UNIVERSITY OF CALIFORNIA,
IRVINE

Statistical Issues in Measurement with Applications in Forensics and Methylomics

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Hina Manojbhai Arora

Dissertation Committee:
Professor Hal Stern, Chair
Professor Daniel Gillen
Professor Babak Shahbaba

2023

DEDICATION

To my wonderful family,
my parents, Manoj and Minakshi,
my siblings, Dipali and Vandit, and my partner, Eric.

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisor, Prof. Hal Stern for helping me sharpen my skills as a statistician, researcher, and scientific writer. I am very grateful for having had the opportunity to work with him and look forward to applying the knowledge I have gained during my Ph.D. to my academic and professional career.

I am also thankful to my committee members, Prof. Daniel Gillen and Prof. Babak Shahbaba for their valuable feedback, support, and guidance through the different stages of my Ph.D. I am also thankful to the faculty and administrative staff at UCI Statistics. I would like to thank my friend and collaborator, Naomi Kaplan-Damary. She provided me with guidance that was invaluable for me as a novice researcher. Additionally, working with the collaborators at Conte Center has been very rewarding for me.

It would not have been possible for me to keep going without the steady support of my exceptional family: my parents, Manoj and Minakshi, my siblings, Dipali and Vandit, and my partner, Eric M. I am endlessly grateful for them and dedicate all my accomplishments to them.

This work was partially funded by the National Institute of Mental Health, Grant/Award Number: P50MH096889 and the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreements 70NANB15H176 and 70NANB20H019 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California Irvine, University of Virginia, West Virginia University, University of Pennsylvania, Swarthmore College and University of Nebraska, Lincoln.

Chapter 3 of this dissertation is currently under review in Law, Probability and Risk. The co-authors listed in this publication are Naomi Kaplan-Damary and Hal Stern. Hal Stern, the co-author listed in this publication, directed and supervised research which forms the basis for the dissertation.

Chapter 4 of this dissertation is currently under review in Forensic Science International. The co-authors listed in this publication are Naomi Kaplan-Damary and Hal Stern. Hal Stern, the co-author listed in this publication, directed and supervised research which forms the basis for the dissertation.

VITA

Hina Manojbhai Arora

EDUCATION

Doctor of Philosophy in Statistics **2023**
University of California, Irvine *Irvine, CA*

MS in Statistics **2021**
University of California, Irvine *Irvine, CA*

MS in Applied Mathematics and Statistics **2017**
Stony Brook University, NY *Stony Brook, NY*

B Tech in Electrical Engineering **2015**
Indian Institute of Technology, Indore *Indore, MP*

RESEARCH EXPERIENCE

Graduate Research Assistant **2019–2023**
Conte Center and CSAFE, *PI: Prof. Hal Stern*
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **2018–2020**
University of California, Irvine *Irvine, CA*

RESEARCH ARTICLES

Arora, et al., "Combining Reproducibility and Repeatability Studies with Applications in Forensic Science" **2023**
Law, Probability and Risk (Tentatively Accepted)

Arora, et al., "Reliability of Ordinal Outcomes in Forensic Black-Box Studies" **2023**
Forensic Science International (Submitted)

Arora, et al., "Identifying Clusters of Raters from Ordinal Data" **-**
(Prepared)

Arora, et al., "Latent Factor Analysis for Binomial Data with Applications to DNA Methylation Data" **-**
(In Preparation)

Davis et al., "Early life exposure to unpredictable parental sensory signals shapes cognitive development across three species" **2022**
Frontiers in Behavioral Neuroscience

Arora et al., "Interpolation on Gauss hypergeometric functions with an application" **2018**
Involve

CONFERENCE TALKS

Identifying intra-individual methylation profiles that distinguish infants impacted by adversity **2022**
Conte Center Symposium

Studying reproducibility and repeatability for pattern evidence comparisons **2022**
International Association for Identification

Reliability for binary and ordinal data in forensics **2022**
Joint Statistical Meetings

Reliability for ordinal data in forensics **2021**
Joint Statistical Meetings

Estimating repeatability and reproducibility with limited replications **2020**
Joint Statistical Meetings

INDUSTRY EXPERIENCE

AI-ML Engineering Intern **2022**
LinkedIn

AI-ML Engineering Intern **2021**
LinkedIn

Data Science Intern **2019**
Intuit

Data Science Intern **2018**
Obsidian Security

ABSTRACT OF THE DISSERTATION

Statistical Issues in Measurement with Applications in Forensics and Methylomics

By

Hina Manojbhai Arora

Doctor of Philosophy in Statistics

University of California, Irvine, 2023

Professor Hal Stern, Chair

The reliability of a measurement system is studied as a precursor to establishing the accuracy of the measurement system. Forensic science disciplines that rely on feature-based comparisons (e.g., handwriting analysis, fingerprint analysis) have been criticized for the absence of studies demonstrating reliability and accuracy. This has led to empirical evaluations through the use of "black-box" studies. Typically, data collected from inter-examiner (reproducibility) studies is analyzed separately from studies of intra-examiner (repeatability) studies. Motivated by these forensic studies, this dissertation develops methods to assess reliability for continuous, binary, and ordinal outcomes in forensics by combining inter-examiner and intra-examiner data for efficient estimation of reliability, while accounting for possible examiner-forensic sample interactions. Furthermore, we propose an exploratory method to cluster raters/ examiners to identify subpopulations that appear to apply similar decision-making approaches. The dissertation also includes the development of a statistical model to address measurement variability in methylomic studies.

Chapter 1

Introduction

Systems of measure or decision-making must have scientifically evaluated error and consistency rates. Accuracy is defined by the correctness of the measure and reliability is related to its consistency. Studying measurement reliability and accuracy is crucial in medicine, engineering, forensics, etc. Statistics plays a key role in the inference due to the empirical nature of the studies conducted and the need to have uncertainty bounds for obtained error and consistency rates.

Analyses of forensic evidence often involve subjective feature-based comparison assessments by forensic examiners, for example, latent fingerprint analyses, shoe-print examinations, firearm comparisons, etc. The National Academy of Sciences (NAS) and the President's Council of Advisors on Science and Technology (PCAST) emphasized the need for standardization and formal scientific foundations for forensic science disciplines in their reports (National Research Council, 2009; President's Council of Advisors on Science and Technology, 2016). Black-box studies were recommended in the PCAST report for empirically establishing error rates and reliability in forensic assessments. Following these recommendations, numerous studies have been conducted so far, to empirically establish the error rates

and precision of subjective pattern matching disciplines (Ulery et al., 2011; Ulery et al., 2012; Baldwin et al., 2014; Hicklin et al., 2021; Hicklin et al., 2022a; Hicklin et al., 2022b; Monson et al., 2023a).

Typically, black-box studies have two components; in the first trial, forensic examiners assess selected forensic samples that are presented to them exactly like they would in real casework. In the second trial, examiners re-evaluate a subset of the samples that they observed in the first trial. While reporting the results from these trials, the data collected from the different trials are analyzed separately. Additionally, the agreement between examiners, for categorical conclusions, is reported through contingency tables. In this dissertation, we will develop methods to analyze the data from black-box studies that will enable understanding the variance in decisions across examiners by combining the data from different trials. Additionally, our method will enable accounting for possible examiner-forensic sample interactions.

We then identify measurement issues that arise in exploratory dimension reduction/ factor analysis of proportion data. The motivating setup has a matrix of data $\frac{y_{ij}}{n_{ij}}$, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, d$ ($d >> n$), y_{ij} are the number of successes in n_{ij} counts, and it is of interest to see if a lower q-dimensional space ($d >> q$) can explain most of the variation in the matrix of proportions ($\frac{y_{ij}}{n_{ij}}$). The estimated proportions may not be reliable if they are based on a few counts, n_{ij}, and could influence the process of factor analysis/ dimension reduction. We aim to provide a method that accounts for binomial variation in counts while performing factor analysis. The motivating data is DNA methylation counts in n human infants at d CpG sites. The counts n_{ij} vary between 5-1000s and our proposed method will account for this heterogeneity in counts while exploring whether a few underlying factors can explain the variation in proportions of methylation.

We will now describe the layout of this dissertation. In Chapter 2, we provide background material that supports the manuscripts that follow. We describe issues of reliability in forensic science and the black-box studies being used to estimate reliability. The current

methods that are used to analyze the data from these studies are also reviewed. Limitations of the methods are identified which motivates the work reported in Chapters 3, 4, and 5. The final contribution of the thesis is a novel approach to exploratory factor analysis of proportion data. The motivation for this project is described along with the existing approach for such data. Chapters 3-6 consist are four manuscripts that summarize the thesis results. These manuscripts are in different stages of the preparation process.

Chapter 3 which is the first paper is under review in Law, Probability and Risk. Chapter 4 which is the second paper has been submitted to Forensic Science International. Chapters 5 and 6, which are the third and fourth papers, are in the preparation stage.

In Chapter 3 we apply a two-way ANOVA model with interactions as an approach for modeling reliability for continuous and binary outcomes in forensic science. These models contribute to the literature by allowing us to combine the data from intra-examiner and inter-examiner trials that are used in forensic black-box studies. This approach also allows for assessing examiner-sample interactions. We conduct simulation studies to study the effects of limited intra-examiner trials on the inference for reliability and thereby provide advice for future studies. Additionally, we study the effect of model misspecification on estimates of reliability and variance components. The methods are applied to data from two reliability studies; a signature complexity assessment study (Angel et al., 2017; Stern et al., 2018) and the 2011 FBI latent fingerprint study (Ulery et al., 2011).

Decisions in forensics are often on a categorical scale with a meaningful order to the categories. Chapter 4 begins with an introduction to the methods that are typically used to model ordinal outcomes. We then develop a model for ordinal decisions that accommodates inter-examiner and intra-examiner trials, allows for examiner-sample interactions, and allows for varying thresholds across examiners. We propose variations of the model that are constrained versions with fewer parameters to address settings with limited data. We use simulation studies to assess the effect of limited repeated decisions as well as the effect of a

misspecified model on the estimates of the variance parameters. We then use these methods to obtain inferences for the data from two reliability studies.

Chapter 5 explores a different way to analyze reliability studies with ordinal outcomes. This exploratory approach investigates whether there exist clusters of examiners that tend to make decisions similarly. We extend the model of Chapter 4 to incorporate a Dirichlet process mixture that can cluster examiners based on their outcomes. Alternatively, the same approach can be used to determine if there are clusters of samples that tend to be rated similarly. We use simulation studies to assess the effectiveness of the method and apply it to the data from two forensic reliability studies as well as the data from a maternal depression study.

Chapter 6 describes a project that arose from our work with the Conte Center at UCI. The aim of the Conte Center is to understand the effects of early life adversity in the cognitive development across species. One of the Center's projects is examining the effects of adversity by looking at changes in the DNA methylation. Principal component analysis (PCA) of DNA methylation has been used to quantify the variation that arises across CpG sites (Jiang et al., 2019). The PCA analysis is applied to a matrix with each row corresponding to a different site on the genome, each column identifying a sample (e.g., an individual at a given time point), and the matrix entry reporting the percentage of read at that site which are methylated. PCA ignores the variation in the number of reads which lead to heterogeneous variances across the matrix entries. We develop a factor analysis model that accounts for variation in methylation proportions with heterogeneous sample sizes and allows us to examine whether the variation in methylation proportions may be explained by a few latent variables. We use simulation studies to assess the performance our model and apply it to the motivating data set.

Finally, Chapter 7 summarizes the contributions of the methods proposed in this dissertation along with the conclusions. We suggest some future directions for each of the projects. We

follow that with supporting materials in the Appendix.

Chapter 2

Background and Aims

2.1 Reliability

Reliability is a concept that arises in the context of a measurement or decision process. Reliability focuses on the consistency of measures or decisions. It is distinct from validity/ accuracy which are defined by the correctness of the measures or decisions. Reliability is a precursor to accuracy because the correctness of examiner decisions is limited by whether they are consistent. We are especially interested in two types of reliability. Reproducibility relates to the consistency of measurement or decisions when different examiners are assessing the same item/ sample. Reproducibility refers to inter-examiner reliability. Repeatability is the consistency of decisions from the same examiner on the same item/ sample at two different points in time. Repeatability is intra-examiner reliability.

Reliability is of interest in many fields such as engineering, radiology, biology, etc. For example, Tsai (1988), Vardeman and VanValkenburg (1999), Weaver et al. (2012), Vardeman (2014) have discussed methods to analyze repeatability and reproducibility of gauge measurements. Heydorn et al. (2000), Furlan et al. (2007), Pearson et al. (2011) have provided

methods to analyze reliability in biological applications. Van Wieringen and De Mast (2008) have discussed reliability for binary measurements such as pass/ fail during inspections. Here we briefly introduce some ways that reliability is measured and introduce applications of reliability in forensic science.

2.1.1 Measurements of reliability

There are a variety of approaches to assessing reliability. The methods vary by data type. We provide a brief summary here.

Continuous Data The correlation between two sets of continuous measurements on $j = 1, 2, ..., J$ items/ samples, (Y_{1j}, Y_{2j}) is defined by:

$$\text{Correlation} = \frac{1}{Js^2} \sum_{j=1}^{J} (Y_{1j} - \bar{Y})(Y_{2j} - \bar{Y}), \tag{2.1}$$

where, $\bar{Y} = \frac{1}{2J} \sum_{j=1}^{J} (Y_{1j} + Y_{2j})$ and $s^2 = \frac{1}{2J} \sum_{j=1}^{J} (Y_{1j} - \bar{Y})^2 + (Y_{2j} - \bar{Y})^2$. Note here that correlation is one type of reliability but it does not judge exact agreement. For example $\mathbf{Y}_1 = (1, 2, 3, 4, 5)$ and $\mathbf{Y}_2 = (2, 3, 4, 5, 6)$ have perfect correlation but never agree.

The intraclass correlation coefficient (Shrout and Fleiss, 1979), also denoted as ICC, is used to evaluate the correlation between measurements from I raters on J items. Shrout and Fleiss (1979) provided ICC evaluations for three cases: **a**. each sample is assessed by I different set of examiners from a larger population of examiners, **b**. I random examiners are samples from the population and they each assess the same set of J samples, **c**. The population of interest consists of I examiners and they each rate the same set of J samples. Let Y_{ij} denote the measurement from examiner i on sample j, then for case **a**, the model and ICC were evaluated as follows:

$$
\begin{aligned}
Y_{ij} &= \mu + \gamma_j + \epsilon_{ij} \\
\text{ICC} &= \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_\epsilon^2}
\end{aligned}
\tag{2.2}
$$

Here, μ is the population mean and $\mu + \gamma_j$ is the true measurement from sample j. γ_j is assumed to be a random effect from population $N(0, \sigma_\gamma^2)$. ϵ_{ij} is random noise modeled to be drawn $N(0, \sigma_\epsilon^2)$. For case **b**, the following model and ICC are proposed:

$$
\begin{aligned}
Y_{ij} &= \mu + \alpha_i + \gamma_j + \delta_{ij} + \epsilon_{ij} \\
\text{ICC} &= \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_\alpha^2 + \sigma_\delta^2 + \sigma_\epsilon^2}
\end{aligned}
\tag{2.3}
$$

The interpretation for μ and γ are the same as the case **a**. However, α_i is examiner i tendency to rate a sample which is assumed to belong to the population $N(0, \sigma_\alpha^2$ and δ_{ij} is the interaction between examiner i and sample j assumed to be drawn from the population σ_δ^2. Finally, for case **c**, the same model is used as **b** given in equations (2.3), however, α_i are fixed effects so that $\sum_i \alpha_i = 0$, additionally, $\sum_i \delta_{ij} = 0$. ICC for case **c** is given as follows:

$$
\text{ICC} = \frac{\sigma_\gamma^2 - \frac{\sigma_\delta^2}{I-1}}{\sigma_\gamma^2 + \sigma_\delta^2 + \sigma_\epsilon^2} \tag{2.4}
$$

Categorical data For categorical data where each measurement results in one of $l = 1, 2, ..., L$ categories, percentage agreement is a common reliability measure. Again, assume that $= 1, 2, \ldots, I$ be the examiners/ judges and $j = 1, 2, \ldots, J$ are the samples and Y_{ij} is the categorical outcome. Denote n_{jl} as the number of decisions across examiners for sample j in category l for $l = 1, 2, \ldots, L$ and $n_{j\cdot}$ is the total number of decisions on sample j. Then, the percentage agreement is given as follows:

$$p_j = \frac{1}{n_{j.}(n_{j.}-1)} \sum_{l=1}^{L} n_{jl}(n_{jl}-1)$$
$$P = \frac{1}{J} \sum_{j=1}^{J} p_j, \tag{2.5}$$

where, p_j is the percentage agreement on sample j and P is the mean percentage agreement across samples.

Percentage agreement may be optimistic while evaluating agreements because it does account for the possibility that agreements may have happened by chance. Cohen's κ and its weighted versions (Cohen, 1960; 1968;Fleiss, 1971) use corrections to percentage agreements to account for chance agreements. There are other proposed measures such as polychoric correlation (Pearson, 1900) that uses the correlation between latent variables that are used to model ordinal data, Cronbach's α (Cronbach, 1951), Krippendorff's α (Krippendorff, 2011) that prioritizes disagreements, etc. Refer to Hallgren (2012), Gadermann et al. (2012), Nelson and Edwards (2015), Raadt et al. (2021) for further discussion on these measures.

2.2 Reliability in Forensic Science

Forensic evidence has a notable effect on case proceedings (Peterson et al., 2013). Forensic science is the application of scientific means to investigate a crime through the evidence that is collected at the scene of crime, for example, latent fingerprints, shoe prints, gunshot residue, etc. However, the specific evidence and the testimony provided is governed by Federal Rule of Evidence 702 as well as the holdings in *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923) and *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579, 113 S. Ct. 2786 (1993), in order for it to be admissible. The Federal Rule of Evidence 702 states that expert testimony about forensic evidence may be admissible if the expert testimony is based on reliable methods applied reliably to the case evidence. *Frye v. United States*,

293 F. 1013 (D.C. Cir. 1923) states that expert testimony must be based on scientifically established methods. *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579, 113 S. Ct. 2786 (1993) additionally states that the techniques used by the examiner must be generally accepted in the scientific community with peer-reviewed studies and known error rates.

The National Academy of Sciences (NAS) published a report titled "Strengthening Forensic Science in the United States: A Path Forward" (National Research Council, 2009) detailing the steps that must be taken to standardize and establish scientific foundations for forensic science disciplines. Among other recommendations, the report emphasized the need for studies assessing the reliability and validity of forensic science decisions. This recommendation also specified the need for quantifiable measures and uncertainties for reliability and accuracy.

Several years later, the President's Council of Advisors on Science and Technology prepared a report "Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods" (President's Council of Advisors on Science and Technology, 2016) assessing the scientific evidence regarding a number of disciplines such as latent fingerprint analysis, footwear analysis, and firearm analysis. The report recommended the use of "black-box" studies to provide evidence regarding accuracy and reliability of pattern comparison disciplines. Such studies provide a series of examples (questioned and known pairs), for which the ground truth is known by the study designers, to a sample of examiners. The examiners are told to use their usual process and to provide their conclusion regarding the evidence. In essence, they are treated as a black box, taking in evidence as input and outputting a conclusion.

2.2.1 Black-box studies

In 2010, the FBI conducted the first large-scale study to assess decision-making in latent fingerprint comparisons (Ulery et al., 2011; Ulery et al., 2012). Following this, there were two studies for assessing reliability and validity for bullet and cartridge case comparisons (Baldwin et al., 2014; Monson et al., 2023a), a study in bloodstain pattern analysis (Hicklin et al., 2021), a study for handwriting analysis (Hicklin et al., 2022a), and a study for footwear analysis (Hicklin et al., 2022b).

Typically, examiners participating in black-box studies work for various national, state, and local laboratories. Forensic examiners can differ in their training backgrounds, certifications, and years of experience. Examiners are assigned a number of forensic samples, for which the ground truth is known by the study designers, and they are asked to make forensic assessments on a pre-defined outcome scale just like they would in real casework. After some time has passed from this first assessment/ study, some of the examiners are asked to re-assess a subset of the samples that they initially observed. We call the first part of the study, the reproducibility trial because the aim is to assess the consistency of decisions across examiners and to assess the accuracy of the disciplines on average. The second part of the study is known as the repeatability component as its aim is to study is to assess the consistency of decisions by the same examiner.

2.2.2 Limitations in Analyses

Black-box studies are expensive and time intensive, and due to this fact, in the studies conducted thus far, the reproducibility trial is much larger in terms of the total number of decisions compared to the repeatability trial. For example, in the FBI latent print black-box study (Ulery et al., 2011) there were $\approx$ 17,000 decisions in the reproducibility part of the study and the repeatability trial had $\approx$ 1900 decisions.

The data collected from the reproducibility and repeatability trials in black-box studies are typically analyzed separately. For example, in the results from latent fingerprint examination decisions which were analyzed in Ulery et al. (2011; 2012), reproducibility is reported using the data from the first trial conditionally on mated and non-mated pairs. Additionally, repeatability was reported using a subset of 72 out of 169 examiners that participated in the repeatability trial. Similarly, in the results from handwritten signature complexity data analyzed by Stern et al. (2018), repeatability is assessed only through the repeated decisions that were made on less than 6% of signature samples in the reproducibility trial.

Reliability, for categorical outcomes in black-box studies, is typically evaluated through percentage agreement or Cohen's κ (Cohen, 1960). Note that this aggregate measure does not account for differences among examiners and samples. Additionally, it ignores the ordering of categories if there is one.

It is worth noting that in the black-box studies conducted so far, covariates related to examiners or samples are not provided, though some aggregate survey information about their education, experiences, employer agencies, etc. may be available.

2.3 Statistical Models for Reliability in Forensic Science

A common approach in reliability studies across disciplines is the two-way random effects ANOVA. For example, it has been applied in reliability studies in manufacturing, radiology, etc. (Vardeman and VanValkenburg, 1999; Pearson et al., 2011). This method can be applied to forensics as well but there are several challenges such as limited data and non-continuous/ Gaussian outcomes. We briefly discuss the two-way ANOVA approach and discuss the limitations in the application to data from black-box studies.

2.3.1 Two-way random effects ANOVA

The two-way ANOVA model (Tabachnick and Fidell, 2007) is used to model outcomes that vary according to the levels of two categorical variables. For example, let there be $i = 1, 2, ..., I$ examiners and $j = 1, 2, ..., J$ samples and each examiner provides K decisions on each sample, then the outcomes Y_{ijk} can be modeled as follows:

$$Y_{ijk} = \mu + \alpha_i + \gamma_j + \delta_{ij} + \epsilon_{ijk} \tag{2.6}$$

Here, the outcome Y_{ijk} is assumed to depend linearly on population mean μ, and the effects α_i, γ_j, and δ_{ij}. α_i is an examiner effect that dictates how examiners are expected to deviate from the population mean, γ_j is a sample effect that dictates how samples deviate from the population mean, and δ_{ij} is an interaction effect between examiners and samples which dictates how examiner i deviates from their tendency α_i for sample j. ϵ_{ijk} is random noise that is assumed to follow $N(0, \sigma_\epsilon^2)$ distribution. The effects may be modeled as random effects when the examiners and samples belong to a larger population and do not need to be estimated individually. A common distribution used to model random effects is the normal distribution:

$$\begin{aligned} \alpha_i &\overset{i.i.d.}{\sim} N(0, \sigma_\alpha^2) \\ \gamma_j &\overset{i.i.d.}{\sim} N(0, \sigma_\gamma^2) \\ \delta_{ij} &\overset{i.i.d.}{\sim} N(0, \sigma_\delta^2) \end{aligned} \tag{2.7}$$

This method is typically applied to continuous data, for example, gauge measurements in manufacturing (Vardeman and VanValkenburg, 1999; Vardeman, 2014) and agriculture (Aguirre et al., 2020). In forensics, this model can be directly applied to some complexity data such that handwriting complexity data discussed in Alewijnse et al. (2011) that were on a scale of 0-100 that may be approximated to a continuous scale.

As discussed previously in equations (2.3), in a two-way random effects ANOVA with interactions setting given by the equations (2.6), reliability can be estimated as follows:

$$\begin{aligned} \text{Reproducibility} = \text{corr}(Y_{ijk}, Y_{i'jk'}) &= \frac{\sigma_\gamma^2}{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\delta^2 + \sigma_\epsilon^2} \\ \text{Repeatability} = \text{corr}(Y_{ijk}, Y_{ijk'}) &= \frac{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\delta^2}{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\delta^2 + \sigma_\epsilon^2} \end{aligned} \tag{2.8}$$

.

We now discuss some limitations of the application of the two-way random effects ANOVA model in forensics.

2.3.2 Applications in Forensic Science

Outcomes in forensics are often non-continuous, they may be categorical/ ordinal and sometimes binary. In this dissertation, we will model such binary and ordinal outcomes by assuming that they depend on a latent scale.

It is interesting to estimate possible examiner-sample interactions by combining the data from the reproducibility and repeatability trials in the black-box studies. We would like to understand how interactions limit reliability. Note that with plenty of repeated decisions, σ_δ^2 can be estimated very well. However, as previously mentioned, black-box studies typically collect repeated decisions on a very limited number of samples which limits the ability to address interactions.

In this dissertation, Chapter 3 applies the two-way random effects ANOVA model with interactions to combine the reproducibility and repeatability studies for studies that collect continuous or binary outcomes such as signature complexity data, match/ no match outcomes on fingerprint comparisons. Reliability is evaluated using intraclass correlations and typical agreement statistics on posterior predictive data sets.

Chapter 4 extends the methods in Chapter 3 to model ordinal outcomes. We propose two methods to that end, one method enables the estimation and inference for examiner thresholds to make decisions in a specific category and the other method is a constrained version of the first method that is more appropriate when interactions need to be estimated with limited data.

The methods in Chapters 3 and 4 we have looked at a few methods to model variability in continuous, binary, and ordinal outcomes in black-box studies. In Chapter 5 we propose a method that encourages parameter sharing between raters/ samples by clustering examiners based on their ratings. This technique is exploratory in nature and can be useful for hypothesis generation with the covariates.

2.4 Early Life Adversity and DNA Methylation

The Conte Center at the University of California, Irvine is interested in exploring the effects of early-life adversity on cognitive and emotional outcomes later in life. They have conducted studies across species to explore how maternal unpredictability, a specific form of early life adversity, can affect development in infants (Baram et al., 2012; Davis et al., 2017; Davis et al., 2019; Short and Baram, 2019).

2.4.1 Entropy

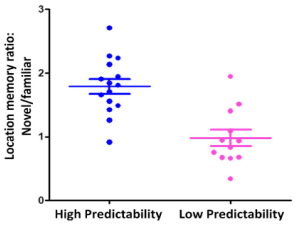
Conte Center researchers have found that entropy rate for sequences of maternal behaviors can be used to assess unpredictability across species. Entropy (Shannon, 1948) is a concept in information theory that can be used to quantify the predictability of random variables. Let X be a discrete random variable with probability mass function defined by $P(X = x_i) = p_i$ for $i = 1, 2, ..., n$, then the entropy of the variable X is defined by:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i). \tag{2.9}$$

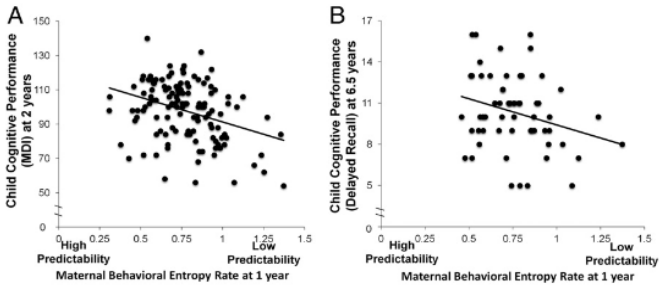
The concept can be expanded to address the entropy rate for a stochastic process $\{X_t : t = 1, 2, \dots\}$. For example, for a first-order Markov chain with transition probabilities p_{ij} and stationary distribution $\pi = \{\pi_i\}_{i=1}^n$, entropy is defined as (Vegetabile et al., 2019):

$$H(X) = -\sum_{i,j} \pi_i p_{ij} \log_2(p_{ij}). \tag{2.10}$$

Davis et al. (2017) used this concept of entropy to model maternal sensory input to infants. Higher entropy scores indicate less predictable sequences. They found that in rats and human beings, low predictability of maternal behaviors was associated with poorer cognitive outcomes as seen in Figures 2.1 from the paper.



(a)



(b)

Figure 2.1: Figures show that infants that experience higher unpredictability have poorer cognitive development. In subfigure 2.1a, they observed that rats that experienced low predictability performed worse on a spatial memory task compared to rats that experienced more predictable maternal care. In subfigure 2.1b they noted that human infants that experienced low predictability at 1 year of age had a worse Mental Development Index (MDI) at 2 years of age and 6.5 years of age compared to infants with a more predictable input (figures from Davis et al., 2017, used with permission from PNAS).

2.4.2 DNA Methylation

Various Conte Center projects explore possible mechanisms using rodents as an animal model and study the consequences of unpredictable maternal signals in human cohorts. One project investigates whether early life unpredictability may leave a certain biological marker. DNA methylation (Moore et al., 2013) data were analyzed to test this hypothesis.

DNA samples are collected for $i = 1, 2, \ldots, n$ subjects and methylation counts y_{ij} for CpG sites $j = 1, 2, \ldots, d$ are recorded from n_{ij} reads for individual i at CpG site j. These reads are collected at least two points in time. Due to limitations in sequencing technologies and the quality of DNA samples, it is difficult to observe reads for each site j in each subject i. Furthermore, due to intra-individual variation in CpG changes across time points it is difficult to isolate sites that have significant changes in methylation across individuals due to unpredictability.

Jiang et al. (2019) conducted a study with infant rats and analyzed the changes in their methylation profile when they were subjected to different early life experiences. As seen in Figure 2.2 that is from the paper, DNA samples were collected from rats on post-natal day two (P2), and then they were divided into Limited Bedding and Nesting (LBN) or control groups. Then DNA samples were collected again on post-natal day ten (P10). Differentially methylated sites (DMSs) were identified when they were significantly methylated in two or more pups. DMSs were then tiled into regions of 100 base pairs to obtain differentially methylated regions (DMRs).

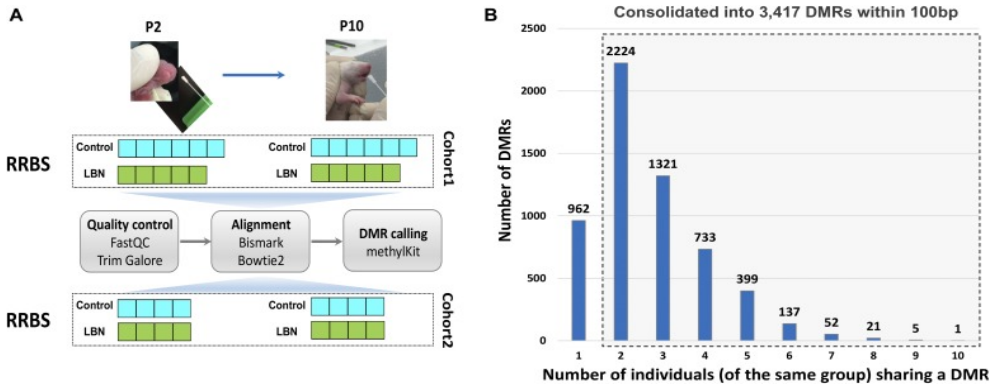


Figure 2.2: Figures show that DNA samples were collected from rats at ages P2 and P10. Differentially methylated sites were identified when they were shared by at least two infant rats and then tiled into differentially methylated regions (figures from Jiang et al., 2019, used with permission from Life Science Alliance).

2.4.3 Principal Component Analysis

Principal component analysis (Abdi and Williams, 2010) is a dimensionality reduction technique used to explain variation in high dimensional data with fewer dimensions. The method relies on linear algebra techniques and identifies eigenvectors that correspond to the q-largest eigenvalues of the variance-covariance matrix of the data.

As seen in Figure 2.3 PCA on differentially methylated regions in rat pups was able to differentiate between age (P2 vs P10) but not able to differentiate between LBN and control groups.

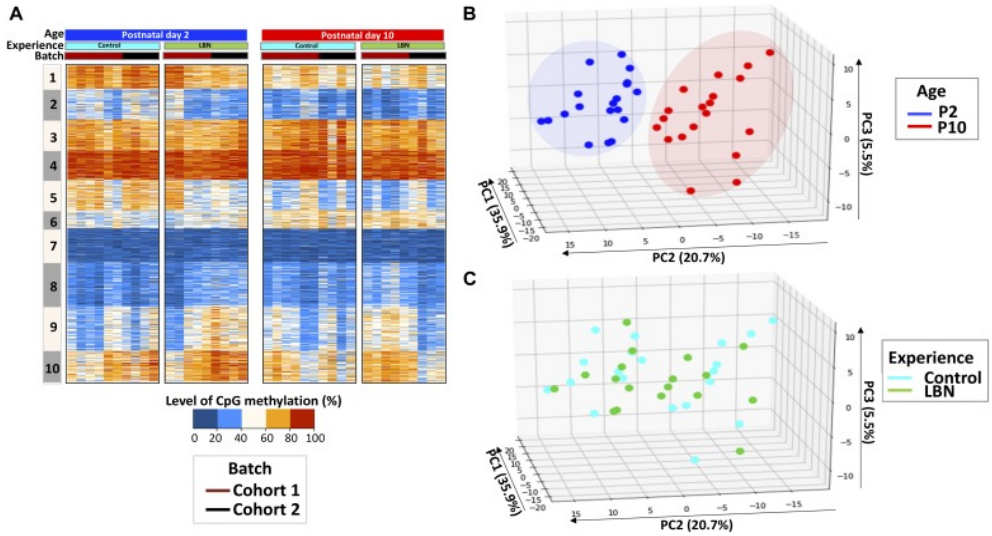


Figure 2.3: Figures show that PCA on differentially methylated regions was not able to differentiate between experiences (figures from Jiang et al., 2019, used with permission from Life Science Alliance).

Jiang et al. (2019) then defined the quantity "δ-methylation" $= \log_2(\frac{P_{10}}{P_2})$ to account for variation in intra-individual methylation. The principal component analysis on δ-methylation was able to differentiate between pups that experienced early life adversity and control groups as demonstrated in Figure 2.4.

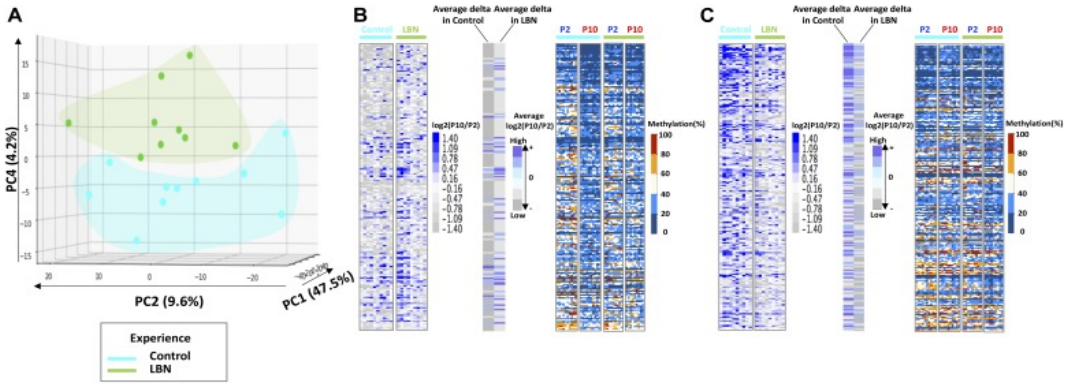


Figure 2.4: Figures show that δ-methylation was able to differentiate between control and LBN experiences (figures from Jiang et al., 2019, used with permission from Life Science Alliance).

Unlike in rats, PCA in human beings has not been as successful in accounting for variation in data. Also, in human beings there are no clear early life adversity and control groups. Therefore, childhood experiences will be quantified through entropy of maternal sensory input. We note that PCA on differentially methylated regions or δ-methylation ignores the variation that arises due to heterogeneity in sample sizes that are used in estimating the proportion of methylation. For example, we observed that in over 71% of the DMSs the number of reads used for obtaining the proportion of methylation was less than 50. We would like to account for such variation while analyzing methylation patterns.

2.4.4 Factor Analysis

Factor analysis (Gorsuch, 2014) is a classic dimensionality reduction technique used widely in psychology and sociological sciences to explain high dimensional data with fewer underlying factors. For example, if the variation in the data $Y_{d\times n}$, where n is the sample size and d is the dimension of the observations, can be explained by a q-dimensional space, where $q < d$, then:

$$\begin{aligned} Y_{d\times n} &= W_{d\times q} X_{q\times n} + \mu_{d\times 1} 1_{1\times n} + E_{d\times n} \\ \tilde{x_i} &\sim N(0_{q\times 1}, I_{q\times q}) \end{aligned} \tag{2.11}$$

W is known as the factor loading matrix and $\tilde{x_i}$ is a q-dimensional vector that represents observations $\tilde{y_i}$ with fewer dimensions. X is independent of E. The total variation in the data is $\text{Cov}(Y) = \Sigma = WW^T + \text{Cov}(E)$.

PCA and FA have similarities, for example, PCA also seeks to perform dimension reduction. Additionally, Tipping and Bishop (1999) discuss how $E = \sigma^2 I$ in 2.11 is a probabilistic PCA model. However, there are key differences between these models, such as PCA assumes that there is no specific variation E that is attributed to the features. Additionally, FA and probabilistic PCA are generative models, i.e., they assume that data is generated from the specified distribution. However, PCA is just a method to project the features on a lower dimensional space. FA is more interpretable compared to PCA. E in FA literature is typically assumed to be a diagonal matrix so that each feature has a specific variance and conditional on x_i, y_{ij} are independent for all j. PCA assumes no such structure.

Confirmatory factor analysis (CFA) is conducted with a specific hypothesis in mind and involves choosing a q as well as investigating whether W has a specific structure. We are interested in exploratory factor analysis (EFA) because although we hope to find associations between the factors and entropy measures, we do not a priori know the number of desired factors q and the specific methylated regions that W should highlight.

Multiplying W to an orthogonal matrix $L_{q\times q}$ and multiplying $L^T_{q\times q}$ to X in equations (2.11) leads to equivalent parametrizations of the model. Therefore, fitting factor analysis models typically involves imposing certain constraints on W, for example fixing the upper triangular part of W to 0 is a popular technique (Geweke and Zhou, 1996; Bernardo et al., 2003; Lopes and West, 2004). This constraint may be too restrictive and recently sparsity inducing priors have been used to avoid identification issues in addition to encouraging factors to load onto fewer dimensions.

Chapter 3

Combining Reproducibility and Repeatability Studies with Applications in Forensic Science

3.1 Introduction

Presentation of scientific evidence in U.S. courts is governed by *Daubert v. Merrell Dow Pharmaceuticals, Inc.,* 509 U.S. 579 (1993) and Federal Rule of Evidence 702. These say that the expert testimony based on scientific knowledge must have "a standard of evidentiary reliability", which means that the testimony should be valid and reliable, and must be the "product of reliable principles and methods". A 2009 report from the National Academy of Sciences (National Research Council, 2009) identified concerns with the scientific foundations of some forensic disciplines and called for scientific studies to establish their validity and accuracy. A subsequent 2016 report from the President's Council of Advisors on Science and Technology (President's Council of Advisors on Science and Technology, 2016), focused on feature comparison methods, re-iterated the concerns raised by the NAS report,

and identified strategies for establishing reliability and validity of forensic analysis. We note here that the European guidelines (Willis et al., 2015) recommend using probabilistic statements while evaluating forensic evidence and the reporting of likelihood ratios rather than categorical conclusions. The methods we develop are based on the current approach of the U.S. for assessing forensic evidence.

We focus here on the reliability of assessments for the feature-based comparison methods discussed by the PCAST report. Forensic disciplines that rely on feature-based comparisons include fingerprints, shoeprints, firearms, etc. In a standard forensic examination, expert examiners are asked to provide assessments of samples collected from a crime scene, e.g., a bullet cartridge case, and a second analogous sample collected from a known source, e.g., a suspected firearm. The decision process, based on the training, experience, and expertise of the examiner, is subjective and thus prone to psychometric variation across examiners. We provide a general-purpose statistical methodology to assess the reliability of a subjective workflow of this type.

Reliability, in the context of subjective forensic examiner decisions, refers to the consistency of decisions made for the same sample. Reliability is distinct from accuracy, which refers to the correctness of the decisions. There are two types of reliability that are of interest to the forensics community. Repeatability refers to the consistency of the decisions made by the same examiner in judging the same sample (or evidence) at two different times. Reproducibility refers to the consistency of the decisions made by different examiners in judging the same sample. Reliability is of interest in its own right. Reliability is necessary but not sufficient for accuracy.

The PCAST report emphasizes the importance of conducting "black-box" studies for evaluating the reliability and accuracy of feature-based comparison methods. In a typical study, forensic examiners, recruited from various government and private agencies, are provided with forensic samples similar to the ones they would see in real case work. They are pro-

vided with questioned samples (pieces of evidence) and exemplar samples which are collected under ideal circumstances from known sources. The examiners are asked to make source determinations for the questioned sample just like they would in practice. The ground truth for the comparisons, that is whether the questioned sample and exemplar sample come from the same source, is known in a black-box study. Studies of this type are called black-box studies because the decision-making process is subjective and the individual steps in the decision-making process are unspecified. The decision-making process is treated like a "black box". In many of the black-box studies conducted so far, data has been collected such that it enables two different types of reliability assessments. The primary design of the study can be thought of as a reproducibility study, a number of examiners make decisions on a certain number of samples and the goal is to assess the consistency of examiner decisions on the same samples. The studies also include a second component, which can be thought of as a repeatability study, wherein examiners give repeated decisions on a subset of the samples they encountered in the reproducibility study. It has also been observed that the repeatability trials rely on much smaller samples compared to the reproducibility trial. For example, in the latent fingerprint examination study of Ulery et al. (2011, 2012) described below in Section 3.2.2, the reproducibility trial had 17121 total decisions and the repeatability trial had 1906 decisions.

In response to the NRC (2009) and PCAST (2016) reports, black-box studies have been conducted for many forensic disciplines. This includes latent print examinations (Ulery et al., 2011, 2012), blood stain pattern analysis (Hicklin et al., 2021), bullet and cartridge case comparisons (Baldwin et al., 2014; Monson et al., 2023a), and footwear analysis (Hicklin et al., 2022b). Additional studies are underway. This paper demonstrates and evaluates a statistical model that can be used to analyze the reliability data from black-box studies.

Measurement reliability is also a key concept in engineering, radiology, and many other disciplines (Weaver et al., 2012; Vardeman and VanValkenburg, 1999; Vardeman, 2014; Pearson

et al., 2011; Furlan et al., 2007; Van Wieringen and De Mast, 2008). Researchers, in some applications, have also identified instances in which there appear to be interactions such that the performance of examiners can vary based on the characteristics of the object. For example, Tsai (1988) studies variability in gauge measurements accounting for the interaction between operator and parts. Heydorn et al. (2000) studied reproducibility in biofilm experiments while accounting for the interaction between bacterial strain and experiment round.

Other related work considers the reliability of forensic examination from a decision-making rather than a measurement perspective. Item response theory (IRT) models are often used for measuring the responses from individuals to a set of questions. Luby and Kadane (2018) and Luby et al. (2020) have employed Rasch models and item response theory (IRT)-like models for the analysis of the data from proficiency testing of forensic examiners. Luby and Kadane (2018) uses an IRT model to understand the variation in examiner behavior while accounting for item difficulty. Luby et al. (2020, 2021) extends the previous work and provides a framework for assessing proficiency with a decision tree-like model for the sequential decision-making process in fingerprint comparisons.

We focus here on developing a general statistical approach to quantify the variation in subjective forensic determinations. We are motivated by several recent black-box studies which include a large reproducibility study and a smaller repeatability study. Thus far the two components have been analyzed separately (Stern et al., 2018; Ulery et al., 2012). Combining the two should provide a method to assess for the possible presence of examiner-sample interactions, as well as, provide greater precision in estimating examiner and sample-specific tendencies in reliability inferences. An interaction between examiner and sample implies that examiners have different tendencies for rating forensic samples or that examiner abilities/ thresholds change depending on the forensic sample. Some results in Ulery et al. (2012) and Hicklin et al. (2020) suggest possible interactions in latent print analysis with higher

repeatability and reproducibility for easier prints than for prints that were rated to be difficult by examiners. It is essential to understand the magnitude of these interactions and understand how they may limit overall repeatability and reproducibility. We begin our discussion by briefly describing in Section 3.2 the data from two forensic studies that motivate our work. In Section 3.3, we introduce the statistical method used to assess reliability for decisions that can be approximated to a continuous scale and then extend to address binary decisions. Section 3.4 has results from simulation studies of different designs where continuous and binary data are generated from a known distribution and the aforementioned model is used to fit the data. Section 3.5, presents analyses of the data sets described in Section 3.2, handwritten signature complexity data (Angel et al., 2017), and fingerprint comparisons data (Ulery et al., 2011, 2012). Finally, Section 3.6 discusses the results, limitations of the approach, and future work.

3.2 Data

This section describes data collected from two forensic science studies that incorporated reliability assessments. The first is a study regarding assessments of the complexity of handwritten signatures (Angel et al., 2017; Stern et al., 2018). The second is a large-scale study conducted by the FBI to investigate the accuracy and reliability of fingerprint comparison decisions (Ulery et al., 2011, 2012).

3.2.1 Handwritten Signature Complexity

Found and Rogers (1996) and Found et al. (1998) provided a statistical method to define complexity for handwritten signatures. Dewhurst et al. (2007) reported that complex signatures are difficult to imitate. Forensic document examiners are more confident and accurate in their decisions while judging handwriting of higher complexity as compared to their decisions while judging handwriting of lower complexity (Sita et al., 2002). This suggests that

assessing the complexity of handwritten samples is a key step for an examiner.

Signatures were collected in a Los Angeles Police Department (LAPD) and Los Angeles County Sheriff's Department (LASD) study described by Angel et al. (2017) and Stern et al. (2018). The study was focused on assessing the reliability of complexity assessments and the characteristics associated with those judgments. The data are also intended as a resource for a future study of the effect of complexity on examiner decisions. A total of 123 participants, ages $21 - 70$, submitted 5 samples of their signatures on paper in the study. A total of five forensic document examiners (FDE) with an average experience of 28.8 years (s.d.= 8.9 years) provided complexity assessments based on images of 300 dpi (dots per inch) resolution for each of the 123 signers using both a 3 point rating scale and a 5 point rating scale. The complexity rating reflected the examiner's judgment of the difficulty with which the signature could be replicated, with the 3-point scale corresponding to the choice of fairly easy, medium, or difficult and the 5-point scale including easy, fairly easy, medium, difficult, and very difficult as options. The five examiners provided repeated decisions on a very small subset, 7 out of the 123 signatures.

Stern et al. (2018) analyzed the reproducibility study (123 signatures assessed by five examiners) and the repeatability study (seven signatures assessed twice by each examiner) separately. They analyzed the 5-point scale data and treated the outcome as a continuous measure. The repeatability was found to be quite similar to the reproducibility. This is a bit surprising but may also reflect the small number of observations in the repeatability study. We are interested for: i) combining the data sets from the two types of trials for a more efficient estimation of reliability, and ii) deriving information about possible examiner-sample interactions from this very limited repeatability data set.

3.2.2 Latent Print Comparisons Reliability

The analysis of latent print evidence has a long history in the United States, dating back to 1911 (see *People v. Jennings*). In the face of high-profile errors such as the misidentification in the Mayfield case (Office of the Inspector General, 2006), there was a need to formally assess the accuracy and reliability for the friction ridge examination procedure. The first large-scale black-box study for evaluating the accuracy and reliability of latent fingerprint analysis was conducted by the FBI (Ulery et al., 2011, 2012). Fingerprint examination is a sequential, multi-part process and it is important that the outcomes from each step of the process are reliable and accurate. Most U.S. agencies use the ACE-V procedure in which the following steps are executed: Analysis, Comparison, Evaluation, and Verification (ACE-V).

3.2.2.1 ACE-V workflow

We begin with a brief description of the ACE-V approach to fingerprint analysis (see Figure 3.1). As a first step, latent prints collected from the crime scene are analyzed by forensic examiners to assess the quality of the prints. Ideally, the collected print should have sufficient distinguishing marks, patterns, or minutiae to make source determinations. Based on the operating procedures of the agency, this analysis step can have a binary outcome (Value for Individualization/ Not Value for Individualization) or a trinary outcome (Value for Individualization, Value for Exclusion Only, No Value). If a latent print is not deemed suitable for comparison i.e., Not Value for Individualization on the binary scale or No Value on the trinary scale, it is not used for comparisons. If a latent is "Of Value" (VID/ VEO) in the analysis step, exemplars are provided to the examiner in the comparison step, where examiners compare the latent print with the exemplar for levels of details such as ridge flow, minutiae, and pattern types. The exemplar prints are comparison prints collected under different circumstances and may be obtained from suspects or from an Integrated Automated

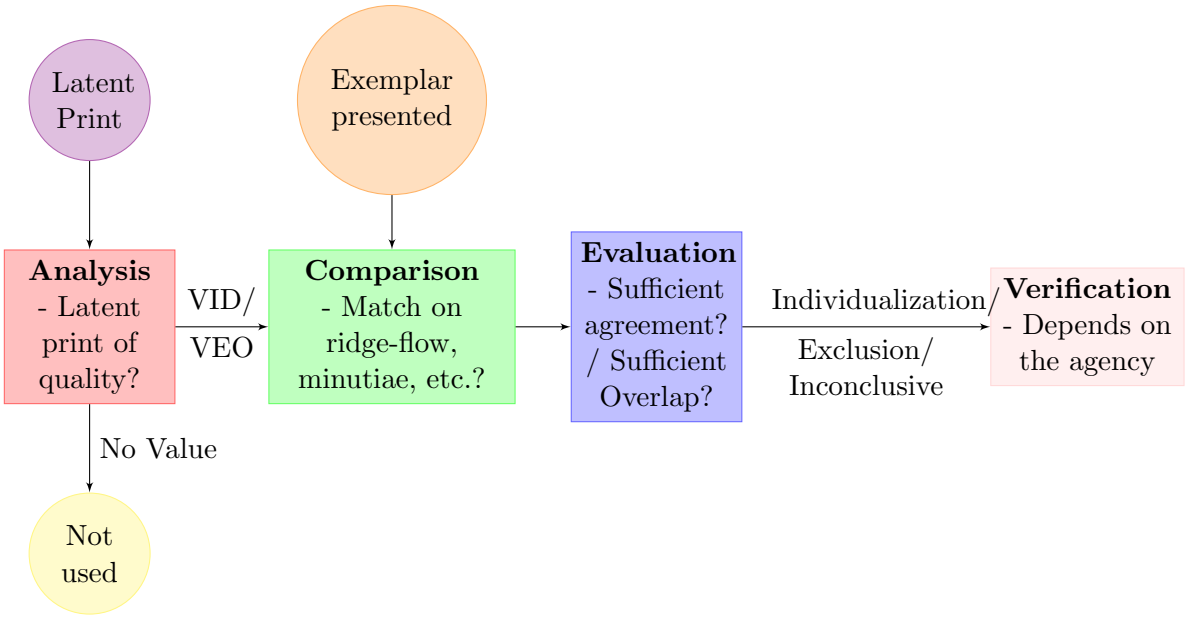


Figure 3.1: Simplified version of ACE-V workflow.

