Evaluating the Probability of Identification in the Forensic Sciences

Sargur N. Srihari

University at Buffalo, The State University of New York

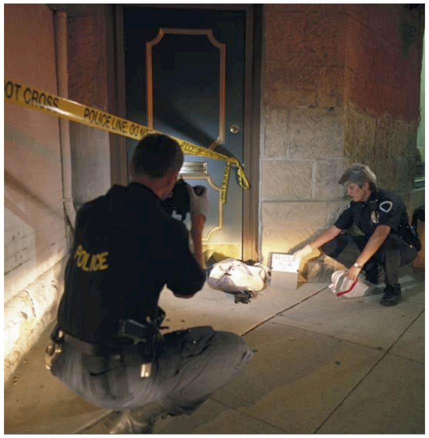
USA

ICFHR Plenary, Bari, Italy

September 2012

Plan of Discussion

- Forensic Identification
- Probability of Identification
- Computational Intractability
- Distance Methods
- Distance and Rarity
- Application to Handwriting

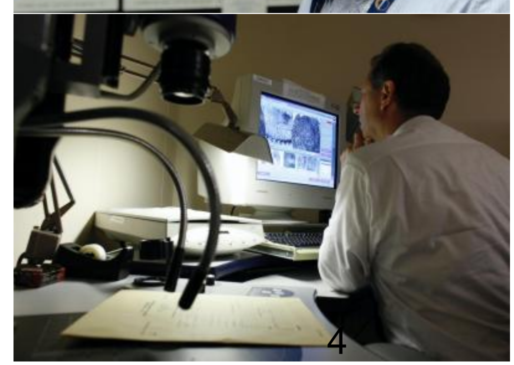


The CSI Effect

- CSI: Popular Television
  - On CBS for 12 years
- Perceptions:
  - Forensics uses high technology
    - Little subjectivity
    - No false positives
- Consequences
  - Jury expectations high
  - Glamorous Field

Controversies

- Innocence Project
  – Several exonerations based on DNA
- Cameron Todd Willingham
  – Possibly Innocent of Arson, Executed 2004
    - New Yorker Sept 7, 2009
- San Jose police withheld uncertainties in fingerprint cases
    - Mercury News Mar 7, 2009



2004 Madrid train bombings

- On the morning of March 11, 2004
- Series of coordinated bombings against commuter train system (4 trains)
- Killing 191 people and injuring 2,000 others

The curious case of Brandon Mayfield

- Spanish National Police (SNP) recovered a latent fingerprint (LFP 17)
  - Partial on a plastic bag of detonators in van used by perpetrators
- FBI IAFIS: LFP 17 has potential match with Mayfield, a lawyer in Oregon
  - Prints from 1984 when arrested for burglary as a teenager
  - Confirmed by 3 FBI examiners + an outside consultant
- Mayfield had never been to Spain
  - Passport at the time had expired
  - Converted to Islam, married an Egyptian
  - Represented man in child custody who turned out to be a jihadist
- On May 6, 2004, FBI arrested Mayfield
- Meanwhile, SNP announced that LFP 17 was sourced to Ounahne Daoud
  - Algerian with criminal record
  - Spanish residency
  - Terrorist links
- FBI concluded that earlier individualization was in error
  - Released 3 weeks after arrest
  - $2 million for mistake
  - apologized to Mayfield

Latent and Known Prints in Mayfield Case

- LFP 17

15 level 2 similarities

- Mayfield fingerprint

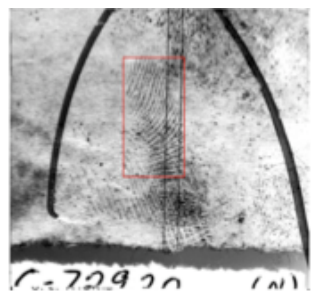
- Daoud fingerprint

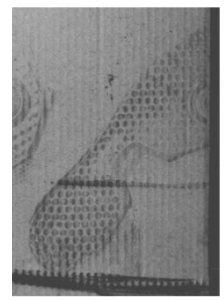
Need for Probabilistic Analysis

- Mayfield highlights need for quantifying uncertainty in latent print analysis
  - Rather than binary decision: individualization or not
- Two types of uncertainty
  - Rarity (how unusual is the evidence)
  - Similarity (between evidence and known)

Forensic Modalities

- Biological Evidence
  - DNA, Blood, Hair
- Impression Evidence
  - Latent Prints, Handwriting, Shoe Prints





- Trace Evidence
  - Pollen, Fiber, Paint, Glass

National Academy of Sciences

Committee on identifying the needs of the forensic science community (2007-09)

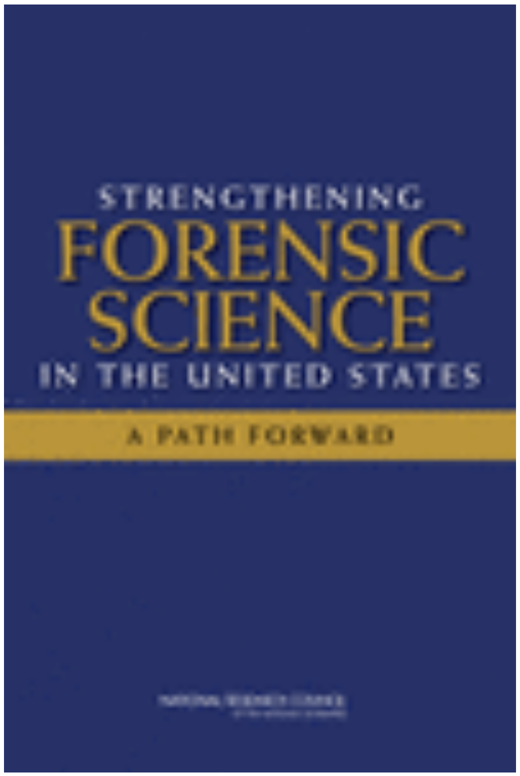
1. **HARRY T. EDWARDS**, (*Co-chair*), Judge, U.S. Court of Appeals, District of Columbia Circuit
2. **CONSTANTINE GATSONIS**, (*Co-chair*), Director, Center for Statistical Sciences, Brown University

3. **MARGARET A. BERGER**, Suzanne J. and Norman Miles Professor of Law, Brooklyn Law School
4. **JOE S. CECIL**, Project Director, Program on Scientific and Technical Evidence, Federal Judicial Center
5. **M. BONNER DENTON**, Professor of Chemistry, University of Arizona
6. **MARCELLA F. FIERRO**, Medical Examiner of Virginia
7. **KAREN KAFADAR**, Rudy Professor of Statistics and Physics, Indiana University

8. **PETE M. MARONE**, Director, Virginia Department of Forensic Science
9. **GEOFFREY S. MEARNS**, Dean, Cleveland-Marshall College of Law, Cleveland State University
10. **RANDALL S. MURCH**, Director, Research, Virginia Polytechnic Institute and State University
11. **CHANNING ROBERTSON**, Bowes Professor, Dean of Faculty, Dept Chemical Engg, Stanford University
12. **MARVIN E. SCHECHTER**, Attorney

13. **ROBERT SHALER**, Professor, Biochemistry, The Pennsylvania State University
14. **JAY A. SIEGEL**, Professor, Forensic Program, Indiana University-Purdue University
15. **SARGUR N. SRIHARI**, SUNY Distinguished Professor, Dept Computer Scien & Engg, University at Buffalo
16. **SHELDON M. WIEDERHORN** (NAE), Senior Fellow, National Institute of Standards and Technology
17. **ROSS E. ZUMWALT**, Chief Medical Examiner, State of New Mexico

NAS Committee Report



Released March 2009
National Academies Press

Committee at
NAS, Woods Hole, MA

NAS Recommendations (13)

(Four Relevant to Computational Forensics)

