Computational Strategies for Toolmarks: Principal Component Analysis and Other Methods
Outline

• Introduction
• Details of Our Approach
  • Data acquisition
  • Methods of statistical discrimination
  • Error rate estimates
  • Measures of association quality
  • Future directions
• All impressions made by tools and firearms can be represented as numerical patterns
  – Machine learning trains a computer to recognize patterns
    • Can give “…the quantitative difference between an identification and non-identification”\textsuperscript{Moran}
    • Can yield identification error rate estimates
    • May be even confidence measures for I.D.s…….
Data Acquisition

- Obtain striation/impression patterns from 3D confocal microscopy
- Store files in ever expanding database
- Data files are available to practitioner and researcher community through web interface
Glock 19 fired cartridge cases
Screwdriver Striation Patterns in Lead

2D profiles

3D surfaces (interactive)
Mean total profile:

Mean “waviness” profile:

Mean “roughness” profile:
Profile Simulator

- We can simulate profiles as well
- Based on DWT MRA
  - May shed light on processed generating surfaces
  - Should be extendable to 2D striations/impressions…

![Graphs showing mean profile, MRA coefficients, real profiles, and simulated profiles.](image)
What Statistics Can Be Used?

- Multivariate statistical pattern comparison!
- Modern algorithms are called machine learning
- Idea is to measure features of the physical evidence that characterize it
- Train algorithm to recognize “major” differences between groups of features while taking into account natural variation and measurement error.
Setup for Multivariate Analysis

- Need a data matrix to do machine learning

\[
X = \begin{bmatrix}
X_{i1} & \ldots & X_{ij} & \ldots & X_{ip} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
X_{ni} & \ldots & X_{nj} & \ldots & X_{np}
\end{bmatrix}
\]

Represent as a vector of values

\{-4.62, -4.60, -4.58, \ldots\}

- Each profile or surface is a row in the data matrix
- Typical length is \(~4000\) points/profile
- 2D surfaces are far longer
- PCA can:
  - Remove much of the redundancy
  - Make discrimination computations far more tractable
  - HIGHESTLY REDUNDANT representation of surface data
• 3D PCA 24 Glocks, 720 simulated and real primer shear profiles:

• ~47% variance retained

• How many PCs should we use to represent the data??
  • No unique answer

• FIRST we need an algorithm to I.D. a toolmark to a tool
Support Vector Machines

- Support Vector Machines (SVM) determine efficient association rules
  - In the absence of any knowledge of probability densities

SVM decision boundary
How many Principal Components should we use?

With 7 PCs, expect ~3% error rate

With 13 PCs, expect ~1% error rate
Error Rate Estimation

- **Cross-Validation**: hold-out chunks of data set for testing
  - Known since 1940s
  - Most common: **Hold-one-out**

- **Bootstrap**: Randomly selection of observed data (with replacement)
  - Known since the 1970s
  - Can yield *confidence intervals around error rate estimate*

- **The Best**: Small training set, BIG test set
18D PCA-SVM Primer Shear I.D. Model, 2000 Bootstrap Resamples

Refined bootstrapped I.D. error rate for primer shear striation patterns = 0.35%
95% C.I. = [0%, 0.83%]
(sample size = 720 real and simulated profiles)
How good of a “match” is it?
Conformal Prediction

• Can give a judge or jury an easy to understand measure of reliability of classification result
  • Confidence on a scale of 0%-100%

• This is an orthodox “frequentist” approach

• Developed from principals known since the 1930s
Empirical Bayes’

• Computer outputs a “match”
  • What’s the probability it is truly not a “match”?

Get it from Bayes’ Rule:

\[
\Pr(S^- | t^+) = \frac{\Pr(t^+ | S^-) \Pr(S^-)}{\Pr(t^+)}
\]

Probability of no actual association given a test/algorithm indicates a positive ID

Name: Posterior error probability (PEP)
Empirical Bayes’

- Use Brad Efron’s machinery for “empirical Bayes’ two-groups model”
  - Get a calibrated PEP model

The SVM alg got these Primer shear IDs wrong
Empirical Bayes’

- Model’s use with crime scene “unknowns”:

This is the estimated posterior probability of no association:

\[ P(\text{S} \perp \text{I}, \text{est}) = 0.00027 = 0.027\% \]

Computer outputs “match” for:
unknown crime scene toolmarks-with knowns from “Bob the burglar” tools

This is an uncertainty in the estimate.
Future Directions

- **Extend ImageJ** surface metrology functionality
- **Eliminate alignment** step
  - Try invariant feature extraction
- **Parallel** implementation of computationally intensive routines
- **Standards board** to review statistical methodology/algorithms
Acknowledgements

• Research Team:
  Practitioners/academics
  • Mr. Peter Diaczuk
  • Ms. Carol Gambino
  • Dr. James Hamby
  • Dr. Brooke Kammrath
  • Dr. Thomas Kubic
  • Mr. Chris Lucky
  • Off. Patrick McLaughlin
  • Dr. Linton Mohammed
  • Mr. Jerry Petillo
  • Mr. Nicholas Petraco
  • Dr. Graham Rankin
  • Dr. Jacqueline Speir
  • Dr. Peter Shenkin
  • Mr. Peter Tytell

  Grad/Undergrad students
  • Helen Chan
  • Julie Cohen
  • Aurora Dimitrova
  • Frani Kammerman
  • Loretta Kuo
  • Dale Purcel
  • Stephanie Pollut
  • Chris Singh
  • Melodie Yu
Website Information and Reprints/Preprints:

toolmarkstatistics.no-ip.org/
npetraco@gmail.com
npetraco@jjay.cuny.edu