Fingerprint Identification System (IAFIS) which relies on pattern matching and other relevant information to produce top potential matches for a latent print. The evaluation step involves an assessment of the similarities and differences between the latent print and the exemplar. This step has three outcomes based on the features shared between the latent print and the exemplar: Individualization (questioned print and exemplar are believed to have come from the same source), Exclusion (questioned print and exemplar are from different sources), Inconclusive (cannot conclude whether the questioned print and exemplar are from the same source or not). Depending on the agency that the examiner belongs to, certain outcomes from the evaluation step warrant review by another examiner in the verification step. Some agencies have a verification step for only individualization decisions, to reduce the potential for false identifications. Other agencies have a verification step irrespective of the outcome of the evaluation step. Verification can be blind in which case the verifier carries out a separate independent examination or unblinded.

3.2.2.2 Latent print FBI Black-Box Study

In 2009, the FBI recruited a total of 169 examiners from federal, state, and private agencies to participate in the first large-scale black-box study for latent fingerprint comparison decisions. Examiners were asked to fill out an anonymous survey about their demographic information, type of training or certification, type of agency (employer), years of experience, etc. The survey responses were only used in the aggregate sense, and not matched with the examiners with any identifying information due to restrictions in the Institutional Review Board (IRB) approval for this study. About 83% of the examiners were certified as latent print examiners by the International Association for Identification (IAI) or other agencies, and the median years of experience was 10 years. More information about this survey can be found in the appendix to Ulery et al. (2011) and on the FBI website (*Black Box Study Results* 2017).

Twenty-one individuals provided a total of 356 latent prints deposited on various surfaces and processed by different techniques as well as clear exemplars. The latent print and exemplars were then combined to form 744 total pairs, 520 mated pairs (latent and exemplar are from the same source) and 224 non-mated pairs (latent and exemplar are from different sources). The quality of latent prints and the difficulty of comparisons were meant to mimic real cases. The ground truth for comparison decisions were known. Each examiner was given an average of 100 latent-exemplar pairs, with roughly the same ratio of mated and non-mated pairs, for making quality and source determinations. Examiners were asked to assess samples like they would in real case work.

The data from the first phase of the study provide inferences about accuracy and reproducibility. To study the repeatability of decisions, 72 of the 169 examiners participated in a second phase of the study about 7 months from the first study (Ulery et al., 2012). Each examiner was assigned 25 pairs out of the samples they observed in the first phase of the study. Out of the 25 pairs, 9 were non-mated, and 16 were mated. The assigned pairs were randomly selected except that there was also an attempt to include pairs that

examiners possibly made false negative errors on in the first study. Ulery et al. (2011, 2012, 2014, 2016) report high accuracy and good reliability for source determinations. They also report less reliability in the analysis phase of the ACE-V process compared to the source determinations.

3.3 Statistical Models for Reliability

Measurement reliability in the physical sciences is often assessed via a two-factor analysis of variance (ANOVA) model: $Y_{ijk} = \mu + \alpha_i + \gamma_j + \epsilon_{ijk}$, where, Y_{ijk} is the measurement or decision for examiner i on sample j in the k^{th} repetition; α_i is the examiner effect or examiner tendency, γ_j is the sample effect or sample complexity, and ϵ_{ijk} is the random noise in the outcomes. These models have been used previously, for example, in the study of inter-rater reliability in engineering, by Vardeman and VanValkenburg (1999), Vardeman (2014), and Weaver et al. (2012). Two-way ANOVA models are easy to fit, intuitive, and used extensively in scientific applications. If there is a reason to believe that there may be an interaction between the two factors, an interaction effect δ_{ij} can be used in addition to the effects for the two main factors (examiner and sample).

3.3.1 Continuous Data

It is most straightforward to introduce the use of the two-way random effects model for outcomes that may be approximated as continuous. In the forensic context, this could be a complexity assessment score or could also be a subjective assessment of the degree of similarity. For example, Alewijnse et al. (2011) analyzed the data from signature complexity assessments on a scale of 0-100. Incorporating an interaction within the model, we get:

$$Y_{ijk} \sim N(\mu + \alpha_i + \gamma_j + \delta_{ij}, \sigma^2_\epsilon). \tag{3.1}$$

We assume that the continuous decisions Y_{ijk}, from examiner i for sample j in repetition k, follow a normal distribution around a mean that depends on a grand mean, examiner-specific and sample-specific random effects, and their interaction, that is μ, α_i, γ_j, and δ_{ij} respectively. Let the total number of examiners be I and the number of samples be J. The normal distribution, though common, is just illustrative. It simplifies computation but other continuous distributions are possible.

More specifically, in the case of handwriting complexity examination, we assume that examiners have an individual tendency to rate samples higher or lower in complexity compared to the overall mean, and we let this tendency for examiner i be denoted by α_i. The parameter, α_i, informs whether a certain examiner is more or less prone to see complexity as compared to other examiners while assessing the same sample. The examiner effect, α_i can be modeled to depend on examiner features such as years of experience or the employer agency; if no such information is collected during the study, α_i can be a proxy for such characteristics. Similarly, the parameter related to the sample informs whether the sample tends to receive higher or lower complexity ratings compared to the mean. The complexity for sample j is denoted by γ_j. There may be additional information about the samples that can be used to model γ_j. Finally, δ_{ij} is an interaction effect between examiners and signature samples. Without interactions there is an additive examiner and sample effect; however interactions change that so there is a differential effect of a sample on examiner effects, i.e., examiner effects change with the sample and vice versa. Inference for δ_{ij} is challenging, especially in a high dimensional case when the level of factors I and J are large. Ideally, we need several repeated decisions for examiner-sample pairs to enable estimation.

In applying the model given by equation (3.1) to reliability studies, we model α_i, γ_j, and δ_{ij} as random effects.

$$\begin{aligned} \alpha_i &\overset{i.i.d.}{\sim} N(0, \sigma_\alpha^2) \\ \gamma_j &\overset{i.i.d.}{\sim} N(0, \sigma_\gamma^2) \\ \delta_{ij} &\overset{i.i.d.}{\sim} N(0, \sigma_\delta^2) \end{aligned} \tag{3.2}$$

This part of the model treats examiners and samples as subsets of a large population of examiners and samples. Therefore, there is an additional level of hierarchy; α_i, γ_j, and δ_{ij} are modeled as normal distributions with means 0 and variance parameters σ_α^2, σ_γ^2, and σ_δ^2 respectively. These variance parameters describe how examiners, samples, and interaction respectively contribute to the observed variation in scores. The variance of decision scores, σ_ϵ^2 is the random noise or the variation that exists in repeated observations of the same sample by the same examiner. The choice of normal distribution here is common for random effects (e.g., examiner tendencies) being modeled as draws from a large population. Schielzeth et al. (2020) provide evidence that conclusions are generally robust to the choice of distributions in (3.2). We address the situations when data deviates from the distributional assumptions with some simulation studies, the results for which can be found in the Supplemental material.

3.3.2 Binary Data

The determination of value in a latent print examination is an example where the forensic decision can be thought of as binary or categorical. Here, we focus on the binary decision scale where samples are assessed for quality (value/ no value). Another example of binary data in forensic examination is match or non-match (excluding Inconclusives). For binary data, percentage agreement is often used to assess reliability. However, percentage agreement or the Cohen's κ (Cohen, 1960) statistic, does not account for the difficulty of samples or variation across examiners. Hence, we are interested in developing a methodology for assessing the reliability for such binary data, while accounting for sample difficulty and

allowing for interactions.

Albert and Chib (1993) proposed a methodology for analyzing binary or categorical response data by modeling it in terms of an underlying latent continuous variable. Their approach can be used to generalize the two-way ANOVA model that was discussed in the previous subsection. An underlying continuous variable, denoted as Z_{ijk}, is modeled using the model (3.1) that includes an examiner-specific effect, a sample-specific effect, and interactions between examiner and sample. The binary decision or outcome is represented by Y_{ijk} where examiner i is making a binary decision for sample j in repetition k. The binary decision is assumed to depend on a latent variable Z_{ijk} with $Y_{ijk} = 1$ when $Z_{ijk} > 0$, and $Y_{ijk} = 0$ otherwise. The model we propose can be found below:

$$Z_{ijk} \sim N(\mu + \alpha_i + \gamma_j + \delta_{ij}, 1)$$
$$Y_{ijk} = \begin{cases} 1, & \text{if } Z_{ijk} \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

Under this model, $P(Y_{ijk} = 1) = \Phi(\mu + \alpha_i + \gamma_j + \delta_{ij})$, where Φ is the standard normal cumulative distribution function. Similar to the model given by equation (3.1), α_i, γ_j, and δ_{ij} are assumed to be drawn from a normal distribution with mean 0 and variances σ^2_α, σ^2_γ, and σ^2_δ respectively.

A model is said to be non-identifiable when two or more parametrizations of the model yield the same likelihood. In other words, if modifying some or all parameters in the model, for example by adding a certain constant to each or multiplying them by a certain factor, yields the same statistical representation of the model, the model is not identifiable. A few different constraints are introduced in the above model to make it identifiable. The cut-point on the latent scale Z_{ijk} that determines the value of the observed binary variable to be either 1 or 0, is assumed to be known and fixed at zero. Note that a non-zero cutpoint would merely

change the meaning of the intercept μ. We also fix the variance for Z_{ijk}, to be equal to 1. This is done because a variance parameter other than one would merely change the scale of the random effects.

3.3.3 Assessing Reliability

One way of assessing reliability for continuous decisions, with the setup in equation (3.1), is to use the intraclass correlation (Shrout and Fleiss, 1979). Reproducibility may be estimated by looking at correlations between outcomes for the same sample across examiners. Repeatability may be estimated by looking at correlations between outcomes for the same sample by the same examiner. It is expected that repeatability is higher than reproducibility because while assessing reproducibility one needs to account for variance across examiners. For continuous data, the intraclass correlations are the following:

$$\text{Reproducibility} = \text{corr}(Y_{ijk}, Y_{i'jk'}) = \frac{\sigma_\gamma^2}{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\delta^2 + \sigma_\epsilon^2} \tag{3.4}$$

$$\text{Repeatability} = \text{corr}(Y_{ijk}, Y_{ijk'}) = \frac{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\delta^2}{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\delta^2 + \sigma_\epsilon^2} \tag{3.5}$$

Similarly, for the binary data model given by equation (3.3), we can assess reliability through intraclass correlation for the underlying latent variable Z_{ijk} by the same equations as (3.4) and (3.5) but with σ_ϵ^2 replaced by 1.

Combining the data sets from reproducibility and repeatability trials are especially beneficial for repeatability assessments when the size of the repeatability data set is much smaller compared to the reproducibility data set. If the data sets are not combined in the analyses, the estimated random effects α_i, γ_j, and δ_{ij} as well as the variance components, using the models given by the equations (3.1) or (3.3), will have a lower precision. Additionally, if there is an examiner-sample interaction, using the reproducibility data set alone cannot

estimate it. Therefore, combining the data sets is beneficial for both reproducibility and repeatability.

3.3.4 Bayesian Inference and Computation

We use a Bayesian approach to fit the models given by equations (3.1) and (3.3), where the model parameters have an associated prior distribution before observing the data (Gelman et al., 2013, Carlin and Louis, 2008). A Bayesian approach to modeling provides a way to incorporate prior information about the parameters, allows us to deal with missing data naturally, and easily provides credible intervals for all parameters including reproducibility and repeatability. Posterior distributions for the model parameters are obtained after incorporating information from the data through the likelihood. Markov chain Monte Carlo (MCMC) algorithms are used to obtain samples from the posterior distribution. Specifically, we use a Gibbs sampling algorithm (Geman and Geman, 1984) where parameter values are drawn conditional on other parameters (fixed) through their conditional posterior distributions. The full conditional distribution for fitting the models (3.1) and (3.3) through Gibbs sampling can be found in Appendix A. The posterior samples are summarized through their posterior medians and credible intervals.

3.4 Simulation Studies

Before applying the models of Section 3.3 to the data from Section 3.2, we present results from simulation studies for the continuous model and binary model. As stated above, it is common that the repeatability study is much smaller compared to the reproducibility study. This presents a challenge for our analysis in that there are limited data for studying interactions. This is a key element that we explore via simulation.

3.4.1 Continuous data

Although we experimented with a range of study sizes, the core design of the simulation study involves $I = 50$ examiners and $J = 80$ samples. As a first step, we simulate one decision per examiner-sample pair, so we have a simulated matrix of 50×80 outcomes generated according to the continuous model given by the equation (3.1). This is a reproducibility data set. The second step incorporates data from a repeatability study. This is described further below.

The value of parameters chosen are $\mu = 3.5$, $\sigma_\alpha^2 = 1$, $\sigma_\gamma^2 = 4$, $\sigma_\delta^2 = 0.5$ and $\sigma_\epsilon^2 = 0.2$. We use these specific values because $\sigma_\gamma^2 > \sigma_\alpha^2 > \sigma_\delta^2 > \sigma_\epsilon^2$ and the value for μ is motivated by the handwritten signature complexity data set. The intercept or grand mean μ is given a relatively uninformative prior distribution, a normal distribution centered around 0 and having a large variance (100). The standard deviation components σ_α, σ_γ, σ_δ and σ_ϵ are given uniform prior distributions as suggested in Gelman (2006). The half-Cauchy prior distribution suggested in Gelman (2006) also provided similar results.

We have used the libraries "rstan" (Stan Development Team, 2022) as well as "rjags" (Plummer et al., 2019) in the R language to analyze the simulated data sets. We have used three MCMC chains for obtaining posterior inference. We incorporate a burn-in period for each chain, samples obtained during this initial period are not used for posterior inference. A burn-in allows for the algorithm to converge to the posterior distribution. We use $20,000$ draws with a burn-in of $10,000$ draws. We evaluate the convergence of the chains for each parameter using the potential scale reduction factor (PSRF) also known as the Gelman-Rubin statistic (Gelman and Rubin, 1992). If the chains have not converged, it indicates that the draws have not found the stationary target distribution and we need to run the chains for longer. Each chain has a different starting point and PSRF informs on whether the different chains have converged to the same distribution.

In practice, the repeatability component of a black-box study is small and usually only a subset of the examinations are repeated. We explore the impact of limited repetitions through a simulation study. In the reproducibility trial, a decision is simulated for each examiner-sample pair (50 x 80 decisions). In the repeatability trial, we investigate four scenarios; a second decision is simulated for each examiner for a total of: i) 80 samples (100% repeated samples), ii) 40 samples (50% repeated samples), iii) 20 samples (25% repeated samples) or iv) 10 samples (12.5% repeated samples). The subset of the samples for which a second decision is obtained varies across examiners. For each of the four settings that are described, 25 simulated data sets are generated. Figure 3.2 provides the posterior medians and 95% credible intervals for parameters μ, σ^2_α, σ^2_γ, σ^2_δ, reproducibility, and repeatability for each of these 25×4 data sets. Table 3.1, provides summaries such as average posterior median and average credible interval limits.

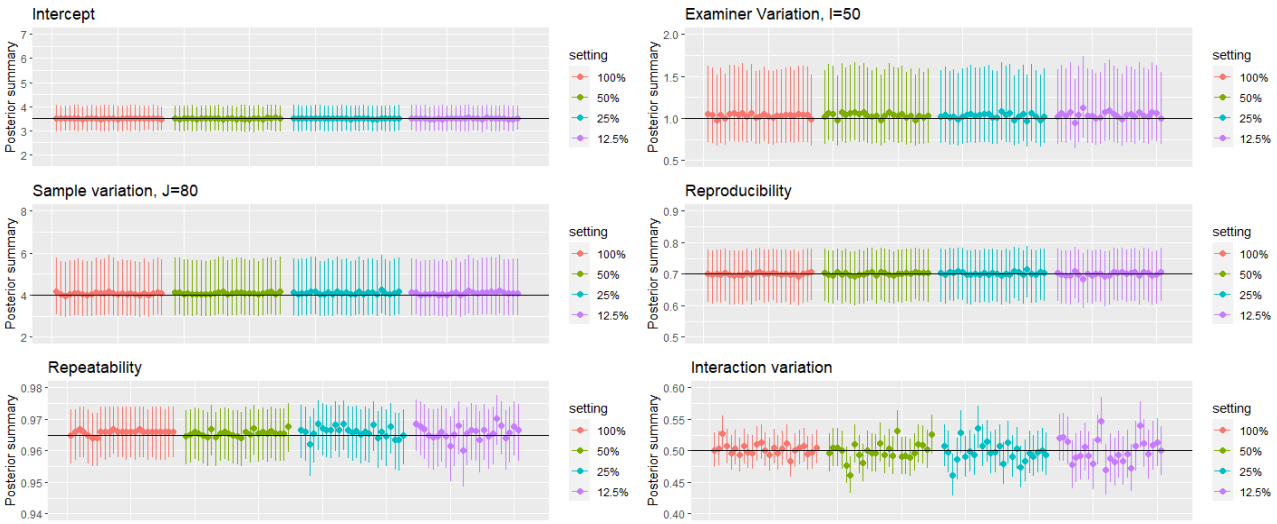


Figure 3.2: Posterior medians with 95% credible intervals for 25 simulated data sets in each case are shown with the black line indicating the true value. The results from different simulated data sets are represented along the x-axis. Here, the setting indicates the percentage of samples that received repeated assessments by the examiner.

Repeat. Setting	$\mu = 3.5$	$\sigma^2_\alpha = 1$	$\sigma^2_\gamma = 4$	$\sigma^2_\delta = 0.5$	$\sigma^2_\epsilon = 0.2$	$R_1 = 0.70$	$R_2 = 0.97$
100%	3.50	1.03	4.07	0.50	0.20	0.70	0.97
	(2.97, 4.03)	(0.71, 1.59)	(3.02, 5.68)	(0.48, 0.53)	(0.19, 0.21)	(0.61, 0.78)	(0.96, 0.97)
50%	3.49	1.04	4.09	0.50	0.20	0.70	0.97
	(2.96, 4.03)	(0.71, 1.60)	(3.03, 5.71)	(0.47, 0.53)	(0.19, 0.22)	(0.61, 0.78)	(0.96, 0.97)
25%	3.49	1.03	4.10	0.50	0.20	0.70	0.97
	(2.96, 4.03)	(0.70, 1.59)	(3.04, 5.72)	(0.47, 0.53)	(0.18, 0.22)	(0.61, 0.78)	(0.96, 0.97)
12.5%	3.50	1.04	4.09	0.50	0.20	0.70	0.97
	(2.97, 4.03)	(0.71, 1.61)	(3.03, 5.71)	(0.46, 0.54)	(0.18, 0.23)	(0.61, 0.78)	(0.96, 0.97)

Table 3.1: Results from 25 simulation data sets with continuous data. Posterior median estimates with the average lower 2.5% quantile and the average upper 97.5% quantile (up to 2 decimal places) are presented. R_1 denotes reproducibility and R_2 denotes repeatability.

In Figure 3.2, we observe little variance in posterior median estimates for examiner variance, sample variance, and reliability components across the simulated data sets, even with few (12.5%) repeated decisions. The credible intervals are also comparable across simulations. This is promising and ensures that when model assumptions hold, good estimates for the variance parameters can be obtained by using the model given by equation (3.1) even with limited data. This is encouraging for designing black-box studies; especially when an interaction between examiners and samples is not expected.

The results presented in Table 3.1 and Figure 3.2 were obtained when the simulated data was generated with the distributions specified in the model given by the equations (3.1) and (3.2). In practice, it is difficult to check model assumptions and the data often deviates from these distributional assumptions. In the supplemental material accompanying this paper, we present the posterior medians and credible intervals for variance parameters, reproducibility, and repeatability when the error distribution is different from Gaussian such as Student's t-distribution, bimodal distribution, or Laplace distribution. We conclude that in several cases our model is robust against violations in model assumptions.

3.4.2 Binary data

For the model given by the equation (3.3), the simulation studies mimic the settings that were described in subsection 3.4.1. We simulate binary decisions according to the model given by the equation (3.3), for $I = 50$ examiners and $J = 80$ samples; the data is simulated with interactions. As an initial step focused on reproducibility, decisions are simulated for each examiner-sample pair so that there are 50×80 total binary decisions. For the repeated decisions (same examiner and same sample), we have four settings with decreasing fractions of repeated decisions just as in subsection 3.4.1. The value of parameters chosen are $\mu = 1$, $\sigma^2_\alpha = 1$, $\sigma^2_\gamma = 4$ and $\sigma^2_\delta = 0.5$. We present the results from these simulations in Figure 3.3 and provide limited summaries in Table 3.2.

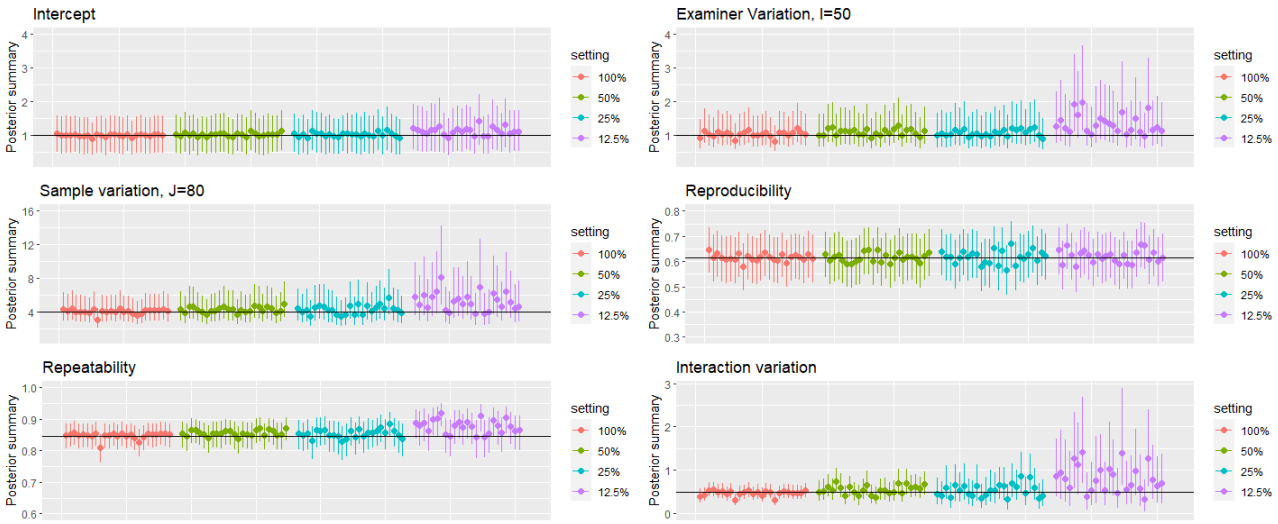


Figure 3.3: Posterior medians with 95% credible intervals for 25 simulated data sets in each case are shown with the black line indicating the true value.

Repeatability Setting	$\mu = 1$	$\sigma^2_\alpha = 1$	$\sigma^2_\gamma = 4$	$\sigma^2_\delta = 0.5$	R_1=0.62	R_2=0.85
100%	1.00	1.02	4.04	0.47	0.62	0.85
	(0.46, 1.54)	(0.68, 1.62)	(2.83, 5.94)	(0.33, 0.64)	(0.52, 0.71)	(0.81, 0.88)
50%	1.02	1.09	4.28	0.56	0.62	0.86
	(0.47, 1.60)	(0.71, 1.77)	(2.93, 6.45)	(0.34, 0.83)	(0.52, 0.71)	(0.81, 0.89)
25%	1.02	1.07	4.29	0.55	0.62	0.85
	(0.47, 1.61)	(0.67, 1.78)	(2.84, 6.71)	(0.26, 0.96)	(0.52, 0.71)	(0.79, 0.90)
12.5%	1.13	1.34	5.27	0.83	0.62	0.88
	(0.52, 1.82)	(0.79, 2.41)	(3.24, 8.99)	(0.33, 1.68)	(0.52, 0.71)	(0.82, 0.92)

Table 3.2: Results from 25 simulated data sets with binary data. Posterior median estimates with the average lower 2.5% quantile and the average upper 97.5% quantile (up to 2 decimal places) are shown above. R_1 denotes reproducibility and R_2 denotes repeatability on the latent scale.

As demonstrated by Figure 3.3 and Table 3.2, with a decrease in the number of samples that have repeated decisions, we obtain wider credible intervals and more bias in posterior median estimates. This difference is most notable for σ^2_δ. The case with 12.5% repetitions exhibits the worst performance. However, it is noteworthy that the inference for reproducibility and repeatability is quite robust even with 12.5% repetitions. While designing black-box studies, we recommend that at least 25% samples should require repeated decisions on them for appropriate inference regarding variance components and reliability estimates.

The challenge in fitting the binary model with limited data from repeated trials is that there are not enough data to draw reliable inferences about the interactions and the interaction variance components. This was confirmed by repeating the simulation scenario for data generated by a model with no interactions.

3.4.3 A note on computation

Fitting the binary outcomes model can be computationally challenging, especially for cases with a small number of repeated trials. Convergence of the MCMC was slower in such cases and it can be a challenge for some starting values. This is especially the case observed when σ_δ^2 is small. In such cases, it may be helpful to re-parametrize the model (Gelman et al., 2013).

3.5 Forensics Data Results

We next present the results of applying the models to the data from the reliability studies described in Section 3.2.

3.5.1 Signatures data set

The handwritten signature data of Section 3.2.1 describes the complexity assessments for 123 signatures by 5 examiners. The prompt for the examiners used a five-point scale. Though this is not necessarily a large enough scale for a continuous model, we follow Stern et al. (2018) and use the continuous data model given by equation (3.1) in Section 3.3.1. We note that complexity or quality measures are often on continuous scales (e.g., Alewijnse et al., 2011) and this example is helpful for seeing how the model performs.

Parameters	μ	σ_α^2	σ_γ^2	σ_δ^2	σ_ϵ^2
Estimates	3.55	0.06	0.80	0.02	0.36
	(3.21, 3.89)	(0.01, 0.65)	(0.61, 1.06)	(0.00, 0.11)	(0.27, 0.42)

Table 3.3: Posterior medians with 95% credible intervals for the combined reproducibility and repeatability handwritten signature complexity data sets. The 5-point complexity scale is approximated to a continuous scale like in Stern et al.(2018).

The estimated value of grand mean μ is 3.55, which means that on average the signature complexity is between 3-4. The examiner variation is much smaller than the signature (sample) variation. The examiners are trained experts and though their scores vary, there is much greater variation across the signers. The interpretation of the interaction component is that it allows examiner effects on signature complexity to depend on the signature. It is noteworthy that the interaction variance component is quite small. The lower bound for the credible interval is close to zero and hence this could be a sign that there is little evidence that examiner effects vary across samples.

Stern et al. (2018) previously analyzed the signature complexity data. However, their methodology is different from ours in the following ways: i) as opposed to the Bayesian setting presented here, they work in a frequentist setting, ii) they do not account for interactions, iii) they analyze the data collected in the two trials separately, i.e., for the inference on repeatability they only use the signatures that have repeated decisions on them (7 signatures, 5 examiners). They provided the following estimates for the reproducibility data: variation in the complexity decisions attributed to the signatures was 0.79, the variation explained by examiners was 0.04, and the residual variation was 0.38. These estimates are close to the posterior medians in our case. This also indicates that the interaction variation is low because if there was a high interaction variance, these two approaches would result in different estimates.

Reliability	Methods or data used in Stern et al. (2018)	Stern et al. (2018)	Our method
Reproducibility	two-way ANOVA 123 123 ×5	0.65 (0.58, 0.72)	0.64 (0.43, 72)
Repeatability	Fisher z-transformation (Snedecor and Cochran, 1989) $7 \times 5 \times 2$	0.67 (0.36, 0.85)	
Repeatability	two-way ANOVA $7 \times 5 \times 2$	0.57 (0.28, 0.85)	0.72 (0.64, 0.82)
Repeatability	two-way ANOVA (inferred from reproducibility study)	0.68 (0.62, 0.74)	

Table 3.4: Posterior medians with 95% credible intervals for reproducibility and repeatability obtained by our method compared to Stern et al.(2018).

In Table 3.4, we have presented comparisons of reliability estimates obtained by our method compared with the results in Stern et al. (2018). The reproducibility estimates are extremely close. The credible interval for reproducibility obtained in our case is considerably wider, this may be due to the fact that Stern et al. (2018) used the delta method (Casella and Berger, 2021) to approximate the confidence interval and we used no approximations. The repeatability estimate obtained by our method is a bit higher and the repeatability credible interval obtained is much smaller compared to the confidence interval from the first two methods in Stern et al. (2018). This is expected since we leverage a lot more data to make inferences about the variance components.

3.5.2 Fingerprint data set

The black-box study of Ulery et al. (2011, 2012) examined the reliability and accuracy of latent print examiners. They find good accuracy and reliability for final decisions but noted lower reliability (more variation in decisions) of the initial assessment of prints. Decisions

regarding the suitability of latent print examination can be viewed as being made on a binary (e.g., VID or not VID) or trinary (e.g., VID, VEO, or NV) scale. We work with the binary scale and apply the binary model of Section 3.3.2 given by equation (3.3). As a first step, we focus on the quality determinations from the analysis phase of the latent print examination using VID and not VID as the possible determinations. Here, α_i would represent the tendency of examiner i to rate latent prints as VID, and γ_j would represent a measure of the suitability of the latent print j relative to the average print. Table 3.5 presents the results from the analysis.

Parameters	μ	σ^2_α	σ^2_γ	σ^2_δ	Reproducibility	Repeatability
Estimates for Analysis phase	0.90 (0.40, 1.42)	1.38 (1.04, 1.87)	18.22 (13.84, 24.46)	0.27 (0.07, 0.53)	0.85 (0.81, 0.88)	0.95 (0.94, 0.96)

Table 3.5: Results from fitting the binary model to the data from the Analysis phase of the latent print examination process. Posterior medians with 95% credible intervals are presented. Reproducibility and repeatability results are provided on the latent scale.

It is important to note that the variance estimates must be interpreted in the context of the model. They refer to the latent scale where the error variance was fixed to one and all variance components are estimated in relation to that. Again, note that the variation among examiners (σ^2_α) is much smaller compared to the variation among latent prints (σ^2_γ). This confirms the intuition that the latent prints have a lot of variation in their quality or their tendencies to be declared of value (γ_j) and there is less variation in examiners' tendencies to declare value/ no value decisions (α_i). Interestingly, there seems to be little evidence of interactions present in the decisions (σ^2_δ is quite small). The estimates for reproducibility and repeatability on the latent scale indicate very good reliability for analysis decisions on the binary scale.

Ulery et al. (2012) used percentage agreement and κ (Fleiss, 1971) for the estimation of

reliability. They evaluated reliability on the quality determinations from 72 out of the 169 examiners who participated in the repeatability study and provided the following percent agreement estimates for reproducibility and repeatability for the analysis phase: 0.85 (0.82, 0.87) (90% confidence interval) and 0.90 (0.87, 0.91) (95% profile likelihood confidence interval) respectively. Note that our reliability estimates are model-based and account for variations in examiner tendencies and latent print complexities. Thus, they are not directly comparable with percentage agreement or κ. To better understand the differences, we fit the binary model on the subset of the data (72 examiners) used by Ulery et al. (2012) and then used posterior draws for μ, σ^2_α, σ^2_γ, and σ^2_δ to generate new decisions and obtain posterior predictive (Gelman et al., 2013) estimates for reproducibility and repeatability. These posterior predictive percent agreement measures yield very similar median estimates to those obtained by Ulery et al. (2012) for reproducibility, with the posterior median estimate of 0.85 and 95% credible interval (0.81, 0.88), through this procedure, and repeatability, with the posterior median 0.90 with 95% credible interval (0.88, 0.92).

Although the estimate of the interaction variance is small, we further investigate the distribution of the interactions to provide insights into what they can tell us about the data. Figure 3.4 presents the distribution of the interaction effects across examiners in four panels, with examiners sorted in quartiles by their estimated effects. The lower quartiles are least likely to judge prints as suitable and the higher quartile are most likely. Figure 3.5 provides a heatmap of the value and sign of the interaction effects.

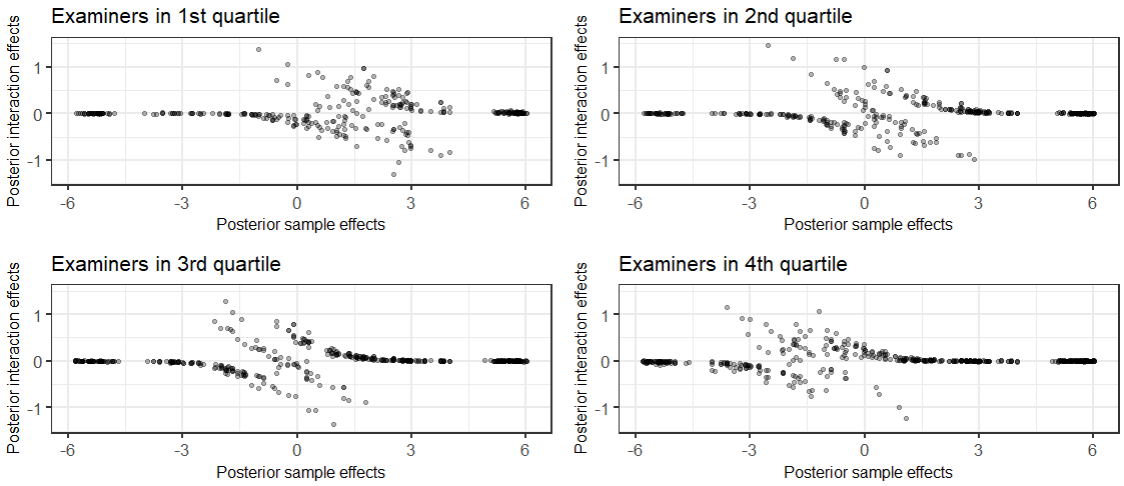


Figure 3.4: Distribution of interaction effects across examiners and samples. Examiner effects were ordered in increasing order of posterior medians ($\alpha_i|Y_{ijk}$), which is the estimate for their tendencies to see value in latent prints and further divided examiners in four quartiles.

Figure 3.4 reveals that examiners that are least likely to find value in latent prints (top left panel) had more non-zero interactions for higher quality prints ($\gamma_j > 0$). Additionally, examiners that are more likely to see value (bottom right panel) have more non-zero interaction effects for low-quality prints ($\gamma_j < 0$). The primarily positive interaction effects for higher quality prints $0 < \gamma_j < 3$ among examiners that are least likely to find value suggests that in these combinations, the quality of the print has a bigger impact on the suitability determination. The pattern in the plot showing examiners in the upper quartile seems to support the notion that their tendency to see value is accentuated for these lower-quality prints.

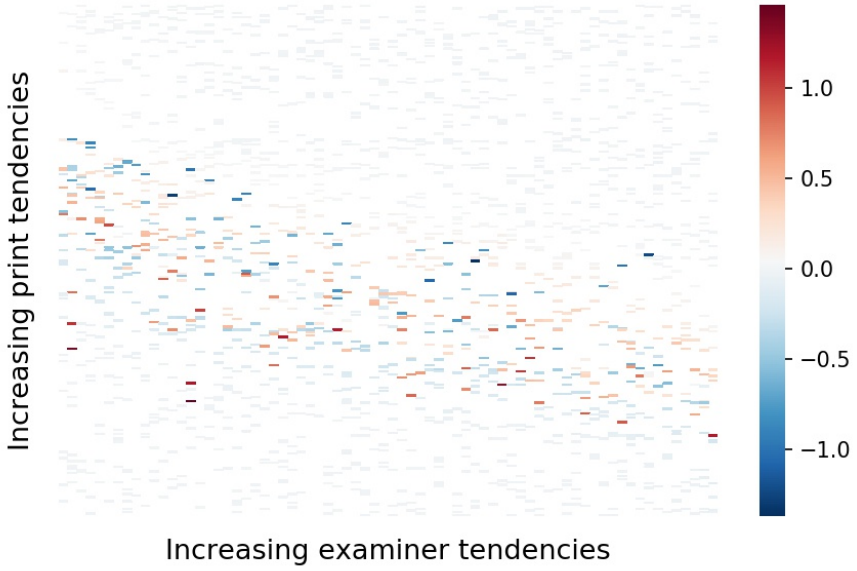


Figure 3.5: A heatmap showing the posterior medians for δ's across examiners and samples. The horizontal axis shows examiner effects from least likely to see value (on the left) to most likely to see value (on the right) in latent prints. The vertical axis represents latent print effects from least likely to receive VID decisions(on the bottom) to most likely to receive VID decisions (on the top) for value decisions. The blank spaces are missing values since the interactions are only plotted for the examiner-sample pairs that have repeated decisions on them.

From the heatmap in Figure 3.5, we observe higher absolute value for posterior interactions ($|\delta_{ij}|$) which are the blue and red values in the plot, for prints that are of mediocre value. These findings are consistent with the results of Hicklin et al. (2020), where they found that examiners had more disagreements amongst each other for value determinations and comparison decisions with mediocre quality prints.

We also analyzed the conclusions in the Evaluation phase of the latent print examinations. We change the trinary decision scale of Individualization, Exclusion, and Inconclusive to a binary scale. Reliability is separately evaluated within mated pairs by treating the Individualization decision as $Y_{ijk} = 1$ and the other conclusions, Exclusion, Inconclusive, and Latent No Value are considered as $Y_{ijk} = 0$. We fit the model and present the results in Table 3.6. In this case, α_i would be the tendency of an examiner to give Individualization decisions

for mated pairs and γ_j would be the tendency of a mated latent-exemplar pair to receive Individualization decisions based on shared characteristics. As in our other results, there is much more variability among pairs than among examiners.

Parameters	μ	σ^2_α	σ^2_γ	σ^2_δ
Estimates for Evaluation phase (Mates)	-2.40 (-3.00, -1.87)	0.89 (0.63, 1.28)	19.02 (13.68, 26.61)	0.38 (0.09, 0.77)

Table 3.6: Results from fitting the binary model to the data from the Evaluation phase of the latent print examination process on known mated pairs. Posterior medians with 95% credible intervals are presented.

The posterior median for reproducibility on the latent scale is 0.89 with 95% credible interval of (0.87, 0.91) and the posterior median for repeatability is 0.95 with 95% credible interval (0.94, 0.97).

3.6 Conclusions

A two-way ANOVA random effects model is widely used to model the reproducibility and repeatability of measurements in engineering, medicine, and other fields. The model is applied here to analyze data from forensic science studies of reproducibility and repeatability with ordinal outcomes approximated as continuous outcomes and binary outcomes such as value/ no-value. It provides a number of benefits in this context: i) The model can combine reproducibility trials (different examiners assess the same set of samples) and repeatability trials (examiners re-assess samples), ii) Variation due to inter-examiner differences and sample differences can be accommodated, iii) When there are sufficient (25%) repeated decisions, the model can allow us to draw inferences about examiner-sample interactions. The model works well in the ideal setting where there are enough repeated comparisons. We observed

that if the percentage of repeated analyses in the study decreases, then the credible interval for variance components widens as would be expected.

The model currently models examiner tendencies (α) and sample characteristics (γ) as random effects. A natural extension would allow these to depend on the measured characteristics of the examiner and sample. Unfortunately, our motivating data does not include covariates measured on examiners and samples. We do not have access to the actual latent prints or signatures or covariates on examiners which can explain their decisions better.

The primary goal of black-box studies is to get estimates of the reliability and validity of forensic sciences. This paper facilitates and furthers that goal by implementing and assessing a model that can be applied to continuous as well as binary data in an incomplete or sparse data setting and enables pooling data from two or more repeated comparisons. Due to this flexibility, this model can be applied to many forensic fields. With the above results, we show that variance in decisions can be explained by the examiners, prints, and their interactions. In the future, we will be extending this methodology to multi-categorical and ordinal decisions without the need for continuous approximations.

Chapter 4

Reliability of Ordinal Outcomes in Forensic Black-Box Studies

4.1 Introduction

Expert decisions on forensic evidence are admissible in a U.S. federal court of law provided that the assessment is a "product of reliable principles and methods" (Federal Rule of Evidence 702) and the testimony is valid and reliable (*Frye v. United States,* 293 F. 1013 (D.C. Cir. 1923); *Daubert v. Merrell Dow Pharmaceuticals, Inc.,* 509 U.S. 579 (1993)). A substantial share of forensic science disciplines require subjective assessments in all or some steps of the evidence analysis process. It has been observed that forensic experts can vary amongst each other while making decisions on the same piece of evidence. For example, in the analysis of latent fingerprints, examiners often differ in the count and type of minutiae that are marked (Ulery et al. 2014; 2015; 2016). Furthermore, there have been numerous cases where erroneous findings in forensic science procedures have resulted in wrongful convictions (Hsu, 2012; Federal Bureau of Investigation, 2015; Bonventre, 2021).

In 2009, the National Academy of Science (NAS) prepared a report (National Research Council, 2009) that emphasized the need for establishing scientific foundations for forensic science disciplines. In 2016, the President's Council of Advisors on Science and Technology (PCAST) prepared a report (President's Council of Advisors on Science and Technology, 2016), assessing the strength of the scientific evidence regarding the reliability and validity of forensic disciplines that rely on feature-based comparisons (pattern matching). The PCAST report recommended conducting black-box studies to deduce the validity and reliability of forensic science analyses. In a black-box study, enrolled forensic experts are asked to make decisions on forensic samples, for which the ground truth is known, just like they would in practice. The steps taken by an examiner to reach a decision are not explicitly defined and hence these studies are called "black-box" studies. In the last decade many black-box studies have been conducted including for latent fingerprint comparison decisions (Ulery et al., 2011, 2012), bloodstain pattern analysis (Hicklin et al., 2021), handwritten signature comparisons (Hicklin et al., 2022a), footwear comparisons (Hicklin et al., 2022b), and firearms examination (Monson et al., 2023b).

Black-box studies provide information about validity and reliability, with the primary focus on validity. Validity or accuracy of decisions, in that, validity relates to the correctness of the decisions and reliability relates to the consistency of decisions. Reliability is a precursor to validity because the decision-making process cannot be correct unless it is consistent, which is why we focus on providing methods to assess reliability. We consider two different components of reliability. Reproducibility, also known as inter-rater reliability, is defined as the consistency of decisions when different examiners provide assessments on the same sample. Repeatability, also known as intra-rater reliability, refers to the consistency of decisions when the same examiner provides assessments of the same sample at two different points in time. Black box studies often have two phases: a reproducibility study, where different examiners provide assessments on a set of samples, with at least a few shared samples between examiners. The reproducibility study is followed by a repeatability study

where examiners give repeated decisions on a subset of the samples they judged in the first study. The repeatability portion of the study is usually much smaller. Arora et al. (2022) provides methods to assess the reliability from such studies for continuous and binary outcomes.

Decisions from forensic examination procedures are often reported as categorical conclusions. These categories often follow a meaningful order. Some examples of ordinal data in forensic decision-making are handwritten signature complexity, quality assessments for latent fingerprints, conclusions for shoe-print comparisons, etc. In latent fingerprint assessments, quality may be judged on a three-category ordinal scale such as VID (Value for Individualization), VEO (Value for Exclusion Only), and NV (No Value). In footwear comparisons, the conclusion scale may be on a seven-point scale based on the degree of match: Exclusion, Indications of Non-Association, Inconclusive, Limited Association of Class Characteristics, Association of Class Characteristics, High Degree of Association, and Identification.

Reliability from black-box studies is usually reported through contingency tables conditional on different categories of examples (Ulery et al., 2012; Hicklin et al., 2022b; Hicklin et al., 2022a). Ulery et al. (2012) also used mean percentage agreement across fingerprint samples. However, these measures do not account for examiner tendencies or sample difficulties. Other measures of inter-rater agreement that are used for categorical data, and that have largely not been used to analyze black-box studies, include the κ-statistic (Cohen, 1960), and its variations such as the weighted-κ (Cohen, 1968), Fleiss' κ (Fleiss, 1971), ordinal alpha (Zumbo et al., 2007), and Krippendorff's α (Krippendorff, 2011). Although κ has been used widely to understand the agreement in various scientific fields, it has drawbacks such as the bias effect and prevalence effect (Feinstein and Cicchetti, 1990; Byrt et al., 1993; Delgado and Tibau, 2019). Nelson and Edwards (2015) provided a method to assess reproducibility for ordinal data but they did not account for interactions.

We briefly describe previous work that developed models for ordinal data as well as reliability

for ordinal data. Albert and Chib (1993) proposed a latent variable approach to model binary and polychotomous data. Johnson (1996) proposed an analysis of ordinal data with applications to automatic essay grading. The methods proposed in this work enable the assessment of inter-rater reliability through the rater variances. Johnson and Albert (2006) also provides methods for ordinal data regression as well as multi-rater ordinal data models. The models described in the aforementioned work also assumed that the ordinal data depends on an underlying continuous variable. Bradlow (1994) and Bradlow and Zaslavsky (1999) extended the method in Albert and Chib (1993) by modeling missing data in ordinal customer survey data. Johnson et al. (2002) proposes an algorithm for fitting a hierarchical model to a multi-rater ranking data with applications to primate intelligence ranking using a similar latent variable approach. Luby et al. (2020) proposes a cumulative logit model for examiner-reported difficulties for latent-exemplar comparison decisions.

We provide a methodology to analyze and assess reliability for the ordinal data collected from black-box studies by leveraging the latent variable approach to ordinal data modeling used by Albert and Chib (1993) and other references cited above. This extends the work of Arora et al. (2022) to ordinal data. Our model offers several contributions: we are able to combine the data collected from reproducibility and repeatability studies and we are able to account for the possibility of examiner-sample interactions. This method also enables an exploratory approach to infer different examiner thresholds for making categorical decisions. Such an exploratory analysis can provide insights into whether examiners differ significantly in their tendencies to rate samples into a certain category, and could be used to assess whether examiners that belong to the same agency or have similar training tend to make decisions similarly.

We begin by introducing three data sets that motivate our work in Section 4.2. We propose a statistical model for modeling ordinal decisions in Section 4.3 and discuss some special cases of the model. We also discuss methods to assess reliability with our model. This dis-

cussion is followed by simulation studies in Section 4.4 to assess performance of the proposed methodology on data that is generated with known parameters in various settings, including different study designs and different mixes of reproducibility and repeatability data sets. We present the results from using these methods for inferences on the motivating data in Section 4.5. Section 4.6 summarizes the advantages of using the proposed methods and discusses limitations and future work.

4.2 Data

We describe three data sets that motivate this work. The first data set is from a handwritten signature complexity study, the second is from a latent fingerprint analysis study, and the third is from a handwriting comparison study.

4.2.1 Signature Complexity Data

Found and Rogers (1996) and Found et al. (1998) developed a statistical method to define the complexity of handwritten signatures in terms of a number of signature features. Using this complexity measure, Sita et al. (2002) concluded that forensic experts are more accurate and confident about signature assessments when the questioned signature is of higher complexity. Additionally, signatures with higher complexity are more difficult to reproduce (Dewhurst et al., 2007). This means that any study of handwriting analysis should try to account for complexity but to do that we need to know whether examiners can reliably judge complexity.

Angel et al. (2017) describes data collected by the Los Angeles Police Department (LAPD) and the Los Angeles County Sheriff's Department (LASD) where 123 participants submitted 5 copies of their signatures. Five forensic document examiners (FDEs) provided complexity assessments on each of the 123 signature samples on two scales, a 3-point scale and a 5-point scale. On the 5-point scale, a signature with a complexity rating 1 reflects the examiner's

belief that the questioned signature is easy to imitate, a signature complexity of 2 reflects the examiner's belief that the questioned signature is fairly easy to imitate, and so on up to a signature with complexity rating 5 which reflects the examiner's belief that the questioned signature is very difficult to imitate. The three-point scale was similar but only allowed assessments of easy, medium, and difficult to imitate. Repeated decisions by all five FDEs were collected for a small subset of seven signature samples.

Stern et al. (2018) assessed the data from this study by treating the 5-point scale data as continuous data. They reported that the reproducibility and repeatability, calculated through intraclass correlations (Shrout and Fleiss, 1979) were 0.65 and 0.67 respectively. It is more appropriate to treat the 3-point scale and 5-point scale data as ordered categories. We re-analyze the data treating the responses as ordinal variables.

4.2.2 Latent Fingerprint Comparisons Reliability

The FBI conducted the first large-scale black box study of latent print analysis (Ulery et al., 2011, 2012). The latent print examination process involves ordinal decisions in two different steps. We briefly review the process here.

The multi-step process of friction ridge examination is known as ACE-V, which is an acronym for Analysis, Comparison, Evaluation, and Verification (Ulery et al., 2011). In the Analysis step, an expert provides a quality determination for the questioned latent print. In some forensic labs, a 3-point ordinal scale is used with possible outcomes that the print has enough distinguishing features to make an individualization or identification decision, known as Value for Individualization (VID); has enough features only to support an exclusion, known as Value for Exclusion Only (VEO); or does not have enough information to be useful, known as No Value (NV). Although some agencies combine VID and VEO to form a single category, the FBI study asked examiners to use these three categories for the Analysis Phase. If the latent print is found to have value, then in the Comparison phase an exemplar (print collected

under ideal circumstances) is presented to the examiner for comparison with the latent print. In the Evaluation phase, examiners provide the results of their comparison on a 3-point scale: Exclusion (questioned print and exemplar are believed to have come from different sources), Inconclusive (the examiner cannot make an Individualization or Exclusion), and Individualization (questioned print and exemplar are believed to have come from the same source). We would like to assess reliability of these ordered categorical decisions in the Analysis and Evaluation phases.

For the reproducibility part of the black box study, 169 examiners were recruited from government and private laboratories. A total of 744 distinct latent-exemplar pairs (520 mated pairs, 224 non-mated pairs, 356 distinct latent prints) were used in the study. Each examiner made quality and source determinations on a 100 latent-exemplar pairs, which were selected to have a balance of mated and non-mated pairs and a similar level of difficulty across examiners. Approximately seven months after the first study, 72 of the 169 examiners participated in a repeatability study and each examiner was presented with a subset of 25 latent-exemplar pairs that they made assessments on in the first study. It was reported in Ulery et al. (2011, 2012), that examiners were very accurate (0.1% false positives and 7.5% false negatives) and their comparison decisions had good reliability. However, it has been established that a lot of inter-examiner variation is observed in the Analysis phase (Ulery et al., 2011, 2012, 2014, 2015, 2016).