1. New federal entity, the National Institute of Forensic Sciences (NIFS) to focus on:

   peer-reviewed research
   support to forensic practices
   development of new technologies
   standardized terminology and reporting

3. Studies on validity of forensic methods

   development and establishment of quantifiable measures of reliability and accuracy

NAS Committee Recommendations

5. Research on human observer bias

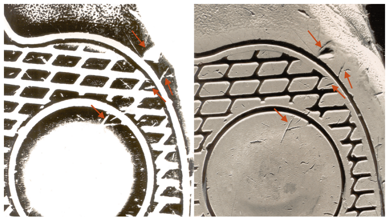
   and human error in forensic examinations

12. Baseline standards for fingerprints

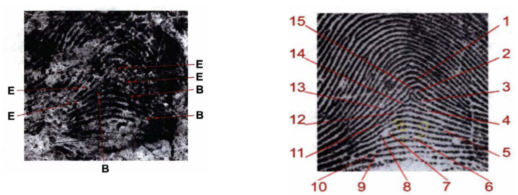
    to map, record, and recognize features
    improvement and characterization of
    accuracy of algorithms

The Identification Task

- Impression Evidence:
  materials with characteristics of impressed objects
  - Footwear impressions

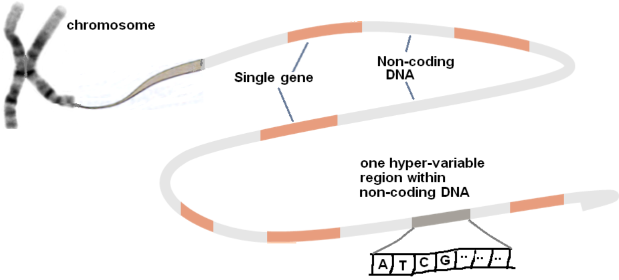


  - latent fingerprints



  - Handwriting samples

- Biological Evidence
  - DNA



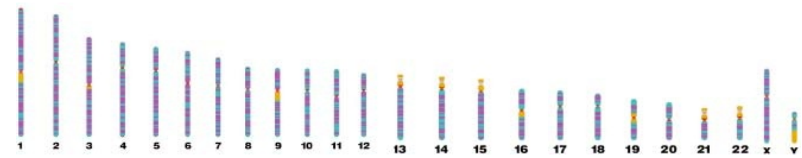
  - Blood type

Forensic Opinion and Individualization

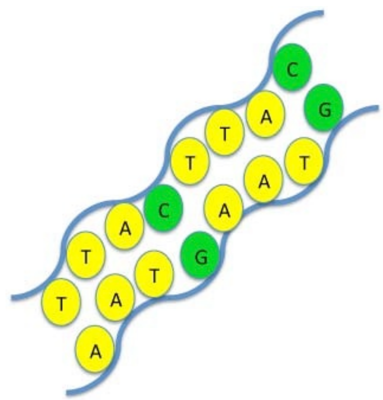
- Courts allow opinion on individualization
  - Evidence attributed to a one individual and no other
  - Three possible opinions for evidence
    - Individualization
      - » No other individual on earth
    - Inconclusive
    - Exclusion
      - » Definitely not this individual
- Need for characterizing degree of uncertainty
  - Degree of uniqueness: rarity (1 in a billion)
  - Probability of identification

DNA Evidence

Genome: sequence of 3×10^9 base-pairs (nucleotides A,C,G,T)
Represents full set of chromosomes



Actual Electron photomicrograph



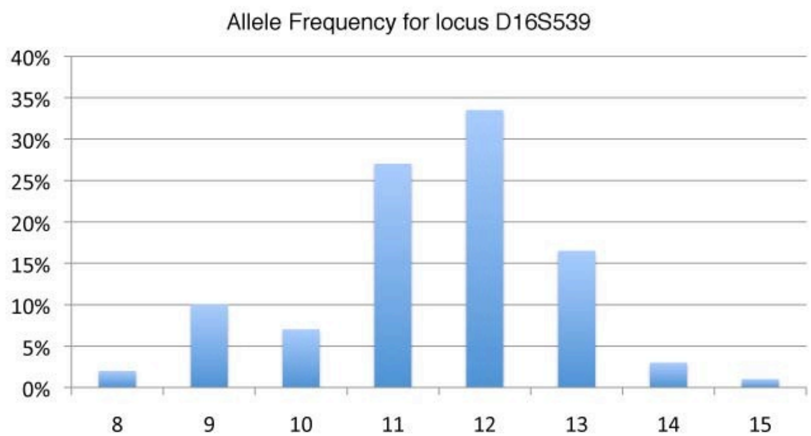
Single Chromosome: ~10^8 base-pairs

Genome has 46 chromosomes
(22 are repeated plus XX and XY)

Large portions of DNA have no survival function (98.5%) called “junk DNA” and have variations useful for identification
Combined DNA Identification System (CODIS) identifies 13 markers CSF1PO,D3S1358,D7S820,..
TH01 is a location on short arm of chromosome 11:
short tandem repeats (STR) of same base pair AATG
Variant forms (alleles) different for different individuals

DNA Profile Probability

Allele Frequency of single locus for 200 individuals



Individual alleles are assumed Independent.

Probability of profile obtained by multiplying individual probabilities

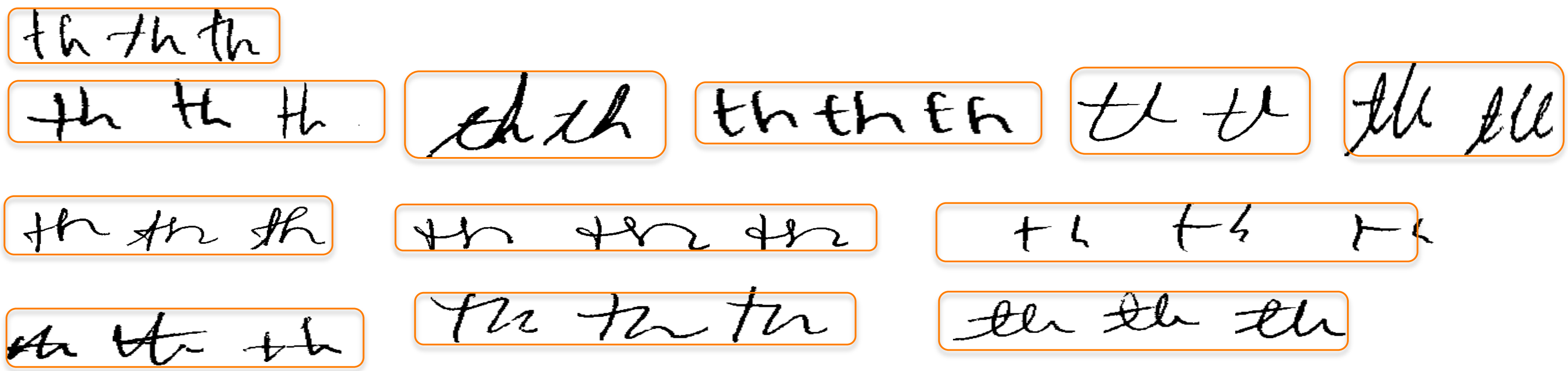
DNA profile of 13 loci:

Average match probability (PRC) is 0.1 per locus, 10^{-13} for a profile

DNA Profile		Allele frequency from database				Genotype frequency for locus	
Locus	Alleles	Times allele observed	Size of database	Frequency		formula	number
CSF1PO	10	109	432	p =	0.25	$2pq$	0.16
	11	134		q =	0.31		
TPOX	8	229	432	p =	0.53	p^2	0.28
	8						
THO1	6	102	428	p =	0.24	$2pq$	0.07
	7	64		q =	0.15		
vWA	16	91	428	p =	0.21	p^2	0.05
	16						
			Profile frequency =				0.00014

