1" represents the number of common observations between the two subsamples. In general, $\xi_{k,s}$ represents the covariance of the same form for k observations in common.

Hájek Projection. Hájek (1968) established the Hájek projection principle, which states that the Hájek projection of $U_n \in L_2(P)$ is

$$\hat{U_n} = \sum_{i=1}^{n} \mathrm{E}\left(U_n - \theta | X_i\right) - (n-1)\mathrm{E}\left(U_n\right). \tag{D.8}$$

Then, if $E[g(X_{1_k}, \ldots, X_{i_k})]^2 < \infty$ and $\mathrm{Var}\left(g(X_{1_k}\right) > 0$, then

$$\sqrt{n}(U_n - \theta - \hat{U}_n) \xrightarrow{p} 0. \tag{D.9}$$

The proof shows that the Hájek projection $\hat{U}_n$ is of the claimed form, assuming that the X_i's are independent and g is permutation symmetric. We check that $U_n - \hat{U}_n$ is orthogonal to each $g_i(X_i)$. Then, we verify that

$$\frac{\mathrm{Var}\left(\hat{U}_n\right)}{\mathrm{Var}\left(U_n\right)} \to 1. \tag{D.10}$$

This is done by deriving the variances for U_n and $\hat{U}_n$. The CLT and finiteness of $\mathrm{Var}\left(\hat{U}_n\right)$ implies that

$$\sqrt{n} \cdot \hat{U}_n \xrightarrow{d} N(0, \zeta_1), \tag{D.11}$$

where ζ_1 is the derived variance. The derived variance for U_n is also found to be ζ_1,

thus showing that the ratio converges to 1. Then,

$$\frac{U_n - \theta}{\sqrt{\text{Var}\,(U_n)}} - \frac{\hat{U}_n}{\sqrt{\text{Var}\,(\hat{U}_n)}} \xrightarrow{p} 0, \tag{D.12}$$

which implies

$$\sqrt{n}(U_n - \theta - \hat{U}_n) \xrightarrow{p} 0. \tag{D.13}$$

Therefore,

$$\sqrt{n}(U_n - \theta) \xrightarrow{d} N(0, \zeta_1). \tag{D.14}$$

Part V

References

List of Abbreviations

ANN adaptive nearest neighbors.

ANOVA analysis of variance.

AUC area under the curve.

Bern Bernoulli.

Bin Binomial.

BR base rate.

CA confidence-accuracy.

CART classification and regression tree.

CFMT Cambridge Face Memory Test.

CL composite likelihood.

CLT central limit theorem.

CSAFE Center for Statistics and Applications in Forensic Evidence.

DFSC Defense Forensic Science Center.

DOJ Department of Justice.

DPP deviation from perfect performance.

DR diagnosticity ratio.

ECL expressed confidence level.

EDA exploratory data analysis.

EE ellipse-envelope.

EW eyewitness.

EWID eyewitness identification.

FAR false alarm rate.

FN false negative.

FP false positive.

FPR false positive rate.

FWB fixed-width confidence bands.

GLM generalized linear models.

HR hit rate.

ID identification.

IF influence function.

IFJ infinitesimal jackknife.

KS Kolmogorov-Smirnov.

LDA linear discriminant analysis.

LEO law enforcement officer.

LIME local interpretable model-agnostic explanations.

LOO leave-one-out.

LR likelihood ratio.

MARS multivariate adaptive regression splines.

MCMC Markov chain Monte Carlo.

MDA mean decreased accuracy.

ML maximum likelihood.

MLE maximum likelihood estimation.

MRI magnetic resonance imaging.

MSE mean squared error.

MVN multivariate normal.

MVUE minimum variance unbiased estimator.

NAS National Academy of Sciences.

NPV negative predictive value.

NRC National Research Council.

OR odds ratio.

pAUC partial area under the curve.

PDP partial dependence plots.

PNN predictive nearest neighbors.

PPV positive predictive value.

PROC predictive receiver operating characteristic.

QDA quadratic discriminant analysis.

RF random forest.

ROC receiver operating characteristic.

SDT signal detection theory.

Se sensitivity.

SEQ sequential.

SHAP Shapley additive explanations.

SIM simultaneous.

SJR simultaneous joint confidence regions.

SLR simple linear regression.

Sp specificity.

SVM support vector machines.

THA threshold averaging.