4.2.3 Handwriting Comparisons Data

Previously, the reliability and accuracy of handwritten signatures was studied in Kam et al.1994, Kam et al. (1997), Kam et al. (2001), Kam and Lin (2003), Durina and Caligiuri (2009), Mitchell (2016). Recently, a large-scale black-box study was conducted to assess the subjective handwritten comparisons discipline (Hicklin et al., 2022a). Handwriting comparisons follow the ACE-V procedure similar to the one described for latent fingerprint

comparisons. Therefore, this black-box study was conducted similarly to the one described in Ulery et al.(2011).

Here, 86 forensic document examiners that worked for federal, state, and local agencies participated in the study. Each examiner was assigned about 100 questioned and known (QK) sets from among 180 possible pairs over the course of 10 months. Ninety of the 100 QK sets were distinct and 10 of the QK sets were repeats of sets previously examined by the individual. Examiners were given a five-point scale to assess comparisons: "Written", when the examiner believes that questioned and known come from the same source; "ProbWritten", when the examiner believes that questioned and known probably come from the same source; "NoConc", when the examiner is not confident about whether the sources for the questioned and known are same or different; "ProbNot", when the examiner believes that questioned and known probably come from different sources; "NotWritten", when the examiner believes that questioned and known come from different sources.

4.3 Methods

In this section, we develop a probability model for ordinal outcomes that can accommodate data from intra-individual (repeatability) and inter-individual (reproducibility) reliability studies.

4.3.1 Category Unconstrained Thresholds (CUT) Model

Consider subjective outcomes Y_{ijk}, on an ordinal scale with M levels, with i denoting examiner, j denoting sample or example (typically a questioned sample and a known sample), and k denoting the repetition (if any). In a standard reproducibility study $k = 1$ for all made assessments in that examiner i sees sample j and draws conclusion Y_{ij1}. If the study includes repeatability trials then the same examiner/ sample assessment may be observed more than once. For the data in Section 4.2, we only see $k = 1$ or $k = 2$. For some simula-

tion scenarios, we use $k > 2$. We will assume that Y_{ijk} depends on an underlying continuous random variable Z_{ijk} as proposed in Albert and Chib (1993). The underlying latent score is modeled as a continuous random variable with Gaussian distribution that depends on the sample j through γ_j and also allows for the possibility of an interaction δ_{ij}. This appears as the first equation of the model (4.1). The variance of Z_{ijk} is fixed at 1, because this is a latent scale the data does not identify the scale. In the handwriting example, the parameter γ_j can be interpreted as a measure of the intrinsic complexity of the sample. The presence of a non-zero interaction would indicate that inter-individual differences in the outcomes for a given sample can be expected to vary from sample to sample. Although, this is not a desirable feature for a forensic examination process, Hicklin et al. (2020) have indicated evidence for the possible presence of interactions in the latent fingerprint examination process. They found that examiners had more disagreements for latent prints that were of mediocre quality. There is also likely to be variation among examiners. This is modeled through variation in the thresholds, $\tau_{i,m}$, that map the underlying continuous scores into the categorical outcomes Y_{ijk} as shown in the second equation in the model (4.1). The sample effects γ_j and the interaction effects δ_{ij} are modeled as Gaussian random effects. Henceforth, we will refer to this model as the Category Unconstrained Thresholds (CUT) model and it is presented below:

$$
\begin{aligned}
& Z_{ijk} \,|\, \gamma_j, \delta_{ij} \sim N(\gamma_j + \delta_{ij}, 1) \\
& P(Y_{ijk} = m) = P(\tau_{i,m} < Z_{ijk} \leq \tau_{i,m+1}); \quad m = 1, 2, \ldots, M \\
& \gamma_j,\ j = 1, 2, \ldots, J \,|\, \sigma_\gamma^2 \overset{i.i.d.}{\sim} N(0, \sigma_\gamma^2) \\
& \delta_{ij},\ i = 1, 2, \ldots, I;\ j = 1, 2, \ldots, J \,|\, \sigma_\delta^2 \overset{i.i.d.}{\sim} N(0, \sigma_\delta^2) \\
& -\infty \equiv \tau_{i,1} < \tau_{i,2} \leq \ldots \leq \tau_{i,M} < \tau_{i,M+1} \equiv \infty
\end{aligned}
\tag{4.1}
$$

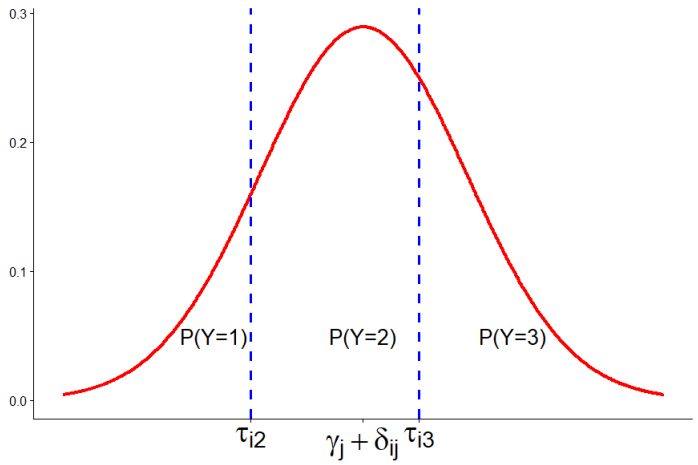


Figure 4.1: A visual presentation for how the cutpoints affect the decision category through Z_{ijk} for M=3.

Figure 4.1 provides a graphical representation of the CUT model (4.1) for $M = 3$. Each examiner is assumed to observe a latent score Z_{ijk} depending on sample j, with an interaction term δ_{ij} causing the curve for individuals to vary around the sample mean γ_j. Additionally, the cutpoints $\tau_{i,2}$ and $\tau_{i,3}$ determine the category for the sample. The cutpoints for an examiner do not change across samples, however, the interaction term affects how the examiners view each sample. Note that to fit this model, we need at least one decision in each category for every examiner. The forensics community has been interested in accounting for individual differences in thresholds between examiners. This model may be used as an exploratory means to account for sample difficulty while quantifying differences among examiners.

If covariates related to the examiners and samples are available, for example, examiner experience or a quantitative measure of sample complexity, then they can be incorporated in the CUT model directly. This could be accommodated by incorporating sample covariates in the model for Z_{ijk} or by incorporating examiner covariates in a model for thresholds $\tau_{i,m}$. However, this is not the case for the motivating examples described in Section 4.2.

4.3.2 Constrained model

If there are sufficient data, several samples per examiner for each category, along with several repeated decisions per examiner-sample pair, separate thresholds may be estimated for each examiner in the CUT model (4.1) along with interactions between examiner and samples. However, with limited repetitions the CUT model is too complex and is likely to overfit the data which may negatively impact inferences for interactions as well as examiner thresholds $\tau_{i,m}$. We discuss this issue in more detail in Section 4.4. If it is known that there are no interactions between examiners and samples or it is not interesting to estimate interactions, then the CUT model may be fit without interactions. However, as we are interested in interactions, we will explore more parsimonious models that have fewer parameters that can be estimated with limited data.

One possible parsimonious model is obtained by introducing some parameter sharing between examiners. For example, we can assume that the cutpoints for all examiners are spaced equally. The constrained CUT model (4.2) below uses the same structure as model (4.1) but replaces the I vectors ($\tau_{i,2}$, $\tau_{i,3}$, ..., $\tau_{i,M}$) with I parameters $\tau_{i,2}$ (one for each examiner) and M-2 cutpoint distances (τ^*_2, τ^*_3, ..., τ^*_{M-1}). The inter-cutpoint distances are assumed to be the same for each examiner.

$$
\begin{aligned}
& Z_{ijk} \,|\, \gamma_j, \delta_{ij} \sim N(\gamma_j + \delta_{ij}, 1) \\
& P(Y_{ijk} = m) = P(\tau_{i,m} < Z_{ijk} \leq \tau_{i,m+1}); \quad m = 1, 2, \ldots, M \\
& \gamma_j,\ j = 1, 2, \ldots, J \,|\, \sigma^2_\gamma \overset{i.i.d.}{\sim} N(0, \sigma^2_\gamma) \\
& \delta_{ij},\ i = 1, 2, \ldots, I;\ j = 1, 2, \ldots, J \,|\, \sigma^2_\delta \overset{i.i.d.}{\sim} N(0, \sigma^2_\delta) \\
& \tau_{i,m} = \tau_{i,m-1} + \tau^*_{m-1},\ \forall\, i,\ 2 < m \leq M
\end{aligned}
\tag{4.2}
$$

This model may be thought of as assuming a fixed distributions of perceptions (centered around γ_j when marginalized over δ_{ij}) with individual thresholds that have a structure over

them. $\tau_{i,2}$ are allowed to vary across examiners.

The constrained CUT model (4.2) can be rewritten as a two-way random effects analysis of variance model similar to the one used by Arora et al. (2022) for assessing reliability for continuous and binary subjective decisions. This setup can also inform about the variation in outcomes attributed to the examiners, the samples and a possible interaction between examiners and samples. Henceforth, we will address this model as the Shared Examiner Thresholds (SET) model to differentiate it from the CUT model that has different category thresholds for each examiner.

$$
\begin{aligned}
& Z_{ijk} \,|\, \alpha_i, \gamma_j, \delta_{ij} \sim N(\alpha_i + \gamma_j + \delta_{ij}, 1) \\
& P(Y_{ijk} = m) = P(\kappa_m < Z_{ijk} \leq \kappa_{m+1}); \quad m = 1, 2, \ldots, M \\
& \alpha_i,\ i = 1, 2, \ldots, I \,|\, \sigma_\alpha^2 \overset{i.i.d.}{\sim} N(0, \sigma_\alpha^2) \\
& \gamma_j,\ j = 1, 2, \ldots, J \,|\, \sigma_\gamma^2 \overset{i.i.d.}{\sim} N(0, \sigma_\gamma^2) \\
& \delta_{ij},\ i = 1, 2, \ldots, I;\ j = 1, 2, \ldots, J \,|\, \sigma_\delta^2 \overset{i.i.d.}{\sim} N(0, \sigma_\delta^2) \\
& -\infty \equiv \kappa_1 < \kappa_2 \leq \ldots \leq \kappa_M < \kappa_{M+1} \equiv \infty
\end{aligned}
\tag{4.3}
$$

The cutpoints κ_m in the SET model (4.3) are shared between all examiners. This model can be thought of as fixed thresholds shared across examiners with individual shifts/ biases (the α_i parameters). The parameterization (4.2) of the constrained version of the CUT model is equivalent to the parameterization of the SET model (4.3) stated above, if $\tau_{i,2}$ have a normal prior. For $M = 3$, for example, $\tau_{i,2} = \kappa_2 - \alpha_i$, $\tau_{i,3} = \kappa_3 - \alpha_i$, and $\tau_2^* = \kappa_3 - \kappa_2$. We demonstrate this through comparing conditional probabilities $P(Y_{ijk} = m \,|\, \text{model parameters})$ in Appendix B.

4.3.3 Bayesian Computation

We use a Bayesian paradigm to fit the previous models. For the CUT model (4.1), the priors for the variance components σ_γ^2 and σ_δ^2 we used are proportional to the inverse of the standard deviations σ_γ and σ_δ respectively (Gelman, 2006). These are improper prior distributions that yield proper posterior distributions as long as $J > 2$. Examiner thresholds $(\tau_{i,m})$ have a uniform prior subject to ordering constraints:

$$\begin{aligned} p(\tau_{i,2}, \tau_{i,3}, \ldots, \tau_{i,M}) &\propto 1_{\tau_{i,2} \leq \tau_{i,3} \leq \ldots \leq \tau_{i,M}} \\ p(\sigma_\gamma^2) &\propto \frac{1}{\sigma_\gamma} \\ p(\sigma_\delta^2) &\propto \frac{1}{\sigma_\delta} \end{aligned} \tag{4.4}$$

Alternatively, examiner thresholds $(\tau_{i,m})$ can also be given a normal prior though they must still obey the ordering constraints. A Gibbs sampling (Geman and Geman, 1984) technique may be used for obtaining draws from the posterior distribution of the parameters; the full conditionals for parameters are derived in Appendix B for the case with $M = 3$.

For the constrained version of the CUT model (4.2), we assume a normal prior over $\tau_{i,2}$:

$$\begin{aligned} \tau_{i,2} \,|\, \mu_{\tau_2}, \sigma_{\tau_2}^2 &\overset{i.i.d.}{\sim} N(\mu_{\tau_2}, \sigma_{\tau_2}^2) \\ p(\tau^*) &\propto 1_{\tau^* > 0} \\ p(\mu_{\tau_2}) &\propto 1 \\ p(\sigma_{\tau_2}^2) &\propto \frac{1}{\sigma_{\tau_2}} \end{aligned}$$

The priors for σ_γ^2, σ_δ^2 are the same as in the equations (4.4). The SET model (4.3), which is a different parameterization to the constrained version of the CUT model (4.2), uses the following priors for κ_m and σ_α^2:

$$p(\kappa_2, \kappa_3, \ldots, \kappa_M) \propto 1_{\kappa_2 \leq \kappa_3 \leq \ldots \leq \kappa_M}$$
$$p(\sigma_\alpha^2) \propto \frac{1}{\sigma_\alpha}$$

The full conditionals for a Gibbs sampling algorithm are provided in Appendix B.

4.3.4 Assessing Reliability

In a black-box reproducibility study, examiners are assigned roughly similar number of samples balanced in difficulty. Although the set of samples vary across examiners, every sample receives multiple assessments by different examiners and some samples receive repeated assessments in the repeatability study. Reliability in black-box studies have been widely reported through summaries of raw data, for example with contingency tables (Ulery et al., 2012; Hicklin et al., 2021; Hicklin et al., 2022b; Hicklin et al., 2022a) that enable inferences of the following nature: "*Reproducibility of VID Individualization decisions was 78.5% on mated pairs*" (Ulery et al., 2012). These contingency tables merely provide summaries of the comparisons conditional on ground truth but they fail to account for the "difficulty" of the samples or the individual tendencies of the examiners. Additionally, they may be difficult to report to the jury due to the absence of a single overall reproducibility or repeatability score.

Additionally, Ulery et al. (2012) uses percentage agreement and Cohen's κ to report reliability. This example was not followed by other black-box studies (Hicklin et al., 2021; Hicklin et al., 2022b; Hicklin et al., 2022a; Monson et al., 2023b). Percentage agreement is a reliability measure defined on categorical data. Ulery et al. (2012) defined it as follows:

$$\begin{aligned} \bar{p_j} &= \frac{1}{n_{j\cdot}(n_{j\cdot} - 1)} \sum_{m=1}^{M} n_{jm}(n_{jm} - 1) \\ \bar{p} &= \frac{1}{J} \sum_{j=1}^{J} \bar{p_j}. \end{aligned} \tag{4.5}$$

Here, $\bar{p_j}$ is the percentage agreement on sample j defined through n_{jm} and $n_{j\cdot}$ which are decisions on sample j made in category m and the total decisions made on sample j respectively. $\bar{p}$ is the average percentage agreement across samples j. Percentage agreement, again, does not account for sample difficulties as well as examiner tendencies or interactions. Additionally, it ignores the possibility of chance agreements (McHugh, 2012). Cohen's κ (Cohen, 1960), defined below, provides a way to account for chance agreement as follows:

$$\kappa = \frac{\bar{p} - p_e}{1 - p_e},$$

with p as the observed percentage agreement between raters and p_e is the agreement expected by chance. Cohen's κ suffers from prevalence and bias effects (Feinstein and Cicchetti, 1990; Byrt et al., 1993; Delgado and Tibau, 2019) which arise from uneven distribution of categories in the population and individual examiner tendencies respectively. Some practitioners recommend using the weighted-κ in practice to overcome limitations of the κ-statistic (Cohen, 1968; Jakobsson and Westergren, 2005).

The latent variable models offer an alternative approach to measuring reliability by using inter-rater reliability tools for continuous data on the latent scale. Bradlow et al. (1999) also used this method to evaluate inter-rater reliability. Using the parameterization in the SET model (4.3) which is the same as the constrained version of the CUT model (4.2):

$$\begin{aligned} \text{Latent Reproducibility}\,(R_1) &= \text{corr}(Z_{ijk}, Z_{i'jk'}) = \frac{\sigma_\gamma^2}{1 + \sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\delta^2} \\ \text{Latent Repeatability}\,(R_2) &= \text{corr}(Z_{ijk}, Z_{ijk'}) = \frac{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\delta^2}{1 + \sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\delta^2} \end{aligned} \tag{4.6}$$

The reliability given by (4.6) are intra-class correlation coefficients for the latent variables; which indicate the degree of similarity of values in the same group. Values closer to 0 indicate very little reliability and values closer to 1 indicate high reliability. However, these have the

disadvantage of not addressing the observed outcomes Y_{ijk}. We further evaluate a model-based reproducibility by extending the statistic provided in Nelson and Edwards (2015) to incorporate interactions, and also derive a model-based repeatability as follows:

$$\begin{aligned}
\text{Reproducibility} &= \sum_{m=1}^{M} \int_{-\infty}^{\infty} \left[\Phi\left(\frac{\kappa^*_{m+1} - x\sqrt{R_1}}{\sqrt{1-R_1}} \right) - \Phi\left(\frac{\kappa^*_{m} - x\sqrt{R_1}}{\sqrt{1-R_1}} \right) \right]^2 \phi(x)\, dx \\
\text{Repeatability} &= \sum_{m=1}^{M} \int_{-\infty}^{\infty} \left[\Phi\left(\frac{\kappa^*_{m+1} - x\sqrt{R_2}}{\sqrt{1-R_2}} \right) - \Phi\left(\frac{\kappa^*_{m} - x\sqrt{R_2}}{\sqrt{1-R_2}} \right) \right]^2 \phi(x)\, dx
\end{aligned} \tag{4.7}$$

In the model-based reliability expressions (4.7), Φ is the standard normal cumulative density function and ϕ is the standard normal probability density function, κ^*_m are given by $\frac{\kappa_m}{1+\sigma^2_\alpha+\sigma^2_\gamma+\sigma^2_\delta}$ in the SET model (4.3). R_1 and R_2 are the reliabilities on the latent scale given by expressions(4.6). The estimates in (4.7) have been derived in Appendix B. This measure of reliability has several advantages over percentage agreement as well as Cohen's κ as it does not suffer from prevalence and bias effects, accounts for examiner-sample interactions, and generalizes across the population of examiners and samples that have not been observed. However, it does not account for chance agreement.

An additional approach to assess reliability is to use the posterior predictive distributions (Gelman et al., 2013) of some of the more traditional discrete data reliability metrics. Data sets are generated given posterior draws from the model fit. For example, for the SET model (4.3) for M=3, data sets for outcomes can be generated through κ_2, κ_3, σ^2_α, σ^2_γ, σ^2_δ and we can evaluate percent agreement or the κ statistic on generated data sets. We are also able to get a posterior distribution for percentage agreement or κ values through the generated data sets and thus better understand the uncertainty. The posterior predictive approach provides a way of summarizing model-based reliability and can also be compared to the measures obtained for the observed data (although the latter do not account for sample variation).

4.4 Simulation Studies

The novel aspects of this approach to ordinal data for forensic studies include the ability to incorporate reproducibility and repeatability study data and the inclusion of interactions. A significant challenge is that repeatability studies are typically small relative to the reproducibility studies. This limits information about interactions. We now present the results from simulation studies where data is generated using our models to: i) check if the posterior draws obtained by using a Markov chain Monte Carlo (MCMC, Gelman et al., 2013) algorithm provide accurate estimates for model parameters and reliability components; and ii) check if accurate estimates are obtained in the presence of potential interaction terms when there are limited repeated decisions. The results provide advice for the design of future forensic studies. The simulation setup is similar to that of Arora et al. (2022); we have created data sets under a variety of scenarios for the CUT model in (4.1) and the constrained version of the CUT model/ SET model in (4.2, 4.3).

We start by generating data sets with the CUT model (4.1). We generate 5 random data sets for ordinal data with $M = 3$ outcome categories, for I=30 examiners and J=50 samples under two scenarios. In the first scenario, each examiner provides 5 decisions for each of the $J = 50$ samples. In the second scenario, each examiner provides two decisions for each of the $J = 50$ samples. It is important to note that typical forensic black-box studies have two decisions for a subset of the examiner-sample pairs (those re-assessed in the repeatability study) and one decision for the rest of the pairs in the reproducibility study. The large number of decisions in these initial simulations are to confirm that the CUT model (4.1) has the potential to address interactions given enough data. The values for $\sigma^2_\gamma = 10$ and $\sigma^2_\delta = 0.5$ were chosen based on values obtained in the analysis of the latent fingerprint examination data. The cutpoints $\tau_{i,2}$ were generated from a uniform distribution $(-3, 1)$ and $\tau_{i,3}$ were generated from a uniform distribution $(-1, 3)$ for each data set, subject to the constraint that $\tau_{i,2} < \tau_{i,3}$.

These data sets were then fit using Gibbs sampling as described in Section 4.3.3 and Appendix B. We ran 4 chains for 100,000 iterations and used every hundredth sample for inference. Convergence was assessed using the potential scale reduction factor (PSRF, Gelman and Rubin, 1992). Figure 4.2 presents the distribution of the differences between estimated examiner thresholds (posterior medians) and true examiner thresholds.

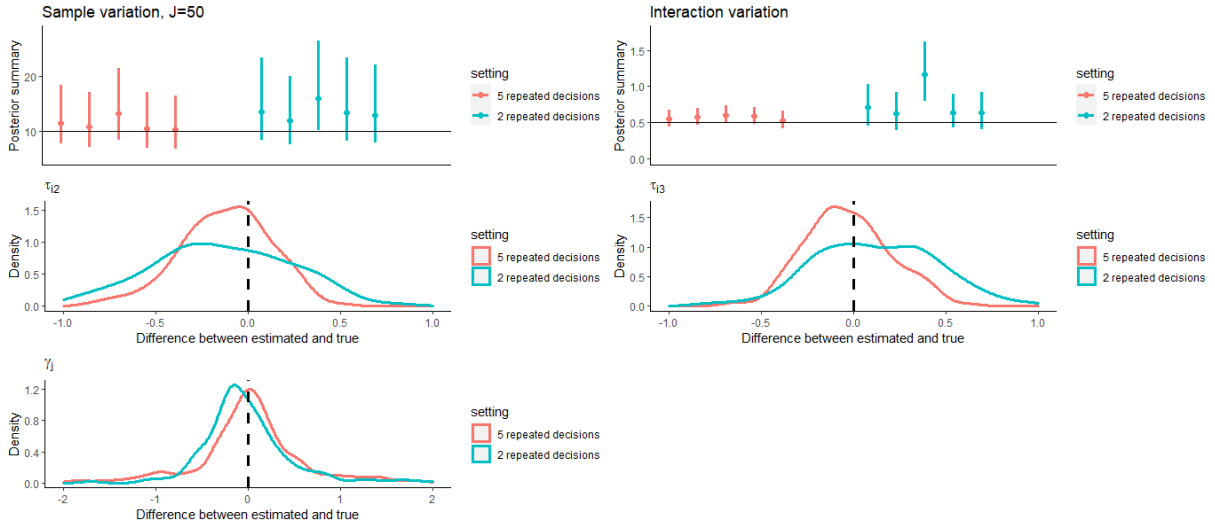


Figure 4.2: Results from fitting the CUT model (4.1) to five simulated random data sets with 5 decisions per examiner-sample pair (in red) and with 2 decisions per examiner-sample pair (in blue). Posterior medians with 95% credible intervals for σ^2_γ, the sample variation, and σ^2_δ, the interaction variation, are presented in the first row. The horizontal black line indicates the true value. The next three plots are density plots for the differences between the true value and posterior medians for $\tau_{i,2}$, $\tau_{i,3}$, and γ_j for all five data sets pooled together.

The results in Figure 4.2 demonstrate that for the CUT model (4.1), inference for σ^2_γ, σ^2_δ, as well as for the examiner thresholds $\tau_{i,2}$, $\tau_{i,3}$ is done well in the scenario with 5 repeated decisions per examiner-sample pair. However, in the second scenario, with only two repeated decisions, the inferences for all parameters are less accurate and have more variance compared to the previous setting. Clearly, it is difficult to estimate σ^2_δ with fewer repetitions when also estimating all examiner thresholds. Assuming it is not possible to obtain many repetitions, there are two strategies that may be useful for analyzing forensic studies with ordinal outcomes. If we are confident that there are no interactions, then we may use

$\delta_{ij} = \sigma_\delta^2 = 0$ in the CUT model (4.1). Otherwise, it can be useful to introduce constraints as in the SET model (4.2, 4.3).

To demonstrate the second approach, we simulated 25 random data sets for ordinal data with $M = 3$ categories with $I = 50$ total examiners and $J = 80$ samples under the model assumptions in the SET model (4.3) for four different scenarios regarding the repeatability study. For each of the four scenarios, decisions are generated for each examiner-sample pair yielding a total of $I \times J$ decisions, these correspond to the first examinations that would occur as part of the reproducibility part of a forensic study. For the repeatability trial, however, we present four cases: i) decisions re-generated for each examiner-sample pair (100% repetitions), ii) decisions re-generated for half of the samples an examiner encountered in the first trial; the subset of samples for different examiners is different (50% repetitions), iii) decisions re-generated for a quarter of the samples an examiner encountered in the first trial; the subset of samples for different examiners is different (25% repetitions), iv) decisions re-generated for an eighth of the samples an examiner encountered in the first trial; the subset of samples for different examiners is different (12.5% repetitions). Stan is a probabilistic programming language that uses No U-Turn (NUTS) or Hamiltonian Monte Carlo (HMC) sampling for a fully Bayesian inference. RSTAN (Stan Development Team, 2022) is an R interface for Stan and we have used it to obtain results for the simulation studies below.

The following parameter values were chosen for the simulation studies: $\kappa_2 = -2$, $\kappa_3 = 2$, $\sigma_\alpha^2 = 2$, $\sigma_\gamma^2 = 10$, and $\sigma_\delta^2 = 0.5$. The random effects α_i, γ_j, and δ_{ij} are generated separately for each data set. The generated data sets are fit using the SET model (4.3). Figure 4.3 provides a visual summary for the variation in the posterior distributions across different simulations by showing posterior medians and credible intervals compared against the true value for all parameters.

Figure 4.3: Results from fitting the SET model (4.3) on 25 simulated data sets for each of four settings are presented. The posterior median and 95% credible intervals for each parameter are shown. The black line indicates the true value for the parameter. The first two plots are κ_2 and κ_3 respectively. Latent reproducibility (R_1) and latent repeatability (R_2) specified in expressions (4.6) are also shown.

We also present, in Table 4.1, numerical summaries aggregated across the 25 repeated datasets for each scenario from Figure 4.3.

Setting	$\kappa_2 = -2$	$\kappa_3 = 2$	$\sigma^2_\gamma = 10$	$\sigma^2_\alpha = 2$	$\sigma^2_\delta = 0.5$	$R_1 = 0.74$	$R_2 = 0.93$
100%	-2.02	1.99	10.05	2.04	0.52	0.74	0.93
	(-2.86, -1.20)	(1.17, 2.82)	(7.42, 14.87)	(1.41, 3.29)	(0.38, 0.67)	(0.65, 0.81)	(0.91, 0.95)
50%	-2.00	1.98	10.09	2.03	0.48	0.74	0.93
	(-2.83, -1.18)	(1.17, 2.81)	(7.12, 14.75)	(1.34, 3.21)	(0.30, 0.70)	(0.65, 0.81)	(0.90, 0.95)
25%	-2.02	2.01	10.54	2.13	0.54	0.74	0.93
	(-2.90, -1.19)	(1.18, 2.89)	(7.21, 15.85)	(1.38, 3.45)	(0.28, 0.88)	(0.65, 0.81)	(0.90, 0.95)
12.5%	-2.05	2.07	11.32	2.23	0.58	0.75	0.93
	(-3.04, -1.18)	(1.17, 3.05)	(7.35, 18.17)	(1.38, 3.82)	(0.21, 1.18)	(0.66, 0.82)	(0.90, 0.96)

Table 4.1: Results from Figure 4.3 are summarized for investigating overall behavior. The estimate is the mean of the posterior medians across the 25 data sets and the average credible interval is obtained by finding average lower 2.5% quantile and average upper 97.5% quantile.

The results in Figure 4.3 and Table 4.1 reinforce the intuition informed by the results in Arora et al. (2022) that the credible intervals for variance components, especially σ^2_δ, are much wider as the number of repeated decisions decrease. Also, the posterior median average for σ^2_γ is farther away from the true value as the number of repeated decisions decrease. However, we observe that even with repeated decisions on 25% of the samples observed in the reproducbility trial, all parameter values are estimated well with the model when the data generating mechanism is correctly specified. The case with 12.5% demonstrates poor performance especially for σ^2_δ. These results indicate that repeated decisions should be collected for more than 12.5% (possibly 25%) of the samples to draw reliable inferences for examiner-sample interactions. This observation can inform the design of black-box studies.

In practice, the data may violate model assumptions. We check the robustness of our model by generating data under a variety of scenarios and then fitting these data using our SET model (4.3). We observe that in most cases when the model generating assumptions are misspecified, the model is still able to estimate variance parameters reasonably well. This exercise assures us that our model is reasonably robust to model misspecifications. The results are presented in detail in Appendix B.

4.5 Forensics Data Results

Motivated by the results from the simulation studies, we now use the models of Section 4.3 for the data from the handwritten signature complexity data set, the latent fingerprint data set, and the handwriting comparisons data set described in Section 4.2.

4.5.1 Signature Complexity Data

The signature complexity data has complexity assessments for signatures from 123 signers evaluated by 5 examiners. Signatures were assessed using a 3-point scale and then again using a 5-point scale. Stern et al. (2018) analyzed these data, primarily focused on reliability using

Data	κ_2	κ_3	κ_4	κ_5	Examiner Variation σ^2_α	Sample Variation σ^2_γ	Interaction Variation σ^2_δ
Signature Complexity (3-point scale)	-2.17 (-3.21, -1.43)	0.02 (-0.73, 0.75)	-	-	0.25 (0.04, 3.03)	3.55 (2.24, 6.74)	0.13 (0.00, 1.03)
Signature Complexity (5-point scale)	-4.14 (-6.07, -3.12)	-2.00 (-3.18, -1.21)	-0.25 (-1.07, 0.51)	1.82 (1.04, 2.92)	0.29 (0.06, 3.48)	3.48 (2.13, 7.14)	0.33 (0.00, 1.66)
Handwriting Comparisons Evaluation Phase	-2.47 (-2.84, -2.10)	-0.52 (-0.86, -0.19)	0.49 (0.16, 0.83)	1.93 (1.58, 2.30)	0.32 (0.23, 0.47)	4.01 (3.12, 5.13)	0.86 (0.64, 1.14)
Latent Prints Analysis phase	-3.18 (-3.76, -2.63)	-0.89 (-1.43, -0.38)	-	-	1.01 (0.79, 1.32)	17.22 (13.76, 21.77)	0.53 (0.37, 0.72)
Latent Prints Evaluation Phase	-1.08 (-1.29, -0.88)	1.57 (1.36, 1.79)	-	-	0.10 (0.08, 0.14)	5.77 (4.91, 6.68)	0.23 (0.13, 0.34)

Table 4.2: Results from fitting the SET model (4.3) to the 3-point scale complexity data from the signature complexity study, the 5-point scale complexity data from the signature complexity study, the data from the Analysis phase of the latent print examination process, the comparison decisions of the latent print examination process, and the handwriting comparison decisions. κ_4 and κ_5 are estimated for data sets with the number of ordinal categories, $M = 5$. Posterior medians with 95% credible intervals are presented.

the 5-point scale which was more appropriate to be approximated by a continuous scale. Our method enables this analysis using both scales as ordinal scales. Since we are interested in interactions between examiners and signature samples and we have very limited repetitions, we will be using the SET model (4.3) instead of the CUT model (4.1). We report results from fitting the method given by the SET model (4.3) to the 3-point scale data, reporting posterior medians and 95% credible intervals for the parameters in Table 4.2. Note that the posterior median estimates for σ^2_α and σ^2_δ are very small with the lower 2.5% quantile for the credible interval close to zero. This suggests that there is much more variance in signatures than examiners or interactions.

Table 4.3 presents reliability estimates on the latent and original scale that are derived from the SET model using expressions (4.6) and (4.7) and it suggests good reproducibility and repeatability on the latent and original scale. Additionally, we generate posterior predictive data sets to perform model checking and compare results with traditional methods of

assessing reliability using the original data set. Stern et al. (2018) reported that on the 1-3 point scale, exact agreement occurs between examiners for 63% of the signatures and examiners differed by a single category for 33% of the signatures. We generated 1000 posterior predictive data sets through posterior draws for σ^2_α, σ^2_γ, σ^2_δ and found that the mean exact agreement through the 1000 data sets was 62% with 95% credible interval (48%, 71%). Similarly, the mean number of signatures for which the examiners differed by exactly 1 point was 35% with the 95% credible interval (27%, 43%). This suggests that our model fits the data well and enables us to obtain uncertainties for the agreement statistics obtained from the data.

The 5-point scale data is again modeled with the method given by the SET model (4.3). The posterior medians and 95% credible intervals for parameters are presented in Table 4.2. We notice again that σ^2_γ is much larger in comparison to the other variance components. We present the reliability estimates derived from equations (4.6) and (4.7) in Table 4.3. We observe that even though the latent scale has good reliability on the 5-point complexity data, the agreement on the original scale is considerably lower. This is expected because when the number of ordinal categories, M, is larger, examiner agreement is smaller. Stern et al. (2018) reported that on the 1-5 point scale, exact agreement occurs between examiners for 45% of the signatures and examiners differed in their conclusions by more than 1 point in about 9% of the signatures. We generated 1000 posterior predictive data sets through posterior draws for σ^2_α, σ^2_γ, σ^2_δ and found that the posterior predictive statistics for percent agreement closely match those in the observed data.

It is interesting to note that, the estimated variance components for σ^2_γ and σ^2_α are comparable across the analysis from the 3-point data and the 5-point data. This could imply that the variance σ^2_γ is indeed capturing something intrinsic to the signatures irrespective of the scale of decisions that are being used to estimate the variance. Similarly, σ^2_α is also probably the variance related to some examiner characteristics. We do not have any identifying

Data	Latent reproducibility (given by (4.6))	Reproducibility on original scale (given by (4.7))	Latent repeatability (given by (4.6))	Repeatability on original scale (given by (4.7))
Signature Complexity (3-point scale)	0.70 (0.46, 0.79)	0.62 (0.52, 0.68)	0.80 (0.71, 0.90)	0.68 (0.63, 0.77)
Signature Complexity (5-point scale)	0.67 (0.43, 0.76)	0.47 (0.34, 0.49)	0.81 (0.71, 0.90)	0.53 (0.46, 0.64)
Handwriting Comparisons Evaluation Phase	0.65 (0.59, 0.69)	0.39 (0.36, 0.41)	0.84 (0.80, 0.87)	0.51 (0.48, 0.54)
Latent Prints Analysis phase	0.87 (0.85, 0.89)	0.74 (0.73, 0.76)	0.95 (0.94, 0.96)	0.83 (0.81, 0.84)
Latent Prints Evaluation Phase	0.81 (0.79, 0.83)	0.66 (0.65, 0.68)	0.86 (0.84, 0.87)	0.71 (0.69, 0.72)

Table 4.3: Reliability on the latent and original scale for the 3-point scale complexity data from the signature complexity study, the 5-scale complexity data from the signature complexity study, the data from the Analysis phase of the latent print examination process, the comparison decisions of the latent print examination process, and the handwriting comparison decisions are presented with 95% credible intervals. Note that credible intervals for the reliability on the latent scale are used for producing the credible intervals for the reliability on the original scale as per the expressions (4.7).

information about examiners or signature samples. If there were any available features, it would be interesting to check for any associations between examiner or sample features with α_i or γ_j respectively.

4.5.2 Latent Fingerprint Data

We next present results from different phases of the latent fingerprint black box study described in Section 4.2 and compare the results obtained here with those of Ulery et al. (2011, 2012).

4.5.2.1 Analysis Phase

Forensic labs use different approaches for latent fingerprint comparisons. However, we will refer to the approach described in Ulery et al. (2011). Quality determinations for latent fingerprints used the ordinal scale NV, VEO, and VID that suggest increasing quality. We have used the SET model (4.3) to fit the data from the latent print examination black-box study (Ulery et al., 2011). The posterior medians and 95% credible intervals are reported below in Table 4.2. Note that there are some interactions in the data although latent fingerprint variance is much higher in comparison to the examiner and interaction variance. This contributes to very high reliability on the latent scale as shown in Table 4.3. Additionally, the model-based agreement on the original scale is also good in Table 4.3.

Again, we perform posterior predictive analysis to check the model fit as well as obtain uncertainties around the usual agreement statistics obtained with the original data. Ulery et al. (2012) reported the inter-examiner percentage agreement to be 0.76 and intra-examiner percentage agreement to be 0.88 on a subset of the data set that consisted of the 72 examiners that participated in the repeatability study. We re-fit this subset of the data and generated posterior predictive data sets and found that the mean reproducibility with these data sets was 0.75 with 95% credible interval of (0.70, 0.79). This shows that our method is a good

fit for the data. The model-based reproducibility and repeatability on the original scale in Table 4.3 can also be compared to the agreement reported in Ulery et al. (2012). We observe that the model-based reproducibility estimate, 0.74, and repeatability estimate, 0.83, are both lower compared to the estimates reported in Ulery et al. (2012). One explanation is that Ulery et al. (2012) only used a subset of the data as described previously. Our approach used all the data which is more efficient. We have also accounted for possible interactions between examiner and samples. Our agreement measure is not prone to bias due to examiner preferences, as it is evaluated across examiners and it is not prone to bias due to imbalance of the print qualities that each examiner observed as, again, we have adjusted for sample difficulties. Our method also provides uncertainties around these reliability scores.

To better showcase the value of the latent model, we consider the results in terms of the alternate parameterization given by the constrained version of the CUT model (4.2). This provides inferences on individual examiner thresholds for rating samples. Figure 4.4 provides the posterior medians for the cutpoints $\tau_{i,2}$ and 95% credible intervals plotted against the percentage of No Value (NV) decisions given by an examiner. There are two noteworthy findings in Figure 4.4. First, there is a positive trend indication that a higher threshold corresponds to more NV decisions. The correlation between the %NV decisions and the posterior median for $\tau_{i,2}$ is 0.77. A second finding demonstrates the advantage of explicitly accounting for the samples assigned to each examiner. In Figure 4.4, we identify examiners A and B that have similar estimated thresholds, posterior medians $\hat{\tau_{A,1}} = \hat{\tau_{B,1}} = -3.65$ (these are identified in red in Figure 4.4), but different % NV decisions of 27% and 16% respectively. The average sample difficulty, as estimated by the mean of the posterior median print qualities (γ_j), is lower for the examples seen by examiner A (-0.32) compared to examiner B (0.47). Examiner A saw more low quality prints. Our method is able to account for such differences to identify examiners that have similar tendencies.

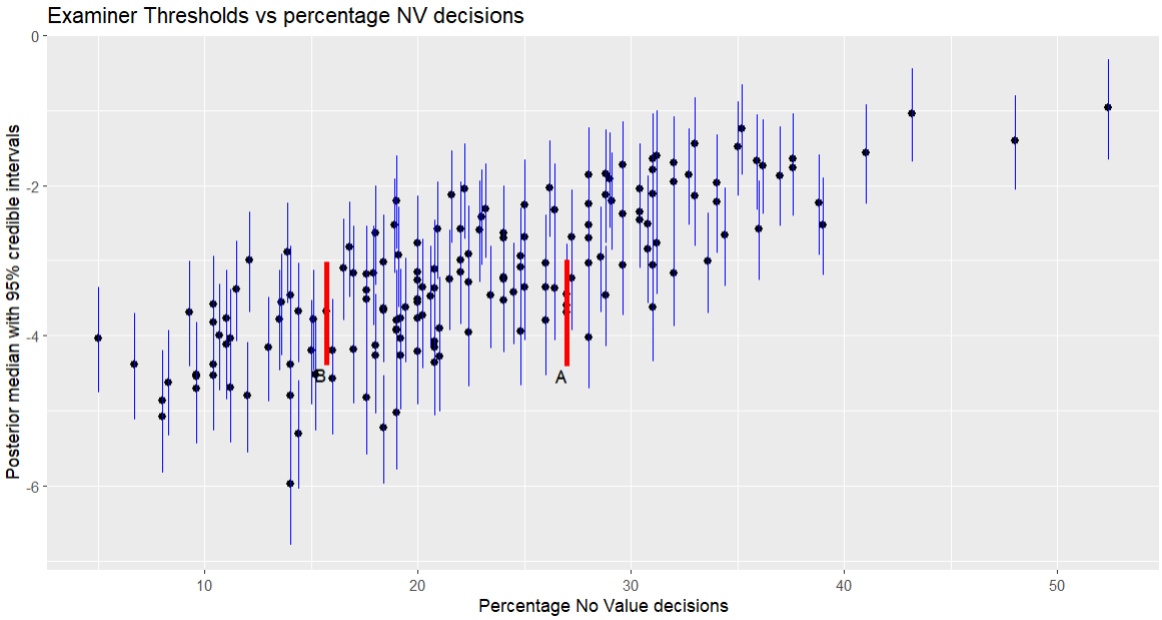


Figure 4.4: Figure presents the posterior median and 95% credible intervals for examiner thresholds $\tau_{i,2}$ for analysis decisions on the scale of VID, VEO, and NV plotted against the percentage of NV decisions given by examiner i. Examiners highlighted in red have similar estimated thresholds but have a very different percentage of NV decisions which is attributed to the fact that the difficulty of prints they analyzed were different.

4.5.2.2 Analysis Phase - Effect of imposing constraints

So far we have focused on the SET model (4.3) instead of the more flexible CUT model (4.1) because we did not have enough repeated decisions and we were interested in inferences about interactions. Given that the interaction variance is very small, it is possible to fit the CUT model (4.1) on the quality assessments data by eliminating the interaction terms.

Parameter	CUT model (4.1) (without δ_{ij})	Constrained model (4.2) (without δ_{ij})
σ^2_γ	14.77 (11.81, 18.67)	12.18 (9.82, 15.26)

Table 4.4: The posterior median estimates and credible intervals for σ^2_γ from the CUT model and the constrained version of the model (4.2) are compared here.

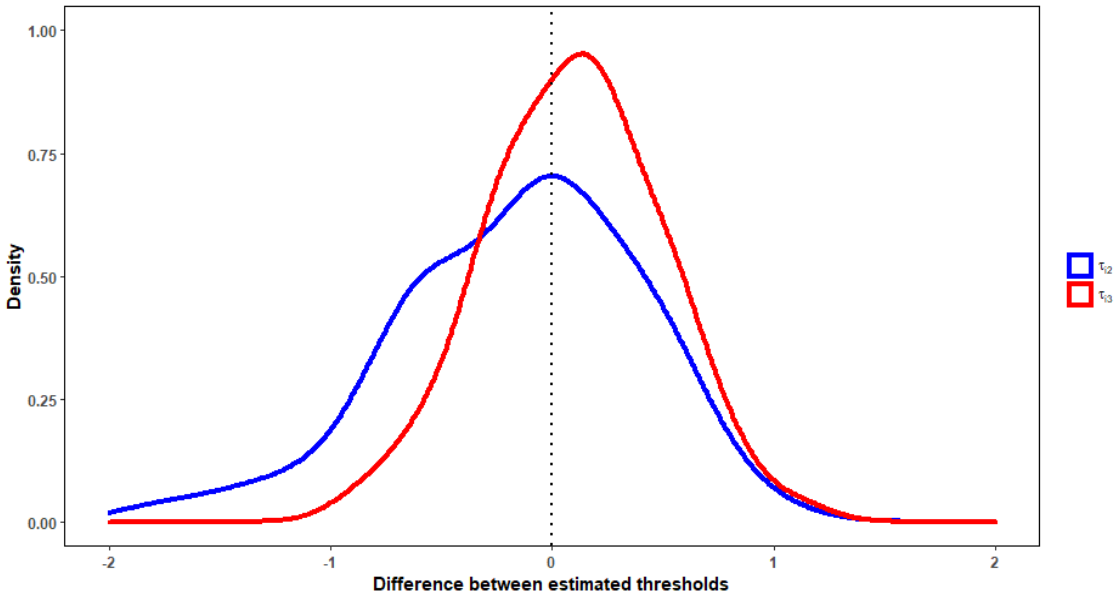


Figure 4.5: Differences between the estimated thresholds ($\tau'_{i,2}$ - $\tau_{i,2}$ and $\tau'_{i,3}$ - $\tau_{i,3}$) obtained by fitting the CUT model (4.1) and constrained version of the model (4.2) model (4.2).

Figure 4.5 presents the differences between the thresholds, $\tau'_{i,2}$ - $\tau_{i,2}$ and $\tau'_{i,3}$ - $\tau_{i,3}$, where τ' denote posterior medians using CUT model (4.1) with no interactions and τ denotes the posterior medians with the contrained version of the CUT model (4.2) with no interactions. We observe that for most examiners the differences between estimated thresholds from the two methods is small.

Figure 4.6 plots the distribution of $\tau'_{i,3}$ - $\tau'_{i,2}$ obtained using the CUT model (4.1) with no interactions and compares it to the estimated difference between cutpoints τ^* from the constrained version of the model (4.2) without interactions. We observe that some examiners have a difference $\tau'_{i,3}$ - $\tau'_{i,2}$ smaller than the difference $\tau_{i,3}$ - $\tau_{i,2} = \tau^* \approx 1.9$ obtained by the constrained version of the model (4.2) and some examiners have a greater difference between thresholds than $\tau_{i,3}$ - $\tau_{i,2} = \tau^*$. However, most examiners have a difference that is close to τ^*, this assures us that enforcing the constraint between examiner thresholds does not vastly affect the inference.

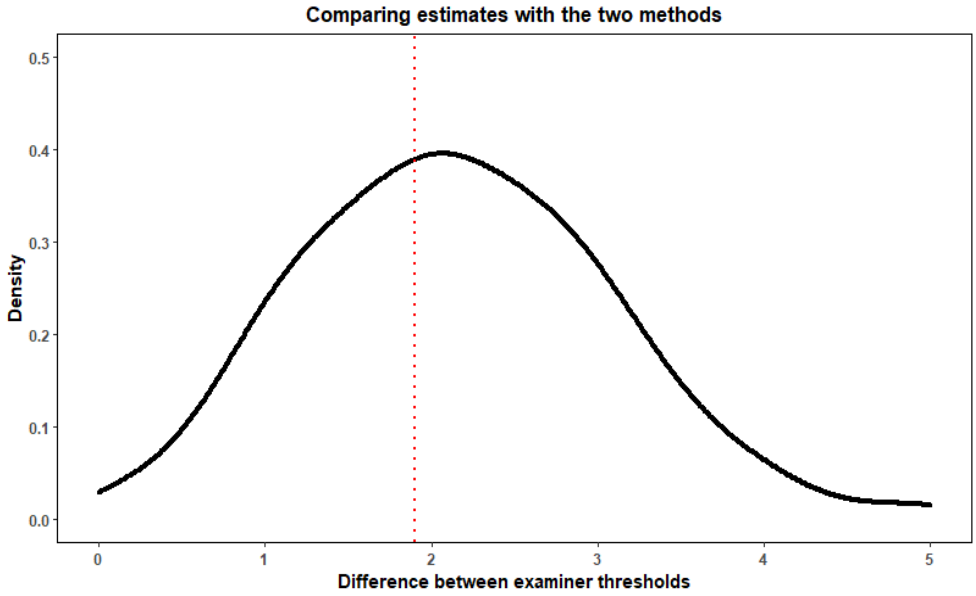


Figure 4.6: Distribution of $\tau'_{i,3}$ - $\tau'_{i,2}$ by fitting the CUT model (4.1) compared against estimated $\tau_{i,3}$ - $\tau_{i,2} = \tau^*$ from the constrained version of the model (4.2).

4.5.2.3 Comparison Decisions

We next analyze the decisions from the Evaluation phase of fingerprint comparisons. The decisions were on a scale of Exclusion, Inconclusive, and Individualization based on the degree of similarity between the latent print and exemplar print. We fit the data from this phase with the SET model (4.3). Here α_i represents the tendency of an examiner to declare Individualization decisions with higher values representing greater tendency to declare individualization and lower values representing higher tendency to exclude. Similarly, γ_j represents the degree of match between latent-exemplar pairs with higher values representing higher degree of similarity and lower values representing very little similarity. According to this interpretation for γ_j, an Inconclusive may arise when there are not enough similarities or differences to allow for an Individualization or Exclusion respectively.

We observe in Table 4.2 that the posterior estimates and credible interval limits for σ^2_α are quite small. This suggest limited evidence for variation among examiners in terms of any tendency toward individualization. The interaction variance is also small, though larger

than the examiner variance. Figure 4.7 presents the posterior distributions of γ_j against the percentage of exclusion decisions. We can observe that for mated pairs (in blue) the estimated γ_j are much higher compared to the non-mated pairs (in red). In both populations (mates and non-mates), lower values of γ_j are associated with more exclusions.

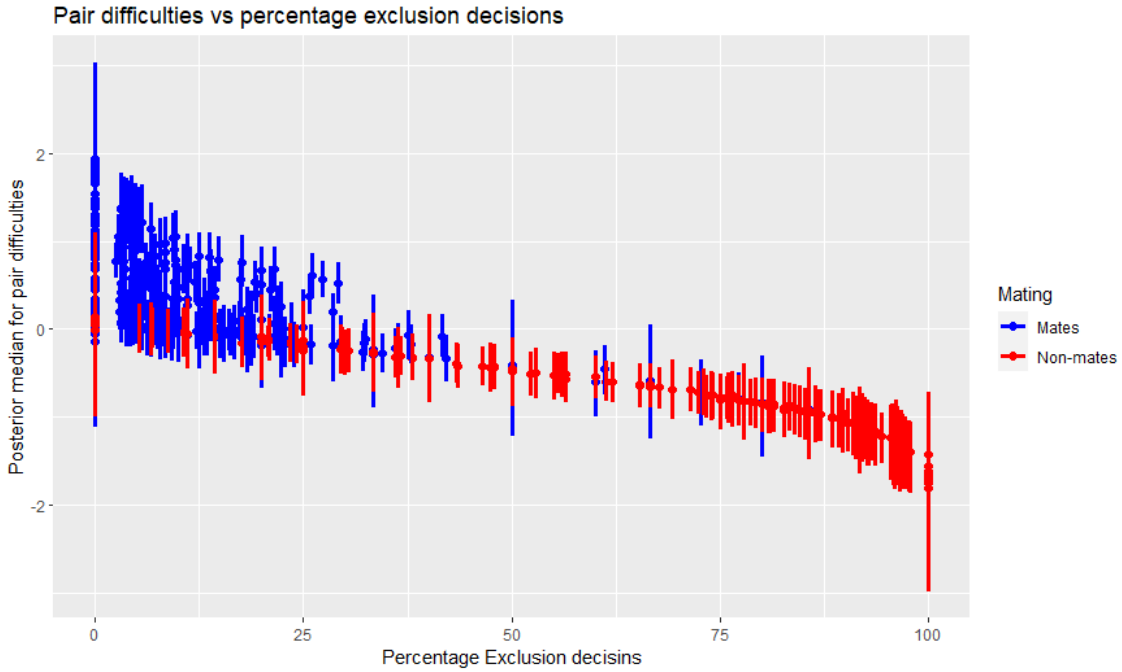


Figure 4.7: Posterior medians with 95% credible intervals for γ_j plotted against percentage exclusion decisions for mated and non-mated pairs.