Probability of one in 10 trillion of another individual with this profile

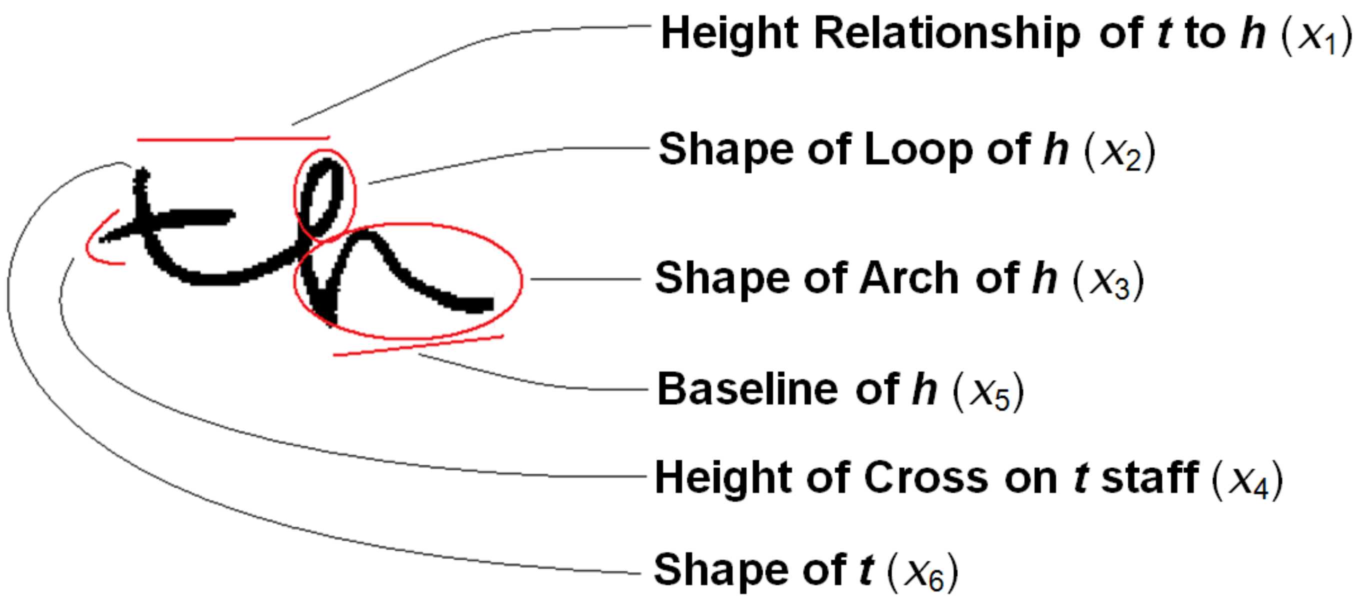
Handwriting Evidence in Forensic QD



Task is to determine whether two writing samples originated from the same individual

We have to take into account both similarity of Known and Questioned as well as the rarity

Handwriting Features (QDE)



R = **Height Relationship of *t* to *h***	L = **Shape of Loop of *h***	A = **Shape of Arch of *h***	C = **Height of Cross on *t* staff**	B = **Baseline of *h***	S = **Shape of *t***
r^0 = *t* shorter than *h*	l^0 = retraced	a^0 = rounded arch	c^0 = upper half of staff	b^0 = slanting upward	s^0 = tented
r^1 = *t* even with *h*	l^1 = curved right side and straight left side	a^1 = pointed	c^1 = lower half of staff	b^1 = slanting downward	s^1 = single stroke
r^2 = *t* taller than *h*	l^2 = curved left side and straight right side	a^2 =no set pattern	c^2 = above staff	b^2 = baseline even	s^2 = looped
r^3 = no set pattern	l^3 = both sides curved		c^3 = no fixed pattern	b^3 = no set pattern	s^3 = closed
	l^4 = no fixed pattern				s^4 = mixture of shapes

Defines a multinomial distribution

Probabilistic Approach to Forensic Identification

- Object (Known): o
- Evidence (Questioned): e
- Models:
  - M_o: Object and evidence are from the same source
  - M_1: Object and evidence are from different sources
- Model priors: $p(M_0)$ and $p(M_1)$ $\quad \sum_i p(M_i) = 1$
- Need to determine the posteriors $p(M_i | o,e)$
- Referred to as the probability of identification

Probability of Identification

Likelihood Ratio: $LR(o,e) = \dfrac{p(o,e \mid M_0)}{p(o,e \mid M_1)}$

Prior Odds: $O_{prior} = \dfrac{p(M_0)}{p(M_1)}$

Posterior Odds: $O_{posterior} = O_{prior} \times LR(o,e)$

Probability of Identification under equal priors:

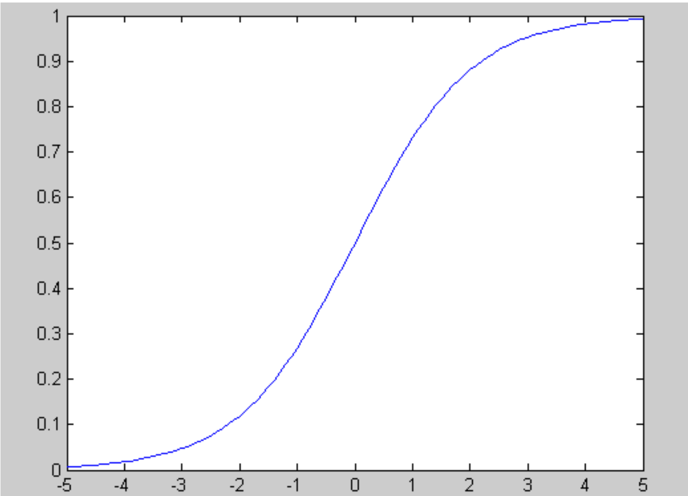
$$p(M_0 \mid o,e) = \frac{LR(o,e)}{1 + LR(o,e)} = \frac{\exp(LLR(o,e))}{1 + \exp(LLR(o,e))}$$

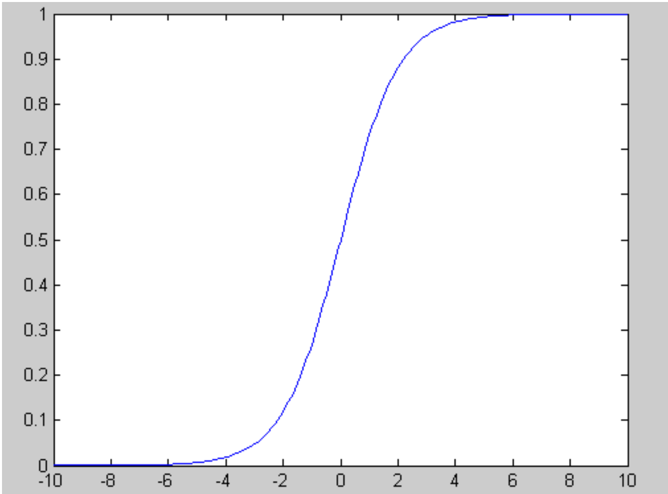
Values of Probability

$$\text{Log-likelihood Ratio: } LLR(o,e) = \frac{\ln p(o,e \mid M_0)}{\ln p(o,e \mid M_1)}$$

