Statistics and R in Forensic Genetics

UseR! 2016, Stanford University, USA

Mikkel Meyer Andersen
mikl@math.aau.dk

Department of Mathematical Sciences, Aalborg University, Denmark

AALBORG UNIVERSITY
DENMARK

Forensic genetics

- Aims: Identify people and investigate legal issues using genetic evidence
  - Legal issues: criminal, paternity and immigration cases
  - Genetic evidence: blood, saliva, semen, ...
- Unbiased evidence evaluation (using statistics, not subjective assessments)
- Rule out suspects

Trace found at crime scene

1. Trace of genetic evidence from the perpetrator found at crime scene
2. Suspect arrested
3. DNA profiles are compared

Evidential weight

- E: evidence (e.g. DNA profile from crime scene)
- Weight of the evidence (likelihood ratio):

$$LR = \frac{P(E \mid H_p)}{P(E \mid H_d)},$$

- H_p (prosecutor's hypothesis) is 'the suspect is the donor of the genetic data' (often assumed equal to 1)
- H_d (defence attorney's hypothesis) is 'the suspect is unconnected to the crime'
- $P(E \mid H_d)$: Match probability $\approx$ match by chance $\approx$ 'How probable it is that some random man's DNA profile matches the DNA profile found at the crime scene?' (population frequency)

Evidential weight interpretation

- E: evidence (e.g. DNA profile from crime scene)
- H_p (prosecutor's hypothesis) is 'the suspect is the donor of the genetic data'
- H_d (defence attorney's hypothesis) is 'the suspect is unconnected to the crime'
- Ideal usage of LR:

$$\underbrace{\frac{P(H_p \mid E)}{P(H_d \mid E)}}_{\text{Posterior odds}} = \underbrace{\frac{P(E \mid H_p)}{P(E \mid H_d)}}_{\text{LR}} \times \underbrace{\frac{P(H_p)}{P(H_d)}}_{\text{Prior odds}},$$

- Toss a coin 10 times to obtain $E = \{4 \text{ heads}, 6 \text{ tails}\}$
- $H_1 : \theta = 0.5$ vs $H_2 : \theta = 0.9$ ($\theta = P(\text{heads})$)
- $P(\theta = 0.5 \mid E) / P(\theta = 0.9 \mid E)$?
- $P(E \mid \theta = 0.5) = 20.51\%$ and $P(E \mid \theta = 0.9) = 0.01\%$
- $LR = P(E \mid \theta = 0.5) / P(E \mid \theta = 0.9) = 1488$
- $P(H_1) / P(H_2)$ must be known to say anything about posterior odds

DNA

- Bases: A, T, C, G (A-T and C-G)
- 3.3 billion base pairs (3.3 billion = 3,300,000,000)
- 23 chromosome pairs
- In each pair: One chromosome inherited from mother and one from father

From www.wikimedia.org

DNA profiles

Based on short tandem repeats, STRs

- Method used today in forensic genetics: short tandem repeat (STR)
- Locus (*loci* in plural): Location at a certain chromosome (e.g. D3S1358, DYS391)
- Allele: The number of times a *motif* (short sequence of 3-5 base pairs) repeats itself
- An example of an allele of 3:

$$\underbrace{AGAT}_{motif}\ AGAT\ AGAT = [AGAT]_3$$

- STR's can mutate during meiosis causing variation (e.g. 11 $\rightarrow$ 10)

DNA profiles

Based on short tandem repeats, STRs

- Traditional DNA profile: Based on autosomal (non-sex) chromosomes
- DNA profile consists of 10-30 loci
- Example of autosomal STR DNA profile (only three loci shown):

$$\texttt{D3S1358} = \{15, 18\},\ \texttt{D5S818} = \{12, 12\},\ \texttt{D7S820} = \{10, 11\}$$

- Other types (lineage markers): e.g. Y chromosomal
  - Y-STR haplotypes: DNA profiles from the Y chromosome using STR
  - Example of Y-STR DNA profile (only three loci shown):

$$\texttt{DYS391} = 10,\ \texttt{DYS437} = 15,\ \texttt{DYS635} = 22$$

DNA profiles: autosomal vs Y profiles

Why bother using anything else than traditional autosomal STR DNA profiles?