TIA target is absent.

TIP target is present.

TN true negative.

TP true positive.

TPR true positive rate.

VA vertical averaging.

WHB Working-Hotelling based bands.

WROC weighted receiver operating characteristic.

References

K. L. Amendola and J. T. Wixted. Comparing the diagnostic accuracy of suspect identifications made by actual eyewitnesses from simultaneous and sequential lineups in a randomized field trial. *Journal of Experimental Criminology*, 2, October 2014.

K. L. Amendola and J. T. Wixted. The role of site variance in the American Judicature Society field study comparing simultaneous and sequential lineups. *Journal of Quantitative Criminology*, December 2015.

S. M. Andersen, C. A. Carlson, M. A. Carlson, and S. D. Gronlund. Individual difference predict eyewitness identification performance. *Personality and Individual Differences*, 60:36–40, 2014.

G. Biau and L. Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101(10):2499–2518, 2010.

G. Biau and E. Scornet. A random forest guided tour. *arXiv*, November 2015.

G. E. P. Box, W. G. Hunter, and J. S. Hunter. *Statistics for Experimenters: Design, Innovation, and Discover*. Wiley, 2nd edition edition, 2005.

L. Breiman. Random forests. *Machine Learning*, 45(5-32), 2001.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Chapman & Hall, 1984.

N. Brewer and G. L. Wells. The confidence-accuracy relationship in eyewitness identification: effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology*, 12(1):11–30, 2006.

J. C. Brigham, C. A. Meissner, and A. W. Wasserman. Applied issues in the construction and expert assessment of photo lineups. *Applied Cognitive Psychology*, 13:S73–S92, 1999.

A. M. Burton, D. White, and A. McNeill. The Glasgow Face Matching Test. *Behavior Research Methods*, 42(1):286–291, 2010.

C. A. Carlson and M. A. Carlson. An evaluation of lineup presentation, weapon presence, and a distinctive feature using ROC analysis. *Journal of Applied Research in Memory and Cognition*, 2014.

C. A. Carlson, S. D. Gronlund, and S. E. Clark. Lineup composition, suspect position and the sequential lineup advantage. *Journal of Experimental Psychology: Applied*, 14 (2):118–128, 2008.

C. A. Carlson, J. L. Dias, D. Weatherford, and M. A. Carlson. An investigation on the weapon focus effect and the confidence-accuracy relationship for eyewitness identification. *Journal of Applied Research in Memory and Cognition*, May 2016a.

C. A. Carlson, D. F. Young, D. Weatherford, M. A. Carlson, J. E. Bednarz, and A. R. Jones. The influence of perpetrator exposure time and weapon presence/timing on eyewitness confidence and accuracy. *Applied Cognitive Psychology*, September 2016b.

R. J. Carroll, D. Ruppert, L. A. Stefanksi, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall/CRC, 2006.

Y. Chen, Y. Liu, J. Ning, L. Nie, H. Zhu, and H. Chu. A composite likelihood method for bivariate meta-analysis in diagnostic systematic reviews. *Statistical Methods in Medical Research*, 26(2):914–930, April 2017.

S.-J. Cho, J. Wilmer, G. Herzmann, R. McGugin, D. Fiset, A. E. V. Gulick, K. Ryan, and I. Gauthier. Item response theory analyses of the Cambridge Face Memory Test (CFMT). *Psychological Assessment*, 27(2):552–566, June 2015.

H. Chu and S. R. Cole. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology*, 59: 1331–1333, 2006.

H. Chu, L. Nie, S. R. Cole, and C. Poole. Meta-analysis of diagnostic accuracy studies

accounting for disease prevalence: alternative parameterizations and model selection. *Statistics in Medicine*, 28:2384–2399, 2009.

S. E. Clark. A re-examination of the effects of biased lineup instructions in eyewitness identification. *Law and Human Behavior*, 29(4):395–424, August 2005.

S. E. Clark, A. S. Benjamin, J. T. Wixted, L. Mickes, and S. D. Gronlund. Eyewitness identification and the accuracy of the criminal justice system. *Policy Insights from the Behavioral and Brain Sciences*, 2(1):175–186, 2015.

M. F. Colloff, K. A. Wade, and D. Strange. Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychological Science*, 27(9):1227–1239, 2016.