Table 4.3 presents the reliability on the latent scale and original scale obtained by fitting the SET model (4.3) to comparison decisions. The reliability on the original scale is a bit lower than expected. Again, we conduct posterior predictive analysis to check if the model is a good fit for the data and to obtain credible intervals for percentage agreement statistics. The inter-examiner percentage agreement for the comparison decisions in the original data set was 0.76. We generated posterior predictive data sets based on the MCMC draws for the parameters σ^2_γ, σ^2_α, and σ^2_δ. In the posterior predictive data sets, the posterior median for inter-examiner reliability was 0.68 with 95% credible interval of (0.66, 0.69). A further investigation into the reason for this disparity revealed that the posterior predictive data set

had more latent-exemplar pairs for which all three decisions (Exclusion, Inconclusive, and Individualization) are observed compared to the original data. The original data usually includes only Exclusion/ Inconclusive conclusions or Inconclusive/ Individualization conclusions. A possible reason for this issue might be that a proxy for degree of similarity between latent-exemplar pairs through γ_j might not be sufficient to explain comparison decisions. Another possible reason could be that the fixed thresholds imposed with this model expect some examiners to use more Inconclusive decisions compared to the truth.

We further investigate whether our approach is better fitted to subsets of the data: mated and non-mated pairs modeled separately with the SET model (4.3). Within non-mated pairs, the inter-examiner agreement in the original data set was 0.79. We fit the SET model (4.3) to just the non-mated pairs and then generated posterior predictive data sets to check if the model is a better fit to the decisions on the non-mated pairs. We found that the posterior median for inter-examiner percentage agreement on the posterior predictive data sets was 0.77 with a 95% credible interval of (0.73, 0.81). This indicates that our model fits much better to non-mated pairs as compared to all the data. Within mated pairs, the percentage agreement in the original data set is 0.75. The SET model (4.3) fit only to the mated pairs suffered from lower posterior predictive percentage agreement like observed in the overall data set. The posterior median for percentage agreement was 0.60 with 95% credible interval of (0.58, 0.63). A possible reason for this is, as stated before, is that the model with fixed thresholds assumes that some examiners provide more Inconclusive decisions than reality.

4.5.3 Handwriting Comparisons

We next apply the SET model (4.3) to assess reliability in the handwriting black-box study (Hicklin et al., 2022a). The handwriting comparison decisions were reported on a 5-point ordinal scale: "NotWritten","ProbNot","NoConc","ProbWritten", and "Written". Table 4.2 shows these results. We note in the table that there seems to be some evidence for

interactions between examiners and handwriting samples. σ_α^2 is smaller than σ_δ^2, and σ_γ^2 is again much higher than both σ_α^2 and σ_δ^2. We present reliability on the latent scale and original scale in Table 4.3. We note that the reliability on the latent scale is good, however, model-based agreement on the original scale is poor. This is expected; percentage agreement is smaller when the number of categories, M is greater. Note that the only reliability measures provided in Hicklin et al. (2022) are through contingency tables that are conditional and not able to provide an overall estimate for reliability. To check if the model is a good fit for the data, we again generate posterior predictive data sets through σ_δ^2, σ_α^2, and σ_γ^2. The inter-examiner percentage agreement on the true data set was 0.41. With 1000 posterior predictive data sets we found that the posterior median for inter-examiner percentage agreement was and the 0.39 with a 95% credible interval of (0.38, 0.41). This indicates that our method is a good fit to these comparisons.

4.6 Conclusions

Decisions on an ordinal scale continue to be an important part of the forensic examination process. We have proposed a latent variable framework for assessing the reliability of forensic examination decisions using the data from reproducibility and repeatability studies. This modeling approach explains the variation in decisions through contributions from the samples, examiners, and a possible interaction between examiners and samples. The most flexible version of the model, incorporating separate decision thresholds for each examiner, is difficult to fit given that the typical study design includes fewer repeated decisions. We introduce a restricted model for use with limited repeated decisions that still enables quantifying different examiner thresholds for ordinal decisions. We applied these methods to three data sets, a signature complexity assessments study and two black box studies, a latent fingerprint analysis study and a handwriting comparisons study. The previously reported analyses on these data sets considered reproducibility and repeatability separately

and used contingency tables or percentage agreement on subsets of the data set to report reliability. These approaches ignore important information. The model here incorporates reproducibility and repeatability data and derives reliability estimates that adjust for the samples seen and the tendencies of examiners. The model developed also explicitly accounts for the possibility of interactions. We have established through simulation studies that these methods can provide a fair assessment of reliability even with as few as 25% repeated decisions. We also saw some evidence of interactions in latent fingerprint quality determinations and handwriting comparison decisions. We also noticed that accounting for interactions as well as using both the reproducibility and repeatability data sets, produced model-based agreement measures that were lower compared to the percentage agreement calculated using the black-box study data. This indicates that ignoring interactions or using only a portion of the data to calculate agreement may provide a false sense of higher reliability.

The results of the latent variable model can be used in an exploratory fashion to assess examiner and sample characteristics. For example, we identify examiners that have similar tendencies while accounting for the different print difficulties that were encountered by the examiners. If examiner covariates or sample covariates are collected, it will be possible to see how the examiner effects or sample effects are associated with such covariates. Examiner and sample features can also be added to the model.

Chapter 5

Identifying Clusters of Raters from Ordinal Data

5.1 Introduction

The design in various scientific and analytical studies involves multiple raters/ respondents that evaluate the same sets of items or answer the same set of questions on an ordinal scale. Examples include forensic science studies (Ulery et al., 2011; Hicklin et al., 2022a; Hicklin et al., 2022b), psychological studies (Glynn et al., 2018), course evaluations (Johnson and Albert, 2006), radiology or medical studies (Spoorenberg et al., 2004; Jones et al., 2007), and customer satisfaction questionnaires (Bradlow, 1994; Bradlow et al., 1999). The motivation for such studies typically includes evaluating the accuracy and reliability of decisions, assessing the distribution of rated items against other covariates associated with the raters, or conducting an analysis to explore whether there exist subpopulations of the raters that respond to items similarly.

Assessing whether there are extant subpopulations within the population of raters is of ex-

ploratory interest. Additionally, it may be helpful for hypothesis generation. For example, in forensic black-box studies that are aimed at assessing the reliability and accuracy of subjective feature-based comparison decisions by expert examiners, it is interesting to explore whether forensic examiners that make decisions similarly share certain covariates such as years of training/ experience, employer agency, etc. This can be explored directly if these data are recorded in advance, but this is not the case in common black-box study designs. Furthermore, certain forensic examination procedures consist of multiple steps. Latent fingerprint and footwear comparisons have a quality determination step that occurs before final comparisons occur between a questioned print and an exemplar print. In those settings, it is interesting to question whether examiners that, say, make similar decisions during quality assessments, have higher reproducibility in comparison decisions.

We propose a method that can cluster raters/ examiners based on their tendency to answer a similar set of questions on an ordinal scale. This method may also be applied to binary outcomes. The motivating data sets, described in Section 5.4, are two forensic black-box studies as well as a maternal depression study. Our method is an extension of a two-way random effects ANOVA model with interactions that has been previously used by Arora et al. (2022, 2023) to model continuous, binary, and ordinal outcomes from black-box studies. We use a Dirichlet process prior (Ferguson, 1973, 1974) on the examiner effects to encourage parameter sharing between examiners that make decisions similarly.

A number of authors have used finite mixture models to model ordinal data. For example, Breen and Luijkx (2010) have used a latent variable approach to model ordinal responses with a finite mixture of ordered logit models. Ranalli and Rocci (2016) used finite mixtures of Gaussian distributions to model ordinal responses. Their approach is dependent on evaluating an information criterion to select among different values for the number of mixture components. Matechou et al. (2016) used finite mixture models to bicluster rows and columns for ordinal data using a variational approximation. Mixtures of Dirichlet process

(Ferguson, 1973; Blackwell and MacQueen, 1973; Ferguson, 1974; Antoniak, 1974; MacEachern, 1994; Escobar, 1994; Escobar and West, 1995; Neal, 2000) have also been used to model discrete (non-continuous) data. Erkanli et al. (1993) used a mixture of Dirichlet processes (MDP) of probit links for ordinal data regression to predict class probabilities in a more flexible way. Mukhopadhyay and Gelfand (1997) have developed methods for using MDPs and overdispersed MDPs for generalized linear models (GLMs) and have shown how they may be efficiently used in place of GLMs and overdispersed GLMs for the purposes of prediction. Ibrahim and Kleinman (1998) suggested the use of a Dirichlet process prior in place of a normal distribution for modeling random effects in a mixed effect model. Shahbaba and Neal (2009) and Hannah et al. (2011) have proposed a generative Dirichlet process mixture method to model GLMs non-parametrically so that items in a mixture share the covariate distribution function and the parameters in the link function that connects the covariates with the items. The methods are shown to be superior to other prediction methods for GLMs. The approach we propose in this paper is most closely related to the one suggested in Ibrahim and Kleinman (1998), however, our goal is clustering raters/ items for the setting where raters mark the same or similar sets of items, which has not been addressed by any of these previous works. We hypothesize that raters that mark questions in a similar way may share some observed or unobserved covariates.

This paper is structured in the following way. In Section 5.2 we propose our method for clustering examiners/ samples using a Dirichlet process mixture setup. Computational methods are addressed there, as are approaches to inferring the number of clusters. Section 5.3 presents the results from a set of simulation studies. In Section 5.4, we use the proposed method to investigate the data from two forensic black-box studies and data regarding maternal depression. Finally, in Section 5.5 we summarize the obtained results, and discuss future applications and extensions of the proposed work.

5.2 Methods

Mixture models (McLachlan and Basford, 1988) are strong modeling candidates when the population under study is believed to contain subpopulations with different distributions. Finite mixture models typically require pre-specifying the number of mixture components. Nonparametric models such as Dirichlet process mixtures (Escobar, 1994; MacEachern, 1994; MacEachern and Müller, 1998) have become a popular choice due to their flexibility for density estimation and generalizability on new data. Rasmussen (1999) argue that infinite mixture models outperform finite mixture models due to avoiding the need to find the right number of clusters. We now provide a brief background on Dirichlet processes (DP) and methods of formulating DPs.

5.2.1 Dirichlet process and stick-breaking representation

Dirichlet processes (DP) (Ferguson, 1973; Blackwell and MacQueen, 1973; Ferguson, 1974) are nonparametric stochastic processes that define a probability model over distributions. Let, G_0 be a probability distribution over a measurable set, Ω, and let λ be a positive real number, then a Dirichlet process defined by (λ, G_0), defines random probability distributions over Ω as follows. For any finite partition $A_1, A_2, \ldots, A_k$ of Ω, if G is a DP defined by (λ, G_0), then:

$$(G(A_1),\, G(A_2),\, \ldots,\, G(A_k)) \sim \text{Dirichlet}\,(\lambda\, G_0(A_1),\, \lambda\, G_0(A_2),\, \ldots,\, \lambda\, G_0(A_k)),$$

where, Dirichlet$(\omega_1, \omega_2, \ldots, \omega_k)$ is the Dirichlet distribution with parameters $(\omega_1, \omega_2, \ldots, \omega_k)$. If a DP is used as a prior for the distribution of a set of independent and identically dis-

tributed data X_i, for $i \in \{1, 2, \ldots, n\}$,

$$\begin{aligned} X_i \,|\, G &\overset{i.i.d.}{\sim} G \\ G \,|\, \lambda, G_0 &\sim \mathrm{DP}(\lambda, G_0), \end{aligned} \tag{5.1}$$

then the posterior distribution of G is also a DP,

$$G \,|\, X_i, \lambda, G_0 \sim \mathrm{DP}\left(\lambda + n, \frac{\lambda}{\lambda + n} G_0 + \frac{1}{\lambda + n} \sum_{i=1}^{n} \delta_{X_i}\right).$$

where δ_{X_i} is the Dirac delta function having unit probability at X_i. The distribution G is infinite-dimensional and therefore is a nonparametric process. If G is integrated out of the specification given in equations (5.1), then X_i are no longer independent in their marginal distribution. It is then convenient to write the joint probability distribution, $p(X_1, X_2, \ldots, X_n)$ through the product of conditional distributions $p(X_j | X_1, X_2, ..., X_{j-1})$, as follows (Blackwell and MacQueen, 1973):

$$X_j | X_1, X_2, ..., X_{j-1} \begin{cases} = X_i, & \text{with probability } \frac{1}{j-1+\lambda},\ i = 1, 2, ..., j-1 \\ \sim G_0, & \text{with probability } \frac{\lambda}{j-1+\lambda} \end{cases} . \tag{5.2}$$

The parameter λ is known as a concentration parameter and it determines (probabilistically) how many distinct values are drawn from the base distribution G_0, with bigger values of λ supporting more distinct values.

Sethuraman (1994) proposed a constructive way to define G known as the "stick-breaking" process. Define a set of probabilities v_1, v_2, ... that are independent and identically distributed with a Beta(1, λ) distribution. Say there is a stick of length of 1, that is infinitely broken as follows. Let v_i define the proportion of the remaining stick that is broken at the i^{th} step. A stick-breaking prior is defined for the total proportion of the stick broken at each

step, w_1, w_2, ..., through v_1, v_2, ... in the following way:

$$w_1 = v_1$$

$$w_2 = v_2(1 - v_1)$$

$$w_n = v_n \prod_{i=1}^{n-1}(1 - v_i).$$

Then, if Y_1, Y_2, ... are drawn from the base distribution G_0, the stick-breaking representation of G is given through w_i and Y_i as follows:

$$G = \sum_{j=1}^{\infty} w_j \delta_{Y_j}, \tag{5.3}$$

where δ_z is again the Dirac delta function which represents a unit probability at z. We use the stick-breaking construction for sampling from a DP distribution in Sections 5.3 and 5.4.

5.2.2 Mixtures of Dirichlet processes

The DP described in Section 5.2.1 is often used as a prior distribution for the parameters in a hierarchical Bayesian model. Suppose Y_i, $i \in \{1, 2, \ldots\}$ are modeled as independent observations with distributions $F(.|\theta_i)$, where θ_i are assumed to have a DP prior, then the resulting model is known as a mixture of Dirichlet processes (MDP, Antoniak, 1974; Escobar, 1994; MacEachern, 1994; MacEachern and Müller, 1998). The MDP uses densities from a parametric family $F = \{f_\theta \mid \theta \in \Theta\}$, where the components θ have a DP prior over them. One way to describe the MDP is to first consider a finite mixture model with L components. Consider Y_i for $i \in \{1, 2, \ldots, n\}$ that are independently drawn from distributions $F(Y_i \,|\, \theta^*_{c_i})$, where c_i takes values in $\{1, 2, \ldots, L\}$ and follows a multinomial distribution with probability vector $\mathbf{p}$ that has length L, and θ^*_l are drawn from a distribution G_0, then this setup defines

a finite mixture model as follows:

$$\begin{aligned} \theta_l^* \,|\, G_0, \overset{i.i.d.}{\sim} G_0 \quad \forall\, l \in \{1, 2, ..., L\} \\ \mathbf{p} \,|\, \lambda \sim \text{Dirichlet}(\frac{\lambda}{L}, ..., \frac{\lambda}{L}) \\ c_i \,|\, \mathbf{p} \sim \text{Multinomial}(\mathbf{p}) \\ Y_i \,|\, c_i, \theta_1^*, ...\theta_L^* \sim F(Y_i \,|\, \theta_{c_i}^*). \end{aligned} \tag{5.4}$$

A mixture of Dirichlet processes (MDP) is obtained as the limit of the finite mixture model $L \to \infty$ (Teh, 2010) in the model (5.4). Model (5.5) presents an example of an MDP:

$$\begin{aligned} Y_i \,|\, \theta_i \overset{\text{ind.}}{\sim} F(Y_i \,|\, \theta_i), \quad \forall\, i \in \{1, 2, ..., n\} \\ \theta_i \,|\, G, \overset{\text{i.i.d.}}{\sim} G \\ G \,|\, \lambda, G_0 \sim DP(\lambda, G_0). \end{aligned} \tag{5.5}$$

As stated before, samples from a DP are a discrete distribution with probability one. It is observed in the representation of the DP that is marginalized over G in (5.2), there is a non-zero probability that $\theta_i = \theta_j$ are equal for $i \neq j$. Let there be K ($\leq n$) distinct values of θ_i across the n data points Y_i, $i \in \{1, 2, \ldots, n\}$, denoted as $\{\theta_1^*, \theta_2^*, \ldots, \theta_K^*\}$. Define $B_k = \{i : \theta_i = \theta_k^*\}$ or following the notation in the finite mixture model (5.4) $B_k = \{i : c_i = k\}$, for all $k \in \{1, 2, \ldots, K\}$; then the B_k, $k = 1, 2, \ldots, K$, define a partition over the set $\{1, 2, \ldots, n\}$ that can be viewed as a clustering. This is how MDPs are used to define clusters across the data set. The cluster memberships are random because the θ_i are random. Bush and MacEachern (1996) proposed a method to sample θ_i from its marginal distribution when G is integrated out (so that θ_i are dependent):

$$\theta_i \,|\, \theta_{-i}, \mathbf{x} \begin{cases} = \theta_k^{*-},\ k = 1, 2, \ldots, K \quad \text{with probability} \propto n_k^- \, p(Y_i \,|\, \theta_k^{*-}) \\ \sim H_1, \quad \text{with probability} \propto \lambda \int p(Y_i \,|\, \theta) dG_0(\theta) \end{cases},$$

where, θ_k^{*-} are the unique values of θ amongst $\theta_{-i} = \{\theta_1, \theta_2, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n\}$; $n_k^- = |\{j : j \neq i, \theta_{c_j}^{*-} = \theta_k^{*-}\}|$ (the number of θ_j that are equal to θ_k^{*-} for $j \neq i$); $H_1(\theta) \propto p(Y_i \,|\, \theta) dG_0(\theta)$. When $p(.)$ and G_0 are conjugate, evaluating H_1 and subsequently the sampling for θ_i is fairly straightforward. Otherwise, it can be difficult to sample from $H_1(\theta)$. We will now describe our proposed method to cluster raters based on ordinal decisions.

5.2.3 Ordinal Data Model

The aim of this paper is to apply the MDP for clustering in an ordinal data setting. We will use the language of the forensic science black-box studies that motivated this work.

Assume Y_{ijk} are ordinal outcomes on a scale of {1, 2, ..., M}, given by examiner i, on sample j, in the k^{th} trial. In a standard forensic reproducibility study, $k = 1$. Often, examiners are asked to provide repeated assessments on a subset of the samples ($k > 1$) that they observed in the reproducibility study to study the intra-examiner repeatability of judgments. We will assume that there are I examiners and J samples. As proposed in Albert and Chib (1993), we can model the ordinal data through a latent continuous variable, Z_{ijk}. The ordinal model data is written as:

$$\begin{aligned} P(Y_{ijk} = m) &= P(\kappa_m < Z_{ijk} \leq \kappa_{m+1}) \\ Z_{ijk} &\sim N(\alpha_{c_i}^* + \gamma_j, 1) \\ -\infty \equiv \kappa_1 \leq \kappa_2 \leq \ldots \leq \kappa_M \leq \kappa_{M+1} &\equiv \infty \end{aligned} \tag{5.6}$$

where, Z_{ijk} is the latent continuous variable corresponding to outcome Y_{ijk}, κ_2, κ_2, ..., κ_M are cutpoints that define the ordinal outcomes, γ_j is a sample or example effect that controls the tendency of the sample to be rated into higher or lower categories, and $\alpha_{c_i}^*$ is an examiner effect when examiner i belongs to a mixture component that is indicated by c_i. Examiners in a cluster are assumed to share their tendencies to rate samples into higher or lower categories.

The example effects γ_j are traditional random effects distribution centered at 0:

$$\gamma_j \,|\, \sigma_\gamma^2 \overset{i.i.d.}{\sim} N(0, \sigma_\gamma^2).$$

Arora et al. (2022, 2023) have used a latent variable approach to combine ordinal outcomes from the reproducibility and repeatability parts of forensic black-box studies while allowing for possible examiner-sample interaction. The presence of interactions implies that there is a differential effect of samples on examiner decisions/ ratings. If it is suspected that there are interactions between examiners and examples, i.e., examiner tendencies to rate samples are not constant across samples, we could extend the method given in model (5.6) to incorporate interactions as follows:

$$\begin{aligned} P(Y_{ijk} = m) &= P(\kappa_m < Z_{ijk} \leq \kappa_{m+1}) \\ Z_{ijk} &\sim N(\alpha^*_{c_i} + \gamma_j + \zeta_{ij}, 1) \end{aligned} \tag{5.7}$$

where, ζ_{ij}, is an interaction effect between examiner i and sample j. The assumption in model (5.7) is that examiners in a cluster still share tendencies to rate samples, even though the tendencies vary slightly (assuming that $|\zeta_{ij}| < |\alpha^*_{c_i}|$) based on the samples. In the forensic black-box studies presented in Section 5.4, we will model an interaction effect between examiners and forensic samples. We assume a standard random effects distribution on $\zeta_{ij} \sim N(0, \sigma_\zeta^2)$.

A DP prior is used on examiner tendencies $\alpha^*_{c_i}$. Ishwaran and James (2001), Ishwaran and Zarepour (2002) and Ishwaran and James (2002) proposed a blocked Gibbs sampling technique for fitting a non-conjugate mixture of DPs. Their method relies on a truncation of the stick-breaking representation of G in equation (5.3) at a value T. They provided an error bound for the absolute difference in the marginal densities of the outcomes ($p(x_i)$ in

model (5.5)) when the DP model is not truncated and in the truncated model ($\int_{-\infty}^{\infty} | \, p_{\infty}(x_i) - p_T(x_i) \, | \, dx_i$, that is approximately $4\, n \exp\left(\frac{-(T-1)}{\lambda}\right)$ for a normal mixture model with sample size n when the infinite representation of G is truncated at a value T. This bound applies to $\int_{-\infty}^{\infty} | \, p_{\infty}(Z_{ijk}) - p_T(Z_{ijk}) \, | \, dZ_{ijk}$ in the model (5.6) as well. Ishwaran and James (2002) noted that even with $n = 1000$ samples, and $\lambda = 3$, $T = 50$ leads to an error bound of 3.2×10^{-4}. We will be using $T = I$, which is the largest possible value for our applications.

The truncated stick-breaking priors are described in expression (5.8). Choosing the final stick-breaking probability $v_T = 1$ ensures that $\sum_{t=1}^{T} w_t = 1$. The stick-breaking weights are defined as follows:

$$\begin{aligned} v_t \overset{i.i.d.}{\sim} \text{Beta}(1, \lambda), \; v_T = 1 \quad \forall \, t \, \in \, \{1, 2, ..., T-1\} \\ w_1 = v_1, \; w_t = v_t \prod_{l=1}^{t-1} (1 - v_l), \quad \forall \, t \, \in \, \{2, ..., T\}. \end{aligned} \tag{5.8}$$

The weights w_t are used to define an approximate Dirichlet process that are used to cluster examiners in (5.9). The membership indicator c_i takes values in $\{1, 2, \ldots, T\}$, according to:

$$c_i \, | \, \mathbf{w} = \sum_{t=1}^{T} w_t \, \delta_t. \tag{5.9}$$

The examiner effects, α_t^*, are drawn from the base distribution as follows:

$$\alpha_t^* \, | \, \mu_0, \sigma_0^2 \overset{i.i.d.}{\sim} N(\mu_0, \sigma_0^2) \equiv G_0 \quad \forall \, t \, \in \, \{1, 2, ..., T\} \tag{5.10}$$

We finally complete the specification of the model with hyperpriors on $\sigma_0^2, \sigma_\gamma^2$ (according to

Gelman, 2006), and μ_0:

$$p(\sigma_0^2, \sigma_\gamma^2) \propto \frac{1}{\sigma_0 \sigma_\gamma}$$

$$p(\mu_0) \propto \mathbf{1}$$

We use a mean parameter with the base distribution μ_0. With this parameter, κ_2 is set to zero because it would otherwise not be identified. We use a noninformative uniform prior on the ordered thresholds κ_m:

$$\kappa_2 = 0\,;\, p(\kappa_3, \ldots, \kappa_M) \propto 1_{0 \leq \kappa_3 \leq \ldots \leq \kappa_M} \tag{5.11}$$

5.2.4 Bayesian Computation

A Gibbs algorithm (Geman and Geman, 1984) is used to sample from the posterior distribution of the full conditionals for Z_{ijk}, γ_j, and σ_γ^2 (5.12):

$$Z_{ijk} \,|\, \{\alpha_t^*\}_{t=1}^T, \{c_i\}_{i=1}^I, \{\gamma_j\}_{j=1}^J, \{Y_{ijk}\}, \{\kappa_m\}_{m=2}^M \sim \begin{cases} N(\alpha_{c_i}^* + \gamma_j, 1)\, I(-\infty, \kappa_2 = 0), & \text{if } Y_{ijk} = 1 \\ N(\alpha_{c_i}^* + \gamma_j, 1)\, I(\kappa_2 = 0, \kappa_3), & \text{if } Y_{ijk} = 2 \\ \vdots & \\ N(\alpha_{c_i}^* + \gamma_j, 1)\, I(\kappa_M, \infty), & \text{if } Y_{ijk} = M \end{cases}$$

$$\{\gamma_j\}_{j=1}^J \,|\, \{Z_{ijk}\}, \{\alpha_t^*\}_{t=1}^T, \{c_i\}_{i=1}^I, \sigma_\gamma^2 \sim N\left(\frac{\sum_i \sum_k (Z_{ijk} - \alpha_{c_i}^*)}{\frac{1}{\sigma_\gamma^2} + \sum_i \sum_k \mathbb{1}_{\mathbb{ijk}}}, \frac{1}{\frac{1}{\sigma_\gamma^2} + \sum_i \sum_k \mathbb{1}_{\mathbb{ijk}}}\right) \tag{5.12}$$

$$\sigma_\gamma^2 \,|\, \{\gamma_j\}_{j=1}^J \sim \text{Inv-Gamma}\left(\frac{J-1}{2}, \frac{\sum_j \gamma_j^2}{2}\right)$$

where, $\mathbb{1}_{\mathbb{ijk}}$ is an indicator function that is equal to 1 if Y_{ijk} is observed in the data, 0 otherwise. Sampling the thresholds κ_m through a typical Gibbs sampling strategy conditional on

the Z_{ijk} mixes very slowly. To remedy that we use a Metropolis-Hastings update as suggested in Cowles (1996) which works with the densities for $\{\kappa_m\}_{m=3}^{M} \,|\, \{Y_{ijk}\}, \{\alpha_t^*\}_{t=1}^{T}, \{\gamma_j\}_{j=1}^{J}, \{c_i\}_{i=1}^{I}$ instead of $\kappa_m \,|\, \{Z_{ijk}\}, \{\alpha_t^*\}_{t=1}^{T}, \{\gamma_j\}_{j=1}^{J}, \{c_i\}_{i=1}^{I}$. We note that the joint distribution of $\{Z_{ijk}\}$ and $\{\kappa_m\}_{m=3}^{M}$ can be written as:

$$
\begin{aligned}
p(\{Z_{ijk}\}, \{\kappa_m\}_{m=3}^{M} \,|\, \{Y_{ijk}\}, \{\alpha_t^*\}_{t=1}^{T}, \{\gamma_j\}_{j=1}^{J}, \{c_i\}_{i=1}^{I}) \propto \\
p(\{Z_{ijk}\} \,|\, \{\kappa_m\}_{m=3}^{M}, \{Y_{ijk}\}, \{\alpha_t^*\}_{t=1}^{T}, \{\gamma_j\}_{j=1}^{J}, \{c_i\}_{i=1}^{I}) \\
p(\{\kappa_m\}_{m=3}^{M} \,|\, \{Y_{ijk}\}, \{\alpha_t^*\}_{t=1}^{T}, \{\gamma_j\}_{j=1}^{J}, \{c_i\}_{i=1}^{I})
\end{aligned}
$$

The first term is the Gibbs sampling update as per (5.12). Updates for κ_m are done as follows. During the Gibbs/ Metropolis-Hastings updates let the current values of κ_m be denoted as κ_m^{old}. The proposal distribution for candidates for κ_m is a truncated normal $k_m \sim N(\kappa_m^{\text{old}}, \sigma^2_{\text{prop}})\, I(k_{m-1}, \kappa_{m+1}^{\text{old}})$ for $m = 3, 4, \ldots, M$ with $\kappa_1^{\text{old}} = k_1 = -\infty$, $\kappa_2^{\text{old}} = k_2 = 0$, and $\kappa_{M+1}^{\text{old}} = k_{M+1} = \infty$. Define R as follows:

$$
R = \prod_{i,j,k} \frac{\Phi(k_{Y_{ijk}+1} - \alpha_{c_i}^* - \gamma_j) - \Phi(k_{Y_{ijk}} - \alpha_{c_i}^* - \gamma_j)}{\Phi(\kappa_{Y_{ijk}+1}^{\text{old}} - \alpha_{c_i}^* - \gamma_j) - \Phi(\kappa_{Y_{ijk}}^{\text{old}} - \alpha_{c_i}^* - \gamma_j)} \prod_{m=3}^{M} \frac{\Phi((\kappa_{m+1}^{\text{old}} - \kappa_m^{\text{old}})/\sigma_{\text{prop}}) - \Phi((k_{m-1} - \kappa_m^{\text{old}})/\sigma_{\text{prop}})}{\Phi((k_{m+1} - k_m)/\sigma_{\text{prop}}) - \Phi((\kappa_{m-1}^{\text{old}} - k_m)/\sigma_{\text{prop}})}
$$

where $\Phi(.)$ is the cumulative density function for standard normal distribution. We accept k_m as the new updates for κ_m with probability min(R,1), otherwise $\kappa_m = \kappa_m^{\text{old}}$. Since the likelihood terms in R can cause underflow issues while sampling, we calculate the logarithm of R during the computation. The variance of the proposal distribution σ^2_{prop} is chosen so that the acceptance rate for the new candidates k_m ranges between 0.25-0.5 (Gelman et al., 1996). We start with a value of $\sigma_{\text{prop}} = \frac{0.5}{M}$ and the value may be decreased or increased if the acceptance rate is less than 0.25 or greater than 0.5 respectively (Johnson and Albert, 2006).

The method to sample membership labels $\{c_i\}_{i=1}^{I}$ is presented next:

$$c_i \,|\, \{Z_{ijk}\}, \{\alpha_t^*\}, \{v\} \overset{ind}{\sim} \sum_{t=1}^{T} \beta_{t,i}\delta_t$$

$$\beta_{t,i} \,|\, \{w_t\}_{t=1}^{T}, \{Z_{ijk}\}, \{\alpha_t^*\}_{t=1}^{T}, \{\gamma_j\}_{j=1}^{J} \propto w_t \prod_j \prod_k \phi(Z_{ijk}|\alpha_t^* + \gamma_j,\, 1)$$

$\beta_{t,i}$ are the probabilities that examiner i belongs to the cluster t. The product in the update for $\beta_{t,i}$ is evaluated for all the samples j that are observed by the examiner i in repetition k. Here, $\phi(.|\omega, \tau^2)$ is the Gaussian probability distribution function with mean ω and variance τ^2. Also note that $\sum_t \beta_{t,i} = 1$. The the stick-breaking parameters $\mathbf{v}$, $\mathbf{w}$ are updated as follows:

$$v_t \,|\, \{c_i\}_{i=1}^{I} \sim \text{Beta}\,(\,1 + S_t\,,\, \lambda + \sum_{l=t+1}^{T} S_l\,),\; v_T = 1, \quad S_t = \sum_{i\,:\,c_i=t} \mathbf{1} \tag{5.13}$$

$$w_1 = v_1,\; w_t = v_t \prod_{l=1}^{t-1}(1 - v_l)$$

S_t are defined as the number of examiners in cluster t. The cluster effects α_t^* are updated through the densities proportional to the following expressions, depending on the examiners that belong in cluster t:

$$p(\alpha_t^* \,|\, \mu_0, \sigma_0^2, \{c_i\}_{i=1}^{I}, \{Z_{ijk}\}) \propto \begin{cases} \phi(\alpha_t^*|\mu_0,\, \sigma_0^2), & \text{if no examiners are associated with component } t \\ \left(\prod_{i:c_i=t} \prod_j \prod_k \phi(\alpha_t^*|Z_{ijk} - \gamma_j\,,\, 1)\right)\, \phi(\alpha_t^*|\mu_0,\, \sigma_0^2), & \text{otherwise} \end{cases} \tag{5.14}$$

The conditionals for the hyperparameters μ_0 and σ_0^2 use α_t^* from clusters that have at least

one examiner associated with them:

$$\mu_0 \,|\, \{c_i\}_{i=1}^{I}, \{\alpha_t^*\}_{t=1}^{T}, \sigma_0^2 \sim N\left(\frac{\sum_{t'} \alpha_{t'}^*}{\sum_{t'} \mathbf{1}_{t'}}, \frac{\sigma_0^2}{\sum_{t'} \mathbf{1}_t}\right), \quad t' \in \{t : \exists \text{ i}, \ c_i = t\} \tag{5.15}$$

$$\sigma_0^2 \,|\, \{c_i\}_{i=1}^{I}, \{\alpha_t^*\}_{t=1}^{T}, \mu_0 \sim \text{Inv-Gamma}\left(\frac{\sum_{t'} \mathbf{1}_{t'} - 1}{2}, \frac{\sum_{t'} (\alpha_{t'}^* - \mu_0)^2}{2}\right), \quad t' \in \{t : \exists \text{ i}, \ c_i = t\}$$

In a non-truncated DP representation, Escobar and West (1995) and Görür and Rasmussen (2010) have proposed methods of sampling the concentration parameter λ. However, we may not be able to utilize those because we have used the truncated stick-breaking process. We instead use the method suggested in Ishwaran and Zarepour (2000). If a Gamma prior is chosen over λ then the posterior is also Gamma:

$$\lambda \,|\, a_1, a_2 \sim \text{Gamma}(a_1, a_2)$$

$$\lambda \,|\, \mathbf{w} \sim \text{Gamma}(T + a_1 - 1, \ a_2 - \log(w_T))$$

Alternatively,

$$\lambda \,|\, \mathbf{v} \sim \text{Gamma}(T + a_1 - 1, \ a_2 - \sum_{t=1}^{T-1} \log(1 - v_t)) \tag{5.16}$$

More discussion on the prior chosen for λ can be found in Section 5.3.

5.2.5 Posterior Inference - Consensus clustering

The Gibbs algorithm described in Section 5.2.4 will produce samples from the posterior distribution for the parameters as well as the cluster memberships c_i for each examiner. Posterior inferences for the cluster memberships can be challenging to interpret due to the randomness in c_i as well as label-switching (Stephens, 2000). For example, Figure 5.1 presents the posterior draws for the number of clusters for a simulated scenario (Scenario E in Table 5.1) with $I = 50$ raters that belonged to 3 clusters during the data generating process. We observe that the mode of the posterior distribution for the number of clusters is 3 which is the true value for the number of clusters. However, we also observe that the distribution

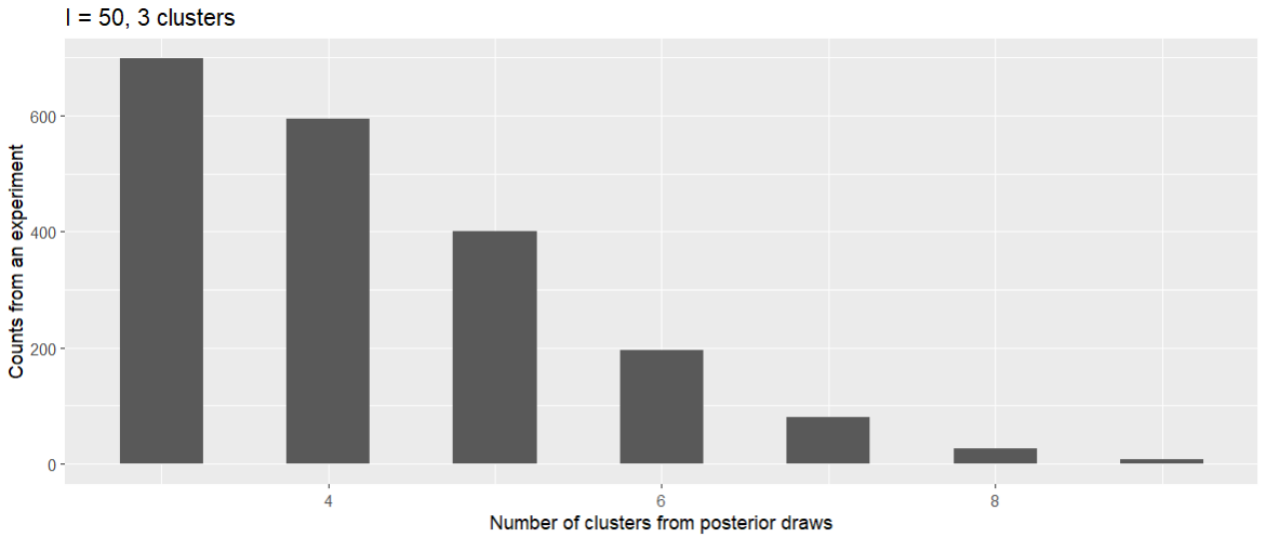


Figure 5.1: Distribution for the number of clusters based on posterior draws from one of the five simulated data sets in I = 50 and 3 clusters settings (Scenario E). Note the long tail for the draws.

has a long tail. One approach to posterior inference would be to report this posterior distribution and the posterior inferences for parameters and clusterings conditional on the number of clusters. We chose not to do this because the distribution of clusterings exhibits great variation and is difficult to interpret. Instead, we draw inferences based on a consensus cluster obtained using the algorithm of Dahl et al. (2022).

Rastelli and Friel (2018), Wade and Ghahramani (2018), and Dahl et al. (2022) provide different algorithms for obtaining point estimates for a consensus clustering through the posterior draws of the membership labels c_i. The algorithms minimize a posterior loss such as Binder loss (Binder, 1978), variation of information (Meilă, 2007), and their variants. We use variation of information (defined below), which is invariant to label-switching and is therefore particularly attractive as a loss function. First, given two ways to cluster data $\{1, 2, \ldots, N_0\}$ namely the clustering $\mathbf{a}$ with clusters $d_1, d_2, \ldots, d_{T_1}$ and the clustering $\mathbf{b}$

with clusters $e_1, e_2, \ldots, e_{T_2}$, define:

$$n^{\mathbf{a},\mathbf{b}}_{d_j,e_k} = \sum_{i=1}^{N_0} I_{c_i=d_j} I_{c'_i=e_k}$$

$$n^{\mathbf{a}}_{d_j} = \sum_{k=1}^{T_2} n^{\mathbf{a},\mathbf{b}}_{d_j,e_k}$$

where $c_i \in \{d_1, d_2, \ldots, d_{T_1}\}$ defines cluster membership for data i under $\mathbf{a}$ and $c'_i \in \{e_1, e_2, \ldots, e_{T_2}\}$ defines cluster membership for data under $\mathbf{b}$. The entropy for clustering $\mathbf{a}$ is defined as:

$$H(\mathbf{a}) = -\sum_{j=1}^{T_1} \frac{n^{\mathbf{a}}_{d_j}}{N_0} log(\frac{n^{\mathbf{a}}_{d_j}}{N_0})$$

The joint entropy between $\mathbf{a}$ and $\mathbf{b}$ and the variation of information are defined as follows:

$$H(\mathbf{a},\mathbf{b}) = -\sum_{j,k} \frac{n^{\mathbf{a},\mathbf{b}}_{d_j,e_k}}{N_0} log\left(\frac{n^{\mathbf{a},\mathbf{b}}_{d_j,e_k}}{N_0}\right)$$

$$L_{VI}(\mathbf{a},\mathbf{b}) = 2H(\mathbf{a},\mathbf{b}) - H(\mathbf{a}) - H(\mathbf{b})$$

We use the greedy algorithm provided in Dahl et al. (2022) that finds the consensus clusterings for examiners $\hat{c} = \{\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_{N_0}\}$ that minimizes the variation of information loss given the posterior draws for these clusterings. Let, the number of MCMC draws be $d = 1, 2, \ldots, D$, then the consensus clusters ($\hat{c}$) are obtained as follows:

$$\hat{c} = \text{argmin}_{\hat{c}} \sum_i \log_2\left(\sum_j I_{\hat{c}_i=\hat{c}_j}\right) - 2\sum_i \log_2\left(\sum_j \pi^*_{ij} I_{\hat{c}_i=\hat{c}_j}\right)$$

$$\text{where, } \pi^*_{ij} = \frac{1}{D}\sum_{d=1}^{D} I(c^{(d)}_j = c^{(d)}_i), \quad \forall j \neq i$$

Scenario	Number of clusters	Cluster 1 mean (Cluster 1 size I=50/ I=150)	Cluster 2 mean (Cluster 2 size I=50/ I=150)	Cluster 3 mean (Cluster 3 size I=50/ I=150)	Cluster 4 mean (Cluster 4 size I=50/ I=150)	Cluster 5 mean (Cluster 5 size I=50/ I=150)	κ_2	κ_3
A	3	-2.00 (16/ 50)	0.00 (18/ 50)	2.00 (16/ 50)	-	-	-1.00	1.00
B	3	-2.00 (16/ 50)	0.00 (18/ 50)	1.00 (16/ 50)	-	-	-1.00	0.50
C	3	-1.00 (16/ 50)	0.00 (18/ 50)	2.00 (16/ 50)	-	-	0.00	1.00
D	3	-2.00 (16/ 50)	0.00 (18/ 50)	1.00 (16/ 50)	-	-	-1.00	1.00
E	3	-3.00 (16/ 50)	0.00 (18/ 50)	2.00 (16/ 50)	-	-	-2.00	1.00
F	5	-2.00 (10/ 30)	-1.00 (10/ 30)	0.00 (10/ 30)	1.00 (10/ 30)	2.00 (10/ 30)	-1.00	1.00
G	5	-3.00 (10/ 30)	-1.75 (10/ 30)	0.00 (10/ 30)	0.75 (10/ 30)	1.50 (10/ 30)	-2.00	0.50
H	5	-1.50 (10/ 30)	-1.00 (10/ 30)	0.00 (10/ 30)	0.75 (10/ 30)	1.25 (10/ 30)	-0.50	1.00
I	5	-1.50 (10/ 30)	-1.00 (10/ 30)	0.50 (10/ 30)	1.00 (10/ 30)	2.00 (10/ 30)	-1.00	1.50
J	5	-1.75 (10/ 30)	-0.75 (10/ 30)	0.00 (10/ 30)	0.75 (10/ 30)	2.00 (10/ 30)	-0.50	1.50
K	5	-2.00 (-/ 12)	-1.00 (-/ 13)	0.00 (-/ 15)	1.00 (-/ 10)	2.00 (-/ 100)	-1.00	1.00
L	5	-1.50 (-/ 100)	-0.75 (-/ 10)	0.00 (-/ 15)	0.75 (-/ 13)	1.50 (-/ 12)	-1.00	1.00
M	5	-1.50 (-/ 20)	-1.00 (-/ 45)	0.00 (-/ 55)	0.75 (-/ 20)	1.25 (-/ 10)	-1.00	1.00
N	5	-1.50 (-/ 10)	-1.00 (-/ 50)	0.50 (-/ 50)	1.00 (-/ 30)	2.00 (-/ 10)	-1.00	1.00
O	5	-1.75 (-/ 10)	-0.75 (-/ 40)	0.00 (-/ 50)	0.75 (-/ 40)	2.00 (-/ 10)	-1.00	1.00

Table 5.1: The different generated simulation scenarios are detailed. There are 3 or 5 clusters in the generated data set with $I = 50$ or $I = 150$ examiners. The cluster means and the number of examiners in each cluster are indicated for each design. The cutpoints κ_2 and κ_3 vary across the scenarios. $J = 50$ γ_j's are generated from N(0, $\sigma^2_\gamma = 10$) separately for all scenarios.

5.3 Simulation studies

We will now present the results from simulation studies that were conducted to check how effectively the model and algorithm presented in Section 5.2 are able to differentiate between clusters in different settings.

Data generating process

Data is generated for all of the simulation studies through the ordinal data model in (5.6) with $M = 3$ ordinal outcome categories. The number of examiners/ participants in the simulated data sets are either $I = 50$ or $I = 150$. These sizes are based on the experiments discussed in Section 5.4. The number of samples assessed by each examiner was fixed at $J = 50$ for all experiments. In generating the data, we vary the cluster means as described below and set the variances of the sample random effects $\sigma^2_\gamma = 10$ (we do not simulate interactions). In all there are 10 scenarios for the cluster means, five with three clusters and five with five clusters. These are denoted as Scenarios A through J in Table 5.1. We carry out a simulation for each scenario with $I = 50$ and $I = 150$.

Scenario	Values for α_i (I=50)	κ_2	κ_3
P	{-1.00, -0.96, ..., 0.92, 0.96}	-0.50	0.50
Q	{-2.00, -1.92, ..., 1.84, 1.92}	-1.00	1.00
R	{-3.00, -2.88, ..., 2.76, 2.88}	-1.50	1.50
S	{-4.00, -3.84, ..., 3.68, 3.84}	-2.00	2.00
T	{-5.00, -4.80, ..., 4.60, 4.80}	-2.50	2.50

Table 5.2: Data generating Scenarios P-T with $I = 50$ examiners where there are no examiner clusters in the data generating model. The examiner effects are increasingly more separated as we move down the rows. The cutpoints κ_2 and κ_3 vary across the scenarios. $J = 50$ γ_j's are generated from N(0, $\sigma^2_\gamma = 10$) separately for all scenarios.

The cluster means are shared within the cluster and the cluster sizes across the scenarios are balanced for $I = 50$ and $I = 150$ designs. The distances between clusters are varied as demonstrated in Table 5.1. Additionally, the thresholds κ_2 and κ_3 are also different across the scenarios. In practice, there may be an imbalance in the number of examiners in a cluster. We need to test our method for cases that have an uneven number of examiners in certain clusters. We carried out 5 additional simulations with uneven cluster sizes identified as Scenarios K through O for $I = 150$ examiners in Table 5.1. Furthermore, we generate data sets with no clusters, in a design with $I = 50$ total examiners and $J = 50$ samples. In the unclustered simulations, the examiner parameters α_i are evenly spaced out over a range. These are denoted as scenarios P-T in Table 5.2.

Model fitting process

For each of the simulated data sets, we use the hybrid Gibbs/ Metropolis-Hastings algorithm sampler of Section 5.2.4 for 2000 iterations and 4 chains. It is tricky to assess the convergence in an MDP model due to label-switching (Stephens, 2000), which arises from the fact that the likelihood for an MDP model is invariant to permutation between the cluster labels. It was recommended in Gelman et al. (2013) that the convergence should be checked on the parameters not related to the mixture components. The convergence was assessed through the potential scale reduction factor (PSRF, Gelman and Rubin, 1992) on the parameters σ^2_γ, κ_3, μ_0, σ^2_0 as well as the log-likelihood $\log\Big(\prod_{i,j} p(Y_{ij} \mid \alpha^*_{c_i}, \gamma_j)\Big)$. We found that 2000 iterations

were sufficient to fit the data sets. The first 1000 draws in each chain were discarded and posterior inference is based on the second half.

The truncated stick-breaking process was used with $T = I = 50$ or 150. Examiners were initially assigned to separate clusters in the initial step. The initial values for each chain in the sampler are set as follows: the parameters σ_γ^2, σ_0^2, were drawn from a Uniform(0,2) distribution; μ_0 was set to be 0 initially for each chain; κ_3 is initially drawn from Unif(0,1); initial value for λ was chosen to be 2; and v was generated with Beta(1,2).

The value of λ controls the distribution of the number of clusters and can have a big impact on posterior inferences with higher values of λ supporting more clusters. The prior for λ requires careful consideration. We made the choice of the prior based on some simulating data sets where we monitored the number of clusters obtained by a DP prior with different sizes of the data sets and different values of λ. The results of these simulation runs are reported in the Supplemental material. Based on the runs, we found that the range of λ between 0.5 - 3 provides enough flexibility for the number of clusters that seem likely to be interesting and scientifically relevant for the data sets of Section 5.4. By this we mean, a small enough number of clusters so that there are multiple examiners per cluster. This led us to choose a Gamma(2,2) prior distribution. The rate parameter in gamma distribution controls the spread of the distribution with higher values meaning that the variance is lower. A large portion (95%) of the selected gamma distribution is between 0.1-2.8. These values support a prior expected range of up to 20 clusters but typically 3-10 clusters.

As described in Section 5.2.5, we use the algorithm presented in Dahl et al. (2022) to obtain a consensus clustering for the raters in each simulation. Dahl et al. (2022) provided a parallel implementation of their algorithm which is available through the CRAN as an R package (Dahl et al., 2021). Consensus clusterings are compared to the clusters that the raters belonged to in the data-generating process.

Results

In all, we have 30 simulation scenarios as described in Table 5.1. We summarize results separately for six different situations: the five Scenarios A-E with I=50 (3 clusters), the five Scenarios F-J with I=50 (5 clusters), the five Scenarios A-E with I=150 (3 clusters), the five Scenarios F-J with I=150 (5 clusters), the five Scenarios K-O with I=150 (5 clusters) with an uneven number of examiners between the clusters, and the five Scenarios P-T with I=50 and no clusters in the data with α_i having a linear structure. Figure 5.2 presents the posterior medians for σ_γ as well the 95% credible intervals for the 30 simulated scenarios. Note that the prior specification (5.11) assumes $\kappa_2 = 0$; in reality, the data has been simulated with a non-zero κ_2. Therefore, we also report the differences between the posterior median estimates for the distance between the cutpoints (κ_3^{est}-($\kappa_2^{\text{est}} = 0$)) and the true distance between the cutpoints ($\kappa_3 - \kappa_2$). Figure 5.2 demonstrates that the model parameters are reliably estimated.

