Studying Reliability for Pattern Evidence Comparisons

Hina Arora
Dr. Naomi Kaplan-Damary
Prof. Hal Stern

University of California, Irvine

08/05/2021

Motivation

- The justice system relies heavily on forensic science.
- In many cases, evidence requires subjective examination from forensic experts.
- Examiner experience, human error and the quality of evidence introduces variance in examination decisions.
- It is important to study variation for accurate assessments and avoid unwarranted conclusions.

Motivation

- Different examiners may not agree on decisions and we want to study these variations.
- We must study the forensic examination process to assess the correctness and variations in decisions.
- Focus on sub-disciplines like fingerprint examination, shoeprint examination etc. where patterns are compared.

Definitions

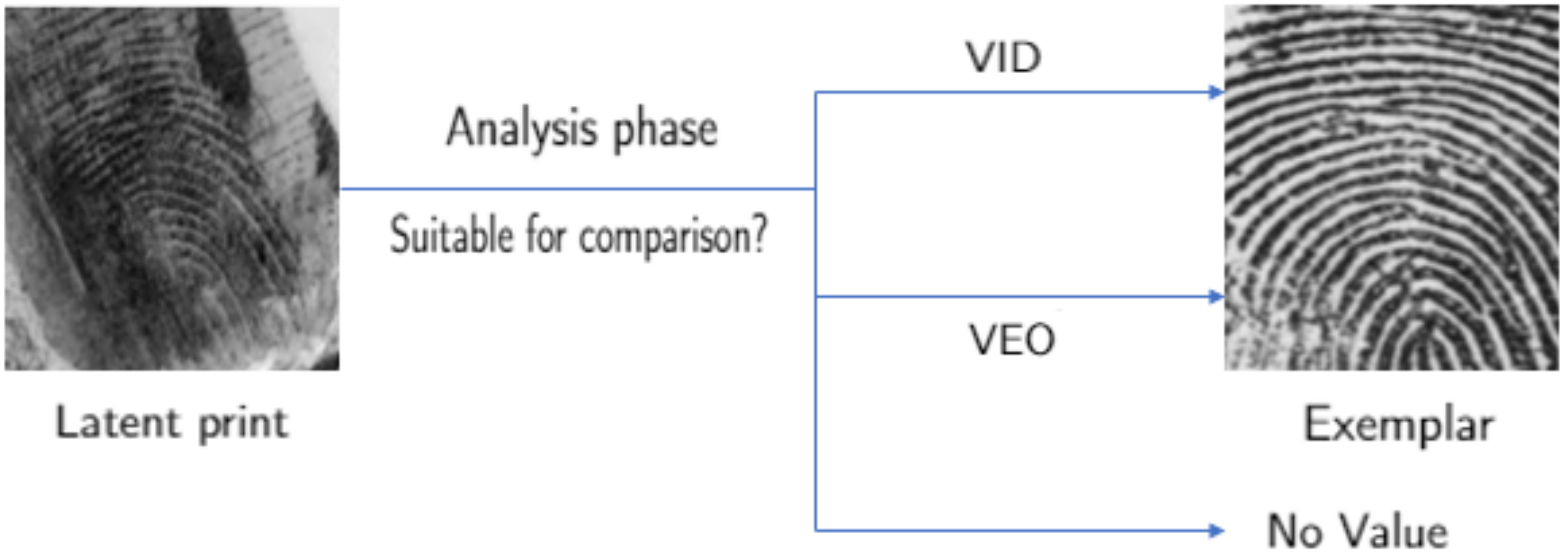
- Reliability: Related to the consistency of a measure or a procedure. Two types of reliability:
  - Repeatability: When the same examiner gives reliable conclusions under the same conditions. Within examiner reliability.
  - Reproducibility: When different examiners give reliable conclusions under the same conditions. Between examiner reliability.
- Validity: Related to the correctness (accuracy) of the procedure.
- Reliability is a pre-cursor for accuracy.