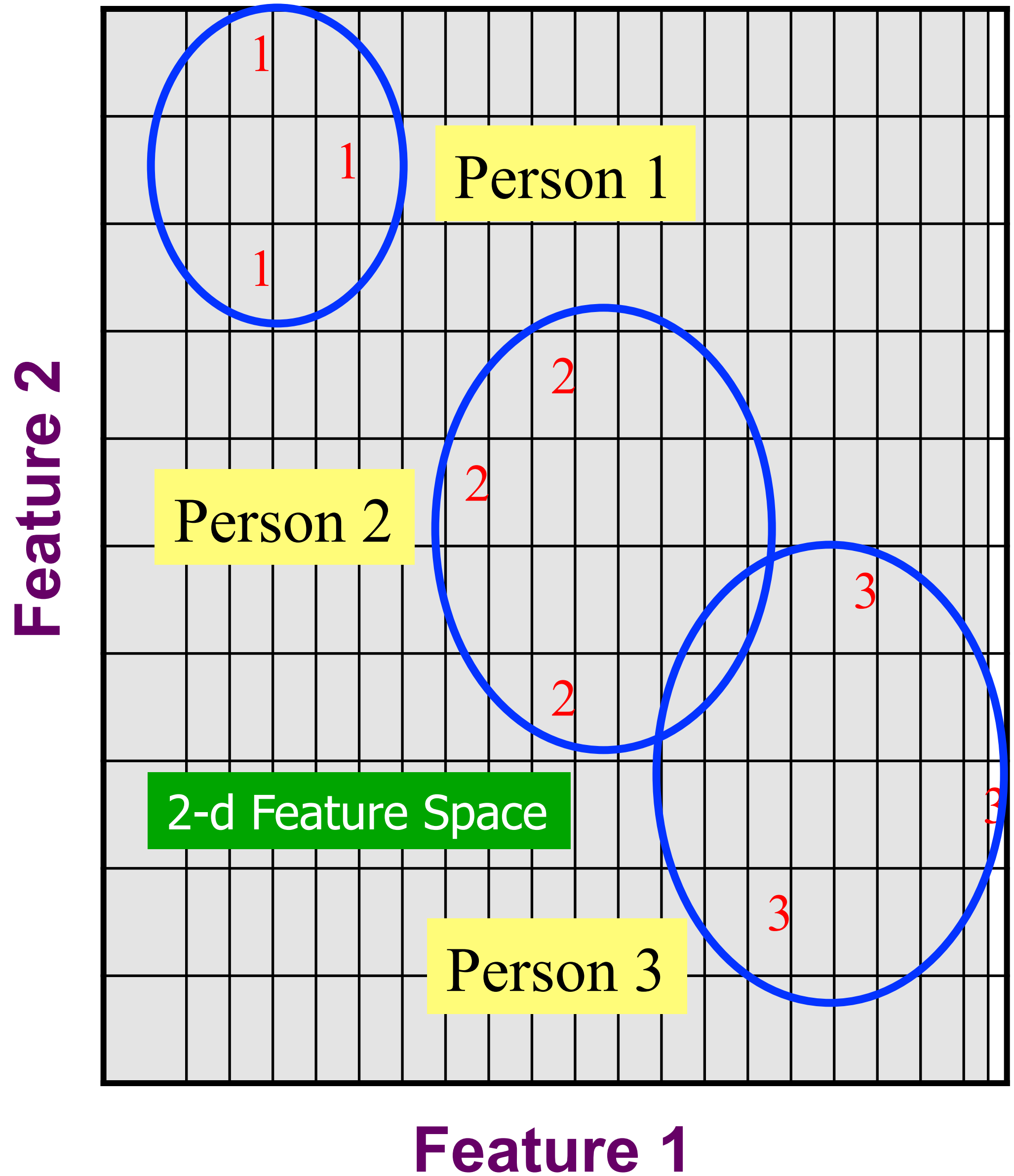
$$p(M_0 \mid o,e) = \frac{\exp(LLR(o,e))}{1+\exp(LLR(o,e))} = \frac{1}{1+\exp(-LLR(o,e))}$$

$$= sigmoid(-LLR(o,e))$$





Feature Space View



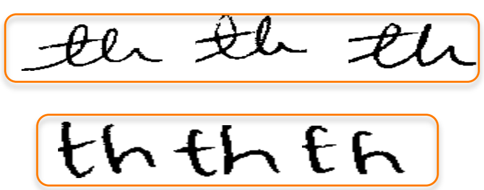
- Consider joint distribution of $p(o,e|M_i)$
- When M_0 holds distribution is defined over pairs from same person
- M_1 will consider pairs from different persons

Intractability of Computing LR from joint distributions

$$\text{Likelihood Ratio: } LR_J(o,e) = \frac{p(o,e \mid M_0)}{p(o,e \mid M_1)}$$

- If o and e have d features each,
  each feature has K discrete values,
  no. of parameters needed is $2K^{2d}$

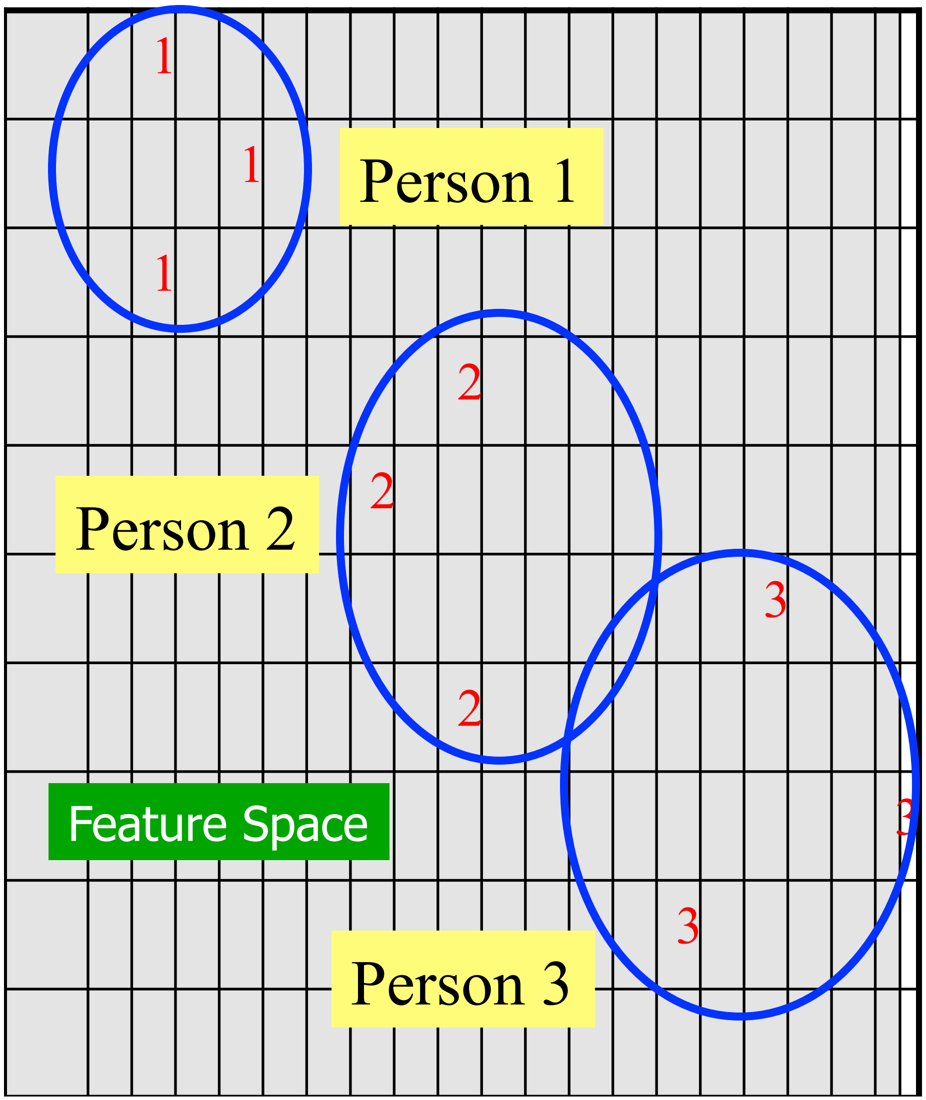
d=6, , *K*=4 or 5, No of parameters= *4,799*