We compare the clustering results obtained by our method against those obtained with k-means clustering (Hartigan and Wong, 1979). The k-means clustering is a technique used for partitioning continuous data into k clusters. The method works by finding k "centers" that define k clusters based on minimizing the Euclidean distances between the points in a cluster and their centers. For selecting the value of k we use silhouette clustering (Rousseeuw, 1987), which compares the distances between the points in a cluster against the distances between points in different clusters. The silhouette values are calculated for each cluster and range between -1 to 1, where 1 indicates perfect clustering and -1 indicates that the points belong to incorrect clusters. For k-means clustering, we start by centering and scaling our simulated data sets within a sample j. We fit the k-means model for $k = 3, 4, \ldots, 9$ for Scenarios A-O, and for $k = 3, 4, \ldots, 49$ for the linear α_i Scenarios P-T and select the k that maximizes the mean silhouette values across the obtained clusters.

In Tables 5.3 and 5.4, we report the misclassification rates (MCR, Rand, 1971) with different techniques, and the number of clusters obtained across the 25 simulated data sets (Scenarios A-O) in each scenario based on whether the consensus clustering matches the ground truth in the following way:

$$\text{Misclassification rate (MCR)} = \frac{c+d}{\frac{I(I-1)}{2}}$$

where, c = pairs of examiners in the same cluster that appear in different consensus clusters

d = pairs of examiners in different clusters that appear in the same consensus cluster.

The classifications are highly accurate with our model for Scenarios A-O. The number of clusters estimated with our method are generally accurate, but sometimes off by one. Most misclassifications occur in cases like Scenarios H and I where the cluster means are close to each other ($\alpha_1^* = -1.50$ and $\alpha_1^* = -0.75$). Even with the simulated data sets with an imbalance in the number of raters across clusters, the misclassifications rate is low and misclassifications happened in cases with clusters that were close to each other. Our method does much better compared to the baseline method of k-means clustering.

In Scenarios P-T in Table 5.2, where α_i are linear with increasing separation between the points, our method finds 4, 6, 8, 9, and 11 clusters respectively. The k-means clustering method finds 5, 7, 3, 3, and 3 clusters. We notice that our method finds more clusters as the separation between examiner effects increases although it still finds clusters when in reality there are no clusters in the underlying data set.

5.4 Experiments

We apply the methods described in Section 5.2 to three data sets: two from black-box studies conducted for latent fingerprint examination and handwriting comparison procedures; and

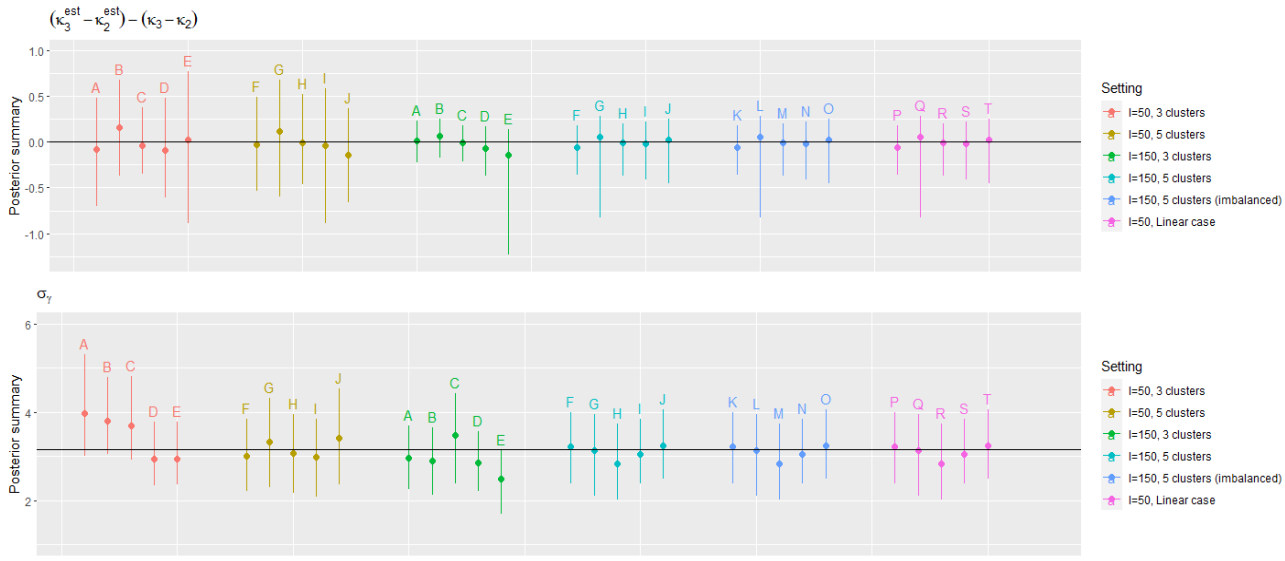


Figure 5.2: Posterior medians and 95% credible intervals for the difference between the true and estimated distance between cutpoints and for the parameter σ_γ obtained by fitting the model given by equations (5.6) on the data simulated in $I = 50, 150$ settings with 3 or 5 clusters in the underlying data, the case when there is an imbalance between the number of examiners/ raters in each cluster as well as the case where there are no clusters in the raters/ examiners. There were 5 simulated data sets in each setting. The black line represents the true value of the parameters.

Design	True no. clus	Proc.	Scenario	MCR	Scenario	MCR	Scenario	MCR	Scenario	MCR	Scenario	MCR	Mean MCR
I = 50	3	MDP	A	0.012	B	**0.027**	C	**0.000**	D	**0.052**	E	**0.000**	**0.018**
		k-means		**0.000**		0.076		0.112		0.091		**0.000**	0.056
I = 50	5	MDP	F	**0.016**	G	**0.060**	H	**0.137**	I	**0.143**	J	**0.071**	**0.085**
		k-means		0.171		0.265		0.180		0.326		0.179	0.224
I = 150	3	MDP	A	**0.000**	B	**0.000**	C	**0.034**	D	**0.000**	E	**0.000**	**0.006**
		k-means		0.054		0.026		0.094		0.152		0.254	0.116
I = 150	5	MDP	F	**0.005**	G	**0.029**	H	**0.140**	I	0.161	J	**0.056**	**0.078**
		k-means		0.171		0.242		0.177		**0.153**		0.219	0.192
I = 150 (imbalanced)	5	MDP	K	**0.013**	L	**0.048**	M	**0.144**	N	0.177	O	**0.083**	**0.093**
		k-means		0.066		0.097		0.156		**0.161**		0.240	0.144

Table 5.3: The misclassification error rates (indicated by MCR) are presented for each scenario. The column "Proc." indicates the method used to fit the data: mixtures of Dirichlet processes (MDP, our method) or k-means clustering. Average misclassification is presented in the last column. The better results for each scenario are highlighted with bold text.

Design	True no. clus	Proc.	Scenario	no. clus	Scenario	no. clus	Scenario	no. clus	Scenario	no. clus	Scenario	no. clus
I = 50	3	MDP k-means	A	4 **3**	B	**3** **3**	C	**3** 7	D	**3** 6	E	**3** **3**
I = 50	5	MDP k-means	F	**5** 3	G	**5** 3	H	7 7	I	4 3	J	**5** 3
I = 150	3	MDP k-means	A	**3** 4	B	**3** **3**	C	**3** **3**	D	**3** 8	E	**3** **3**
I = 150	5	MDP k-means	F	**5** 3	G	**5** 3	H	4 4	I	**4** **5**	J	**5** 3
I = 150 (imbalanced)	5	MDP k-means	K	6 3	L	**5** 3	M	4 3	N	4 **5**	O	**5** 4

Table 5.4: The number of clusters obtained through the techniques are presented for each scenario. The column "Proc." indicates the method used to fit the data: mixtures of Dirichlet processes (MDP, our method) or k-means clustering. The correct results for each scenario are highlighted with bold text.

one psychological study of maternal depression and its impact on child development.

5.4.1 Latent Fingerprint Examination

Reliability and accuracy for feature-based comparison disciplines in forensic science such as fingerprint examination, footwear comparison, bloodstain pattern analysis, and handwriting comparisons is measured through black-box studies. In a typical black-box study, examiners from various agencies, that may have different standards for forensic analyses, are asked to analyze forensic samples with known ground truth. Forensic experts provide assessments of the samples, based on the standards/ decision scale provided by the designers of the study, just like they would in real casework. The steps taken by an examiner to reach a specific conclusion are not defined and hence the decision-making process is treated like a "black box". In the black-box studies conducted so far, data is not collected in a way that allows examiner performance to be connected to their characteristics. This is due to the protections provided in IRB protocols.

The accuracy or validity is defined by the correctness of the decisions. Reliability is defined by the consistency of the decisions. The forensics community is interested in two types of reliability; reproducibility and repeatability. Reproducibility or inter-examiner reliability is the consistency of decisions across examiners for the same forensic evidence sample. Repeatability is the consistency of decisions made by the same examiner for the same forensic evidence sample at different points in time. Previously, Arora et al. (2022) and Arora et al. (2023) have proposed statistical methods to assess the reliability and variability in decisions that are collected in black-box studies. Our method provides another way to model variability in decisions by encouraging parameter sharing among examiners in a cluster.

The Forensic Bureau of Investigation (FBI) conducted the first large-scale black-box study to establish a scientific foundation for latent fingerprint examinations (Ulery et al., 2011, 2012). The study included 169 examiners from different laboratories all across the United States participated in this study. There were a total of 356 latent prints used in the study design. These were used to develop 744 latent print-exemplar pairs (520 known mates, 224 non-mates). Each examiner was presented with about a hundred latent print-exemplar pairs and they were asked to conduct the examination just as they would in real casework.

Latent print examination is typically a multi-step process and most laboratories follow the ACE-V procedure. The steps are analysis, comparison, evaluation, and verification. More specifically, they involve: Analysis, when the latent print is analyzed for quality; Comparison, an exemplar print is presented if the latent print is deemed suitable for comparisons; Evaluation, examiners present conclusions based on all features observed during comparison; Verification, some agencies present the latent print-exemplar pair independently to another examiner for re-examination. Examiners were asked to provide their conclusion for the analysis phase using the following ordinal categories: Value for Individualization (VID) when the latent print has enough features to support an individualization decision; Value for Exclusion Only (VEO) when the latent print has enough features to support an exclusion

but not an identification; and No Value (NV) when the latent print does not have sufficient information to carry out a comparison. The evaluation phase also has a three-point ordinal decision scale as well; Individualization, when the examiner believes that the latent print and exemplar come from the same source; Exclusion, when the examiner believes that the latent and exemplar come from different sources; and Inconclusive, when the examiner is not able to reach either of the other conclusions. We think of these as being ordered as Exclusion, Inconclusive, and Individualization.

5.4.1.1 Clustering Examiners

We are interested in whether there exist clusters of examiners that make similar decisions while accounting for variation in the samples examined. As a first step, we focus on conclusions in the analysis stage. The decisions were on an ordinal scale of NV, VEO, and VID. There were 169 examiners that participated in the study and 356 distinct latent prints in the data. We fit the model with interactions given in the equations (5.7). We found that the posterior median for σ_γ was 3.46 with 95% credible interval of (3.14, 3.82), the posterior median for σ_ζ was 0.54 with 95% credible interval of (0.43, 0.66), and the posterior median for κ_3 was 1.90 with a 95% credible interval (1.76, 2.06). We observe that the interaction variation is much smaller compared to the latent print variance. Note it is difficult to interpret κ_3.

We present the inference for the examiner clusters through the consensus clusterings obtained using the approach described in Section 5.2.5. We found that there were 7 clusters of examiners in the data set based on their quality assessments. Figure 5.3 presents the distribution of percentages of decisions that are Value for Individualization (VID) assigned by examiners in different consensus clusters. Similarly, Figures 5.4 and 5.5 present the percentages of decisions that are Value for Exclusion Only (VEO) and No Value (NV) respectively. We observe a clear difference in the frequency of use for the categories across the clusters, especially VID and NV categories. Figure 5.6 presents the combined frequencies of NV and

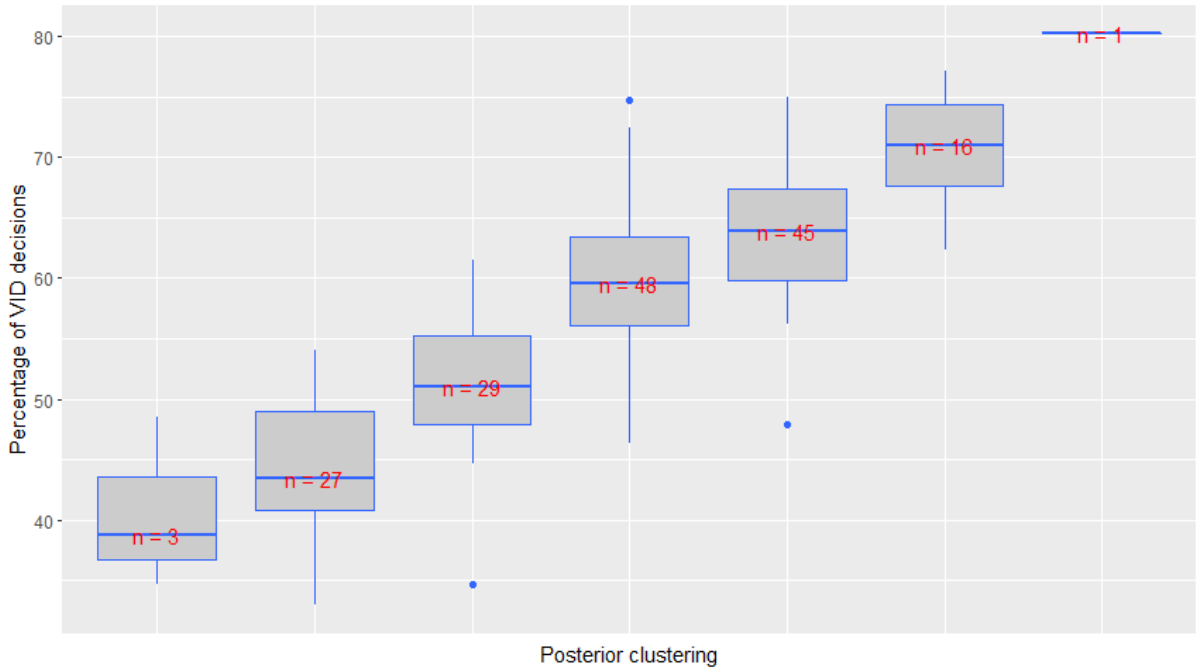


Figure 5.3: Differences in percentages of Value for Individualization (VID) decisions provided by the examiners in different consensus clusters.

VEO decisions across consensus clusters.

Figure 5.7 is a heatmap of the analysis decisions of examiners across the samples. The examiners (rows) are grouped by their consensus clusters indicated by the grayscale colors on the left vertical axis of the plot, which are in turn ordered by the mean decisions within a consensus cluster. The samples (columns) are ordered by the average decisions on the sample increasing from left to right. The NV decisions are in red, VEO decisions are in green and VID decisions are in blue. The top cluster has the most NV decisions and the least VID decisions. The clusters below show fewer NV decisions and more VID decisions. There is still variability within clusters.

Figure 5.8 is a comparison of decisions within a cluster against all the other clusters on ten randomly chosen latent prints. Random noise has been added to the decisions to be able to visualize how the ratings received by a sample in the same cluster are more similar compared to the ratings across all other clusters.

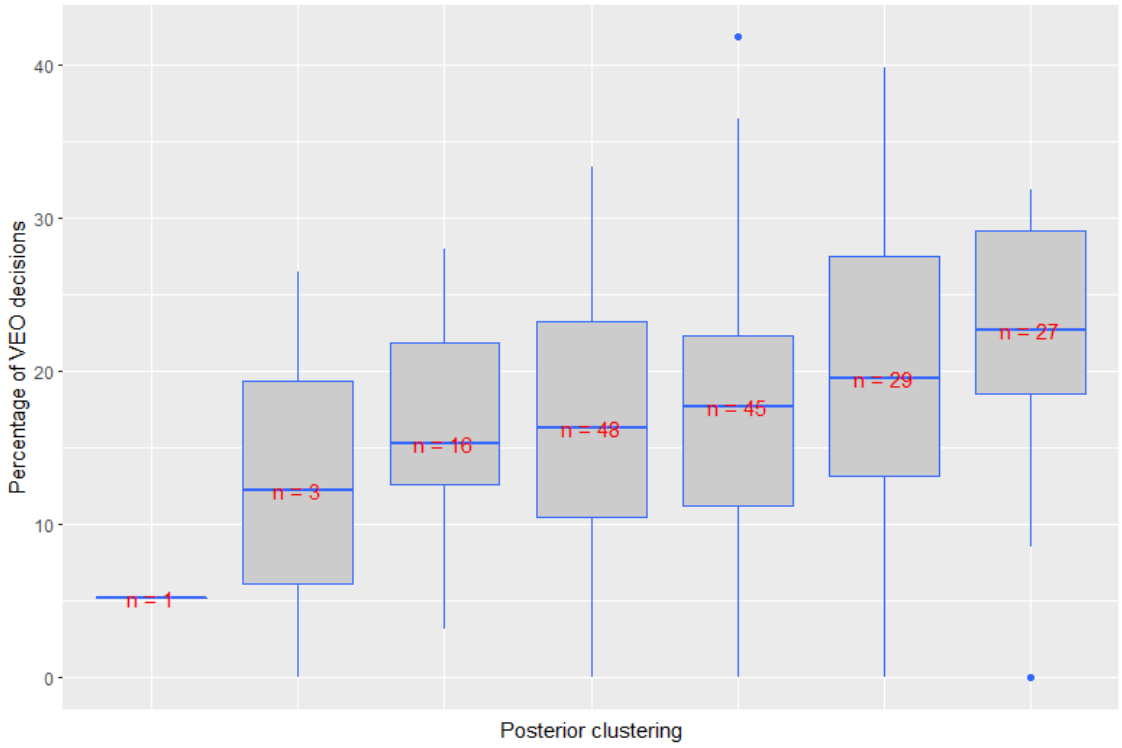


Figure 5.4: Differences in percentages of Value for Exclusion Only (VEO) decisions provided by the examiners in different consensus clusters.

One hypothesis is that assessment will be more consistent (reliable) within clusters. We explore this using average percentage agreement as reported in Ulery et al. (2012). It is calculated as follows: let n_{jm} denote the number of times sample j was placed in category m and $n_{j.}$ denote the total decisions on sample j. The $\bar{P}_j$, the percentage agreement for sample j is:

$$\bar{P}_j = \frac{1}{n_{j.}(n_{j.} - 1)} \sum_{m=1}^{M} n_{jm}(n_{jm} - 1)$$

and the average percentage agreement across samples is $\bar{P}$:

$$\bar{P} = \frac{1}{J} \sum_{j=1}^{J} \bar{P}_j.$$

Percentage agreement ranges from 0 to 1 and is a method to evaluate the reproducibility of decisions across raters. We expect that within the consensus clusters, the reproducibility

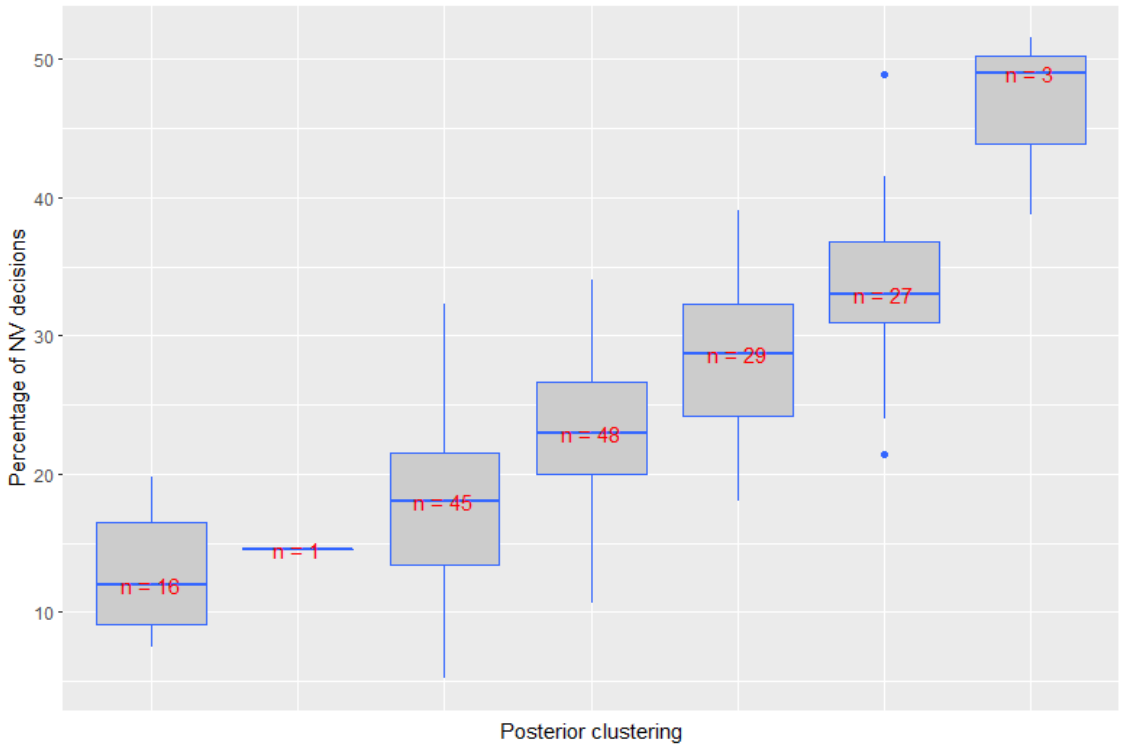


Figure 5.5: Differences in percentages of Value for No Value (NV) decisions provided by the examiners in different consensus clusters.

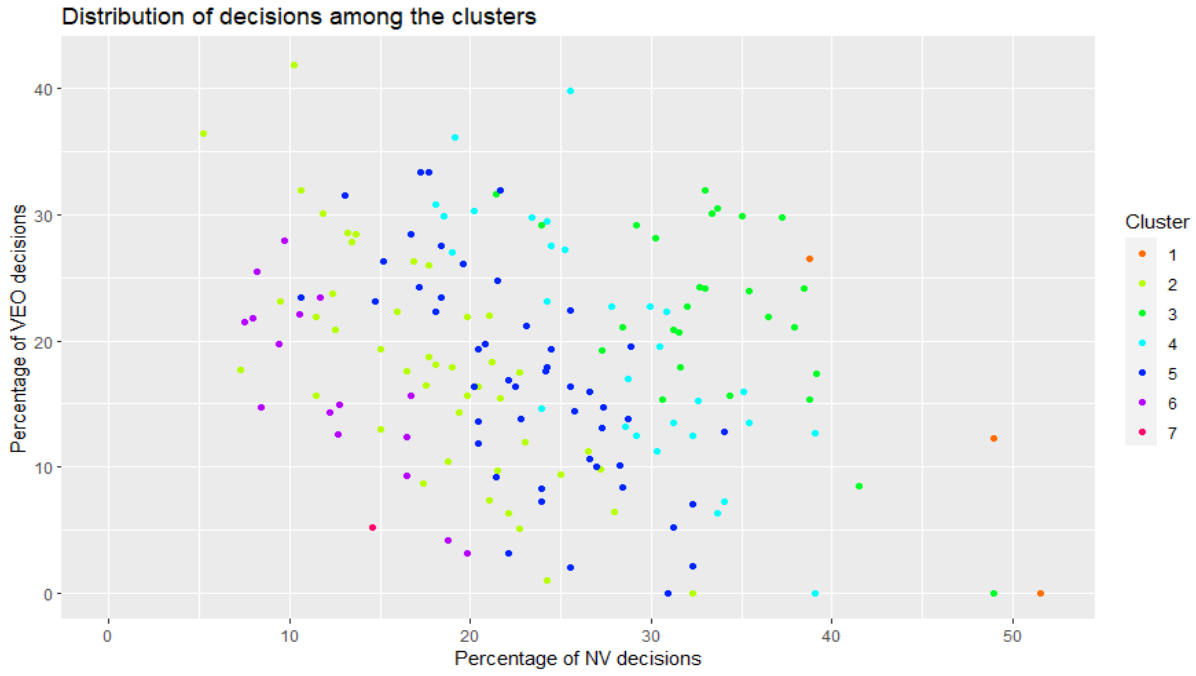


Figure 5.6: Distribution of percentages of (No Value indicated as NV, Value for Exclusion Only indicated as VEO) decisions provided by the examiners in different consensus clusters that are indicated by the different colors.

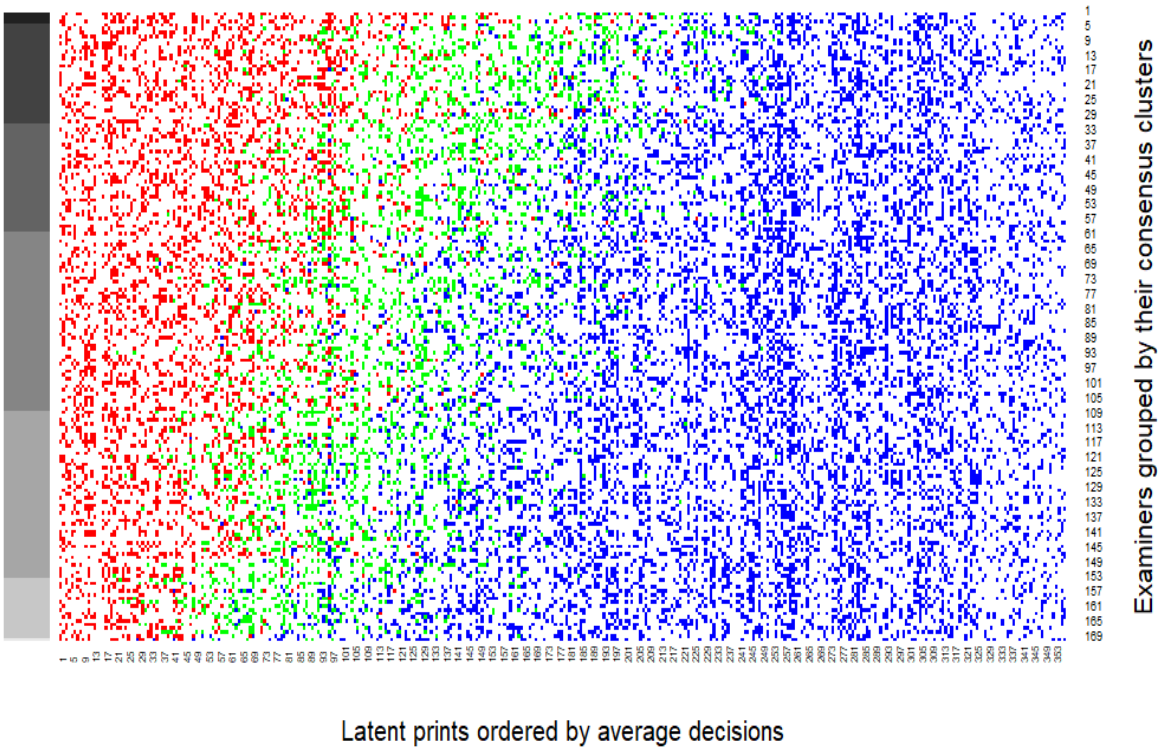


Figure 5.7: Heatmap of analysis decisions is presented. Red indicates NV decisions, green indicates VEO, and blue indicates VID decisions. The examiners (rows) are grouped by the consensus clusters indicated by the grayscale colors in the left vertical axis in the plot. The prints (columns) are ordered by the average decision on the print with NV=1, VEO=2, VID=3.

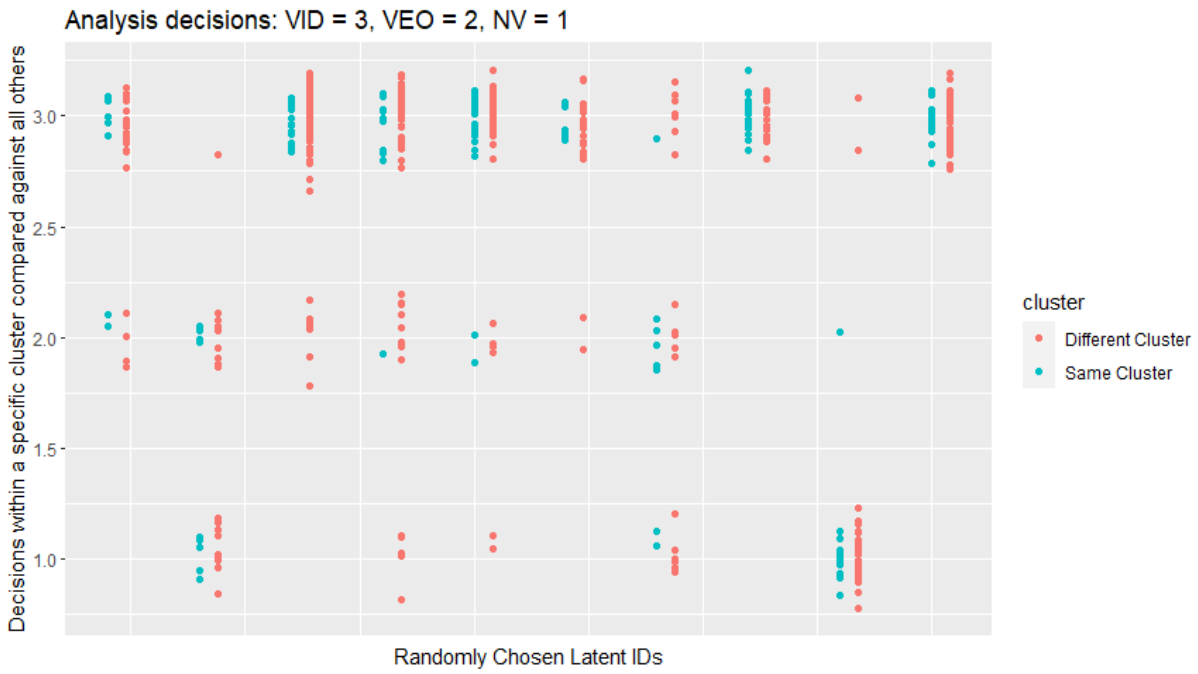


Figure 5.8: Decisions compared within a cluster (blue) against decisions across all other clusters (pink) on ten randomly chosen latent prints.

of decisions will be higher when compared to the overall reproducibility. The percentage agreement across all latent prints is 0.76. As expected, the percentage agreement within clusters is higher with values 0.77 (n=27), 0.80 (n=29), 0.80 (n=48), 0.83 (n=45), 0.86 (n=16) for the five sizeable clusters in Figure 5.3. It is also interesting to see if these clusters are meaningful outside of the analysis stage of decision-making. For example, do examiners within a cluster also make evaluation decisions similarly? We analyzed the reproducibility of the evaluation decisions of examiners within these clusters and compared it to the overall reproducibility. The overall reproducibility, assessed through percentage agreement, for the evaluation decisions was 0.76 on a total of 13174 decisions on the scale of Exclusion, Inconclusive, and Individualization. The reproducibility within 5 of the 7 clusters (2 clusters have < 5 examiner each) was 0.77 (n=27), 0.79 (n=29), 0.78 (n=48), 0.78 (n=45), 0.78 (n=16). The percentage agreement for evaluation decisions within the clusters of examiners that make similar quality determinations is a bit higher compared to the overall agreement for evaluation decisions though the differences are relatively small.

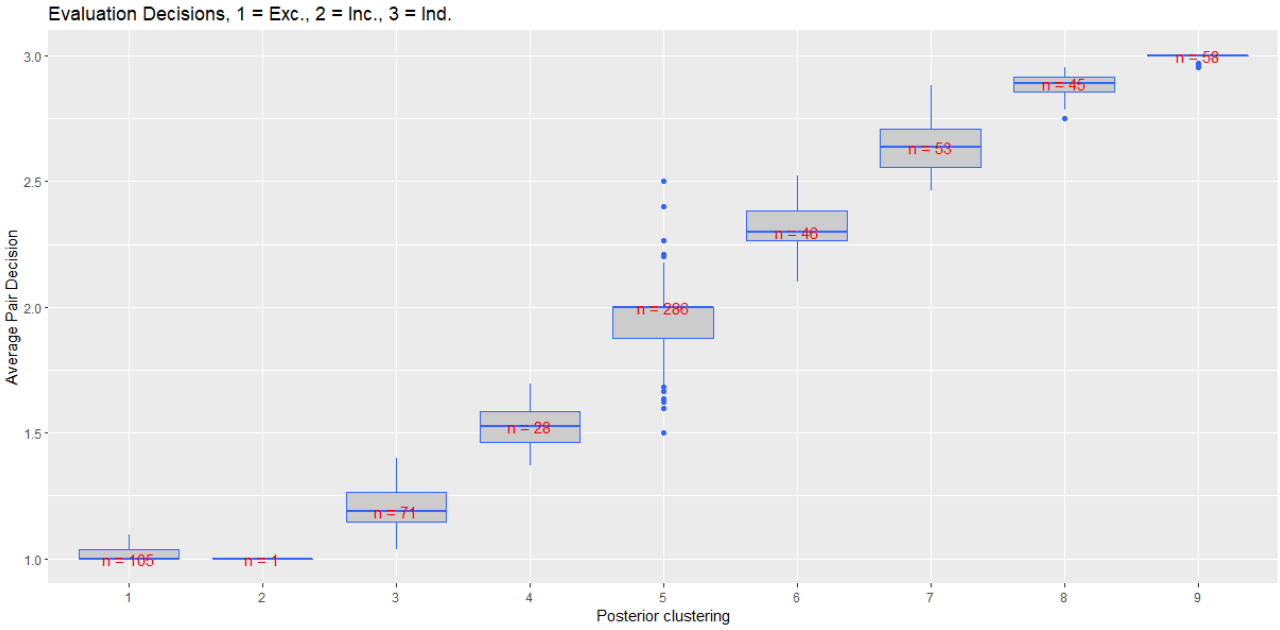


Figure 5.9: Average pair decisions plotted against posterior clustering. As expected, the latent-exemplar pairs are being clustered based on their tendencies to receive decisions.

5.4.1.2 Clustering Samples

So far we have discussed clustering raters based on their tendencies to rate samples in similar ways. We are also able to cluster items or samples based on their tendencies to be rated in the same way by examiners. We cluster the 744 latent-exemplar pairs in the FBI data set based on their tendency to receive Exclusion, Inconclusive, or Individualization decisions with the model (5.6). The truncation value for the stick-breaking procedure was chosen to be $T = 100$ because it may be difficult to draw out meaningful observations if there are too many clusters.

The posterior median for σ_α was 0.28 with a 95% credible of (0.23, 0.33). The examiner variation is small which is expected due to the fact that the latent-exemplar pairs will have a lot more variation compared to the variation in trained and practicing examiners. The posterior median for κ_3 with the 95% credible interval was 2.35 (2.11, 2.56),

There were 9 clusters of latent-exemplar pairs in the consensus clustering based on the poste-

Cluster index	No. of mates	No. of non-mates
1	4	101
2	1	0
3	7	64
4	4	24
5	251	35
6	46	0
7	53	0
8	45	0
9	58	0

Table 5.5: Distribution of mated and non-mated pairs within consensus clusters.

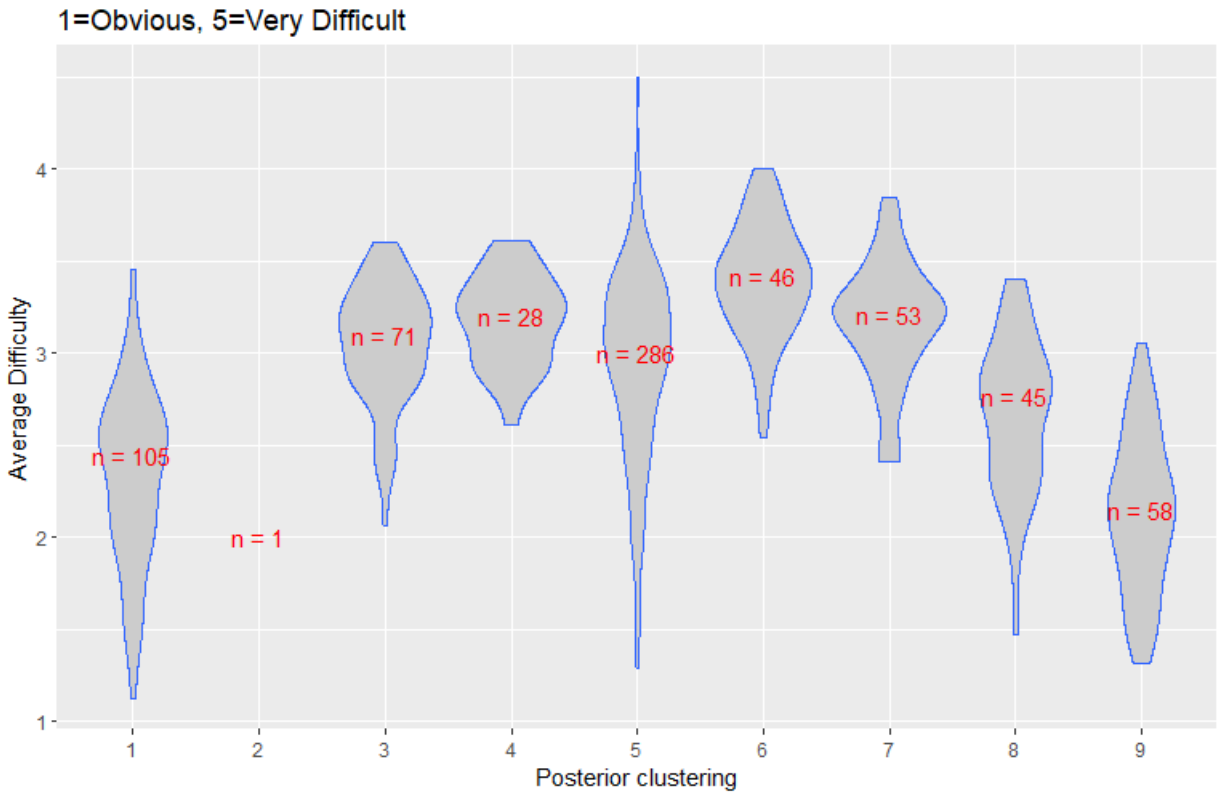


Figure 5.10: Average examiner-reported difficulty of the comparison decision plotted against the posterior clusters.

rior samples. Figure 5.9 plots the spread of the average evaluation decisions received by the pairs in the clusters. Cluster 5 seems to consist primarily of mated pairs that mostly receive Inconclusive decisions (see Figure 5.9). Categorizing inconclusive decisions is a controversial issue in assessing examiner proficiency and reliability; it is unclear whether they should be a separate category, discarded, or treated as errors (Scurich, 2022). One suggestion has been to identify samples for which there is a consensus that Inconclusive is the best answer. Cluster 5 in Figure 5.9 may provide a relevant set of examples. Inconclusives in these samples could be interpreted as correct, while Inconclusives in other clusters as incorrect.

Table 5.5 presents the ground truth for the samples across the consensus clusters. Clusters 2, 5, 6, 7, 8, and 9 in Table 5.5 seem to have mostly mated pairs, and Clusters 1, 3, and 4 seem to have mostly non-mated pairs.

Figure 5.10 presents the average examiner-reported difficulty for the comparisons across the samples in the clusters. There is a lot of variation in the type of conclusions received in Clusters 4, 5, 6, and 7 in Figure 5.9 and they also receive higher examiner-reported comparison difficulties as seen in Figure 5.10.

5.4.2 Handwriting Comparisons

The reliability and accuracy of handwriting comparisons have been previously studied by many authors (Durina and Caligiuri, 2009; Kam et al., 1997, 2001; Kam and Lin, 2003; Kam et al., 1994; Mitchell, 2016). Handwriting comparisons also follow the ACE-V procedure described in subsection 5.4.1. The questioned handwriting sample should ideally contain sufficient quality and quantity of information to qualify for comparisons and evaluations with exemplar documents.

Hicklin et al. (2022) described the results from a black box study conducted to establish a scientific foundation for handwriting comparison decisions. This study was conducted

Parameters	κ_3	κ_4	κ_5	σ_γ	σ_ζ
Handwriting Comparisons	1.47 (1.42, 1.53)	2.26 (2.18, 2.33)	3.34 (3.26, 3.43)	1.53 (1.37, 1.71)	0.77 (0.66, 0.91)

Table 5.6: Results from fitting the model (5.7) to the comparison decisions in handwriting black-box study with posterior medians for parameters and 95% credible intervals.

in a similar manner to the FBI latent fingerprint examination study (Ulery et al., 2011). Eighty-six forensic document examiners from federal, state, and local agencies participated in the study. Each examiner was assigned about 90 distinct questioned and known (QK) sets (each set included a questioned sample and a known sample) from among 180 distinct pairs that were prepared. Examiners carried out the analysis over a 10-month period. Ten of the 90 QK sets were re-assigned to each examiner so that repeatability could be assessed. Thus we have a total of 100 assessments for each examiner. Examiners were asked to make assessments on a five-category ordinal scale: *Written* when the QK set is believed to come from a single writer *ProbWritten*, when the examiner believes that the QK set was probably written by a single writer; *NoConc*, when the examiner is not able to make a decision either way; *ProbNot*, when the QK set is believed to probably not have been written the same writer; *NotWritten*, when the QK set is believed to have different writers.

We cluster the 86 examiners based on their tendencies to rate samples based on the model (5.7) with interactions. Table 5.6 presents posterior estimates of the parameters that are not related to examiner clusters. We observe that there are some interactions between examiners and handwriting samples as was observed in Arora et al. (2023).

Figure 5.11 provides the consensus clustering of the examiners based on the posterior draws. There are six clusters in the data set with three clusters having one examiner each. Figure 5.11 demonstrates that the examiners in Clusters 4, 5, and 6 possibly provide a lot of *NoConc*, *ProbWritten*, or *Written* conclusions compared to other examiners. The other clusters show a clear tendency to rate samples similarly, assuming the type of pairs assigned to each examiner is balanced.

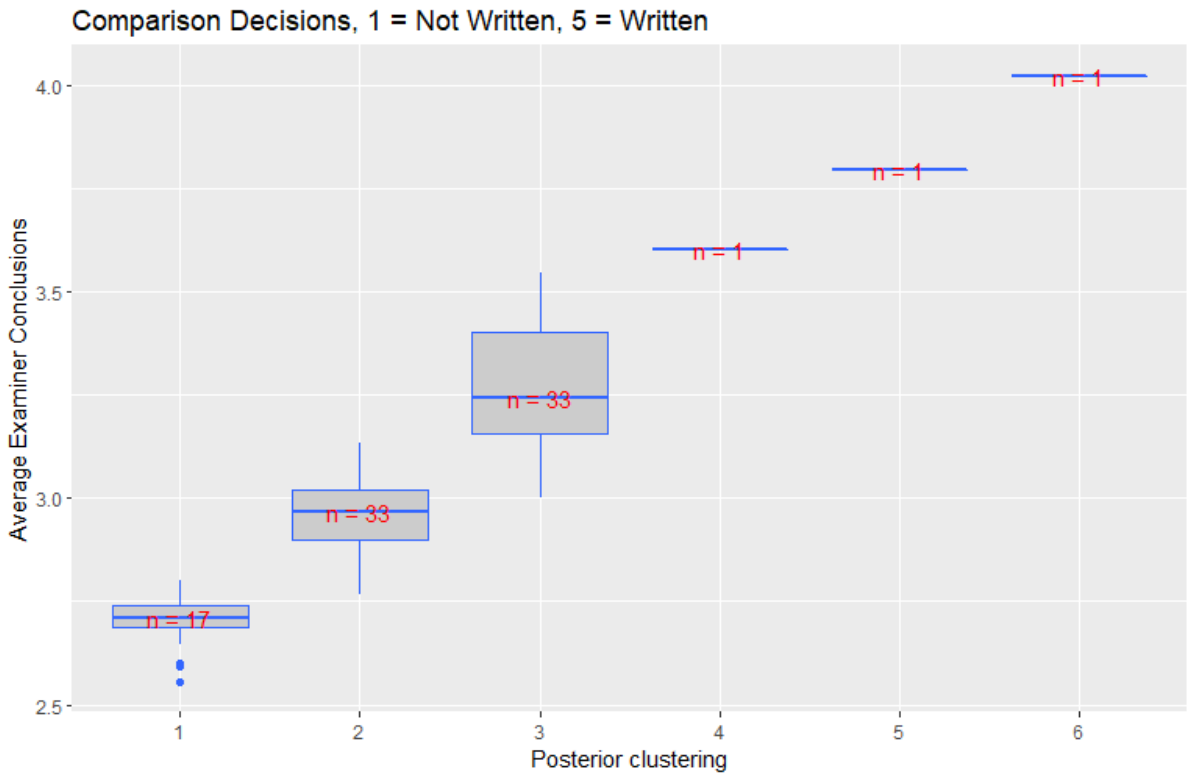


Figure 5.11: Average examiner conclusions plotted against posterior clusterings.

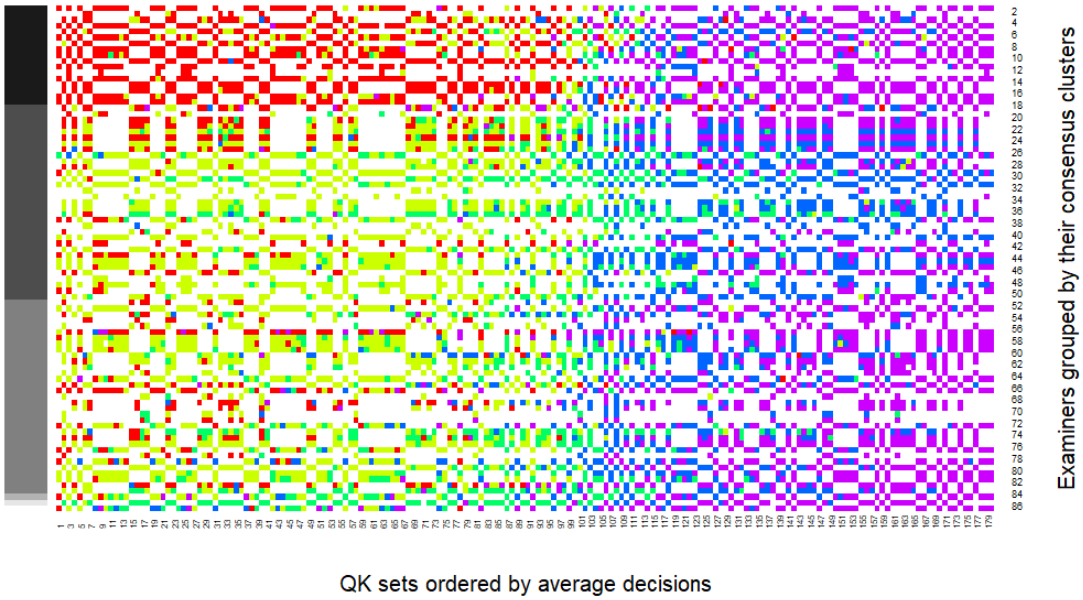


Figure 5.12: Heatmap of decisions across QK sets are shown for examiners in consensus clusters, indicated by the grayscale colors on the left vertical axis. The QK sets (columns) are ordered in increasing order of average decisions provided on the QK set. Red indicates *NotWritten*, yellow indicates *ProbNot*, green indicates *NoConc*, blue indicates *ProbWritten*, and violet indicates *Written* decisions. Clusters are also ordered by their average decisions.

Figure 5.12, is a heatmap of the decisions on the QK sets (columns) within and across the consensus clusters indicated by the grayscale vertical axis on the left of the heatmap. The columns are ordered by average decisions in an increasing order and the examiners are grouped by consensus clusters which are ordered by average decisions. Red indicates *NotWritten*, yellow indicates *ProbNot*, green indicates *NoConc*, blue indicates *ProbWritten*, and violet indicates *Written* decisions. Clusters are also ordered by their average decisions.

Although we do not have covariate information for raters, we further explore the clusters based on the information provided in Hicklin et al. (2022). They indicate that examiners that had more than 2 years of formal training (73% $\approx$ 62 examiners) made less definitive conclusions compared to examiners that had less than 2 years of formal training (27% $\approx$ 24 examiners). However, the examiners with more than 2 years of formal training also made more accurate decisions on the samples for which they made definitive conclusions compared to the examiners with less than 2 years of formal training. In Figure 5.13, we have plotted the most frequent decisions provided by examiners in a cluster where 0 indicated more definitive statements such as *Written* and *NotWritten*, 1 indicated probabilistic statements such as *ProbWritten* and *ProbNot*, 2 indicated *NoConc* decisions. Figure 5.13 indicates that the first cluster (n=17) made more definitive decisions, the pattern expected by less experienced examiners. Clusters 2 and 3 (total n=66) tend to use probabilistic conclusions more often.