Background

- An important goal is to establish reliability and validity for feature-based comparison methods.
- For subjective feature-based comparison methods, "black-box" (empirical) studies.
- The process of reaching a decision is not clearly defined, and hence the process is treated like a black-box.

Reliability of fingerprint evidence

- In late 2000's, the FBI conducted a study to establish reliability of latent fingerprint examination decisions (Ulery et al., 2011, 2012).
- Latent print examination workflow- Analysis, Comparison, Evaluation and Verification (ACE-V).



Decisions in Analysis phase

Reliability of fingerprint evidence

- Analysis phase may have binary or ternary decisions that depend on the quality of the latent. For example, VID (Value for Individualization), VEO (Value for Exclusion Only) and NV (No Value).
- Evaluation phase has ternary decisions as well: Individualization (coming from the same source), Exclusion (coming from different sources) and Inconclusive decisions.
- Decisions of this nature are crucial and need to be studied.

Reliability of fingerprint evidence

- Reproducibility study - 169 examiners, 744 latent prints and corresponding exemplars (mated and non-mated), ≈ 100 pairs per examiner.
- Prints selected to contain features that mimic real cases.
- Repeatability study - About 7 months after the first study, a subset of prints were re-examined by a subset of these examiners.
- Studies with a large repeatability components are rare because they are expensive and use precious examiner time. Hence, this incomplete design is common in many forensic studies.

Reliability of signature complexity decisions

- Signature complexity is an important factor for an examiner to differentiate between a real signature and a simulation.
- Data collected by LAPD/ LASD: 123 signatures, 5 forensic document examiners (FDEs). Repetitions on $n = 7$ signatures. Signature complexity assessed on two different scales (a 1-5 point scale and a 1-3 point scale).
- Stern et al. (2018) analyzed reliability for handwritten signature complexity scores based on this study.
- 1-5 point scale data was treated as continuous. Also the data collected in the repeatability study was analyzed separately.

Example of complexity data

Sample	Examiner A	Examiner B
#1	4	3
#2	3	4
#3	2	3

Sample	Examiner A	Examiner B
#1	3	4
#2	3	4
#3	NA	NA

Example of data collected from two examiners in two reliability trials. The size of the first dataset is 123×5 and the second dataset is 7×5.

Questions of interest

- Interested in finding out if and how we can analyze reproducibility and repeatability data together to draw better inference. Important to do that for better inference and get the most information from the data.
- Establish reliability for studies where there may be a presence of an examiner-sample interaction.
- Start with modelling variability in continuous and binary decisions.
- Explore methodology for ordinal decisions with incomplete replicates.

Continuous data model

- Start with continuous data model to explain approach.
- Let Y_{ijk} be a continuous decision for sample (evidence) j by examiner i on the k^{th} examination.
- This setting is the most intuitive to model. In the past, continuous measurements, in engineering, have been modeled with an ANOVA (analysis of variance) model.
- An ANOVA model is used when comparing an outcome of interest across different groups/ classes. For example, examiners in our setting.

Continuous data model

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

α_i = examiner i effect

γ_j = sample j effect

δ_{ij} = possible interaction of i and j

- α_i is an examiner specific effect, that affects the decision through an examiner's ability or experience.
- γ_j is a sample (evidence) specific effect that affects the decision because of the quality/ complexity of the sample and say δ_{ij} is an interaction between sample and examiner.

Continuous data model

- If δ_{ij} is non-zero, it means that the examiner specific effect changes according to the sample.
- Say, α_i come from a population with variance σ^2_α, then σ^2_α represents the variance of examiner effects.
- Similarly, let γ_j come from a population with variance σ^2_γ, then σ^2_γ represents the variance of sample effects.
- We expect σ^2_α, to be lower compared to σ^2_γ, because forensic experts may not have as much variation amongst themselves but the sample effects can be very different.
- Say, δ_{ij} have a variance σ^2_δ.