- Unbalanced mixture of female/male DNA (minor male component masked), e.g. sexual assault cases:
  - touch DNA / male DNA under the fingernails of a victim
  - rape without ejaculation or by a vasectomised male
- Extract (biochemically) Y chromosomal DNA to obtain Y chromosomal DNA profile

DNA profiles: autosomal vs Y profiles

From www.wikimedia.org

Statistical properties (due to genetic inheritance)

- **Autosomal**: 2 alleles per locus inherited independently between and within loci from each parent
  - Widely used and a lot of statistics for that area exist
  - Match probability (grossly simplified) of DNA profile: Product of the allele frequencies at each locus
- **Y chromosomal**: 1 allele per locus inherited as a whole from the father
  - Strong dependency between loci
  - Match probability of DNA profile: Very different than for autosomal DNA profiles

Match probability

- Match probability $\approx$ DNA profile frequency
- Count method (works for any trait, e.g. blood type)
  - n: Database (DB) size
  - n_x: Number of times x is observed in the database
  - $P(X = x) = n_x/n$ such that $LR = n/n_x$
- Problem: Singletons (haplotypes only observed once) are common (a lot of rare variants), $> 90\%$ of observed haplotypes are singletons
  - $\sum_{x \in \text{DB}} n_x/n = 1$, hence $P(X = x) = 0$ for $x \notin$ DB
  - $1/n$ overestimates the match probability for singletons
- Many suggestions (not probability distributions on all haplotypes)

Discrete Laplace distribution

Discrete Laplace distributed $X \sim DL(p, \mu)$:

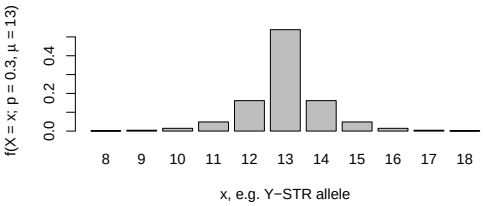
- Dispersion parameter $0 < p < 1$ and
- Location parameter $\mu \in \mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$

Probability mass function:

$$f(X = x; p, \mu) = \frac{1 - p}{1 + p} \cdot p^{|x - \mu|} \quad \text{for } x \in \mathbb{Z}.$$

Perfectly homogeneous population with 1-locus haplotypes:

$$P(X = x) = f(X = x; p, \mu)$$



Discrete Laplace exponential family

- Andersen (2013): Exponential family for known location parameter ($\theta = \log p$ and $d = x - \mu$):

$$f(d; \theta) = \exp\left(\theta |d| - A(\theta)\right) \quad \text{with } A(\theta) = \log\left(\frac{1 + e^{\theta}}{1 - e^{\theta}}\right).$$

- `R` family object for generalized linear model implemented in `R` library `disclap` (also {`d`, `p`, `r`}`disclap`)
- `glm(d ~ 1, dat, family = DiscreteLaplace())`

Statistical model for Y-STR haplotypes

Perfectly homogeneous population with r-locus haplotypes:

$$P(X = (x_1, x_2, \ldots, x_r)) = \prod_{k=1}^{r} f(x_k - \mu_k; p_k)$$

- $\mu = (\mu_1, \mu_2, \ldots, \mu_r)$: central haplotype
- $p = (p_1, p_2, \ldots, p_r)$: discrete Laplace parameters (one for each locus)
- Mutations happen independently across loci (relative to μ)

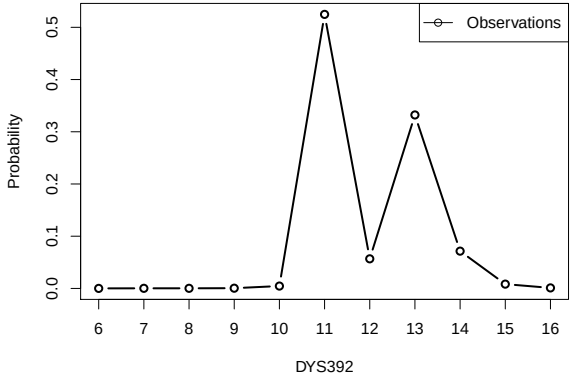
Statistical model for Y-STR haplotypes

Non-homogeneous population with c subpopulations and r-locus haplotypes:

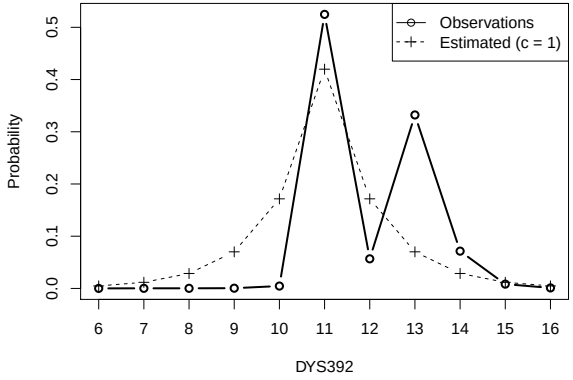
$$P(X = (x_1, x_2, \ldots, x_r)) = \sum_{j=1}^{c} \tau_j \prod_{k=1}^{r} f(x_k - \mu_{jk}; p_{jk})$$