5.4.3 Maternal Depression Data

We now apply the method to the data from an application in psychology. The Conte Center at the University of California, Irvine aims at discovering the effects of early life adversity on cognitive and emotional development in infants across species. In one study, they examined the effects of maternal mood on a child's mental health (Glynn et al., 2018). The maternal

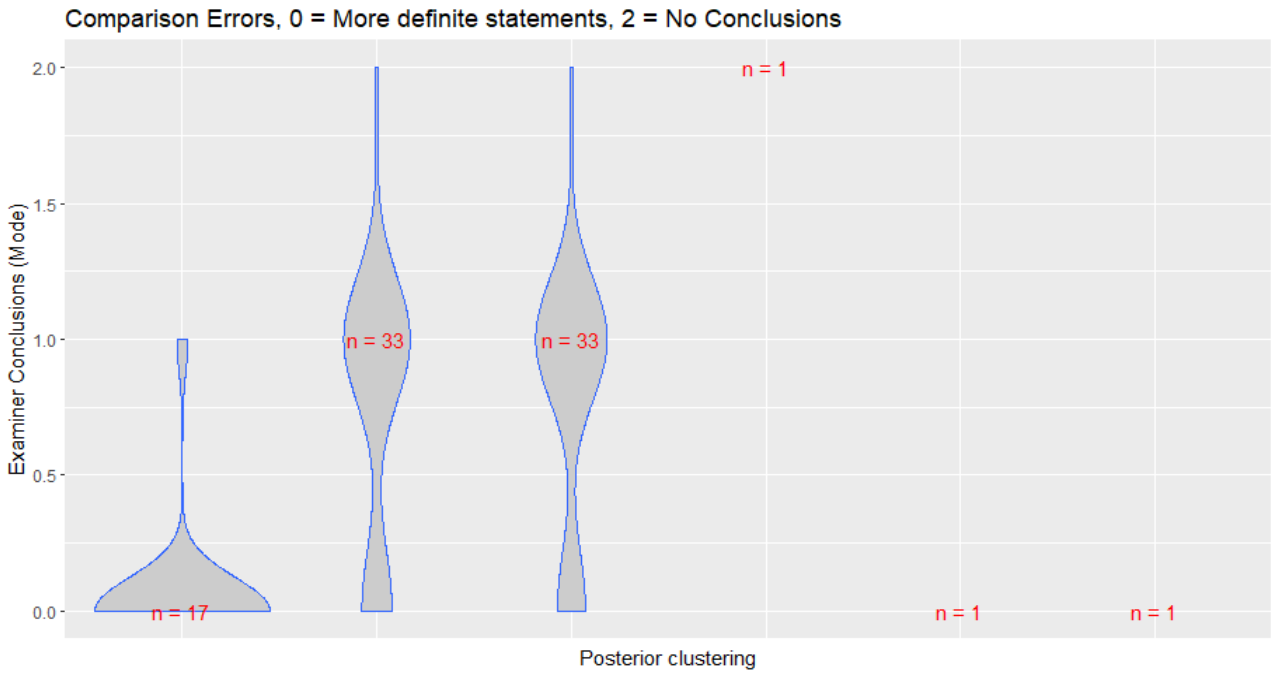


Figure 5.13: Tendency to make probabilistic statements in different clusters through examiner modes.

mood was assessed in 934 mothers who were asked to fill the Center for Epidemiologic Studies Depression Scale Short Form (CES-D SF) questionnaire (Radloff, 1977; Santor and Coyne, 1997; Glynn et al., 2018). There are 9 questions that track depressive symptoms of mothers on a Likert scale of 0-3 where 0 indicated no presence of depressive symptoms such as lack of happiness, restless sleep, etc.; 1 indicated feeling the symptom sometimes; 2 indicated feeling the symptom occasionally; and 3 indicated feeling the symptoms all the time. Additionally, covariates such as household income, household income to needs ratio, marital status of the mother, and education level were also collected. This data set is interesting for us because it provides an opportunity to cluster respondents and see if the clusters are related to measured covariates. However, note that this data is different from the data in the forensic black-box study setting because all mothers rated the same items which are aimed at assessing depression.

We use the model (5.6) to fit these data. Note the slight change in category notation that

range from $0-3$ instead of $1-4$ and the notation for cutpoints κ_m accordingly dictates that: $\kappa_0 = -\infty, \kappa_1 = 0$. Again, with the aim of extracting meaningful clusters in the data, we truncated the stick-breaking process at $T = 100$. The posterior median for the standard deviation of question effects, σ_γ, has the posterior median of 0.47 with a 95% credible interval (0.30, 0.84). Note that the estimated standard deviation of question effects is very low which can be explained by the fact that the questions assess very similar traits in the mothers, for example, how depressed they felt or whether they had difficulty enjoying life. The posterior median for κ_2 (cutpoint between category 1 and 2) is 1.27 with 95% credible interval (1.23, 1.3); κ_3 (cutpoint between category 2 and 3) has the posterior median estimate of 2.30 with a 95% credible interval (2,22, 2.35).

Next, we look at the consensus clustering of the mothers. Figure 5.14 presents the distribution of the percentage of questions that were answered with a 3 on the Likert scale, which indicates feeling depressive symptoms all the time, across the clusters. We notice that there is a clear difference between the frequency with which the mothers in different clusters respond with a 3 to the questions. Similarly, Figure 5.15 plots the distribution of the percentage of questions that were answered with a 0 on the Likert scale, which indicates no depressive symptoms, across the clusters. We observe a difference between clusters similar to that seen in Figure 5.14.

We further investigate whether the posterior clusters have associations with underlying covariates. We hypothesize that the clusters may be related to covariates such as household income and the marital status of the mother. Figure 5.16 plots the average log of household income across the clusters. We see that there is almost a decreasing trend in the mean of the log household income. Medians of log household income across the clusters were 11.08, 10.92, 10.71, and 10.71 respectively in the clusters. We also conducted an ANOVA test for the log household income across the clusters and it was statistically significant with a p-value of 0.0019.

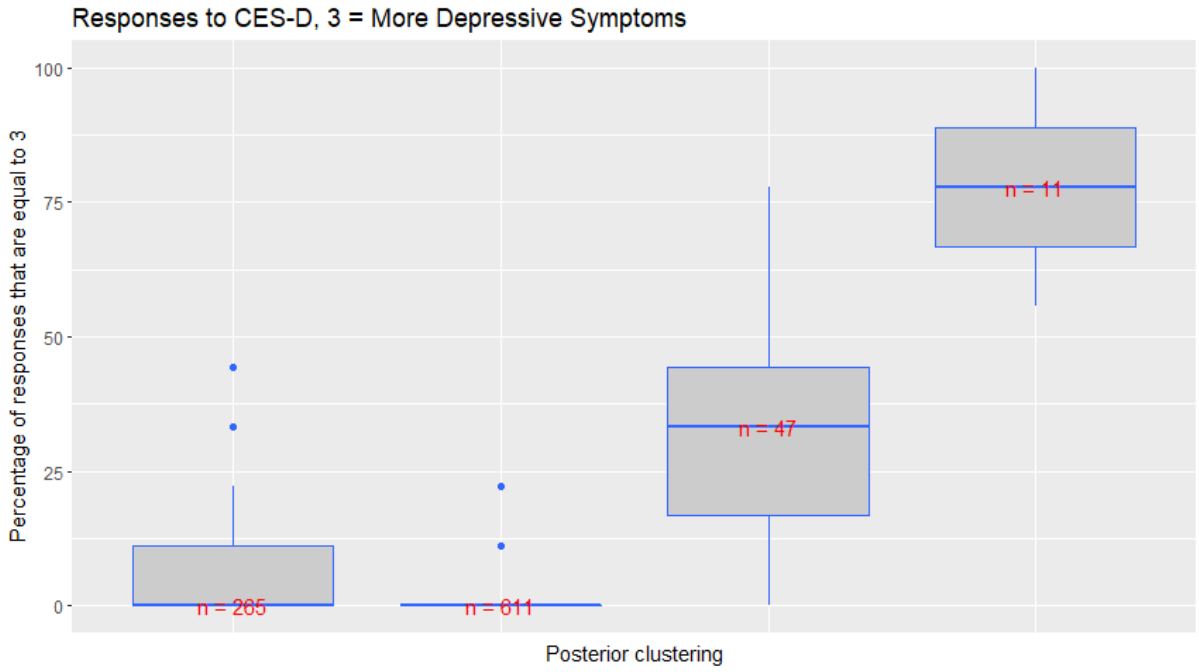


Figure 5.14: Distribution of the percentage of questions that were answered with a 3 across clusters. On the Likert scale, 3 indicates feeling depressive symptoms all the time.

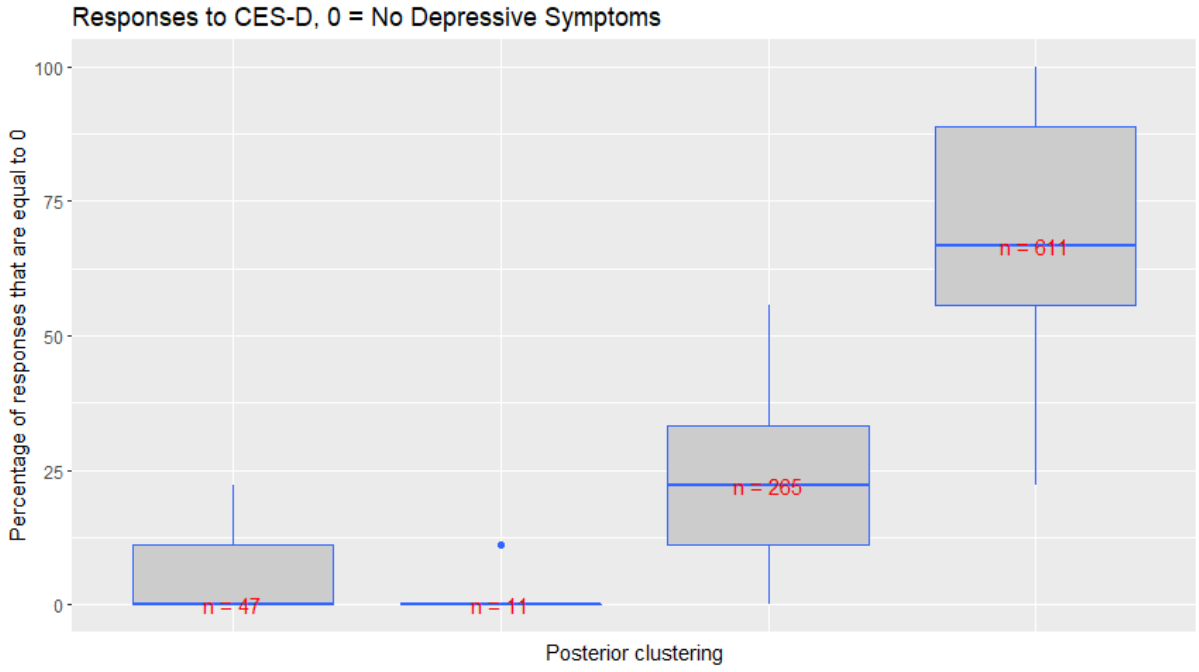


Figure 5.15: Distribution of the percentage of questions that were answered with a 0 across clusters. On the Likert scale, 0 indicates feeling no depressive symptoms.

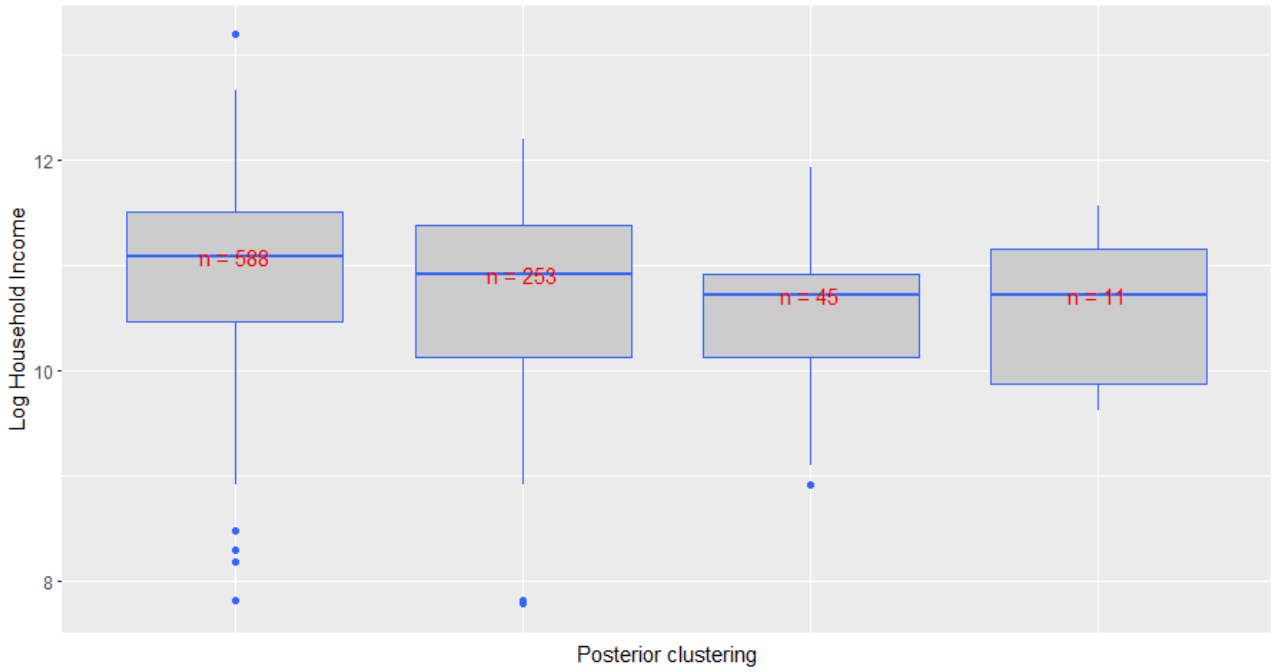


Figure 5.16: Average of log household income across the posterior clusters. We did not have the household income for all mothers which is why not all 934 mothers are included in this plot.

5.5 Conclusions

We have presented a method to cluster raters based on their tendencies to classify items of varying "difficulties" on an ordinal scale. Our approach is able to adjust for different examples while clustering raters. We deploy an MDP model, which encourages parameter sharing between raters and has the advantage of not needing to pre-specify the number of clusters in the data set. This method may also be used to cluster items. We demonstrated in Section 5.3 that our proposed method is able to correctly place examiners in clusters even when there is an imbalance between the number of examiners in each cluster. Most misclassifications were in cases when the clusters were not well separated from each other.

Our method has several applications: we are able to identify clusters of raters that make decisions similarly, we are also able to generate further hypotheses based on this exploratory technique. Another application of our method is using clusters of questions/items (based on

their tendencies to receive similar decisions) to inform future study design. Our approach was demonstrated with the experiments in Section 5.4 with latent fingerprint comparisons, handwriting comparisons, and maternal depression data. In the examples, we gained insights about the raters and items through the consensus clusterings based on the ordinal ratings. For example, we found in the FBI study that the tendencies shared by examiners for the analysis stage of latent fingerprint examination might generalize to their tendency to make comparison decisions. We also clustered latent-exemplar pairs based on their tendencies to be assessed similarly by examiners. Typically, it is tricky to evaluate whether inconclusive decisions in black-box studies are errors or should be discarded before assessing examiner proficiency. Our approach may provide a relevant set of examples that have a consensus of Inconclusive decisions. The inconclusive decisions within these examples could be interpreted as "true inconclusives" and inconclusive decisions on other examples may be considered as incorrect. We hope that the findings in this paper are able to motivate future studies to collect examiner and sample covariates. Forensic black-box studies are an integral technique for assessing the reliability and validity for subjective forensic examination procedures. Typically, the reliability and validity are reported as an aggregate across all examiners and samples. However, as previously observed in Hicklin et al. (2022), the accuracy of decisions depends on examiner covariates such as years of formal training. Additionally, it is expected that reliability on samples that are more difficult to assess must be lower than for the easier assessments. Previously, Arora et al. (2022, 2023) have presented a method to model variability in forensic black-box studies. Our model is another way to model variability while accounting for possible interactions. In the future, we could bicluster based on both raters as well as items; similar work that has been done in other settings (Rost, 1990; Duong, 2013; Guo and Kwok, 2016).

Chapter 6

Latent Factor Analysis for Binomial Data with Applications to DNA Methylation Data

6.1 Introduction

DNA methylation profiles of CpG islands (Moore et al., 2013) are known to be altered as a result of the environment (Smith et al., 2020; Katrinli et al., 2022). However, the variation in DNA methylation that is associated with changes in the environment is highly confounded by inter-individual variation and variation in epigenome due to age and tissue (Hüls and Czamara, 2020; Czamara et al., 2021). Exploratory dimension reduction techniques are often used for analyzing epigenetic data due to the high-dimensional nature of such data (Richardson et al., 2016). For example, Jiang et al. (2019) have used principal component analysis (PCA) for analysis of DNA methylation data in rats to differentiate between pups that were exposed to early life adversity and the control group. Short et al. (2023) also used PCA in DNA methylation data for human infants. However, it was observed that the first

several principal components did not explain a significant portion of the data.

We posit that part of the reason that PCA explains limited variation in human data is due to the quality of the samples and the limitations in sequencing technologies that entail that not all subjects have sufficient reads for all CpG sites. Additionally, PCA aims to extract lower dimensional representations of the data that are obtained with linear combinations of the features. This method may be restrictive because it does not explicitly account for measurement errors; methods such as factor analysis (FA) although similar to PCA (Tipping and Bishop, 1999) may have an advantage over PCA by accounting for measurement error (Harman, 1976; Kim et al., 1978; Rummel, 1988).

In this paper we introduce a Binomial Latent Factor Analysis (BLFA) model, an exploratory method that accounts for variation in the number of reads through a binomial distribution; the true proportions of methylation profiles are further assumed to depend on a small dimensional space through a factor analysis (FA) model. We encourage sparse representations of the factor analysis model with a spike-and-slab prior on the elements of the factor loading matrix (George and McCulloch, 1997). The Binomial Latent Factor Analysis (BLFA) model has several contributions: we are able to account for heterogeneity in sample sizes and account for measurement variation; our model is also able to extract sparse representations of the factor loading matrix so that the pathways that contain the methylated sites that explain more variation in the data can be deduced.

This paper is structured in the following way. We begin Section 6.2 with a description of the typical DNA methylation data setup that motivated the proposed method. We also describe the variation in the reads obtained across the CpG sites and across subjects and how that can affect the statistical inference of the data. Section 6.3 begins with a discussion of factor analysis models followed by the Binomial Latent Factor Analysis (BLFA) model and the algorithm used for fitting the method. In Section 6.4, we present some results from applying the BLFA model to simulated data sets and compare the BLFA method with a

baseline model. This discussion is followed by applying our model in Section 6.5 to a data set collected from 107 infants that had different early life experiences. Section 6.6 presents a discussion and proposes future directions for this work.

6.2 DNA Methylation Data

We briefly describe the data setup for methylation data. DNA samples are collected from different subpopulations of subjects that share characteristics such as disease status or environmental factors such as early life adversity. DNA samples are processed using methylation sequencing techniques such as the reduced representation bisulfite sequencing (RRBS) technique. Denote sites along the DNA as $j = 1, 2, \ldots, d$ and subjects by $i = 1, 2, \ldots, N$, then let n_{ij} be the reads obtained for subject i on site j and let y_{ij} be the methylated reads within the n_{ij} total reads. The collected data can be arranged in two matrices of size $d \times N$; one matrix can contain the reads n_{ij} and the second matrix can have the methylated reads y_{ij}.

Such data is typically analyzed through $\hat{p}_{ij} = \frac{y_{ij}}{n_{ij}}$ also known as β-values in the literature. Due to the high dimensional nature of omics data, the analysis is focused on the sites that show significant change across subpopulations, for example, disease/ control sites, time point 1/ time point 2 sites, etc. The sites that show significant changes in methylation across subpopulations of interest are called differentially methylated sites (DMS). Due to the limitations in sequencing technology and the quality of the samples collected, there is often significant variation in the number of reads n_{ij} that are used to estimate $\hat{p}_{ij}$.

Statistical techniques for DNA methylation data

Statistical techniques that are appropriate for analyzing DNA methylation data have been developed and compared in numerous studies. Siegmund et al. (2004) used a Bernoulli-lognormal mixture model that used a Bernoulli model to account for zeroes that are observed

and a mixture model to separate aberrant DNA methylation patterns. Du et al. (2010) compared the performance of β-values $\left(\frac{y_{i,meth}}{y_{i,\text{meth}}+y_{i,\text{unmeth}}+\alpha_0}\right)$ and M-values $\left(\log_2\left(\frac{y_{i,meth}+\alpha_0}{y_{i,\text{unmeth}}+\alpha_0}\right)\right)$ for quantifying the methylation and concluded that M-values are more suited to differential analysis of methylation levels as they are more robust to heteroskedasticity and beta-values are more biologically interpretable. Zhuang et al. (2012) found that M-values aggravated the effects of outliers on inference for methylation levels and preferred β-values for principal component analysis. We use β-values which are more interpretable and additionally account for the difference in sample sizes used to estimate the proportions. Ma and Teschendorff (2013) applied a variational Bayes beta mixture model to perform feature selection and avoid false positives in detecting biomarkers in DNA methylation data. Dimension reduction techniques on DNA methylation data have been compared by Ma et al. (2014). Non-Gaussian and Gaussian dimension reduction techniques, followed by clustering were compared and it was concluded that non-Gaussian techniques that accounted for the bounded nature of β-values had the better performance in terms of clustering. Hubin et al. (2020) has been the only work so far that has accounted for the count variation in the methylation reads. They modeled the methylation counts in a regression model with auto-correlated errors between neighboring sites.

Jiang et al. (2019) used PCA to study the differences in DNA methylation profiles between rodent pups in limited bedding and nesting (LBN) and control groups. They collected buccal swabs from pups on post-natal day 2 and then they were randomly assigned to control and LBN groups. Buccal swabs were again collected on post-natal day 10. Principal component analysis on the DMS (differentially methylated sites that showed significant changes between post-natal day 2 and post-natal day 10) was able to distinguish between methylation profiles at different ages but not able to differentiate between different experiences. They further used an intra-individual approach called δ-methylation scores ($\log_2(\frac{p_{10}}{p_2})$, p_n=proportion of methylation at CpG site at postnatal day n) to differentiate between pups in LBN and control groups. It was found through the principal component weights that differentiated

between experiences, that rats in the LBN group had more methylation in genes that were responsible for important metabolic functions and less methylation in genes related to inflammation.

In human subjects, PCA techniques have not been successful in explaining a lot of the variation in the data. We would like to account for the heterogeneity in the number of reads n_{ij} across individuals while possibly extracting latent representations of the methylation proportions that explain the variation in the profiles. PCA aims to extract lower dimensional representations of the data that are obtained with linear combinations of the features. This method may be restrictive because it does not explicitly account for measurement errors; methods such as factor analysis (FA) although similar to PCA (Tipping and Bishop, 1999) may have an advantage over PCA (Harman, 1976; Kim et al., 1978; Rummel, 1988).

6.3 Methods

A principal component analysis (PCA) might have limitations when applied to empirical proportions of methylations due to the high variability in counts. We develop a method that accounts for the variation in empirical proportions due to limited counts and accounts for measurement error. We assume a factor analysis model on unobserved methylation proportion profiles that influence the observed methylated counts y_{ij} and observed proportions $\frac{y_{ij}}{n_{ij}}$. We briefly review factor analysis methods before introducing our proposed BLFA model.

6.3.1 Bayesian Factor Analysis

Bayesian factor analysis has been widely studied and applied to scientific and sociological studies. Factor analysis is known to have infinite solutions due to the model being invariant to multiplication with an orthogonal matrix. This problem is known as rotational invariance and a lot of literature focuses on fitting factor analysis models that introduce the uniqueness

of solutions, for example, some methods impose a positive lower triangular structure on the factor loadings (Geweke and Zhou, 1996; Bernardo et al., 2003; Lopes and West, 2004), and Frühwirth-Schnatter and Lopes (2018) recommended the use of a generalized lower-triangular representation of the factor loading matrix that avoids overfitting in a sparse model and correctly recovered the unknown number of factors.

There have also been efforts to perform sparse BFA so that the factor consists of fewer features for an interpretable model. Also since the number of factors are unknown *a priori*, methods have been explored that can either compare the fit between models that are fit with different number of factors or estimate the number of factors. Bernardo et al. (2003), Carvalho et al. (2008), and Bhattacharya and Dunson (2011) used different sparsity priors on the factor loading matrix to encourage sparsity, estimate the number of factors and perform Bayesian variable selection. Bai and Ng (2002) proposed an information criteria that can be used in large n and large p situations to compare models with different number of factors. Lopes and West (2004) proposed a reversible jump Markov chain Monte Carlo technique that can vary between different number of factors. Conti et al. (2014) proposed a Bayesian factor analysis model where each item loads onto at most one factor which produces sparse representations and can estimate the number of factors in the factor loading matrix. Finally, Ročková and George (2016) used a new sparsity prior (spike-and-slab Lasso, Ročková and George, 2018) coupled with an Indian Buffet process prior to avoid pre-specifying the number of factors.

6.3.2 Binomial Latent Factor Analysis (BLFA) Model

Let $i = 1, 2, \ldots, N$, denote the samples (subject), $j = 1, 2, \ldots, d$, denote the sites. The number of sites could include all CpG sites or just a subset based on a preliminary statistical analysis (i.e., DMS). For the j^{th} site in individual i, let n_{ij} be the number of observed reads in the DMS, and y_{ij} be the number of methylated reads. A binomial distribution is assumed for

y_{ij} with n_{ij} trials and an unknown probability of methylation, p_{ij}. Denote the d-dimensional vector of methylation probabilities across sites for a sample i as $\mathbf{p}_i$. We assume that p_{ij} depend on a latent variable Z_{ij} through a probit link. We assume that the vector of latent variables $\mathbf{Z}_i$ can be represented through a factor analysis model with a q-dimensional factor representation $\mathbf{x}_{i,q\times 1}$ ($q << d$). We write the model as follows:

$$\begin{aligned} y_{ij} \,|\, p_{ij} &\sim \text{Binomial}(n_{ij}, p_{ij}) \\ p_{ij} &= \Phi(Z_{ij}) \\ (\mathbf{Z}_i)_{d\times 1} \,|\, \mathbf{W},\, \mathbf{x},\, \boldsymbol{\mu} &\sim \text{MVN}(\mathbf{W}_{d\times q}\mathbf{x}_{i,q\times 1} + \boldsymbol{\mu}_{d\times 1}, \boldsymbol{\Sigma}), \end{aligned} \tag{6.1}$$

where $\Phi(.)$ is the standard Gaussian cumulative density function and the final equation uses a traditional FA setup (Harman, 1976; Rummel, 1988). The factor loading matrix $\mathbf{W}_{d\times q}$ defines the mapping from the d-dimensional vector $\mathbf{Z}_i$ to a lower-dimensional representation $\mathbf{x}_i$. The factor loading matrix $\mathbf{W}$, the intercept vector $\boldsymbol{\mu}_{d\times 1}$, and the variance-covariance matrix $\boldsymbol{\Sigma}$ is shared across individuals. If a significant fraction of the variation in $\mathbf{Z}_i$ can be captured in a few underlying continuous latent variables for all individuals, then the sites loading on these factors may shed insights on the samples. We assume that conditional on $\mathbf{W}$, $\mathbf{x}_i$, and $\boldsymbol{\mu}$, Z_{ij} are independent for all j, so that $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2, \ldots, \sigma_d^2)$ is a diagonal matrix. It is a typical assumption in factor analysis that $\mathbf{x}_i \sim \text{MVN}(\mathbf{0}_{q\times 1}\,, \mathbb{I}_{q\times q})$. This assumption means that the factors are independent and centered around zero. The variance of x_{ik}, $k = 1, 2, \ldots, q$ are fixed at the identity for identifiability of the factor loading matrix $\mathbf{W}$.

The key differences between a traditional factor analysis model and BLFA is that our model makes the distinction between p_{ij} and $\hat{p}_{ij}$ and assumes the factor analysis model applies to the unobserved methylation proportions.

6.3.3 Prior Distributions

We now discuss the prior distributions for the parameters of the BLFA model (6.1).

The model attempts to explain high-dimensional data in terms of a smaller number of factors with a subset of sites associated with each factor. This preference for sparse factor loading matrices $\mathbf{W}$ is expressed through the prior distribution on the elements of W. George and McCulloch (1997) introduced a spike-and-slab (SS) prior that uses a mixture of Gaussian distributions with different variance parameters for the "spike" (small variance) and "slab" (large variance) part:

$$
\begin{aligned}
&\pi(w_{jk} \,|\, \gamma_{jk}, \sigma^2_{j,\,\text{spike}}, \sigma^2_{j,\,\text{slab}}) = (1-\gamma_{jk})\,\phi(w_{jk} \,|\, 0, \sigma^2_{j,\,\text{spike}}) + \gamma_{jk}\,\phi(w_{jk} \,|\, 0, \sigma^2_{j,\,\text{slab}}) \\
&\sigma^2_{j,\,\text{slab}} = \sigma^2_j\,\sigma^2_{\text{slab}} \\
&\sigma^2_{j,\,\text{spike}} = \sigma^2_j\,\sigma^2_{\text{spike}} \\
&\phi(w_{jk} \,|\, 0, \sigma^2) = \frac{1}{\sqrt{2\pi\,\sigma^2}} \exp\left(\frac{-\,w^2_{jk}}{2\,\sigma^2}\right) \\
&\gamma_{jk} \,|\, \theta_k \sim \text{Bernoulli}\,(\theta_k) \quad j = 1, 2, \ldots, d;\; k = 1, 2, \ldots, q
\end{aligned}
\tag{6.2}
$$

In the priors for the factor loadings, the elements w_{jk} are selected from either a "spike" Gaussian distribution with parameter σ^2_{spike} or a "slab" Gaussian distribution with parameter σ^2_{slab}, where $\sigma^2_{\text{slab}} >> \sigma^2_{\text{spike}}$. The γ_{jk} are indicators as to whether the spike or the slab distribution is used for the site j for factor k respectively. The indicators γ_{jk} follow a Bernoulli distribution with success probability θ_k, shared by the factor loadings in a column. The hyperparameters σ^2_{slab} and σ^2_{spike} are not estimated and are specified such that $\sigma^2_{\text{slab}} >> \sigma^2_{\text{spike}}$. We use a simple prior for θ_k, such as a uniform distribution on the unit interval:

$$
\theta_k \sim \text{Beta}(1, 1) \tag{6.3}
$$

Non-informative priors are used on both μ_j and σ_j^2:

$$\pi(\mu_j) \propto 1$$
$$\pi(\sigma_j^2) \propto \text{Inv-Gamma}(\frac{1}{2}, \frac{1}{2}) \quad \forall\, j \in \{1, 2, \ldots, d\}.$$

6.3.4 Choosing the dimension q

The model described in Sections 6.3.2 and 6.3.3 depends on the dimension of the underlying factors q. It is common in PCA or FA to fit the model with different values of q and compare the results through metrics such as cross-validation loss or variance explained. We briefly describe the Indian Buffet process prior that can automatically infer the number of factors q in the factor analysis model.

The Indian Buffet Processes (IBP, Ghahramani and Griffiths, 2005; Teh et al., 2007; Knowles and Ghahramani, 2007; Griffiths and Ghahramani, 2011) are stochastic processes that define distributions over sparse binary matrices with a finite number of rows and potentially an infinite number of columns. The IBP prior is used in situations when the number of columns in a matrix are not assumed to be known. Teh et al. (2007) proposed a stick-breaking formulation for the IBP. First, consider a beta-Bernoulli prior on the elements of a finite binary matrix Γ with d rows and q columns as in our specification (6.3). For each column $k = 1, 2, \ldots, q$, θ_k is defined as the probability that $\gamma_{jk} = 1$ for the elements in column k. The θ_k are modeled as draws from a beta distribution:

$$\theta_k \overset{\text{i.i.d}}{\sim} \text{Beta}\left(\frac{\alpha}{q}, 1\right), \forall k = 1, 2, \ldots, q$$

Teh et al. (2007) define the IBP after marginalizing θ_k and taking the limit $q \to \infty$. Additionally, a stick-breaking construction for IBPs was presented which is very similar to the stick-breaking construction for Dirichlet Processes (Sethuraman, 1994). We next define the

stick-breaking construction for the IBP. Define, the ordered elements of the vector $\boldsymbol{\theta}_q$ as $\theta_{(1)} > \theta_{(2)} > \theta_{(3)} > \ldots > \theta_{(q)}$. Then, in the limit $q \to \infty$, the following stick-breaking law holds:

$$\begin{aligned} \theta_{(k)} &= \prod_{k_0=1}^{k} v_{k_0} \\ v_{k_0} &\overset{\text{i.i.d.}}{\sim} \text{Beta}(\alpha, 1), \quad \forall\, k_0. \end{aligned} \tag{6.4}$$

The model specification in (6.3) incorporates a fixed uniform prior on the set of θ_k's with known q. We can replace that prior with (6.4) to obtain the IBP prior on the binary matrix that represents whether the factor loading matrix has non-zero elements. This prior on the binary matrix, $\Gamma_W = [\gamma_{jk}]_{d\times q} = [\, I(|w_{jk}| > 0)\,]_{d\times q}$ avoids the need to pre-specify the number of factors in the sparse factor loading matrix. This setup has also been used in Ročková and George (2016) for a sparse factor analysis model.

For the preliminary results provided here, we fix q and compare results across different values.

6.3.5 Computation

The BLFA setup presented above with the model (6.1) and priors (6.2, 6.3) are fit to the data by using an expectation-maximization algorithm to obtain maximum a posteriori (MAP) parameter estimates. Note that the number of factors q is pre-specified here. The Expectation Maximization (EM) algorithm is an iterative approach for obtaining maximum likelihood (ML)/ maximum a posteriori (MAP) parameter estimates in the presence of missing or latent variables (Dempster et al., 1977).

We start by specifying the joint distribution L (up to a constant) of the data y_{ij}, latent variables $\mathbf{x}, \boldsymbol{\Gamma}$, and the parameters $\mathbf{W}, \mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\Sigma}$, that is proportional to the product of the

conditional distributions specified in (6.1), (6.2), and (6.3):

$$
\begin{aligned}
L \propto \pi(y \,|\, p)\ \pi(p \,|\, W, \mu, \Sigma)\ \pi(W \,|\, \Gamma, \sigma^2_{j,\,\text{spike}}, \sigma^2_{j,\,\text{slab}})\ \pi(\Gamma \,|\, \theta)\ \pi(x_i)\ \pi(\mu)\,\pi\,(\theta)\,\pi(\Sigma) \\
&\propto \prod_{i=1}^{N}\prod_{j=1}^{d} p_{ij}^{y_{ij}}(1-p_{ij})^{n_{ij}-y_{ij}} \times \\
&\prod_{i=1}^{N} \frac{1}{\sqrt{|\Sigma|}} \exp\!\left(\frac{-(\Phi^{-1}(p_i) - Wx_i - \mu)^T\Sigma^{-1}(\Phi^{-1}(p_i) - Wx_i - \mu)}{2}\right)\exp\!\left(-\frac{x_i^T x_i}{2}\right)\times \\
&\prod_{j=1}^{d}\prod_{k=1}^{q} \gamma_{jk}\,\frac{1}{\sqrt{\sigma^2_{j,\,\text{slab}}}}\exp\left(\frac{-w_{jk}^2}{2\sigma^2_{j,\,\text{slab}}}\right) + 1 - \gamma_{jk}\,\frac{1}{\sqrt{\sigma^2_{j,\,\text{spike}}}}\exp\left(\frac{-w_{jk}^2}{2\sigma^2_{j,\,\text{spike}}}\right) \times \\
&\prod_{j=1}^{d}\prod_{k=1}^{q}(1-\theta_k)^{1-\gamma_{jk}}\theta_k^{\gamma_{jk}} \times \prod_{k=1}^{q}\theta_k^{a-1}(1-\theta_k)^{b-1} \times \prod_{j=1}^{d}\frac{1}{(\sigma_j^2)^{\frac{3}{2}}}\exp\left(\frac{-1}{2\sigma_j^2}\right)
\end{aligned}
\tag{6.5}
$$

The expression L in (6.5) is maximized with respect to the parameters of interest $\mathbf{\Omega} = (\mathbf{p}, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\theta}, \mathbf{\Sigma})$, with the components $\mathbf{x}, \mathbf{\Gamma}$ are treated as latent or missing. In the first (expectation) step, we derive the expected value of the complete data likelihood L with respect to the conditional distribution of the latent variables $\mathbf{x}$ and $\mathbf{\Gamma}$ given the data and the parameters $\mathbf{\Omega} = (\mathbf{p}, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\theta}, \mathbf{\Sigma})$. In the second (maximization) step, the expression obtained from the expectation step is then maximized with respect to the parameters $\mathbf{\Omega} = (\mathbf{p}, \mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\theta}, \mathbf{\Sigma})$. This may be done simultaneously if possible or we can find maxima of each parameter in $\mathbf{\Omega}$ one at a time. The latter approach is called Expectation Conditional Maximization (ECM, Meng and Rubin, 1993). These two steps are repeated until convergence. The general EM steps are defined below where t denotes the iteration number:

$$
\begin{aligned}
\text{E-step: } Q(\Omega \,|\, \Omega^{(t)}) &\equiv E_{\pi(x,\Gamma \,|\, \Omega^{(t)})}\left[\log(L)\right] \\
\text{M-step: } \Omega^{(t+1)} &\equiv \arg\max_{\Omega} Q(\Omega \,|\, \Omega^{t})
\end{aligned}
\tag{6.6}
$$

The EM steps for fitting the model (6.1) through the expression proportional to the log posterior given in (6.5) are discussed in more detail here. In the t^{th} iteration, conditional on

$\mathbf{\Omega^{(t)}}$, the E-step replaces functions of the augmented variables $\mathbf{x}_i$ and $\mathbf{\Gamma}$ with their expected values with respect to the conditional distribution $\pi(\mathbf{x}, \mathbf{\Gamma} \,|\, \mathbf{\Omega^{(t)}}, y)$. The binary elements of $\mathbf{\Gamma}$ appear in $\log L$, thus the E-step computes $E(\gamma_{jk} \,|\, \Omega)$, denoted as $\hat{\gamma}_{jk}^{(t)}$ in the t^{th} iteration:

$$\hat{\gamma}_{jk}^{(t)} = \frac{\phi(w_{jk}^{(t-1)} \,|\, 0, \sigma_{j,\,\text{slab}}^{2\,(t-1)})\, \theta_k^{(t-1)}}{\phi(w_{jk}^{(t-1)} \,|\, 0, \sigma_{j,\,\text{slab}}^{2\,(t-1)})\, \theta_k^{(t-1)} + \phi(w_{jk}^{(t-1)} \,|\, 0, \sigma_{j,\,\text{spike}}^{2\,(t-1)})\, (1 - \theta_k^{(t-1)})} \quad \forall\, j = 1, 2, \ldots, d;\; k = 1, 2, \ldots, q, \tag{6.7}$$

where $\phi(\,.\,|\,\mu,\, \sigma^2)$ denotes the Gaussian density function for a distribution with mean μ and variance σ^2. The conditional distribution of $\mathbf{x}_i$ given $\mathbf{\Omega}$ is multivariate Gaussian distribution with mean $(I_{q\times q} + W^T\Sigma^{-1}W)^{-1}W^T\Sigma^{-1}(Z_i - \mu)$ and variance $(I_{q\times q} + W^T\Sigma^{-1}W)^{-1}$. These are used to calculate the $E(x_i \,|\, \Omega)$ denoted as $\hat{x}_i^{(t)}$ and $E(x_i x_i^T \,|\, \Omega)$ denoted as $\hat{x^2}_i^{(t)}$ which are needed to compute $E\,(\log L)$.

$$\begin{aligned} \hat{x}_i^{(t)} &= (I_{q\times q} + W^{T\,(t-1)}\Sigma^{-1\,(t-1)}W^{(t-1)})^{-1}W^{T\,(t-1)}\Sigma^{-1\,(t-1)}(Z_i^{(t-1)} - \mu^{(t-1)}) \\ \hat{x^2}_i^{(t)} &= (I_{q\times q} + W^{T\,(t-1)}\Sigma^{-1\,(t-1)}W^{(t-1)})^{-1} + \hat{x}_i^{(t)}\hat{x}_i^{T\,(t)}. \end{aligned} \tag{6.8}$$

The maximization step involves maximizing the result from the expectation step with respect to the parameters in $\mathbf{\Omega}$. This can be done analytically for W, μ, Σ, θ after some linear algebra

manipulation in $Q(W, \Sigma \,|\, W^{(t)}, \Sigma^{(t)})$:

$$
\begin{aligned}
Q(W, \Sigma \,|\, W^{(t)}, \Sigma^{(t)}) \propto & \sum_{i=1}^{N} \Bigg(\log(\sqrt{|\Sigma|}) - \left(\frac{-(\Phi^{-1}(p_i) - W\hat{x}_i^{(t)} - \mu)^T \Sigma^{-1} (\Phi^{-1}(p_i) - W\hat{x}_i^{(t)} - \mu)}{2} \right) \\
& + \left(\frac{\hat{x}_i^{T\,(t)} W^T \Sigma^{-1} W \hat{x}_i^{(t)} - \hat{x^2}_i^{(t)} W^T \Sigma^{-1} W}{2} \right) \Bigg) \\
& + \sum_{k=1}^{q} \sum_{j=1}^{d} \frac{1}{\sigma_j^2} \left(\hat{\gamma}_{jk}^{(t)} \frac{-w_{jk}^2}{2\sigma_{\text{slab}}^2} + 1 - \hat{\gamma}_{jk}^{(t)} \frac{-w_{jk}^2}{2\sigma_{\text{spike}}^2} \right) - \frac{1}{2} \sum_{k=1}^{q} \sum_{j=1}^{d} \log(\sigma_j^2) - \sum_{j=1}^{d} \frac{1}{2\sigma_j^2} \\
= & \sum_{i=1}^{N} \Bigg(\log(\sqrt{|\Sigma|}) - \left(\frac{-(\Phi^{-1}(p_i) - W\hat{x}_i^{(t)} - \mu)^T \Sigma^{-1} (\Phi^{-1}(p_i) - W\hat{x}_i^{(t)} - \mu)}{2} \right) \\
& + \left(\frac{W^T \Sigma^{-1} W \left(I_{q\times q} + W^{T\,(t-1)} \Sigma^{-1\,(t-1)} W^{(t-1)} \right)^{-1}}{2} \right) \Bigg) \\
& + \sum_{k=1}^{q} \sum_{j=1}^{d} \frac{1}{\sigma_j^2} \left(\hat{\gamma}_{jk}^{(t)} \frac{-w_{jk}^2}{2\sigma_{\text{slab}}^2} + 1 - \hat{\gamma}_{jk}^{(t)} \frac{-w_{jk}^2}{2\sigma_{\text{spike}}^2} \right) - \frac{1}{2} \sum_{k=1}^{q} \sum_{j=1}^{d} \log(\sigma_j^2) - \sum_{j=1}^{d} \frac{1}{2\sigma_j^2}
\end{aligned}
$$

The second term in the expression $(\sum_{i=1}^{N} \frac{\hat{x}_i^{T\,(t)} W^T \Sigma^{-1} W \hat{x}_i^{(t)} - \hat{x^2}_i^{(t)} W^T \Sigma^{-1} W}{2})$ is a trace of a matrix and can be written as a function of $N \sum_{j=1}^{d} \tilde{w_j}^T (I_{q\times q} + W^{T\,(t-1)} \Sigma^{-1\,(t-1)} W^{(t-1)}) \tilde{w_j}$. Define,

$$
\mathbf{Z}_0 = \begin{pmatrix} (\mathbf{Z}_{N\times d}^T - \mu^T) = (\Phi^{-1}(\mathbf{p}) - \mu^T) \\ \mathbf{0}_{q\times d} \end{pmatrix}_{(N+q)\times d} \quad \text{and}
$$

$$
\mathbf{x}_0 = \begin{pmatrix} \hat{\mathbf{x}}_{N\times q}^T \\ \sqrt{N} (I_{q\times q} + W^{T\,(t-1)} \Sigma^{-1\,(t-1)} W^{(t-1)})_{q\times q}^{-\frac{1}{2}} \end{pmatrix}_{(N+q)\times q}, \text{ then the maximization steps are:}
$$

$$
\begin{aligned}
& \tilde{w_j}^{(t)} = (x_0^T x_0 + \Lambda_{j,\,q\times q}^{(t)})^{-1} x_0^T Z_{0,j}, \quad \text{where, } \Lambda_j^{(t)} = \text{ diag } \left(\frac{\hat{\gamma}_{jk}^{(t)}}{\sigma_{\text{slab}}^2} + \frac{1 - \hat{\gamma}_{jk}^{(t)}}{\sigma_{\text{spike}}^2} \right), \, k = 1, 2, \ldots, q \\
& \mu_j^{(t)} = \frac{\sum_{i=1}^{N} (\Phi^{-1}(p_{ij}^{(t)}) - \hat{x}_i^T w_j^{(t)})}{N}, \quad j = 1, 2, \ldots, d \\
& \sigma_j^{2\,(t)} = \frac{\sum_{i=1}^{N+q} (Z_{0,i,j} - x_{0,i} w_j^{(t)})^2 + \sum_{k=1}^{q} \Lambda_{j,kk}^{(t)} w_{jk}^{2\,(t)} + 1}{N + 2\,q + 1}, \quad j = 1, 2, \ldots, d \qquad (6.9) \\
& \theta_k^{(t)} = \frac{\sum_{j=1}^{d} \hat{\gamma}_{jk}^{(t)}}{d}, \quad k = 1, 2, \ldots, q
\end{aligned}
$$

Note that in the maximization steps for the loadings $\tilde{w_j}$, Λ_j is a diagonal matrix consisting of

elements that are a function of $\langle\gamma_{jk}\rangle$, σ_{spike}, and σ_{slab}. The maximization expression for $\tilde{w_j}$ is obtained as the closed-form estimate of ridge regression (Ročková and George, 2014).

The maximization step for methylation proportions p_{ij} involves finding the maximizing value of the function specified below:

$$y_{ij}\log(p_{ij}) + (n_{ij} - y_{ij})\log(1 - p_{ij}) - \frac{(\Phi^{-1}(p_{ij}) - \hat{x}_i\tilde{w_j}^{(t)} - \mu_j^{(t)})^2}{2\,\sigma_j^{2\,(t)}} \tag{6.10}$$

We achieve this by using the *optimize* function in the R programming language. Alternatively, the Newton-Raphson algorithm (Ypma, 1995) can also be employed.

Additionally, at the end of each ECM step, we perform varimax rotations (Kaiser, 1958) on the factor loading matrix W, similar to the technique used in Ročková and George (2016). This rotation allows variables to load onto fewer factors and makes the results easier to interpret.

6.4 Simulation Studies

We demonstrate the analysis approach on simulated data.

6.4.1 Data Generation Technique

We describe our approach for generating simulated data for d sites, N samples, and q factors. In all simulations the number of samples, N was fixed to 50. The factor analysis parameters $\sigma_j^2 = 1$ and $\mu_j = 0.5$, $j = 1, 2, \ldots, d$, are also fixed for all simulations. To generate the latent scores $\mathbf{Z}_i$ we need to specify W. We simulate different W values for different values of $d = 100$, 500, 1000 and $q = 3$, 5, 7. In all cases, W is a block-diagonal matrix so that starting with the first column, adjacent rows in W are set to 1 and the rest of the elements in the column are set to 0. An example of a block-diagonal factor loading matrix is demonstrated in Figure 6.1. $Z_i = \Phi^{-1}(p_i)$ are then generated with a multivariate normal (MVN) distribution having mean μ and variance-covariance matrix $W^TW + \Sigma$, the marginal distribution of Z_i once x_i is integrated out. Finally, our approach is

intended to address binomial sampling variability; we introduce this variation by randomly drawing $y_{ij} \sim \text{Bin}(n_{ij}, p_{ij})$, with the binomial counts n_{ij} uniformly sampled from [5, 50].

Fitting the data

The input to the algorithm are y_{ij}, n_{ij} and starting values for parameters $\mathbf{\Omega}^0$, which are generated randomly. We use a Gaussian distribution for W, μ and σ_j^2. The proportions θ_k and p are uniformly generated from the interval (0,1). The hyperparameters are set as $\sigma^2_{\text{spike}} = 0.1$ and $\sigma^2_{\text{slab}} = 100$. Note we assume that the number of latent factors (q) is specified. The ECM steps provided in (6.7, 6.8, 6.9, 6.10) are repeated until the maximum absolute difference in W values, $\max_{j,k} |w_{jk}^{(t+1)} - w_{jk}^{(t)}|$ are less than a very small number, e.g., $\epsilon = 0.05$. Additionally, we monitor the log posterior probability (up to a proportionality constant) given by log (L) (6.5).

Baseline Model

In previous exploratory analyses of DNA methylation data (Du et al., 2010; Ma et al., 2014; Jiang et al., 2019) the methylation proportions are used without accounting for binomial variation. In order to assess the contribution of accounting for binomial variation, we compare the BLFA model to a baseline model that uses factor analysis on the probit transformed proportions.

$$\begin{aligned} \hat{p}_{ij} &= \frac{y_{ij}}{n_{ij}} \\ \Phi^{-1}(\hat{p}_{ij}) &= Z_{ij}^b \\ Z_i^b \,|\, W^b, \mu^b, x^b, \Sigma^b &\sim \text{MVN}(W^b x_i^b + \mu^b, \Sigma^b) \end{aligned} \tag{6.11}$$

We assume a spike-and-slab prior on the factor loading matrix for the baseline model similar to the setup presented earlier. The only change in this method is ignoring the binomial variation. We compare metrics such as root mean squared error (RMSE) between the true methylation proportions and the estimated methylation proportions, and the Frobenius norm $\sqrt{\sum_{ij}(a_{ij} - b_{ij})^2}$ between the true and estimated variance-covariance matrices of the distributions for $\Phi^{-1}(p)$ marginalized over x $(W^T W + \Sigma)$.

Results

Data was generated for six total designs: i) $d = 100$ and $q = 3, 5$, ii) $d = 500$ and $q = 5, 7$, iii) $d = 1000$ and $q = 5, 7$. Details for the W used to simulate the data have been provided in Section 6.4.1. For each design (i.e., each choice of d and q), we generate 5 data sets and report results combined over the five repetitions. Figure 6.2 shows an example of the recovered factor loading matrix after convergence for one of the simulated cases with the simulated block diagonal matrix given in Figure 6.1. Note that the columns get re-ordered because the factor loading matrices are only identifiable up to a rotation of columns. We can see that factor 2 in the recovered loading matrices corresponds to factor 5 in the original. The same is true for other factors. Figure 6.3 shows the trend in the log posterior (6.5) as a function of iteration for a simulated example. These figures demonstrate that our algorithm is able to recover the factor structure pretty well.

Table 6.1 reports the mean RMSE and Frobenius norm for 5 simulated data sets in each setting with the BLFA method (6.1) and baseline method (6.11). It is important to note that the Frobenius norm increases with d because it has not been normalized. We notice that for most settings, our method performs better compared to the baseline model in terms of RMSE and Frobenius norm. The RMSE is better with our model for $q = 3$ and 5, and worse for $q = 7$ compared to the baseline across all examples. The Frobenius norm obtained by the BLFA method was lower for $d = 500$, and 1000 compared to the baseline for 4 out of the 5 simulated data sets and higher for $d = 100$ in all 5 simulated data sets.

We also study the effect of misspecifying q in two settings: i) $d = 100$ and $q_{\text{true}} = 5$ and ii) $d = 100$ and $q_{\text{true}} = 7$. For both these designs, 5 data sets were generated and then fit using 3 values of q_{fit} each so that: $q_{\text{fit},1} < q_{\text{fit},2} = q_{\text{true}} < q_{\text{fit},3}$. Table 6.2 reports the mean RMSE, mean Frobenius norm, and the value of log(L) in expression (6.5) over the 5 simulated data sets. We notice that RMSE is worse when $q_{\text{fit}} > q_{\text{true}}$ and Frobenius norm is worse when $q_{\text{fit}} < q_{\text{true}}$. The criteria log(L) increases with q_{fit}. These results indicate that the method does not necessarily identify the true number of underlying factors.