M. F. Colloff, K. A. Wade, J. T. Wixted, and E. A. Maylor. A signal-detection analysis of eyewitness identification across the adult lifespan. *Psychology and Aging*, 32(3):243–258, 2017.

G. D'Agostini. Basic probabilistic issues in sciences and in forensics (hopefully) clarified by a toy experiment modeled by a BN. Presented at Isaac Newton Institute for Mathematical Sciences at the conference on Bayesian networks and argumentation in evidence analysis, September 2016.

A. P. Dawid and J. Mortera. *Graphical Models for Forensic Analysis*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods Series, 2017.

E. Demidenko. Confidence intervals and bands for the binormal ROC curve revisited. *Journal of Applied Statistics*, 39(1):67–79, January 2012.

Department of Justice. Eyewitness evidence: A guide for law enforcement, 1999.

N. L. Dimou, M. Adam, and P. G. Bagos. A multivariate method for meta-analysis and comparison of diagnostic tests. *Statistics in Medicine*, March 2016.

C. S. Dodson and D. G. Dobolyi. Confidence and eyewitness identifications: the cross-race effect, decision time, and accuracy. *Applied Cognitive Psychology*, 30:113–125, 2016.

I. E. Dror and D. Charlton. Why experts make errors. *Journal of Forensic Identification*, 56(4):600–616, 2006.

B. Duchaine and K. Nakayama. The Cambridge Face Memory Test: results for neurologically intact individuals and an investigation of its validity using inverted face stimuli

and prosopagnosic participants. *Neuropsychologia*, 44:576–585, 2006.

W. DuMouchel. Hierarchical Bayes linear models for meta-analysis. Technical Report 27, National Institute of Statistical Sciences, 19 T.W. Alexander Drive, PO Box 14006, Research Triangle Park, NC 27709-4006, September 1994.

B. Efron. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Conference Board of the Mathematical Sciences, 1982.

B. Efron. Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007, 2014.

B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3): 586–596, 1981.

P. Eusebi, J. B. Reitsma, and J. K. Vermunt. Latent class bivariate model for the meta-analysis of diagnostic test accuracy studies. *BioMed Central Medical Research Methodology*, 14(88), 2014.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

S. French and D. Rickles. Understanding permutation symmetry. *arXiv*, January 2003.

J. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19(1):1–67, 1991.

J. Friedman and P. Hall. On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, 137(3):669–683, 2007.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, 2000.

P. Garbolino. Bayesian networks for the evaluation of testimony. Presented at Isaac Newton Institute for Mathematical Sciences at the conference on Bayesian networks and argumentation in evidence analysis, September 2016.

B. L. Garrett. *Convicting the Innocent: Where Criminal Prosecutions Go Wrong*. Harvard University Press, 2012.

M. Georgeson. Sensitivity and bias - an introduction to signal detection theory. Electronic.

J. Grabman, D. G. Dobolyi, N. L. Berelovich, and C. S. Dodson. Predicting high confidence

errors in eyewitness memory: the role of face recognition ability, decision-time, and justifications. *Journal of Applied Research in Memory and Cognition*, 8(2):233–243, June 2019.

S. D. Gronlund and A. S. Benjamin. The new science of eyewitness memory. *The Psychology of Learning and Motivation*, 69:241–284, 2018.

S. D. Gronlund and J. S. Neuschatz. Eyewitness identification discriminability: ROC analysis versus logistic regression. *Journal of Applied Research in Memory and Cognition*, 3:54–57, 2014.

S. D. Gronlund, J. T. Wixted, and L. Mickes. Evaluating eyewitness identification procedures using receiver operating characteristic analysis. *Current Directions in Psychological Science*, 23(1):3–10, 2014.

S. D. Gronlund, L. Mickes, J. T. Wixted, and S. E. Clark. *Conducting an Eyewitness Lineup: How the Research Got It Wrong*, volume 63. Elsevier, 2015.

J. Hájek. Asymptotic normality of simple linear rank statistics under alternatives. *The Annals of Mathematical Statistics*, 39(2):325–345, 1968.

K. Hajian-Tilaki. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J intern Med*, 4(2):627–635, 2013.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Machine Learning*. Springer, second edition, 2013.

W. Hoeffding. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19(3):293–325, 1948.

C. Hong. Bivariate ROC curve. *The Korean Journal of Applied Statistics*, 19:277–286, 2012.

C. Hong and J. Jeong. Bivariate ROC curve and optimal classification function. *Communications for Statistical Applications and Methods*, 19:629–638, 2012.