- τ_j: a priori probability for originating from the j'th subpopulation ($\sum_{j=1}^{c} \tau_j = 1$)
- $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jr})$: central haplotype for j'th subpopulation
- $p_j = (p_{j1}, p_{j2}, \ldots, p_{jr})$: parameters for all loci at j'th subpopulation
- Parameter estimation explanation coming up! (Implemented in `R` library `disclapmix`)

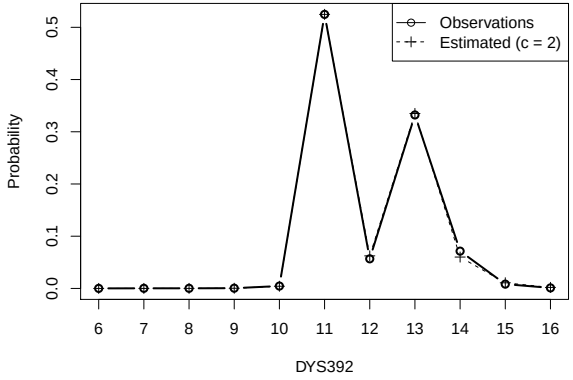
Data and fit


Observations
Probability
0.0
0.1
0.2
0.3
0.4
0.5
6
7
8
9
10
11
12
13
14
15
16
DYS392

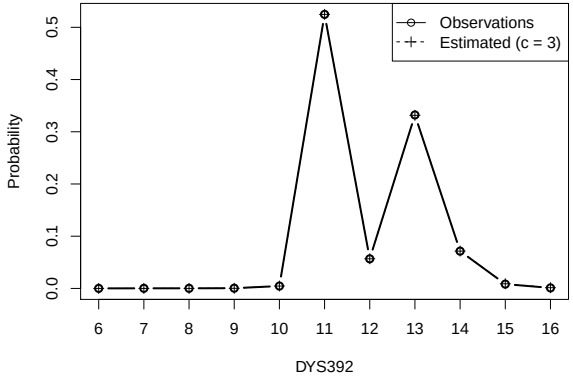
Data and fit


Observations
Estimated (c = 1)
Probability
0.0
0.1
0.2
0.3
0.4
0.5
6
7
8
9
10
11
12
13
14
15
16
DYS392

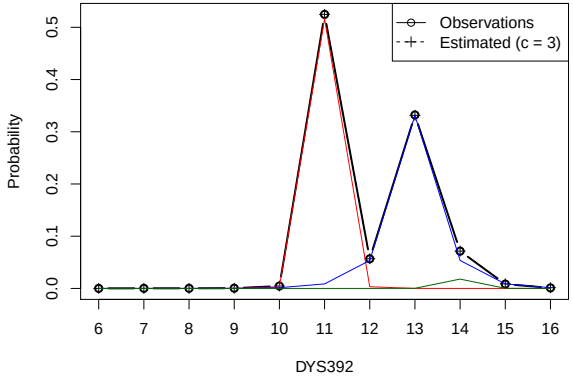
Data and fit


Observations
Estimated (c = 2)
Probability
0.0
0.1
0.2
0.3
0.4
0.5
6
7
8
9
10
11
12
13
14
15
16
DYS392

Data and fit


Observations
Estimated (c = 3)
Probability
0.0
0.1
0.2
0.3
0.4
0.5
6
7
8
9
10
11
12
13
14
15
16
DYS392

Data and fit


Observations
Estimated (c = 3)
Probability
0.0
0.1
0.2
0.3
0.4
0.5
6
7
8
9
10
11
12
13
14
15
16
DYS392

Parameter estimation

- Maximise the full likelihood of the n independent observations $\{x_i\}_{i=1}^n$:

$$L_f = L_f\left(\{p_{jk}\}_{j,k}, \{\mu_j\}_j, \{\tau_j\}_j, \{v_{ij}\}_{i,j}; \{x_i\}_i\right) \tag{1}$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{c} \prod_{k=1}^{r} \left(\tau_j^{1/r} f(|x_{ik} - \mu_{jk}|; p_{jk})\right)^{v_{ij}}, \tag{2}$$

- n individuals, c subpopulations/clusters, r loci
- Wedel and DeSarbo (1995): 'power v_{ij} is equivalent to fixed, known weights in a GLM likelihood'
- Finite mixture model of generalized linear models (e.g. `R` library `FlexMix`)
- GLIMMIX models in the marketing literature