R = **Height Relationship of *t* to *h***	L = **Shape of Loop of *h***	A = **Shape of Arch of *h***	C = **Height of Cross on *t* staff**	B = **Baseline of *h***	S = **Shape of *t***
r^0= *t* shorter than *h*	l^0 = retraced	a^0 = rounded arch	c^0 = upper half of staff	b^0 = slanting upward	s^0 = tented
r^1 = *t* even with *h*	l^1 = curved right side and straight left side	a^1 = pointed	c^1 = lower half of staff	b^1 = slanting downward	s^1 = single stroke
r^2 = *t* taller than *h*	l^2 = curved left side and straight right side	a^2 =no set pattern	c^2 = above staff	b^2 = baseline even	s^2 = looped
r^3 = no set pattern	l^3 = both sides curved		c^3 = no fixed pattern	b^3 = no set pattern	s^3 = closed
	l^4 = no fixed pattern				s^4 = mixture of shapes

No of letter pairs is 325. If we assume 200 legal ones, need 10^6 parameters
100 milllion letter pairs to be manually truthed, features/hw may change

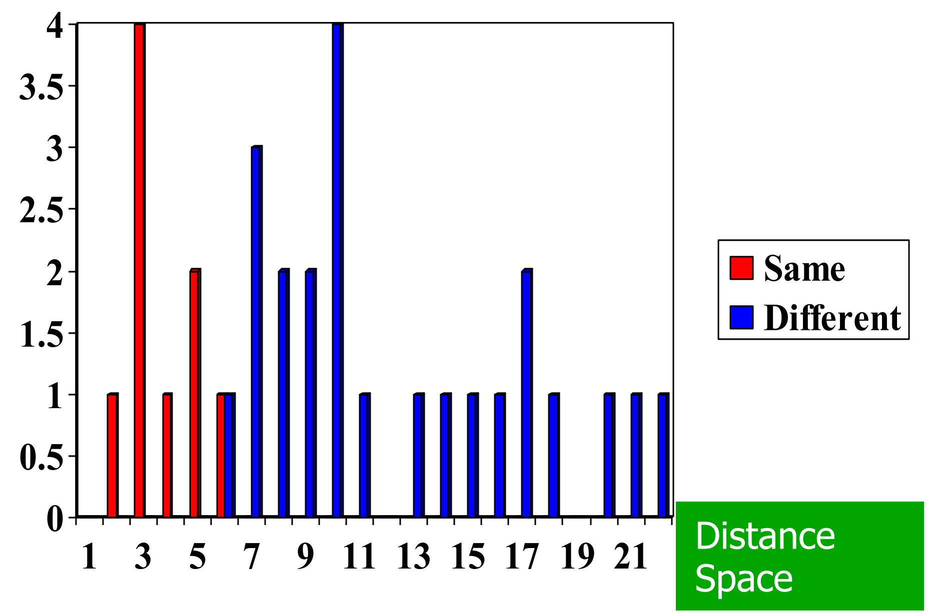
From 2-d *Feature* Space to 1-d *Distance* Space



Manhattan Distances

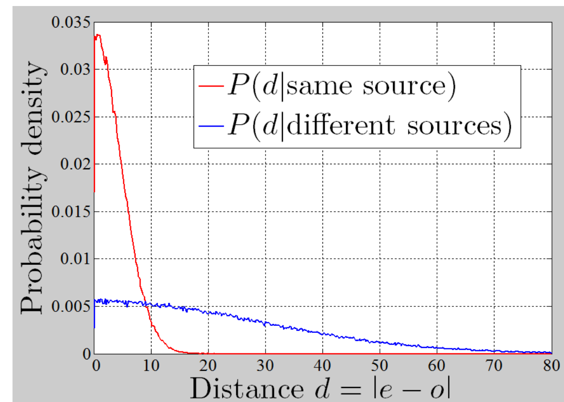
Intra-class (9):
2,3,3, 3,3,4, 5,5,6

Inter-class (27):
10,9,13,19,24,20, 7,6,10,16,17,21
8,7,11,17,18,22, 9,14,10,10,15,11,8,11,7



Distance or Similarity Method

Likelihood Ratio: $LR_D(o,e) = \dfrac{p(d(o,e) \mid M_0)}{p(d(o,e) \mid M_1)}$

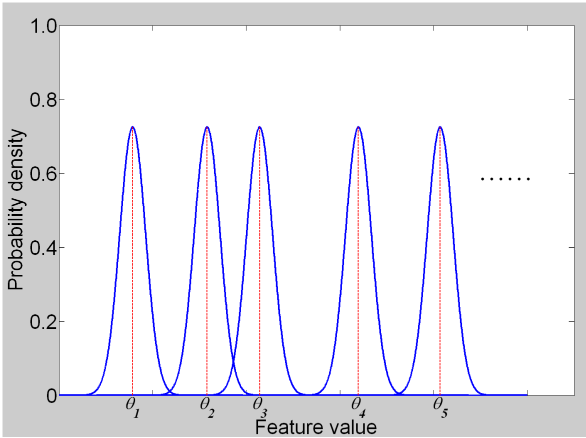


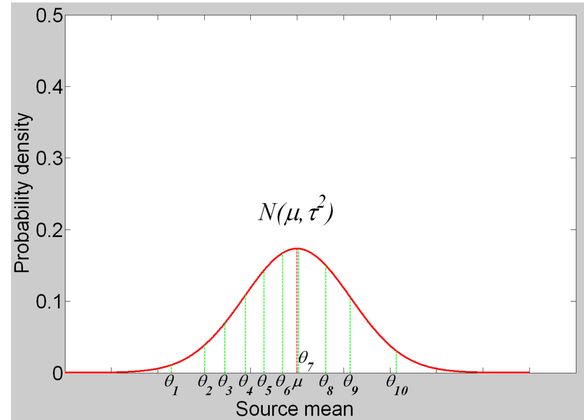
- Maps two multivariate distributions of $2n$ variables each into two univariate ones
- Severe loss of information
- Natural extension is to use vector distance

$$LR_{VD}(o,e) = \frac{p(\mathrm{d}(o,e) \mid M_0)}{p(\mathrm{d}(o,e) \mid M_1)}$$

  – Still there is information loss

Lindley's Result





Each source (individual) is Normally distributed

Sources (individual means) are normally distributed

$$LR = \frac{\tau}{\sqrt{2}\sigma} \exp\left\{-\frac{(o-e)^2}{4\sigma^2}\right\} \exp\left\{\frac{(m-\mu)^2}{2\tau^2}\right\}$$

where $m = (o+e)/2$ *is the mean of* o *and* e.

Product of *Difference* Term and *Rarity* Term

Generalization of Lindley's Result

- Generalize to mutivariate

$$LR_{DR} = P(\mathrm{d}(o,e) \mid M_0) * \frac{1}{P(\mathrm{m}(o,e))}$$

–$d(o,e)$ is vector difference and $m(o,e)$ is mean

- Other data types

1. Binary
   - Difference is 0,1 or -1
   - Mean is 0 if bits are different, 1 otherwise
2. Multinomial
   - Difference of categorical values
3. Graph
   - Difference of features of matching nodes/edges
   - Mean of feature vector of matching nodes/edges