M. L. Howe and L. M. Knott. The fallibility of memory in judicial processes: lessons from the past and their modern consequences. *Memory*, 23(5):633–656, July 2015.

A. Hoyer and O. Kuss. Meta-analysis for the comparison of two diagnostic tests to a common gold standard: a generalized linear mixed model approach. *Statistical Methods*

in Medical Research, 0(0):1–12, 2016.

J. E. Humphries and H. D. Flowe. Receiver operating characteristic analysis of age-related changes in lineup performance. *Journal of Experimental Child Psychology*, 132:189–204, 2015.

H. L. R. III, J. H. Wixted, and K. A. Desoto. *The Curious Complexity Between Confidence and Accuracy in Reports From Memory*. Oxford University Press, 2012.

L. Irwig, A. N. Tosteson, C. Gatsonis, J. Lau, G. Colditz, T. C. Chalmers, and F. Mosteller. Guidelines for meta-analyses evaluating diagnostic tests. *Annals of Internal Medicine*, 120(8):667–676, April 1994.

L. A. Jaeckel. The infinitesimal jackknife. Memorandum, Bell Laboratories, June 1972.

H. Jin and Y. Lu. The ROC region of a regression tree. *Statistics and Probability Letters*, 79:936–942, 2009.

Junaidi, D. Nur, and E. Stojanovski. Bayesian estimation of a meta-analysis model using Gibbs sampler. In *Proceedings of the Fifth Annual ASEARC Conference - Looking to the future - Programme and Proceedings*, 2012.

P. Juslin, N. Olsson, and A. Winman. Calibration and diagnosticity of confidence in eyewitness identification: comments on what can be inferred from low confidence-accuracy correlation: comments on what can be inferred from low confidence-accuracy correlation. *Journal of Experimental Psychology*, 22(5):1304–1316, 1996.

K. Kafadar. Statistical issues and reliability of eyewitness identification as a forensic tool. Presentation, September 2015.

J. Kantner and I. G. Dobbins. Partitioning the sources of recognition confidence: the role of individual differences. *Psychonomic Bulletin & Review*, 2019.

K. Krug. The relationship between confidence and accuracy: current thoughts of the literature and a new area of research. *Applied Psychology in Criminal Justice*, 3(1), 2007.

J. Kruppa, A. Schwarz, G. Arminger, and A. Ziegler. Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, 40: 5125–5131, 2013.

J. Kruppa, Y. Liu, G. Biau, M. Kohler, I. R. König, J. D. Malley, and A. Ziegler. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. *Biometrical Journal*, 56(4):534–563, 2014.

J. M. Lampinen, A. M. Smith, and G. L. Wells. Four utilities in eyewitness identification practice: dissociations between receiver operating characteristic (ROC) analysis and expected utility analysis. *Law and Human Behavior*, 43(1):26–44, 2019.

K. Lange and E. Brunner. Sensitivity, specificity, and ROC curves in multiple reader diagnostic trials - a unified nonparametric approach. *Statistical Methodology*, 9:490–500, 2012.

A. J. Lee. *U-Statistics: Theory and Practice*. Boca Raton: Routledge, 2019.

M. M. Leeflang, J. J. Deeks, A. W. Rutjes, J. B. Reitsma, and P. M. Bossuyt. Bivariate meta-analysis of predictive values of diagnostic tests can be an alternative to bivariate meta-analysis of sensitivity and specificity. *Journal of Clinical Epidemiology*, 65: 1088–1097, 2012.

E. L. Lehmann. *Elements of Large-Sample Theory*. Springer, 2004.

C. Li. Probability estimation in random forests. Master's thesis, Utah State University, May 2013.

A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3): 18–22, 2002.

Y. Lin and Y. Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, June 2006.

R. Lindsay and G. L. Wells. Improving eyewitness identifications from lineups: simultaneous versus sequential lineup presentation. *Journal of Applied Psychology*, 70(3): 556–564, 1985.

A. J. Liu, K. Kafadar, B. L. Garrett, and J. Yaffe. Bringing new statistical approaches to eyewitness evidence. In D. L. Banks, K. Kafadar, D. H. Kaye, and M. Tackett, editors, *Handbook of Forensic Statistics*, chapter 21. Chapman & Hall/CRC Handbooks of Modern Statistical Methods Series, 2020.