Continuous data model

- One way of assessing reliability in this setting, would be to calculate correlation.
- Reproducibility can be thought of the correlation (or association) of decisions of different examiners for the same sample or evidence.
- Repeatability can be thought of the correlation (or association) of decisions the same examiner for the same sample or evidence.
- Turns out that these quantities can be estimated with the variance components.

$$\text{Reproducibility} = \frac{\sigma_\gamma^2}{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\delta^2 + \sigma_\epsilon^2}$$

$$\text{Repeatability} = \frac{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\delta^2}{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\delta^2 + \sigma_\epsilon^2}$$

Interaction visualization

Sample ——————— Examiner	Sample difficulty= -1	Sample difficulty= 1	Difference in outcomes across samples
Examiner ability= -0.5	1.5	3.5	2
Examiner ability= 0.5	2.5	4.5	2
Difference in examiner decisions	1	1	

Difference in outcomes across examiners and samples in absence of interaction. Value for $\mu = 3$.

Sample ——————— Examiner	Sample difficulty= -1	Sample difficulty= 1	Difference in outcomes across samples
Examiner ability= -0.5	1.3	3.7	2.4
Examiner ability= 0.5	2.7	4.3	1.6
Difference in examiner decisions	1.4	0.6	

Difference in outcomes across examiners and samples in presence of interaction. Value for $\mu = 3$.

Signature data results

- Results were obtained from the handwritten signature complexity study.
- The data was on 1-5 point scale which was apporximated to a continuous scale.
- The posterior mean and 95% credible interval of μ is 3.56 $(3.24, 3.89)$, examiner variation σ^2_α is 0.10 $(0.01, 0.47)$, signature (sample) variation σ^2_γ is 0.80 $(0.61, 1.05)$, random noise variation σ^2_ϵ is 0.36 $(0.27, 0.42)$ and interaction variation σ^2_δ is 0.03 $(0.00, 0.12)$.

Signature data results

- Stern et al. (2018) estimated the reproducibility with 95% confidence interval is 0.65 $(0.58, 0.72)$ and repeatability with 95% confidence interval was 0.67 $(0.36, 0.85)$.

- Comparing these results to our methodology the reproducibility with 95% credible interval is 0.63 $(0.47, 0.71)$ and the repeatability with 95% credible interval was 0.72 $(0.64, 0.81)$.

- The reproducibility estimates as well as the intervals are comparable. However, using our methodology we obtain a smaller credible interval for repeatability.

Binary data model

- Assume Y_{ijk} are binary decisions (match/ non-match or value/ no-value) on sample j by examiner i on the k^{th} trial.
- Albert and Chib (1993) proposed an algorithm that extends the model that was proposed for continuous data to binary data by introducing a latent (underlying) continuous variable Z_{ijk} corresponding to binary variable Y_{ijk}.

$$Y_{ijk} = \mathbb{1}_{(Z_{ijk}>0)}$$

$$Z_{ijk}|\beta_0, \alpha_i, \gamma_j, \delta_{ij} \sim N(\beta_0 + \alpha_i + \gamma_j + \delta_{ij}, 1)$$

$$\alpha_i|\sigma^2_\alpha \overset{i.i.d.}{\sim} N(0, \sigma^2_\alpha)$$

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

$$\delta_{ij}|\sigma^2_\delta \overset{i.i.d.}{\sim} N(0, \sigma^2_\delta)$$

Interaction visualization

Sample ——————— Examiner	Sample difficulty= -1	Sample difficulty= 1	Difference across samples
Examiner ability= -0.5	0.07	0.69	0.62
Examiner ability= 0.5	0.31	0.93	0.62
Difference across examiners	0.24	0.24	

Difference in probabilities of outcomes across examiners and samples in absence of interaction. Value for $\mu = 0$

Sample ——————— Examiner	Sample difficulty= -1	Sample difficulty= 1	Difference across samples
Examiner ability= -0.5	0.05	0.76	0.71
Examiner ability= 0.5	0.38	0.90	0.52
Differnce across examiners	0.33	0.14	