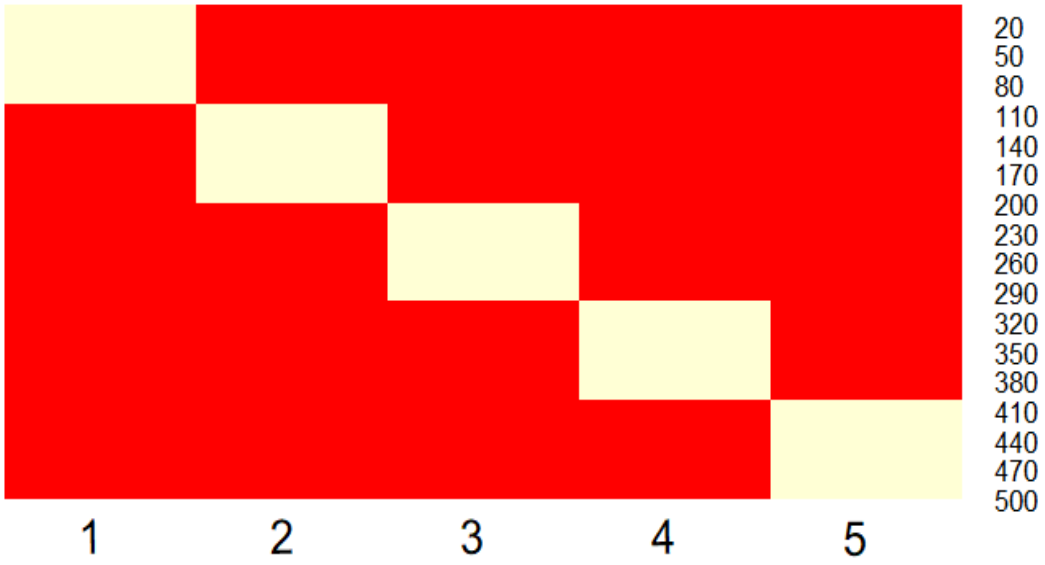


Figure 6.1: An example of the simulated block diagonal factor loading matrix W that was generated for simulations for $d = 500$ sites and $q = 5$ underlying factors.

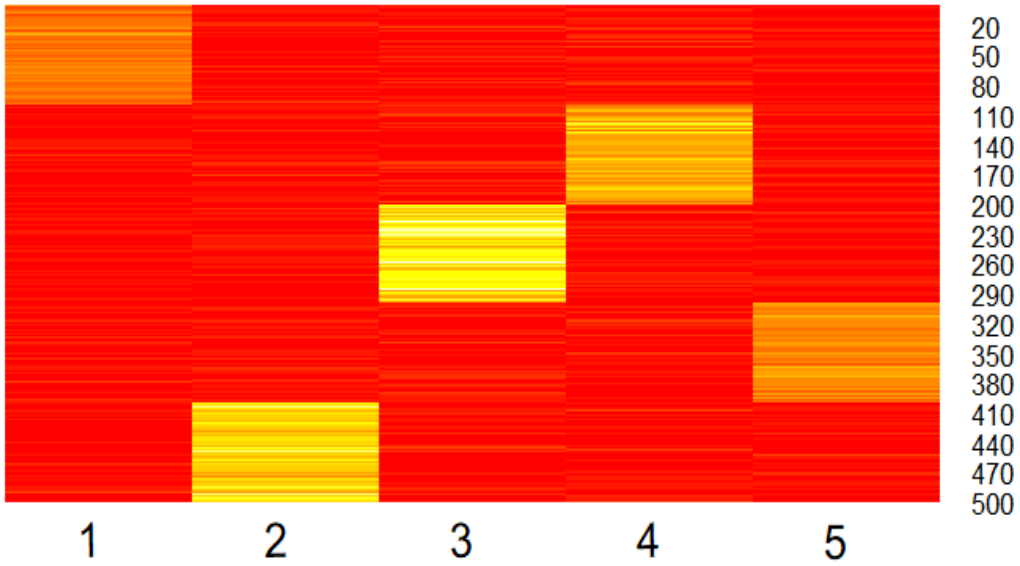


Figure 6.2: The recovered factor loading matrix. An example of the resulting factor loading matrix obtained after fitting the simulated data with the simulated block diagonal matrix in Figure 6.1.

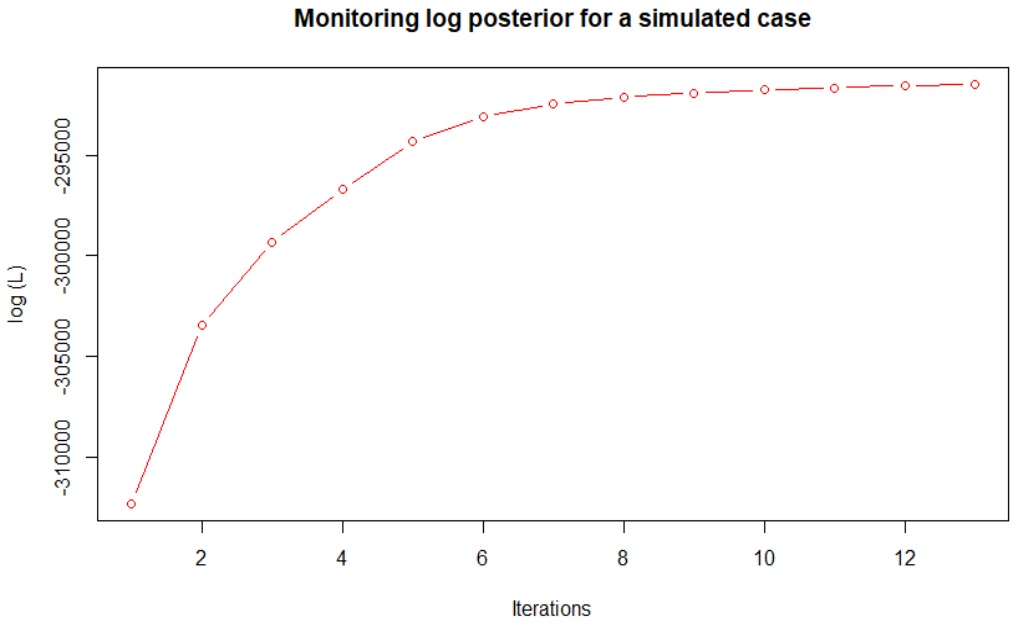


Figure 6.3: Log posterior of L in expression (6.5) is plotted before convergence for a simulated data set.

Metric Setting	Mean RMSE (BLFA)	Mean RMSE (baseline)	Mean Frobenius norm (BLFA)	Mean Frobenius norm (baseline)
$d = 100$ $q_{\text{true}} = 3$	0.0775	0.0816	46.8976	42.8977
$d = 100$ $q_{\text{true}} = 5$	0.0795	0.0823	35.723	33.0995
$d = 500$ $q_{\text{true}} = 5$	0.0805	0.0825	171.1929	185.2211
$d = 500$ $q_{\text{true}} = 7$	0.0821	0.0813	147.3723	159.0564
$d = 1000$ $q_{\text{true}} = 5$	0.0810	0.0821	358.9809	378.2567
$d = 1000$ $q_{\text{true}} = 7$	0.0834	0.0820	287.4971	318.2121

Table 6.1: Average RMSE and Frobenius norm between true $W^T W + \Sigma$ and estimated $W_{\text{est}}^T W_{\text{est}} + \Sigma_{\text{est}}$ are reported across 5 simulated cases for each design (d and q). Data is fit using BLFA method (6.1) and baseline method (6.11).

Metric Setting	Mean RMSE (BLFA)	Mean Frobenius norm (BLFA)	Mean log(L) (BLFA)
$d = 100$ $q_{\text{true}} = 5$, $q_{\text{fit}} = 4$	0.0791	37.2928	-58957.21
$d = 100$ $q_{\text{true}} = 5$, $q_{\text{fit}} = 5$	0.0795	35.7230	-58674.73
$d = 100$ $q_{\text{true}} = 5$, $q_{\text{fit}} = 7$	0.0822	36.4136	-58398.18
$d = 500$ $q_{\text{true}} = 7$, $q_{\text{fit}} = 5$	0.0796	162.8881	-291079.2
$d = 500$ $q_{\text{true}} = 7$, $q_{\text{fit}} = 7$	0.0821	147.3723	-288246.6
$d = 500$ $q_{\text{true}} = 7$, $q_{\text{fit}} = 10$	0.0880	137.2893	-286012.4

Table 6.2: Effect of misspecifying q is compared through RMSE, Frobenius norm, and mean log posterior. q_{fit} indicates the q that was assumed and q_{true} indicates the true number of factors that were used to generate the data.

6.5 Studying DNA methylation in Human Subjects

Data

The Conte Center at the University of California Irvine aims to study the effects of early life adversity (ELA) on cognitive and emotional development. DNA methylation data is being explored to assess whether there is an epigenetic signature of ELA. The analysis of these data motivated the development of the BLFA method.

We briefly describe the data that were gathered. DNA samples were collected through buccal swabs from N=107 infants at one month of age and then again at one year post birth. DNA sequencing was done through reduced representation bisulfite sequencing (RRBS) technique followed by aligning the reads with the reference genome with Bismark 0.16.3 (Short et al., 2023, Krueger and Andrews, 2011). This procedure identified $> 1.6 \times 10^6$ distinct methylated or unmethylated CpG sites across individuals and time points.

We focus on methylated sites that show significant changes in methylation across time points of 1 month of age and 1 year of age; firstly, sites that had more than $\pm 5\%$ change in methylation were

identified. These sites were then tested for change in the methylation proportions with a Fisher's exact test and the logarithm of these p-values for each site was added across individuals. Define, $T_j = -2\sum_{i=1}^{N} \ln(\rho_{ij})$, ρ_{ij} denotes the p-value for Fisher's exact test (Fisher, 1922) for subject i at site j (Dai et al., 2014). Under the null hypothesis that $\rho_{ij} \sim \text{Unif}(0,1)$, statistic T_j follows a χ^2_{2N} distribution (Fisher, 1992); so for each site we obtained a new p-value with $\rho_{\chi,j} = P(T > T_j \mid H_0)$. We then used the Benjamini-Hochberg procedure to control for false discovery rate q=0.1 with the p-values $\rho_{\chi,j}$ (Benjamini and Hochberg, 1997). Sites that were selected after this process were used for downstream analysis with the BLFA model proposed in Section 6.5. These sites were called differentially methylated sites (DMS). There were 14,103 DMS in our data.

Due to limitations in sequencing techniques and the quality of samples collected, note that the resulting DMS data has a lot of variation in the number of reads n_{ij}. We are interested in using β-values $\left(\frac{y_{i,\text{meth}}}{y_{i,\text{meth}} + y_{i,\text{unmeth}} + \alpha_0}\right)$ which can display a lot of variation especially when there are an insufficient number of reads, i.e., the denominator is small. For example, in our data, $\approx 40.4\%$ of the DMS have less than 30 reads to estimate the methylation proportion (β value) and $\approx 70.9\%$ of the DMS have less than 50 reads. However, some sites have more than 1000 reads for some individuals. Note that for $< 1\%$ of the data across individuals and sites, there are 0 reads for a site. This situation arises when a site is a DMS but there were no reads for an individual for that site during the sequencing. The BLFA model presented in Section 6.3 addresses the heterogeneity in read counts.

Quantifying Childhood Unpredictability

Along with DNA samples, some other covariates such as sex, household income, etc. were collected for each infant child. One of the main aims of this study was to understand the associations between childhood unpredictability and DNA methylation. One of the methods used for quantifying unpredictability is entropy (Davis et al., 2017; Vegetabile et al., 2019; Davis et al., 2022). For a discrete random variable X with K states and probability mass function: $\pi_1, \pi_2, \ldots, \pi_K$, the entropy is defined as follows:

$$H(X) = -\sum_{k=1}^{K} \pi_k \log_2(\pi_k)$$

Interactions between a mother and child were recorded and sensory input to infants was measured through combinations of auditory, tactile, and visual input ($2^3 = 8$ total states). These interactions were then modeled with a first-order Markov chain and the transition matrix that has elements P_{kl} that indicate the probability of jumping from state k to state l. A first-order Markov chain is fully specified with a transition matrix and under a few regularity conditions it has a stable long-term behavior in terms of the frequency with which the stochastic process visits the different states. Denote this stable frequency as π'_k, then the entropy for the behavioral interactions may be calculated as follows:

$$H(X) = -\sum_{k=1}^{K} \sum_{l \neq k} \pi'_k P_{kl} \log_2(P_{kl})$$

which is a sum of entropies across the rows of the transition matrix weighted by the long-term behavior π'. Higher entropy indicates more unpredictability and it ranges between 0 and $\log_2 K$. Entropy has been shown to be associated with cognitive functions later in life (Davis et al., 2017; Davis et al., 2019; Davis et al., 2022). We will use the entropy measure to check if latent factor representations are associated with early life experience.

Results

We analyze the DNA methylation data collected in infants with the BLFA method. We use $\sigma^2_{\text{spike}} = 0.1$ and $\sigma^2_{\text{slab}} = 100$ as initial hyperparameter values. The initial values for the parameters are chosen randomly as described in the simulation studies in Section 6.4. We start with $q = 5$ latent factors because we are interested in capturing the variation in the data with a few latent variables that load on some DMSs each so that we may be able to interpret the significance of these DMSs.

Results from a spike-and-slab analysis involve thresholding the resulting output to select the factor loadings that are larger/ more significant than others (George and McCulloch, 1997; Ishwaran and Rao, 2005). We use an ad-hoc approach to identify loadings that are significant with absolute value > 0.3.

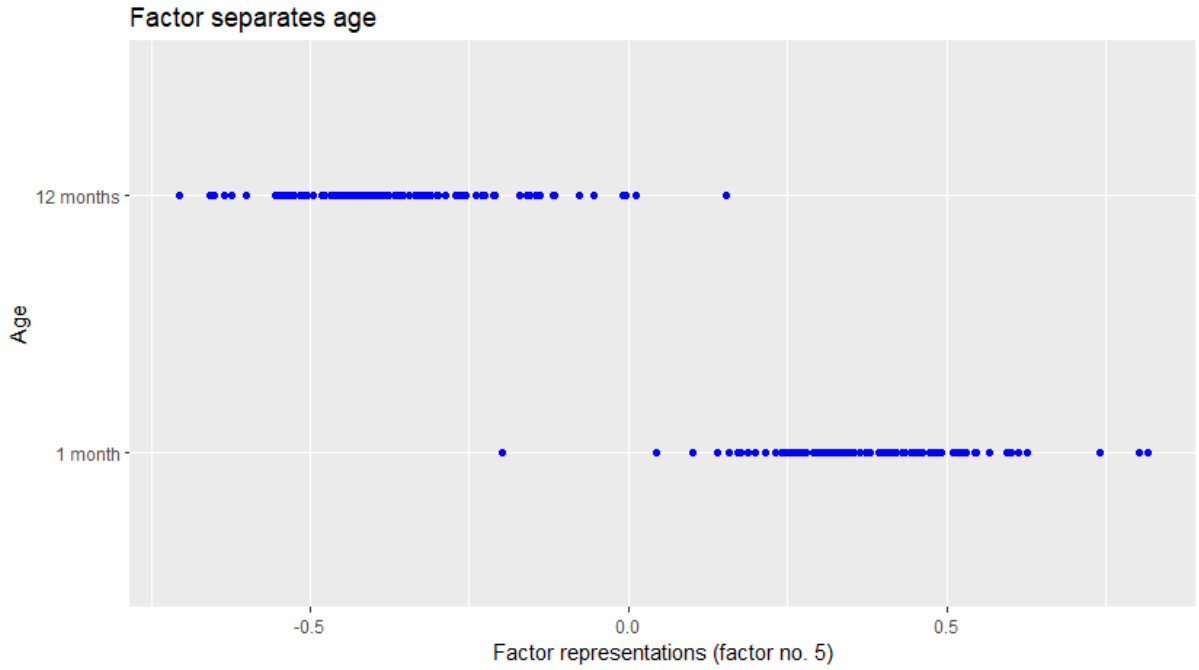


Figure 6.4: One of the factor representations x_i (fifth factor) is able to differentiate between ages the samples at age 1 month and 1 year.

We observed that one of the five factors is able to differentiate between the ages at which the samples were collected as demonstrated in Figure 6.4. Similar results were also found with the baseline model. We further analyze if there may be any relationships between the latent factors and covariates that indicate childhood unpredictability. However, we found only a weak association for one of the factors with entropy. One possibility is that, as in Jiang et al. (2019) rodent study, it would be better to focus on change in methylation between the two time points.

6.6 Conclusion

In this work, we have developed the BLFA method to account for the variation in DNA methylation counts while conducting exploratory analyses of methylation proportions. We accomplished that through a binomial model on methylation counts along with an underlying factor analysis model with a spike-and-slab sparsity prior on the elements of the factor loading matrix. We have detailed the steps for an ECM algorithm to fit this model along with presenting the efficacy of this technique with simulation studies. We demonstrated with simulation studies that our method outperforms

the typical method of ignoring count variation when using dimension reduction techniques.

Additionally, we analyzed a DNA methylation data set from a study of early life adversity and found that the BLFA method was able to obtain sparse representations of the methylation proportions. We observed that one of the extracted factors was able to differentiate between age groups. However, we did not observe a strong correlation between the extracted factors and childhood unpredictability. Our results for these data are quite similar to those obtained by the baseline model that ignores variation in read counts. In the future, we would like to extend our BLFA method to accommodate an intra-individual design like the one applied in Jiang et al. (2019) with the δ-methylation measure. Additionally, we would like to interpret the results from our sparse factor analysis approach by recording the CpG sites that the factors load onto and understanding the role of the genes that are associated with the respective sites.

Chapter 7

Discussion and Future Work

The reliability of measurements is critical for scientifically well-founded inference. In forensics, subjective decisions will continue to be part of evidence assessments for the foreseeable future. Well-studied error rates and reliability for each forensic evidence discipline is key for interpreting case evidence correctly as well as fair and just course proceedings.

In this work, the proposed statistical methods address numerous questions that are interesting to the forensics community. We started out by developing a method to assess reliability for continuous and binary outcomes while combining the reproducibility and repeatability black-box study data sets that are generally analyzed separately. We also provided a method to infer possible examiner-sample interactions. Even though continuous outcomes are rare in forensics, examples such as handwriting complexity scores can be modeled with this method. Binary outcomes such as value/no-value, match/no-match can also be assessed for reliability with the applied method.

Ordinal outcomes are ubiquitous in forensics. We extended the latent variable model applied to the binary data for ordinal outcomes. Examiners differ in their tendencies to rate forensic samples and the CUT model is again able to combine reliability studies. Additionally, we are able to quantify examiner thresholds, while adjusting for sample difficulties, that enable comparisons across examiners. The SET model, a simpler version of the CUT model, assumes that examiners

share these tendencies but enables the estimation of possible examiner-sample thresholds. These methods are applicable to many black-box studies that have been conducted so far and they provide an efficient estimation for reliability while accounting for examiner thresholds and interactions. We also developed a method that can cluster examiners based on their tendencies to rate varying subsets of samples while accounting for their sample difficulties. We achieved this goal with a mixture of Dirichlet processes (MDP) model on examiner tendencies. A similar structure can also cluster samples based on their tendencies to receive decisions. This model has applications in exploratory analysis and hypothesis generation. One can explore whether there exist unobserved covariates (black-box studies do not collect examiner information such that they can be compared against performance) that explain the clusters. This could imply that reliability and accuracy are dependent on differences in such covariates. This motivates black-box studies that collect more information.

We have also looked at measurement reliability in the application area of DNA methylation. Our proposed BLFA model is able to account for the variation that arise due to limited reads while estimating methylation proportion at a CpG site. We are also able to assess whether there exists an underlying low-dimensional subspace that explains most of the variation in the methylation proportions. Our method encourages sparse representations of the factor loading matrix which entails that few CpG sites load to each factor.

7.1 Future Work

Our proposed methods enable efficient analysis of the data from black-box studies while accounting for examiner tendencies, sample difficulties and examiner-sample interactions. There is, however, more work that will follow these analyses. The jury should be informed of the empirical reliability and accuracy of the forensic science discipline that is being used. The method and language that is used to report conclusions to the jury is an ongoing area of research (Thompson, 2017; Thompson et al., 2018).

In the future, if examiner and sample covariates are made available in black-box studies, it will be

important to account for these variables while assessing accuracy and reliability. Our clustering technique enables inferring groups of examiners or samples, however, there is the possibility of performing biclustering. Such a technique would extract clusters of examiners and samples simultaneously. These biclusters may provide exploratory insights about the decision-making process and might also be helpful for predictions.

It may also be helpful to identify examiners that differ significantly from other examiners while making assessments. Holsclaw et al. (2012) used a spike-and-slab distribution as a base distribution for a Dirichlet process mixture. A similar method may also be applied in the setting that was motivated in Chapter 5, for example:

$$P(Y_{ijk} = m) = P(\kappa_m < Z_{ijk} \leq \kappa_{m+1})$$

$$Z_{ijk} \sim N(\alpha^*_{c_i} + \gamma_j, 1)$$

$$\alpha^*_t \sim a_t \, N(\mu_0, \sigma_0^2) + (1 - a_t) \, \delta_{\alpha_0}$$

Here, a_t is the probability that cluster t samples from the slab component $N(\mu_0, \sigma_0^2)$ and the spike component is the Dirac-delta function δ_{α_0} centered around α_0.

The sparse factor loading structure for methylation data that is identified with the BLFA model is helpful while identifying CpGs that the factors load onto. In Chapter 6, we were interested in associations between underlying factors and entropy (a proxy for early life predictability). We would like to find the functions of the corresponding genes and find their biological functions. Furthermore, we would like to extend our method to automatically select the number of factors q. This can be accomplished using the Indian buffet prior (IBP) as described in Chapter 6. However, we leave the implementation as future work. Additionally, we would like to extend our method to account for intra-individual methylation as developed in Jiang et al. (2019).

Bibliography

Abdi, H. and L. J. Williams (2010). "Principal component analysis". *Wiley interdisciplinary reviews: computational statistics* 2:4, 433–459.

Aguirre, A. J., G. E. Guevara-Viera, C. S. Torres-Inga, R. V. Guevara-Viera, A. Bone, M. Vidal, and F. J. Garcia-Ramos (2020). "Analysis of Fluid Velocity inside an Agricultural Sprayer Using Generalized Linear Mixed Models". *Applied Sciences* 10:15, 5029.

Albert, J. H. and S. Chib (1993). "Bayesian analysis of binary and polychotomous response data". *Journal of the American Statistical Association* 88:422, 669–679.

Alewijnse, L., E. Van Den Heuvel, R. Stoel, and F. K. (2011). "Analysis of signature complexity". *Journal of Forensic Document Examiners* 21, 37–49.

Angel, M., M. Caligiuri, and M. Cavanaugh (2017). "Kinematic models of subjective complexity in handwritten signatures". *Journal of the American Society of Questioned Document Examiners, Inc* 20:2, 3–10.

Antoniak, C. E. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems". *The Annals of Statistics* 2:6, 1152–1174.

Arora, H., N. Kaplan-Damary, and H. Stern (2022). "Combining Reproducibility and Repeatability Studies in Forensics". *Technical Report (submitted to journal).*

Arora, H., N. Kaplan-Damary, and H. Stern (2023). "Reliability of Ordinal Outcomes in Forensic Black-Box Studies". *Technical Report (submitted to journal).*

Bai, J. and S. Ng (2002). "Determining the number of factors in approximate factor models". *Econometrica* 70:1, 191–221.

Baldwin, D. P., S. J. Bajic, M. Morris, and D. Zamzow (2014). *A study of false-positive and false-negative error rates in cartridge case comparisons*. Tech. rep. AMES LAB IA.

Baram, T. Z., E. P. Davis, A. Obenaus, C. A. Sandman, S. L. Small, A. Solodkin, and H. Stern (2012). "Fragmentation and unpredictability of early-life experience in mental disorders". *American Journal of Psychiatry* 169:9, 907–915.

Benjamini, Y. and Y. Hochberg (1997). "Multiple hypotheses testing with weights". *Scandinavian Journal of Statistics* 24:3, 407–418.

Bernardo, J., M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West (2003). "Bayesian factor regression models in the "large p, small n" paradigm". *Bayesian Statistics* 7, 733–742.

Bhattacharya, A. and D. B. Dunson (2011). "Sparse Bayesian infinite factor models". *Biometrika* 98:2, 291–306.

Binder, D. A. (1978). "Bayesian cluster analysis". *Biometrika* 65:1, 31–38.

Black Box Study Results (July 2017).

Blackwell, D. and J. B. MacQueen (1973). "Ferguson distributions via Pólya urn schemes". *The Annals of Statistics* 1:2, 353–355.

Bonventre, C. L. (2021). "Wrongful convictions and forensic science". *Wiley Interdisciplinary Reviews: Forensic Science* 3:4, e1406.

Bradlow, E. T., H. Wainer, and X. Wang (1999). "A Bayesian random effects model for testlets". *Psychometrika* 64:2, 153–168.

Bradlow, E. T. and A. M. Zaslavsky (1999). "A hierarchical latent variable model for ordinal data from a customer satisfaction survey with "no answer" responses". *Journal of the American Statistical Association* 94:445, 43–52.

Bradlow, E. T. (1994). *Analysis of ordinal survey data with "no answer" responses*. Harvard University.

Breen, R. and R. Luijkx (2010). "Mixture Models for Ordinal Data". *Sociological Methods & Research* 39:1, 3–24. DOI: 10.1177/0049124110366240.

Bush, C. A. and S. N. MacEachern (1996). "A semiparametric Bayesian model for randomised block designs". *Biometrika* 83:2, 275–285.

Byrt, T., J. Bishop, and J. B. Carlin (1993). "Bias, prevalence and kappa". *Journal of Clinical Epidemiology* 46:5, 423–429.

Carlin, B. P. and T. A. Louis (2008). *Bayesian Methods for Data Analysis*. CRC Press, Boca Raton.

Carvalho, C. M., J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West (2008). "High-dimensional sparse factor modeling: applications in gene expression genomics". *Journal of the American Statistical Association* 103:484, 1438–1456.

Casella, G. and R. L. Berger (2021). *Statistical Inference*. Cengage Learning, Boston.

Cohen, J. (1960). "A coefficient of agreement for nominal scales". *Educational and Psychological Measurement* 20:1, 37–46.

Cohen, J. (1968). "Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit." *Psychological bulletin* 70:4, 213.

Conti, G., S. Frühwirth-Schnatter, J. J. Heckman, and R. Piatek (2014). "Bayesian exploratory factor analysis". *Journal of Econometrics* 183:1, 31–57.

Cowles, M. K. (1996). "Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models". *Statistics and Computing* 6, 101–111.

Cronbach, L. J. (1951). "Coefficient alpha and the internal structure of tests". *Psychometrika* 16:3, 297–334.

Czamara, D., E. Tissink, J. Tuhkanen, J. Martins, Y. Awaloff, A. J. Drake, B. Khulan, A. Palotie, S. M. Winter, C. B. Nemeroff, et al. (2021). "Combined effects of genotype and childhood adversity shape variability of DNA methylation across age". *Translational Psychiatry* 11:1, 88.

Dahl, D. B., D. J. Johnson, and P. Müller (2022). "Search algorithms and loss functions for Bayesian clustering". *Journal of Computational and Graphical Statistics* 31:4, 1189–1201.

Dahl, D. B., D. J. Johnson, and P. Müller (2021). *salso: "Search Algorithms and Loss Functions for Bayesian Clustering"*. R package version 0.3.0.

Dai, H., J. S. Leeder, and Y. Cui (2014). "A modified generalized Fisher method for combining probabilities from dependent tests". *Frontiers in Genetics* 5, 32.

Davis, E. P., R. Korja, L. Karlsson, L. M. Glynn, C. A. Sandman, B. Vegetabile, E.-L. Kataja, S. Nolvi, E. Sinervä, J. Pelto, et al. (2019). "Across continents and demographics, unpredictable maternal signals are associated with children's cognitive function". *EBioMedicine* 46, 256–263.

Davis, E. P., K. McCormack, H. Arora, D. Sharpe, A. K. Short, J. Bachevalier, L. M. Glynn, C. A. Sandman, H. S. Stern, M. Sanchez, et al. (2022). "Early life exposure to unpredictable parental sensory signals shapes cognitive development across three species". *Frontiers in Behavioral Neuroscience* 16.

Davis, E. P., S. A. Stout, J. Molet, B. Vegetabile, L. M. Glynn, C. A. Sandman, K. Heins, H. Stern, and T. Z. Baram (2017). "Exposure to unpredictable maternal sensory signals influences cognitive development across species". *Proceedings of the National Academy of Sciences* 114:39, 10390–10395.

Delgado, R. and X.-A. Tibau (2019). "Why Cohen's Kappa should be avoided as performance measure in classification". *PloS one* 14:9, e0222916.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". *Journal of the Royal Statistical Society: Series B (Methodological)* 39:1, 1–22.

Dewhurst, T., B. Found, D. Rogers, et al. (2007). "The relationship between quantitatively modelled signature complexity levels and forensic document examiners' qualitative opinions on casework". *Journal of Forensic Document Examination* 18, 21–40.

Du, P., X. Zhang, C.-C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin (2010). "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis". *BMC Bioinformatics* 11:1, 1–9.

Duong, T. (2013). *Generalized Probabilistic Biclustering for Pattern Recognition*. University of California, Irvine.

Durina, M. and M. Caligiuri (2009). "The determination of authorship from a homogenous group of writers". *Journal of the American Society of Questioned Document Examiners* 12:2, 77–90.

Erkanli, A., D. Stangl, and P. Müller (1993). "A Bayesian analysis of ordinal data using mixtures".

Escobar, M. D. (1994). "Estimating normal means with a Dirichlet process prior". *Journal of the American Statistical Association* 89:425, 268–277.

Escobar, M. D. and M. West (1995). "Bayesian density estimation and inference using mixtures". *Journal of the American Statistical Association* 90:430, 577–588.

Federal Bureau of Investigation (2015). *Testimony on Microscopic Hair Analysis Contained Errors in at Least 90 Percent of Cases in Ongoing Review*.

Feinstein, A. R. and D. V. Cicchetti (1990). "High agreement but low kappa: I. The problems of two paradoxes". *Journal of Clinical Epidemiology* 43:6, 543–549.

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems". *The Annals of Statistics* 1:2, 209–230.

Ferguson, T. S. (1974). "Prior distributions on spaces of probability measures". *The Annals of Statistics* 2:4, 615–629.

Fisher, R. A. (1922). "On the interpretation of χ 2 from contingency tables, and the calculation of P". *Journal of the Royal Statistical Society* 85:1, 87–94.

Fisher, R. A. (1992). *Statistical Methods for Research Workers*. Springer, New York, 66–70.

Fleiss, J. L. (1971). "Measuring nominal scale agreement among many raters." *Psychological bulletin* 76:5, 378–382.

Found, B. and D. Rogers (1996). "The forensic investigation of signature complexity". *Handwriting and Drawing Research: Basic and Applied Issues*, 483–492.

Found, B., D. Rogers, V. Rowe, and D. Dick (1998). "Statistical modelling of experts' perceptions of the ease of signature simulation". *Journal of Forensic Document Examination* 11, 73–99.

Frühwirth-Schnatter, S. and H. F. Lopes (2018). "Sparse Bayesian factor analysis when the number of factors is unknown". *arXiv preprint arXiv:1804.04231*.

Furlan, J. C., M. G. Fehlings, E. M. Massicotte, B. Aarabi, A. R. Vaccaro, C. M. Bono, I. Madrazo, C. Villanueva, J. N. Grauer, and D. Mikulis (2007). "A quantitative and reproducible method to assess cord compression and canal stenosis after cervical spine trauma: a study of interrater and intrarater reliability". *Spine* 32:19, 2083–2091.

Gadermann, A. M., M. Guhn, and B. D. Zumbo (2012). "Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide". *Practical Assessment, Research, and Evaluation* 17:1, 3.

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)". *Bayesian Analysis* 1:3, 515–534.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis (3rd edition)*. CRC press, Boca Raton.

Gelman, A., G. O. Roberts, W. R. Gilks, et al. (1996). "Efficient Metropolis jumping rules". *Bayesian Statistics* 5, 599–607.

Gelman, A. and D. B. Rubin (1992). "Inference from iterative simulation using multiple sequences". *Statistical Science* 7:4, 457–472.

Geman, S. and D. Geman (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 721–741.

George, E. I. and R. E. McCulloch (1997). "Approaches for Bayesian variable selection". *Statistica Sinica* 7:2, 339–373.

Geweke, J. and G. Zhou (1996). "Measuring the pricing error of the arbitrage pricing theory". *The review of financial studies* 9:2, 557–587.

Ghahramani, Z. and T. Griffiths (2005). "Infinite latent feature models and the Indian buffet process". *Advances in Neural Information Processing Systems* 18.

Glynn, L. M., M. A. Howland, C. A. Sandman, E. P. Davis, M. Phelan, T. Z. Baram, and H. S. Stern (2018). "Prenatal maternal mood patterns predict child temperament and adolescent mental health". *Journal of Affective Disorders* 228, 83–90.

Gorsuch, R. L. (2014). *Factor Analysis: Classic Edition*. Routledge, New York.

Görür, D. and C. E. Rasmussen (2010). "Dirichlet process gaussian mixture models: Choice of the base distribution". *Journal of Computer Science and Technology* 25:4, 653–664.

Griffiths, T. L. and Z. Ghahramani (2011). "The Indian Buffet Process: An Introduction and Review." *Journal of Machine Learning Research* 12:4.

Guo, X. and J. T. Kwok (2016). "Aggregating crowdsourced ordinal labels via Bayesian clustering". *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*. Springer, 426–442.

Hallgren, K. A. (2012). "Computing inter-rater reliability for observational data: an overview and tutorial". *Tutorials in Quantitative Methods for Psychology* 8:1, 23.

Hannah, L. A., D. M. Blei, and W. B. Powell (2011). "Dirichlet process mixtures of generalized linear models." *Journal of Machine Learning Research* 12:6, 1923–1953.

Harman, H. H. (1976). *Modern Factor Analysis*. University of Chicago Press, Chicago.

Hartigan, J. A. and M. A. Wong (1979). "Algorithm AS 136: A k-means clustering algorithm". *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28:1, 100–108.

Heydorn, A., B. K. Ersbøll, M. Hentzer, M. R. Parsek, M. Givskov, and S. Molin (2000). "Experimental reproducibility in flow-chamber biofilms". *Microbiology* 146:10, 2409–2415.

Hicklin, R. A., L. Eisenhart, N. Richetelli, M. D. Miller, P. Belcastro, T. M. Burkes, C. L. Parks, M. A. Smith, J. Buscaglia, E. M. Peters, et al. (2022a). "Accuracy and reliability

of forensic handwriting comparisons". *Proceedings of the National Academy of Sciences* 119:32, e2119944119.

Hicklin, R. A., B. C. McVicker, C. Parks, J. LeMay, N. Richetelli, M. Smith, J. Buscaglia, R. S. Perlman, E. M. Peters, and B. A. Eckenrode (2022b). "Accuracy, Reproducibility, and Repeatability of Forensic Footwear Examiner Decisions". *Forensic Science International* 339, 111418.

Hicklin, R. A., B. T. Ulery, M. Ausdemore, and J. Buscaglia (2020). "Why do latent fingerprint examiners differ in their conclusions?" *Forensic Science International* 316, 110542.

Hicklin, R. A., K. R. Winer, P. E. Kish, C. L. Parks, W. Chapman, K. Dunagan, N. Richetelli, E. G. Epstein, M. A. Ausdemore, and T. A. Busey (2021). "Accuracy and reproducibility of conclusions by forensic bloodstain pattern analysts". *Forensic Science International* 325, 110856.

Holsclaw, T., B. Shahbaba, and D. Gillen (2012). "Quantifying the association between longitudinal changes in serum albumin and mortality via a Gaussian process model".

Hsu, S. S. (2012). "Justice Dept., FBI to review use of forensic evidence in thousands of cases". *The Washington Post*.

Hubin, A., G. O. Storvik, P. E. Grini, and M. A. Butenko (2020). "A Bayesian binomial regression model with latent Gaussian processes for modelling DNA methylation". *arXiv preprint arXiv:2004.13689*.

Hüls, A. and D. Czamara (2020). "Methodological challenges in constructing DNA methylation risk scores". *Epigenetics* 15:1-2, 1–11.

Ibrahim, J. G. and K. P. Kleinman (1998). "Semiparametric Bayesian methods for random effects models". *Practical Nonparametric and Semiparametric Bayesian Statistics*, 89–114.

Ishwaran, H. and L. F. James (2001). "Gibbs sampling methods for stick-breaking priors". *Journal of the American Statistical Association* 96:453, 161–173.

Ishwaran, H. and L. F. James (2002). "Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information". *Journal of Computational and Graphical Statistics* 11:3, 508–532.

Ishwaran, H. and J. S. Rao (2005). "Spike and slab variable selection: frequentist and Bayesian strategies". *The Annals of Statistics* 33:2, 730–773.

Ishwaran, H. and M. Zarepour (2000). "Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models". *Biometrika* 87:2, 371–390.

Ishwaran, H. and M. Zarepour (2002). "Exact and approximate sum representations for the Dirichlet process". *Canadian Journal of Statistics* 30:2, 269–283.

Jakobsson, U. and A. Westergren (2005). "Statistical methods for assessing agreement for ordinal data". *Scandinavian Journal of Caring Sciences* 19:4, 427–431.

Jiang, S., N. Kamei, J. L. Bolton, X. Ma, H. S. Stern, T. Z. Baram, and A. Mortazavi (2019). "Intra-individual methylomics detects the impact of early-life adversity". *Life Science Alliance* 2:2, e201800204.

Johnson, V. E. and J. H. Albert (2006). *Ordinal Data Modeling*. Springer Science & Business Media, New York.

Johnson, V. E., R. O. Deaner, and C. P. Van Schaik (2002). "Bayesian analysis of rank data with application to primate intelligence experiments". *Journal of the American Statistical Association* 97:457, 8–17.

Johnson, V. E. (1996). "On Bayesian Analysis of Multirater Ordinal Data: An Application to Automated Essay Grading". *Journal of the American Statistical Association* 91:433, 42–51. ISSN: 01621459.

Jones, K., J. Buckwalter, E. McCarthy, B. DeYoung, G. El-Khoury, L. Dolan, F. Gannon, C. Inwards, M. Klein, M. Kyriakus, A. Rosenberg, G. Siegal, K. Unni, L. Fayad, M. Kransdorf, M. Murphey, D. Panicek, D. Rubin, M. Sundararri, and D. Vanel (2007). "Reliability of histopathologic and radiologic grading of cartilaginous neoplasms in long

bones". English (US). *Journal of Bone and Joint Surgery* 89:10, 2113–2123. ISSN: 0021-9355. DOI: 10.2106/JBJS.F.01530.

Kaiser, H. F. (1958). "The varimax criterion for analytic rotation in factor analysis". *Psychometrika* 23:3, 187–200.

Kam, M., G. Fielding, and R. Conn (1997). "Writer identification by professional document examiners". *Journal of Forensic Sciences* 42:5, 778–786.

Kam, M., K. Gummadidala, G. Fielding, and R. Conn (2001). "Signature authentication by forensic document examiners". *Journal of Forensic Science* 46:4, 884–888.

Kam, M. and E. Lin (2003). "Writer identification using hand-printed and non-hand-printed questioned documents." *Journal of Forensic Sciences* 48:6, 1391–1395.

Kam, M., J. Wetstein, and R. Conn (1994). "Proficiency of professional document examiners in writer identification". *Journal of Forensic Science* 39:1, 5–14.

Katrinli, S., A. X. Maihofer, A. H. Wani, J. R. Pfeiffer, E. Ketema, A. Ratanatharathorn, D. G. Baker, M. P. Boks, E. Geuze, R. C. Kessler, et al. (2022). "Epigenome-wide meta-analysis of PTSD symptom severity in three military cohorts implicates DNA methylation changes in genes involved in immune system and oxidative stress". *Molecular Psychiatry* 27:3, 1720–1728.

Kim, J.-O., O. Ahtola, P. E. Spector, J.-O. Kim, and C. W. Mueller (1978). *Introduction to Factor Analysis: What it is and how to do it*. 13. Sage, New York.

Knowles, D. and Z. Ghahramani (2007). "Infinite sparse factor analysis and infinite independent components analysis". *Independent Component Analysis and Signal Separation: 7th International Conference, ICA 2007, London, UK, September 9-12, 2007. Proceedings 7*. Springer, 381–388.

Krippendorff, K. (2011). "Computing Krippendorff's alpha-reliability".

Krueger, F. and S. R. Andrews (2011). "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications". *Bioinformatics* 27:11, 1571–1572.

Lopes, H. F. and M. West (2004). "Bayesian model assessment in factor analysis". *Statistica Sinica* 14:1, 41–67.

Luby, A., A. Mazumder, and B. Junker (2020). "Psychometric analysis of forensic examiner behavior". *Behaviormetrika* 47:2, 355–384.

Luby, A., A. Mazumder, and B. Junker (2021). "Psychometrics for Forensic Fingerprint Comparisons". *Quantitative Psychology*, 385–397.

Luby, A. S. and J. B. Kadane (2018). "Proficiency testing of fingerprint examiners with Bayesian Item Response Theory". *Law, Probability and Risk* 17:2, 111–121.

Ma, Z. and A. E. Teschendorff (2013). "A variational Bayes beta mixture model for feature selection in DNA methylation studies". *Journal of Bioinformatics and Computational Biology* 11:04, 1350005.

Ma, Z., A. E. Teschendorff, H. Yu, J. Taghia, and J. Guo (2014). "Comparisons of non-gaussian statistical models in DNA methylation analysis". *International Journal of Molecular Sciences* 15:6, 10835–10854.

MacEachern, S. N. (1994). "Estimating normal means with a conjugate style Dirichlet process prior". *Communications in Statistics-Simulation and Computation* 23:3, 727–741.

MacEachern, S. N. and P. Müller (1998). "Estimating mixture of Dirichlet process models". *Journal of Computational and Graphical Statistics* 7:2, 223–238.

Matechou, E., I. Liu, D. Fernández, M. Farias, and B. Gjelsvik (2016). "Biclustering models for two-mode ordinal data". *Psychometrika* 81:3, 611–624.

McHugh, M. L. (2012). "Interrater reliability: the kappa statistic". *Biochemia Medica* 22:3, 276–282.

McLachlan, G. J. and K. E. Basford (1988). *Mixture Models: Inference and Applications to Clustering*. Vol. 38. Marcel Dekker, New York.

Meilă, M. (2007). "Comparing clusterings—an information based distance". *Journal of multivariate analysis* 98:5, 873–895.

Meng, X.-L. and D. B. Rubin (1993). "Maximum likelihood estimation via the ECM algorithm: A general framework". *Biometrika* 80:2, 267–278.

Mitchell L. L., M. (2016). "A blind study on the reliability of hand printing identification by forensic document examiners". *Journal of the American Society of Questioned Document Examiners* 19:1, 25–31.

Monson, K. L., E. D. Smith, and E. M. Peters (2023a). "Accuracy of comparison decisions by forensic firearms examiners". *Journal of Forensic Sciences* 68:1, 86–100.

Monson, K. L., E. D. Smith, and E. M. Peters (2023b). "Accuracy of comparison decisions by forensic firearms examiners". *Journal of Forensic Sciences* 68:1, 86–100.

Moore, L. D., T. Le, and G. Fan (2013). "DNA methylation and its basic function". *Neuropsychopharmacology* 38:1, 23–38.

Mukhopadhyay, S. and A. E. Gelfand (1997). "Dirichlet process mixed generalized linear models". *Journal of the American Statistical Association* 92:438, 633–639.

National Research Council (2009). *Strengthening Forensic Science in the United States: A Path Forward*. National Academies Press. bit.ly/2EF1uKC.

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models". *Journal of Computational and Graphical Statistics* 9:2, 249–265.

Nelson, K. P. and D. Edwards (2015). "Measures of agreement between many raters for ordinal classifications". *Statistics in Medicine* 34:23, 3116–3132.

Office of the Inspector General (2006). "Review of the FBI's Handling of the Brandon Mayfield Case". *Office of the Inspector General, Oversight and Review Division, US Department of Justice*, 1–330.

Pearson, K. (1900). "I. Mathematical contributions to the theory of evolution.—VII. On the correlation of characters not quantitatively measurable". *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 195:262-273, 1–47.

Pearson, R., T. Kurien, K. Shu, and B. Scammell (2011). "Histopathology grading systems for characterisation of human knee osteoarthritis–reproducibility, variability, reliability, correlation, and validity". *Osteoarthritis and Cartilage* 19:3, 324–331.

Peterson, J. L., M. J. Hickman, K. J. Strom, and D. J. Johnson (2013). "Effect of forensic evidence on criminal justice case processing". *Journal of Forensic Sciences* 58, S78–S90.

Plummer, M., A. Stukalov, M. Denwood, and M. M. Plummer (2019). "Package 'rjags'". *Update* 1.

President's Council of Advisors on Science and Technology (2016). "Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods".

Raadt, A. de, M. J. Warrens, R. J. Bosker, and H. A. Kiers (2021). "A comparison of reliability coefficients for ordinal rating scales". *Journal of Classification*, 1–25.

Radloff, L. S. (1977). "The CES-D scale: A self-report depression scale for research in the general population". *Applied Psychological Measurement* 1:3, 385–401.

Ranalli, M. and R. Rocci (2016). "Mixture models for ordinal data: a pairwise likelihood approach". *Statistics and Computing* 26, 529–547.

Rand, W. M. (1971). "Objective criteria for the evaluation of clustering methods". *Journal of the American Statistical Association* 66:336, 846–850.

Rasmussen, C. (1999). "The infinite Gaussian mixture model". *Advances in Neural Information Processing Systems* 12, 554–560.

Rastelli, R. and N. Friel (2018). "Optimal Bayesian estimators for latent variable cluster models". *Statistics and Computing* 28:6, 1169–1186.

Richardson, S., G. C. Tseng, and W. Sun (2016). "Statistical methods in integrative genomics". *Annual Review of Statistics and its Application* 3, 181–209.

Ročková, V. and E. I. George (2014). "EMVS: The EM approach to Bayesian variable selection". *Journal of the American Statistical Association* 109:506, 828–846.

Ročková, V. and E. I. George (2016). "Fast Bayesian factor analysis via automatic rotations to sparsity". *Journal of the American Statistical Association* 111:516, 1608–1622.

Ročková, V. and E. I. George (2018). "The spike-and-slab lasso". *Journal of the American Statistical Association* 113:521, 431–444.

Rost, J. (1990). "Rasch models in latent classes: An integration of two approaches to item analysis". *Applied Psychological Measurement* 14:3, 271–282.

Rousseeuw, P. J. (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". *Journal of Computational and Applied Mathematics* 20, 53–65.

Rummel, R. J. (1988). *Applied Factor Analysis*. Northwestern University Press, Evanston.

Santor, D. A. and J. C. Coyne (1997). "Shortening the CES–D to improve its ability to detect cases of depression". *Psychological Assessment* 9:3, 233–243.

Schielzeth, H., N. J. Dingemanse, S. Nakagawa, D. F. Westneat, H. Allegue, C. Teplitsky, D. Réale, N. A. Dochtermann, L. Z. Garamszegi, and Y. G. Araya-Ajoy (2020). "Robustness of linear mixed-effects models to violations of distributional assumptions". *Methods in Ecology and Evolution* 11:9, 1141–1152.

Scurich, N. (2022). "Inconclusives in firearm error rate studies are not "a pass"". *Law, Probability, and Risk* 21:2, 123–127.

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors". *Statistica Sinica* 4:2, 639–650.

Shahbaba, B. and R. Neal (2009). "Nonlinear models using Dirichlet process mixtures". *Journal of Machine Learning Research* 10:8, 1829–1850.

Shannon, C. E. (1948). "A mathematical theory of communication". *The Bell system technical journal* 27:3, 379–423.

Short, A. K., R. Weber, N. Kamei, C. Wilcox-Thai, H. Arora, A. Mortazavi, H. Stern, L. Glynn, and T. Z. Baram (2023). "Novel, within-subject methylome analyses identify individual signatures of early-life adversities, predicting neuropsychiatric outcome." *Unpublished Manuscript*.

Short, A. K. and T. Z. Baram (2019). "Early-life adversity and neurological disease: age-old questions and novel answers". *Nature Reviews Neurology* 15:11, 657–669.

Shrout, P. E. and J. L. Fleiss (1979). "Intraclass correlations: uses in assessing rater reliability." *Psychological Bulletin* 86:2, 420–428.

Siegmund, K. D., P. W. Laird, and I. A. Laird-Offringa (2004). "A comparison of cluster analysis methods using DNA methylation data". *Bioinformatics* 20:12, 1896–1904.

Sita, J., B. Found, and D. K. Rogers (2002). "Forensic handwriting examiners' expertise for signature comparison". *Journal of Forensic Sciences* 47:5, 1117–1124.

Smith, A. K., A. Ratanatharathorn, A. X. Maihofer, R. K. Naviaux, A. E. Aiello, A. B. Amstadter, A. E. Ashley-Koch, D. G. Baker, J. C. Beckham, M. P. Boks, et al. (2020). "Epigenome-wide meta-analysis of PTSD across 10 military and civilian cohorts identifies methylation changes in AHRR". *Nature Communications* 11:1, 5965.

Snedecor, G. W. and W. G. Cochran (1989). "Statistical Methods". *Ames: Iowa State Univ. Press Iowa* 54, 71–82.

Spoorenberg, A., K. De Vlam, S. van der Linden, M. Dougados, H. Mielants, H. van de Tempel, and D. van der Heijde (2004). "Radiological scoring methods in ankylosing spondylitis. Reliability and change over 1 and 2 years." *The Journal of Rheumatology* 31:1, 125–132.

Stan Development Team (2022). *RStan: the R interface to Stan*. R package version 2.21.5.

Stephens, M. (2000). "Dealing with label switching in mixture models". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62:4, 795–809.

Stern, H. S., M. Angel, M. Cavanaugh, S. Zhu, and E. L. Lai (2018). "Assessing the complexity of handwritten signatures". *Law, Probability and Risk* 17:2, 123–132.

Tabachnick, B. G. and L. S. Fidell (2007). *Experimental Designs Using ANOVA*. Vol. 724. Thomson/Brooks/Cole, Belmont.

Teh, Y. W. (2010). "Dirichlet Process". *Encyclopedia of machine learning* 1063, 280–287.

Teh, Y. W., D. Grür, and Z. Ghahramani (2007). "Stick-breaking construction for the Indian buffet process". *Artificial Intelligence and Statistics*. PMLR, 556–563.

Thompson, W. C. (2017). "How should forensic scientists present source conclusions". *Seton Hall L. Rev.* 48, 773.