We now have three LR methods

1. Based on Joint Distributions

$$LR_J(o,e) = \frac{p(o,e \mid M_0)}{p(o,e \mid M_1)}$$

2. Based on Distance Distributions

$$LR_D(o,e) = \frac{p(d(o,e) \mid M_0)}{p(d(o,e) \mid M_1)}$$

3. Based on Distance and Rarity Distributions

$$LR_{DR}(o,e) = P(\mathrm{d}(o,e) \mid M_0) * \frac{1}{P(\mathrm{m}(o,e))}$$

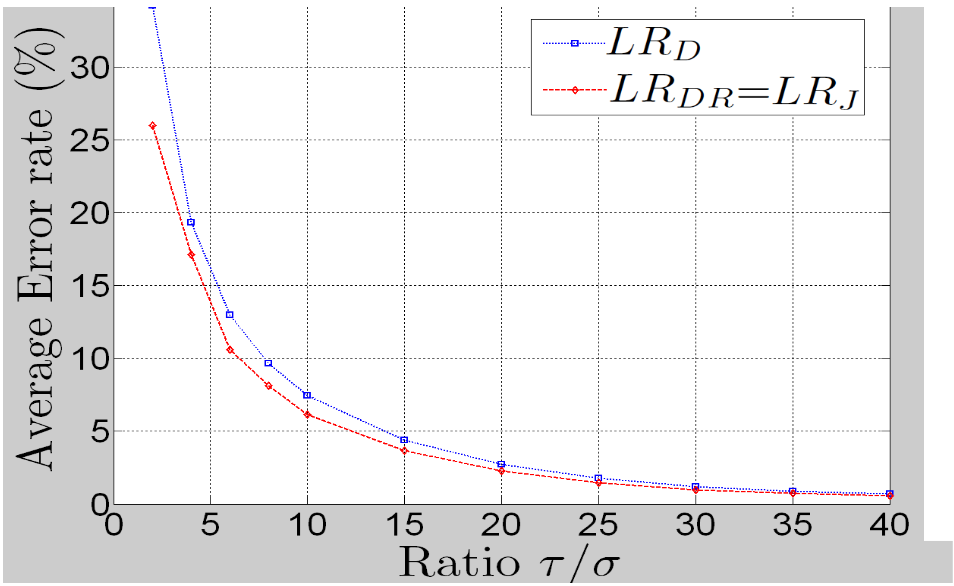
Comparison Metrics

1. Probability of misclassification
   – Determine error based on whether LLR is positive or negative
2. Computation time
3. Scalability
   – Tractability with increasing number of variables

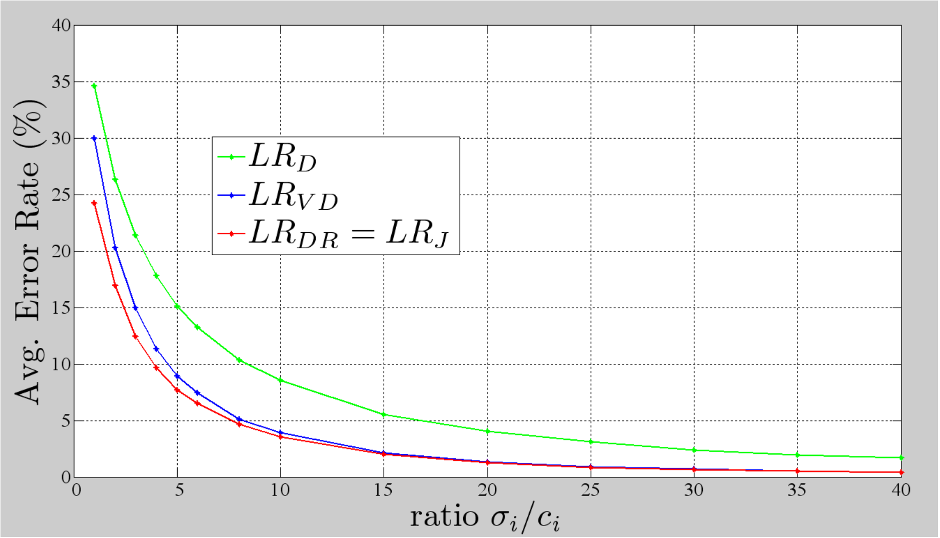
Error Rates of LR methods with Gaussian Data

Distribution of object is:

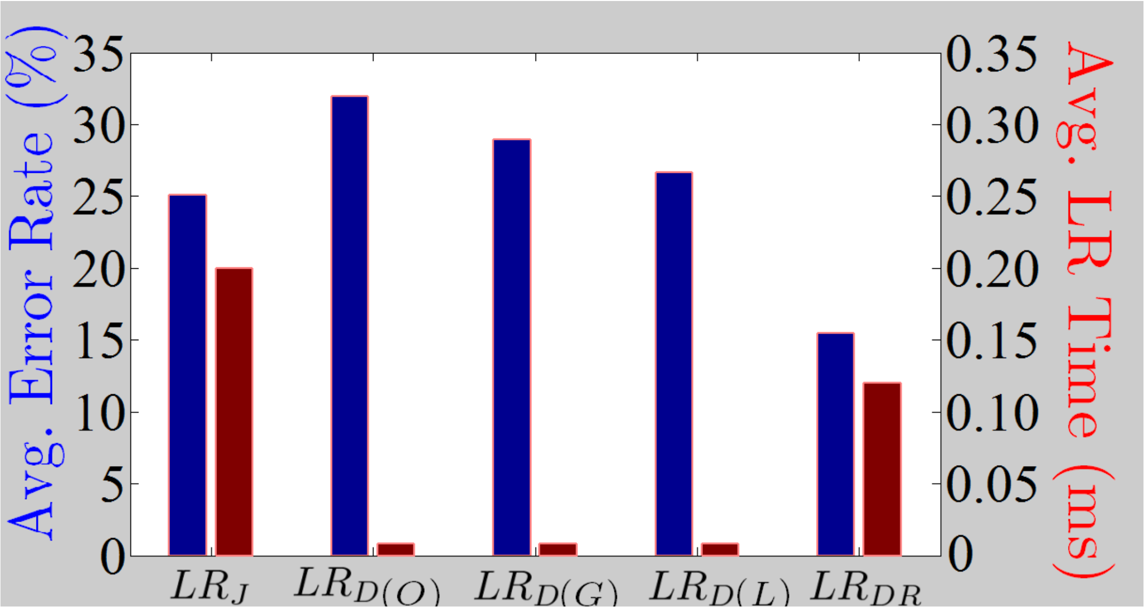
Univariate



5-variate



Error/Speed with Multinomial Data ("*th*")