Parameter estimation

- $\{x_i\}_{i=1}^n$: database of n Y-STR haplotypes
- $\hat{v}_{ij} = P(\text{From subpopulation } j \mid \text{Haplotype} = x_i)$
- Initial μ_{jk}'s from e.g. partitioning around medoids (PAM) with L_1 norm

Repeat until convergence:

- $d_{ijk} = |x_{ik} - \mu_{jk}|$
- EM-algorithm to estimate $\{\hat{p}_{jk}\}_{j,k}$, $\{\hat{\tau}_j\}_j$ and $\{\hat{v}_{ij}\}_{i,j}$
  - Repeat until convergence:
    - Estimate $\{p_{jk}\}_{j,k}$ using GLM model $d_{ijk} \sim \omega_j + \lambda_k$ with discrete Laplace family and weights $\hat{v}_{ij}$ ($p_{jk} = \exp(\omega_j + \lambda_k)$):
      `glm(d ~ locus + cluster, dat, family = DiscreteLaplace(), weights = v)`
    - Update $\hat{v}_{ij}$ and $\hat{\tau}_j = \frac{\hat{v}_{+j}}{n}$
- Move subpopulation centers, $\{\hat{\mu}_{jk}\}_{j,k}$, if others are more optimal
  - Update $d_{ijk} = |x_{ik} - \mu_{jk}|$

Estimation

- `glm(d ~ cluster + locus, dat, family = DiscreteLaplace(), weights = ...)`
- Design matrix of dimension $(n \cdot c \cdot r) \times (c + r - 1)$
  - 20,000 DNA profiles (n), 20 loci (r), 150 mixture components (c), $n \cdot c \cdot r = 6 \times 10^7 = 60{,}000{,}000$ and $c + r - 1 = 169$

All individuals (balanced design), no matter DNA profiles (response vector)

```
> model.matrix(~ cluster + locus - 1,
      data = expand.grid(cluster = factor(1:2), locus = factor(1:3)))
  cluster1 cluster2 locus2 locus3
1        1        0      0      0
2        0        1      0      0
3        1        0      1      0
4        0        1      1      0
5        1        0      0      1
6        0        1      0      1
```

Estimation

- $\left(X^\top W^{(m+1)} X\right)^{-1} X^\top W^{(m+1)} \vec{y}^{(m+1)}$ must be calculated
- Calculate $\left(X^\top W^{(m+1)} X\right)^{-1}$ without constructing X...
- It turns out (D_k is diagonal $k \times k$ and H is $c \times (r-1)$) that

$$X^\top W^{(m+1)} X = \begin{bmatrix} D_c & H \\ H^\top & D_{r-1} \end{bmatrix}. \tag{3}$$

- According to Seber (1984), the inverse of this is

$$\left(X^\top W^{(m+1)} X\right)^{-1} = \begin{bmatrix} D_c & H \\ H^\top & D_{r-1} \end{bmatrix}^{-1} = \begin{bmatrix} D_c^{-1} + FE^{-1}F^\top & -FE^{-1} \\ -E^{-1}F^\top & E^{-1} \end{bmatrix}, \tag{4}$$

  where

$$E = D_{r-1} - H^\top D_c^{-1} H \quad \text{and} \quad F = D_c^{-1} H. \tag{5}$$

- Details not (yet?) published (except in my PhD thesis)

Notes

- Deviance is expensive, measure changes in paramter vector first, and then deviance when that's converged

Speed-up

		Method		
		Efficient IRLS (coef.)	Efficient IRLS (dev.)	`glm.fit` (dev.)
$c = 1$	Speed-up	19x	11x	1x
	Total time	0.07 sec	0.12 sec	1.27 sec
$c = 5$	Speed-up	43x	12x	1x
	Total time	3.38 sec	11.78 sec	145.59 sec
$c = 10$	Speed-up	49x	13x	1x
	Total time	2.70 sec	10.12 sec	131.84 sec
$c = 20$	Speed-up	66x	17x	1x
	Total time	11.33 sec	43.13 sec	748.08 sec
$c = 30$	Speed-up	84x	21x	1x
	Total time	24.32 sec	95.62 sec	2,041.64 sec
$c = 40$	Speed-up	101x	26x	1x
	Total time	43.14 sec	169.09 sec	4,348.99 sec
$c = 50$	Speed-up	118x	31x	1x
	Total time	69.14 sec	267.99 sec	8,192.30 sec

- $n = 1{,}690$ DNA profiles (with $r = 23$ Y-STR loci)
- dev.: deviance as convergence criterium
- coef.: relative change in the coefficient vector
- Speed-up: compared to `glm.fit` (dev.)
- Total time: time for the entire EM algorithm (many IRLS's) to converge