# Guidelines for Appropriate Use of Simulated Data from Bio-authentication Research [a b]

Yan Ma[1], Michael E. Schuckers[2], Bojan Cukic[1]

[1] West Virginia University

[2] St. Lawrence University and

the Center for Identification Technology Research (CITeR)

AutoID 2005

[a]NSF grant CNS-0325640 which is cooperative funded by the National Science Foundation and the United States Department of Homeland Security

[b]Additional funding from Center for Identification Technology Research

# Outline

1. Introduction to Simulation
2. Taxonomy of Data
3. Guidelines
4. Illustrations

# Simulation

Current Debate:

**One side:** Simulation is *very, very bad.*

**Other side:** Simulation is *very, very good.*

**This talk:** We offer a third way.

Simulation can be good but . . .

under what conditions?

# Taxonomy

If we want to simulate data, what data?

- Images
- Features
- Match/Similarity Scores
- Decisions

# Taxonomy of Biometric Data

- **Image data** - Collection of image
  E.g. raw 'picture' of biometric image
- **Feature data** - Measurements of features
  E.g. Iris Densities, FP minutiae, intra-pupil distance
- **Match Score data** - Distance metric
  E.g. Match Scores, Normalized scores, Hamming distance, Multi-modal
- **Decision data** - Binary Decision
  Accept or Reject, Allow or Deny Access

# Notation

Let $\mathbf{X} \sim F(\mathbf{X} \mid \boldsymbol{\theta})$ represent the cdf of our simulation model where $\mathbf{X}$ is a RV representing the data and

$\boldsymbol{\theta}$ represents the parameters of the simulation model

Let $\hat{F}(\mathbf{x} \mid \hat{\boldsymbol{\theta}})$ represent the estimated cdf where $\mathbf{x}$ is the realized data and

$\hat{\boldsymbol{\theta}}$ represents estimates of $\boldsymbol{\theta}$ using the data, $\mathbf{x}$.

# Guidelines

Three criteria for simulation

1. Flexibility
2. Parsimonious
3. Goodness-of-Fit

# Flexibility and Parsimony

Simulation needs

- Random generation via cdf say $F(\mathbf{X} \mid \boldsymbol{\theta})$
- Enough parameters to capture data complexity

But ...

- Simple as need be

# Goodness-of-fit

**Idea:**
Is $\hat{F}(\mathbf{x} \mid \hat{\boldsymbol{\theta}})$ similar to $F(\mathbf{X} \mid \boldsymbol{\theta})$?

Examples

- Kolmogorov-Smirnov
- Anderson-Darling
- QQ-plot
- Garren *et al.* (2001)

# Illustrations

Data: Genuine Facial Matching Scores

Source: Michigan State University, Ross and Jain (2003)

Model: $X \sim log - normal(\mu, \sigma)$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(log x - \mu)^2}{2\sigma^2}}, x \in (-\infty, \infty) \qquad (1)$$

Estimation via MLE

# Illustrations

Data: Imposter Facial Matching Scores

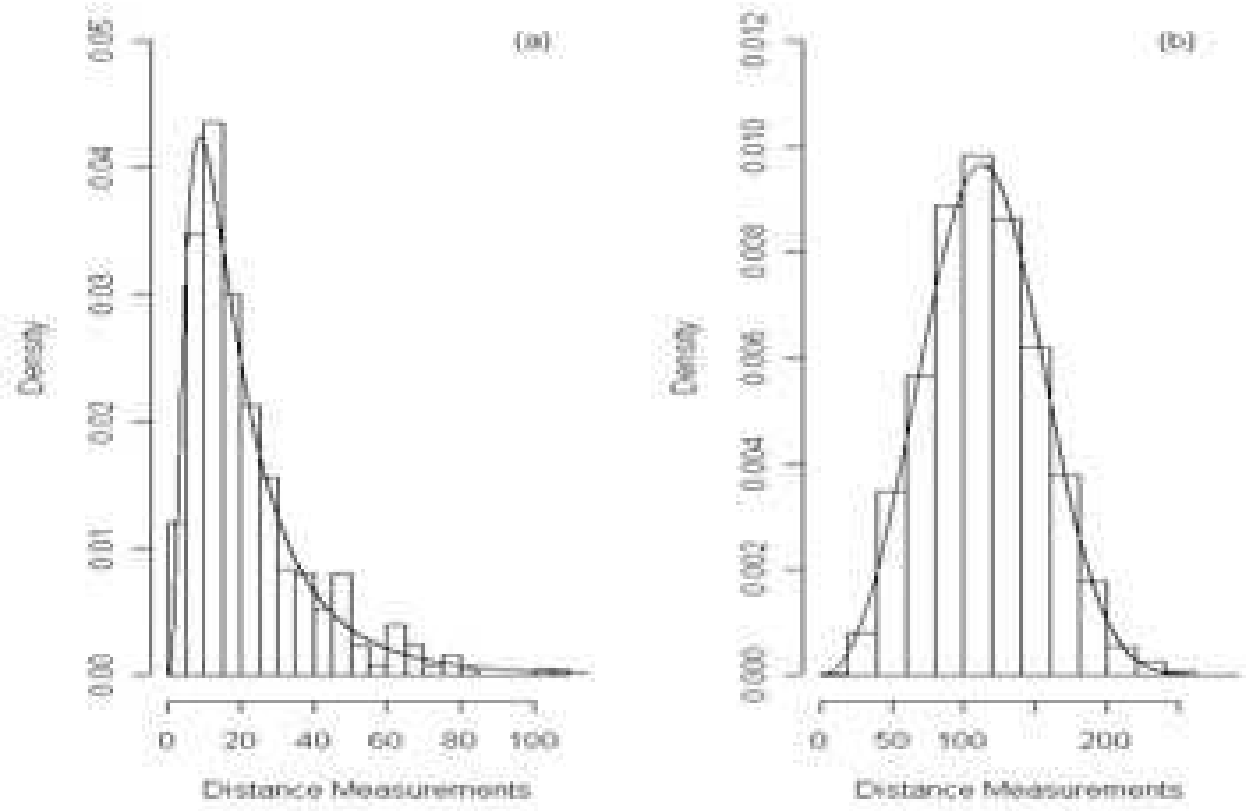
Source: Michigan State University, Ross and Jain (2003)