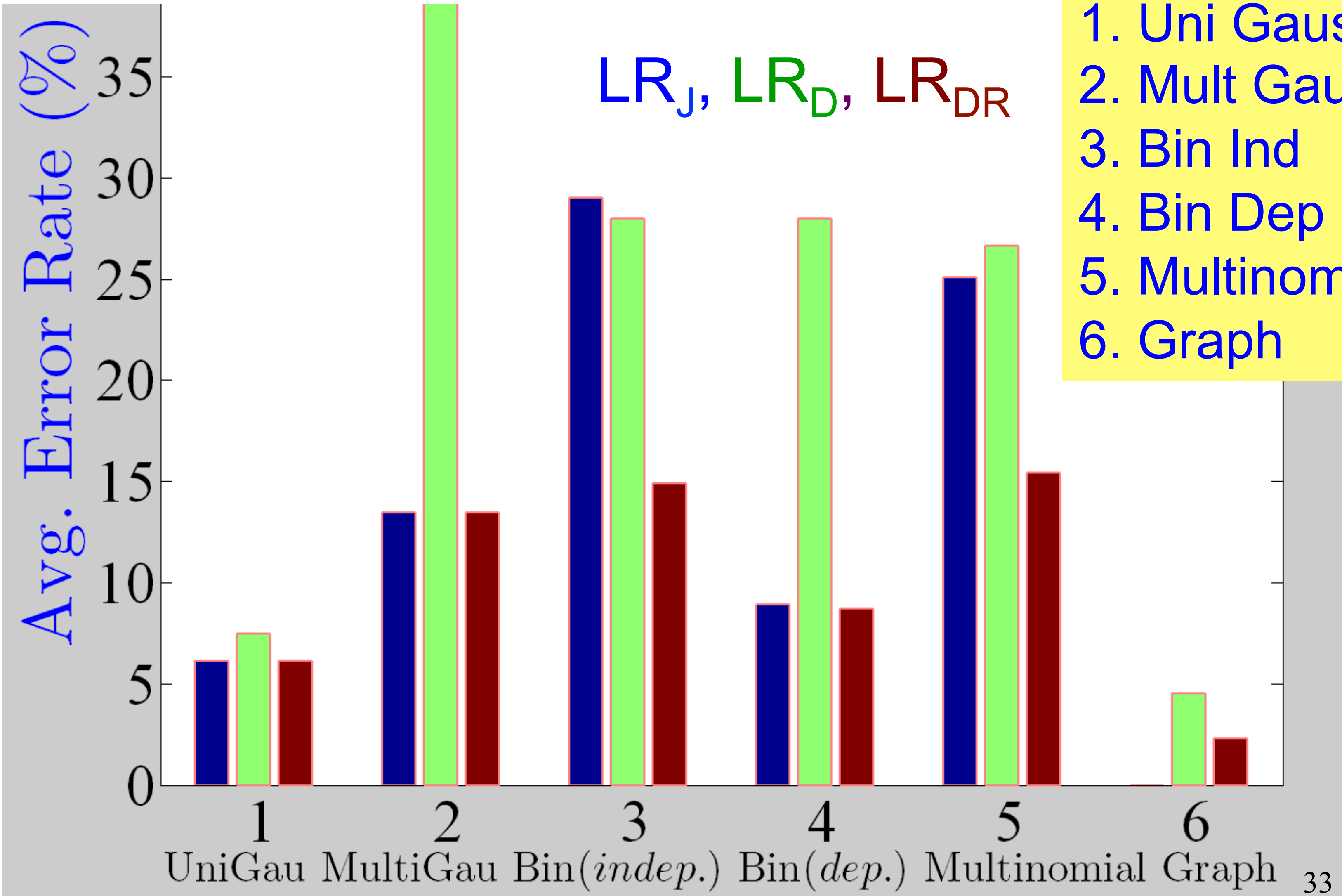


Joint Distribution: assumed independence due to intractability.

Three distance methods:

L: Lin
O: overall frequency
G: Goodall

Three LR methods with six data types



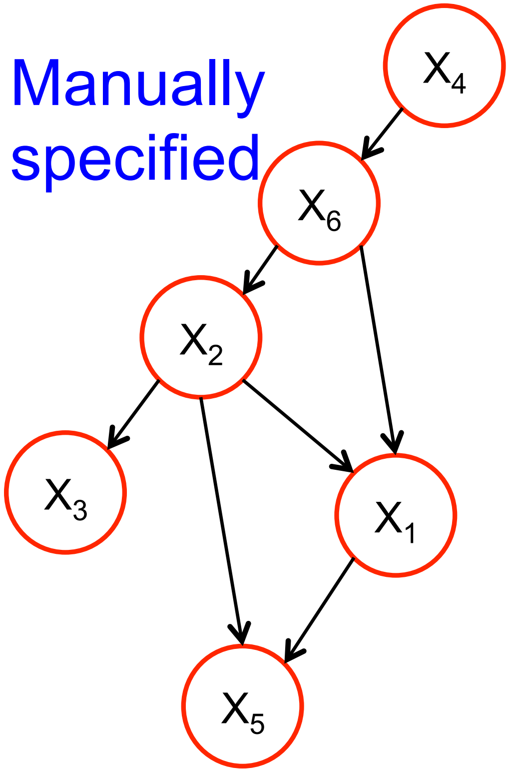
1. Uni Gauss
2. Mult Gauss
3. Bin Ind
4. Bin Dep
5. Multinomial
6. Graph

Computation of LR_{DR}

- Need two distributions (difference and mean) each with n variables
  - No. of parameters is $2K^n$ (where K=no. of values)
  - As opposed to $2K^{2n}$ with LR_J
  - We are dealing with two n-dimensional distributions rather than two $2n$ dimensional distributions
- Still exponential with n: scalability is an issue
- Solution is to use PGMs of feature variables
  - Bayesian networks
  - Markov Networks

Parameter Learning For BNs

Manually specified



Conditional Probability Distributions

Max Likelihood Estimates

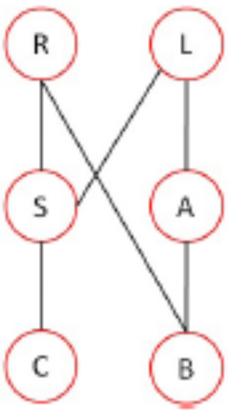
	$X_5 = 0$	$X_5 = 1$	$X_5 = 2$	$X_5 = 3$
$X_1 = 0, X_2 = 0$	0.50	0	0	0.50
$X_1 = 0, X_2 = 1$	0	1.00	0	0
$X_1 = 0, X_2 = 2$	0.18	0.36	0.27	0.18
$X_1 = 0, X_2 = 3$	0.27	0.40	0.30	0.03
$X_1 = 0, X_2 = 4$	0.22	0.45	0.28	0.05
$X_1 = 1, X_2 = 0$	0.43	0	0.28	0.29
$X_1 = 1, X_2 = 1$	NaN	NaN	NaN	NaN
$X_1 = 1, X_2 = 2$	0.39	0.06	0.33	0.22
$X_1 = 1, X_2 = 3$	0.33	0.17	0.33	0.17
$X_1 = 1, X_2 = 4$	0.42	0.11	0.29	0.18
……	……	……	……	……

Bayesian Estimates (Dirichlet priors)

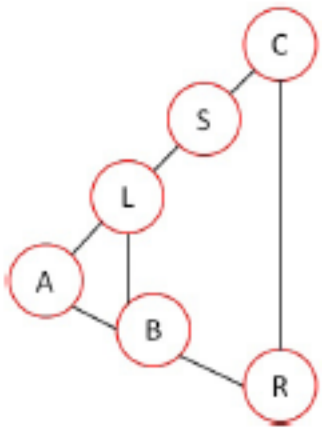
	$X_5 = 0$	$X_5 = 1$	$X_5 = 2$	$X_5 = 3$
$X_1 = 0, X_2 = 0$	0.29	0.14	0.29	0.29
$X_1 = 0, X_2 = 1$	0.25	0.25	0.25	0.25
$X_1 = 0, X_2 = 2$	0.25	0.38	0.25	0.12
$X_1 = 0, X_2 = 3$	0.22	0.41	0.31	0.06
$X_1 = 0, X_2 = 4$	0.16	0.52	0.25	0.07
$X_1 = 1, X_2 = 0$	0.29	0.14	0.29	0.29
$X_1 = 1, X_2 = 1$	0.25	0.25	0.25	0.25
$X_1 = 1, X_2 = 2$	0.37	0.05	0.47	0.11
$X_1 = 1, X_2 = 3$	0.33	0.22	0.33	0.11
$X_1 = 1, X_2 = 4$	0.38	0.13	0.29	0.20
……	……	……	……	……

