Estimating repeatability and reproducibility with limited replications

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

University of California, Irvine

Tuesday 4th August, 2020

1 Motivation
Background
Some definitions
Reliability studies
Our contribution

2 Continuous data
Continuous Data Model
Simulation studies
Results on Signature Data

3 Binary Data
Binary Data Model
Simulation studies

4 Conclusions

Table of Contents

Motivation
- Background
- Some definitions
- Reliability studies
- Our contribution

Continuous data
- Continuous Data Model
- Simulation studies
- Results on Signature Data

Binary Data
- Binary Data Model
- Simulation studies

Conclusions

Background

- The justice system relies heavily on forensic science.
- However, innocent people have been held or imprisoned because of incorrect application of forensic science. This is a serious problem.

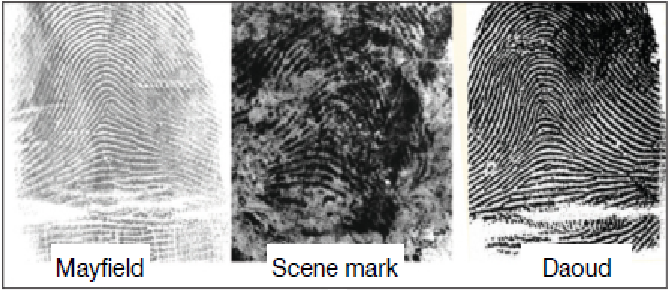
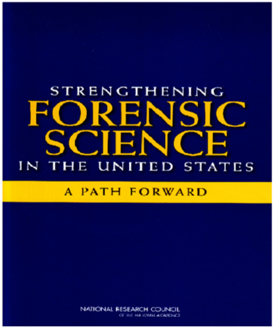
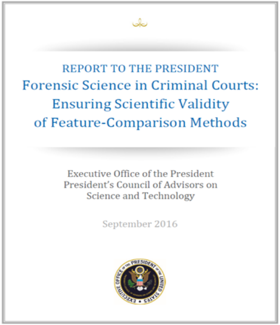


Figure: Brandon Mayfield's fingerprint compared to the mark found at the scene.

1 Motivation
Background
Some definitions
Reliability studies
Our contribution
2 Continuous data
Continuous Data Model
Simulation studies
Results on Signature Data
3 Binary Data
Binary Data Model
Simulation studies
4 Conclusions

Background





- National Academy of Sciences (NAS) prepared a report in 2009 emphasizing the need for formal assessment of scientific foundations of forensic science.
- In 2016, President's Council of Advisors on Science and Technology (PCAST) prepared a report assessing progress and highlighting the need to ensure validity and reliability of feature-based comparison methods.

Important definitions

- Reliability: Related to the consistency of a measure or a procedure. Two types of reliability:

  - Repeatability: Consistency of repeated measurements obtained from the same examiner under the same conditions. Intra-examiner reliability.

  - Reproducibility: Consistency of measurements obtained from different examiners under the same conditions. Inter-examiner reliability.

- Validity: Related to the correctness (accuracy) of the procedure.

1 Motivation
Background
Some definitions
Reliability studies
Our contribution
2 Continuous data
Continuous Data Model
Simulation studies
Results on Signature Data
3 Binary Data
Binary Data Model
Simulation studies
4 Conclusions

Background

- An important goal is to establish reliability and validity for feature-based comparison methods through empirical studies.
- For objective measurement methods, it is easier to establish reliability and validity.
- For subjective feature-based comparison methods, "black-box" studies.
- The decision process is subjective and treated like a black-box.

1 Motivation
Background
Some definitions
Reliability studies
Our contribution
2 Continuous data
Continuous Data Model
Simulation studies
Results on Signature Data
3 Binary Data
Binary Data Model
Simulation studies
4 Conclusions

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

1 Motivation
Background
Some definitions
Reliability studies
Our contribution
2 Continuous data
Continuous Data Model
Simulation studies
Results on Signature Data
3 Binary Data
Binary Data Model
Simulation studies
4 Conclusions

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.