Model: $X \sim Weibull(\alpha, \beta)$

$$f(x) = \frac{\alpha}{\beta}(\frac{x}{\beta})^{\alpha-1}e^{-(\frac{x}{\beta})^{\alpha}} \tag{2}$$

Estimation via MLE

# Histogram of facial image match scores




(a). Genuine matching scores (b). Imposter matching scores.

# Illustrations

Data: Genuine/Imposter Face, Finger, Hand Geometry Matching Scores

Source: Michigan State University, Ross and Jain (2003)

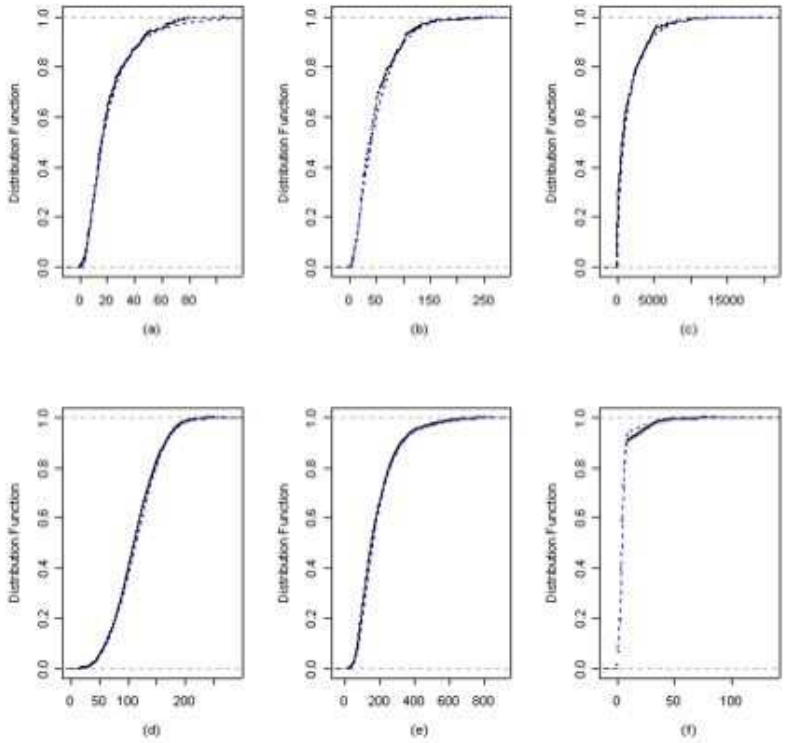
Model: Various Models (see next slide)

Estimation via MLE

# Kolmogorov Goodness-of-Fit Tests

| **Modality** | | |
|---|---|---|
| **Population** | **Fitted Distribution** | ***p*-value** |
| Face | | |
| Genuine | Log-Normal | 0.3030 |
| Imposter | Weibull | 0.1800 |
| Hand Geometry | | |
| Genuine | Gamma | 0.1640 |
| Imposter | Log-Normal | 0.8830 |
| Fingerprint | | |
| Genuine (transformed) | Gamma | 0.6960 |
| Imposter | Truncated Mixture Normal | 0.2010 |

# Empirical CDF and theoretical CDF




(a). Face genuine (b). Hand geometry genuine (c). Fingerprint genuine

(d). Face imposter (e). Hand geometry imposter (f). Fingerprint imposter

# Decision Data

Schuckers (2003) used Beta-binomial to model Decision Data

Data: Face, Fingerprint, Hand geometry Decision data

Source: Michigan State University, Ross and Jain (2003)

Model: $X \sim Betabin(m, \pi, \rho)$

# Goodness-of-fit Hand Geometry FMR

| Threshold | $\hat{\pi}$ | p-value |
|---|---|---|
| 80 | 0.1136 | 0.0017 |
| 70 | 0.0637 | 0.1292 |
| 60 | 0.0272 | 0.6945 |
| 50 | 0.0098 | 0.9998 |
| 40 | 0.0016 | 0.9972 |
| 30 | 0.0008 | 0.9996 |

# Goodness-of-fit Facial FNMR

| Threshold | $\hat{\pi}$ | p-value |
|---|---|---|
| 45 | 0.1060 | 0.2614 |
| 50 | 0.0660 | 0.9509 |
| 55 | 0.0540 | 0.5353 |
| 60 | 0.0500 | 0.5885 |
| 65 | 0.0300 | 0.9216 |
| 70 | 0.0180 | 0.9067 |
| 75 | 0.0140 | 0.9067 |
| 80 | 0.0060 | 0.9985 |
| 85 | 0.0040 | 0.9996 |
| 90 | 0.0040 | 0.9996 |
| 95 | 0.0040 | 0.9996 |
| 100 | 0.0040 | 0.9996 |
| 105 | 0.0020 | 1.0000 |

# Summary

- Third approach to use of simulation: model but verify
- Guidelines: Flexible, parsimonious, consistent
- Taxonomy of data
- Illustrated methods

***Thank You***

schuckers@stlawu.edu