Thompson, W. C., R. H. Grady, E. Lai, and H. S. Stern (2018). "Perceived strength of forensic scientists' reporting statements about source conclusions". *Law, Probability and Risk* 17:2, 133–155.

Tipping, M. E. and C. M. Bishop (1999). "Probabilistic principal component analysis". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61:3, 611–622.

Tsai, P. (1988). "Variable gauge repeatability and reproducibility study using the analysis of variance method". *Quality Engineering* 1:1, 107–115.

Ulery, B. T., R. A. Hicklin, J. Buscaglia, and M. A. Roberts (2011). "Accuracy and reliability of forensic latent fingerprint decisions". *Proceedings of the National Academy of Sciences* 108:19, 7733–7738.

Ulery, B. T., R. A. Hicklin, J. Buscaglia, and M. A. Roberts (2012). "Repeatability and reproducibility of decisions by latent fingerprint examiners". *PloS One* 7:3, e32800.

Ulery, B. T., R. A. Hicklin, M. A. Roberts, and J. Buscaglia (2014). "Measuring what latent fingerprint examiners consider sufficient information for individualization determinations". *PloS One* 9:11, e110179.

Ulery, B. T., R. A. Hicklin, M. A. Roberts, and J. Buscaglia (2015). "Changes in latent fingerprint examiners' markup between analysis and comparison". *Forensic Science International* 247, 54–61.

Ulery, B. T., R. A. Hicklin, M. A. Roberts, and J. Buscaglia (2016). "Interexaminer variation of minutia markup on latent fingerprints". *Forensic Science International* 264, 89–99.

Van Wieringen, W. N. and J. De Mast (2008). "Measurement system analysis for binary data". *Technometrics* 50:4, 468–478.

Vardeman, S. B. (2014). "Gauge Repeatability and Reproducibility (R & R) Studies". *Wiley StatsRef: Statistics Reference Online*.

Vardeman, S. B. and E. S. VanValkenburg (1999). "Two-way random-effects analyses and gauge R&R studies". *Technometrics* 41:3, 202–211.

Vegetabile, B. G., S. A. Stout-Oswald, E. P. Davis, T. Z. Baram, and H. S. Stern (2019). "Estimating the entropy rate of finite Markov chains with application to behavior studies". *Journal of Educational and Behavioral Statistics* 44:3, 282–308.

Wade, S. and Z. Ghahramani (2018). "Bayesian cluster analysis: Point estimation and credible balls (with discussion)". *Bayesian Analysis* 13:2, 559–626.

Weaver, B. P., M. S. Hamada, S. B. Vardeman, and A. G. Wilson (2012). "A Bayesian approach to the analysis of gauge R&R data". *Quality Engineering* 24:4, 486–500.

Willis, S., L. McKenna, S. McDermott, G. O'Donell, A. Barrett, B. Rasmusson, A. Nordgaard, C. Berger, M. Sjerps, J. Lucena-Molina, et al. (2015). "Strengthening the Evaluation of Forensic Results Across Europe (STEOFRAE), ENFSI Guideline for Evaluative Reporting in Forensic Science".

Ypma, T. J. (1995). "Historical development of the Newton-Raphson method". *SIAM review* 37:4, 531–551.

Zhuang, J., M. Widschwendter, and A. E. Teschendorff (2012). "A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform". *BMC bioinformatics* 13, 1–14.

Zumbo, B. D., A. M. Gadermann, and C. Zeisser (2007). "Ordinal versions of coefficients alpha and theta for Likert rating scales". *Journal of Modern Applied Statistical Methods*.

Appendix A

Appendix to Chapter 3

A.1 Full conditionals for Continuous Data

As described in Section 3.3.1, we take a Bayesian approach to inference using MCMC to obtain samples from the posterior distribution. For the continuous data model given by the equation (3.1) we apply a Gibbs sampling algorithm (Geman and Geman, 1984). This appendix provides the full conditional distributions of each parameter conditional on all the other parameters and the data. A flat (uninformative) prior distribution was assumed for μ and the standard deviation parameters σ_α, σ_γ and σ_δ.

$$p(\mu, \sigma_\alpha, \sigma_\gamma, \sigma_\delta) \propto 1$$

The full conditionals for μ, α_i, γ_j, δ_{ij}, σ^2_α, σ^2_γ and σ^2_δ are:

$$\mu|\text{the rest} \sim N(\frac{\sum_i \sum_j \sum_k (Y_{ijk} - \alpha_i - \gamma_j - \delta_{ij})}{\sum_i \sum_j \sum_k \mathbb{1}_{ijk}}, \frac{\sigma^2_\epsilon}{\sum_i \sum_j \sum_k \mathbb{1}_{ijk}})$$

$$\alpha_i|\text{the rest} \sim N(\frac{\sum_j \sum_k (Y_{ijk} - \mu - \gamma_j - \delta_{ij})}{\frac{1}{\sigma^2_\alpha} + \sum_j \sum_k \frac{\mathbb{1}_{ijk}}{\sigma^2_\epsilon}}, \frac{1}{\frac{1}{\sigma^2_\alpha} + \sum_j \sum_k \frac{\mathbb{1}_{ijk}}{\sigma^2_\epsilon}})$$

$$\gamma_j|\text{the rest} \sim N(\frac{\sum_i \sum_k (Y_{ijk} - \mu - \alpha_i - \delta_{ij})}{\frac{1}{\sigma^2_\gamma} + \sum_i \sum_k \frac{\mathbb{1}_{ijk}}{\sigma^2_\epsilon}}, \frac{1}{\frac{1}{\sigma^2_\gamma} + \sum_i \sum_k \frac{\mathbb{1}_{ijk}}{\sigma^2_\epsilon}})$$

$$\delta_{ij}|\text{the rest} \sim N(\frac{\sum_k (Y_{ijk} - \mu - \alpha_i - \gamma_j)}{\frac{1}{\sigma^2_\delta} + \sum_k \frac{\mathbb{1}_{ijk}}{\sigma^2_\epsilon}}, \frac{1}{\frac{1}{\sigma^2_\delta} + \sum_k \frac{\mathbb{1}_{ijk}}{\sigma^2_\epsilon}})$$

$$\sigma^2_\alpha|\text{the rest} \sim \text{Inv-Gamma}(\frac{I-}{2}, \frac{\sum_i \alpha_i^2}{2})$$

$$\sigma^2_\gamma|\text{the rest} \sim \text{Inv-Gamma}(\frac{J-1}{2}, \frac{\sum_j \gamma_j^2}{2})$$

$$\sigma^2_\delta|\text{the rest} \sim \text{Inv-Gamma}(\frac{IJ-1}{2}, \frac{\sum_i \sum_j \delta_{ij}^2}{2})$$

$$\sigma^2_\epsilon|\text{the rest} \sim \text{Inv-Gamma}(\frac{\sum_i \sum_j \sum_k \mathbb{1}_{ijk} - 1}{2}, \frac{\sum_i \sum_j \sum_k (Y_{ijk} - \mu - \alpha_i - \gamma_j - \delta_{ij})^2}{2})$$

Note that $\mathbb{1}_{ijk}$ is an indicator function based on whether Y_{ijk} is an available observation for examiner i, sample j in the kth repetition. Given these full conditionals, a Gibbs sampler can be used to iterate over these with a systematic scan.

A.2 Full conditionals for Binary Data

We now describe the full conditionals to sample from the posterior distribution for the binary data model given by the equation (3.3). The goal is to sample from the posterior distribution of the

latent variables and model parameters:

$$p(Z_{ijk}, \mu, \alpha_i, \gamma_j, \delta_{ij}, \sigma^2_\alpha, \sigma^2_\gamma, \sigma^2_\delta | Y_{ijk}) \propto$$

$$p(Y_{ijk}, Z_{ijk}, \mu, \alpha_i, \gamma_j, \delta_{ij}, \sigma^2_\alpha, \sigma^2_\gamma, \sigma^2_\delta) =$$

$$\prod_{i=1}^{I}\prod_{j=1}^{J_i}\prod_{k=1}^{K_i} p(Y_{ijk}|Z_{ijk}) \prod_{i=1}^{I}\prod_{j=1}^{J_i}\prod_{k=1}^{K_i} p(Z_{ijk}|\mu, \alpha_i, \gamma_j, \delta_{ij})$$

$$\prod_{i=1}^{I} p(\alpha_i|\sigma^2_\alpha) \prod_{j=1}^{J_i} p(\gamma_j|\sigma^2_\gamma) \prod_{i=1}^{I}\prod_{j=1}^{J} p(\delta_{ij}|\sigma^2_\delta) p(\mu, \sigma^2_\alpha, \sigma^2_\gamma, \sigma^2_\delta)$$

The distribution of latent variables conditional on the observed Y_{ijk} are truncated normal distributions:

$$Z_{ijk}|Y_{ijk}, \mu, \alpha_i, \gamma_j, \delta_{ij} = \begin{cases} N(\mu + \alpha_i + \gamma_j + \delta_{ij}, 1)\, \mathbb{I}_{Z_{ijk}<0}\,, & \text{if } Y_{ijk} = 0 \\ N(\mu + \alpha_i + \gamma_j + \delta_{ij}, 1)\, \mathbb{I}_{Z_{ijk}>0}\,, & \text{if } Y_{ijk} = 1 \end{cases}$$

The full conditional distributions for other parameters are based on the latent variables Z_{ijk} and can be found using the approach of the previous section:

$$\mu|\text{the rest} \sim N(\overline{Z}_{...} - \overline{\alpha}_{.} - \overline{\gamma}_{.} - \overline{\delta}_{..}, \frac{1}{\sum_i \sum_j \sum_k \mathbb{1}_{ijk}})$$

$$\alpha_i|\text{the rest} \sim N(\frac{\sum_j \sum_k (Z_{ijk} - \mu - \gamma_j - \delta_{ij})}{\frac{1}{\sigma^2_\alpha} + \sum_j \sum_k \mathbb{1}_{ijk}}, \frac{1}{\frac{1}{\sigma^2_\alpha} + \sum_j \sum_k \mathbb{1}_{ijk}})$$

$$\gamma_j|\text{the rest} \sim N(\frac{\sum_i \sum_k (Z_{ijk} - \mu - \alpha_i - \delta_{ij})}{\frac{1}{\sigma^2_\gamma} + \sum_i \sum_k \mathbb{1}_{ijk}}, \frac{1}{\frac{1}{\sigma^2_\gamma} + \sum_i \sum_k \mathbb{1}_{ijk}})$$

$$\delta_{ij}|\text{the rest} \sim N(\frac{\sum_k (Z_{ijk} - \mu - \alpha_i - \gamma_j)}{\frac{1}{\sigma^2_\delta} + \sum_k \mathbb{1}_{ijk}}, \frac{1}{\frac{1}{\sigma^2_\delta} + \sum_k \mathbb{1}_{ijk}})$$

$$\sigma^2_\alpha|\text{the rest} \sim \text{Inv-Gamma}(\frac{I-1}{2}, \frac{\sum_i \alpha_i^2}{2})$$

$$\sigma^2_\gamma|\text{the rest} \sim \text{Inv-Gamma}(\frac{J-1}{2}, \frac{\sum_j \gamma_j^2}{2})$$

$$\sigma^2_\delta|\text{the rest} \sim \text{Inv-Gamma}(\frac{IJ-1}{2}, \frac{\sum_i \sum_j \delta_{ij}^2}{2})$$

A.3 Results under model misspecification

The models of Section 3 assume a Gaussian distribution for the outcomes Y or the latent variable Z. Therefore, it is crucial to assess the robustness of the presented methodology when the data-generating mechanism deviates from the model assumptions. Additionally, we must study how the violation of model assumptions affects our inferences about reliability. To do this we consider alternatives to the Gaussian when generating the data and use the normal model for analysis. We briefly describe statistical distributions that we consider as alternatives. We will introduce some statistical distributions before presenting the results under model misspecifications.

The Laplace distribution is a mirrored exponential distribution and even though the probability distribution for Laplace distribution looks similar to that of a normal distribution, it has lighter tails. Lighter tails mean that it is much less probable for a sample to be far away from the mean for a Laplace distribution compared to the normal distribution.

Student's t-distribution arises naturally for hypothesis testing about the mean of a normal distribution when the number of samples is small. It is also useful as an alternative to the Gaussian because it is bell-shaped like a normal distribution but has heavier tails. Heavier tails imply that it is more probable to sample away from the mean as compared to normal distributions. The t-distribution has one parameter, the degrees of freedom, with smaller values indicating heavier tails.

The normal, Laplace, and t-distributions are all symmetric around the mean. It is also important to study what happens when the data-generating mechanism deviates from symmetry. A generalized extreme value distribution (GEV) allows for the distribution to be asymmetric.

To assess the impact of model misspecification, we change different parts of the data-generating mechanism and try to fit the model as if the model assumptions are true. We compare the results obtained from the MCMC with the original parameter values and also compare the results for reliability. We simulate 25 data sets with a data-generating mechanism that is different from the model assumptions. We generate two decisions per examiner-sample pair so that there are repeated decisions on 100% of the samples. Firstly, we change the $Y_{ijk} \sim N(\mu + \alpha_i + \gamma_j + \delta_{ij}, \sigma^2_\epsilon)$ part of the

model given by the equation (2) by changing the error distribution from normal to: i) Laplace, ii) Student's t-distribution (ν= 5), iii) generalized extreme value distribution (GEV), iv) a bimodal distribution as a mixture of two normal distributions with different means and different variances. Additionally, we change the distribution of the examiner and sample random effects in the model given by equations (3) from Gaussian to Student's t-distributed random effects and fit them using the model in Section 3.1. The degrees of freedom chosen for the t-distribution were $\nu_\alpha = 4$ for examiner random effects and $\nu_\gamma = \frac{8}{3}$ so that the variances of these distributions were 2 and 4 respectively.

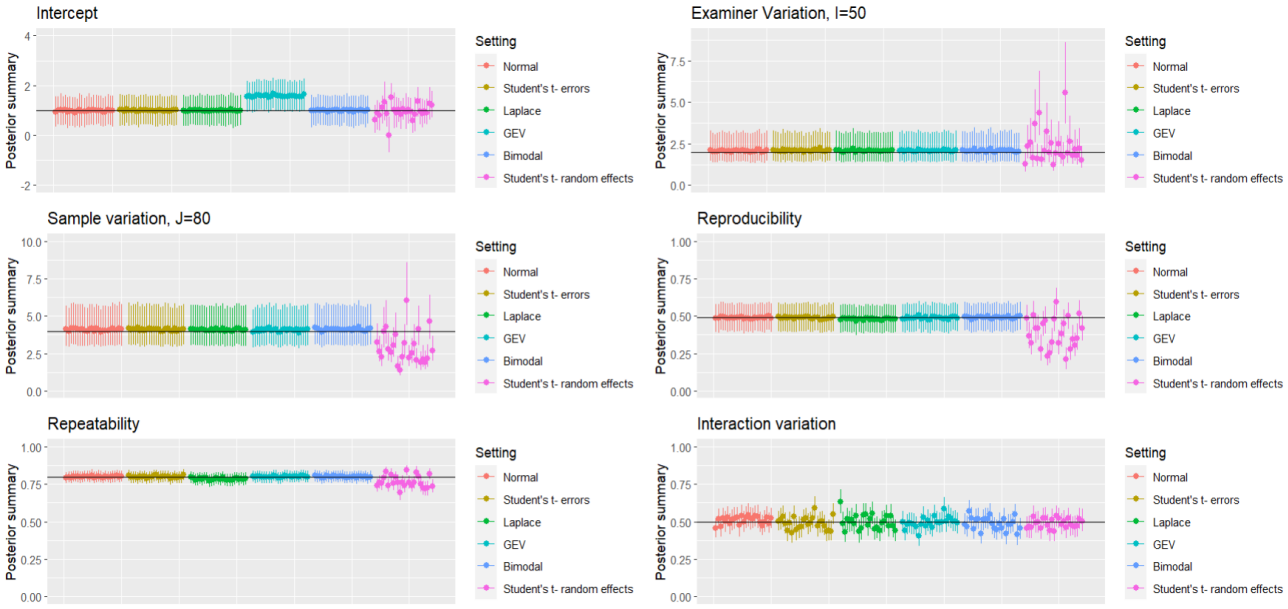


Figure A.1: Posterior medians and confidence intervals obtained by generating continuous data from alternate distributions and fitting them by the model given by the equations (2). The Normal (Gaussian) case represents the results from the example when the data is generated from the model assumptions in equations (2 and 3).

Misspecification	μ=1	σ^2_α=2	σ^2_γ =4	σ^2_δ =0.5	R_1=0.49	R_2=0.80
Normal	0.98	2.06	4.09	0.51	0.49	0.80
	(0.39, 1.58)	(1.41, 3.19)	(3.03, 5.72)	(0.44, 0.58)	(0.40, 0.58)	(0.76, 0.84)
Student's t	1.00	2.10	4.09	0.49	0.49	0.80
	(0.41, 3.21)	(1.44, 3.21)	(3.03, 5.74)	(0.42, 0.56)	(0.40, 0.58)	(0.76, 0.84)
Laplace	1.00	2.07	4.09	0.50	0.48	0.79
	(0.39, 1.61)	(1.41, 3.20)	(3.03, 5.70)	(0.43, 0.57)	(0.39, 0.57)	(0.75, 0.83)
GEV	1.58	2.08	4.06	0.50	0.49	0.80
	(0.98, 2.20)	(1.42, 3.21)	(3.01, 5.67)	(0.43, 0.57)	(0.40, 0.58)	(0.76, 0.84)
Bimodal	1.00	2.08	4.12	0.49	0.49	0.80
	(0.39, 1.59)	(1.42, 3.22)	(3.04, 5.75)	(0.42, 0.56)	(0.40, 0.59)	(0.76, 0.84)
Student's t random effects	0.96	2.31	2.90	0.49	0.39	0.77
	(0.38, 1.53)	(1.58, 3.58)	(2.14, 4.05)	(0.42, 0.56)	(0.31, 0.48)	(0.72, 0.81)

Table A.1: Effect of model misspecification on variance and reliability components. A total of 25 simulated data sets were used for inference in each case.

The inference for the intercept is comparable to the Normal case in all settings except when the errors have a generalized extreme value distribution. We also notice that the examiner variance, the sample variance, as well as the reliability components, are robust against the choice of error distribution. However, the posterior medians and credible intervals for σ^2_δ have more variance when the error distribution is changed from the Gaussian distribution. When the random effects are generated from a t-distribution, our model is unable to obtain good inference for variance components which affect the reliability components. This effect is observed especially for the sample variation. This issue is observed due to heavier tails in a t-distributed random variable and because ν_γ is smaller in comparison to ν_α, sample variation (σ^2_γ) has poorer inference compared to examiner variance (σ^2_α). It is reassuring that in most cases our model is robust to deviations from assumptions.

A.4 Effect of combining data sets on reliability

Since the repeatability studies (the second trials) are generally much smaller than the reproducibility studies (the first trials), it is interesting to assess how combining the data sets helps the inference for reproducibility. One might assume that since repeated decisions inform about interactions and add to the amount of data collected, the reproducibility estimate will have lower bias and narrower credible intervals. We conducted an experiment to answer the above question and found that the results confirm with our intuition. The inference for reproducibility is actually better (less bias and smaller credible intervals) when the data from both trials are used.

The experiment is conducted with $3^3 = 27$ combinations of variance values for examiner variation, sample variation, and interaction variation: low (0.2), medium (1), and high (5). The error variance is fixed to be 1 for all 27 data sets. One estimation technique only uses the reproducibility study, one uses only repeated measurements and one uses the combined data. Repetitions are obtained on 25% samples.

Reliability	Metric (average)	Data from first trial	Data from repeated decisions	Combined data set
Reproducibility	Absolute bias	0.072	0.097	**0.071**
Reproducibility	Range of credible interval	0.069	0.116	**0.067**
Repeatability	Absolute bias	0.359	0.067	**0.058**
Repeatability	Range of credible interval	0.120	0.067	**0.061**

Table A.2: Average absolute bias and average range for repeatability and reproducibility for different experiments.

A.5 Latent Print Analysis Results

We present additional results that compare the parameter estimates obtained from fitting the latent print analysis data with some raw data summaries. We plot the following: i) the mean VID

decisions by an examiner ($Y_{i..}$) against the posterior median for examiner tendencies α_i, ii) the mean VID decisions on a sample ($Y_{.j.}$) against the posterior median for sample tendencies γ_j, and iii) the interaction estimate in a general continuous data ANOVA setting ($Y_{ij.} - Y_{i..} - Y_{.j.} - Y_{...}$) against the posterior median for interaction effects δ_{ij}. Note that $Y_{ij.}$ is the mean decisions on an examiner and sample pair and $Y_{...}$ is the mean of all decisions.

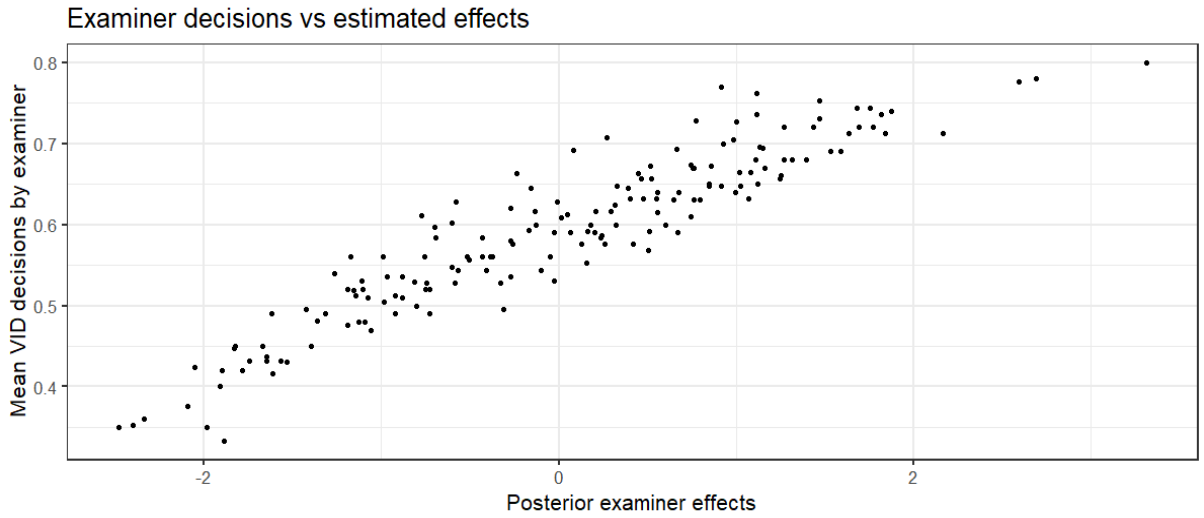


Figure A.2: $Y_{i..}$ v/s posterior median for α_i for the results from the analysis phase of the latent print examination.

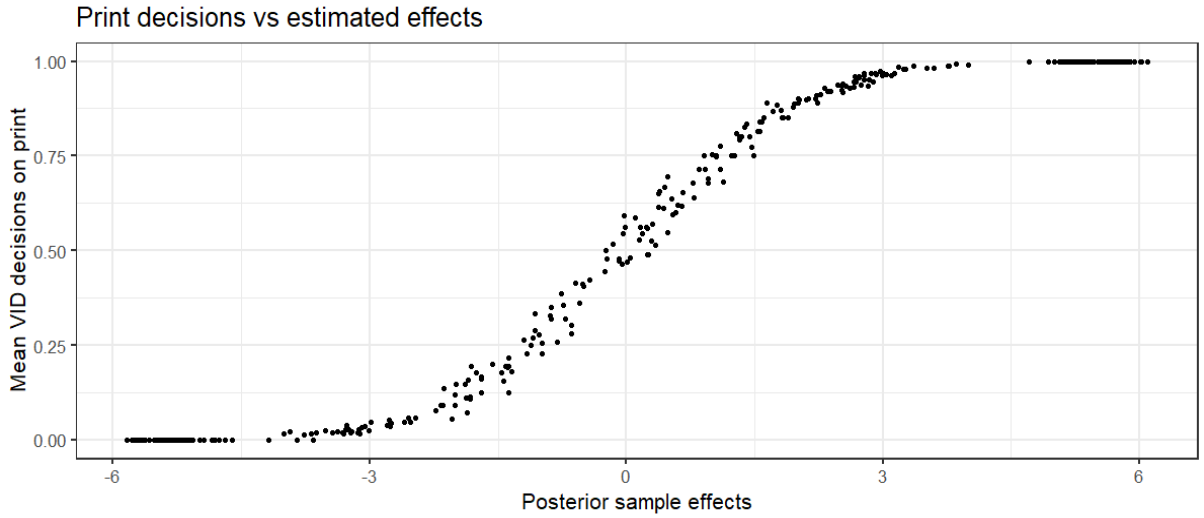


Figure A.3: $Y_{.j.}$ v/s posterior median for γ_j for the results from the analysis phase of the latent print examination.

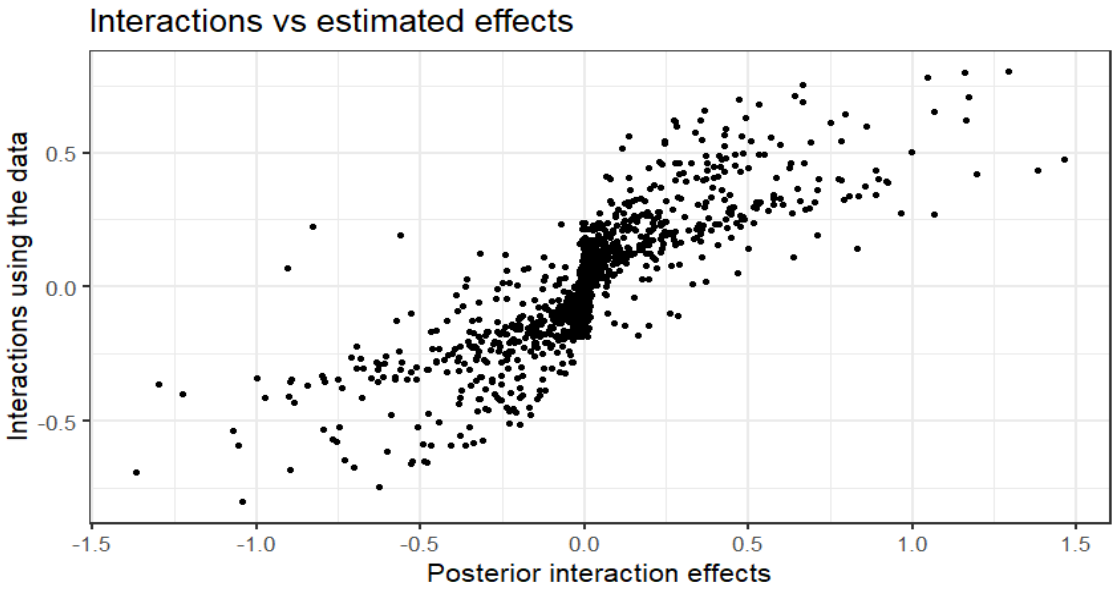


Figure A.4: $Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...}$ v/s posterior median for δ_{ij} for the results from the analysis phase of the latent print examination.

We observe strong associations in all these plots which provides some evidence to the hypothesis that the model is a good fit for the data.

Appendix B

Appendix to Chapter 4

B.1 Equivalent parameterizations

We demonstrate that the constrained CUT model (4.2) and the SET model (4.3) are different parameterizations of the same model when $\tau_{i,2}$ have a normal prior. Consider the conditional probabilities for Y_{ijk} given the parameters in the constrained CUT model (4.2) with $M = 3$ categories:

$$\begin{aligned} P(Y_{ijk} = 1) &= \Phi(\tau_{i,2} - \gamma_j - \delta_{ij}) \\ P(Y_{ijk} = 2) &= \Phi(\tau_{i,3} - \gamma_j - \delta_{ij}) - \Phi(\tau_{i,2} - \gamma_j - \delta_{ij}) \\ P(Y_{ijk} = 3) &= 1 - \Phi(\tau_{i,3} - \gamma_j - \delta_{ij}) \end{aligned}$$

The analogous probabilities for the SET model (4.3), are:

$$\begin{aligned} P(Y_{ijk} = 1) &= \Phi(\kappa_2 - \alpha_i - \gamma_j - \delta_{ij}) \\ P(Y_{ijk} = 2) &= \Phi(\kappa_3 - \alpha_i - \gamma_j - \delta_{ij}) - \Phi(\kappa_2 - \alpha_i - \gamma_j - \delta_{ij}) \\ P(Y_{ijk} = 3) &= 1 - \Phi(\kappa_3 - \alpha_i - \gamma_j - \delta_{ij}) \end{aligned}$$

Comparing these expressions, we find that the probabilities are the same with $\tau_{i,2} = \kappa_2 - \alpha_i$,

$\tau_{i,3} = \kappa_3 - \alpha_i$, and $\tau^* = \kappa_3 - \kappa_2$.

B.2 Full conditionals for Gibbs sampling CUTs and SETs model

We provide here the full conditional posterior distributions for sampling from the posterior distributions for the CUT model (4.1) parameters for $M = 3$. We derive these using a uniform prior on the thresholds as in the equations (4.4). Note that $\mathbb{1}_{\mathbb{ijk}}$ is an indicator function that is equal to 1 if decision Y_{ijk} is observed for examiner i, sample j, in repetition k, and is 0 otherwise.

$$\gamma_j|\text{the rest} \sim N(\frac{\sum_i \sum_k (Z_{ijk} - \delta_{ij})}{\frac{1}{\sigma_\gamma^2} + \sum_i \sum_k \mathbb{1}_{\mathbb{ijk}}}, \frac{1}{\frac{1}{\sigma_\gamma^2} + \sum_i \sum_k \mathbb{1}_{\mathbb{ijk}}})$$

$$\delta_{ij}|\text{the rest} \sim N(\frac{\sum_k (Z_{ijk} - \gamma_j)}{\frac{1}{\sigma_\delta^2} + \sum_k \mathbb{1}_{\mathbb{ijk}}}, \frac{1}{\frac{1}{\sigma_\delta^2} + \sum_k \mathbb{1}_{\mathbb{ijk}}})$$

$$\sigma_\gamma^2|\text{the rest} \sim \text{Inv-Gamma}(\frac{J-1}{2}, \frac{\sum_j \gamma_j^2}{2})$$

$$\sigma_\delta^2|\text{the rest} \sim \text{Inv-Gamma}(\frac{IJ-1}{2}, \frac{\sum_i \sum_j \delta_{ij}^2}{2})$$

$$Z_{ijk}|\text{the rest} \sim N(\gamma_j + \delta_{ij}, 1)\, I(\tau_{i,Y_{ijk}}, \tau_{i,Y_{ijk}+1})$$

$$\tau_{i,2}|\text{the rest} \sim \mathbf{I}\,(\min_{\text{lim,i,2}} < \tau_{i,2} < \max_{\text{lim,i,2}})$$

$$\min_{\text{lim,i,2}} = \max_{j,k}(Z_{ijk} \,|\, Y_{ijk} = 1)$$

$$\max_{\text{lim,i,2}} = \min_{j,k}(Z_{ijk} \,|\, Y_{ijk} = 2)$$

$$\tau_{i,3}|\text{the rest} \sim \mathbf{I}\,(\min_{\text{lim,i,3}} < \tau_{i,3} < \max_{\text{lim,i,3}})$$

$$\min_{\text{lim,i,3}} = \max_{j,k}(Z_{ijk} \,|\, Y_{ijk} = 2)$$

$$\max_{\text{lim,i,3}} = \min_{j,k}(Z_{ijk} \,|\, Y_{ijk} = 3)$$

If $M > 3$ the expressions for the thresholds are derived using an analogous approach. For the constrained CUT model (4.2), the full-conditional posterior distributions for Z_{ijk}, γ_j, σ_γ^2, δ_{ij}, σ_δ^2

are the same as derived. The full conditionals for $\tau_{i,2}$, τ^*, μ_{τ_2}, and $\sigma^2_{\tau_2}$ are provided here for the $M = 3$ case.

$$\tau_{i,2}|\text{the rest} \sim N(\mu_{\tau_2}, \sigma^2_{\tau_2})\,\mathbf{I}\left(\min_{\text{lim,i}} < \tau_{i,2} < \max_{\text{lim,i}}\right)$$

$$\min_{\text{lim,i}} = \max_{j,k}(\max(Z_{ijk} \,|\, Y_{ijk} = 1), \max((Z_{ijk} - \tau^*) \,|\, Y_{ijk} = 2))$$

$$\max_{\text{lim,i}} = \min_{j,k}(\min(Z_{ijk} \,|\, Y_{ijk} = 2), \min((Z_{ijk} - \tau^*) \,|\, Y_{ijk} = 3))$$

$$\tau^*|\text{the rest} \sim \text{Unif}\left(\min^*_{\text{lim}}, \max^*_{\text{lim}}\right)$$

$$\min^*_{\text{lim}} = \max_{i,j,k}(0, Z_{ijk} - \tau_{i,2} \,|\, Y_{ijl} = 2)$$

$$\max^*_{\text{lim}} = \min_{i,j,k}(Z_{ijk} - \tau_{i,2} \,|\, Y_{ijl} = 3)$$

$$\mu_{\tau_2}|\text{the rest} \sim N(\frac{\sum_i \tau_{i,2}}{I}, \frac{1}{I})$$

$$\sigma^2_{\tau_2}|\text{the rest} \sim \text{Inv-Gamma}(\frac{I-1}{2}, \frac{\sum_i \tau^2_{i,2}}{2})$$

For the SETs model (4.3), the full conditional distributions for κ_2, κ_3, α_i, γ_j, and δ_{ij} have been derived below:

$$\alpha_i|\text{the rest} \sim N(\frac{\sum_j \sum_k (Z_{ijk} - \gamma_j - \delta_{ij})}{\frac{1}{\sigma^2_\alpha} + \sum_j \sum_k \mathbb{1}_{\mathbb{ijk}}}, \frac{1}{\frac{1}{\sigma^2_\alpha} + \sum_j \sum_k \mathbb{1}_{\mathbb{ijk}}})$$

$$\gamma_j|\text{the rest} \sim N(\frac{\sum_i \sum_k (Z_{ijk} - \alpha_i - \delta_{ij})}{\frac{1}{\sigma^2_\gamma} + \sum_i \sum_k \mathbb{1}_{\mathbb{ijk}}}, \frac{1}{\frac{1}{\sigma^2_\gamma} + \sum_i \sum_k \mathbb{1}_{\mathbb{ijk}}})$$

$$\delta_{ij}|\text{the rest} \sim N(\frac{\sum_k (Z_{ijk} - \alpha_i - \gamma_j)}{\frac{1}{\sigma^2_\delta} + \sum_k \mathbb{1}_{\mathbb{ijk}}}, \frac{1}{\frac{1}{\sigma^2_\delta} + \sum_k \mathbb{1}_{\mathbb{ijk}}})$$

$$\sigma^2_\alpha|\text{the rest} \sim \text{Inv-Gamma}(\frac{I-1}{2}, \frac{\sum_i \alpha^2_i}{2})$$

$$\kappa_2|\text{the rest} \sim \text{Unif}(\max_{i,j,k}(Z_{ijk}|Y_{ijk} = 1), \min_{i,j,k}(Z_{ijk}|Y_{ijk} = 2))$$

$$\kappa_3|\text{the rest} \sim \text{Unif}(\max_{i,j,k}(Z_{ijk}|Y_{ijk} = 2), \min_{i,j,k}(Z_{ijk}|Y_{ijk} = 3))$$

$$Z_{ijk}|\text{the rest} \sim N(\alpha_i + \gamma_j + \delta_{ij}, 1)\, I(\kappa_{Y_{ijk}}, \kappa_{Y_{ijk}+1})$$

B.3 Model-based Reliability

We derive the model-based agreement on the original scale as defined in the expressions (4.7), here for the SET model (4.3). First, note that marginalizing over α_i and δ_{ij} from the distribution of Z_{ijk}, we get:

$$Z_{ijk}|\gamma_j \sim N(\gamma_j, 1 + \sigma_\alpha^2 + \sigma_\delta^2).$$

Additionally, define, $X_{ijk} = Z_{ijk} - \gamma_j$, then:

$$X_{ijk}|\gamma_j \overset{i.i.d.}{\sim} N(0, 1 + \sigma_\alpha^2 + \sigma_\delta^2) \qquad \forall\, i, k$$

If we define reproducibility as the probability of agreement, then:

$$\begin{aligned}
\text{Reproducibility} &= \sum_{m=1}^{M} P((Y_{ijk} = m) \cap (Y_{i'jk} = m)) \\
&= \sum_{m=1}^{M} \int_{-\infty}^{\infty} P((\kappa_m < Z_{ijk} \leq \kappa_{m+1}) \cap (\kappa_m < Z_{i'jk} \leq \kappa_{m+1})|\gamma_j) f(\gamma_j) d\gamma_j \quad f(\gamma_j) \sim N(0, \sigma_\gamma^2) \\
&= \sum_{m=1}^{M} \int_{-\infty}^{\infty} (P(X_{ijk} \leq \kappa_{m+1} - \gamma_j) - P(X_{ijk} \leq \kappa_m - \gamma_j))^2 f(\gamma_j) d\gamma_j \\
&\qquad X_{ijk} \text{ are independent} \\
&= \sum_{m=1}^{M} \int_{-\infty}^{\infty} \left(\Phi\left(\frac{\kappa_{m+1} - x\sigma_\gamma}{\sqrt{1 + \sigma_\alpha^2 + \sigma_\delta^2}} \right) - \Phi\left(\frac{\kappa_m - x\sigma_\gamma}{\sqrt{1 + \sigma_\alpha^2 + \sigma_\delta^2}} \right) \right)^2 \phi(x) dx \qquad \text{where, } x = \frac{\gamma_j}{\sigma_\gamma} \\
&= \sum_{m=1}^{M} \int_{-\infty}^{\infty} \left[\Phi\left(\frac{\kappa_{m+1}^* - x\sqrt{R_1}}{\sqrt{1 - R_1}} \right) - \Phi\left(\frac{\kappa_m^* - x\sqrt{R_1}}{\sqrt{1 - R_1}} \right) \right]^2 \phi(x)\, dx \\
&\qquad \text{Divide by } \sqrt{1 + \sigma_\alpha^2 + \sigma_\delta^2 + \sigma_\gamma^2}
\end{aligned}$$

Here, $R_1 = \frac{\sigma_\gamma^2}{1+\sigma_\alpha^2+\sigma_\gamma^2+\sigma_\delta^2}$, $\kappa_m^* = \frac{\kappa_m}{\sqrt{1+\sigma_\alpha^2+\sigma_\delta^2+\sigma_\gamma^2}}$, x is a standard normal random variable, Φ is the cumulative density function of the standard normal distribution, and ϕ is the probability density function of standard normal distribution.

Similarly, for repeatability, note that $Z_{ijk}|\alpha_i, \gamma_j, \delta_{ij} \overset{i.i.d.}{\sim} N(\alpha_i + \gamma_j + \delta_{ij}, 1)$, $\forall k$. Define, $T_{ij} = \alpha_i + \gamma_j + \delta_{ij}$. Then, with the SET model (4.3), $T_{ij}|\sigma^2_\alpha, \sigma^2_\delta, \sigma^2_\gamma \overset{i.i.d.}{\sim} N(0, \sigma^2_\alpha + \sigma^2_\delta + \sigma^2_\gamma)$. Given, examiner i and sample j, $Z_{ijk} - T_{ij}|\alpha_i, \gamma_j, \delta_{ij} \sim N(0,1)$ are independent for all k:

$$
\begin{aligned}
\text{Repeatability} &= \sum_{m=1}^{M} P((Y_{ijk} = m) \cap (Y_{ijk'} = m)) \\
&= \sum_{m=1}^{M} \int_{-\infty}^{\infty} P((\kappa_m < Z_{ijk} \leq \kappa_{m+1}) \cap (\kappa_m < Z_{ijk'} \leq \kappa_{m+1}) | \gamma_j, \alpha_i, \delta_{ij}) f(t_{ij}) dt_{ij} \\
&= \sum_{m=1}^{M} \int_{-\infty}^{\infty} (P(x' \leq \kappa_{m+1} - T_{ij}) - P(x' \leq \kappa_m - T_{ij}))^2 f(t_{ij}) dt_{ij} \qquad x' = Z_{ijk} - T_{ij} \sim N(0,1) \\
&= \sum_{m=1}^{M} \int_{-\infty}^{\infty} \left(\Phi(\kappa_{m+1} - x\sqrt{\sigma^2_\gamma + \sigma^2_\delta + \sigma^2_\delta}) - \Phi(\kappa_m - x\sqrt{\sigma^2_\gamma + \sigma^2_\delta + \sigma^2_\alpha}) \right)^2 \phi(x) dx \\
&\quad \text{where, } x = \frac{T_{ij}}{\sqrt{\sigma^2_\gamma + \sigma^2_\delta + \sigma^2_\alpha}} \\
&= \sum_{m=1}^{M} \int_{-\infty}^{\infty} \left[\Phi\left(\frac{\kappa^*_{m+1} - x\sqrt{R_2}}{\sqrt{1 - R_2}} \right) - \Phi\left(\frac{\kappa^*_m - x\sqrt{R_2}}{\sqrt{1 - R_2}} \right) \right]^2 \phi(x)\, dx \\
&\qquad \text{Divide by } \sqrt{1 + \sigma^2_\alpha + \sigma^2_\delta + \sigma^2_\gamma}
\end{aligned}
$$

Here, $R_2 = \frac{\sigma^2_\alpha + \sigma^2_\gamma + \sigma^2_\delta}{1 + \sigma^2_\alpha + \sigma^2_\gamma + \sigma^2_\delta}$.

B.4 Effects of Model Misspecification

A statistical model should be robust to deviations in assumptions. In practice, the data might deviate from the data-generating mechanism that the model assumes. Similar to the discussion in the Supplemental material in Arora et al. (2022), we present results from fitting the SET model (4.3) when the data is generated from a model that differs from the model generating assumptions. Similar results apply to the model specifications given by equations (4.1) and (4.2).

The distribution of Z_{ijk} are changed in the following ways while generating the data:

- Normal - Data is generated with the SET model (4.3) for comparison with the other cases.

- Student's t-error - Data is generated from a model where $Z_{ijk} \,|\, \alpha_i, \gamma_j, \delta_{ij} \sim \alpha_i + \gamma_j + \delta_{ij} + t_6$, where t_6 is a Student's t-distribution with 6 degrees of freedom with variance 1.5. The parameters κ_m, σ_α, σ_γ, and σ_δ are scaled (by the factor $\sqrt{\frac{6}{4}}$ which is the variance of t_6) for comparison with other cases. This alternative allows us to consider a case with heavier tails for the latent variables Z_{ijk}, so it is more likely to sample values that are farther away from the mean.
- Laplace - Data is generated from a model where $Z_{ijk} \,|\, \alpha_i, \gamma_j, \delta_{ij} \sim \alpha_i + \gamma_j + \delta_{ij} + \text{Laplace}(\mu = 0, \text{scale} = 0.5)$. The Laplace distribution has lighter tails compared to the normal distribution. The variance of the Laplace distribution with a scale 0.5 is 1, so the other parameters do not need to be rescaled.
- GEV (Generalized Extreme Value) distribution - Data is generated from a model where $Z_{ijk} \,|\, \alpha_i, \gamma_j, \delta_{ij} \sim \alpha_i + \gamma_j + \delta_{ij} + \text{GEV}(\mu = 0, \text{scale} = \frac{\sqrt{6}}{\pi}, \text{shape} = 0)$. GEV is a distribution that is not centered around its mean unlike all other distributions described in this Appendix.
- Bimodal case- Data is generated from a model where $Z_{ijk} \,|\, \alpha_i, \gamma_j, \delta_{ij}, p_{ijk} \sim p_{ijk} N(\alpha_i + \gamma_j + \delta_{ij} - 0.5, 0.75) + (1 - p_{ijk}) N(\alpha_i + \gamma_j + \delta_{ij} + 0.5, 0.75)$, here, $p_{ijk} \sim \text{Bernoulli}(0.5)$. This latent distribution has two modes.

Additionally, we also consider a model specification in which the random effects do not follow a normal distribution:

- Student's t- random effects- The random effects for α_i and γ_j are generated from t-distributions of 4 and $\frac{8}{3}$ degrees of freedom. Student's t-distribution has heavier tails compared to a normal distribution. So, it is more likely to sample values that are farther away from the mean.

We generated 25 data sets for each of these cases, for I=50 examiners and J=80 samples where each examiner observes each sample twice (100% repeated decisions). We fit these generated data sets to the SET model (4.3) and have presented the posterior medians and 95% credible intervals for κ_2, κ_3, σ^2_α, σ^2_γ, σ^2_δ, and reproducibility and repeatability on the latent scale in Figure B.1.

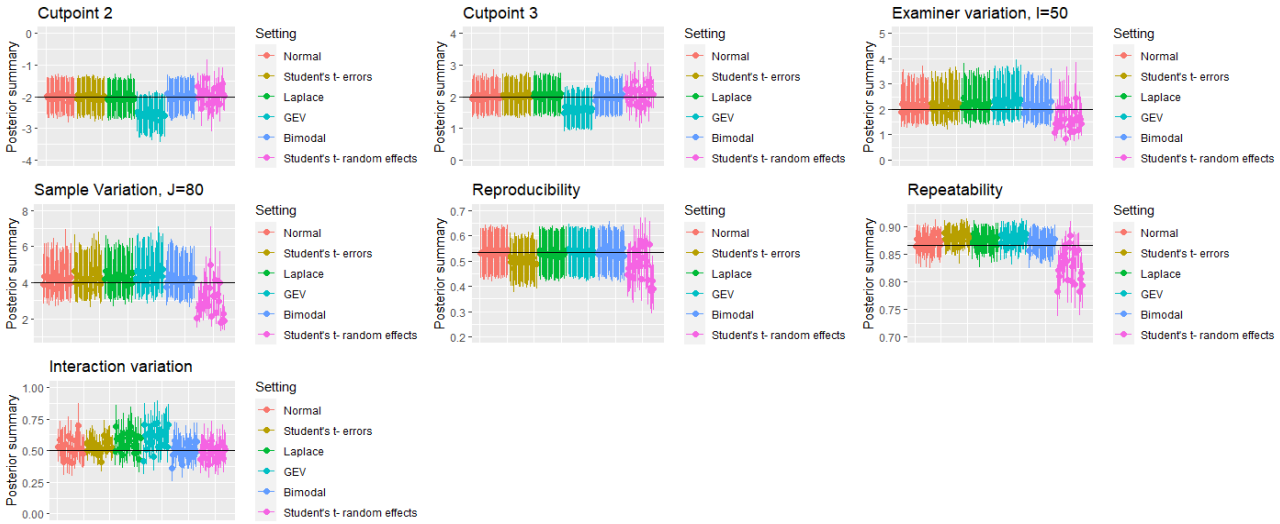


Figure B.1: Posterior medians and 95% credible intervals for estimated parameters with a misspecified model.

From Figure B.1, we can observe that in the GEV case, there is a bias in estimating the values for κ_2, κ_3 through posterior medians. In the Student's t-random effects case, there is slightly more variation in the difference between the true values and the posterior medians for κ_2 and κ_3 but in all other cases they are well estimated. The variance parameters σ_α^2 and σ_γ^2 are estimated well in all cases when the distribution of Z_{ijk} deviates from the normal distribution. The interaction variation is estimated well in most cases, it is overestimated slightly in the bimodal case.

We also consider alternative data generation for the random effects. The variance parameters σ_α^2 and σ_γ^2 are not estimated well when the random effects are Student's t-distributed. Note that the random effects in this case were generated with a Student's t-distribution with small degrees of freedom (4 and $\frac{8}{3}$) which implies that the distribution for the random effects had very heavy tails. Given this fact, it makes sense that the estimation of the variance components is difficult. Given these observations, we can conclude that our model is robust to these deviations in model assumptions.

Appendix C

Appendix to Chapter 5

C.1 Studying the Distribution of Clusters

We have used a Gamma(2,2) prior on the concentration parameter λ and we justify that choice here by simulating data from the stick-breaking process. For different values of I (number of examiners), we have generated clusters with the stick-breaking process for 10,000 data sets. We report the mean, median, and the range of the number of clusters generated for each λ in the next few tables for different choices of I. It is important to consider different choices of I because the maximum number of clusters of potential interest can increase with I.

λ	Range	Mean	Median	Standard Deviation
0.1	(1, 7)	1.43	1.00	0.65
0.25	(1, 7)	2.03	2.00	0.98
0.4	(1, 8)	2.59	2.00	1.19
0.5	(1, 9)	2.94	3.00	1.30
0.6	(1, 11)	3.29	3.00	1.41
0.8	(1, 11)	3.90	4.00	1.56
1.0	(1, 11)	4.49	4.00	1.67
1.5	(1, 16)	5.83	6.00	1.93
2.0	(1, 16)	7.02	7.00	2.12
2.5	(1, 18)	8.15	8.0	2.30
4.0	(3,22)	10.88	11.0	2.61
5.0	(3, 23)	12.46	12.0	2.68
10.0	(6, 29)	18.03	18.0	2.92

Table C.1: I=50. Summary from 10,000 draws for each λ.

λ	Range	Mean	Median	Standard Deviation
0.1	(1, 6)	1.49	1.00	1.69
0.25	(1, 8)	2.21	2.00	1.06
0.4	(1, 10)	2.85	3.00	1.30
0.5	(1, 11)	3.26	3.00	1.43
0.6	(1, 11)	3.72	4.00	1.55
0.8	(1, 13)	4.47	4.00	1.73
1.0	(1, 14)	5.18	5.00	1.89
1.5	(1, 18)	6.88	7.00	2.20
2.0	(2, 19)	8.42	8.00	2.43
2.5	(2, 20)	9.83	10.0	2.61
3.0	(3,24)	11.11	11.0	2.75
4.0	(4, 28)	13.53	13.0	3.03
5.0	(6, 29)	15.70	16.0	3.20

Table C.2: I=100. Summary from 10,000 draws.

λ	Range	Mean	Median	Standard Deviation
0.1	(1, 6)	1.55	1.0	0.74
0.25	(1, 8)	2.33	2.0	1.12
0.5	(1, 11)	3.54	3.0	1.52
1.0	(1, 16)	5.69	6.0	2.00
2.0	(2, 20)	9.39	9.0	2.61
3.0	(3, 26)	12.63	13.0	3.00
4.0	(5, 30)	15.62	15.0	3.36
5.0	(6, 32)	18.23	18.00	3.60
10.0	(14, 48)	29.30	29	4.39

Table C.3: I=169. 10,000 draws.

As demonstrated in the tables, it seems that for most designs $\lambda \in (0, 2.5)$, gives enough flexibility to infer the number of clusters that may be scientifically interesting in a data set.