1 Motivation
Background
Some definitions
Reliability studies
Our contribution
2 Continuous data
Continuous Data Model
Simulation studies
Results on Signature Data
3 Binary Data
Binary Data Model
Simulation studies
4 Conclusions

Our contribution

- Develop methods that allow for assessing reliability that:
  - Combine data from reproducibility studies and (incomplete) repeatability studies.
  - Allow for investigating examiner-sample interactions.

1 Motivation
Background
Some definitions
Reliability studies
Our contribution
2 Continuous data
Continuous Data Model
Simulation studies
Results on Signature Data
3 Binary Data
Binary Data Model
Simulation studies
4 Conclusions

Table of Contents

Motivation
Background
Some definitions
Reliability studies
Our contribution

Continuous data
Continuous Data Model
Simulation studies
Results on Signature Data

Binary Data
Binary Data Model
Simulation studies

Conclusions

Continuous data model

- Let Y_{ijk} be continuous measurements (e.g., signature complexity) for sample j by examiner i on the k^{th} trial.
- Assume that α_i is a random effect for examiner ability, γ_j is random effect for sample difficulty, and δ_{ij} is an interaction between sample and examiner.

$$p(Y_{ijk}|\beta_0, \alpha_i, \gamma_j, \delta_{ij}) \sim N(\beta_0 + \alpha_i + \gamma_j + \delta_{ij}, \sigma_\epsilon^2)$$

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

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

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

- A Bayesian approach was used to fit these models.

1 Motivation
Background
Some definitions
Reliability studies
Our contribution
2 Continuous data
Continuous Data Model
Simulation studies
Results on Signature Data
3 Binary Data
Binary Data Model
Simulation studies
4 Conclusions

Continuous data model

- A diffuse prior was used on β_0 and a half-t prior was used on standard deviations σ_α, σ_γ, σ_δ.
- One way of assessing reliability in this setting, would be to calculate correlation: σ^2_α,σ^2_γ,σ^2_δ.

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

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

- Computation for the reliability components is straightforward through MCMC. Credible intervals are also easily obtained.

1 Motivation
Background
Some definitions
Reliability studies
Our contribution
2 Continuous data
Continuous Data Model
Simulation studies
Results on Signature Data
3 Binary Data
Binary Data Model
Simulation studies
4 Conclusions

Continuous Data Simulations

- For simulation studies we try to replicate studies with smaller repeatability studies as compared to the reproducibility study.
- We simulate 100 datasets for fixed parameters β_0, σ_α^2, σ_γ^2, σ_δ^2 for total number of examiners $I = 10$ and total number of samples $J = 120$.
- Four scenarios:
  - 100% replicates - ($120 \times 10 \times 2$)
  - 50% replicates - (60 samples re-examined)
  - 25% replicates - (30 samples re-examined)
  - 12.5% replicates - (15 samples re-examined)
- We report average posterior median and average lower 2.5% quantiles and average upper 97.5% quantiles.