Difficult to specify causality of handwriting features

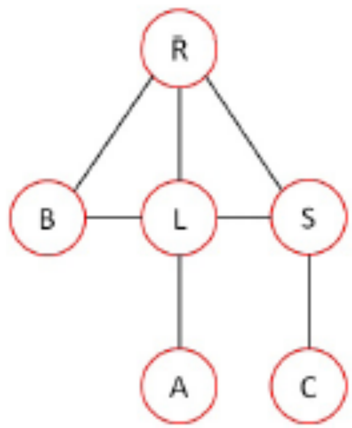
Markov Structure Learning



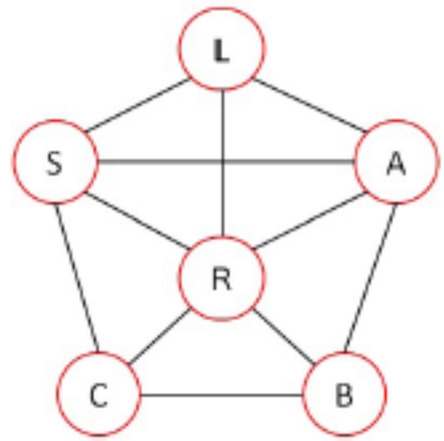
Manual



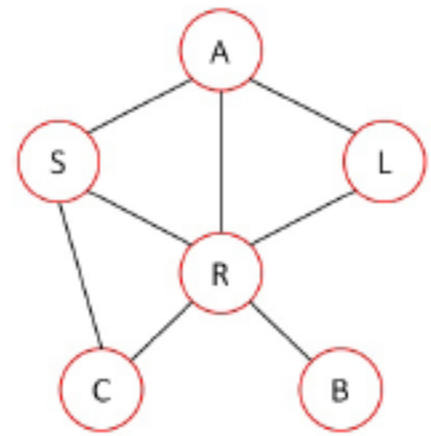
Manual



Modified Chow-Liu

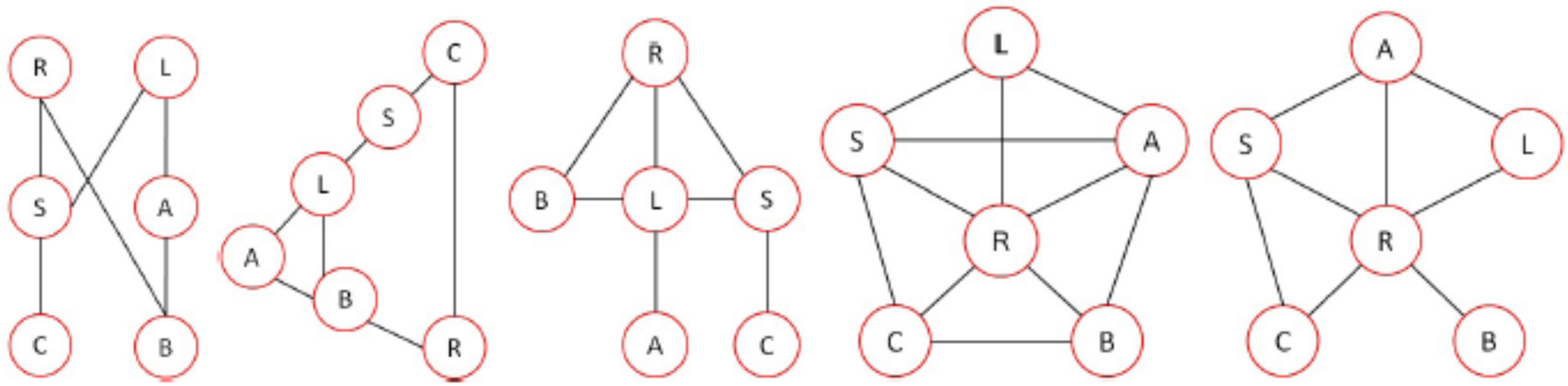


Greedy L_1-Reg.



Fast Greedy.

Evaluation of Rarity using MNs



MRF index	Construction method	Time of structure learning (sec)	Average log-loss	Probability of 'th' in Fig. 3(f)	Probability of 'th' in Fig. 3(g)
MRF_1	Manual	n/a	6.428	$1.59 \cdot 10^{-2}$	$16 \cdot 10^{-6}$
MRF_2	Manual	n/a	6.464	$1.82 \cdot 10^{-2}$	$6 \cdot 10^{-6}$
MRF_3	Mod. Chow–Liu	2	6.426	$1.7 \cdot 10^{-2}$	$5 \cdot 10^{-6}$
MRF_4	Greedy w. L_1–reg	53	6.326	$1.94 \cdot 10^{-2}$	$25 \cdot 10^{-6}$
MRF_5	FGA	7	6.328	$1.8 \cdot 10^{-2}$	$21 \cdot 10^{-6}$

Rare and Common Style Conditional nPRCs

Rare Styles : Looped or tented ‘t’, loop of ‘h’ with both sides curved

Author:1025c
nPRC: 4.17×10^{-8}

Author: 1409c
nPRC: 1.23×10^{-6}

Author: 1479b
nPRC: 1.66×10^{-5}

Common Styles: Single stroke ‘t’, retraced ‘h’, pointed arch of ‘h’, baseline of ‘h’ slanting down, ‘t’ taller, cross of ‘t’ below

Author: 40b
nPRC: 0.76

Author: 130b

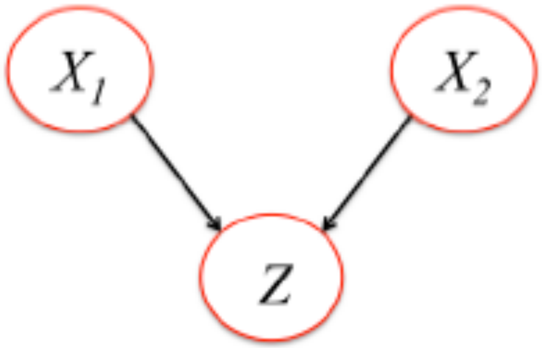
Author: 1007c

Author: 685a

PRC = 0.0043 for the dataset. Population size n=100

Rarity Metrics: PRC in database of size n

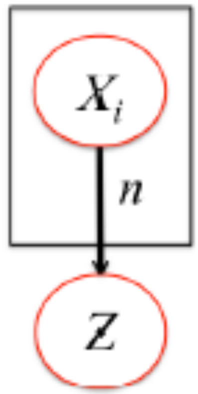
PRC



$$\rho = P(z^0) = \sum\sum P(z^0|X_1,X_2)P(X_1)P(X_2)$$

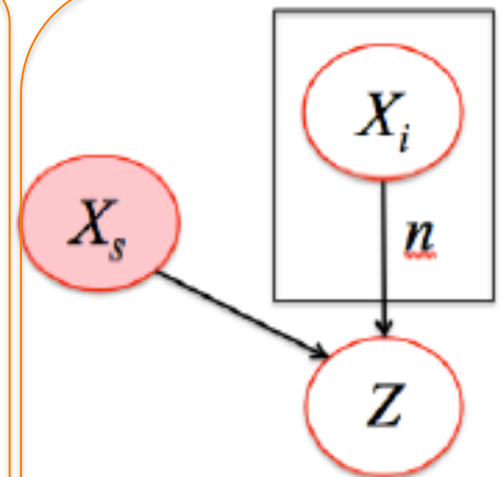
$$P(z^0|X_1,X_2) = \begin{cases} 1 \text{ if } d(X_1,X_2) \le \epsilon \\ 0, otherwise, \end{cases}$$

nPRC



$$\rho[n] = 1 - (1-\rho)^{\frac{n(n-1)}{2}}.$$

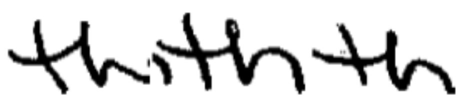
Conditional nPRC



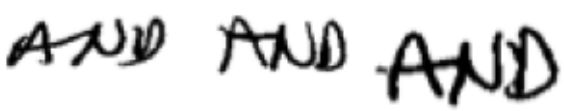
$$p(Z=1|X_s) = \sum_{\mathbf{X}} p(Z=1|X_s,\mathbf{X})p(\mathbf{X})$$

For identical match $1-(1-P(X_s))^n$

Rare



nPRC=1.17 x10^{-5}



nPRC=2.14 x 10^{-8}

Common

nPRC=0.156

nPRC=0.166

Summary

1. Forensic Identification facing court scrutiny
   - Expressing Uncertainty is necessary
     - More uncertainty in impression evidence than DNA
2. LR based on Joint distributions is exact but is intractable
   - Distance based LR are a rough approximation
3. Distance and rarity methods are more accurate
   - But still intractable
4. PGMs provide a solution
     - Markov network structure learning is needed

Credits

- LR Computation methods: Yi Tang
- Markov Structure Learning: Dmitry Kovalenko
- Rarity Evaluation: Chang Su
- Funding from: National Institute of Justice

Word Cloud of IEEE Spectrum Paper December 2010