E. F. Loftus. Planting misinformation in the human mind: a 30-year investigation of the

malleability of memory. *Learning & Memory*, 12:361–366, 2005.

A. S. Luby. A graphical model approach to eyewitness identification. Presented at Isaac Newton Institute for Mathematical Sciences at the conference on Bayesian networks and argumentation in evidence analysis, September 2016.

A. S. Luby. Strengthening analyses of line-up procedures: a log-linear model framework. *Law, Probability and Risk*, 16:241–257, 2017.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *31st Conference on Neural Information Processing System (NIPS 2017)*, 2017.

X. Ma, M. F. K. Suri, and H. Chu. A trivariate meta-analysis of diagnostic studies accounting for prevalence and non-evaluable subjects: re-evaluation of the meta-analysis of coronary CT angiography studies. *BioMed Central Medical Research Methodology*, 14(128), 2014.

P. Macaskill, C. A. Gatsonis, J. J. Deeks, R. Harbord, and Y. Takwoingi, editors. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*, chapter Analysing and Presenting Results. The Cochrane Collaboration, 2010.

S. A. Macskassy and F. Provost. Confidence bands for ROC curves: methods and an empirical study. In *Conference: ROC Analysis in Artificial Intelligence, 1st International Workshop*, 2008.

J. D. Malley, J. Kruppa, A. Dasgupta, K. Malley, and A. Ziegler. Probability machines: Consistent probability estimation using nonparametric learning machines. *Methods of Information in Medicine*, 51(1):74–81, January 2012.

E. McKone, S. Stokes, J. Liu, S. Cohan, C. Fiorentini, M. Pidcock, G. Yovel, M. Broughton, and M. Pelleg. A robust method of measuring other-race and other-ethnicity effects: the Cambridge Face Memory Test format. *PLoS One*, 7(10), October 2012.

N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7: 983–999, June 2006.

C. A. Meissner and J. C. Brigham. Thirty years of investigating the own-race bias in memory for faces: a meta-analytic review. *Psychology, Public Policy, and Law*, 7:3–35,

2001.

L. Mentch and G. Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(26): 1–41, 2016.

L. Mickes, H. D. Flowe, and J. T. Wixted. Receiver operating characteristic analysis of eyewitness memory: comparing the diagnostic accuracy of simultaneous vs. sequential lineups. *Journal of Experimental Psychology: Applied*, 18(4):361–376, 2012.

L. Mickes, M. B. Moreland, S. E. Clark, and J. T. Wixted. Missing the information needed to perform ROC analysis? then compute d', not the diagnosticity ratio. *Journal of Applied Research in Memory Cognition*, 3:58–62, 2014.

L. Mickes, T. M. Seale-Carlisle, S. A. Wetmore, S. D. Gronlund, S. E. Clark, C. A. Carlson, C. A. Goodsell, D. Weatherford, and J. T. Wixted. ROCs in eyewitness identification: instructions versus confidence ratings. *Applied Cognitive Psychology*, 31(5):467–477, September/October 2017.

National Research Council, editor. *Identifying the Culprit: Assessing Eyewitness Identification*. The National Academies Press, 2014.

A. K. Nikoloulopoulos. On composite likelihood in bivariate meta-analysis of diagnostic test accuracy studies. *Advances in Statistical Analysis*, 102(2):211–227, April 2018.

M. Olson. *Essays On Random Forest Ensembles*. PhD thesis, University of Pennsylvania, 2018.

M. Olson and A. J. Wyner. Making sense of random forest probabilities: a kernel perspective. *arXiv*, 2018.

M. A. Palmer, N. Brewer, N. Weber, and A. Nagesh. The confidence-accuracy relationship for eyewitness identification decisions: effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied*, 19(1):55–71, 2013.

S. H. Park, J. M. Goo, and C.-H. Jo. Receiver operating characteristic (ROC) curve: practical review for radiologists. *Korean Journal of Radiology*, 5:11–18, 2004.

M. S. Pepe. Receiver operating characteristic methodology. *Journal of the American Statistical Association*, 95(449):308–311, 2000.

H. L. Price, K. C. Bruer, and M. C. Adkins. Using machine learning analyses to explore relations between eyewitness lineup looking behaviors and suspect guilt. *Law and Human Behavior*, February 2020.