1 Motivation
- Background
- Some definitions
- Reliability studies
- Our contribution

2 Continuous data
- Continuous Data Model
- Simulation studies
- Results on Signature Data

3 Binary Data
- Binary Data Model
- Simulation studies

4 Conclusions

Continuous Data Simulations

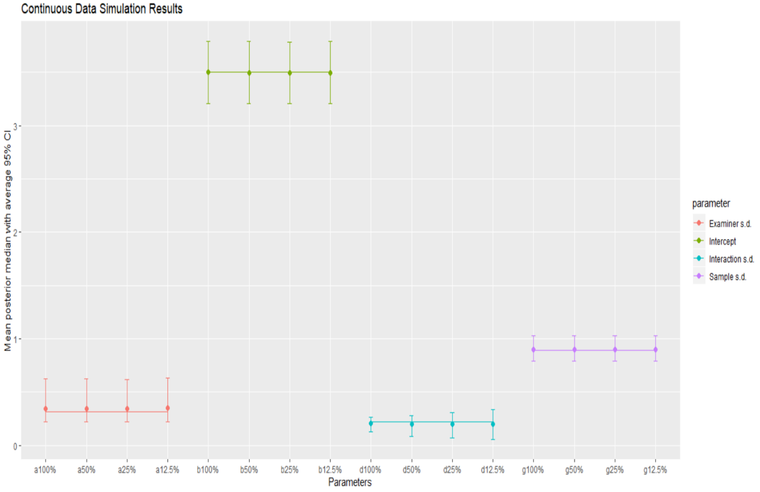


Figure: Simulation results from 100 datasets (continuous data) for each setting of replicates.

1 Motivation
Background
Some definitions
Reliability studies
Our contribution
2 Continuous data
Continuous Data Model
Simulation studies
Results on Signature Data
3 Binary Data
Binary Data Model
Simulation studies
4 Conclusions

Signature data results

- Analyzed the signature 1-5 point complexity data.

Parameters	σ^2_α Examiner variation	σ^2_γ Signature variation	σ^2_δ Interaction variation
Estimates	0.09 (0.01, 0.40)	0.80 (0.61, 1.05)	0.02 (0.00, 0.09)

Table: Posterior means with 95% credible intervals. A low interaction is a positive for an measurement process.

Model type	Reproducibility corr($Y_{ijk}, Y_{i'jk'}$)	Repeatability corr($Y_{ijk}, Y_{ijk'}$)
Separate Analysis	0.65 (0.58, 0.72)	0.67 (0.36, 0.85)
Combining datasets (using posterior mean)	0.63 (0.49, 0.71)	0.71 (0.64, 0.80)

Table: Changes in reliability after combining datasets and accounting for interaction variance. Note that the credible interval is much narrower for repeatability.

1 Motivation
Background
Some definitions
Reliability studies
Our contribution
2 Continuous data
Continuous Data Model
Simulation studies
Results on Signature Data
3 Binary Data
Binary Data Model
Simulation studies
4 Conclusions

Table of Contents

Motivation
- Background
- Some definitions
- Reliability studies
- Our contribution

Continuous data
- Continuous Data Model
- Simulation studies
- Results on Signature Data

Binary Data
- Binary Data Model
- Simulation studies

Conclusions

1 Motivation
Background
Some definitions
Reliability studies
Our contribution
2 Continuous data
Continuous Data Model
Simulation studies
Results on Signature Data
3 Binary Data
Binary Data Model
Simulation studies
4 Conclusions

Binary data model

- Assume Y_{ijk} are binary measurements (e.g., match/ non-match or value/ no-value) on sample j by examiner i on the k^{th} trial.
- Albert and Chib (1993) proposed an MCMC algorithm for Bayesian analysis for binary and polychotomous data based on latent variables Z_{ijk}.

$$Y_{ijk} = \mathbb{1}_{(Z_{ijk}>0)}$$
$$p(Z_{ijk}|\beta_0, \alpha_i, \gamma_j, \delta_{ij}) \sim N(\beta_0 + \alpha_i + \gamma_j + \delta_{ij}, 1)$$
$$p(\alpha_i|\sigma^2_\alpha) \overset{i.i.d.}{\sim} N(0, \sigma^2_\alpha)$$
$$p(\gamma_j|\sigma^2_\gamma) \overset{i.i.d.}{\sim} N(0, \sigma^2_\gamma)$$
$$p(\delta_{ij}|\sigma^2_\delta) \overset{i.i.d.}{\sim} N(0, \sigma^2_\delta)$$

1 Motivation
Background
Some definitions
Reliability studies
Our contribution
2 Continuous data
Continuous Data Model
Simulation studies
Results on Signature Data
3 Binary Data
Binary Data Model
Simulation studies
4 Conclusions