Difference in probabilities of outcomes across examiners and samples in

Binary Data Simulations

- We simulate data to test our methodology again.
- We simulate 100 datasets for fixed parameters μ, σ^2_α, σ^2_γ, σ^2_δ for total number of examiners $I = 20$ and total number of samples $J = 60$.
- Four scenarios:
  - 100% replicates - $(60 \times 20 \times 2)$
  - 50% replicates - (30 samples re-examined)
  - 25% replicates - (15 samples re-examined)
  - 12.5% replicates - (7 samples re-examined)

Binary Data Simulations

- Placeholder for simulation tables.

Results from Ulery et al., 2011-12

- Ulery et al. (2011) analyzed the data from the first trial for assessing accuracy of comparison decisions.
- They found that examiners had a false-positive rate of 0.1% and a false-negative rate of 7.5%.
- Ulery et al. (2012) analyzed the reproducibility and repeatability of decisions from first and second trial.
- They concluded that 89.1% of individualization decisions and 90.1% of exclusion decisions were repeated. Repeatability was 90.0% for mated pairs, and 85.9% for nonmated pairs.

Fingerprint data results

- The outcome (decisions) were VID and not VID. There were a total of 169 examiners and they examined a subset from a total of 744 latent prints each.
- The posterior mean and 95% credible interval of μ is 3.78 $(3.21, 4.49)$, examiner variation is 1.86 $(1.34, 2.62)$, latent print (sample) variation is σ^2_γ is 24.19 $(18.00, 32.35)$ and interaction variation σ^2_δ is 0.53 $(0.22, 0.94)$.
- The posterior mean for reproducibility on latent scale with 95% credible interval is 0.88 $(0.85, 0.90)$ and the posterior mean for repeatability with 95% credible interval is 0.96 $(0.95, 0.97)$.

Ordinal data model

- In forensic examination process, decisions are often made on an ordinal scale.
- Ordinal data is categorical data where the categories have a certain order to them.
- For example, signature complexity decisions are made on a $1-5$ point. We treated these measurements as continuous data however,
- Latent fingerprint decisions: Value for Individualization (VID), Value for Exclusion Only (VEO) and No Value (NV).