S. Pundir and R. Amala. Detecting diagnostic accuracy of two biomarkers through a bivariate log-normal ROC curve. *Journal of Applied Statistics*, 42(12):2671–2685, 2015.

J. B. Reitsma, A. S. Glas, A. W. Rutjes, R. J. Scholten, P. M. Bossuyt, and A. H. Zwinderman. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*, 58:982–990, 2005.

M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *arXiv*, August 2016.

C. M. Rotello, E. Heit, and C. Dube. When more data steer us wrong: replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review*, November 2014.

A. J. Russ, M. Sauerland, C. E. Lee, and M. Bindermann. Individual differences in eyewitness accuracy across multiple lineups of faces. *Cognitive Research: Principles and Implications*, 3(30), 2018.

C. M. Rutter and C. A. Gatsonis. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine*, 20:2865–2884, 2001.

J. Sauer, N. Brewer, T. Zweck, and N. Weber. The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law and Human Behavior*, 34:337–347, 2010.

J. D. Sauer, M. A. Palmer, and N. Brewer. Pitfalls in using eyewitness confidence to diagnose the accuracy of an individual identification decision. *Psychology, Public Policy, and Law*, 2019.

M. Sauerland, A. Sagana, S. L. Sporer, and J. T. Wixted. Decision time and confidence predict choosers' identification performance in photographic showups. *PLoS One*, 13 (1), 2018.

R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

E. Scornet. On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146,

2016a.

E. Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theor*, 62(3):1485–1500, March 2016b.

T. M. Seale-Carlisle, S. A. Wetmore, H. D. Flowe, and L. Mickes. Designing police lineups to maximize memory performance. *Journal of Experimental Psychology*, February 2019.

C. Semmler, M. Kaesler, and J. Dunn. An introduction to maximum likelihood estimation for signal detection theory applications to eyewitness identification data. In *SARMAC Regional Meeting*, Adelaide, South Australia, February 2018. SARMAC.

S.-Y. Shiu and C. A. Gatsonis. The predictive receiver operating characteristic curve for the joint assessment of the positive and negative predictive values. *Philosophical Transactions of the Royal Society A*, 366:2313–2333, 2008.

A. M. Smith, G. L. Wells, R. Lindsay, and S. Penrod. Fair lineups are better than biased lineups and showups, but not because they increase underlying discriminability. *Law and Human Behavior*, 2016.

A. M. Smith, J. M. Lampinen, G. L. Wells, L. Smalarz, and S. Mackovichova. Deviation from perfect performance measures the diagnostic utility of eyewitness lineups but partial area under the ROC curve does not. *Journal of Applied Research in Memory and Cognition*, 2018.

S. L. Sporer. The cross-race effect: beyond recognition of faces in the laboratory. *Psychology, Public Policy, and Law*, 7(1):170–200, 2001.

S. L. Sporer, S. Penrod, D. Read, and B. Cutler. Choosing, confidence, and accuracy: a meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin*, 118(3):315–327, 1995.

N. M. Steblay. Social influence in eyewitness recall: a meta-analytic review of lineup instruction effects. *Law and Human Behavior*, 21(3):283–297, 1997.

D. L. Streiner and J. Cairney. What's under the ROC? An introduction to receiver operating characteristics curves. *The Canadian Journal of Psychiatry*, 52(2):121–126, February 2007.

A. J. Sutton and K. R. Abrams. Bayesian methods in meta-analysis and evidence synthesis.

Statistical Methods in Medical Research, 10:277–303, 2001.

R. C. Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

S. Wager. Asymptotic theory for random forests. *arXiv*, May 2016.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 2018.

S. Wager, T. Hastie, and B. Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15(1):1625–1651, May 2014.

M. Wallace. Analysis in an imperfect world. *Significance*, 17(1):14–19, February 2020.

S. D. Walter. The partial area under the summary ROC curve. *Statistics in Medicine*, 24: 2025–2040, 2005.

F. Wang and C. A. Gatsonis. Hierarchical models for ROC curve summary measures: Design and analysis of multi-reader, multi-modality studies of medical tests. *Statistics in Medicine*, 27:243–256, 2008.

M.-C. Wang and S. Li. Bivariate marker measurements and ROC analysis. *Biometrics*, 68:1207–1281, December 2012.