Binary Data Simulations

- For simulation studies we try to replicate studies with smaller repeatability studies as compared to the reproducibility study.
- We simulate 100 datasets for fixed parameters β_0, σ^2_α, σ^2_γ, σ^2_δ for total number of examiners $I = 10$ and total number of samples $J = 120$.
- Four scenarios:
  - 100% replicates - $(120 \times 10 \times 2)$
  - 50% replicates - (60 samples re-examined)
  - 25% replicates - (30 samples re-examined)
  - 12.5% replicates - (15 samples re-examined)
- We report average posterior median and average lower 2.5% quantiles and average upper 97.5% quantiles.

1 Motivation
- Background
- Some definitions
- Reliability studies
- Our contribution

2 Continuous data
- Continuous Data Model
- Simulation studies
- Results on Signature Data

3 Binary Data
- Binary Data Model
- Simulation studies

4 Conclusions

Binary Data Simulations

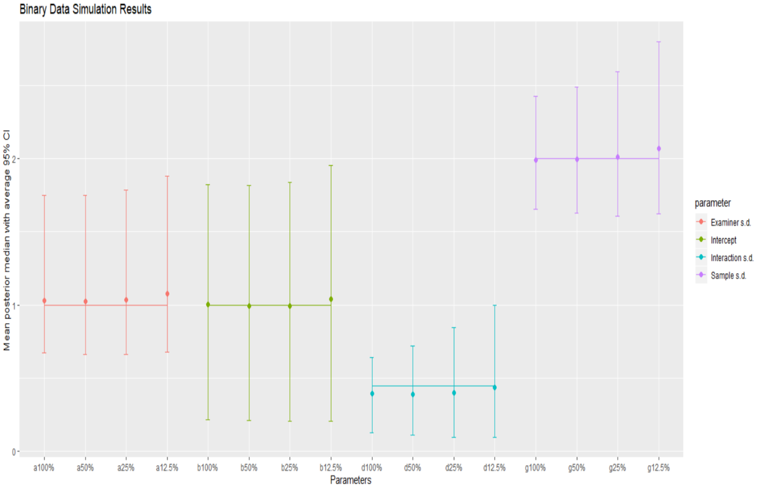


Figure: Simulation results from 100 datasets (binary data) for each setting of replicates.

1 Motivation
Background
Some definitions
Reliability studies
Our contribution
2 Continuous data
Continuous Data Model
Simulation studies
Results on Signature Data
3 Binary Data
Binary Data Model
Simulation studies
4 Conclusions

Table of Contents

Motivation
- Background
- Some definitions
- Reliability studies
- Our contribution

Continuous data
- Continuous Data Model
- Simulation studies
- Results on Signature Data

Binary Data
- Binary Data Model
- Simulation studies

Conclusions

Conclusions

- A single model can be used to simultaneously analyze data from reproducibility and repeatability studies.
- Have been able to show through simulation studies that parameters can be recovered even with 25% repeated decisions.
- In forensic signature analysis no evidence for examiner-signature interactions.
- Some convergence issues with small interaction variance and higher computational time for large data sets.

1 Motivation
Background
Some definitions
Reliability studies
Our contribution
2 Continuous data
Continuous Data Model
Simulation studies
Results on Signature Data
3 Binary Data
Binary Data Model
Simulation studies
4 Conclusions

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.
- Stern, Hal S., Maria Cuellar, and David Kaye. "Reliability and validity of forensic science evidence." *Significance* 16.2 (2019): 21-24.
- Albert, James H., and Siddhartha Chib. "Bayesian analysis of binary and polychotomous response data." Journal of the American statistical Association 88.422 (1993): 669-679.

1 Motivation
Background
Some definitions
Reliability studies
Our contribution
2 Continuous data
Continuous Data Model
Simulation studies
Results on Signature Data
3 Binary Data
Binary Data Model
Simulation studies
4 Conclusions