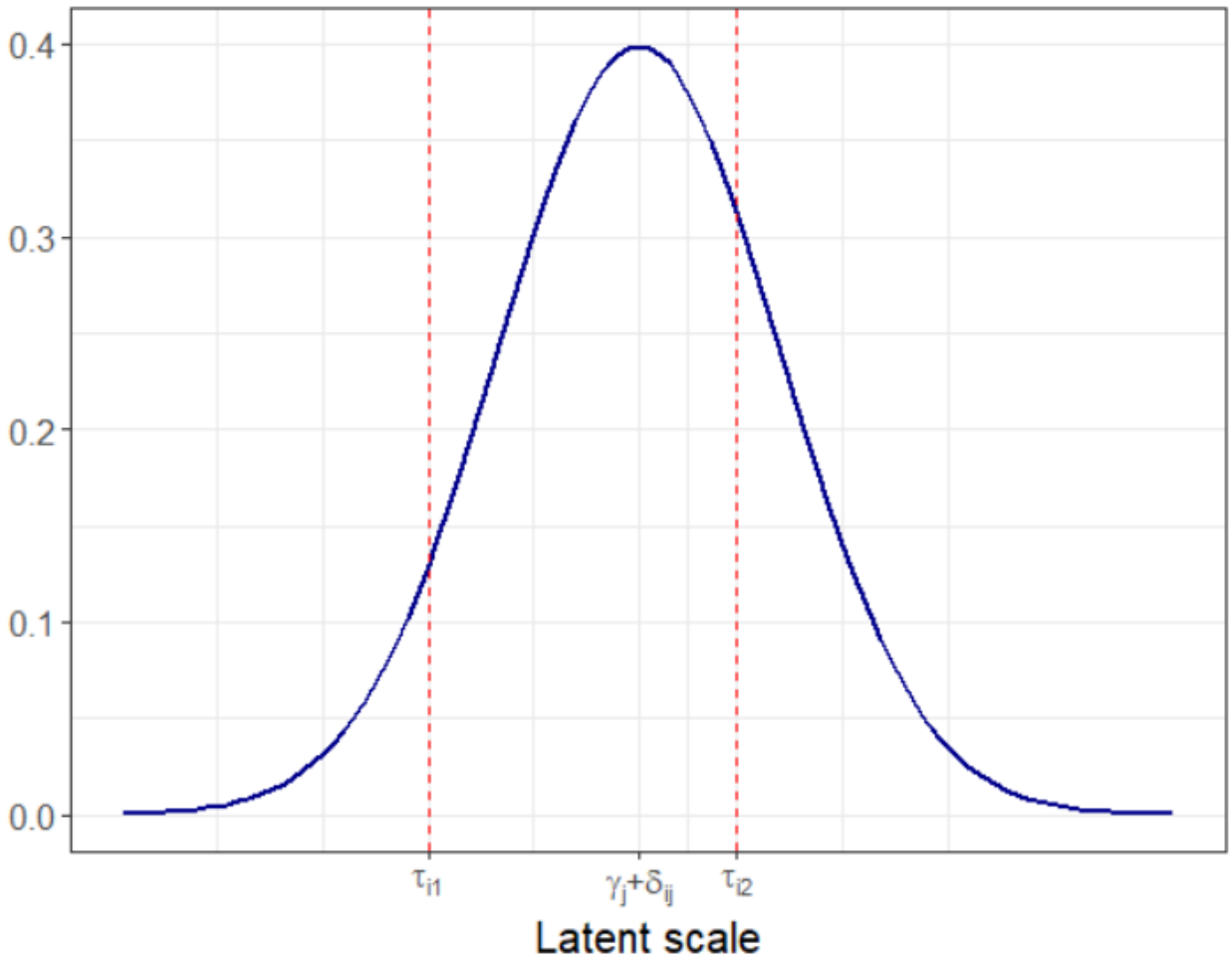
Ordinal data model

- Assume that the decisions can be one of M ordinal categories.
- Model decisions to depend on an underlying continuous random variable.
- Examiners assign the sample to a particular category as long as it is above a certain threshold for them.
- Since decisions are subjective, thresholds are assumed to depend on the examiner.

Ordinal data model

- Let i be the index for examiner, j be the index for sample, and k is the trial. Categories go from 1 through M. Assume the decisions are denoted by Y_{ijk}.
- Decisions are assumed to depend on an underlying continuous variable Z_{ijk}.
- The continuous variable is centered around the sample difficulty and an interaction between sample and examiner.
- $M-1$ ordered cut-points on the real line represent examiner thresholds for each category.

Cut-points depend on examiner



Conclusions and Future Work

- Our results suggest that we have found a way to integrate different reliability studies.
- Repetitions on 25% of samples can give reasonable estimates for reliability for binary and continuous data.
- Some computational issues:
  - Small interaction variance
  - Time
- Alternative interpretable models for interactions.

References

- Stern, Hal S., et al. "Assessing the complexity of handwritten signatures." *Law, Probability and Risk* 17.2 (2018): 123-132.
- Ulery, Bradford T., et al. "Accuracy and reliability of forensic latent fingerprint decisions." *Proceedings of the National Academy of Sciences* 108.19 (2011): 7733-7738.
- Ulery, Bradford T., et al. "Repeatability and reproducibility of decisions by latent fingerprint examiners." *PloS one* 7.3 (2012): e32800.
- Johnson, Valen E., and James H. Albert. *Ordinal data modeling*. Springer Science & Business Media, 2006.
- Stern, Hal S., Maria Cuellar, and David Kaye. "Reliability and validity of forensic science evidence." *Significance* 16.2 (2019): 21-24.