M.-C. Wang and S. Li. ROC analysis for multiple markers with tree-based classification. *Lifetime Data Analysis*, 19:257–277, 2013.

E. K. Warrington. *Recognition Memory Test*. Nfer-Nelson, 1984.

G. L. Wells and A. L. Bradfield. "Good, you identified the suspect": feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology*, 83(3):360–376, 1998.

G. L. Wells and R. Lindsay. On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin*, 88(3):776–784, 1980.

G. L. Wells and E. A. Olson. Eyewitness identification: information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied*, 8(3): 155–167, 2002.

G. L. Wells, N. K. Steblay, and J. E. Dysart. A test of the simultaneous vs. sequential

lineup methods. Technical report, American Judicature Society, 2011.

G. L. Wells, L. Smalarz, and A. M. Smith. ROC analysis of lineups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory and Cognition*, 5:313–317, 2015a.

G. L. Wells, A. M. Smith, and L. Smalarz. ROC analysis of lineups obscures information that is critical for both theoretical understanding and applied purposes. *Journal of Applied Research in Memory and Cognition*, 4:324–328, 2015b.

S. A. Wetmore, J. S. Neuschatz, S. D. Gronlund, A. Wooten, C. A. Goodsell, and C. A. Carlson. Effect of retention interval on showup and lineup performance. *Journal of Applied Research in Memory and Cognition*, 4:8–14, 2015.

J. Wilson, K. Hugenbert, and M. Bernstein. The cross-race effect and eyewitness identification: how to improve recognition and reduce decision errors in eyewitness situations. *Social Issues and Policy Review*, 7(1):83–113, 2013.

J. T. Wixted and L. Mickes. A continuous dual-process model of remember/know judgments. *Psychological Review*, 117(4):1025–1054, 2010.

J. T. Wixted and L. Mickes. The field of eyewitness memory should abandon probative value and embrace receiver operating characteristic analysis. *Perspectives on Psychological Sciences*, 7(3):275–278, 2012.

J. T. Wixted and L. Mickes. Evaluating eyewitness identification procedures: ROC analysis and its misconceptions. *Journal of Applied Research in Memory Cognition*, 4: 318–323, 2015a.

J. T. Wixted and L. Mickes. ROC analysis measures objective discriminability for any eyewitness identification procedure. *Journal of Applied Research in Memory and Cognition*, 4(4):329–334, 2015b.

J. T. Wixted and G. L. Wells. The relationship between eyewitness confidence and identification accuracy: a new synthesis. *Psychological Science in the Public Interest*, 18(1): 10–65, 2017.

J. T. Wixted, L. Mickes, S. E. Clark, S. D. Gronlund, and H. L. R. III. Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist*,

70(6):515–526, September 2015.

J. T. Wixted, L. Mickes, J. C. Dunn, S. E. Clark, and W. Wells. Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences of the United States of America*, 113(2):304–309, January 2016a.

J. T. Wixted, J. D. Read, and D. S. Lindsay. The effect of retention interval on the eyewitness identification confidence-accuracy relationship. *Journal of Applied Research in Memory and Cognition*, 5:192–203, 2016b.

J. T. Wixted, L. Mickes, and R. P. Fisher. Rethinking the reliability of eyewitness memory. *Perspectives on Psychological Sciences*, November 2017a.

J. T. Wixted, L. Mickes, S. A. Wetmore, S. D. Gronlund, and J. S. Neuschatz. ROC analysis in theory and practice. *Journal of Applied Research in Memory and Cognition*, 6(3):343–351, September 2017b.

J. T. Wixted, E. Vul, L. Mickes, and B. M. Wilson. Models of lineup memory. *Cognitive Psychology*, 105:81–114, September 2018.

S. Q. Yates. Eyewitness identification: procedures for conducting photo arrays, January 2017.

J. Yin. Overview of inference about ROC curve in medical diagnosis. *Biometrics and Biostatistics International Journal*, 1(3), 2014.

A. Zapf, A. Hoyer, K. Kramer, and O. Kuss. Nonparametric meta-analysis for diagnostic accuracy studies. *Statistics in Medicine*, 34:3831–3841, 2015.

M. Zhao, W. G. Hayward, and I. Bulthoff. Holistic processing, contact, and the other-race effect in face recognition. *Vision Research*, 105:61–69, 2014.